Revisiting the Omitted Variables Argument: Substantive vs. Statistical Adequacy*

Aris Spanos†
Department of Economics,
Virginia Tech,
Blacksburg, VA 24061,
e-mail: <aris@vt.edu>

Abstract

The problem of omitted variables is commonly viewed as a statistical misspecification issue which renders the inference concerning the influence of $\mathbf{X}_t$ on y_t unreliable, due to the exclusion of certain relevant factors $\mathbf{W}_t$. That is, omitting certain potentially important factors $\mathbf{W}_t$ may confound the influence of $\mathbf{X}_t$ on y_t. The textbook omitted variables argument attempts to assess the seriousness of this unreliability using the sensitivity of the estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\mathbf{y}$ to the inclusion/exclusion of $\mathbf{W}_t$, by tracing that effect to the potential *bias/inconsistency* of $\widehat{\boldsymbol{\beta}}$. It is argued that the confounding problem is one of *substantive inadequacy* in so far as the potential error concerns subject-matter, not statistical, information. Moreover, the textbook argument in terms of the sensitivity of point estimates provides a poor basis for addressing the confounding problem. The paper reframes the omitted variables question into a hypothesis testing problem, supplemented with a post-data evaluation of inference based on severe testing. It is shown that this testing perspective can deal effectively with assessing the problem of confounding raised by the omitted variables argument. The assessment of the confouding effect using hypothesis testing is related to the conditional independence and faithfulness assumptions of graphical causal modeling.

[*] Forthcoming in the *Journal of Economic Methodology.*

[†] I would like to thank Kevin D. Hoover for numerous valuable comments/suggestions on a previous draft which helped to improve the paper considerably.

1 Introduction

Evaluating the effects of omitting relevant factors on the reliability of inference constitutes a very important problem in fields like econometrics where the overwhelming majority of the available data are observational in nature. To place this problem in a simple context, consider estimating the *Linear Regression model*:

$$\boxed{M_0:}\quad y_t = \mathbf{x}_t^{\top}\boldsymbol{\beta} + u_t, \quad u_t \backsim \mathsf{NIID}(0, \sigma^2),\ t = 1, ..., T, \tag{1}$$

where '$^{\top}$' denotes the transpose of a vector/matrix, $u_t \backsim \mathsf{NIID}(0, \sigma^2)$ reads 'u_t is distributed as a Normal, Independent and Identically Distributed (NIID) process with mean 0 and variance σ^2'; (1) is viewed as explaining the behavior of y_t in terms of the exogenous variables in $\mathbf{X}_t$ - note that $\mathbf{X}_t$ denotes the random vector and $\mathbf{x}_t$ its observed value. The problem is motivated by concerns that the estimated model M_0 could have omitted certain potentially important factors $\mathbf{W}_t$, which may confound the influence of $\mathbf{X}_t$ on y_t. This contemplation gives rise to a potentially 'true' model:

$$\boxed{M_1:}\quad y_t = \mathbf{x}_t^{\top}\boldsymbol{\beta} + \mathbf{w}_t^{\top}\boldsymbol{\gamma} + \varepsilon_t, \quad \varepsilon_t \backsim \mathsf{NIID}(0, \sigma^2),\ t = 1, ..., T, \tag{2}$$

whose comparison with (1) indicates that inferences concerning the influence of $\mathbf{X}_t$ on y_t, based on M_0, might be unreliable because one of the assumptions of the error term, $E(u_t)$=0, is false. That is, in view of (2), $E(u_t)$= $\mathbf{w}_t^{\top}\boldsymbol{\gamma} \neq$0.

These concerns are warranted because the multitude of potential factors influencing the behavior of y_t is so great that no logically stringent case can be made that 'all relevant factors have been included in $\mathbf{X}_t$'. Indeed, economic theories are notorious for invoking *ceteris paribus* clauses which are usually invalid for observational data. As a result of this insuperable obstacle, the problem is commonly confined to assessing the sensitivity (fragility/robustness) of the estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\intercal}\mathbf{X})^{-1}\,\mathbf{X}^{\intercal}\mathbf{y}$ to the inclusion/exclusion of certain specific factors $\mathbf{W}_t$ as suggested by some theory/conjecture. As such the issue is primarily one of *substantive inadequacy* in so far as certain substantive subject-matter information raises the prospect that certain influential factors might have been omitted from explaining the behavior of y_t, leading to misleading inferences.

The textbook omitted variables argument attempts to assess the sensitivity of the estimator $\widehat{\boldsymbol{\beta}}$ to the inclusion/exclusion of certain specific factors $\mathbf{W}_t$, by tracing the effects of these omitted variables to its *bias/inconsistency*; see Johnston (1984), Greene (1997), Stock and Watson (2002), Wooldridge (2003), inter alia. The sensitivity is often evaluated in terms of the difference in the numerical values of the estimators of $\boldsymbol{\beta}$ in the two models (M_0, M_1). Large discrepancies are interpreted as indicating the presence of bias, reflecting the fragility of inference, which in turn indicates that a confounding problem endangers the assessment of the real influence of $\mathbf{X}_t$ on y_t.

The question posed in this paper is the extent to which the textbook omitted variables argument can help shed light on this confounding problem. It is argued that, as it stands, this argument raises a fundamental issue in empirical modeling, but provides a rather poor basis for addressing it. The deficiency of this argument stems from the fact that (a) the framing of the problem in terms of the sensitivity of point estimators is inadequate for the task, and (b) the problem concerning the relevance of omitted variables should be best viewed as one of substantive, not statistical, inadequacy.

Broadly speaking, *statistical adequacy* concerns the validity of the *statistical model* (the probabilistic assumptions constituting the model - $\varepsilon_t \backsim \mathsf{NIID}(0, \sigma^2)$) vis-a-vis the observed data. *Substantive adequacy* concerns the validity of the *structural model* (the inclusion of relevant and the exclusion of irrelevant variables, the functional relationships among these variables, confounding factors, causal claims, external validity, etc.) vis-a-vis the phenomenon of interest that gave rise to the data. The two premises are related in so far as the statistical model provides the operational context in which the structural model can be analyzed, but the nature of errors associated with the two premises is very different. Moreover, a structural model gains statistical 'operational meaning' when embedded into a statistically adequate model. As argued below, this perspective suggests that the sensitivity of point estimators provides a poor basis for assessing either statistical or substantive inadequacies. A more effective assessment is provided by recasting both, the substantive and statistical adequacy problems, in terms of classical (frequentist) testing. However, the sensitivity to statistical inadequacies is gauged in terms of the difference between nominal and actual error probabilities, but the sensitivity to substantive inadequacies could only be assessed in terms of statistical procedures which can evaluate reliably the validity of claims concerning unknown parameters.

2 The Omitted Variables Argument revisited

For reference purposes, let us summarize the textbook omitted variables argument, as a prelude to the critical discussion that follows. The ugliness of the matrix notation is needed in order to point out some crucial flaws in the argument.

2.1 Summary of the textbook argument

Consider a situation where the following Linear Regression model was estimated:

$$\boxed{M_0:} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} \backsim \mathsf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_T), \tag{3}$$

where $\mathbf{y}: T \times 1,\ \mathbf{X}: T \times k,\ \mathbf{I}_T: T \times T$ (identity matrix), but the 'true' model is:

$$\boxed{M_1:} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \backsim \mathsf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_T), \tag{4}$$

where $\mathbf{W}: T \times m$. The question posed is how the 'optimal' properties of the least squares estimators $(\widehat{\boldsymbol{\beta}}, \mathbf{s}^2)$:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{y}, \quad s^2 = \tfrac{1}{T-k} \widehat{\mathbf{u}}^\intercal \widehat{\mathbf{u}}, \tag{5}$$

of $(\boldsymbol{\beta},\sigma^2)$, respectively, are affected by 'omitting' the variables $\mathbf{W}$ from (3).

Substituting the 'true' model (4) into the estimator $\widehat{\boldsymbol{\beta}}$ and taking expectations $(E(.))$ and limits in probability $(\underset{T\to\infty}{\mathbb{P}\lim})$, one can show that, in general, $\widehat{\boldsymbol{\beta}}$ is a *biased and inconsistent estimator* of $\boldsymbol{\beta}$:

$$E(\widehat{\boldsymbol{\beta}})= \boldsymbol{\beta}+ (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\boldsymbol{\gamma},\ \ \underset{T\to\infty}{\mathbb{P}\lim}(\widehat{\boldsymbol{\beta}})= \boldsymbol{\beta}+\lim_{T\to\infty}\left([(\mathbf{X}^\top\mathbf{X})/T]^{-1}[(\mathbf{X}^\top\mathbf{W})/T]\,\boldsymbol{\gamma}\right), \tag{6}$$

assuming that $\lim\limits_{T\to\infty}(\frac{\mathbf{X}^\top\mathbf{W}}{T})=Cov(\mathbf{X}_t,\mathbf{W}_t)\neq\mathbf{0}$. Similarly, given that:

$$\widehat{\mathbf{u}}=\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}=\mathbf{M}_{\mathbf{X}}(\mathbf{W}\boldsymbol{\gamma}+\boldsymbol{\varepsilon}),\ \ \text{where } \mathbf{M}_{\mathbf{X}}=\mathbf{I}-\mathbf{X}\,(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top,$$

s^2 will, in general, be a *biased* and *inconsistent* estimator of σ^2 :

$$E(s^2)=\sigma^2+\tfrac{1}{T-k}\left(\boldsymbol{\gamma}^\top\mathbf{W}^\top\mathbf{M}_{\mathbf{X}}\mathbf{W}\boldsymbol{\gamma}\right),\ \ \underset{T\to\infty}{\mathbb{P}\lim}(s^2)=\sigma^2+\lim_{T\to\infty}[\boldsymbol{\gamma}^\top\mathbf{W}^\top\,(\mathbf{M}_{\mathbf{X}}/T)\,\mathbf{W}\boldsymbol{\gamma}]. \tag{7}$$

It is interesting to note that the bias/inconsistency of $(\widehat{\boldsymbol{\beta}},\mathbf{s}^2)$ vanishes when $\boldsymbol{\gamma}=\mathbf{0}$, but that of $\widehat{\boldsymbol{\beta}}$ also vanishes when $Cov(\mathbf{X}_t,\mathbf{W}_t)=\mathbf{0}$. As shown below, $\boldsymbol{\gamma}=\mathbf{0}$ and $Cov(\mathbf{X}_t,\mathbf{W}_t)=\mathbf{0}$ are *not* equivalent**.**

The impression given by the textbook omitted variables argument is that the difference between the numerical values of the two estimators:

$$\widehat{\boldsymbol{\beta}}=(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y},\qquad \widetilde{\boldsymbol{\beta}}=(\mathbf{X}^\top\mathbf{M}_W\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}_W\mathbf{y},$$

arising from (3) and (4), respectively, provide a way to assess this sensitivity. The rationale behind this view is that, since:

$$\widehat{\boldsymbol{\beta}}=\boldsymbol{\beta}+(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{W}\boldsymbol{\gamma}+\boldsymbol{\varepsilon}),\ \ \widetilde{\boldsymbol{\beta}}=\boldsymbol{\beta}+(\mathbf{X}^\top\mathbf{M}_W\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}_W\boldsymbol{\varepsilon},$$

$$(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})\simeq(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\boldsymbol{\gamma}, \tag{8}$$

the value of $(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})$ reflects the bias/inconsistency. As argued in the sequel, (8) provides very poor grounds for assessing the presence/absence of estimation bias.

2.2 A preliminary critical view

Let us take a preliminary look at the textbook omitted variables argument as a way to assess the sensitivity (fragility/robustness) of the estimator $\widehat{\boldsymbol{\beta}}$ vis-a-vis the presence of potentially relevant factors $\mathbf{W}_t$.

To begin with, the omitted variables argument does not distinguish between statistical and substantive inadequacy. Indeed, by calling it 'specification error', the textbook interpretation gives the clear impression that the issue is one of statistical misspecification, i.e. 'all relevant factors have been included in $\mathbf{X}_t$' forms an integral part of the statistical premises; see Johnston (1984). It what follows, it is argued that the statistical premises include no such assumption. The problem of confounding is an issue of substantive inadequacy because it concerns the probing for errors in bridging the gap between the structural model and the phenomenon of interest.

The confusion between the statistical and structural premises arises primarily because the textbook specification of a statistical model is given in terms of the error term: $\mathbf{u}\backsim\mathsf{N}(\mathbf{0},\sigma^2\mathbf{I}_T)$. This, in turn, imposes a substantive perspective on the statistical analysis which often leads to confusions; see Spanos and McGuirk (2002). Instead, the substantive and statistical perspectives should be viewed as two separate but complemental standpoints. As argued in Spanos (1986), the error assumptions provide a somewhat misleading and incomplete picture of the probabilistic structure indirectly imposed on the observable random variables involved; the only structure that matters for statistical inference purposes. In order to provide a more well-defined statistical perspective, and draw the line between the statistical and substantive premises more clearly, Spanos (1986) proposed a recasting of the probabilistic assumptions of the error term into assumptions concerning the distribution of the observable random variables involved $D(y_t|\ \mathbf{X}_t;\boldsymbol{\theta})$. The recasting revealed that assumptions [**1**]-[**4**] (see table 1) are equivalent to the error assumptions $(u_t|\ \mathbf{X}_t\!=\mathbf{x}_t)\backsim\mathsf{NIID}(0,\sigma^2),\ t\in\mathbb{T}$ (see table 2), where $\mathbb{T}:=\{1,2,...,T,...\}$ ('$:=$' denotes 'is defined as') is an index set with t often, but not exclusively, indicating time. (The index could also refer to the order of cross-section data.)

Table 1 - The Normal/Linear Regression Model

Statistical GM:	$y_t=\beta_0+\boldsymbol{\beta}_1^\top\mathbf{x}_t+u_t,\ t\in\mathbb{T}$,
[**1**] **Normality:**	$(y_t\|\ \mathbf{X}_t\!=\mathbf{x}_t)\backsim\mathsf{N}(.,.)$,
[**2**] **Linearity:**	$E(y_t\|\ \mathbf{X}_t\!=\mathbf{x}_t)=\beta_0+\boldsymbol{\beta}_1^\top\mathbf{x}_t$, linear in $\mathbf{x}_t$,
[**3**] **Homoskedasticity:**	$Var(y_t\|\ \mathbf{X}_t\!=\mathbf{x}_t)=\sigma^2$, free of $\mathbf{x}_t$,
[**4**] **Independence:**	$\{(y_t\|\ \mathbf{X}_t\!=\mathbf{x}_t),\ t\in\mathbb{T}\}$ is an independent process,
[**5**] **t-invariance:**	$\boldsymbol{\theta}:=(\beta_0,\boldsymbol{\beta}_1,\sigma^2)$ do not vary with t.
$\beta_0:=\mu_1-\boldsymbol{\beta}_1^\top\boldsymbol{\mu}_2$,	$\boldsymbol{\beta}_1:=\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}$, $\sigma^2:=\sigma_{11}-\boldsymbol{\sigma}_{21}^\top\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}$.
$\mu_1\!=\!E(y_t),\ \boldsymbol{\mu}_2\!=\!E(\mathbf{X}_t),\ \sigma_{11}\!=\!Var(y_t),\ \boldsymbol{\sigma}_{21}\!=\!Cov(\mathbf{X}_t,y_t),\ \boldsymbol{\Sigma}_{22}\!=\!Cov(\mathbf{X}_t)$.	

Table 2: Equivalence of probabilistic assumptions

Process $\{(u_t\|\ \mathbf{X}_t\!=\mathbf{x}_t),\ t\in\mathbb{T}\}$	vs.	Process $\{(y_t\|\ \mathbf{X}_t\!=\mathbf{x}_t),\ t\in\mathbb{T}\}$
(1) $(u_t\mid\mathbf{X}_t\!=\mathbf{x}_t)\backsim\mathsf{N}(.,.)$	$\Leftrightarrow$	[1] $(y_t\mid X_t=x_t)\backsim\mathsf{N}(.,.)$,
(2) $E(u_t\mid\mathbf{X}_t\!=\mathbf{x}_t)=0$	$\Leftrightarrow$	[2] $E(y_t\mid\mathbf{X}_t\!=\mathbf{x}_t)=\beta_0+\boldsymbol{\beta}_1^\top\mathbf{x}_t$,
(3) $E(u_t^2\mid X_t=x_t)=\sigma^2$	$\Leftrightarrow$	[3] $Var(y_t\mid\mathbf{X}_t\!=\mathbf{x}_t)=\sigma^2$,
(4) Independent over $t\in\mathbb{T}$	$\Leftrightarrow$	[4] Independent over $t\in\mathbb{T}$.

Assumption [**5**], however, turns out to be 'veiled' because it is not entailed by assumptions (1)-(4), even though it is (implicitly) part of the statistical premises. NOTE that the homoskedasticity assumption [**3**] refers to $Var(y_t|\ \mathbf{X}_t\!=\mathbf{x}_t)$ being free of $\mathbf{x}_t$, not t. Hence, statistical inadequacy for the Linear Regression model concerns departures from assumptions [**1**]-[**5**] only; no assumption concerning the inclusion of all 'relevant' variables in $\mathbf{X}_t$ is made by the statistical premises – it's part of the structural premises.

As argued in section 3, when the statistical premises of the models M_0 and M_1 are respecified in terms of the observable random variables involved (see table 1), it becomes clear that the claim '$\widehat{\boldsymbol{\beta}}$ is a bad estimator of the structural parameter $\boldsymbol{\beta}$ (reflecting the effect of $\mathbf{X}_t$ on y_t)', constitutes a highly equivocal assertion. The (statistical) operational meaning of $\boldsymbol{\beta}$ in M_0 is different from that in M_1, say $\boldsymbol{\alpha}$**,** where, in general, $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$. Similarly, the conditional variances of the two models M_0 and M_1 are also different, say σ_0^2 and σ_1^2, respectively; see section 3.2.

To simplify the preliminary discussion let us consider the case where X_t is a single variable and one is interested in the sensitivity of the estimated coefficient β with respect to the inclusion of a potentially relevant set of variables $\mathbf{W}_t$ by comparing the following two models:

$$\begin{aligned} &\boxed{M_0 :} \quad y_t = \delta_0 + x_t\beta + \varepsilon_{1t}, && \varepsilon_{1t} \backsim \mathsf{NIID}(0, \sigma_0^2),\ t \in \mathbb{T}, \\ &\boxed{M_1 :} \quad y_t = \delta_1 + x_t\alpha + \mathbf{w}_t^{\top}\boldsymbol{\gamma} + \varepsilon_{2t}, && \varepsilon_{2t} \backsim \mathsf{NIID}(0, \sigma_1^2),\ t \in \mathbb{T}. \end{aligned} \tag{9}$$

The idea behind the omitted variables argument is that when the numerical values of the two least-squares estimators, say $\widehat{\beta}$ and $\widehat{\alpha}$, corresponding to models (M_0, M_1), are similar, the estimator $\widehat{\beta}$ will be *robust* to the omitted variables $\mathbf{W}_t$, and thus reliable; otherwise the inference is considered fragile and unreliable; see Leamer (1978), Leamer and Leonard (1983). Although this argument appears to be intuitively obvious, on closer examination it turns out to be seriously flawed. Assessing the sensitivity (fragility/robustness) of inference in terms of the observed difference $(\widehat{\beta} - \widehat{\alpha})$ leaves a lot to be desired in addressing the substantive unreliability issue raised by the omitted variables problem. It is demonstrated below that this strategy is ineffective in assessing the reliability of either statistical or substantive inferences. What is more, it is important to emphasize that the reliability of statistical inferences is assessed using error probabilities, not sensitivity analysis, and the reliability of substantive claims can only be properly assessed by using reliable testing procedures, as opposed to point estimation fragility arguments.

2.2.1 The sensitivity of frequentist statistical inference

The sensitivity of (frequentist) inference to departures from the statistical premises is *not* measured in terms of parameter estimates and their numerical differences, or/and by attaching probabilities to different values of the unknown parameters as in Bayesian inference; see Leamer (1978). It is appraised via changes in the method's *trustworthiness*: its capacity to give rise to valid inferences. This is calibrated in terms of the associated *error probabilities*: how often these procedures lead to erroneous inferences. In the case of confidence-interval estimation the calibration is commonly gauged in terms of minimizing the coverage error probability: the probability that the interval does not contain the 'true' value of the unknown parameter. In the case of hypothesis testing the calibration is ascertained in terms of minimizing the type II error probability: the probability of accepting the null hypothesis when false, for a given type I error probability; see Cox and Hinkley (1974).

In the presence of departures from the model assumptions, the difference between the *nominal* (based on the assumed premises) and *actual* (based on the actual premises) error probabilities provides a measure of the sensitivity of inference to such departures; see Spanos (2005a). For example, if the nominal type I error probability of a t-test is 5%, but the actual turns out to be 95%, the sensitivity of inference is very high and the inference based on such a test is said to be fragile; if the nominal and actual error probabilities are, say 5% and 7%, respectively, the test is said to be robust. It should be noted that changes in type II error probabilities (or power) are also an integral part of the assessment of the test's sensitivity; see Spanos (2005a).

2.2.2 The limited scope of the estimated difference $(\widehat{\beta}-\widehat{\alpha})$

For the sake of the preliminary discussion that follows, let us assume that both models (M_0, M_1) are statistically adequate. It is argued below that this is a crucial assumption which, when false, will invalidate the whole analysis.

The *first problem* with the omitted variables argument is that the estimated difference $(\widehat{\beta}-\widehat{\alpha})$, by itself, provides poor grounds for assessing the similarity between the underlying unknown parameters (β, α), and even poorer grounds for assessing the relevance of the omitted variables $\mathbf{W}_t$; i.e. $\boldsymbol{\gamma}=\mathbf{0}$. Let us flesh this out in some detail.

Consider two estimates $\widehat{\beta}$=.796 and $\widehat{\alpha}$=.802 which are close in numerical values; $(\widehat{\beta}-\widehat{\alpha})=-.06$. Viewed in the context of the omitted variables argument these estimates are likely to be interpreted as indicating that the omitted variables bias is rather small and thus immaterial, ensuring the robustness of any inference based on $\widehat{\beta}$. In what follows we will consider three different scenarios, which are based on the same estimates but license diametrically different conclusions.

In *scenario 1* the estimates $\widehat{\beta}$=.796, $\widehat{\alpha}$=.802 are supplemented with their estimated standard errors $SE(\widehat{\beta})$=.012 and $SE(\widehat{\alpha})$=1.2, and the information that the coefficients of $\mathbf{W}_t$ are significantly different from zero $(\boldsymbol{\gamma}\neq\mathbf{0})$ on the basis of a joint 5% significance level F-test. (NOTE that both the standard errors and the F-test require estimating the variances (σ_0^2, σ_1^2)). In this case the estimators $(\widehat{\beta}, \widehat{\alpha})$ give rise to completely dissimilar inferences concerning the effect of X_t on y_t. The t-ratios $\tau(\beta)=\frac{.796}{.012}$=66.3[.000], and $\tau(\alpha)=\frac{.802}{1.2}=.668$[.504] with the p-values in square brackets, can be interpreted as indicating support for the claims (a) $0<\beta\leq .82$, and (b) α=0. Interpreted in the context of the substantive reliability issue, scenario 1 seems to suggest that the relationship between y_t and X_t in M_0 is *spurious* because in the presence of $\mathbf{W}_t$, X_t becomes redundant.

In *scenario 2,* the same estimates ($\widehat{\beta}$=.796, $\widehat{\alpha}$=.802) are supplemented with different estimated standard errors $SE(\widehat{\beta})$=.012, $SE(\widehat{\alpha})$=.010, and the information that the coefficients of $\mathbf{W}_t$ are *not* significantly different from zero (i.e. $\boldsymbol{\gamma}=\mathbf{0}$) on the basis of a joint 5% significance level F-test. Given that $\tau(\beta)=\frac{.796}{.012}$=66.3[.000], and $\tau(\alpha)=\frac{.802}{.01}$=80.2[.000], the evidence might be interpreted as indicating a *non-spurious* relationship which is robust to the inclusion of other potentially relevant variables; the exact opposite to the conclusions in scenario 1.

In *scenario 3* we retain the same estimates ($\widehat{\beta}$=.796, $\widehat{\alpha}$=.802) and standard errors ($SE(\widehat{\beta})$=.012, $SE(\widehat{\alpha})$=.010) as in scenario 2, but $\boldsymbol{\gamma}\neq\mathbf{0}$ on the basis of a joint 5% significance level F-test. This information seems to indicate a substantive relationship between y_t and X_t which captures the systematic information in the behavior of y_t, but at the same time $\mathbf{W}_t$ seems to be relevant. How can one interpret such evidence? As argued below, for a proper interpretation of these evidence one needs to investigate the relationship between X_t and $\mathbf{W}_t$, and in particular the claim $Cov(\mathbf{X}_t, \mathbf{W}_t)=\mathbf{0}$.

The above discussion indicates that on the basis of the same point estimates of the effect of X_t on y_t, $\widehat{\beta}$=.796 and $\widehat{\alpha}$=.802, very different substantive inferences can be drawn depending on other information relating to the estimated statistical models, including the statistics $SE(\widehat{\beta})$ and $SE(\widehat{\alpha})$ and the parameters $\boldsymbol{\gamma}$ and $Cov(\mathbf{X}_t, \mathbf{W}_t)$.

By the same token, two numerically dissimilar point estimates $\widehat{\beta}$=.796 and $\widehat{\alpha}$=.532, where $(\widehat{\beta}-\widehat{\alpha})=.264$, might provide support for similar claims concerning the underlying 'true' β. Viewed in the context of the omitted variables argument, the apparent discrepancy between the two estimates will be interpreted as indicating that the omitted variables bias is substantial. This argument, however, can be very misleading because, if we combine these estimates with $SE(\widehat{\beta})$=.012 and $SE(\widehat{\alpha})$=.144 to define *scenario 4,* the evidence seems to support the claim $0<\beta\leq .82$, as in scenario 1. How these claims are substantiated using severe testing reasoning (see Mayo, 1996) will be discussed in section 4.2. At this stage it suffices to state that such claims cannot be legitimated on the basis of confidence-interval estimation.

In summary, viewing the problem as one of *omitted variables bias* for $\widehat{\beta}$ and revolving around the difference in the numerical values of the two point estimates $\widehat{\beta}$ and $\widehat{\alpha}$, does not do justice to the problem of assessing the substantive reliability issue. As the above scenarios 1-4 indicate, for a more informative assessment of the relationship between y_t and X_t and the role of $\mathbf{W}_t$, one needs additional information concerning the other parameters $\boldsymbol{\gamma}$, σ_1^2 and $Cov(\mathbf{X}_t, \mathbf{W}_t)$. A more informative perspective is provided by *hypothesis testing* where the emphasis is placed on warranted claims concerning these parameters in light of the data; see section 4.

2.2.3 Statistical inadequacy and the assessment of substantive claims

This brings up the *second problem* with the textbook omitted variables argument. Given that any assessment of substantive reliability relies on statistical inference procedures, such as estimation and testing, *statistical adequacy* constitutes a necessary condition for such an assessment to be reliable. Statistical inferences such as (a) $0 < \beta \leq .82$, and (b) α=0 in scenario 1 above, depend crucially on the validity of the probabilistic assumptions [**1**]-[**5**]; see table 1. When any of these assumptions is invalid the statistical procedures cannot be relied upon to assess the substantive reliability issue in question. For instance, if assumption [**5**] in M_0 is false, say δ_0 is not t-invariant, but instead:

$$\delta_0(t) = \mu_0 + \mu_1 t, \tag{10}$$

and the omitted variables $\mathbf{W}_t$ are *trending*, then $\boldsymbol{\gamma}$ is likely to appear statistically significant in M_1 (i.e. $\boldsymbol{\gamma} \neq \mathbf{0}$), *irrespective* of its substantive significance. This is because $\mathbf{W}_t$ will act as a proxy for the missing $(\mu_1 t)$; $\mathbf{W}_t$ need not have the same trending structure to appear statistically significant. This suggests that before one can draw *valid substantive inferences* concerning the sensitivity of β or whether the additional factors in $\mathbf{W}_t$ play a significant role in explaining the behavior of y_t, one should ensure the statistical adequacy of the estimated model M_1 for data $(\mathbf{y}, \mathbf{x}, \mathbf{W})$; see Mayo and Spanos (2004).

The need to distinguish between substantive and statistical reliability assessments raises the problem of delineating the two issues to avoid Duhemian problems for apportioning blame. In order to indicate how such problems arise in theory testing more generally, let us consider the following simple *modus tollens* argument for falsifying a theory, where a theory T is embedded in a statistical model S entailing evidence $\mathbf{e}$:

$$\begin{array}{c} \text{If } T \text{ and } S, \quad \text{then } \mathbf{e} \\ \text{not-}\mathbf{e}, \\ \hline \therefore \text{ not-}T \textit{ or } \text{not-}S \end{array}$$

It's clear from this argument that one cannot distinguish between not-T (the theory is false) *and* not-S (the statistical model is misspecified) when not-$\mathbf{e}$ is observed. However, if one is able to establish the statistical adequacy of S (separate from the adequacy of T), then the above argument can be modified to yield a deductively valid argument for falsifying a theory:

$$\begin{array}{c} \text{If } T \text{ and } S, \quad \text{then } \mathbf{e} \\ S \text{ \& not-}\mathbf{e}, \\ \hline \therefore \text{ not-}T \end{array}$$

Returning to the omitted variables argument, T denotes model M_0, S stands for model M_1 and the presence/absence of $\mathbf{W}_t$ should be viewed as a substantive issue posed in the context of M_1, assuming that the latter is statistically adequate. Because of the necessity of ensuring the statistical adequacy of S, substantive and statistical adequacy assessments are very different in nature and are posed in the context of different models, the structural (M_0) and statistical (M_1), respectively.

In the next section it is argued that the best way to assess substantive adequacy is to view the structural model as embedded into a statistical model, with the two models retaining their own ontology, despite the fact that sometimes they seem to coincide, or take on different roles depending on the nature of the question posed.

3 The Probabilistic Reduction perspective

The aim of this section is to provide a summary of the Probabilistic Reduction framework in the context of which the problems and issues raised above in relation to the textbook omitted variables argument can be delineated and addressed. An important feature of this framework is the sequence of interrelated models (theory, structural, statistical and empirical), which are used to bridge the gap between a theory and data; see Spanos (1986, 1989, 1995a, 2005c). This sequence of models is in the spirit of the hierarchy of models (primary, experimental, data) proposed by Mayo (1996), ch. 5, designed to 'localize' the probing for different types of errors arising at different levels of empirical modeling; see Spanos (2006c). In the discussion that follows we focus primarily on distinguishing between structural and statistical models and their associated errors in an attempt to place the omitted variables argument in a proper perspective and separate the substantive from the statistical adequacy issues.

3.1 Structural vs. Statistical models

In an attempt to distinguish more clearly between the statistical and substantive perspectives we briefly discuss the nature and ontology of structural and statistical models in empirical modeling; see Spanos (2005d) for a more detailed discussion.

It is widely recognized that most economic phenomena are influenced by a very large number of contributing factors, and that explains why economic theories are dominated by *ceteris paribus* clauses. The idea behind postulating a theory is that in explaining the behavior of an economic variable, say y_k, one demarcates the segment of reality to be modeled by selecting the primary influencing factors $\mathbf{x}_k$, cognizant of the fact that there might be numerous other potentially relevant factors $\boldsymbol{\xi}_k$ (observable and unobservable) that jointly determine the behavior of y_k via:

$$y_k = h^*(\mathbf{x}_k, \boldsymbol{\xi}_k), \quad k \in \mathbb{N}, \tag{11}$$

where $h^*(.)$ represents the true behavioral relationship for y_k. The guiding principle in selecting the variables in $\mathbf{x}_k$ is to ensure that they collectively account for the *systematic* behavior of y_k, and the unaccounted factors $\boldsymbol{\xi}_k$ represent non-essential disturbing influences which have only a non-systematic effect on y_k. This line of reasoning gives rise to a *structural (estimable) model* of the form:

$$y_k = h(\mathbf{x}_k; \boldsymbol{\phi}) + \epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k), \quad k \in \mathbb{N}, \tag{12}$$

where $h(.)$ denotes the postulated functional form, $\boldsymbol{\phi}$ stands for the structural parameters of interest, and $\epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k)$ represents the structural error term, viewed as a function of both $\mathbf{x}_k$ and $\boldsymbol{\xi}_k$. By definition the error term process is:

$$\left\{\epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k) = y_k - h(\mathbf{x}_k; \boldsymbol{\phi}),\ k \in \mathbb{N}\right\}, \tag{13}$$

and represents all unmodeled influences, intended to be a *white-noise* (non-systematic) stochastic process $\{\epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k),\ k\in\mathbb{N}\}$ satisfying the properties:

$$\left.\begin{array}{ll}\text{(i)} & E(\epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k))=0,\\ \text{(ii)} & E(\epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k)\cdot\epsilon(\mathbf{x}_\ell,\boldsymbol{\xi}_\ell))=\left\{\begin{array}{ll}\sigma^2, & k=\ell\\ 0, & k\neq\ell\end{array}\right.,\ k,\ell\in\mathbb{N},\\ \text{(iii)} & E(\epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k)\cdot h(\mathbf{x}_k;\boldsymbol{\phi}))=0,\end{array}\right\}\ \forall(\mathbf{x}_k,\boldsymbol{\xi}_k)\in\mathbb{R}_{\mathbf{x}}\times\mathbb{R}_{\boldsymbol{\xi}}. \tag{14}$$

In summary, a *structural model* provides an 'idealized' substantive description of the phenomenon of interest, in the form of a 'nearly isolated' mathematical system; see Spanos (1986). As such, a structural model (a) influences the demarcation of the segment of reality to be modeled, (b) suggests the aspects of the phenomenon to be measured, with a view to (c) 'capture' the systematic (recurring) features of the phenomenon of interest giving rise to data $\mathbf{Z}:=(\mathbf{z}_1,\mathbf{z}_2,...,\mathbf{z}_n)$. The kind of *errors* one should probe for in the context of a structural model concern the bridging of the gap between the phenomenon of interest and the assumed structural model. These concerns include the form of $h(\mathbf{x}_k;\boldsymbol{\phi})$ and the circumstances that render the error term $\epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k)$ potentially systematic, such as the presence of relevant factors, say $\mathbf{w}_k$ in $\boldsymbol{\xi}_k$, that might have a systematic influence on the behavior of y_t.

It is important to emphasize that (12) depicts a 'factual' generating mechanism, which aims to approximate the *actual data generating mechanism.* As the assumptions (i)-(iii) of the structural error stand, are statistically non-testable because their assessment involves the unspecified/unobserved $\boldsymbol{\xi}_k$. To render (i)-(iii) statistically testable one needs to embed this structural model into a statistical one. Not surprisingly, the nature of the embedding itself depends crucially on whether the data $\mathbf{Z}:=(\mathbf{z}_1,\mathbf{z}_2,...,\mathbf{z}_n)$ are the result of an experiment or they are non-experimental (observational) in nature. To bring out the vulnerability to the problem of omitted variables when modeling with observational data, we need to compare the two cases.

3.1.1 Experimental data

In the case where one can perform experiments, 'experimental design' techniques (such as randomization, blocking and replication), might allow one to operationalize the 'near isolation' condition, including the *ceteris paribus* clauses. The objective is to ensure that the *error term* is no longer a function of $(\mathbf{x}_k,\boldsymbol{\xi}_k)$, but for all values $(\mathbf{x}_k,\boldsymbol{\xi}_k)\in\mathbb{R}_{\mathbf{x}}\times\mathbb{R}_{\boldsymbol{\xi}}$:

$$\epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k)=\varepsilon_k\backsim\mathsf{NIID}(0,\sigma^2),\quad k=1,...,n. \tag{15}$$

For instance, *randomization* and *blocking* are often used to 'neutralize' and 'isolated' the phenomenon from the potential effects of $\boldsymbol{\xi}_k$ by ensuring that the uncontrolled factors cancel each other out; see Fisher (1935).

As a direct result of the experimental 'control' via (15) the structural model (12) is essentially transformed into a *statistical model*:

$$y_k=h(\mathbf{x}_k;\boldsymbol{\theta})+\varepsilon_k,\quad \varepsilon_k\backsim\mathsf{NIID}(0,\sigma^2),\ k=1,2,...,n, \tag{16}$$

where the statistical error term ε_k in (16) is qualitatively very different from the structural error term $\epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k)$ in (12) because ε_k is no longer a function of $\boldsymbol{\xi}_k$. For more precise inferences one needs to be more particular about the probabilistic assumptions defining the statistical model, including the functional form $h(.)$. This is because the more finical the probabilistic assumptions (the more constricting the statistical premises), the more precise the inferences.

The most important aspect of embedding the structural (12) into the statistical model (16) is that, in contrast to (i)-(iii) in (14), the probabilistic assumptions concerning the *statistical error term* ε_k are rendered *testable*. By operationalizing the 'near isolation' condition via (15), the error term has been tamed and (16) can now be viewed as a *statistical generating mechanism* that 'schematically' describes how the data $(y_1, y_2, \dots, y_n)$ could have been generated using $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ in conjunction with pseudo-random numbers for $(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$. In principle, when the probabilistic assumptions constituting the statistical model in question are valid, one is able to use (16) to simulate 'faithful replicas' of the data $(y_1, y_2, \dots, y_n)$.

The ontological status of the statistical model (16) is different from that of the structural model (12) in so far as (15) has operationalized the 'near isolation' condition. The statistical model has been 'created' as a result of the experimental design/control. As a consequence of (15) the informational universe of discourse for the statistical model (16) has been confined to the probabilistic information relating to the observables $\mathbf{Z}_k := (y_k, \mathbf{X}_k)$. The 'taming' of the error term using (15) has introduced the *probabilistic perspective*, which considers data $\mathbf{Z} := (\mathbf{z}_1, \dots, \mathbf{z}_n)$ as a 'realization' of a (vector) stochastic process $\{\mathbf{Z}_t,\ t \in \mathbb{T}\}$ whose probabilistic structure is such that would render data $\mathbf{Z}$ a 'truly typical' realization thereof. This probabilistic structure, according to Kolmogorov's theorem, can be fully described, under certain mild regularity conditions, in terms of the joint distribution $D(\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_T; \boldsymbol{\phi})$; see Doob (1953). It turns out that a statistical model can be viewed as a parameterization of the presumed probabilistic structure of $\{\mathbf{Z}_t,\ t \in \mathbb{T}\}$, arising as a reduction from this joint distribution; see Spanos (2006c).

In summary, a *statistical model* constitutes an 'idealized' probabilistic description of a stochastic process $\{\mathbf{Z}_t,\ t \in \mathbb{T}\}$, giving rise to data $\mathbf{Z}$, in the form of an internally consistent set of probabilistic assumptions, chosen to ensure that this data represent a 'truly typical realization' of $\{\mathbf{Z}_t,\ t \in \mathbb{T}\}$.

In contrast to a *structural model*, which relies on substantive subject matter information, a statistical model relies on the statistical information in $D(\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_T; \boldsymbol{\phi})$, that 'reflects' the chance regularity patterns exhibited by the data. Hence, once $\mathbf{Z}_t$ is chosen, a statistical model takes on 'a life of its own' in the sense that it constitutes a self-contained data generating mechanism defined exclusively in terms of probabilistic assumptions pertaining to the observables $\mathbf{Z}_k := (y_k, \mathbf{X}_k)$. An important example of such a statistical model arises when $h(.)$ is assumed to be linear, say $h(\mathbf{x}_k; \boldsymbol{\phi}) = \beta_0 + \boldsymbol{\beta}_1^\top \mathbf{x}_t$. When this is combined with the error assumptions $\mathsf{NIID}(0, \sigma^2)$, the result is the *Gauss linear model* given in table 3; see Spanos (1986), ch. 18. It is important to emphasize that the experimental control over the values of $\mathbf{X}_k$ renders these variables *non-stochastic,* and thus y_k is the only random variable in terms of whose distribution the statistical model is specified. Given the designated values $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $y_t \backsim \mathsf{NI}(\beta_0 + \boldsymbol{\beta}_1^\top \mathbf{x}_t,\ \sigma^2),\ t = 1, 2, ..., T$. It is interesting to note that the above procedure can be used to view the specification of a wide variety of statistical models of interest in economics from the Probit/Logit and *Poisson regressions* to the *Cox proportional hazard regression model*; see Spanos (2006a).

In concluding this sub-section it is important to emphasize that the kind of *errors* one can probe in the context of a statistical model, such as the Gauss Linear model (table 3), are concerned with departures from assumptions [**1**]-[**5**] for data $\mathbf{Z}$.

Table 3 - The Gauss Linear Model

Statistical GM:	$y_t = \beta_0 + \boldsymbol{\beta}_1^\top \mathbf{x}_t + \varepsilon_k,\ t \in \mathbb{T}$,	
[**1**] **Normality:**	$y_t \backsim \mathsf{N}(.,.)$,	
[**2**] **Linearity:**	$E(y_t) = \beta_0 + \boldsymbol{\beta}_1^\top \mathbf{x}_t$, linear in parameters,	(17)
[**3**] **Homoskedasticity:**	$Var(y_t) = \sigma^2$, free of $\mathbf{x}_t$,	
[**4**] **Independence:**	$\{y_t,\ t \in \mathbb{T}\}$ is an independent process,	
[**5**] **t-invariance:**	$\boldsymbol{\theta} := (\beta_0, \boldsymbol{\beta}_1, \sigma^2)$ do not vary with t.	

3.1.2 Observational data

This is the case where the observed data on $(y_k, \mathbf{x}_k)$ are the result of an ongoing actual data generating process, undisturbed by any experimental control or intervention. In this case the route followed in (15) in order to render the statistical error term (a) free of $(\mathbf{x}_k, \boldsymbol{\xi}_k)$, and (b) non-systematic in a statistical sense, is no longer feasible. One needs to find a different way to secure the non-systematic nature of the statistical error term without controls and intervention. It turns out that *conditioning* supplies the primary tool in dealing with modeling observational data because one assumes that behavior is influenced by information available at the time of making decisions.

As shown in Spanos (1986), sequential conditioning provides a general way to transform an arbitrary stochastic process $\{\mathbf{Z}_t,\ t \in \mathbb{T}\}$ into a *martingale difference process.* This provides the key to an alternative approach to specifying statistical models in the case of non-experimental data by replacing the controls and interventions with the choice of the *relevant conditioning information set* $\mathfrak{D}_t$ that would render the error term non-systematic.

As in the case of experimental data the universe of discourse for a statistical model is the joint distribution of the observables $D(\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_T; \boldsymbol{\phi})$, where the vector process $\{\mathbf{Z}_t := \left(y_t, \mathbf{X}_t^\top\right)^\top,\ t \in \mathbb{T}\}$ is defined on the probability space $(S, \mathfrak{F}, \mathbb{P}(.))$, where S denotes an outcomes set, $\mathfrak{F}$, a set of subsets of S defining a sigma-field, and $\mathbb{P}(.)$, a probability set function defined on $\mathfrak{F}$; see Spanos (1999), ch.3. Assuming that $\{\mathbf{Z}_t,\ t\in\mathbb{T}\}$ has a bounded mean, we can choose the conditioning information set to be:

$$\mathfrak{D}_{t-1}=\sigma(\mathbf{X}_t,\mathbf{Z}_{t-1}^0)\subset\mathfrak{F},\quad t\in\mathbb{T}. \tag{18}$$

where $\mathbf{Z}_{t-1}^0:=(\mathbf{Z}_{t-1},...,\mathbf{Z}_1)$. This renders the error process $\{u_t,\ t\in\mathbb{T}\}$, defined by:

$$u_t=y_t-E(y_t|\mathfrak{D}_{t-1}), \tag{19}$$

a *martingale difference process*, irrespective of the probabilistic structure of $\{\mathbf{Z}_t,\ t\in\mathbb{T}\}$, i.e.

$$\text{(I}^*\text{)}\ E(u_t|\mathfrak{D}_{t-1})=0,\quad \text{for all}\quad t\in\mathbb{T}; \tag{20}$$

see Spanos (1999). That is, the error process $\{(u_t|\mathfrak{D}_{t-1}),\ t\in\mathbb{T}\}$ is *non-systematic* (it has mean zero) relative to the conditioning information in $\mathfrak{D}_{t-1}$. The notion of a martingale difference process generalizes the notion of white-noise in the sense that the former requires the existence of the first moment only, but in cases where second moments exist, one can show that:

$$\text{(II}^*\text{)}\quad E(u_t{\cdot}u_s|\mathfrak{D}_{s-1})=\left\{\begin{array}{cc} g(\mathbf{X}_t,\mathbf{Z}_{t-1}^0), & t=s\\ 0, & t<s\end{array}\right.,\ \text{for}\ t,s\in\mathbb{T},$$

i.e. it is uncorrelated $E(u_t{\cdot}u_s|\mathfrak{D}_{s-1}){=}0$, but potentially heteroskedastic $E(u_t^2|\mathfrak{D}_{t-1})=g(\mathbf{X}_t,\mathbf{Z}_{t-1}^0)$; see Spanos (1999), ch. 8. By construction, it is also true that:

$$\text{(III}^*\text{)}\ E(u_t\cdot E(y_t|\mathfrak{D}_{t-1})|\mathfrak{D}_{t-1})=E(y_t|\mathfrak{D}_{t-1})E(u_t|\mathfrak{D}_{t-1})=0,\quad \text{for}\ t\in\mathbb{T}. \tag{21}$$

The error process (19) is based on $D(y_t|\ \mathbf{X}_t,\mathbf{Z}_{t-1}^0;\boldsymbol{\psi}_{1t})$, which constitutes a reduction from $D(\mathbf{Z}_1,\mathbf{Z}_2,...,\mathbf{Z}_T;\boldsymbol{\phi})$ via the sequential conditioning:

$$D(\mathbf{Z}_1,\mathbf{Z}_2,...,\mathbf{Z}_T;\boldsymbol{\phi}){=}D(\mathbf{Z}_1;\boldsymbol{\psi}_1)\prod_{t=2}^T D_t(y_t|\ \mathbf{X}_t,\mathbf{Z}_{t-1}^0;\boldsymbol{\psi}_{1t}){\cdot}D_t(\mathbf{X}_t|\ \mathbf{Z}_{t-1}^0;\boldsymbol{\psi}_{2t}). \tag{22}$$

This gives rise to a generic generating mechanism of the form:

$$y_t=E(y_t\mid\mathfrak{D}_{t-1})+u_t,\quad t\in\mathbb{T}, \tag{23}$$

which is *non-operational* as it stands because without further restrictions on the process $\{\mathbf{Z}_t,\ t\in\mathbb{T}\}$, the systematic component $E(y_t\mid\mathfrak{D}_{t-1})$ cannot be specified explicitly. For operational models one needs to postulate some probabilistic structure for $\{\mathbf{Z}_t,\ t\in\mathbb{T}\}$ that would render the data $\mathbf{Z}$ a 'truly typical' realization thereof.

Example. *The Normal/Linear Regression model* results from the reduction (22) by assuming that $\{\mathbf{Z}_t,\ t\in\mathbb{T}\}$ is a NIID vector process; see Spanos (1986). As a result of the NIID assumptions the relevant conditioning information set is reduced from (18) to:

$$\mathfrak{D}_{t-1}^0:=\{\mathbf{X}_t{=}\ \mathbf{x}_t\}.$$

Hence, the reduction in (22) gives rise to a model specified exclusively in terms of $D(y_t|\ \mathbf{X}_t;\boldsymbol{\psi}_1)$, with an operational statistical generating mechanism based on:

$$y_t = E(y_t \mid \mathbf{X}_t{=}\ \mathbf{x}_t) + u_t, \quad t \in \mathbb{T},$$

where $\mu_t{:=}E(y_t|\ \mathbf{X}_t{=}\ \mathbf{x}_t){=}\beta_0{+}\boldsymbol{\beta}_1^{\top}\mathbf{x}_t$ and $u_t = y_t{-}E(y_t|\ \mathbf{X}_t{=}\ \mathbf{x}_t)$ denote the *systematic and non-systematic components*, satisfying the following properties:

$$\begin{array}{ll}\text{(I)} & E(u_t|\ \mathbf{X}_t{=}\ \mathbf{x}_t){=}0,\\ \text{(II)} & E(u_t{\cdot}u_s|\ \mathbf{X}_t{=}\ \mathbf{x}_t){=}\left\{\begin{array}{ll}\sigma^2 & t=s\\ 0, & t\neq s,\end{array}\right.\ ,\ \forall(t,s)\in\mathbb{T}.\\ \text{(III)} & E(u_t{\cdot}\mu_t|\ \mathbf{X}_t{=}\ \mathbf{x}_t){=}0.\end{array} \tag{24}$$

These properties parallel the non-testable properties of the structural error term (i)-(iii) in (14), but (I)-(III) in (24) are rendered testable by the choice of the relevant conditioning information set. The statistical error term u_t takes the form:

$$(u_t|\ \mathbf{X}_t{=}\ \mathbf{x}_t) \backsim \mathsf{NIID}(0,\sigma^2),\ k=1,2,...,n. \tag{25}$$

This is also analogous to (15) in the case of experimental data, but now the the error term has been operationalized by a judicious choice of the conditioning information set $\mathfrak{D}_{t-1}$. The complete specification of the Linear Regression model is given in table 1 where assumptions [**1**]-[**5**] are purely probabilistic assumptions pertaining to the structure of the observable process $\{(y_t|\ \mathbf{X}_t{=}\ \mathbf{x}_t),\ t{\in}\mathbb{T}\}$. In view of the fact that $u_t = y_t{-}E(y_t|\ \mathbf{X}_t{=}\ \mathbf{x}_t)$, the error assumptions $\mathsf{NIID}(0,\sigma^2)$ are equivalent to assumptions [**1**]-[**4**] (see table 1), with assumption [**5**] being implicitly made.

The probabilistic perspective gives a statistical model 'a life of its own' in the sense that assumptions [**1**]-[**5**] pertain only to the probabilistic structure of $\{(y_t|\ \mathbf{X}_t{=}\ \mathbf{x}_t),\ t\in\mathbb{T}\}$, including the unknown parameters $\boldsymbol{\theta}$, which are specified exclusively in terms of the primary statistical parameters $\boldsymbol{\psi}$ of $D(y_t,\mathbf{X}_t;\boldsymbol{\psi})$. In this sense, it brings into the modeling the *statistical information* which supplements the *substantive subject matter information* carried by the structural model. For example in the context of the latter the functional form $h(\mathbf{x}_k;\boldsymbol{\phi})$ is determined by a theory, but in the context of the statistical model $h(\mathbf{x}_k;\boldsymbol{\theta})$ is determined by the probabilistic structure of the process $\{\mathbf{Z}_t,\ t\in\mathbb{T}\}$. In view of the fact that $h(\mathbf{x}_k;\boldsymbol{\theta}) = E(y_t|\ \mathbf{X}_t{=}\ \mathbf{x}_t)$, and $g(\mathbf{x}_t;\boldsymbol{\theta}) = Var(y_t|\ \mathbf{X}_t{=}\ \mathbf{x}_t)$, where $\boldsymbol{\theta}$ denotes the unknown *statistical parameters*, the functional forms $h(.)$ and $g(.)$ are determined by the joint distribution $D(y_t,\mathbf{X}_t;\boldsymbol{\psi})$. In particular, knowing $D(y_t,\mathbf{X}_t;\boldsymbol{\psi})$ one can proceed to derive $D(y_t|\ \mathbf{X}_t;\boldsymbol{\theta})$, and then derive the regression and skedastic functions:

$$E(y_t|\ \mathbf{X}_t{=}\ \mathbf{x}_t) = h(\mathbf{x}_k;\boldsymbol{\theta})\quad,\quad Var(y_t|\ \mathbf{X}_t{=}\ \mathbf{x}_t) = g(\mathbf{x}_t;\boldsymbol{\theta})\quad\text{for all } \mathbf{x}_t\in\mathbb{R}^k, \tag{26}$$

using exclusively statistical information. For instance, if the scatter plot of the data $\{(x_t,y_t),\ t=1,...,T\}$ exhibits the elliptical shape shown in fig. 1, the linearity and homoskedasticity assumptions [**2**]-[**3**] seem reasonable, but if the data exhibit the triangular shape shown in fig. 2, then both assumptions are likely to be false, irrespective of whether the substantive information supports assumptions [**2**]-[**3**]. This is because the distribution underlying the data in figure 2 is bivariate exponential whose $h(\mathbf{x}_k;\boldsymbol{\theta})$ is non-linear and $g(\mathbf{x}_t;\boldsymbol{\theta})$ is heteroskedastic; see Spanos (1999), ch. 6-7. In this sense the probabilistic perspective on statistical models brings to the table statistical information which is distinct from the substantive information. Indeed, the distinctness of the two types of information is at the heart of the problem of confronting a theory with observed data enabling one to assess its empirical validity.

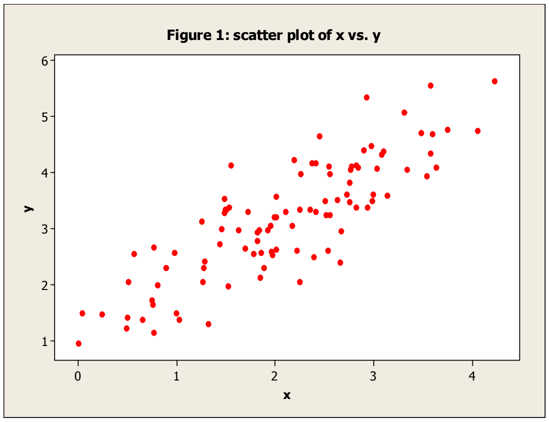
Figure 1: scatter plot of x vs. y

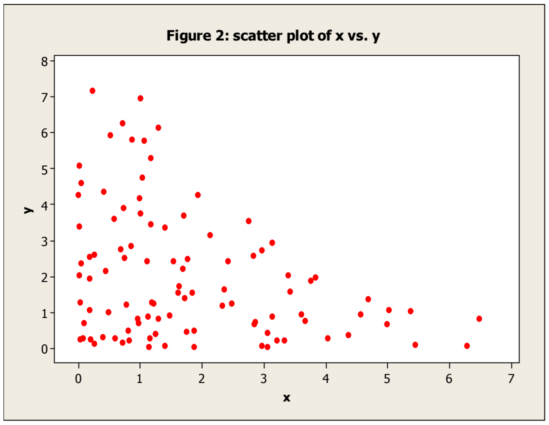
Figure 2: scatter plot of x vs. y

An important aspect of embedding a structural into a statistical model is to ensure that the former can be viewed as a *reparameterization/restriction* of the latter. The idea is that the statistical model summarizes the statistical information in a form which enables one to assess the substantive information in its context. The adequacy of the structural model is then tested against the benchmark provided by a statistically adequate model. Often, there are more statistical parameters $\boldsymbol{\theta}$ than structural parameters $\boldsymbol{\phi}$ and the embedding enables one to test the validity of the additional (over-identifying) restrictions $\boldsymbol{\phi}$ imposes on $\boldsymbol{\theta}$; see Spanos (1990).

Returning to the omitted variables argument, one can view (3) as the structural model with parameters $\boldsymbol{\phi}:=(\boldsymbol{\beta},\sigma_u^2)$, and (4) as the statistical model, with parameters $\boldsymbol{\theta}:=(\boldsymbol{\alpha},\boldsymbol{\gamma},\sigma_\varepsilon^2)$, in the context of which (3) is embedded; see the details in the next section. The *restrictions* on the statistical parameters $\boldsymbol{\theta}$ which define the structural parameters $\boldsymbol{\phi}$ come in the form of: $\boldsymbol{\gamma}=\mathbf{0}$, i.e. $\boldsymbol{\theta}|_{\boldsymbol{\gamma}=\mathbf{0}}=\boldsymbol{\phi}$. These (over-identifying) restrictions are testable and their assessment provides a way to evaluate the substantive adequacy of model (3) when model (4) is statistically adequate.

3.1.3 Experimental vs. Observational data

Despite the fact that the situation under which the experimental data are generated is fundamentally very different from that of observational data, the presence/absence of $\mathbf{w}_t$ from the behavioral relationship between y_t and $\mathbf{x}_t$ is similarly probed in the two cases. The above discussion makes it clear that the question of substantive adequacy cannot be addressed fully unless one is able to embed the structural model into an adequate statistical model. In the case of experimental data the question can be viewed as one of probing for errors concerning the effectiveness of the experimental design/control, and in the case of observational data as probing for omitted factors that can potentially invalidate the empirical modeling.

How does one decide which factors $\mathbf{w}_t$ in $\boldsymbol{\xi}_t$ to probe for? In answering this question one is guided by both substantive as well as statistical information. The relevance of the substantive information is well understood but that of the statistical information is less apparent. It turns out that certain forms of statistical misspecification in M_0, which cannot be addressed using just the statistical information in $D(\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_T; \boldsymbol{\phi})$, provide good grounds to suspect that some key variables influencing the behavioral relationship of y_t might be missing. Hence, when repeated attempts to respecify M_0 to account for all the statistical information in $D(\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_T; \boldsymbol{\phi})$ fail, there is a good chance that certain important factors $\mathbf{w}_t$ in $\boldsymbol{\xi}_t$ have been mistakenly omitted. Moreover, the form and nature of the detected misspecification can often be of great value in guiding substantive subject matter information toward potentially important factors.

3.2 Delineating the omitted variables argument

The textbook omitted variables argument makes perfectly good sense from the substantive perspective in so far as one is comparing two structural explanations for the behavior of y_t and posing a question concerning the presence of potential confounding factors. To answer this question one needs to embed it in the context of a coherent statistical framework. A closer look at the textbook argument reveals that, in addition to the inadequacies raised in section 2.2, the argument is also highly equivocal from the statistical perspective.

The above discussion of the probabilistic perspective makes it clear that the statistical parameters are interwoven with the underlying distribution of the observable random variables involved, and thus one needs to distinguish between the parameters of the two models (3) and (4):

$$\boxed{M_0:} \quad y_t = \mathbf{x}_t^\top \boldsymbol{\beta} + u_t, \qquad\qquad (u_t|\mathbf{X}_t{=}\,\mathbf{x}_t) \backsim \mathsf{NIID}(0, \sigma_u^2), \; t \in \mathbb{T}, \tag{27}$$

$$\boxed{M_1:} \quad y_t = \mathbf{x}_t^\top \boldsymbol{\alpha} + \mathbf{w}_t^\top \boldsymbol{\gamma} + \varepsilon_t, \; (\varepsilon_t|\mathbf{X}_t{=}\,\mathbf{x}_t, \mathbf{W}_t{=}\,\mathbf{w}_t) \backsim \mathsf{NIID}(0, \sigma_\varepsilon^2), \; t \in \mathbb{T}, \tag{28}$$

where, in general, *by assumption*: $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$ and $\sigma_u^2 \neq \sigma_\varepsilon^2$.
The statistical parameterizations associated with the two models, $\boldsymbol{\phi}{:=}(\boldsymbol{\beta},\sigma_u^2)$ and $\boldsymbol{\theta} :=(\boldsymbol{\alpha}, \boldsymbol{\gamma},\sigma_\varepsilon^2)$, can be derived directly from the joint distribution via (26), or indirectly via the error term assumptions; Spanos (1995b). For M_0 the underlying statistical parameterization is:

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}, \quad \sigma_u^2 = \sigma_{11} - \boldsymbol{\sigma}_{21}^\top \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\sigma}_{21}, \tag{29}$$

and for M_1 (see Spanos, 1986, p. 420):

$$\boldsymbol{\alpha}=\boldsymbol{\Sigma}_{2.3}^{-1}(\boldsymbol{\sigma}_{21}-\boldsymbol{\Sigma}_{23}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\sigma}_{31}),\quad \boldsymbol{\gamma}=\boldsymbol{\Sigma}_{3.2}^{-1}(\boldsymbol{\sigma}_{31}-\boldsymbol{\Sigma}_{32}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}), \tag{30}$$

where $\boldsymbol{\Sigma}_{2.3}=\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{23}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\Sigma}_{23}$, $\boldsymbol{\Sigma}_{3.2}=\boldsymbol{\Sigma}_{33}-\boldsymbol{\Sigma}_{32}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{23}$,

$$\sigma_{\varepsilon}^2=\sigma_{11}-\begin{pmatrix}\boldsymbol{\sigma}_{21} & \boldsymbol{\sigma}_{31}\end{pmatrix}\begin{pmatrix}\boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23}\\ \boldsymbol{\Sigma}_{32} & \boldsymbol{\Sigma}_{33}\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{\sigma}_{21}\\ \boldsymbol{\sigma}_{31}\end{pmatrix}, \tag{31}$$

where $\boldsymbol{\sigma}_{31}=Cov(\mathbf{W}_t,y_t)$, $\boldsymbol{\Sigma}_{23}=Cov(\mathbf{X}_t,\mathbf{W}_t)$, $\boldsymbol{\Sigma}_{33}=Cov(\mathbf{W}_t)$.

In view of the statistical parameterizations of the two models, the conclusion of the textbook omitted variables argument can be stated unequivocally as follows:

$(\widehat{\boldsymbol{\beta}},\mathbf{s}^2)$ in (27) are *biased and inconsistent* estimators of $(\boldsymbol{\alpha},\sigma_{\varepsilon}^2)$ in (28).

Given that this statement is almost self-evident, the question arises, ‘how does the equivocation affect the statistical aspects of the original argument?’

By focusing on the error terms, the textbook argument ignores the fact that the two models, (3) and (4), have different underlying distributions; $D(y_t|\mathbf{X}_t;\boldsymbol{\theta})$ and $D(y_t|\mathbf{X}_t,\mathbf{W}_t;\boldsymbol{\phi})$, respectively. It easy to see that when $E(.)$ and $\mathbb{P}\lim_{T\to\infty}(.)$ are defined with respect to $D(y_t|\mathbf{X}_t;\boldsymbol{\theta})$, $(\widehat{\boldsymbol{\beta}},\mathbf{s}^2)$ are unbiased and consistent estimators of $(\boldsymbol{\beta},\sigma_u^2)$ as defined in (29). That is, when the properties of $(\widehat{\boldsymbol{\beta}},\mathbf{s}^2)$ are assessed in the context of the underlying statistical model, they are perfectly ‘good’ estimators of its parameters from the statistical viewpoint, irrespective of whether they are substantively adequate. Hence, statistical and substantive inadequacies raise very different issues.

By the same token, when $E(.)$ and $\mathbb{P}\lim_{T\to\infty}$ are defined with respect to $D(y_t|\mathbf{X}_t,\mathbf{W}_t;\boldsymbol{\phi})$, then the result is biased and inconsistent estimators:

$$\underset{y|\mathbf{X},\mathbf{W}}{E}(\widehat{\boldsymbol{\beta}})=\boldsymbol{\alpha}+\left(\mathbf{X}^{\intercal}\mathbf{X}\right)^{-1}\mathbf{X}^{\intercal}\mathbf{W}\boldsymbol{\gamma},\quad \underset{y|\mathbf{X},\mathbf{W}}{\mathbb{P}}\lim_{T\to\infty}(\widehat{\boldsymbol{\beta}})=\boldsymbol{\alpha}+\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{23}\boldsymbol{\gamma}, \tag{32}$$

$$\underset{y|\mathbf{X},\mathbf{W}}{E}(\mathbf{s}^2)=\sigma_{\varepsilon}^2+\tfrac{1}{T-k}\left(\boldsymbol{\gamma}^{\top}\mathbf{W}^{\top}\mathbf{M}_X\mathbf{W}\boldsymbol{\gamma}\right),\quad \underset{y|\mathbf{X},\mathbf{W}}{\mathbb{P}}\lim_{T\to\infty}(s^2)=\sigma_{\varepsilon}^2+\boldsymbol{\gamma}^{\top}\boldsymbol{\Sigma}_{3.2}\boldsymbol{\gamma}, \tag{33}$$

A direct comparison between (6)-(7) and (32)-(33) indicates most clearly that a crucial equivocation arises from the fact that the textbook omitted variables argument in (6)-(7) does not distinguished between the implicit parameterizations $(\boldsymbol{\beta},\sigma_u^2)$and $(\boldsymbol{\alpha},\sigma_{\varepsilon}^2)$.

The question that naturally arises is whether the unambiguous results in (32)-(33) can be used to shed light on the problem of confounding. From the statistical perspective, the textbook omitted variables argument, restated unequivocally in (32)-(33), has two major weaknesses:

(a) it seems to amount to a comparison between ‘apples’ ($\boldsymbol{\beta}$) and ‘oranges’ ($\boldsymbol{\alpha}$),

(b) it offers no reliable way to evaluate the sensitivity of point estimators.

In the next section it is argued that both of these weaknesses can be addressed by posing the confounding question in the context of a *nesting* Neyman-Pearson framework. The nesting is achieved by ensuring that the two models (3) and (4) are based on the same statistical information, and the hypothesis testing set up enables one to transform the bias/inconsistency terms into claims concerning unknown parameters which can be reliably evaluated.

4 A testing perspective for omitted variables

4.1 Assessing substantive adequacy

In this subsection we discuss the case where the variables $\mathbf{W}_t$ are observed and the relevant observable vector process is $\{\mathbf{Z}_t,\ t{\in}\mathbb{T}\}$ where $\mathbf{Z}_t{:=}(y_t, \mathbf{X}_t^\top, \mathbf{W}_t^\top)^\top$. Consider two models based on the same underlying joint distribution $D(\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_T; \boldsymbol{\varphi})$:

$$\begin{aligned} &\boxed{M_0:} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, && (\mathbf{u}|\mathbf{X}) \backsim \mathsf{N}(\mathbf{0}, \sigma_u^2\mathbf{I}_T), \\ &\boxed{M_1:} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, && (\boldsymbol{\varepsilon}|\mathbf{X}, \mathbf{W}) \backsim \mathsf{N}(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I}_T). \end{aligned} \tag{34}$$

This ensures that models (M_0, M_1) are based on the same statistical information, but M_0 is a special case of M_1 subject to the *substantive restrictions* $\boldsymbol{\gamma} = \mathbf{0}$. In the terminology of the previous section, M_0 can be viewed as a structural model embedded into the statistical model M_1.

From this *nesting perspective* one can deduce, using (30)-(31), that the parameterizations $\boldsymbol{\phi} := (\boldsymbol{\beta}, \sigma_u^2)$ and $\boldsymbol{\theta}{:=} (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma_\varepsilon^2)$ of the two models are interrelated via:

$$\boldsymbol{\alpha} = \boldsymbol{\beta} - \boldsymbol{\Delta}\boldsymbol{\gamma}, \quad \boldsymbol{\gamma} = \boldsymbol{\delta} - \mathbf{D}\boldsymbol{\alpha}, \tag{35}$$

$$\sigma_\varepsilon^2 \ = \sigma_u^2 - \left[(\boldsymbol{\sigma}_{13} - \boldsymbol{\sigma}_{12}\boldsymbol{\Delta})\, \boldsymbol{\Sigma}_{3.2}^{-1}\, (\boldsymbol{\sigma}_{13} - \boldsymbol{\sigma}_{12}\boldsymbol{\Delta})^\top\right], \tag{36}$$

where the parameters $(\boldsymbol{\beta}, \boldsymbol{\Delta}, \mathbf{D}, \sigma_u^2)$ take the form:

$$\begin{aligned} &\boldsymbol{\beta} := \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21},\ \boldsymbol{\delta} := \boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\sigma}_{31},\ \boldsymbol{\Delta} := \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{23},\ \mathbf{D} := \boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\Sigma}_{32}, \\ &\sigma_u^2 = \sigma_{11} - \boldsymbol{\sigma}_{21}^\top\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21},\ \boldsymbol{\Sigma}_{3.2}{:=}\ \boldsymbol{\Sigma}_{33} - \boldsymbol{\Delta}^\top\boldsymbol{\Sigma}_{23}, \boldsymbol{\Sigma}_{2.3}{:=}\ \boldsymbol{\Sigma}_{22} - \mathbf{D}^\top\boldsymbol{\Sigma}_{32}. \end{aligned} \tag{37}$$

These statistical parameterizations can be used to assess the relationship between $\mathbf{X}_t$ and y_t by evaluating the broader issues of *confounding* and *spuriousness*. Although these issues are directly or indirectly related to the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the appraisal of the broader issues depends crucially on the 'true' values of all three covariances:

$$\boldsymbol{\sigma}_{21}{=}Cov(\mathbf{X}_t, y_t),\ \boldsymbol{\sigma}_{31}{=}Cov(\mathbf{W}_t, y_t),\ \boldsymbol{\Sigma}_{23}{=}Cov(\mathbf{X}_t, \mathbf{W}_t), \tag{38}$$

and in particular whether they are zero or not. In what follows it is assumed that:

$$\boldsymbol{\Sigma}_{33}{=}Cov(\mathbf{W}_t) > \mathbf{0},\ \boldsymbol{\Sigma}_{22}{=}Cov(\mathbf{X}_t) > \mathbf{0},$$

in order to ensure that the encompassing model M_1 is *identified*; see Spanos and McGuirk (2002). It turns out that in testing the various combinations of covariances in (38) being zero or not one needs to consider other linear regression models directly related to the two in (34). The most interesting scenarios, created by allowing different combinations of covariances to be zero, are considered below, beginning with scenarios A and B where the parameterization of $\boldsymbol{\alpha}$ reduces to that of $\boldsymbol{\beta}$.

4.1.1 Scenario A: $\boldsymbol{\sigma}_{31} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{23}= \mathbf{0}$, given $\boldsymbol{\sigma}_{21}\neq \mathbf{0}$ (no confounding)

Under these covariance restrictions, the parameters $\boldsymbol{\theta}:= (\boldsymbol{\alpha}, \boldsymbol{\gamma},\sigma_{\varepsilon}^{2})$ take the form:

$$\boldsymbol{\alpha}|_{\substack{\boldsymbol{\Sigma}_{32}=0\\ \boldsymbol{\sigma}_{31}=0}}= \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21} = \boldsymbol{\beta}, \quad \boldsymbol{\gamma}|_{\substack{\boldsymbol{\Sigma}_{32}=0\\ \boldsymbol{\sigma}_{31}=0}}= \mathbf{0}, \quad \boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}|_{\substack{\boldsymbol{\Sigma}_{32}=0\\ \boldsymbol{\sigma}_{31}=0}} = \sigma_{11}-\boldsymbol{\sigma}_{21}^{\top}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21} = \sigma_{u}^{2}. \tag{39}$$

In scenario A model M_1 reduces to model M_0.

The 'given' restrictions $\boldsymbol{\sigma}_{21} \neq \mathbf{0}$ should be tested first using an F-test (see Spanos (1986), p. 399) for the hypotheses:

$$\left(\boxed{M_0}\right) \quad H_0 : \boldsymbol{\beta} = \mathbf{0} \;\text{ vs. }\; H_1 : \boldsymbol{\beta} \neq \mathbf{0}, \tag{40}$$

where $\left(\boxed{M_0}\right)$ indicates that the testing takes place in the context of M_0. Assuming that H_0 was rejected, one can proceed to test $\boldsymbol{\sigma}_{31}= \mathbf{0}$ and $\boldsymbol{\Sigma}_{23}= \mathbf{0}$.

The restriction $\boldsymbol{\sigma}_{31} \neq \mathbf{0}$ can be similarly tested using the hypotheses:

$$\left(\boxed{M_2}\right) \quad H_0 : \boldsymbol{\delta} = \mathbf{0} \;\text{ vs. }\; H_1 : \boldsymbol{\delta} \neq \mathbf{0}. \tag{41}$$

in the context of the linear regression:

$$\boxed{M_2 :} \quad \mathbf{y} = \mathbf{W}\boldsymbol{\delta} + \mathbf{v}, \quad (\mathbf{v}|\mathbf{W}) \backsim \mathsf{N}(\mathbf{0}, \sigma_{v}^{2}\mathbf{I}_{T}), \tag{42}$$

$\sigma_{v}^{2} = \sigma_{11}-\boldsymbol{\sigma}_{31}^{\top}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\sigma}_{31}$. In view of the fact that $\boldsymbol{\Delta} := \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{23}$, the restriction $\boldsymbol{\Sigma}_{23}= \mathbf{0}$ can be tested using the hypotheses:

$$\left(\boxed{M_3}\right) \quad H_0 : \boldsymbol{\Delta} = \mathbf{0}, \text{ vs. } H_1 : \boldsymbol{\Delta} \neq \mathbf{0}, \tag{43}$$

in the context of the multivariate linear regression:

$$\boxed{M_3 :} \quad \mathbf{W} = \mathbf{X}\boldsymbol{\Delta} + \mathbf{U}_2, \tag{44}$$

(see Spanos, 1986, ch. 24). It is important to note that the bias/inconsistency term for $\widehat{\boldsymbol{\beta}}$ in (32) has been transformed into the hypotheses (43), since $\widehat{\boldsymbol{\Delta}}= (\mathbf{X}^{\intercal}\mathbf{X})^{-1} (\mathbf{X}^{\intercal}\mathbf{W})$. The appropriate test for (43) is a multivariate F-type test (see Spanos, 1986, p. 593-5), which is more difficult to apply than the F-tests for (40) and (41).

Searching for a simpler way to test the restrictions $(\boldsymbol{\Sigma}_{23}= \mathbf{0},\ \boldsymbol{\sigma}_{31}= \mathbf{0})$, we note that in view of (30), ($\boldsymbol{\Sigma}_{23}= \mathbf{0}$ and $\boldsymbol{\sigma}_{31}= \mathbf{0}$) imply $\boldsymbol{\gamma} = \mathbf{0}$ (given $\boldsymbol{\sigma}_{21}\neq \mathbf{0}$), and thus they can can be tested using:

$$\left(\boxed{M_1}\right) \quad H_0 : \boldsymbol{\gamma} = \mathbf{0} \;\text{ vs. }\; H_1 : \boldsymbol{\gamma} \neq \mathbf{0}. \tag{45}$$

The F-test for (45) takes the form (see ibid., pp. 399-402):

$$F(\mathbf{y}) = \frac{\left(\widehat{\mathbf{u}}^{\top}\widehat{\mathbf{u}} - \widehat{\boldsymbol{\varepsilon}}^{\top}\widehat{\boldsymbol{\varepsilon}}\right)}{\widehat{\boldsymbol{\varepsilon}}^{\top}\widehat{\boldsymbol{\varepsilon}}} \left(\frac{T-k-m}{m}\right) \backsim \mathsf{F}(d; m, T-k-m), \tag{46}$$

$$C_1 = \{\mathbf{y} : F(\mathbf{y}) > c_{\alpha}\} \quad \text{being the rejection region,}$$

where T is the number of observations; k, the number of included variables; m, the number of omitted variables; $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\alpha}} - \mathbf{W}\widehat{\boldsymbol{\gamma}}$, the residuals from the two models; $d = \frac{(\boldsymbol{\gamma}^{\top}\mathbf{W}^{\intercal}\mathbf{M}_X\mathbf{W}\boldsymbol{\gamma})}{\sigma_{\varepsilon}^2}$, the *non-centrality* parameter. The form of the F-test is particularly interesting in this case because the recasting of the confounding question in a testing framework has transformed the estimation bias of s^2 in (33) into d which determines the F-test's capacity (power) to detect discrepancies from the null when present.

It is interesting to note that testing (45) is directly related to the testing of the *conditional independence* assumption in the context of *graphical causal modeling*, associated with Spirtes et al (2000) and Pearl (2000); see Hoover (2001). That is, the F-test for (45) provides an effective way to test the hypothesis that y_t conditional on $\mathbf{X}_t$ is independent of $\mathbf{W}_t$. At the same time the above discussion brings out the possibility of highly unreliable inferences when one focuses exclusively on the conditional independence restriction $\boldsymbol{\gamma} = \mathbf{0}$, ignoring the covariances $(\boldsymbol{\sigma}_{21}, \boldsymbol{\sigma}_{31}, \boldsymbol{\Sigma}_{23})$. Indeed, a question that naturally arises at this stage is whether testing (45) is equivalent to testing (41) and (43) simultaneously. It turns out that equivalence holds only if one excludes the possibility that $\boldsymbol{\delta} = \mathbf{D}\boldsymbol{\alpha}$; see (35). This possibility arises when the null in (45) is rejected because $\boldsymbol{\Sigma}_{23} \neq \mathbf{0}$ and $\boldsymbol{\sigma}_{31} \neq \mathbf{0}$ hold. This case is excluded by the so-called *faithfulness assumption* invoked in *graphical causal modeling*; see Spirtes et al (2000) and Pearl (2000). As argued below under scenario F, however, there is no need to impose the faithfulness assumption a priori because in that scenario the restrictions $\boldsymbol{\delta} = \mathbf{D}\boldsymbol{\alpha}$ turn out to be testable.

Acceptance. In the case where the null hypotheses in (41) and (43) are accepted, but the null in (40) is rejected, one can infer (under certain circumstances, to be discussed in section 4.2) that the $\mathbf{W}_t$ variables *do not confound* the effect of $\mathbf{X}_t$ on y_t. These circumstances concern primarily the question of *substantive vs. statistical insignificance,* which need to be investigated further using Mayo's (1991) post-data evaluation of inference based on severe testing. Accepting the null hypotheses in (41) and (43) establishes *statistically insignificance,* but the question of *substantive insignificance (no confounding)* needs to be investigated further; it could be the case that the test applied had no power to detect discrepancies of substantive interest. In the case where substantive insignificance can be established (see section 4.2), one can deduce that the estimated model M_0 will give rise to more *precise inferences* than model M_1 because: $Cov(\widehat{\boldsymbol{\alpha}}) - Cov(\widehat{\boldsymbol{\beta}}) \geq \mathbf{0}$.

Rejection. The case where all null hypotheses in (41), (43) and (40) are rejected, gives rise to scenario F to be discussed below. The case where the null hypotheses in (41) and (43) are rejected, but the null in (40) is accepted, gives rise to scenario D.

4.1.2 Scenario B: $\boldsymbol{\Sigma}_{23} = \mathbf{0}$, given $\boldsymbol{\sigma}_{31} \neq \mathbf{0}$, $\boldsymbol{\sigma}_{21} \neq \mathbf{0}$ (no confounding)

Under these covariance restrictions, the parameters $\boldsymbol{\theta} := (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma_{\varepsilon}^2)$ take the form:

$$\boldsymbol{\alpha}|_{\boldsymbol{\Sigma}_{32}=0} = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21} = \boldsymbol{\beta}, \quad \boldsymbol{\gamma}|_{\boldsymbol{\Sigma}_{32}=0} = \boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\sigma}_{31} = \boldsymbol{\delta}, \quad \sigma_{\varepsilon}^2|_{\boldsymbol{\Sigma}_{32}=0} = \sigma_u^2 - \boldsymbol{\sigma}_{13}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\sigma}_{13}^{\top} = \sigma_3^2. \tag{47}$$

In scenario B, M_1 reduces to:

$$\boxed{M_{13}:}\quad \mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{W}\boldsymbol{\delta}+\boldsymbol{\varepsilon}_3,\quad (\boldsymbol{\varepsilon}_3|\mathbf{X},\mathbf{W})\backsim \mathsf{N}(\mathbf{0},\sigma_3^2\mathbf{I}_T), \tag{48}$$

where the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ coincide with the regression coefficients in models M_0 and M_2, respectively; note, however, that $\sigma_3^2\neq\sigma_\varepsilon^2\neq\sigma_u^2\neq\sigma_v^2$.

The 'given' restrictions $\boldsymbol{\sigma}_{21}\neq\mathbf{0}$, $\boldsymbol{\sigma}_{31}\neq\mathbf{0}$, should be established first by testing them in the context of models M_0 and M_3, respectively using (40) and (41). Assuming that both null hypotheses in (40) and (41) have been rejected, one can proceed to test $\boldsymbol{\Sigma}_{23}=\mathbf{0}$ in the context of model M_2 using (43).

Acceptance. When the null hypothesis in (43) is accepted and those in (40) and (41) are rejected, one can infer (under certain circumstances, to be discussed in section 4.2) that $\widehat{\boldsymbol{\alpha}}$ in model M_1 constitutes a 'good' estimator of $\boldsymbol{\beta}=\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}$ in model M_0, i.e. the omitted variables $\mathbf{W}_t$ *do not confound* the effect of $\mathbf{X}_t$ on y_t, even though $\mathbf{W}_t$ is relevant in explaining y_t. More generally, the effect of $\mathbf{X}_t$ on y_t can be *reliably estimated* when one can establish the *non-correlation* between the included ($\mathbf{X}_t$) and omitted ($\mathbf{W}_t$) variables. It is important to emphasize that this result is confined to the *point estimation* and does not extend to the reliability of other forms of inference concerning $\boldsymbol{\beta}$, such as confidence-intervals and hypothesis testing because the latter require a consistent estimator of σ_3^2.

Rejection. The case where all the null hypotheses in (43), (40) and (41) are rejected, gives rise to scenario F below. In the context of the latter scenario one can infer that the omitted variables $\mathbf{W}_t$ are likely to confound the effect of $\mathbf{X}_t$ on y_t, but the extent of the confounding needs to be investigated further using severe testing; see section 4.2. The case where the null hypothesis in (43) is rejected, and those in (40) and (41) are accepted, gives rise to a *scenario G* ($\boldsymbol{\sigma}_{21}=\mathbf{0}$, $\boldsymbol{\sigma}_{\mathbf{31}}=\mathbf{0}$, $\boldsymbol{\Sigma}_{32}\neq\mathbf{0}$). This and *scenario H* ($\boldsymbol{\sigma}_{21}=\mathbf{0}$, $\boldsymbol{\sigma}_{31}=\mathbf{0}$, $\boldsymbol{\Sigma}_{32}=\mathbf{0}$) are uninteresting from our perspective because neither set of variables $\mathbf{X}_t$ or $\mathbf{W}_t$ is relevant in explaining y_t.

4.1.3 Scenario C: $\boldsymbol{\sigma}_{31}=0$, given $\boldsymbol{\sigma}_{21}\neq\mathbf{0}$, $\boldsymbol{\Sigma}_{32}\neq\mathbf{0}$ ('apparent' confounding)

Under these covariance restrictions, the parameters $\boldsymbol{\theta}:=(\boldsymbol{\alpha},\boldsymbol{\gamma},\sigma_\varepsilon^2)$ take the form:

$$\begin{aligned}
&\boldsymbol{\alpha}|_{\boldsymbol{\sigma}_{31}=0}=\boldsymbol{\Sigma}_{2.3}^{-1}\boldsymbol{\sigma}_{21}=\boldsymbol{\alpha}_1,\quad \boldsymbol{\gamma}|_{\boldsymbol{\sigma}_{31}=0}=-\boldsymbol{\Sigma}_{3.2}^{-1}\boldsymbol{\Sigma}_{32}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}=-\mathbf{D}\boldsymbol{\alpha}_1=\boldsymbol{\gamma}_1,\\
&\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}|_{\boldsymbol{\sigma}_{31}=0}=\sigma_{11}-\boldsymbol{\sigma}_{21}^\top\left[\boldsymbol{\Sigma}_{22}^{-1}-\boldsymbol{\Delta}\boldsymbol{\Sigma}_{3.2}^{-1}\boldsymbol{\Delta}^\top\right]\boldsymbol{\sigma}_{21}=\sigma_1^2,
\end{aligned}\tag{49}$$

In scenario C model M_1 reduces to:

$$\boxed{M_{11}:}\quad \mathbf{y}=\mathbf{X}\boldsymbol{\alpha}_1+\mathbf{W}\boldsymbol{\gamma}_1+\boldsymbol{\varepsilon}_1,\quad (\boldsymbol{\varepsilon}_1|\mathbf{X},\mathbf{W})\backsim \mathsf{N}(\mathbf{0},\sigma_1^2\mathbf{I}_T). \tag{50}$$

The 'given' restrictions $\boldsymbol{\sigma}_{21}\neq\mathbf{0}$ and $\boldsymbol{\Sigma}_{32}\neq\mathbf{0}$ can be tested in the context of models M_3 and M_0, respectively, using (40) and (43). Assuming that the null hypotheses in (43) and (40) have been rejected, one can proceed to test $\boldsymbol{\sigma}_{31}=0$ using (41) in the context of model M_3 to establish scenario C.

Similarly to scenario A, searching for a more convenient way to test these restrictions, one can see that in view of (30), ($\boldsymbol{\Sigma}_{23}= \mathbf{0}$ and $\boldsymbol{\sigma}_{21}= \mathbf{0}$) imply $\boldsymbol{\alpha} = \mathbf{0}$ (given $\boldsymbol{\sigma}_{31} \neq \mathbf{0}$ - see scenario E), and thus one can test both coefficient restrictions simultaneously using the F-test for the hypotheses:

$$\left(\boxed{M_1}\right) \quad H_0 : \boldsymbol{\alpha} = \mathbf{0} \ \text{ vs. } \ H_1 : \boldsymbol{\alpha} \neq \mathbf{0}, \tag{51}$$

whose form is analogous to that in (46). This test is equivalent to testing (41) and (43) simultaneously only if one were to exclude the possibility that $\boldsymbol{\beta} = \boldsymbol{\Delta\gamma}$; this case arises from rejecting the null in (35) when $\boldsymbol{\Sigma}_{23}\neq \mathbf{0}$ and $\boldsymbol{\sigma}_{21} \neq \mathbf{0}$ hold. In *graphical causal modeling* this equivalence holds because this case is excluded by the *faithfulness assumption*; see Spirtes et al (2000). When $\boldsymbol{\sigma}_{21} \neq \mathbf{0}$, $\boldsymbol{\sigma}_{31}\neq \mathbf{0}$, $\boldsymbol{\Sigma}_{23}\neq \mathbf{0}$ scenario F holds, and the restrictions $\boldsymbol{\beta} = \boldsymbol{\Delta\gamma}$ turn out to be testable in its context.

It is instructive to consider an alternative (but related) way to assess the validity of the hypotheses in (41) arising from the fact that when the restrictions $\boldsymbol{\gamma}_1 = -\mathbf{D}\boldsymbol{\alpha}_1$ are imposed on M_{11} it reduces to $\mathbf{y} = [\mathbf{X} - \mathbf{WD}]\,\boldsymbol{\alpha}_1 + \boldsymbol{\varepsilon}_1$. Hence, an estimation-based sensitivity assessment of the validity of (41) can be based on comparing the estimates of $\boldsymbol{\alpha}_1$ in (50) with those of $\boldsymbol{\alpha}_1^*$ from the restricted model:

$$\boxed{M_{11}^* :} \quad \mathbf{y} = \widehat{\mathbf{U}}_{\mathbf{1}}\boldsymbol{\alpha}_1^* + \boldsymbol{\varepsilon}_1^*, \quad \widehat{\mathbf{U}}_{\mathbf{1}} = \mathbf{X} - \mathbf{W}\widehat{\mathbf{D}}, \tag{52}$$

where $\widehat{\mathbf{D}} = \left(\mathbf{W}^\intercal\mathbf{W}\right)^{-1}\mathbf{W}^\intercal\mathbf{X}$, and $\widehat{\mathbf{U}}_{\mathbf{1}}$ denotes the matrix of the residuals from the multivariate regression:

$$\boxed{M_4 :} \quad \mathbf{X} = \mathbf{WD} + \mathbf{U}_1. \tag{53}$$

The comparison of the two estimators $\widehat{\boldsymbol{\alpha}}_1$ (based on M_{11}) and $\widehat{\boldsymbol{\alpha}}_1^*$ (based on M_{11}^*) will involve, not only the estimates but also their standard errors, as well as the standard errors of the two regressions, restricted and unrestricted. This estimation-based evaluation differs crucially from the textbook omitted variables argument, based on the difference ($\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\alpha}}$) in (8), in so far as $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\alpha}}_1^*$ are estimating the same parameter $\boldsymbol{\alpha}_1$. In this sense, ($\widehat{\boldsymbol{\alpha}}_1 - \widehat{\boldsymbol{\alpha}}_1^*$) provides a more appropriate basis for assessing the sensitivity of these estimates than ($\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\alpha}}$). Indeed, the difference ($\widehat{\boldsymbol{\alpha}}_1 - \widehat{\boldsymbol{\alpha}}_1^*$) is directly related to the F-test based on the difference between restricted and unrestricted residual sums of squares arising from the two models.

Acceptance. When the null hypothesis in (41) is accepted, but those in (40) and (43) are rejected, one can infer (under certain circumstances - see section 4.2) that the textbook omitted variables argument based on the difference ($\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}$) in (8) will be very misleading because it will (erroneously) indicate the presence of confounding. This is because ($\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}$) is likely to be sizeable, reflecting $(\boldsymbol{\beta} - \boldsymbol{\alpha}_1) = \left(\boldsymbol{\Sigma}_{22}^{-1} - \boldsymbol{\Sigma}_{2.3}^{-1}\right)\boldsymbol{\sigma}_{21} \neq \mathbf{0}$, but in reality $\mathbf{W}_t$ has no role to play in establishing the 'true' relationship between $\mathbf{X}_t$ and y_t since $Cov(\mathbf{W}_t, y_t) = \mathbf{0}$.

Rejection. The case where all null hypotheses in (41), (40) and (43) are rejected gives rise to scenario F (see below), which, in turn calls for further probing in order to establish the extent of the confounding effect using severe testing (see section 4.2). The case where the null hypothesis in (41) is rejected, but those in (40) and (43) are accepted, gives rise to scenario E below.

4.1.4 Scenario D: $\boldsymbol{\sigma}_{21}=\mathbf{0}$ given $\boldsymbol{\Sigma}_{32}\neq\mathbf{0}$, $\boldsymbol{\sigma}_{31}\neq\mathbf{0}$ (spurious)

Under these covariance restrictions, the parameters $\boldsymbol{\theta}{:=}\,(\boldsymbol{\alpha},\boldsymbol{\gamma},\sigma_{\varepsilon}^{2})$ take the form:

$$\begin{aligned}&\boldsymbol{\alpha}|_{\boldsymbol{\sigma}_{21}=\mathbf{0}}=-\boldsymbol{\Sigma}_{2.3}^{-1}\boldsymbol{\Sigma}_{23}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\sigma}_{31}=-\boldsymbol{\Delta}\boldsymbol{\gamma}_{2}=\boldsymbol{\alpha}_{2},\quad \gamma|_{\boldsymbol{\sigma}_{21}=\mathbf{0}}=\boldsymbol{\Sigma}_{3.2}^{-1}\boldsymbol{\sigma}_{31}=\boldsymbol{\gamma}_{2},\\ &\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}|_{\boldsymbol{\sigma}_{21}=0}=\sigma_{11}-\boldsymbol{\sigma}_{13}\boldsymbol{\Sigma}_{3.2}^{-1}\boldsymbol{\sigma}_{13}^{\top}=\sigma_{2}^{2}.\end{aligned}\tag{54}$$

In scenario D model M_1 reduces to:

$$\boxed{M_{12}:}\quad \mathbf{y}=\mathbf{X}\boldsymbol{\alpha}_{2}+\mathbf{W}\boldsymbol{\gamma}_{2}+\boldsymbol{\varepsilon}_{2},\quad (\boldsymbol{\varepsilon}_{2}|\mathbf{X},\mathbf{W})\backsim\mathsf{N}(\mathbf{0},\sigma_{2}^{2}\mathbf{I}_{T}).\tag{55}$$

The 'given' restrictions $\boldsymbol{\Sigma}_{32}\neq\mathbf{0}$, $\boldsymbol{\sigma}_{31}\neq\mathbf{0}$ should be tested first in the context of M_3 and M_2, respectively. Assuming that the null hypotheses in (43) and (41) are rejected, one can proceed to test the restrictions $\boldsymbol{\sigma}_{21}=\mathbf{0}$ in the context of model M_0 using (40).

In direct analogy to scenario C, an alternative (but related) way to assess the validity of the hypotheses in (40) be based on imposing the restrictions $\boldsymbol{\alpha}_{2}=-\boldsymbol{\Delta}\boldsymbol{\gamma}_{2}$, which will reduce M_{12} to $\mathbf{y}=[\mathbf{W}-\mathbf{X}\boldsymbol{\Delta}]\,\boldsymbol{\gamma}_{2}+\boldsymbol{\varepsilon}_{2}$. Hence, an estimation-based sensitivity assessment of the validity of (40) can be performed by comparing the estimates of $\boldsymbol{\gamma}_{2}$ in (55) with those from the restricted model:

$$\boxed{M_{12}^{*}:}\quad \mathbf{y}=\widehat{\mathbf{U}}_{2}\boldsymbol{\gamma}_{2}+\boldsymbol{\varepsilon}_{2}^{*},\quad \widehat{\mathbf{U}}_{2}=\mathbf{W}-\mathbf{X}\widehat{\boldsymbol{\Delta}},\tag{56}$$

where $\widehat{\boldsymbol{\Delta}}=(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\mathbf{W}$ and $\widehat{\mathbf{U}}_{2}$ denotes the matrix of the residuals from the multivariate regression (44).

Acceptance. When the null hypothesis in (40) is accepted but those in (41 and (43) are rejected, one can infer (under certain circumstances, to be discussed in section 4.2) that the relationship between $\mathbf{X}_t$ and y_t in M_0 is *spurious*. Moreover, the textbook omitted variables argument based on the difference $(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\alpha}})$ in (8) can be misleading. This is because the difference $(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\alpha}})$ will reflect $(\boldsymbol{\beta}-\boldsymbol{\alpha}_{2})=-\boldsymbol{\Sigma}_{2.3}^{-1}\boldsymbol{\Sigma}_{23}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\sigma}_{31}\neq\mathbf{0}$, which should not be interpreted as indicating the presence of confounding, but the result of a spurious relationship between $\mathbf{X}_t$ and y_t.

Rejection. The case where all null hypotheses in (40), (41 and (43) are rejected gives rise to scenario F discussed below. The case where the null hypothesis in (40) is rejected but those in (41 and (43) are accepted, gives rise to scenario A above.

4.1.5 Scenario E: $\boldsymbol{\Sigma}_{23}{=}\,\mathbf{0}$, $\boldsymbol{\sigma}_{\mathbf{21}}=\mathbf{0}$, given $\boldsymbol{\sigma}_{\mathbf{31}}\neq\mathbf{0}$ (spurious)

Under these covariance restrictions, the parameters $\boldsymbol{\theta}{:=}\,(\boldsymbol{\alpha},\boldsymbol{\gamma},\sigma_{\varepsilon}^{2})$ take the form:

$$\boldsymbol{\alpha}|_{\substack{\boldsymbol{\Sigma}_{32}=0\\ \boldsymbol{\sigma}_{21}=0}}{=}\,\mathbf{0},\quad \gamma|_{\substack{\boldsymbol{\Sigma}_{32}=0\\ \boldsymbol{\sigma}_{21}=0}}{=}\,\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\sigma}_{31}=\boldsymbol{\delta},\quad \boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}|_{\substack{\boldsymbol{\Sigma}_{32}=0\\ \boldsymbol{\sigma}_{21}=0}}{=}\sigma_{11}{-}\boldsymbol{\sigma}_{13}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\sigma}_{13}^{\top}{=}\sigma_{v}^{2}.\tag{57}$$

In scenario E model M_1 reduces to:

$$\boxed{M_{2}:}\quad \mathbf{y}=\mathbf{W}\boldsymbol{\delta}+\mathbf{v},\quad (\mathbf{v}|\mathbf{W})\backsim\mathsf{N}(\mathbf{0},\sigma_{v}^{2}\mathbf{I}_{T}).\tag{58}$$

The 'given' restrictions $\boldsymbol{\sigma}_{31} \neq \mathbf{0}$ can be tested in the context of model M_2 using (41); see scenario B. Assuming the null hypothesis in (41) has been rejected, one can proceed to test the restrictions $\boldsymbol{\Sigma}_{23}= \mathbf{0}$, $\boldsymbol{\sigma}_{21}= \mathbf{0}$ in the context of models M_3 and M_0, respectively, using (43) and (40). As argued above, (57) suggests that ($\boldsymbol{\Sigma}_{23}= \mathbf{0}$, $\boldsymbol{\sigma}_{21}= \mathbf{0}$) implies $\boldsymbol{\alpha} = \mathbf{0}$, and if one were to exclude the possibility that $\boldsymbol{\beta} = \boldsymbol{\Delta\gamma}$ (see scenario F), testing ($\boldsymbol{\Sigma}_{23}= \mathbf{0}$, $\boldsymbol{\sigma}_{21}= \mathbf{0}$) is equivalent to using the F-test for the hypotheses (51) in the context of M_1; see scenario C. Note that under scenario D, M_1 reduces to M_2 in (42).

Acceptance. When the null hypotheses in (43) and (40) are accepted, and the null in (41) is rejected, one can infer (under certain circumstances, to be discussed in section 4.2) that the relationship between $\mathbf{X}_t$ and y_t in M_0 is *spurious.* Moreover, the textbook omitted variables argument based on the difference ($\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\alpha}}$) in (8) can be misleading. This is because ($\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\alpha}}$) will reflect ($\boldsymbol{\beta} - \boldsymbol{\alpha}_2$)$= \boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\sigma}_{31} \neq \mathbf{0}$, which should not be interpreted as indicating the presence of confounding, but as a result of a spurious relationship between $\mathbf{X}_t$ and y_t.

Rejection. The case where all null hypotheses in (43), (40) and (41) are rejected, gives rise to scenario F, discussed next. The case where the null hypotheses in (43) and (40) are rejected but the null in (41) is accepted, gives rise to scenario C above.

4.1.6 Scenario F: $\boldsymbol{\Sigma}_{23}\neq \mathbf{0}$, $\boldsymbol{\sigma}_{31} \neq \mathbf{0}$, $\boldsymbol{\sigma}_{21} \neq \mathbf{0}$ (possible confounding)

Under these covariance restrictions model M_1 in (34) remains unchanged and the parameters $\boldsymbol{\theta}$:$= (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma_\varepsilon^2)$ take the original form given in (35)-(36). This raises the possibility of confounding but the extent of the effect needs to be quantified. In this scenario one needs to establish that all three covariances ($\boldsymbol{\Sigma}_{23}$, $\boldsymbol{\sigma}_{31}, \boldsymbol{\sigma}_{21}$) are non-zero by testing them separately in their respective statistical models M_3, M_2 and M_0 by rejecting the null hypotheses. Let us assume that $\boldsymbol{\Sigma}_{23}\neq \mathbf{0}$, $\boldsymbol{\sigma}_{31} \neq \mathbf{0}$, $\boldsymbol{\sigma}_{21} \neq \mathbf{0}$ have been established. Does that mean that there is confounding? Not necessarily, because the null the hypothesis in:

$$\left(\boxed{M_1}\right) \quad H_0 : \boldsymbol{\gamma} = \mathbf{0} \text{ vs. } H_1 : \boldsymbol{\gamma} \neq \mathbf{0}, \tag{59}$$

can still be true in the case where (see (35)):

$$\boldsymbol{\delta} = \mathbf{D}\boldsymbol{\alpha}, \tag{60}$$

(assuming that $\boldsymbol{\alpha} \neq \mathbf{0}$) because, in view of (35), (60) implies $\boldsymbol{\gamma} = \mathbf{0}$. As mentioned above, in the graphical causal modeling literature the restrictions (60) are ruled out by the assumption of the *causal faithfulness*; see Spirtes et al (2000) and Pearl (2000). The above discussion, however, indicates most clearly that one need not impose the causal faithfulness assumption a priori because the implicit restrictions are testable in this context of scenario F. That is, given $\boldsymbol{\Sigma}_{23}\neq \mathbf{0}$, $\boldsymbol{\sigma}_{31} \neq \mathbf{0}$, $\boldsymbol{\sigma}_{21} \neq \mathbf{0}$, this assumption becomes testable via (59). When the null hypothesis $\boldsymbol{\gamma} = \mathbf{0}$ is accepted the restrictions excluded by the faithfulness assumption do, in fact, hold. When the null hypothesis

$\boldsymbol{\gamma} = \mathbf{0}$ is rejected, one can proceed to quantify the extent of the confounding effect, using the results of the next subsection.

Similarly, the restrictions $\boldsymbol{\Sigma}_{23}\neq\mathbf{0},\ \boldsymbol{\sigma}_{31}\neq\mathbf{0},\ \boldsymbol{\sigma}_{21}\neq\mathbf{0}$ do not necessarily imply that $\boldsymbol{\alpha}\neq\mathbf{0}$. It follows from (35) that in scenario F $\boldsymbol{\alpha}=\mathbf{0}$ when the restrictions (see (35)):

$$\boldsymbol{\beta}=\boldsymbol{\Delta\gamma}, \tag{61}$$

excluded by faithfulness, hold; assuming that $\boldsymbol{\gamma}\neq\mathbf{0}$.

In concluding this subsection, it is important to re-iterate that the above discussion of the Neyman-Pearson (N-P) testing revolving around scenarios A-F has been heavily qualified. The assessment of confounding and spuriousness needs to go beyond the traditional N-P testing 'accept/reject' decisions, and consider the question of whether such decisions provide evidence for the hypothesis that 'passed', which we consider next.

4.2 Severe testing: assessing substantive significance

The testing results of the previous subsection have been heavily qualified because the Neyman-Pearson testing procedure, based on the coarse 'accept/reject' rules, is susceptible to both *the fallacy of acceptance* (interpreting 'accept H_0' as evidence for a substantive claim associated with H_0), and *the fallacy of rejection* (interpreting 'reject H_0' as evidence for a substantive claim associated with H_1). In order to address these fallacies Mayo (1991, 1996) proposed a post-data assessment of the Neyman-Pearson 'accept/reject' decisions based on evaluating the severity with which a claim passes the particular test with data $\mathbf{z}_0$. In particular, the severe testing reasoning can be used to assess the substantive significance of claims discussed in relation to the confounding problem. This post-data evaluation, based on severe testing, supplements the Neyman-Pearson decision rules with an evidential interpretation that renders the inference considerably more informative by distinguishing clearly the claims that are or are not warranted on the basis of the particular data; see Mayo and Spanos (2006) for further details.

A hypothesis H (H_0 or H_1) passes a *severe test* $\tau(\mathbf{Z})$ with data $\mathbf{z}_0$ if,
(S-1) $\mathbf{z}_0$ agrees with H, and
(S-2) with very high probability, test $\tau(\mathbf{Z})$ would have produced a result that accords less well with H than $\mathbf{z}_0$ does, if H were false.

To illustrate the use of the evaluation of severity we consider a numerical example.

Example. Consider estimating (9) using simulated data with $T=120$:

$$\boxed{M_0:}\qquad y_t=\underset{(5.497)}{10.617}+\underset{(0.039)}{0.777}x_t+\widehat{u}_t,\qquad R_0^2{=}.770,\ s_u{=}5.816, \tag{62}$$

$$\boxed{M_1:}\quad y_t=\underset{(8.896)}{48.562}+\underset{(0.041)}{0.671}x_t-\underset{(0.043)}{0.223}w_t+\widehat{\varepsilon}_t,\qquad R_1^2{=}.813,\ s_\varepsilon{=}5.274. \tag{63}$$

Let us assume that economic theory indicates that the range of values for the three coefficients of interest that would render them *substantively significant* are:

$$h_\gamma:\gamma<-.05,\quad h_\beta:\beta>.1,\quad h_\alpha:\alpha>.1 \tag{64}$$

The estimates of (γ, β, α) based on (62)-(63), are: $\widehat{\gamma} = -.223$, $\widehat{\beta} = .777$, $\widehat{\alpha} = .671$.

What inferences concerning substantive significance and the presence of confounding effects are warranted on the basis of the estimated models (62)-(63)?

To begin with, both estimated models (62) and (63) are statistically adequate (assumptions [1]-[5] in table 1 are valid for the data in question). This provides a clear answer to the argument that omitted variables is a statistical misspecification issue; both models can be statistically but not substantively adequate. Having established statistical adequacy one can proceed to compare the two models on other grounds, including goodness of fit and substantive adequacy. For example, M_1 is clearly better on goodness of fit grounds since $R_1^2 > R_0^2$, but is it better on substantive grounds? To establish that one needs to assess the claims in (64) using Neyman-Pearson testing, supplemented with a post-data evaluation of inference based on the notion of severe testing; see Mayo (1996).

Testing the hypotheses:

$$\left(\boxed{M_1}\right) \quad H_0 : \gamma = 0, \text{ vs. } H_1 : \gamma < 0, \tag{65}$$

using a t-test based on $\tau_\gamma(\mathbf{Z}) = \frac{\widehat{\gamma} - 0}{\widehat{SE(\widehat{\gamma})}}$, with a rejection region $C_1 = \{\mathbf{z} : \tau_\gamma(\mathbf{z}) < c_\alpha\}$, yields $\tau_\gamma(\mathbf{z}_0) = \frac{-0.223}{0.043} = -\ 5.186[.000005]$. The p-value (in square brackets) indicates a rejection of $H_0 : \gamma = 0$. In relation to severity, condition (S-1) the p-value indicates that $\mathbf{z}_0$ agrees with H_1. This suggests that γ is *statistically different from zero* for *some* discrepancy $\gamma_1 < 0$, but does not establish the *size* γ_1 of the warranted discrepancy. When the null is rejected, the goal is to be able to license as large a discrepancy γ_1 from the null as possible. Severity reasoning establishes the warranted discrepancies by evaluating the probability that 'test $\tau_\gamma(\mathbf{Z})$ would have produced a result that accords less well with H_1 than $\mathbf{z}_0$ does, if H_1 were false', i.e.

$$\begin{aligned} SEV(\tau_\gamma(\mathbf{z}_0); \gamma < \gamma_1) \quad &= \mathbb{P}(\tau_\gamma(\mathbf{Z}) > \tau_\gamma(\mathbf{z}_0); \gamma < \gamma_1 \text{ is false}) \\ &= \mathbb{P}(\tau_\gamma(\mathbf{Z}) > \tau_\gamma(\mathbf{z}_0); \gamma \geq \gamma_1). \end{aligned}$$

Choosing different values for γ_1 one can show[1] that:

Table 4: Severity evaluations for $\gamma < \gamma_1$

γ_1	$-.050$	$-.100$	$-.150$	$-.200$	$-.300$
$SEV(\tau_\gamma(\mathbf{z}_0); \gamma < \gamma_1)$	1.000	.997	.954	.703	.038

which suggest that the severity of inferring $\gamma < \gamma_1$, for $-.150 < \gamma_1 < 0$, with test $\tau_\gamma(\mathbf{Z})$ and data $\mathbf{z}_0$, is very high, i.e. the claim is warranted. Moreover, the severity of inferring the substantive claim $h_\gamma : \gamma < -.05$ in (64) is 1, on the basis of which one can infer the substantive significance of γ. That is, the test value $\tau_\gamma(\mathbf{z}_0) =$

[1] The actual evaluations are based on: $\mathbb{P}\left(\tau_\gamma(\mathbf{Z}) > \left[\left(\widehat{\gamma}/\widehat{SE(\widehat{\gamma})}\right) - \left(\gamma_1/\widehat{SE(\widehat{\gamma})}\right)\right]; \gamma = \gamma_1\right)$; the evaluation is at $\gamma = \gamma_1$, because for values $\gamma \geq \gamma_1$ the severity increases.

$-5.186[.000005]$ establishes the statistical significance of γ, and the severity evaluation in table 4 establishes its substantive significance in relation to (64).

Having established the substantive significance of the coefficient of W_t, we can now proceed to consider the question whether W_t confounds the effect of X_t on y_t. The null hypothesis of interest:

$$\left(\boxed{M_0}\right) \quad H_0: \beta=0, \text{ vs. } H_1: \beta>0, \tag{66}$$

is rejected on the basis of the t-test: $\tau_\beta(\mathbf{Z})=\frac{\widehat{\beta}-0}{\widehat{SE(\widehat{\beta})}}$, $C_1=\{\mathbf{z}: \tau_\beta(\mathbf{z})>c_\alpha\}$, since it yields $\tau_\beta(\mathbf{z}_0)=\frac{.777}{.039}=19.923[.000000]$. The severity evaluation for the claim $\beta>\beta_1$, for some $\beta_1>0$, takes the form (see Mayo and Spanos, 2006):

$$\begin{aligned} SEV(\tau_\beta(\mathbf{z}_0); \beta>\beta_1) &= \mathbb{P}(\tau_\beta(\mathbf{Z}) \leq \tau_\beta(\mathbf{z}_0); \beta>\beta_1 \text{ is false}) \\ &= \mathbb{P}(\tau_\beta(\mathbf{Z}) \leq \tau_\beta(\mathbf{z}_0); \beta \leq \beta_1), \end{aligned} \tag{67}$$

giving rise to the following error probabilistic assessments[2]:

Table 5: Severity evaluations for $\beta>\beta_1$

β_1	.600	.650	.700	.750
$SEV(\tau_\beta(\mathbf{z}_0); \beta>\beta_1)$	1.00	.999	.975	.755

In view of $h_\beta: \beta>.1$ in (64) one can argue that the claim $\beta>\beta_1$ with $\beta_1 \leq .7$ passes severely with data $\mathbf{z}$ and the above t-test, and that establishes the substantive significance of β.

Applying the same t-test to the hypotheses:

$$\left(\boxed{M_1}\right) \quad H_0: \alpha=0, \text{ vs. } H_1: \alpha>0, \tag{68}$$

yields $\tau_\alpha(\mathbf{z}_0)=\frac{.671}{.041}=16.366[.000000]$, which also rejects H_0. The same severity evaluations as in (67) for the claim $\alpha>\alpha_1$ give rise to:

Table 6: Severity evaluations for $\alpha>\alpha_1$

α_1	.500	.550	.590	.600	.650
$SEV(\tau_\alpha(\mathbf{z}_0); \alpha>\alpha_1)$	1.00	.998	.975	.957	.695

On the basis of the evaluations in table 5, the claim $\alpha>.6$, is warranted with test $\tau_\gamma(\mathbf{Z})$ and data $\mathbf{z}_0$. Moreover, the severity for inferring the substantive claim $h_\alpha: \alpha>.1$ in (64) is 1, on the basis of which one can infer the substantive significance of α.

Using the evaluations of severity in tables 5-6, one can make a case that there is a deviation between the warranted discrepancies associated with rejecting the null

[2] The actual evaluations are based on:
$\mathbb{P}\left(\tau_\beta(\mathbf{Z}) \leq \left[\left(\widehat{\beta}/\widehat{SE(\widehat{\beta})}\right)-\left(\beta_1/\widehat{SE(\widehat{\beta})}\right)\right]; \beta=\beta_1\right)$.

in the case of the hypotheses (66) and (68). By choosing a high enough severity threshold, say .975, one can argue that the warranted discrepancy from β=0, associated with the claim $\beta > \beta_1$ is β_1=.700, but that from α=0, associated with the claim $\alpha > \alpha_1$ is α_1=.590. An informal comparison between the two claims (ad hoc t-test) indicates that the difference $(\beta_1 - \alpha_1) = .110$ is likely to be statistically different from zero, and the severity evaluations in tables 5-6 suggest that it's also likely to be substantively significant. A more formal assessment of these claims can be based on $(\beta - \alpha) = (\beta_1 - \alpha_1) \neq 0$, but the resulting test is rather involved because it requires reparameterizing M_1.

The above severity evaluations provide support for *scenario F:* $\boldsymbol{\sigma}_{21} \neq \mathbf{0}$, $\boldsymbol{\sigma}_{31} \neq \mathbf{0}$, $\boldsymbol{\Sigma}_{32} \neq \mathbf{0}$. The estimated models M_0 and M_1 in (62)-(63) provide evidence that when W_t is omitted (a) the effect of X_t on y_t is over-estimated, and (b) the structural model M_0 is *substantively inadequate*.

4.2.1 The severe testing reasoning vs. other tail area evaluations

In concluding this subsection it is important to emphasize that severe testing, puts forward a post-data evaluation of a specific claim in accordance with the null or the alternative hypotheses decided on the basis of a specific Neyman-Pearson test result. In this sense the underlying reasoning is different from that implicit in other error probabilities such as type I and II error probabilities, power, confidence level and p-values. In view of the fact that all these error probabilities are based on the evaluation of tail areas of the same sampling distributions (t-tests in this case), there is always some algebraic relationship that enables one to relate, in some ad hoc way, all these error probabilities. However, it is a mistake to confuse these ad hoc relationships with the different rationales underlying the various error probabilities.

What is important is the underlying reasoning that enables one to assess claims concerning the parameters of interest on the basis of the data. It is well-known that the type I and II (or power) error probabilities are pre-data error probabilities based on a predesignated rejection value c_α, determined by the chosen significance level α. As shown above, the severity evaluation is based on the observed value of the test statistic $\tau(\mathbf{z}_0)$ and not c_α. Moreover, it is well-known that no error probability can be attached to observed confidence-intervals, say $(.699 \leq \beta < .855)$; see Cox and Hinkley (1974). Hence, the probabilistic reasoning associated with the severity evaluation is very different from that of confidence-interval estimation, which revolves around defining a random interval that would cover the 'true' value of the parameter θ (whatever that happens to be) with a predesignated probability $(1-\alpha)$. This does not preclude the existence of some relationship between $(1-\alpha)$ and the severity evaluation since, as mentioned above, they are both based on the same sampling distribution. Similarly, the severe testing reasoning is different from the p-value rationale, even though one can always find an algebraic relationship between the two tail areas; see Mayo (1996), Mayo and Spanos (2006) for further details.

4.3 Unobserved omitted variables

Consider the case where no data on the potentially crucial factors $\mathbf{W}_t$ are available. In such a case one might still be able to assess whether $\widehat{\boldsymbol{\beta}}$ might provide an *underestimation/overestimation* of $\boldsymbol{\alpha}$ using certain qualitative information. To that end, the relationship between the parameterizations of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ (see (35)):

$$\boldsymbol{\beta} = \boldsymbol{\alpha} + \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{23}\boldsymbol{\gamma}, \tag{69}$$

provides the most promising way to shed light on the issue. Assuming that one has some qualitative information concerning the sign and approximate magnitudes of $(\boldsymbol{\Sigma}_{23}, \boldsymbol{\gamma})$, one can put forward a qualitative assessment argument concerning the potential inaccuracies of $\widehat{\boldsymbol{\beta}}$ as an estimator of $\boldsymbol{\alpha}$. For example, in the case where $\mathbf{W}_t$ comprises a single factor W_t, and there is qualitative information that $\boldsymbol{\Sigma}_{23}{>}0,\ \gamma{>}0$, then one can deduce that $\widehat{\boldsymbol{\beta}}$ is likely to underestimate $\boldsymbol{\alpha}$. Similarly, when:
$\boldsymbol{\Sigma}_{23}{>}0,\ \gamma{<}0,\ \ \widehat{\boldsymbol{\beta}}$ is likely to overestimate $\boldsymbol{\alpha}$; see Stock and Watson (2002), p. 146.

A closer look at this qualitative assessment argument, however, reveals that it needs to be qualified. As argued above, appraising the sign and magnitude of a coefficient could not rely exclusively on the value of an estimate, but should also take account for the relevant standard error derived from $Cov(\widehat{\boldsymbol{\alpha}}) = \sigma_\varepsilon^2 \left(\mathbf{X}^\intercal \mathbf{P}_W \mathbf{X}\right)^{-1}$, where $\mathbf{P}_\mathbf{W} = \mathbf{W}\left(\mathbf{W}^\intercal\mathbf{W}\right)^{-1}\mathbf{W}^\intercal$, which includes the conditional variance as given in (36), i.e. the coefficient can be insignificantly different from zero. A glance at these formulae suggests that it will be a non-trivial matter to assess that. Hence, even for the simplest cases where there is only one missing factor, the assessment based on how $\widehat{\boldsymbol{\beta}}$ is likely to underestimate/overestimate $\boldsymbol{\alpha}$ can easily lead one astray.

A clearer qualitative assessment is possible when there is qualitative information that strongly indicates $Cov(\mathbf{X}_t, W_t) = \mathbf{0}$. In that case, the estimator $\widehat{\boldsymbol{\beta}}$ based on (27) will provide a reliable point estimator of the effect of $\mathbf{X}_t$ on y_t, even when $\mathbf{W}_t$ is substantively significant in explaining y_t.

4.4 Omitted variables and misspecification testing

An interesting variation on the testing framework created by the two nested models (27) and (28) in sections 4.1-4.2 arises when $\mathbf{W}_t$ does not represent additional variables, but functions of the original variables $(y_t, \mathbf{X}_t)$ such as higher powers of the x_{it}'s, say their second order functions $\boldsymbol{\psi}_t{:=}(x_{it}{\cdot}x_{jt})_{i,j}$, $i, j = 2, ..., k$, lags $\mathbf{z}_{t-i} := (y_{t-i}, \mathbf{x}_{t-i})$, $i = 1, 2, ..., \ell$, and trends $\mathbf{t} := (t, t^2, .., t^m)$. The discussion above can be adapted to provide a general framework for constructing misspecification tests based on artificial regressions; see Spanos (2005b) for more details.

For instance, one can construct a misspecification test for departures from the linearity assumption [2] (see table 1) by using $E(y_t|\ \mathbf{X}_t{=}\ \mathbf{x}_t){=}\alpha_0{+}\boldsymbol{\alpha}_1^\top\mathbf{x}_t{+}\boldsymbol{\alpha}_2^\top\boldsymbol{\psi}_t$, as an approximation to a non-linear regression function. The two regression functions give rise to two competing models:

$$\boxed{M_0:}\ y_t = \beta_0 + \boldsymbol{\beta}_1^\top\mathbf{x}_t + u_t, \quad \boxed{M_1:}\ y_t = \alpha_0 + \boldsymbol{\alpha}_1^\top\mathbf{x}_t + \boldsymbol{\alpha}_2^\top\boldsymbol{\psi}_t + \varepsilon_t, \tag{70}$$

which are comparable because they share the same statistical information based on $D(\mathbf{Z}_1, ..., \mathbf{Z}_T; \boldsymbol{\varphi})$, where $\mathbf{Z}_t := (y_t, \mathbf{X}_t)$. A comparison between them (subtracting M_0 from M_1) gives rise to the artificial regression:

$$u_t = (\alpha_0 - \beta_0) + (\boldsymbol{\alpha}_1^\top - \boldsymbol{\beta}_1^\top)\mathbf{x}_t + \boldsymbol{\alpha}_2^\top \boldsymbol{\psi}_t + \varepsilon_t, \tag{71}$$

or equivalently, its operational form:

$$\widehat{u}_t = (\alpha_0 - \widehat{\beta}_0) + (\boldsymbol{\alpha}_1^\top - \widehat{\boldsymbol{\beta}}_1^\top)\mathbf{x}_t + \boldsymbol{\alpha}_2^\top \boldsymbol{\psi}_t + \varepsilon_t, \tag{72}$$

based on the estimated M_0. (72) provides the basis for a misspecification test for the linearity assumption [2] using the hypotheses:

$$\left(\boxed{M_1}\right) \; H_0 : \boldsymbol{\alpha}_2 = 0 \quad \text{vs.} \quad H_1 : \boldsymbol{\alpha}_2 \neq 0. \tag{73}$$

The set up in (70) can be easily extended to derive misspecification tests for the other assumptions using artificial regressions, as well as joint misspecification tests for assessing the validity of several assumptions simultaneously; see Spanos (2005b) for further details.

In concluding this subsection it is interesting to mention that an early attempt to use the omitted variables argument as a basis for constructing misspecification tests was Pagan (1984), whose use of the equivocal notation in (3)-(4) gave rise to the inappropriate artificial regression $\mathbf{u} = \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$. The difference in the two parameterizations in M_0 and M_1 gives rise to the additional terms $(\alpha_0 - \beta_0)$ and $(\boldsymbol{\alpha}_1^\top - \boldsymbol{\beta}_1^\top)\mathbf{x}_t$ which contribute significantly to the power of the misspecification test based on (73); see Spanos (2005b).

4.5 Omitted variables as Instruments

Another interesting variation on the two nested models (27) and (28) in section 4.1 arises when (27) is a structural model of the form:

$$\boxed{M_0^* :} \quad y_t = \mathbf{X}_t^\top \boldsymbol{\delta} + v_t, \; (v_t | \mathbf{W}_t = \mathbf{w}_t) \backsim \mathsf{NIID}(0, \sigma_v^2), \; t \in \mathbb{T}, \tag{74}$$

where (i) $E(\mathbf{X}_t v_t) \neq \mathbf{0}$, and $\mathbf{W}_t$ represents a $m \times 1$ $(m \geq k)$ vector of *instrumental variables* such that:

$$\begin{array}{ll} \text{(ii)} \;\; E(\mathbf{W}_t v_t) = 0, & \text{(iii)} \; Cov(\mathbf{X}_t, \mathbf{W}_t) = \boldsymbol{\Sigma}_{23} \neq \mathbf{0}, \\ \text{(iv)} \; Cov(\mathbf{W}_t) = \boldsymbol{\Sigma}_{33} > 0, & \text{(v)} \; Cov(y_t, \mathbf{W}_t) = \boldsymbol{\sigma}_{13} \neq \mathbf{0}. \end{array}$$

This model is widely used in econometric modeling because it encompasses the case of a single equation from a Simultaneous Equations Model. The moment restrictions (ii)-(v) define $\mathbf{W}_t$ as correlated with y_t, but y_t and $\mathbf{W}_t$ are *conditionally uncorrelated* given $\mathbf{X}_t$; see Spanos (2006b). Hence, $\mathbf{W}_t$ can be viewed as a set of (legitimately) omitted variables which can be used as instruments for estimating the structural parameters $(\boldsymbol{\delta}, \sigma_v^2)$. As shown in Spanos (1986), the statistical model in the context of which the structural model (74) is embedded is not (28), but the *implicit reduced form:*

$$\boxed{M_1^*:}\ y_t = \mathbf{w}_t^\top \boldsymbol{\beta}_1 + u_{1t}, \qquad \mathbf{X}_t = \mathbf{w}_t^\top \mathbf{B}_2 + \mathbf{u}_{2t},$$

$$\begin{pmatrix} (u_{1t}|\mathbf{W}_t\!= \mathbf{w}_t) \\ (\mathbf{u}_{2t}|\mathbf{W}_t\!= \mathbf{w}_t) \end{pmatrix} \sim \mathsf{NIID}\left(\begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \omega_{11} & \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix} \right). \tag{75}$$

This statistical model, based on $D(y_t, \mathbf{X}_t|\ \mathbf{w}_t; \boldsymbol{\psi})$ is related to (28), based on $D(y_t|\ \mathbf{X}_t, \mathbf{w}_t; \boldsymbol{\varphi}_1)$, via:

$$D(y_t, \mathbf{X}_t|\ \mathbf{w}_t; \boldsymbol{\psi}) = D(y_t|\ \mathbf{X}_t, \mathbf{w}_t; \boldsymbol{\varphi}_1) \cdot D(\mathbf{X}_t|\ \mathbf{w}_t; \boldsymbol{\varphi}_2), \tag{76}$$

giving rise to the *reparameterized* statistical model:

$$\boxed{M_2^*:}\ y_t = \mathbf{X}_t^\top \boldsymbol{\alpha} + \mathbf{w}_t^\top \boldsymbol{\gamma} + \varepsilon_t, \qquad \mathbf{X}_t = \mathbf{w}_t^\top \mathbf{B}_2 + \mathbf{u}_{2t},$$

$$\begin{pmatrix} (\varepsilon_t|\mathbf{X_t}, \mathbf{W}_t\!= \mathbf{w}_t) \\ (\mathbf{u}_{2t}|\mathbf{W}_t\!= \mathbf{w}_t) \end{pmatrix} \sim \mathsf{NIID}\left(\begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_{22} \end{pmatrix} \right), \ t \in \mathbb{T}. \tag{77}$$

In contrast to scenarios A ($\boldsymbol{\gamma} = \mathbf{0}$) and B ($\boldsymbol{\Sigma}_{23}\!= \mathbf{0}$) of the previous section, the restrictions imposed on (77) are:

$$\left(\boxed{M_2^*}\right)\ \boldsymbol{\gamma} = \mathbf{0}, \ \text{ subject to } \ \ \boldsymbol{\beta}_1 \neq \mathbf{0}, \ \mathbf{B}_2 \neq \mathbf{0}. \tag{78}$$

That is, the implicit *structural parameterization* ($\boldsymbol{\delta}$,σ_v^2) in (74) takes the form:

$$\begin{array}{ll} \boldsymbol{\alpha}|_{\ [\boldsymbol{\gamma} = \mathbf{0}, \ \ \boldsymbol{\beta}_1 \neq \mathbf{0} \ \ \mathbf{B}_2 \neq \mathbf{0}]} & = \boldsymbol{\delta} = (\boldsymbol{\Sigma}_{23}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\Sigma}_{23})^{-1}\boldsymbol{\Sigma}_{23}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\sigma}_{31}, \\ \sigma_\varepsilon^2|_{\ [\boldsymbol{\gamma} = \mathbf{0}, \ \ \boldsymbol{\beta}_1 \neq \mathbf{0} \ \ \mathbf{B}_2 \neq \mathbf{0}]} & = \sigma_v^2 = \sigma_{11} - \boldsymbol{\sigma}_{12}(\boldsymbol{\Sigma}_{23}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\Sigma}_{23})^{-1}\boldsymbol{\Sigma}_{23}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\sigma}_{31}. \end{array} \tag{79}$$

This shows most clearly that $\boldsymbol{\delta}$ in (74) and the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ in (34) can represent very different effects of $\mathbf{X}_t$ on y_t; Spanos (1986, 2006).

Noting that $\boldsymbol{\gamma} = \boldsymbol{\beta}_1 - \mathbf{B}_2\boldsymbol{\alpha}$, one can see that the restrictions (78) amount to $(m-k)$ restrictions on $(\boldsymbol{\beta}_1, \mathbf{B}_2)$ that transform the statistical parameter $\boldsymbol{\alpha}$ into the structural parameter $\boldsymbol{\delta}$. These are the so-called overidentifying restrictions which are testable, but the reliability of such a test depends crucially on the statistical adequacy of the statistical model (75). Indeed, the choice of 'optimal' instruments is depends crucially on the statistical adequacy of (75). For instance, if the vector stochastic process $\{\mathbf{Z}_t, \ t \in \mathbb{T}\}$, where $\mathbf{Z}_t := (y_t, \mathbf{X}_t^\top, \mathbf{W}_t^\top)^\top$, is Markov dependent (see Spanos, 1999, ch. 8), the set of optimal instruments that would render (75) statistically adequate is likely to include lags of this process, even if the relevant static economic theory might suggest that the past history of the process is irrelevant; see Spanos (1986, ch. 25). It is important to emphasize that the restrictions $\boldsymbol{\beta}_1 \neq \mathbf{0}, \ \mathbf{B}_2 \neq \mathbf{0}$ are testable, and need to be tested, in the context of the statistical model (75) before testing the overidentifying restrictions.

If we return to the original question of substantive adequacy, as it relates to (78), we can see that the numerical difference between the values of the two least-squares estimators:

$$\widehat{\boldsymbol{\delta}}=\left(\mathbf{X}^{\intercal}\mathbf{X}\right)^{-1}\mathbf{X}^{\intercal}\mathbf{y}, \qquad \widehat{\boldsymbol{\alpha}}=\left(\mathbf{X}^{\intercal}\mathbf{M}_W\mathbf{X}\right)^{-1}\mathbf{X}^{\intercal}\mathbf{M}_W\mathbf{y},$$

where $\mathbf{M}_W=\mathbf{I}-\mathbf{W}\left(\mathbf{W}^{\intercal}\mathbf{W}\right)^{-1}\mathbf{W}^{\intercal}$, provide no valuable information concerning the sensitivity of this inference. This is because this difference has no direct bearing on the substantive hypotheses (78). What might be more informative in this case is the difference between the OLS and the Instrumental Variables estimators**:**

$$\widehat{\boldsymbol{\delta}}_{IV}=\left(\mathbf{X}^{\intercal}\mathbf{P}_W\mathbf{X}\right)^{-1}\mathbf{X}^{\intercal}\mathbf{P}_W\mathbf{y}, \text{ where } \mathbf{P}_W=\mathbf{I}-\mathbf{M}_W.$$

Indeed, the difference $(\widehat{\boldsymbol{\delta}}-\widehat{\boldsymbol{\delta}}_{IV})$, which is related to (78), provides the basis of the Durbin-Wu-Hausman test for exogeneity; see Spanos (2006c) for further details.

An important conclusion following from the above discussion is that the numerical similarity between two Instrumental Variables estimators of $\boldsymbol{\delta}$, based on two different sets of instruments, say $\mathbf{W}_{1t}$ and $\mathbf{W}_{2t}$, does *not* indicate the appropriateness of either set of instruments; they can both lead to equally unreliable inferences. What establishes appropriateness and inappropriateness in this context is the statistical adequacy of the embedding statistical model (75) and the substantive adequacy of the structural model in question, not the apparent numerical insensitivity of the estimates to the choice of instruments.

5 Statistical inadequacy and substantive inference

Throughout the discussion of assessing substantive adequacy in sections 2-4, it was assumed that both models M_0 (27) and M_1 (28) were *statistically adequate*; assumptions [**1**]-[**5**] (see table 1) are valid for M_0 (with $\mathbf{X}_t$) and M_1 (with $\mathbf{X}_t^*:=(\mathbf{X}_t,\mathbf{W}_t)$). When this is not the case, the discrepancy between *nominal* and *actual error probabilities* will often give rise to unreliable inferences; see Spanos (2005a).

Setting 1: Models M_0 and M_1 are both statistically adequate. This is the case where we know that both sets of parameters $(\boldsymbol{\beta},\sigma_u^2)$ and $(\boldsymbol{\alpha},\boldsymbol{\gamma},\sigma_\varepsilon^2)$ will be *statistically* well estimated, and any tests will have the correct size and power characteristics; the discussion concerning scenarios A-E in section 4.1, go through unaltered.

Setting 2: M_1 is statistically adequate but M_0 is statistically inadequate. In this case the discussion concerning scenarios A-E needs to be qualified. Any tests of hypotheses concerning $\boldsymbol{\beta}$ in the context of model M_0 are likely to be affected by its statistical inadequacy. However the F-test for hypotheses in the context of M_1 is robust to the statistical inadequacy of M_0 because, in the case of testing the restrictions $\boldsymbol{\gamma}=\mathbf{0}$, it coincides with the Wald (W) test which is based exclusively on the estimators of M_1. In particular, one can show that the F-test statistic in (46), reduces to:

$$F(\mathbf{y})=\frac{(\mathbf{R}\widehat{\boldsymbol{\delta}})^{\top}\left[\mathbf{R}\left(\mathbf{X}^{*\intercal}\mathbf{X}^{*}\right)^{-1}\mathbf{R}^{\top}\right]^{-1}(\mathbf{R}\widehat{\boldsymbol{\delta}})}{ms_\varepsilon^2}=W(\mathbf{y}), \tag{80}$$

where $\mathbf{R}=(\mathbf{0}_{m\times k}\colon\mathbf{I}_{m\times m})$, $\widehat{\boldsymbol{\delta}}=\left(\widehat{\boldsymbol{\alpha}}^{\top},\widehat{\boldsymbol{\gamma}}^{\top}\right)^{\top}$, $\mathbf{X}^{*}{:=}\left(\mathbf{X}\colon\mathbf{W}\right)$, $s_\varepsilon^2{=}(\widehat{\boldsymbol{\varepsilon}}^{\top}\widehat{\boldsymbol{\varepsilon}}/T{-}k{-}m)$.

Given that M_1 is statistically adequate, the actual power properties of both tests will coincide with the nominal, but the nominal type I error probabilities might differ from the actual ones because the latter are evaluated under M_0; this is less of a problem for testing $\mathbf{\Sigma}_{23} = 0$ because the relevant statistical model under this restriction remains M_1.

The danger with this case is that when one does *not* do a sufficiently competent job in respecifying the original model with the view to reach a statistically adequate model using data $\{(\mathbf{W}_t, y_t)\ t = 1, 2, ..., T\}$, the presence of departures from the probabilistic assumptions might be erroneously interpreted as an indication of missing factors. Such a misleading inference is likely to be confirmed when such omitted variables are included in the model because, more often than not, $\mathbf{W}_t$ can play the role of a 'proxy' for certain missing statistical information such as trends, temporal dependence and non-linearities; irrespective of their substantive significance. This problematic situation can be prevented by establishing the statistical adequacy of M_1 itself, and not relying on the statistical inadequacy of M_0 as providing evidence for M_1.

Setting 3: M_1 is statistically inadequate but M_0 is statistically adequate. In this case the testing procedure concerning scenarios A-E need to be modified. Given that M_0 is statistically adequate, the type I error probability of the test for $\boldsymbol{\gamma} = \mathbf{0}$ will coincide with the actual, but its nominal power is likely to differ from the actual. Hence, in testing $\boldsymbol{\alpha}= \mathbf{0}$ or $\boldsymbol{\gamma} = \mathbf{0}$, one might supplement the F-test with the *Lagrange Multiplier (LM) test*, which is less vulnerable to the statistical inadequacy of M_1, because it is based exclusively on the estimators of M_0. In particular, one can show that the LM test (see Spanos, 1986, ch. 20) takes the form:

$$LM(\mathbf{y}){=}\frac{(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^{\top}\left[\mathbf{W}\left(\mathbf{W}^{\intercal}\mathbf{M}_X\mathbf{W}\right)^{-1}\mathbf{W}^{\intercal}\right](\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})}{ms_u^2} \backsim \mathsf{F}(\delta^*; m, T{-}k), \tag{81}$$

$$C_1 = \{\mathbf{y} : LM(\mathbf{y}) > c_\alpha\} \;\text{ being the rejection region,}$$

$\delta^* = \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\top}\left[\mathbf{W}(\mathbf{W}^{\intercal}\mathbf{M}_X\mathbf{W})^{-1}\mathbf{W}^{\intercal}\right](\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma_u^2}$; compare the denominator of (80) and (81).

The primary problem with the case where M_1 is statistically inadequate is that no positive answer can be given to the question of substantive relevance for $\mathbf{W}_t$; one can only establish the inadequacy of M_0. To say anything positive one needs to probe the statistical inadequacy of M_1 in order to establish whether the substantive factors $\mathbf{W}_t$ are relevant, but (i) they do not enter the model as assumed, or (ii) other additional factors are still missing.

Setting 4: M_1 and M_0 are both statistically inadequate. This is the most difficult case for inference purposes because one would have a hard time distinguishing between statistical and substantive inadequacy. In this case the testing discussion concerning scenarios A-E is likely to give rise to unreliable inferences. Hence, the first task of the modeler is to respecify at least M_1, in order to ensure statistical adequacy and use the discussion in setting 2.

Setting 5: Data $\{w_t,\ t{=}1, 2, ..., T\}$ are unavailable. In such a case inference has to be based on M_0 by necessity. If all tenacious attempts to find a statistically adequate model based exclusively on the data $\{(\mathbf{x}_t, y_t)\ t{=}1, 2, ..., T\}$ fail, then a natural conclusion is that some missing factors is the likeliest source of the systematic information in the residuals of M_0; a conjecture that needs to be probed. This is a situation where an anomaly has been detected (systematic residuals) and one can use the statistical information as a source of guidance in finding the relevant omitted factor(s). The situation is less transparent in cases where M_0 turns out to be statistically adequate.

6 Conclusion

The problem of omitted variables is traditionally viewed as a statistical misspecification issue which renders unreliable any inference concerning the influence of $\mathbf{X}_t$ on y_t due to the exclusion of certain factors $\mathbf{W}_t$. The textbook omitted variables argument attempts to assess the seriousness of this unreliability using the sensitivity of the estimator $\widehat{\boldsymbol{\beta}}$ to the inclusion/exclusion of $\mathbf{W}_t$ as evidencing its *bias/inconsistency.*

In this paper, it is argued that the textbook argument in terms of the sensitivity of point estimates provides a poor basis for the task. The confounding problem should not be viewed as a form of statistical misspecification, but as a departure from the substantive premises, which is concerned with the adequacy of the structural model in explaining the behavior of y_t. This issue can only be addressed using statistically reliable procedures when the structural model is embedded into a statistical model whose premises are adequate. By separating the statistical from the substantive adequacy, the paper recasts the confounding question into a Neyman-Pearson hypothesis testing problem, supplemented with a post-data evaluation of inference based on severe testing. Using an empirical example it is shown how this testing perspective can deal effectively with assessing the confounding issue. In the case where one of the two models, estimated with and without the variables $\mathbf{W}_t$, is statistically misspecified, the choice of inference methods becomes important. This provides a way to motivate the use of different test procedures, such as the likelihood ratio, Wald and Lagrange Multiplier methods, under different misspecification scenarios.

References

[1] Cox, D. R. and D. V. Hinkley (1974), *Theoretical Statistics*, Chapman & Hall, London.

[2] Doob, J. L. (1953), *Stochastic Processes*, Wiley, New York.

[3] Fisher, R. A. (1935), *The Design of Experiments*, Oliver and Boyd, Edinburgh.

[4] Greene, W. H. (1997), *Econometric Analysis*, 3rd Ed., Prentice Hall, NJ.

[5] Hoover, K. D. (2001), *Causality in Macroeconomics*, Cambridge University Press, Cambridge.

[6] Johnston, J. (1984), *Econometric methods*, 3rd edition, McGraw-Hill, New York.

[7] Leamer, E. E. (1978), *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, Wiley, New York.

[8] Leamer, E. E. and H. B. Leonard (1983), "Reporting the fragility of regression estimates, *Review of Economics and Statistics*, **65**, 306-317.

[9] Mayo, D. G. (1991), "Novel Evidence and Severe Tests," *Philosophy of Science*, **58,** 523-552.

[10] Mayo, D. G. (1996), *Error and the Growth of Experimental Knowledge*, The University of Chicago Press, Chicago.

[11] Mayo, D. G. and A. Spanos (2004), "Methodology in Practice: Statistical Misspecification Testing", *Philosophy of Science*, **71**, 1007-1025.

[12] Mayo, D. G. and A. Spanos (2006), "Severe Testing as a Basic Concept in the Neyman-Pearson Philosophy of Induction," forthcoming in *The British Journal for the Philosophy of Science.*

[13] Pagan, A. (1984), "Model Evaluation by Variable Addition," pp. 103-135 of Hendry, D. F. and K. F. Wallis (eds.) (1984), *Econometrics and Quantitative Economics*, Basil Blackwell, Oxford.

[14] Pearl, J. (2000), *Causality: Models, Reasoning and Inference*, Cambridge University Press, Cambridge.

[15] Spanos, A., (1986), *Statistical Foundations of Econometric Modelling*, Cambridge University Press, Cambridge.

[16] Spanos, A. (1989), "On re-reading Haavelmo: a retrospective view of econometric modeling", *Econometric Theory*, **5**, 405-429.

[17] Spanos, A. (1990), "The Simultaneous Equations Model revisited: statistical adequacy and identification", *Journal of Econometrics*, **44**, 87-108.

[18] Spanos, A. (1995a), "On theory testing in Econometrics: modeling with nonexperimental data", *Journal of Econometrics,* 67:189-226.

[19] Spanos, A. (1995b), "On Normality and the Linear Regression model", *Econometric Reviews*, **14**, 195-203.

[20] Spanos, A. (1999), *Probability Theory and Statistical Inference: econometric modeling with observational data*, Cambridge University Press, Cambridge.

[21] Spanos, A. (2005a), "Misspecification and the Reliability of Inference: the t-test in the presence of Markov dependence," Virginia Tech Working Paper.

[22] Spanos, A. (2005b), "Omitted Variables and Artificial Regressions: A General Approach to Misspecification Testing," Virginia Tech Working Paper.

[23] Spanos, A. (2005c) "Structural Equation Modeling, Causal Inference and Statistical Adequacy," pp. 639-661, *Logic, Methodology and Philosophy of Science: Proceedings of the Twelfth International Congress,* Editors, P. Hajek, L. Valdes-Villanueva and D. Westerstahl, King's College, London.

[24] Spanos, A. (2005d), "Structural vs. Statistical Models in Empirical Modeling: Kepler's first law of planetery motion revisited," Virginia Tech working paper.

[25] Spanos, A. (2006a), "Econometrics in Retrospect and Prospect," in the *Palgrave Handbook of Econometrics, vol. 1: Theoretical Econometrics,* London: MacMillan, pp. 3-58.

[26] Spanos, A. (2006b), "The Instrumental Variables Method revisited: On the Nature and Choice of Optimal Instruments," forthcoming in *Essays in Memory of Michael Magdalinos*, ed. by G. D. A. Phillips, Cambridge University Press, Cambridge.

[27] Spanos, A. (2006c), "Where Do Statistical Models Come From? Revisiting the Problem of Specification," forthcoming in *The Second Erich L. Lehmann Symposium*, vol. ##, Institute of Mathematical Statistics.

[28] Spanos, A. and A. McGuirk (2002), "The problem of near-multicollinearity revisited: erratic vs systematic volatility", *The Journal of Econometrics*, **108**, 365-393.

[29] Spirtes, P., C. Glymor and R. Scheines (2000), *Causation, Prediction and Search*, 2nd edition, The MIT Press, Cambridge.

[30] Stock, J. H. and M. W. Watson (2002), *Introduction to Econometrics*, Addison Wesley, Boston, MA.

[31] Wooldridge, J. M. (2003), *Introductory Econometrics: A Modern Approach*, 2nd ed., Thomson, South-Western, OH.