UNIVERSITY OF LEICESTER

Department of Economics

Removing Specification Errors from the Usual Formulation of Binary Choice Models*

P. A. V. B. Swamy, Federal Reserve Board
I-Lok Chang, American University
Jatinder S. Mehta, Philadelphia
William H. Greene, New York University
Stephen G. Hall, University of Leicester
George S. Tavlas, Bank of Greece

Working Paper No. 16/11

Removing Specification Errors from the Usual Formulation of Binary Choice Models[*]

P. A. V. B. Swamy[a], I-Lok Chang[b], Jatinder S. Mehta[c], William H. Greene[d], Stephen G. Hall[e], and George S. Tavlas[f]

Abstract

We develop a procedure for removing four major specification errors from the usual formulation of binary choice models. The model that results from this procedure is different from the conventional probit and logit models. This difference arises as a direct consequence of our relaxation of the usual assumption that omitted regressors constituting the error term of a latent linear regression model do not introduce omitted regressor biases into the coefficients of the included regressors.

[*] We are grateful to four referees for their thoughtful comments.

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

[b] Department of Mathematics (Retired), American University, Washington, DC, e-mail: ilchang@american.edu

[c] Department of Mathematics (Retired), Philadelphia, PA 19122, e-mail: mehta1007@comcast.net

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

[e] Leicester University, Deputy pro vice chancellor and Bank of Greece, Room Astley Clarke 116, University Road, Leicester LEI 7RH, UK, e-mail: s.g.hall@le.ac.uk

[f] Member, Monetary Policy Council, Bank of Greece, 21 El. Venizelos Ave. 102 50, Athens, Greece, e-mail: gtavlas@bankofgreece.gr

1. Introduction

It is well-known that binary choice models are subject to certain specification errors. It can be shown that the usual approach of adding an error term to a mathematical function leads to a model with nonunique coefficients and error term. In this model, the conditional expectation of the dependent variable given the included regressors does not always exist. Even when it exists, its functional form may be unknown. The nonunique error term is interpreted as representing the net effect of omitted regressors on the dependent variable. Pratt and Schlaifer (1988, p. 34) pointed out that omitted regressors are not unique and as a result, the condition that the included regressors be independent of 'the' excluded variables themselves is "meaningless". There are cases where the correlation between the nonunique error term and the included regressors can be made to appear and disappear at the whim of an arbitrary choice between two observationally equivalent models. To avoid these problems, we specify models with unique coefficients and error terms without misspecifying their correct functional forms. The unique error term of a model is a function of certain 'sufficient sets' of omitted regressors. We derive these sufficient sets for a binary choice model in this paper. In the usual approach, omitted regressors constituting the error term of a model do not introduce omitted-regressor biases into the coefficients of the included regressors. In our approach, they do so.

Following the usual approach, Yatchew and Griliches (1984)[1] showed that if one of two uncorrelated regressors included in a simple binary choice model is omitted, then the estimator of the coefficient on the remaining regressor will be inconsistent. They also showed that, if the disturbances in a latent regression model are heteroscedastic, then the maximum likelihood estimators that assume homoscedasticity are inconsistent and the covariance matrix is inappropriate. In this paper, we show that the use of a latent regression model with unique

[1] See, also, Greene (2012, chapter 17, p. 713).

coefficients and error term changes their results. Our binary choice model is different from those of such researchers as Yatchew and Griliches, Cramer (2006/07), and Wooldridge (2002, Chapter 15). The concept of unique coefficients and error term is distinctive to our work. Specifically, we do not assume any incorrect functional form, and we account for relevant omitted regressors, measurement errors, and correlations between excluded and the included regressors. Our model features varying coefficients (VCs) in which we interpret the VC on a continuous regressor as a function of three quantities: (i) bias-free partial derivative of the dependent variable with respect to the continuous regressor, (ii) omitted-regressor biases, and (iii) measurement-error biases. This interpretation of the VCs is unique to our work and allows us to focus on the bias-free (*i.e.,* partial derivatives) parts of the VCs.

The remainder of this paper is comprised of three sections. Section 2 summarizes a recent derivation of Swamy, Mehta, Tavlas and Hall (2014) of all the terms involved in a binary choice model with unique coefficients and error term. The section also provides the conditions under which such a model can be consistently estimated. Section 3 presents an empirical example. Section 4 concludes. An Appendix at the end of the paper has two sections. The first section compares the relative generality of assumptions underlying different linear and nonlinear models. The second section derives the information matrix for a binary choice model with unique coefficients and error term.

2. Methods of Correctly Specifying Binary Choice Models and Their Estimation

2.1 Model for a Cross-section of Individuals

Greene (2012, pp. 681-683) described various situations under which the use of discrete choice models is called for. In what follows, we develop a discrete choice model that is free of several specification errors. To explain, we begin with the following specification:

$$y_i^* = \psi_i(x_{i1}^*,...,x_{iL_i}^*) \quad (1)$$

where i indexes n individuals, $\psi_i(x_{i1}^*,...,x_{iL_i}^*)$ is a mathematical function, and its arguments are mathematical variables. Let $\psi_i(.)$ be short hand for this function. We do not observe y_i^* but view the outcome of a discrete choice as a reflection of the underlying mathematical function in equation (1). We only observe whether a choice is made or not (see Greene 2012, p. 686). Therefore, our observation is

$$y_i = 1 \quad \text{if } y_i^* > 0,$$

$$y_i = 0 \quad \text{if } y_i^* \leq 0 \quad (2)$$

where the choice "not made" is indicated by the value 0 and the choice "made" is indicated by the value 1, i.e., y_i takes either 0 or 1. An example of model (2), provided in Greene (2012, Example 17.1, pp. 683-684), is a situation involving labor force participation where a respondent either works or seeks work (y_i = 1) or does not (y_i = 0) in the period in which a survey was taken.[2]

In equation (1), $\boldsymbol{x_i^*} = (x_{i1}^*,...,x_{i,L_i}^*)'$ is $L_i \times 1$, L_i denotes the total number of the arguments of $\psi_i(x_{i1}^*,...,x_{i,L_i}^*) = \psi_i(.)$; there are no omitted arguments needing an error term. Equation (1) is a mathematical equation that holds exactly. Inexactness and stochastic error term enters into (1) when we derive the appropriate error term and make distributional assumption about it. The reasons for this are explained below. The number L_i may depend on i. This dependence occurs when the number of arguments of $\psi_i(.)$ is different for different individuals. For example, before deciding whether or not to make a large purchase, each consumer makes a marginal

[2] We will show below that the inconsistency problems Yatchew and Griliches (1984, p. 713) pointed out with the probit and logit models are eliminated by replacing these models by the model in (1) and (2).

benefit/marginal cost calculations based on the utilities achieved by making the purchase and by not making the purchase, and by using the available funds for the purchase of something else. The difference between benefit and cost as an unobserved variable y_i^* can vary across consumers if they have different utility functions with different arguments. These variations can show up in the form of different arguments of $\psi_i(.)$ for different consumers.

It should be noted that (1) represents our departure from the usual approach of adding a nonunique error term to a mathematical function and making a "meaningless" assumption about the error term. Pratt and Schlaifer (1988, p. 34) severely criticized this approach. To avoid this criticism, what we have done in (1) is that we have taken all the **x**-variables and the variables constituting the error term ε in Cramer's (2006/07) (or e in Wooldridge's (2002, p. 457)) latent regression model, and included them in $\psi_i(.)$ as its arguments. In addition, we also included in $\psi_i(.)$ all relevant pre-existing conditions as its arguments.[3]

The problem that researchers face is that of uncovering the correct functional form of $\psi_i(.)$.[4] However, any false relations can be shown to have been eliminated when we control for all relevant pre-existing conditions. To make use of this observation due to Skyrms (1988, p. 59), we incorporate these pre-existing conditions into $\psi_i(.)$ by letting some of the elements of $\boldsymbol{x}_i^*$ represent these conditions.[5] Clearly, we have no way of knowing what these pre-existing conditions might be, how to measure them (if we knew them), or how many there may be. To control for these conditions, we use the following approach. We assume that all relevant pre-existing conditions appear as arguments of the function $\psi_i(.)$ in equation (1). This is a

[3] We explain in the next paragraph why we have included these conditions.
[4] Some researchers may believe that there is no such thing as the true functional form of (1). Whenever we talk of the correct functional form of (1), we mean the functional form of (1) that is appropriate to the particular binary choice in (2).
[5] Here we are using Skyrms' (1988, p. 59) definition of the term 'all relevant pre-existing conditions.'

straightforward approach. Therefore, when we take the partial derivatives of $\psi_i(.)$ with respect to x_{ij}^*, a determinant of y_i^* included in $\psi_i(.)$ as its argument, the values of all pre-existing conditions are automatically held constant. This action is important because it sets the partial derivative $\partial y_i^* / \partial x_{ij}^*$ equal to zero whenever the relation of y_i^* to x_{ij}^*, an element $\boldsymbol{x}_i^*$, is false (see Skyrms 1988, p. 59).

The function $\psi_i(.)$ in (1) is exact and mathematical in nature, without any relevant omitted arguments. Moreover, its unknown functional form is not misspecified. Therefore, it does not require an error term; indeed, it would be incorrect to add an error term to $\psi_i(.)$. We refer to (1) as "a minimally restricted mathematical equation," the reason being that no restriction other than the normalization rule, that the coefficient of y_i^* is unity, is imposed on (1). Without this restriction, the function $\psi_i(.)$ is difficult to identify. The reason why no other restriction is imposed on it is that we want (1) to be a real-world relationship. With such a relationship we can estimate the causal effects of a treatment. Basmann (1988, pp. 73 and 99) argued that causality is a property of the real world. We define that real-world relations are those that are not misspecified. Causal relations are unique in the real world. This is the reason why we insist that the coefficients and error term of our model be unique. From Basmann's (1988, p. 98) definition it follows that (1) is free of the most serious objection, i.e., non-uniqueness, which occurs when stationarity producing transformation of observable variables are used.[6] We do not use such variables in (1).

[6] This is Basmann's (1988, pp. 73 and 99) statement.

2.2 Unique Coefficients and Error Terms of Models

2.2.1 Causal relations: Basmann (1988, pp. 73 and 99) emphasized that the word 'causality' designates a property of the real world." Hence we work only with appropriate real-world relationships to evaluate causal effects.

We define that the real-world relationships are those that are not subject to any specification errors. It is possible to avoid some of those errors, as we show below. The real-world relationships and their properties are always true and unique. Such relationships cannot be found, however, by imposing severe restrictions because they can be false. Examples of these restrictions are given in the Appendix to the paper. For example, certain separability conditions are imposed on (1) to obtain (A1) in the Appendix. As a result of these conditions, (A1) may not be a real-world relationship and may not possess the causality property. Again in the Appendix, (A2) is a general condition of statistical independence which is very strong. Model (A5) of the Appendix with a linear functional form could be misspecified.[7]

2.2.2 Derivation of a model from (1) without committing a single specification error: To avoid misspecifications of the unknown correct functional form of (1), we change the problem of estimating (1) to the problem of estimating some of its partial derivatives in

$$y_i^* = \alpha_{i0}^* + x_{i1}^*\alpha_{i1}^* + \cdots + x_{iL_i}^*\alpha_{iL_i}^* \qquad (3)$$

where, for $\ell = 1, \ldots, L_i$, $\alpha_{i\ell}^* = \dfrac{\partial \psi_i(.)}{\partial x_{i\ell}^*}$ if $x_{i\ell}^*$ is continuous and $= \dfrac{\Delta \psi_i(.)}{\Delta x_{i\ell}^*}$ with the right sign if $x_{i\ell}^*$ is discrete having zero as one of its possible values, Δ is the first-difference operator, and the intercept $\alpha_{i0}^* = y_i^* - \sum_{\ell=1}^{L_i} x_{i\ell}^*\alpha_{i\ell}^*$. In words, this intercept is the error of approximation due to

[7] Cramer (2006/07, p.2) and Wooldridge (2002, p. 457) assume that their latent regression models are linear. This is a usual assumption. We are simply justifying our unusual assumptions without criticizing the usual assumptions.

approximating y_{it}^{*} by ($\sum_{\ell=1}^{L_i} x_{i\ell}^{*}\alpha_{i\ell}^{*}$). Therefore, model (3) with zero intercept does not misspecify the unknown functional form of (1) if the error of approximation is truly zero and model (3) with nonzero intercept is the same as (1) with no misspecifications of its functional form, since y_i^{*} - $\sum_{\ell=1}^{L_i} x_{i\ell}^{*}\alpha_{i\ell}^{*}$ + $\sum_{\ell=1}^{L_i} x_{i\ell}^{*}\alpha_{i\ell}^{*}$ = y_i^{*}. This is how we deal with the problem of the unknown functional form of (1). Note that no separability conditions need to be imposed on (1) to write it in the form of (3). This is the advantage of (3) over (A1).

Note that in the above definition of the partial derivative ($\alpha_{i\ell}^{*}$), the values of all the arguments of $\psi_{it}(.)$ (including all relevant pre-existing conditions) other than $x_{i\ell}^{*}$ are held constant. These partial derivatives are different from those that can be derived from (A1) with ε_i suppressed. This is because in taking the latter derivatives the values of $x_{i,K+1}^{*},...,x_{iL_t}^{*}$ are not held constant.

Equation (3) is not a false relationship, since we held the values of all relevant pre-exiting conditions constant in deriving its coefficients. The regression in (3) has the minimally restricted equation in (1) as its basis. The coefficients of (3) are constants if (1) is linear and are variables otherwise. In the latter case, the coefficients of (3) can be the functions of all of the arguments of $\psi_{it}(.)$. Any function of the form (1) with unknown functional form can be written as linear in variables and nonlinear in coefficients, as in (3). We have already established that this linear-in-variables and nonlinear-in-coefficients model has the correct functional form if its intercept is zero and is the same as (1) otherwise. In either case, (3) does not have a misspecified functional form. In this paper, we take advantage of this procedure.

Not all elements of $\boldsymbol{x}_i^*$ are measured; suppose that the first *K* of them are measured. This assumption needs the restriction that min($L_1,...,L_n$) > *K*. Even these measurements may contain errors so that the observed argument x_{ij} is equal to the sum of the true value x_{ij}^* and a measurement error, denoted by v_{ij}^*.[8] The arguments, x_{ig}^*, $g = K+1, \ldots, L_i$, for which no data are available, are treated as regressors omitted from (3).[9] These are of two types: (i) unobserved determinants of y_i^* and (ii) all relevant pre-existing conditions. We know nothing of these two types of variables. With these variables being present in (3), we cannot estimate it. Again without misspecifying (1) these variables should be eliminated from (3). To do so, we consider the "auxiliary" relations of each x_{ig}^* to $x_{i1}^*,...,x_{iK}^*$. Such relations are: For $g = K+1 \ldots, L_i$,

$$x_{ig}^* = \lambda_{ig0}^* + x_{i1}^*\lambda_{ig1}^* + \cdots + x_{iK}^*\lambda_{igK}^* \qquad (4)$$

where $\lambda_{igj}^* = \dfrac{\partial x_{ig}^*}{\partial x_{ij}^*}$ if x_{ij}^* is continuous holding the values of all the regressors of (A8) other than that of x_{ij}^* constant and = $\dfrac{\Delta x_{ig}^*}{\Delta x_{ij}^*}$ with the right sign if x_{ij}^* is discrete taking zero as one of its possible values and $\lambda_{ig0}^* = x_{ig}^* - \sum_{j=1}^K x_{tj}^*\lambda_{igj}^*$. This intercept is the portion of x_{ig}^* remaining after the effect ($\sum_{j=1}^K x_{ij}^*\lambda_{igj}^*$) of $x_{i1}^*,...,x_{iK}^*$ on x_{ig}^* has been subtracted from it.

In (4), there are L_i - *K* relationships. The intercept λ_{ig0}^* is the error due to approximating the relationship between the gth omitted regressor and all the included regressors ("the correct relationship") by $\sum_{j=1}^K x_{tj}^*\lambda_{igj}^*$. If this error of approximation is truly zero, then equation (4) with

[8] We postpone making stochastic assumptions about measurement errors.
[9] The label 'omitted' means that we would remove them from (3).

zero intercept has the same functional form as the correct relationship. In the alternative case where the error of approximation is not zero, (4) is the same as the correct relationship, i.e., x_{ig}^* - $\sum_{j=1}^{K} x_{tj}^* \lambda_{igj}^*$ + $\sum_{j=1}^{K} x_{tj}^* \lambda_{igj}^*$. In either case (4) does not misspecify the correct functional form. According to Pratt and Schlaifer (1988, p. 34), the condition that the included regressors be independent of 'the' omitted regressors themselves is meaningless. This statement supports (4) but not the usual assumption that the error term of a latent regression model is uncorrelated with or independent of the included regressors.[10] The problem is that omitted regressors are not unique, as Pratt and Schlaifer (1988, p. 34) proved.

2.2.3 A latent regression model with unique coefficients and error terms: Substituting the right-hand side of equation (4) for x_{ig}^* in (3) gives

$$y_i^* = \alpha_{i0}^* + \sum_{g=K+1}^{L_i} \lambda_{tg0}^* \alpha_{ig}^* + \sum_{j=1}^{K} x_{ij}^* (\alpha_{ij}^* + \sum_{g=K+1}^{L_i} \lambda_{igj}^* \alpha_{ig}^*) \quad (5)$$

The error term, the intercept, and the coefficients of the nonconstant regressors of this model are $\sum_{g=K+1}^{L_{it}} \lambda_{ig0}^* \alpha_{ig}^*$, α_{i0}^*, and $(\alpha_{ij}^* + \sum_{g=K+1}^{L_i} \lambda_{igj}^* \alpha_{ig}^*)$, $j = 1, \ldots, K$, respectively.

Bias-free partial derivatives: $\alpha_{ij}^* = \frac{\partial \psi_i(.)}{\partial x_{ij}^*}$ or $= \frac{\Delta \psi_i(.)}{\Delta x_{ij}^*}$, $j = 1, \ldots, K$. (6)

These partial derivatives have the correct functional form if the x_{ij}^*'s are continuous.

'Sufficient' sets of omitted regressors: The regressors, $x_{i,,K+1}^*, \ldots, x_{iL_i}^*$, are called "omitted regressors" because they are included in (3) but not in equation (5). The regressors $x_{i1}^*, \ldots, x_{iK}^*$ are called "the included regressors." It can be seen from (5) that the portions $\lambda_{i,K+1,0}^*, \ldots, \lambda_{iL_i 0}^*$ of

[10] Cramer (2006/07, p. 4) and Wooldridge (2002, p. 457) make the usual assumption.

omitted regressors, $x^*_{i,K+1},...,x^*_{iL_i}$, respectively, in conjunction with the included regressors $x^*_{i1},...,x^*_{iK}$ are sufficient to determine the value of y^*_i exactly. For this reason, Pratt and Schlaifer (1988, p. 34) called $\lambda^*_{i,K+1,0},...,\lambda^*_{i,L_i,0}$ "certain 'sufficient sets' of omitted regressors." The second term ($\sum_{g=K+1}^{L_i}\lambda^*_{ig0}\alpha^*_{ig}$) on the right-hand side of (5) is called "a function of these sufficient sets of omitted regressors $x^*_{i,K+1},...,x^*_{iL_i}$." Pratt and Schlaifer (1984, 1988) pointed out that this function can be taken as the error term of (5). It remains as a mathematical function until we make a distributional assumption about it. Note that the problem with the error terms of the usual latent regression models including those of (A1), Karlsen, Myklebust and Tjøstheim's (2007) and White's (1980, 1982) models is that they are not the appropriate functions of the sufficient sets of omitted regressors and hence are not unique and/or are arbitrary.[11]

Deterministic omitted-regressors bias: The term $\sum_{g=K+1}^{L_i}\lambda^*_{igj}\alpha^*_{ig}$ contained in the coefficient of x^*_{tj} in (5) measures such a bias.

Swamy et al. (2014, pp. 197, 199 and 217-219) proved that the coefficients $(\alpha^*_{ij}+\sum_{g=K+1}^{L_{it}}\lambda^*_{igj}\alpha^*_{ig})$ and the error term $\sum_{g=K+1}^{L_i}\lambda^*_{ig0}\alpha^*_{ig}$ of (5) are unique in the following sense:

Uniqueness: The coefficients and error term of model (5) are unique if they are invariant under the addition and subtraction of the coefficient of any regressor omitted from (5) times any regressor included in (5) on the right-hand side of (3) (see Swamy et al. 2014, pp. 199 and 219).

The equations in (4) play a crucial role in Swamy et al.'s (2014) proof of the uniqueness of the coefficients and error term of (5). If we had taken the sum $x^*_{i,K+1}\alpha^*_{i,K+1}+\cdots+x^*_{iL_i}\alpha^*_{iL_i}$ of the

[11] Cramer (2006/07, p. 4) and Wooldridge (2002, p. 457) adopt the usual latent regression models with nonunique coefficients and error terms.

last L_i - K terms on the right-hand side of (3) as its error term, then we would have obtained a nonunique error term. The reason why this would have happened is that omitted regressors are not unique. What (4) has done is that it has split each *g*th omitted regressor into a 'sufficient set' and an included-regressors' effect. This sufficient set times the coefficient of the *g*th omitted regressor has become a term in the (unique) error term of (5) and the included-regressors' effect times the coefficient of the *g*th omitted regressor has become a term in omitted-regressor biases of the coefficients of (5). This is not the usual procedure where the whole of each omitted regressor goes into the formation of an error term and no part of it becomes a term in omitted-regressor biases. The usual procedure leads to nonunique coefficients and error term. In the YG procedure, only some of the included regressors which, when omitted, introduce omitted-regressor biases into the (nonunique) coefficients on the remaining included regressors. Yatchew and Griliches (1984), Wooldridge (2002), and Cramer (2006/07) followed the usual procedure. Without using (4) it is not possible to derive a model with unique coefficients and error term.

2.2.4 A correctly specified latent regression model: Substituting the terms on the right-hand sides of equations $x_{ij}^* = x_{ij} - v_{ij}^*$, $j = 1, \ldots, K$, for x_{ij}^*, $j = 1, \ldots, K$, respectively, in (5) gives a model of the form

$$y_i^* = \gamma_{i0} + x_{i1}\gamma_{i1} + \cdots + x_{iK}\gamma_{iK} \tag{7}$$

where the intercept is defined as

$$\gamma_{i0} = \alpha_{i0}^* + \sum_{g=K+1}^{L_i} \lambda_{ig0}^* \alpha_{ig}^* - \sum_{v_{ij}^* \in S_2} v_{ij}^* (\alpha_{ij}^* + \sum_{g=K+1}^{L_i} \lambda_{igj}^* \alpha_{ig}^*) \tag{8}$$

and the other terms are defined as

$$x_{ij}\gamma_{ij} = x_{ij}(1-\frac{\nu^*_{ij}}{x_{ij}})(\alpha^*_{ij} + \sum_{g=K+1}^{L_i} \lambda^*_{igj}\alpha^*_{ig}) \text{ if } x_{ij} \in S_1$$

$$= x_{ij} (\alpha^*_{ij} + \sum_{g=K+1}^{L_i} \lambda^*_{igj}\alpha^*_{ig}) \text{ if } x_{ij} \in S_2 \quad (9)$$

where the set S_1 contains all the regressors of equation (7) that are continuous, the set S_2 contains all the regressors of (7) that can take the value zero with positive probability, the ratio of $(1-\frac{\nu^*_{ij}}{x_{ij}})$ in the first line of equation (9) comes from the equation $x^*_{ij} = x_{ij} - \nu^*_{ij} = (1 - \frac{\nu^*_{ij}}{x_{ij}})x_{ij}$, and this ratio does not appear in the second line of equation (9) because $x_{ij} \in S_2$ can take the value zero with positive probability.

Equation (7) implies that a model is correctly specified if it is derived by inserting measurement errors at the appropriate places in a model with unique coefficients and error term (see Swamy et al. 2014, p. 199).

Deterministic measurement-error biases: The formula $-\sum_{\nu^*_{ij} \in S_2} \nu^*_{ij}(\alpha^*_{ij} + \sum_{g=K+1}^{L_i} \lambda^*_{igj}\alpha^*_{ig})$ in (8) measures the sum of measurement-error biases in the coefficients of $x_{ij} \in S_2$ and the formula $(-\frac{\nu^*_{ij}}{x_{ij}})(\alpha^*_{ij} + \sum_{g=K+1}^{L_i} \lambda^*_{igj}\alpha^*_{ig})$ in the first line of (9) measures such a bias in the coefficient γ_{ij} of $x_{ij} \in S_1$.

Under our approach, measurement errors do not become random variables until distributional assumptions are made about them.

2.2.5 What specification errors are (3)-(8) free from? (i) The unknown functional form of (1) is not allowed to become the source of a specification error in (3). (ii) The uniqueness of the coefficients and error term of (5) eliminates the specification error resulting from non-unique coefficients and error term. (iii) Pratt and Schlaifer (1988, p. 34) pointed out that the requirement that the included regressors be independent of the excluded regressors themselves is "meaningless". The specification error introduced by making this meaningless assumption is avoided by taking a correct function of certain 'sufficient sets' of omitted regressors as the error term of (5). (iv) The specification error of ignoring measurement errors when they are present is avoided by placing them at the *appropriate* places in (5). It should be noted that when we affirm that (7) is free of specification errors, we mean that it is free of specification-errors (i)-(iv). Using (3)-(6) we have derived a real-world relationship in (7) that is free of specification-errors (i)-(iv). Thus, our approach affirms that any relationship suffering from anyone of these specification errors is definitely not a real-world relationship.

2.3 Comparison of (7) with the Yatchew and Griliches (1985), Wooldridge (2002), and Cramer (2006/07) Latent Regression Models

In Section 2.2.3, we have seen that the relationships between each omitted regressor and the included regressors in (4) introduce omitted-regressor biases into the coefficients on the regressors of (5). We have pointed out in the last paragraph of that section that this is not how Yatchew and Griliches **(**YG) derived omitted-regressor biases. They work with models in which the coefficients and error terms do not satisfy our definition of uniqueness. YG considered a simple binary choice model and omitted one of its two included regressors. According to YG, this omission introduces omitted regressor bias into the coefficient on the regressor that is allowed to remain. The results proved by YG are: (i) even if the omitted regressor is uncorrelated

with the remaining included regressor, the coefficient on the latter regressor will be inconsistent. (ii) If the errors in the underlying regression are heteroscedastic, then the maximum likelihood estimators that ignore the heteroscedasticity are inconsistent and the covariance matrix is inappropriate (see also Greene 2012, p. 713). We do not omit any of the included regressors from (5) to generate omitted-regressor biases. For YG, omitted regressors in (4) are those that generate the error term in their latent regression model. Equation (5)'s error term is a function of those variables that satisfy Pratt and Schlaifer's definition of 'sufficient sets' of our omitted regressors. Thus, YG' concepts of omitted-regressors, included regressors, and error terms are different from ours. Their model is subject to specification errors (i)-(iv) listed in the previous section. YG's assumptions about the error term of their model are questionable because of its non-uniqueness. Unless its coefficients and error term are unique no model can represent any real-world relationship which is unique. According to YG, Wooldridge, and Cramer, the regressors constituting the error term of a latent regression model do not produce omitted-regressors biases. Their omitted-regressor bias is not the same as those in (5). YG's results cannot be obtained from our model (7). Their nonunique heteroscedastic error term is different from our unique heteroscedastic error term in (7). It can be shown that the results of YG arose as a direct consequence of ignoring our omitted-regressor and measurement-error biases in (9). Omitted regressors constituting the YG model's (non-unique) error term also introduce omitted-regressor biases in our sense but not in their sense. Furthermore, the YG model suffers from all the four specification errors (i)-(iv) which equations (3)-(5), (7)-(9) avoid.

To recapitulate, misspecifications of the correct functional form of (1) are avoided by expressing it in the form of equation (3). If the sum, $\sum_{g=K+1}^{L_i} x_{ig}^{*}\alpha_{ig}^{*}$, of the last $L_i - K$ terms on the

right-hand side of (3) is treated as its error term, then this error term is not unique (see Swamy et al. 2014, p. 197). Suppose that the coefficients of (3) are constants. Then the correlation between the nonunique error term and the first *K* regressors of (3) can be made uncertain and certain, at the whim of an arbitrary choice between two observationally equivalent forms of (3), as shown by Swamy et al. 2014, pp. 217-218). To eliminate this difficulty, a model with unique coefficients and error term is derived by substituting the right-hand side of equation (4) for the omitted regressor, x_{ig}^{*}, in (3) for every $g = K+1, \ldots, L_i$. Equation (7) shows how the terms of an equation look like if this equation is made free of specification errors (i)-(iv). For each continuous x_{ij} with $j > 0$ in (7), its coefficient contains the bias-free partial derivative $(\frac{\partial \psi_i(.)}{\partial x_{ij}^{*}})$ and omitted-regressor and measurement-error biases.

2.3.1 Parameterization of model (7): The partial derivative $(\partial y_i^{*} / \partial x_{ij}^{*})$ components of the coefficients (γ_{ij}, $j = 1, \ldots, K$) of (7) are the objects of our estimation. For this purpose, we parameterize (7) using our knowledge of the probability model governing the observations in (7). We assume that for $j = 0, 1, \ldots, K$:

$$\gamma_{ij} = z_{i0}\pi_{j0} + z_{i1}\pi_{j1} + \cdots + z_{ip}\pi_{jp} + \varepsilon_{ij} \tag{10}$$

where $z_{i0} \equiv 1$, the π's are fixed parameters, the z's drive the coefficients of (7) and are, therefore, called "coefficient drivers." These drivers are observed. We will explain below how to select these drivers. The errors (ε_{ij}'s) are included in equation (10) because the $p + 1$ drivers may not be able to explain all variation in γ_{ij}.

Admissible drivers: For $j = 0, 1, \ldots, K$, the vector $Z_i = \{ Z_{i0} \equiv 1, Z_{i1}, ..., Z_{ip} \}'$ in (10) is an admissible set of coefficient drivers if given Z_i, the value that the vector of the coefficients of (7)

would take in unit i had $X_i = \{1, X_{i1},...,X_{iK}\}'$ been $\boldsymbol{x_i} = (1, x_{i1},...,x_{iK})'$ is independent of X_i for all i.[12]

We use the following matrix notation: $\boldsymbol{x_i} = (1, x_{i1},...,x_{iK})'$ is $(K+1)\times 1$, $\boldsymbol{z_i} = (1, z_{i1},...,z_{ip})'$ is $(p+1)\times 1$, $\boldsymbol{\pi}_j' = (\pi_{j0},\pi_{j1},...,\pi_{jp})$ is $1\times(p+1)$, $\gamma_{ij} = \boldsymbol{\pi}_j'\boldsymbol{z_i}$ is a scalar, Π is the $(K+1)\times(p+1)$ matrix having $\boldsymbol{\pi}_j'$ as its jth row, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i0},...,\varepsilon_{iK})'$ is $(K+1)\times 1$. Substituting the right-hand side of (10) for γ_{ij} in (7) gives

$$y_i^* = \boldsymbol{x_i}'\Pi\boldsymbol{z_i} + \boldsymbol{x_i}'\boldsymbol{\varepsilon_i} \tag{11}$$

Assumption I: The regressors of equation (7) are conditionally independent of their coefficients given the coefficient drivers.

Assumption II: For all i, let $g(\boldsymbol{x_i},\boldsymbol{z_i})$ be a Borel function of $(\boldsymbol{x_i},\boldsymbol{z_i})$, $E|y_i^*| < \infty$, and $E|\, y_i^* \; g(\boldsymbol{x_i},\boldsymbol{z_i})| < \infty$.

Assumption III: For $i, i' = 1,\ldots,n$, $\boldsymbol{E}(\boldsymbol{\varepsilon_i} \mid \boldsymbol{x_i},\boldsymbol{z_i}) = 0$, $\boldsymbol{E}(\boldsymbol{\varepsilon_i}\boldsymbol{\varepsilon_i}' \mid \boldsymbol{x_i},\boldsymbol{z_i}) = \boldsymbol{\sigma_\varepsilon^2}\Delta_{\boldsymbol{\varepsilon}}$, and $\boldsymbol{E}(\boldsymbol{\varepsilon_i}\boldsymbol{\varepsilon_{i'}}' \mid \boldsymbol{x_i},\boldsymbol{z_i}) = 0$ if $i \neq i'$.

In terms of homoscedastic error term, equation (11) can be written as

$$y_i^* / \sigma_\varepsilon\sqrt{\boldsymbol{x_i}'\Delta_\varepsilon\boldsymbol{x_i}} = \boldsymbol{x_i}'\Pi\boldsymbol{z_i} / \sigma_\varepsilon\sqrt{\boldsymbol{x_i}'\Delta_\varepsilon\boldsymbol{x_i}} + \boldsymbol{x_i}'\varepsilon_i / \sigma_\varepsilon\sqrt{\boldsymbol{x_i}'\Delta_\varepsilon\boldsymbol{x_i}} \tag{12}$$

where Δ_ε is positive definite.

[12] A similar admissibility condition for covariates is given in Pearl (2000, p. 79). Pearl (2000, p. 99) also gives an equation that forms a connection between the opaque phrase "the value that the coefficient vector of (3) would take in unit i, had $\boldsymbol{X_i} = (X_{i1},...,X_{iK})'$ been $\boldsymbol{x_i} = (x_{i1},...,x_{iK})'$" and the physical processes that transfer changes in $\boldsymbol{X_i}$ into changes in y_i^*.

Under Assumptions I, II, and III, the conditional expectation

$$E(y_i^* \mid \boldsymbol{x}_i, \boldsymbol{z}_i) = \boldsymbol{x}_i' \Pi \boldsymbol{z}_i \qquad (13)$$

exists (see Rao 1973, p. 97).

2.4 Derivation of the Likelihood Function for (11)

The parameters of model (11) to be estimated are Π and $\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^2 \Delta_{\boldsymbol{\varepsilon}}$. Due to the lack of observations on the dependent variable y_i^* not all of these parameters are identified. Therefore, we need to impose some restrictions. The following two restrictions are imposed on model (11):

(i) The σ_ε^2 in $\sigma_\varepsilon^2 \Delta_\varepsilon$ cannot be estimated, since there is no information about it in the data. To solve this problem, we set σ_ε^2 equal to 1.

(ii) From (2) it follows that the conditional probability that $y_i = 1$ (or $y_i^* > 0$) given $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ is

$$\text{Prob}(y_i^* > 0 \mid \boldsymbol{x}_i, \boldsymbol{z}_i) = \text{Prob}(\boldsymbol{x}_i' \boldsymbol{\varepsilon}_i > - \boldsymbol{x}_i' \Pi \boldsymbol{z}_i \mid \boldsymbol{x}_i, \boldsymbol{z}_i) \qquad (14)$$

where the information about the constant term is contained in the proportion of observations for which the dependent variable is equal to 1.

For symmetric distributions like normal,

$$\text{Prob}(y_i^* > 0 \mid \boldsymbol{x}_i, \boldsymbol{z}_i) = \text{Prob}(\boldsymbol{x}_i' \boldsymbol{\varepsilon}_i < \boldsymbol{x}_i' \Pi \boldsymbol{z}_i \mid \boldsymbol{x}_i, \boldsymbol{z}_i) = F(\boldsymbol{x}_i' \Pi \boldsymbol{z}_i / \sqrt{\boldsymbol{x}_i' \Delta_\varepsilon \boldsymbol{x}_i} \mid \boldsymbol{x}_i, \boldsymbol{z}_i) \qquad (15)$$

where $F(.|.)$ is the conditional distribution function of $\boldsymbol{x}_i' \boldsymbol{\varepsilon}_i$. The conditional probability that $y_i = 0$ (or $y_i^* \leq 0$) given $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ is $1 - F(\boldsymbol{x}_i' \Pi \boldsymbol{z}_i / \sqrt{\boldsymbol{x}_i' \Delta_\varepsilon \boldsymbol{x}_i} \mid \boldsymbol{x}_i, \boldsymbol{z}_i)$. The conditional probability that $y_i = 1$ (or $y_i^* > 0$) given $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ is $F(\boldsymbol{x}_i' \Pi \boldsymbol{z}_i / \sqrt{\boldsymbol{x}_i' \Delta_\varepsilon \boldsymbol{x}_i} \mid \boldsymbol{x}_i, \boldsymbol{z}_i)$. $F(.|.)$ in (15) denotes the

conditional normal distribution function of the random variable $\boldsymbol{x}_i'\boldsymbol{\varepsilon}_i / \sqrt{\boldsymbol{x}_i'\Delta_\varepsilon \boldsymbol{x}_i}$ with mean zero and unit variance; $\boldsymbol{f}_i$ is the density function of the standard normal. Let $\boldsymbol{\delta}_\varepsilon$ be the column stack of Δ_ε. To exploit the symmetry property of Δ_ε, we add together the two elements of $(\boldsymbol{x}_i' \otimes \boldsymbol{x}_i')$ corresponding to the (j, j') and (j', j) elements of Δ_ε in $\boldsymbol{x}_i'\Delta_\varepsilon \boldsymbol{x}_i$ and eliminate the (j', j) element of Δ_ε from $\boldsymbol{\delta}_\varepsilon$ for j = 0, 1, …, K. These operations change the $(1 \times (K+1)^2)$ vector $(\boldsymbol{x}_i' \otimes \boldsymbol{x}_i')$ to the $(1 \times (K+1)(K+2)/2)$ vector, denoted by $(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})$, and change the $(K+1)^2 \times 1$ vector $\boldsymbol{\delta}_\varepsilon$ to the $[(K+1)(K+2)/2] \times 1$ vector, denoted by $\overline{\boldsymbol{\delta}}_\varepsilon$.

The maximum likelihood (ML) method is used to estimate the elements of Π and Δ_ε. To do so, each observation is treated as a single draw from a binomial distribution. The model with success probability $F(\boldsymbol{x}_i'\Pi\boldsymbol{z}_i / \sqrt{\boldsymbol{x}_i'\Delta_\varepsilon \boldsymbol{x}_i} \mid \boldsymbol{x}_i, \boldsymbol{z}_i)$ and independent observations leads to the likelihood function,

$$\text{Prob}(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n \mid \boldsymbol{x}_i, \boldsymbol{z}_i) = \prod_{y=0} [1 - F(\boldsymbol{x}_i'\Pi\boldsymbol{z}_i / \sqrt{\boldsymbol{x}_i'\Delta_\varepsilon \boldsymbol{x}_i} \mid \boldsymbol{x}_i, \boldsymbol{z}_i)] \prod_{y=1} F(\boldsymbol{x}_i'\Pi\boldsymbol{z}_i / \sqrt{\boldsymbol{x}_i'\Delta_\varepsilon \boldsymbol{x}_i} \mid \boldsymbol{x}_i, \boldsymbol{z}_i) \quad (16)$$

The likelihood function for a sample of n observations can be written as

$$L(\Pi, \Delta_\varepsilon \mid \text{data}) = \prod_{i=1}^{n} [F(\boldsymbol{x}_i'\Pi\boldsymbol{z}_i / \sqrt{\boldsymbol{x}_i'\Delta_\varepsilon \boldsymbol{x}_i} \mid \boldsymbol{x}_i, \boldsymbol{z}_i)]^{y_i} [1 - F(\boldsymbol{x}_i'\Pi\boldsymbol{z}_i / \sqrt{\boldsymbol{x}_i'\Delta_\varepsilon \boldsymbol{x}_i} \mid \boldsymbol{x}_i, \boldsymbol{z}_i)]^{1-y_i} \quad (17)$$

This equation gives

$$\ln L = \sum_{i=1}^{n}\left\{ y_i \ln \boldsymbol{F}\left(\frac{(\boldsymbol{z}_i' \otimes \boldsymbol{x}_i')\boldsymbol{\pi}^{Long}}{\sqrt{(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})\,\bar{\boldsymbol{\delta}}_{\varepsilon}}} \right) + (1\text{-}y_i)\ln\left[1 - \boldsymbol{F}\left(\frac{(\boldsymbol{z}_i' \otimes \boldsymbol{x}_i')\boldsymbol{\pi}^{Long}}{\sqrt{(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})\,\bar{\boldsymbol{\delta}}_{\varepsilon}}} \right) \right] \right\} \tag{18}$$

where $\otimes$ is a Kronecker product and $\boldsymbol{\pi}^{Long}$ is the column stack of Π.

2.4.1 Unconstrained and constrained maximum likelihood estimation: In the case where Δ_{ε} is identified, then its positive definite estimate may not be obtained unless the log likelihood function in (18) is maximized subject to the restriction that Δ_{ε} is positive definite. Furthermore, these constrained estimates of $\boldsymbol{\pi}^{Long}$ and $\bar{\boldsymbol{\delta}}_{\varepsilon}$ do not satisfy the following likelihood equations.

$$\frac{\partial \ln L}{\partial \boldsymbol{\pi}^{Long}} = \sum_{i=1}^{n}\left[\frac{f_i(y_i - F_i)}{F_i(1 - F_i)} \right] \frac{(\boldsymbol{z}_i \otimes \boldsymbol{x}_i)}{\sqrt{(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})\,\bar{\boldsymbol{\delta}}_{\varepsilon}}} = 0 \tag{19}$$

$$\frac{\partial \ln L}{\partial \bar{\boldsymbol{\delta}}_{\varepsilon}} = \sum_{i=1}^{n}\left[\frac{\boldsymbol{f}_i(y_i - \boldsymbol{F}_i)}{\boldsymbol{F}_i(1 - \boldsymbol{F}_i)} \right](\boldsymbol{z}_i' \otimes \boldsymbol{x}_i')\boldsymbol{\pi}^{Long}(-\frac{1}{2})\left[\frac{1}{(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})\bar{\boldsymbol{\delta}}_{\varepsilon}} \right]^{\frac{3}{2}} (\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})' = 0 \tag{20}$$

where F_i stands for $\boldsymbol{F}\left(\frac{(\boldsymbol{z}_i' \otimes \boldsymbol{x}_i')\boldsymbol{\pi}^{Long}}{\sqrt{(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})\,\bar{\boldsymbol{\delta}}_{\varepsilon}}} \right)$ and f_i is the derivative of F_i.

We now show that Δ_{ε} is not identified. The log likelihood function in (18) has the property that it does not change when $\boldsymbol{\pi}^{Long}$ is multiplied by a positive constant κ and $\bar{\boldsymbol{\delta}}_{\varepsilon}$ inside the square root by κ^2. This can be seen clearly from

$$\ln L = \sum_{i=1}^{n}\left\{ y_i \ln \boldsymbol{F}\left(\frac{(\boldsymbol{z}_i' \otimes \boldsymbol{x}_i')(\kappa\boldsymbol{\pi}^{Long})}{\sqrt{(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})(\kappa^2\bar{\boldsymbol{\delta}}_{\varepsilon})}} \right) + (1\text{-}y_i)\ln\left[1 - \boldsymbol{F}\left(\frac{(\boldsymbol{z}_i' \otimes \boldsymbol{x}_i')(\kappa\boldsymbol{\pi}^{Long})}{\sqrt{(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})(\kappa^2\bar{\boldsymbol{\delta}}_{\varepsilon})}} \right) \right] \right\} \tag{21}$$

An implication of this property is that if ln L in (18) attains a maximum value at $(\hat{\boldsymbol{\pi}}^{Long\prime}, \hat{\bar{\boldsymbol{\delta}}}_{\boldsymbol{\varepsilon}}')'$, then $(\hat{\boldsymbol{\pi}}^{Long\prime}\ \kappa\ ,\ \hat{\bar{\boldsymbol{\delta}}}_{\boldsymbol{\varepsilon}}'\ \kappa^2\ )'$ yields another point at which ln L attains its maximum value. Consequently, solving equations (19) and (20) for $\boldsymbol{\pi}^{Long}$ and $\bar{\boldsymbol{\delta}}_{\boldsymbol{\varepsilon}}$ gives an infinity of solutions, respectively. None of these solutions is consistent because Δ_ε is not identified. For this reason we set $\Delta_\varepsilon = I$. After inserting this value in equation (19), it is solved for $\boldsymbol{\pi}^{Long}$. This solution is taken as the maximum likelihood estimate of $\boldsymbol{\pi}^{Long}$.

The information matrix, denoted by $I(\boldsymbol{\pi}^{Long})$, is

$$E\left[-\frac{\partial^2 \ln \boldsymbol{L}}{\partial \boldsymbol{\pi}^{Long}\partial(\boldsymbol{\pi}^{Long})'}\right] \tag{22}$$

where the elements of this matrix are given in equation (A8) with $\Delta_\varepsilon = I$.

Suppose that $\Delta_\varepsilon = I$. Then the positive definiteness of (22) is a necessary condition for $\boldsymbol{\pi}^{Long}$ to be identifiable on the basis of the observed variables in (17). If the likelihood equations in (19) have a unique solution, then the inverse of the information matrix in (22) will give the covariance matrix of the limiting distribution of the ML estimator of $\boldsymbol{\pi}^{Long}$. Suppose that the solution of (19) is not unique. In this case, if Lehmann and Casella's (1998, p. 467, (5.5)) method of solving (19) for $\boldsymbol{\pi}^{Long}$ is followed, then the square roots of the diagonal elements of (22) when evaluated at Lehmann and Casella's solutions of (19), give the large sample standard errors of the estimate of $\boldsymbol{\pi}^{Long}$.

2.5 Estimation of the Components of the Coefficients of (7)

The estimates of the coefficients of (7) are obtained by replacing the π's and ε_{ij} of (10) by their maximum likelihood estimates and the mean value zero, respectively. We do not get the correct estimates of the components of γ_{ij} in (9) from its estimate unless its two different functional forms in (10) and (8) and (9) are reconciled. For a continuous x_{ij} with $j > 0$, we recognize that its coefficient γ_{ij} in (9) and γ_{ij} in (10) are the same. Therefore, the sum $z_{i0}\pi_{j0} + z_{i1}\pi_{j1} + \cdots + z_{ip}\pi_{jp} + \varepsilon_{ij}$ in (10) is equal to the function $(1 - \frac{v^*_{ij}}{x_{ij}})(\alpha^*_{ij} + \sum_{g=K+1}^{L_i} \lambda^*_{igj}\alpha^*_{ig})$ in (9). We have already shown that α^*_{ij} is equal to $\frac{\partial \psi_i(.)}{\partial x^*_{ij}}$ and the sum of omitted-regressor and measurement-error biases (ORMEB) is equal to $\{ \sum_{g=K+1}^{L_i} \lambda^*_{igj}\alpha^*_{ig} - \frac{v^*_{ij}}{x_{ij}} (\alpha^*_{ij} + \sum_{g=K+1}^{L_i} \lambda^*_{igj}\alpha^*_{ig})\}$.

Equation (9) has the form

$$\gamma_{ij} = (1 - \mathrm{D}_{ij})(\mathrm{A}_{ij} + \mathrm{B}_{ij}) \tag{23}$$

where $\mathrm{D}_{ij} = (\frac{v^*_{ij}}{x_{ij}})$, $\mathrm{A}_{ij} = \alpha^*_{ij}$, $\mathrm{B}_{ij} = \sum_{g=K+1}^{L_i} \lambda^*_{igj}\alpha^*_{ig}$.

Equations (9) and (10) imply that

$$\hat{\pi}_{j0} + \sum_{h=1}^{p} z_{ih}\hat{\pi}_{jh} = (1 - \hat{\mathrm{D}}_{ij})(\hat{\mathrm{A}}_{ij} + \hat{\mathrm{B}}_{ij}) \tag{24}$$

where the $\hat{\pi}$'s are the ML estimates of the π's derived in Section 2.4.1. We do not know how to predict ε_{ij} and, therefore, we set it equal to its mean value which is equal to zero. Equation (24) reconciles the discrepancies between the functional forms of (9) and (10). We have the ML

estimates of all the unknown parameters on the left-hand side of equation (24). From these estimates, it can be determined that for individual *i* and regressors x_{ij}, $j = 1,\ldots K$:

The estimate of the partial derivative $(\alpha_{ij}^{*}) = \hat{\mathrm{A}}_{ij} = (1 - \hat{\mathrm{D}}_{ij})^{-1} (\hat{\pi}_{j0} + \sum_{h \in G_1} z_{ih}\hat{\pi}_{jh})$ (25)

The estimate of omitted-regressor bias $(\sum_{g=K+1}^{L_i} \lambda_{igj}^{*}\alpha_{ig}^{*}) = \hat{\mathrm{B}}_{ij} = (1 - \hat{\mathrm{D}}_{ij})^{-1} (\sum_{h \in G_2} z_{ih}\hat{\pi}_{jh})$ (26)

The estimate of measurement-error bias $[-(\frac{\nu_{ij}^{*}}{x_{ij}})(\alpha_{ij}^{*} + \sum_{g=K+1}^{L_i} \lambda_{igj}^{*}\alpha_{ig}^{*})] = -\hat{\mathrm{D}}_{ij}(\hat{\mathrm{A}}_{ij} + \hat{\mathrm{B}}_{ij})$ (27)

where the $p + 1$ coefficient drivers are allocated either to a group, denoted by G_1, or to a group, denoted by G_2; $G_2 = p + 1 - G_1$. The unknowns in formulas (25)-(27) are $\hat{\mathrm{D}}_{ij}$ and G_1. We discuss how to determine these unknowns below.

The type of data Greene (2012, pp. 244-246, Example 8.9) used can tell us about $\hat{\mathrm{D}}_{ij}$. Which of the terms in $\hat{\pi}_{j0} + \sum_{h=1}^{p} z_{ih}\hat{\pi}_{jh}$, should go into $(\hat{\pi}_{j0} + \sum_{h \in G_1} z_{ih}\hat{\pi}_{jh})$, can be decided after examining the sign and magnitude of each term in $\hat{\pi}_{j0} + \sum_{h=1}^{p} z_{ih}\hat{\pi}_{jh}$. If we are not sure of any particular value of G_1, then we can present the estimated kernel functions for α_{ij}^{*}, $i = 1, \ldots n$, for various values of G_1.

Regarding D_{ij} we can make the following assumption:

Assumption IV: For all *i* and *j*: (i) The measurement error v_{ij}^{*} forms a negligible proportion $(\frac{v_{ij}^{*}}{x_{ij}})$ of x_{ij}.

Alternatively, the percentage point $(\frac{v_{ij}^{*}}{x_{ij}}) \times 100$ can be specified if we have the type of data Greene (2012, pp. 244-246, Example 8.9) had. If such data are not available, then we can make Assumption IV. Under this assumption, $(1-\hat{\mathrm{D}}_{ij})^{-1}$ in (25) and (26) gets equated to 1 and the number of unknown quantities in formulas (25) and (26) is reduced to 1.

Under these assumptions, we can obtain the estimates of $\hat{\mathrm{A}}_{ij}$ and their standard errors. These standard errors are based on those of $\hat{\pi}$'s involved in $\hat{\mathrm{A}}_{ij}$. If the estimate of A_{ij} given by formula (25) is accurate, then our estimate of the partial derivative α_{ij}^{*} is free of omitted-regressor and measurement-error biases, and also of specification errors (i)-(iv) listed in Section 2.2.5.

2.5.1 How to select the regressors and coefficient drivers appearing in (11)? The choice of the dependent variable and regressors to be included in (7) is entirely dictated by the partial derivatives we want to learn. The learning of a partial derivative, say $\alpha_{ij}^{*} = \frac{\partial y_i^{*}}{\partial x_{ij}^{*}}$, requires (i) the use of y_i^{*} and x_{ij}^{*} as the dependent variable and a regressor of (7), respectively, (ii) the use of z's in (25) and (26) as the coefficient drivers in (10), and (iii) the use of the values of G_1 and D_{ij} in (25). These requirements show that the learning about one partial derivative is more straightforward than learning about more than one partial derivative. Therefore, in our practical

work we will include in our basic model (7) only one non-constant regressor besides the intercept.

It should be remembered that the coefficient drivers in (10) are different from the regressors in (7). There are also certain requirements that the coefficient drivers should satisfy. They explain variations in the components of the coefficients of (7), as is clear from equations (25) and (26). After deciding that we want to learn about $\alpha_{ij}^{*} = \frac{\partial y_i^{*}}{\partial x_{ij}^{*}}$ and knowing from (23) that this α_{ij}^{*} is only a part of the coefficient γ_{ij} of the regressor x_{ij}^{*} in the (y_i^{*}, x_{ij}^{*})-relationship, we need to include in (10) those coefficient drivers that facilitate accurate evaluation of the formulas (25)-(27). Initially, we do not know what such coefficient drivers are. We have decided to use as coefficient drivers those variables that economists include in their models of the (y_i^{*}, x_{ij}^{*})-relationship as additional explanatory variables. Specifically, instead of using them as additional regressors we use them as coefficient drivers in (10).[13, 14] It follows from equations (25) and (26) that among all the coefficient drivers included in (10) there should be one subset of G_1 coefficient drivers that is highly correlated with the bias-free partial derivative part and another subset of G_2 coefficient drivers that is highly correlated with the omitted-regressor bias of the *j*th coefficient of (7).[15]

[13] We illustrate this procedure in Section 3 below.

[14] Pratt and Schlaifer (1988) consider what they call "concomitants" that absorb 'proxy effects' and include them as additional regressors in their model. The result in (9) calls for equation (10) which justifies our label for its right-hand side variables.

[15] An important difference between coefficient drivers and instrumental variables is that a valid instrument is one that is uncorrelated with the error term, which often proves difficult to find, particularly when the error term is nonunique. For a valid driver we need variables which should satisfy equations (25) and (26). On the problems with instrumental variables, see Swamy, Tavlas, and Hall (2015).

If G_1 and D_{ij} are unknown, as they usually are, then we should make alternative assumptions about them and compare the results obtained under these alternative assumptions.

2.5.2 Impure marginal effects: The marginal effect of any one of the included regressors on the probability that $y_i = 1$ is

$$\frac{\partial \text{Prob}(y_i = 1 \mid x, z_1)}{\partial x_i} = f_i\left(x_i'\Pi z_i / \sqrt{x_i'\Delta_\varepsilon x_i} \mid x_i, z_i \right) \left(\frac{z_i'\Pi'}{(x_i'\Delta_\varepsilon x_i)^{(1/2)}} - \frac{x_i'\Pi z_i (\Delta_\varepsilon x_i)}{(x_i'\Delta_\varepsilon x_i)^{(3/2)}} \right) \tag{28}$$

where we set $\Delta_\varepsilon = I$.

These effects are impure because they involve omitted-regressor and measurement-error biases. It is not easy to integrate omitted-regressor and measurement-error biases out of the probability in (15).[16]

3. Earnings and Education Relationship

This section is designed to give some specific empirical examples on the type of misspecification that usually found in actual data sets. Several authors studied this relationship. We are also interested in learning about the partial derivative of earnings with respect to education of individuals. For this purpose, we set up the model

$$y_i^* = \gamma_{i0} + x_{i1}^* \gamma_{i1} \tag{29}$$

where y_i^* denotes unobserved earnings, x_{i1}^* denotes unobserved education, and the components of γ_{i0} and γ_{i1} are given in (8) and (9), respectively. Equation (29) is derived in the same way

[16] These biases are not involved in Wooldridge's marginal effects because according to that researcher omitted regressors constituting his model's error term do not introduce omitted-regressor biases into the coefficients of the included regressors.

that (7) is derived. Like (7), equation (29) is devoid of four specification errors. In our empirical work, we use x_{i1} = years of schooling as a proxy for education.

Greene (2012, p. 14) used model (29) after changing it to a fixed coefficient model with added error term, in which the dependent variable is the log of earnings (hourly wage times hours worked). He pointed out that this model neglects the fact that most people have higher incomes when they are older than when they are young, regardless of their education. Therefore, Greene argued that the coefficient on education will overstate the marginal impact of education on earnings. He further pointed out that if age and education are positively correlated, then his regression model will associate all the observed increases in income with increases in education. Greene concluded that a better specification would account for the effect of age. He also pointed out that income tends to rise less rapidly in the later earning years than in the early ones. To accommodate this phenomena and the age effect, Greene (2012, p. 14) included in his model the variables age and age^2.

Recognizing the difficulties in measuring education pointed out by Greene (2012, p. 221), we measure education as hours of schooling plus measurement error. Another problem Greene discussed is that of the endogeneity of education. We handle this problem by making Assumptions II and III. Under these assumptions, the conditional expectation $\mathrm{E}(y_i^* \mid \boldsymbol{x_i}, \boldsymbol{z_i}) = \boldsymbol{x_i'}\Pi\boldsymbol{z_i}$ exists. Other regressors Greene (2012, p. 708) included in his labor supply model include kids, husband's age, husband's education and family income.

The question that arises is the following: How should we handle the variables mentioned in the previous two paragraphs? Researchers who studied earnings-education relationship have often included these variables as additional explanatory variables in earnings and education

equation with fixed coefficients. Greene (2012, p. p. 699) also included the interaction between age and education as an additional explanatory variable. Previous studies, however, have dealt with fixed coefficient models and did not have anything to do with VCs of the type in (29). The coefficients of the earnings-education relationship in (29) have unwanted omitted-regressor and measurement-error biases as their portions. We need to separate them from the corresponding partial derivatives, as shown in (7) and (9). How do we preform this separation? Based on the above derivation in (1)-(9), we use the variables identified in the previous two paragraphs as the coefficient drivers.

When these coefficient drivers are included, the following two equations get added to equation (29):

$$\gamma_{ij} = z_{i0}\pi_{j0} + z_{i1}\pi_{j1} + \cdots + z_{i6}\pi_{j6} + \varepsilon_{ij} \quad (j = 0, 1) \tag{30}$$

where $z_{i0} = 1$ for all i, z_{i1} = Wife's Age, z_{i2} = Wife's Age^2, z_{i3} = Kids, z_{i4} = Husband's age, z_{i5} = Husband's education, and z_{i6} = Family income.

It can be seen from (11) that equation (30) with $j = 0$ makes the coefficient drivers act as additional regressors in (29) and equation (30) with $j = 1$ introduces the interactions between education and each of the coefficient drivers. Greene (2012, p. 699) informed us that binary choice models with interaction terms received considerable attention in recent applications. Note that for $j = 1$, $h = 1, \ldots, 6$, π_{jh} should not be equated to $\frac{\partial^2 y_i^*}{\partial x_{i1}^* \partial z_{ih}}$ because γ_{i1} is not equal to $\frac{\partial y_i^*}{\partial x_{i1}^*}$.

Appendix Table F5.1 of Greene (2012) contains 753 observations used in the Mroz study of the labor supply behavior of married women. We use these data in this section. Of the 753 married women in the sample, 428 were participants and the remaining 325 were nonparticipants

in the formal labor market. This means that y_i = 1 for 428 observations and y_i = 0 for 325 observations. The data on x_{i1} and the z's for these 753 married women are obtained from Greene's Appendix Table F5.1. Using these data and applying an iteratively rescaled generalized least squares method to (29) and (30) we obtain

$$\hat{\gamma}_{i0} = \underset{(81.0820)}{27.6573} + \underset{(3.9504)}{0.1316}\, z_{i1} - \underset{(0..0441)}{0.0049}\, z_{i2} - \underset{(7.0849)}{11.9494}\, z_{i3} - \underset{(0.6261)}{0.4414}\, z_{i4} - \underset{(0.7910)}{1.4708}\, z_{i5} + \underset{(0.0002)}{0.0003}\, z_{i6} \qquad (31)$$

$$\hat{\gamma}_{i1} = - \underset{(6.9397)}{4.2328} + \underset{(0.3405)}{0.1696}\, z_{i1} - \underset{(0.0038)}{0.0019}\, z_{i2} + \underset{(0.6076)}{0.5168}\, z_{i3} + \underset{(0.0550)}{0.0261}\, z_{i4} + \underset{(0.0676)}{0.0702}\, z_{i5} - \underset{(0.000019)}{0.000013}\, z_{i6} \qquad (32)$$

Table 1 Estimates of γ_{i0} and γ_{i1} with their Standard Errors for Five Married Women[17]

$\hat{\gamma}_{i0}$ (standard error)	$\hat{\gamma}_{i1}$ (standard error)
-13.280 (6.5787)	1.2811 (0.5535)
-5.2539 (9.1608)	0.7930 (0.7830)
-15.221 (4.9202)	1.5029 (0.4197)
-21.577 (11.213)	1.8393 (0.9942)
-9.2086 (7.8416)	1.0386 (0.6504)

From (23) we obtain

$$\hat{\gamma}_{i1} = (1 - \hat{D}_{i1})(\hat{A}_{i1} + \hat{B}_{i1}) \qquad (33)$$

Our interest is in the partial derivative $\hat{A}_{i1} = \alpha^*_{i1} = \dfrac{\partial y^*_i}{\partial x^*_{i1}}$ which is the bias-free portion of $\hat{\gamma}_{i1}$. This partial derivative measures the "impact" of the *i*th married woman's education on her earnings.

[17] The standard errors of estimates are given in parentheses below the estimates for five married women. The estimates and their standard errors for other married women are available from the authors upon request.

Our prior belief is that the right sign for this bias-free portion is positive. Now it is appropriate to use the formula $(\alpha_{i1}^{*}) = \hat{A}_{i1} = (1-\hat{D}_{i1})^{-1}(\hat{\pi}_{10} + \sum_{h\in G_1} z_{ih}\hat{\pi}_{1h})$ in (25) with $j = 1$ to estimate $\alpha_{i1}^{*} = \frac{\partial y_i^{*}}{\partial x_{i1}^{*}}$. We assume that $\hat{D}_{i1}$ is negligible. We need to choose the terms in the sum $(\hat{\pi}_{10} + \sum_{h\in G_1} z_{ih}\hat{\pi}_{1h})$ from the terms on the right-hand side of equation (32). It can be seen from this equation that if we retain the estimate $\hat{\pi}_{10} = \underset{(6.9397)}{-4.2328}$ in the sum $(\hat{\pi}_{10} + \sum_{h\in G_1} z_{ih}\hat{\pi}_{1h})$, then this sum does not give positive estimate of α_{i1}^{*} for any combination of the six coefficient drivers in (32). Therefore, we remove $\hat{\pi}_{10}$ from $(\hat{\pi}_{10} + \sum_{h\in G_1} z_{ih}\hat{\pi}_{1h})$. We expect the impact of education on earnings to be small. To obtain the smallest possible positive estimate of α_{i1}^{*}, we choose the smallest positive term on the right-hand side of (32). This term is $+\underset{(0.0676)}{0.0702}\, z_{i5}$. Hence we set the z's other than z_{i5} in $\sum_{h\in G_1} z_{ih}\hat{\pi}_{1h}$ equal to zero. Thus, we obtain $G_1 = 1$ and $\hat{\alpha}_{i1}^{*} = +\underset{(0.0676)}{0.0702}\, z_{i5}$. The value of z_{i5} times 0.0702 gives the estimate of the impact of the ith married woman's education on her earnings.

Table 2 Estimates of the Bias-free Portion of γ_{i1} for Five Married Women[18]

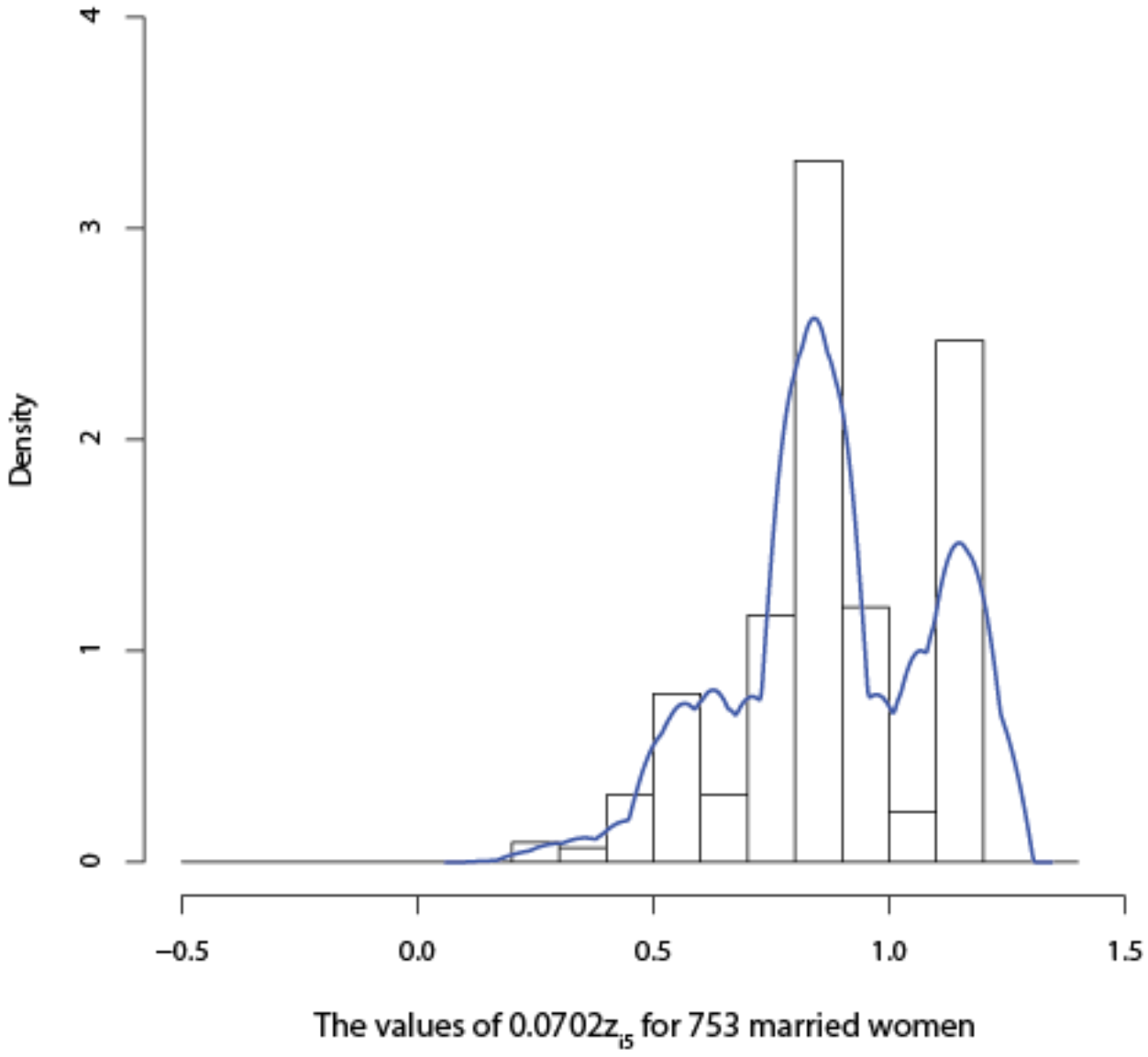
$0.0702z_{i5}$
0.8419 (0.8110)
0.6314 (0.6083)
0.8419 (0.8110)
0.7016 (0.6758)
0.8419 (0.8110)

[18] The standard errors of estimates are given in parentheses below the estimates for five married women. These estimates and standards errors for other married women are available from I-Lok Chang upon request.

To conserve space, we present the values of $\hat{\alpha}_{i1}^{*} = +\underset{(0.0676)}{0.0702}\, z_{i5}$ only for only i = 1, …, 5 in Table 2. The impact estimates for all 753 married women are presented in the form of a histogram or a kernel density function in Figure 1 below. We interpret the estimate $\underset{(0.0676)}{0.0702}\, z_{i5}$ to imply that an additional year of schooling is associated with a $\underset{(0.0676)}{0.0702}\, z_{i5} \times 100$ percent increase in earnings. This impact of education on earnings is different for different married women. The impact of a wife's education on her earnings is 0.0702 times her husband's education.[19] Our results in Table 2 and Figure 1 below show that the greater are the years of schooling of a husband, the larger is the impact of his wife's education on her earnings. However, the estimates of α_{i1}^{*} appear to be high at least for some married women whose husbands had larger years of schooling. Therefore, they may contain some omitted-regressor biases.

Figure 1: Estimates of the "Impacts" of Education on the Earnings for 753 Married Women

[19] According to Geene (2012, p, 708), it would be natural to assume that all the determinants of a wife's labor force participation would be correlated with the husband's hours which is defined as a linear stochastic function of the husband's age and education and the family income. Our inclusion of husband's variables in (32) is consistent with this assumption.

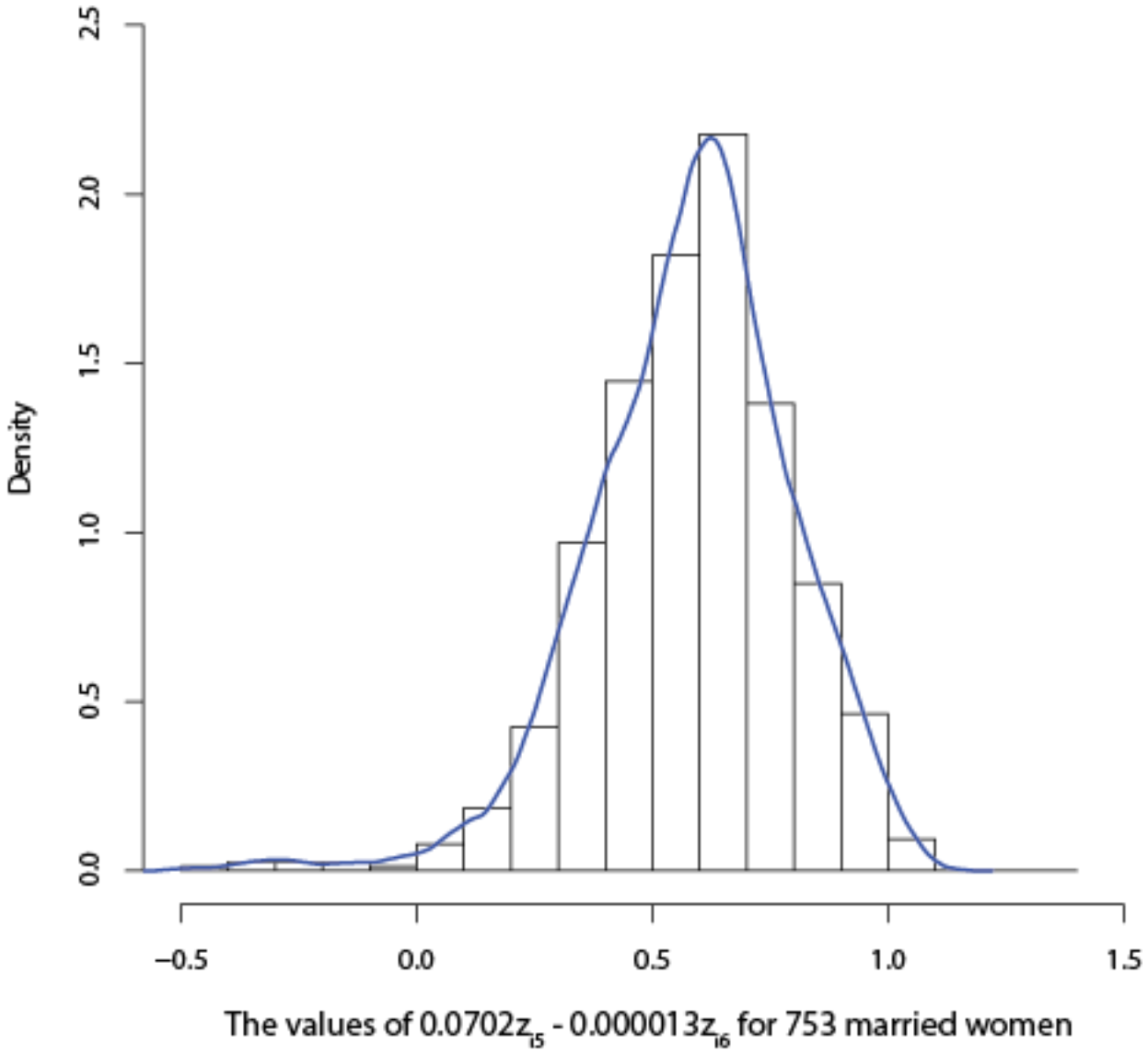


A histogram and a kernel density function presented in Figure 1 are much more revealing than a table containing the values, $0.0702z_{i5}$, $i = 1,\ldots, 753$, and their standard errors would. Also, such a table occupies a lot of space without telling us mush. We are using Figure 1 as a descriptive

device. The kernel function in Figure 1 is multimodal. All the estimates in this figure have the correct signs.

Greene's (2012, p. 708) estimate 0.0904 of the coefficient of education in his estimated labor supply model is not comparable to the estimates in Figure 1 because (i) his model is different from our model, (ii) the dependent variable of his labor supply model in Greene (2012, p.683) is the log of the dependent variable of our model (29), and (iii) our definition of $\frac{\partial y_i^*}{\partial x_{i1}^*}$ in (3) is different from Greene's (2012, p. 14) definition of $\frac{\partial y_i}{\partial x_{i1}}$. Greene's estimate is some kind of average estimate applicable to all 753 married women. It is unreasonable to expect his average estimate to be close to the estimate for each married women. We will now show that given the 6 coefficient drivers in (32), it is not possible to reduce the magnitudes of all the estimates in Figure 1 without changing the positive sign of some of these estimates in the left tail end of Figure 1 to the negative sign. This is what has happened in Figure 2. To reduce the magnitudes of the estimates of bias-free parts ($\hat{A}_{i1} = \alpha_{i1}^* = \frac{\partial y_i^*}{\partial x_{i1}^*}$) of the $\hat{\gamma}_{i1}$'s given in Figure 1 for all *i*, we use the alternative estimates, $0.0702 z_{i5}$ - $0.000013\, z_{i6}$ of the bias-free parts of the $\hat{\gamma}_{i1}$'s called "modified $\hat{A}_{i1}$, $i = 1,\ldots, 753$." The histogram and kernel density function for the modified $\hat{A}_{i1}$ is given in Figure 2 below.

Figure 2: Modified Estimates of the "Impacts" of Education on the Earnings for 753 Married Women



Five of the estimates in the left-tail end of this figure have the wrong (negative) sign. More number of wrong signs will occur if we try to further reduce the magnitudes of the modified estimates. The kernel density function of the modified estimates is unimodal unlike the kernel density function in Figure 1. The range of the modified estimates is smaller than that of the estimates in Figure 1.

From these results it is incorrect to conclude that the conventional discrete choice models and their method of estimation give better and unambiguous results than the latent regression

model in (11) and (14) and formula (25). The reasons for this circumstance are the following: (i) The conventional models including discrete choice models suffer from four specification errors listed in Section 2.2.5 and the model in (11) and (14) is free of these errors; (ii) The conventional latent regression models have nonunique coefficients and error terms and the model in (11) and (14) is based on model (5) which has unique coefficients and error term. How can a model with nonunique coefficients and error term give unambiguous results? (iii) The conventional method of estimating the discrete choice models appears to be simple because these models are based on the assumption that 'the' omitted regressors constituting their error terms do not introduce omitted-regressor biases into the coefficients of their included regressors. The model in (11) and (14) is not based on any such assumption. (iv) Pratt and Schlaifer pointed out that in the conventional model the condition that its regressors be independent of 'the' omitted regressors constituting its error term is meaningless. The error terms of the model in (11) and (14) are not the functions of 'the' omitted-regressors.

4. Conclusions

We have removed four major specification errors from the conventional formulation of probit and logit models. A reformulation of Yatchew and Griliches' probit model so that it is devoid of these specification errors changes their results. We also find that their model has nonunique coefficients and error term. YG make the assumption that omitted regressors constituting the error term of their model do not introduce omitted-regressor biases into the coefficients of the included regressors. We have developed a method of calculating the bias-free partial derivative portions of the coefficients of a correctly specified probit model.

Appendix

In this Appendix, we show that any of the models estimated in the econometric literature is more restrictive than (1). We also show that these restrictions, when imposed on (1), lead to several specification errors.

I. Derivation of Linear and Nonlinear Regressions with Additive Error Terms

I.1 Nonunique Coefficients and Error Terms

I.1.1 Beginning problems – Rigorous derivation of models with additive error terms: It is widely assumed that the error term in an econometric model arises because of omitted regressors influencing the dependent variable. We can use appropriate Felipe and Fisher's (2003) separability and other conditions to separate the included regressors, $x_{i1}^{*},...,x_{iK}^{*}$, from omitted regressors, $x_{i,K+1}^{*},...,x_{iL_i}^{*}$, so that (1) can be written as

$$y_i^{*} = \psi_{i1}(x_{i1}^{*},...,x_{iK}^{*}) + \psi_{i2}(x_{i,K+1}^{*},...,x_{i,L_i}^{*}) = \psi_{i1}(x_{i1}^{*},...,x_{iK}^{*}; \beta_1, ..., \beta_p) + \varepsilon_i \quad \text{(A1)}$$

where $\varepsilon_i = \psi_{i2}(x_{i,K+1}^{*},...,x_{i,L_i}^{*})$ is a function of omitted regressors. Let ε_i be the random error term and let $\psi_{i1}(x_{i1}^{*},...,x_{iK}^{*})$ be equal to $\psi_{i1}(x_{i1}^{*},...,x_{iK}^{*}; \beta_1, ..., \beta_p)$ which is an unknown function of $x_{i1}^{*},...,x_{iK}^{*}$.[20] Let $\beta_1, ..., \beta_p$ be the fixed parameters representing the constant features of model (A1). From the above derivation we know what type of conditions which, when imposed on (1), give exactly the model in Greene (2012, p. 181, (7-3)).

[20] Another widely cited work that utilized a set of separability conditions is that of Heckman and Schmierer (2010). These authors postulated a threshold crossing model which assumes separability between observables Z that affect choice and an unobservable V. They used a function of Z as an instrument and used the distribution of V to define a fundamental treatment parameter known as the marginal treatment effect.

The separability conditions used to rewrite (1) in the form of (A1) are very restrictive, as shown by Felipe and Fisher (2003). Furthermore, in his scrutiny of the Rotterdam School demand models, Goldberger (1987) pointed out that the treatment of any features of (1) as constant parameters such as $\beta_1, ..., \beta_p$ may be questioned and these parameters are not unique.[21] Use of non-unique parameters is a specification error. Therefore, the functional form of (A1) is most probably misspecified.

Skyrms (1988, p. 59) made the important point that spurious correlations disappear when we control for all relevant pre-existing conditions.[22] Even though some of the regressors, $x^*_{t,K+1},...,x^*_{t,L_t}$, represent all relevant pre-existing conditions in our formulation of (A1), they cannot be controlled for, as we should to eliminate false (spurious) correlations, since they are included in the error term of (A1). Therefore, in (A1), the correlations between y^*_t and some of $x^*_{t1},...,x^*_{tK}$ can be spurious.

Karlsen, Myklebust and Tjøstheim (KMT) (2007) considered a model of the type (A1) for time series data. They assumed that $\{\varepsilon_t\}$ is an unobserved stationary process and $\{X^*_{t1},...,X^*_{tK}\}$ and $\{Y^*_t\}$ are both observed nonstationary processes and are of unit-root type. White (1980, 1982) also considered (A1) for time-series data and assumed that the ε_t's are serially independent and are distributed with mean zero and constant variance. He also assumed that ε_t is uncorrelated with $\{X^*_{t1},...,X^*_{tK}\}$ for all *t*. Pratt and Schlaifer (1984, 1988) criticized that these assumptions are meaningless because they are about ε_t which is not unique and is

[21] The 'uniqueness' is defined in Section 2.2.3.
[22] We have been using the cross-sectional subscript *i* so far. We change this subscript to the time subscript *t* wherever the topic under discussion requires the use of the latter subscript.

composed of variables of which we know nothing. Any distributional assumption about a nonunique error term is arbitrary.

I.1.2 Full independence and the existence of conditional expectations: Consider (A1) again. Let $X = \psi_{i1}(x_{i1}^*,...,x_{iK}^*)$, $Y = y_i^*$ and $M = \varepsilon_i$, be three random variables. Then X and M are statistically independent if their joint distribution can be expressed as the product of their marginal distributions. It is not possible to verify this condition.

Let $H(X)$ and $K(M)$ be the functions of X and ε_i, respectively. As Whittle (1976) pointed out, we must live with the idea that, for the given random variables like M and X, we may be only able to assert the validity of the condition

$$\mathrm{E}[H(X)K(M)] = \mathrm{E}[H(X)]\mathrm{E}[K(M)] \qquad \text{(A2)}$$

where the functions H and K are such that $\mathrm{E}[H(X)] < \infty$ and $\mathrm{E}[K(M)] < \infty$. If condition (A2) holds only for certain functions, H and K, then we cannot say that X and M are independent. Suppose that equation (A2) holds only for linear K, so that $\mathrm{E}[H(X)M] = \mathrm{E}(M)\mathrm{E}[H(X)]$ for any H for which $\mathrm{E}[H(X)] < \infty$. This equation is equivalent to $\mathrm{E}(\varepsilon_i \mid x) = \mathrm{E}(\varepsilon_i)$ which shows that the disturbance at observation *i* is mean independent of x at *i*. This may be true for all *i* in the sample. This mean independence implies Greene's (2012, p. 183) assumption (3).

Let us now drop condition (A2) and let us assume instead that

$H(X)$ be a Borel function of X, (A2.1)

$\mathrm{E}|Y| < \infty$, (A2.2)

$\mathrm{E}|Y\mathrm{H(X)}| < \infty$. (A2.3)

Using these assumptions, Rao (1973, p. 97) proved that

$$\mathrm{E}[H(X)Y \mid X = x] = H(x)\mathrm{E}(Y \mid x) \qquad \text{(A3)}$$

$$\mathrm{E}\{H(X)[Y - \mathrm{E}(Y \mid X)]\} = 0 \qquad \text{(A4)}$$

Equations (A3) and (A4) prove that under conditions (A2.1)-(A2.3), E($Y|x$) exists.[23]

I.1.3 Linear conditional means and variances: Under the necessary and sufficient conditions of Kagan, Linnik and Rao's (KLR's) (1973, pp. 11-12) lemma (reproduced in Swamy and von zur Muehlen 1988, pp. 114-115), the following two equations hold almost certainly:

$$\mathrm{E}(Y_i^* \mid x_{i1}^*,...,x_{iK}^*) = \sum_{j=0}^{K} x_{ij}^* \beta_j \text{ with } x_{i0}^* \equiv 1 \text{ for } i = 1,\ldots,\mathrm{n} \tag{A5}$$

and

$$\mathrm{Var}(Y_i^* \mid x_{i1}^*,...,x_{iK}^*) = \sigma_\varepsilon^2, \text{ a finite, positive constant for all } i = 1,\ldots,n \tag{A6}$$

If the conditions of KLR's lemma are not satisfied, then (A5) and (A6) are not the correct first and second conditional moments of Y_i^*. The problem is that we cannot know a priori whether or not these conditions are satisfied. The conditions of KLR's Lemma are not satisfied if $\{x_{t1}^*,...,x_{tK}^*\}$ and $\{y_t^*\}$ are integrated series.[24] Furthermore, $\{y_t^*\}$ cannot be made stationary by first differencing it once or more than once because of the nonlinearity of $\psi_{t1}(x_{t1}^*,...,x_{tK}^*)$. In these cases, we can use, as Berenguer-Rico and Gonzalo (2013) do, the concepts of summability, cosummability and balanced relationship to analyze model (A1). Clearly the conditions of KLR's lemma are stronger than White's assumptions which, in turn, are stronger than KMT's (2007) assumptions. It is clear that KMT's assumptions are not always satisfied.

II Derivation of the Information Matrix for (10)

[23] This proof is relevant to Heckman's interpretation that in any of his models, the error term is the deviation of the dependent variable from its conditional expectation (see Heckman and Vytlacil 2005). Conditions (A2.1)-(A2.3) do not always hold and hence this conditional expectation does not always exist.

[24] A nonstationary series is integrated of order *d* if it becomes stationary after being first differenced *d* times (see Greene 2012, p. 943). If $\{y_t^*\}$ in (A1) is a nonstationary series of this type, then it cannot be made stationary by first differencing it once or more than once if $\psi_{t1}(x_{t1}^*,...,x_{tK}^*)$ is nonlinear. Basmann (1988, p. 98) acknowledged that a model representation is not free of the most serious objection, i.e., nonuniqueness, if stationarity producing transformations of its observable dependent variable are used.

Consider the log likelihood function in (18). For this function,

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\pi}^{Long} \partial (\boldsymbol{\pi}^{Long})'} = \frac{\partial}{\partial (\boldsymbol{\pi}^{Long})'} \sum_{i=1}^{n} \left[\frac{y_i \boldsymbol{f}_i}{\boldsymbol{F}_i} - \frac{(1-y_i)\boldsymbol{f}_i}{(1-\boldsymbol{F}_i)} \right] \frac{(\boldsymbol{z}_i \otimes \boldsymbol{x}_i)}{\sqrt{(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})\bar{\boldsymbol{\delta}}_{\varepsilon}}}$$

$$= \sum_{i=1}^{n} \left[y_i \left(\frac{\boldsymbol{f}_i'}{\boldsymbol{F}_i} - \frac{\boldsymbol{f}_i^2}{\boldsymbol{F}_i^2} \right) - (1-y_i) \left(\frac{\boldsymbol{f}_i'}{(1-\boldsymbol{F}_i)} + \frac{\boldsymbol{f}_i^2}{(1-\boldsymbol{F})^2} \right) \right] \frac{(\boldsymbol{z}_i \otimes \boldsymbol{x}_i)(\boldsymbol{z}_i' \otimes \boldsymbol{x}_i')}{(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})\bar{\boldsymbol{\delta}}_{\varepsilon}} \qquad \text{(A7)}$$

where $\boldsymbol{f}_i'$ is the partial derivative of $\boldsymbol{f}_i$ with respect to $\boldsymbol{\pi}^{Long}$.

$\mathrm{E}(y_i) = 1 \times F_i + 0 \times (1 - F_i) = F_i$. Using this result in (A7) gives

$$\mathrm{E}\left[-\frac{\partial^2 \ln L}{\partial \boldsymbol{\pi}^{Long} \partial (\boldsymbol{\pi}^{Long})'} \right] = \sum_{i=1}^{n} \frac{\boldsymbol{f}_i^2}{\boldsymbol{F}_i(1-\boldsymbol{F}_i)} \frac{(\boldsymbol{z}_i \otimes \boldsymbol{x}_i)(\boldsymbol{z}_i' \otimes \boldsymbol{x}_i')}{(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})\bar{\boldsymbol{\delta}}_{\varepsilon}} \qquad \text{(A8)}$$

where the condition that n > (K+1)(p+1) is needed for the matrix on the right-hand side of this equation to be positive definite.

$$\frac{\partial^2 \ln \mathrm{L}}{\partial \boldsymbol{\pi}^{Long} \partial \bar{\boldsymbol{\delta}}_{\varepsilon}'} = \frac{\partial}{\partial \bar{\boldsymbol{\delta}}_{\varepsilon}'} \sum_{i=1}^{n} \frac{(\boldsymbol{z}_i \otimes \boldsymbol{x}_i)}{\sqrt{(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})\bar{\boldsymbol{\delta}}_{\varepsilon}}} \left[\frac{y_i \boldsymbol{f}_i}{\boldsymbol{F}_i} - \frac{(1-y_i)\boldsymbol{f}_i}{(1-\boldsymbol{F}_i)} \right]$$

$$= \sum_{i=1}^{n} \frac{(\boldsymbol{z}_i \otimes \boldsymbol{x}_i)}{\sqrt{(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})\bar{\boldsymbol{\delta}}_{\varepsilon}}} \left[y_i \left(\frac{\boldsymbol{f}_i'}{\boldsymbol{F}_i} - \frac{\boldsymbol{f}_i^2}{\boldsymbol{F}_i^2} \right) - (1-y_i) \left(\frac{\boldsymbol{f}_i'}{(1-\boldsymbol{F}_i)} + \frac{\boldsymbol{f}_i^2}{(1-\boldsymbol{F})^2} \right) \right]$$

$$\times \left(-\frac{1}{2} \right) \frac{(\boldsymbol{z}_i' \otimes \boldsymbol{x}_i')\boldsymbol{\pi}^{Long} (\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})}{((\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})\bar{\boldsymbol{\delta}}_{\varepsilon})^{3/2}} +$$

$$\sum_{i=1}^{n} \left[\frac{y_i \boldsymbol{f}_i}{\boldsymbol{F}_i} - \frac{(1-y_i)\boldsymbol{f}_i}{(1-\boldsymbol{F}_i)} \right] \left(-\frac{1}{2} \right) \frac{(\boldsymbol{z}_i \otimes \boldsymbol{x}_i)(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})}{((\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})\bar{\boldsymbol{\delta}}_{\varepsilon})^{3/2}} \qquad \text{(A9)}$$

$$\mathrm{E}\left[-\frac{\partial^2 \ln \mathrm{L}}{\partial \boldsymbol{\pi}^{Long} \partial \boldsymbol{\delta}_{\varepsilon}'}\right] = \sum_{i=1}^{n} \frac{\boldsymbol{f}_i^2}{\mathrm{F}_i(1-\mathrm{F}_i)} \frac{(\boldsymbol{z}_i \otimes \boldsymbol{x}_i)}{\sqrt{(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})\,\bar{\boldsymbol{\delta}}_{\varepsilon}}} (-\frac{1}{2}) \frac{(\boldsymbol{z}_i' \otimes \boldsymbol{x}_i')\boldsymbol{\pi}^{Long}\,(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})}{((\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})\,\bar{\boldsymbol{\delta}}_{\varepsilon})^{3/2}} \tag{A10}$$

$$\frac{\partial^2 \ln L}{\partial \bar{\boldsymbol{\delta}}_{\varepsilon} \partial \bar{\boldsymbol{\delta}}_{\varepsilon}'} = \frac{\partial}{\partial \bar{\boldsymbol{\delta}}_{\varepsilon}'} \sum_{i=1}^{n} \left[\frac{\boldsymbol{y}_i \boldsymbol{f}_i}{\boldsymbol{F}_i} - \frac{(1-\boldsymbol{y}_i)\boldsymbol{f}_i}{(1-\boldsymbol{F}_i)}\right] (\boldsymbol{z}_i' \otimes \boldsymbol{x}_i')\boldsymbol{\pi}^{Long} (-\frac{1}{2}) \left[\frac{1}{(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})\,\bar{\boldsymbol{\delta}}_{\varepsilon}}\right]^{\frac{3}{2}} (\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})'$$

$$= \sum_{i=1}^{n} \left[\boldsymbol{y}_i (\frac{\boldsymbol{f}_i'}{\boldsymbol{F}_i} - \frac{\boldsymbol{f}_i^2}{\boldsymbol{F}_i^2}) - (1-\boldsymbol{y}_i)(\frac{\boldsymbol{f}_i'}{(1-\boldsymbol{F}_i)} + \frac{\boldsymbol{f}_i^2}{(1-\boldsymbol{F}_i)^2})\right]$$

$$\times [(\boldsymbol{z}_i' \otimes \boldsymbol{x}_i')\boldsymbol{\pi}^{Long}]^2 (\frac{1}{4}) \left[\frac{1}{(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})\,\bar{\boldsymbol{\delta}}_{\varepsilon}}\right]^3 (\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})'(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})$$

$$+ \sum_{i=1}^{n} \left[\frac{\boldsymbol{y}_i \boldsymbol{f}_i}{\boldsymbol{F}_i} - \frac{(1-\boldsymbol{y}_i)\boldsymbol{f}_i}{(1-\boldsymbol{F}_i)}\right]$$

$$\times (\boldsymbol{z}_i' \otimes \boldsymbol{x}_i')\boldsymbol{\pi}^{Long} (-\frac{1}{2})(-\frac{3}{2}) \left[\frac{1}{(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})\,\bar{\boldsymbol{\delta}}_{\varepsilon}}\right]^{\frac{5}{2}} (\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})'(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'}) \tag{A11}$$

Taking the expectation of both sides of equation (A11) gives

$$\mathrm{E}\left[-\frac{\partial^2 \ln \mathrm{L}}{\partial \bar{\boldsymbol{\delta}}_{\varepsilon} \partial \bar{\boldsymbol{\delta}}_{\varepsilon}'}\right] = \sum_{i=1}^{n} \frac{\boldsymbol{f}_i^2}{\mathrm{F}_i(1-\mathrm{F}_i)} (\frac{1}{4}) \frac{[(\boldsymbol{z}_i' \otimes \boldsymbol{x}_i')\boldsymbol{\pi}^{Long}]^2}{[(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})\,\bar{\boldsymbol{\delta}}_{\varepsilon}]^3} (\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'})'(\overline{\boldsymbol{x}_i' \otimes \boldsymbol{x}_i'}) \tag{A12}$$

where the condition that n > (K+1)(K+1) is needed for the matrix on the right-hand side of this equation to be positive definite.

References

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

Berenguer-Rico, V. and J. Gonzalo (2013), Summability of Stochastic Processes: A Generalization of Integration and Co-integration valid for Non-linear Processes, Unpublished manuscript.

Cramer, J. S. (2006/07) Robustness of Logit Analysis: Unobserved Heterogeneity and Misspecified Disturbances, Discussion Paper 2006/07, Amsterdam School of Economics, Department of Quantitative Economics, Amsterdam, Netherlands.

Felipe, J. and F.M. Fisher (2003), Aggregation in Production Functions: What Applied Economists Should Know, Metroeconomica, 54, 208–262.

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

Greene, W. (2012), Econometric Analysis, 7th edition, Upper Saddle River, New Jersey: Pearson, Prentice Hall.

Heckman, J. J. and E. J. Vytlacil (2005), Structural Equations, Treatment Effects and Econometric Policy Evaluation, Econometrica, 73, 669-738.

Heckman, J. J. and D. Schmierer (2010), Tests of Hypotheses Arising in the Correlated Random Coefficient Model, Economic Modelling, 27, 1355-1367.

Kagan, A. M., Y.V. Linnik and C. R. Rao (1973), Characterization Problems in Mathematical Statistics, New York: John Wiley & Sons.

Karlsen, H. A., T. Myklebust and D. Tjøstheim (2007), Nonparametric Estimation in a Nonlinear Cointegration Type Model, The Annals of Statistics, 35, 252–299.

Lehmann, E. L. and G. Casella (1998), Theory of Point Estimation, New York: Spriger Verlag, Inc.

Pearl, J. (2000), Causality, Cambridge, UK: Cambridge University Press.

Pratt, J. W. and R. Schlaifer (1984), On the Nature and Discovery of Structure (with discussion), Journal of the American Statistical Association, 79, 9-21.

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

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

Skyrms, B. (1998), Probability and Causation, Journal of Econometrics 39, 53-68.

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

Swamy, P. A. V. B., J. S. Mehta, G. S. Tavlas and S. G. Hall (2014), Small Area Estimation with Correctly Specified Linking Models, 193-228 in: J. Ma and M. Wohar (eds.), Recent Advances in Estimating Nonlinear Models, With Applications in Economics and Finance, New York: Springer.

Swamy, P.A.V.B. and P. von zur Muehlen (1988), Further Thoughts on Testing for Causality with Econometric Models, Journal of Econometrics, 39, 105-147.

White, H. (1980), Using Least Squares to Approximate Unknown Regression Functions, International Economic Review 21, 149-170.

White, H. (1982), Maximum Likelihood Estimation of Misspecified Models, Econometrica 50, 1-25.

Whittle, P. (1976), Probability, New York: John Wiley & Sons.

Wooldridge, J. M. (2002), Econometric Analysis of Cross-Section and Panel Data, Cambridge, Massachusetts: The MIT Press.

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