OMITTED VARIABLE BIAS
AND
CROSS SECTION REGRESSION

by

Thomas M. Stoker

July 1983

WP #1460-83

Thomas M. Stoker is Assistant Professor, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139. The author wishes to thank A. Deaton, T. Gorman, J. Hausmann, J. Heckman, D. Jorgenson, A. Lewbel, J. Muellbauer, J. Powell and J. Rotemberg for helpful comments on this and related work. All errors, etc., remain the responsibility of the author.

ABSTRACT

This paper reinterprets and explains the standard omitted variable bias formula in the context of cross section regression when the true model underlying behavior is unknown and possibly nonlinear. The vehicle employed to analyze cross section regression in this case is the macroeconomic interpretation of cross section OLS coefficients established in Stoker (1982a).

The exposition begins by indicating precisely the distributional assumptions underlying a correctly specified linear cross section regression equation when the true model is nonlinear and possibly unknown. By considering the case of too many regressors, we show that the omitted variable bias formula reflects constraints in distribution movement, which alternatively allow the bias formula to be derived as a total derivative formula among macroeconomic effects. By considering the case of too few regressors, we show that the macroeconomic impact of the omitted variables can be measured by their partial contribution to the variance of the dependent variable in a cross section regression. Some practical implications of these results are discussed and an illustrative example is given.

I. Introduction

The purpose of this paper is to reinterpret and explain standard omitted variable bias formulae in the context of cross section regression when the true model underlying behavior is unknown and possible nonlinear. The vehicle employed to analyze cross section regression in this case is the macroeconomic interpretation of cross section OLS coefficients established in Stoker (1982a).

The omitted variable bias formula is a very useful tool for judging the impact on regression analysis of omitting important influences on behavior which are not observed in the data set. In small sample form, the bias formula was developed and popularized by Thiel (1957, 1971), and has been used extensively in empirical research.[1] The bias interpretation of the formula, however, relies exclusively on the assumed linearity of the included and omitted variables in the equation modeling the dependent variable.

The formula itself has an empirical counterpart which holds an identity among computed OLS regression coefficients from equations with different subsets of regressors.[2] The question of interest here is whether this regression coefficient relationship can be interpreted when the behavioral model is general and possibly unknown. A macroeconomic interpretation for cross section OLS coefficients in this case was established by Stoker (1982a). In this paper we will extend the interpretation to the standard omitted variable bias formula.

The precise issue addressed can be described in more detail as follows. Stoker (1982a) established that OLS slope coefficients obtained from regressing a dependent variable y on predictor variables X computed using cross section data will consistently estimate the effects of changing mean X, E(X) on mean y, E(y), provided that the X distribution varies through time via the

exponential family form. This latter condition is of interest because it implies no testable restrictions on the cross section data, and in particular does not rely on a particular functional form of the relationship between y and X. But suppose that X is partitioned as $X = (X_1, X_2)$. The above result can also be applied to say that the OLS coefficients of y on X_1 consistently estimate the effects of changing $E(X_1)$ on $E(y)$. In this paper, we will explain exactly how the assumptions underlying the macroeconomic interpretations of these two regressions differ. In so doing, we obtain a general interpretation of the omitted variable bias formula, which connects the coefficients of these two regressions.

The results of the paper shift the misspecification question from the behavioral model to the assumptions which control the way the population distribution evolves through time. If the driving variables (to be defined) of the predictor distribution are X_1, then the proper macroeconomic effects are estimated by the cross section regression of y on X_1 only. The bias formula connecting these coefficients to those of regression of y on X_1 and X_2 just reflects the induced effect of $E(X_1)$ on $E(X_2)$. The development showing this can be regarded as an alternative proof of the omitted variable bias formula, obtained by manipulating derivatives of macroeconomic functions.

Alternatively, if the driving variables of the predictor variable distribution are X_1 and X_2, then a cross section regression of y on X_1 will not uncover the full impact of distribution change on the mean of y, $E(y)$. In this case we show that the additional effect due to distribution change can be measured by the partial contribution of X_2 to the variance of y holding X_1 constant in a cross section regression. This result has some practical implications for regression analysis when the true behavioral model is unknown.

The appeal of our results derives from two sources. First, the omitted variable bias formula is an important tool for econometricians, which is covered in virtually every intermediate level econometrics textbook, and should be at work in judging coefficient robustness in any good empirical analysis. It is therefore useful to understand its applicability in misspecified nonlinear circumstances. Second, the results indicate how to use cross section data to determine exactly what distributional influences are important to macroeconomic equations. For instance, the auxiliary equations of the omitted variables regressed on included ones provide observable constraints which will hold if the omitted variable means can be correctly excluded from the model explaining mean y.

To see this latter point, consider the simple example of characterizing income and family size effects in the demand for a commodity such as food. Supposing prices are constant for simplicity, it is certainly the case that for each family size, the individual Engel curve relating food to income is nonlinear. In studying this relationship with traditional modeling techniques, the omitted variable bias formula cannot be used to indicate the bias in the income effects induced by omitting family size unless a) the true food equation is intrinsically linear, depending on prespecified nonlinear income terms (logs, etc.) with only coefficients to be estimated and b) exactly the right nonlinear income terms have been specified.

The developments of this paper indicate how to interpret simple cross section regression analysis results in a way which does not require the correct specifications of the behavioral model; the family size augmented Engel curve for food in this example. If the exponential family structure is adopted for changes in the joint income - family size distribution, the cross section OLS coefficients of food on income and family size consistently estimate the effects of changing average income and average family size on average food.

Now, average family size may be correctly excluded from an average food equation if family size has a zero cross section food regression coefficient or if the conditional distribution of family size given income is constant through time. This latter condition says that average food is a function only of average income, with average family size having no independent effect. It is for checking this latter possibility that the omitted variable bias calculations are useful. In particular, the estimated coefficients of the auxiliary equation of family size regressed on income indicates the effect of average income changes on average family size. If two or more time series observations on average income and average family size are consistent with the estimated effects, then omitting average family size from the average food model is suggested. If the estimated effects bear no relation to the time series patterns of average income and average family size, and if family size has a nonzero cross section regression coefficient in a food equation, then average family size has an independent influence on average food demand.

We begin with the notation, a discussion of the omitted variable bias formula and a review of the OLS coefficient results of Stoker (1982a). In Section 3 we consider the case of too many regressors in the cross section equation, and present the alternative derivation of the omitted variable bias formulae using macroeconomic derivatives. In Section 4 we consider the case of too few regressors , indicating the macroeconomic analogue of coefficient bias. In Section 5 we present an algebraic example, and in Section 6 discuss some related work.

2. Notation and Background Results

2.1 Individual Models and Cross Section Data

All of our results will concern interpretations of OLS regression coefficients computed with cross section data observed at a particular time period, say $t = t_o$. Denote by y a dependent variable of interest, and by X

an M vector of predictor variables. The cross section data consists of K observations on these variables y_k, X_k, k=1 ..., K, which are assumed to represent a random sample from a distribution with density $P_o(y,X)$. Moreover, the entire population at $t = t_o$ of (say) N observations is assumed to be a random sample from the same distribution with $N >> K$.[4] The following assumption characterizes the cross section structure.

ASSUMPTION 1: The means, variances and covariances of y and x exist, and the variance-covariance matrix of X is non-singular and positive definite. The conditional distribution of y given X exists, with density $q_o(y|X)$, as does the mean of y given X, denoted $E(y|X) \equiv F(X)$.

For the purpose of considering omitted variables, we suppose that X is partitioned into an M_1 vector X_1 and an M_2 vector X_2 as $X' = (X_1', X_2')$ where $M_1 + M_2 = M$. Denote the means of y and X by $E_o(y) \equiv \mu_o^y$ and $E_o(X)' = \mu_o' = (\mu_o^{1'}, \mu_o^{2'})$, the variance of y by σ_{yy}^o, the variance-covariance matrix of X by

$$\Sigma_{XX}^o = \begin{bmatrix} \Sigma_{11}^o & \Sigma_{12}^{o'} \\ \Sigma_{12}^o & \Sigma_{22}^o \end{bmatrix} \tag{2.1}$$

and the covariance matrix between y and X as

$$\Sigma_{Xy}^o = \begin{bmatrix} \Sigma_{1y}^o \\ \Sigma_{2y}^o \end{bmatrix} \tag{2.2}$$

when the notation corresponds to the partitioning of X. The overall density $P_o(y|X)$ which underlies the cross section can be factored as $P_o(y|X) = q_o(y|X)\, p_o(X)$, where $p_o(X)$ is the marginal distribution of X.

The conditional density $q_o(y|X)$ corresponds to the true econometric model relating y and X for individual observations. In standard practice, in order to study the relationship between y and X, one would specify a behavioral model $y = f_\gamma(X,u)$, where u represents unobserved individual heterogenteity and γ a set of parameters to be estimated, together with the stochastic distribution of u given X, say with density $\tilde{q}(u|X)$. Combining the behavioral function and the heterogeneity distribution gives the conditional density $q_o(y|X)$. We assume γ is equal to its true value, and thus suppress it in the notation. For concreteness, consider the example given in the introduction, where y denotes the demand for food by individual families, X_1 income and X_2 family size. The true Engel curve with family size is represented by $E(y|X) = F(X)$, and $q(y|X)$ reflects the Engel curve together with the stochastic specification of the deviation $y - F(X)$. If the true behavioral function was linear with additive disturbance - i.e., $y = f_\gamma(X,u) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + u$ - and the distribution of u conditional on X was normal with mean 0 and variance σ^2, then $q_o(y|X)$ denotes a normal distribution with mean $F(X) = \gamma_o + \gamma_1' X_1 + \gamma_2' X_2$ and variance σ^2. Alternatively, the framework will accomodate many other standard econometric modeling situations - for example if y takes on only a finite number of values and behavior is described by a discrete choice model, then $q_o(y|X)$ gives the choice probabilities for each of the possible values of y given X, which could be of the probit or logit form with appropriate specification of the distribution of unobserved individual influences on the choice process.

All of the exposition is concerned with interpretations of regressions performed using the cross section data. The regression of y on X is represented as

$$y_k = \hat{a}_{y.12} + X_k'\hat{b}_{y.12} + \hat{\varepsilon}_k$$

$$= \hat{a}_{y.12} + X_{1k}'\hat{b}_{y.1(2)} + X_{2k}'\hat{b}_{y.2(1)} + \hat{\varepsilon}_k \qquad (2.3)$$

where $\hat{b}_{y.12}' = (\hat{b}_{y.1(2)}', \hat{b}_{y.2(1)}')$ are computed using ordinary least squares (OLS) and the notation reflects the partitioning of $X' = (X_1', X_2')$. We denote the large sample (probability limit) values of the statistics from this regression as in

$$\operatorname*{plim}_{K\to\infty} \hat{b}_{y.12} = (\Sigma^o_{XX})^{-1}\Sigma^o_{Xy} = \beta_{y.12} = (\beta_{y.1(2)}', \beta_{y.2(1)}')'$$

$$\operatorname*{plim}_{K\to\infty} \hat{a}_{y.12} = \mu^y_o - \beta_{y.12}'\mu_o = \alpha_{y.12} \qquad (2.4)$$

$$\operatorname*{plim}_{K\to\infty} \Sigma\varepsilon_k^2/K = \sigma^o_{yy} - \Sigma^{o\prime}_{Xy}(\Sigma^o_{XX})^{-1}\Sigma^o_{Xy}$$

$$= \sigma_{y.12}$$

Also of interest is the regression of y on X_1 only, which we denote as

$$y_k = \hat{a}_{y.1} + X_{1k}'\hat{b}_{y.1} + \hat{w}_k \qquad (2.5)$$

$$k = 1, \ldots, k.$$

The large sample values of these statistics are denoted as in:

$$\operatorname*{plim}_{K\to\infty} \hat{b}_{y.1} = (\Sigma^o_{11})^{-1}\Sigma^o_{1y} = \beta_{y.1}$$

$$\operatorname*{plim}_{K\to\infty} \hat{a}_{y.1} = \mu^y_o - \mu_o^{1\prime}\beta_{y.1} = \alpha_{y.1} \qquad (2.6)$$

$$\operatorname*{plim}_{K\to\infty} \frac{\Sigma w_k^2}{k} = \sigma^o_{yy} - \Sigma^{o\prime}_{1y}(\Sigma^o_{11})^{-1}\Sigma^o_{1y} = \sigma_{y.1}$$

The slope regression coefficients of (2.3) and (2.5) are connected by the identity

$$\hat{b}_{y.1} = \hat{b}_{y.1(2)} + \hat{B}_{2.1}\,\hat{b}_{y.2(1)} \tag{2.7}$$

where $\hat{B}_{2.1}$ is the $M_1 \times M_2$ matrix of OLS coefficients of the auxiliary regression

$$X_{2k}' = \hat{A}_{2.1}' + X_{1k}'\hat{B}_{2.1} + \hat{v}_k' \qquad k=1,\ldots,K \tag{2.8}$$

The version of (2.7) relating the large sample values of the coefficients is

$$\beta_{y1} = \beta_{y.1(2)} + B_{2.1}\beta_{y.2(1)} \tag{2.9}$$

where $B_{2.1} = (\Sigma_{11}^{o})^{-1}\Sigma_{12}^{o}{}' = \text{plim}\ \hat{B}_{2.1}$.

2.2 The Omitted Variable Bias Formula

The standard omitted variable bias formula is an equation explaining the small sample expectation of $\hat{b}_{y.1}$ when the true behavioral model specifies y as a linear function of X_1 and X_2 with additive residual. The equation is formally quite similar to (2.7) and (2.9), and we introduce it separately here for later comparison with our general development.

We begin by assuming that $q_o(y|X)$ is a distribution with mean $E(y|X) = F(X) = \gamma_o + \gamma_1'X_1 + \gamma_2'X_2$, or equivalently that the true behavioral model is

$$y = \gamma_o + \gamma_1'X_1 + \gamma_2'X_2 + u \tag{2.10}$$

where u has zero expectation conditional on X. In this case, it is easy to verify that $\alpha_{y.12} = \gamma_o$, $\beta_{y.1(2)} = \gamma_1$ and $\beta_{y.2(1)} = \gamma_2$, using our previous notation.

The omitted variable bias formula is derived by inserting (2.10) evaluated for y_k into the OLS formula defining $\hat{b}_{y.1}$ of (2.5) and taking its expectation. This yields

$$E(\hat{b}_{y.1}|X \text{ data}) = \gamma_1 + \hat{B}_{2.1}\gamma_2 \qquad (2.11)$$

where $\hat{B}_{2.1}$ is defined as the OLS coefficients of (2.8) and "X data" denotes that the expectation is taken conditional on $X_{1k}, X_{2k}, k=1,\ldots,K$. (2.11) is the omitted variable bias formula.

The practical usefulness of this formula can be illustrated using our previous example. Suppose y is food expenditure, X_1 is income, X_2 is family size and (2.10) is the true demand equation, with γ_1 and γ_2 positive. (2.11) says if one regresses food y on income X_1 only (omitting family size X_2) that $\hat{b}_{y.1}$ will on average overestimate (underestimate) the income effect γ_1 if the regression coefficient $\hat{B}_{2.1}$ of family size X_2 on income X_1 is positive (negative). The magnitude of the bias $E(\hat{b}_{y.1}|X \text{ data}) - \gamma_1$ depends on the size of the true family size effect γ_2 and the amount of the correlation between family size and income in the data.

Perhaps better use of the equation (2.11) occurs when X_2 is not observed in the data. Suppose for instance that y and X_1 are as above, but X_2 now represents an unobserved variable, say the amount of gambling done by each family. If we suppose that gambling has a negative effect on food expenditure, $\gamma_2 < 0$, then (2.11) says that $\hat{b}_{y.1}$ will on average overestimate (underestimate) the true income effect if income and amount of gambling are negatively (positively) correlated in the sample. If the analyst has outside information that gambling activity is weakly correlated with income level, then (2.11) provides an argument for robustness, namely that $\hat{b}_{y,1}$ will on average equal the true income effect γ_1.

In our general framework, where we relax the linear model assumption (2.10), it is difficult to characterize the small sample properties of $\hat{b}_{y.1}$, and so we lose the omitted variable bias formula (2.11) as a tool for analysis. We will instead concentrate on interpreting (2.9), the large sample version of (2.7) and (2.11). For this task, we must first review the macroeconomic interpretation of cross section regression coefficients, which characterizes the large sample values $\beta_{y.1}$ and $\beta_{y.12}$.

2.3 Macroeconomic Effects and Regression Coefficients

The results of Stoker (1982a) (reviewed below) establish that cross section OLS regression coefficients consistently estimate the macroeconomic effects of changing the mean of X on the mean of y. In this section we review the exponential family assumptions which are sufficient for the result. We then provide an immediate proof of the result for the exponential family case.

In order to discuss a relationship between the mean of y and the mean of X for a general behavioral model $q_o(y|X)$, we must specify precisely how the population density $P_o(y,X)$ changes through time. We assume that the behavioral model $q_o(y|X)$ is stable through time, so that we can focus attention on how the marginal X density $p_o(X)$ varies. In this paper we will employ a particular structure for the X distribution, known as exponential family structure, which is introduced through the following assumptions:[5]

ASSUMPTION 2: The La Place Transform of $p_o(X)$:

$$L(\Pi) = \frac{1}{C(\Pi)} = \int p_o(X) \exp(\Pi'X)\, dX \qquad (2.12)$$

exists for Π in a convex open neighborhood Γ of the origin in R^M.

DEFINITION: The exponential family generated by $p_o(X)$ with driving variables X is the family defined by

$$p^*(X|\Pi) = C(\Pi)\ p_o(X)\ \exp\ (\Pi' X) \tag{2.13}$$

where $\Pi \varepsilon \Gamma$ and $C(\Pi)$ is defined via (2.12).

As given the exponential family form is a standard distribution form known to statistics, which encompasses virtually all of the "textbook" distribution forms, such as Poisson, gamma, beta, multivariate normal and lognormal distributions among others, found by appropriate specification of the generating distribution and driving variables.[6] Notice for our purposes that the natural parameters Π serve to index movements in the X distribution, with $\Pi = 0$ corresponding to the cross section density $p_o(X) = p^*(X|0)$. We formalize this as

ASSUMPTION 3A: For each time period $t \neq t_o$, there exists Π_t such that the marginal X density at time t is given via the exponential family form with driving variables X and parameter Π_t; i.e., $p_t(X) = p^*(X|\Pi_t)$ of (2.12). The joint distribution of y and X at time t has density $P_t(y,X) = q_o(y|X)\ p^*(X|\Pi_t)$.

Assumption 3A provides sufficient structure to determine the means of y and X as functions of the natural parameters Π of the X distribution. By direct integration, we have

$$E(y) = \mu^y = \int y\ q(y|X)\ p^*(X|\Pi)\ dX \equiv \Phi^*(\Pi) \tag{2.14}$$

$$E(X) = \mu = \int X\ p^*(X|\Pi)\ dX \equiv H(\Pi) \tag{2.15}$$

where, for $\Pi = 0$ we have the cross section values of $\mu_o^y = \Phi^*(0)$ and $\mu_o = H(0)$. The problem with (2.14-15) is inconvenience, for it is not clear how to behaviorally interpret the natural parameters Π. To overcome this, we

reparameterize the X distribution by $\mu = E(X)$, and derive the relation between $E(y) = \mu^y$ and $E(X) = \mu$ induced by (2.14-15). This is possible because $H(\Pi)$ of (2.15) is invertible, and so we can redefine $p^*(X|\Pi)$ as

$$p(X|\mu) = p^*(X|H^{-1}(\mu)) \tag{2.16}$$

and derive the (macroeconomic) aggregate function between μ^y and μ as

$$E(y) = \mu^y = \int y\ q(y|X)\ p(X|\mu)\ dX \equiv \Phi(\mu) \tag{2.17}$$

Of course, we have $\mu_o^y = \Phi(\mu_o)$ for the cross section parameter values. Parenthetically, to see that $H(\Pi)$ is invertible, note that[7]

$$\mu = H(\Pi) = -\frac{\partial \ln C(\Pi)}{\partial \Pi} \tag{2.18}$$

(where $\frac{\partial}{\partial \Pi}$ is the gradient operator) is invertible locally at $\Pi = 0$ if and only if its differential (Jacobian) matrix is nonsingular. This matrix is easily seen to be the covariance matrix of X via

$$-\frac{\partial^2 \ln C}{\partial \Pi \partial \Pi'} = \Sigma_{XX} \tag{2.19}$$

which is assumed nonsingular at $\Pi = 0$, the cross section value.

The aggregate function $\mu^y = \Phi(\mu)$ represents the model of macroeconomic behavior in our framework, corresponding with the individual behavioral model $q_o(y|X)$ and Assumption 3A on the X distribution. The macroeconomic effects of $\mu = E(X)$ on $\mu^y = E(y)$ are defined as the first derivatives of $\Phi(\mu)$, denoted by $\frac{\partial \Phi}{\partial \mu}$. Our results are concerned with the value of these derivatives at $\mu = \mu_o$, the cross section parameter values.

As a final bit of background notation, it is useful to introduce formulae which capture the local behavior of the expectations (2.14), (2.15) at $\Pi = 0$. For (2.15) we have that changes in μ, $d\mu$, are related to changes in Π, $d\Pi$, at $\Pi = 0$ as in

$$d\mu = \Sigma^{o}_{XX}\, d\Pi \tag{2.20}$$

which is obvious from (2.19). Similarly, for (2.14), it is easy to show that changes in μ^{y}, $d\mu^{y}$, are related to $d\Pi$ at $\Pi = 0$ as in

$$d\mu^{y} = \Sigma^{o'}_{Xy}\, d\Pi \tag{2.21}$$

We refer to (2.20) and (2.21) as the "local equations" corresponding to (2.15) and (2.14) respectively.

The local equations provide very convenient methods for manipulating derivatives of expectations in our framework. For an illustration, we provide an immediate proof of the result of Stoker (1982a) that cross section (OLS) coefficients always consistently estimate macroeconomic effects under exponential family structure on the distribution of X. To see this, invert (2.20) and insert into (2.21) as

$$\begin{aligned} d\mu^{y} &= \Sigma^{o'}_{Xy}\, d\Pi = \Sigma^{o'}_{Xy}\, \Sigma^{o\,-1}_{XX}\, d\mu \\ &= (\beta_{y.12})'\, d\mu \end{aligned} \tag{2.22}$$

and so plim $\hat{b}_{y.12} = \beta_{y.12} = \frac{\partial \Phi}{\partial \mu}(\mu_0)$, the macroeconomic effects. The above manipulations just reflect application of the chain rule to the aggregate function (2.17).

Before proceeding to discuss omitted variables, it is useful to point out some salient aspects of the development preceding the OLS coefficient result (2.22). First, the result holds for a virtually arbitrary behavioral model $q_o(y|X)$ and cross section distribution $p_o(X)$, which are restricted by only the innocuous Assumptions 1 and 2. Second, the driving variables X of the exponential family play an important role,[8] as they constitute the proper regressors in the cross section equation whose coefficients consistently estimate the macroeconomic effects. Elaboration of this relation is what permits analysis of omitted variables and specification error, to which we now turn.

3. Too Many Regressors

The result of Stoker (1982a) reviewed above provides a macroeconomic interpretation of the OLS coefficients of any cross section regression performed, under the corresponding set of distribution movement assumptions. In this section we consider the case where the regression (2.5) of y on X_1 is the correct one for estimating macroeconomic effects, as opposed to the regression (2.3) of y on X_1 and X_2.

The local equations (2.20), (2.21) and (2.22) are derived under the structure where the latter regression (2.3) is appropriate. The equations (2.20) and (2.21) rewritten to reflect the $X' = (X_1', X_2')$ partitioning are

$$d\mu^1 = \Sigma^o_{11} d\Pi_1 + \Sigma^o_{12}{}' d\Pi_2 \tag{3.1a}$$

$$d\mu^2 = \Sigma^o_{12} d\Pi_1 + \Sigma^o_{22} d\Pi_2 \tag{3.1b}$$

and

$$d\mu^y = \Sigma^o_{1y} d\Pi_1 + \Sigma^o_{2y}{}' d\Pi_2 \tag{3.2}$$

where $\Pi' = (\Pi_1', \Pi_2')$ is partitioned into the natural parameters corresponding to X_1 and X_2. Equation (2.22), which established that plim $\hat{b}_{y.12}' = \left(\frac{\partial \Phi'}{\partial \mu}, \frac{\partial \Phi'}{\partial \mu_2}\right)$, is written in partitioned form as

$$d\mu^y = \beta_{y.1(2)}' d\mu^1 + \beta_{y.2(1)}' d\mu^2 \tag{3.3}$$

As noted at the end of Section 2, for the regression coefficients of y on X_1 from equation (2.5) to consistently estimate macroeconomic effects, we must adopt the corresponding assumption that the X distribution changes via the exponential family with driving variables X_1 only, as in

$$p_1^*(X|\Pi_1) = C_1(\Pi_1)\, p_o(X) \exp(\Pi_1 X_1) \tag{3.4}$$

where $C_1(\Pi_1) = C(\Pi)$, the latter evaluated at $\Pi_2 = 0$. The parallel assumption is written out as

> ASSUMPTION 3B: For each time period $t \neq t_o$ there exists Π_{1t} such that the marginal distribution of X at time t is given via the exponential family with driving variables X_1 and parameters Π_{1t} ; i.e., $p_t(X) = p_1^*(X|\Pi_{1t})$ of (3.4). The joint distribution of y and X at time t has density $P_t(y,X) = q_o(y|X)\, p_1^*(X|\Pi_{1t})$.

Under Assumption 3B, we can compute the mean of y and X_1 as functions of Π_1 as before

$$E(y) = \mu^y = \Phi_1^*(\Pi_1) \tag{3.5}$$

$$E(X_1) = \mu^1 = H_1(\Pi_1) \tag{3.6}$$

and find the induced relation between μ^y and μ^1 as

$$\mu^y = \phi_1^*(H_1^{-1}(\mu^1)) = \phi_1(\mu^1) \tag{3.7}$$

the pertinant aggregate function for this case. The OLS coefficient result now says that plim $\hat{b}_{y.1} = \beta_{y.1} = \frac{\partial \phi_1}{\partial \mu^1}(\mu_o^1)$, where $\hat{b}_{y.1}$ are the coefficients of y on X_1 in equation (2.5). The result can be verified easily as above by directly deriving the local equations pertinant to (3.5) and (3.6), which are

$$d\mu^y = \Sigma_{1y}^{o\prime} d\Pi_1 \tag{3.8}$$

$$d\mu^1 = \Sigma_{11}^{o} d\Pi_1 \tag{3.9}$$

and solving them for the induced local relation between μ^y and μ^1 as

$$d\mu^y = \Sigma_{1y}^{o\prime}(\Sigma_{11}^{o})^{-1} d\mu_1 = \beta_{y.1}' d\mu_1 \tag{3.10}$$

The large sample omitted variable bias formula (2.9) arises out of the differences between Assumptions 3A and 3B. A moments reflection indicates that Assumption 3B is just Assumption 3A with the proviso that Π_2 is held constant at $\Pi_2 = 0$. This is reflected in the fact that the local equations (3.8), (3.9) coincide with (3.2) and (3.1a) when $d\Pi_2 = 0$. Consequently, (3.10) and (3.2) must coincide when $d\Pi_2 = 0$. Detailing this correspondence yields formula (2.9).

By requiring $d\Pi_2 = 0$, Assumption 3B also structures the mean $\mu^2 = E(X_2)$ as a function of Π_1. By factoring the base density $p_o(X)$ into $p_o(X) = p_2(X_2|X_1)p_{10}(X_1)$ where $p_{10}(X)$ is the marginal distribution of X_1, we have

$$E(X_2) = \mu^2 = \int C_1(\Pi_1)\ p_2(X_2|X_1)\ p_{10}(X_1)\ \exp\ (\Pi_1'X_1)\ d\ X_1$$

$$= G^*(\Pi_1) \tag{3.11}$$

or in terms of μ^1;

$$\mu^2 = G^*(H_1^{-1}(\mu^1)) = G(\mu^1) \tag{3.12}$$

The local behavior of G^* and G at $\Pi_1 = 0$ can be derived directly as before, or equivalently by setting $d\Pi_2 = 0$ in (3.1a-b). The local behavior of G^* is found from (3.1b) to be

$$d\mu^2 = \Sigma_{12}^o\ d\Pi_1 \tag{3.13}$$

Inverting (3.1a) and inserting into (3.13) gives the local behavior of G as

$$d\mu^2 = \Sigma_{12}^o\ d\Pi_1 = \Sigma_{12}^o\ (\Sigma_{11}^o)^{-1}\ d\mu_1 \tag{3.14}$$

$$= B_{2.1}'\ d\mu_1$$

Consequently, equation (3.3) under $d\Pi_2 = 0$ is

$$\begin{aligned} d\mu^y &= \beta_{y.1(2)}'d\mu^1 + \beta_{y.2(1)}'d\mu^2 \\ &= \beta_{y.1(2)}'d\mu^1 + \beta_{y.2(1)}'B_{2.1}'d\mu^1 \\ &= (\beta_{y.1(2)} + B_{2.1}\beta_{y.2(1)})'d\mu_1 \\ &= \beta_{y.1}'\ d\mu_1 \end{aligned} \tag{3.15}$$

establishing the equivalence between (3.3) and (3.10).

This development yields several interpretations of standard specification analysis calculus in the context of a general population model. Equation (3.12) points out the macroeconomic interpretation of the auxiliary regressions (2.9) of X_2 and X_1; namely that $\hat{B}_{2.1}$ consistently estimates the induced effects of μ^1 on μ^2. The development (3.13) just says that the overall effect of μ^1 on μ^y under Assumption 3B is the direct effect $\beta_{y.1(2)}$ plus the direct effect $\beta_{y.2(1)}$ of μ^2 on μ^y multiplied by the induced effect of μ^1 on μ^2. Consequently, the large sample omitted variable bias formula (2.9) is just the total derivative of $\mu^y = \Phi(\mu)$ with respect to μ^1 under the constraint that $d\Pi_2 = 0$. Thus this development can be regarded as an alternative proof of equation (2.9) found by taking macroeconomic derivatives.

The question of whether Assumption 3B is correct versus Assumption 3A cannot be decided with cross section data, since each restricts only the way the distribution changes away from the cross section. However, the auxiliary equation coefficients $\hat{B}_{2.1}$ do provide consistent estimates of the induced effects on changes in μ^2 due to μ^1 changes when Assumption 3B holds. Consequently, if small changes in μ^1 and μ^2 are observed (via time series) which are consistent with $\hat{B}_{2.1}$, then Assumption 3B is not rejected.[9]

Moreover, the development shows that including too many variables in a cross section regression is not a problem in our general format. In particular, if equation (2.3) of y regressed on X_1 and X_2 is estimated, the coefficients will still estimate the independent effects of μ^1 and μ^2 on μ^y. By recognizing the dependence of the erroneously included variable means μ^2 on μ^1, the overall effect on μ^y indicated is the same as that estimated by the properly specified equation (2.5) of y on X_1.

4. Too Few Regressors

In this section we consider the classical omitted variables problem of omitting pertinant regressors, in the context of a general behavioral model. In the macroeconomic interpretation of cross section regression coefficients, the pertinant regressors correspond with the driving variables of the exponential family. Consequently, here we take that Assumption 3A represents population movements, and consider the implications of performing the regression (2.5) of y on X_1, omitting X_2.

The full impact of distribution movements under Assumption 3A is represented by (3.3), reproduced here as

$$d\mu^y = \beta_{y.1(2)}' d\mu^1 + \beta_{y.2(1)}' d\mu^2 \tag{4.1}$$

Changes in μ^2 are no longer constrained as with Assumption 3B. Consequently, the misspecified regression (2.5) cannot adequately estimate all possible distribution effects, and the question becomes how to measure the extent of what $\hat{b}_{y.1}$ of equation (2.5) misses.

We can find such a measure by again adjusting the parameterization of distribution movements under Assumption 3A. We introduced the exponential family using the natural parameters $\Pi' = (\Pi_1', \Pi_2')$ in (2.12) and then considered the mean parameterization $\mu' = (\mu^{1\prime}, \mu^{2\prime})$ in (2.16). Now we reparameterize locally with the mixed parameter (μ^1, Π_2).[10] This is accomplished by manipulating the local equations (3.1a) and (3.2) as follows. Solve (3.1a) for $d\Pi_1$ as

$$d\Pi_1 = (\Sigma_{11}^o)^{-1} d\mu_1 + (\Sigma_{11}^o)^{-1} \Sigma_{12}^{o'} d\mu_2 \tag{4.2}$$

and insert into (3.2) as

$$d\mu^y = \Sigma^o_{1y}{}' (\Sigma^o_{11})^{-1} d\mu_1 + (\Sigma^o_{2y}{}' - \Sigma^o_{1y}{}' (\Sigma^o_{11})^{-1}\Sigma^o_{12}{}')d\Pi_2 \qquad (4.3)$$

$$= \beta_{y.1}{}' d\mu_1 + (\Sigma^o_{2y} - \Sigma^o_{12}(\Sigma^o_{11})^{-1}\Sigma^o_{1y})' d\Pi_2$$

This equation says that the misspecified regression coefficients of y on X_1 consistently estimate the effects of μ^1 changes on μ^y holding Π_2 constant, an obvious finding in light of Section 3. The remaining distributional effects arise from changes in Π_2, with their relative importance measured by the coefficient of Π_2 in (4.3). This coefficient is easily seen to be

$$\Sigma^o_{2y} - \Sigma^o_{12} (\Sigma^o_{11})^{-1} \Sigma^o_{1y} = \Sigma^o_{2y} - B_{2.1}{}' \Sigma^o_{1y}$$

$$= \text{Cov}_0 (X_2 - B_{2.1}{}' X_1, y) \qquad (4.3)$$

$$= \sigma_{y.2(1)}$$

the partial covariance between X_2 and y holding X_1 constant. Consequently, the local importance of Π_2 deviations in the mean of y (given μ^1) is directly measured by the independent contribution of X_2 to the explanation of y in the true cross section regression (2.3). This covariance can alternatively be written as $\sigma_{y.2(1)} = \sigma_{2.1}\beta_{y.2(1)}$ where $\sigma_{2.1} = \text{plim } \Sigma v_k v_k'/K$ is the large sample residual variance matrix from the auxiliary regression (2.8) and $\beta_{y.2(1)}$ is the true macroeconomic effect of μ^2 on y holding μ^1 constant.

This analysis, along with analysis of Section 3, provides an alternative justification of some common practice techniques of regression analysis in the context of unknown functional form of the true behavioral model. In particular, for the purpose of characterizing macroeconomic effects, this work suggests performing relatively large regressions (many X's), and choosing variables via their importance in the explanation of the variance of y. Section 3 says when

the list of regressors is too large, there will exist induced constraints between the means of erroneously included variables and means of the correctly included ones. The local version of these constraints are given as the large sample omitted variables formula (2.9), which equate the macroeconomic effects on mean y indicated by the properly specified regression to those from the regression with too many regressors. This section shows that omitting proper variables has an impact on mean y which can be measured by the partial covariance between the omitted variables and y in a cross section regression. Consequently, in a circumstance of unknown functional form, this analysis suggests including all variables which have large partial impacts, since including too many will be reconciled by the induced constraints.

The other suggestion of this development concerns the characterization of distribution movements, say with panel data or aggregate time series. For an exponential family structure, the candidates of most interest for driving variables are those which exhibit substantial contributions to y in a cross section regression framework.[11]

5. An Example

In this section we add some concreteness to the general development by displaying the various regression and omitted variable formulae for a specific behavioral function with normally distributed regressors. Suppose that the true model gives y as a quadratic function of two scalar variables X_1 and X_2 and an independent (mean zero) disturbance u as

$$y = \gamma_o + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_{11} X_1^2 + \gamma_{12} X_1 X_2 + \gamma_{22} X_2^2 + u \qquad (5.1)$$

If the true form above were known, one would perform a regression including linear and quadratic terms to estimate all the γ parameters. Here, we consider

the regressions of y on X_1 and X_2, and y on X_1 with (5.1) as the true model.

Suppose that in the cross section $(X_1,X_2)'$ is joint normally distributed with mean $\mu_o = (\mu_o^1,\mu_o^2)'$ and covariance matrix

$$\Sigma^o = \begin{bmatrix} \sigma_{11}^o & \sigma_{12}^o \\ \sigma_{12}^o & \sigma_{22}^o \end{bmatrix} \tag{5.2}$$

The exponential family with driving variables X_1 and X_2 (Assumption 3A) consists of the normal distributions with varying means $\mu = (\mu^1, \mu^2)'$ and fixed covariance matrix Σ^o. The aggregate function relating $E(y) = \mu^y$ to $E(X_1) = \mu^1$ and $E(X_2) = \mu^2$ is

$$\begin{aligned} E(y) = \mu^y = \Phi(\mu) = \gamma_o + \gamma_1\mu^1 + \gamma_2\mu^2 + \gamma_{11}((\mu^1)^2 + \sigma_{11}^o) \\ + \gamma_{12}\,(\mu^1\mu^2 + \sigma_{12}^o) + \gamma_{22}\,((\mu^2)^2 + \sigma_{22}^o) \end{aligned} \tag{5.3}$$

Given the model (5.1), the following covariances can be verified[12]

$$\begin{aligned} \mathrm{Cov}(y,X_1) = \sigma_{1y}^o &= \gamma_1\sigma_{11}^o + \gamma_2\sigma_{12}^o + 2\gamma_{11}\mu_o^1\,\sigma_{11}^o \\ &\quad + \gamma_{12}\,(\sigma_{12}^o\mu_o^1 + \mu_o^2\sigma_{11}^o) + 2\gamma_{22}(\mu_o^2\sigma_{12}^o) \\ \mathrm{Cov}(y,X_2) = \sigma_{2y}^o &= \gamma_1\sigma_{12}^o + \gamma_2\sigma_{22}^o + 2\,\gamma_{11}\mu_o^2\sigma_{12}^o \\ &\quad + \gamma_{12}\,(\sigma_{12}^o\mu_o^2 + \mu_o^1\sigma_{22}^o) + 2\,\gamma_{22}(\mu_o^1\sigma_{12}^o) \end{aligned} \tag{5.4}$$

The cross section OLS coefficients of y on X_1 and X_2 from (2.3) consistently estimate $\frac{\partial\Phi}{\partial\mu}$ evaluated at $\mu = \mu_o$. Using the covariance formulae (5.4), this is directly verified as

$$\text{plim } \hat{b}_{y.12} = \beta_{y.12} = \begin{pmatrix} \beta_{y.1(2)} \\ \beta_{y.2(1)} \end{pmatrix} = (\Sigma^o)^{-1} \begin{bmatrix} \sigma^o_{1y} \\ \sigma^o_{2y} \end{bmatrix}$$

$$= \begin{bmatrix} \gamma_1 + 2\gamma_{11}\mu^1_o + \gamma_{12}\mu^2_o \\ \gamma_2 + 2\gamma_{22}\mu^2_o + \gamma_{12}\mu^1_o \end{bmatrix} \tag{5.5}$$

$$= \frac{\partial \Phi}{\partial \mu}(\mu_o)$$

For the regression (2.5) of y on X_1 only, we must characterize the exponential family with driving variable X_1 in correspondence with Assumption 3B. As can be easily verified, under this family the marginal distribution of X_1 is normal with varying mean μ_1 and fixed variance σ^o_{11}. The conditional distribution $p_2(X_2|X_1)$ is stable over time under this assumption and is given by a normal distribution with mean $E(X_2|X_1) = c + pX_1$ and variance $\sigma^o_{22} - (\sigma^o_{12})^2/\sigma^o_{11}$, where $p = \sigma^o_{12}/\sigma^o_{11}$.
Thus, under Assumption 3B, the mean of X_2 is given in terms of μ_1 as

$$\mu^2 = c + p\,\mu^1 \tag{5.6}$$

The macroeconomic function between μ^y and μ^1 under Assumption 3B is now

$$E(y) = \mu^y = \Phi_1(\mu^1) = \gamma_1\mu^1 + \gamma_2(c + p\,\mu^1) + \gamma_{11}((\mu^1)^2 + \sigma^o_{11})$$

$$+ \gamma_{12}(c\,\mu^1 + p(\mu^1)^2 + \sigma^o_{12}) \tag{5.7}$$

$$+ \gamma_{22}(c^2 + 2cp\mu^1 + p^2(\mu^1)^2 + \sigma^o_{22})$$

Under this assumption, the cross section OLS regression coefficient of y on X_1 consistently estimates $\frac{\partial \Phi}{\partial \mu^1}$ evaluated at $\mu^1 = \mu^1_o$. To verify this, use (5.4) as

$$\text{plim } \hat{b}_{y.1} = \beta_{y.1} = \frac{\sigma^o_{1y}}{\sigma^o_{11}}$$

$$= \gamma_1 + \gamma_2 \frac{\sigma^o_{12}}{\sigma^o_{11}} + 2\gamma_{11}(\mu^1_o) + \gamma_{12}\left(\frac{\sigma^o_{12}\mu^1_o}{\sigma^o_{11}} + \mu^2_o\right)$$

$$+ 2\gamma_{22}\left(\frac{\sigma^o_{12}}{\sigma^o_{22}}\ \mu^2_o\right) \qquad (5.8)$$

$$= \frac{\partial \Phi_1}{\partial \mu^1}(\mu^1_o)$$

the latter equality following from $\mu^2_o = c + p\mu^1_o$ and $p = \sigma^o_{12} / \sigma^o_{12}$; $c = \mu^2_o - p\mu^1_o$.

The large sample omitted variable bias formula (2.9) can easily be verified using formulae (5.5) and (5.9), while recognizing that the large sample value of the OLS auxiliary coefficient of X_2 on X_1 is $\text{plim } \hat{B}_{2.1} = B_{2.1} = \frac{\sigma^o_{12}}{\sigma^o_{11}} = p$, which similarly is the induced effect of μ^1 on μ^2 from (5.6).

6. Related Work

In this paper we have reinterpreted the standard omitted variable bias formula for cross section regression in the context of a general model between dependent and independent variables. We have utilized the macroeconomic interpretation of cross section regression coefficients to show how the omitted variable bias formula and auxiliary regression coefficients reflect induced constraints between the means of included and excluded regressors and their

macroeconomic effects on the dependent variable mean. Moreover, we have shown that the impact of excluding variables on the mean of the dependent variable can be measured in general by the partial contribution of the excluded variables to the cross section regression on y.

The "properly specified" regression for estimating macroeconomic effects was seen to be determined by the driving variables of the exponential family, or in more general terms, the score vector of distribution movements. The general theory of the score vector and its role in the cross section estimation of macroeconomic effects is given in Stoker (1983a), which establishes the correspondence between cross section regression and the efficiency properties of data aggregates. This general development yields a macroeconomic interpretation of cross section instrumental variables estimators, which is then extended and developed by Lewbel and Stoker (1983). Finally, the impact of structural changes in individual behavior on macroeconomic functions is treated in Stoker (1983b).

NOTES

1. As with any standard tool, to attempt to cite all references to the omitted variable bias formula would result in a bibliography much longer than this paper. For an introduction to the skillful use of the formula, the work of Zvi Griliches is strongly recommended - some good examples are Griliches (1957,1971) and Griliches and Ringstad (1971).

2. These relationships have a long history, dating back at least to Frisch (1934), and are included as standard material in textbooks on regression analysis - see Kendall and Stuart (1967) and Rao (1973) for example.

3. The terms "predictor variable" and "regressor" are used interchangeably to describe X in the regression $y = \hat{a} + \hat{b}X + \hat{\varepsilon}$.

4. This feature eliminates sample selection problems from our framework.

5. For a treatment of cross section regression and macroeconomic effects for general movements in the predictor variable distribution, see Stoker (1983a).

6. For standard textbook treatments of the exponential family, see Ferguson (1967) and Lehmann (1959). For modern treatment, see Barndorff Neilson (1978) and Efron (1978).

7. For derivatives of expectations taken over an exponential family, see Stoker (1982a), Lemma 6.

8. In other words, by altering the driving variables of the exponential family, one obtains a different sequence of marginal distributions of X through time, and if F(X) is nonlinear, a different set of macroeconomic effects.

9. In particular, if μ_1^1 and μ_1^2 are the means of X_1 and X_2 observed in a time period adjacent to the cross section, we would expect $\mu_1^2 - \mu_o^2 \tilde{=} \hat{B}_{2.1}(\mu_1^1 - \mu_o^1)$ under Assumption 3B.

10. For a theoretical treatment of mixed parameterizations of exponential families, see Barndorff Neilson (1978).

11. This idea suggests studying the possibility that y itself is a driving variable. This is pursued in Stoker (1983b), and gives rise to an interesting characterization of residual variance $\sigma_{y.12}$ in a general functional form framework.

12. These are easily found by differentiating and evaluating the moment generating function of X_1 and X_2.

REFERENCES

Barndorff-Neilson, O. (1978), Information and Exponential Families in Statistical Theory, Wiley, New York.

Efron, B. (1978), "The Geometry of Exponential Families," Annals of Statistics 6, pp. 362 - 376.

Ferguson, T.S. (1967), Mathematical Statistics, A Decision Theoretic Approach, Academic Press, New York.

Frisch, R. (1934), Statistical Confluence Analysis by Means of Complete Regression Systems, Oslo, Universitets Økonomiske Institutt.

Griliches, Z. (1957), "Specification Bias in Estimates of Production Functions," Journal of Form Economics, 39, 1, pp. 8 - 20.

Griliches, Z. (1971), "Hedonic Price Indexes for Automobiles: An Econometric Analysis of Quality Change," chapter 3 of Price Indices and Quality Change, Z. Griliches, ed., Harvard University Press, pp. 55 - 87.

Griliches, Z. and V. Ringstad, (1971), Economies of Scale and the Form of the Production Function: An Econometric Study of Norwegian Manufacturing Establishment Data, North Holland, Amsterdam.

Kendall, M.G. and A. Stuart (1967), The Advanced Theory of Statistics, Volume 2, Hafner Publishing Co., New York.

Lehmann, E.L. (1959), Testing Statistical Hypotheses, Wiley, New York.

Stoker, T.M. (1982a), "The Use of Cross Section Data to Characterize Macro Functions," Journal of the American Statistical Association, June, pp. 369 - 380.

Stoker, T.M. (1982b), "Completeness, Distribution Restrictions and the Form of Aggregate Functions," MIT Sloan School of Management Working Paper WP 1345-82, August.

Stoker, T.M. (1983a), "Aggregation, Efficiency and Cross Section Regression," MIT Sloan School of Management Working Paper No. WP 1453-83, June.

Stoker, T.M. (1983b), "Aggregation, Structural Change and Cross Section Regression," draft, July.

Theil, H. (1957), "Specification Errors and the Estimation of Economic Relationships," Review of the International Statistical Institute, 25, pp. 41 - 51.

Theil, H. (1971), Principles of Econometrics, John Wiley and Sons, Amsterdam.