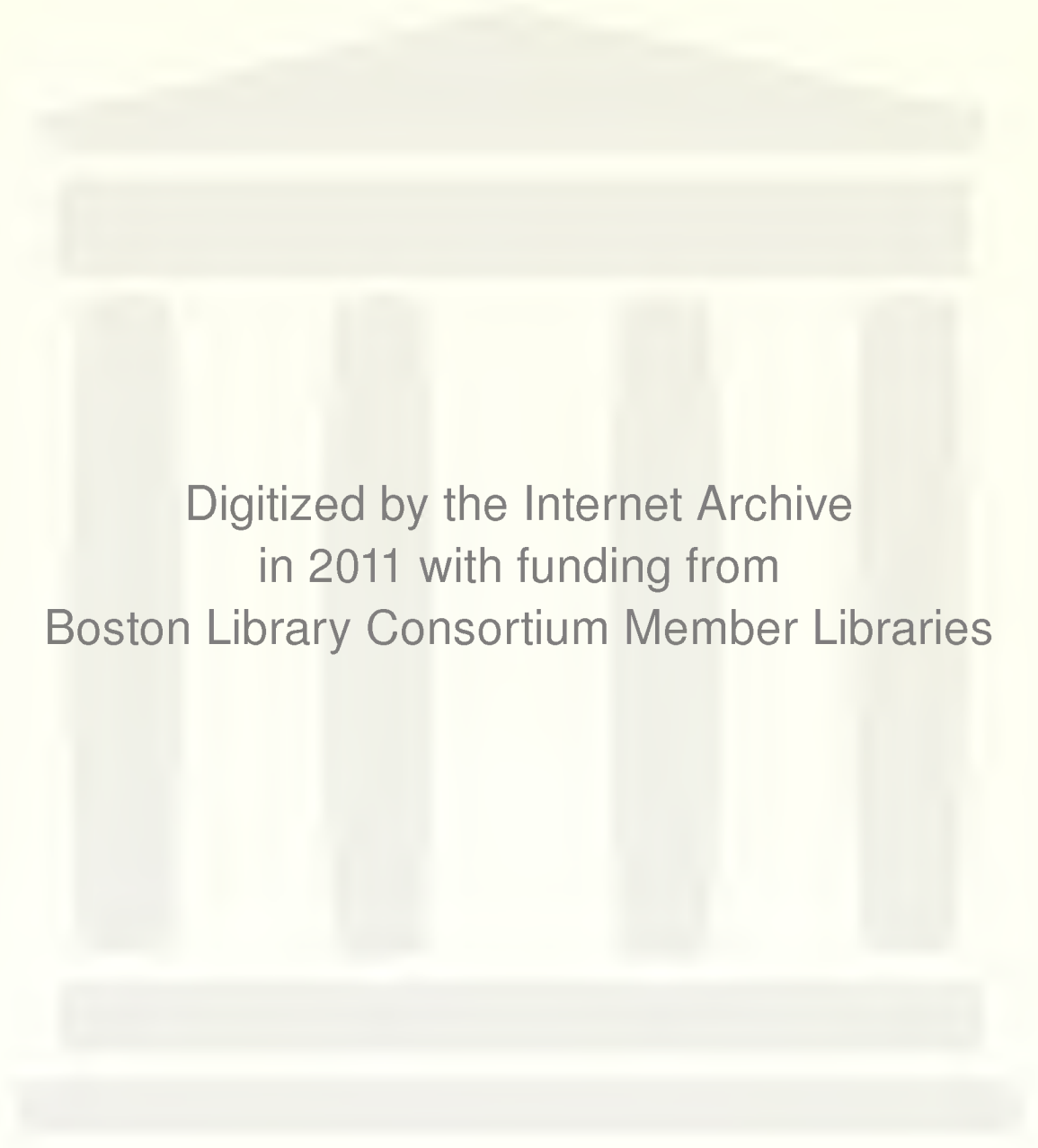

working paper
department
of economics

NONLINEAR ERRORS IN VARIABLES: ESTIMATION OF SOME ENGEL CURVES

J. A. Hausman
W. K. Newey
J. L. Powel

No. 504 November 1988

massachusetts
institute of
technology

50 memorial drive
cambridge, mass. 02139

NONLINEAR ERRORS IN VARIABLES: ESTIMATION OF SOME ENGEL CURVES

J. A. Hausman
W. K. Newey
J. L. Powel

No. 504 November 1988

Revised Draft
November 1, 1988
Comments Welcome

Nonlinear Errors in Variables: Estimation of Some Engel Curves

by J.A. Hausman, W.K. Newey, and J.L. Powel

1988 Jacob Marschak Lecture of the Econometric Society

Australian Economics Congress

Canberra, Australia

August 31, 1988

Nonlinear Errors in Variables: Estimation of Some Engel Curves

by J.A. Hausman, W.K. Newey, and J.L. Powell[1]

The errors in variables problem has been long known in statistics; Adcock (1878) is perhaps the first reference which points out the problem. In the simple bivariate regression model the result of errors in variable is a downward bias (in magnitude) of the estimated regression coefficient: the "iron law" of econometrics as known to MIT students.[2] During the formative period of econometrics in the 1930's, considerable attention was given to the errors in variable problem. However, with the subsequent emphasis on aggregate time series research the errors in variables problem decreased in importance to most econometric research. In the past decade as econometric research on micro data has increased dramatically, the errors in variables problem has once again moved to the forefront of econometric research.[3]

Solutions to the errors in variables problem for the linear regression model have been well explored and are often used by econometricians. The most common solution is the use of instrumental variable estimation (IV) which depends on the existence of an appropriate instrument or repeated observation of the

---

[1] MIT, Princeton University, and University of Wisconsin. We thank Greg Leonard for excellent research assistance and the National Science Foundation for financial support. A. Deaton, A. Lewbel, R. Pollak, D. Jorgenson and J. Poterba made helpful suggestions. Presented as the Jacob Marschak Lecture of the Econometric Society at the 1988 Australian Economics Congress.

[2] The notion of the "iron law" is that the estimated effect is (almost) never as large as economic theory or the applied researcher expects it to be. Of course, in the multiple regression situation with many right hand side variables the result need no longer hold true. Nevertheless, the folklore in econometrics plus years of reading students econometrics papers results in the belief that downward bias in coefficients estimates is a pervasive problem in micro data parameter estimates.

[3] Griliches (1986) discusses micro data problems which lead to errors in variables problems in many typical econometric data sets.

variable measured with error.[4] Two other solutions exist, but they have only infrequently been used by econometricians. The first alternative solution involves knowledge or an estimate of the variance of the measurement error(s) of the right hand side variable(s) or its relative size compared to the variance of the stochastic disturbance, which can be partly or entirely composed of the measurement error in the left hand side variable. This type of knowledge is usually not available to econometricians. The second alternative solution is to use distributional properties of the right hand side variables or to use higher order moments which depend on distributional assumptions, beyond the first two moments, to estimate the parameters. This approach again has only rarely been used by econometricians.[5]

Thus, the IV appropach is by far the most widely used technique for dealing with errors in variables problems in linear multiple regression problems. The linear model with measurement error is isomorphic to a linear simultaneous equation model, so that two stage least squares or a closely related estimator is

---

[4] Zellner (1970) and Goldberger (1972) extend the single equation errors in variables problem to the multiple equation context. Geraci (1977), Hausman (1977) and Hsiao (1976) consider the errors in variables problem in the simultaneous equations situation. An excellent survey is given by Aigner, Hsiao, Kapteyn, and Wansbeek (1984).

[5] Reiersol (1950) demonstrated identification of the errors in variables problem using distributional assumptions. Kapteyn and Wansbeek (1983) generalize the result to the multiple regression context. Bickel and Ritov (1987) apply the method in an adaptive estimation framework. Geary (1942) originally proposed using higher order moments in estimation in the errors in variables problem. Higher order moments may offer a useful methodology given the increasingly large data sets of many thousands of observations used by econometricians. They can be applied in a straghtforward method of moments procedure. However, the technique has been little used to date and J. Hausman has been unsuccessful in a few previous attempts.

used.[6] However, this relationship no longer holds in the nonlinear regression framework as recently noted by Y. Amemiya (1985). The reason that 2SLS no longer leads to a consistent estimator in the nonlinear errors in variables problem is because the error of measurement is no longer additively separable from the true variable in the nonlinear regression model. Application of 2SLS or nonlinear 2SLS (N2SLS) leads to inconsistent estimates.[7]

A staightforward way in which to see why 2SLS or N2SLS does not yield consistent estimators in the nonlinear errors in variable model is to consider the linear in parameters and nonlinear in variables specification:

$$(1.1) \quad y_i = \beta_0 + \beta_1 g(z_i) + \epsilon_i \qquad i = 1,\ldots,n$$

where g(z) is a sufficiently smooth function to do Taylor approximations. As in the linear errors in variables framework, we assume that z is unobservable; instead the observed varible x_i takes the form:

$$(1.2) \quad x_i = z_i + \eta_i \qquad i = 1,\ldots,n$$

where η_i is assumed to be uncorrelated with z_i. Replacing the unobservable z with x in equation (1.1) and taking a Taylor expansion leads to:

---

[6] Fuller (1987) discusses many of these other estimators which may improve the finite sample performance of IV-type estimators in the errors in variables context.

[7] The failure of additive separability also arises in the reduced from of the nonlinear simultaneous equations problem where the additive stochastic disturbance of the structural form enters nonlinearly into the reduced form. This situation leads to N2SLS and N3SLS being inefficient relative to ML in the nonlinear simultaneous equations model as demonstrated by T. Amemiya (1977).

$$(1.3)\quad y_i = \beta_0 + \beta_1 g(x_i) + \epsilon_i - \beta_1 g^1(x_i)\eta_i - \beta_1 \sum_{j=2}^{\infty} g^{[j]}(x_i)\,\eta_i^j/j!$$

where [j] denotes the jth derivative of g. Inspection of the first term of the Taylor expansion in equation (1.3) demonstrates the fundamental problem with N2SLS or other IV techniques. The instrument must be correlated with $g(x_i)$, but be uncorrelated with η_i and ϵ_i. However, the first term of the Taylor expansion contains both η_i and the derivative of $g(x_i)$. In the linear errors in variables framework the first derviative of $g(x_i)$ is unity so the observation error is linearly separable from the right hand side variable. This linear separation is not present in equation (1.3) so that, even if the higher order terms of the Taylor expansion were absent, it is unlikely that an appropriate instrumental variable would exist. However, the additional factor of the added terms in the Taylor expansion beyond the first make the problem even more unwieldy to solve with the usual instrumental variable techniques.

To date the methods proposed to estimate the nonlinear errors in variable model depend on very strong restrictions on the distribution of the measurement errors of the unknown regression coefficients. However, knowledge of the parametric form of the distribution function of the measurement errors is not sufficient for consistent estimation. An additional assumption is needed that the true values of the regressors, e.g. the z_i in equation (1.1), are also assumed to be random drawings from a distribution with a known parametric form. Instead, if the true regressors are treated as fixed but unknown constants, then maximum likelihood estimation is inconsistent due to the "incidental parameters" problem of Neyman and Scott (1948).[8]

---

[8] Aigner et. al. (1984) have a brief discussion on ML for nonlinear errors in variables models.

An alternative approach assumes that a large number of measurements on each true regressor exist, so that the average of these measurements closely approximates the true regressors. Consistent estimation of nonlinear errors in varibles models then follows because the covariance matrix of the measurement errors for the regressors approaches zero as the sample size increases. Estimators under this type of assumption have been proposed by Villegas (1969), Dolby and Lipton (1972), Wolter and Fuller (1982b), Powell and Stoker (1984), and Y. Amemiya (1985). This situation seems unlikely to occur very often in econometrics.

Lastly, Griliches and Ringstad (1970) analyzed a quadratic specification and demonstrated that the bias of least squares can be exacerbated by the nonlinearity. Wolter and Fuller (1982a) propose a consistent estimator for the quadratic specification so long as the errors are normally distributed. Neither instrumental variables nor additional measurements are required for estimation.

In this paper we discuss consistent estimators for nonlinear regression specifications when errors in variables are present. Our estimators depend on the existence of instrumental variables or a single repeated observation. Thus we do not require the large number of measurements or shrinking covariance matrix assumption of much previous research.

In Section 2 we discuss an estimator for polynomial specifications in the true regressors. This estimator proposed by Hausman, Ichimura, Newey, and Powell (HINP)(1986) leads to consistent and asympotically normal estimators so long as either instrumental variables or an additional measurement of each true regressor are present. An interesting result emerges in the instrumental

variable case because the model turns out to be overidentified. Thus, tests of the model specification are possible.

In Section 3 we discuss an estimator for the general nonlinear specification when errors in variables are present. This estimator proposed by Hausman, Newey, and Powell (HNP)(1988) yields a consistent estimator when an additional measurement of each true regressor is present. To date, we have not been able to establish asymptotic normality of the estimator or to extend it to the instrumental variable situation. However, Monte Carlo evidence provides some indication that the distribution of the estimator is not badly behaved so that we use bootstrap estimates of the precision of our estimates.

In Section 4 we apply our methodology to estimation of Engel curves on household data, a problem which econometricians have done considerable previous research on. Here we find a number of interesting results. First, we find that the "Leser-Working" specification of budget shares regressed on the log of income or expenditure should be generalized to higher order terms in log income. Also, we find that errors in variables in either reported income or expenditure should be accounted for. However, we do not find evidence that more general functional forms beyond polynomial specifications in income improve estimation of the Engel curve significantly. Lastly, and perhaps most interesting, we find rather strong support for the Gorman (1981) rank restriction on the matrix of coefficients for the polynomial terms in income. Thus, after over 100 years of Engel curve analysis, a restriction from economic theory may affect the econometric estimation. This result is remarkable if future research leads to similar findings.

II. Identification and Estimation of the Polynomial Functional Model

We first consider estimation of the parameters of the polynomial specification

$$(2.1) \quad y_i = \sum_{j=0}^{K} \beta_j (z_i)^j + \epsilon_i \qquad i = 1,\ldots n$$

which is a kth order polynomial in the unobservable variable z_i. We will treat z_i as a random variable with unknown distribution function; alternatively, the $\{z_i\}$ can be interpreted as a sequence of fixed constants with appropriate modifications in the regularity conditions. The observed variable x_i has the same relationship to the unobserved variable z_i as in Section I:

$$(2.2) \quad x_i = z_i + \eta_i \quad i = 1,\ldots,n.$$

Our first estimator uses the additional information of a single repeated measurement w_i of z_i with an additional measurement error v_i defined by

$$(2.3) \quad w_i = z_i + v_i \quad i = 1,\ldots,n.$$

Two points of interest arise from the specification and assumptions of equation (2.3). First, we will assume that v_i is uncorrelated with ϵ_i and η_i and is independent of z_i. The independence assumption is required by the nonlinearity. These assumptions are analogous to the linear case so that w_i could be used as an instrumental variable if the specification of equation (2.1) were linear. Second, we will not impose the usual restriction $E(v_i) = 0$ so that a constant term can be present in the measurement equation (2.3). Alternatively, v_i can be assumed to have zero mean, but the slope coefficient of z_i in equation

(2.3) can then be non-unity. The details of the derivation of the estimator in this latter case is left to the interested reader.

We now turn to sufficient regularity conditions to allow identification and estimation of the β_j in equation (2.1). Define the matrix norm $||A|| = \max_{i,j}|a_{ij}|$. We make

Assumption 1: The random variables ϵ_i, η_i, v_i, and z_i are jointly i.i.d. with

(i) $E(\epsilon_i \mid z_i, v_i) = E(\eta_i \mid z_i, v_i) = 0$

(ii) v_i is independent of z_i

(iii) $E\,||(\epsilon_i, \eta_i, v_i^{2K}, z_i^{2K})||^2 < \infty$

(iv) All necessary moment matrices are nonsingular.

The i.i.d. assumption can be relaxed to allow for either dependence or heterogeneity or both. Assumptions A.1(i), (iii), and (iv) are standard assumptions to allow derivation of both identification and the asymptotic distribution of the estimator. Only assumption A.2(ii) is stronger than in the usual linear case. Independence is necessary because of the nonlinear specification. Note that assumption A.1(i) could be strengthened to independence for purposes of symmetry; we require only the weaker no correlation assumption.

For identification we consider the population analogue of the normal equations. Define the moments $\xi_p = E[y_i (z_i)^p]$ for $p = 0,\ldots,K$ and $\Phi_m = E[(z_i)^m]$ for $m = 0,\ldots,2K$. The normal equations $z'y = (z'z)\beta$ take the form

$$(2.4) \quad \xi_p = \sum_{j=0}^{K} \beta_j \Phi_{j+p} \qquad p = 0,\ldots,K$$

Both sides of equation (2.4) depend on the unobservable variables z_i. However, the unobservable moments can be derived from the observable moments $E[x_i (w_i)^P]$, $E[(w_i)^P]$, and $E[y_i (w_i)^P]$. We now use assumption A.1 and the fact that $E[(w_i)^0] = E[(z_i)^0] = E[(v_i)^0] = 1$ to find:

$$(2.5) \quad E[x_i (w_i)^{j-1}] = \sum_{p=0}^{j-1} \binom{j-1}{p} \Phi_{p+1} \nu_{j-p-1} \quad \text{for } j = 1,\ldots,2K \text{ where } \nu_j = E[(v_i)^j].$$

$$(2.6) \quad E[(w_i)^j] = \sum_{p=0}^{j} \binom{j}{p} \Phi_p \nu_{j-p} \quad \text{for } j = 1,\ldots,2K.$$

$$(2.7) \quad E[y_i (w_i)^j] = \sum_{p=0}^{j} \binom{j}{p} \xi_p \nu_{j-p} \quad \text{for } j = 0,\ldots,K.$$

Equations (2.5) and (2.6) allow identification of $z'z$, and equation (2.7) then allows identification of $z'y$. That is, equations (2.5)-(2.7) defined $(5K + 1)$ equations which have a one-to-one relationship between the moments of the observable variables and the $(5K + 1)$ elements of the unobservable moment vectors Φ and ν each of which have $2K$ elements and ξ which has K elements. HINP (1986) derive recursive relationships which permit convenient solution of the elements of the parameter vector $\theta = (\Phi', \nu', \xi')$. Once θ is computed, β is then identifiable as a solution to the normal equations (2.4).

To derive the asymptotic distribution of the estimator define the $(5K + 1)$ dimensional data vector

$$(2.8) \quad m_i = [x_i,\ldots,x_i(w_i)^{2K-1}, \; w_i,\ldots,(w_i)^{2K}, \; y_i,\ldots,y_i(w_i)^K].$$

Define the population moments to be $\mu = E[m_i]$. The moment vector μ is

consistently estimated by the sample average moment vector $\hat{m}$ and application of the Lindeberg-Levy CLT yields the asymptotic distribution of $\hat{m}$

$$(2.9) \quad \sqrt{n}\,(\hat{m} - \mu) \xrightarrow{d} N(o, \Omega) \quad \text{for } \Omega = E[m_i m_i'] - \mu\mu'.$$

The elements of θ are then estimated by the continuous and continuously diffentiable relationship $\theta = h(\mu)$. First order approximations, also known as the delta method, lead to the asymptotic distribution of θ

$$(2.10) \quad \sqrt{n}\,(\hat{\theta} - \theta) \xrightarrow{d} N(0, H\Omega H') \text{ for } H = \partial h(\mu)/\partial \mu'.$$

The elements of the Jacobian matrix H can be calculated recursively with computational details given in HINP (1986).

Lastly, we solve for β using the normal equations (2.4) and the estimated θ. Let D be the second moment matrix of $(1, z_i, \ldots, (z_i)^K)$ and $\hat{D} = D(\hat{\theta})$ based on the estimated θ. We estimate β by

$$(2.11) \quad \hat{\beta} = \hat{D}^{-1}\,\hat{\xi}.$$

Define S_ξ as the selection matrix which gives $S_\xi\,\theta = \xi$ and S_Φ as the selection matrix which gives $S_\Phi\,\theta = \text{vec}(D)$. HINP (1986) derive the asymptotic distribution

$$(2.12) \quad \sqrt{n}\,(\hat{\beta} - \beta) \xrightarrow{d} N(0, V) \text{ where}$$

$$V = D^{-1}\,[\,S_\xi - (\beta' \otimes I_{K+1})\,S_\Phi\,]\,H\,\Omega\,H'\,[S_\xi - (\beta' \otimes I_{K+1})\,S_\Phi]'D^{-1}$$

We have demonstrated identification and derived a consistent and asymptotically

normal estimator in the case of a single replicated measurement for the unobservable variable z_i.

We now turn to identification and estimation when instrumental variables which allow prediction of the unobserved regressor z_i are available. Thus, we assume that z_i is determined by the p dimensional vector of instruments q_i

$$(2.13) \quad z_i = q_i\alpha + v_i \qquad i = 1,\ldots,n.$$

To state sufficient conditions for identification and estimation we change assumption A.1.(ii) to an assumption that v_i and the instruments q_i are independent. Otherwise the sufficient conditions are quite similar to the previous case that we considered:

Assumption 2: The random variables ϵ_i, η_i, v_i, and q_i are jointly i.i.d. with

(i) $E(\epsilon_i \mid q_i, v_i) = E(\eta_i \mid q_i, v_i) = 0$, $E(\epsilon_i\eta_i \mid q_i, v_i) = \sigma_{\epsilon\eta}$

(ii) v_i is independent of q_i with $E[v_i] = 0$

(iii) $E[||(\epsilon_i, \eta_i)||^2] < \infty$, $E[||v_i, q_i||^{2(K+1)}] < \infty$

(iv) All necessary moment matrices are nonsingular.

Again, the i.i.d. assumption can be relaxed to more general situations.

For purposes of identification we assume that α is known since it is identified from

$$(2.14) \quad x_i = q_i\alpha + \eta_i - v_i \qquad i = 1,\ldots,n.$$

Let $w_i = q_i\alpha$ and again denote $\nu_j = E[(v_i)^j]$. We must again determine the ν_j for identification and estimation.

Substitution of the instrumental variables into equation (2.1) yields:

$$(2.15)\quad (i)\qquad y_i = \sum_{j=0}^{K} \gamma_j\,(w_i)^j + e_i \quad \text{where}$$

$$(ii)\qquad \gamma_j = \sum_{p=j}^{K} \binom{p}{j} \beta_p\,\nu_{p-j} \qquad j = 0,\ldots,K$$

$$(iii)\qquad e_i = \epsilon_i + \sum_{j=0}^{K} \sum_{p=j}^{K} \binom{p}{j} \beta_p\,[(v_i)^{p-j} - \nu_{p-j}]\,(w_i)^j$$

Equation (2.15) (iii) implies that $E(e_i \mid w_i) = 0$ so that γ is identified by the least equares projection of equation (2.15) (i).

We now have the convolution of β and ν in γ which must be sorted out for identification. Before proceeding to do so, note that since $\nu_0 = 1$ and $\nu_1 = 0$, we have $\gamma_k = \beta_k$ and $\gamma_{k-1} = \beta_{k-1}$. Thus, the two highest elements of β are identified from equation (2.15) (i) alone which will subsequently lead to overidentification.

To complete the identification, we now multiply through equation (2.1) by the observable variable x_i and we substitute $z_i = w_i + v_i$:

$$(2.16)\quad (i)\qquad x_iy_i = \sum_{j=0}^{K+1} \delta_j(w_i)^j + u_i \quad \text{where}$$

$$(ii)\qquad \delta_0 = \sum_{p=0}^{K} \beta_p\,\nu_{p+1} + \sigma_{\xi\eta}$$

$$\delta_j = \sum_{p=j}^{K+1} \binom{p}{j} \beta_{p-1}\,\nu_{p-j} \qquad j = 1,\ldots,K+1$$

$$(iii) \qquad u_i = \sum_{j=0}^{K+1} \sum_{p=j}^{K+1} \binom{p}{j} \beta_{p-1} \left[(v_i)^{p-j} - \nu_{p-j} \right] (w_i)^j + \left[\eta_i y_i - \sigma_{\xi\eta} \right]$$

$$+ z_i \epsilon_i \qquad i = 1, \ldots, n.$$

Again, the disturbance term in equation (2.16) (i) has zero conditional expectation so that $E[u_i \mid w_i, v_i] = 0$. The estimate of δ follows from the least squares projection of equation (2.16) (i). We can again identify the two highest order terms of β from $\delta_{K+1} = \beta_K$ and $\delta_K = \beta_{K-1}$. Overidentification of these two parameters follows from the γ and δ coefficients.

Thus, we have (2K + 3) reduced form coefficients $\gamma_0, \ldots, \gamma_K$, $\delta_0, \ldots, \delta_{K+1}$. We have K + 1 unknown β coefficients and K unknown nuisance parameters in ν. Thus, we have overidentification of order 2, or equivalently, we can discard two equations and still identify the unknown parameters. The solution to the equations once again follows a convenient recursion relationship. HINP (1986) give the recursion formulae.

Estimation proceeds from initial estimation of the reduced form parameters γ and δ. Given γ and δ, we then estimate β and ν. This two step procedure need not be asymptotically efficient; the topic is left to future research. The derivation of the asymptotic distribution of the resulting estimator is straightforward, but tedious. We give only a sketch here and direct the interested reader to HINP (1986) who give the complete derivation. Taking account of the fact that α must be estimated, the asymptotic distribution of the reduced form parameter is normal, say

$$(2.17) \qquad \sqrt{n} \begin{bmatrix} \hat{\gamma} - \gamma \\ \hat{\delta} - \delta \end{bmatrix} \xrightarrow{d} N \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, M \right\} .$$

Given equation (2.17) we can then obtain efficient estimates of the β's by minimum chi-square estimation. Denote the (2K + 2) vector of reduced form coefficients, after elimination of δ_0 by

$$(2.18)\quad \hat{\Pi} = (\hat{\gamma},\ \hat{\delta})$$
$$= (\hat{\Pi}_1,\ \hat{\Pi}_2) \text{ where } \hat{\Pi}_1 = (\hat{\gamma}_K,\ \hat{\gamma}_{K-1},\ \hat{\delta}_{K+1},\ \hat{\delta}_K).$$

Similarly, denote the 2K vector of b and ν parameters as

$$(2.19)\quad \Theta = (\beta,\ \nu)$$
$$= (\Theta_1,\ \Theta_2) \text{ where } \Theta_1 = (\beta_K,\ \beta_{K-1}).$$

The unknown Θ parameters follow from the reduced form Π parameters

$$(2.20)\quad \Pi = h(\Theta)$$

so rearrange the covariance matrix M from equation (2.17) to conform to equation (2.20) and denote a consistent estimate of the rearranged matrix, V.
The minimum chi square estimator is then

$$(2.21)\quad Q = \min_{\Theta} \left[\hat{\Pi} - h(\Theta)\right]' \hat{V}^{-1} \left[\hat{\Pi} - h(\Theta)\right].$$

The value of Q from equation (2.21) can be used to test the overidentifying restrictions since it is distributed as a central chi square random variable with 2 degrees of freedom if the specification is correct. The test of identification

is equivalent to a test of equation (2.2): a non-zero mean and non-unity slope coefficient of z_i cause the system to be just identified.[9] The asymptotic covariance matrix of the minimum chi square estimator takes the usual form

$$(2.22) \quad \sqrt{n} \begin{bmatrix} \hat{\beta} - \beta \\ \hat{\nu} - \nu \end{bmatrix} \xrightarrow{d} N(0, (HV^{-1}H')^{-1}) \text{ where } H = \partial h(\Theta)/\partial \Theta'.$$

The specification of equation (2.1) does not contain other variables besides the polynomial terms. However, in many applications we might expect the appropriate specification to be

$$(2.23) \quad y_i = \sum_{j=0}^{K} \beta_j (z_i)^j + R_i\phi + \epsilon_i \qquad i = 1,\ldots,n.$$

where we assume that the R_i are measured without error. The usual partialing out technique does not work for equation (2.23) because of the nonlinear errors in variables specification. Instead, we apply two different approaches for the replicated measurement and instrumental variables setups. The additional R_i variables are accounted for in the replicated measurement case by considering equation (2.4), the normal equations. We need to augment z'y and z'z to include R. Thus we need to form the matrices R'z and R'y. The latter matrix is directly observable. The matrix R'z depends on the unobservable variable z; however,

---

[9] Because Θ_2 is just identified, equation (2.21) may be further simplified using partitioned inverses. Θ_2 follows from Π_2, while the overidentified parameters Θ_1 are estimated from Π_1. See HINP(1986) for computational details.

equation (2.5) with R_i included permits computation of an estimate of the required moment matrix.[10]

Inclusion of additional regressors in the instrumental variables case is also quite straightforward. The addition of $R_i\phi$ to equation (2.15) (i) has no effect on the disturbance e_i so that γ_j in equation (2.15) (ii) remains the same. Similarly, δ_j is estimated from equation (2.16) (i) after the term $w_iR_i\phi$ is added to the right hand size of the equation. In both cases least squares projections yield consistent estimates of γ_j and δ_j so that estimates of β and ν can be calculated from the estimated reduced form parameters.

In this section we have proven identification and developed consistent estimators for the polynomial specification when either a single replicated measurement or instrumental variables are present. Additional replicated measurements can be included easily; a minimum chi square combination of the estimated parameters offers a convenient approach. Similarly, a replicated measurement and instrumental variable situation can be combined using a minimum chi square approach. In both cases we increase the efficiency of our estimator, or alternatively, we can test the specification of our model. However, we do not claim to have found the most efficient estimator since there exists an infinite class of unconditional moment restrictions which can, in principle, be used in estimation. We leave the construction of feasible efficient estimators which attain an efficiency bound as a topic for future research.

---

[10] This approach is equivalent to using the transformed replicated measurement $\tilde{w} = (I - R(R'R)^{-1}R)w$ together with equations (2.4)-(2.7).

III. Errors in Variables in a General Nonlinear Specification

We now discuss identification and consistent estimation of a general nonlinear errors in variables specification with errors in variables following Hausman, Newey, and Powell (1988). Our estimator is limited to the replicated measurement case; we have been unable to extend estimation as yet to the instrumental variables case. Also, we do not currently have an asymptotically normal distribution for the estimator. We lack the centering correction for the estimator which would permit derivation of the asymptotic distribution. However, limited Monte Carlo investigations indicate that the bias of the estimator is small and that bootstrap estimates of the sampling distribution of the estimator provide a good indication of the precision of the estimates.

We consider the general nonlinear regression model with additive disturbance:

$$(3.1) \quad y_i = f(z_i, \beta) + \epsilon_i \qquad i = 1,\ldots,n.$$

where we take z to be a scalar. Inclusion of additional variables measured without error is straightforward using the approach of the last section. The variable z_i is unobservable; instead we observe x_i as in equation (2.2). The replicated measurement w_i is determined similarly by equation (2.3). Lastly, we make assumption A.1 (i)-(ii) of Section II where the assumption that v_i is independent of z_i is crucial for our estimator. The moment assumptions that we make are

Assumption 3: $E[|\epsilon_i|^2]$ and $E[|f(z_i, \beta)|^2]$ are finite and there exists

$T>0$ such that $E[\exp\{T|(z_i, \eta_i, v_i)|\}]$ is finite.

Using these assumptions and the results of Section II, we can estimate the coefficients of the linear projection of y_i on polynomials of the true, but unobservable, regressors.

We denote this polynomial approximation using the estimated moments for the normal equations (2.4) as

$$(3.2) \quad \hat{\xi}_p = \sum_{j=0}^{K} \hat{\Pi}_{jK} \hat{\Phi}_{j+p} \qquad p = 0,\ldots K.$$

where $\Pi(K)$ stands for a Kth order polynomial. Once we have the estimated $\Pi(K)$ projection coefficients we can estimate the true regression function $f_0(z) = f(z,\beta)$ by

$$(3.3) \quad \hat{f}_K(z) = \sum_{j=0}^{K} \hat{\Pi}_{jK} (z^j)$$

which is a nonparametric estimate of the regression function. Equation (3.3) provides an estimate of the true least squares projection

$$(3.4) \quad f_K(z) = \sum_{j=0}^{K} \Pi_{jK} (z^j)$$

of y_i on $z_i(K)$ and also provides an estimate of the least squares projection of $f_0(z_i)$ on $z_i(K)$ because e_i has zero mean conditional on z_i by Assumption A.1 (i). Furthermore, existence of the moment generating function of z_i, given assumption A.3, is sufficient for polynomials to form a basis for the space of square integrable functions, which implies that

$$(3.5) \quad \lim_{K\to\infty} E\,[\{f_0(z_i) - f_K(z_i)\}^2] = 0$$

For K large enough $f_K(z_i)$ provides an arbirarily good mean square approximation of $f_0(z_i)$.

Given the nonparametric estimate of the true regression function, $f_K(z)$, we estimate β by imposing the restrictions implied by the parametric model on $f_K(z)$. We use a minimum distance approach and obtain β from

$$(3.6) \quad \hat{\beta} = \text{argmin}_{\beta \epsilon B} \sum_{i=1}^{n} \omega(x_i) [\hat{f}_J(x_i) - f(x_i, \beta)]^2$$

where $\omega(x_i)$ is a nonnegative weight function and B is a feasible set of parameter values. Note that the observed variable x_i is used in equation (3.6) in place of the unobserved variable z_i. Other observed variables could also be used which could lead to more efficient estimators. We restrict our attention here to the x_i with an analysis of other variables left to future research. The purpose of the weight function $\omega(x_i)$ is to take account of the substitution of x_i for the unobservable z_i so that the mean square approximation holds

$$(3.7) \quad \lim_{K \to \infty} E [\omega(x_i) \{f_0 (x_i) - f_K (x_i)\}^2] = 0$$

Also, we require regularity conditions and an identification assumption to proceed

Assumption 4: (i) B is compact. (ii) $f(x_i, \beta)$ is continuous at each β in B with probability one. (iii) There exists d(.) such that $\sup_{\beta \epsilon B} | f(x_i, \beta) | \leq d(x_i)$ and $E [d(x_i)^2]$ is finite. (iv) for all β in B such that $\beta \neq \beta_0$, $E [\omega(x_i)\{f(x_i, \beta_0) - f(x_i, \beta)\}^2] > 0$ where β_0 is the true β.

Our last assumption imposes a condition on the weight function

Assumption 5: The distributions of z_i and x_i are absolutely continuous with densities $g_Z(.)$ and $g_X(.)$ respectively such that there is a positive constant C with $\omega(.)g_Z(.) < C\ g_X(.)$.

While estimation of the weight function may sometimes require careful attention, in the common case where the density of z_i is continuous and nonzero everywhere and the the density of x_i is bounded, then any $\omega(x_i)$ which is zero outside some bounded set will suffice.

Hausman, Newey and Powell (1988) then prove that given the assumptions and the additional condition that the density of z_i is bounded away from zero on an interval, then β determined from equation (3.6) is consistent, plim $\beta = \beta_o$, if K which is chosen as a function of the sample size K(n) has the properties that $K(n) \to \infty$ and $K(n)^2\ \ln[K(n)]/\ln(n) \to 0$. Note that the growth rate for K is somewhat slower than the square root of the natural log of the sample size.

In this section we have discussed a consistent estimator for the general nonlinear errors in variables specification. We now apply this estimator together with the estimators of Section II to estimate Engel curves on micro data. Derivation of an estimator with two or more mismeasured variables and derivation of the asymptotic distribution of the estimator of this Section are both quite complicated problems which we defer to future research.

IV. Estimation of Some Engel Curves

Estimation of Engel curves has long been an area of interest among econometricians. Many of the early investigations used British data, and the detailed cross section information collected in the annual Family Expenditure Surveys has led to considerable further investigation. Many of these studies investigate the best specification for the form of the Engel curves; Prais and Houtthaker (1955,1971) and Leser (1963) are notable examples. The "Leser-Working" form of Engel curve in which budget shares are regressed on the log of income or expenditure has been widely adopted in recent research. Both the translog specifications of Engel curves, e.g. Jorgenson, Lau and Stoker (1982), and the AIDS specification of Deaton and Muellbauer (1980) use this specification. An alternative specification, the quadratic expenditure system, which specifies budget shares as a function of both the inverse of expenditure and it square has been estimated by Pollak and Wales (1980).

Economic theory gives almost no general guidance in specification of Engel curves. "Adding-up" of budget shares to one is the only restriction, and this restriction is typically enforced in the data. However, in a notable paper Gorman (1981) considered Engel curves in which either expenditure or budget shares are specified as polynomials in functions of expenditure, e.g. log of expenditure. Gorman makes the key assumption, as does most of the previous demand curve literature, that the polynomial functions which contain expenditure do not depend on price in the demand curve specifications. Given this "exactly aggregable function", Gorman demonstrates that the rank of the matrix of coefficients for the polynomial terms in income is at most three.[11] We

---

[11] Lewbel (1986,1987) further considers Gorman's results for additional Engel curve specifications.

investigate Engel curve specifications of the Gorman form and provide tests of his rank three restriction.

Few studies of Engel curves have used estimators other than least squares or nonlinear least squares. The most notable exception is Liviatan (1961). Liviatan noted that Friedman's (1957) relabelling of the classic errors in variables model into "permanent" income and "transitory" income made the use of current income as a predetermined variable inappropriate in family budget studies. Liviatan also noted Summer's (1959) objection to the use of current expenditure as the predetermined variable because of reasons of joint endogeneity. He used instrumental variables with current income used as an instrument for current expenditure.[12] Livitan's assumption that current income is uncorrelated with the stochastic disturbances in an Engel curve specification seems highly questionable. He justified the assumption based on Friedman's assertion that "permanent" income and "transitory" consumption are uncorrelated with each other. However, subsequent econometric research, e.g. Attfield (1978), has demonstrated that the Friedman assumption is unlikely to hold true in family budget data. Thus, alternative instrumental variables are necessary. Here we investigate two alternative sets of instruments: expenditure in other periods or determinants of income and expenditure such as education and age. Neither of these alternative sets of instrumental variables should be correlated with the stochastic disturbance in the Engel curve specifications, although we test the assumptions subsequently.

---

[12] Liviatan used IV on a linear Engel curve specification. Leser (1963) applied a variant of Liviatan's procedure. However, straightforward IV is inapplicable to all of Leser's Engel curve specifications, except his first two linear specifications, because of the nonlinearity of his specifications. Inconsistent estimates will result for reasons discussed in Section I. In particular, Leser's best fitting specification (1963, p. 699) contains terms in both log expenditure and the inverse of expenditure which makes application of regular IV inappropriate.

We first consider a quadratic generalization of the Leser-Working Engel curve specification where the budget share of commodity i is a function of both log expenditure and the square of log expenditure:

$$(4.1) \quad w_i = \beta_0 + \beta_1 \log(z) + \beta_2 \log^2(z) + \epsilon_i$$

where ϵ_i is the stochastic disturbance. However, we do not observe actual expenditure, but we instead have data on $\log x = \log z + \eta$ where we assume that the error of observation satisfies the properties of Assumption A.1 (ii). Alternatively, a permanent income explanation can be attached to equation (4.1); however, we are unwilling to make any assumption about the relationship of permanent income and transitory consumption. Note that equation (4.1) satisfies the Gorman rank 3 condition, while the usual translog-AIDS specifications based on the Leser-Working specification have rank 2 coefficient matrices.

Our first results are from the 1982 Consumer Expenditure Survey (CES). The CES collects data from families over 4 quarters so that we can apply the repeated measurement techniques discussed in Section II. The basic data we use are budget share and total expenditure for each family from 1982:1. We initially use as the repeated measurement total expenditure from 1982:2. The repeated observation estimator of equations (2.5)-(2.7) is used. We estimate Engel curves on 5 commodity groups: food, clothing, recreation, health care, and transportation. We report elasticity estimates and asymptotic standard errors at 3 quartiles so that the shape of the Engel curves can be compared:

Table 4.1: Expenditure Elasticity Estimates Using 1982 CES Data
Repeated Measurement Estimator using 1982:2
(Asymptotic Standard Errors)

Commodity	IV Estimates Percentile			OLS Estimates Percentile		
	25th	50th	75th	25th	50th	75th
Food	.83	.74	.63	.73	.68	.60
	(.06)	(.04)	(.05)	(.03)	(.02)	(.03)
Clothing	1.44	1.43	1.41	1.29	1.14	.99
	(.16)	(.08)	(.13)	(.06)	(.04)	(.04)
Recreation	1.47	1.28	1.12	1.51	1.28	1.08
	(.18)	(.09)	(.14)	(.09)	(.06)	(.06)
Health	.009	.10	.44	.50	.56	.68
	(.21)	(.14)	(.21)	(.09)	(.07)	(.09)
Transpor.	1.19	1.11	1.02	1.18	1.35	1.48
	(.11)	(.05)	(.12)	(.06)	(.04)	(.04)

Number of observations= 1324

Overall, with the exception of health the IV and OLS estimates are reasonably close and both accurately estimate the elasticities. A joint test of the Leser-Working specification, that all the β_2's are zero, is computed to be 9.75 for the IV estimates. The marginal significance level for a χ^2 random variable with 5 degrees of freedom is about .08. Similarly, a test based on the OLS estimates is computed to be 73.1 which has a marginal significance level of less than .001. Thus, both the estimates, especially for food and recreation, and the statistical tests give some indication that a quadratic term gives better estimates of Engel curves on individual data. The usual assumption of constant budget share elasticities which the Leser-Working specification imposes appears inconsistent with the 1982 CES data.

While the IV and OLS estimates are reasonable close, some sizeable differences do occur. For instance the estimated food elasticity at the 25th

percentile and the clothing elasticities at the 50th and 75th percentile are quite different with both sets of elasticities estimated with a high degree of accuracy. Also, the estimated transportation elasticities differ by a range of 25% to 50% at the 50th and 75th percentile between OLS and IV. To test for a possible difference we do a Hausman (1978) type specification test of the IV estimates versus the OLS estimates. The estimated statistic is 87.39 which is distributed as a χ^2 random variable with 15 degrees of freedom. Thus, we find strong evidence that use of current expenditure in estimation of Engel curves on micro data leads to errors in variables problems. An alternative method to consider the problem is to note that the estimated var(η) is .108 while the estimated var(z) is .150. Thus, the measurement error in current expenditure is about 42% of the total variance of .258 of the logarithm of measured expenditure. The substantial proportion of measurement error in measured expenditure can lead to significant problems in OLS-type estimators.

We now explore the Gorman results. Gorman's theorem implies that higher order polynomial terms in log income will have a linear relationship to the lower order terms since the matrix of coefficients is at most of rank three. For the polynomial generalization of equation (4.1), the rank restriction takes the form that the ratio of coefficients of the cubic terms to the coefficients of the quadratic terms will be constant across equations.[13] First, we estimate a generalization of equation (4.1) with a third degree term in log income included:

$$(4.2) \quad w_i = \beta_0 + \beta_1 \log(z) + \beta_2 \log^2(z) + \beta_3 \log^3(z) + \epsilon_i.$$

---

[13] This rank restriction result follows from Gorman (1981), p. 16, equation (1).

The estimated Engel curve and elasticities are quite similar to those based on the quadratic specification of equation (4.1) The χ^2 statistic that the third order terms are all zero for the IV estimator is 2.59 with five degrees of freedom; the corresponding statistic for the OLS estimates is 10.57. Thus, the IV estimates do not demonstrate much evidence for more than a quadratic term in the budget share specification. The OLS estimates, with a marginal significance level of about .07, are more ambiguous about the cubic terms. However, we will use the quadratic specification in our subsequent estimation because we believe that the IV estimates are likely to be better than the OLS estimates.

We then estimate the "Gorman statistic" to see whether the coefficients in the cubic specific of equation (4.2) have rank three. We find a rather remarkable result (which we hope is not due to computational error). Despite considerable variation in the estimates of the β_2's and the β_3's, we find their ratios to be remarkably close in actual values and estimated precisely although we cannot estimate the individual coefficients very precisely. Thus, we find a more special result than Gorman's result--not only is the coefficient matrix of rank three, but the linear dependency takes on a remarkably simple form.

Table 4.2: Estimated Ratios of β_3/β_2 for Equation (4.2)

Commodity	IV Ratio	OLS Ratio
Food	-24.98	-129.35
	(0.47)	(10739)
Clothing	-25.24	-22.6
	(0.56)	(2.46)
Recreation	-25.12	-22.97
	(0.32)	(2.17)
Health Care	-23.31	-28.57
	(5.18)	(2.38)
Transportation	-25.57	-34.09
	(0.31)	(9.05)
$\chi^2(4)$ Statistic	1.60	7.85

Thus, both the IV results and the OLS results demonstrate that the Gorman results on rank 3 holds in the 1982 CES data. The one anomalous result for OLS is for food where the estimated quadratic term is very near zero. This estimate leads to the high estimated Gorman ratio as well as the very high estimated standard error of the ratio. Note that the OLS results are not as good as the IV results since the test statistic has a marginal significance level of about .10. However, as before, we tend to prefer the IV estimates. Perhaps the results are "too good" given the variation in prices faced by families in the sample which we have no data on.

We now reestimate equations (4.1) and (4.2) by IV using expenditure from 1982:3 in place of expenditure from 1982:2 to form the instrument. The results are very similar to the results in Table 4.1:

Table 4.3: Expenditure Elasticity Estimates Using 1982 CES Data
Repeated Measurement Estimator using 1982:3
IV Estimates

Commodity	Percentile			Gorman Statistic
	25th	50th	75th	
Food	.72	.69	.65	-25.18
	(.06)	(.04)	(.06)	(.38)
Clothing	1.50	1.44	1.38	-25.18
	(.12)	(.07)	(.13)	(2.23)
Recreation	1.41	1.47	1.50	-24.65
	(.17)	(.11)	(.20)	(1.01)
Health	.10	.23	.60	-28.34
	(.18)	(.13)	(.23)	(7.37)
Transpor.	1.23	1.09	.94	-26.60
	(.09)	(.05)	(.10)	(7.10)

$\chi^2(4)$ Statistic 0.44 Number of Obs=1324

The χ^2 statistic for the Gorman ratios again shows no evidence of rejecting the rank 3 restriction. A Hausman (1978) specification test type statistic for IV

versus OLS is estimated to be 135.71, which is distributes as χ^2 with 15 degrees of freedom. Again, a strong indication is found of the importance of measurement error in the micro data and potential problems with the use of OLS-type estimators. The estimated variance of the measurement error is .106 which is extremely close to previous estimate and which represents 41% of the total variance of the logarithm of measured expenditure. Thus, both sets of repeated measurement estimates yield numerically consistent sets of coefficient estimates.

Up to this point, we have used the replicated measurement estimator for our Engel curve specifications. Here we use the predicted IV estimator of Section II, equations (2.13) and equations (2.15)-(2.16), where the instruments used include age, education, race, union membership, spouse age and employment, and region and industry dummy variables. Thus, we use a "predicted value" for expenditure to form the instruments to use in the nonlinear specifications where the R^2 of the prediction equation is about 0.3.

Table 4.4: Expenditure Elasticity Estimates Using 1982 CES Data
Using Predicted Expenditure Estimator
IV Estimates

Commodity	Percentile			Gorman Statistic	Overid Statistic
	25th	50th	75th		
Food	.69	.62	.51	-25.39	3.12
	(.15)	(.06)	(.15)	(.15)	
Clothing	1.71	1.47	1.28	-25.22	1.35
	(.35)	(.01)	(.25)	(.21)	
Recreation	2.35	1.51	.93	-25.94	0.59
	(.45)	(.11)	(.31)	(.66)	
Health	.15	.24	.50	-25.26	4.68
	(.31)	(.15)	(.34)	(.15)	
Transpor.	1.59	1.02	.39	-26.61	1.46
	(.23)	(.08)	(.25)	(1.84)	

$\chi^2(4)$ Statistic 2.43 Number of obs=1324

Except for recreation and for transportation at the 75th percentile the estimated elasticities are quite close to the repeated measurement estimates. A test that all the quadratic terms are zero is estimated to be 11.78 which has a marginal significance level about .04. Thus, again we find evidence that higher order terms should be included in the Engel curve specification. A Hausman specification test statistic is calculated to be 73.34 which again indicates that the IV estimates are better than the OLS estimates. The $\chi^2(2)$ test for correct specification from equation (2.21) is well below its critical value of 6.0 at the 5 percent level for each commodity. The test for overidentification does not reject our specification.

We now do a χ^2 test that the two sets of repeated measurement IV estimates are the same. This test is equivalent to a test of the overidentifying restrictions on the instruments. The χ^2 statistic is estimated to be 12.4, and since it has 15 degrees of freedom, no evidence is found to reject the hypothesis of orthogonality of the instruments. However, the equivalent tests for the predicted expenditure form of the IV estimator in relation to the repeated measurement IV estimators yield χ^2 statistics of 35.6 and 91.6, respectively, both of which indicate that either the repeated measurement or the predicted value instruments are not mutually orthogonal to the stochastic disturbance in the Engel curve specifications. Since the repeated measurement estimators are mutually consistent with each other, we tend to believe that they are the superior estimators in the current situation. We cannot be more specific about the relative superiority of the estimators without further research.

We repeat the IV estimation of equations (4.1) and (4.2) using 1972 CES data where we predict expenditure using similar instruments.[14] Repeated

---

[14] These data were kindly provided to us by Professor Dale Jorgenson.

observations on family expenditure are not available for 1972. The 1972 CES data set is sufficiently large that we estimated the Engel curve only on 4 person families to minimize potential family size effects on the estimates. The estimated elasticities are reported in Table 4.5:

Table 4.5: Expenditure Elasticity Estimates Using 1972 CES Data
Predicted Expenditure Estimator
IV Estimates

Commodity	Percentile			Gorman Statistic		Overid Statistic
	25th	50th	75th	IV	OLS	
Food	.76	.67	.54	-42.1	-44.5	1.23
	(.12)	(.07)	(.13)	(.48)	(4.73)	
Clothing	1.43	1.36	1.22	-41.8	-42.8	3.82
	(.89)	(.83)	(.80)	(1.14)	(.66)	
Recreation	1.32	1.41	1.49	-41.3	-43.5	2.39
	(.13)	(.10)	(.17)	(2.23)	(1.65)	
Health	1.07	.78	.44	-41.1	-42.0	0.41
	(.20)	(.11)	(.21)	(1.67)	(.50)	
Transpor.	.54	.59	.65	-42.2	-40.6	2.31
	(.18)	(.10)	(.19)	(.22)	(2.60)	

$\chi^2(4)$ Statistic 0.54 Number of obs = 992

The estimated elasticity for transportation is below the 1982 estimates which may well arise from the extremely large rise in gasoline prices between 1972 and 1982. The estimated IV Gorman ratios are again very close, and a χ^2 test fails to come close to a rejection of equality.[15] Note that the estimated values of the Gorman ratios differ from their estimated values in 1982. This result is to be expected since the slope coefficients are, in general, nonlinear functions of prices. The CPI increased by over 130% between 1972 and 1982 with significant

[15] Note that the estimated OLS Gorman ratios are also very close. The $\chi^2(4)$ statistic for the OLS estimates is 1.87 which indicates no grounds for rejection. For completeness, the $\chi^2(5)$ statistic that all the quadratic terms are zero is 17.2 which is strong evidence against the Leser-Working Engel curve specification on micro data.

differences in increases across expenditure categories. Thus, the ratios of nonlinear functions of prices would be expected to change as prices change. A Hausman (1978) type specification test of IV versus OLS is estimated to be 117.8. Thus, we again find strong evidence of the importance of measurement error in the 1972 CES data as we did in the 1982 data. Lastly, the $\chi^2(2)$ of overidentification of equation (2.21) once more does not reject our specification of the Engel curve for any commodity.

One potential problem that we have not yet accounted for is errors in variables in the left hand side variable, the budget shares, in equations (4.1) and (4.2). To the extent that the errors in variables occurs in expenditure on a given commodity, which forms the numerator of the budget share, no special problem arises. However, to the extent that the denominator of the budget share, total expenditure, is measured with error, estimation problems arise. Since the measurement error enters the problem in a non-polynomial variable, no obvious solution exists. But the problem can be eliminated by respecifying equations (4.1) and (4.2) with commodity expenditure as the left hand side variable instead of the budget share. The estimation procedure remains the same except for an adjustment to the estimated standard errors of the coefficients to account for heteroscedasticity.

The results are presented in Table 4.6. We do not find that errors in variables in the left hand side variable presents a significant problem although this Engel curve specification does not fit the data as well as the earlier specifications.

Table 4.6: Expenditure Elasticity Estimates Using 1982 CES Data
Commodity Expenditure is Left Hand Side Variable
IV Estimates

Commodity	Percentile			Gorman Statistic
	25th	50th	75th	
Food	.89	.73	.62	-33.38
	(.08)	(.04)	(.07)	(33.18)
Clothing	1.72	1.61	1.36	-15.26
	(.23)	(.11)	(.16)	(1.49)
Recreation	1.55	1.26	1.05	-14.59
	(.17)	(.11)	(.20)	(27.04)
Health	-.41	.15	.68	-16.61
	(.39)	(.24)	(.32)	(1.13)
Transpor.	1.06	1.22	1.16	-17.69
	(.33)	(.13)	(.30)	(0.82)

$\chi^2(4)$ Statistic 2.29 Number of Obs=1324

The elasticity estimates are quite close to the elasticity estimates derived from the budget share results of Tables 4.1 and 4.3. The only exception is the estimated health elasticity at the 25th percentile which is estimated very imprecisely. The Gorman ratios are all quite close to one another with the exception of food, which again is measured quite imprecisely. The χ^2 statistic does not come close to a rejection of the Gorman restriction. Thus, when we estimate the Engel curves in commodity expenditure form, rather than budget share form, the results remain essentially unchanged. We again find support for the Gorman restriction on the specification of the Engel curve.

Up to this point we have considered only polynomial Engel curves for budget share data. However, Leser (1963) found evidence which indicated that the following Engel curve specification was superior to the Leser-Working specification:

$$(4.3) \quad w_i = \beta_0 + \beta_1 \log(z) + \beta_2/z + \epsilon_i.$$

Thus, he generalized the Working specification to include the inverse of income as well as its logarithm. We consider this extended Leser specification as well as another generalization of the Leser-Working specification:

$$(4.4) \quad w_i = \beta_0 + \beta_1 \log(z) + \beta_2 \, z\log(z) + \epsilon_i .$$

Note that both equations (4.3) and (4.4) are rank two specifications. The coefficients of these generalized Engel curve specifications are estimated using the general nonlinear errors in variables estimator of Section III, equation (3.6). Recall that the estimation strategy of Section III involves fitting the Engel curves with polynomials followed by estimation of the coefficients of the nonlinear specifications using the predicted values of the budget shares from the polynomial coefficient estimates. The estimated coefficients of equations (4.3) and (4.4) follow from the best fitting polynomial. In 9 out of 10 cases the best fitting polynomial is a second degree polynomial with the sole exception being health care for the specification of equation (4.4) which uses a third degree polynomial.

The estimates of the nonlinear Engel specifications are given in Table 4.7

Table 4.7: Estimates Using 1982 CES Data--General Nonlinear Specifications[16]

Commodity	Equation (4.3) Percentile			Equation (4.4) Percentile		
	25th	50th	75th	25th	50th	75th
Food	.82	.71	.59	.82	.76	.67
	(.060)	(.031)	(.068)	(.057)	(.043)	(.045)
Clothing	1.45	1.45	1.42	1.45	1.41	1.39
	(.13)	(.11)	(.15)	(.13)	(.076)	(.084)
Recreation	1.43	1.23	1.09	1.45	1.33	1.20
	(.15)	(.11)	(.17)	(.17)	(.10)	(.10)
Health	.06	.28	.60	.12	.04	.15
	(.19)	(.13)	(.27)	(.15)	(.17)	(.22)
Transpor.	1.16	1.08	1.01	1.17	1.12	1.07
	(.08)	(.07)	(.14)	(.09)	(.06)	(.08)

Note that the estimates of the elasticities are quite similar between the two nonlinear specifications. Furthermore, the estimated elasticities are close to the estimated elasticities for the polynomial specifications in Table 4.1. We compare the closeness of fit of the nonlinear specifications to the predicted values of the underlying polynomials since that is the criterion used to estimate the coefficients in equation (3.6). For 4 of the 5 commodities, the extended Leser specification of equation (4.3) fits better than the generalized specification of equation (4.4). The exception is health care where none of the Engel curve specifications do very well. However, to the extent that the estimated elasticities are so similar, the choice of the "best" specification is probably a rather unimportant exercise.

We now consider an additional nonlinear specification which accounts for possible measurement errors in the left hand side variables, the budget shares. We take the quadratic generalization of the Leser-Working Engel curve of equation (4.1) and multiply both sides of the equation by total expenditure:

---

16 Standard errors are calculated by the bootstrap method here.

$$(4.5) \quad e_i = \beta_0 z + \beta_1 z\log(z) + \beta_2 z\log^2(z) + \epsilon_i.$$

In equation (4.5) commodity expenditure is now the left hand side variable which eliminates possible problems from measurement error in the denominator of the budget shares in equations (4.1) and (4.2). Our estimates of equation (4.5) are very similar to the estimates in Table 4.7 and earlier tables. For instance, the estimated elasticities at the 50th percentile are (0.69, 1.46, 1.17, 0.37, 1.13). Thus, the nonlinear specification of the quadratic version of the Leser-Working Engel curve yields estimates very close to our previous estimates so that measurement error in the left hand side variable again does not seem to be an important problem.

Our final exploration of the Engel curve specification involves the addition of demographic variables in equation (4.1). Differences in household size have often been a focus of attention in the specification of Engel curves; here we also include region of the U.S. to account for regional price differences of the commodities as well as age of the household head to account for life cycle effects. The demographic variables are all entered as indicator (dummy) variables with 4 family size groups, 4 region groups, and 4 age groups. We believe that this specification is preferable to the non-identified approach of family equivalence scale specifications. The approach of equation (2.23) is used to include the additional right hand side variables in the Engel curve specifications.

We now reestimate Tables 4.1 and 4.2 where we include the demographic variables and used the repeated measurement estimator.

Table 4.8: Expenditure Elasticity Estimates Using 1982 CES Data
Repeated Measurement Estimator using 1982:2
IV Estimates

Commodity	Percentile			Gorman Statistic
	25th	50th	75th	
Food	.72	.66	.59	-9.93
	(.05)	(.06)	(.07)	(6.17)
Clothing	1.47	1.43	1.39	-10.44
	(.14)	(.13)	(.12)	(8.23)
Recreation	1.17	1.16	1.15	-14.80
	(.17)	(.17)	(.17)	(.43)
Health	.21	-.09	-.33	-14.01
	(.17)	(.23)	(.39)	(.52)
Transpor.	1.07	1.07	1.06	-15.34
	(.10)	(.10)	(.11)	(.93)

$\chi^2(4)$ Statistic 1.90 Number of Obs=1321

3 Family Size, 3 Age, and 3 Region variables are included

The estimated quartile elasticities change very little from the specification which omits demographics, with the sole exception of the health equation. The health equation elasticities are estimated very imprecisely with the negative coefficient estimates accompanied by quite large asymptotic standard error estimates. The Gorman ratios are not as close as in Table 4.2 although the test statistic takes on almost the same value which indicates no reason to reject the Gorman restrictions. The food and clothing ratios are smaller in magnitude than the other three commodities. However, the Gorman ratio for food and clothing are estimated very imprecisely. We again find strong evidence against the Leser-Working specification of equation (4.1) and evidence in favor of the cubic specification of equation (4.2). The $\chi^2(5)$ statistic is estimated to be 48.2. The Hausman (1978) test of IV versus OLS with demographics is 473.0 which indicates strong evidence of measurement error since it is distributed as a $\chi^2(10)$ random variable under the null hypothesis.

Despite the closeness of the quartile elasticity estimates without and without demographic variables included, we do find a quite significant influence of demographic variables on expenditures shares. We present the estimates results in Table 4.9:

Table 4.9: Coefficient Estimates for Demographic Variables

Commodity	Food	Clothing	Recreation	HC	Transportation
Age1 (19-29)	-.04	-.01	-.00	-.02	.01
	(.01)	(.01)	(.01)	(.01)	(.02)
Age2 (30-39)	.04	.01	.01	.07	.03
	(.02)	(.01)	(.01)	(.02)	(.03)
Age3 (40-49)	-.00	.00	.00	.01	.01
	(.01)	(.00)	(.01)	(.01)	(.01)
Reg1 (NE)	.01	.00	.01	-.00	-.00
	(.01)	(.00)	(.00)	(.01)	(.01)
Reg2 (W)	.01	.01	-.01	-.08	-.01
	(.03)	(.01)	(.02)	(.03)	(.04)
Reg3 (MW)	-.01	.01	.02	.01	.01
	(.03)	(.01)	(.02)	(.03)	(.04)
Size1 (2)	-.00	-.00	-.01	.01	.01
	(.01)	(.00)	(.00)	(.01)	(.01)
Size2 (3)	.01	-.01	-.01	.02	.01
	(.01)	(.00)	(.01)	(.01)	(.01)
Size3 (4)	.01	-.00	.00	.00	.00
	(.00)	(.00)	(.00)	(.00)	(.01)
Mean Share	.22	.06	.05	.06	.21

We find fairly sizeable age effects in the estimated share equations. The region effects are not particularly large which helps support the necessary assumption of constant prices across the US in the Engel curve specification. The notable

exception is the Western region for health care. The health care equation is difficult to estimate overall; the estimated effect here may arise from the much larger share of health maintenance organizations in the West in 1982. The family size effects are statistically significant although they have only a small effect compared to the shares in expenditure of the five commodities.

We then reestimated the Engel curve specifications for 1982 using the second repeated measurement. The estimated elasticities, not reported here, are quite similar to Table 4.7 and the demographic effects are quite similar to Table 4.8. The 5 Gorman ratios are estimated to be (-8.26, -8.10, -9.51, -7.68,-11.86). Thus, the ratios are once again quite close to each other with a $\chi^2(4)$ test statistic of 3.97. The test for the quadratic against the cubic specification is 86.4 which once again gives strong evidence against the Leser-Working specification. The Hausman test statistic of IV versus OLS is estimated to be 675.8. However, when we include the demographic variables the two sets of replicated measurement results are no longer nearly so mutually consistent as before. The $\chi^2(10)$ statistic for the test of overidentification is estimated to be 50.4 which easily rejects the overidentifying restrictions.

In this section we have estimated various Engel curve specifications where we have taken account of errors in measurement in income or expenditure. We find very strong evidence that substantial measurement error exists in the CES data. We also find strong evidence that equation (4.2) is preferable to the Leser-Working specification of equation (4.1). However, we do not find support for the hypothesis that more general nonlinear specifications or higher order polynomial terms than cubic are needed. We find strong support for the Gorman rank condition which limits polynomial Engel curve specifications to rank 3. Our results also show that demographic variables have significant effects in Engel

curve specifications of the share. Lastly, we have demonstrated the feasibility and importance of using consistent instrumental variable estimators in nonlinear econometric model specifications.

References

Adcock, R. (1887), "A Problem in Least Squares", The Analyst, 5, 53-54

Aigner, D.J., Hsiao, C., Kapteyn A., Wansbeek, T. (1984), "Latent Variable Models in Econometrics", in Z. Griliches and M. D. Intriligator, Handbook of Econometrics, vol. II, 1323-1393

Amemiya, T. [1974], "The Nonlinear Two Stage Least Squares Estimator," Journal of Econometrics, 2: 105-110

Amemiya, T. (1977), "The Maximum Likelihood Estimator and the Nonlinear Three-Stage Least Squares Estimator in the General Nonlinear Simultaneous Equation Model," Econometrica, 45, 955-968

Amemiya, Y. [1985], "Instrumental Variable Estimator for the Nonlinear Errors-in-Variables Model," Journal of Econometrics, 28: 273-289.

Bickel, P.J. and Y. Ritov (1987), "Efficient Estimation in the Errors in Variables Model", Annals of Statistics, 15, 513-540

Deaton, A. and J.S. Muellbauer (1980), An Almost Ideal Demand System, American Economic Review, 70, 312-326

Dolby, G.R. and S. Lipton [1972], "Maximum Likelihood Estimation of the Generalized Nonlinear Functional Relationship with Replicated Observations and Correlated Errors," Biometrika, 59: 121-129.

Freidman, M. (1957), A Theory of the Consumption Function, (Princeton U.P.)

Fuller, W. (1987), Measurement Error Models, (Wiley: New York)

Gallant, A.R., 1980, "Explicit Estimators of Parametric Functions in Nonlinear Regression, Journal of the American Statistical Association, 75: 182-193.

Geary, R.C. (1942), "Inherent Relations Between Random Variables", Proceedings of the Royal Irish Academy, A, 47, 63-67

Geraci, V. (1977), "Estimation of Simultaneous Equation Models with Measurement Error", Econometrica, 45, 1243-1256

Goldberger, A.S. (1972), "Maximum-Likelood Estimation of Regressions Containing Unobservable Independent Variables", International Economic Review, 13, 1-15

Gorman, W.M. (1981), "Some Engel Curves,", in A. Deaton ed., Essays in the Theory and Measurement of Consumer Behaviour in Honor of Sir Richard Stone, (Cambridge U.P.)

Griliches, Z. and V. Ringstad [1970], "Errors-in-Variables Bias in Nonlinear Contexts," Econometrica, 38: 368-370.

Griliches, Z. (1986), "Economic Data Issues", in Z. Griliches and M. D. Intriligator, Handbook of Econometrics, vol. III, 1465-1514.

Hausman, J.A. (1978), "Specification Tests in Econometrics", Econometrica, 46, 1251-1271

Hausman, J.A. (1978), "Errors in Variables in Simultaneous Equation Models", Journal of Econometrics, 5, 389-401

Hausman, J., H. Ichimura, W. Newey, and J. Powell (1986), "Measurement Errors in Polynomial Regression Models," MIT mimeo

Hausman, J., W. Newey, and J. Powell (1988), "Consistent Estimation of Nonlinear Errors-In-Variables Models with Few Measurements", MIT mimeo

Hsiao, C. (1976), "Identification and Estimation of Simultaneous Equation Models with Measurement Error", International Economic Review, 17, 319-339

Jorgenson, D.W., Lau, L.L, and Stoker, T. M. (1982), "The Transcendental Logarithmic Model of Aggregate Consumer Behavior", Advances in Econometrics, 1, 97-238.

Kapteyn, A. and T.J. Wansbeek (1983), "Identification in the Linear Errors in Variables Model", Econometrica, 51, 1847-1849

Leser, C.E.V. (1963), "Forms of Engel Functions", Econometrica, 31, 694-703

Lewbel A. (1986), "Characterizing Some Gorman Engel Curves", Brandeis Univ. mimeo

Lewbel A. (1987), Characterization and Rank Theorems for Deflated Income Demand Systems, Brandeis Univ. mimeo

Livitan N. (1961), "Errors in Variables and Engel Curve Analysis", Econometrica, 29, 336-362

Neyman, J. and E.L. Scott (1948), "Consistent Estimates Based on Partially Consistent Observations", Econometrica, 16, 1-32

Pollak, R.A. and T.J. Wales, "Comparison of the Quadratic Expenditure System and Translog Demand Systems with Alternative Specifications of Demographic Effects", Econometrica, 595-612

Powell, J.L. and T.M. Stoker [1986], "The Estimation of Complete Aggregation Structures," Journal of Econometrics, 30: 317-341.

Prais, S.J. and H.S. Houthakker (1955), The Analysis of Family Budgets, Cambridge U.P., (2d Ed. 1971)

Reiersol, O. (1950), "Identifiability of a Linear Relation Between Variables which are Subject to Error, Econometrica, 18, 375-389

Summers, R. (1959), "A Note on Least Squares Bias in Household Expenditure Analysis," Econometrica, 27, 121-134

Villegas, C. [1969], "On the Least Squares Estimation of Non-Linear Relations," Annals of Mathematical Statistics, 11, 284-300.

Wolter, K.M. and W.A. Fuller [1982b], "Estimation of Nonlinear Errors-in-Variables Models," Annals of Statistics, 10, 539-548.

Wolter, K.M. and W.A. Fuller [1982a], "Estimation of the Quadratic Error-in-Variables Model," Biometrika, 69: 175-182.

Zellner, A. (1970), "Estimation of Regression Relationships Containing Unobservable Independent Variables", International Economic Review, 11, 441-454

3-22-89