working paper
department
of economics

THE ESTIMATION OF COMPLETE AGGREGATION STRUCTURES

James L. Powell and Thomas M. Stoker*

M.I.T. Working Paper #346

April 1984

massachusetts
institute of
technology

50 memorial drive
cambridge, mass. 02139

THE ESTIMATION OF COMPLETE AGGREGATION STRUCTURES

James L. Powell and Thomas M. Stoker*

M.I.T. Working Paper #346

April 1984

* Department of Economics and Sloan School of Management, M.I.T. The authors gratefully acknowledge support from the National Science Foundation for this research.

THE ESTIMATION OF COMPLETE AGGREGATION STRUCTURES

James L. Powell and Thomas M. Stoker

1. Introduction

The purpose of this paper is to establish several results which permit efficient estimation of behavioral parameters in models of aggregate data. The generic econometric situation of interest occurs when one observes data aggregates for a given number of groups, which for example could be separate time periods. The model for these data accounts for the problem of aggregation over individuals - it is constructed from a parameterized microeconometric model integrated over an assumed distribution form, in a way which permits identification of the micro behavioral parameters. The question addressed in this paper is how one can consistently and efficiently estimate the values of these behavioral parameters, using only the aggregate data.

The issues involved in constructing models which account for the problem of aggregation over individuals are extensively covered in Stoker (1984). The empirical implications of individual heterogeneity and distribution shifts on aggregate data are formally modeled by an aggregate function, which is constructed from a microeconometric model integrated over an assumed distribution form. Completeness of the aggregation structure (or the absence of an aggregation problem) occurs when the parameters of micro behavior are identified by the aggregate function, which requires specific restrictions on the form of micro behavior and/or the form of the predictor variable distribution. Stoker (1984) presents several examples of aggregate functions and characterizing results for complete aggregation structures, which include

models with intrinsically nonlinear microeconometric equations and specific restrictions on the predictor variable distribution.

In this paper we show several results which facilitate the use of complete aggregation structure modeling to study observed aggregate data from large populations. In particular, we begin by showing that natural consistent parameter estimators can be obtained via weighted least squares (WLS), and that the proper choice of weights provides efficient estimators, so that the best WLS estimator is first order efficient in the much broader class of minimum distance estimators, which in general contains maximum likelihood. The results are in part based on an appropriate modification of the principal finding of the theory of minimum chi-square estimation; namely the first order equivalence of minimum chi-square and other minimum distance estimators. As part of the exposition, we provide examples of estimation problems with aggregate data, and connect the optimal selection of weights to cross section regression results of Stoker (1983).

The major advantages of our results lie in their generality and computational simplicity. The results are applicable to a broad range of different empirical situations, being valid for: (i) many different types of aggregate data on both dependent and predictor variables - averages, medians, more general order statistics, etc., (ii) virtually arbitrary forms of micro behavior -- linear or nonlinear, continuous or discrete variables -- and (iii) general specifications of the predictor variable distribution. The computational simplicity of the results lies in the first-order optimality of weighted least squares, which is a now standard (nonlinear) estimation technique.

We begin with the notation and basic framework in Section 2. Section 3 gives conditions under which the weighted least squares estimator of the

behavioral parameters is consistent and asymptotically normal, discusses estimation of the asymptotic covariance matrix of the estimator, and shows how the optimal weighting scheme is related to a particular cross-section regression. Section 4 shows the first-order equivalence between general minimum distance estimators and weighted least squares estimators with appropriately chosen weights. Section 5 suggests some natural extensions of the framework, and gives some concluding remarks.

2. Definitions and Notation

We assume throughout that the object of statistical analysis is the estimation (with associated hypothesis tests) of an unknown parameter vector δ characterizing a single behavioral equation,

$$(2.1) \qquad y_{it} = f(x_{it}, u_{it}, \rho_t; \delta),$$

where y_{it} is a variable representing a behavioral response of an individual agent i in a group of agents indexed by t, x_{it} is a k-vector of (potentially) observable characteristics of that individual, and u_{it} is a disturbance embodying unobserved differences in agents not captured by x_{it}; the components of ρ_t, which may be observable or unobservable, represent variables which are common to all agents in a particular group but which vary across groups. The probability law generating a particular realization of y_{it} is thus determined by the true value δ_o of the unknown parameter δ and by the distribution of characteristics (x_{it}, ρ_t, u_{it}) in the population being studied. In the aggregation problems considered here, we assume that x_{it} and u_{it} are distributed independently of ρ_t, and that the distribution of (x_{it}, u_{it}) within each group t has a density $h_t(x, u)$ which is absolutely continuous with respect to some sigma-finite measure λ, invariant across groups; further, we suppose that this density can be parameterized as

$$(2.2) \qquad h_t(x, u) = g_t(u \mid x, \sigma_o)p_t(x \mid \theta_t),$$

where the forms of $g_t(\cdot)$ and $p_t(\cdot)$ are known up to the unknown parameters σ_o and θ_t.

This setup, while not completely general in its treatment of individual and group heterogeneity, characterizes a broad class of economic models of interest. The groups denoted by t may represent cross-section aggregation units, i.e., observations for particular states, countries, or industries in a particular period of time, or the groups may correspond to populations

varying across time. When the group specific characteristics $\{\rho_t\}$ are completely observable, the parameters θ_t determining the distribution of the observable individual characteristics x_{it} completely characterize the unknown aspects of distributional differences across groups; throughout the analysis below, we will assume that the $\{\rho_t\}$ are observable, and thus rewrite the behavioral model as

$$(2.1') \qquad y_{it} = f_t(x_{it}, u_{it}; \delta_o),$$

with the sequence of functions f_t indexed by the observable ρ_t.

If a random sample of N_t observations on (y_{it}, x_{it}) were available for each group $t=1,\ldots,T$, maximum likelihood estimation of the unknown parameters δ_o, σ_o and θ_t would be feasible; because $\{x_{it}, i=1,\ldots,N_t\}$ is sufficient for each θ_t by the structure imposed in (2.2), the principle of ancillarity would lead to estimation of the parameters $\gamma_o' = [\delta_o', \sigma_o']$ through maximization of the conditional likelihood function

$$(2.3) \qquad L_\gamma \equiv \sum_{t=1}^{T} \sum_{i=1}^{N_t} \log q_t(y_{it}|x_{it}, \gamma),$$

where $q_t(\cdot)$ is the conditional density of y_{it} given x_{it} and the observable ρ_t. However, we suppose that these "micro samples" for each group are not available for statistical analysis, perhaps due to prohibitive costs of complete data reporting or confidentiality considerations for the individuals sampled; instead, only aggregates $\hat{Y}_t$ and $\hat{X}_t$ representing the "typical" values of y_{it} and x_{it} are available for each group. We further assume that the aggregates constructed from a random sample of (y_{it}, x_{it}') within each group (for $i=1,\ldots N_t$ and $t=1,\ldots,T$) and are "democratic" to the extent that they depend only on the respective empirical distribution functions, i.e. $\hat{Y}_t$ and $\hat{X}_t$ are functionals of the form

(2.3) $\hat{Y}_t = Y(\hat{Q}_t),\ \hat{X}_t = X(\hat{P}_t),$

where

(2.4) $$\hat{Q}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} 1(y_{it} \leq y) \equiv \hat{Q}_t(y),$$

$$\hat{P}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} 1(x_{it} \leq x) \equiv \hat{P}_t(x),$$

with "1(A)" denoting the indicator function of the event "A" and the inequality "$x_{it} \leq x$" being interpreted as holding coordinate-by-coordinate. To these empirical aggregates $\hat{Y}_t$ and $\hat{X}_t$ correspond population aggregates $Y(Q_t)$ and $X(P_t)$, where Q_t and P_t are the marginal distributions of y_{it} and x_{it} in the $t^{\underline{th}}$ group; writing these explicitly as functions of the unknown parameters, we have

(2.5) $$Y(Q_t) = Y(Q_t(y|\theta_t, \gamma)) \equiv \Lambda_t(\theta_t, \gamma) \text{ and}$$

$$X(P_t) = X(P_t(x|\theta_t)) \equiv \Psi_t(\theta_t).$$

A key assumption in our approach is that $\Psi_t(\theta_t)$ is a known and invertible function of θ_t. This implies that we may reparameterize the distribution of x_{it} in group t by $\mu_t \equiv \Psi_t(\theta_t)$, i.e., $p_t(x \mid \Psi_t^{-1}(\mu_t)) = \tilde{p}_t(x \mid \mu_t)$; further, with an assumption of Fisher consistency of the function $X(\cdot)$--that is, $X(\hat{P}_t) \to X(P_t)$ as $\hat{P}_t \to P_t$, where the latter convergence is appropriately defined--the empirical aggregate $\hat{X}_t$ will consistently estimate the only unknown source of variability of the joint distribution of y_{it} and x_{it} across groups.

An important special case of this general framework occurs when $Y(\cdot)$ and $X(\cdot)$ are expectations operators,

(2.6) $\hat{Y}_t = \int y d\hat{Q}_t(y) \equiv \overline{Y}_t$ and

$$\hat{X}_t = \int x d\hat{P}_t(x) \equiv \overline{X}_t.$$

Identification of the unknown parameters γ in this case has been extensively investigated in Stoker (1984), who relates identification of γ to completeness of the family $\{p_t(x \mid \theta)\}$ for θ in some parameter space Θ. As Stoker's analysis demonstrates, the assumption that $\Psi_t(\theta)$ is invertible, while a strong restriction on the availability of aggregate data on characteristics, is essential if the behavioral parameters γ are to be estimable.

The general approach to aggregation adopted here also applies to order statistic data. If $\hat{Y}_t$ is the pth quantile (p x 100th percentile) of $\{y_{it}, i=1, \ldots, N_t\}$, then

(2.7) $$Y(\hat{Q}_t) = \hat{Q}_t^{-1}(p),$$

where, as usual, some rule for choice of a particular value of $\hat{Y}_t$ when this inverse is set-valued must be specified. The population aggregate associated with Y_t is then

(2.8) $$Y(Q_t) = Q_t^{-1}(p \mid \gamma, \theta_t)$$

which is uniquely defined when Q_t is continuously differentiable with positive density in a neighborhood of $m(p) = \inf \{y: Q_t(y) \geq p\}$.

With these preliminaries, we can now turn to the central question of this paper, namely, estimation of the parameter vector γ when only T aggregates $\{(\hat{Y}_t, \hat{X}_t'), t=1, \ldots, T\}$ are available for analysis. One suggestive approach would be estimation by the method of maximum likelihood; that is, given the functional forms of the behavioral function in (2.1), the density of the characteristics in (2.2), and the form of the aggregation functionals $Y(\cdot)$ and $X(\cdot)$ in (2.3), the joint density function $d_t(y, x \mid \gamma, \mu_t, N_t)$ of the induced probability law for $\hat{Y}_t$ and $\hat{X}_t$ can be

calculated, and γ and $\mu_o = (\mu_1', \ldots, \mu_T')'$ can be estimated by

$$(2.9) \qquad \begin{bmatrix} \hat{\gamma} \\ \hat{\mu} \end{bmatrix} = \underset{\gamma, \mu}{\operatorname{argmax}} \sum_{t=1}^{T} \log d_t(\hat{Y}_t, \hat{X}_t \mid \gamma, \mu_t, N_t).$$

Upon further consideration, though, two drawbacks in this strategy are apparent. First calculation of the joint density function d_t of $\hat{Y}_t$ and $\hat{X}_t$ is far from trivial in general; for example if $\hat{Y}_t \equiv \overline{Y}_t$ and $\hat{X}_t \equiv \overline{X}_t$, computation of $d_t(\cdot)$ involves an N_t-fold convolution of the joint density of (y_{it}, x_{it}'), which may be intractable if f_t and h_t of (2.1) and (2.2) are not of simple form. Another, more fundamental issue is the applicability of the usual justification for maximum likelihood estimation in the present circumstance. The usual large-sample properties of maximum likelihood -- consistency, asymptotic normality, and asymptotic efficiency of the estimator -- could be demonstrated (under suitable regularity conditions) for T tending to infinity as each group size N_t is held fixed for each group t; however, for typical problems involving aggregates, the group sizes N_t are large relative to T. A more appropriate asymptotic theory then would take N_t tending to infinity for fixed T; the likelihood function, rather than being a sum of an increasing number of log density functions, would be a fixed sum of log density functions, each of which varies as N_t increases. Since the standard asymptotic results on maximum likelihood estimation do not apply in this case (which we assume here), and since calculation of the density functions is problematic, this does not appear to be the most promising approach to estimation with aggregate data.

A more attractive approach is based upon estimation of the population aggregates in (2.5). Reparameterizing by $\mu_t = \Psi_t(\theta_t)$, we can rewrite (2.5) as

$$(2.5') \qquad Y(Q_t) = \Lambda_t(\Psi^{-1}(\mu_t), \gamma) \equiv \Phi_t(\mu_t, \gamma) \text{ and}$$

$$X(P_t) \equiv \mu_t,$$

so that

$$(2.10) \qquad Y(Q_t) = \Phi_t[X(P_t), \gamma].$$

This latter relation, plus the assumed consistency of $\hat{Y}_t$ and $\hat{X}_t$ for $Y(Q_t)$ and $X(P_t)$ suggests estimation of γ by minimizing the average "distance" between $\hat{Y}_t$ and $\Phi_t(\hat{X}_t, \gamma)$ over γ, where the average is taken over the number of groups T. A convenient measure of distance is a weighted squared difference; thus, we will consider estimation of γ by

$$(2.11) \qquad \hat{\gamma}_w \equiv \underset{\gamma}{\operatorname{argmin}} \sum_{t=1}^{T} \hat{w}_t \, (\hat{Y}_t - \Phi_t \, (\hat{X}_t, \gamma))^2,$$

where $\{\hat{w}_t\}$ is some set of (possibly stochastic) nonnegative weights. The large sample properties of this estimator will be derived in the following section.

This approach is quite closely related to minimum chi square and related minimum distance estimation methods for multinomial models (see, for example, Rao (1973), or Bishop, Feinberg, and Holland (1975)). Nonetheless, the analysis here differs in some aspects, most notably in the generality of the aggregates $\hat{Y}_t$ and $\hat{X}_t$ considered. Furthermore, unlike the usual analysis of cell count data, it is neither necessary to suppose that the characteristics within each group (which are analagous to, say, dosage levels in bio-assay) are fixed for all observations in that group nor known with certainty; it suffices here that consistent estimates of the parameters governing the differences in the distributions of characteristics across groups are available.

Still, one of the principal findings of the theory of minimum chi-square estimation -- namely, the first-order equivalence of minimum chi-square,

maximum likelihood, and other minimum distance estimators using cell proportions -- carries over, with some modification, to the aggregate estimation framework considered here. In section 4 below, conditions are given under which the weighted least squares estimator $\hat{\gamma}_w$ defined in (2.11) above will be first order efficient (with appropriately chosen weights) among minimum distance estimates, including the maximum likelihood estimator of (2.9). In that section we also consider the relation of the type of available aggregation functionals $\hat{Y}_t$ and $\hat{X}_t$ to the precision of the weighted least squares estimator of γ.

Before turning to the large sample properties of weighted least squares estimation, we present examples of aggregate estimation problems for two particular economic models.

Example 1: (Continuous Demand Model)

Suppose that y represents the expenditure on a given commodity, which is determined as

$$\begin{aligned} y_{it} &= f(x_{it}, u_{it}, \rho_t ; \delta) \\ &= A(\rho_t, \delta)x_{it} + B(\rho_t, \delta)x_{it}\log x_{it} + u_{it}, \end{aligned}$$

where ρ_t is a vector of relative prices, presumed constant and observable for each group t, x_{it} is individual income (expenditures on all commodities), and u_{it} is a disturbance term, assume to have zero mean conditional on x_{it} and variance σ_o^2 (assumed fixed across individuals and groups for simplicity). This demand function is of the PIGLOG form of Muellbauer (1975), and includes the AIDS system of Deaton and Muellbauer (1980) and the translog model of Jorgenson, Lau, and Stoker (1982) without attributes; δ represents individual

preference parameters. We take $\delta = \gamma_o$.

Supposing $\hat{Y}_t = \overline{Y}_t = \frac{1}{N_t} \sum_{t=1}^{T} y_{it}$ is the aggregate available for consumption, and noting that

$$E(y_{it} \mid x_{it}, \rho_t, \gamma) = A_\gamma(\rho_t)x_{it} + B_\gamma(\rho_t)\, x_{it} \log x_{it},$$

we obtain the population aggregate $Y_t(Q_t)$ by taking the expectation of this formula with respect to the distribution of x_{it}. Again for simplicity we suppose the income distribution is log-normal, i.e., the density of x_{it} (with respect to Lebesque measure) is

$$P(x \mid \theta_t, \tau_t) = \frac{1(x>0)}{2\pi\tau_t x} \exp\left(\frac{-(\ln x-\theta_t)^2}{2\tau_t^2}\right).$$

Since this distribution has two time-varying parameters, we require two aggregates for the characteristics to identify γ. If

$$\hat{X}_t = \left(\frac{1}{N_t} \sum_{i=1}^{N_t} x_{it},\ \text{median}\{x_{it},\ i=1, \ldots, N_t\}\right)$$

$$\equiv (\overline{X}_t, \tilde{X}_t)$$

is available, the corresponding population aggregate characteristics for the tth group are

$$X_t(P_t) = (E(x_{it}),\ P_t^{-1}(1/2))$$

$$= (\exp(\theta_t + \tfrac{1}{2}\tau_t^2),\ \exp(\theta_t))$$

$$= \mu_t,$$

which is clearly an invertible function of (θ_t, τ_t).

For this distribution of characteristics,

$$Y_t(Q_t) = E(y_{it})$$

$$= \exp(\theta_t + \tfrac{1}{2}\tau_t^2)(A_\gamma(\rho_t) + B_\gamma(\rho_t)(\theta_t + \tau_t^2))$$

$$= \Lambda_t(\theta_t, \tau_t, \gamma)$$

or, in terms of μ_t,

$$Y_t(Q_t) = \mu_{1t}\,[A_\gamma(\rho_t) + B_\gamma(\rho_t)(2 \log \mu_{1t} - \log \mu_{2t})].$$
$$= \Phi_t\,(\mu_t, \gamma)$$

With these calculations, computation of the weighted least squares estimator $\hat{\gamma}_w$ of (2.11) for given weights $\{\hat{w}_t\}$ can be carried out with standard nonlinear optimization techniques.

Example 2: (Discrete Choice Model)

Suppose now that y is 1 or 0 according to whether a commodity (say a car) is purchased or not. In this setup, the behavioral model may be of the form

$$y_{it} = f(x_{it}, u_{it}, \rho_t; \delta)$$
$$= 1(u_{it} \leq \delta_1 + \delta_2 \log x_{it} + \delta_3\rho_t),$$

where x_{it} denotes income, ρ_t denotes the commodity price and u_{it} represents individual heterogeneity, distributed normally with mean 0 and variance 1. Again supposing the aggregate $\hat{Y}_t = \overline{Y}_t$, the sample proportion of individuals in group t who purchase the commodity, and that x_{it} is lognormally distributed,

$$p(x \mid \theta_t, \tau_t) = \frac{1(x>0)}{2\pi\tau_t x} \exp\left(\frac{-(\log x-\theta_t)^2}{2\tau_t^2}\right),$$

the population aggregate $Y(Q_t)$ can be written as

$$Y(Q_t) = F\left[\frac{\delta_1 + \sigma_2\theta_t + \delta_3\rho_t}{\sqrt{1 + \delta_2^2\tau_t^2}}\right],$$

where F is the standard normal cumulative, as shown by McFadden and Reid (1975). In terms of the previous characteristic aggregates

$$\mu_t = (E(x_{it}), P_t^{-1}(1/2)) = (\text{mean}(x_{it}), \text{median}(x_{it}))$$

3. Large Sample Properties of Weighted Least Squares

3.1 Consistency and Asymptotic Normality

For convenience, we summarize the discussion of the previous section in the following assumptions:

Assumption A1: For each group t (t=1,...,T), N_t underlying observations (y_{it}, x_{it}, u_{it}), i=1, ..., N_t, are generated from the behavioral relation

$$y_{it} = f_t(x_{it}, u_{it}, \delta_o)$$

and i.i.d. draws of (x'_{it}, u_{it}) according to the probability law

$$\Pr((x'_{it}, u_{it}) \varepsilon A) = \int 1((x', u) \varepsilon A) g_t(u \mid x, \sigma_o) p_t(x \mid \theta_t) d\lambda$$

for λ-measurable $f_t(\cdot)$ and A, and for $\gamma_o \equiv (\delta'_o, \sigma'_o)'$ a fixed, finite dimensional vector of unknown parameters.

Assumption A2: For each group t = 1,...,T, the individual observations are summarized by empirical aggregates $\hat{Y}_t = Y(\hat{Q}_t)$ and $\hat{X}_t = X(\hat{P}_t)$. where Y and X are functionals defined on the space of distribution functions of y_{it} and x_{it}, respectively, and $\hat{Q}_t$ and $\hat{P}_t$ are empirical distribution functions for $(y_{it}, i = 1,..., N_t)$ and $(x_{it}, i = 1,..., N_t)$, respectively.

Assumption A3: The population aggregates $Y_t(Q_t)$ and $X_t(P_t)$ satisfy the relation

$$Y_t(Q_t) = \Phi_t(\mu_t, \gamma_o)$$

$$X_t(P_t) = \mu_t,$$

where the functional form of Φ_t is known for each t, and where Q_t and P_t are the marginal d.f's of y_{it} and x_{it} in group t.

To demonstrate strong consistency of the weighted least squares estimator $\hat{\gamma}_w$ of (2.11) as $N_t \to \infty$, we impose further regularity conditions on the unknown parameters, the aggregation functionals, and the group weights.

<u>Assumption A4</u>: The parameter vector γ_o is an element of a compact parameter space Γ.

<u>Assumption A5</u>: The aggregation functionals Y and X are Fisher consistent at Q_t and P_t for each t; i.e., if $\hat{Q}_t$ and $\hat{P}_t$ converge weakly to Q_t and P_t, then $Y(\hat{Q}_t) \to Y(Q_t)$ and $X(\hat{P}_t) \to X(P_t)$.

<u>Assumption A6</u>: For each t, the function $\Phi_t(\mu, \gamma)$ is continuous in γ for any μ, and Φ_t is continuous at μ_t uniformly in $\gamma \varepsilon \Gamma$, i.e.,

$$\sup_{\gamma \varepsilon \Gamma} | \Phi_t(\mu, \gamma) - \Phi_t(\mu_t, \gamma) | \to 0 \text{ as } \|\mu_t - \mu\| \to 0.$$

<u>Assumption A7</u>: The (possibly stochastic) weights $\{\hat{w}_t\}$ converge almost surely (as $N_t \to \infty$) to some nonnegative constants $\{w_t\}$.

<u>Assumption A8</u>: If $\gamma \neq \gamma_o$, $\gamma \varepsilon \Gamma$, then $\sum_{t=1}^{T} w_t \, (\Phi_t(\mu_t, \gamma) - \Phi(\mu_t, \gamma_o))^2 > 0$, where the weights $\{w_t\}$ are given in A7.

Considering each of these conditions in turn, assumption A4 is typically imposed in such nonlinear estimation problems as considered here; with the continuity of Φ_t in γ (given in A6), it ensures the existence and measurability of $\hat{\gamma}_w$, and is imposed in the lemma we use to show strong consistency. Assumption A5 and the uniform convergence (with probability

one) of $\hat{Q}_t$ and $\hat{P}_t$ to Q_t and P_t imply the almost sure convergence of $\hat{Y}_t$ and $\hat{X}_t$ to their respective population aggregates, as noted by Rao (1973, p.346). Conditions A6 and A7 allow us to demonstrate the uniform almost-sure convergence of the WLS minimand to the sum given in assumption A8; the identification condition in A8 is a stronger version of the notion of a "complete aggregation structure" introduced in Stoker (1984) (which, in this context, would require only that $\Phi_t(\mu, \gamma_1) - \Phi_t(\mu, \gamma_o)$ not be identically zero in μ for any $\gamma_1 \neq \gamma_o$).

With these preliminary conditions, strong consistency of $\hat{\gamma}_w$ is easily demonstrated.

<u>Theorem 3.1</u>: Under A1 - A7, the estimator $\hat{\gamma}_w$ of γ_o is strongly consistent, i.e., if

$$N \equiv \min\{N_t,\ t=1,\ \ldots,\ T\} \text{ and}$$

$$\hat{\gamma}_w = \underset{\gamma\varepsilon\Gamma}{\operatorname{argmin}} \sum_{t=1}^{T} w_t(\hat{Y}_t - \Phi_t(\hat{X}_t,\ \gamma))^2$$

$$\equiv \underset{\gamma\varepsilon\Gamma}{\operatorname{argmin}}\ S_N(\gamma),$$

then $\hat{\gamma}_w \to \gamma_o$ a.s. as $N \to \infty$.

<u>Proof</u>: As noted above, A2, A5, and the Glivenko-Cantelli lemma imply $\hat{X}_t \to \mu_t$ and $\hat{Y}_t \to \Phi_t(\mu_t,\ \gamma_o)$ a.s. as $N \to \infty$, and $\hat{w}_t \to w_t$ by assumption A7. Writing

$$(3.1) \qquad S(\gamma) = \sum_{t=1}^{T} w_t(\Phi_t(\mu_t,\ \gamma_o) - \Phi_t(\mu_t,\ \gamma))^2,$$

we clearly have

(3.2) $\sup_{\gamma \varepsilon \Gamma} | S_N(\gamma) - S(\gamma) | \to 0$ a.s. as $N \to \infty$,

since, for example,

$$\sup_{\gamma \varepsilon \Gamma} \sum_{t=1}^{T} \hat{w}_t(\Phi_t(\hat{X}_t, \gamma) - \Phi_t(\mu_t, \gamma))^2 \to 0, \text{ a.s.},$$

by assumption A6 and the strong consistency of $\hat{X}_t$ and $\hat{w}_t$. Because γ_o uniquely minimizes $S(\gamma)$ by assumption A8, Lemma 2 of Amemiya (1973) yields the almost sure convergence of $\hat{\gamma}_w$ to γ_o.

Before turning to conditions for asymptotic normality of $\hat{\gamma}_w$, we should reiterate the essential difference between this approach and the usual demonstration of consistency of extremum estimators. By taking T fixed as $N \to \infty$, we can write the estimator $\hat{\gamma}_w$ as a fixed, continuous function of $\{(\hat{Y}_t, \hat{X}_t', \hat{w}_t)\}$ for suitably large N, so consistency of $\hat{\gamma}_w$ follows from consistency of these aggregates and the properties of the function defining $\hat{\gamma}_w$. In contrast, the usual approach would have the minimized S_N and the resulting estimator $\hat{\gamma}_w$ varying with the sample size T, requiring stronger conditions (such as bounded moments of the aggregates $\hat{Y}_t$ and $\hat{X}_t$) to establish strong convergence. Indeed, taking N_t fixed for all t and assuming $T \to \infty$, we would not expect $\hat{\gamma}_w$, as defined above, to be consistent if Φ_t is nonlinear in $\hat{X}_t$. The appropriate regression function in the latter case would not be Φ_t but $E_t(\hat{Y}_t \mid \hat{X}_t)$, which, in addition to the computational problems noted in section 2, would in general involve the distribution parameters μ_t if the behavioral function f_t of A1 were nonlinear in (x_{it}, u_{it}). Consistent estimation in this context would appear to require an instrumental variables approach or strong restrictions on the incidental parameters $\{\mu_t\}$.

To show asymptotic normality of the estimator $\hat{\gamma}_w$, we further restrict the model with the following conditions:

Assumption A9: The parameter vector γ_o is an interior point of the parameter space Γ.

Assumption A10: The empirical aggregates $\hat{Y}_t$ and $\hat{X}_t$ are of the form

$$\hat{Y}_t = \Phi_t(\mu_t, \gamma_o) + \frac{1}{N_t} \sum_{i=1}^{N_t} \xi_t(y_{it}, \mu_t, \gamma_o) + o_p(1/\sqrt{N_t})$$

$$\equiv \Phi_t(\mu_t, \gamma_o) + \frac{1}{N_t} \sum_{i=1}^{N_t} \xi_{it} + o_p(1/\sqrt{N_t})$$

and $$\hat{X}_t = \mu_t + \frac{1}{N_t} \sum_{i=1}^{N_t} \zeta_t(x_{it}, \mu_t) + o_p(1/\sqrt{N_t})$$

$$\equiv \mu_t + \frac{1}{N_t} \sum_{i=1}^{N_t} \zeta_{it} + o_p(1/\sqrt{N_t}),$$

with $E_t\,(\xi_{it}, \zeta'_{it}) = 0$ and

$$E_t \begin{bmatrix} \xi_{it} \\ \zeta_{it} \end{bmatrix} (\xi_{it}, \zeta'_{it}) \equiv \Sigma_t \equiv \begin{bmatrix} \sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix}_t .$$

Assumption A11: The function $\Phi_t\,(\mu, \gamma)$ is continuously differentiable at $\mu = \mu_t$, $\gamma = \gamma_o$ for each t.

Assumption A12: For $N = \min\{N_t, t=1, \ldots, T\}$,

$$\lim_{N\to\infty} (N/N_t) = \kappa_t \;\; \varepsilon\; (0, 1] \text{ for each } t.$$

Of these additional conditions, only A10 requires further elaboration. This assumption requires the aggregation functionals $\hat{Y}_t$ and $\hat{X}_t$ to be sufficiently regular to admit a Taylor's series expansion at $\hat{Q}_t$ and $\hat{P}_t$; the functions $\xi_t(y, \mu_t, \gamma_o)$ and $\zeta_t(x, \mu_t)$ are then the influence curves

associated with $\hat{Y}_t$ and $\hat{X}_t$ (see Huber [1982], Chapter 1). When $Y(\cdot)$ and $X(\cdot)$ are expectation operators, the representation in A10 is trivial, while if $\hat{Y}_t = \text{median}\ \{y_{it},\ i=1,\ \ldots,\ N_t\}$, for example, it can be shown that

$$(3.3) \qquad \hat{Y}_t = \Phi_t\ (\mu_t,\ \gamma_o) + \frac{1}{N_t} \sum_{i=1}^{N_t} [2\ \phi_t(\mu_t,\ \gamma_o)]^{-1}\ \text{sgn}(y_{it} - \Phi_t(\mu_t,\ \gamma_o)) + o_p(1/\sqrt{N_t})$$

if y_{it} is continuously distributed with positive density at $\Phi_t(\mu_t,\ \gamma_o) = Q_t^{-1}(1/2)$, where $\phi_t(\mu_t,\ \gamma_o)$ is the density evaluated at Φ_t.

<u>Theorem 3.2</u>: Under A1 - A12, the weighted least squares estimator $\hat{\gamma}_w$ is asymptotically normal,

$$\sqrt{N}\ (\hat{\gamma}_w - \gamma_o) \xrightarrow{d} N(0,\ M^{-1} V\ M^{-1}) \text{ as } N \to \infty,$$

where

$$M = \sum_{t=1}^{T} w_t \left[\frac{\partial \Phi_t}{\partial \gamma} \cdot \frac{\partial \Phi_t}{\partial \gamma'} \Bigg|_{\gamma_o,\ \mu_t} \right],$$

$$V = \sum_{t=1}^{T} w_t^2\ \kappa_t\ (v_t'\ \Sigma_t\ v_t) \left[\frac{\partial \Phi_t}{\partial \gamma} \cdot \frac{\partial \Phi_t}{\partial \gamma'} \Bigg|_{\gamma_o,\ \mu_t} \right],$$

for w_t, κ_t, and σ_t defined above and

$$v_t' \equiv \left(1,\ -\frac{\partial \Phi}{\partial \mu'} \Bigg|_{\gamma_o, \mu_t}\right).$$

<u>Proof</u>: We first note that, for each t,

$$(3.4) \qquad \sqrt{N} \begin{bmatrix} \hat{Y}_t - \Phi_t(\mu_t,\ \gamma_o) \\ \hat{X}_t - \mu_t \end{bmatrix} \xrightarrow{d} N(0,\ \kappa_t \Sigma_t),$$

independently across t, by application of the Lindeberg-Levy central limit theorem. Since γ_o is interior to Γ and $\hat{\gamma}_w \to \gamma_o$ a.s. by Theorem 1, the first-order condition

$$(3.5) \qquad 0 = \sum_{t=1}^{T} \hat{w}_t(\hat{Y}_t - \Phi_t(\hat{X}_t, \hat{\gamma}_w))\left[\frac{\partial\Phi_t}{\partial\gamma}\bigg|_{\hat{X}_t,\ \hat{\gamma}_w}\right]$$

must be satisfied for all N suitably large by the definition of $\hat{\gamma}$. Making a further mean-value expression,

$$(3.6) \quad 0 = \sum_{t=1}^{T} \hat{w}_t \left[(\hat{Y}_t - \Phi_t(\mu_t, \gamma_o) - \frac{\partial\Phi_t}{\partial\mu_t'}\bigg|_{\mu_t^*, \gamma_t^*}(\hat{X}_t - \mu_t) + \frac{\partial\Phi_t}{\partial\gamma'}\bigg|_{\mu_t^*, \gamma_t^*}(\hat{\gamma}_w - \gamma_o)\right]\left[\frac{\partial\Phi_t}{\partial\gamma}\bigg|_{\hat{X}_t, \hat{\gamma}_w}\right],$$

for μ_t^*, γ_t^* a convex combination of $\hat{X}_t$, $\hat{\gamma}_w$ and μ_t, γ_o. Since $\hat{w}_t$, $\hat{X}_t$ and $\hat{\gamma}_w$ converge a.s. to w_t, μ_t, γ_o, and the matrix M defined above is nonsingular by A8 and A11, the expression above implies

$$(3.7) \qquad \sqrt{N}\,(\hat{\gamma}_w - \gamma_o) = M^{-1} \sum_{t=1}^{T} w_t \sqrt{N}\left[(\hat{Y}_t - \Phi_t(\mu_t, \gamma_o)) - \frac{\partial\Phi_t}{\partial\mu_t'}\bigg|_{\mu_t, \gamma_t}(\hat{X}_t - \mu_t)\right]\left[\frac{\partial\Phi_t}{\partial\gamma}\bigg|_{\mu_t,\ \gamma_o}\right] + o_p(1)$$

by the continuous diffentiability of Φ_t. The result then follows by the usual properties of the multivariate normal distribution.

3.2 Estimation of the Asymptotic Covariance Matrix

The proof of Theorem 3.2 shows that, for large N, the estimator $\hat{\gamma}_w$ behaves like a weighted linear regression of $\{\sqrt{N}[\hat{Y}_t - \Phi_t(\mu_t, \gamma_o), \hat{X}_t' - \mu_t']v_t\}$ on the vectors $\{\partial\Phi_t(\mu_t, \gamma_o)/\partial\gamma\}$. If κ_t and $v_t'\Sigma_t v_t$ were known (or could be consistently estimated as $N \to \infty$), an efficient choice of weights $\{\hat{w}_t\}$ (in the Gauss-Markov sense) would satisfy

$$(3.8) \qquad \lim_{N\to\infty} \frac{\hat{w}_t}{\hat{w}_s} = \left[\frac{\kappa_s\, v_s'\, \Sigma_s\, v_s}{\kappa_t\, v_t'\, \Sigma_t\, v_t}\right]^{1/2}, \text{ a.s.}$$

In some circumstances, the matrices $\{\Sigma_t\}$ will be completely determined by the unknown behavioral parameters γ_o and the population characteristic aggregates $\{\mu_t\}$ (e.g. in Example 2 of Section 2 above); in this situation, construction of optimal weights $\hat{w}_t$ using $\hat{\Sigma}_t = \Sigma_t(\hat{X}_t, \hat{\gamma}_w)$ is straightforward. In other cases, knowledge of the population aggregates $\{Y(Q_t)\}$ and $\{X(P_t)\}$ will be insufficient to determine the elements of $\{\Sigma_t\}$. In the continuous demand example of section 2, the covariance matrices $\{\Sigma_t\}$ involve the unknown variance σ_o^2 of the unobserved heterogeneity terms $\{u_{it}\}$; this parameter is clearly unidentified given only the location parameters of the data, i.e., given only the population means of $\{y_{it}\}$ and the population means and medians of $\{x_{it}\}$. In general, dependence of the distribution of y_{it} on unknown nuisance parameters not included in γ_o inhibits not only construction of efficient weights but also consistent (for T fixed) estimation of the covariance matrix of $\hat{\gamma}_w$, which will depend upon these nuisance parameters (through Σ_t).

Consistent estimation of the covariance matrix of $\hat{\gamma}_w$ in this situation does appear feasible if T, the number of aggregation groups, also tends to infinity with N, the minimum group size. While we do not explore the details here, it seems reasonable to expect that the nuisance parameters in Σ_t can be consistently estimated using method-of-moments estimation applied to the "between group" averages of squared values of the residuals $\sqrt{N}\,(\hat{Y}_t - \Phi_t(\hat{X}_t, \hat{\gamma}_w)) \equiv \hat{e}_t$, whose limiting distributions have covariances involving Φ_t, μ_t, γ_o, and the nuisance parameters. Alternatively, following Eicker (1967) and White (1980), we could estimate V, the matrix given in the

statement of Theorem 3.2, by

$$(3.9) \qquad \hat{V} = \sum_{t=1}^{T} \hat{w}_t \left(\frac{N_t}{N}\right) \hat{e}_t^{\,2} \left[\frac{\partial \Phi_t}{\partial \gamma} \cdot \frac{\partial \Phi_t}{\partial \gamma'} \Big| \hat{X}_t, \hat{\gamma}_w \right] ,$$

for e_t defined above and $\{\hat{w}_t\}$ a given (not necessarily optimal) set of weights. Because the proofs of Theorems 3.1 and 3.2 used the assumption of fixed T, the foregoing results cannot immediately be generalized to permit $T \to \infty$; still, the results seem very likely to hold in this case, with further regularity conditions, and we conjecture that, under such conditions,

$$(3.10) \qquad \operatorname*{plim}_{T \to \infty} T^{-1}[\hat{V} - V] = 0$$

provided $T/N \to 0$ as $T \to \infty$, so that large sample inference on γ_0 will be feasible if the number of groups is large.

3.3 Relation to Cross Section Regression

It is clear from the above development that $v_t' \Sigma_t v_t$ (normalized by N_t^{-1}) is the relevant "true" aggregate residual variance of departures of $\hat{Y}_t$ from $\Phi_t(\hat{X}_t, \gamma_0)$. In this section we point out how $v_t' \Sigma_t v_t$ can be regarded as the residual variance from cross section linear regression, which indicates how nonlinearities in micro behavior affect the size of the aggregate residual variance.

For the remainder of this section we focus attention on a single time period t. Consider the individual i.i.d. influence terms ξ_{it} and ζ_{it} of assumption A10, for $i=1,\ldots,N_t$ (or a smaller random sample) and define $s_{yy} = \sum_i \xi_{it}^{*\,2}/N_t$, $S_{xy} = \sum_i \zeta_{it}^{*} \xi_{it}^{*}/N_t$ and $S_{xx} = \sum_i \zeta_{it}^{*} \zeta_{it}^{*\,\prime}/N_t$ as the sample covariances, with

$$S = \begin{bmatrix} s_{yy} & S_{xy}' \\ S_{xy} & S_{xx} \end{bmatrix}$$

where $\xi^*_{it} = \xi_{it} - \bar{\xi}_{it}$, $\zeta^*_{it} = \zeta_{it} - \bar{\zeta}_{it}$ are the deviations from sample averages. By a straightforward application of the (i.i.d.) Strong Law of Large Numbers, we have that $\lim S = \Sigma_t$ a.s. under A1, A10 and A11.

We can obtain several interpretative results by considering cross section linear equations of the form

$$(3.11) \qquad \xi^*_{it} = b'\zeta^*_{it} + u_{it} \qquad i=1,\ldots N_t$$

where B could be considered as the slope coefficient of the regression of ξ_{it} on ζ_{it} which includes a constant term. We denote the OLS coefficient estimates as $\hat{\beta}_t = (S_{xx})^{-1}S_{xy}$, and $\beta_t = \lim \hat{\beta}_t = (\Sigma_{xx,t})^{-1}\Sigma_{xy,t}$. Of special interest is the large sample residual variance from (3.11), which we denote as the almost sure limit $\sigma(b) = \lim (\sum(\xi^*_{it} - b'\zeta^*_{it})^2/N_t)$. Clearly $\sigma(b)$ is minimized at $b = \beta_t$, and we denote the minimum as $\sigma(\beta_t) = \tilde{\sigma}$.

The first interpretative result is available by direct computation -- namely, under assumptions A1, A10, A11, $\sigma\left[\frac{\partial\Phi}{\partial\mu}\Big|_{\mu_t,\gamma_o}\right] = v_t'\Sigma_t v_t$.

Consequently, $v_t'\Sigma_t v_t$ can be viewed as the cross section variance of the difference $\xi^*_{it} - \left[\frac{\partial\Phi}{\partial\mu}\Big|_{\mu_t,\gamma_o}\right]\zeta^*_{it}$, or as the large sample residual variance of equation (3.11) where b is set to $b = \frac{\partial\Phi}{\partial\mu}\Big|_{\mu_t,\gamma_o}$, the macroeconomic effects.

Of more interest is the condition under which $v_t'\Sigma_t v_t = \tilde{\sigma}$, or that $v_t'\Sigma_t v_t$ represents the (least squares) cross section residual variance of ξ_{it} regressed on ζ_{it}. In view of the above observation, an obvious sufficient condition occurs when $\lim \hat{\beta}_t = \beta_t = \frac{\partial\Phi}{\partial\mu}\Big|_{\mu_t,\gamma_o}$, or when $\hat{\beta}_t$ consistently estimates the macroeconomic effects. This property is studied extensively by

Stoker (1982, 1983). We state the following proposition, which easily follows from the development of Stoker (1983).

Theorem 3.3: Under Assumptions A1-A6, A10, and A11, a sufficient condition for $v_t'\Sigma_t v_t = \tilde{\sigma}$ occurs when the influence function $\zeta(x,\mu)$ can be written as

(3.12) $\zeta(x,\mu) = B(\mu)\ell(x,\mu)$

for all x and for μ in a neighborhood of μ_o, where $B(\mu)$ is a nonsingular matrix and $\ell(x,\mu)$ is the score vector

(3.2) $$\ell(x,\mu) = \frac{\partial \ln \tilde{p}(x|\ \mu_t)}{\partial \mu}.$$

Condition (3.12) is valid (locally) if and only if the aggregate X_t is a (locally) efficient estimator of μ_t.

Thus the true aggregate residual variance $v_t'\Sigma_t v_t$ coincides with the linear cross section residual variance when $\zeta(x,\mu)$ is (locally) collinear with the score $\ell(x,\mu)$. In particular, when $B(\mu)$ is the inverse of the information matrix $I_\mu = E(\ell\ell'|\mu)$, as in

(3.14) $\zeta(x,\mu) = (I_\mu)^{-1}\ell(x,\mu)$,

then we have $\lim \hat{\beta}_t = \beta_t = \frac{\partial\Phi}{\partial\mu}\Big|_{\mu_t,\gamma_o}$. Moreover, the score condition is associated with the efficiency properties of the consistent aggregate $\hat{X}_t$, being globally valid when $\hat{X}_t$ is efficient (and hence a sufficient statistic) for μ. The condition holds locally if and only if $\hat{X}_t$ is locally sufficient and efficient in a neighborhood of $\mu=\mu_t$. While we do not provide proof here, it is easy to see that if $\hat{X}_t$ is consistent and locally efficient for μ, then the influence function can be written in the restricted form (3.14).

The sufficient conditions of Theorem 3.3 are of interest because they characterize the aggregate residual variance $v_t'\Sigma_t v_t$ as a cross section OLS residual variance, regardless of the true form of the micro behavioral model $y_{it} = f_t(x_{it}, u_{it})$. Thus, when the aggregates $\hat{X}_t$ are efficient estimators of the parameters μ_t, the aggregate residual variance $v_t'\Sigma_t v_t = \sigma(\beta_t)$ does not depend on whether the cross section OLS residuals $\xi^*_{it}-\beta_t'\zeta^*_{it}$ arise from structural nonlinearities of f_t in x_{it} or from random disturbances. While other sufficient conditions can be found, they will necessarily depend on restrictions on the functional form of micro behavior $y_{it} = f_t(x_{it}, u_{it})$, such as conditions which would imply that ξ_{it} is a linear function of ζ_{it} plus an independent disturbance.

For motivation, it may be useful to consider these interpretations when the aggregates $\hat{X}_t$ and $\hat{Y}_t$ are simple averages of the respective micro variables. For $\hat{X}_t = \overline{X}_t = \sum x_{it}/N_t$, (3.12) is valid (locally) if and only if $p(x \mid \mu)$ is (locally) in exponential family form with driving variables X:

(3.15) $\quad \tilde{p}(x \mid \mu) = c(\mu)h(x)\exp[\pi(\mu)'x]$

In this case $\zeta_{it} = x_{it} - \mu_t$ and $\hat{\beta}_t$ is the OLS slope coefficient vector of ξ_{it} regressed on x_{it}. When in addition, $\hat{Y}_t = \overline{Y}_t = \sum y_{it}/N_t$, we have $\xi_{it} = y_{it}-\Phi(\mu_t, \gamma_o)$, so that β_t is the OLS slope coefficient of y_{it} regressed on x_{it}, and $\tilde{\sigma} = v_t'\Sigma_t v_t$ is the large sample residual variance from the regression.

4. First-Order Equivalence of Weighted Least Squares and Minimum Distance Estimation

In the previous section we discussed how the efficiency of the weighted least squares estimator $\hat{\gamma}_w$ would be maximized through judicious choice of weights $\{\hat{w}_t\}$. In this section we compare the large sample behavior of $\hat{\gamma}_w$ to a broader class of "minimum distance" estimators, such as the maximum likelihood estimator $\tilde{\gamma}_w$ of (2.9) above. Under appropriate conditions, it will be shown that such estimators will be first order equivalent to a weighted least squares estimator, with weights depending upon the form of the minimand defining the estimator. An interesting result is that the corresponding estimators of μ_t need not be first order equivalent to $\hat{X}_t$, even though the estimator of γ_o is first-order equivalent to some $\hat{\gamma}_w$ which implicitly uses $\mu_t = \hat{X}_t$ in its construction.

Initially, we restrict attention to a narrower class of estimators than one which includes maximum likelihood, namely, estimators of γ_o and $\mu_o = (\mu_1, \ldots, \mu_T)'$ defined by

$$(4.1) \quad \begin{matrix} \tilde{\gamma} \\ \tilde{\mu} \end{matrix} = \underset{\gamma, \mu}{\operatorname{argmin}} \sum_{t=1}^{T} \nu_t (\hat{Y}_t, \hat{X}_t, \gamma, \mu_t),$$

where the $\{\nu_t(\cdot)\}$ are fixed "distance functions" with properties to be specified below. The maximum likelihood estimator of (2.9) is not in this class, since its "distance function" depends upon the group sample size N_t as well as the group index t; however, we will extend our results to include this case as well.

We impose the following conditions on the parameter vector μ_o and the criterion functions $\{\nu_t\}$:

Assumption B1: The vector μ_o of true characteristic aggregates is an interior point of a compact parameter space M.

Assumption B2: For each t and all $\gamma \varepsilon$ int Γ, $\mu \varepsilon$ int M,

$\nu_t(y, x, \gamma, \mu) = 0$ if $y = \Phi_t(\mu, \gamma)$ and $x = \mu$ and

$\nu_t(y, x, \gamma, \mu) > 0$ otherwise.

Assumption B3: The $\{\nu_t\}$ are fixed functions which are twice continuously differentiable in their arguments.

Assumption B4: Under assumptions A1 through A12, the minimum distance estimators $\tilde{\gamma}$ and $\tilde{\mu}$ are $\sqrt{N}$ - consistent, i.e., $\sqrt{N}(\tilde{\gamma} - \gamma_o) = O_p(1)$ and $\sqrt{N}(\tilde{\mu} - \mu_o) = O_p(1)$.

The points $\{\mu_t\}$ are assumed to be interior points of M in B1 so that Taylor's series arguments are applicable; compactness of M is not used in the theorem which follows, but is needed to extend the results when the distance functions depend upon N_t. B2 makes more precise the sense in which $\nu_t(\cdot)$ reflects the "distance" between the population aggregates $(Y(Q_t), X(P_t))$ and any arbitrary (y,x), while B3 permits criterion functions to be approximated by a quadratic form in a neighborhood of the true values $(y, x, \gamma, \mu) = (\Phi_t(\mu_t, \gamma_o), \mu_t, \gamma_o, \mu_t)$. Finally, condition B4 restricts attention to mimimum distance estimators which are interesting alternatives to $\hat{\gamma}_w$, ruling out superefficient or inconsistent estimators; the condition is imposed in lieu of a list of assumptions which would imply it, since our objective is only to show that such estimators are of weighted least squares form asymptotically.

We note that $\hat{\gamma}_w$ and the $\{\hat{X}_t\}$ are not in the class of mimimum distance estimators defined here unless the weights w_t are fixed and nonstochastic and the functions $\Phi_t(\mu, \gamma)$ are twice differentiable, in which case

$$(4.2) \qquad \nu_t(y, x, \gamma, \mu) = w_t(y-\Phi_t(x, \gamma))^2 + \|x-\mu\|^2$$

would generate $\tilde{\gamma} = \hat{\gamma}_w$ and $\tilde{\mu}_t = \hat{X}_t$ in (4.1).

<u>Theorem 4.1</u>: Under A1 - A12 and B1 - B4, the minimum distance estimator $\tilde{\gamma}$ of γ_o given by (4.1) is first-order equivalent to a weighted least squares estimator,

$$\sqrt{N}\,(\tilde{\gamma} - \hat{\gamma}_w) = o_p(1),$$

where $\hat{\gamma}_w$ is of the form (2.11) with weights proportional to

$$w_t = \left[\frac{\partial^2 \nu_t}{\partial y^2} + \frac{\partial^2 \nu_t}{\partial y \partial \mu'}\left[\frac{\partial^2 \nu_t}{\partial \mu \partial \mu'}\right]^{-1}\frac{\partial^2 \nu_t}{\partial \mu \partial y}\right],$$

with the derivatives being evaluated at

$$(y, x, \gamma, \mu) = [\Phi_t(\mu_t, \gamma_o), \mu_t, \gamma_o, \mu_t].$$

<u>Proof</u>: Let H denote the matrix of second partials of $\nu_t(y, x, \gamma, \mu)$ evaluated at $[\Phi_t(\mu_t, \gamma_o), \mu_t, \gamma_o, \mu_t)]$. Since $\nu_t(\Phi_t(\mu,\gamma), \mu, \gamma, \mu) = 0$ identically in γ and μ, H_o must satisfy the restrictions

$$(4.3) \qquad \begin{bmatrix} \partial\Phi_t/\partial\gamma & 0 & I_l & 0 \\ \partial\Phi_t/\partial\mu & I_k & 0 & I_k \end{bmatrix} H = 0,$$

where the derivatives are evaluated at μ_t and γ_o, $k = \dim(\mu_t)$, $l = \dim(\gamma)$, and I_m is the identity matrix of order m. We denote the various component submatrices of H by a double subscript, e.g.,

$$H_{\mu\gamma} = \left.\frac{\partial^2 \nu_t}{\partial \mu \partial \gamma'}\right|_{(\Phi_t(\mu_t, \gamma_o), \mu_t, \gamma_o, \mu_t)}, \text{ etc.}$$

With this notation, we observe that, with arbitrarily high probability, the minimum distance estimators must satisfy the first-order condition

$$(4.4) \qquad 0 = \sum_{t=1}^{T} \frac{\partial \nu_t}{\partial \gamma} (\hat{Y}_t, \hat{X}_t, \tilde{\gamma}, \tilde{\mu}_t)$$

for N suitably large. Making a mean-value expansion of the left-hand side of (4.4) at the limiting values, we have

$$(4.5) \qquad 0 = \sum_{t=1}^{T} H_{\gamma y} \sqrt{N} (\hat{Y}_t - \Phi_t(\mu_t, \gamma_o)) + H_{\gamma x} \sqrt{N} (\hat{X}_t - \mu_t)$$
$$+ H_{\gamma\gamma} \sqrt{N} (\tilde{\gamma} - \gamma_o) + H_{\gamma\mu} \sqrt{N} (\tilde{\mu}_t - \mu_t) + o_p(1),$$

where the convergence of the remainder term to zero in probability follows from the continuity of the hessian matrix and the $\sqrt{N}$ - consistency of $\hat{Y}_t$, $\hat{X}_t$, $\tilde{\gamma}$, and $\tilde{\mu}_t$. Now, using the restrictions given in (4.3),

$$(4.6) \qquad 0 = \sum_{t=1}^{T} H_{yy} \left[\frac{\partial \Phi_t}{\partial \gamma}\right] \left[\sqrt{N} (\hat{Y}_t - \Phi_t(\mu_t, \gamma_o)) - \left[\frac{\partial \Phi_t}{\partial \mu'}\right] \sqrt{N} (\hat{X}_t - \mu_t) - \left[\frac{\partial \Phi_t}{\partial \gamma'}\right] (\tilde{\gamma} - \gamma_o) \right]$$
$$+ \sum_{t=1}^{T} \left[\frac{\partial \Phi_t}{\partial \gamma}\right] H_{y\mu} \sqrt{N} (\hat{X}_t - \tilde{\mu}_t) + o_p^{*}(1),$$

where, as always, all derivative are evaluated at the true values.

To obtain an expression for the term involving $(\hat{X}_t - \tilde{\mu}_t)$, we use the first-order conditions for $\tilde{\mu}_t$,

$$(4.7) \qquad 0 = \frac{\partial \nu_t}{\partial \mu} (\hat{Y}_t, \hat{X}_t, \tilde{\gamma}, \tilde{\mu}_t) \quad ,$$

$$= H_{\mu y} \sqrt{N} (\hat{Y}_t - \Phi_t(\mu_t, \gamma_o)) + H_{\mu x} \sqrt{N} (\hat{X}_t - \mu_t)$$
$$+ H\mu\gamma \sqrt{N} (\tilde{\gamma} - \gamma_o) + H_{\mu\mu} \sqrt{N} (\hat{X}_t - \tilde{\mu}_t) + o_p(1)$$

which also must hold in probability for large N by B1 and B4. Again using the total-differential restrictions of (4.3), we can express the difference $\sqrt{N}\,(\hat{X}_t - \tilde{\mu}_t)$ as

$$(4.8)\qquad \sqrt{N}\,(\hat{X}_t - \tilde{\mu}_t) = H_{\mu\mu}^{-1} H_{\mu y} \Big[\sqrt{N}\,(\hat{Y}_t - \Phi_t(\mu_t, \gamma_o)) + \frac{\partial \Phi_t}{\partial \mu'} \sqrt{N}\,(X_t - \mu_t) + \frac{\partial \Phi_t}{\partial \gamma'} \sqrt{N}(\tilde{\gamma} - \gamma_o) \Big] + o_p(1),$$

where the nonsingularity of $H_{\mu\mu}$ follows from the assumed $\sqrt{N}$ - consistency of $\tilde{\mu}_t$. Combining (4.5) and (4.7) yields

$$(4.9)\qquad 0 = \sum_{t=1}^{T} \left[H_{yy} + H_{y\mu} H_{\mu\mu}^{-1} H_{\mu y}\right] \left[\frac{\partial \Phi_t}{\partial \gamma}\right] \Big[\sqrt{N}\,(\hat{Y}_t - \Phi_t(\mu_t, \gamma_o)) + \frac{\partial \Phi_t}{\partial \mu'} \sqrt{N}\,(\hat{X}_t - \mu_t) + \frac{\partial \Phi_t}{\partial \gamma'} \sqrt{N}\,(\tilde{\gamma} - \gamma_o) \Big] + o_p(1).$$

Hence $\sqrt{N}\,(\tilde{\gamma} - \gamma_o)$ can be written exactly in the form of equation (3.7) above, with

$$(4.10)\qquad w_t = \left[H_{yy} + H_{y\mu} H_{\mu\mu}^{-1} H_{\mu y}\right],$$

which is nonnegative by B2 and B3.

As noted above, Theorem 3 applies only to estimators with distance functions which do not vary with N_t; thus, both the maximum likelihood and weighted least squares estimators are excluded from this class (the latter due to its dependence on estimated weights $\hat{w}_t$). To extend the result of Theorem 3 to the case in which the distance functions $\hat{\nu}_t(y, x, \gamma, \mu)$ are (possibly stochastic) functions of N_t, we impose the following conditions:

<u>Assumption B5</u>: The functions $\{\hat{\nu}_t\}$ are twice continuously differentiable in their arguments (with probability one) for all N_t.

Assumption B6: The minimum distance estimators based upon $\{\hat{\nu}_t\}$ are $\sqrt{N}$ - consistent.

Assumption B7: The matrix of second partials of $\hat{\nu}_t$ converges to the corresponding hessian of some function ν_t satisfying B2 and B3, uniformly in $\gamma\varepsilon\Gamma$, μ and $x\varepsilon M$, and $y\varepsilon\{\Phi_t(\mu, \gamma): \mu\varepsilon M, \gamma\varepsilon\Gamma\}$, with probability one.

Corollary 1: Under A1 - A12 and B1 - B7, the minimum distance estimator $\tilde{\gamma}$ based upon the distance function $\hat{\nu}_t$ satisfies the conclusion of Theorem 3.

Proof: The expressions (4.5) and (4.7) above for $\tilde{\gamma}$ and $\tilde{\mu}_t$ are still valid, since the hessian of $\hat{\nu}_t$ evaluated at $(\hat{Y}_t, \hat{X}_t, \tilde{\gamma}, \tilde{\mu}_t)$ converges in probability to H by Lemma 4 of Amemiya [1973].

This corollary extends the result of Theorem 3 to a much broader class of estimators, including the WLS estimator $\hat{\gamma}_W$ (which trivially satisfied the conclusion of Theorem 3, and now satisfies the hypothesis B5 - B7 as well) and the maximum likelihood estimation $\hat{\gamma}_M$ of (2.9). Strictly speaking, application of the corollary to maximum likelihood estimation requires that the log density, $\log d_t(\hat{Y}_t, \hat{X}_t \mid \gamma, \mu, N_t)$, can be transformed into a distance function $\hat{\nu}_t(\hat{Y}_t, \hat{X}_t, \gamma, \mu)$ satisfying B5 - B7 through proper normalization (e.g., multiplication by $-N^{-1/2}$); while we suspect such a transformation is feasible for most applied problems, it is beyond the scope of this paper to provide sufficient conditions for equivalence of maximum likelihood and minimum distance estimation because of the generality of the aggregates $\hat{Y}_t$ and $\hat{X}_t$ considered.

An interesting aspect of Theorem 4.1 is that the asymptotic

distributions of $\tilde{\mu}_t$, the minimum distance estimators of the population aggregates, and $\hat{X}_t$, the empirical characteristic aggregate, need not coincide. Equation (4.8) implies

$$(4.10) \qquad \sqrt{N}(\hat{X}_t - \tilde{\mu}_t) = \sqrt{N}(\hat{Y}_t - \Phi_t(\hat{X}_t, \tilde{\gamma}))\alpha_t + o_p(1),$$

where $\alpha_t = [\partial^2 \nu_t/\partial\mu\partial\mu']^{-1} [\partial^2 \nu_t/\partial\mu\partial y]$. That is, the estimators $\tilde{\mu}_t$ may exploit the dependence of the distribution of $\hat{Y}_t$ (as well as $\hat{X}_t$) on the population characteristics μ_t in its estimation. For maximum likelihod estimates, it is clear that α_t will equal zero if $\hat{X}_t$ is a sufficient statistic for μ_t, as it will, for example, if $\hat{X}_t = \bar{X}_t$ and the distributions of the characteristics x_{it} are in the exponential family given in (3.15); the likelihood function will then factor into the sum of the conditional log likelihood of $\hat{Y}_t$ given $\hat{X}_t$ and γ (which will not depend on μ_t by sufficiency) and the marginal log density of $\hat{X}_t$ given μ_t. In general, though, when $\hat{X}_t$ is not (locally) sufficient for μ_t -- that is, when the ζ_{it} are not the scores corresponding to maximum likelihood estimation of μ_t based upon cross-section data -- we may expect the optimal α_t to differ from zero for estimation of the population characteristic vector.

Finally, we note that the results in this and the previous section do not depend upon the special form of the empirical aggregates $\hat{Y}_t = Y(\hat{Q}_t)$ and $\hat{X}_t = X(\hat{P}_t)$ assumed here; all that is required for Theorems 3.1, 3.2, and 4.1 are the consistency and asymptotic normality (of order $\sqrt{N}$) of the aggregates $\hat{Y}_t$ and $\hat{X}_t$ about $\Phi_t(\mu_t, \gamma_o)$ and μ_t respectively. Thus, for example, the results will still hold if $\hat{Y}_t$ is constructed from a stratified sample of the $\{y_{it}\}$, with stratification on the basis of the characteristics $\{x_{it}\}$. We have chosen to focus on the more restrictive forms of $\hat{Y}_t$ and $\hat{X}_t$ because they represent the most common type of aggregate available, that is, simple

averages or percentiles from marginal tabulations of $\{y_{it}\}$ and $\{x_{it}\}$. It should be clear that more efficient aggregates could be computed given the original micro-data and knowledge of the micro-model (Assumption A1); in such a case, though, no aggregation problem would be present, and the usual results on efficiency of maximum likelihood estimation of γ_0 would apply.

5. Conclusion

The theorems given in sections 3 and 4 above provide the theoretical basis for large-sample inference on microeconomic behavioral parameters using empirical aggregates. Viewed in a broader context, the results of this paper indicate that the most difficult step in constructing and implementing models of aggregate data (that properly account for the problem of aggregation over individuals) is the construction of the appropriate aggregate function through integration (when the aggregates are sample means) or more general determination of the population aggregates $Y(Q_t)$ and $X(P_t)$ as functions of the micro parameters. Once the aggregate function is formulated, parameters can be consistently estimated by just inserting aggregate data and performing least squares. Efficiency gains are potentially available by optimally weighting observations, as indicated above; moreover, using the optimal weights exhausts all of the (first order) efficiency gains available. Given the validity of the conjectures of section 3.2, hypothesis tests for the micro parameters can be performed in entirely standard fashion.

While we have indicated the generality of the framework, several extensions are of interest to future work. With regard to applications with time series aggregate data, a natural questions regards how to incorporate autocorrelated individual stochastic terms into estimation, as suggested by the standard interpretation of the disturbances as unobserved attributes of

individual agents (for example, demographic effects which are not explicitly modeled). It is clear that this complication would only affect the efficiency discussions above, with the consistency of the estimators proposed here not affected. A second extension concerns how to incorporate the endogeneity of predictor variables at the micro and/or macro level, which is indicated for problems where simultaneous equations modeling and/or expectations modeling is important. It should be noted here that our results clearly would be applicable to the estimation of reduced form equations in such circumstances.

REFERENCES

Amemiya, T. [1973], "Regression Analysis When the Dependent Variable is Truncated Normal," Econometrica, 41: 997-1016.

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland [1975], Discrete Multivariate Analysis: Thoery and Practice. Cambridge, MA: MIT Press.

Deaton, A. and J. Muellbauer [1980], "An Almost Ideal Demand System," American Economic Review, 70: 312-326.

Eicker, F. [1967], "Limit Theorems for Regressions with Unequal and Dependent Errors," Proceedings of the Fifth Berkeley Symposium, 1: 59-82.

Huber, P.J. [1981], Robust Statistics. New York: John Wiley and Sons.

Jorgenson, D.W., L.J. Lau, and T.M. Stoker [1982], "The Transcendental Logarithmic Model of Aggregate Consumer Behavior," in R. Basmann and G. Rhodes, eds., Advances in Econometrics. Greenwich, CT: JAI Press.

McFadden, D. and F. Reid [1975], "Aggregate Travel Demand Forecasting From Disaggregated Behavioral Models," Transportation Research Record, No. 534.

Muellbauer, J. [1975], "Aggregation, Income Distribution and Consumer Demand," Review of Economic Studies, 42: 525-543.

Rao, C.R. [1973], Linear Statistical Inference and Its Applications, Second Edition. New York: Wiley.

Stoker, T.M. [1982], "The Use of Cross-Section Data to Characterize Macro Functions," Journal of the American Statistical Association, 77: 369-380.

Stoker, T.M. [1983], "Aggregation, Efficiency, and Cross Section Regression," manuscript, MIT Sloan School of Management.

Stoker, T.M. [1984], "Completeness, Distribution Restrictions and the Form of Aggregate Functions," Econometrica, forthcoming.

White, H. [1980a], "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," Econometrica, 48: 817-838.