

University of Connecticut

Department of Economics Working Paper Series

GMM Based Inference with Standard Stratified Samples when the Aggregate Shares are Known

Gautam Tripathi
University of Connecticut

Working Paper 2008-31R

September 2008, revised November 2009

341 Mansfield Road, Unit 1063
Storrs, CT 06269–1063
Phone: (860) 486–3022
Fax: (860) 486–4463
http://www.econ.uconn.edu/

This working paper is indexed on RePEc, http://repec.org/

Abstract

We show how to do efficient moment based inference using the generalized method of moments (GMM) when data is collected by standard stratified sampling and the maintained assumption is that the aggregate shares are known.

Journal of Economic Literature Classification: C30

Keywords: Generalized method of moments, GMM, standard stratified sampling.

1. Introduction

Let Z^* be a $d \times 1$ random vector that denotes an observation from the population of interest (henceforth called the "target" population) and suppose there exists a parameter $\theta^* \in \Theta \subset \mathbb{R}^p$ satisfying the moment condition

$$\mathbb{E}_{P^*}[g(Z^*, \theta^*)] = 0. \tag{1.1}$$

The moment function g is a $q \times 1$ vector of functions known up to θ^* such that $q \geq p$, i.e., overidentification is allowed, and P^* is the unknown probability distribution of Z^* (note that Z^* can have discrete components). The notation $\mathbb{E}_{P^*}$ indicates that expectation is with respect to P^*. Cf. Section 3.4 for some illustrative examples.

If data is collected by random sampling, so that observations from the target population have the same chance of being represented in the realized sample, then it is well known how to efficiently estimate θ^* using the generalized method of moments (GMM); cf. Newey and McFadden (1994). However, as with many large datasets, if data is collected by stratified sampling so that units from the target population have unequal chances of being selected, then the realized sample consists of observations drawn from the distribution induced by the sampling scheme rather than the target distribution P^* — and in general the two distributions are not the same.

Therefore, since the parameter of interest θ^* is a function of P^* (cf. (1.1)) and not the distribution induced by the sampling scheme, statistical procedures that do not account for the consequences of stratification are not guaranteed to produce reliable inference about θ^*. For instance, letting $Z_1, \ldots, Z_n$ denote the stratified sample, the sample average $\sum_{j=1}^n Z_j/n$ will in general not be a consistent estimator of $\theta^* := \mathbb{E}_{P^*}[Z^*]$, the mean of the target population, because $\operatorname{plim}_{n\to\infty} n^{-1} \sum_{j=1}^n Z_j = \mathbb{E}_{P_{\text{obs}}}[Z]$ by a weak law of large numbers, where P_{obs} denotes the distribution induced by the sampling scheme, but $\mathbb{E}_{P_{\text{obs}}}[Z] \neq \mathbb{E}_{P^*}[Z^*]$ because $P_{\text{obs}} \neq P^*$.

Standard stratified (SS) sampling, discussed in the next section, is a scheme where a predetermined number of observations are randomly sampled from each stratum of the target population. Wooldridge (2001) examines asymptotic properties of M-estimators when data is collected by SS sampling and his approach can be used to handle models where parametric assumptions are made about the conditional density of variables in the target population, linear and nonlinear regression models, etc., that have been studied previously in the literature on SS sampling (cf., e.g., Manski and Lerman (1977), Cosslett (1981a, 1981b, 1993), Hausman and Wise (1981), Manski and McFadden (1981), Jewell (1985), Quesenberry and Jewell (1986), Imbens (1992), Imbens and Lancaster (1996), and the references therein).

However, the parameters of interest in Wooldridge's model are exactly identified whereas we allow θ^* to be overidentified. Therefore, (1.1) nests his setup as a special case. [Recall that the moment conditions that identify parameters in M-estimation problems are the first order

conditions necessary for maximizing or minimizing an objective function at the population level; e.g., for conditional maximum likelihood the moment condition is that the mean of the conditional score in the target population is zero.] Note that since the moment conditions in Wooldridge's paper are exactly identified, their validity cannot be tested, at least unless additional moment conditions are added. In contrast, we also investigate specification testing under stratification. Finally, unlike Wooldridge who does not address efficiency issues in his paper, we obtain the efficiency bound for estimating θ^* (a nonstandard problem because observations collected by standard stratified sampling are independently but not identically distributed) and propose an estimator of θ^* that is asymptotically efficient, i.e., its variance matches the efficiency bound as the sample size becomes arbitrarily large. An additional benefit of our efficiency bound result is that it can be used to show that the M-estimators in Wooldridge (2001) are asymptotically efficient within their class.

The treatment in this paper is general enough to allow for different sources of stratification. For instance, in models where Z^* can be decomposed into endogenous and exogenous components, the approach taken here can handle stratification based only on the endogenous variables, or on the exogenous variables alone, or stratification that is based on a subset of these variables, in a straightforward manner (cf. Example 3.2). The stratifying variable(s) can be discrete or continuously distributed (or both). We have also taken special care to derive intuitive closed form expressions for the asymptotic variances of estimators proposed here so that standard errors are easily obtained.

In a recent paper, Tripathi (2007) treats the aggregate shares (defined subsequently) as unknown parameters but requires an additional random sample to deal with their consequent lack of identification whereas in the present paper — as well as in Wooldridge (2001) — the aggregate shares are known, an assumption that is justifiable for many datasets (cf. Section 2). The proof of the efficiency bound presented here is also different than the proofs of the efficiency bounds in Tripathi (2007). In short, to the best of our knowledge, the results obtained here are not to be found elsewhere in the literature.

The remainder of the paper is organized as follows. In Section 2 we describe standard stratified sampling and the statistical consequences of collecting data by such a sampling scheme. Estimators of θ^*, their asymptotic properties, inference, and some useful examples that illustrate the insights obtained in this paper are discussed in Section 3. Section 4 concludes. All proofs are in the appendices.

2. Standard stratified sampling

Let the support of Z^* be partitioned into L nonempty disjoint strata $\mathbb{C}_1, \ldots, \mathbb{C}_L$. In SS sampling, used to collect most large datasets, the number of observations drawn from each stratum is fixed in advance and data is sampled randomly within each stratum. In

particular, suppose that n observations $Z_1, \ldots, Z_n$ are collected by SS sampling with $n_l := \sum_{j=1}^{n} \mathbb{1}(Z_j \in \mathbb{C}_l)$ defined to be the (predetermined) number of observations drawn from the lth stratum, $l = 1, \ldots, L$, so that the "sampling fractions" (namely, the n_l/n's) sum to one, i.e., $(n_1/n) + \ldots + (n_L/n) = 1$. The distribution induced by the SS sampling scheme, denoted by P_n, is then given by

$$P_n(Z \in B) := \sum_{l=1}^{L} \frac{(n_l/n)}{Q_l^*} \int_B \mathbb{1}(z \in \mathbb{C}_l)\, dP^*(z), \tag{2.1}$$

where B is any Borel subset of $\mathbb{R}^d$ and $Q_l^* := P^*(Z^* \in \mathbb{C}_l)$ is the probability that a randomly chosen observation from the target population lies in the lth stratum. [For the sake of completeness, a short proof of (2.1) is provided in Appendix A.] The Q_l^*'s are popularly called "aggregate shares" because $Q_1^* + \ldots + Q_L^* = 1$. Notice that (2.1) implies that the density of P_n with respect to any Borel measure on $\mathbb{R}^d$ that dominates P^* is given by

$$dP_n(z) := \sum_{l=1}^{L} \frac{(n_l/n)}{Q_l^*} \mathbb{1}(z \in \mathbb{C}_l) dP^*(z), \qquad z \in \mathbb{R}^d.$$

As noted by Wooldridge (2001, p. 453), the aggregate shares $Q^* := (Q_1^*, \ldots, Q_L^*)_{L\times 1}$ being unconditional probabilities can often be estimated easily and extremely precisely from large surveys such as the census. So it is not very surprising that researchers working with stratified datasets often disregard the estimation uncertainty that comes from estimating the aggregate shares and simply assume that they are known. Therefore, as in Wooldridge (2001), we also maintain the assumption that the Q_l^*'s are known. [By contrast, severe identification problems arise if the aggregate shares are unknown; cf. Tripathi (2007) for a discussion regarding these problems and their resolution.]

Observations collected by SS sampling are independently but not identically distributed (inid) because the n_l's are treated as nonstochastic constants. This complicates the problem of calculating the efficiency bounds which are much easier to obtain in an iid setting. Fortunately, this technical hurdle can be bypassed with the following trick: First, letting $K^0 := (K_1^0, \ldots, K_L^0)$ denote an $L \times 1$ vector of unknown parameters in $(0,1)^L$ such that $\sum_{l=1}^{L} K_l^0 = 1$, assume (counterfactually) that observations in the stratified sample are iid draws from the density

$$dP(z) := \sum_{l=1}^{L} \frac{K_l^0}{Q_l^*} \mathbb{1}(z \in \mathbb{C}_l) dP^*(z), \qquad z \in \mathbb{R}^d. \tag{2.2}$$

Next, estimate K^0 jointly and efficiently with θ^*. We show in Section 3 that estimating K^0 — which can be thought of as the vector of "limiting" sampling fractions — jointly and efficiently with θ^* leads to asymptotic inference that is conditional on the observed values of the n_l's. In other words, treating the sampling fractions as unknown parameters to be estimated (even

though they are known!) has the effect of asymptotically conditioning on the number of observations lying in each stratum of the stratified sample. Therefore, our asymptotic results are valid under the inid framework though we derive them in an artificially created iid environment.

3. Efficient estimation

3.1. **Motivation.** Our estimator is easy to motivate. Let $K^0_{-L} := (K^0_1, \ldots, K^0_{L-1})_{(L-1)\times 1}$ and $K^0_L := 1 - \sum_{l=1}^{L-1} K^0_l$. Then, by (2.2),

$$dP^*(z) = \sum_{l=1}^{L} \frac{Q^*_l}{K^0_l} \mathbb{1}(z \in \mathbb{C}_l) dP(z) = dP(z)/b_{Q^*}(z, K^0_{-L}), \tag{3.1}$$

where

$$b_{Q^*}(z, K^0_{-L}) := \sum_{l=1}^{L} \frac{K^0_l}{Q^*_l} \mathbb{1}(z \in \mathbb{C}_l).$$

Hence, (3.1) implies that

$$\mathbb{E}_{P^*}[g(Z^*, \theta^*)] = \mathbb{E}_P[g(Z, \theta^*)/b_{Q^*}(Z, K^0_{-L})]. \tag{3.2}$$

Moreover, since $\mathbb{E}_P[\mathbb{1}(Z \in \mathbb{C}_l)] = K^0_l$ by (2.2), it follows that $K^0_1, \ldots, K^0_{L-1}$ are exactly identified by the $L-1$ moment conditions

$$\mathbb{E}_P[s(Z) - K^0_{-L}] = 0, \tag{3.3}$$

where $s(Z) := (\mathbb{1}(Z \in \mathbb{C}_1), \ldots, \mathbb{1}(Z \in \mathbb{C}_{L-1}))_{(L-1)\times 1}$. Therefore, by (3.2) and (3.3),

$$(1.1) \iff \mathbb{E}_P \begin{bmatrix} g(Z, \theta^*)/b_{Q^*}(Z, K^0_{-L}) \\ s(Z) - K^0_{-L} \end{bmatrix} = 0.$$

Hence, $\beta_0 := (\theta^*, K^0_{-L})_{(p+L-1)\times 1}$ can be efficiently estimated by doing optimal GMM on the $(q + L - 1) \times 1$ transformed moment function

$$\rho(Z, \beta) := \begin{bmatrix} g(Z, \theta)/b_{Q^*}(Z, K_{-L}) \\ s(Z) - K_{-L} \end{bmatrix} := \begin{bmatrix} \rho_1(Z, \beta) \\ \rho_2(Z, K_{-L}) \end{bmatrix}, \tag{3.4}$$

where $\rho_1(Z, \beta) := g(Z, \theta)/b_{Q^*}(Z, K_{-L})$ and $\rho_2(Z, K_{-L}) := s(Z) - K_{-L}$.

The two-step optimal GMM estimator of β_0, denoted by $\tilde{\beta} := (\tilde{\theta}, \tilde{K}_{-L})_{(p+L-1)\times 1}$, is given by

$$\tilde{\beta} := \underset{\beta \in \mathcal{B}}{\operatorname{argmin}}\, \hat{\rho}'(\beta) \hat{V}^{-1}_{\rho}(\check{\beta}) \hat{\rho}(\beta),$$

where $\mathcal{B} := \Theta \times [0,1]^{L-1}$, $\hat{\rho}(\beta) := \sum_{j=1}^n \rho(Z_j, \beta)/n$, and $\hat{V}_{\rho}(\check{\beta}) := \sum_{j=1}^n \rho(Z_j, \check{\beta}) \rho'(Z_j, \check{\beta})/n$ estimates $\mathbb{E}_P[\rho(Z, \beta_0)\rho'(Z, \beta_0)]$ with a preliminary estimator $\check{\beta} := \operatorname{argmin}_{\beta \in \mathcal{B}} \hat{\rho}'(\beta)\hat{\rho}(\beta)$.

3.2. **Asymptotic normality and efficiency.** Let $\|\cdot\|$ denote the Euclidean norm. The following standard regularity conditions ensure that GMM estimators are consistent and asymptotically normal.

Assumption 3.1. (i) $\beta_0 \in \mathcal{B}$ *is the unique solution to* $\mathbb{E}_P[\rho(Z,\beta)] = 0$*;* (ii) $\mathcal{B}$ *is compact;* (iii) $\rho(Z,\beta)$ *is continuous at each* $\beta \in \mathcal{B}$ *w.p.1;* (iv) $\mathbb{E}_P[\sup_{\beta\in\mathcal{B}} \|\rho(Z,\beta)\|^2] < \infty$*;* (v) *The matrix* $\mathbb{E}_P[\rho(Z,\beta_0)\rho'(Z,\beta_0)]$ *is nonsingular;* (vi) $\beta_0 \in \text{int}(\mathcal{B})$*;* (vii) $\rho(Z,\beta)$ *is continuously differentiable in a neighborhood* $\mathcal{N}$ *of* β_0 *and* $\mathbb{E}_P[\sup_{\beta\in\mathcal{N}} \|\partial\rho(Z,\beta)/\partial\beta\|] < \infty$*;* (viii) *The matrix* $\mathbb{E}_P[\partial\rho(Z,\beta_0)/\partial\beta]$ *is of full column rank.*

(i)–(v) can be used to prove consistency and (vi)–(viii) to prove the asymptotic normality of GMM estimators as in Newey and McFadden (1994).

Let ε denote the residual that results when $\rho_1(Z,\beta_0)$ is orthogonally projected onto the space spanned by the coordinates of $\rho_2(Z,K^0_{-L})$; i.e.,

$$\varepsilon := \rho_1(Z,\beta_0) - \Sigma_{12}V_2^{-1}\rho_2(Z,K^0_{-L}),$$

where $\Sigma_{12} = \mathbb{E}_P[\rho_1(Z,\beta_0)\rho_2'(Z,K^0_{-L})]$ and $V_2 := \mathbb{E}_P[\rho_2(Z,K^0_{-L})\rho_2'(Z,K^0_{-L})]$. Then, letting $0_{k_1\times k_2}$ denote the $k_1 \times k_2$ matrix of zeros, we can show the following result.

Theorem 3.1. *Let Assumption* 3.1 *hold. Then, as the size of the stratified sample* $n \to \infty$*,*

$$\begin{bmatrix} n^{1/2}(\tilde\theta - \theta^*) \\ n^{1/2}(\tilde K_{-L} - K^0_{-L}) \end{bmatrix} \xrightarrow{d} \text{N}(0_{(p+L-1)\times 1}, \begin{bmatrix} (D'\Omega^{-1}D)^{-1} & 0_{p\times(L-1)} \\ 0'_{p\times(L-1)} & V_2 \end{bmatrix}),$$

where $D := \mathbb{E}_P[\partial\rho_1(Z,\beta_0)/\partial\theta]$ *and* $\Omega := \mathbb{E}_P[\varepsilon\varepsilon']$*.*

From the proof of Theorem 3.1 it is clear that $\tilde\theta$ is asymptotically linear with influence function $-(D'\Omega^{-1}D)^{-1}D'\Omega^{-1}\varepsilon$. But since ε is orthogonal to $\rho_2(Z,K^0_{-L})$ by definition, the central limit theorem reveals that $\tilde\theta$ is asymptotically independent of $\sum_{j=1}^n s(Z_j) = (n_1,\ldots,n_{L-1})_{(L-1)\times 1}$. Therefore, as emphasized in Section 2, inference using the asymptotic distribution of $\tilde\theta$ is equivalent to inference based on the asymptotic distribution of $\tilde\theta$ conditional on the number of observations lying in each stratum of the stratified sample.

Let $V_1 := \mathbb{E}_P[\rho_1(Z,\beta_0)\rho_1'(Z,\beta_0)]$. From the definition of ε, it is immediate that

$$\Omega = V_1 - \Sigma_{12}V_2^{-1}\Sigma_{12}'.$$

Since D and Ω can be estimated by replacing population expectations with their sample analogs, standard errors of $\tilde\theta$ are straightforward to obtain.

The next result shows that the asymptotic variances of $\tilde\theta$ and $\tilde K_{-L}$ coincide with the efficiency bounds for estimating θ^* and K^0_{-L}, respectively. Therefore, $\tilde\theta$ and $\tilde K_{-L}$ are asymptotically efficient.

Theorem 3.2. *Let Assumption* 3.1 *hold. Then, the efficiency bound for estimating* θ^* *and* K^0_{-L} *are given by* $(D'\Omega^{-1}D)^{-1}$ *and* V_2*, respectively.*

Theorem 3.2 also implies that the M-estimators in Wooldridge (2001) are asymptotically efficient within their class. To see this, suppose that θ^* is just identified, i.e., $q = p$. Then, $\tilde{\theta}$ and $\tilde{K}_{-L}$ are obtained by setting the sample analog of $\mathbb{E}_P[\rho(Z, \beta_0)]$ to zero; i.e., $\tilde{\theta}$ solves $\sum_{j=1}^n g(Z_j, \tilde{\theta})/b_{Q^*}(Z_j, \tilde{K}_{-L}) = 0$, where $\tilde{K}_{-L} = (n_1/n, \ldots, n_{L-1}/n)$. Hence, by Theorem 3.1, the asymptotic variance of $n^{1/2}(\tilde{\theta} - \theta^*)$ reduces to $D^{-1}\Omega D'^{-1}$. But, as shown in Appendix A,

$$\begin{aligned} D &= \sum_{l=1}^{L} Q^*_l \mathbb{E}_{P^*}[\frac{\partial g(Z^*, \theta^*)}{\partial \theta} | Z^* \in \mathbb{C}_l] = \mathbb{E}_{P^*}[\frac{\partial g(Z^*, \theta^*)}{\partial \theta}] \\ \Omega &= \sum_{l=1}^{L} \frac{Q^{*2}_l}{K^0_l} \operatorname{Var}_{P^*}[g(Z^*, \theta^*) | Z^* \in \mathbb{C}_l]. \end{aligned} \tag{3.5}$$

Therefore, notational differences aside, comparing (3.5) above with equations (3.2), (3.7), and (3.8) in Wooldridge (2001) reveals that $D^{-1}\Omega D'^{-1}$ matches the asymptotic variance in Theorem 3.2 of Wooldridge's paper. Hence, the M-estimators proposed there are asymptotically efficient.

Remarks. (i) The known aggregate shares satisfy the moment condition

$$\mathbb{E}_P[(s(Z) - Q^*_{-L})/b_{Q^*}(Z, K^0_{-L})] = 0,$$

where $Q^*_{-L} := (Q^*_1, \ldots, Q^*_{L-1})_{(L-1)\times 1}$. Similarly, because (3.1) defines a density,

$$\mathbb{E}_P[1/b_{Q^*}(Z, K^0_{-L}) - 1] = 0.$$

However, since $\rho_3(Z, K^0_{-L}) := (s(Z) - Q^*_{-L})/b_{Q^*}(Z, K^0_{-L})$ and $\rho_4(Z, K^0_{-L}) := 1/b_{Q^*}(Z, K^0_{-L}) - 1$ are linear transformations of $\rho_2(Z, K^0_{-L})$, cf. (A.5) in the appendix, these moment conditions are automatically satisfied by (3.4).

(ii) Notice that $\mathcal{B}$ is compact if and only if Θ is compact. Furthermore, Assumption 3.1(viii) holds if and only if D or equivalently, by (3.5), $\mathbb{E}_{P^*}[\partial g(Z^*, \theta^*)/\partial\theta]$ is full rank (this follows because $K^0_1, \ldots, K^0_{L-1}$ are just identified). The latter is of course a well known sufficient condition for θ^* to be locally identified (Newey and McFadden, 1994, p. 2127). □

3.3. **A computational simplification.** As mentioned earlier, if θ^* is just identified then no optimization is necessary to obtain $\tilde{\beta}$ because then $\tilde{K}_{-L} = (n_1/n, \ldots, n_{L-1}/n)$, $\tilde{K}_L = n_L/n$, and $\tilde{\theta}$ solves $\sum_{j=1}^n \rho_1(Z_j, \tilde{\theta}, \tilde{K}_{-L}) = 0$. By contrast, if θ^* is overidentified and enters the moment function nonlinearly, then implementing $\tilde{\beta}$ requires searching over a subset of $\mathbb{R}^{p+L-1}$. However, taking advantage of the fact that $(n_1/n, \ldots, n_L/n)$ is an asymptotically efficient estimator of K^0, it is possible to construct a GMM estimator of θ^* such that the dimensionality of the optimization problem is reduced to $\mathbb{R}^p$ without compromising its asymptotic efficiency.

To see this, let $\hat{K}_{-L} := (n_1/n, \ldots, n_{L-1}/n)_{(L-1)\times 1}$, $\hat{K} := (\hat{K}_{-L}, n_L/n)_{L\times 1}$, and

$$\hat{\theta} := \underset{\theta\in\Theta}{\operatorname{argmin}}\, \hat{\rho}_1'(\theta, \hat{K}_{-L})\hat{\Omega}^{-1}(\check{\theta}, \hat{K}_{-L})\hat{\rho}_1(\theta, \hat{K}_{-L}), \tag{3.6}$$

where $\hat{\Omega}(\theta, K_{-L}) := \hat{V}_1(\theta, K_{-L}) - \hat{\Sigma}_{12}(\theta, K_{-L})\hat{V}_2^{-1}(K_{-L})\hat{\Sigma}_{12}'(\theta, K_{-L})$,

$$\hat{V}_1(\theta, K_{-L}) := n^{-1}\sum_{j=1}^{n} \rho_1(Z_j, \theta, K_{-L})\rho_1'(Z_j, \theta, K_{-L}),$$

$$\hat{\Sigma}_{12}(\theta, K_{-L}) := n^{-1}\sum_{j=1}^{n} \rho_1(Z_j, \theta, K_{-L})\rho_2'(Z_j, K_{-L}),$$

$$\hat{V}_2(K_{-L}) := n^{-1}\sum_{j=1}^{n} \rho_2(Z_j, K_{-L})\rho_2'(Z_j, K_{-L}),$$

and $\check{\theta} := \operatorname{argmin}_{\theta\in\Theta} \hat{\rho}_1'(\theta, \hat{K}_{-L})\hat{\rho}_1(\theta, \hat{K}_{-L})$ is a preliminary estimator of θ^*.

Since $\hat{K}$ estimates a nuisance parameter, it makes sense to think of $\hat{\theta}$ as a "plug-in" GMM estimator of θ^*. Note that if θ^* is just identified, then $\hat{\theta} = \tilde{\theta}$; but they will be different in finite samples whenever θ^* is overidentified. However, the following result shows that $\hat{\theta}$ and $\tilde{\theta}$ are always asymptotically equivalent.

Lemma 3.1. $n^{1/2}(\hat{\theta} - \tilde{\theta}) = o_P(1)$ *under Assumption* 3.1.

Lemma 3.1 implies that $\hat{\theta}$ is also asymptotically efficient. Therefore, since it is computationally less expensive than $\tilde{\theta}$, for the remainder of Section 3 we focus on $\hat{\theta}$.

3.4. **Examples.** In this section we look at some illustrative examples. Henceforth, let I_k be the $k \times k$ identity matrix and $\hat{D}(\theta, K_{-L}) := \partial\hat{\rho}_1(\theta, K_{-L})/\partial\theta$. The support of a random vector A is denoted by $\operatorname{supp}(A)$.

Example 3.1 (Estimating the population mean)**.** Let θ^* denote the mean of the target population, i.e., $\mathbb{E}_{P^*}[Z^* - \theta^*] = 0 \implies g(Z^*, \theta^*) := Z^* - \theta^*$. Therefore, since θ^* is just identified,

$$\hat{\theta} = \frac{\sum_{j=1}^n Z_j b^{-1}(Z_j, \hat{K}_{-L})}{\sum_{j=1}^n b^{-1}(Z_j, \hat{K}_{-L})} = \sum_{l=1}^{L} Q_l^* \bar{Z}_l,$$

where $\bar{Z}_l := \sum_{j=1}^n Z_j \mathbb{1}(Z_j \in \mathbb{C}_l)/n_l$ is the lth stratum sample average and the second equality follows because $\sum_{j=1}^n b^{-1}(Z_j, \hat{K}_{-L}) = n$ and $\sum_{j=1}^n Z_j b^{-1}(Z_j, \hat{K}_{-L}) = n\sum_{l=1}^L Q_l^* \bar{Z}_l$. The estimated asymptotic variance of $\hat{\theta}$ in this example is given by $\widehat{\operatorname{asvar}}(\hat{\theta}) = n^{-1}\hat{\Omega}(\hat{\theta}, \hat{K}_{-L})$ due to the fact that here $\hat{D}(\hat{\theta}, \hat{K}_{-L}) = -n^{-1}\sum_{j=1}^n b^{-1}(Z_j, \hat{K}_{-L})I_p = -I_p$. □

Example 3.2 (Linear instrumental variables (IV))**.** Suppose that $Y^* = X^{*\prime}\theta^* + u^*$ and some of the regressors are endogenous. Assume there exists a $q \times 1$ vector of instrumental variables

W^* satisfying $\mathbb{E}_{P^*}[u^*|W^*] = 0$ w.p.1. This leads to an IV model of the form

$$\mathbb{E}_{P^*}[W^*(Y^* - X^{*\prime}\theta^*)] = 0 \implies g(Z^*, \theta^*) := W^*(Y^* - X^{*\prime}\theta^*),$$

where $Z^* := (Y^*, X^*, W^*)_{(1+p+q)\times 1}$. Because g is linear in parameters, the first order condition for (3.6) has a closed form solution. Thus no optimization is necessary to obtain $\hat{\theta}$ even when θ^* is overidentified and it is easy to verify that $\hat{\theta}$ takes the familiar form of an IV estimator with a correction for stratification, i.e.,

$$\hat{\theta} = ((\sum_{j=1}^{n} \frac{X_j W_j'}{b_{Q^*}(Z_j, \hat{K}_{-L})})\hat{\Omega}^{-1}(\check{\theta}, \hat{K}_{-L})(\sum_{j=1}^{n} \frac{W_j X_j'}{b_{Q^*}(Z_j, \hat{K}_{-L})}))^{-1} \times (\sum_{j=1}^{n} \frac{X_j W_j'}{b_{Q^*}(Z_j, \hat{K}_{-L})})\hat{\Omega}^{-1}(\check{\theta}, \hat{K}_{-L})(\sum_{j=1}^{n} \frac{W_j Y_j}{b_{Q^*}(Z_j, \hat{K}_{-L})}). \quad (3.7)$$

Since here $\hat{D}(\hat{\theta}, \hat{K}_{-L}) = -n^{-1}\sum_{j=1}^{n} W_j X_j' / b_{Q^*}(Z_j, \hat{K}_{-L})$,

$$\widehat{\text{asvar}}(\hat{\theta}) = n^{-1}(\hat{D}'(\hat{\theta}, \hat{K}_{-L})\hat{\Omega}^{-1}(\hat{\theta}, \hat{K}_{-L})\hat{D}(\hat{\theta}, \hat{K}_{-L}))^{-1}.$$

Notice that (3.7) implicitly assumes that Y^*, X^*, W^* were all collected by SS sampling. But if stratification is based only on Y^* (models where stratification is based only on the response variable are often said to be "endogenously" stratified), then $\hat{\theta}$ can be obtained by simply letting $\mathbb{C}_l := \mathbb{C}_l^{Y^*} \times \text{supp}(X^*) \times \text{supp}(W^*)$, where $\mathbb{C}_l^{Y^*}$ denotes the lth stratum of the support of Y^*; i.e., $\hat{\theta}$ can be obtained by replacing the $b_{Q^*}(Z_j, \hat{K}_{-L})$ in (3.7) with $b_{Q^*}(Z_j, \hat{K}_{-L}) := \sum_{l=1}^{L}(n_l/n)\mathbb{1}(Y_j \in \mathbb{C}_l^{Y^*})/Q_l^*$. Similarly, for stratification based only on (Y^*, X^*) use $b_{Q^*}(Z_j, \hat{K}_{-L}) := \sum_{l=1}^{L}(n_l/n)\mathbb{1}((Y_j, X_j) \in \mathbb{C}_l^{Y^* \times X^*})/Q_l^*$ to construct $\hat{\theta}$ with $\mathbb{C}_l^{Y^* \times X^*}$ now the lth stratum of the support of $Y^* \times X^*$. Modifications to b_{Q^*} needed to account for other sources of stratification follow mutatis mutandis. □

Example 3.3 (Box-Cox type transformation model)**.** Let $h_1(Y^*, \theta_1^*) = h_2(X^*, \theta_2^*) + u^*$, where h_1 and h_2 are real-valued functions known up to the θ^*'s and $\mathbb{E}_{P^*}[u^*|X^*] = 0$ w.p.1. Since least squares will typically not consistently estimate $\theta^* := (\theta_1^*, \theta_2^*)$, we propose an IV estimator instead. So, letting $A(X^*)$ denote a vector of instruments that just identify or overidentify θ^*, we have an IV model of the form

$$\mathbb{E}_{p^*}[A(X^*)(h_1(Y^*, \theta_1^*) - h_2(X^*, \theta_2^*))] = 0 \implies g(Z^*, \theta^*) := A(X^*)(h_1(Y^*, \theta_1^*) - h_2(X^*, \theta_2^*)),$$

where $Z^* := (Y^*, X^*)$. If h_1 or h_2 are nonlinear in parameters then, unlike the previous examples, $\hat{\theta}$ is not available in closed-form but has to be computed numerically as described in (3.6). As in the previous example, depending upon what variables are used to stratify the target population, b_{Q^*} has to be defined appropriately when implementing $\hat{\theta}$ and computing its standard errors. □

Before ending this section, we look at a special case of SS sampling that is often encountered in applied work.

Example 3.4 (Proportional allocation)**.** This refers to the case when the predetermined sampling fractions are chosen to be equal to the known aggregate shares, i.e., $\hat{K} = Q^*$ for each n. The plug-in GMM estimator of θ^* under proportional allocation is $\hat{\theta}_{\text{PA}} := \hat{\theta}|_{Q^*=\hat{K}}$, i.e., simply replace the b_{Q^*} in (3.6) by $b_{\hat{K}}$. For instance, since $b_{\hat{K}}(Z, \hat{K}_{-L}) = 1$, it is easy to see that the $\hat{\theta}_{\text{PA}}$'s for Examples 3.1 and 3.2 are, respectively, $\sum_{j=1}^n Z_j/n$ and

$$((\sum_{j=1}^n X_j W_j')\hat{\Omega}^{-1}(\check{\theta}, \hat{K}_{-L})(\sum_{j=1}^n W_j X_j'))^{-1}(\sum_{j=1}^n X_j W_j')\hat{\Omega}^{-1}(\check{\theta}, \hat{K}_{-L})(\sum_{j=1}^n W_j Y_j).$$

$n^{1/2}(\hat{\theta}_{\text{PA}} - \theta^*)$ is asymptotically normal with mean zero and variance $(D_{\text{PA}}'\Omega_{\text{PA}}^{-1}D_{\text{PA}})^{-1}$, where $D_{\text{PA}} := D|_{K^0=Q^*} = \mathbb{E}_P[\partial g(Z, \theta^*)/\partial\theta]$ and $\Omega_{\text{PA}} := \Omega|_{K^0=Q^*}$; in particular, by (3.5),

$$D_{\text{PA}} = \sum_{l=1}^L Q_l^* \mathbb{E}_{P^*}[\frac{\partial g(Z^*, \theta^*)}{\partial\theta}|Z^* \in \mathbb{C}_l] \quad \& \quad \Omega_{\text{PA}} = \sum_{l=1}^L Q_l^* \operatorname{Var}_{P^*}[g(Z^*, \theta^*)|Z^* \in \mathbb{C}_l].$$

Standard errors of $\hat{\theta}_{\text{PA}}$ are easily obtained because D_{PA} and Ω_{PA} can be consistently estimated by $\hat{D}(\hat{\theta}_{\text{PA}}, \hat{K}_{-L})|_{Q^*=\hat{K}}$ and $\hat{\Omega}(\hat{\theta}_{\text{PA}}, \hat{K}_{-L})|_{Q^*=\hat{K}}$, respectively. Note that since $D_{\text{PA}} = \mathbb{E}_{P^*}[\partial g(Z, \theta^*)/\partial\theta]$ and $\Omega_{\text{PA}} - \mathbb{E}_{P^*}[g(Z^*, \theta^*)g'(Z^*, \theta^*)]$ is negative definite (cf. Lemma B.4), proportional allocation leads to a more efficient GMM estimator than random sampling. This result, well known in the context of estimating population means (Cochran, 1963, Theorem 5.8, p. 98), is often cited as the raison d´être for proportional allocation. □

3.5. **Inference.** Finally, a brief comment regarding hypothesis and specification tests. Suppose we want to test the parametric restriction $H(\theta^*) = 0$ against the alternative that it is false, where H is a $h \times 1$ vector of twice continuously differentiable functions such that $\partial H(\theta^*)/\partial\theta$ has rank $h \leq p$. As described in Newey and McFadden (1994, Theorem 9.2), a variety of statistics based on $\hat{\theta}$ can be used to test this hypothesis. In each case, the test statistic is asymptotically χ_h^2 under the null. Confidence regions can be obtained by inverting these test statistics.

Next, assume that $q > p$. Since inference based on the estimated θ^* is sensible only if (1.1) is true, it is important to test it against the alternative that it is false. It is straightforward to show that $\hat{\mathcal{J}} := n\hat{\rho}_1'(\hat{\theta}, \hat{K}_{-L})\hat{\Omega}^{-1}(\hat{\theta}, \hat{K}_{-L})\hat{\rho}_1(\hat{\theta}, \hat{K}_{-L})$, the J-statistic corresponding to the plug-in GMM estimator $\hat{\theta}$, is asymptotically χ_{q-p}^2 under the null hypothesis that (1.1) is true. Therefore, rejecting (1.1) whenever $\hat{\mathcal{J}} \geq Q_{\chi_{q-p}^2}(1-\alpha)$ yields an asymptotic size-α specification test for (1.1), where $Q_{\chi_{q-p}^2}(t)$ denotes the tth quantile of a χ_{q-p}^2 random variable.

4. Conclusion

We have shown how to do efficient GMM based inference when data is collected by standard stratified sampling and the aggregate shares are assumed to be known.

Acknowledgements

We thank the associate editor, an anonymous referee, Paul Devereux, Guido Imbens, and Yuichi Kitamura for helpful comments and suggestions.

Appendix A. Proofs

Additional notation used throughout the proofs: $\mathcal{K} := \text{diag}(K_1^0, \ldots, K_{L-1}^0)$ and $\mathcal{Q} := \text{diag}(Q_1^*, \ldots, Q_{L-1}^*)$ are $(L-1) \times (L-1)$ diagonal matrices, $L_2(Z, P)$ is the set of real-valued functions of Z that are square-integrable with respect to P, the operator $\mathscr{P}_A$ denotes orthogonal projection onto $A \subset L_2(Z, P)$ using the inner product $\langle a, b \rangle_P := \mathbb{E}_P[ab]$, the induced P-norm is $\|\cdot\|_{2,P} := \langle \cdot, \cdot \rangle_P^{1/2}$, the range and null space of D are $\mathcal{R}(D)$ and $\mathcal{N}(D)$, respectively, and $\tilde{1}$ denotes the $(L-1) \times 1$ vector of ones.

Proof of (2.1). Let Z denote an observation collected by SS sampling. Then, by the definition of SS sampling, $\text{Law}(Z|Z \in \mathbb{C}_l) = \text{Law}(Z^*|Z^* \in \mathbb{C}_l)$ for $l = 1, \ldots, L$. But,

$$\text{Prob}(Z \in B | Z \in \mathbb{C}_l) = \frac{\text{Prob}(Z \in B \cap \mathbb{C}_l)}{n_l/n}$$

and

$$\text{Prob}(Z^* \in B | Z^* \in \mathbb{C}_l) = \frac{P^*(Z^* \in B \cap \mathbb{C}_l)}{Q_l^*}.$$

Therefore, for $l = 1, \ldots, L$,

$$\frac{\text{Prob}(Z \in B \cap \mathbb{C}_l)}{n_l/n} = \frac{P^*(Z^* \in B \cap \mathbb{C}_l)}{Q_l^*},$$

implying that

$$\text{Prob}(Z \in B) = \sum_{l=1}^{L} \frac{(n_l/n)}{Q_l^*} P^*(Z^* \in B \cap \mathbb{C}_l).$$

The desired result follows since $P^*(Z^* \in B \cap \mathbb{C}_l) = \int_B \mathbb{1}(z \in \mathbb{C}_l)\, dP^*(z)$. □

Proof of Theorem 3.1. By standard GMM theory, $\tilde{\beta}$ is consistent and $n^{1/2}(\tilde{\beta} - \beta_0)$ is asymptotically normal with mean zero and variance $(D_\rho' V_\rho^{-1} D_\rho)^{-1}$, where

$$D_\rho := \mathbb{E}_P[\partial \rho(Z, \beta_0)/\partial \beta] \quad \& \quad V_\rho := \mathbb{E}_P[\rho(Z, \beta_0)\rho'(Z, \beta_0)] = \begin{bmatrix} V_1 & \Sigma_{12} \\ \Sigma_{12}' & V_2 \end{bmatrix}.$$

Now, with $D_2 := \mathbb{E}_P[\partial\rho_1(Z,\beta_0)/\partial K_{-L}]$,

$$D_\rho = \begin{bmatrix} D & D_2 \\ 0_{(L-1)\times p} & -I_{L-1} \end{bmatrix}.$$

Also, by the partitioned inverse formula,

$$V_\rho^{-1} = \begin{bmatrix} \Omega^{-1} & -\Omega^{-1}\Sigma_{12}V_2^{-1} \\ -V_2^{-1}\Sigma_{12}'\Omega^{-1} & V_2^{-1} + V_2^{-1}\Sigma_{12}'\Omega^{-1}\Sigma_{12}V_2^{-1} \end{bmatrix}.$$

Hence, letting $J := D_2 + \Sigma_{12}V_2^{-1}$, some straightforward algebra reveals that

$$D_\rho' V_\rho^{-1} D_\rho = \begin{bmatrix} D'\Omega^{-1}D & D'\Omega^{-1}J \\ J'\Omega^{-1}D & V_2^{-1} + J\Omega^{-1}J \end{bmatrix} \overset{\text{Lemma B.1}}{=} \begin{bmatrix} D'\Omega^{-1}D & 0_{p\times(L-1)} \\ 0_{p\times(L-1)}' & V_2^{-1} \end{bmatrix}.$$

The desired result follows. □

Proof of Theorem 3.2. Let $t \mapsto P_t$ denote a curve from I_0, an interval containing zero, into the set of probability distributions of Z such that $P_t|_{t=0} = P$. Then the score function for the loglikelihood $\log dP_t$ is $\dot{S} \in \{h \in L_2(Z,P) : \mathbb{E}_P[h(Z)] = 0\}$. Also, let θ_t and $K_{-L,t}$ be curves through θ^* and K_{-L}^0, respectively, such that $\mathbb{E}_{P_t}[\rho_1(Z,\theta_t,K_{-L,t}^0)] = 0$, $\mathbb{E}_{P_t}[\rho_2(Z,K_{-L,t}^0)] = 0$, $\mathbb{E}_{P_t}[\rho_3(Z,K_{-L,t}^0)] = 0$, and $\mathbb{E}_{P_t}[\rho_4(Z,K_{-L,t}^0)] = 0$ for $t \in I_0$. Differentiating these moment conditions with respect to t and evaluating the resulting derivatives at $t = 0$, we can use Lemma B.1 and (B.1) to show that

$$D\dot{\theta} - \Sigma_{12}V_2^{-1}\dot{K}_{-L} + \mathbb{E}_P[\rho_1(Z,\beta_0)\dot{S}] = 0 \tag{A.1}$$

$$\dot{K}_{-L} - \mathbb{E}_P[\rho_2(Z,K_{-L}^0)\dot{S}] = 0 \tag{A.2}$$

$$\mathbb{E}_P[\rho_3(Z,K_{-L}^0)\rho_3'(Z,K_{-L}^0)](\Omega^{-1} + \frac{\tilde{1}\tilde{1}'}{Q_L^*})\dot{K}_{-L} - \mathbb{E}_P[\rho_3(Z,K_{-L}^0)\dot{S}] = 0 \tag{A.3}$$

$$\mathbb{E}_P[\rho_4(Z,K_{-L}^0)\rho_3'(Z,K_{-L}^0)](\Omega^{-1} + \frac{\tilde{1}\tilde{1}'}{Q_L^*})\dot{K}_{-L} - \mathbb{E}_P[\rho_4(Z,K_{-L}^0)\dot{S}] = 0, \tag{A.4}$$

where $\dot{\theta}$ and $\dot{K}$ are the tangent vectors to θ_t and $K_{-L,t}$, respectively, at $t = 0$.

Now, it can be shown that

$$\begin{aligned} \rho_3(Z,K_{-L}^0) &= (\Omega - Q_{-L}^*{Q_{-L}^*}')V_2^{-1}\rho_2(Z,K_{-L}^0) \\ \rho_4(Z,K_{-L}^0) &= \tilde{1}'(\Omega\mathcal{K}^{-1} - \frac{Q_L^*}{K_L^0}I_{L-1})\rho_2(Z,K_{-L}^0). \end{aligned} \tag{A.5}$$

Hence,

$$\begin{aligned} \mathrm{Var}_P[\rho_3(Z,K_{-L}^0)] &= (\Omega - Q_{-L}^*{Q_{-L}^*}')V_2^{-1}(\Omega - Q_{-L}^*{Q_{-L}^*}') \\ \mathbb{E}_P[\rho_3(Z,K_{-L}^0)\dot{S}] &= (\Omega - Q_{-L}^*{Q_{-L}^*}')V_2^{-1}\mathbb{E}_P[\rho_2(Z,K_{-L}^0)\dot{S}]. \end{aligned}$$

Therefore, since $(\mathcal{Q} - Q^*_{-L}Q^*_{-L}{}')^{-1} = \mathcal{Q}^{-1} + \tilde{1}\tilde{1}'/Q^*_L$,

$$\text{(A.2)} \iff \text{(A.3)}. \tag{A.6}$$

Again, by (A.5),

$$\mathbb{E}_P[\rho_4(Z, K^0_{-L})\rho_3'(Z, K^0_{-L})] = \tilde{1}'(\mathcal{Q}\mathcal{K}^{-1} - \frac{Q^*_L}{K^0_L} I_{L-1})(\mathcal{Q} - Q^*_{-L}Q^*_{-L}{}')$$
$$\mathbb{E}_P[\rho_4(Z, K^0_{-L})\dot{S}] = \tilde{1}'(\mathcal{Q}\mathcal{K}^{-1} - \frac{Q^*_L}{K^0_L} I_{L-1})\mathbb{E}_P[\rho_2(Z, K^0_{-L})\dot{S}].$$

Hence,

$$\text{(A.4)} \iff \tilde{1}'(\mathcal{Q}\mathcal{K}^{-1} - \frac{Q^*_L}{K^0_L} I_{L-1})(\dot{K}_{-L} - \mathbb{E}_P[\rho_2(Z, K^0_{-L})\dot{S}]) = 0,$$

which means that

$$\text{(A.2)} \implies \text{(A.4)}. \tag{A.7}$$

Thus, (A.6) and (A.7) together imply that (A.3) and (A.4) do not affect the efficiency bound for estimating θ^* or K^0_{-L}.

Now, by (A.1) and (A.2),

$$D\dot{\theta} + \mathbb{E}_P[\varepsilon\dot{S}] = 0. \tag{A.8}$$

Therefore, the tangent space of score functions is given by

$$\dot{\mathcal{M}} := \{\dot{S} \in L_2(Z, P) : \mathbb{E}_P[\dot{S}] = 0 \ \&\ \mathbb{E}_P[\varepsilon\dot{S}] \in \mathcal{R}(D)\}. \tag{A.9}$$

Suppose we want to obtain the efficiency bound for estimating $\lambda'\theta^*$, where $\lambda \in \mathbb{R}^p$ is chosen arbitrarily. Then, thinking of $\lambda'\theta_t$ as some functional η of the loglikelihood $\log dP_t$, by (A.8) it follows that, for every $\dot{S} \in \dot{\mathcal{M}}$,

$$\nabla\eta(\dot{S}) := -\lambda' D^+ \mathbb{E}_P[\varepsilon\dot{S}] = \langle -\lambda' D^+\varepsilon, \dot{S}\rangle_P,$$

where $\nabla\eta$ is the pathwise derivative of η and D^+ the Moore-Penrose generalized inverse of D. But, since $\dot{S} \in \dot{\mathcal{M}}$,

$$\nabla\eta(\dot{S}) = \langle -\lambda' D^+\varepsilon, \mathscr{P}_{\dot{\mathcal{M}}}(\dot{S})\rangle_P = \langle -\mathscr{P}_{\dot{\mathcal{M}}}(\lambda' D^+\varepsilon), \dot{S}\rangle_P.$$

Note that $\mathscr{P}_{\dot{\mathcal{M}}}$ exists, and thus is uniquely defined, because $\dot{\mathcal{M}}$ is closed in the norm topology; cf. Lemma B.2. Hence, following the argument in Severini and Tripathi (2001), the efficiency bound for estimating $\lambda'\theta^*$ is given by $\mathbb{E}_P[\mathscr{P}_{\dot{\mathcal{M}}}(\lambda' D^+\varepsilon)]^2$, the squared operator norm of $\nabla\eta$. But, by Lemma B.3,

$$\begin{aligned}\mathscr{P}_{\dot{\mathcal{M}}}(\lambda' D^+\varepsilon) &= \lambda' D^+\varepsilon - \mathbb{E}_P[\lambda' D^+\varepsilon] - \varepsilon'(I_q - \Omega^{-1}D(D'\Omega^{-1}D)^{-1}D')\Omega^{-1}\mathbb{E}_P[\varepsilon\lambda' D^+\varepsilon] \\ &= \lambda' D^+\varepsilon - \varepsilon'(I_q - \Omega^{-1}D(D'\Omega^{-1}D)^{-1}D')\Omega^{-1}\mathbb{E}_P[\varepsilon\lambda' D^+\varepsilon].\end{aligned}$$

Hence, since $\mathbb{E}_P[\varepsilon\lambda' D^+\varepsilon] = \mathbb{E}_P[\varepsilon\varepsilon']D^{+\prime}\lambda = \Omega D^{+\prime}\lambda$,

$$\begin{aligned}\mathscr{P}_{\dot{\mathcal{M}}}(\lambda' D^+\varepsilon) &= \varepsilon'\Omega^{-1}D(D'\Omega^{-1}D)^{-1}(D^+D)'\lambda \\ &= \varepsilon'\Omega^{-1}D(D'\Omega^{-1}D)^{-1}\lambda,\end{aligned}$$

where the second equality follows because the operator D^+D is a projection onto the orthogonal complement of $\mathcal{N}(D)$ — a well known property of generalized inverses (Engl, Hanke, and Neubauer, 2000, Proposition 2.3, p. 33) — and D is full rank by Assumption 3.1(viii). Therefore, the efficiency bound for estimating $\lambda'\theta^*$ is given by $\lambda'(D'\Omega^{-1}D)^{-1}\lambda$. Since λ was chosen arbitrarily, it follows that the efficiency bound for estimating θ^* is given by $(D'\Omega^{-1}D)^{-1}$.

A similar argument, but now using (A.2) instead of (A.8), shows that the efficiency bound for estimating K^0_{-L} is given by V_2. □

Proof of (3.5). Observe that

$$b_{Q^*}(Z, K^0_{-L}) := \sum_{l=1}^{L} \frac{K^0_l}{Q^*_l}\mathbb{1}(Z \in \mathbb{C}_l) \implies \frac{1}{b_{Q^*}(Z, K^0_{-L})} = \sum_{l=1}^{L} \frac{Q^*_l}{K^0_l}\mathbb{1}(Z \in \mathbb{C}_l).$$

Hence, since $\mathbb{E}_P[\mathbb{1}(Z \in \mathbb{C}_l)] = K^0_l$ by (2.2),

$$\begin{aligned}D := \mathbb{E}_P[\frac{\partial}{\partial\theta}\frac{g(Z,\theta^*)}{b_{Q^*}(Z, K^0_{-L})}] &= \sum_{l=1}^{L} \frac{Q^*_l}{K^0_l}\mathbb{E}_P[\frac{\partial g(Z,\theta^*)}{\partial\theta}\mathbb{1}(Z \in \mathbb{C}_l)] \\ &= \sum_{l=1}^{L} Q^*_l\mathbb{E}_P[\frac{\partial g(Z,\theta^*)}{\partial\theta}|Z \in \mathbb{C}_l] \\ &= \sum_{l=1}^{L} Q^*_l\mathbb{E}_{P^*}[\frac{\partial g(Z^*,\theta^*)}{\partial\theta}|Z^* \in \mathbb{C}_l] \qquad \text{(A.10)} \\ &= \mathbb{E}_{P^*}[\frac{\partial g(Z^*,\theta^*)}{\partial\theta}],\end{aligned}$$

where (A.10) follows because $\text{Law}(Z|Z \in \mathbb{C}_l) = \text{Law}(Z^*|Z^* \in \mathbb{C}_l)$ for each l by the definition of SS sampling. Next, a similar argument shows that

$$\begin{aligned}V_1 &= \mathbb{E}_P[\frac{g(Z,\theta^*)g'(Z,\theta^*)}{b^2(Z, K^0_{-L})}] \\ &= \sum_{l=1}^{L} \frac{Q^{*2}_l}{K^0_l}\mathbb{E}_P[g(Z,\theta^*)g'(Z,\theta^*)|Z \in \mathbb{C}_l].\end{aligned}$$

Moreover, since $\Sigma_{12} = \mathbb{E}_P[\rho_1(Z,\beta_0)s'(Z)]$, $V_2^{-1} = \mathcal{K}^{-1} + \tilde{1}\tilde{1}'/K^0_L$, and $s'(Z)\tilde{1} = 1 - \mathbb{1}(Z \in \mathbb{C}_L)$, some laborious but straightforward matrix algebra reveals that

$$\Sigma_{12}V_2^{-1}\Sigma'_{12} = \sum_{l=1}^{L} \frac{Q^{*2}_l}{K^0_l}\mathbb{E}_P[g(Z,\theta^*)|Z \in \mathbb{C}_l]\mathbb{E}_P[g'(Z,\theta^*)|Z \in \mathbb{C}_l].$$

Therefore,

$$\begin{aligned}\Omega &:= V_1 - \Sigma_{12}V_2^{-1}\Sigma_{12}' \\ &= \sum_{l=1}^{L} \frac{Q_l^{*2}}{K_l^0}(\mathbb{E}_P[g(Z,\theta^*)g'(Z,\theta^*)|Z \in \mathbb{C}_l] - \mathbb{E}_P[g(Z,\theta^*)|Z \in \mathbb{C}_l]\mathbb{E}_P[g'(Z,\theta^*)|Z \in \mathbb{C}_l]) \\ &= \sum_{l=1}^{L} \frac{Q_l^{*2}}{K_l^0} \operatorname{Var}_P[g(Z,\theta^*)|Z \in \mathbb{C}_l] \\ &= \sum_{l=1}^{L} \frac{Q_l^{*2}}{K_l^0} \operatorname{Var}_{P^*}[g(Z^*,\theta^*)|Z^* \in \mathbb{C}_l]. \end{aligned}$$ □

Proof of Lemma 3.1. Recall that $\hat{\theta}$ satisfies the first order necessary condition

$$\hat{D}'(\hat{\theta}, \hat{K}_{-L})\hat{\Omega}^{-1}(\check{\theta}, \hat{K}_{-L})\hat{\rho}_1(\hat{\theta}, \hat{K}_{-L}) = 0_{p\times 1}.$$

Now, by a mean value expansion,

$$\hat{\rho}_1(\hat{\theta}, \hat{K}_{-L}) = \hat{\rho}_1(\theta^*, \hat{K}_{-L}) + \hat{D}(\bar{\theta}, \hat{K}_{-L})(\hat{\theta} - \theta^*),$$

where $\bar{\theta}$ lies between $\hat{\theta}$ and θ^*. Therefore, $n^{1/2}(\hat{\theta} - \theta^*)$ equals

$$-(\hat{D}'(\hat{\theta}, \hat{K}_{-L})\hat{\Omega}^{-1}(\check{\theta}, \hat{K}_{-L})\hat{D}(\bar{\theta}, \hat{K}_{-L}))^{-1}\hat{D}'(\hat{\theta}, \hat{K}_{-L})\hat{\Omega}^{-1}(\check{\theta}, \hat{K}_{-L})n^{1/2}\hat{\rho}_1(\theta^*, \hat{K}_{-L}).$$

Similarly, by another mean value expansion,

$$\hat{\rho}_1(\theta^*, \hat{K}_{-L}) = \hat{\rho}_1(\theta^*, K_{-L}^0) + \hat{D}_2(\theta^*, \bar{K}_{-L})(\hat{K}_{-L} - K_{-L}^0),$$

where $\hat{D}_2(\theta, K_{-L}) := \partial\hat{\rho}_1(\theta, K_{-L})/\partial K_{-L}$ and $\bar{K}_{-L}$ lies between $\hat{K}_{-L}$ and K_{-L}^0. But $\hat{K}_{-L} - K_{-L}^0 = \hat{\rho}_2(K_{-L}^0)$ because ρ_2 is linear in parameters. Therefore, by a uniform weak law of large numbers (Newey and McFadden, 1994, Lemma 2.4),

$$n^{1/2}(\hat{\theta} - \theta^*) = -(D'\Omega^{-1}D)^{-1}D'\Omega^{-1}n^{1/2}(\hat{\rho}_1(\beta_0) + D_2\hat{\rho}_2(K_{-L}^0)) + o_P(1).$$

Hence, by Lemma B.1,

$$n^{1/2}(\hat{\theta} - \theta^*) = \sum_{j=1}^{n} -(D'\Omega^{-1}D)^{-1}D'\Omega^{-1}\varepsilon_j + o_P(1). \tag{A.11}$$

From the proof of Theorem 3.1 it is clear that the influence function of $\tilde{\theta}$ is also given by $-(D'\Omega^{-1}D)^{-1}D'\Omega^{-1}\varepsilon$, i.e.,

$$n^{1/2}(\tilde{\theta} - \theta^*) = n^{-1/2}\sum_{j=1}^{n} -(D'\Omega^{-1}D)^{-1}D'\Omega^{-1}\varepsilon_j + o_P(1). \tag{A.12}$$

Therefore, the desired result follows from (A.11) and (A.12). □

Appendix B. Some useful results

Lemma B.1. *$D_2 = -\Sigma_{12}V_2^{-1}$. Therefore, $J = 0_{q\times(L-1)}$.*

Proof of Lemma B.1. We show that $D_2 = -\Sigma_{12}V_2^{-1}$. The consequence that $J = 0_{q\times(L-1)}$ then follows from the definition of J. Begin by observing that

$$D_2 = -\mathbb{E}_P[\frac{\rho_1(Z,\beta_0)}{b_{Q^*}(Z,K^0_{-L})}\frac{\partial b_{Q^*}(Z,K^0_{-L})}{\partial K_{-L}}].$$

But we can show that

$$\begin{aligned}\frac{\partial b_{Q^*}(Z,K^0_{-L})}{\partial K_{-L}} &= (s(Z)-Q^*_{-L})'(\mathcal{Q}^{-1}+\frac{\tilde{1}\tilde{1}'}{Q^*_L}) \\ &= \rho_3'(Z,K^0_{-L})(\mathcal{Q}^{-1}+\frac{\tilde{1}\tilde{1}'}{Q^*_L})b_{Q^*}(Z,K^0_{-L}).\end{aligned} \tag{B.1}$$

Hence, by (A.5) and the fact that $(\mathcal{Q}-Q^*_{-L}{Q^*_{-L}}')^{-1} = \mathcal{Q}^{-1}+\tilde{1}\tilde{1}'/Q^*_L$,

$$D_2 = -\mathbb{E}_P[\rho_1(Z,\beta_0)\rho_3'(Z,K^0_{-L})](\mathcal{Q}^{-1}+\frac{\tilde{1}\tilde{1}'}{Q^*_L}) = -\Sigma_{12}V_2^{-1}. \qquad \square$$

Lemma B.2. *$\dot{\mathcal{M}}$, defined in* (A.9)*, is closed in the $\|\cdot\|_{2,P}$ norm.*

Proof of Lemma B.2. Let $\dot{m} \in \mathrm{cl}(\dot{\mathcal{M}})$. Then, there exists a sequence $(m_j)_{j\in\mathbb{N}} \subset \dot{\mathcal{M}}$ such that $\|m_j - \dot{m}\|_{2,P} \to 0$ as $j \to \infty$. Clearly, $\dot{m} \in L_2(Z,P)$ and

$$\lim_{j\to\infty}\|m_j-\dot{m}\|_{2,P} = 0 \implies \lim_{j\to\infty}\mathbb{E}_P[m_j] = \mathbb{E}_P[\dot{m}] \implies \mathbb{E}_P[\dot{m}] = 0.$$

Moreover, by Cauchy-Schwarz,

$$\|\mathbb{E}_P[\varepsilon m_j] - \mathbb{E}_P[\varepsilon\dot{m}]\|^2 \le \mathrm{trace}(\Omega)\|m_j-\dot{m}\|^2_{2,P} \implies \mathbb{E}_P[\varepsilon m_j] \xrightarrow[j\to\infty]{} \mathbb{E}_P[\varepsilon\dot{m}].$$

But since $\mathbb{E}_P[\varepsilon m_j] \in \mathcal{R}(D)$ for every $j \in \mathbb{N}$ and $\mathcal{R}(D)$ is finite dimensional hence closed, it follows that $\mathbb{E}_P[\varepsilon\dot{m}] \in \mathcal{R}(D)$. Therefore, $\dot{m} \in \dot{\mathcal{M}}$ and the desired result follows. $\square$

Lemma B.3. *Let $h \in L_2(Z,P)$. Then,*

$$\mathscr{P}_{\dot{\mathcal{M}}}(h) = h - \mathbb{E}_P[h] - \varepsilon'(I_q - \Omega^{-1}D(D'\Omega^{-1}D)^{-1}D')\Omega^{-1}\mathbb{E}_P[\varepsilon h].$$

Proof of Lemma B.3. Let

$$\pi^* := h - \mathbb{E}_P[h] - \varepsilon'(I_q - \Omega^{-1}D(D'\Omega^{-1}D)^{-1}D')\Omega^{-1}\mathbb{E}_P[\varepsilon h].$$

Note that $\pi^* \in L_2(Z,P)$, $\mathbb{E}_P[\pi^*] = 0$, and

$$\mathbb{E}_P[\varepsilon\pi^*] = D(D'\Omega^{-1}D)^{-1}D'\Omega^{-1}\mathbb{E}_P[\varepsilon h] \in \mathcal{R}(D).$$

Thus, $\pi^* \in \dot{\mathcal{M}}$. Next, let $\dot{m} \in \dot{\mathcal{M}}$. Then,

$$\begin{aligned}\langle h - \pi^*, \dot{m} \rangle_P &= \langle \varepsilon'(I_q - \Omega^{-1}D(D'\Omega^{-1}D)^{-1}D')\Omega^{-1}\mathbb{E}_P[\varepsilon h], \dot{m} \rangle_P \\ &= \mathbb{E}_P[\varepsilon' h]\Omega^{-1}(I_q - D(D'\Omega^{-1}D)^{-1}D'\Omega^{-1})\mathbb{E}_P[\varepsilon \dot{m}].\end{aligned}$$

But

$$\dot{m} \in \dot{\mathcal{M}} \implies \mathbb{E}_P[\varepsilon \dot{m}] \in \mathcal{R}(D) \iff \mathbb{E}_P[\varepsilon \dot{m}] = D\alpha \text{ for some } \alpha \in \mathbb{R}^p.$$

Therefore,

$$(I_q - D(D'\Omega^{-1}D)^{-1}D'\Omega^{-1})\mathbb{E}_P[\varepsilon \dot{m}] = (I_q - D(D'\Omega^{-1}D)^{-1}D'\Omega^{-1})D\alpha = 0_{q\times 1}.$$

Hence, $\langle h - \pi^*, \dot{m} \rangle_P = 0$ for every $\dot{m} \in \dot{\mathcal{M}}$. The desired result follows. □

Lemma B.4. $\sum_{l=1}^{L} Q_l^* \operatorname{Var}_{P^*}[g(Z^*, \theta^*)|Z^* \in \mathbb{C}_l] - \operatorname{Var}_{P^*}[g(Z^*, \theta^*)]$ *is negative definite.*

Proof of Lemma B.4. Let $\alpha \in \mathbb{R}^q$ and $g^* := g(Z^*, \theta^*)$. Since

$$\begin{aligned}\operatorname{Var}_{P^*}[\alpha' g^*|Z^* \in \mathbb{C}_l] &= \mathbb{E}_{P^*}[(\alpha' g^*)^2|Z^* \in \mathbb{C}_l] - (\mathbb{E}_{P^*}[\alpha' g^*|Z^* \in \mathbb{C}_l])^2 \\ &= \frac{\mathbb{E}_{P^*}[(\alpha' g^*)^2 \mathbb{1}(Z^* \in \mathbb{C}_l)]}{Q_l^*} - \left(\frac{\mathbb{E}_{P^*}[\alpha' g^* \mathbb{1}(Z^* \in \mathbb{C}_l)]}{Q_l^*}\right)^2\end{aligned}$$

and $x \mapsto x^2$ is strictly convex, by Jensen's inequality we have that

$$\begin{aligned}\sum_{l=1}^{L} Q_l^* \operatorname{Var}_{P^*}[\alpha' g^*|Z^* \in \mathbb{C}_l] &= \mathbb{E}_{P^*}[\alpha' g^*]^2 - \sum_{l=1}^{L} Q_l^* \left(\frac{\mathbb{E}_{P^*}[\alpha' g^* \mathbb{1}(Z^* \in \mathbb{C}_l)]}{Q_l^*}\right)^2 \\ &< \mathbb{E}_{P^*}[\alpha' g^*]^2 - (\sum_{l=1}^{L} \mathbb{E}_{P^*}[\alpha' g^* \mathbb{1}(Z^* \in \mathbb{C}_l)])^2 \\ &= \mathbb{E}_{P^*}[\alpha' g^*]^2 - (\mathbb{E}_{P^*}[\alpha' g^*])^2 \\ &= \operatorname{Var}_{P^*}[\alpha' g^*].\end{aligned}$$

□

References

Cochran, W. G. (1963): *Sampling techniques*, John Wiley & Sons, 2nd ed.

Cosslett, S. (1981a): "Efficient estimation of discrete choice models," in *Structural analysis of discrete data with econometric applications*, ed. by C. F. Manski and D. McFadden, Cambridge, MA: MIT Press, 51–111.

——— (1981b): "Maximum likelihood estimation for choice-based samples," *Econometrica*, 49, 1289–1316.

——— (1993): "Estimation from endogenously stratified samples," in *Handbook of Statistics*, ed. by G. Maddala, C. Rao, and H. Vinod, Elsevier Science, vol. 11, 1–43.

Engl, H. W., M. Hanke, and A. Neubauer (2000): *Regularization of inverse problems*, Kluwer.

Hausman, J. A. and D. A. Wise (1981): "Stratification on endogenous variables and estimation: The Gary income maintenance experiment," in *Structural analysis of discrete data with econometric applications*, ed. by C. F. Manski and D. McFadden, Cambridge: MIT Press, 365–391.

Imbens, G. W. (1992): "An efficient method of moments estimator for discrete choice models with choice-based sampling," *Econometrica*, 60, 1187–1214.

Imbens, G. W. and T. Lancaster (1996): "Efficient estimation and stratified sampling," *Journal of Econometrics*, 74, 289–318.

Jewell, N. P. (1985): "Least squares regression with data arising from stratified samples of the dependent variable," *Biometrika*, 72, 11–21.

Manski, C. F. and S. R. Lerman (1977): "The estimation of choice probabilities from choice based samples," *Econometrica*, 45, 1977–1988.

Manski, C. F. and D. McFadden (1981): "Alternative estimators and sample design for discrete choice analysis," in *Structural analysis of discrete data with econometric applications*, ed. by C. F. Manski and D. McFadden, Cambridge, MA: MIT Press, 2–50.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," in *Handbook of Econometrics, vol. IV*, ed. by R. Engle and D. McFadden, Elsevier Science B.V., 2111–2245.

Quesenberry, C. P. and N. P. Jewell (1986): "Regression analysis based on stratified samples," *Biometrika*, 73, 605–614.

Severini, T. A. and G. Tripathi (2001): "A simplified approach to computing efficiency bounds in semiparametric models," *Journal of Econometrics*, 102, 23–66.

Tripathi, G. (2007): "Moment based inference with stratified data," Accepted for publication in *Econometric Theory*.

Wooldridge, J. M. (2001): "Asymptotic properties of weighted M-estimators for standard stratified samples," *Econometric Theory*, 17, 451–470.

Department of Economics, University of Connecticut, Storrs, CT - 06269.
E-mail address: gautam.tripathi@uconn.edu