January 2007

Moment Based Inference with Stratified Data

Gautam Tripathi
University of Connecticut

University of Connecticut

Department of Economics Working Paper Series

Moment Based Inference with Stratified Data

Gautam Tripathi
University of Connecticut

Working Paper 2005-38R

September 2005, revised January 2007

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

Many datasets used by economists and other social scientists are collected by stratified sampling. The sampling scheme used to collect the data induces a probability distribution on the observed sample that differs from the target or underlying distribution for which inference is to be made. If this effect is not taken into account, subsequent statistical inference can be seriously biased. This paper shows how to do efficient semiparametric inference in moment restriction models when data from the target population is collected by three widely used sampling schemes variable probability sampling, multinomial sampling, and standard stratified sampling.

Journal of Economic Literature Classification: C14

Keywords: Empirical likelihood, Moment conditions, Stratified sampling.

I thank the co-editors and two anonymous referees for comments that greatly improved this paper. I also thank Paul Devereux and seminar participants at several universities for helpful suggestions and conversations. Financial support for this project from NSF grant SES-0214081 is gratefully acknowledged.

1. Introduction

The process of doing applied research in economics and other social sciences can be divided into three distinct yet equally important steps. First, a model is written in terms of the target population for which inference is to be made. Next, data is collected. Finally, the resulting data is used to draw inference about the target population.

If data is collected by random sampling, so that observations from the target population have the same chance of being represented in the sample, then there is no distinction between the target and observed data distributions and statistical inference is straightforward. However, for administrative convenience or to increase statistical precision by oversampling rare but informative outcomes, in many applications data is collected by stratified sampling so that observations from the target population have unequal chances of being selected. Hence, the sampling scheme used to collect the data induces a probability distribution on the observed sample that differs from the target or underlying distribution for which inference is to be made. Subsequent inference can, therefore, be seriously biased if this effect is not taken into account.

In this paper we show how to do efficient inference in models defined via unconditional moment restrictions when data from the target population is collected by stratified sampling. Earlier works in the literature, with few exceptions, either make parametric assumptions about the conditional density of variables in the target population or look at linear regression or non-linear discrete response models; see, e.g., DeMets and Halperin (1977), Manski and Lerman (1977), Holt, Smith, and Winter (1980), Cosslett (1981a, 1981b), Hausman and Wise (1981), Manski and McFadden (1981), DuMouchel and Duncan (1983), Jewell (1985), Quesenberry and Jewell (1986), Scott and Wild (1986), Bickel and Ritov (1991), Imbens (1992), Imbens and Lancaster (1996), and Butler (2000).

Unlike these papers, the class of overidentified models examined here subsumes linear regression and discrete choice models as special cases; e.g., our ability to handle instrumental variables (IV) models allows semiparametric inference in Box-Cox type models using stratified datasets, an important advantage because it is well known that least squares is not consistent for estimating such models. The unified approach proposed in this paper can deal with different kinds of sampling schemes and our treatment is general enough to handle stratification based only on the response variables, or on the explanatory variables alone, or stratification that is based on a subset of these variables; the stratifying variables can be discrete or continuously distributed. We have taken special care to derive intuitive closed form expressions for the asymptotic variances of estimators so that standard errors are easily obtained.

Wooldridge (1999, 2001) also leaves the target density completely unspecified and provides asymptotic theory for M-estimators. However, his model is defined in terms of a set of just identified moment conditions whereas we deal with possibly overidentified moment restrictions; therefore, our model nests his moment conditions as a special case. Since the moment conditions in Wooldridge's papers are exactly identified, their validity cannot be tested unless

additional moment conditions are added. In contrast, specification testing under stratification is examined in this paper. For standard stratified sampling, Wooldridge (2001) assumes that the aggregate shares (defined in Section 2.2) are known, whereas we treat the aggregate shares as unknown parameters but require an additional random sample to deal with the consequent lack of identification; see Section 4.1 for details. Qin (1993) uses data collected by variable probability sampling, along with an independent sample from the target population, to construct empirical likelihood based confidence intervals for the population mean of the target population. El-Barmi and Rothmann (1998) generalize Qin's treatment to handle nonlinear overidentified models; they also use two independent samples whereas we only need a single sample to do inference when data is collected by variable probability sampling. Unlike us, Qin or El-Barmi and Rothmann do not investigate other kinds of sampling schemes; nor do the latter consider testing the overidentifying restrictions.

2. Stratification in a moment based framework

2.1. **The model.** Let Z^* be a $d \times 1$ random vector that denotes an observation from the target population and Θ a subset of $\mathbb{R}^p$ such that

$$\mathbb{E}_{f^*}\{g(Z^*, \theta^*)\} = 0 \quad \text{for some } \theta^* \in \Theta, \tag{2.1}$$

where g is a $q \times 1$ vector of functions known up to θ^* such that $q \geq p$, i.e., overidentification is allowed, and f^* is the unknown density of Z^* with respect to a dominating measure μ which need not be the Lebesgue measure so that Z^* can have discrete components. The notation $\mathbb{E}_{f^*}$ indicates that expectation is with respect to f^*. Henceforth, "vector" means a column vector.

A familiar example of (2.1) is the linear model $Y^* = X^{*\prime}\theta^* + \varepsilon^*$, where $\mathbb{E}_{f^*}\{X^*\varepsilon^*\} = 0$; here, $g(Z^*, \theta^*) = X^*(Y^* - X^{*\prime}\theta^*)$ and $Z^* = (Y^*, X^*)_{(p+1)\times 1}$. Extensions include nonlinear regression or simultaneous equations models. We can also handle conditional moment restrictions in an IV framework; e.g., if $\mathbb{E}_{Y^*|X^*}\{\tilde{g}(Y^*, X^*, \theta^*)|X^*\} = 0$ w.p.1, where $\tilde{g}$ is a vector of functions known up to θ^*, then (2.1) holds with $g(Z^*, \theta^*) = A(X^*)\tilde{g}(Y^*, X^*, \theta^*)$ for a conformable matrix of instruments $A(X^*)$. Although it is possible to improve upon IV estimators, because conditional moment restrictions are stronger than unconditional ones, such an extension is beyond the scope of this paper; see, e.g., Tripathi (2002).

If data is collected by random sampling, then (2.1) is easily handled; see, e.g., Newey and McFadden (1994). However, if data is collected by stratified sampling, then the sample consists of iid observations $Z_1, \ldots, Z_n$ generated from f, the density induced by the sampling scheme, instead of iid observations from the target density f^*. Hence, unless proper precautions are taken, statistical inference using stratified data is about f and not f^*; e.g., the sample average $\sum_{j=1}^n Z_j/n$ is not a consistent estimator of the mean of the target population because $\sum_{j=1}^n Z_j/n \xrightarrow{p} \mathbb{E}_f\{Z\}$ by the weak law of large numbers (throughout the paper, all limits are taken as the total sample size $n \uparrow \infty$), but $\mathbb{E}_f\{Z\} \neq \mathbb{E}_{f^*}\{Z^*\}$.

2.2. **Some commonly used sampling schemes.** Let the target population be partitioned into L nonempty disjoint strata $\mathbb{C}_1, \ldots, \mathbb{C}_L$. Depending upon the manner in which the observations are actually drawn from the strata, we study three general sampling schemes: variable probability (VP) sampling, multinomial (MN) sampling, and standard stratified (SS) sampling. Good descriptions of these stratified sampling schemes can be found in Jewell (1985), Cosslett (1993), Imbens and Lancaster (1996), and Wooldridge (1999).

In VP sampling, typically used when data is collected by telephone surveys, an observation is first drawn randomly from the target population. If it lies in stratum $\mathbb{C}_l$ it is retained with known probability P_l; if it is discarded, all information about the observation is lost. Hence, instead of observing a random variable Z^* drawn from the target density f^*, we observe a random variable Z drawn from the density

$$f(z) = \frac{\sum_{l=1}^{L} P_l \mathbb{1}(z \in \mathbb{C}_l) f^*(z)}{\sum_{l=1}^{L} P_l Q_l^*} \stackrel{\text{def}}{=} \frac{b(z) f^*(z)}{b^*}, \tag{2.2}$$

where $b(z) = \sum_{l=1}^{L} P_l \mathbb{1}(z \in \mathbb{C}_l)$, $Q_l^* = \int_{\mathbb{C}_l} f^*(z)\, d\mu$, $b^* = \sum_{l=1}^{L} P_l Q_l^*$, and $\mathbb{1}$ is the indicator function. Q_l^* denotes the probability that a randomly chosen observation from the target population lies in the l^{th} stratum; i.e., the "demand" for the l^{th} stratum. The Q_l^*'s, popularly called "aggregate shares", are unknown parameters of interest and will be estimated along with the structural parameter θ^*. The parameter b^* also has a practical interpretation. It is the probability that an observation from the target population is ultimately retained in the sample.

In MN sampling, the researcher first selects a stratum, say $\mathbb{C}_l$, with known probability H_l so that $H_1 + \ldots + H_L = 1$. Then, an observation is drawn randomly from the selected stratum. Hence, instead of observing Z^* from the target density f^*, we observe Z from the density $f(z) = \sum_{l=1}^{L} (H_l / Q_l^*) \mathbb{1}(z \in \mathbb{C}_l) f^*(z)$.

In SS sampling, used for most large datasets, the number of observations drawn from each stratum is fixed in advance and data is sampled randomly within each stratum. Suppose that n observations $Z_1, \ldots, Z_n$ are collected by SS sampling. The density for a single observation is given by $f_n(z) = \sum_{l=1}^{L} (n_l / n) \mathbb{1}(z \in \mathbb{C}_l) f^*(z) / Q_l^*$, where $n_l = \sum_{j=1}^{n} \mathbb{1}(Z_j \in \mathbb{C}_l)$ is the number of observations lying in the l^{th} stratum of the stratified dataset.

Unlike MN sampling, observations collected by SS sampling are independently but not identically distributed (inid) because in SS sampling the n_l's are treated as nonstochastic constants whereas in MN sampling they are random variables. Thus statistical inference under SS sampling should be done conditional on the observed values of the n_l's. This can be achieved in a simple manner by the following trick: Let $\tilde{K} = (\tilde{K}_1, \ldots, \tilde{K}_L)$ denote an $L \times 1$ vector of unknown parameters in $(0,1)^L$ such that $\sum_{l=1}^{L} \tilde{K}_l = 1$ and assume, counterfactually, that observations collected by SS sampling are iid draws from the density

$$f(z) = \sum_{l=1}^{L} \frac{\tilde{K}_l \mathbb{1}(z \in \mathbb{C}_l) f^*(z)}{Q_l^*} \stackrel{\text{def}}{=} b(z, Q^*, \tilde{K}) f^*(z), \tag{2.3}$$

where $b(z, Q^*, \tilde{K}) = \sum_{l=1}^{L} (\tilde{K}_l/Q_l^*)\mathbb{1}(z \in \mathbb{C}_l)$ and $Q^* = (Q_1^*, \ldots, Q_L^*)_{L\times 1}$. In Section 4.3, we show that estimating $\tilde{K}$ jointly and efficiently with θ^* and Q^* leads to asymptotic inference that is conditional on the number of observations lying in each stratum of the stratified sample. Therefore, although we work in an artificially created iid environment, because in an iid setting it is easier to do efficiency bound calculations, apply standard statistical arguments to prove our results, etc., the results we obtain are identical to those under the inid framework.

Since the densities for MN and SS schemes are observationally equivalent conditional on the number of observations lying within each stratum, inference for them will be the same provided we condition on the number of observations lying in each stratum of the stratified dataset. Therefore, without loss of generality, henceforth we only consider SS sampling.

3. Inference when data is collected by variable probability sampling

In this section we investigate estimating and testing (2.1) when data is collected by VP sampling. We begin with an example.

Example 3.1 (Linear regression)**.** Let $Y^* = X^{*\prime}\theta^* + \varepsilon^*$, where $\mathbb{E}_{f^*}\{X^*\varepsilon^*\} = 0$. Instead of $Z^* = (Y^*, X^*)$ from the target density, we observe $Z = (Y, X)$ from (2.2). The least squares estimator that ignores stratification, denoted by $\hat{\theta}_{LS} = (\sum_{j=1}^{n} X_j X_j')^{-1} \sum_{j=1}^{n} X_j Y_j$, is not a consistent estimator of θ^*. To see this, observe that $\text{plim}(\hat{\theta}_{LS}) = (\mathbb{E}_f XX')^{-1}(\mathbb{E}_f XY) \overset{(2.2)}{=} \theta^* + (\mathbb{E}_f XX')^{-1}\mathbb{E}_{f^*}\{b(Z^*)X^*\varepsilon^*\}/b^*$. But since $\mathbb{E}_{f^*}\{X^*\varepsilon^*\} = 0$ does not imply $\mathbb{E}_{f^*}\{b(Z^*)X^*\varepsilon^*\} = 0$, it follows that $\hat{\theta}_{LS}$ is not consistent for θ^*. Furthermore, since the asymptotic bias depends upon the distribution of Z^* and the retention probabilities, the decision to ignore stratification can only be made on a case by case basis; see, e.g., Imbens and Lancaster (1996). The least squares estimator remains inconsistent even if stratification is based only upon X^*. However, as pointed out by Wooldridge (1999, 2001) and Tripathi (2002), if the identifying assumption $\mathbb{E}_{f^*}\{X^*\varepsilon^*\} = 0$ is replaced by the stronger condition $\mathbb{E}_{Y^*|X^*}\{\varepsilon^*|X^*\} = 0$ w.p.1, then ignoring stratification based on the explanatory variables does not affect the consistency of $\hat{\theta}_{LS}$ although it will still affect its asymptotic variance. □

3.1. Identification.

Since we use f to do inference on f^*, before proceeding any further we first have to investigate whether f^* can be recovered in terms of f. If there is no way of going from the stratified sample density (loosely speaking, the "reduced form") to the target density (the "structural form"), then moment based inference about f^* is impossible. In other words, we first have to examine whether f^* is identified. The density f is of course identified by definition since it generates the data.

Fortunately, there are no identification issues for VP sampling. As discussed later in Section 4.1, this is in sharp contrast to SS sampling where ignorance of Q^* leads to serious identification problems. All parameters of interest associated with VP sampling are identifiable from the stratified sample alone; namely, b^* is identified because $b^* = 1/\mathbb{E}_f\{1/b(Z)\}$, the

aggregate shares are identified because, for each l,

$$\mathbb{E}_f\{\mathbb{1}(Z \in \mathbb{C}_l)/b(Z)\} = Q_l^* \, \mathbb{E}_f\{1/b(Z)\} \iff Q_l^* = \mathbb{E}_f\{\mathbb{1}(Z \in \mathbb{C}_l)/b(Z)\}/\mathbb{E}_f\{1/b(Z)\}, \quad (3.1)$$

and identification of f^* follows from the fact that $f^*(z) = f(z)/[b(z)\mathbb{E}_f\{b^{-1}(Z)\}]$. Therefore, b^*, Q^*, and f^* can all be explicitly written in terms of f.

3.2. **Efficient estimation.** The inference in this paper is based on the empirical likelihood (EL) approach proposed by Owen (1988), although the results obtained here also hold for the generalized method of moments (GMM) used widely in econometrics. EL, however, has lately begun to emerge as a serious contender to GMM; see, e.g., Qin and Lawless (1994), Imbens (1997), Kitamura (1997, 2001, 2006), Smith (1997, 2005), Imbens, Spady, and Johnson (1998), and Owen (2001). Although EL and GMM based inference is asymptotically equivalent up to a first order analysis, recent research by Newey and Smith (2004) has shown that under certain regularity conditions EL has better second order properties than GMM; e.g., unlike GMM, the second order bias of EL does not depend upon the number of moment conditions which makes it very attractive for estimating models with large q, such as panel data models with long time dimension, where GMM is known to perform poorly in small samples.

Our estimator for θ^* is easy to motivate: Since $\mathbb{E}_{f^*}\{g(Z^*, \theta^*)\} = 0$ if and only if $\mathbb{E}_f\{g(Z, \theta^*)/b(Z)\} = 0$, we can efficiently estimate θ^* by doing EL on the transformed moment function $g(Z, \theta)/b(Z)$.[1] Technically, this is a change of measure result; i.e., since f^* can be expressed in terms of f by inverting the mapping in (2.2), dividing $g(Z, \theta^*)$ by $b(Z)$ allows (2.1) to be rewritten in terms of f without loss of information. More intuitively, since $1/b(Z) = \sum_{l=1}^{L} \mathbb{1}(Z \in \mathbb{C}_l)/P_l$, this transformation represents an "inverse probability" weighting scheme in which oversampled strata are assigned smaller weights than the undersampled strata, thereby correcting the effects of stratification.

Example 3.2 (Population mean)**.** Since $\mathbb{E}_{f^*}\{Z^* - \theta^*\} = 0$ if and only if $\mathbb{E}_f\{(Z - \theta^*)/b(Z)\} = 0$, the EL estimator of the mean of the target population is $\hat{\theta} = \sum_{j=1}^{n} Z_j b^{-1}(Z_j) / \sum_{j=1}^{n} b^{-1}(Z_j)$. This can be written more revealingly as $\hat{\theta} = \sum_{l=1}^{L} \hat{Q}_l \bar{Z}_l$, where $\hat{Q}_l = (n_l/P_l)/\sum_{l=1}^{L}(n_l/P_l)$ estimates the l^{th} aggregate share and $\bar{Z}_l = \sum_{j=1}^{n} Z_j \mathbb{1}(Z_j \in \mathbb{C}_l)/n_l$ is the l^{th} stratum sample average. It can be directly shown that $n^{1/2}(\hat{\theta} - \theta^*)$ is asymptotically normal with mean zero and variance $b^{*2}\mathbb{E}_f\{(Z - \theta^*)(Z - \theta^*)'/b^2(Z)\}$, which agrees with the result in Theorem 3.1. □

The aggregate shares can be estimated jointly with θ^* by including additional moment conditions. In fact, since they add up to one, it suffices to estimate $Q_{-L}^* = (Q_1^*, \ldots, Q_{L-1}^*)_{(L-1)\times 1}$. So let $\beta^* = (\theta^*, Q_{-L}^*)_{(p+L-1)\times 1}$ and $s(Z) = (\mathbb{1}(Z \in \mathbb{C}_1), \ldots, \mathbb{1}(Z \in \mathbb{C}_{L-1}))_{(L-1)\times 1}$. Following (3.1), define the $(q + L - 1) \times 1$ transformed moment function

$$\rho(Z, \beta) = \begin{bmatrix} g(Z, \theta)/b(Z) \\ \{s(Z) - Q_{-L}\}/b(Z) \end{bmatrix} \stackrel{\text{def}}{=} \begin{bmatrix} \rho_1(Z, \theta) \\ \rho_2(Z, Q_{-L}) \end{bmatrix}, \quad (3.2)$$

where $\rho_1(Z,\theta) = g(Z,\theta)/b(Z)$ and $\rho_2(Z,Q_{-L}) = \{s(Z) - Q_{-L}\}/b(Z)$. An asymptotically efficient estimator of β^* can be obtained by doing EL on (3.2) as follows: For a fixed β, construct the nonparametric loglikelihood for the observed sample by solving

$$\max_{p_1,\ldots,p_n} \sum_{j=1}^{n} \log p_j \quad s.t. \quad p_j \geq 0,\ \sum_{j=1}^{n} p_j = 1,\ \sum_{j=1}^{n} \rho(Z_j,\beta) p_j = 0.$$

The solution to this optimization problem is given by $\hat{p}_j(\beta) = n^{-1}\{1 + \lambda'(\beta)\rho(Z_j,\beta)\}^{-1}$, where $j = 1,\ldots,n$ and $\lambda(\beta)$ satisfies $\sum_{j=1}^{n} \rho(Z_j,\beta)/\{1 + \lambda'(\beta)\rho(Z_j,\beta)\} = 0$. Now let

$$\mathrm{EL}(\beta) = \sum_{j=1}^{n} \log \hat{p}_j(\beta) = -\sum_{j=1}^{n} \log\{1 + \lambda'(\beta)\rho(Z_j,\beta)\} - n\log n \tag{3.3}$$

and, for $\mathcal{B} = \Theta \times [0,1]^{L-1}$, define the empirical likelihood estimator of β^* as

$$\hat{\beta} = \underset{\beta\in\mathcal{B}}{\operatorname{argmax}}\, \mathrm{EL}(\beta).$$

Let $\|\cdot\|$ be the Euclidean norm and $\partial\rho(Z,\beta)/\partial\beta$ the $(q+L-1)\times(p+L-1)$ Jacobian matrix. The regularity conditions below ensure that $\hat{\beta}$ is consistent and asymptotically normal.

Assumption 3.1. (i) $\beta^* \in \mathcal{B}$ *is the unique solution to* $\mathbb{E}_f\{\rho(Z,\beta)\} = 0$*;* (ii) $\mathcal{B}$ *is compact;* (iii) $\rho(Z,\beta)$ *is continuous at each* $\beta \in \mathcal{B}$ *with probability one;* (iv) $\mathbb{E}_f\{\sup_{\beta\in\mathcal{B}} \|\rho(Z,\beta)\|^\alpha\} < \infty$ *for some* $\alpha > 2$*;* (v) $\mathbb{E}_f\{\rho(Z,\beta^*)\rho'(Z,\beta^*)\}$ *is nonsingular;* (vi) $\beta^* \in int(\mathcal{B})$*;* (vii) $\rho(Z,\beta)$ *is continuously differentiable in a neighborhood* $\mathcal{N}$ *of* β^* *and* $\mathbb{E}_f\{\sup_{\beta\in\mathcal{N}} \|\partial\rho(Z,\beta)/\partial\beta\|\} < \infty$*;* (viii) $\mathbb{E}_f\{\partial\rho(Z,\beta^*)/\partial\beta\}$ *is of full column rank.*

Newey and Smith (2004, page 226) use (i)–(v) to show the consistency and (vi)–(viii) to prove the asymptotic normality of EL estimators. In particular, letting $D = \mathbb{E}_f\{\partial\rho_1(Z,\theta^*)/\partial\theta\}$, $V_1 = \mathbb{E}_f\{\rho_1(Z,\theta^*)\rho_1'(Z,\theta^*)\}$, $V_2 = \mathbb{E}_f\{\rho_2(Z,Q_{-L}^*)\rho_2'(Z,Q_{-L}^*)\}$, $\Sigma_{12} = \mathbb{E}_f\{\rho_1(Z,\theta^*)\rho_2'(Z,Q_{-L}^*)\}$, $M_{V_1} = V_1^{-1} - V_1^{-1}D(D'V_1^{-1}D)^{-1}D'V_1^{-1}$, and $0_{k_1\times k_2}$ the $k_1 \times k_2$ matrix of zeros, we can show the following result.

Theorem 3.1. *Let Assumption* 3.1 *hold. Then,*

$$\begin{bmatrix} n^{1/2}(\hat{\theta} - \theta^*) \\ n^{1/2}(\hat{Q}_{-L} - Q_{-L}^*) \end{bmatrix} \xrightarrow{d} \mathrm{N}(0_{(p+L-1)\times 1}, \begin{bmatrix} (D'V_1^{-1}D)^{-1} & -b^*(D'V_1^{-1}D)^{-1}D'V_1^{-1}\Sigma_{12} \\ -b^*\Sigma_{12}'V_1^{-1}D(D'V_1^{-1}D)^{-1} & b^{*2}(V_2 - \Sigma_{12}'M_{V_1}\Sigma_{12}) \end{bmatrix}).$$

The estimators $\hat{\theta}$ and $\hat{Q}_{-L}$ are asymptotically efficient because it can be shown that $(D'V_1^{-1}D)^{-1}$ and $b^{*2}(V_2 - \Sigma_{12}'M_{V_1}\Sigma_{12})$ coincide with the efficiency bounds for estimating θ^* and Q_{-L}^*, respectively. Notice that if $b(Z)$ is constant, so that stratification disappears, then $(D'V_1^{-1}D)^{-1}$ becomes the well known asymptotic variance for estimating θ^* in the absence of stratification. Similarly, if there is no auxiliary information, e.g., if g is identically zero or if there are no overidentifying restrictions, then the asymptotic variance of $n^{1/2}(\hat{Q}_{-L} - Q_{-L}^*)$ reduces to $b^{*2}V_2$. Therefore, imposing the overidentified model leads to an efficiency gain in

estimating the aggregate shares. Theorem 3.1 also reveals that if θ^* is the only parameter of interest then it is not necessary to jointly estimate Q^* in order to obtain an efficient estimator of θ^* because, as mentioned at the beginning of Section 3.2, the EL estimator of θ^* based on the moment condition $\mathbb{E}_f\{g(Z,\theta^*)/b(Z)\} = 0$ alone will be asymptotically efficient, i.e., have asymptotic variance $(D'V_1^{-1}D)^{-1}$.

Let $1_{k\times 1}$ be the $k \times 1$ vector of ones and $\hat{Q} = (\hat{Q}_{-L}, 1 - 1'_{(L-1)\times 1}\hat{Q}_{-L})_{L\times 1}$ denote the EL estimator of Q^* for the remainder of the paper.

Example 3.3 (Example 3.1 cont.)**.** Since β^* is just identified, the EL estimators of θ^* and Q_l^* are given by $\hat{\theta} = \{\sum_{j=1}^n X_jX_j'/b(Z_j)\}^{-1}\sum_{j=1}^n X_jY_j/b(Z_j)$ and $\hat{Q}_l = (n_l/P_l)/\sum_{l=1}^L(n_l/P_l)$, respectively. By Theorem 3.1, $n^{1/2}(\hat{\theta} - \theta^*)$ is asymptotically normal with mean zero and variance covariance matrix $\{\mathbb{E}_f XX'/b(Z)\}^{-1}\mathbb{E}_f\{XX'(Y - X'\beta^*)^2/b^2(Z)\}\{\mathbb{E}_f XX'/b(Z)\}^{-1}$, which resembles the Eicker-White heteroscedasticity consistent asymptotic variance with a correction for stratification. A little simplification reveals that each $n^{1/2}(\hat{Q}_l - Q_l^*)$ is asymptotically normal with mean zero and variance $b^*\{Q_l^* - 2Q_l^{*2} + \bar{k}P_lQ_l^{*2}\}/P_l$, where $\bar{k} = \sum_{l=1}^L(Q_l^*/P_l)$. □

Let us now see how b^* can be efficiently estimated. Since $b^* = \sum_{l=1}^L P_lQ_l^*$, its EL estimator is given by $\hat{b} = \sum_{l=1}^L P_l\hat{Q}_l$. Hence, using the asymptotic distribution of $\hat{Q}_{-L}$ given in Theorem 3.1, some straightforward algebra shows that

$$n^{1/2}(\hat{b} - b^*) = n^{-1/2}\sum_{j=1}^n\{m(Z_j) - \bar{d}'M_{V_1}\rho_1(Z_j,\theta^*)\} + o_p(1),$$

where $m(Z) = b^*\{b(Z) - b^*\}/b(Z)$ and $\bar{d} = \mathbb{E}_f\{m(Z)\rho_1(Z,\theta^*)\}$. The next result is immediate.

Theorem 3.2. *Let Assumption* 3.1 *hold. Then,*

$$n^{1/2}(\hat{b} - b^*) \xrightarrow{d} \mathrm{N}(0, \mathbb{E}_f\{m^2(Z)\} - \bar{d}'M_{V_1}\bar{d}).$$

Since $\hat{b}$ is a known linear function of $\hat{Q}$ and the latter is asymptotically efficient, it follows that $\hat{b}$ is also asymptotically efficient. If there is no overidentification, its asymptotic variance becomes $b^{*2}\mathbb{E}_f\{[b(Z) - b^*]/b(Z)\}^2$. This makes sense because $\hat{Q}_l = (n_l/P_l)/\sum_{l=1}^L(n_l/P_l)$ when $q = p$ and, hence, $\hat{b} = \sum_{l=1}^L P_l\hat{Q}_l = n/\sum_{j=1}^n\{1/b(Z_j)\}$ is just the sample analog of $1/\mathbb{E}_f\{1/b(Z)\}$. Using $\hat{b}$, the asymptotic variances in Theorems 3.1–3.2 and other results can be estimated in the obvious manner by replacing population means with their sample analogs.

In addition to the aggregate shares, other unconditional probabilities can also be of interest in applied work; e.g., descriptive statistics for the target population, typically reported unconditionally, can include probabilities; for instance, estimating the proportion of individuals in the target population with 11 or fewer years of education. Hence, we next consider efficient estimation of the cumulative distribution function (cdf) $F^*(\cdot) = \Pr_{f^*}\{Z^* \le \cdot\}$. See Efromovich (2004) for some additional cross-disciplinary examples where estimation of F^* may be of interest. Efficient estimation of F^* may also be relevant if one wants to bootstrap from the

target population. When prior information about the target population is available, merely using a consistent estimator of F^* can lead to poor inference from the bootstrap. Hence, Brown and Newey (2002) suggest that resampling be done using $\hat{F}^*$, an estimator of F^* that incorporates the stochastic restrictions imposed by the model (2.1). For the sake of completeness, we also efficiently estimate $F(\cdot) = \Pr_f\{Z \leq \cdot\}$. Contrasting $\hat{F}^*$ and $\hat{F}$ (the estimator of F), a useful diagnostic tool, can reveal the extent of stratification; $\hat{F}^*$ can also be compared with the empirical distribution but since $\hat{F}$ takes the model into account, it is more precise.

So let $\hat{F}^*(\xi) = \hat{b}\sum_{j=1}^n \hat{p}_j(\hat{\beta})\mathbb{1}(Z_j \leq \xi)/b(Z_j)$ and $\hat{F}(\xi) = \sum_{j=1}^n \hat{p}_j(\hat{\beta})\mathbb{1}(Z_j \leq \xi)$, where ξ is a fixed evaluation point in $\mathbb{R}^d$. The asymptotic distributions of $\hat{F}^*(\xi)$ and $\hat{F}(\xi)$ are given by the following results.

Theorem 3.3. *Let Assumption* 3.1 *hold. Then,*

$$n^{1/2}\{\hat{F}^*(\xi) - F^*(\xi)\} \xrightarrow{d} \mathrm{N}(0, \mathbb{E}_f\{m^2(Z,\xi)\} - d_\xi' M_{V_1} d_\xi),$$

where $m(Z,\xi) = b^*\{\mathbb{1}(Z \leq \xi) - F^*(\xi)\}/b(Z)$ *and* $d_\xi = \mathbb{E}_f\{m(Z,\xi)\rho_1(Z,\theta^*)\}$.

Theorem 3.4. *Let Assumption* 3.1 *hold. Then,*

$$n^{1/2}\{\hat{F}(\xi) - F(\xi)\} \xrightarrow{d} \mathrm{N}(0, F(\xi)[1-F(\xi)] - \mathbb{E}_f\{\mathbb{1}(Z \leq \xi)\rho_1'(Z,\theta^*)\}M_{V_1}\mathbb{E}_f\{\mathbb{1}(Z \leq \xi)\rho_1(Z,\theta^*)\}).$$

The asymptotic variances in Theorems 3.3 and 3.4 correspond to the efficiency bounds for estimating $F^*(\xi)$ and $F(\xi)$; hence, these estimators are asymptotically efficient. If $f^* = f$, i.e., no stratification, then the asymptotic variances become

$$F(\xi)[1-F(\xi)] - \mathbb{E}\{\mathbb{1}(Z \leq \xi)g'(Z,\theta^*)\}\{V^{-1} - V^{-1}D(D'V^{-1}D)^{-1}D'V^{-1}\}\mathbb{E}\{\mathbb{1}(Z \leq \xi)g(Z,\theta^*)\},$$

where $D = \mathbb{E}\{\partial g(Z,\theta^*)/\partial\theta\}$ and $V = \mathbb{E}\{g(Z,\theta^*)g'(Z,\theta^*)\}$, which is the asymptotic variance for estimating $F(\xi)$ under (2.1) in the absence of stratification (Brown and Newey, 1998).

Example 3.4. Suppose we know a priori that $\mathbb{E}_{f^*}\{g(Z)\} = 0$, where g is a vector of known functions. These types of auxiliary information models, which are a special case (2.1), have been investigated by Imbens and Lancaster (1994), Hellerstein and Imbens (1999), and Nevo (2003), although these authors do not consider efficient estimation of Q^*, b^*, F^*, or F for such models. The asymptotic distributions for EL estimators of Q^*, b^*, $F^*(\xi)$, and $F(\xi)$ follow from Theorems 3.1–3.4 by replacing $g(Z,\theta^*)/b(Z)$ with $g(Z)/b(Z)$ and setting $D = 0$. □

3.3. **Hypothesis testing.** Suppose we want to test the parametric restriction $H(\theta^*) = 0$ against the alternative that it is false, where H is an $h \times 1$ vector of twice continuously differentiable known functions such that $\partial H(\theta^*)/\partial\theta$ has rank $h \leq p$. Since $\hat{\theta}$ is asymptotically normal, the Wald statistic $\hat{W} = nH'(\hat{\theta})\{[\partial H(\hat{\theta})/\partial\theta](\hat{D}'\hat{V}_1^{-1}\hat{D})^{-1}[\partial H(\hat{\theta})/\partial\theta]'\}^{-1}H(\hat{\theta})$ is asymptotically χ_h^2 under the null, where $\hat{D}$ and $\hat{V}_1$ are consistent estimators of D and V_1, respectively. Alternatively, the test can be based on the objective function itself. Letting

$\bar{\beta} = \text{argmax}_{\{\beta \in \mathcal{B}: H(\theta)=0\}} \text{EL}(\beta)$ denote the restricted estimator, define the likelihood ratio statistic $\text{LR} = 2\{\text{EL}(\hat{\beta}) - \text{EL}(\bar{\beta})\}$. A test for $H(\theta^*) = 0$ can be based upon LR; critical values follow from Qin and Lawless (1994, Theorem 2) who show that $\text{LR} \xrightarrow{d} \chi^2_h$ under the null.

Since $\hat{W}$ and LR are asymptotically equivalent, the decision to use a particular test depends upon computational and other considerations; e.g., though both can be inverted to obtain asymptotically valid confidence regions, LR based regions are invariant to the formulation of the null hypothesis and automatically satisfy natural range restrictions. Furthermore, unlike $\hat{W}$, the likelihood ratio statistic LR is internally studentized, i.e., it does not require preliminary estimation of any variance terms. This guarantees that confidence regions based on LR are also invariant to nonsingular transformations of the moment conditions. Internal studentization may also lead to better finite sample properties for LR; see, e.g., Fisher, Hall, Jing, and Wood (1996).

3.4. **Specification testing.** Assume that $q > p$. In this section we describe an EL based specification test of (2.1) against the alternative that it is false. Besides being internally studentized and invariant to nonsingular and algebraic transformations of the moment conditions, Kitamura (2001) has shown this test to be optimal in terms of a large deviations criterion. So let $\hat{\beta}$ denote a $n^{1/2}$-consistent preliminary estimator of β; e.g., $\hat{\beta}$ can be the EL estimator defined previously. The restricted, i.e., under (2.1), EL is $\text{EL}^r = \sum_{j=1}^n \log \hat{p}_j(\hat{\beta})$, where $\hat{p}_j$'s are the EL probabilities; the unrestricted, i.e., when the model is not imposed, nonparametric likelihood is $\text{EL}^{ur} = -n \log n$. Now define $\text{ELR} = 2(\text{EL}^{ur} - \text{EL}^r) = 2\sum_{j=1}^n \log\{1 + \lambda'(\hat{\beta})\rho(Z_j, \hat{\beta})\}$, where $\lambda(\beta)$ was defined earlier in Section 3.2. ELR can be regarded as an analog of the usual parametric likelihood ratio test statistic; i.e., (2.1) is rejected if ELR is large enough. Critical values for ELR are easily obtained because $\text{ELR} \xrightarrow{d} \chi^2_{q-p}$ under (2.1) by Qin and Lawless (1994, Corollary 4).

4. Inference when data is collected by standard stratified sampling

We now consider the estimation and testing of (2.1) using data collected by SS sampling. As shown subsequently, the major difference between the VP and SS sampling schemes is that the unknown aggregate shares create a lack of identification for the target density when data is collected by SS sampling.

4.1. **Identification.** Although we can write $f^*(z) = f(z)/b(z, Q^*, \tilde{K})$ by (2.3), we cannot recover f^* in terms of f alone because, unlike VP sampling, data collected by SS sampling cannot identify the aggregate shares Q^*.[2] Therefore, the target density is also unidentified. To overcome this lack of identification, suppose that in addition to the stratified sample we also have some additional observations that were collected by random sampling. Since the second sample is not stratified, we can use it to recover the aggregate shares and, as shown later, combining the stratified and random samples allows us to completely recover f^*.

The existence of such additional random samples should not be regarded as being an overly restrictive requirement. For instance, Manski and Lerman (1977) suggest carrying out a small random survey to gather a supplementary sample in order to estimate the aggregate shares. Indeed, some widely used stratified datasets such as the Panel Study of Income Dynamics (PSID) and the National Longitudinal Survey (NLS) automatically provide an additional random sample that can be used for this purpose.

4.2. **Data combination.** As in Devereux and Tripathi (2006), the process of combining the stratified and random samples is modelled as follows. Let Z denote an observation from the combined sample. Along with Z, we observe a dummy variable R that indicates whether Z comes from the random or the stratified sample; i.e., $R = 1$ if Z is from the random sample and $R = 0$ if Z belongs to the stratified sample. Hence, for $r \in \{0, 1\}$, the conditional density of $Z|R = r$ is given by

$$f_{Z|R=r}(z) = f^*(z)r + f(z)(1 - r), \tag{4.1}$$

where f is defined in (2.3). Next, since R is a binary random variable, assume that $R \overset{d}{=} \text{Bernoulli}(\kappa_0)$, where $\kappa_0 \in (0, 1)$ is an unknown nuisance parameter that will be estimated along with the parameters of interest. Therefore, by (4.1), the joint density of Z and R is

$$f_e(z, r) = \kappa_0 f^*(z)r + (1 - \kappa_0)f(z)(1 - r). \tag{4.2}$$

Henceforth, we refer to f_e as the density of an observation from the "enriched" sample, i.e., the random and stratified samples combined together. f_e is a density with respect to the dominating measure $\mu \otimes \bar{c}$, where $\bar{c}$ denotes the counting measure on $\{0, 1\}$.

To see how combining the datasets identifies f^*, note that by (2.3) and (4.2) we have

$$f^*(z) = \frac{\int_{r\in\{0,1\}} f_e(z, r)\, d\bar{c}}{\kappa_0 + (1 - \kappa_0)b(z, Q^*, \tilde{K})}. \tag{4.3}$$

But the aggregate shares are identified from the random sample by the moment conditions

$$Q_l^* = \mathbb{E}_{f_e}\{\mathbb{1}(Z \in \mathbb{C}_l)|R = 1\} \iff \mathbb{E}_{f_e}\{\mathbb{1}(Z \in \mathbb{C}_l) - Q_l^*\}R = 0 \tag{4.4}$$

for each l. Similarly, the $\tilde{K}_l$'s, which were also assumed to be unknown, are identified from the stratified sample via the moment conditions

$$\tilde{K}_l = \mathbb{E}_{f_e}\{\mathbb{1}(Z \in \mathbb{C}_l)|R = 0\} \iff \mathbb{E}_{f_e}\{\mathbb{1}(Z \in \mathbb{C}_l) - \tilde{K}_l\}(1 - R) = 0, \tag{4.5}$$

and κ_0, which can be loosely described as the probability of randomly sampling from the target population, is identified via the moment condition

$$\kappa_0 = \mathbb{E}_{f_e}\{R\} \iff \mathbb{E}_{f_e}\{R - \kappa_0\} = 0. \tag{4.6}$$

Since (4.4)–(4.6) imply that (4.3) can be written in terms of f_e alone, it follows that the target density can be fully recovered from the enriched density and is, therefore, identified.

For the remainder of Section 4, let n denote the size of the enriched sample. Observations $(Z_1, R_1), \ldots, (Z_n, R_n)$ from the enriched dataset are regarded as iid draws from f_e and all limits are taken as the combined sample size n approaches infinity. In the next section we show how the enriched dataset can be used to estimate and test (2.1).

We end this section with a brief technical remark: Although the introduction of R allows the combined sample to be treated as a collection of iid draws from the enriched density f_e, which greatly simplifies the mathematical treatment, it makes $\sum_{j=1}^n R_j$, the size of the randomly sampled dataset, a random variable. However, as shown in Section 4.3, asymptotic inference about θ^* is conditional on the observed value of $\sum_{j=1}^n R_j$ because we estimate θ^* jointly and efficiently with κ_0. Therefore, our results coincide with those obtained in a setting where the size of the random sample is non-stochastic and observations from the combined sample are regarded as being independently but not identically distributed.

4.3. **Efficient estimation and inference.** Recalling that the aggregate shares and the $\tilde{K}_l$'s sum to one, by (4.3) we can express (2.1) in terms of the enriched density as

$$\mathbb{E}_{f_e}\{g(Z, \theta^*)/c(Z, Q^*_{-L}, \tilde{K}_{-L}, \kappa_0)\} = 0, \tag{4.7}$$

where $\tilde{K}_{-L} = (\tilde{K}_1, \ldots, \tilde{K}_{L-1})_{(L-1)\times 1}$ and $c(Z, Q^*_{-L}, \tilde{K}_{-L}, \kappa_0) = \kappa_0 + (1 - \kappa_0)b(Z, Q^*, \tilde{K})$.

To estimate $\beta^* = (\theta^*, Q^*_{-L}, \tilde{K}_{-L}, \kappa_0)_{(p+2L-1)\times 1}$, use (4.4)–(4.7) to define the $(q+2L-1)\times 1$ moment function

$$\rho(Z, R, \beta) = \begin{bmatrix} g(Z, \theta)/c(Z, Q_{-L}, K_{-L}, \kappa) \\ \{s(Z) - Q_{-L}\}R \\ \{s(Z) - K_{-L}\}(1 - R) \\ R - \kappa \end{bmatrix} \stackrel{\text{def}}{=} \begin{bmatrix} \rho_1(Z, \beta) \\ \rho_2(Z, R, Q_{-L}) \\ \rho_3(Z, R, K_{-L}) \\ \rho_4(R, \kappa) \end{bmatrix}, \tag{4.8}$$

where $s(Z)$ was defined earlier in Section 3.2, $\rho_1(Z, \beta^*)$ is the moment function in (4.7), $\rho_2(Z, R, Q_{-L}) = \{s(Z) - Q_{-L}\}R$, $\rho_3(Z, R, K_{-L}) = \{s(Z) - K_{-L}\}(1 - R)$, and $\rho_4(R, \kappa) = R - \kappa$. Since ρ_2, ρ_3, and ρ_4 just identify $(Q^*_{-L}, \tilde{K}_{-L}, \kappa_0)$, it follows that (2.1) holds if and only if $\mathbb{E}_{f_e}\{\rho(Z, R, \beta^*)\} = 0$. Hence, θ^* can be efficiently estimated from the latter moment condition. Using notation developed earlier, the EL estimator of β^* is given by $\hat{\beta} = \operatorname{argmax}_{\beta\in\mathcal{B}} \mathrm{EL}(\beta)$, where $\mathcal{B} = \Theta \times [0, 1]^{L-1} \times [0, 1]^{L-1} \times [0, 1]$ and the objective function $\mathrm{EL}(\beta)$ is defined as in (3.3) with the moment function given in (4.8).

We need some additional notation to describe the asymptotic distribution of $\hat{\beta}$. So let $\mathrm{Proj}\{\rho_1(Z, \beta^*)\big|1, \rho_2(Z, R, Q^*_{-L}), \rho_3(Z, R, \tilde{K}_{-L}), \rho_4(R, \kappa_0)\}$ denote the orthogonal projection of $\rho_1(Z, \beta^*)$ onto the span of $\{1, \rho_2(Z, R, Q^*_{-L}), \rho_3(Z, R, \tilde{K}_{-L}), \rho_4(R, \kappa_0)\}$ using the inner product $\langle a, b\rangle = \mathbb{E}_{f_e}\{a'b\}$, and let ε be the residual from this projection; i.e.,

$$\varepsilon = \rho_1(Z, \beta^*) - \mathrm{Proj}\{\rho_1(Z, \beta^*)\big|1, \rho_2(Z, R, Q^*_{-L}), \rho_3(Z, R, \tilde{K}_{-L}), \rho_4(R, \kappa_0)\}.$$

Since $\rho_2(Z, R, Q^*_{-L})$, $\rho_3(Z, R, \tilde{K}_{-L})$, and $\rho_4(R, \kappa_0)$ are mean zero and mutually orthogonal,

$$\varepsilon = \rho_1(Z, \beta^*) - \Sigma_{12}V_2^{-1}\rho_2(Z, R, Q^*_{-L}) - \Sigma_{13}V_3^{-1}\rho_3(Z, R, \tilde{K}_{-L}) - \Sigma_{14}\rho_4(R, \kappa_0)/V_4,$$

where, as in Section 3.2, $\Sigma_{12} = \mathbb{E}_{f_e}\{\rho_1(Z,\beta^*)\rho_2'(Z,R,Q_{-L}^*)\}$, $\Sigma_{13} = \mathbb{E}_{f_e}\{\rho_1(Z,\beta^*)\rho_3'(Z,R,\tilde{K}_{-L})\}$, $\Sigma_{14} = \mathbb{E}_{f_e}\{\rho_1(Z,\beta^*)\rho_4(R,\kappa_0)\}$, $V_2 = \mathbb{E}_{f_e}\{\rho_2(Z,R,Q_{-L}^*)\rho_2'(Z,R,Q_{-L}^*)\}$, $V_4 = \mathbb{E}_{f_e}\{\rho_4^2(R,\kappa_0)\}$, and $V_3 = \mathbb{E}_{f_e}\{\rho_3(Z,R,\tilde{K}_{-L})\rho_3'(Z,R,\tilde{K}_{-L})\}$.

Next, define $J = \Sigma_{12}V_2^{-1} + (1/\kappa_0)\mathbb{E}_{f_e}\{\partial\rho_1(Z,\beta^*)/\partial Q_{-L}\}$ and $V = \mathbb{E}_{f_e}\{vv'\}$, where $v = \varepsilon + J\rho_2(Z,R,Q_{-L}^*)$. Letting $D = \mathbb{E}_{f_e}\{\partial\rho_1(Z,\beta^*)/\partial\theta\}$ and $M_V = V^{-1} - V^{-1}D(D'V^{-1}D)^{-1}D'V^{-1}$, we have the following result.

Theorem 4.1. *Let Assumption* 3.1 *hold with the moment function* $\rho(Z,R,\beta^*)$ *defined in* (4.8) *and expectations with respect to* f_e. *Then,* $n^{1/2}(\hat{\theta}-\theta^*)$, $n^{1/2}(\hat{Q}_{-L}-Q_{-L}^*)$, $n^{1/2}(\hat{K}_{-L}-\tilde{K}_{-L})$, *and* $n^{1/2}(\hat{\kappa}-\kappa_0)$ *converge jointly in distribution to a* $(p+2L-1)\times 1$ *normal random vector with mean zero and variance-covariance matrix*

$$\begin{bmatrix} (D'V^{-1}D)^{-1} & -(D'V^{-1}D)^{-1}D'V^{-1}JV_2/\kappa_0 & 0_{p\times(L-1)} & 0_{p\times 1} \\ -V_2J'V^{-1}D(D'V^{-1}D)^{-1}/\kappa_0 & (V_2 - V_2J'M_VJV_2)/\kappa_0^2 & 0_{(L-1)\times(L-1)} & 0_{(L-1)\times 1} \\ 0'_{p\times(L-1)} & 0_{(L-1)\times(L-1)} & V_3/(1-\kappa_0)^2 & 0_{(L-1)\times 1} \\ 0'_{p\times 1} & 0'_{(L-1)\times 1} & 0'_{(L-1)\times 1} & \kappa_0(1-\kappa_0) \end{bmatrix}.$$

As shown in Appendix B, $(D'V^{-1}D)^{-1}$ coincides with the efficiency bound for estimating θ^*; hence, $\hat{\theta}$ is asymptotically efficient. Following the proof of Theorem 4.1, we can also show that $\hat{\theta}$ is asymptotically linear with influence function $-(D'V^{-1}D)^{-1}D'V^{-1}v$. But since v is orthogonal to $\rho_3(Z,R,\tilde{K}_{-L})$ and $\rho_4(R,\kappa_0)$, an application of the Cramér-Wold device and the central limit theorem immediately reveals that $\hat{\theta}$ is asymptotically independent of $\sum_{j=1}^n s(Z_j)(1-R_j)$ and $\sum_{j=1}^n R_j$. Therefore, as emphasized earlier in Sections 2.2 and 4.2, inference using the asymptotic distribution of $\hat{\theta}$ is equivalent to inference based on the asymptotic distribution of $\hat{\theta}$ conditional on the number of observations lying in each stratum of the stratified sample and the size of the random sample.

Asymptotic efficiency of $\hat{Q}_{-L}$ is demonstrated in Appendix B; similarly, we can also show that $\hat{K}_{-L}$ and $\hat{\kappa}$ are asymptotically efficient. Since the aggregate shares are estimated from the random sample alone, the asymptotic variance of $n^{1/2}(\hat{Q}_{-L}-Q_{-L}^*)$ when there is no overidentification is given by V_2/κ_0^2; as expected, overidentification of θ^* leads to a better estimator of Q^*.

Using the definitions of v and ε, it immediately follows that $V = \Omega + JV_2J'$, where

$$\Omega \stackrel{\text{def}}{=} \mathbb{E}_{f_e}\{\varepsilon\varepsilon'\} = V_1 - \Sigma_{12}V_2^{-1}\Sigma_{12}' - \Sigma_{13}V_3^{-1}\Sigma_{13}' - \Sigma_{14}\Sigma_{14}'/V_4 \tag{4.9}$$

and $V_1 = \mathbb{E}_{f_e}\{\rho_1(Z,\beta^*)\rho_1'(Z,\beta^*)\}$. Hence, the asymptotic variances can be estimated as before by replacing population expectations with their sample analogs.

For the remainder of the paper, let $\gamma^* = (Q_{-L}^*, \tilde{K}_{-L}, \kappa_0)_{(2L-1)\times 1}$ and $\hat{\gamma} = (\hat{Q}_{-L}, \hat{K}_{-L}, \hat{\kappa})$.

Example 4.1 (Population mean)**.** Suppose we want to estimate θ^*, the mean of the target population. Since θ^* is just identified and $\sum_{j=1}^n 1/c(Z_j,\hat{\gamma}) = n$,[3] we have $\hat{\theta} = n^{-1}\sum_{j=1}^n Z_j/c(Z_j,\hat{\gamma})$, where $\hat{Q}_l = \sum_{j=1}^n \mathbb{1}(Z_j \in \mathbb{C}_l)R_j/\sum_{j=1}^n R_j$ is the fraction of observations lying in the l^{th} stratum

of the random sample, $\hat{K}_l = \sum_{j=1}^n \mathbb{1}(Z_j \in \mathbb{C}_l)(1 - R_j)/\sum_{j=1}^n (1 - R_j)$ the fraction of observations in the l^{th} stratum of the stratified sample, and $\hat{\kappa} = \sum_{j=1}^n R_j/n$ the size of the random sample relative to the enriched sample. As in Example 3.2, a little algebra shows that we can express $\hat{\theta}$ more intuitively as $\hat{\theta} = \sum_{l=1}^L \hat{Q}_l \bar{Z}_l$. The asymptotic distribution of $\hat{\theta}$ follows from Theorem 4.1 upon noting that D is the $p \times p$ identity matrix. □

Example 4.2 (Linear regression)**.** For the model in Example 3.1, assume that Z and R are drawn from the enriched density f_e defined in (4.2). Since θ^* is again just identified, $\hat{\theta} = \{\sum_{j=1}^n X_j X_j'/c(Z_j, \hat{\gamma})\}^{-1} \sum_{j=1}^n X_j Y_j/c(Z_j, \hat{\gamma})$ with $\hat{Q}$, $\hat{K}$, and $\hat{\kappa}$ as in the previous example. By Theorem 4.1, $n^{1/2}(\hat{\theta} - \theta^*)$ is asymptotically normal with mean zero and variance $\{\mathbb{E}_{f_e} XX'/c(Z, \gamma^*)\}^{-1} V \mathbb{E}_{f_e}\{XX'/c(Z, \gamma^*)\}$. □

We now show that, even asymptotically, it never makes sense to throw away data and use only the randomly sampled dataset to estimate θ^*. So let $\hat{\theta}_R$ denote the EL estimator of θ^* obtained using only the random sample; i.e., $\hat{\theta}_R$ is based on the moment condition

$$\mathbb{E}_{f_e}\{g(Z, \theta^*)|R = 1\} = 0 \iff \mathbb{E}_{f_e}\{g(Z, \theta^*)R\} = 0. \tag{4.10}$$

The next result demonstrates that $\hat{\theta}_R$ is asymptotically inefficient relative to $\hat{\theta}$. Therefore, θ^* should be estimated using the enriched dataset and not just the random sample alone.

Theorem 4.2. *Let* $D_* = \mathbb{E}_{f^*}\{\partial g(Z^*, \theta^*)/\partial\theta\}$ *and* $V_* = \mathbb{E}_{f^*}\{g(Z^*, \theta^*)g'(Z^*, \theta^*)\}$. *Then,* (i) $n^{1/2}(\hat{\theta}_R - \theta^*)$ *is asymptotically normal with mean zero and variance* $(D_*' V_*^{-1} D_*)^{-1}/\kappa_0$*; and* (ii) $(D_*' V_*^{-1} D_*)^{-1}/\kappa_0 > (D' V^{-1} D)^{-1}$*, i.e.,* $(D_*' V_*^{-1} D_*)^{-1}/\kappa_0 - (D' V^{-1} D)^{-1}$ *is positive definite.*

The inflation factor $1/\kappa_0$ appears in the asymptotic variance of $\hat{\theta}_R$ because it only makes use of a fraction of the enriched sample. As stressed earlier, (ii) makes clear the penalty for throwing away data.

Next, let $\hat{F}^*(\xi) = \sum_{j=1}^n \hat{p}_j(\hat{\beta})\mathbb{1}(Z_j \le \xi)/c(Z_j, \hat{\gamma})$ and $\hat{F}_e(\xi) = \sum_{j=1}^n \hat{p}_j(\hat{\beta})\mathbb{1}(Z_j \le \xi)$ denote estimators of the target cdf $F^*(\xi)$ and the enriched cdf $F_e(\xi)$, respectively, where $\hat{p}_j$'s are the EL probabilities. Also, define $I_c(Z, \xi) = \{\mathbb{1}(Z \le \xi) - F^*(\xi)\}/c(Z, \gamma^*)$,

$$u = I_c(Z, \xi) - \text{Proj}\{I_c(Z, \xi)\big|1, \rho_2(Z, R, Q_{-L}^*), \rho_3(Z, R, \tilde{K}_{-L}), \rho_4(R, \kappa_0)\},$$

and $\tilde{J}' = \mathbb{E}_{f_e}\{I_c(Z, \xi)\rho_2'(Z, R, Q_{-L}^*)\}V_2^{-1} + (1/\kappa_0)\mathbb{E}_{f_e}\{\partial I_c(Z, \xi)/\partial Q_{-L}\}$. The asymptotic distributions of $\hat{F}^*(\xi)$ and $\hat{F}_e(\xi)$ are given below.

Theorem 4.3. *Let Assumption* 3.1 *hold with the moment function* $\rho(Z, R, \beta^*)$ *defined in* (4.8) *and expectations with respect to* f_e*. Then, letting* $w = u + \tilde{J}'\rho_2(Z, R, Q_{-L}^*)$*,*

$$n^{1/2}\{\hat{F}^*(\xi) - F^*(\xi)\} \xrightarrow{d} \text{N}(0, \mathbb{E}_{f_e}\{w^2\} - \mathbb{E}_{f_e}\{wv'\}M_V\mathbb{E}_{f_e}\{wv\}).$$

Theorem 4.4. *Let Assumption* 3.1 *hold with the moment function* $\rho(Z, R, \beta^*)$ *defined in* (4.8) *and expectations with respect to* f_e*. Then,*

$$n^{1/2}\{\hat{F}_e(\xi) - F_e(\xi)\} \xrightarrow{d} \text{N}(0, F_e(\xi)[1 - F_e(\xi)] - \mathbb{E}_{f_e}\{\mathbb{1}(Z \le \xi)v'\}M_V\mathbb{E}_{f_e}\{\mathbb{1}(Z \le \xi)v\}).$$

Theorems 4.3 and 4.4 again reveal that imposing the overidentified model leads to an efficiency gain in estimating F^* and F_e. The efficiency bounds derived in Appendix B show that $\hat{F}^*$ and $\hat{F}_e$ are asymptotically efficient.

Hypotheses of the form $H(\theta^*) = 0$ can be tested using the Wald or LR statistics as described in Section 3.3 by basing the test on (4.8); in each case, the test statistic is asymptotically distributed as a χ^2_h random variable under the null hypothesis. If $q > p$, then EL based specification testing of (2.1) can also be done using (4.8), the details being analogous to those in Section 3.4; i.e., the test statistic is asymptotically χ^2_{q-p} under (2.1).

5. Conclusion

This paper develops efficient empirical likelihood based inference for moment restriction models using stratified datasets. Since the aggregate shares are assumed to be unknown, the target density is unidentified when data is collected by standard stratified (but not variable probability) sampling, a problem we overcome by combining the original stratified sample with an additional random sample in an optimal manner. We show that correcting for the effects of stratification is straightforward; namely, an appropriate transformation of the moment conditions ensures that all standard empirical likelihood based inference goes through. No special software is required to implement the procedures developed in this paper; any computer package that can do empirical likelihood based estimation and testing will be able to do the same with stratified data.

Notes

[1]The M-estimators in Wooldridge (1999) can be motivated in a similar manner. Suppose that θ^* is identified as $\theta^* = \text{argmin}_{\theta\in\Theta}\, \mathbb{E}_{f^*}\{\psi(Z,\theta)\}$, where ψ is a real-valued objective function. Since $\mathbb{E}_{f^*}\{\psi(Z,\theta)\} = b^*\mathbb{E}_f\{\psi(Z,\theta)/b(Z)\}$ and b^* does not depend upon θ^*, it follows that $\theta^* = \text{argmin}_{\theta\in\Theta}\, \mathbb{E}_f\{\psi(Z,\theta)/b(Z)\}$. Hence, the M-estimator of θ^* is given by $\hat{\theta}_M = \text{argmin}_{\theta\in\Theta} \sum_{j=1}^n \psi(Z_j,\theta)/b(Z_j)$. A similar argument works for SS sampling when the aggregate shares are assumed known as in Wooldridge (2001).

[2]By (2.3), we have $\mathbb{E}_f\{\mathbb{1}(Z \in \mathbb{C}_l)\} = \tilde{K}_l$ for each l; hence, with SS sampling we can only recover the sampling fractions, not the aggregate shares, from the stratified sample.

[3]In the proof of Theorem 4.3 we show that $\sum_{j=1}^n \hat{p}_j(\hat{\beta})/c(Z_j,\hat{\gamma}) = 1$. But when θ^* is just identified, $\hat{p}_j(\beta) = 1/n$ for each j and β. Hence, $\sum_{j=1}^n 1/c(Z_j,\hat{\gamma}) = n$ whenever $q = p$.

References

Bickel, P., C. Klassen, Y. Ritov, and J. Wellner (1993): *Efficient and adaptive estimation for semiparametric models*, Johns Hopkins Press.

Bickel, P. J. and J. Ritov (1991): "Large sample theory of estimation in biased sampling regression models," *The Annals of Statistics*, 19, 797–816.

Brown, B. W. and W. K. Newey (1998): "Efficient semiparametric estimation of expectations," *Econometrica*, 66, 453–464.

——— (2002): "GMM, efficient bootstrapping, and improved inference," *Journal of Business and Economic Statistics*, 20, 507–517.

Butler, J. (2000): "Efficiency results of MLE and GMM estimation with sampling weights," *Journal of Econometrics*, 96, 25–37.

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, 1–43.

DeMets, D. and M. Halperin (1977): "Estimation of a simple regression coefficient in samples arising from a subsampling procedure," *Biometrics*, 33, 47–56.

Devereux, P. and G. Tripathi (2006): "Optimally combining censored and uncensored datasets," Manuscript. Department of Economics, University of Connecticut-Storrs.

DuMouchel, W. H. and G. J. Duncan (1983): "Using sample survey weights in multiple regression analysis of stratified samples," *Journal of the American Statistical Association*, 78, 535–543.

Efromovich, S. (2004): "Distribution estimation for biased data," *Journal of Statistical Planning and Inference*, 124, 1–43.

El-Barmi, H. and M. Rothmann (1998): "Nonparametric estimation in selection biased models in the presence of estimating equations," *Nonparametric Statistics*, 9, 381–399.

Fisher, N. I., P. Hall, B.-Y. Jing, and A. T. Wood (1996): "Improved pivotal methods for constructing confidence regions with directional data," *Journal of the American Statistical Association*, 91, 1062–1070.

Harville, D. A. (1997): *Matrix algebra from a statistician's perspective*, Springer-Verlag.

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.

Hellerstein, J. and G. W. Imbens (1999): "Imposing moment restrictions from auxiliary data by weighting," *Review of Economics and Statistics*, 81, 1–14.

Holt, D., T. Smith, and P. Winter (1980): "Regression analysis of data from complex surveys," *Journal of The Royal Statistical Society, Series A*, 143, 474–487.

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

——— (1997): "One-step estimators for over-identified generalized method of moments models," *Review of Economic Studies*, 64, 359–383.

Imbens, G. W. and T. Lancaster (1994): "Combining micro and macro data in microeconomic models," *Review of Economic Studies*, 61, 655–680.

——— (1996): "Efficient estimation and stratified sampling," *Journal of Econometrics*, 74, 289–318.

Imbens, G. W., R. H. Spady, and P. Johnson (1998): "Information theoretic approaches to inference in moment condition models," *Econometrica*, 66, 333–357.

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

Kitamura, Y. (1997): "Empirical likelihood methods with weakly dependent processes," *Annals of Statistics*, 25, 2084–2102.

——— (2001): "Asymptotic optimality of empirical likelihood for testing moment restrictions," *Econometrica*, 69, 1661–1672.

——— (2006): "Empirical Likelihood Methods in Econometrics: Theory and Practice," Invited symposium on Weak Instruments and Empirical Likelihood at the 9th World Congress of the Econometric Society.

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.

Nevo, A. (2003): "Using weights to adjust for sample selection when auxiliary information is available," *Journal of Business and Economic Statistics*, 21, 43–52.

Newey, W. K. (1990): "Semiparametric efficiency bounds," *Journal of Applied Econometrics*, 5, 99–135.

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.

Newey, W. K. and R. J. Smith (2004): "Higher order properties of GMM and generalized empirical likelihood estimators," *Econometrica*, 72, 219–255.

Owen, A. (1988): "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, 75, 237–249.

——— (2001): *Empirical likelihood*, Chapman and Hall/CRC.

Qin, J. (1993): "Empirical likelihood in biased sample problems," *Annals of Statistics*, 21, 1182–1196.

Qin, J. and J. Lawless (1994): "Empirical likelihood and general estimating equations," *Annals of Statistics*, 22, 300–325.

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

Scott, A. and C. Wild (1986): "Fitting logistic models under case-control or choice based sampling," *Journal of The Royal Statistical Society, Series B*, 48, 170–182.

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

Smith, R. J. (1997): "Alternative semi-parametric likelihood approaches to generalized method of moments estimation," *Economic Journal*, 107, 503–519.

——— (2005): "Weak Instruments and Empirical Likelihood: A Discussion of the Papers by D.W.K. Andrews and J. H. Stock and Y. Kitamura," Invited discussion of the symposium on Weak Instruments and Empirical Likelihood at the 9th World Congress of the Econometric Society.

Tripathi, G. (2002): "Inference in conditional moment restriction models when there is selection due to stratification," Manuscript. Department of Economics, University of Wisconsin-Madison.

Wooldridge, J. M. (1999): "Asymptotic properties of weighted M-estimators for variable probability samples," *Econometrica*, 67, 1385–1406.

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

Appendix A. Proofs

We only provide proofs for the results in Section 4 because SS sampling is the hardest to handle. Results for VP sampling described in Section 3 can be shown in a similar manner.

In addition to the earlier notation, let $\mathcal{Q} = \text{diag}(Q_1^*, \ldots, Q_{L-1}^*)$, $\mathcal{K} = \text{diag}(\tilde{K}_1, \ldots, \tilde{K}_{L-1})$, and $\mathcal{A} = \text{diag}(\alpha_1^*, \ldots, \alpha_{L-1}^*)$ be $(L-1)\times(L-1)$ diagonal matrices, where $\alpha_l^* = \kappa_0 Q_l^* + (1-\kappa_0)\tilde{K}_l$, and $I_{k\times k}$ denote the $k \times k$ identity matrix.

Proof of Theorem 4.1. From standard EL theory we know that $n^{1/2}(\hat{\beta}-\beta^*)$ is asymptotically normal with mean zero and variance $(D_{f_e}' V_{f_e}^{-1} D_{f_e})^{-1}$, where $D_{f_e} = \mathbb{E}_{f_e}\{\partial\rho(Z,R,\beta^*)/\partial\beta\}$ and $V_{f_e} = \mathbb{E}_{f_e}\{\rho(Z,R,\beta^*)\rho'(Z,R,\beta^*)\}$. Letting $\Sigma = [\,\Sigma_{12}\ \Sigma_{13}\ \Sigma_{14}\,]$, we can write

$$V_{f_e} = \begin{bmatrix} V_1 & \Sigma \\ \Sigma' & V_{-1} \end{bmatrix}, \quad \text{where} \quad V_{-1} = \begin{bmatrix} V_2 & 0_{(L-1)\times(L-1)} & 0_{(L-1)\times 1} \\ 0_{(L-1)\times(L-1)} & V_3 & 0_{(L-1)\times 1} \\ 0_{(L-1)\times 1}' & 0_{(L-1)\times 1}' & V_4 \end{bmatrix}.$$

Next, by Lemma C.1 and the partitioned inverse formula,

$$D_{f_e} = \begin{bmatrix} D & A \\ 0_{(2L-1)\times p} & B \end{bmatrix} \quad \text{and} \quad V_{f_e}^{-1} = \begin{bmatrix} \Omega^{-1} & -\Omega^{-1}\Sigma V_{-1}^{-1} \\ -V_{-1}^{-1}\Sigma'\Omega^{-1} & V_{-1}^{-1} + V_{-1}^{-1}\Sigma'\Omega^{-1}\Sigma V_{-1}^{-1} \end{bmatrix}.$$

Hence, some straightforward matrix algebra reveals that

$$D'_{f_e}V_{f_e}^{-1}D_{f_e} = \begin{bmatrix} D'\Omega^{-1}D & \kappa_0 D'\Omega^{-1}J & 0_{p\times(L-1)} & 0_{p\times 1} \\ \kappa_0 J'\Omega^{-1}D & \kappa_0^2(J'\Omega^{-1}J + V_2^{-1}) & 0_{(L-1)\times(L-1)} & 0_{(L-1)\times 1} \\ 0'_{p\times(L-1)} & 0_{(L-1)\times(L-1)} & (1-\kappa_0)^2 V_3^{-1} & 0_{(L-1)\times 1} \\ 0'_{p\times 1} & 0'_{(L-1)\times 1} & 0'_{(L-1)\times 1} & 1/V_4 \end{bmatrix}.$$

The desired result now follows by applying the partitioned inverse formula and using Woodbury's formula, see, e.g., Harville (1997, Page 424), to simplify the resulting terms. □

Proof of Theorem 4.2. Since $\hat{\theta}_R$ is based on the moment condition $\mathbb{E}_{f_e}\{g(Z,\theta^*)R\} = 0$, we know that $n^{1/2}(\hat{\theta}_R - \theta^*)$ is asymptotically normal with mean zero and variance $(D'_R V_R^{-1} D_R)^{-1}$, where $D_R = \mathbb{E}_{f_e}\{\partial g(Z,\theta^*)R/\partial\theta\}$ and $V_R = \mathbb{E}_{f_e}\{g(Z,\theta^*)g'(Z,\theta^*)R\}$. But

$$D_R = \kappa_0 \mathbb{E}_{f_e}\{\partial g(Z,\theta^*)/\partial\theta | R = 1\} \overset{(4.1)}{=} \kappa_0 \mathbb{E}_{f^*}\{\partial g(Z^*,\theta^*)/\partial\theta\} = \kappa_0 D_*,$$

and, in a similar manner, $V_R = \kappa_0 V_*$. Hence, $(D'_R V_R^{-1} D_R)^{-1} = (D'_* V_*^{-1} D_*)^{-1}/\kappa_0$ and (i) follows. Next, since $D_* = D$ by (4.3), to prove (ii) it suffices to show that $V_*/\kappa_0 - V$ is positive definite. We proceed as follows. First, using (4.3) to further simplify (C.2), we can show that

$$\Sigma_{12} = \mathbb{E}_{f^*}\{g(Z^*,\theta^*)s'(Z^*)\}\{\mathcal{A}^{-1} + (1/\alpha_L^*)1_{(L-1)\times 1}1'_{(L-1)\times 1}\}V_2.$$

Similarly, we can also show that

$$\mathbb{E}_{f_e}\{\partial\rho_1(Z,\beta^*)/\partial Q_{-L}\} = (1-\kappa_0)\mathbb{E}_{f^*}\{g(Z^*,\theta^*)s'(Z^*)\}\{\mathcal{K}\mathcal{Q}^{-1}\mathcal{A}^{-1} + \frac{\tilde{K}_L}{Q_L^*\alpha_L^*}1_{(L-1)\times 1}1'_{(L-1)\times 1}\}.$$

Using these results, some straightforward algebra reveals that

$$\begin{aligned} J &= \mathbb{E}_{f^*}\{g(Z^*,\theta^*)s'(Z^*)\}V_2^{-1}, \\ \Sigma_{13} &= \mathbb{E}_{f^*}\{g(Z^*,\theta^*)s'(Z^*)\}\{\mathcal{A}^{-1} + (1/\alpha_L^*)1_{(L-1)\times 1}1'_{(L-1)\times 1}\}V_3, \\ \Sigma_{14} &= V_4\mathbb{E}_{f^*}\{g(Z^*,\theta^*)s'(Z^*)\}\{\mathcal{A}^{-1} + (1/\alpha_L^*)1_{(L-1)\times 1}1'_{(L-1)\times 1}\}(Q_{-L}^* - \tilde{K}_{-L}). \end{aligned}$$

Hence, by (4.9), we can write

$$V = V_1 - \mathbb{E}_{f^*}\{g(Z^*,\theta^*)s'(Z^*)\}\Delta\mathbb{E}_{f^*}\{s(Z^*)g'(Z^*,\theta^*)\}, \tag{A.1}$$

where

$$\begin{aligned} \Delta = \{\mathcal{A}^{-1} + (1/\alpha_L^*)1_{(L-1)\times 1}1'_{(L-1)\times 1}\}\{V_2 + V_3 + V_4(Q_{-L}^* - \tilde{K}_{-L})(Q_{-L}^* - \tilde{K}_{-L})'\} \\ \times \{\mathcal{A}^{-1} + (1/\alpha_L^*)1_{(L-1)\times 1}1'_{(L-1)\times 1}\} - V_2^{-1}. \end{aligned}$$

Further calculations show that Δ can be expressed in a compact manner as

$$\Delta = -\kappa_0^{-1}(1-\kappa_0)\{\mathcal{K}\mathcal{Q}^{-1}\mathcal{A}^{-1} + \frac{\tilde{K}_L}{Q_L^*\alpha_L^*}1_{(L-1)\times 1}1'_{(L-1)\times 1}\}.$$

Hence, adding and subtracting V_*/κ_0 to the right hand side of (A.1) and simplifying the resulting terms, we obtain that

$$V = V_*/\kappa_0 - \kappa_0^{-1}(1-\kappa_0)\sum_{l=1}^{L}(\tilde{K}_l Q_l^*/\alpha_l^*)\operatorname{var}_{f^*}\{g(Z^*,\theta^*)|Z^* \in \mathbb{C}_l\}.$$

Therefore, assuming that $\operatorname{var}_{f^*}\{g(Z^*,\theta^*)|Z^* \in \mathbb{C}_l\}$ is positive definite for at least one stratum, the desired result follows since $\kappa_0 \in (0,1)$. □

Proof of Theorem 4.3. Since $\sum_{j=1}^n \hat{p}_j(\hat{\beta})/c(Z_j,\hat{\gamma}) = 1$ [because $1/c(Z,\gamma) - 1$ is spanned by the coordinates of $\rho_2(Z,R,Q_{-L})$, $\rho_3(Z,R,K_{-L})$, $\rho_4(R,\kappa)$; $\sum_{j=1}^n \hat{p}_j(\hat{\beta})\rho_2(Z_j,R_j,\hat{Q}_{-L}) = 0$, $\sum_{j=1}^n \hat{p}_j(\hat{\beta})\rho_3(Z_j,R_j,\hat{K}_{-L}) = 0$, $\sum_{j=1}^n \hat{p}_j(\hat{\beta})\rho_4(R_j,\hat{\kappa}) = 0$; and the $\hat{p}_j$'s add up to one],

$$\hat{F}^*(\xi) - F^*(\xi) = \sum_{j=1}^n \hat{p}_j(\hat{\beta})\{\mathbb{1}(Z_j \le \xi) - F^*(\xi)\}/c(Z_j,\hat{\gamma}).$$

Hence, a Taylor expansion reveals that

$$n^{1/2}\{\hat{F}^*(\xi) - F^*(\xi)\} = n^{1/2}\sum_{j=1}^n \hat{p}_j(\hat{\beta})I_c(Z_j,\xi) + \sum_{j=1}^n \hat{p}_j(\hat{\beta})\frac{\partial I_c(Z_j,\xi)}{\partial\gamma}\, n^{1/2}(\hat{\gamma}-\gamma^*) + O_p(n^{-1/2}). \quad \text{(A.2)}$$

But by a uniform weak law of large numbers as in Newey and McFadden (1994, Lemma 2.4),

$$\sum_{j=1}^n \hat{p}_j(\hat{\beta})\frac{\partial I_c(Z_j,\xi)}{\partial\gamma} = \mathbb{E}_{f_e}\{\frac{\partial I_c(Z,\xi)}{\partial\gamma}\} + o_p(1).$$

Furthermore, since $\hat{Q}_{-L} = \sum_{j=1}^n \hat{p}_j(\hat{\beta})s(Z_j)R_j/\sum_{j=1}^n \hat{p}_j(\hat{\beta})R_j$, we have

$$n^{1/2}(\hat{Q}_{-L} - Q_{-L}^*) = n^{1/2}\sum_{j=1}^n \hat{p}_j(\hat{\beta})\rho_2(Z_j,R_j,Q_{-L}^*)/\kappa_0 + o_p(1).$$

Similarly, we can show that

$$n^{1/2}(\hat{K}_{-L} - \tilde{K}_{-L}) = n^{1/2}\sum_{j=1}^n \hat{p}_j(\hat{\beta})\rho_3(Z_j,R_j,\tilde{K}_{-L})/(1-\kappa_0) + o_p(1)$$

and $n^{1/2}(\hat{\kappa}-\kappa_0) = n^{1/2}\sum_{j=1}^n \hat{p}_j(\hat{\beta})\rho_4(R_j,\kappa_0)$. Using these results, some algebra shows that (A.2) can be written as

$$n^{1/2}\{\hat{F}^*(\xi) - F^*(\xi)\} = n^{1/2}\sum_{j=1}^n \hat{p}_j(\hat{\beta})(w_j + \delta_{3j} + \delta_{4j}) + o_p(1),$$

where $\delta_{3j} = \operatorname{Proj}\{I_c(Z_j,\xi)|1,\rho_3(Z_j,R_j,\tilde{K}_{-L})\} + \mathbb{E}_{f_e}\{\partial I_c(Z_j,\xi)/\partial K_{-L}\}\rho_3(Z_j,R_j,\tilde{K}_{-L})/(1-\kappa_0)$ and $\delta_{4j} = \operatorname{Proj}\{I_c(Z_j,\xi)|1,\rho_4(R_j,\kappa_0)\} + \mathbb{E}_{f_e}\{\partial I_c(Z_j,\xi)/\partial\kappa\}\rho_4(R_j,\kappa_0)$. But replacing $\rho_1(Z,\beta^*)$

in the proof of Lemma C.2 with $I_c(Z,\xi)$, we can show that δ_{3j} and δ_{4j} are identically zero for each j. Hence, we have that

$$n^{1/2}\{\hat{F}^*(\xi) - F^*(\xi)\} = n^{1/2}\sum_{j=1}^{n}\hat{p}_j(\hat{\beta})w_j + o_p(1).$$

Therefore, the asymptotic distribution of $n^{1/2}\{\hat{F}^*(\xi) - F^*(\xi)\}$ follows from the proof of Theorem 4.4 upon replacing $\mathbb{1}(Z_j \leq \xi) - F_e(\xi)$ with w_j. □

Proof of Theorem 4.4. Since $n^{1/2}\{\hat{F}_e(\xi) - F_e(\xi)\} = n^{1/2}\sum_{j=1}^{n}\hat{p}_j(\hat{\beta})[\mathbb{1}(Z_j \leq \xi) - F_e(\xi)]$, from Brown and Newey (2002, Theorem 1) we know that the latter is asymptotically normal with mean zero and variance

$$\mathbb{E}_{f_e}\{\mathbb{1}(Z \leq \xi) - F_e(\xi)\}^2 - \mathbb{E}_{f_e}\{\mathbb{1}(Z \leq \xi)\rho'(Z,R,\beta^*)\}M_{f_e}\mathbb{E}_{f_e}\{\mathbb{1}(Z \leq \xi)\rho(Z,R,\beta^*)\}, \quad \text{(A.3)}$$

where $M_{f_e} = V_{f_e}^{-1} - V_{f_e}^{-1}D_{f_e}(D_{f_e}'V_{f_e}^{-1}D_{f_e})^{-1}D_{f_e}'V_{f_e}^{-1}$. Using expressions for D_{f_e}, $V_{f_e}^{-1}$, and $(D_{f_e}'V_{f_e}^{-1}D_{f_e})^{-1}$ in the proof of Theorem 4.1, straightforward calculations show that

$$M_{f_e} = \begin{bmatrix} M_V & M_V D_2/\kappa_0 & -M_V\Sigma_{13}V_3^{-1} & -M_V\Sigma_{14}V_4^{-1} \\ D_2'M_V/\kappa_0 & D_2'M_V D_2/\kappa_0^2 & -D_2'M_V\Sigma_{13}V_3^{-1}/\kappa_0 & -D_2'M_V\Sigma_{14}V_4^{-1}/\kappa_0 \\ -V_3^{-1}\Sigma_{13}'M_V & -V_3^{-1}\Sigma_{13}'M_V D_2/\kappa_0 & V_3^{-1}\Sigma_{13}'M_V\Sigma_{13}V_3^{-1} & V_3^{-1}\Sigma_{13}'M_V\Sigma_{14}V_4^{-1} \\ -V_4^{-1}\Sigma_{14}'M_V & V_4^{-1}\Sigma_{14}'M_V D_2/\kappa_0 & V_4^{-1}\Sigma_{14}'M_V\Sigma_{13}V_3^{-1} & V_4^{-1}\Sigma_{14}'M_V\Sigma_{14}V_4^{-1} \end{bmatrix},$$

where $D_2 \stackrel{\text{def}}{=} \mathbb{E}_{f_e}\{\partial\rho_1(Z,\beta^*)/\partial Q_{-L}^*\}$ for notational convenience. But since

$$v = \begin{bmatrix} I_{q\times q} & D_2/\kappa_0 & -\Sigma_{13}V_3^{-1} & -\Sigma_{14}/V_4 \end{bmatrix}\rho(Z,R,\beta^*),$$

the second term in (A.3) is equal to $\mathbb{E}_{f_e}\{\mathbb{1}(Z \leq \xi)v'\}M_V\mathbb{E}_{f_e}\{\mathbb{1}(Z \leq \xi)v\}$. The desired result follows. □

Appendix B. Efficiency bounds

As in Appendix A, we only obtain efficiency bounds for estimating parameters when data is collected by SS sampling; bounds for parameters when data is collected by VP sampling can be shown in a similar manner. We use the methodology of Severini and Tripathi (2001) to calculate the efficiency bounds. For a comprehensive treatment of efficiency bounds and the related literature see, e.g., Newey (1990) and Bickel, Klassen, Ritov, and Wellner (1993).

Begin by writing the enriched density as $f_e(z,r) = a_0^2(z,r)$. This ensures that a_0 lies in $L_2(z,r)$, the set of real-valued functions on $\mathbb{R}^d \times \{0,1\}$ that are square-integrable with respect to $\mu \otimes \bar{c}$. Now, suppose that we want to calculate the efficiency bound for estimating $\eta(a_0)$, a pathwise differentiable functional of a_0; see Severini and Tripathi (2001) for technical definitions and details. We proceed as follows. Let $t \mapsto a_t$ be a curve from an interval containing zero into the unit ball of $L_2(z,r)$ such that $a_t|_{t=0} = a_0$. Since the observed loglikelihood for t in this submodel is $\log a_t^2(z,r)$, the Fisher information for a single observation is given by

$i_{\mathcal{F}} = 4\int_{\mathbb{R}^d\times\{0,1\}} \dot{a}^2(z,r)\,d\mu\,d\bar{c}$, where $\dot{a}$ denotes the tangent vector to a_t at $t=0$; i.e., $\dot{a}\in L_2(z,r)$ satisfies $\int_{\mathbb{R}^d\times\{0,1\}} a_0(z,r)\dot{a}(z,r)\,d\mu\,d\bar{c}=0$. Note that $i_{\mathcal{F}}$ is induced by the Fisher inner-product $\langle \dot{a}_1,\dot{a}_2\rangle_{\mathcal{F}} = 4\int_{\mathbb{R}^d\times\{0,1\}} \dot{a}_1(z,r)\dot{a}_2(z,r)\,d\mu\,d\bar{c}$. Thus $i_{\mathcal{F}} = \|\dot{a}\|^2_{\mathcal{F}}$, where $\|\cdot\|_{\mathcal{F}}$ denotes the norm generated by the Fisher inner-product. Let $\mathcal{T}$ denote the collection of tangent vectors, i.e., the tangent space; namely,

$$\mathcal{T} = \{\dot{a}\in L_2(z,r) : \int_{\mathbb{R}^d\times\{0,1\}} a_0(z,r)\dot{a}(z,r)\,d\mu\,d\bar{c}=0\}.$$

By (4.7), we know that (2.1) is equivalent to $\mathbb{E}_{f_e}\{g(Z,\theta^*)/c(Z,Q^*_{-L},\tilde{K}_{-L},\kappa_0)\}=0_{q\times 1}$. Hence, we have to incorporate this additional information when calculating the efficiency bound for estimating $\eta(a_0)$. To do so, let $t\mapsto\theta_t$ denote a curve in $\mathbb{R}^p$ passing through θ^* at $t=0$ such that, for all t in a neighborhood of zero,

$$\int_{\mathbb{R}^d\times\{0,1\}} g(z,\theta_t)a_t^2(z,r)/c(z,Q_{-L,t},K_{-L,t},\kappa_t)\,d\mu\,d\bar{c}=0_{q\times 1} \tag{B.1}$$

and, following (4.4)–(4.6), $Q_{-L,t}$, $K_{-L,t}$, and κ_t are curves passing through Q^*_{-L}, $\tilde{K}_{-L}$, and κ_0 at $t=0$ given by the following moment conditions:

$$\int_{\mathbb{R}^d\times\{0,1\}} (s(z)-Q_{-L,t})ra_t^2(z,r)\,d\mu\,d\bar{c}=0_{(L-1)\times 1}, \tag{B.2}$$

$$\int_{\mathbb{R}^d\times\{0,1\}} (s(z)-K_{-L,t})(1-r)a_t^2(z,r)\,d\mu\,d\bar{c}=0_{(L-1)\times 1}, \tag{B.3}$$

$$\int_{\mathbb{R}^d\times\{0,1\}} (r-\kappa_t)a_t^2(z,r)\,d\mu\,d\bar{c}=0. \tag{B.4}$$

Hence, using (B.1)–(B.4), some algebra shows that the tangent vectors $\dot{a}$ and $\dot{\theta}$ must satisfy

$$\begin{aligned}
&D\dot{\theta} + 2\int_{\mathbb{R}^d\times\{0,1\}} \varepsilon a_0(z,r)\dot{a}(z,r)\,d\mu\,d\bar{c}\\
&\quad + 2[\Sigma_{12}V_2^{-1} + \kappa_0^{-1}\mathbb{E}_{f_e}\{\partial\rho_1(Z,\beta^*)/\partial Q_{-L}\}]\int_{\mathbb{R}^d\times\{0,1\}} \rho_2(z,r,Q^*_{-L})a_0(z,r)\dot{a}(z,r)\,d\mu\,d\bar{c}\\
&\quad + 2[\Sigma_{13}V_3^{-1} + (1-\kappa_0)^{-1}\mathbb{E}_{f_e}\{\partial\rho_1(Z,\beta^*)/\partial K_{-L}\}]\int_{\mathbb{R}^d\times\{0,1\}} \rho_3(z,r,\tilde{K}_{-L})a_0(z,r)\dot{a}(z,r)\,d\mu\,d\bar{c}\\
&\quad + 2[\Sigma_{14}/V_4 + \mathbb{E}_{f_e}\{\partial\rho_1(Z,\beta^*)/\partial\kappa\}]\int_{\mathbb{R}^d\times\{0,1\}} \rho_4(r,\kappa_0)a_0(z,r)\dot{a}(z,r)\,d\mu\,d\bar{c}\\
&\quad = 0_{q\times 1}.
\end{aligned}$$

Therefore, by Lemma C.2, it follows that

$$D\dot{\theta} + 2\int_{\mathbb{R}^d\times\{0,1\}} va_0(z,r)\dot{a}(z,r)\,d\mu\,d\bar{c}=0_{q\times 1}. \tag{B.5}$$

Let W be a $q \times q$ symmetric positive-definite matrix. Premultiplying (B.5) by $(D'WD)^{-1}D'W$ and solving for $\dot{\theta}$, we obtain that

$$\dot{\theta} = -2(D'WD)^{-1}D'W \int_{\mathbb{R}^d \times \{0,1\}} v a_0(z,r)\dot{a}(z,r)\, d\mu\, d\bar{c}. \tag{B.6}$$

Finally, substituting (B.6) in (B.5), we get that

$$(I_{q\times q} - D(D'WD)^{-1}D'W) \int_{\mathbb{R}^d \times \{0,1\}} v a_0(z,r)\dot{a}(z,r)\, d\mu\, d\bar{c} = 0_{q\times 1}. \tag{B.7}$$

Note that (B.7) represents the restriction on the tangent space due to the presence of overidentifying moment restrictions (because if $q = p$, then (B.7) holds for all $\dot{a} \in \mathcal{T}$ and $W \in \mathcal{W}$, where $\mathcal{W}$ denotes the set of $q \times q$ symmetric positive-definite matrices). Furthermore, since the map $x \mapsto D(D'WD)^{-1}D'Wx$ represents orthogonal projection onto the column space of D using the weighted inner product $\langle x_1, x_2\rangle = x_1'Wx_2$, it follows that (B.7) is satisfied by only those tangent vectors for which $\int_{\mathbb{R}^d\times\{0,1\}} v a_0(z,r)\dot{a}(z,r)\, d\mu\, d\bar{c}$ lies in the column space of D. Let $\mathcal{T}_W$ denote the set of tangent vectors that satisfy (B.7); i.e.,

$$\mathcal{T}_W = \{\dot{a} \in \mathcal{T} : (I_{q\times q} - D(D'WD)^{-1}D'W) \int_{\mathbb{R}^d \times \{0,1\}} v a_0(z,r)\dot{a}(z,r)\, d\mu\, d\bar{c} = 0_{q\times 1}\}.$$

Following Severini and Tripathi (2001), the efficiency bound for estimating $\eta(a_0)$ is given by $\sup_{W\in\mathcal{W}} \|\nabla\eta\|_W^2$, where $\|\nabla\eta\|_W = \sup_{\{\dot{a}\in\mathcal{T}_W : \dot{a}\neq 0\}} |\nabla\eta(\dot{a})|$ and $\nabla\eta$ denotes the pathwise derivative of η. To calculate the bound we do the following. First, for any $W \in \mathcal{W}$, we employ a guess-and-verify strategy to find an $a_W^* \in \mathcal{T}$ satisfying

$$\nabla\eta(\dot{a}) = \langle \dot{a}, a_W^*\rangle_{\mathcal{F}} \quad \text{for all } \dot{a} \in \mathcal{T}_W. \tag{B.8}$$

Next, we pick a $W^* \in \mathcal{W}$ such that $\int_{\mathbb{R}^d\times\{0,1\}} v a_0(z,r) a_{W^*}^*(z,r)\, d\mu\, d\bar{c}$ lies in the column space of D. This means that $a_{W^*}^*$ lies in $\mathcal{T}_{W^*}$ and we can use this fact to show that $\|\nabla\eta\|_{W^*} = \|a_{W^*}^*\|_{\mathcal{F}}$. [By (B.8), $a_{W^*}^*$ satisfies $\nabla\eta(\dot{a}) = \langle \dot{a}, a_{W^*}^*\rangle_{\mathcal{F}}$ for all $\dot{a} \in \mathcal{T}_{W^*}$. Hence, $\|\nabla\eta\|_{W^*} \leq \|a_{W^*}^*\|_{\mathcal{F}}$ by Cauchy-Schwarz. But since $a_{W^*}^* \in \mathcal{T}_{W^*}$, we also have $\|a_{W^*}^*\|_{\mathcal{F}}^2 = \nabla\eta(a_{W^*}^*) \leq \|\nabla\eta\|_{W^*}\, \|a_{W^*}^*\|_{\mathcal{F}}$; i.e., $\|\nabla\eta\|_{W^*} \geq \|a_{W^*}^*\|_{\mathcal{F}}$.] But as shown in the proofs of Theorems B.1–B.4, the matrix W^* is uniquely determined up to scale. Hence, the efficiency bound for estimating $\eta(a_0)$ under (2.1) is given by $\|a_{W^*}^*\|_{\mathcal{F}}^2$.

We use this procedure in Theorems B.1–B.4 to obtain the efficiency bounds for estimating θ^*, Q_{-L}^*, $F^*(\xi)$, and $F_e(\xi)$. Comparing them with Theorems 4.1–4.4 immediately shows that the estimators $\hat{\theta}$, $\hat{Q}_{-L}$, $\hat{F}(\xi)$, and $\hat{F}_e(\xi)$ are asymptotically efficient.

Theorem B.1. *The efficiency bound for estimating θ^* is given by $(D'V^{-1}D)^{-1}$.*

Proof of Theorem B.1. Let $\zeta \in \mathbb{R}^p$ be arbitrary. To obtain the efficiency bound for estimating $\eta(a_0) = \zeta'\theta^*$, the tangent vectors $\dot{a}$ and $\dot{\theta}$ must satisfy $\nabla\eta(\dot{a}) = \zeta'\dot{\theta}$. Hence, by (B.6), for

any $W \in \mathcal{W}$ we have that

$$\nabla\eta(\dot{a}) = -2\zeta'(D'WD)^{-1}D'W \int_{\mathbb{R}^d\times\{0,1\}} va_0(z,r)\dot{a}(z,r)\,d\mu\,d\bar{c}.$$

By (B.8), we have to find a $a^*_W \in \mathfrak{T}$ such that

$$\int_{\mathbb{R}^d\times\{0,1\}} \{a^*_W(z,r) + 0.5\zeta'(D'WD)^{-1}D'Wva_0(z,r)\}\dot{a}(z,r)\,d\mu\,d\bar{c} = 0 \quad \text{for all } \dot{a} \in \mathfrak{T}_W. \tag{B.9}$$

We claim that $a^*_W(z,r) = -0.5\zeta'(D'WD)^{-1}D'Wva_0(z,r)$. It is easily verified that $a^*_W \in \mathfrak{T}$ and satisfies (B.9). Hence, we only have to determine W^* such that $\int_{\mathbb{R}^d\times\{0,1\}} va_0(z,r)a^*_{W^*}(z,r)\,d\mu\,d\bar{c}$ lies in the column space of D. But since

$$\int_{\mathbb{R}^d\times\{0,1\}} va_0(z,r)a^*_W(z,r)\,d\mu\,d\bar{c} = -0.5VWD(D'WD)^{-1}\zeta,$$

it follows that $\int_{\mathbb{R}^d\times\{0,1\}} va_0(z,r)a^*_{W^*}(z,r)\,d\mu\,d\bar{c}$ lies in the column space of D if and only if $VW^* \propto I_{q\times q}$. Hence, $a^*_{W^*}(z,r) = -0.5\zeta'(D'V^{-1}D)^{-1}D'V^{-1}va_0(z,r)$, and the efficiency bound for estimating $\zeta'\theta^*$ is given by

$$4\int_{\mathbb{R}^d\times\{0,1\}} \{a^*_{W^*}(z,r)\}^2\,d\mu\,d\bar{c} = \zeta'(D'V^{-1}D)^{-1}\zeta.$$

The desired result follows since ζ was arbitrary. □

Theorem B.2. *The efficiency bound for estimating Q^*_{-L} is given by $(V_2 - V_2J'M_VJV_2)/\kappa_0^2$.*

Proof of Theorem B.2. Let $\phi \in \mathbb{R}^{L-1}$ be arbitrary. Since by (4.4) we can express Q^*_{-L} in terms of a_0, we have to find the efficiency bound for estimating $\eta(a_0) = \phi'Q^*_{-L}$. Thus, by (B.2),

$$\nabla\eta(\dot{a}) = 2\int_{\mathbb{R}^d\times\{0,1\}} \kappa_0^{-1}\phi'\rho_2(z,r,Q^*_{-L})a_0(z,r)\dot{a}(z,r)\,d\mu\,d\bar{c}. \tag{B.10}$$

Comparing (B.10) with (B.12), we see that the term $\phi'\rho_2(z,r,Q^*_{-L})/\kappa_0$ in (B.10) corresponds to $\mathbb{1}(z \le \xi) - F_e(\xi)$ in (B.12). Therefore, the efficiency bound for estimating $\phi'Q^*_{-L}$ is easily obtained by replacing $\mathbb{1}(Z \le \xi) - F_e(\xi)$ in (B.14) with $\phi'\rho_2(Z,R,Q^*_{-L})/\kappa_0$. The desired result follows since ϕ was arbitrary. □

Theorem B.3. *The efficiency bound for estimating $F^*(\xi)$ is given by*

$$\mathbb{E}_{f_e}\{w^2\} - \mathbb{E}_{f_e}\{wv'\}M_V\mathbb{E}_{f_e}\{wv\}.$$

Proof of Theorem B.3. Since $F^*(\xi) = \mathbb{E}_{f^*}\{\mathbb{1}(Z^* \le \xi)\}$, by (4.3) it follows that we can identify $F^*(\xi) \overset{\text{def}}{=} \eta(a_0)$ via the moment condition

$$\int_{\mathbb{R}^d\times\{0,1\}} \{\frac{\mathbb{1}(z \le \xi) - \eta(a_0)}{c(z,Q^*_{-L},\tilde{K}_{-L},\kappa_0)}\}a_0^2(z,r)\,d\mu\,d\bar{c} = 0.$$

Hence, similar to the manner in which we derived (B.5), we can show that

$$\nabla\eta(\dot{a}) = 2\int_{\mathbb{R}^d\times\{0,1\}} wa_0(z,r)\dot{a}(z,r)\,d\mu\,d\bar{c}. \tag{B.11}$$

But w in (B.11) corresponds to $\mathbb{1}(z \leq \xi) - F_e(\xi)$ in (B.12). Therefore, the efficiency bound for estimating $F^*(\xi)$ is obtained by replacing $\mathbb{1}(Z \leq \xi) - F_e(\xi)$ in (B.14) with w. □

Theorem B.4. *The efficiency bound for estimating $F_e(\xi)$ is given by*

$$F_e(\xi)[1 - F_e(\xi)] - \mathbb{E}_{f_e}\{\mathbb{1}(Z \leq \xi)v'\}M_V\mathbb{E}_{f_e}\{\mathbb{1}(Z \leq \xi)v\}.$$

Proof of Theorem B.4. Since here $\int_{\mathbb{R}^d\times\{0,1\}}\{\mathbb{1}(z \leq \xi) - \eta(a_0)\}a_0^2(z,r)\,d\mu\,d\bar{c} = 0$,

$$\nabla\eta(\dot{a}) = 2\int_{\mathbb{R}^d\times\{0,1\}}[\mathbb{1}(z \leq \xi) - F_e(\xi)]a_0(z,r)\dot{a}(z,r)\,d\mu\,d\bar{c}. \tag{B.12}$$

By (B.8), we have to find a $a_W^* \in \mathcal{T}$ such that

$$\int_{\mathbb{R}^d\times\{0,1\}}\{a_W^*(z,r) - 0.5[\mathbb{1}(z \leq \xi) - F_e(\xi)]a_0(z,r)\}\dot{a}(z,r)\,d\mu\,d\bar{c} = 0 \quad \text{for all } \dot{a} \in \mathcal{T}_W. \tag{B.13}$$

Define $c_\xi = \mathbb{E}_{f_e}\{[1(Z \leq \xi) - F_e(\xi)]v\}$. We claim that

$$a_W^*(z,r) = 0.5\{\mathbb{1}(z \leq \xi) - F_e(\xi) - c_\xi' V^{-1}(I - D(D'WD)^{-1}D'W)v\}a_0(z,r).$$

Using (B.7), it is easily verified that a_W^* satisfies (B.13) and that it lies in $\mathcal{T}$. Next, since

$$\int_{\mathbb{R}^d\times\{0,1\}} va_0(z,r)a_W^*(z,r)\,d\mu\,d\bar{c} = 0.5VWD(D'WD)^{-1}D'V^{-1}c_\xi,$$

it follows that $\int_{\mathbb{R}^d\times\{0,1\}} va_0(z,r)a_{W^*}^*(z,r)\,d\mu\,d\bar{c}$ lies in the column space of D if and only if $W^* \propto V^{-1}$. Therefore, $a_{W^*}^*(z,r) = 0.5\{\mathbb{1}(z \leq \xi) - F_e(\xi) - c_\xi^{*\prime}M_V v\}a_0(z,r)$ and the efficiency bound for estimating $F^*(\xi)$ is given by

$$\begin{aligned}4\int_{\mathbb{R}^d\times\{0,1\}}\{a_{W^*}^*(z,r)\}^2\,d\mu\,d\bar{c} = \mathbb{E}_{f_e}\{\mathbb{1}(Z \leq \xi) - F_e(\xi)\}^2 \\ - \mathbb{E}_{f_e}\{[1(Z \leq \xi) - F_e(\xi)]v'\}M_V\mathbb{E}_{f_e}\{[1(Z \leq \xi) - F_e(\xi)]v\}.\end{aligned} \tag{B.14}$$

The desired result follows since $\mathbb{E}_{f_e}\{v\} = 0$. □

Appendix C. Some useful results

Lemma C.1. *Let $D_{f_e} = \mathbb{E}_{f_e}\{\partial\rho(Z,R,\beta^*)/\partial\beta\}$. Then*

$$D_{f_e} = \begin{bmatrix} D & A_{q\times(2L-1)} \\ 0_{(2L-1)\times p} & B_{(2L-1)\times(2L-1)}\end{bmatrix},$$

where $A = \Big[\mathbb{E}_{f_e}\{\partial\rho_1(Z,\beta^*)/\partial Q_{-L}\} \quad -(1-\kappa_0)\Sigma_{13}V_3^{-1} \quad -\Sigma_{14}/V_4\Big]$ *and*

$$B = \begin{bmatrix} -\kappa_0 I_{(L-1)\times(L-1)} & 0_{(L-1)\times(L-1)} & 0_{(L-1)\times 1} \\ 0_{(L-1)\times(L-1)} & -(1-\kappa_0) I_{(L-1)\times(L-1)} & 0_{(L-1)\times 1} \\ 0'_{(L-1)\times 1} & 0'_{(L-1)\times 1} & -1 \end{bmatrix}.$$

Proof of Lemma C.1. From (4.8) it is immediate that $D_{f_e} = \left[\begin{smallmatrix} D \\ 0_{(2L-1)\times p} \end{smallmatrix} \begin{smallmatrix} A \\ B \end{smallmatrix}\right]$, where

$$A = \Big[\mathbb{E}_{f_e}\{\partial\rho_1(Z,\beta^*)/\partial Q_{-L}\} \quad \mathbb{E}_{f_e}\{\partial\rho_1(Z,\beta^*)/\partial K_{-L}\} \quad \mathbb{E}_{f_e}\{\partial\rho_1(Z,\beta^*)/\partial\kappa\}\Big].$$

The desired result now follows by Lemma C.2. □

Lemma C.2. $(1-\kappa_0)\Sigma_{13}V_3^{-1} + \mathbb{E}_{f_e}\{\partial\rho_1(Z,\beta^*)/\partial K_{-L}\} = 0_{q\times(L-1)}$ *and*

$$\Sigma_{14}/V_4 + \mathbb{E}_{f_e}\{\partial\rho_1(Z,\beta^*)/\partial\kappa\} = 0_{q\times 1}.$$

Proof of Lemma C.2. Use the definition of $c(Z,\gamma^*)$ to observe that

$$\mathbb{E}_{f_e}\{\partial\rho_1(Z,\beta^*)/\partial K_{-L}\} = -(1-\kappa_0)\mathbb{E}_{f_e}\Big\{\frac{g(Z,\theta^*)}{c^2(Z,\gamma^*)}\frac{\partial b(Z,Q^*,\tilde{K})}{\partial K_{-L}}\Big\}.$$

Doing a little simplifying, we can show that $\partial b(Z,Q^*,\tilde{K})/\partial K_{-L} = \kappa_0[s(Z) - Q^*_{-L}]'V_2^{-1}$. Hence,

$$\mathbb{E}_{f_e}\{\partial\rho_1(Z,\beta^*)/\partial K_{-L}\} = -\kappa_0(1-\kappa_0)\mathbb{E}_{f_e}\{g(Z,\theta^*)[s(Z) - Q^*_{-L}]'/c^2(Z,\gamma^*)\}V_2^{-1}. \tag{C.1}$$

Now, by (4.3), it is easy to see that

$$\Sigma_{12} = \kappa_0\mathbb{E}_{f_e}\{g(Z,\theta^*)[s(Z) - Q^*_{-L}]'/c^2(Z,\gamma^*)\}. \tag{C.2}$$

Therefore, the first result follows by (C.1), (C.2), and Lemma C.3. For the second result, note that $\partial c(Z,\gamma^*)/\partial\kappa = [1 - c(Z,\gamma^*)]/(1-\kappa_0)$. Therefore,

$$(1-\kappa_0)\mathbb{E}_{f_e}\{\partial\rho_1(Z,\beta^*)/\partial\kappa\} = -\mathbb{E}_{f_e}\{g(Z,\theta^*)[1-c(Z,\gamma^*)]/c^2(Z,\gamma^*)\} = -\mathbb{E}_{f_e}\{g(Z,\theta^*)/c^2(Z,\gamma^*)\}.$$

The second result follows since $\Sigma_{14} = \kappa_0\mathbb{E}_{f_e}\{g(Z,\theta^*)/c^2(Z,\gamma^*)\}$ and $V_4 = \kappa_0(1-\kappa_0)$. □

Lemma C.3. $\Sigma_{12}V_2^{-1} = \Sigma_{13}V_3^{-1}$.

Proof of Lemma C.3. Begin by observing that $\Sigma_{13} = (1-\kappa_0)\mathbb{E}_f\{\rho_1(Z,\beta^*)[s(Z) - \tilde{K}_{-L}]'\}$, where f is defined in (2.3). A little algebra shows that

$$\begin{aligned}\mathbb{E}_f\{\rho_1(Z,\beta^*)s'(Z)\} &= \mathbb{E}_{f_e}\{g(Z,\theta^*)[s(Z) - Q^*_{-L}]'/c^2(Z,\gamma^*)\}\mathcal{K}\mathcal{Q}^{-1} + \mathbb{E}_{f_e}\{g(Z,\theta^*)/c^2(Z,\gamma^*)\}\tilde{K}'_{-L} \\ &= (\Sigma_{12}\mathcal{K}\mathcal{Q}^{-1} + \Sigma_{14}\tilde{K}'_{-L})/\kappa_0.\end{aligned}$$

Hence, since $\mathbb{E}_f\{\rho_1(Z,\beta^*)\} = -\Sigma_{14}/(1-\kappa_0)$, we get that

$$\Sigma_{13} = \{(1-\kappa_0)\Sigma_{12}\mathcal{K}\mathcal{Q}^{-1} + \Sigma_{14}\tilde{K}'_{-L}\}/\kappa_0. \tag{C.3}$$

Next, since $\rho_2'(Z,R,Q^*_{-L})[\mathcal{Q}^{-1}\mathcal{A} - (\alpha^*_L/Q^*_L)I_{(L-1)\times(L-1)}]1_{(L-1)\times 1} = [c(Z,\gamma^*) - 1]R$, we have

$$\Sigma_{12}\{\mathcal{Q}^{-1}\mathcal{A} - (\alpha^*_L/Q^*_L)I_{(L-1)\times(L-1)}\}1_{(L-1)\times 1} = -\Sigma_{14}. \tag{C.4}$$

Therefore, using (C.4) to substitute for Σ_{14} in (C.3) and simplifying further, we obtain that $\Sigma_{13} = \Sigma_{12} V_2^{-1} V_3$. The desired result follows. □