INFERENCE IN CONDITIONAL MOMENT RESTRICTION MODELS WHEN THERE IS SELECTION DUE TO STRATIFICATION

ANTONIO COSMA[1], ANDREÏ VICTOROVITCH KOSTYRKA[2], AND GAUTAM TRIPATHI[3]

[1,2,3]Center for Research in Economics and Management (CREA)
Faculty of Law, Economics and Finance
University of Luxembourg
L-1511, Luxembourg

antonio.cosma@uni.lu, andrei.kostyrka@uni.lu, gautam.tripathi@uni.lu

Running head: Selection due to stratification

Abstract. We show how to use a smoothed empirical likelihood approach to conduct efficient semiparametric inference in models characterized as conditional moment equalities when data is collected by variable probability sampling. Results from a simulation experiment suggest that the smoothed empirical likelihood based estimator can estimate the model parameters very well in small to moderately sized stratified samples.

Keywords: Conditional moment models, Smoothed empirical likelihood, Stratification, Variable probability sampling.

31st July, 2018

1. INTRODUCTION

The gold standard for collecting data, at least for the ease of doing subsequent statistical analysis, is simple random sampling, whereby each observation in the "target" population, namely, the population of interest, has an equal chance of being chosen. Consequently, the probability distribution of the chosen observation, regarded as belonging to a "realized" population, is the same as the probability distribution of an observation in the target population, which facilitates statistical analysis.

However, when estimating or testing economic relationships, economists often discover that the data they plan to use is not drawn from the target population they wish to study. Instead, the observations are found to be sampled from a related but different population. Sometimes this is done deliberately to make the sample more informative. E.g., when studying the impact of welfare legislation, it is desirable to oversample minorities and low income families. Similarly, if we want to examine the effect of disability laws on demand for public transportation, it makes sense to oversample households with disabled members. At other times, a distinction between the target and realized populations can be created unintentionally. E.g., in sampling the duration of unemployment at a randomly chosen time, economists are more likely to observe longer unemployment spells than shorter ones. Using a dataset to answer questions for which it was not originally designed, a typical situation in economics where data is often costly to collect, may also lead to such a situation (Newey, 1993, p. 419). For instance, if the reason for collecting data is to estimate mean income for an underlying population, oversampling low income and undersampling high income families can improve the precision of estimators. However, at some later stage this income data can be used by another researcher as the dependent variable in a regression model without realizing that the original sample was drawn from a distribution other than the target population.

Whatever its cause, if the distinction between the target and realized populations is not taken into account when analyzing the data, statistical inference can be seriously off the mark. This phenomenon is commonly called selection bias. Cf. Heckman (1976, 1979) and Manski (1989, 1995) for a classic exposition of the selection problem.

In this paper, we describe an efficient semiparametric approach for conducting inference in conditional moment restriction models when data is collected by a variable probability sampling scheme such that the observations from the target population have unequal chances of being chosen. In other words, we show how to efficiently deal with the selection bias caused by the sampling scheme used to collect the data, because the sampling scheme induces a probability distribution on the realized population which differs from the target distribution for which inference is to be made.

The remainder of the paper is organized as follows. In Section 2, we describe the conditional moment restriction model and the variable probability sampling scheme. Section 3 discusses how to do inference using the smoothed empirical likelihood approach, and finite sample properties of the proposed estimator are examined in Section 4. Section 5 concludes the paper. Related technical details are in the appendices.

2. THE MODEL

2.1. **Conditional moment equalities.** Let $Z^* := (Y^*, X^*)_{\dim(Y^*)+\dim(X^*)\times 1}$ be a random (column) vector that denotes an observation from the target population, where Y^* is the vector of endogenous variables and X^* the vector of exogenous variables. Assume that

$$\text{H}_0: \quad \exists\theta^* \in \mathbb{R}^{\dim(\theta^*)} \quad \text{s.t.} \quad \mathbb{E}_{P^*_{Y^*|X^*}}[g(Z^*, \theta^*)|X^*] = 0 \qquad P^*_{X^*}\text{-a.s.,} \tag{2.1}$$

where g is a vector of functions known up to θ^*, the notation $\mathbb{E}_{P^*_{Y^*|X^*}}$ indicates that the conditional expectation is with respect to the conditional distribution $P^*_{Y^*|X^*} := \text{Law}(Y^*|X^*)$, and $P^*_{X^*}$ denotes the marginal distribution of X^*. The conditional distribution of $Y^*|X^*$ and the marginal distribution of X^* are unknown.[1] Throughout the paper, random variables and probability measures associated with the target population appear with the superscript "$*$." The parameter of interest θ^* has an asterisk attached to it because it is a functional of $P^*_{Y^*|X^*}$.[2]

A large class of models in applied economics can be characterized in terms of conditional moment equalities of the form (2.1). E.g., in linear regression models where some or all of the regressors are endogenous, we have $g(Z^*, \theta^*) := Y_1^* - \alpha^* - X_1^{*\prime}\beta^* - Y_2^{*\prime}\delta^*$, where $Y^* := (Y_1^*, Y_2^*)$ with Y_1^* the outcome variable and Y_2^* the vector of endogenous regressors; $X^* := (X_1^*, X_2^*)$ with X_1^* the exogenous regressors, i.e., the "included instruments," and X_2^* the "excluded instruments" for Y_2^*; and $\theta^* := (\alpha^*, \beta^*, \delta^*)$. If all the regressors are endogenous, then X_1^* is the empty vector and the definition of θ^* has to be adjusted accordingly by dropping β^*. Similarly, for nonlinear regression models, $g(Z^*, \theta^*) := Y_1^* - \psi(Y_2^*, X_1^*, \theta^*)$, where the nonlinear function $\psi(Y_2^*, X_1^*, \cdot)$ is known up to θ^*. Multivariate extensions include systems of equations or transformation models, linear or nonlinear, of the form $g(Z^*, \theta^*) = \varepsilon^*$, where g is a vector of known functions and the identifying assumption is that $\mathbb{E}_{P^*_{Y^*|X^*}}[\varepsilon^*|X^*] = 0$ $P^*_{X^*}$-a.s.. Several examples of econometric models defined via conditional moment restrictions may be found in Newey (1993, Section 3), Pagan & Ullah (1999, Chapter 3), and Wooldridge (2010).

[1] If X^* is constant $P^*_{X^*}$-a.s., then there is no conditioning and (2.1) reduces to a system of unconditional moment equalities. These models are studied in Tripathi (2011a,b).

[2] Similar notation, but without the "$*$" superscript, applies to the random variables and probability measures in the realized population.

2.2. **Variable probability sampling.** Instead of observing Z^* directly from the target population, we possess a random vector $Z := (Y, X)$ that is collected by variable probability (VP) sampling, also known as Bernoulli sampling. For more on VP and other stratified sampling schemes, cf., e.g., DeMets & Halperin (1977), Manski & Lerman (1977), Holt, Smith, & Winter (1980), Cosslett (1981a,b, 1991, 1993), Manski & McFadden (1981), Jewell (1985), Quesenberry & Jewell (1986), Scott & Wild (1986), Kalbfleisch & Lawless (1988), Bickel & Ritov (1991), Imbens (1992), Imbens & Lancaster (1996), Deaton (1997), Wooldridge (1999, 2001), Butler (2000), Bhattacharya (2005, 2007), Hirose (2007), Hirose & Lee (2008), Tripathi (2011a,b), and Severini & Tripathi (2013).

Let the support of Z^*, denoted by $\text{supp}(Z^*)$, be partitioned into L nonempty disjoint strata $\mathbb{C}_1, \dots, \mathbb{C}_L$. 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 $p_l \in (0, 1]$. If it is discarded, all information about the observation is lost. Hence, instead of observing a random vector Z^* drawn from the target distribution $P^* := \text{Law}(Z^*)$, we observe a random vector Z, with $\text{supp}(Z) = \text{supp}(Z^*)$, drawn from the realized distribution $P := \text{Law}(Z)$ given by[3]

$$P(Z \in B) := \sum_{l=1}^{L} \frac{p_l}{b^*} \int_B \mathbb{1}_{\mathbb{C}_l}(z)\, dP^*(z), \qquad B \in \mathscr{B}(\mathbb{R}^{\dim(Z^*)}), \tag{2.2}$$

where $\mathscr{B}(\mathbb{R}^{\dim(Z^*)})$ is the Borel sigma-field of $\mathbb{R}^{\dim(Z^*)}$, $b^* := \sum_{l=1}^{L} p_l Q_l^*$, and $Q_l^* := P^*(Z^* \in \mathbb{C}_l) > 0$ denotes the probability that a randomly chosen observation from the target population lies in the lth stratum.

Since Q_l^* represents the probability mass of the lth stratum in the target population, the Q_l^*'s are popularly called "aggregate shares." The aggregate shares, which add up to one, i.e., $\sum_{l=1}^{L} Q_l^* = 1$, are unknown parameters of interest to be estimated along with the structural parameter θ^*. The parameter b^* also has a practical interpretation, namely, it is the probability that an observation drawn from the target population during the sampling process is ultimately retained in the sample.

It is immediate from (2.2) that the density of P, with respect to any measure on $\mathscr{B}(\mathbb{R}^{\dim(Z^*)})$ that dominates P^*, is given by

$$\begin{aligned} dP(z) &:= \sum_{l=1}^{L} \frac{p_l}{b^*} \mathbb{1}_{\mathbb{C}_l}(z) dP^*(z) \qquad (z \in \mathbb{R}^{\dim(Z^*)}) \\ &= \frac{b(z)}{b^*} dP^*(z), \end{aligned} \tag{2.3}$$

[3]Cf. Severini & Tripathi (2013, Appendix H) for a short proof of (2.2).

where $b(z) := \sum_{l=1}^{L} p_l \mathbb{1}_{\mathbb{C}_l}(z)$. Following Imbens & Lancaster (1996, p. 296), $b(\cdot)/b^*$ is referred to as a bias function because it determines the selection bias due to stratified sampling, i.e., the extent to which P differs from P^*. For instance, it is easy to see that if the sampling probabilities $p_1, \ldots, p_L$ are all equal, then there is no selection bias, i.e., $P = P^*$, because $b(\cdot)/b^* = 1$ irrespective of the values taken by the aggregate shares.

The marginal density of X is given by

$$\begin{aligned} dP_X(x) &:= \int_{y \in \mathbb{R}^{\dim(Y^*)}} dP(y,x) && (x \in \mathbb{R}^{\dim(X^*)}) \\ &= \int_{y \in \mathbb{R}^{\dim(Y^*)}} \frac{b(y,x)}{b^*} \, dP^*_{Y^*|X^*=x}(y) \, dP^*_{X^*}(x) && ((2.3)) \\ &= \frac{\gamma^*(x)}{b^*} \, dP^*_{X^*}(x), && (2.4) \end{aligned}$$

where $\gamma^*(x) := \mathbb{E}_{P^*_{Y^*|X^*}}[b(Y^*,x)|X^* = x]$. Throughout the paper, we maintain the assumption that $\gamma^* > 0$ on $\text{supp}(X^*)$.[4] Under this condition, the probability distributions P_X and $P^*_{X^*}$ are mutually absolutely continuous, which we denote by writing $P^*_{X^*} \ll P_X \ll P^*_{X^*}$.

Since $\text{supp}(Y,X) = \text{supp}(Y^*,X^*)$ and $\gamma^* > 0$ on $\text{supp}(X^*)$, the conditional density of $Y|X$ is given by

$$\begin{aligned} dP_{Y|X=x}(y) &:= \frac{dP(y,x)}{dP_X(x)} && ((y,x) \in \text{supp}(Y^*) \times \text{supp}(X^*)) \\ &= \frac{b(y,x)}{\gamma^*(x)} \, dP^*_{Y^*|X^*=x}(y), && (2.5) \end{aligned}$$

where (2.5) follows from (2.3) and (2.4).

By (2.5), $dP_{Y|X=x}(y) = dP^*_{Y^*|X^*=x}(y)$ if and only if $b(y,x) = \gamma^*(x)$ for all $(y,x) \in \text{supp}(Y^*) \times \text{supp}(X^*)$. However, as discussed subsequently, the condition $b(y,x) = \gamma^*(x)$ holds only in a special case. Therefore, in general, $dP_{Y|X} \neq dP^*_{Y^*|X^*}$. Consequently, estimating (2.1) using the realized sample without accounting for the fact that it was obtained by stratified sampling, i.e., ignoring stratification, will generally not lead to a consistent estimator of θ^*.

2.3. **Identification.** In contrast to some other stratified sampling schemes (Tripathi, 2011b, Sections 3.1 and 4.1), identification, i.e., uniqueness, of θ^* cannot be lost because of VP sampling. To see this, begin by recalling that the assumption that $\gamma^* > 0$ on $\text{supp}(X^*)$ implies

[4] A sufficient condition for this is that $P^*_{Y^*|X^*}((Y^*,x) \in \mathbb{C}_l | X^* = x) > 0$ for each l and $x \in \text{supp}(X^*)$.

that the distributions P_X and $P^*_{X^*}$ are mutually absolutely continuous. Hence,

$$
\begin{aligned}
(2.1) \iff & \mathbb{E}_{P^*_{Y^*|X^*}}[g(Y^*, x, \theta^*)|X^* = x] = 0 \quad \text{for } P^*_{X^*}\text{-a.a. } x \in \operatorname{supp}(X^*) \\
\iff & \gamma^*(x)\mathbb{E}_{P_{Y|X}}[\frac{g(Y, x, \theta^*)}{b(Y, x)}|X = x] = 0 \quad \text{for } P^*_{X^*}\text{-a.a. } x \in \operatorname{supp}(X^*) && ((2.5)) \\
\iff & P^*_{X^*}\{x \in \operatorname{supp}(X^*) : \mathbb{E}_{P_{Y|X}}[\frac{g(Y, x, \theta^*)}{b(Y, x)}|X = x] \neq 0\} = 0 && (\gamma^* > 0) \\
\iff & P_X\{x \in \operatorname{supp}(X^*) : \mathbb{E}_{P_{Y|X}}[\frac{g(Y, x, \theta^*)}{b(Y, x)}|X = x] \neq 0\} = 0 && (P^*_{X^*} \ll P_X \ll P^*_{X^*}) \\
\iff & \mathbb{E}_{P_{Y|X}}[\frac{g(Y, x, \theta^*)}{b(Y, x)}|X = x] = 0 \quad \text{for } P_X\text{-a.a. } x \in \operatorname{supp}(X^*).
\end{aligned}
$$

Therefore, we have that

$$
(2.1) \iff \mathbb{E}_{P_{Y|X}}[\frac{g(Z, \theta^*)}{b(Z)}|X] = 0 \qquad P_X\text{-a.s.}. \tag{2.6}
$$

Since $b(Z)$ does not depend on θ^*, the equivalence in (2.6) reveals that θ^* in (2.1) is uniquely defined if and only if θ^* in $\mathbb{E}_{P_{Y|X}}[g(Z, \theta^*)/b(Z)|X] = 0$ (P_X-a.s.) is uniquely defined. That is, any condition that leads to the identification of θ^* in (2.1) will also ensure identification of θ^* in the right hand side of (2.6) and vice-versa. To illustrate this, assume that the columns of the partial derivative $\partial_\theta \mathbb{E}_{P^*_{Y^*|X^*}}[g(Z^*, \theta^*)|X^*]$ are linearly independent $P^*_{X^*}$-a.s.. As shown in Cosma, Kostyrka, & Tripathi (2018), this condition is sufficient to ensure that θ^* is locally identified.[5] However, since b does not depend on θ (which implies that γ^* does not depend on θ), we have that

$$
\partial_\theta \mathbb{E}_{P^*_{Y^*|X^*}}[g(Z^*, \theta^*)|X^* = x] \overset{(2.5)}{=} \gamma^*(x)\partial_\theta \mathbb{E}_{P_{Y|X}}[\frac{g(Z, \theta^*)}{b(Z)}|X = x], \qquad x \in \operatorname{supp}(X^*).
$$

Therefore, since $\gamma^* > 0$ on $\operatorname{supp}(X^*)$, the columns of $\partial_\theta \mathbb{E}_{P^*_{Y^*|X^*}}[g(Z^*, \theta^*)|X^*]$ are linearly independent $P^*_{X^*}$-a.s. if and only if the columns of $\partial_\theta \mathbb{E}_{P_{Y|X}}[g(Z, \theta^*)/b(Z)|X]$ are linearly independent P_X-a.s. (because P_X and $P^*_{X^*}$ are mutually absolutely continuous).

Since identification of θ^* cannot be lost because of VP sampling, for the remainder of the paper we maintain that θ^* is identified.

2.4. **Endogenous and exogenous stratification.** As noted by Wooldridge (1999, p. 1385), VP sampling is employed when it is cheaper to obtain information on a subset of variables in the target population. Hence, it may happen that in certain datasets only Y^* is stratified (endogenous stratification),[6] or only X^* is stratified (exogenous stratification), or both Y^* and X^* are stratified. To see how all these cases can be handled in a unified manner in our

[5] The same condition leads to global identification of θ^* whenever $g(Z^*, \theta^*)$ is linear in θ^*.

[6] In the econometrics literature, stratification based on a finite set of response variables is often referred to as choice based sampling.

framework, let the support of Y^* be partitioned into J nonempty disjoint strata $\mathbb{A}_1, \ldots, \mathbb{A}_J$, and the support of X^* be partitioned into M nonempty disjoint strata $\mathbb{B}_1, \ldots, \mathbb{B}_M$. Then, since $\cup_{j=1}^J \mathbb{A}_j \times \cup_{m=1}^M \mathbb{B}_m = \cup_{j=1}^J \cup_{m=1}^M \mathbb{A}_j \times \mathbb{B}_m$,

$$\operatorname{supp}(Y^*, X^*) = \begin{cases} \cup_{j=1}^J \cup_{m=1}^M \mathbb{A}_j \times \mathbb{B}_m & \text{if both } Y^* \text{ and } X^* \text{ are stratified} \\ \cup_{j=1}^J (\mathbb{A}_j \times \operatorname{supp}(X^*)) & \text{if only } Y^* \text{ is stratified} \\ \cup_{m=1}^M (\operatorname{supp}(Y^*) \times \mathbb{B}_m) & \text{if only } X^* \text{ is stratified.} \end{cases}$$

Therefore, if both Y^* and X^* are stratified, then $\operatorname{supp}(Z^*) = \cup_{l=1}^L \mathbb{C}_l$ with $L = JM$ and each $\mathbb{C}_l = \mathbb{A}_j \times \mathbb{B}_m$ for some $(j, m) \in \{1, \ldots, J\} \times \{1, \ldots, M\}$. This is the most general case for which $P_{Y|X}$ is given by (2.5).[7]

In contrast, simplifications occur if the stratification is endogenous or exogenous. For instance, if only Y^* is stratified, then $\operatorname{supp}(Z^*) = \cup_{l=1}^L \mathbb{C}_l$ with $L = J$ and $\mathbb{C}_l = \mathbb{A}_l \times \operatorname{supp}(X^*)$, which implies that, for $(y, x) \in \operatorname{supp}(Y^*) \times \operatorname{supp}(X^*)$,

$$b(y, x) = \sum_{l=1}^J p_l \mathbb{1}_{\mathbb{A}_l \times \operatorname{supp}(X^*)}(y, x) = \sum_{l=1}^J p_l \mathbb{1}_{\mathbb{A}_l}(y) =: b_{\text{endog}}(y).$$

Hence, by (2.5), we have that, for $(y, x) \in \operatorname{supp}(Y^*) \times \operatorname{supp}(X^*)$,

$$\text{endogenous stratification} \implies dP_{Y|X=x}(y) = \frac{b_{\text{endog}}(y)}{\gamma^*_{\text{endog}}(x)} \, dP^*_{Y^*|X^*=x}(y), \tag{2.7}$$

where $\gamma^*_{\text{endog}}(x) := \mathbb{E}_{P^*_{Y^*|X^*}}[b_{\text{endog}}(Y^*)|X^* = x]$.

If only X^* is stratified, then $\operatorname{supp}(Z^*) = \cup_{l=1}^L \mathbb{C}_l$ with $L = M$ and $\mathbb{C}_l = \operatorname{supp}(Y^*) \times \mathbb{B}_l$. Consequently, for $(y, x) \in \operatorname{supp}(Y^*) \times \operatorname{supp}(X^*)$,

$$b(y, x) = \sum_{l=1}^M p_l \mathbb{1}_{\operatorname{supp}(Y^*) \times \mathbb{B}_l}(y, x) = \sum_{l=1}^M p_l \mathbb{1}_{\mathbb{B}_l}(x) =: b_{\text{exog}}(x),$$

which implies that $\gamma^*_{\text{exog}}(x) := \mathbb{E}_{P^*_{Y^*|X^*}}[b_{\text{exog}}(X^*)|X^* = x] = b_{\text{exog}}(x)$. Hence, by (2.5),

$$\text{exogenous stratification} \implies dP_{Y|X=x}(y) = dP^*_{Y^*|X^*=x}(y) \tag{2.8}$$

for $(y, x) \in \operatorname{supp}(Y^*) \times \operatorname{supp}(X^*)$. Consequently, exogenous stratification can be ignored, at least as far as consistent estimation is concerned. However, as the following example demonstrates, ignoring endogenous stratification does not lead to a consistent estimator.

Example 2.1 (Linear regression with exogenous regressors)**.** Consider the linear regression model $Y^* = \widetilde{X^*}'\theta^* + \varepsilon^*$, where $\widetilde{X^*} := (1, X^*)$. Assume that the regressors are exogenous with respect to the model error in the target population, i.e., $\mathbb{E}_{P^*_{Y^*|X^*}}[\varepsilon^*|X^*] = 0$ $P^*_{X^*}$-a.s..

[7] Unless mentioned otherwise, it is assumed throughout the paper that both Y^* and X^* are stratified.

Suppose that only Y^* is stratified. If we ignore the fact that the data were collected by VP sampling and simply regress the observed Y on the observed X and the constant regressor, then θ^* cannot be consistently estimated by the least-squares (LS) estimator. Indeed, letting $\hat{\theta}_{\text{LS}}$ denote the LS estimator obtained by regressing Y on $\tilde{X} := (1, X)$, we have that

$$\begin{aligned}\text{plim}_{n\to\infty}\,\hat{\theta}_{\text{LS}} &= \text{plim}_{n\to\infty}(n^{-1}\sum_{j=1}^{n}\tilde{X}_j\tilde{X}_j')^{-1}(n^{-1}\sum_{j=1}^{n}\tilde{X}_jY_j)\\ &= (\mathbb{E}_{P_X}\tilde{X}\tilde{X}')^{-1}(\mathbb{E}_P\tilde{X}Y)\\ &= (\mathbb{E}_{P_X}\tilde{X}\tilde{X}')^{-1}(\mathbb{E}_{P_X}\tilde{X}\mu(X)),\end{aligned} \tag{2.9}$$

where $\mu(X) := \mathbb{E}_{P_{Y|X}}[Y|X]$. But, $\mathbb{E}_{P_X}\tilde{X}\tilde{X}' \overset{(2.4)}{=} \mathbb{E}_{P^*_{X^*}}\gamma^*_{\text{endog}}(X^*)\widetilde{X^*}\widetilde{X^*}'/b^*$ and

$$\begin{aligned}\mu(x) &:= \mathbb{E}_{P_{Y|X}}[Y|X=x] && (x \in \text{supp}(X^*))\\ &= \frac{1}{\gamma^*_{\text{endog}}(x)}\mathbb{E}_{P^*_{Y^*|X^*}}[Y^*b_{\text{endog}}(Y^*)|X^*=x] && ((2.7))\\ &= \frac{1}{\gamma^*_{\text{endog}}(x)}\mathbb{E}_{P^*_{Y^*|X^*}}[(\widetilde{X^*}'\theta^*+\varepsilon^*)b_{\text{endog}}(Y^*)|X^*=x]\\ &= \tilde{x}'\theta^* + \frac{1}{\gamma^*_{\text{endog}}(x)}\mathbb{E}_{P^*_{Y^*|X^*}}[\varepsilon^*b_{\text{endog}}(Y^*)|X^*=x].\end{aligned} \tag{2.10}$$

Hence, writing (2.9) in terms of $P^*_{X^*}$,

$$\begin{aligned}\text{plim}_{n\to\infty}\,\hat{\theta}_{\text{LS}} &= (\mathbb{E}_{P^*_{X^*}}\frac{\gamma^*_{\text{endog}}(X^*)}{b^*}\widetilde{X^*}\widetilde{X^*}')^{-1}(\mathbb{E}_{P^*_{X^*}}\frac{\gamma^*_{\text{endog}}(X^*)}{b^*}\widetilde{X^*}\mu(X^*)) && ((2.9)\ \&\ (2.4))\\ &\overset{(2.10)}{=} (\mathbb{E}_{P^*_{X^*}}\frac{\gamma^*_{\text{endog}}(X^*)}{b^*}\widetilde{X^*}\widetilde{X^*}')^{-1}\\ &\qquad\times(\mathbb{E}_{P^*_{X^*}}\frac{\gamma^*_{\text{endog}}(X^*)}{b^*}\widetilde{X^*}[\widetilde{X^*}'\theta^* + \frac{1}{\gamma^*_{\text{endog}}(X^*)}\mathbb{E}_{P^*_{Y^*|X^*}}[\varepsilon^*b_{\text{endog}}(Y^*)|X^*]])\\ &= \theta^* + (\mathbb{E}_{P^*_{X^*}}\gamma^*_{\text{endog}}(X^*)\widetilde{X^*}\widetilde{X^*}')^{-1}(\mathbb{E}_{P^*}\widetilde{X^*}\varepsilon^*b_{\text{endog}}(Y^*))\\ &\neq \theta^*,\end{aligned}$$

because $\mathbb{E}_{P^*_{Y^*|X^*}}[\varepsilon^*|X^*] = 0$ ($P^*_{X^*}$-a.s.) does not imply that $\mathbb{E}_{P^*}\widetilde{X^*}\varepsilon^*b_{\text{endog}}(Y^*) = 0$.

If, however, stratification is exogenous, then

$$\begin{aligned}\mu(x) = \mathbb{E}_{P_{Y|X}}[Y|X=x] &\overset{(2.8)}{=} \mathbb{E}_{P^*_{Y^*|X^*}}[Y^*|X^*=x] && (x \in \text{supp}(X^*))\\ &= \mathbb{E}_{P^*_{Y^*|X^*}}[\widetilde{X^*}'\theta^* + \varepsilon^*|X^*=x]\\ &= \tilde{x}'\theta^*.\end{aligned} \tag{2.11}$$

Hence, ignoring exogenous stratification does not affect the consistency of $\hat{\theta}_{\text{LS}}$ because

$$\begin{aligned}
\text{plim}_{n\to\infty}\,\hat{\theta}_{\text{LS}} &= (\mathbb{E}_{P_X}\tilde{X}\tilde{X}')^{-1}(\mathbb{E}_{P_X}\tilde{X}\mu(X)) && ((2.9))\\
&= (\mathbb{E}_{P^*_{X^*}}\frac{\gamma^*_{\text{exog}}(X^*)}{b^*}\widetilde{X^*}\widetilde{X^*}')^{-1}(\mathbb{E}_{P^*_{X^*}}\frac{\gamma^*_{\text{exog}}(X^*)}{b^*}\widetilde{X^*}\mu(X^*)) && ((2.4))\\
&= (\mathbb{E}_{P^*_{X^*}}\frac{\gamma^*_{\text{exog}}(X^*)}{b^*}\widetilde{X^*}\widetilde{X^*}')^{-1}(\mathbb{E}_{P^*_{X^*}}\frac{\gamma^*_{\text{exog}}(X^*)}{b^*}\widetilde{X^*}\widetilde{X^*}'\theta^*) && ((2.11))\\
&= \theta^*.
\end{aligned}$$

However, as shown subsequently (cf. Example 3.1), $\hat{\theta}_{\text{LS}}$ is not asymptotically efficient. Hence, ignoring exogenous stratification does not affect the consistency of the LS estimator,[8] but it does affect its efficiency. □

3. INFERENCE

3.1. **Related literature and our contribution.** There is a large literature on estimation and testing models using data collected by various types of stratified sampling schemes; cf. the papers cited at the beginning of Section 2.2, and the references therein. In this section, we briefly describe only some of the works that consider VP sampling.

Earlier papers in the literature on estimating models with conditioning variables assume that $P^*_{Y^*|X^*}$ is known up to a finite dimensional parameter; only $P^*_{X^*}$ is left completely unspecified. E.g., a well known application of VP sampling can be found in Hausman & Wise (1981). Imbens & Lancaster (1996) extend the maximum likelihood approach of Hausman and Wise to a moment based methodology that allows for VP sampling, mixed response variables, and stratification on exogenous covariates. Regression under VP sampling and a parametric $P^*_{Y^*|X^*}$ has also been investigated. E.g., Jewell (1985) and Quesenberry & Jewell (1986) propose iterative estimators of regression coefficients under VP sampling without imposing normality or independence, though they do not provide any asymptotic theory for their estimators.

The papers described above impose strong conditions on the distribution of $Y^*|X^*$. Exceptions include Wooldridge (1999) and Tripathi (2011b), who leave both $P^*_{Y^*|X^*}$ and $P^*_{X^*}$ completely unspecified. Wooldridge provides asymptotic theory for M-estimation under VP sampling for a model defined in terms of a set of just-identified unconditional moment equalities, whereas Tripathi considers optimal generalized method of moments (GMM) estimation in unconditional moment restriction models that allow for the parameter of interest to be overidentified. The major difference between (2.1) and the models in the papers of Wooldridge and Tripathi is that (2.1) is a conditional moment restriction, whereas the moment conditions in

[8] Tripathi (2011b) shows that in unconditional moment restriction models even exogenous stratification cannot be ignored.

the aforementioned papers are all unconditional. Therefore, (2.1) nests the moment conditions of Wooldridge and Tripathi as a special case.

In this paper, we show how to efficiently estimate θ^* and the aggregate shares using a smoothed empirical likelihood based approach. The results presented here answer the question posed in Wooldridge (1999, p. 1402) by providing efficiency bounds for models with conditional moment restrictions under VP sampling and showing that these bounds are attainable.

Furthermore, the results in this paper are also directly applicable to a class of "biased sampling" problems. To see this, recall that the phenomenon where the realized probability distribution P differs from the target probability distribution P^* is generically referred to as selection bias.[9] It is useful to note that the class of problems that can be handled when selection is modeled using (2.3) includes more than just those involving stratified sampling. For instance, consider the so called "length biased sampling" problem where the probability of observing a random variable is proportional to its "size." E.g., economists are more likely to observe longer unemployment spells if they are sampled at a randomly chosen time. Similarly, as Owen (2001, p. 127) points out, if internet log files are sampled randomly then longer sessions are likely to be over-represented. It is useful to examine length biased sampling in the context of VP sampling, because in length biased sampling we have

$$dP(z) := \frac{\|z\|}{\mathbb{E}_{P^*}\|Z^*\|}\, dP^*(z), \qquad z \in \mathbb{R}^{\dim(Z^*)},$$

where $\|\cdot\|$ is the Euclidean norm. That is, length biased sampling can be expressed as (2.3) with $b(z) := \|z\|$ and $b^* := \mathbb{E}_{P^*}\|Z^*\|$. Therefore, with only minor notational changes, the results obtained in this paper can be extended to length biased sampling as well.

Length biased sampling has been extensively studied for the parametric case, i.e., where dP^* is specified up to a finite dimensional parameter. Cf., e.g., Patil & Rao (1977, 1978), Bickel, Klassen, Ritov, & Wellner (1993, Section 4.4), and Owen (2001, Chapter 6). As far as a nonparametric treatment of length biased sampling is concerned, Vardi (1982) deals with the case when P^* is unknown. Vardi assumes that both P^* and P can be sampled with positive probability. Using two independent samples (one each from P^* and P), he shows how to construct the nonparametric maximum likelihood estimators (NPMLE) of P^* and P, and also obtains their asymptotic distributions. Vardi (1985), and Gill, Vardi, & Wellner (1988), provide conditions for the existence and uniqueness of the NPMLE of P^* in a general setup when more than two independent samples from F^* and F are available. These papers concentrate on the distributions P^* and P; there are no other parameters to estimate. Qin (1993) uses the empirical likelihood approach to construct a nonparametric likelihood ratio confidence interval for $\theta^* := \mathbb{E}_{P^*}Z^*$, i.e., a just-identified unconditional moment equality, using an independent

[9] Hence, for the LS estimator in Example 2.1, one can say that it is inconsistent because of selection bias due to endogenous stratification, whereas exogenous stratification does not lead to any selection bias.

sample from P^* and P. El-Barmi & Rothmann (1998) generalize Qin's treatment to handle models with overidentified unconditional moment restrictions of the form $\mathbb{E}_{P^*}g(Z,\theta^*) = 0$. They also obtain efficient estimators of P^* and P. However, they do not consider the testing of overidentifying restrictions.

3.2. **Efficiency bounds.** The efficiency bounds for estimating θ^* and related functionals have been derived in Severini & Tripathi (2013, Section 14.3). In this section, we describe some of these bounds and discuss their salient features. Construction of estimators that achieve these bounds is considered in the next section.

For the remainder of the paper, let $\rho_1(Z,\theta) := g(Z,\theta)/b(Z)$. Since the right hand side of (2.6) is a conditional moment equality with respect to the realized conditional distribution $P_{Y|X}$, the efficiency bound for θ^* follows from Chamberlain (1987). Namely, the efficiency bound for estimating θ^* is given by[10]

$$\text{l.b.}(\theta^*) := (\mathbb{E}_{P_X} D'(X)V_1^{-1}(X)D(X))^{-1}, \tag{3.1}$$

where $D(X) := \partial_\theta \mathbb{E}_{P_{Y|X}}[\rho_1(Z,\theta^*)|X]$ and $V_1(X) := \mathbb{E}_{P_{Y|X}}[\rho_1(Z,\theta^*)\rho_1'(Z,\theta^*)|X]$.

The efficiency bound in (3.1), given as a functional of the realized distribution P, can be used to determine whether an estimator of θ^* is semiparametrically efficient by comparing its asymptotic variance with l.b.(θ^*). However, as the moment condition model (2.1) is specified in terms of the target distribution P^*, in order to answer questions such as how the efficiency bound for θ^* changes if stratification is purely endogenous (or purely exogenous) or if the error term in a regression model is conditionally homoskedastic in the target population, it is helpful to rewrite (3.1) in terms of P^*. To do so, observe that, by (2.5), we have

$$\begin{aligned} D(x) &= \frac{1}{\gamma^*(x)}\partial_\theta \mathbb{E}_{P^*_{Y^*|X^*}}[g(Z^*,\theta^*)|X^*=x], \qquad x \in \text{supp}(X^*), \\ V_1(x) &= \frac{1}{\gamma^*(x)}\mathbb{E}_{P^*_{Y^*|X^*}}[\frac{g(Z^*,\theta^*)g'(Z^*,\theta^*)}{b(Y^*,x)}|X^*=x]. \end{aligned} \tag{3.2}$$

Hence, by (2.4) and (3.2), the efficiency bound in (3.1) can be written as

$$\text{l.b.}(\theta^*) = b^*(\mathbb{E}_{P^*_{X^*}}(\partial_\theta \mathbb{E}_{P^*_{Y^*|X^*}}[g(Z^*,\theta^*)|X^*])'V_b^{*-1}(X^*)(\partial_\theta \mathbb{E}_{P^*_{Y^*|X^*}}[g(Z^*,\theta^*)|X^*]))^{-1}, \tag{3.3}$$

where $V_b^*(X^*) := \mathbb{E}_{P^*_{Y^*|X^*}}[g(Z^*,\theta^*)g'(Z^*,\theta^*)/b(Z^*)|X^*]$.

We can use (3.3) to determine the efficiency bound for estimating θ^* if stratification is purely endogenous or purely exogenous. For instance, the efficiency bound when only Y^* is stratified follows from (3.3) on replacing $b(Z^*)$ in the definition of $V_b^*(X^*)$ with $b_{\text{endog}}(Y^*)$. Similarly, the bound when only X^* is stratified follows from (3.3) on replacing $b(Z^*)$ in the definition of $V_b^*(X^*)$ with $b_{\text{exog}}(X^*)$.

[10]The abbreviation "l.b." stands for "lower bound," because the efficiency bound is the greatest lower bound for the asymptotic variance of any $n^{1/2}$-consistent regular estimator.

If there is no conditioning in (2.1), i.e., X^* is constant $P^*_{X^*}$-a.s., and $\dim(g) \geq \dim(\theta^*)$, then (3.3) reduces to the efficiency bound for estimating θ^* in unconditional moment restriction models when observations are collected by VP sampling (Severini & Tripathi, 2013, Section 14.2.1). At the other extreme, if there is no stratification, i.e., $L = 1 = p_1$ and $\mathbb{C}_1 = \operatorname{supp}(Z^*)$, so that $Z^* = Z$ and $P^* = P$, then the efficiency bound in (3.3) becomes

$$(\mathbb{E}(\partial_\theta \mathbb{E}[g(Z^*, \theta^*)|X^*])'(\mathbb{E}[g(Z^*, \theta^*)g'(Z^*, \theta^*)|X^*])^{-1}(\partial_\theta \mathbb{E}[g(Z^*, \theta^*)|X^*]))^{-1},$$

which is Chamberlain's 1987 bound for estimating θ^* in the absence of any selection.

The next example uses (3.3) to determine the efficiency bound for θ^* under various scenarios.

Example 3.1 (Example 2.1 contd.)**.** Here, $g(Z^*, \theta) = Y^* - \widetilde{X^*}'\theta$ and the efficiency bound for estimating θ^* is given by

$$\text{l.b.}(\theta^* \text{ in Example 2.1}) \overset{(3.3)}{=} b^* \left(\mathbb{E}_{P^*_{X^*}} \frac{\widetilde{X^*}\widetilde{X^*}'}{V^*_b(X^*)} \right)^{-1} \overset{(2.4),(2.5)}{=} \left(\mathbb{E}_{P_X} \frac{\tilde{X}\tilde{X}'}{\gamma^{*2}(X) V_1(X)} \right)^{-1}. \tag{3.4}$$

If stratification is endogenous, then

$$\text{l.b.}(\theta^* \text{ in Example 2.1})\big|_{\text{endog. strat.}} \overset{(3.4)}{=} \left(\mathbb{E}_{P_X} \frac{\tilde{X}\tilde{X}'}{\gamma^{*2}_{\text{endog}}(X) V_{1,\text{endog}}(X)} \right)^{-1},$$

where $V_{1,\text{endog}}(X) := \mathbb{E}_{P_{Y|X}}[(Y - \tilde{X}'\theta^*)^2 / b^2_{\text{endog}}(Y)|X]$.

In contrast, if stratification is exogenous then

$$\text{l.b.}(\theta^* \text{ in Example 2.1})\big|_{\text{exog. strat.}} \overset{(3.4)}{=} \left(\mathbb{E}_{P_X} \frac{\tilde{X}\tilde{X}'}{V_{1,\text{exog}}(X)} \right)^{-1}, \tag{3.5}$$

where $V_{1,\text{exog}}(X) := \mathbb{E}_{P_{Y|X}}[(Y - \tilde{X}'\theta^*)^2|X]$.

Recall from Example 2.1 that, under exogenous stratification, the LS estimator consistently estimates θ^*. Since $n^{1/2}(\hat{\theta}_{\text{LS}} - \theta^*)$ is asymptotically (as $n \to \infty$) normal with mean zero and variance $V_{\text{LS,exog}} := (\mathbb{E}_{P_X}\tilde{X}\tilde{X}')^{-1}(\mathbb{E}_{P_X}\tilde{X}\tilde{X}'V_{1,\text{exog}}(X))(\mathbb{E}_{P_X}\tilde{X}\tilde{X}')^{-1}$, an application of a matrix version of the Cauchy-Schwarz inequality reveals that

$$\text{l.b.}(\theta^* \text{ in Example 2.1})\big|_{\text{exog. strat.}} \leq_L V_{\text{LS,exog}},$$

where $\leq_L$ is the usual (Löwner) order on the set of symmetric matrices.[11] Therefore, under exogenous stratification, the LS estimator is consistent but not asymptotically efficient. However, if stratification is exogenous and ε^* is conditionally homoskedastic in the target population, then (3.5) and (A.3) imply that the LS estimator is asymptotically efficient.

[11]Namely, $M_1 \leq_L M_2$ for symmetric matrices M_1, M_2 means that $M_1 - M_2$ is negative semidefinite.

Even under endogenous stratification, it is not difficult to obtain an estimator of θ^* that is consistent but asymptotically inefficient. To see this, assume that only Y^* is stratified. Then,

$$\begin{aligned}
&\mathbb{E}_{P^*_{Y^*|X^*}}[Y^* - \widetilde{X^*}'\theta^* | X^*] = 0 \quad P^*_{X^*}\text{-a.s.} \\
&\iff \mathbb{E}_{P_{Y|X}}[\frac{Y - \tilde{X}'\theta^*}{b_{\text{endog}}(Y)} | X] = 0 \quad P_X\text{-a.s.} && ((2.6)\ \&\ (2.7)) \\
&\implies \mathbb{E}_P \tilde{X}[\frac{Y - \tilde{X}'\theta^*}{b_{\text{endog}}(Y)}] = 0.
\end{aligned}$$

Hence, it is straightforward to show that the GMM estimator

$$\hat{\theta}_{\text{GMM,endog}} := \left(\sum_{j=1}^{n} \frac{\tilde{X}_j \tilde{X}_j'}{b_{\text{endog}}(Y_j)} \right)^{-1} \left(\sum_{j=1}^{n} \frac{\tilde{X}_j Y_j}{b_{\text{endog}}(Y_j)} \right)$$

is consistent for θ^*.[12] However, $\hat{\theta}_{\text{GMM,endog}}$ is not asymptotically efficient because its asymptotic variance is $V_{\text{GMM,endog}} := (\mathbb{E}_P \tilde{X}\tilde{X}'/b_{\text{endog}}(Y))^{-1}(\mathbb{E}_{P_X}\tilde{X}\tilde{X}'V_{1,\text{endog}}(X))(\mathbb{E}_P \tilde{X}\tilde{X}'/b_{\text{endog}}(Y))^{-1}$ but

$$\text{l.b.}(\theta^* \text{ in Example 2.1})\big|_{\text{endog. strat.}} \leq_L V_{\text{GMM,endog}}.$$

Analogous to $\hat{\theta}_{\text{GMM,endog}}$, the GMM estimator under exogenous stratification is

$$\hat{\theta}_{\text{GMM,exog}} := \left(\sum_{j=1}^{n} \frac{\tilde{X}_j \tilde{X}_j'}{b_{\text{exog}}(X_j)} \right)^{-1} \left(\sum_{j=1}^{n} \frac{\tilde{X}_j Y_j}{b_{\text{exog}}(X_j)} \right),$$

which is also not asymptotically efficient because its asymptotic variance is $V_{\text{GMM,exog}} := (\mathbb{E}_{P_X}\tilde{X}\tilde{X}'/b_{\text{exog}}(X))^{-1}(\mathbb{E}_{P_X}\tilde{X}\tilde{X}'V_{1,\text{exog}}(X)/b^2_{\text{exog}}(X))(\mathbb{E}_{P_X}\tilde{X}\tilde{X}'/b_{\text{exog}}(X))^{-1}$ but

$$\text{l.b.}(\theta^* \text{ in Example 2.1})\big|_{\text{exog. strat.}} \leq_L V_{\text{GMM,exog}}.$$

Constructing efficient estimators requires more effort. For instance, suppose that stratification is purely exogenous. Then, following Robinson (1987), it can be shown that the asymptotic variance of $\hat{\theta}_{\text{Robinson}} := (\sum_{j=1}^{n} \tilde{X}_j\tilde{X}_j'/\hat{\sigma}^2(X_j))^{-1}(\sum_{j=1}^{n} \tilde{X}_j Y_j/\hat{\sigma}^2(X_j))$ equals (3.5), where $\hat{\sigma}^2$ denotes a consistent estimator of $V_{1,\text{exog}}$. Hence, $\hat{\theta}_{\text{Robinson}}$ is an asymptotically efficient estimator of θ^* under exogenous stratification. A general approach, which can be used to construct efficient estimators irrespective of whether stratification is endogenous, exogenous, or both, is discussed in Section 3.3. □

[12] The estimator $\hat{\theta}_{\text{GMM,endog}}$ is an example of an inverse probability weighted (IPW) estimator, which uses the weights $1/b_{\text{endog}}(Y_1), \ldots, 1/b_{\text{endog}}(Y_n)$ to correct the selection bias due to stratification by downward weighting the strata that are oversampled and upward weighting the strata that are undersampled.

Since the aggregate shares add up to one, it suffices to determine the efficiency bound for estimating $Q^*_{-L} := (Q^*_1, \dots, Q^*_{L-1})_{(L-1)\times 1} \in (0,1)^{L-1}$. The aggregate shares are identified in the realized population by the moment condition

$$\mathbb{E}_P[\frac{s(Z) - Q^*_{-L}}{b(Z)}] = 0, \tag{3.6}$$

where $s(Z) := (\mathbb{1}_{\mathbb{C}_1}(Z), \dots, \mathbb{1}_{\mathbb{C}_{L-1}}(Z))_{(L-1)\times 1}$. The moment conditions in (3.6) modify accordingly if stratification is endogenous or exogenous; namely,

$$\begin{aligned} &\text{endog. strat.} \implies \begin{cases} \mathbb{E}_{P_Y}[\frac{s_{\text{endog}}(Y) - Q^*_{-L}}{b_{\text{endog}}(Y)}] = 0 \\ s_{\text{endog}}(Y) := (\mathbb{1}_{\mathbb{A}_1}(Y), \dots, \mathbb{1}_{\mathbb{A}_{J-1}}(Y))_{(J-1)\times 1} \end{cases} \\ &\text{exog. strat.} \implies \begin{cases} \mathbb{E}_{P_X}[\frac{s_{\text{exog}}(X) - Q^*_{-L}}{b_{\text{exog}}(X)}] = 0 \\ s_{\text{exog}}(X) := (\mathbb{1}_{\mathbb{B}_1}(X), \dots, \mathbb{1}_{\mathbb{B}_{M-1}}(X))_{(M-1)\times 1}. \end{cases} \end{aligned} \tag{3.7}$$

Let $\rho_2(Z, Q^*_{-L}) := (s(Z) - Q^*_{-L})/b(Z)$, and $\Sigma_{12}(X) := \mathbb{E}_{P_{Y|X}}[\rho_1(Z,\theta^*)\rho_2'(Z, Q^*_{-L})|X]$ be the conditional (on X) covariance between $\rho_1(Z,\theta^*)$ and $\rho_2(Z, Q^*_{-L})$. Then, under (2.1), the efficiency bound for estimating Q^*_{-L} is given by

$$\begin{aligned} \text{l.b.}(Q^*_{-L}) := b^{*2}[&\text{var}_P(\rho_2(Z, Q^*_{-L})) - (\mathbb{E}_{P_X}\Sigma_{12}'(X)V_1^{-1}(X)\Sigma_{12}(X)) \\ &+ (\mathbb{E}_{P_X}\Sigma_{12}'(X)V_1^{-1}(X)D(X))(\text{l.b.}(\theta^*))(\mathbb{E}_{P_X}D'(X)V_1^{-1}(X)\Sigma_{12}(X))], \end{aligned} \tag{3.8}$$

where $\text{l.b.}(\theta^*)$ is the efficiency bound for estimating θ^* given in (3.1).

In the absence of (2.1), the efficiency bound for Q^*_{-L} is given by $b^{*2}\,\text{var}_P(\rho_2(Z, Q^*_{-L}))$, which follows from standard GMM theory applied to (3.6). Hence, estimating the aggregate shares in the presence of (2.1) leads to efficiency gains under endogenous stratification. There are no efficiency gains for estimating Q^*_{-L} under exogenous stratification because

$$\begin{aligned} \text{exog. strat.} \implies \Sigma_{12}(X) &= \mathbb{E}_{P_{Y|X}}[\frac{g(Z,\theta^*)}{b_{\text{exog}}(X)}\frac{(s_{\text{exog}}(X) - Q^*_{-L})'}{b_{\text{exog}}(X)}|X] \\ &= \mathbb{E}_{P^*_{Y^*|X^*}}[\frac{g(Z^*,\theta^*)}{b_{\text{exog}}(X^*)}\frac{(s_{\text{exog}}(X^*) - Q^*_{-L})'}{b_{\text{exog}}(X^*)}|X^*] && \text{((2.8))} \\ &= \mathbb{E}_{P^*_{Y^*|X^*}}[g(Z^*,\theta^*)|X^*]\frac{(s_{\text{exog}}(X^*) - Q^*_{-L})'}{b^2_{\text{exog}}(X^*)} \\ &= 0 \qquad P^*_{X^*}\text{-a.s.} && \text{((2.1))} \\ &= 0 \qquad P_X\text{-a.s..} && (P^*_{X^*} \ll P_X \ll P^*_{X^*}) \end{aligned}$$

3.3. **Efficient estimation.** The estimation and testing techniques demonstrated here extend Kitamura, Tripathi, & Ahn (2004) and Tripathi & Kitamura (2003). These papers, which are based on a generalization of the empirical likelihood approach of Owen (1988), develop an asymptotically efficient methodology for estimating and testing models with conditional moment restrictions when the data are collected by random sampling.

In the papers of Kitamura, Tripathi, & Ahn, and Tripathi & Kitamura, kernel smoothing is used to efficiently incorporate the information implied by conditional moment restrictions into an empirical likelihood, which is henceforth referred to as the "smoothed empirical likelihood (SEL)." As shown in these papers, maximizing the SEL leads to one-step estimators which avoid any preliminary estimation of optimal instruments. It also yields internally studentized likelihood ratio-type statistics for testing H_0 and parametric restrictions on θ^* that do not require preliminary estimation of any variance terms. Moreover, the resulting estimation and testing procedures are invariant to normalizations of H_0. Simulation results presented in the aforementioned papers suggest that the SEL based approach can work very well in finite samples.

The advantages of the SEL approach described above extend to the case when the observations are collected by VP sampling. Furthermore, it leads to a unified approach of estimating and testing models using stratified samples, which should appeal to applied economists and practitioners in the field. Therefore, we now demonstrate how to use the SEL approach to construct asymptotically efficient estimators, i.e., estimators with asymptotic variance equal to the efficiency bounds in Section 3.2.

If the focus is on efficient estimation of θ^* alone, then the equivalence in (2.6) reveals that replacing the moment function in Kitamura, Tripathi, and Ahn (Equation 2.1) with $\rho_1(Z,\theta^*)$ will deliver an asymptotically efficient estimator of θ^*.

But what about Q^*_{-L}? Although the aggregate shares $Q^*_{-L} \overset{(3.6)}{=} \mathbb{E}_P[s(Z)]/\mathbb{E}_P[1/b(Z)]$ can be simply estimated by their sample analogs, this estimator will not be efficient because it does not take (2.1) into account; cf. the discussion after (3.8). To construct an estimator of Q^*_{-L} that accounts for (2.1), we have to jointly estimate θ^* and Q^*_{-L}, which we do using the SEL approach.

For the remainder of the paper, assume that we have independent observations $Z_1,\dots,Z_n$ collected by VP sampling. Hence, these are i.i.d. draws from the realized density dP in (2.3). Our estimation approach relies on a smoothed version of empirical likelihood. This smoothing, or localization, is carried out using positive kernel weights $w_{ij} := \dfrac{\mathcal{K}_{b_n}(X_i - X_j)}{\sum_{k=1}^n \mathcal{K}_{b_n}(X_i - X_k)}$, $i,j = 1,\dots,n$, where $\mathcal{K}$ is a second order kernel, $\mathcal{K}_{b_n}(\cdot) := \mathcal{K}(\cdot/b_n)$, and b_n the bandwidth.

For $i,j = 1,\dots,n$, let p_{ij} denote the probability mass placed at (X_i, Z_j) by a discrete distribution with support $(X_1,\dots,X_n)\times(Z_1,\dots,Z_n)$. The collection of probabilities $(p_{ij})_{i,j=1}^n$ can be thought of as a set of nuisance parameters that includes the empirical distribution of the data. Using the kernel weights (w_{ij}) and the distribution (p_{ij}) construct the smoothed loglikelihood $\sum_{i=1}^n \sum_{j=1}^n w_{ij}\log p_{ij}$. Then, given (θ, Q_{-L}), concentrate out (p_{ij}) by solving the following optimization problem:

$$\max_{(p_{ij})} \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log p_{ij}$$

$$\text{s.t.} \quad p_{ij} \geq 0 \text{ for } i,j=1,\ldots,n, \quad \sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij} = 1, \tag{3.9}$$

$$\sum_{j=1}^{n}\rho_1(Z_j,\theta)p_{1j} = 0, \ \ldots, \ \sum_{j=1}^{n}\rho_1(Z_j,\theta)p_{nj} = 0, \ \sum_{i=1}^{n}\sum_{j=1}^{n}\rho_2(Z_j,Q_{-L})p_{ij} = 0.$$

If the convex hulls of $\{\rho_1(Z_1,\theta),\ldots,\rho_1(Z_n,\theta)\}$ and $\{\rho_2(Z_1,Q_{-L}),\ldots,\rho_2(Z_n,Q_{-L})\}$ contain the origin, then (3.9) can be solved by using Lagrange multipliers. In this case, it can be verified that the solution to (3.9) is given by

$$\hat{p}_{ij}(\theta,Q_{-L}) := \frac{1}{n}\left(\frac{w_{ij}}{1+\lambda_i'\rho_1(Z_j,\theta)+\mu'\rho_2(Z_j,Q_{-L})}\right), \qquad i,j=1,\ldots,n,$$

where the multipliers $\lambda_i := \lambda_i(\theta,Q_{-L})$ and $\mu := \mu(\theta,Q_{-L})$ solve

$$\begin{aligned} \sum_{j=1}^{n}\frac{w_{ij}\rho_1(Z_j,\theta)}{1+\lambda_i'\rho_1(Z_j,\theta)+\mu'\rho_2(Z_j,Q_{-L})} &= 0, \qquad i=1,\ldots,n, \\ \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{w_{ij}\rho_2(Z_j,Q_{-L})}{1+\lambda_i'\rho_1(Z_j,\theta)+\mu'\rho_2(Z_j,Q_{-L})} &= 0. \end{aligned} \tag{3.10}$$

The smoothed empirical loglikelihood of (θ,Q_{-L}) is given by

$$\begin{aligned} \mathrm{SEL}(\theta,Q_{-L}) &:= \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log\hat{p}_{ij}(\theta,Q_{-L}) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log\left(\frac{w_{ij}/n}{1+\lambda_i'\rho_1(Z_j,\theta)+\mu'\rho_2(Z_j,Q_{-L})}\right), \end{aligned} \tag{3.11}$$

where the multipliers solve (3.10).

The estimators of θ^* and Q^*_{-L} can, in principle, be defined to be the maximizers of $\mathrm{SEL}(\theta,Q_{-L})$. However, this leads to a constrained optimization problem because the Lagrange multipliers in $\mathrm{SEL}(\theta,Q_{-L})$ have to satisfy (3.10). To ease computation, it is better to convert the constrained optimization problem into an unconstrained optimization problem as follows. Begin by observing that, by (3.11),

$$\mathrm{SEL}(\theta,Q_{-L}) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log(w_{ij}/n) - \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log(1+\lambda_i'\rho_1(Z_j,\theta)+\mu'\rho_2(Z_j,Q_{-L})).$$

Furthermore,[13]

$$\lambda_1,\ldots,\lambda_n,\mu = \operatorname*{argmax}_{\tilde\lambda_1,\ldots,\tilde\lambda_n,\tilde\mu} \sum_{i=1}^n \sum_{j=1}^n w_{ij} \log(1+\tilde\lambda_i'\rho_1(Z_j,\theta)+\tilde\mu'\rho_2(Z_j,Q_{-L})). \tag{3.12}$$

Therefore, the estimators of θ^* and Q^*_{-L} are defined to be

$$(\hat\theta,\hat Q_{-L}) := \operatorname*{argmax}_{\theta,Q_{-L}} \mathrm{SEL}_{\mathbb{T}}(\theta,Q_{-L}), \tag{3.13}$$

where the "trimmed" SEL objective function

$$\begin{aligned}\mathrm{SEL}_{\mathbb{T}}(\theta,Q_{-L}) &:= -\max_{\tilde\lambda_1,\ldots,\tilde\lambda_n,\tilde\mu} \sum_{i=1}^n \mathbb{T}_{i,n} \sum_{j=1}^n w_{ij}\log(1+\tilde\lambda_i'\rho_1(Z_j,\theta)+\tilde\mu'\rho_2(Z_j,Q_{-L}))\\ &= -\max_{\tilde\mu}\sum_{i=1}^n \mathbb{T}_{i,n}\max_{\tilde\lambda_i}\sum_{j=1}^n w_{ij}\log(1+\tilde\lambda_i'\rho_1(Z_j,\theta)+\tilde\mu'\rho_2(Z_j,Q_{-L})).\end{aligned} \tag{3.14}$$

The trimming indicator $\mathbb{T}_{i,n} := \mathbb{1}(\hat h(X_i) \geq b_n^\tau)$, where $\hat h(X_i) := \sum_{j=1}^n \mathcal{K}_{b_n}(X_i - X_j)/(nb_n^{\dim(X)})$ and $\tau \in (0,1)$ is a trimming parameter, is incorporated in (3.14) to deal with the "denominator problem," namely, the instability of the local empirical loglikelihood $\sum_{j=1}^n w_{ij}\log(1+\tilde\lambda_i'\rho_1(Z_j,\theta)+\tilde\mu'\rho_2(Z_j,Q_{-L}))$ caused by the density of the conditioning variables becoming too small in the tails. Since $\mathbb{T}_{i,n} \xrightarrow{\mathrm{p}} 1$ as $n\to\infty$, this trimming scheme ensures that asymptotically no data is lost.

Following Kitamura, Tripathi, & Ahn, it can be shown that, under some regularity conditions, $\hat\theta$ and $\hat Q_{-L}$ are consistent, asymptotically normal, and asymptotically efficient, i.e., their asymptotic variances match the efficiency bounds.

3.4. **Testing.** The empirical likelihood approach provides a convenient unified environment for testing. For instance, suppose we want to test the parametric restriction $\tilde{\mathrm{H}}_0 : R(\theta^*) = 0$ against the alternative that $\tilde{\mathrm{H}}_0$ is false, where R is a vector of twice continuously differentiable functions such that $\operatorname{rank}\partial_\theta R(\theta^*) = \dim(R)$. Let

$$(\hat\theta_R,\hat Q_{-L,R}) := \operatorname*{argmax}_{\theta,Q_{-L}} \mathrm{SEL}_{\mathbb{T}}(\theta,Q_{-L}) \quad \text{s.t.} \quad R(\theta)=0.$$

A version of the likelihood ratio statistic for testing $\tilde{\mathrm{H}}_0$ is given by

$$\mathrm{LR} := 2[\mathrm{SEL}_{\mathbb{T}}(\hat\theta,\hat Q_{-L}) - \mathrm{SEL}_{\mathbb{T}}(\hat\theta_R,\hat Q_{-L,R})].$$

It can be shown that, under some regularity conditions, $\mathrm{LR} \xrightarrow[n\to\infty]{\mathrm{d}} \chi^2_{\dim(R)}$ whenever $\tilde{\mathrm{H}}_0$ is true. This result can be used to obtain the critical values for LR. Although a Wald statistic can also be constructed, it is less attractive than LR because the latter is internally studentized. As in parametric situations, LR can be inverted to obtain asymptotically valid confidence

[13]To see this, compare the first order conditions for (3.12) with (3.10).

intervals. A nice property of confidence intervals based on LR is that they are invariant to nonsingular transformations of the moment conditions. They also automatically satisfy natural range restrictions.

Since inference based on $\hat{\theta}$ is sensible only if (2.1) is true, it is important to devise a test for H_0 against the alternative that it is false. As we are dealing with conditional moment restrictions, any specification test which first converts (2.1) into a finite set of unconditional moment restrictions will not be consistent for testing H_0. However, using the equivalence in (2.6), a consistent test of H_0 is easily obtained by replacing moment function in Tripathi & Kitamura (2003, Equation 1.1) with $\rho_1(Z, \theta^*)$. Note that since (3.6) just identifies the aggregate shares, testing the specification of (2.1) and (3.6) jointly is equivalent to testing (2.1).

4. Simulation study

We now examine the finite sample behavior of the LS, GMM, and SEL estimators to illustrate the effects of estimating a simple linear regression model specified for the target population, when data is collected by VP sampling and stratification is either endogenous or exogenous. Code for the simulations is written in `R`, and the SEL estimator of the model parameters and aggregate shares defined in (3.13) is implemented using the algorithm in Owen (2013); cf. Appendix B for details.

4.1. **Design.** We consider the design in Kitamura et al. (Section 5), which has been used earlier by Cragg (1983) and Newey (1993). The model to be estimated is

$$Y^* = \beta_0^* + \beta_1^* X^* + \sigma^*(X^*)\varepsilon^*, \tag{4.1}$$

where $\mathbb{E}_{P^*_{Y^*|X^*}}[\varepsilon^*|X^*] = 0$ $P^*_{X^*}$-a.s., $\theta^* := (\beta_0^*, \beta_1^*) = (1,1)$, and $(\varepsilon^*, \log X^*) \overset{\mathrm{d}}{=} \mathrm{NIID}(0,1)$. We consider two specifications for the skedastic function in the target population: a (conditional) heteroskedastic design, relevant for applications, with $\sigma^*(X^*) := \sqrt{0.1 + 0.2X^* + 0.3X^{*2}}$; and a (conditional) homoskedastic design, essentially of theoretical interest, with $\sigma^*(X^*) := 1$.

The target population is partitioned into two strata. Under endogenous stratification, $\mathbb{A}_1 = (-\infty, 1.4)$ and $\mathbb{A}_2 = [1.4, \infty)$. Under exogenous stratification, $\mathbb{B}_1 = \mathbb{A}_1$ and $\mathbb{B}_2 = \mathbb{A}_2$. The aggregate shares for the four configurations are given in Table 1. The VP sampling probabilities are $(p_1, p_2) = (0.9, 0.3)$; i.e., the first stratum is heavily oversampled, irrespective of whether the stratification is endogenous or exogenous. Since it is typically strata with small aggregate shares that are oversampled, this sampling design focuses on endogenous stratification, which is the object of attention in most applications.

Tables 2 and 3 reports the summary statistics averaged across 1000 Monte Carlo replications for the LS estimator $\hat{\theta}_{\mathrm{LS}}$, the GMM estimators $\hat{\theta}_{\mathrm{GMM,endog}}$ and $\hat{\theta}_{\mathrm{GMM,exog}}$, and the SEL estimator $\hat{\theta}$.[14] Three sample sizes are considered, namely, $n = 50, 150, 500$. Tables 4 and 5, which summarize the simulation results for estimating Q_1^*, compare the GMM estimators based on the moment conditions in (3.7) with the SEL estimator $\hat{Q}_1$.

4.2. **Discussion.** Recall that the LS estimator is inconsistent under endogenous stratification, and consistent but generally inefficient under exogenous stratification; the GMM estimators are consistent but inefficient under endogenous and exogenous stratification; and the SEL estimator is consistent and asymptotically efficient irrespective of whether the stratification is endogenous or exogenous. Tables 2–5 largely confirm these results, at least as far as estimating the model parameters is concerned.

The inconsistency of the LS estimator of β_1^* under endogenous stratification is apparent from Tables 2 and 3 because the bias of the LS estimator, as a fraction of β_1^*, remains greater than 9% in magnitude under heteroskedasticity, and greater than 6% under homoskedasticity, as the sample size increases from 50 to 500.[15] In contrast, in both designs, the LS and GMM estimators under exogenous stratification are practically unbiased even when $n = 50$. Under exogenous stratification, the LS estimator has smaller sampling variance than the GMM estimator for each sample size. However, this finding can be mathematically justified only for homoskedastic designs (recall from Example 3.1 that the LS estimator is asymptotically efficient when stratification is exogenous and the error term in the regression model is conditionally homoskedastic in the target population). Indeed, as shown in Appendix A, cf. Example A.1, counterexamples can be constructed to show that in heteroskedastic designs the LS estimator can have higher sampling variance than the GMM estimator when stratification is exogenous.[16] Under endogenous stratification, the GMM estimator of the slope coefficient exhibits some bias

[14]The SEL estimator is implemented with $\mathbb{T}_{i,n} := 1$. To the best of our knowledge, how to choose an optimal data driven bandwidth for the SEL estimator remains an open problem. Consequently, we naively chose the bandwidth by repeating the simulation experiment on a grid of bandwidths and picking the one that minimized the average (across the simulation replications) RMSE of the SEL estimator of β_1^*. The naively chosen bandwidth, labeled c_n, is reported in Tables 2–5. For the sake of comparison, we also report the SEL estimator when the bandwidth is chosen using Silverman's rule of thumb, namely, $b_n = 1.06\,\widehat{\mathrm{sd}}(X)\,n^{-1/5}$. Since $\widehat{\mathrm{sd}}(X)$ depends on the data, the b_n reported in the tables is the value averaged across the simulations.

[15]This is even more so for the LS estimator of the intercept because, under endogenous stratification, the bias of the LS estimator of β_0^* is $\approx 18\%$ (resp. $\approx 41\%$) in magnitude for the heteroskedastic (resp. homoskedastic) design even when $n = 500$. For the reminder of this section, however, we only discuss the simulation results for the slope coefficient because it can be interpreted as an average partial effect. Results for the intercept, which is a pure level effect, are qualitatively very similar.

[16]It is shown in Appendix A, cf. (A.1), that $\mathrm{asvar}(n^{1/2}(\hat{\theta}_{\mathrm{GMM,exog}} - \theta^*)) - \mathrm{asvar}(n^{1/2}(\hat{\theta}_{\mathrm{LS}} - \theta^*)) = A + B$ holds under exogenous stratification, where the matrix A is positive semidefinite and the matrix B is negative semidefinite. Therefore, in general, it is not clear which estimator has smaller asymptotic variance. However, since $B = 0$ under conditional homoskedasticity, cf. (A.4), $\mathrm{asvar}(n^{1/2}(\hat{\theta}_{\mathrm{LS}} - \theta^*)) \leq_L \mathrm{asvar}(n^{1/2}(\hat{\theta}_{\mathrm{GMM,exog}} - \theta^*))$ holds under exogenous stratification and conditional homoskedasticity. Alternatively, under conditional

($\approx$ 2–4% in both designs) when $n = 50$, but the bias is very close to zero when $n = 500$. This is true whether the design is homoskedastic or heteroskedastic, although the magnitude of the bias is higher under heteroskedasticity.

Tables 2–5 reveal that the SEL estimator using the naively chosen bandwidth (c_n), described in Footnote 14, behaves very similarly to the SEL estimator using the Silverman's rule of thumb bandwidth (b_n). Hence, subsequent discussion regarding the SEL estimator is based on its implementation using the naively chosen bandwidth.

The SEL estimator of β_1^* is consistent whether stratification is exogenous or endogenous. In the heteroskedastic design, the SEL estimator exhibits some bias ($\approx$ 1%) under endogenous stratification when $n = 500$, although its bias under exogenous stratification is close to zero. Moreover, in the heteroskedastic design, the SEL estimator beats the GMM estimator in terms of the RMSE under each stratification scheme and for each sample size. Not surprisingly, the contrast between the two is most pronounced when $n = 500$; e.g., irrespective of the stratification scheme, the RMSE of the GMM estimator is at least 65% larger than the RMSE of the SEL estimator.

In the homoskedastic design, even though it exhibits some bias under endogenous and exogenous stratification when $n = 50$, the bias of the SEL estimator is close to zero for $n = 500$. However, its RMSE is larger than that of the GMM estimator even when $n = 500$. This finding, which corroborates the simulation results in Kitamura et al. (p. 1682), is likely due to the fact that the SEL estimator internally estimates the skedastic function nonparametrically to achieve semiparametric efficiency and is thus unable to take advantage of conditional homoskedasticity in small samples.

Tables 4 and 5 reveal that the GMM estimator of Q_1^* is consistent whether stratification is endogenous or exogenous. It exhibits some upward bias ($\approx$ 1–2%) in both designs and for both types of stratification when $n = 50$, but the bias is very close to zero when $n = 500$.[17] In both designs, the RMSE of the SEL estimator of Q_1^* is always slightly larger than the RMSE of the of the GMM estimator under endogenous stratification, implying that in small samples there appears to be no efficiency gain in estimating Q_1^* jointly with the model parameters. As can be seen from Tables 4 and 5, the increase in the RMSE of $\hat{Q}_1$ is due to its bias, because RMSE $\approx$ SE whenever the bias is small. This becomes clear on comparing the bias of $\hat{Q}_1$ under endogenous and exogenous stratification: the latter is always larger. The higher bias of $\hat{Q}_1$ under exogenous stratification is likely a design effect.

homoskedasticity, the Gauss-Markov theorem implies the same result because $\hat{\theta}_{\text{GMM,exog}}$ and $\hat{\theta}_{\text{LS}}$ are both linear and unbiased when stratification is exogenous.

[17]In Tables 4 and 5, the results under exogenous stratification are almost identical for the heteroskedastic and homoskedastic designs because $P^*(X^* \in \mathbb{B}_1)$ is not affected by conditional heteroskedasticity in Y^* (cf. Table 1).

5. CONCLUSION

When estimating or testing economic relationships, economists often discover that the data they plan to use is not drawn randomly from the target population for which they wish to draw an inference. Instead, the observations are found to be sampled from a related but different distribution. If this feature is not taken into account when doing statistical analysis, subsequent inference can be severely biased. In this paper, we show how to use a smoothed empirical likelihood approach to conduct efficient semiparametric inference in models characterized as conditional moment equalities when data is collected by variable probability sampling. Results from a simulation experiment suggest that the smoothed empirical likelihood based estimator can estimate the model parameters very well in small to moderately sized stratified samples.

ACKNOWLEDGEMENTS

We thank two anonymous referees and seminar participants at the 2017 "Econometrics of Complex Survey Data: Theory and Applications" workshop organized by the Bank of Canada, Ottawa, Canada, for helpful comments. The simulation experiments reported in this paper were carried out using the HPC facilities of the University of Luxembourg (Varrette et al., 2014, `http://hpc.uni.lu`).

Appendix A. COMPARING THE ASYMPTOTIC VARIANCE OF LS AND GMM ESTIMATORS UNDER EXOGENOUS STRATIFICATION

We begin by proving the assertion in Footnote 16, namely, that, under exogenous stratification, $\operatorname{asvar}(n^{1/2}(\hat\theta_{\mathrm{GMM,exog}} - \theta^*)) - \operatorname{asvar}(n^{1/2}(\hat\theta_{\mathrm{LS}} - \theta^*)) = A + B$, where the matrix A is positive semidefinite, the matrix B is negative semidefinite, and $B = 0$ under conditional homoskedasticity.

Recalling the expressions for $V_{\mathrm{GMM,exog}}$ and $V_{\mathrm{LS,exog}}$ in Example 3.1, we can write

$$\begin{aligned}
&\operatorname{asvar}(n^{1/2}(\hat\theta_{\mathrm{GMM,exog}} - \theta^*)) - \operatorname{asvar}(n^{1/2}(\hat\theta_{\mathrm{LS}} - \theta^*)) \\
&\qquad\qquad = \left(\mathbb{E}_{P_X} \frac{\tilde X \tilde X'}{b_{\mathrm{exog}}(X)}\right)^{-1} \Omega \left(\mathbb{E}_{P_X} \frac{\tilde X \tilde X'}{b_{\mathrm{exog}}(X)}\right)^{-1},
\end{aligned}$$

where

$$\begin{aligned}
\Omega := &\left(\mathbb{E}_{P_X} \tilde X \tilde X' \frac{V_{1,\mathrm{exog}}(X)}{b^2_{\mathrm{exog}}(X)}\right) \\
&- \left(\mathbb{E}_{P_X} \frac{\tilde X \tilde X'}{b_{\mathrm{exog}}(X)}\right) (\mathbb{E}_{P_X} \tilde X \tilde X')^{-1} (\mathbb{E}_{P_X} \tilde X \tilde X' V_{1,\mathrm{exog}}(X)) (\mathbb{E}_{P_X} \tilde X \tilde X')^{-1} \left(\mathbb{E}_{P_X} \frac{\tilde X \tilde X'}{b_{\mathrm{exog}}(X)}\right).
\end{aligned}$$

Next, letting $a_1 := \tilde{X}\sqrt{V_{1,\mathrm{exog}}(X)}/b_{\mathrm{exog}}(X)$ and $a_2 := (\mathbb{E}_{P_X}\tilde{X}\tilde{X}')^{-1}\tilde{X}/\sqrt{V_{1,\mathrm{exog}}(X)}$, we have

$$\begin{aligned}\Omega &= \mathbb{E}_{P_X}a_1a_1' - (\mathbb{E}_{P_X}a_1a_2')(\mathbb{E}_{P_X}\tilde{X}\tilde{X}'V_{1,\mathrm{exog}}(X))(\mathbb{E}_{P_X}a_2a_1') \\ &= \mathbb{E}_{P_X}a_1a_1' - (\mathbb{E}_{P_X}a_1a_2')(\mathbb{E}_{P_X}a_2a_2')^{-1}(\mathbb{E}_{P_X}a_2a_1') \\ &\qquad + (\mathbb{E}_{P_X}a_1a_2')[(\mathbb{E}_{P_X}a_2a_2')^{-1} - (\mathbb{E}_{P_X}\tilde{X}\tilde{X}'V_{1,\mathrm{exog}}(X))](\mathbb{E}_{P_X}a_2a_1').\end{aligned}$$

Consequently, under exogenous stratification we can write

$$\operatorname{asvar}(n^{1/2}(\hat{\theta}_{\mathrm{GMM,exog}} - \theta^*)) - \operatorname{asvar}(n^{1/2}(\hat{\theta}_{\mathrm{LS}} - \theta^*)) = A + B, \tag{A.1}$$

where

$$A := \left(\mathbb{E}_{P_X}\frac{\tilde{X}\tilde{X}'}{b_{\mathrm{exog}}(X)}\right)^{-1}[\mathbb{E}_{P_X}a_1a_1' - (\mathbb{E}_{P_X}a_1a_2')(\mathbb{E}_{P_X}a_2a_2')^{-1}(\mathbb{E}_{P_X}a_2a_1')]\left(\mathbb{E}_{P_X}\frac{\tilde{X}\tilde{X}'}{b_{\mathrm{exog}}(X)}\right)^{-1}$$

and

$$\begin{aligned}B := {}&\left(\mathbb{E}_{P_X}\frac{\tilde{X}\tilde{X}'}{b_{\mathrm{exog}}(X)}\right)^{-1}(\mathbb{E}_{P_X}a_1a_2') \\ &\times[(\mathbb{E}_{P_X}a_2a_2')^{-1} - (\mathbb{E}_{P_X}\tilde{X}\tilde{X}'V_{1,\mathrm{exog}}(X))]\left(\mathbb{E}_{P_X}\frac{\tilde{X}\tilde{X}'}{b_{\mathrm{exog}}(X)}\right)^{-1}(\mathbb{E}_{P_X}a_2a_1').\end{aligned}$$

It remains to show that A is positive semidefinite and B is negative semidefinite. To do so, recall the matrix version of the Cauchy-Schwarz inequality (Tripathi, 1999), namely,

$$(\mathbb{E}GH')(\mathbb{E}HH')^{-1}(\mathbb{E}HG') \le_L \mathbb{E}GG', \tag{A.2}$$

where G and H are random column vectors. Then, letting $G := a_1$ and $H := a_2$, it is immediate from (A.2) that A is positive semidefinite. Next,

$$\begin{aligned}(\mathbb{E}_{P_X}a_2a_2')^{-1} &= (\mathbb{E}_{P_X}\tilde{X}\tilde{X}')\left(\mathbb{E}_{P_X}\frac{\tilde{X}\tilde{X}'}{V_{1,\mathrm{exog}}(X)}\right)^{-1}(\mathbb{E}_{P_X}\tilde{X}\tilde{X}') \\ &\le_L \mathbb{E}_{P_X}\tilde{X}\tilde{X}'V_{1,\mathrm{exog}}(X)\end{aligned}$$

follows from (A.2) on letting $G := \tilde{X}\sqrt{V_{1,\mathrm{exog}}(X)}$ and $H := \tilde{X}/\sqrt{V_{1,\mathrm{exog}}(X)}$. Hence, B is negative semidefinite. Consequently, as A is positive semidefinite and B is negative semidefinite, it is not clear from (A.1) which estimator has smaller asymptotic variance.

However, if conditional homoskedasticity holds in the target population, then

$$\operatorname{var}_{P^*}(Y^*|X^*) = \sigma^{*2} \qquad P^*_{X^*}\text{-a.s.}$$

for some constant $\sigma^{*2} > 0$. Moreover, under exogenous stratification,

$$\operatorname{var}_{P^*}(Y^*|X^* = x) \overset{(2.8)}{=} \operatorname{var}_P(Y|X = x), \qquad x \in \operatorname{supp}(X^*).$$

Hence, since $P^*_{X^*} \ll P_X \ll P^*_{X^*}$, conditional homoskedasticity and exogenous stratification together imply that

$$V_{1,\text{exog}}(X) = \text{var}_P(Y|X) = \sigma^{*2} \qquad P_X\text{-a.s..} \tag{A.3}$$

Therefore, under conditional homoskedasticity and exogenous stratification,

$$\begin{aligned}
&(\mathbb{E}_{P_X} a_2 a_2')^{-1} - \mathbb{E}_{P_X} \tilde{X}\tilde{X}' V_{1,\text{exog}}(X) \\
&\quad = (\mathbb{E}_{P_X} \tilde{X}\tilde{X}') \left(\mathbb{E}_{P_X} \frac{\tilde{X}\tilde{X}'}{V_{1,\text{exog}}(X)} \right)^{-1} (\mathbb{E}_{P_X} \tilde{X}\tilde{X}') - \mathbb{E}_{P_X} \tilde{X}\tilde{X}' V_{1,\text{exog}}(X) \\
&\quad = \sigma^{*2} [\mathbb{E}_{P_X} \tilde{X}\tilde{X}' - \mathbb{E}_{P_X} \tilde{X}\tilde{X}'] \qquad ((\text{A.3})) \\
&\quad = 0.
\end{aligned}$$

It follows from the definition of B that

$$\text{conditional homoskedasticity and exogenous stratification} \implies B = 0. \tag{A.4}$$

Hence, $\text{asvar}(n^{1/2}(\hat{\theta}_{\text{LS}} - \theta^*)) \leq_L \text{asvar}(n^{1/2}(\hat{\theta}_{\text{GMM,exog}} - \theta^*))$ holds under exogenous stratification and conditional homoskedasticity.

However, as demonstrated in the next example, this result may not hold under conditional heteroskedasticity.

Example A.1. Consider (4.1) with $\beta_0 := 0$, i.e., a simple linear regression through the origin. As before, $\mathbb{E}_{P^*_{Y^*|X^*}}[\varepsilon^*|X^*] = 0$ $P^*_{X^*}$-a.s.. Assume that only $X^* := \begin{cases} c & \text{w.p. } 1-r \\ d & \text{w.p. } r \end{cases}$ is stratified with $L = 2$, where $\mathbb{B}_1 = (-\infty, 0)$ and $\mathbb{B}_2 = [0, +\infty)$.

Under exogenous stratification,

$$\text{asvar}(n^{1/2}(\hat{\beta}_{1,\text{LS}} - \beta_1^*)) = \frac{\mathbb{E}_{P_X} X^2 V_{1,\text{exog}}(X)}{(\mathbb{E}_{P_X} X^2)^2} = \frac{\mathbb{E}_{P_X} X^2 \sigma^{*2}(X)}{(\mathbb{E}_{P_X} X^2)^2}$$

$$\text{asvar}(n^{1/2}(\hat{\beta}_{1,\text{GMM}} - \beta_1^*)) = \frac{\mathbb{E}_{P_X}[X^2 V_{1,\text{exog}}(X)/b^2_{\text{exog}}(X)]}{(\mathbb{E}_{P_X} X^2/b_{\text{exog}}(X))^2} = \frac{\mathbb{E}_{P_X}[X^2 \sigma^{*2}(X)/b^2_{\text{exog}}(X)]}{(\mathbb{E}_{P_X} X^2/b_{\text{exog}}(X))^2}.$$

Let $r = 1/3$, $c = -1$, $d = 2$, $\sigma^{*2}(c) = 1$, $\sigma^{*2}(d) = 4$, $p_1 = 0.2$, and $p_2 = 0.8$. Note that $b_{\text{exog}}(c) = p_1 \mathbb{1}_{\mathbb{B}_1}(c) + p_2 \mathbb{1}_{\mathbb{B}_2}(c) = p_1$ because $c < 0$, and $b_{\text{exog}}(d) = p_1 \mathbb{1}_{\mathbb{B}_1}(d) + p_2 \mathbb{1}_{\mathbb{B}_2}(d) = p_2$ because $d > 0$. Then, it can be verified that

$$\mathbb{E}_{P_X} X^2 \sigma^{*2}(X) = 6, \quad \mathbb{E}_{P_X} X^2 = 2, \quad \mathbb{E}_{P_X}[X^2 \sigma^{*2}(X)/b^2_{\text{exog}}(X)] = 25, \quad \mathbb{E}_{P_X}[X^2/b_{\text{exog}}(X)] = 5.$$

Consequently,

$$\text{asvar}(n^{1/2}(\hat{\beta}_{1,\text{LS}} - \beta_1^*)) = 1.5 \quad > \quad \text{asvar}(n^{1/2}(\hat{\beta}_{1,\text{GMM}} - \beta_1^*)) = 1.$$

This shows that the LS estimator may be asymptotically inefficient compared to the GMM estimator under conditional heteroskedasticity and exogenous stratification. □

Appendix B. computation

In this appendix, we describe how the SEL estimator was implemented by adapting the code of Owen (2017). The `R` function `cemplik` in Owen (2017) was originally written for count random variables, and allows for ties in the data. Let $Z_j := (Y_j, X_j)$ be i.i.d. draws from the realized density dP, and assume that each of the n distinct values of Z_j can be taken by c_j distinct draws, so that the total sample size is $N := \sum_{j=1}^n c_j$. If we impose on the data the vector of unconditional moment equalities $\mathbb{E}_P m(Z, \theta) = 0$, then Owen (2017, p. 2) shows that the empirical loglikelihood, as a function of θ, and modulo constants not depending on θ, is obtained by solving (in our notation)

$$-\max_{\tilde{\lambda}} \sum_{j=1}^{n} c_j \log(1 + \tilde{\lambda}' m(Z_j, \theta)). \tag{B.1}$$

Note how in (B.1) the original sample size N has disappeared, and only the number n of distinct values of Z_j remains. The function `cemplik` asks for $\mathfrak{m} := (m(Z_1, \theta), \ldots, m(Z_n, \theta))$ and a vector $\mathfrak{c} := (c_1, \ldots, c_n)$ as inputs, and delivers three outputs:

1). The empirical loglikelihood (EL) for a given value of θ, computed at the vector $\lambda_{\dim(m)\times 1}$ of Lagrange multipliers that maximize (B.1), i.e.,

$$\mathrm{EL}_{\mathfrak{m}}(\theta; \mathfrak{c}, \lambda) := -\sum_{j=1}^{n} c_j \log(1 + \lambda' m(Z_j, \theta)).$$

2). The vector λ used to compute $\mathrm{EL}_{\mathfrak{m}}(\theta; \mathfrak{c}, \lambda)$.

3). The unconditional empirical probabilities

$$p_j := \frac{c_j}{n} \frac{1}{1 + \lambda' m(Z_j, \theta)}, \qquad j = 1, \ldots, n.$$

We now describe how to compute $\mathrm{SEL}_{\mathbb{T}}(\theta)$ when only the conditional moment restriction $\mathbb{E}_{P_{Y|X}}[\rho_1(Z, \theta)|X] = 0$ is imposed on the data. In the following, we do not deal with ties in the data.[18] Instead, we take advantage of the formal resemblance of the optimization problem in (B.1) with the one that leads to the smoothed empirical loglikelihood. Indeed, obtaining $\mathrm{SEL}_{\mathbb{T}}(\theta)$ only under $\mathbb{E}_{P_{Y|X}}[\rho_1(Z, \theta)|X] = 0$ is equivalent to solving (3.14) with $\rho_2 := 0$, i.e.,

$$\mathrm{SEL}_{\mathbb{T}}(\theta)\big|_{\rho_2=0} := -\max_{\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n} \sum_{i=1}^{n} \mathbb{T}_{i,n} \sum_{j=1}^{n} w_{ij} \log(1 + \tilde{\lambda}_i' \rho_1(Z_j, \theta)). \tag{B.2}$$

[18] In our setup, all the components of Z are continuous random variables, so that ties in the data occur with probability (P) zero .

From the first order conditions, it is clear that the maximizers in (B.2) can be recovered as solutions to n independent maximization problems, namely,

$$\lambda_i := \underset{\tilde{\lambda}_i}{\operatorname{argmax}} \sum_{j=1}^{n} w_{ij} \log(1 + \tilde{\lambda}_i' \rho_1(Z_j, \theta)), \qquad i = 1, \ldots, n. \tag{B.3}$$

The elements of $\mathfrak{c}$ in (B.1) are not constrained to be integers, but are only supposed to be positive. Hence, comparing (B.1) with (B.3), each λ_i can be obtained by invoking `cemplik` n times with $\mathfrak{c}_i := (w_{i1}, \ldots, w_{in})$ as the weights and $\mathfrak{m}$ replaced with $\rho_1 := (\rho_1(Z_1, \theta), \ldots, \rho_1(Z_n, \theta))$. Consequently,

$$\mathrm{SEL}_{\mathbb{T}}(\theta)\Big|_{\rho_2=0} = \sum_{i=1}^{n} \mathbb{T}_{i,n} \mathrm{EL}_{\rho_1}(\theta; \mathfrak{c}_i, \lambda_i) \tag{B.4}$$

with $\mathrm{EL}_{\rho_1}(\theta; \mathfrak{c}_i, \lambda) = \sum_{j=1}^{n} w_{ij} \log(1+\lambda' \rho_1(Z_j, \theta))$. The `R` commands used to implement (B.4) are as follows. Let `rho1` denote $(\rho_1(Z_1, \theta), \ldots, \rho_1(Z_n, \theta))$, `sel.weights` be the $n \times n$ matrix whose elements are the kernel weights w_{ij}, and `trim` the trimming vector $\mathbb{T}_{i,n}$. Then, $\mathrm{SEL}_{\mathbb{T}}(\theta)\big|_{\rho_2=0}$ is obtained with the following code:

```
emplik.list = apply(sel.weights, MARGIN=1, function(w) cemplik(rho1, w))
SEL = trim %*% unlist(lapply(emplik.list, '[[', 1))
```

Finally, we show how to impose a conditional and an unconditional moment restriction on the data, i.e., compute the objective function $\mathrm{SEL}_{\mathbb{T}_{i,n}}(\theta, Q_{-L})$ defined in (3.14). We treat the optimization problem in (3.14) as a two-step maximization. In the first step, we fix $\bar{\mu}$ and solve the n independent maximization problems

$$\max_{\tilde{\lambda}_i} \sum_{j=1}^{n} w_{ij} \log(1 + \tilde{\lambda}_i' \rho_1(Z_j, \theta) + \bar{\mu}' \rho_2(Z_j, Q_{-L})), \qquad i = 1, \ldots, n. \tag{B.5}$$

To carry out the maximizations in (B.5), we need to slightly modify Owen's `cemplik`. We wrote a function `cemplik2`, which receives an extra argument $\bar{\mu}' \rho_2(Z_j, Q_{-L})$. The new function `cemplik2` evaluates the logarithm in (B.3) at $1 + \tilde{\lambda}_i' \rho_1(Z_j, \theta) + \bar{\mu}' \rho_2(Z_j, Q_{-L})$ instead of at $1 + \tilde{\lambda}_i' \rho_1(Z_j, \theta)$. The second step needed to compute $\mathrm{SEL}_{\mathbb{T}_{i,n}}(\theta, Q_{-L})$ is a maximization over $\bar{\mu}$ as shown in (3.14), which can be carried out by a standard optimization routine as follows:[19]

```
muopt = optim(0, SmoothEmplik, rho1, rho2, sel.weights,
              method="Brent", lower=0, upper=1)$value
SmoothEmplik = function(mu, rho1, rho2, sel.weights){
  smooth.emplik.list = apply(sel.weights, MARGIN=1,
                             function(w) cemplik2(rho1, mu*rho2, w))
  SEL = trim %*% unlist(lapply(smooth.emplik.list, '[[', 1))
```

[19]This is a simplified but working version of the code we actually used. The complete code is available from GitHub at `https://github.com/Fifis/SELshares`.

```
    return(SEL)
}
```

The finite sample performance of the SEL estimator, implemented as described above, is discussed in Section 4. The simulation experiments in Section 4 were carried out on the high performance computing clusters at the University of Luxembourg. Table 6 gives some idea about the average time taken to complete one Monte-Carlo replication for the heteroskedastic design (the execution times under endogenous and exogenous stratification are very similar).

References

Bhattacharya, D. (2005). Asymptotic inference from multi-stage samples. *Journal of Econometrics*, *126*, 145-171.

Bhattacharya, D. (2007). Inference on inequality from household survey data. *Journal of Econometrics*, *137*, 674-707.

Bickel, P. J., Klassen, C. A. J., Ritov, Y., & Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models.* Johns Hopkins University Press, Baltimore, MD, USA.

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

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

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, *34*, 305–334.

Cosma, A., Kostyrka, A. V., & Tripathi, G. (2018). Smoothed empirical likelihood based inference with missing endogenous variables. In progress.

Cosslett, S. R. (1981a). Efficient estimation of discrete choice models. In C. F. Manski & D. McFadden (Eds.), *Structural analysis of discrete data with econometric applications* (pp. 51–111). MIT Press, Cambridge, MA, USA.

Cosslett, S. R. (1981b). Maximum likelihood estimation for choice-based samples. *Econometrica*, *49*, 1289–1316.

Cosslett, S. R. (1991). Efficient estimation from endogenously stratified samples with prior information on marginal probabilities. Manuscript. Retrieved from `economics.sbs.ohio-state.edu/scosslett/papers/cbsample1.pdf`.

Cosslett, S. R. (1993). Estimation from endogenously stratified samples. In G. Maddala, C. Rao, & H. Vinod (Eds.), *Handbook of Statistics, vol. 11* (pp. 1–43). Elsevier, The Netherlands.

Cragg, J. G. (1983). More efficient estimation in the presence of heteroscedasticity of unknown form. *Econometrica*, *49*, 751–764.

Deaton, A. (1997). *The analysis of household surveys.* Johns Hopkins University Press, Baltimore, MD, USA.

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

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

Gill, R. D., Vardi, Y., & Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Annals of Statistics*, *16*, 1069–1112.

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

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, *5*, 475–492.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*, 153–161.

Hirose, Y. (2007). M-Estimators in semi-parametric multi-sample models. Manuscript. Retrieved from `sms.victoria.ac.nz/foswiki/pub/Main/ResearchReportSeries/mscs08-05.pdf`.

Hirose, Y., & Lee, A. J. (2008). Semi-parametric efficiency bounds for regression models under generalised case-control sampling: The profile likelihood approach. *Annals of the Institute of Statistical Mathematics*, *62*, 1023–1052.

Holt, D., Smith, T. M. F., & Winter, P. D. (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.

Imbens, G. W., & Lancaster, T. (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.

Kalbfleisch, J. D., & Lawless, J. F. (1988). Estimation of reliability in field-performance studies (with discussion). *Technometrics*, *30*, 365–388.

Kitamura, Y., Tripathi, G., & Ahn, H. (2004). Empirical likelihood based inference in conditional moment restriction models. *Econometrica*, *72*, 1667–1714.

Manski, C. F. (1989). Anatomy of the selection problem. *Journal of Human Resources*, *24*, 343–360.

Manski, C. F. (1995). *Identification problems in the social sciences*. Harvard University Press, Cambridge, MA, USA.

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

Manski, C. F., & McFadden, D. (1981). Alternative estimators and sample design for discrete choice analysis. In C. F. Manski & D. McFadden (Eds.), *Structural analysis of discrete data with econometric applications* (pp. 2–50). MIT Press, Cambridge, MA, USA.

Newey, W. K. (1993). Efficient estimation of models with conditional moment restrictions.

In G. S. Maddala, C. R. Rao, & H. D. Vinod (Eds.), *Handbook of Statistics, vol. 11* (pp. 2111–2245). Elsevier, The Netherlands.

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

Owen, A. B. (2001). *Empirical likelihood.* Chapman & Hall/CRC, New York, NY, USA.

Owen, A. B. (2013). Self-concordance for empirical likelihood. *Canadian Journal of Statistics*, *41*, 387–397.

Owen, A. B. (2017). A weighted self-concordant optimization for empirical likelihood. Manuscript. Retrieved from `http://statweb.stanford.edu/~owen/empirical/countnotes.pdf`.

Pagan, A., & Ullah, A. (1999). *Nonparametric econometrics.* Cambridge University Press, Cambridge, UK.

Patil, G. P., & Rao, C. R. (1977). The weighted distributions: A survey of their applications. In P. R. Krishnaiah (Ed.), *Applications of Statistics* (pp. 383–405). North Holland, Amsterdam, The Netherlands.

Patil, G. P., & Rao, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, *34*, 179–189.

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

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

Robinson, P. M. (1987). Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica*, *55*, 875–891.

Scott, A. J., & Wild, C. J. (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., & Tripathi, G. (2013). Semiparametric efficiency bounds for microeconometric models: A survey. *Foundations and Trends in Econometrics*, *6*, 163–397.

Tripathi, G. (1999). A matrix extension of the Cauchy-Schwarz inequality. *Economics Letters*, *63*, 1–3.

Tripathi, G. (2011a). Generalized method of moments (GMM) based inference with stratified samples when the aggregate shares are known. *Journal of Econometrics*, *165*, 258–265.

Tripathi, G. (2011b). Moment based inference with stratified data. *Econometric Theory*, *27*, 47–73.

Tripathi, G., & Kitamura, Y. (2003). Testing conditional moment restrictions. *Annals of Statistics*, *31*, 2059–2095.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Annals of Statistics*, *10*, 616–620.

Vardi, Y. (1985). Empirical distributions in selection biased models. *Annals of Statistics*, *13*, 178–203.

Varrette, S., Bouvry, P., Cartiaux, H., & Georgatos, F. (2014). Management of an academic HPC cluster: The UL experience. In *Proceedings of the* 2014 *international conference on high performance computing and simulation* (*HPCS* 2014). IEEE, Bologna, Italy.

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

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

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). MIT Press, Cambridge, MA, USA.

Table 1. Aggregate shares for the simulation study.

Stratification	Design	Q_1^*
endogenous	homoskedastic	0.27
	heteroskedastic	0.28
exogenous	homoskedastic	0.63
	heteroskedastic	0.63

Table 2. Simulation summary: Estimated β_0^*, β_1^* under heteroskedasticity.

			Intercept			Slope		
Stratification	n	Estimator	Bias	SE	RMSE	Bias	SE	RMSE
endogenous	50	LS	−.1595	.4417	.4694	−.1076	.4868	.4983
		GMM	.0307	.5165	.5171	−.0408	.4757	.4772
		SEL ($c_n = 0.3$)	−.0033	.2909	.2910	−.0329	.3845	.3859
		SEL ($b_n \approx 0.46$)	.0015	.3916	.3916	−.0213	.4140	.4146
	150	LS	−.1714	.3489	.3885	−.1025	.3514	.3658
		GMM	.0234	.3980	.3985	−.0248	.3325	.3332
		SEL ($c_n = 0.4$)	.0202	.1880	.1891	−.0332	.2394	.2417
		SEL ($b_n \approx 0.39$)	.0224	.2313	.2324	−.0296	.2543	.2560
	500	LS	−.1805	.2894	.3410	−.0906	.2641	.2790
		GMM	.0043	.3304	.3302	−.0061	.2456	.2456
		SEL ($c_n = 0.8$)	.0107	.1316	.1321	−.0130	.1486	.1492
		SEL ($b_n \approx 0.31$)	.0096	.1242	.1246	−.0131	.1454	.1460
exogenous	50	LS	.0038	.3435	.3434	−.0032	.4275	.4273
		GMM	.0080	.4863	.4861	−.0041	.4791	.4789
		SEL ($c_n = 0.3$)	−.0161	.2518	.2523	.0250	.3754	.3762
		SEL ($b_n \approx 0.29$)	−.0098	.3547	.3549	.0117	.4225	.4227
	150	LS	.0021	.2609	.2608	−.0063	.3062	.3061
		GMM	.0042	.3838	.3836	−.0070	.3364	.3363
		SEL ($c_n = 0.4$)	.0010	.1562	.1562	−.0026	.2326	.2326
		SEL ($b_n \approx 0.24$)	.0020	.1910	.1910	−.0034	.2472	.2472
	500	LS	−.0012	.2189	.2188	.0014	.2323	.2322
		GMM	−.0023	.3354	.3352	.0012	.2540	.2539
		SEL ($c_n = 0.8$)	.0006	.1200	.1200	.0017	.1530	.1530
		SEL ($b_n \approx 0.19$)	.0003	.0988	.0988	.0017	.1425	.1425

Table 3. Simulation summary: Estimated β_0^*, β_1^* under homoskedasticity.

			Intercept			Slope		
Stratification	n	Estimator	Bias	SE	RMSE	Bias	SE	RMSE
endogenous	50	LS	−.4576	.2855	.5392	.0991	.2156	.2372
		GMM	−.0431	.3389	.3415	.0158	.2231	.2236
		SEL ($c_n = 0.3$)	−.0790	.4180	.4255	.0379	.3501	.3521
		SEL ($b_n \approx 0.42$)	−.0559	.3766	.3807	.0255	.2793	.2805
	150	LS	−.4273	.1480	.4522	.0708	.0906	.1149
		GMM	−.0053	.1680	.1680	−.0011	.0902	.0902
		SEL ($c_n = 0.4$)	−.0160	.2028	.2034	.0070	.1364	.1366
		SEL ($b_n \approx 0.35$)	−.0135	.1914	.1919	.0069	.1235	.1237
	500	LS	−.4142	.0845	.4227	.0626	.0453	.0772
		GMM	−.0005	.0938	.0938	.0000	.0427	.0427
		SEL ($c_n = 0.8$)	−.0047	.1026	.1027	.0031	.0577	.0578
		SEL ($b_n \approx 0.28$)	−.0061	.1053	.1055	.0037	.0590	.0591
exogenous	50	LS	−.0039	.2432	.2431	.0031	.1984	.1983
		GMM	−.0022	.2622	.2621	.0022	.2084	.2083
		SEL ($c_n = 0.3$)	−.0184	.3153	.3159	.0230	.3214	.3222
		SEL ($b_n \approx 0.29$)	−.0077	.2958	.2959	.0078	.2787	.2788
	150	LS	−.0007	.1260	.1260	−.0022	.0843	.0843
		GMM	.0011	.1314	.1314	−.0024	.0863	.0863
		SEL ($c_n = 0.4$)	.0001	.1502	.1502	−.0007	.1261	.1261
		SEL ($b_n \approx 0.24$)	.0001	.1516	.1516	−.0012	.1199	.1199
	500	LS	.0027	.0703	.0703	−.0006	.0410	.0410
		GMM	.0040	.0755	.0756	−.0008	.0416	.0416
		SEL ($c_n = 0.8$)	.0053	.0785	.0787	−.0029	.0554	.0555
		SEL ($b_n \approx 0.19$)	.0033	.0801	.0802	−.0012	.0595	.0595

Table 4. Simulation summary: Estimated Q_1^* under heteroskedasticity.

Stratification	Sample size	Estimator	Bias	SE	RMSE
endogenous	50	GMM	.0132	.0890	.0900
		SEL ($c_n = 0.3$)	.0178	.0939	.0956
		SEL ($b_n \approx 0.46$)	.0209	.0940	.0963
	150	GMM	.0047	.0504	.0506
		SEL ($c_n = 0.4$)	.0096	.0532	.0540
		SEL ($b_n \approx 0.39$)	.0126	.0534	.0549
	500	GMM	.0014	.0278	.0278
		SEL ($c_n = 0.8$)	.0106	.0294	.0313
		SEL ($b_n \approx 0.31$)	.0092	.0293	.0307
exogenous	50	GMM	.0133	.1070	.1078
		SEL ($c_n = 0.3$)	.0384	.1102	.1167
		SEL ($b_n \approx 0.29$)	.0550	.1032	.1169
	150	GMM	.0050	.0633	.0635
		SEL ($c_n = 0.4$)	.0414	.0655	.0775
		SEL ($b_n \approx 0.24$)	.0523	.0620	.0811
	500	GMM	.0009	.0347	.0347
		SEL ($c_n = 0.8$)	.0719	.0364	.0806
		SEL ($b_n \approx 0.19$)	.0471	.0344	.0583

TABLE 5. Simulation summary: Estimated Q_1^* under homoskedasticity.

Stratification	Sample size	Estimator	Bias	SE	RMSE
endogenous	50	GMM	.0135	.0873	.0883
		SEL ($c_n = 0.3$)	.0204	.0909	.0931
		SEL ($b_n \approx 0.42$)	.0266	.0924	.0962
	150	GMM	.0042	.0492	.0493
		SEL ($c_n = 0.4$)	.0129	.0515	.0531
		SEL ($b_n \approx 0.35$)	.0176	.0519	.0548
	500	GMM	.0013	.0262	.0263
		SEL ($c_n = 0.8$)	.0180	.0284	.0337
		SEL ($b_n \approx 0.28$)	.0150	.0284	.0321
exogenous	50	GMM	.0133	.1070	.1078
		SEL ($c_n = 0.3$)	.0380	.1088	.1153
		SEL ($b_n \approx 0.29$)	.0561	.1020	.1164
	150	GMM	.0050	.0633	.0635
		SEL ($c_n = 0.4$)	.0420	.0654	.0777
		SEL ($b_n \approx 0.24$)	.0530	.0619	.0815
	500	GMM	.0009	.0347	.0347
		SEL ($c_n = 0.8$)	.0720	.0364	.0807
		SEL ($b_n \approx 0.19$)	.0473	.0344	.0585

TABLE 6. Running time (in minutes) to estimate the parameters.

n	(β_0^*, β_1^*)	$(\beta_0^*, \beta_1^*, Q_1^*)$
50	0.036	16.17
150	0.129	45.70
500	0.523	149.4