

Asymptotic Inference from Multi-Stage Samples ^{*}

Debopam Bhattacharya[†]
Department of Economics,
Dartmouth College.

January 28, 2004.

Abstract

I develop a GMM-based framework for asymptotic inference to analyze data from surveys whose designs involve stratification and clustering. I set up the estimation problem, derive the appropriate asymptotic distribution theory as the number of clusters per stratum tends to infinity and compute asymptotic standard errors that are robust to sample-design effects. The analysis is then extended to nonparametric regression and to semiparametric estimation based on U-processes. Empirical illustrations are provided using consumption expenditure data from the complexly designed Indian national household survey.

JEL classification code: C12, C13, C31, C42.

Keywords: Stratification, clustering, generalized method of moments, nonparametric estimation, U-statistics

^{*}I am grateful to Aureo Paula and Professors Adriana Lleras-Muney, Han Hong and Jeffrey Kling for many helpful conversations and especially to Professors Angus Deaton and Bo Honore for constant help, encouragement and support. I have immensely benefitted from discussions with Alessandro Tarozzi. I would also like to thank three anonymous referees, the editor and an associate editor of the Journal of Econometrics for helpful comments. Financial support from the Wilson Fellowship is gratefully acknowledged. All errors are mine.

[†]All correspondences should be addressed to Debopam Bhattacharya, Department of Economics, 301 Rockefeller Hall, Dartmouth College, Hanover, NH 03755, Fax: 603-646-2122, Phone: 603-359-5994. email: debopam.bhattacharya@dartmouth.edu

1 Introduction

Analysis of cross-sectional data typically proceeds by assuming that the data are generated by a simple random sample from the entire population¹. However, large-scale cross-sectional household surveys, the main sources of large cross-section data on individual behavior, are rarely simple random samples and have designs that involve stratification and clustering. Examples include all of the World Bank's multi-country Living Standards Measurement Studies (LSMS)- the primary source of micro-data from developing countries, USA's Current Population Survey (CPS), the (cross-section component of) Panel Study of Income Dynamics (PSID), the Indian National Sample Survey (NSS) among many others. Ignoring the survey design in the estimation process can lead to inconsistent estimates of the population parameters and almost always produces inconsistent estimates of the standard error of these estimates.

The main features of a complex survey design are stratification and clustering. Stratification means that the original population is first (i.e. prior to sampling) divided into several subgroups, based on criteria like area of residence, race, age etc. obtained from the latest census². Sampling is done separately within each stratum, independently across the strata. Clusters are (physically) contiguous groups of households existing within a stratum; in rural areas they are villages, in urban areas, they are blocks or neighborhoods. Typically, a fixed and large number of clusters are sampled randomly within each stratum and then within each selected cluster, a small and fixed number of households are sampled randomly. Such designs are motivated by a variety of financial and administrative considerations (see Cochran (1977) for further details).

Historically, effects of sample design on estimation were analyzed in the sample survey literature (e.g. Cochran (1977)) in statistics since the mid-sixties. This literature is concerned exclusively with the estimation of population means and its exact finite sample distribution theory (assuming finite populations and using combinatorial methods) becomes unwieldy for more complex parameters like the population median for which one

¹To the best of my knowledge, Wooldridge (2001a) is the only textbook on cross-section econometrics that has a limited discussion of alternative sampling designs.

²In the context of regression of y on x , stratification can be either endogenous (on y) or exogenous (on x). This distinction is unimportant for computation of standard errors, given a method of consistent estimation of the parameter of interest.

has to keep track of stratum and cluster identities of the ordered observations. Asymptotic analysis provides an elegant alternative which is applicable to a much larger class of estimation problems and provides reasonably accurate approximation to the exact finite sample results when large sample sizes are available (as is frequently the case for real-life large-scale household surveys). No unified framework currently exists in the econometrics literature for asymptotic analysis with complex survey data³. In this paper, I develop a framework for asymptotic inference, which enables one to handle data from complex surveys. I show how to set up the estimation problem, derive the appropriate asymptotic distribution theory and finally, compute the asymptotic standard errors that are robust to sample-design effects.

The procedure of inference from multi-stage stratified samples involves two distinct stages of correction (relative to simple random samples), one involving the estimation of the parameters and the second involving the computation of standard errors of these estimates. One needs to modify the method of estimation (the first ‘level’ of correction) to account for the fact that the distribution of the sampled observations generally differs from their distribution in the population as a result of the multi-stage design. One can usually achieve this by suitably weighting the data where the weights, computed from the latest census, are typically available in the survey data (see section 1.1 for more on this). At the second ‘level’, one needs to use asymptotic theory for dependent and non-identically distributed observations to derive the asymptotic distribution of the estimates and compute standard errors that are robust to the sample-design effects. Briefly, clustering induces positive correlations in variables and increases standard errors; stratification leads to smaller variability of statistics over repeated samples leading to smaller standard errors (relative to simple random samples). Ignoring the design, therefore, produces inconsistent estimates of the standard errors unless, by rare chance, the two effects exactly cancel.

The phenomenon of clustering and correlation between physically proximate units have

³Wooldridge (2001b) and Sakata (working paper) have analyzed asymptotic properties of M-estimators under stratification. In contrast, this paper covers GMM, a large class of semiparametric estimation problems (all of which are M-estimators, although the U-statistics based estimators considered here could be viewed as an extension of standard M-estimation) as well as purely nonparametric inference. The sampling design considered here, unlike the aforementioned papers, combines stratification with clustering and, therefore, closely resembles the designs of most real-life large-scale household surveys. Also, see footnote 7.

been discussed in the restrictive form of random effects in panel data models (e.g. county-specific random effects in a cross-section regression, see Moulton, 1986 for instance) and more explicitly in the spatial statistics literature (c.f. Conley (1999), Kloek (1981) and Pfeiffermann and Nathan (1981)). Unlike the above instances, in this paper, we do not impose any structure on the nature of this correlation and derive estimates of standard errors that are robust to arbitrary correlation structures (and heteroskedasticity) between units residing in the same cluster.

Stratification (without clustering) has been extensively studied in econometrics in the context of choice-based sampling (e.g. Manski-Lerman (1977), Cosslett (1981), Imbens (1992)). That literature implicitly assumes that the strata are also chosen probabilistically. Therefore, standard errors do not warrant the stratification-correction. In most real surveys, the strata are not chosen probabilistically, the number of units sampled per stratum is fixed by design⁴ and therefore a correction is necessary because the strata remain fixed over repeated samples.

The plan of the paper is as follows: Section 2 introduces the MoM problem, sets out the moment conditions for a two-stage design with stratification, section 2.1 lists the main theorems. Section 2.2 illustrates theoretically the design effects on the variance of the GMM estimator by breaking up the total effect into stratum and cluster effects. Section 3, extends these methods to nonparametric regressions and a class of semiparametric estimators which are defined as minimands of U-processes. Section 4 briefly discusses the actual implementation of the methods for real-life surveys. Section 5 provides a brief illustration of the methods, using data from the complexly designed Indian National Sample Survey. Section 6 concludes. In the next subsection, I provide a brief discussion of sampling weights and their role in estimation from complex surveys.

1.1 A brief note on weighting and consistent estimation

The issue of whether to weigh observations during parameter estimation is distinct from correcting standard errors for survey design. The current paper focuses on the second of these two issues. Nonetheless, in this subsection, I briefly clarify the role of weights in

⁴Some authors have previously noted this difference, including Cosslett (1995), Imbens and Lancaster (1996) and Wooldridge (1999) and have named this "standard stratified sampling".

estimation (see DuMouchel and Duncan (1983) and Wooldridge (1999, 2001a) for further details).

Because of the stratified, clustered design, not all households in the population, in general, have an equal probability of being included in the sample. As a result, different sample observations are usually assigned different weights, with the sampling weight of the observation denoting how many observations in the population it represents.⁵ In general, when the parameter of interest is the census parameter (i.e. the parameter one would get if one performed the same estimation exercise with the entire population), a weighted estimation technique is appropriate. Unweighted estimates will not be consistent for the census parameter. However, when the researcher's model, say for the conditional mean of y given x , is correct and the stratification and clustering are based on the independent variables x , then unweighted estimates will be consistent for the parameters of that model. When comparing standard errors for the estimate of the parameter of interest that are and are not corrected for the sample design (the main focus of the paper), I shall focus on the *same estimate of the parameter* (weighted for the method of moment estimators in section 2 but unweighted for the nonparametric regressions in Section 3).

2 The Method of Moment problem

In this section, I shall set-up the method of moment problem with data from a stratified, multi-stage clustered sample. In the next section, I shall extend the analysis to cover purely nonparametric estimation (of conditional means), followed by U-statistic based estimation for a large class of semiparametric models.

The sampling design I consider is generic and is as follows. The population is divided into S first stage strata. Stratum s contains a mass of H_s clusters. A sample of n_s clusters (indexed by c_s) is drawn via simple random sample with replacement (sampling with or without replacement has no effect on my asymptotic analysis based on increasing number of clusters) from stratum s , for each s . The c_s th sampled cluster in the s th stratum contains

⁵In many real-life surveys, the probability of drawing a cluster is proportional to its (estimated) size. Such designs are called self-weighted, implying that all population units have equal probability of being included in the sample. As a result, all units have identical weights and the weights are dropped from the data set.

a finite population of M_{sc_s} households. A simple random sample of k households (equal for all strata and clusters and indexed by h) is drawn from it. The h th household in the c_s th cluster in the s th stratum has $\nu_{sc_s h}$ members. The joint density of a (per capita) characteristic Y and household size N in the s th stratum is denoted by $dF(y, \nu|s)$ with $F(a, b|s)$ denoting the population proportion of households in stratum s with $Y < a$ and $N < b$. Note that this joint density can differ across strata, so that sampled observations from different strata are independent but in general not identically distributed.

Let

$$n = \sum_{s=1}^S n_s \text{ and } n_s = na_s \text{ with } \sum_{s=1}^S a_s = 1.^6$$

The weight of every member in the h th household in the c_s th sampled cluster in the s th stratum is given by

$$w_{sc_s h} = \frac{M_{sc_s} H_s}{kn_s} \nu_{sc_s h}$$

and equals the number of individuals in the population represented by this particular individual. All expectation and variances are taken with respect to the sampling distribution, which differs in general from the population distribution due to the non-simple random sampling. I shall let $E_{h|c_s, s}(\cdot)$, $Var_{h|c_s, s}(\cdot)$ to denote expectation and variance respectively taken with respect to the second stage of sampling, conditional on stratum s and cluster c_s (analogously, $E_{c_s|s}(\cdot)$ and $Var_{c_s|s}(\cdot)$ for first stage of sampling). When expectations and variances are taken with respect to both the stages of sampling, I simply denote those by $E|_s(\cdot)$ and $V|_s(\cdot)$; $O_p(1)$ and $o_p(1)$ will denote quantities that are respectively (asymptotically) bounded in probability and go to 0 in probability; \rightarrow_d and \rightarrow_P will denote convergence in distribution and probability, respectively.

In most real-life surveys, the number of clusters sampled per stratum is much larger than the number of households sampled per cluster, the latter typically being very small (in the Indian NSS for instance, the numbers are about 120 and 10, respectively). Therefore, asymptotic analysis with the number of clusters (n) going to infinity with number of households staying fixed and finite is likely to yield a more accurate approximation to

⁶Note that this can be equivalently written as

$$\frac{n_s}{n} \rightarrow a_s < \infty \text{ as } n_s, n \rightarrow \infty; \quad s = 1, 2, \dots, S$$

the true distributions of the estimates, and this motivates our fixed- k , large- n asymptotics. But we note in passing that situations could arise where asymptotics on the number of households sampled per cluster could be more appropriate⁷. Secondly, clusters sampled within a stratum are geographically scattered over a large area; households sampled within a cluster are physically close to each other. This motivates my assumption that cluster-level aggregates are independent across clusters within a stratum but household level variables are correlated within a cluster.⁸

Suppose one is interested in estimating a parameter θ_0 of dimension p (typically characterizing an individual level characteristic, e.g. the per person mean consumption in the population), which solves the $l \geq p$ population moment condition

$$0 = \sum_{s=1}^S H_s \int \nu \mathbf{m}(y, \theta_0) dF(y, \nu | s). \quad (1)$$

For instance, the population mean μ_0 solves:

$$\begin{aligned} 0 &= \sum_{s=1}^S H_s \int (y - \mu_0) \nu dF(y, \nu | s) \\ &= \sum_{s=1}^S H_s E_{c_s} \left\{ \sum_{K=1}^{M(s, c_s)} \nu_{sc_s K} (y_{sc_s K} - \mu_0) | s \right\}.^9 \end{aligned}$$

The method of moment estimator of θ_0 is based on the sample analog (corresponding to

⁷Sakata considers asymptotics on the number of strata. His objects of interest are parameters of a superpopulation from which the strata are sampled. So the strata for his analysis are like clusters for our analysis and correction of standard errors (of superpopulation parameter estimates) due to fixed stratification are irrelevant. Also, see the next footnote.

⁸For smaller strata, cluster level variables might be correlated and, as in the spatial statistics literature, one needs this dependence to ‘disappear’ (spatial ergodicity) as the distance between clusters increases, in order for the laws of large number to hold as the number of clusters tends to infinity. The information on spatial distances between clusters is rare if not totally non-existent in survey data, which makes this approach infeasible. A similar consideration holds for asymptotics on the number of (correlated) households per cluster (which would arise in a design where the number of clusters selected per stratum is much smaller relative to the number of households selected per cluster; but such designs are rare).

⁹Consider for a given stratum s ,

$$E_c \left\{ \sum_{K=1}^{M(s, c)} n_{scK} y_{scK} | s \right\}$$

the multi-stage design) of the moment conditions (1), viz. :

$$\sum_{s=1}^S \frac{H_s}{n_s} \sum_{c_s=1}^{n_s} \frac{M(s, c_s)}{k} \sum_{h=1}^k \nu_{sc_s h} \mathbf{m}(\mathbf{y}_{sc_s h}, \boldsymbol{\theta}) \simeq 0. \quad (2)$$

For later use, let us define $\mathbf{z}_{sc_s h} = (\mathbf{y}_{sc_s h}, \nu_{sc_s h})$ and $\tilde{\mathbf{m}}(\mathbf{z}_{sc_s h}, \boldsymbol{\theta}) = \nu_{sc_s h} \mathbf{m}(\mathbf{y}_{sc_s h}, \boldsymbol{\theta})$.

The following analysis characterizes the asymptotic distribution of $\hat{\boldsymbol{\theta}}$. By ‘asymptotic’ I mean that the number of sampled clusters for every stratum goes to infinity at the same rate, so that the quantities a_s ’s stay fixed. I shall re-index clusters by i with i running from 1 to n . n denotes the total number of clusters in the sample. Corresponding to every cluster i is associated the index s_i which denotes the stratum from which i is drawn. Then by definition,

$$\#(i|s_i = s) = n_s \text{ for each } 1 \leq s \leq S. \quad (3)$$

Then (2) reduces to

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{m}}_i(\boldsymbol{\theta}) &\simeq 0 \text{ where} \\ \tilde{\mathbf{m}}_i(\boldsymbol{\theta}) &= \left[\sum_{s=1}^S \frac{H_s}{a_s} \mathbf{1}(s_i = s) \frac{M(s_i, i)}{k} \sum_{h=1}^k \mathbf{m}(\mathbf{y}_{s_i h}, \boldsymbol{\theta}) \nu_{s_i h} \right]. \end{aligned} \quad (4)$$

Define

$$g_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{m}}_i(\boldsymbol{\theta}).$$

The GMM estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ solves

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta} \in \Theta} \{g_n(\boldsymbol{\theta})' A_n g_n(\boldsymbol{\theta})\} \quad (5)$$

where A_n is an appropriate random weighting matrix. Note that the functions $\mathbf{m}_i(\boldsymbol{\theta})$ are independent (though not identically distributed owing to stratification) across i . This makes the asymptotic analysis of the estimator completely standard via the theory of GMM

This equals the expectation (over clusters) of total cluster income (added across all population households in that cluster) whereas

$$E_c \left\{ \sum_{K=1}^{M(s,c)} n_{scK} |s \right\}$$

equals the expectation (over clusters) of total cluster population of individuals.

estimators developed in the econometrics and statistics literature over the last two decades. The first two chapters in the Handbook of Econometrics volume 4, in particular, have a comprehensive treatment of this theory. Note that the proof of consistency uses WLLN for independent non-identically distributed random variables; the proof of asymptotic normality uses the Central limit theorem (Lindeberg-Feller-Lyapunov version) for independent and non identically distributed variables. After stating the relevant theorems (without proofs which are standard), I shall derive the expression for asymptotic variance which takes into account the sample design.

2.1 The main theorems

Assumptions:

A0a. For $s, s' = 1 \dots S$, $(\mathbf{z}_{sc_s h}, \mathbf{z}_{s'c'_s h'})$ are independent unless $s = s'$ and $c_s = c'_s$ for $c_s = 1, \dots, n_s$, $c'_s = 1, \dots, n_{s'}$ and $h, h' = 1 \dots k$.

A0b. For each s , $\{\mathbf{z}_{sc_s h}\}_{c_s=1, \dots, n_s, h=1, \dots, k}$ are identically distributed.¹⁰

A0c. For $s \neq s'$, \mathbf{z}_s and $\mathbf{z}_{s'}$ are independent (but not necessarily identically distributed) where $\mathbf{z}_s \equiv \{\mathbf{z}_{sc_s h}\}_{c_s=1, \dots, n_s, h=1, \dots, k}$.

A1. $\tilde{\mathbf{m}}^j(\cdot, \boldsymbol{\theta})$ is continuous at each $\boldsymbol{\theta}$ with probability 1 (which includes singleton points of discontinuity as in quantiles), for each $j = 1 \dots l$.

A2. $\exists d(\cdot)$ with $E(d(\cdot)) < \infty$ such that $\|\tilde{\mathbf{m}}^j(\mathbf{t}, \boldsymbol{\theta})\| \leq d(\mathbf{t})$ for each $j = 1 \dots l$ for all \mathbf{t} .

A3. The parameter space Θ is compact.

A4. $\boldsymbol{\theta}_0$ solves (1) uniquely in Θ and $\boldsymbol{\theta}_0 \in \text{int}(\Theta)$.

A5. $E(\tilde{\mathbf{m}}(\mathbf{z}, \boldsymbol{\theta}))$ is continuously differentiable at $\boldsymbol{\theta}_0$ and

$$p \lim \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\theta}'} E(\tilde{\mathbf{m}}_i(\boldsymbol{\theta}_0)) = \Gamma, \text{ nonsingular.}$$

A6. The sequence

$$\nu_n(\boldsymbol{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \{\tilde{\mathbf{m}}_i(\boldsymbol{\theta}) - E\tilde{\mathbf{m}}_i(\boldsymbol{\theta})\}$$

¹⁰Note that A0a-b are not conditioned on the clusters; clearly conditional on the clusters, $(\mathbf{z}_{sc_s h}, \mathbf{z}_{s'c'_s h'})$ are independent for all s, s', c_s, c'_s, h, h' and also $\{\mathbf{z}_{sc_s h}, \mathbf{z}_{s'c'_s h'}\}$ may not be identically distributed given c_s and c'_s .

is stochastically equicontinuous¹¹.

A7. $\sup_{\theta \in \Theta} E|\tilde{\mathbf{m}}_i(\boldsymbol{\theta})|^3 < \infty$.

A8. a. $\lim_{n \rightarrow \infty} \sum_{i=1}^n \frac{Var(\tilde{\mathbf{m}}_i(\boldsymbol{\theta}))}{i^2} < \infty$.

b. $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n Var(\tilde{\mathbf{m}}_i(\boldsymbol{\theta})) \equiv \lim_{n \rightarrow \infty} W_n = W_0 < \infty$.

A9. $p \lim_{n \rightarrow \infty} A_n = A_0$ and A_n positive definite with probability 1 for each n .

Proposition 1 *Under assumptions A0 through A4 and A8a, A9.*

$$p \lim_{n \rightarrow \infty} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = 0.$$

Proposition 2 *Additionally, under A0-A7 and A8b, A9,*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \rightarrow_d N(0, V)$$

with

$$V = (\Gamma A_0 \Gamma')^{-1} \Gamma A_0 W_0 A_0 \Gamma' (\Gamma A_0 \Gamma')^{-1}.$$

Choosing $A_n = W_n^{-1}$ yields the efficient estimator with asymptotic covariance matrix $V_{eff} = (\Gamma W_0^{-1} \Gamma')^{-1}$. Moreover, V is consistently estimable. (See the next subsection for an expression for V_{eff} and for consistent estimates of Γ and W_0 , that are robust to the sample design.)

The method of moment framework is flexible enough to encompass linear and IV regressions, maximum likelihood estimation, concentration curves and nearly all measures of poverty and inequality. Several papers have been written on the latter measures that have employed different techniques to prove asymptotic normality¹². The above framework shows that the same technique works for all these measures so that separate asymptotic results are not required. It is standard to verify that assumptions A1-A7 hold for each of the problems mentioned above.

¹¹Sufficient conditions for stochastic equicontinuity can be found in Andrews (1999) and Pakes and Pollard (1989). In most applications like linear and quantile regression, inequality and poverty estimation etc., these sufficient conditions will be met via piecewise linearity of the $m(\cdot)$ functions, boundedness of the parameter space and bounded moments up to order 2 of the m -functions. Assuming finite and positive moments of the Y_i 's is also sufficient to guarantee finite moments of the m -functions.

¹²For instance, Zheng (2002) uses the Bahadur representation of quantiles for complex surveys; Beach and Davidson (1983) used asymptotic results for order statistics for i.i.d. samples.

2.2 Expression for variance and design effects

In this section, I derive an expression for W_n and illustrate theoretically the separate effects of stratification and clustering on estimates of standard error. This decomposition shows the factors on which the stratum and cluster effects depend and therefore suggests some diagnostic checks, that can be made prior to the corrections to assess the degree of inconsistency, absent the corrections.

Note from above that

$$W_0 = \text{Var} \left(\sum_{s=1}^S \sum_{c_s=1}^{n_s} \sum_{h=1}^k w_{sc_s h} \mathbf{m}(y_{sc_s h}, \boldsymbol{\theta}_0) \right).$$

$$\Gamma = p \lim \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} E \tilde{\mathbf{m}}_i(\boldsymbol{\theta}_0).$$

A consistent estimate of Γ is given by

$$\hat{\Gamma} = \frac{\partial}{\partial \theta} \left\{ \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{m}}_i(\boldsymbol{\theta}) \right\}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

Now,

$$\begin{aligned} W_0 &= \text{Var} \left(\sum_{s=1}^S \sum_{c_s=1}^{n_s} \sum_{h=1}^k w_{sc_s h} \mathbf{m}(y_{sc_s h}, \boldsymbol{\theta}_0) \right) \\ &= \sum_{s=1}^S \frac{H_s^2}{n_s} E_{c_s|s} \left(\text{Var}_{h|c_s,s} \left\{ \frac{M_{sc_s}}{k} \sum_{h=1}^k \boldsymbol{\nu}_{sc_s h} \mathbf{m}(y_{sc_s h}, \boldsymbol{\theta}_0) \right\} \right) \\ &\quad + \sum_{s=1}^S \frac{H_s^2}{n_s} \frac{1}{k} \text{Var}_{c_s|s} \{ M_{sc_s} \boldsymbol{\nu}_{sc_s h} \mathbf{m}(y_{sc_s h}, \boldsymbol{\theta}_0) \} \\ &\quad + \sum_{s=1}^S \frac{H_s^2}{n_s} \frac{M_{sc_s}^2}{k^2} \sum_{h=1}^k \sum_{h' \neq h} \text{cov}_{c_s|s} \{ \boldsymbol{\nu}_{sc_s h} \mathbf{m}(y_{sc_s h}, \boldsymbol{\theta}_0), \boldsymbol{\nu}_{sc_s h'} \mathbf{m}(y_{sc_s h'}, \boldsymbol{\theta}_0) \}. \end{aligned}$$

So that a consistent estimate W_n of W_0 is given by:

$$\begin{aligned}
W_n = & \sum_{s=1}^S \sum_{c_s=1}^{n_s} \sum_{h=1}^k w_{sc_s h}^2 \mathbf{m}(y_{sc_s h}, \hat{\theta}) \mathbf{m}(y_{sc_s h}, \hat{\theta})' \\
& + \sum_{s=1}^S \sum_{c_s=1}^{n_s} \sum_{h=1}^k \sum_{h \neq h'}^k w_{sc_s h} w_{sc_s h'} \mathbf{m}(y_{sc_s h}, \hat{\theta}) \mathbf{m}(y_{sc_s h'}, \hat{\theta})' \\
& - \sum_{s=1}^S \frac{1}{n_s} \left(\sum_{c_s=1}^{n_s} \sum_{h=1}^k w_{sc_s h} \mathbf{m}(y_{sc_s h}, \hat{\theta}) \right) \left(\sum_{c_s=1}^{n_s} \sum_{h=1}^k w_{sc_s h} \mathbf{m}(y_{sc_s h}, \hat{\theta}) \right)'.^{13} \quad (6)
\end{aligned}$$

The first term in (6) is the estimate of the variance without taking the sample design into account. This would be the correct expression if the sample were i.i.d. and the parameter of interest solved a weighted moment condition. The design described above warrants two correction terms which are the second and the third terms in (6). The second term is the cluster effect and is a function of the covariance between values obtained from the same cluster. If the covariances are positive on average (which is empirically true and is natural), this term is positive and implies that the wrong estimate of the standard error is an *underestimate* of the true standard error. The greater the degree of correlation between the observations inside a single cluster and the larger the number of observations (k) sampled from each cluster, the larger the degree of underestimation.

The third term is the stratum effect. First note that if there was just one stratum, the original moment conditions would cause this term to be close to zero asymptotically and one could ignore this term. But in general, with multiple strata, the expression within (.) is asymptotically non-zero (a weighted average of these expressions across strata is zero). So that its 'square' is a positive definite matrix. Intuitively, the variance with a stratified design is the sum of within-stratum variances. Ignoring stratification and estimating the

¹³For the sample mean, for instance, the corresponding expression is

$$\begin{aligned}
W_n = & \sum_{s=1}^S \sum_{c_s=1}^{n_s} \sum_{j=1}^k w_{sc_s j}^2 (y_{sc_s j} - \bar{y})^2 \\
& + \sum_{s=1}^S \sum_{c_s=1}^{n_s} \sum_{j=1}^k \sum_{j \neq j'}^k w_{sc_s j} w_{sc_s j'} (y_{sc_s j} - \bar{y})(y_{sc_s j'} - \bar{y}) \\
& - \sum_{s=1}^S \frac{H_s^2}{n_s} \left(\frac{1}{n_s} \sum_{c_s=1}^{n_s} M_{sc_s} (\bar{y}_{sc_s} - \bar{y}) \right)^2
\end{aligned}$$

variance as if it were a simple random sample causes over-estimation by wrongly adding on the between strata-variances. The degree of overestimation is larger the more homogeneous are the units within a stratum and the more heterogeneous the units across the strata (in the extreme case where the distribution of variables in every stratum is identical to that in the population, $\sum_{c_s=1}^{n_s} \sum_{h=1}^k w_{sc_s h} \mathbf{m}(y_{sc_s h}, \hat{\boldsymbol{\theta}}) = 0$ for each s and there is no overestimation).

Since the stratum and cluster effects go in opposite directions, the final effect depends on which effect dominates¹⁴. For the mean, one can easily compute the ratios of these two terms to the naive estimate of variance to get a prior sense of whether correction of the standard errors will alter results significantly. When the parameter of interest is an method of moment type estimator but more complicated than the mean, the calculations for the mean can be a computationally cheaper and yet useful diagnostic tool.

For a standard GMM problem solved with an arbitrary weighting matrix, my procedure will produce the right standard errors. From efficiency considerations, my methods provide the correct optimal weighting matrix viz. the inverse of the second moment matrix of the moment conditions which takes the sample design into account; the naive ‘optimal’ weighting matrix that ignores the survey design will produce a suboptimal weighting matrix and therefore an inefficient estimate.

3 Two extensions

In this section I extend my methods to nonparametric regression estimation and to the estimation of the parametric component in semiparametric models. For the former, I shall use ordinary kernel based estimators and for the latter I shall focus on estimators which are minimands of U-processes. I shall assume the same sample design as in section 2 of the paper. These extensions are not mere applications of the results in section 2 and therefore broaden the scope of my analysis to include nonparametric and semiparametric models.

¹⁴For simple cases like the population mean and linear regression coefficients, empirical studies using the sample survey methods (for exact finite sample standard errors) show that cluster effects usually dominate (cf. Deaton (1997), Howes and Lanjouw (1998)).

3.1 Kernel-based estimation

I only outline the case for nonparametric regression, the case of density estimation is similar and easier. Consider the following model generating the data

$$y_{sc_s h} = \mu(x_{sc_s h}) + \varepsilon_{sc_s h}$$

where for all s, c_s

$$\begin{aligned} E(\varepsilon_{sc_s h} | x_{sc_s h}, s) &= 0 \\ \text{Var}(\varepsilon_{sc_s h} | x_{sc_s h}) &\equiv \sigma^2(x_{sc_s h}). \end{aligned}$$

Also

$$\begin{aligned} \text{cov}(\varepsilon_{sc_s h}, \varepsilon_{s'c'_s h'} | x_{sc_s h}, x_{s'c'_s h'}, s, s') &\neq 0 \text{ for } s = s', c_s = c'_s \\ &= 0, \text{ otherwise.} \end{aligned}$$

One is interested in estimating the function $\mu(\cdot)$ at a given point x_0 . Let, as above, $n_s = a_s n$ with $\sum_s a_s = 1$. The Nadaraya-Watson estimate $\hat{\mu}(x_0)$ satisfies

$$\begin{aligned} &\hat{\mu}(x_0) - \mu(x_0) \\ = &\frac{\frac{1}{h_n} \sum_s \frac{1}{n_s} \sum_{c_s=1}^{n_s} \sum_{h=1}^k (\varepsilon_{sc_s h} + \mu(x_{sc_s h}) - \mu(x_0)) K\left(\frac{x_{sc_s h} - x_0}{h_n}\right) \frac{1}{k}}{\frac{1}{h_n} \sum_s \frac{1}{n_s} \sum_{c_s=1}^{n_s} \sum_{h=1}^k K\left(\frac{x_{sc_s h} - x_0}{h_n}\right) \frac{1}{k}} \end{aligned}$$

where h_n is an appropriate bandwidth and $K(\cdot)$ a standard kernel function¹⁵. Under standard conditions, one can show the consistency of this estimate. I now outline the steps

¹⁵Note that we do not use sampling weights in the estimation since the conditional mean function is assumed to be identical in every stratum and cluster. Unweighted estimates are therefore consistent for the true population regression function.

in deriving the standard error of this estimate. Observe that one can write

$$\begin{aligned}
& (nh_n)^{1/2} (\hat{\mu}(x_0) - \mu(x_0)) \\
= & \frac{\frac{(nh_n)^{1/2}}{h_n} \sum_s \frac{1}{n_s} \sum_{c_s=1}^{n_s} \sum_{h=1}^k \varepsilon_{sc_s h} K\left(\frac{x_{sc_s h} - x_0}{h_n}\right) \frac{1}{k}}{\hat{f}(x_0)} \\
& + \frac{\frac{(nh_n)^{1/2}}{h_n} \sum_s \frac{1}{n_s} \sum_{c_s=1}^{n_s} \sum_{h=1}^k (\mu(x_{sc_s h}) - \mu(x_0)) K\left(\frac{x_{sc_s h} - x_0}{h_n}\right) \frac{1}{k}}{\hat{f}(x_0)} \\
= & \frac{\frac{(nh_n)^{1/2}}{h_n n} \sum_{i=1}^n t_{1i}}{\hat{f}(x_0)} + \frac{\frac{(nh_n)^{1/2}}{h_n n} \sum_{i=1}^n t_{2i}}{\hat{f}(x_0)} \tag{7}
\end{aligned}$$

where I have again re-indexed the clusters to run from 1... n and

$$\begin{aligned}
\hat{f}(x_0) &= \frac{1}{h_n} \sum_s \frac{1}{n_s} \sum_{c_s=1}^{n_s} \sum_{h=1}^k K\left(\frac{x_{sc_s h} - x_0}{h_n}\right) \frac{1}{k} \\
t_{1i} &= \sum_s \frac{1}{a_s} 1(s_i = s) \sum_{h=1}^k \varepsilon_{ih} K\left(\frac{x_{ih} - x_0}{h_n}\right) \frac{1}{k} \\
t_{2i} &= \sum_s \frac{1}{a_s} 1(s_i = s) \sum_{h=1}^k (\mu(x_{ih}) - \mu(x_0)) K\left(\frac{x_{ih} - x_0}{h_n}\right) \frac{1}{k} \\
\hat{f}(x_0) &\equiv \sum_s \hat{f}(x_0|s) \rightarrow_P \sum_s f(x_0|s).
\end{aligned}$$

Note that $f(x_0|s)$ is not the true density of X in the s th stratum (since the $\hat{f}(x_0|s)$ terms are unweighted).

Under standard regularity conditions (see for instance, Pagan-Ullah (1999, pages 110-111), modified for non-identically distributed independent sequences, the first term will converge to a normal distribution with mean 0 and variance given by

$$V = \lim_{n \rightarrow \infty} \frac{1}{h_n} \left\{ \frac{1}{n} \sum_{i=1}^n \text{Var}(t_{1i}) \right\}$$

and the second term will converge to 0 in probability. So, to sum up,

$$(nh_n)^{1/2} (\hat{\mu}(x_0) - \mu(x_0)) \rightarrow_d N\left(0, \frac{V}{(\sum_s f(x_0|s))^2}\right). \tag{8}$$

It is easy to show (see appendix for the intermediate steps) that

$$\begin{aligned}
Var(t_{1i}) &= \sum_s \frac{1(s_i = s)}{a_s^2} Var_{|s} \left(\sum_{h=1}^k \varepsilon_{sih} K \left(\frac{x_{sih} - x_0}{h_n} \right) \frac{1}{k} \right) \\
&= \sum_s \frac{1(s_i = s)}{a_s^2} \frac{1}{k} E_{|s} \left\{ K^2 \left(\frac{x_{sih} - x_0}{h_n} \right) \sigma^2(x_{sih}) \right\} \\
&\quad + \sum_s \frac{1(s_i = s)}{a_s^2} \frac{k-1}{k} \left\{ E_{|s} \left(K \left(\frac{x_{sih} - x_0}{h_n} \right) K \left(\frac{x_{sih'} - x_0}{h_n} \right) \delta_s(x_{sih}, x_{sih'}) \right) \right\}
\end{aligned}$$

where

$$\begin{aligned}
\delta_s(x_{sih}, x_{sih'}) &= E(\varepsilon_{sih} \varepsilon_{sih'} | x_{sih}, x_{sih'}, s). \\
&= h_n \sum_s \frac{1(s_i = s)}{a_s^2} \frac{1}{k} \int (K^2(u) \sigma^2(x_0 + h_n u) f(x_0 + h_n u | s) du) \\
&\quad + h_n^2 \sum_s \frac{1(s_i = s)}{a_s^2} \frac{k-1}{k} \int \int dudv \left[\frac{K(u) K(v) \delta_s(x_0 + h_n u, x_0 + h_n v)}{g(x_0 + h_n u, x_0 + h_n v | s)} \right]
\end{aligned}$$

where $g(x_0 + h_n u, x_0 + h_n v | s)$ denotes the joint density of two sampled x -values from the same cluster in stratum s (note that since clusters are sampled within the strata, this joint density depends only on the strata). Therefore,

$$\begin{aligned}
&\frac{1}{h_n} \frac{1}{n} \sum_{i=1}^n Var(t_{1i}) \\
&= \sum_s \frac{1}{a_s} \frac{1}{k} \int (K^2(u) \sigma^2(x_0 + h_n u) f(x_0 + h_n u | s) du) \\
&\quad + h_n \sum_s \frac{1}{a_s} \frac{k-1}{k} \int \int dudv \left(\frac{K(u) K(v) \delta_s(x_0 + h_n u, x_0 + h_n v)}{g(x_0 + h_n u, x_0 + h_n v | s)} \right). \quad (9)
\end{aligned}$$

which converges to

$$\left\{ \int (K^2(u) du) \right\} \times \sigma^2(x_0) \sum_s \frac{1}{a_s} \frac{1}{k} f(x_0 | s)$$

as $n \rightarrow \infty$ and $h_n \rightarrow 0$. The striking feature of (9) is that the covariance terms are of smaller order than the variance terms, so that asymptotically, under the normalization with $(nh_n)^{1/2}$, the covariance terms vanish, implying that cluster effects do not matter

asymptotically. The intuition for the result is as follows. The asymptotics for the non-parametric estimate is driven by the number of clusters going to infinity, with the number of households staying fixed. Therefore, as the number of clusters increases (and the bandwidth h_n shrinks), in the h_n -neighborhood of a fixed observation x_0 , the proportion of households, from the same cluster as x_0 , goes to zero but the proportion of households from other clusters goes to infinity. This happens since the total number of observations in the neighborhood grows to infinity and X has a density in a neighborhood of x_0 . Results of somewhat similar spirit were derived in the time-series literature (c.f. Robinson, 1983). Note also that it is important for this result that not all population units in the population clusters be identical (existence of the joint density $g(\cdot, \cdot|s)$ and the fact that $\delta_s(x_{sc_s h}, x_{sc_s h'}) \neq \sigma^2(x|s)$ where $x = x_{sc_s h} = x_{sc_s h'}$ guarantees this). If that were true, I would get

$$\begin{aligned} \text{Var}(t_{1i}) &= \sum_s \frac{1}{a_s} \frac{1}{k} \int (K^2(u) \sigma^2(x_0 + h_n u) f(x_0 + h_n u|s) du) \\ &\quad + \sum_s \frac{1}{a_s} \frac{k-1}{k} \int (K^2(u) \sigma^2(x_0 + h_n u) f(x_0 + h_n u|s) du). \end{aligned}$$

and the non-vanishing cluster effects are the (limits of the) second term in this expression.¹⁶

The variance of the estimate for a finite sample (and finite bandwidth) can be estimated using the first term in (7) (after replacing all quantities by their estimates) as the relevant influence function and then applying the expressions in (6) without the cluster terms. Although the cluster correction term vanishes asymptotically, this term may not be ignorable in real applications for a finite sample size corresponding to the choice of a finite and non-zero bandwidth (see section 5 and table 4 for an illustration of this). So it is advisable to retain the cluster correction terms if it makes a significant difference to the results¹⁷.

¹⁶The stratum correction terms would also increase (implying adjusted standard errors would decrease by larger and larger proportions) as the bandwidth increases, but given the large number of observations within a stratum (unlike clusters), this rise will be modest.

¹⁷I am grateful to a referee for pointing this out.

3.2 U-statistics based estimates

Many estimators of the parametric component of a semiparametric model can be interpreted as minimizers of U-processes. See for instance Han (1987), Sherman (1994a, 1994b), Bhattacharya (2003), Honore and Powell (1994). In this subsection, I show how to derive large sample distribution theory for such estimators when the data come from stratified clustered samples.

I shall only consider U-statistics of order two based on all distinct pairs of observations. The sampling design is the same as above. Let N denote the total number of sampled observations (households) in the sample and n the total number of sampled clusters with $N = nk$. I am interested in finding the asymptotic distribution of the estimate (of the true parameter θ_0), which minimizes

$$Q_N(z, \theta) = \frac{1}{N(N-1)} \sum_I \sum_{J \neq I} Q_2(z_I, z_J, \theta)$$

where $z_I = (y_I, x_I)$ and $Q_2(\cdot, \cdot, \theta)$ is symmetric in the first two arguments.¹⁸ I shall re-index clusters from $1, \dots, n$ (keeping track of which stratum each cluster came from) and re-write this objective function as

$$Q_N(\mathbf{z}, \theta) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1, j \neq i}^n G(\mathbf{z}_i, \mathbf{z}_j, \theta) + \frac{1}{(n-1)k^2} \left\{ \frac{1}{n} \sum_{i=1}^n \left\{ \sum_{h=1}^k \sum_{h'=1, h' \neq h}^k Q_2(z_{ih}, z_{ih'}, \theta) \right\} \right\}$$

where i and j denote clusters and

$$\begin{aligned} \mathbf{z}_i &= (y_{i1}, \dots, y_{ik}, x_{i1}, \dots, x_{ik}) \\ G(\mathbf{z}_i, \mathbf{z}_j, \theta) &= \frac{1}{k^2} \sum_{h=1}^k \sum_{h'=1}^k Q_2(z_{ih}, z_{jh'}, \theta), i \neq j. \end{aligned}$$

The idea is to split the objective function into two parts- one involving distinct pairs of households from different clusters (the first term) and the second involving all distinct pairs of households that belong to the same cluster. Under the assumption that $\sum_{i=1}^n \left\{ \sum_{h=1}^k \sum_{h'=1, h' \neq h}^k Q_2(z_{ih}, z_{ih'}, \theta) \right\}$ is $O_p(n)$ uniformly in θ (which holds for instance

¹⁸Note that I do not use sampling weights in the estimation step since such problems usually arise from a specification of conditional mean or median, which is assumed to be identical in every stratum. So an unweighted procedure will yield consistent estimates of the parameter of interest.

if $Q_2(z_{ih}, z_{ih'}, \theta)$ are bounded, uniformly in θ , the second term is asymptotically negligible and one is left with

$$Q_N(\mathbf{z}, \theta) \simeq \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1, j \neq i}^n G(\mathbf{z}_i, \mathbf{z}_j, \theta)$$

which is a U-process on independent (though not identically distributed) observations. Using results from Sherman (1994), which require only the independence of the observations, it is straightforward to show that (under appropriate differentiability conditions and existence of finite moments) the minimizer of $Q_N(\mathbf{z}, \theta)$ is asymptotically equivalent to the minimizer of an empirical process on independent observations, whence \sqrt{n} -consistency and asymptotic normality follow. The minimizer itself will have an empirical process form with influence functions given by

$$\Psi_i(\theta_0) = 2 \sum_{s=1}^S 1(s_i = s) \sum_{h=1}^k \frac{1}{n-1} \frac{\partial}{\partial \theta} E \left\{ \sum_{s'} \sum_{j, (s', j) \neq (s_i, i)} Q_2(z_{s_i h}, z_{s' j 1}, \theta_0 | z_{s_i h}) \right\}.$$

Replacing population quantities by their sample counterparts and using numeric derivatives in place of analytic ones, one can estimate the covariance matrix of the minimizer. One can show

$$\begin{aligned} & \sqrt{n}(\hat{\theta} - \theta_0) \\ &= 2 \left\{ k \frac{\partial^2}{\partial \theta \partial \theta'} E \left[\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} Q_2(z_{ih}, z_{jh'}, \theta) \right] \right\}^{-1} \\ & \times \frac{1}{\sqrt{n}} \sum_{i=1}^n \sum_{s=1}^S 1(s_i = s) \sum_{h=1}^k \left(\frac{1}{n-1} \frac{\partial}{\partial \theta} E \left\{ \sum_{s'} \sum_{j, (s', j) \neq (s_i, i)} Q_2(z_{s_i h}, z_{s' j 1}, \theta_0 | z_{s_i h}) \right\} \right) \\ & + o_p(1) \end{aligned}$$

where $E_{z_j} Q_2(z_{ih}, z_j, \theta_0)$ denotes expectation taken w.r.t. the second random variable in $Q_2(\cdot, \cdot, \cdot)$. Let $\tilde{\nabla}_l Q_2(z_{ih}, z_j, \theta_0)$, $l = 1, 2$ denote 1st & 2nd order numeric derivatives of

$Q_2(z_{ih}, z_j, \theta_0)$,

$$\begin{aligned}\hat{\Psi}_{sih} &= \frac{1}{n-1} \left[\sum_{s' \neq s} \frac{1}{n_{s'}} \sum_{c_s=1}^{n_{s'}} \frac{1}{k} \sum_{h'=1}^k \hat{\nabla}_1 Q_2(z_{s'ih}, z_{s'c_s h'}, \hat{\theta}) \right. \\ &\quad \left. + \frac{1}{n_s-1} \sum_{c_s=1, c_s \neq i}^{n_s} \frac{1}{k} \sum_{h'=1}^k \hat{\nabla}_1 Q_2(z_{s'ih}, z_{s'c_s h'}, \hat{\theta}) \right] \\ \bar{\Psi}_{sh} &= \frac{1}{n_s} \sum_{c_s=1}^{n_s} \hat{\Psi}_{s'ih} \\ \bar{\Psi}_h &= \frac{\sum_{s=1}^S \bar{\Psi}_{sh}}{S} \\ \hat{\Phi} &= \frac{\sum_{I=1}^N \sum_{J \neq I} \hat{\nabla}_2 Q_2(z_I, z_J, \hat{\theta})}{N(N-1)}\end{aligned}$$

where $\sum_{I=1}^N \sum_{J \neq I}$ denotes the sum over all pairs of observations. Also let

$$\begin{aligned}\hat{\Sigma} &= \sum_{s=1}^S \sum_{i=1}^{n_s} \sum_{h=1}^k (\hat{\Psi}_{s'ih} - \bar{\Psi}_{sh}) (\hat{\Psi}_{s'ih} - \bar{\Psi}_{sh})' \\ \hat{\Delta} &= \sum_{s=1}^S \sum_{i=1}^{n_s} \sum_h \sum_{h' \neq h} (\hat{\Psi}_{s'ih} - \bar{\Psi}_{sh}) (\hat{\Psi}_{s'ih'} - \bar{\Psi}_{sh'})'\end{aligned}$$

A consistent estimate of the asymptotic variance of $\hat{\theta}$ is given by $4\hat{\Phi}^{-1} (\hat{\Sigma} + \hat{\Delta}) \hat{\Phi}^{-1}$. Note that

$$\begin{aligned}\hat{\Sigma} &= \sum_{s=1}^S \sum_{i=1}^{n_s} \sum_{h=1}^k (\hat{\Psi}_{s'ih} - \bar{\Psi}_h) (\hat{\Psi}_{s'ih} - \bar{\Psi}_h)' \\ &\quad - \sum_{s=1}^S n_s \sum_{h=1}^k (\bar{\Psi}_{sh} - \bar{\Psi}_h) (\bar{\Psi}_{sh} - \bar{\Psi}_h)'\end{aligned}$$

so that the cluster effect is $\hat{\Delta}$ and the stratum effect is $\sum_{s=1}^S n_s \sum_{h=1}^k (\bar{\Psi}_{sh} - \bar{\Psi}_h) (\bar{\Psi}_{sh} - \bar{\Psi}_h)'$.

4 Comments on Implementation

Data agencies differ in terms of availability of stratum and cluster information in the public-use data files. For almost all developing countries, including the LSMS surveys (which currently cover more than thirty developing countries for multiple years), the stratum and cluster identifiers are available in the public use micro-data. In some cases, these occur as

variables in the data set¹⁹. In other cases, the stratum and cluster identities are contained in the unique household identifier variable which is constructed by concatenating stratum and cluster identity numbers²⁰. Ideally, one should consult the sample design document to see what variables are the stratification and clustering based on and identify them in the micro-data before applying my methods. For several US surveys like the PSID and the HRS, the stratum and cluster information are available upon signing a sensitive data agreement for protection of respondents' privacies.

Different real-life surveys in the real world employ different number of levels of stratification and clustering. With multiple layers of stratification, only the final, i.e. the finest level of stratification matters and that is what should be used as the stratifying variable. e.g. if the stratification is first by state and then by districts within every state, then each state-district cell constitutes one stratum. For multiple layers of clustering (as in the PSID, the LSMS survey for Peru etc.), taking into account correlations between observations from the primary clusters suffices since this also takes into account correlations between units residing in secondary clusters. Thus no changes are warranted in my formulae when there are multiple levels of stratification and clustering; it is enough to set the stratum variable to the 'ultimate' stratifying variable and the cluster variable to the 'primary' level of clustering and then applying my formulae developed above for one stage each of stratification and clustering.

Sampling weights are included in all micro-data files. One needs to use the right weights depending on whether one is performing the analysis at a district level, household level, individual level etc. For instance, for the individual level analysis, household weights should be multiplied by household size.

5 Illustration with Indian NSS data

In this section, I briefly illustrate my procedures with data from the complexly designed Indian National Sample Survey (NSS) for 1993-4. This survey applies several levels of

¹⁹The terminologies vary between surveys: e.g. the LSMS survey for Azerbaijan lists strata as raions and clusters by the variable PPID, that for Pakistan are stratum and psu, for Peru it is regtype and cluster etc.

²⁰e.g. in the Albanian survey of the LSMS, the first two digits of the hhd id represent the bashki (stratum) and the next two represent the village (cluster)

stratification, first by states, then by sector (rural and urban) and finally by districts (rural) and size of town (urban). The final level of stratification is denoted by the variable ‘stratum’ in the data set. Within each stratum, clusters- villages in rural areas and blocks in urban areas- are selected at the first stage. From each selected cluster, a sample of ten households are drawn²¹. The clusters are identified by the variable “fsu_number”.

In Table 1, I report estimates of the mean and the Lorenz share at median (denoting the percentage of total resources accruing to the bottom 50% of the population) for monthly per capita household expenditure (mpce) (in the appendix, I show how Lorenz share estimation can be interpreted as a method of moment problem). I report both the weighted estimates and the unweighted naive ones for all India and the four largest states in the west, north, east and southern parts of the country. Estimates are provided separately for the entire state as well as for the rural and urban sectors.

For the mean, note that the naive estimates always overestimate the population mean. This happens because the NSS consciously tries to oversample wealthier households, relative to their population frequencies. The unweighted estimates therefore load the result disproportionately towards the wealthier households. A similar reasoning shows that unweighted estimates of the Lorenz share will produce systematically lower estimates since the relatively poorer people are under-represented in the sample. Weighting serves to correct for this under-representation and produces the consistent estimates of true population quantities.

In Table 2, I compare two different estimates of the standard error- one taking the design into account and the other not- *for the same estimate of the parameter* viz. the weighted consistent estimates of mean and Lorenz share. Since the results differ in interesting ways between rural and urban sectors, I report the two sectors separately. Columns 1-6 report the numbers for the mean and columns 7-12 for the Lorenz share at median. As explained in the footnote to the table, columns 2 and 3 (8 & 9 resp.) report the standard errors that, respectively, do and do not take the survey design into account and column 6 (12 resp.) reports the % change in standard errors due to overall design effects as percentage of the naive standard error estimates. Column 4 (resp., 10) shows the % change in standard errors

²¹In fact, the clusters are further stratified into a wealthy and a non-wealthy strata. 2 households are drawn from the wealthy strata and 8 from the poorer one. This second level of stratification is identified by the variable “substratum” in the data set.

(negative when the standard error declines) as a result of taking only stratification (and not clustering) into account (for instance, -30.97 in row 4, column 4 means that by taking stratification into account the estimate of the standard errors has fallen by 30.97% of the naive standard error for a consistent estimate of the mean for all of rural India). In terms of the expression in equation (6), this corresponds to the standard error one would get if one ignored the second term but included the third. The idea is to look at the separate contributions of the three terms in (6) to the overall standard error. Similarly column 5 (11, resp.) shows the increase in standard errors as a result of taking only clustering (and not stratification) into account.

Note from Table 2 that in general, cluster effects are larger than stratum effects. They are also much larger in urban areas relative to rural ones. The most likely explanation for this is that due to higher mobility in urban areas (better property markets and no strong attachment to land unlike the agricultural rural population), the urban population sorts itself more efficiently by income. So urban clusters are more homogeneous in terms of income. In other words, there are poor neighborhoods and rich neighborhoods in cities to a larger extent than there are rich villages and poor villages.

Next, note that the cluster effects are much larger for the mean than they are for the Lorenz share. This happens because the Lorenz shares are nonlinear functions of the data and the correlations between these nonlinear functions (of mpce, say) within a cluster tend to be smaller than the correlations between the mpce's themselves.

These results suggest that for countries with greater degrees of segregation, survey design will have stronger effects on standard errors through larger cluster effects. Strata being larger in size are likely to be less homogeneous and therefore will produce relatively smaller stratum effects on estimates of standard errors.

In Table 3, I report the Lorenz shares at medians corresponding to two successive rounds of the NSS survey- 1987-88 and 1993-94. I report the shares, the observed increases in Lorenz share at median in 1993-4, relative to 1987-8 and finally in the last two columns, I report the naive and the design-corrected t-statistics for testing hypotheses regarding the change in Lorenz shares²². The purpose of this table is to demonstrate that the relative

²²We have assumed here that the samples for the two different years are independent so that the variance of the differences in Lorenz shares is the sum of the variances of the shares for each year. Since clusters are sampled independently in the two years, this assumption is plausible.

magnitude of the standard errors become critical when testing *changes* in inequality. It is often the case that sample Lorenz shares actually move very little over long periods of time (c.f. column 3 of Table 3 where the largest change in Lorenz share at median over six years is merely 1.6 percentage points), which reinforces the importance of computing the correct standard errors. I observe from Table 3 that in the urban sector for all India as well as for each of the states, except Andhra Pradesh, corrected t-values would lead one to accept the hypothesis of no change at either 95% or 99% level, whereas the naive t-values would lead to the opposite conclusion. In the rural areas, that is not the case, as expected.

Finally, in Table 4, I provide results from a Nadaraya-Watson regression of (log) per capita calorie consumption on (log) mpce to illustrate the design effects on standard errors in nonparametric regressions, the theory of which is laid out in Section 3.1, above²³. I report the results for urban West Bengal where, as can be seen from Tables 2 and 3, the cluster effects were the largest. The value of (log) mpce at which I compute these regressions equals 10.70 (equivalent to 581.47 rupees), which is the sample median²⁴. I use a quartic kernel, given by

$$\begin{aligned} k(u) &= \frac{15}{16} (1 - u^2)^2, \quad -1 \leq u \leq 1 \\ &= 0, \text{ otherwise.} \end{aligned}$$

The optimal bandwidth (marked with an asterisk in Table 4) was chosen proportional to $n^{-1/5}$ where n equals the number of clusters (335) in the sample, with the constant of proportionality determined by minimizing a (leave-one-out) cross-validation criterion, over a large range of possible bandwidth choices²⁵. I report the results for a range of bandwidth choices ranging between 0.2 times the optimal bandwidth to 1.8 times the optimal bandwidth (in multiplicative increments of 0.2). We notice that at the optimal

²³The choice of this particular example is arbitrary since we are only using it for illustrations. No causal effects are to be inferred from them. For more careful treatments of calorie-income relations, see, for instance, Subramanian and Deaton (1996).

²⁴The choice of this point is motivated by the fact that one would expect the regression estimates at the ‘center’ of the expenditure distribution to be more accurate than at the boundaries. However, this is not to be interpreted as an estimate of the regression function at the population median. Otherwise, the correct standard error would have to account for the fact that the median is estimated.

²⁵Results from a local linear regression or choice of a higher order kernel for the purpose of bias reduction were of very similar orders of magnitude.

bandwidth, cluster effects are 34.16% and the overall increase in standard error 33% of the naive estimate of standard error. This reinforces the point that in finite samples, cluster effects may not be ignorable, although asymptotically, they should vanish. Moreover, as expected, the standard errors decrease and cluster effects increase as bandwidth increases. The effect of stratification is modest.

6 Conclusion

This paper has illustrated method of moments estimation when the data come from stratified, clustered surveys. It has outlined the methods of estimation of the parameters of interest and demonstrated the method of standard error computation that takes into account the survey design. The paper shows that ignoring stratification leads to overestimation of variances and the extent of overestimation increases with the degree of homogeneity inside and the degree of heterogeneity across strata. It also demonstrates that ignoring clustering likely leads to underestimation of variances with the extent of underestimation increasing in homogeneity within clusters. The analysis is presented for GMM-based estimation problems and extended to cover nonparametric regression and U-statistic based estimators for semiparametric models.

A related question, not covered in this paper, is how to best design a survey, given the financial constraints. Clearly, finer stratification and less clustering are desirable but also costlier. The formulas above should convince the reader that cluster and stratum effects upon standard errors also depend on the estimation problem at hand (through the moment functions). The scale of these effects are likely to be different for different types of estimation problems (e.g. means and medians). So in order to design a survey efficiently, one has to make a judgement on both what types of parameters are to be estimated from it and also how much is it worth (in dollars) to reduce standard errors by a certain percentage.

7 Appendix

Variance of nonparametric regressions

$$\begin{aligned}
& \text{Var}(t_{1i}) \\
&= \sum_s \frac{1(s_i = s)}{a_s^2} \text{Var}_{|s} \left(\frac{1}{k} \sum_{h=1}^k \varepsilon_{ih} K \left(\frac{x_{ih} - x_0}{h_n} \right) \right) \\
&= \sum_s \frac{1(s_i = s)}{a_s^2} \frac{1}{k} \text{Var}_{|s} \left(\varepsilon_{sih} K \left(\frac{x_{sih} - x_0}{h_n} \right) \right) \\
&\quad + \sum_s \frac{1(s_i = s)}{a_s^2} \frac{k-1}{k} \text{covar}_{|s} \left(\varepsilon_{sih} K \left(\frac{x_{sih} - x_0}{h_n} \right), \varepsilon_{sih'} K \left(\frac{x_{sih'} - x_0}{h_n} \right) \right) \\
&= \sum_s \frac{1(s_i = s)}{a_s^2} \frac{1}{k} \text{Var}_{i|s} \left(E_h \left(\varepsilon_{sih} K \left(\frac{x_{sih} - x_0}{h_n} \right) \mid s, i \right) \right) \\
&\quad + \sum_s \frac{1(s_i = s)}{a_s^2} \frac{1}{k} E_{i|s} \left(\text{Var}_{h|s,i} \left(\varepsilon_{sih} K \left(\frac{x_{sih} - x_0}{h_n} \right) \right) \right) \\
&\quad + \sum_s \frac{1(s_i = s)}{a_s^2} \frac{k-1}{k} \left\{ E_{|s} \left(\varepsilon_{sih} K \left(\frac{x_{sih} - x_0}{h_n} \right) \varepsilon_{sih'} K \left(\frac{x_{sih'} - x_0}{h_n} \right) \right) \right\} \\
&\quad - \sum_s \frac{1(s_i = s)}{a_s^2} \frac{k-1}{k} \left\{ E_{|s} \left(\varepsilon_{sih} K \left(\frac{x_{sih} - x_0}{h_n} \right) \right) E_{|s} \left(\varepsilon_{sih'} K \left(\frac{x_{sih'} - x_0}{h_n} \right) \right) \right\} \\
&= \sum_s \frac{1(s_i = s)}{a_s^2} \frac{1}{k} E_{i|s} \left(E_X \left(K^2 \left(\frac{x_{sih} - x_0}{h_n} \right) \sigma^2(x_{sih}) \mid i, s \right) \right) \\
&\quad + \sum_s \frac{1(s_i = s)}{a_s^2} \frac{k-1}{k} \left\{ E_{|s} \left(K \left(\frac{x_{sih} - x_0}{h_n} \right) K \left(\frac{x_{sih'} - x_0}{h_n} \right) \delta_s(x_{sih}, x_{sih'}) \right) \right\}
\end{aligned}$$

where

$$\begin{aligned}
& \delta_s(x_{sih}, x_{sih'}) = E_{|s}(\varepsilon_{sih} \varepsilon_{sih'} \mid x_{sih}, x_{sih'}) \\
&= h_n \sum_s \frac{1(s_i = s)}{a_s^2} \frac{1}{k} E_{c_s|s} \left(\left\{ \int (K^2(u) \sigma^2(x_0 + h_n u) f(x_0 + h_n u \mid s, c_s) du) \right\} \right) \\
&\quad + h_n^2 \sum_s \frac{1(s_i = s)}{a_s^2} \frac{k-1}{k} E_{c_s|s} \left\{ \int \int dudv \left[\begin{array}{c} K(u) K(v) \delta_s(x_0 + h_n u, x_0 + h_n v) \\ \times g(x_0 + h_n u, x_0 + h_n v \mid s) \end{array} \right] \right\}
\end{aligned}$$

Therefore,

$$\text{Var} \left(\frac{1}{\sqrt{h_n}} \frac{1}{\sqrt{n}} \sum_{i=1}^n t_{1i} \right) = \frac{1}{h_n} \frac{1}{n} \sum_{i=1}^n \text{Var}(t_{1i})$$

$$\begin{aligned}
&= \sum_s \frac{1}{a_s} \frac{1}{k} E_{c_s|s} \left(\left\{ \int (K^2(u) \sigma^2(x_0 + h_n u) f(x_0 + h_n u | s, c_s) du) \right\} \right) \\
&\quad + h_n \sum_s \frac{1}{a_s} \frac{k-1}{k} E_{c_s|s} \left\{ \int \int dudv \left(\begin{array}{c} K(u) K(v) \delta_s(x_0 + h_n u, x_0 + h_n v) \\ \times g(x_0 + h_n u, x_0 + h_n v | s) \end{array} \right) \right\}.
\end{aligned}$$

Lorenz shares

Here I report the moment functions for Lorenz share estimation. Recall that for a given percentile p , the Lorenz share at p is defined as

$$\phi(p) = \frac{\sum_{s=1}^S H_s E_{|s} \{y 1(y \leq Q(p))\}}{\sum_{s=1}^S H_s E_{|s} y} \equiv \frac{\alpha(p)}{\mu}, \text{ say}$$

where $Q(p)$ satisfies

$$p = \frac{\sum_{s=1}^S H_s \Pr\{y \leq Q(p) | s\}}{\sum_{s=1}^S H_s}$$

The corresponding sample moment conditions are

$$\begin{aligned}
\sum_{s=1}^S \sum_{c_s=1}^{n_s} \sum_{h=1}^k w_{sc_s h} \left(p - 1 \left\{ y_{sc_s h} \leq \hat{Q}(p) \right\} \right) &= 0 \\
\sum_{s=1}^S \sum_{c_s=1}^{n_s} \sum_{h=1}^k w_{sc_s h} \left(\hat{\alpha}(p) - y_{sc_s h} 1 \left\{ y_{sc_s h} \leq \hat{Q}(p) \right\} \right) &= 0 \\
\sum_{s=1}^S \sum_{c_s=1}^{n_s} \sum_{h=1}^k w_{sc_s h} (\hat{\mu} - y_{sc_s h}) &= 0.
\end{aligned}$$

Using the methods for GMM, developed in the text, one gets the joint asymptotic distribution of $(\hat{Q}(p), \hat{\alpha}(p), \hat{\mu})$. Using the usual delta method, the asymptotic distribution of $\hat{\phi}(p)$ follows.

References

- [1] Andrews, D., 1994. Empirical process methods in econometrics, in Handbook of econometrics vol 4, (ed) Engle, R. and McFadden, D., (North-Holland), 2248-2294.
- [2] Beach, C.M., Davidson, R., 1983. Distribution-free statistical inference with Lorenz curves and income shares, *Review of Economic Studies* 50, 723-34.
- [3] Bhattacharya, D., 2003. A simple estimator for monotone-index models; mimeo, Princeton University.
- [4] Butler, J.S., 1999. Efficiency Results of MLE and GMM Estimation using Sampling Weights, *Journal of Econometrics*, 96, Issue 1, 25-37.
- [5] Cochran, W., 1977. *Sampling techniques*, (Wiley, New York).
- [6] Conley, T. G., 1999. GMM Estimation with cross sectional dependence, *Journal of Econometrics*, September 1999; 92(1), 1-45.
- [7] Cosslett, Stephen R. 1981. Maximum Likelihood Estimator for Choice-Based Samples, *Econometrica*, vol. 49, no. 5, 1289-1316.
- [8] Cosslett, Stephen, R., 1993. Estimation from endogenously stratified samples, in G.S. Maddala, C.R. Rao and H.D. Vinod (eds.) *Handbook of statistics* Vol. 11, 1-43.
- [9] Davidson, R., Duclos, J., 2000. Statistical inference for stochastic dominance and for measurement of poverty and inequality, *Econometrica* 68, 1435-64.
- [10] Deaton, A., 1997. *Analysis of household surveys: a microeconomic approach to Development policy* (Johns Hopkins Press).
- [11] DuMouchel, W.H., Duncan G.J., 1983. Using sample survey weights in multiple regression analysis of stratified samples, *Journal of the American Statistical Association*, 78, 535-43.
- [12] Francisco, C., Fuller, W., 1991. Quantile estimation with a complex survey design, *Annals of Statistics*, 19, 454-69.

- [13] Han, A.K., 1987. Non-parametric analysis of a generalized regression model, *Journal of Econometrics*, 35, 303-316.
- [14] Honore, Bo, Powell, J., 1994. Pairwise difference estimators of censored and truncated regression models, *Journal of Econometrics*, 64(1-2), 241-78.
- [15] Howes, S., Lanjouw, J.O., 1998. Making poverty comparisons taking into account survey design, *Review of Income and Wealth*, March, 1998, 99-110.
- [16] Imbens, Guido W., 1992. An efficient method of moments estimator for discrete choice models with choice-based sampling, *Econometrica*, 60, no. 5, 1187-214.
- [17] Imbens, G. and Lancaster, T., 1996. Efficient estimation and stratified sampling; *Journal of econometrics*, 74, 289-318.
- [18] Kish, Leslie., 1965. *Survey Sampling*. New York, NY. John Wiley & Sons.
- [19] Kish, Leslie and Frankel, M., 1970. Inference from complex samples, *Journal of the Royal statistical society, series B*, 36, 1-37.
- [20] Kloek, T., 1981. OLS estimation in a model where a microvariable is explained by aggregates and contemporaneous disturbances are equicorrelated; *Econometrica*, 49(1), 205-07.
- [21] Manski, Charles F., Lerman, Steven R., 1977. The estimation of choice probabilities from choice based samples, *Econometrica*, 45, no. 8, 1977-88.
- [22] Moulton, B., 1986. Random group effects and the precision of regression estimates, *Journal of Econometrics*, 32, 385-97.
- [23] Murthy, M., 1977. *Sampling theory and methods*; Calcutta, statistical publishing company.
- [24] Newey, W., McFadden, D., 1994. Large sample estimation and hypothesis testing, *Handbook of econometrics vol 4*, (ed) Engle, R. and McFadden, D., 2111-2241.

- [25] Pepper, J.V., 2002. Robust inferences from random clustered samples: an application using data from the panel study of income dynamics, *Economics Letters*, 75, Issue 3, 341-345.
- [26] Pfeiffermann, D., Nathan, G., 1981. Regression analysis of data from a cluster sample; *Journal of the American statistical association*; vol 76, no. 375, pages 681-689.
- [27] Robinson, P.M., 1983. Nonparametric estimators for time series. *Journal of Time Series Analysis*, 4, 185-207.
- [28] Sakata, Shinichi, Quasi-Maximum likelihood estimation with complex survey data. (work in progress). University of Michigan.
- [29] Sherman, R.P., 1994a. Maximal inequalities for degenerate U-processes with applications to optimization estimators, *Annals of Statistics*, 22, 439-459.
- [30] Sherman, R.P., 1994b. U-Processes in the analysis of a generalized semiparametric regression estimator, *Econometric Theory*, 10, iss. 2, 372-95.
- [31] Subramanian, S., Deaton, A., 1996. The demand for food and calories, *Journal of Political Economy*, 104, 133-62.
- [32] White, H., 1980. A Heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity; *Econometrica*, 48, 817-838.
- [33] Wooldridge, J., 1999. Asymptotic properties of weighted M-estimators for variable probability samples; *Econometrica*, Vol.67, no. 6, 1385-1406.
- [34] Wooldridge, J., 2001a. *Econometric analysis of cross-section and panel data*, (MIT press).
- [35] Wooldridge, J., 2001b. Asymptotic properties of weighted M-estimators for standard stratified samples, *Econometric Theory*, 17, 451-470.
- [36] Zheng, B., 2002. Testing Lorenz curves with non-simple random samples, *Econometrica*, vol. 70 (3), 1235-1243.