9-2016

Is Predicted Data a Viable Alternative to Real Data?

Tomoki FUJII
Singapore Management University, tfujii@smu.edu.sg

Roy VAN DER WEIDE
World Bank

WPS7841

Policy Research Working Paper 7841

Is Predicted Data a Viable Alternative to Real Data?

Tomoki Fujii
Roy van der Weide

WORLD BANK GROUP
Development Research Group
Poverty and Inequality Team
September 2016

Policy Research Working Paper 7841

Abstract

It is costly to collect the household- and individual-level data that underlies official estimates of poverty and health. For this reason, developing countries often do not have the budget to update their estimates of poverty and health regularly, even though these estimates are most needed there. One way to reduce the financial burden is to substitute some of the real data with predicted data. An approach referred to as double sampling collects the expensive outcome variable for a sub-sample only while collecting the covariates used for prediction for the full sample. The objective of this study is to determine if this would indeed allow for realizing meaningful reductions in financial costs while preserving statistical precision. The study does this using analytical calculations that allow for considering a wide range of parameter values that are plausible to real applications. The benefits of using double sampling are found to be modest. There are circumstances for which the gains can be more substantial, but the study conjectures that these denote the exceptions rather than the rule. The recommendation is to rely on real data whenever there is a need for new data, and use the prediction estimator to leverage existing data.

This paper is a product of the Poverty and Inequality Team, Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at rvanderweide@worldbank.org.

Is predicted data a viable alternative to real data? *

Tomoki Fujii† Roy van der Weide‡

JEL classification c odes: C20, C53, I32.

Keywords: Prediction; Double sampling; Survey costs; Poverty.

*This study has been funded by the World Bank's Knowledge for Change Program for which the authors are grateful.

†Singapore Management University. Email: tfujii@smu.edu.sg

‡World Bank. Email: rvanderweide@worldbank.org

1 Introduction

In economics, health sciences, and other disciplines, data on the outcome variable of interest is often costly to collect. The measurement of poverty for instance relies on household consumption expenditure data. Collecting this data involves long questionnaires administered over an extended period of time which substantially adds to the data collection costs. Individual-level data collection that involves physical examinations also tends to be costly. This has implications for the sample size and the frequency with which the data is collected. In many developing countries for example, particularly in Sub-Saharan Africa, estimates of poverty, malnutrition and health are obtained highly irregularly and as a result are often outdated.[1]

The demand for bigger and more frequent data has motivated researchers to explore ways to substitute predicted data for real data. Consider the application to poverty measurement by Douidich et al. (2015) where household consumption poverty is predicted into annual labor force surveys in order to increase the frequency of poverty estimates from once every 7 years to every year.[2] Similar applications can be found in Stifel and Christiaensen (2007) and Christiaensen et al. (2012). Prediction methods have also been used to dramatically expand the sample size of poverty and health data. A prominent example of this is the small area estimation of welfare where the outcome variable of interest (i.e. consumption poverty) is predicted into a population census which covers (almost) all members of the population unlike typical household surveys. This allows researchers to obtain estimates of poverty at a highly disaggregated level such as districts, communities, and towns. These small-area estimates are often plotted in the form of a map known as a poverty map. The small-area estimation, which provides the methodological foundations for poverty mapping, was pioneered by Elbers et al. (2003) and extended by e.g. Tarozzi and Deaton (2009), Tarozzi (2011) and Elbers and van der Weide (2014). Fujii (2010) has modified the approach to obtain small areas estimates of the prevalence of

[1] For example, across the 26 low-income countries in Sub-Saharan Africa over the period between 1993 and 2012, the national poverty rate and prevalence of stunting for children under five are on average reported only once every five years and once every ten years in the World Development Indicators.

[2] This approach could of course be expanded to other outcome variables of interest (such as health outcomes), while surveys other than labor force surveys could be considered to further extend the frequency of estimates. The same approach could also be adopted to construct a measure of consumption poverty that alleviate concerns of comparability over time when the original consumption data is deemed incomparable due to changes in the questionnaire. See for example the debate on the comparability of poverty estimates in India documented in e.g. Deaton (2003), Deaton and Drèze (2002), Kijima and Lanjouw (2005), Tarozzi (2007), and Deaton (2005).

stunting and underweight of children using predicted data.

When new data is needed, it may be tempting to purposefully only collect the covariates x, which are used to predict the outcome variable of interest y, and to scale down the collection of y itself, particularly if this yields a significant financial cost reduction. It is not uncommon that predictors of household welfare such as demographic and dwelling characteristics, education, employment, and asset ownership, can indeed be collected relatively inexpensively. In health sciences, simple oral questions and anthropometric data taken by a simple non-invasive device too may serve as a predictor of the outcome that is expensive to measure. Collecting y for a sub-sample only, and the covariates x for all, is referred to as "double-sampling",[3] see e.g. Hidiroglou (2001).[4] The advantage of this approach is that the prediction model can be estimated with data for the relevant population and time period (the same population for which the covariate data is available). The alternative where the model is estimated to data from an entirely different dataset that may describe a different population at a different point in time is referred to as "non-nested double sampling", in which case one will have to make assumptions about how the model has evolved between the two different datasets (or assume that the model is invariant), see e.g. Kim and Rao (2012) as well as Douidich et al. (2015) for an empirical application to poverty measurement. We will refer to any estimator that works with predicted data (i.e. imputations of y) as a prediction estimator.

There is a considerable practical interest in adopting a double-sampling approach and relying on predicted data in the hope of reducing financial costs while preserving a reasonable degree of statistical precision. Consider for example the recent initiative by the World Bank under the name of SWIFT (Survey of Well-being via Instant and Frequent Tracking) which "does not collect direct income or consumption data which can be both time-consuming and vulnerable to error without the right know-how and resources; instead, it collects poverty correlates, such as household size, ownership of assets or education levels, and then converts them

[3] It is also known as "two-phase sampling".

[4] The literature on double sampling dates back to Neyman (1938) and Bose (1943). Many of the existing studies, including Hidiroglou (2001), Kim et al. (2006), Palmgren (1987), Rao and Sitter (1995), and Sitter (1997), provide analytical or simulation results on the properties of the estimators based on double sampling. While there are some explicit empirical applications of double sampling such as Tamhane (1978), Hansen and Tepping (1990), and Armstrong et al. (1993), the use of double sampling appears to be comparatively limited. Even fewer studies take into account the costs of data collection (Cochran (1977), Davidov and Haitovsky (2000), Särndal et al. (2003), and Fujii and van der Weide (2013)), despite the fact that one major attraction of double sampling is the reduction in data collection cost.

to poverty statistics using estimation models." (Yoshida et al. (2015)). It employs Computer Assisted Personal Interviewing technology to collect the data and carefully builds and estimates the model that is used for prediction. Both nested and non-nested double sampling approaches are considered.[5] While SWIFT is still relatively young, at the time of writing, it has already been applied 34 times in 27 countries. Assessments of the cost-precision trade-offs however are still limited. Ahmed et al. (2014) denotes an exception which provides a SWIFT-like application of double sampling to poverty estimation in Bangladesh, where a variety of different data collection scenarios are being considered. Their simulation results indicate that substantial cost savings could be achieved with a moderate loss in precision. It is unclear however how much of this may be attributed to the use of double sampling as, in addition to relying on predicted data, they also reduce the number of primary sampling units.[6]

Pape and Mistiaen (2015) apply a double sampling approach to poverty measurement in Mogadishu. They extend the set of predictors by including a sub-set of consumption items. Ligon and Sohnesen (2016) explore a similar strategy and include an empicical illustration using data from Uganda, Tanzania and Rwanda. Both these studies build on the idea that was put forward by Lanjouw and Lanjouw (2001). This approach strengthens the correlation with total household consumption but also adds to the costs, although Pape and Mistiaen (2015) in their application to Mogadishu are able to keep the face-to-face interview time below 60 minutes.[7] Other popular examples of prediction-based poverty estimation include the "Simple Poverty Scorecard" (SPS) project and the "Poverty Assessment Tool" (PAT) developed by the IRIS Center for USAID, see Schreiner (2014a) for a comparison.[8] These approaches predict a household's poverty status on the basis of a small number of questions, mostly relying on non-nested double sampling, and then aggregate these predictions to obtain estimates of poverty at the national level (and possibly other administrative levels). SPS and PAT have been applied to

[5] In its version of the non-nested double sampling approach infrequent conventional surveys, which collect both y and x for all households, are alternated with more frequent low cost surveys that only collect x. The infrequent full surveys are used to estimate (and update) the prediction model which is then used to predict poverty into the subsequent low cost surveys that are conducted for the years in between the full surveys.

[6] Furthermore, they do not provide a full break-down of the costs involved; only selected cost components are reported which excludes transport costs for example.

[7] Security concerns denote an important motivation for reducing the time of data collection by means of face-to-face interviews.

[8] The poverty scorecard takes a more pragmatic approach to building its prediction model as it emphasizes simplicity and ease of implementation. See Schreiner (2014b) for further details on the Simple Poverty Scorecard.

around 60 and 40 countries, respectively. A recent evaluation can be found in Diamond et al. (2015). To the best of our knowledge, no cost-precision assessments are available for any of these initiatives.

The objective of this study is to determine if and when double sampling may reasonably be expected to yield meaningful reductions in the cost of data collection while preserving statistical precision. We do this using analytical calculations that rely on an approximation to the financial cost function and on the asymptotic variance as the measure of statistical precision. This allows us to consider an inclusive set of parameter values. We subsequently make an attempt to calibrate the parameters involved to a variety of real data. Specifically, we solve a cost minimization problem subject to a statistical precision constraint and its dual problem of variance minimization problem subject to a budget constraint. This helps us identify the conditions under which the gains from double sampling are relatively large (and small). To the best of our knowledge, this is the first study to attempt an analytical assessment of how much precision prediction estimators trade for financial cost savings. We treat the sample direct estimator and the prediction estimator under single- and double-sampling in a unified framework and derive the analytic results for cost or variance reduction. The assumed model allows for clustering, which plays an important role in empirical applications but which is often ignored in theoretical work.

We find that the financial gains from double sampling over optimal single sampling tend to be modest for many of the parameter values considered, which are calibrated to real data. The magnitude of the potential discounts are mostly below 20 percent and can be as low as zero percent. There are circumstances in which the gains can be more substantial, but we conjecture that these denote the exceptions rather than the rule. Double sampling is most advantageous when: (a) the marginal cost of collecting y is particularly large, (b) x is highly correlated with y, (c) travel costs between clusters are modest, (d) the sampling error is large relative to the model error, and (e) the spatial correlation in the data is modest. Unfortunately, these conditions are rarely jointly satisfied. When the covariates x are particularly low-cost (as is the case with SPS and PAT), then the correlation between x and y tends to be weak. When the covariates offer exceptionally good predictors (as may be the case in Pape and Mistiaen (2015) and Ligon and Sohnesen (2016)), then the marginal cost of collecting y tends to be low. SWIFT arguably lies

somewhere in between.

We have assumed away any error that may stem from model misspecification, i.e. all results hold under the standard assumption that the prediction model is correctly specified. This means that the real gains realized by the prediction estimators could be more modest yet. Model misspecification would introduce an entirely new source of error which is not accounted for in estimates of statistical precision.[9] Consequently, applied users should always bear in mind that prediction estimators are arguably less precise than is suggested by conventional standard errors.

The financial savings are larger for non-nested double sampling estimators. However these are also based on stronger assumptions. The added assumption is that the model parameters did not change between the dataset used for estimation and the dataset used for prediction which can be multiple years apart.[10] This adds to the risk of model misspecification error.

Our recommendation is to rely on real data on the outcome variable of interest y whenever there is a need for new data. There is an argument for scaling back the collection of y, and collecting covariates of y instead, if there is insufficient budget to accommodate an adequate sample of observations with real data on y (and x). Note that this does not in any way rule out the use of prediction estimators, such as the approaches employed in Douidich et al. (2015) and Elbers et al. (2003). This variety of non-nested double sampling estimators provides researchers with a means of leveraging existing data, which adds to the value of these data.

The remainder of this paper is organized as follows. Section 2 presents the prediction estimators under the single- and double-sampling setup. Then, they are applied to cost-effective sampling in Section 3. We study the conditions under which double sampling is most useful and explore the magnitude of potential gains from double sampling under practical conditions in Section 4. Finally, Section 5 provides some discussion.

[9] As any given model can be misspecified in infinitely different ways, there are currently no methods available that would account for model misspecification error.

[10] One does not necessarily need to assume that the model parameters are time-invariant. Alternatively, one could make assumptions about how the model has evolved exactly between the two datasets, but this assumption is just as strong.

2 Prediction estimator

2.1 Preliminaries

Consider the following data generating process:

$$y_{ch} = x_{ch}^T \beta + u_{ch} = x_{ch}^T \beta + \eta_c + e_{ch},$$

where c and h are the indexes of clusters and households, respectively. The continuous state variable y_{ch}, which may or may not be observable, is related to the observable outcome variable of interest $Y_{ch} = m(y_{ch})$ by some function m. The L-vectors of (observable) covariates and coefficients are denoted by x_{ch} and β, respectively. The idiosyncratic error term $u_{ch}(= \eta_c + e_{ch})$, which consists of the cluster-specific error η_c and the household-specific error e_{ch}, is unobservable. This error structure allows for some degree of spatial correlation in the errors. We denote the variances of error terms by $\sigma_e^2 \equiv \text{var}[e_{ch}]$, $\sigma_\eta^2 \equiv \text{var}[\eta_c]$, and $\sigma_u^2 \equiv \text{var}[u_{ch}](= \sigma_e^2 + \sigma_\eta^2)$.

Each cluster is assumed to consist of K sampled households. We make the following assumptions about x_{ch}, η_c, and e_{ch}:

Assumption 1 *The triple* $(\{x_{ch}\}_{h=1}^K, \{e_{ch}\}_{h=1}^K, \eta_c)$ *is iid across* c.

Assumption 2 x_{ch}, e_{ch}, *and* η_c *are independent of each other for all* c. *Furthermore,* e_{ch} *is independent across* h *for all* c.

It is convenient to denote the stacked error terms for households in cluster c by $U_c \equiv (u_{c1}, u_{c2}, \ldots, u_{cK})^T$ and all households by $U \equiv (U_1^T, U_2^T, \ldots, U_J^T)$, where J is the number of clusters in the sample. Similarly, we denote all x's stacked together in cluster c by X_c, and all X_c stacked together by X. We define $\Omega \equiv E[UU^T]$ and $\Omega_c \equiv E[U_c U_c^T]$.

Let us define the expected value of Y_{ch} conditional on x_{ch} by $g(x_{ch}, \theta) \equiv E_u[Y_{ch}|x_{ch}]$, where θ is a κ-vector of identifiable model parameters and g is assumed differentiable with respect to θ. We also define $\varepsilon_{ch} \equiv Y_{ch} - g(x_{ch}, \theta)$.

The parameter of interest in this study is $\mu \equiv E[Y_{ch}] = E_x[g(x_{ch}, \theta)]$ and not θ. The standard estimator for μ is the sample mean $\bar{Y} = n^{-1} \sum_c \sum_h Y_{ch}$, where $n = JK$ denotes the sample size. This estimator will be referred to as the sample direct estimator.

2.2 Binary outcome variable

For ease of exposition, we specialize in a case where the outcome variable is binary. This is an important special case, because binary outcomes are routinely encountered in empirical applications. Examples include the poverty rate (proportion of the individuals under the poverty line), prevalence of undernourished children, and the share of underemployed people among the employed individuals. We will subsequently use the poverty rate for the purpose of illustration, but it is straightforward to apply our theory to other contexts.

In the context of binary outcome, we maintain the following assumption:

Assumption 3 *The variables Y_{ch} and y_{ch} are related by $Y_{ch} = \text{Ind}(y_{ch} < z)$, where z is a constant. Furthermore, η_c and e_{ch} are normally distributed.*

The normality of η_c and e_{ch} is not essential but it helps us to simplify our presentation as u_{ch} is also a normal random variable in this case. When μ represents the poverty rate, the constant z corresponds to the poverty line. We denote the probability density function and cumulative distribution function for a standard normal random variable by ϕ and Φ, respectively. In this case it follows that: $g(x_{ch}, \theta) = \Phi\left((z - x_{ch}^T\beta)/\sigma_u\right)$.

2.3 Single-sample prediction estimator

The prediction estimator for μ relies on the assumption that the functional form of g is known.[11] This is true whether we use the single-sample or the double-sample estimator. Consider first the single-sample prediction estimator:

$$\hat{\mu}(\hat{\theta}) \equiv n^{-1} \sum_c \sum_h g(x_{ch}, \hat{\theta}), \tag{1}$$

where $\hat{\theta}$ denotes the estimator for θ. All the model uncertainty is captured by the estimator $\hat{\theta}$ of the unknown model parameter θ. Rather than taking the sample average over the actual realizations of Y_{ch}, it estimates the average of the conditional mean g given x_{ch}. For the binary outcome variable the prediction estimator is seen to solve:[12]

[11] We avoid using the term "regression estimator" in this study because it typically refers to the prediction estimator under the assumption of linearity (and often a single covariate).

[12] Notice that the predicted values do not incorporate estimates of the cluster-specific effects η_c conditional on the available data in eq. (2). In other words, it is not an Empirical Bayes estimator (See Elbers and van der

$$\hat{\mu} = \frac{1}{J}\sum_{c}\frac{1}{K}\sum_{h}\Phi(\hat{B}_{ch}) = \frac{1}{J}\sum_{c}\frac{1}{K}\sum_{h}\Phi\left(\frac{z - x_{ch}^{T}\hat{\beta}}{\hat{\sigma}_{u}}\right), \tag{2}$$

where: $B_{ch} \equiv (z - x_{ch}^{T}\beta)/\sigma_{u}$.

The sample direct and prediction estimators are subject to different sources of error. The former is purely a function of the sampling error. The latter trades some of the sampling error for the model error. More precisely, it averages out the error terms η_c and e_{ch} and reduces the sampling error component. However, it introduces the model error instead because the prediction estimator uses an estimate of θ rather than the true parameter value. Put differently, the contributions of η and e to the error of the estimate of μ are "re-packaged" from sampling error to model error in the prediction estimator.

To study the properties of prediction estimators, we need to make some assumptions about $\hat{\theta}$. Let us begin conservatively by merely assuming that $\hat{\theta}$ is a consistent and asymptotically normal estimator for θ. Note that this accommodates practically all commonly used estimators.

Assumption 4 *The estimator $\hat{\theta}$ of the model parameters θ satisfies the following properties:*

$$\hat{\theta} \xrightarrow{p} \theta \quad \text{and} \quad \sqrt{J}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, V_{\theta}/K) \quad \text{as} \quad J \to \infty,$$

where V_θ is a symmetric positive-definite $\kappa \times \kappa$ asymptotic covariance matrix of $\hat{\theta}$.

Remark 5 *Because K is fixed, there is a bijective correspondence between n and J. Therefore, it is also possible to write Assumption 4 as follows:*

$$\hat{\theta} \xrightarrow{p} \theta \quad \text{and} \quad \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, V_{\theta}) \quad \text{as} \quad n \to \infty.$$

Hereafter, we assume that suitable regularity conditions always hold. In particular, we assume the almost-sure existence and non-singularity of relevant moments, which typically pose no problem in empirical applications. With these assumptions, we have the following theorem (all proofs are in the Appendix):

Weide (2014) for an application of Empirical Bayes estimator in a similar context). While the prediction estimator in eq. (2) is generally inefficient when y_{ch} is observed, we chose to use this form for two reasons. First, the exposition is much simpler when eq. (2) is used. Second, the subsequent discussion is applicable without much modifications even when y_{ch} and z are unobservable.

Theorem 6 *Let $M_g \equiv E[\partial g(x_{ch}, \theta)/\partial\theta](\neq 0_\kappa)$, where 0_κ is a column κ-vector of zeros. Then, under Assumptions 1, 2, and 4, we have:*

$$\hat{\mu} \xrightarrow{p} \mu \quad \text{and} \quad \sqrt{J}(\hat{\mu}(\hat{\theta}) - \mu) \xrightarrow{d} \mathcal{N}\left(0, V_g + M_g^T V_\theta M_g / K\right) \quad \text{as} \quad J \to \infty, \tag{3}$$

where $V_g \equiv \text{var}_{x_{ch}}[K^{-1}\sum_{h=1}^{K} g(x_{ch}, \theta)]$.

The factor K^{-1} in the definition of V_g is included to normalize the variance at the cluster level. Notice that $V_g = K^{-1}\text{var}[g(x_{ch}, \theta)]$ holds when x_{ch} is independent across all c and h. It should also be noted that the asymptotic variance of $\hat{\mu}(\hat{\theta})$ can be consistently estimated by replacing V_g, M_g and V_θ with their consistent estimators. For ease of presentation, we hereafter simply write $\hat{\mu}$ dropping the argument $\hat{\theta}$.

As discussed earlier, the prediction estimator $\hat{\mu}$ may be more precise than the sample direct estimator $\bar{Y}$. The improvement in precision achieved by the prediction estimator is not necessarily the main attraction of the prediction estimator; its advantage is that it allows for a double sampling strategy, which potentially leads to the reduction in financial costs for collecting survey data without compromising the statistical precision. We pursue this idea in Section 3.

2.4 Double-sample prediction estimator

The prediction estimator given in eq. (2) uses a single sample. However, it is clear that we only need the observations of x_{ch} once the estimates of β, σ_e, and σ_η are obtained. This, in turn, means that it is not necessary to observe Y (or y) for all sample households; a sub-sample will do. We may still use the full sample to evaluate the mean prediction estimator if x is collected for all. That is, even when data on Y (or y) is collected only for a sub-sample, provided that x is collected for the full sample of households, predicted values of Y may still be evaluated for the full sample. This approach is referred to as "double sampling".

If observing Y is much more expensive than observing x, then double sampling may be preferred to the standard single sampling approach—where both x and Y are observed for all households in the sample; double sampling has the potential to realize a reduction in the financial costs associated with collecting the necessary survey data while maintaining the desired statistical precision.

To formally introduce double sampling, we make a few assumptions.

Assumption 7 *The covariates x are observed for all sample households, while y is observed for the first $k \leq K$ households in all J clusters.*[13]

Let s_{ch} denote the indicator variable that equals one if the household is in the full sample and zero otherwise. We further denote by s^I_{ch} [$s^{II}_{ch} (= s_{ch} - s^I_{ch})$] the indicator variable that the household is in the sub-sample containing data on both y and x [x only]. Using this notation, the number of households included in the former and latter sub-samples equals, respectively, rn and $(1-r)n$, where n is the sample size of the full sample and $r (= k/K)$ is the ratio of sample households with an observation of y.

Assumption 8 *The distribution of (x_{ch}, η_c, e_{ch}) is independent of (s^I_{ch}, s^{II}_{ch}).*

This requires that the selection into either sample carries no information about x_{ch}, η_c, or e_{ch}. This is a reasonable assumption in our setup because the researcher chooses whether to observe y_{ch}.

We use the households with the observations of y and x to compute the estimator $\hat{\theta}^I$ of θ, where $\hat{\theta}^I$ satisfies the following assumption, which is a double-sample analogue of Assumption 4:

Assumption 9 *The estimator $\hat{\theta}^I$ of the model parameters θ satisfies the following properties:*

$$\hat{\theta}^I \xrightarrow{p} \theta \quad \text{and} \quad \sqrt{J}(\hat{\theta}^I - \theta) \xrightarrow{d} \mathcal{N}(0, V^I_\theta / k) \quad \text{as} \quad J \to \infty,$$

where V^I_θ is a symmetric positive-definite $\kappa \times \kappa$ asymptotic covariance matrix of $\hat{\theta}^I$.

[13] Since data on x is already being collected for all households in all clusters of the sample, the marginal cost of collecting data on y from all clusters is identical to the cost of collecting y for the same number of households from half the number of clusters, say. Given that any two households from different clusters carry more information than a pair of households from the same cluster, due to spatial correlation, it is optimal to collect y for a sub-sample of households from all clusters. The added advantage of collecting household expenditure data from all clusters is that it allows the user to construct unit value prices for each cluster which are needed to convert nominal household expenditure data into real terms. Alternatively, the survey could collect its price data by visiting local markets, i.e. by including a community price module. In this study we will abstract away from spatial (and temporal) price adjustments, and assume that this is taken care off. Note however that this is not a trivial matter, see e.g. van Veelen and van der Weide (2008).

Using $\hat{\theta}^I$, we can predict μ by $g(x_{ch}, \hat{\theta}^I)$ for all observations in S. Therefore, the prediction estimator for μ under Assumption 7 is given by:

$$\hat{\mu}^{DS} \equiv n^{-1} \sum_{ch \in S} g(x_{ch}, \hat{\theta}^I).$$

The following is a direct extension of Theorem 6 to the double-sample estimator:

Theorem 10 *Suppose that Assumptions 1, 2, 7, 8, and 9 hold. Then, $\hat{\mu}^{DS}$ satisfies the following properties as $J \to \infty$:*

$$\hat{\mu}^{DS} \xrightarrow{p} \mu, \tag{4}$$

$$\sqrt{J}(\hat{\mu}^{DS} - \mu) \xrightarrow{d} \mathcal{N}(0, V_g + r^{-1} M_g^T V_\theta^I M_g / K) \tag{5}$$

Note that Theorem 6 is a special case of Theorem 10 when $r = 1$ (i.e., $k = K$).

In the binary context, we have the following results:

Theorem 11 *Suppose that Assumptions 1, 2, 3, 7, and 8 hold. Further, $\hat{\theta}^I = (\hat{\beta}^T, \hat{\sigma}_\eta^2, \hat{\sigma}_e^2)^T$ is a maximum-likelihood estimator. Then, $\hat{\mu}^{DS}$ satisfies the following as $J \to \infty$:*

$$\begin{cases} \hat{\mu}^{DS} \xrightarrow{p} \mu \\ \sqrt{J}(\hat{\mu}^{DS} - \mu) \xrightarrow{d} \mathcal{N}\left(0, V_g + \frac{\Sigma_{\phi x}^T E^{-1}[X_c^T \Omega_c^{-1} X_c] \Sigma_{\phi x}}{\sigma_u^2} + \frac{\Sigma_{\phi B}^2 (k\sigma_\eta^4 + 2\sigma_e^2 \sigma_\eta^2 + \sigma_e^4)}{2k\sigma_u^4}\right), \end{cases} \tag{6}$$

where $\Sigma_{\phi x}$ and $\Sigma_{\phi B}$ are defined as follows:

$$\Sigma_{\phi x} \equiv E[\phi(B_{ch}) x_{ch}], \quad \text{and} \quad \Sigma_{\phi B} \equiv E[\phi(B_{ch}) B_{ch}].$$

Further, Ω_c^{-1} can be written as:

$$\Omega_c^{-1} \equiv \frac{1}{\sigma_e^2} \left[I_k - \frac{\sigma_\eta^2}{\sigma_e^2 + k\sigma_\eta^2} \mathbf{1}_k \mathbf{1}_k^T \right],$$

where I_k and $\mathbf{1}_k$ are the $K \times K$-identity matrix and K-vector of ones, respectively.

Let us further develop the expression for the asymptotic variance by making some modest simplifying assumptions that will ease the exposition.

Assumption 12 *The vector of covariates x_{ch} can be written as $x_{ch} = x_c^0 + x_{ch}^1$, where x_c^0 is iid across c, x_{ch}^1 is iid across c and h, $E[x_{ch}^1] = 0$, $E[x_c^0(x_c^0)^T] \equiv \Sigma_{xx}^0$, and $E[x_{ch}^1(x_{ch}^1)^T] \equiv \Sigma_{xx}^1$.*

This assumption essentially states that the covariates can be decomposed into cluster- and household-specific components. The following lemma follows directly from the Law of Total Variance and the definition of V_g.

Lemma 13 *Under Assumptions 1 and 12, the sampling variance component V_g in eq. (6) can be decomposed in the following manner: $V_g = K^{-1}V_g^0 + V_g^1$, where V_g^0 and V_g^1 are normalized variances due to the cluster-level and household-level variations in the sample and have the following definitions:*

$$\begin{aligned} V_g^0 &\equiv KE_{x_c^0}[\text{var}_{x_{ch}^1}[K^{-1}\sum_h g(x_{ch},\theta)|x_c^0]] = E_{x_c^0}[\text{var}_{x_{ch}^1}[g(x_{ch},\theta)|x_c^0]] \\ V_g^1 &\equiv \text{var}_{x_c^0}[E_{x_{ch}^1}[K^{-1}\sum_h g(x_{ch},\theta)|x_c^0]] = \text{var}_{x_c^0}[E_{x_{ch}^1}[g(x_{ch},\theta)|x_c^0]] \end{aligned}$$

In particular, when Assumption 3 holds, V_g^0 and V_g^1 can be written as follows:

$$V_g^0 = E_{x_c^0}\left[\text{var}_{x_{ch}^1}\left[\Phi\left(\frac{z-(x_c^{0T}+x_{ch}^{1T})\beta}{\sigma_u}\right)\right]\right] \simeq E_{x_c^0}\left[\phi^2\left(\frac{z-x_c^{0T}\beta}{\sigma_u}\right)\cdot\frac{\beta^T\Sigma_{xx}^1\beta}{\sigma_u^2}\right] \quad (7)$$

$$V_g^1 = \text{var}_{x_c^0}\left[E_{x_{ch}^1}\left[\Phi\left(\frac{z-(x_c^{0T}+x_{ch}^{1T})\beta}{\sigma_u}\right)\right]\right] \simeq \text{var}_{x_c^0}\left[\Phi\left(\frac{z-x_c^{0T}\beta}{\sigma_u}\right)\right], \quad (8)$$

where the approximation is taken around $x_{ch}^1 = \mathbf{0}_L$ and used to obtain the sample analogues of V_g^0 and V_g^1.

Hereafter, we also make the following relationship to hold:

$$\alpha \equiv \frac{\sigma_\eta^2}{\sigma_e^2} \ll 1. \quad (9)$$

This assumption is valid empirically for the datasets we used to calibrate the parameters for the model error (See Table 4 in the Appendix). This assumption is also found to be valid widely in the small-area estimation literature.

When eq. (9) holds, $E^{-1}[X_c^T \Omega_c^{-1} X_c]$ can be approximated as follows:

$$\begin{aligned} E^{-1}[X_c^T \Omega_c^{-1} X_c] &= \frac{\sigma_e^2}{k}\left[(\Sigma_{xx}^0 + \Sigma_{xx}^1) - \frac{\alpha}{1+k\alpha}(k\Sigma_{xx}^0 + \Sigma_{xx}^1)\right]^{-1} \\ &\simeq \frac{\sigma_e^2}{k}\left[(\Sigma_{xx}^0 + \Sigma_{xx}^1)^{-1} + \alpha(\Sigma_{xx}^0 + \Sigma_{xx}^1)^{-1}(k\Sigma_{xx}^0 + \Sigma_{xx}^1)(\Sigma_{xx}^0 + \Sigma_{xx}^1)^{-1}\right] \\ &= \frac{\sigma_e^2}{k}(\Sigma_{xx}^0 + \Sigma_{xx}^1)^{-1}\left[I_L + \alpha(k\Sigma_{xx}^0 + \Sigma_{xx}^1)(\Sigma_{xx}^0 + \Sigma_{xx}^1)^{-1}\right], \end{aligned} \tag{10}$$

where we have taken a first-order approximation with respect to α in the second line using the formula of the differentiation of an matrix inverse (e.g., p.151 of Magnus and Neudecker (2007)). Plugging eq. (10) in eq. (6) and using Lemma 13, the variance of $\hat{\mu}^{DS}$ can be approximated as follows:

$$\begin{aligned} \text{var}[\hat{\mu}^{DS}] &\simeq \frac{1}{J}\left[\frac{V_g^0}{K} + V_g^1 + \frac{\sigma_e^2 \Sigma_{\phi x}^T (\Sigma_{xx}^0 + \Sigma_{xx}^1)^{-1}\left[I_L + \alpha(k\Sigma_{xx}^0 + \Sigma_{xx}^1)(\Sigma_{xx}^0 + \Sigma_{xx}^1)^{-1}\right]\Sigma_{\phi x}}{k\sigma_u^2}\right. \\ &\quad \left. + \frac{\Sigma_{\phi B}^2(k\sigma_\eta^4 + 2\sigma_e^2\sigma_\eta^2 + \sigma_e^4)}{2k\sigma_u^4}\right] \\ &= \frac{V_i}{n} + \frac{V_h}{nr} + \frac{V_c}{J} (\equiv V), \end{aligned} \tag{11}$$

where V_i, V_h, and V_c are the variance components due to household-specific sampling errors, model errors, and cluster-specific errors, respectively, and have the following definitions:

$$\begin{aligned} V_i &\equiv V_g^0 \\ V_h &\equiv \frac{\Sigma_{\phi x}^T [\Sigma_{xx}^0 + \Sigma_{xx}^1]^{-1}[I_L + \alpha\Sigma_{xx}^1[\Sigma_{xx}^0 + \Sigma_{xx}^1]^{-1}]\Sigma_{\phi x}}{1+\alpha} + \frac{\Sigma_{\phi B}^2(2\alpha+1)}{2(1+\alpha)^2} \\ V_c &\equiv V_g^1 + \frac{\alpha\Sigma_{\phi x}^T[\Sigma_{xx}^0 + \Sigma_{xx}^1]^{-1}\Sigma_{xx}^0[\Sigma_{xx}^0 + \Sigma_{xx}^1]^{-1}\Sigma_{\phi x}}{1+\alpha} + \frac{\Sigma_{\phi B}^2\alpha^2}{2(1+\alpha)^2} \end{aligned}$$

The formula for V_c shows that the variance component due to cluster-specific errors has two important parts: the first term (V_g^1) represents the sampling errors at the cluster level (i.e., errors due to the variations of x_c^0) whereas the second and third terms represent the idiosyncratic errors at the cluster level (i.e., errors due to the variations of η_c).

One important observation to make here is that the total variance V consists of three components, each inversely proportionate to the full sample size (i.e., n), the size of sub-sample with the outcome variable (i.e., nr), and the number of clusters (i.e., J). Note also that eq. (11)

includes the single-sampling prediction estimator as a special case with $r = 1$.

To compare the variances of various estimators, it is useful to show that the survey direct estimator can be written in a form very similar to eq. (11) as the following theorem shows:

Theorem 14 *Under Assumptions 1, 2, and 12, the following holds:*

$$E[\bar{Y}] = \mu \quad \text{and} \quad \text{var}[\bar{Y}] = \frac{\tilde{V}_i}{n} + \frac{\tilde{V}_c}{J}, \tag{12}$$

where $\tilde{V}_i \equiv E_{X_c}\left[E_{\eta_c}[\text{var}_{e_{ch}}[Y_{ch}|\eta_c, X_c]]\right] + V_g^0$ *and* $\tilde{V}_c \equiv E_{X_c}\left[\text{var}_{\eta_c}[E_{e_{ch}}[Y_{ch}|\eta_c, X_c]]\right] + V_g^1$. *Further, under Assumption 3,* $\tilde{V}_i$ *and* $\tilde{V}_c$ *can be written as follows:*

$$\begin{aligned}
\tilde{V}_i &= E_{X_c}\left[E_{\eta_c}\left[\Phi\left(\frac{z_{ch} - x_{ch}^T\beta - \eta_c}{\sigma_e}\right)\left[1 - \Phi\left(\frac{z_{ch} - x_{ch}^T\beta - \eta_c}{\sigma_e}\right)\right]\right]\right] + V_g^0 \\
&\simeq E_{X_c}\left[\Phi\left(\frac{z_{ch} - x_{ch}^T\beta}{\sigma_e}\right)\left[1 - \Phi\left(\frac{z_{ch} - x_{ch}^T\beta}{\sigma_e}\right)\right]\right] + V_g^0 \qquad (13) \\
\tilde{V}_c &= E_{X_c}\left[\text{var}_{\eta_c}\left[\Phi\left(\frac{z_{ch} - x_{ch}^T\beta - \eta_c}{\sigma_e}\right)\right]\right] + V_g^1 \\
&\simeq E_{X_c}\left[\phi^2\left(\frac{z_{ch} - x_{ch}^T\beta}{\sigma_e}\right)\right]\alpha + V_g^1 \qquad (14)
\end{aligned}$$

We hereafter use tilde () to emphasize that it is derived for the sample direct estimator.

This theorem is useful because the optimal single sampling discussed in the next section is directly applicable to the sample direct estimator using eq. (12), even though our primary focus is on the prediction estimator. Furthermore, this result helps us to compare the sample direct estimator with the prediction estimator as we elaborate in Section 4.

For completeness, let us also present the non-nested double sampling estimator (see e.g. Kim and Rao (2012)):

$$\hat{\mu}^{NN,DS} \equiv n^{-1}\sum_{ch\in S} g(x_{ch}, \check{\theta}), \tag{15}$$

where $\check{\theta}$ denotes an estimator for θ that is derived from a secondary non-nested sample. This secondary sample is considered given and its data collection cost is already sunk. The secondary sample often refers to a previous survey of the same type or a contemporaneous survey of a different type in practice.[14] For example, in a poverty measurement application to Morocco,

[14]Note that the imputed household expenditures or welfare indicators could also be used in a regression analysis, in addition to evaluating mean values of the predicted values, see Elbers et al. (2005).

Douidich et al. (2015) estimate the relationship between consumption poverty and household characteristics using a household consumption survey and then use this model to predict consumption poverty into a series of annual labor force surveys. This approach enables users to leverage existing data sources. If one decides to collect new data for the estimation of poverty, one has the option of only collecting the covariates x, since an estimator for θ can be obtained from the secondary data source. Put differently, the covariate-only sample is typically part of data collection planning when $\hat{\mu}^{DS}$ is used but this is not necessarily the case when $\hat{\mu}^{NN,DS}$ is used.

We will not formally derive the precision of $\hat{\mu}^{NN,DS}$ here as this would require us to make assumptions about the dynamics of the model parameters. When a previous survey is used as in Douidich et al. (2015) for example, one would have to make an assumption about how the model has evolved over time (or assume that it is time-invariant), which is not required for $\hat{\mu}^{DS}$. In Section 4.1 we will however provide a brief discussion on the financial costs savings that may be expected when using the non-nested double sampling estimator (under some simplifying assumptions).

3 Cost efficient sampling

To see whether a meaningful reduction in costs is achievable, we examine the trade-offs between financial costs and statistical precision analytically. A stylized yet informative financial cost function is used. For a measure of statistical precision we appeal to the analytic expression for the asymptotic variance. The advantage of studying this trade-off analytically is that it allows us to work out the conditions under which double-sampling may be expected to be most beneficial. And, similarly, under what conditions the benefits will be marginal.

Specifically, we consider the problem of minimizing financial costs given a statistical precision constraint and its dual problem of maximizing statistical precision under a given budget constraint. Formally, we make the following assumption:

Assumption 15 *The cost of collecting only x for any additional household in a given cluster equals $\tau \in (0, 1)$. The travel cost to visit an additional cluster is equal to c.*

Here, we normalize the cost of collecting (x_{ch}, y_{ch}) to be equal to one. As a result, it is reasonable to require $\tau \in (0, 1)$, because it costs something to observe the covariates but not as much as it would if both the covariates and the outcome variable are to be observed. Under Assumption 15, the total variable cost of data collection is given by:

$$C = nr + n(1 - r)\tau + cJ. \tag{16}$$

We ignore the fixed cost of data collection as it does not affect the optimal sampling design. If the fixed cost differs between the single and double sampling, the difference has to be taken into account in the choice of optimal design.

We now consider the optimal sampling under single and double sampling. In the former case, we fix $r = 1$ and choose n and J to minimize the financial cost for a given variance $\bar{V}$ or minimize the variance for a given total cost $\bar{C}$ (i.e. budget for data collection). In the case of the latter, we also allow r to vary.

3.1 Optimal single sampling

To assess how much one stands to gain by adopting double sampling over single sampling, we derive the level of statistical precision and the cost for the optimal single sample case. To provide a competitive benchmark we will consider the optimal single-sampling prediction estimator, which may be a sample direct estimator or a prediction estimator.

Suppose that one wants to minimize the cost of data collection subject to a required accuracy. This formulation is relevant, for example, when the researchers or policy-makers know how accurate the estimate $\hat{\mu}$ should be. Therefore, when the variance has the form of eq. (11) and the cost function is given in eq. (16), the cost minimization problem can be formulated as follows:

$$C_1^* \equiv \min_{n,J} n + cJ \quad \text{s.t.} \quad \frac{V_i + V_h}{n} + \frac{V_c}{J} = \bar{V} \tag{17}$$

Ignoring the integer constraints for n and J for simplicity of presentation, we can obtain the

following minimizing arguments (n_1^*, J_1^*) and minimized cost C_1^*:

$$n_1^* = \frac{H_1}{\bar{V}}\sqrt{V_i + V_h}, \quad J_1^* = \frac{H_1}{\bar{V}}\sqrt{\frac{V_c}{c}}, \quad C_1^* = \frac{H_1^2}{\bar{V}},$$

where $H_1 \equiv \sqrt{V_i + V_h} + \sqrt{V_c c}$.

When the budget for data collection is exogenously given, the following dual problem would be more relevant.

$$V_1^+ = \min_{n,J} \frac{V_i + V_h}{n} + \frac{V_c}{J} \quad \text{s.t.} \quad n + cJ = \bar{C}$$

Solving this yields:

$$n_1^+ = \frac{\bar{C}}{H_1}\sqrt{V_i + V_h}, \quad J_1^+ = \frac{\bar{C}}{H_1}\sqrt{\frac{V_c}{c}}, \quad V_1^+ = \frac{H_1^2}{\bar{C}}.$$

The solution above shows that the optimal total sample size n increases with V_i (as a larger n will be needed to curb the sampling error). Similarly, the optimal number of clusters J increases with V_c (as a larger J in this case is needed to curb the cluster level error component) and decreases with c (i.e. the price tag associated with adding to the number of clusters).

3.2 Optimal double sampling

We now turn to the optimization problem for double sampling, in which $r(\leq 1)$ is also a choice variable. In this case, the cost minimization corresponding to eq. (17) is as follows:

$$C_2^* = \min_{n,r,J} nr + n(1-r)\tau + cJ \quad \text{s.t.} \quad \frac{V_i}{n} + \frac{V_h}{nr} + \frac{V_c}{J} = \bar{V} \tag{18}$$

When we have an interior solution, solving the first order conditions yields:

$$r^* = \sqrt{\frac{\tau w_h}{1-\tau}}, \quad n_2^* = \frac{H_2}{\bar{V}}\sqrt{\frac{V_i}{\tau}}, \quad J_2^* = \frac{H_2}{\bar{V}}\sqrt{\frac{V_c}{c}}, \text{and} \quad C_2^* = \frac{H_2^2}{\bar{V}}. \tag{19}$$

where $H_2 \equiv \sqrt{\tau V_i} + \sqrt{(1-\tau)V_h} + \sqrt{V_c c}$ and $w_h \equiv V_h/V_i$. As with the case of single sampling, we can also consider the dual problem of eq. (18). In this case, the total variance is minimized

under a fixed budget $\bar{C}$ for data collection:

$$V_2^+ = \min_{n,r,J} \frac{V_i}{n} + \frac{V_h}{nr} + \frac{V_c}{J} \quad \text{s.t.} \quad nr + n(1-r)\tau + cJ = \bar{C} \tag{20}$$

Solving this, we obtain:

$$r^+ = \sqrt{\frac{\tau w_h}{1-\tau}}, \quad n_2^+ = \frac{\bar{C}}{H_2}\sqrt{\frac{V_i}{\tau}}, \quad \text{and} \quad J_2^+ = \frac{\bar{C}}{H_2}\sqrt{\frac{V_c}{c}}, \quad \text{and} \quad V_2^+ = \frac{H_2^2}{\bar{C}}.$$

The interpretation of the solutions for n and J is intuitive and similar to the case of single sampling. The optimal solution for r (i.e., the share of observations for which data on both y and x will be collected) is found to be an increasing function of w_h and of the cost parameter τ. The positive relationship with w_h conveys the fact that it takes data on y to reduce the model error component: If the model error is important relative to the sampling error, then it is optimal to collect more data on y (i.e. increase rn). If the sampling error is relatively more important, then it is optimal to expand the total sample size (i.e. n) at the expense of limiting the number of households for which data on y is collected. The positive relationship between r^* (or r^+) and τ conveys the fact that collecting data on y is relatively lighter on the budget when τ is larger.

Note that the solution above does not necessarily satisfy $r \leq 1$. For this to hold, the following condition for an interior solution needs to be satisfied:

$$\frac{1-\tau}{\tau} \geq w_h. \tag{21}$$

4 Evaluating the potential gains from double sampling

A general comparison between sample direct estimators and prediction estimators is complicated by the fact that the latter relies on a prediction model which can be estimated using different methods. Prediction estimators may or may not outperform and the sample direct estimator (see Matloff (1981) and Fujii and van der Weide (2013)). In Section 4.1, we first consider the comparison between the optimal single- and double-sample estimators. This choice has at least two advantages. First, because both use prediction estimators, all the error components (i.e., V_i, V_h, and V_c) are the same. Their differences come only from the differences in sampling (i.e.,

the choices of n, r, and J). This in turn means that the difference can be taken as the pure effect of choosing double sampling. If we compare the optimal double-sampling estimator with the sample direct estimator, then part of the difference must be attributed to the fact that the optimal double-sampling estimator is a prediction estimator (while the direct estimator is not). Second, because the variance components are the same between optimal single- and double-sampling estimators, the comparison provides a clear prediction about the circumstances under which double sampling is most useful.

In Section 4.2, we compute the potential gains from the optimal double sampling estimator not only in comparison with the optimal single sampling estimator but also with the sample direct estimator using empirically-relevant parameter values. This analysis provides plausible ballpark estimates of the gains from optimal double sampling for the estimation of poverty rates.

4.1 Optimal single sampling vs optimal double sampling

One intuitive measure of comparative performance between the single- and double-sampling estimator is the ratio of their respective variances given the same budget. Another candidate measure is the ratio of financial cost between the optimal double- and single-sampling estimators given the same statistical precision. Under our assumptions, it conveniently follows that these two measures coincide and can be expressed as follows:

$$\rho(c, w_c, \tau, w_h) = \frac{C_2^*}{C_1^*} = \frac{V_2^+}{V_1^+} = \frac{H_2^2}{H_1^2} = \left[\frac{\sqrt{\tau} + \sqrt{(1-\tau)w_h} + \sqrt{w_c c}}{\sqrt{1 + w_h} + \sqrt{w_c c}} \right]^2, \tag{22}$$

where $w_c \equiv V_c/V_i$. Note that the relative performance measure satisfies $\rho \in (0, 1]$ as long as eq. (21) is satisfied, where lower values of ρ indicate larger gains from double sampling. It can be verified that $\rho = 1$ holds if and only if eq. (21) is satisfied with equality. This is the threshold case where optimal double sampling reduces to optimal single sampling. When eq. (21) is violated, the double-sampling optimization problems in eqs. (18) and (20) have a corner solution, which is the solution for the comparable single-sampling optimization problems. In this case, double-sampling offers no advantage over single sampling, a situation that occurs when w_h and τ are sufficiently high.

The following lemma, which follows directly from eq. (22), shows under what variance

and cost parameters one stands to gain the most from adopting the optimal double sampling estimator:

Lemma 16 *Suppose that eq. (21) holds. Then, ρ satisfies the following conditions:*

$$\frac{\partial \rho}{\partial c} \geq 0, \quad \frac{\partial \rho}{\partial w_c} \geq 0, \quad \frac{\partial \rho}{\partial \tau} \geq 0, \quad \text{and} \quad \frac{\partial \rho}{\partial w_h} \geq 0. \tag{23}$$

These results are intuitive. First, consider the impact of c on ρ. When the travel costs make up a larger share of total costs, ρ goes up. This makes the optimal double sampling estimator less attractive relative to the optimal single sampling estimator. Second, the impact of w_c on ρ is also positive. When cluster-specific variations make up a larger share of the total variance of the prediction estimator, the gains from double sampling will be smaller as double sampling helps to reduce neither the travel cost nor cluster-specific variations. These two points can also be readily seen from the facts that eq. (22) is an increasing function of $w_c c$ and that ρ tends to 1 in the limit where $w_c c$ tends to infinity. In this case, the number of clusters to be visited is the only thing that matters asymptotically and thus there are no gains from double sampling.

Third, ρ tends to go up when w_h is higher. This essentially means that double sampling is beneficial when the household-specific sampling error is important relative to the model error. This result is also intuitive as the use of double sampling does not reduce the model error but it helps to reduce the household-specific sampling error. Finally, ρ also tends to go up when τ is higher. Hence, the double sampling strategy is most useful when we have covariates that can be collected cheaply. It can be verified that τ functions as a lower bound for ρ.

Lemma 17 *Suppose that eq. (21) holds. Then, $\rho \geq \tau$.*

To further understand the difference between the optimal single- and double-sampling schemes, it is useful to consider the following ratio of cluster sizes for the main and dual problems:

$$r_J^* \equiv \frac{J_2^*}{J_1^*} = \frac{H_2}{H_1} = \sqrt{\rho}, \quad \text{and} \quad r_J^+ \equiv \frac{J_2^+}{J_1^+} = \frac{H_1}{H_2} = 1/\sqrt{\rho}.$$

It is clear that $r_J^* < 1$ whereas $r_J^+ > 1$ when eq. (21) holds with a strict inequality. Therefore, in the main [dual] problem where the cost [variance] is to be minimized, the number of clusters in the optimal double sampling is smaller [larger] from its counterpart in the optimal single

sampling to save the travel cost [cancel out the cluster-specific errors]. Further, because r_J^* [r_J^+] is an increasing [a decreasing] monotonic transformation of ρ, the signs of its partial derivatives with respect to c, w_c, τ, and w_h are the same as [the opposite of] those in Lemma 16.

It is also interesting to note that the number of households to be sampled in each cluster is the same between the main and dual problems as the following equation shows:

$$K_1 \equiv \frac{n_1^*}{J_1^*} = \frac{n_1^+}{J_1^+} = \sqrt{\frac{(1+w_h)c}{w_c}}, \quad K_2 \equiv \frac{n_2^*}{J_2^*} = \frac{n_2^+}{J_2^+} = \sqrt{\frac{c}{\tau w_c}}. \tag{24}$$

Both equations show that the cluster size tends to get smaller when the errors due to the cluster-level variations become more pronounced. On the other hand, the cluster size tends to get larger (and the number of clusters to be visited get smaller) when the travel cost is larger. The ratio $K_2/K_1 = 1/\sqrt{(1+w_h)\tau}$, which is no less than one when eq. (21) is satisfied, represents the change in the cluster size when one switches from optimal single sampling to optimal double sampling. When the parameters are unfavorable to double sampling (i.e., when w_h and τ are high), the ratio of cluster sizes tends to be smaller.

4.2 Realistic estimates of gains from double sampling

Let us now evaluate the potential benefits of double sampling using a realistic set of parameter values (c, τ, w_c, w_h) taken from existing surveys. While the parameter values are clearly context-dependent and cannot be readily extrapolated to other contexts, this exercise is still useful as it gives practitioners a sense of how much they could reasonably expect to gain from double sampling. It also facilitates comparisons between the sample direct estimator and the optimal single sampling estimator.

Ideally, all the parameter values should come from a single survey. However, we are unable to do so due to the lack of data. In particular, the information on survey costs is typically unavailable to the public. Therefore, we collect empirical values of c, τ, ,w_c, and w_h from various sources. For each of these parameters, we specify low, mid, and high values. The low and high values are close to the minimum and maximum observed in our data sources. The mid value is either the arithmetic or geometric mean of minimum and maximum. The parameter values are presented in Table 1. The details of the data sources and assumptions are provided in

Table 1: Set of parameter values used in this study.

Value	Low	Mid	High
c	$c^L = 4$	$c^M = 16$	$c^H = 64$
τ	$\tau^L = 0.06$	$\tau^M = 0.36$	$\tau^H = 0.66$
w_c	$w_c^L = 0.6$	$w_c^M = 1.2$	$w_c^H = 1.8$
w_h	$w_h^L = 0.4$	$w_h^M = 1.2$	$w_h^H = 3.6$

Appendix B.

While the parameter values reported in Table 1 are based on real data and existing studies, they do not necessarily represent the full range of values one might encounter in practice. Furthermore, because the parameter values are taken from different sources, we do not know the correlational structure of these parameters. To make the most of the available data, we choose to calculate ρ for all the $81 (= 3^4)$ combinations.

Table 2 provides the values of ρ under different combinations of parameter values. A few points are worth mentioning here. First, Table 2 shows that eq. (21) is not satisfied for some combinations of parameter values. This occurs when τ and w_h are relatively high. In this case, it does not save much to omit y from observations and the model error is relatively high. Thus, optimal double sampling has no advantage over optimal single sampling. In fact, from a practical perspective, both τ and w_h have to be low for optimal double sampling to have a meaningful advantage over optimal single sampling.

Second, in comparison with τ and w_h, the values of c and w_c have limited impact on ρ for the range of values we consider. Third, the gains from optimal double sampling relative to optimal single sampling appear to be reasonably modest, even when eq. (21) is satisfied. The lowest number reported in Table 2 is 0.776. This means that the cost saving from optimal double sampling is at best 22.4 percent of the cost for optimal single sampling within the set of parameters we considered.

Figure 1 gives an idea of what values ρ might attain by plotting ρ as a function of τ for four different choices of $w_c c$ and w_h. The top (bottom) row corresponds to a relatively low (high) value for w_h, while the left (right) column refers to relatively low (high) values for $w_c c$ (see Table 1). The diagonal line denotes the lower bound for ρ from Lemma 17. Note that ρ is only evaluated for $\tau < \tau_{max} = 1/(1 + w_h)$ (see eq. (21)). Judging by these figures, ρ mostly

Table 2: Values of ρ for different combinations of parameters (τ, c, w_c, w_h).

	(w_c^L, w_h^L)	(w_c^L, w_h^M)	(w_c^L, w_h^H)	(w_c^M, w_h^L)	(w_c^M, w_h^M)	(w_c^M, w_h^H)	(w_c^H, w_h^L)	(w_c^H, w_h^M)	(w_c^H, w_h^H)
(τ^L, c^L)	0.776	0.887	0.968	0.817	0.906	0.972	0.839	0.917	0.975
(τ^L, c^M)	0.854	0.925	0.977	0.887	0.941	0.982	0.903	0.949	0.984
(τ^L, c^H)	0.914	0.955	0.986	0.936	0.966	0.989	0.946	0.971	0.991
(τ^M, c^L)	0.944	0.995	—	0.955	0.996	—	0.960	0.997	—
(τ^M, c^M)	0.964	0.997	—	0.972	0.998	—	0.977	0.998	—
(τ^M, c^H)	0.979	0.998	—	0.985	0.999	—	0.987	0.999	—
(τ^H, c^L)	0.999	—	—	0.999	—	—	0.999	—	—
(τ^H, c^M)	0.999	—	—	0.999	—	—	0.999	—	—
(τ^H, c^H)	0.999	—	—	1.000	—	—	1.000	—	—

— indicates that eq. (21) is not satisfied.

ranges between 0.8 and 1, and it takes incredibly low values of τ to realize reductions in costs (or variance) of 10 percent or more (i.e. values of ρ below 0.9). Assuming that the parameter values for $w_c c$ and w_h are indeed reasonable, the gains from double sampling are expected to be modest unless conditions are particularly favorable.

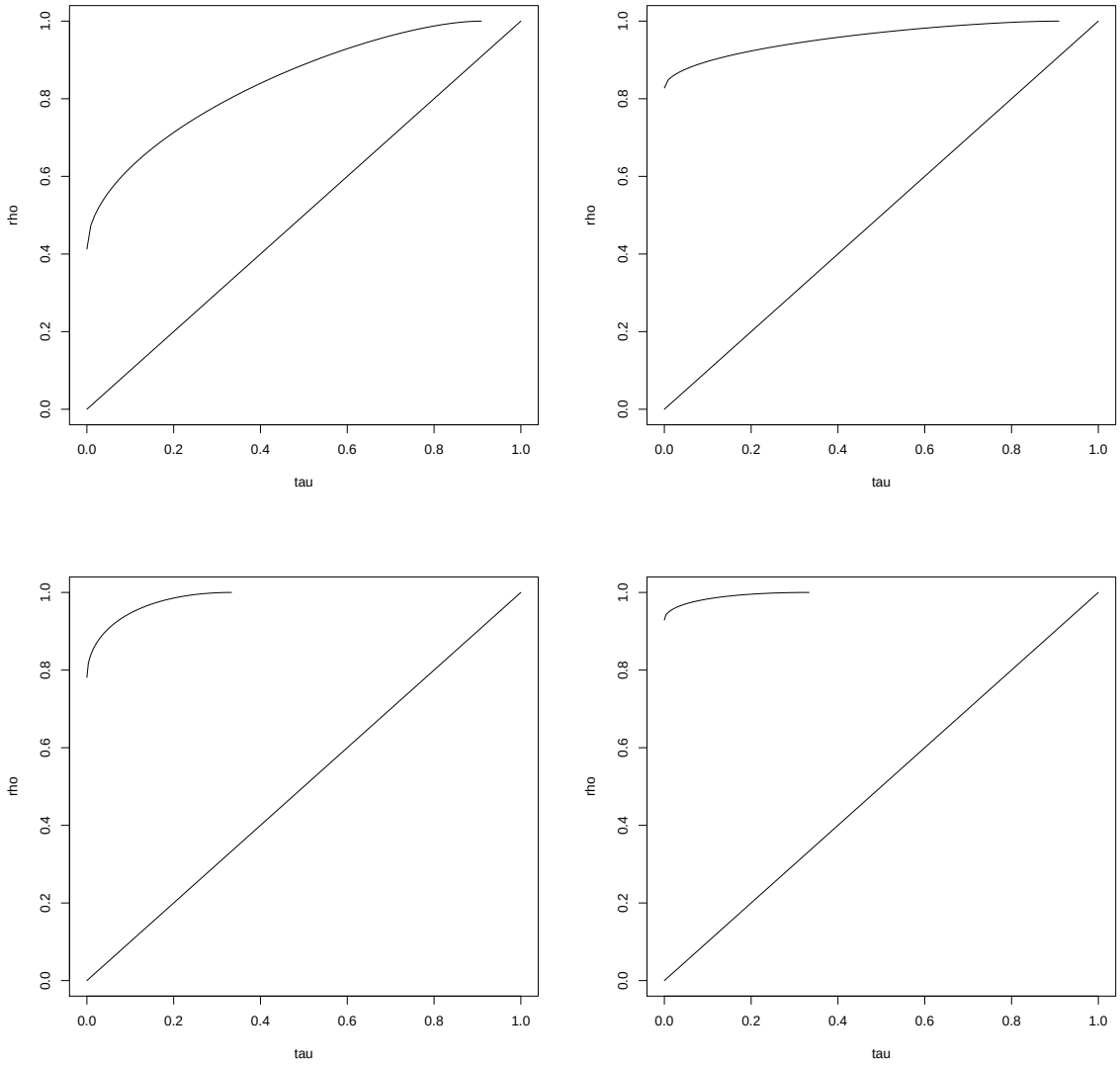

Figure 1: Plots of ρ versus τ for different choices of $w_c c$ and w_h: (a) $w_c c = 1$ and $w_h = 0.1$ (top-left), (b) $w_c c = 50$ and $w_h = 0.1$ (top-right), (c) $w_c c = 1$ and $w_h = 2$ (bottom-left), and (d) $w_c c = 50$ and $w_h = 2$ (bottom-right)

Recall that ρ is the ratio of data collection costs between optimal single and double sampling strategies. One may also be interested in the gains relative to the sample direct estimator. Let ω denote the ratio of data collection costs between the sample direct estimator and the double sample estimator given the required accuracy (variance) constraint.

To this end, we first derive the ratio ψ of data collection costs between the sample direct estimator and the single-sampling prediction estimator (given the required accuracy constraint), which solves:

$$\psi \equiv \frac{C_1^*}{\tilde{C}_1^*} = \frac{V_1^+}{\tilde{V}_1^*} = \frac{\sqrt{V_i + V_h} + \sqrt{V_c c}}{\sqrt{\tilde{V}_i} + \sqrt{\tilde{V}_c c}}. \tag{25}$$

This quantity can be computed when all the relevant inputs (i.e., V_i, V_h, V_c, $\tilde{V}_i$, $\tilde{V}_c$, and c) can be obtained from a single data source. This is indeed the case for Malawi ($\psi = 0.945$) and Niger ($\psi = 0.977$). The gain of using the optimal single-sample estimator instead of the sample-direct estimator is thus found to be very small, at least for these two countries.

If we assume that the choice of K in a given survey is optimized for the sample direct estimator, we are able to derive ψ without data on c. To see this point, notice first that the derivation of eq. (24) does not depend on the nature of the variance components V_i, V_h, and V_c. Therefore, we can apply the variance components of the sample direct estimator defined in Theorem 14 to eq. (24), where we have $V_h = 0$ because there is no model error for the sample direct estimator. Solving for c, we obtain:

$$\tilde{c} = \frac{\tilde{V}_c K^2}{\tilde{V}_i}, \tag{26}$$

where K is the cluster size in the sample. The resulting $\tilde{c}$ can be interpreted as the implied travel cost that justifies the design of the sample, because the observed choice of K is consistent with the optimal design for the sample direct estimator. We can then substitute $\tilde{c}$ in place of c in eq. (25). We denote the left hand side of eq. (25) derived in this way by $\tilde{\psi}$ to distinguish it from ψ.

The values of $\tilde{c}$ tend to be higher than the values of c observed in Malawi and Niger (see Tables 3 and 4). However, the resulting values of $\tilde{\psi}$ are not very different from ψ ($\tilde{\psi} = 0.948$ in Malawi and $\tilde{\psi} = 0.983$ in Niger). Therefore, the alternative derivation of ψ appears to provide a reasonable indication of the gains associated with the optimal single-sample estimator relative to the sample direct estimator.

We also derived $\tilde{\psi}$ for Tanzania ($\tilde{\psi} = 0.961$) as well as the following three strata of Cambodia: Phnom Penh ($\tilde{\psi} = 0.945$), Other Urban ($\tilde{\psi} = 0.977$), and Rural ($\tilde{\psi} = 0.965$). All these

estimates are close to unity. Hence, even if we take the lowest value of ρ and ψ among our estimates, the resulting value of the ratio $\omega \equiv \psi\rho$ of minimized variances [costs] under the optimal double sampling and optimal sample direct estimators for an exogenously given cost [variance] is $0.733 (\simeq 0.776 \times 0.945)$. This suggests that the gain from optimal double sampling relative to the sample direct estimator is only about 27 percent even in the most optimistic case.

Let us also briefly comment on the financial cost savings that may be expected when the poverty rate is estimated using the non-nested double sampling estimator from eq. (15). We expect the reduction in costs to be more substantial in this case. For ease of exposition, suppose that the primary sample (the newly collected data) and the secondary sample (a previous survey, say) would offer similar estimates of the model parameters θ, in terms of precision, if the primary sample indeed includes the observations of y for all households. This ensures that the non-nested double sampling estimator (which predicts poverty into the primary sample using a model that is estimated from the secondary sample) will match the precision of the single sampling estimator (which uses the primary sample only), allowing for a fair comparison of costs. The non-nested double sampling estimator is able to achieve this level of precision without using any of the data on y from the primary sample. By not collecting any y but only the covariates x the data for the primary sample come at a cost of τ rather than unit cost. Note that this coincides with the lower bound value for ρ as derived in Lemma 17. The difference between ρ and τ tends to be substantial, as can be seen in Figure 1, which suggests that the gains from non-nested double sampling are indeed expected to be substantially larger compared to those obtained from nested double sampling. It should be noted however that it is by no means guaranteed that a secondary sample (which may denote a noticeably older survey) can compete with an up-to-date sample as far as the estimation of θ is concerned. A more detailed study of the cost savings that can be realized with non-nested double sampling, under different assumptions of model stability, is recommended but is beyond the scope of this paper.

4.3 When is double sampling promising?

A relatively high value of ω under the most optimistic case indicates that gains from double sampling are rather limited. However, this finding should not be overextrapolated because the

parameter values depend on the application. Different survey designs and data collection techniques may lead to different values of ω. Let us therefore also consider the possibility where ω may be substantially lower than 0.733.

In some contexts, the cost of observing outcomes may be very expensive. For example, if the data collection involves physically invasive techniques (e.g., blood testing), observing y may be costly and thus the value of τ may be substantially lower than 0.06. If we use $\tau^L = 0.001$ instead of $\tau^L = 0.06$, the lowest estimate for ρ in Table 2 would be $\rho = 0.656$ instead of $\rho = 0.776$. The use of new data collection technology may also allow for significantly lower data collection costs. The marginal cost of collecting data online for example is typically much lower than that through traditional face-to-face interviews.

It is conceivable that the values we use for w_h and w_c too may be lower in other contexts, particularly when an obvious and strong proxy for the outcome of interest is available. Consider an application where consumption data derived from a short-form questionnaire serves as a proxy for the complete consumption aggregate that is based on a long-form questionnaire. The predictive power of the model is likely to be strong in that case such that w_c and w_h may be much lower. Just for the purpose of illustration, if we use $(w_c^L, w_h^L) = (0.15, 0.1)$ instead of $(0.6, 0.4)$ while keeping $(\tau_c^L, c^L) = (0.06, 4)$, then ρ can go as low as 0.529. Therefore, one can think of circumstances where the gains from double sampling can be substantial, but these may be the exceptions rather than the rule.

To further elucidate this point, consider a study by Ahmed et al. (2014), which implies considerable gains from using a prediction estimator in an application to poverty measurement in Bangladesh. Unfortunately, their approach does not fit perfectly into our analytical framework such that we are unable to compute the values for (w_c, w_h) that would apply to their data. Specifically, their prediction estimator works with different data (compared to the sample direct estimator), the number of clusters is not necessarily chosen optimally, and neither the level of statistical precision nor the financial costs are fixed. This hampers a fair comparison between the two estimators.

However, the standard errors for the national poverty rate derived from a small sub-sample of 640 households are 2.9 and 2.4 percentage points (see Table 5 of Ahmed et al. (2014)), respectively, for the sample direct and prediction estimators, which corresponds to a 32 percent

($\approx 1 - (2.4/2.9)^2$) reduction in variance. Without further information, however, it is hard to determine how much of this cost-precision trade-off can be attributed to the fact that their double sampling estimator substituted predicted data for real data.[15] It is unlikely that the gains are due to a favorably low level of τ as τ is estimated to be around 0.6 in their case, which is in the mid range. If the gains can indeed be attributed to the use of predicted data, then it is more likely that this is due to favorable values of w_h and w_c.

5 Discussion

The primary motivation for the use of prediction in economics, health sciences, and other disciplines has been to deal with various forms of missing data problems. One could also make a case for adopting prediction estimators to obtain more cost-efficient estimates of the population mean when it is expensive to observe the outcome of interest in comparison with its covariates. For example, consider the estimation of poverty and malnutrition rates. The conventional sample direct estimators in this case require household- and individual-level data on expenditures and health outcomes. Collecting this data is generally costly. It is not uncommon that in developing countries, where poverty and poor health outcomes are most pressing, statistical agencies do not have the budget that is needed to collect these data frequently. As a result, official estimates of poverty and malnutrition are often outdated. This then makes it difficult to monitor progress (or lack thereof) in times when circumstances might be subject to considerable change, such as shocks to international staple prices, domestic climate shocks, etc. Using predicted data as a substitute for real data may then offer a valuable alternative.

In recent years, a number of studies have explored the option of predicting household expenditure data into existing secondary surveys in an effort to supplement existing poverty estimates and increase their frequency (Stifel and Christiaensen, 2007; Douidich et al., 2015). Douidich et al. (2015) for example considers the Labour Force Survey as their secondary survey, which is often available at a higher frequency than household expenditure surveys.

There is also a large literature that predicts household expenditure data into the population census, see for example Elbers et al. (2003) and the references therein. The objective here is to

[15] Also, the study does not report the total cost of data collection, but selected cost components instead.

obtain estimates of poverty at a high level of disaggregation, or at the level of small area such as a district. It would be unpractical to use the sample direct estimator because the data must contain an extraordinarily large number of households to obtain a reliable estimate for each small area, which would be financially infeasible.

It is then a small step to purposefully collect data on covariates that are ideally suited for the prediction of household- or individual-level outcomes of interest. If real data on the variable of interest is collected for a sub-sample of households, then this sub-sample can be used to estimate the model parameters that are used for prediction. The advantage of this double sampling approach is that the prediction model will apply to the population of interest by construction. There is certainly a considerable interest in adopting such an approach in practice in the hope that this will enable a meaningful reduction in financial costs while preserving a reasonable level of statistical precision.

The objective of our study is to investigate the potential gains that might be derived from a double sampling approach under a set of fairly general conditions. We achieve this by analytically deriving the asymptotic variances of the single- and double-sample prediction estimators and by considering approximations to a financial cost function. This allows us to maximize statistical precision [minimize financial costs] under a budget constraint [statistical precision constraint] for a wide set of parameter values. Even though we are working with analytic approximations, we expect that the broad findings coming out of this analysis may carry over to real applications.

When we calibrate the parameters from the variance structure and the financial cost function to real data in the context of the estimation of poverty, we find that the reductions in costs rarely exceed 25 percent and are often below 10 percent. Furthermore, we find that the magnitude of the gains derived from double sampling are primarily determined by the following factors: (a) relative size of the travel costs, (b) degree of spatial correlation between residuals, (c) financial discount obtained by not collecting the outcome variable of interest, and (d) the share of total error that may be attributed to model error (versus sampling error). Double sampling is most effective when a reasonably large geographic coverage can be obtained without having to spend a disproportionate share of the budget on travel, and when the spatial correlation between the residuals is smaller rather than larger. The financial discount obtained by not collecting the

expensive outcome variable of interest is most notable when a larger share of the total error is due to the sampling error (rather than the model error).

It is conceivable that larger gains can be obtained under certain conditions, for example, when the cost of collecting the outcome variable of interest is extraordinarily high, when the cheaply available predictors exhibit an exceptionally high correlation with the outcome variable of interest, and when the data exhibits very little spatial correlation. We conjecture, however, that these circumstances represent the exception rather than the rule. Moreover, we have currently abstracted away from model misspecification error. Ignoring this component of error obviously favors the prediction estimator. Accounting for misspecification error is not obvious; it is hard to quantify since the true model is inherently unknown and any given estimate of the model can be misspecified in infinitely many ways.

Given these observations, when new data is to be collected, we recommend that the outcome variable of interest should be included so that one does not have to rely on predicted data. This does not mean that there is no role for prediction estimators. Under the right circumstances we believe they could be of great value. For one, prediction estimators provide the means of leveraging already existing data (think of non-nested double sampling estimators, see for example Kim and Rao (2012) and Douidich et al. (2015)). Furthermore, if no previous data exists and the budget is particularly constrained such that one may be left with the choice between predicted data or no data, then the former may be preferred over the latter. In such a data-poor environment, which is not unheard of in developing countries, double sampling estimators may continue to provide a valuable option.

References

Ahmed, F., C. Dorji, S. Takamatsu, and N. Yoshida (2014) 'Hybrid survey to improve the reliability of poverty statistics in a cost-effective manner.' World Bank Policy Research Working Paper 6909, The World Bank

Aliaga, A., and R. Ren (2006) 'Optimal sample sizes for two-stage cluster sampling in demographic and health surveys.' DHS Working Papers 2006 No.30, ORC Macro

Armstrong, J., C. Block, and K.P. Srinath (1993) 'Two-phase sampling of tax records for business surveys.' *Journal of Business & Economic Statistics* 11(4), 407–419

Beegle, K., J. De Weerdt, J. Friedman, and J. Gibson (2012) 'Methods of household consumption measurement through surveys: Experimental results from tanzania.' *Journal of Development Economics* 98(1), 3–18

Bose, C. (1943) 'Note on the sampling error in the method of double sampling.' *Sankhyā* 6, 329–330

Christiaensen, L., P. Lanjouw, J. Luoto, and D. Stifel (2012) 'Small area estimation-based prediction methods to track poverty: validation and applications.' *Journal of Economic Inequality* 10(2), 267–297

Cochran, W.G. (1977) *Sampling Techniques,* 3rd edition ed. (John Wiley & Sons)

Davidov, O., and Y. Haitovsky (2000) 'Optimal design for double sampling with continuous outcomes.' *Journal of Statistical Planning and Inference* 86, 253–263

Deaton, A. (2003) 'Adjusted Indian poverty estimates for 1999-2000.' *Economic and Political Weekly* 38(4), 322–326

_ (2005) 'Data and dogma: The great indian poverty debate.' *World Bank Research Observer* 20(2), 177–199

Deaton, A., and J.P. Drèze (2002) 'Poverty and inequality in India: A reexamination.' *Economic and Political Weekly* 37(36), 3729–3748

Diamond, A., M. Gill, M. Dellepiane, E. Skoufias, K. Vinha, and Y. Xu (2015) 'Estimating poverty rates in target populations: An assessment of the simple poverty scorecard and alternative approaches.' *mimeo*

Douidich, M., A. Ezzrari, R. van der Weide, and P. Verme (2015) 'Estimating quarterly poverty rates using labor force surveys: A primer.' *World Bank Economic Review.* Advance Access pulished 2015

Elbers, C., and R. van der Weide (2014) 'Estimation of normal mixtures in a nested error model with an application to small area estimation of poverty and inequality.' World Bank Policy Research Working Paper 6962, The World Bank

Elbers, C., J. Lanjouw, and P. Lanjouw (2003) 'Micro-level estimation of poverty and inequality.' *Econometrica* 71, 355–364

_ (2005) 'Imputed welfare estimates in regression analysis.' *Journal of Economic Geography* 5, 101–118

Fujii, T (2006) 'Community-level estimation of poverty measures and its application in cambodia.' In 'Spatial Disparities in Human Development' (United Nations University Press) pp. 289–314

Fujii, T. (2010) 'Micro-level estimation of child undernutrition indicators in Cambodia.' *World Bank Economic Review* 24(3), 520–553

Fujii, T., and R. van der Weide (2013) 'Cost-effective estimation of the population mean using prediction estimators.' World Bank Policy Research Working Paper 6509, The World Bank

Grosh, M.E., and Muñoz (1996) 'A manual for planning and implementing the living standards measurement study survey.' Living Standards Measurement Study Working Paper 126, The World Bank

Hansen, M.H., and B.J. Tepping (1990) 'Regression estimates in federal welfare quality control programs.' *Journal of the American Statistical Association* 85(411), 856–864

Hidiroglou, M. (2001) 'Double sampling.' *Survey Methodology* 27(2), 143–154

Humphreys, C.P. (1979) 'The cost of sample survey designs.' In 'Proceedings of the Survey Research Methods Section' American Statistical Association pp. 395–400

Kijima, Y., and P. Lanjouw (2005) 'Economic diversification and poverty in rural India.' *Indian Journal of Labour Economics* 48(2), 349–374

Kim, J., A. Navarro, and W. Fuller (2006) 'Replication variance estimation for two-phase stratified sampling.' *Journal of the American Statistical Association* 101(473), 312–320

Kim, J., and J. Rao (2012) 'Combining data from two independent surveys: A model-assisted approach.' *Biometrika* 99, 85–100

Lanjouw, J., and P. Lanjouw (2001) 'How to compare apples and oranges? poverty measurment based on different definitions of consumption.' *Review of Income and Wealth* 47(1), 25–42

Ligon, E., and T. Sohnesen (2016) 'Using reduced consumption aggregates to track and analyze poverty.' *mimeo*

Magnus, J., and H. Neudecker (2007) *Matrix Differential Calculus with Applications in Statistics and Econometrics: Revised Edition* (John Wiley & Sons)

Matloff, N. (1981) 'Use of regression functions for improved estimation of means.' *Biometrika* 68, 685–689

Ministry of Planning, and United Nations World Food Programme (2002) 'Estimation of poverty rates at commune-level in cambodia: Using the small area estimation technique to obtain reliable estimates.' , Ministry of Planning, Royal Government of Cambodia and United Nations World Food Programme, Phnom Penh, Cambodia

Neyman, J. (1938) 'Contribution to the theory of sampling human populations.' *Journal of the American Statistical Association* 33, 101–116

Palmgren, J. (1987) 'Precision of double sampling estimators for comparing two probabilities.' *Biometrika* 74(4), 687–694

Pape, U., and J. Mistiaen (2015) 'Measuring household consumption and poverty in 60 minutes: The Mogadishu high frequency survey.' mimeo, The World Bank

Pettersson, H., and B. Sisouphanthong (2005) 'Cost model for an income and expenditure survey.' In 'Household Sample Surveys in Developing and Transition Countries,' vol. ST/ESA/STAT/SER.F/96 of *Series F* (Department of Economic and Social Affairs, United Nations Statistics Division) studies in methods 13, pp. 267–277

Rao, J., and R. Sitter (1995) 'Variance estimation under two-phase sampling with application to imputation for missing data.' *Biometrika* 82(2), 453–460

Särndal, C.-E., B. Swensson, and J. Wretman (2003) *Model Assisted Survey Sampling* (Springer)

Schreiner, M. (2014a) 'How do the poverty scorecard and the PAT differ?' *mimeo*

_ (2014b) 'The process of poverty-scoring analysis.' *mimeo*

Sitter, R. (1997) 'Variance estimation for the regression estimator in two-phase sampling.' *Journal of the American Statistical Association* 92, 780–787

Stifel, D., and L. Christiaensen (2007) 'Tracking poverty over time in the absence of comparable consumption data.' *World Bank Economic Review* 21(2), 317–341

Tamhane, A.C. (1978) 'Inference based on regression estimator in double sampling.' *Biometrika* 65(2), 419–427

Tarozzi, A. (2007) 'Calculating comparable statistics from incomparable surveys, with an application to poverty in India.' *Journal of Business & Economic Statistics* 25(3), 314–336

_ (2011) 'Can census data alone signal heterogeneity in the estimation of poverty maps?' *Journal of Development Economics* 95(2), 170–185

Tarozzi, A., and A. Deaton (2009) 'Using census and survey data to estimate poverty and inequality for small areas.' *Review of Economics and Statistics* 91(4), 773–792

van Veelen, Matthijs, and Roy van der Weide (2008) 'A note on different approaches to index number theory.' *American Economic Review* 98(4), 1722–1730

Yoshida, N., R. Munoz, A. Skinner, C. Kyung-eun Lee, M. Brataj, W. Durbin, and D. Sharma (2015) 'Swift data collection guidelines version 2.' mimeo, The World Bank

A Proofs

Proof of Theorem 6 Letting $k = K$ (or $r = 1$) and $\hat{\theta}^I = \hat{\theta}$ in the proof of Theorem 10, we obtain the proof of Theorem 6. □

Proof of Theorem 10 By an exact first-order Taylor expansion of $\hat{\theta}^I$ around θ, the Law of Large Numbers, and Assumption 4, we have the following as $J \to \infty$:

$$\hat{\mu}^{DS} = \frac{1}{JK}\sum_c\sum_h g(x_{ch},\theta) + \frac{1}{JK}\sum_c\sum_h \frac{\partial g(x_{ch},\tilde{\theta}^I)}{\partial\theta^T}(\hat{\theta}^I - \theta) \xrightarrow{p} \mu, \tag{27}$$

where $\tilde{\theta}^I$ is between θ and $\hat{\theta}^I$.

By the Central Limit Theorem and Assumption 4, we obtain:

$$\begin{aligned}\sqrt{J}(\hat{\mu}^{DS} - \mu) &= \frac{1}{\sqrt{J}}\sum_c\left[\frac{1}{K}\sum_h g(x_{ch},\theta) - \mu\right] + \frac{1}{J}\sum_c\frac{1}{K}\sum_h\frac{\partial g(x_{ch},\tilde{\theta})}{\partial\theta^T}\sqrt{J}(\hat{\theta}^I - \theta)\\ &\xrightarrow{p} \mathcal{N}(0, V_g + r^{-1}M_g^T V_\theta M_g/K),\end{aligned}$$

as $J \to \infty$. This completes the proof. □

Proof of Theorem 11 Let $\phi(\cdot)$ be the probability density function for the standard normal distribution. Then, the log-likelihood function $l_c(\theta)$ for cluster c and has the following form:

$$\begin{aligned}l_c(\theta) &\equiv \ln\left[\int_{-\infty}^{\infty}\frac{1}{\sigma_\eta}\phi\left(\frac{\eta_c}{\sigma_\eta}\right)\prod_h\frac{1}{\sigma_e}\phi\left(\frac{y_{ch} - x_{ch}^T\beta - \eta_c}{\sigma_e}\right)\mathrm{d}\eta_c\right]\\ &= -\frac{1}{2}\left[\frac{1}{\sigma_e^2}\left[\sum_h[y_{ch} - x_{ch}^T\beta]^2 - \frac{\sigma_\eta^2}{\sigma_e^2 + k\sigma_\eta^2}\left[\sum_h y_{ch} - x_{ch}^T\beta\right]^2\right] + \ln\left[k\frac{\sigma_\eta^2}{\sigma_e^2} + 1\right] + k\ln[2\pi\sigma_e^2]\right],\end{aligned}$$

where the set of parameters to be estimated is $\theta = (\beta^T, \sigma_\eta^2, \sigma_e^2)^T$. Therefore, the log-likelihood satisfies $l(\theta) \equiv \sum_c l_c(\theta)$ and the maximum likelihood estimator is given by $\hat{\theta}^I = l(\theta)$. It is straightforward to show that $\hat{\theta}^I$ satisfies Assumption 9 with V_θ^I given by the following:

$$\begin{aligned}V_\theta^I &= -E^{-1}\left[\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^T}\right]\\ &= -J^{-1}E^{-1}\left[\frac{\partial^2 l_c(\theta)}{\partial\theta\partial\theta^T}\right]\\ &= E\begin{bmatrix} E^{-1}[X_c^T\Omega_c^{-1}X_c] & \mathbf{0}_L^T \\ \mathbf{0}_L & \frac{2\sigma_e^4}{k(k-1)}\begin{bmatrix}\frac{k(k-1)\sigma_\eta^4 + 2(k-1)\sigma_\eta^2\sigma_e^2 + \sigma_e^4}{\sigma_e^4} & -1 \\ -1 & k\end{bmatrix}\end{bmatrix},\end{aligned} \tag{28}$$

where $\mathbf{0}_L$ is an L-vector of zeros.

By the definition of M_g, we have:

$$M_g = -E\left[\frac{\Sigma_{\phi x}^T}{\sigma_u}, \frac{\Sigma_{\phi B}}{2\sigma_u^2}, \frac{\Sigma_{\phi B}}{2\sigma_u^2}\right] \tag{29}$$

Applying eqs. (28) and (29) to Theorem 10, we obtain eq. (6). □

Proof of Theorem 14 By the definition of $\bar{Y}$, the following holds:

$$E[\bar{Y}] = n^{-1}\sum_c \sum_h E[Y_{ch}] = \mu$$

By Assumptions 1 and 2, Lemma 13, and the Law of Iterated Variance, we have:

$$\begin{aligned}
\text{var}[\bar{Y}] &= \frac{1}{J}\text{var}[\bar{Y}_c] \\
&= \frac{1}{J}\left[E_{X_c}[\text{var}[\bar{Y}_c|X_c]] + \text{var}_{x_c}[E[\bar{Y}_c|X_c]]\right] \\
&= \frac{1}{J}\left[E_{X_c}\left[E_{\eta_c}[\text{var}_{e_{ch}}[\bar{Y}_c|\eta_c, X_c]] + \text{var}_{\eta_c}[E_{e_{ch}}[\bar{Y}_c|\eta_c, X_c]]\right] + \text{var}_{x_c}\left[\frac{1}{K}\sum_h g(x_{ch}, \theta)\right]\right] \\
&= \frac{\tilde{V}_i}{n} + \frac{\tilde{V}_c}{J},
\end{aligned}$$

where $\bar{Y}_c \equiv K^{-1}\sum_{h=1}^K Y_{ch}$. □

Proof of Lemma 17 The results follows directly from the fact that $\rho = \tau$ when $w_c c = 0$ and $w_h = 0$ and the fact that ρ is an increasing function of $w_c c$ and w_h for all τ. The latter is established in Lemma 16. □

B Details of the variance and cost parameter estimates

To compute ρ in eq. (22), we need a realistic set of values for the following parameters: τ, c, w_c, and w_h. All these values are context-dependent. Further, w_c, and w_h are also dependent on the choice of covariates and poverty line. However, to see the potential benefits of double sampling, it is useful to have a reasonable range of values that the designers of surveys may encounter. Therefore, we compiled the estimates of these parameter values from various sources to obtain

a plausible range for each of these parameters.

Some empirical values of τ

Beegle et al. (2012) consider the cost implications for various types of questionnaire in Tanzania. For example, Beegle et al. (2012, Table 10) report that four households can be interviewed by an interviewer per day. Assuming that each interviewer interviews for eight hours a day, it takes $8 \cdot 60/4 = 120$ minutes to complete a survey with recall consumption model.

The same study also reports that it takes on average 41 minutes to complete a short consumption module with 17 most important items (Beegle et al., 2012, Figure 1). Therefore, the proportion of time that is spent to collect consumption data is $41/120 \approx 0.34$. If we assume that the data entry cost is roughly proportionate to the time needed for data collection and that there is no other cost items, we have $\tau \approx 1 - 0.34 = 0.66$ in this case.

This estimate may be close to the upper bound of τ. The average time to complete a recall consumption module is longer when a long questionnaire (with 58 items) is used and when a longer recollection period is adopted, even though the difference is modest. When the personal diary format is used, the cost of collecting consumption data can be much higher because it involves frequent visits to the household, As reported in Beegle et al. (2012, Table 10), the number of interviews that can be performed per day may be as low as 0.35. Assuming that the time required to complete the non-consumption component remains the same, the value of τ when personal diary format is used is $\tau \approx 0.66 \cdot 0.35/4 = 0.06$.

It should be noted that Beegle et al. (2012) only consider a relatively short questionnaire with a complete consumption module and not a large multi-topic survey. Therefore, when we consider a typical Living Standard Measurement Survey, the value of τ may be higher as the weights of non-consumption modules become more important. On the other hand, if we are only concerned with getting an accurate estimate of poverty, τ could be lower because we typically need a small fraction of non-consumption modules to predict consumption.

In Bangladesh, Ahmed et al. (2014) adopt the assumption that five enumerators have to stay for two weeks to complete the survey in one primary sampling unit (PSU) to estimate the cost of conducting Household Income Expenditure Survey in Bangladesh. They further assume that

two persons engage in collecting consumption data and three persons collect non-consumption data for two weeks in one PSU. Under these assumptions, we have $\tau = 3/5 = 0.6$. Given these examples, we use $\tau = \{0.06, 0.36, 0.66\}$ as a plausible set of values to consider.

Some empirical values of c

Compared with τ, there are more studies that provide us with some empirical values of c. For Demographic and Health Surveys, Aliaga and Ren (2006, Table 3.1) present an estimate of c based on past eight surveys. The value varies from 10 in Cambodia and Uganda to 52 in Togo. For consumption surveys, we are not aware of a study that provides cost comparisons from multiple surveys. Therefore, we derive the estimates of c from existing studies.

Let us start with Pettersson and Sisouphanthong (2005, p.275), who provide the cost ratio, or the ratio of the cost of adding a PSU to the cost of adding a household, for the third Lao Expenditure and Consumption Survey (LECS-3) conducted in 2002-2003. This ratio is nothing but c in our notation. They report $c = 3.9$ in urban areas and $c = 6.1$ for rural areas. The lower cost of c in urban areas reflects the lower cost of travel to move between primary sampling units in the urban areas.

A few cautions are in order here. First, travel costs and field allowances are included in the calculation of c but the permanent staff salaries are excluded. As Pettersson and Sisouphanthong (2005) claim, the cost ratio may be affected only slightly by the omission of salaries as the omission will have rather similar effects on both the denominator and numerator of the ratio.

Second, as noted by Pettersson and Sisouphanthong (2005), the cost ratios in their study are rather low. This reflects the fact that the survey required considerable time for interview and follow-up per household over the month when the interviewer-supported diary method was used. Therefore, when a simpler consumption module is used, c is likely to be higher.

Next, we also calculated c based on the survey costs reported in Humphreys (1979, p.396) for the Ada Baseline Survey (ABS) conducted in rural Ethiopia. Their first stage sample consists of 87 administrative localities and costs \$5,547, whereas 632 observations were in the second stage costs at the cost of \$5,899.[16] Therefore, we have an estimate of c as follows:

[16] In addition, there were fixed costs of \$3,589. Humphreys (1979) use a slightly different cost function for their analysis, but their cost model reduces to ours when we ignore the terms involving square roots.

$c = (5547/87)/(5899/632) \approx 7$.

The numbers obtained above are calculated from a highly aggregated budgetary figures. Therefore, we consider the generic, all-inclusive budget for a one year, 3,200-household living standards survey presented in Grosh and Muñoz (1996, Table 8.2). While the numbers are hypothetical, they are meant to be used to create a prototype budget and may well serve our purpose to create a ballpark figure. We assume that each cluster has 16 households as with a majority of Living Standards Measurement Study (LSMS) surveys reviewed in Table 4.1 of Grosh and Muñoz (1996), which implies that there are 200 clusters in this hypothetical survey.

There remain a challenge in deriving an estimate of c because we need to assign each cost components to (i) variable costs proportionate to the number of clusters, (ii) variable costs proportionate to the number of households, and (iii) the fixed costs. We decided to assign all costs relating to "travel allowance" and vehicles, fuel, and car maintenance to (i) variable costs proportionate to the number of clusters, which amounts to $345,880 for 200 clusters. For (ii) variable costs proportionate to the number of households, we include all the costs relating to "base salaries" and printing of questionnaire as well as "materials" other than vehicles, fuel, and car maintenance, which in total amount to $475,150 for 3,200 households. The remaining costs including "consultancy and travel" and "other" are taken as (iii) fixed costs. Based on these classifications, an estimate of c can be obtained as follows: $c = (345880/200)/(475150/3200) \approx 12$.

We have also collected similar budgetary information for several household surveys through personal communications. We then computed c in a similar manner. Table 3 summarizes all the estimates of c we have obtained using the budgetary information we obtained from personal communications. Based on this table, we choose to use the following values of the cost ratio: $c \in \{4, 16, 64\}$.

Note that the cost calculations are based on average cost. Because the actual cost function for a survey depends on the logistical arrangement and is in general not exactly equal to eq. (16), the marginal and average costs are likely to differ. For example, we would not need to incur additional training cost to just add one more household to the sample. However, if we scale up the survey significantly, it is likely that we need to increase most of the cost components, such as the costs of personnel, travel, and materials. Therefore, the calculations of c provided in Table 3 should be taken as approximations.

Table 3: Summary of the estimates of c

Country/Area	c	Survey	Source
Cambodia	10	Demographic and Health Survey	Aliaga and Ren (2006)
Uganda	10	Demographic and Health Survey	Aliaga and Ren (2006)
Jordan	12	Demographic and Health Survey	Aliaga and Ren (2006)
Ethiopia	12	Demographic and Health Survey	Aliaga and Ren (2006)
Haiti	15	Demographic and Health Survey	Aliaga and Ren (2006)
Turkey	27	Demographic and Health Survey	Aliaga and Ren (2006)
Burkina Faso	48	Demographic and Health Survey	Aliaga and Ren (2006)
Togo	52	Demographic and Health Survey	Aliaga and Ren (2006)
Laos-Urban	3.9	Third Lao Expenditure and Consumption Survey	Pettersson and Sisouphanthong (2005)
Laos-Rural	6.1	Third Lao Expenditure and Consumption Survey	Pettersson and Sisouphanthong (2005)
Ethiopia	7	Ada Baseline Survey	Humphreys (1979)
$---$	12	Generic LSMS	Grosh and Muñoz (1996)
Malawi	39	Malawi Third Integrated Household Survey 2010/11	Personal communication
Nigeria	24	Nigeria General Household Survey 2012/13	Personal communication
Niger	31	2011 National Survey on Household Living Conditions and Agriculture	Personal communication
Estonia	10	Household Budget Survey 2012	Personal communication

Some empirical values of w_c and w_h

To consider the values of w_c and w_h in a realistic setup, we take several prediction models and poverty lines used in small-area estimation or survey-to-survey imputation in several countries. For Cambodia, we adopt the consumption model and poverty line used in a small-area estimation project detailed in Ministry of Planning and United Nations World Food Programme (2002) and Fujii (2006). This project uses the Cambodia Socioeconomic Survey 1997. In each of the three strata (Phnom Penh, Other Urban and Rural), a separate consumption model and a separate poverty line are used. Our point estimates of β (unreported) are slightly different from those used in this project because the heteroskedasticity of e_{ch} is allowed for in the former and the estimation method is slightly different as a result. However, the difference is not important for our purpose because we are only interested in the plausible range of w_c and w_h.

For Malawi, Niger, and Tanzania, we used a prediction model and national poverty line from an independent survey-to-survey imputation project.[17] The datasets used are Malawi Third Integrated Household Survey 2010-2011 for Malawi, l'Enquête Nationale sur les Conditions de Vie des Ménages et l'Agriculture 2011-12 for Niger, and Tanzania National Panel Survey 2010-11 for Tanzania. All these surveys are a part of the Living Standard Measurement Surveys by the World Bank.

Table 4 provides a summary of the estimates of w_c, w_h, $\tilde{c}$, $\tilde{\psi}$, and α for various surveys. The first two columns report the basic information about the survey such as the number of observations (n) and number of clusters (K). The next two columns show that the values of w_c and w_h vary substantially across countries. Based on this, we use $w_c \in \{0.6, 1.2, 1.8\}$ and $w_h \in \{0.4, 1.2, 3.6\}$. The next two columns report the implied travel cost $tildec$ under the assumption of optimal sampling for the sample direct estimator and the ratio $\tilde{\psi}$ of data collection costs between the sample direct and single-sample prediction estimators with the implied travel cost. As shown in the last column, the ratio α of variances between the cluster- and household-specific error terms satisfies eq. (9), which is the condition for the approximation we use.

[17] Specific details, including the model specifications used, are available upon request.

Table 4: Summary of the estimates of w_c, w_h, $\tilde{c}$, $\tilde{\psi}$, and α.

Country	n	K	w_c	w_h	$\tilde{c}^{\dagger}$	$\tilde{\psi}$	α
Cambodia (Phnom Penh)	1200	120	1.141	3.600	15	0.945	0.073
Cambodia (Other Urban)	1000	100	0.961	1.357	26	0.977	0.013
Cambodia (Rural)	3810	254	1.809	1.999	81	0.965	0.199
Malawi	12239	768	0.612	0.667	67	0.948	0.268
Niger	3961	270	1.415	0.761	117	0.983	0.201
Tanzania	3272	349	0.708	0.990	22	0.961	0.123

† In Malawi, Niger, and Tanzania, there are some variations in the cluster size. Therefore, we use the average cluster size ($\bar{K} \equiv n/J$) in eq. (26) to derive $\tilde{c}$. In Cambodia, all the cluster have exactly the same number of households.