6-2013

Cost-effective estimation of the population mean using prediction estimators

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

Roy VAN DER WEIDE
World Bank

DOI: https://doi.org/10.1596/1813-9450-6509

WPS6509

Policy Research Working Paper 6509

Cost-Effective Estimation of the Population Mean Using Prediction Estimators

Tomoki Fujii
Roy van der Weide

The World Bank
Development Research Group
Poverty and Inequality Team
June 2013

Policy Research Working Paper 6509

Abstract

This paper considers the prediction estimator as an efficient estimator for the population mean. The study may be viewed as an earlier study that proved that the prediction estimator based on the iteratively weighted least squares estimator outperforms the sample mean. The analysis finds that a certain moment condition must hold in general for the prediction estimator based on a Generalized-Method-of-Moment estimator to be at least as efficient as the sample mean. In an application to cost-effective double sampling, the authors show how prediction estimators may be adopted to maximize statistical precision (minimize financial costs) under a budget constraint (statistical precision constraint). This approach is particularly useful when the outcome variable of interest is expensive to observe relative to observing its covariates.

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.

Cost-effective estimation of the population mean using prediction estimators*

Tomoki Fujii† Roy van der Weide‡

JEL classification codes: C20, C53, I32.

Keywords: Prediction; Double sampling; Maximum likelihood; Generalized method of moment; Regression estimator.

*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, it is common to have a situation where the outcome variable of interest is costly to observe while its covariates are relatively inexpensive to observe. For example, measurement of consumption poverty is expensive because it involves a long questionnaire often administered over an extended period of time. However, its covariates, such as asset holdings, water and lighting sources, and housing materials, are relatively inexpensive to observe. In health sciences, simple oral questions and anthropometric data taken by a non-invasive device often serve as a predictor of the outcome that is expensive to measure. In industries, certain non-destructive testing may serve as a cheaper but less accurate alternative of destructive testing.

In these cases, prediction estimators offer a useful alternative. Prediction estimators estimate the mean outcome for the population by evaluating the mean of predicted outcomes, which exploits the information contained in the available covariates. The data on covariates is not utilized when estimating the population mean by the mean of observed outcomes. For applications of this type of estimator to poverty measurement see e.g. Elbers et al. (2003), Stifel and Christiaensen (2007), Christiaensen et al. (2012), and Douidich et al. (2013), where poverty rate estimates are derived from predicted household consumption data. In an application to health measurement, Fujii (2010) estimates the prevalence of stunting and underweight of children using predicted data.

As is shown by Matloff (1981), prediction estimators may be useful even when the outcome variable is in fact observed for all subjects, simply because the mean of predicted outcomes can be a more efficient estimator of the population mean than the sample mean of the observed outcome variable. By extension, this suggests that prediction estimators might also be incorporated into the design of surveys where the outcome variable is collected for a sub-sample of subjects only.

This study can be viewed as a generalization of Matloff (1981), who uses a weighted-least square estimator for prediction, in several directions. First, we establish the properties of the prediction estimator when prediction is based on a general class of consistent and asymptotically normal estimators. Second, we derive the condition under which the prediction estimator

based on Generalized-Method-of-Moment (GMM) estimation is no less efficient than the sample mean and *any* linear combination of the sample mean and the prediction estimator itself. In the special cases where the estimator is a least-squares (LS) estimator or maximum likelihood (ML) estimator, this condition is guaranteed to hold under suitable regularity conditions.

Third, we further specialize in a case where the outcome variable of interest is binary. This is an interesting case as it is often encountered in empirical applications. It also offers an opportunity to compare a number of different prediction strategies that are motivated by different assumptions. We consider two ML estimators; one where the continuous state variable that generates the outcome variable serves as the dependent variable (MLC) and the other where the binary outcome variable serves as the dependent variable (MLB). The OLS estimator is included as a third candidate, which coincides with MLC when the errors are normally distributed. For general errors, OLS may be viewed as pseudo-ML. As expected, prediction based on MLC is most efficient.

Finally, we apply the prediction estimator to cost-effective double-sampling, where the outcome variable is collected only for a subset of subjects, while the covariates are observed for all subjects. This allows us to maximize statistical precision [minimize financial costs] under a budget constraint [statistical precision constraint]. Such a sampling scheme is particularly useful when it is expensive to collect data on the outcome variable while the budget for data collection is limited.

Therefore, besides generalizing Matloff (1981), this study also contributes to the literature on double sampling (also called two-phase sampling), which dates back to Neyman (1938) and Bose (1943). By considering a prediction estimator in the context of GMM estimation, we extend earlier studies of double sampling based on a linear model such as Cochran (1977), Tamhane (1978), Palmgren (1987), Davidov and Haitovsky (2000), and Särndal et al. (2003).[1] Among these studies, this study is particularly closely related to Särndal et al. (2003), who applies double sampling to stratification to economize on the cost of estimation. However, our study is different because we allow for non-linear models and only use predictions for estimation.

With notable exceptions of Cochran (1977) and Särndal et al. (2003), the discussion of cost-

[1]For an empirical application of double sampling, see Hansen and Tepping (1990).

effectiveness is only implicit in most double-sampling studies. By bringing the prediction estimator to a general GMM context and explicitly discussing optimal cost-effective estimation of the population mean, we hope to expand the scope of possible applications of double-sampling. A recent example where double-sampling (without optimization) is explored as an option to reduce the costs of poverty measurement in Bangladesh can be found in Ahmed et al. (2013).

2 Prediction estimator

Suppose that the potentially limited outcome variable Y_i for subject i is related to the continuous state variable y_i by $Y_i = h(y_i)$, where h is a function of y_i. Further, the state variable y_i is related to a vector of covariates x_i and a disturbance term u_i by $y_i = q(x_i, u_i; \Theta)$, where Θ is a vector of parameters. We assume for now that the state variable y_i is unobservable. We define the expected value of Y_i given x_i by $g(x_i, \theta) \equiv E_u[Y_i|x_i]$, where θ is a K-vector of identifiable model parameters with $K \leq \dim(\Theta)$ and g is assumed differentiable with respect to θ. We also define $\epsilon_i \equiv Y_i - g(x_i, \theta)$ and make the following assumptions about x_i and u_i:

Assumption 1 *The pair (x_i, u_i) is independently and identically distributed across i.*

This assumption is relaxed in Section 6.

Assumption 2 *The variables x_i and u_i are independent for all i.*

Notice that the parameter of interest in this study is $\mu \equiv E[Y_i] = E_x[g(x_i, \theta)]$ and not Θ or θ. The standard way to estimate μ is to take the sample mean $\bar{Y} = N^{-1} \sum_i Y_i$, where N denotes the sample size. However, $\bar{Y}$ is not the most efficient estimator in general. Using a weighted least-square (WLS) estimator $\hat{\theta}^{WLS}$ for θ, Matloff (1981) has shown that the sample mean $\hat{\mu}^{WLS} \equiv N^{-1} \sum_i g(x_i, \hat{\theta}^{WLS})$ of predicted values is asymptotically no less efficient than $\bar{Y}$.

It is not obvious whether the results of Matloff (1981) extend to other estimators. Therefore, we first extend Matloff (1981) to a general case where predictions are made using a consistent and asymptotically normal estimator for θ. To this end, we make the following assumption:

Assumption 3 *The estimator $\hat{\theta}$ of the model parameters θ satisfies the following properties:*

$$\hat{\theta} \xrightarrow{p} \theta \quad \text{and} \quad \sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \Omega^{-1}) \quad \text{as} \quad N \to \infty,$$

where Ω is a symmetric positive-definite $K \times K$ matrix.

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 poses no problem in empirical applications.

Using $\hat{\theta}$, we can now construct the prediction estimator $\hat{\mu}(\hat{\theta}) \equiv N^{-1} \sum_i g(x_i, \hat{\theta})$, which satisfies the following properties:[2]

Theorem 4 *Let $M_g \equiv E[\partial g(x_i, \theta)/\partial \theta] (\neq 0_K)$ and $V_g \equiv \text{var}[g(x_i, \theta)]$, where 0_K is a column K-vector of zeros. Then, under Assumptions 1, 2, and 3, we have:*

$$\hat{\mu} \xrightarrow{p} \mu \quad \text{and} \quad \sqrt{N}(\hat{\mu}(\hat{\theta}) - \mu) \xrightarrow{d} \mathcal{N}(0, V_g + M_g^T \Omega^{-1} M_g) \quad \text{as} \quad N \to \infty. \tag{1}$$

Theorem 4 is a direct extension of Matloff (1981) to a very general case where the estimator for the model parameter vector θ used for prediction is consistent and asymptotically normally distributed. Note that the asymptotic variance of $\hat{\mu}(\hat{\theta})$ can be consistently estimated by replacing V_g, M_g and Ω with their consistent estimators. For ease of presentation, we drop the argument $\hat{\theta}$ and simply write $\hat{\mu}$ until Section 4.

We now specialize in a case where the estimator for prediction is the optimal GMM estimator. This is still fairly general and relevant to empirical applications since many of the estimators widely used in practice, including the least-squares (LS) and ML estimators, can be viewed as an optimal GMM estimator, even though the concept of "optimality" is irrelevant when θ is exactly identified.

Henceforward, we maintain the following assumptions:

Assumption 5 *An L-vector of moment functions $m(\theta, x_i, Y_i)$ satisfies $E_u[m(\theta, x_i, Y_i)|x_i] = 0_L$, where m is differentiable with respect to θ, $L \geq K$, and $V_m \equiv \text{var}[m(\theta, x_i, Y_i)]$.*

[2]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). All proofs are provided in Appendix B.

Assumption 6 *The estimator $\hat{\theta}$ of the model parameters is given by:*

$$\hat{\theta} \equiv \arg\min_{\theta} \left[\frac{1}{N} \sum_i m^T(\theta, x_i, Y_i) \right] \hat{W}_N \left[\frac{1}{N} \sum_i m(\theta, x_i, Y_i) \right], \tag{2}$$

where $\hat{W}_N$ is a positive-definite symmetric weighting $L \times L$-matrix satisfying $\hat{W}_N \xrightarrow{p} V_m^{-1}$ as $N \to \infty$.

It is convenient to define $M_m \equiv E[\partial m(\theta, x_i, Y_i)/\partial \theta^T]$, which is the expected gradient of the moment functions. The following results are well-known (Hansen, 1982; Cameron and Trivedi, 2005):

Theorem 7 *Assumptions 1, 2, 5, and 6 imply Assumption 3 with $\Omega = M_m^T V_m^{-1} M_m$.*

Remark 8 *A conventional choice of $\hat{W}_N$ is $[N^{-1} \sum_i m(\tilde{\theta}, x_i, Y_i) m^T(\tilde{\theta}, x_i, Y_i)]^{-1}$, where $\tilde{\theta}$ is a consistent estimator for θ.*

In what follows, we verify whether the following condition holds:

$$\text{acov}[\bar{Y}, \hat{\mu}] = \text{avar}[\hat{\mu}]. \tag{3}$$

To see the significance of eq. (3), the following lemma is useful:

Lemma 9 *Suppose that Assumptions 1, 2, and 3 hold. Then, if eq. (3) holds, the asymptotic variance of the linear combination $\check{\mu} \equiv \alpha \bar{Y} + (1-\alpha)\hat{\mu}$ of the sample average $\bar{Y}$ and prediction estimator $\hat{\mu}$ for $\alpha \in \mathbb{R}$ is minimized when $\alpha = 0$ (i.e., when $\check{\mu} = \hat{\mu}$). Furthermore, eq. (3) implies:*

$$\text{avar}[\hat{\mu}] \leq \text{avar}[\bar{Y}], \tag{4}$$

where the equality holds if and only if $\hat{\mu} = \bar{Y}$ almost surely.

Lemma 9 shows that the prediction estimator $\hat{\mu}$ is asymptotically at least as efficient as the sample average $\bar{Y}$ and any linear combination of $\hat{\mu}$ and $\bar{Y}$, provided that eq. (3) is satisfied. It thus follows that eq. (3) is crucial for establishing the efficiency of $\hat{\mu}$. Note that eq. (3) is exactly what Matloff (1981) has shown to hold for the prediction estimator (based on WLS estimation of the model parameters) to establish its efficiency relative to $\bar{Y}$.

The question now is whether eq. (3) holds in general. It turns out that a moment condition has to be satisfied for eq. (3) to hold:

Theorem 10 *Let $m_i \equiv m(\theta, x_i, Y_i)$. Then, under Assumptions 1, 2, 5, and 6, eq. (3) holds regardless of the distribution of x if and only if the following equation holds:*

$$M_g = -M_m^T V_m^{-1} E[m_i Y_i]. \tag{5}$$

When θ is exactly identified, eq. (5) is equivalent to:

$$E[m_i Y_i] = -M_m^{-T} V_m M_g. \tag{6}$$

Theorem 10 shows that the moment condition eq. (5) must be satisfied for the prediction estimator to be at least as efficient as the sample mean. One important special case of Theorem 10 is when $\hat{\theta}$ is an LS estimator, in which case eq. (5) is automatically satisfied.

Assumption 11 *The estimator $\hat{\theta}$ of θ is an LS estimator such that it satisfies:*

$$\hat{\theta} = \arg\min_{\theta} N^{-1} \sum_i (Y_i - g(x_i, \theta))^2.$$

Further, for $m_i = (Y_i - g_i)\partial g_i / \partial \theta$, Assumption 5 is satisfied.

Under Assumption 11, Theorem 10 leads to the following corollary:

Corollary 12 *Suppose Assumptions 1, 2, and 11 are satisfied. Then, Assumption 6 and eq. (3) hold.*

Note that the results above do not change even when the expectations are taken with weights. Therefore, Corollary 12 is simply a restatement of the results proven by Matloff (1981). It should also be apparent from Theorem 10 and Corollary 12 that the main finding of Matloff (1981) that the WLS-based prediction estimator is at least as efficient as the sample mean does not hold for GMM-based prediction estimators unless the moment condition given in eq. (5) holds.

Another important special case is where $\hat{\theta}$ is an ML estimator.

Assumption 13 *Suppose that the estimator $\hat{\theta}$ of θ is an ML estimator such that it satisfies the following first-order condition:*

$$\sum_i \left.\frac{\partial l_i}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0,$$

where $l_i(\theta, x_i, Y_i)$ is the individual log-likelihood function. Further, for $m_i \equiv \partial l_i/\partial\theta$, Assumption 5 is satisfied.

Assumption 14 *Suppose that ϵ_i has a once-differentiable conditional probability density function $f_\epsilon(\epsilon_i|x_i)$. Further, let $l_i \equiv \ln f_\epsilon(Y_i - g(x_i,\theta)|x_i)$. Then, the standard ML regularity conditions are satisfied for the ML estimator in Assumption 13.*

With these assumptions, we have the following corollary:

Corollary 15 *Suppose that Assumptions 1, 2, 13, and 14 are satisfied. Then, eq. (3) holds.*

So far, we have not specified the functional form of $q(x_i, u_i; \Theta)$. Hereafter, we make the following linearity assumption, which is commonly used:

Assumption 16 *The variables y_i and x_i are related by the following equation:*

$$y_i = q(x_i, u_i; \Theta) = x_i^T \beta + u_i, \tag{7}$$

where $\Theta = \{\beta, \sigma_u\}$, the disturbance term u_i has a twice-differentiable cumulative distribution function F_0 and a probability density function $f_0 (\equiv F_0')$ with $E[u_i|x_i] = 0$ and $\sigma_u^2 \equiv \text{var}[u_i]$.

It is useful to define the normalized disturbance term: $\tilde{u} \equiv u/\sigma_u$, where its probability density function $\tilde{f}$ and cumulative distribution function $\tilde{F}$ satisfy $\tilde{f}(\tilde{u}) = \sigma_u f_0(u)$ and $\tilde{F}(\tilde{u}) = F_0(u)$ for all $\tilde{u} \in \mathbb{R}$, respectively.

One model that satisfies Assumption 16 is the Tobit model. This is an important model as it is widely used and provides a concrete example where the prediction estimator is in general not efficient.

Example 17 *Assume that Assumptions 1, 2, 13, and 16 hold. Further, $\tilde{u}$ has a standard normal distribution and $Y_i = h(y_i) = \max(0, y_i)$. Then, it can be shown that eq. (3) does not hold in*

general. Therefore, there exists a linear combination of $\hat{\mu}$ and $\bar{Y}$ such that it is different from and more efficiency than $\hat{\mu}$.

[3]

3 Binary outcome variable

In this section we further specialize in a case where the outcome variable is binary. This is an important case for two reasons. First, binary outcomes are often encountered in empirical applications which makes the binary model empirically relevant. Second, eq.(3) is seen to hold regardless of the underlying distribution of u in this case, provided that the ML estimator is used for $\hat{\theta}$. To be more specific, we make the following assumption.

Assumption 18 *The variables Y_i and y_i are related by $Y_i = \text{Ind}(y_i > z)$, where z is an unknown constant. Further, the first component of x_i is a constant and β is a column K-vector.*

Note that under Assumption 18, we cannot identify σ_u, β, and z separately. We can only identify $\theta = \tilde{\beta} \equiv ((\beta_1 - z)/\sigma_u, \beta_2/\sigma_u, \beta_3/\sigma_u, \cdots, \beta_K/\sigma_u)$. With this notation, we have the following result:

Theorem 19 *Under Assumptions 1, 2, 13, 16, and 18, eq. (3) holds.*

Matloff (1981) found that $\hat{\mu} \equiv \bar{Y}$ holds almost surely for logistic regressions. It turns out that this is the only case where $\hat{\mu} \equiv \bar{Y}$ holds regardless of the distribution of x_i when the outcome variable is binary.

Theorem 20 *Suppose that Assumptions 1, 2, 13, 16, and 18 hold. Then, $\hat{\mu} = \bar{Y}$ and hence* $\text{avar}[\hat{\mu}] = \text{avar}[\bar{Y}]$ *almost surely for any distribution of x_i if and only if u has the following cumulative distribution function $\tilde{F}$:*

$$\tilde{F}(\tilde{u}) = \frac{1}{1 + \mathrm{e}^{-\tilde{u}}}. \tag{8}$$

[3] It can be shown $\hat{\mu} = N^{-1} \sum_i \tilde{f}(-x_i^T \hat{\beta}) + x_i^T \hat{\beta}/[1 - \tilde{F}(-x_i^T \hat{\beta})]$, where $\hat{\beta}$ is an ML estimate of β. See also Appendix C for further discussion on Example 17.

4 Comparison of prediction estimators

Now, suppose that z is known and both the continuous state variable y_i and the binary outcome variable Y_i are observable. In this case, we can still opt to estimate θ by ML taking Y_i as the (binary) dependent variable (MLB). Alternatively, we can also work with y_i as the (continuous) dependent variable and estimate θ either by ML estimation (MLC) or ordinary least squares (OLS) estimation. To clearly distinguish the prediction estimators based on these different estimators of $\hat{\theta}$, we introduce the following notations:[4]

$$\begin{aligned}
\hat{\theta}^{OLS} &\equiv \begin{pmatrix} \hat{\beta}^{OLS} \\ \hat{\sigma}_u^{OLS} \end{pmatrix} \\
\hat{\theta}^{MLB} &\equiv \arg\max_{\tilde{\beta}} N^{-1} \sum_i Y_i \ln[1 - \tilde{F}(x_i^T \tilde{\beta})] + (1 - Y_i) \ln \tilde{F}(x_i^T \tilde{\beta}) \\
\hat{\theta}^{MLC} &\equiv \arg\max_{\beta, \sigma_u} N^{-1} \sum_i \ln[\tilde{f}((y_i - x_i^T \beta)/\sigma_u)] - \ln \sigma_u,
\end{aligned}$$

where

$$\hat{\beta}^{OLS} \equiv \arg\min_{\beta} N^{-1} \sum_i (y_i - x_i^T \beta)^2 \quad \text{and} \quad \hat{\sigma}_u^{OLS} \equiv \left[N^{-1} \sum_i (y_i - x_i^T \hat{\beta}^{OLS})^2 \right]^{1/2}.$$

Because the estimation of σ_u is not relevant for $\hat{\theta}^{MLB}$, we have $K^{MLB} = J \equiv \dim(x_i)$ and $K^{OLS} = K^{MLC} = J + 1$, where K^a for $a \in \{OLS, MLB, MLC\} (\equiv \mathcal{E})$ represents the dimension of $\hat{\theta}^a$. Using $\hat{\theta}^a$ for $a \in \mathcal{E}$ defined above, we can obtain the estimate $\hat{\tilde{\beta}}^a$ of $\tilde{\beta}^a$.[5] In the case where the outcome variable is binary, we have: $\hat{\mu}(\hat{\theta}^a) = N^{-1} \sum_i \tilde{F}(-x_i^T \hat{\tilde{\beta}}^a) (\equiv \hat{\mu}^a)$ for $a \in \mathcal{E}$.

In this section, we make the following assumption:

Assumption 21 *f is twice differentiable and satisfies $f(u) = f(-u)$ for all $u \in \mathbb{R}$.*

With this assumption, we derive the following results:

[4] To keep the mathematical expressions and proofs simple, we ignore the degree-of-freedom adjustment for $\hat{\sigma}_u^{OLS}$.

[5] That is, $\hat{\tilde{\beta}}^a \equiv ((\hat{\beta}_1^a - z)/\hat{\sigma}_u, \hat{\beta}_2^a/\hat{\sigma}_u, \hat{\beta}_3^a/\hat{\sigma}_u, \cdots, \hat{\beta}_J^a/\hat{\sigma}_u)$ for $a \in \{OLS, MLC\}$ and $\hat{\tilde{\beta}}^{MLB} \equiv \hat{\theta}^{MLB}$.

Theorem 22 *Suppose that Assumptions 1, 2, 13, 16, 18, and 21 hold. Then, the asymptotic variances and covariances of $\hat{\mu}^a$ for $a \in \mathcal{E}$ are as follows:*

$$\begin{aligned}
\text{avar}[\hat{\mu}^{OLS}] &= V_g + \Sigma_{fx}^T \Sigma_{xx}^{-1} \Sigma_{fx} + \frac{(\nu_u - \sigma_u^4)\Sigma_{fB}^2}{4\sigma_u^4} \\
\text{avar}[\hat{\mu}^{MLB}] &= V_g + \Sigma_{fx}^T E_x^{-1} \left[\tilde{f}(B_i)\lambda(B_i) x_i x_i^T \right] \Sigma_{fx} \\
\text{avar}[\hat{\mu}^{MLC}] &= V_g + E_u^{-1} \left[(\tilde{f}'(\tilde{u})/\tilde{f}(\tilde{u}))^2 \right] \Sigma_{fx}^T \Sigma_{xx}^{-1} \Sigma_{fx} \\
&= \text{acov}[\hat{\mu}^{OLS}, \hat{\mu}^{MLC}] = \text{acov}[\hat{\mu}^{MLB}, \hat{\mu}^{MLC}]
\end{aligned}$$

where $\nu_u \equiv E[u^4]$, $B_i \equiv -x_i^T \tilde{\beta}$, $V_g \equiv \text{var}[\tilde{F}(B_i)]$, $\Sigma_{fx} \equiv E_x[\tilde{f}(B_i)x_i]$, $\Sigma_{fB} \equiv E_x[\tilde{f}(B_i)B_i]$, $\Sigma_{xx} \equiv E_x[x_i x_i^T]$, *and* $\lambda(B_i) \equiv \tilde{f}(B_i)/[\tilde{F}(B_i)(1 - \tilde{F}(B_i))]$.

The following corollary immediately follows:

Corollary 23 *Under Assumptions 1, 2, 13, 16, 18, and 21, we have the following relationships concerning the asymptotic variances of $\hat{\mu}$.*

$$\text{avar}[\hat{\mu}^{MLC}] \leq \text{avar}[\hat{\mu}^{OLS}], \quad \text{avar}[\hat{\mu}^{MLC}] \leq \text{avar}[\hat{\mu}^{MLB}]$$

Remark 24 $\text{avar}[\hat{\mu}^{MLC}] \leq \text{avar}[\hat{\mu}^{OLS}]$ *can also be proved by directly comparing them. By Assumption 21,* $E_u[\tilde{f}'(\tilde{u})/\tilde{f}(\tilde{u})] = 0$. *Further, by the definition of* $\tilde{u}$, $\text{var}[\tilde{u}] = 1$. *Hence, we have:*

$$\begin{aligned}
E_u[(\tilde{f}'(\tilde{u})/\tilde{f}(\tilde{u}))^2] &= \text{var}_u[\tilde{f}'(\tilde{u})/\tilde{f}(\tilde{u})] + E_u^2[\tilde{f}'(\tilde{u})/\tilde{f}(\tilde{u})] = \text{var}_u[\tilde{f}'(\tilde{u})/\tilde{f}(\tilde{u})]\text{var}_u[\tilde{u}] \\
&\geq \left| \text{cov}[\tilde{f}'(\tilde{u})/\tilde{f}(\tilde{u}), \tilde{u}] \right| = \left| E_u[\tilde{u}\tilde{f}'(\tilde{u})/\tilde{f}(\tilde{u})] \right| = 1.
\end{aligned}$$

Theorem 22 implies that if y_i is observed and the model is correctly specified, the MLC prediction estimator is most preferable. Also, even though the MLB prediction estimator only uses the information on whether y_i is above z, it is not necessarily less accurate than the OLS prediction estimator. Moreover, the MLB prediction estimator is robust to a misspecification

that does not affect the sign of y_i unlike the OLS and MLC prediction estimators.[6]

Another important point to note is that the MLB prediction estimator can be a useful alternative to the MLC prediction estimator in practice, if Y_i is more cheaply observed than y_i. To highlight this point, suppose that the budget for data collection is fixed and we are only interested in estimating μ. Further, the cost of collecting N_Y observations of (x_i, Y_i) is equal to the cost of collecting N_y observations of (x_i, y_i). Then, if we have $\text{avar}[\hat{\mu}^{MLB}]/N_{MLB} < \text{avar}[\hat{\mu}^{MLC}]/N_{MLC}$, collecting N_Y observations of (x_i, Y_i) and using $\hat{\mu}_{MLB}$ leads to a more accurate estimate of μ than collecting N_y observations of (x_i, y_i) and using $\hat{\mu}_{MLC}$.

5 Application to cost-effective double sampling

In this section, we apply the prediction estimator to cost-effective sampling. If collecting data on x is cheaper than Y, double sampling—where x_i is observed for all subjects but Y_i is observed only for a subset of the sample—may be preferred to the standard single sampling—where both x_i and Y_i are observed for all subjects in the sample. To highlight the benefits of double sampling, we consider the problem of maximizing statistical precision under a given budget constraint and its dual problem of minimizing financial costs given a statistical precision constraint. Formally, we make the following assumption in this section:

Assumption 25 *The covariates x_i are observed for all subjects $i \in \{1, 2, \ldots, N\} \equiv S$, while Y_i is observed only for subjects $i \in \{1, 2, \ldots, \lfloor rN \rfloor\} \equiv S^I$, where $\lfloor \cdot \rfloor$ is the floor operator that gives the maximum integer that does not exceed the argument and $r \in (0, 1]$ denotes the proportion of subjects for which Y_i is observed as $N \to \infty$.*

Let us refer to S^I as Sample 1 and to $S^{II} \equiv S \setminus S^I$ as Sample 2, respectively. Because the indexing of the observations may be changed arbitrarily under Assumption 1, we assume that the first $N^I (\equiv \lfloor rN \rfloor)$ observations are those that contain Y_i. We further make the following assumption:

[6]For example, consider the case where the true relationship between y_i and x_i is given by $y_i = (x_i^T \beta + u_i)^3$ but we use the following misspecified model: $y_i = x_i^T \beta + u_i$. This does not affect $\hat{\mu}^{MLB}$ but it affects $\hat{\mu}^{MLC}$.

Assumption 26 *The distribution of* (x_i, u_i) *is independent of the sample selection. That is:*

$$\Pr(x_i, u_i | i \in S^I) = \Pr(x_i, u_i | i \in S^{II}) = \Pr(x_i, u_i). \tag{9}$$

This requires that the selection into either sample carries no information about x_i or u_i. This is a reasonable assumption when the researcher can decide whether to observe Y_i.

In this setup, we use S^I to compute the estimator $\hat{\theta}^I$ of θ and predict $g(x_i, \hat{\theta}^I)$ for all observations in S. Therefore, the prediction estimator for μ under Assumption 25 is given by:

$$\hat{\mu}^{DS} \equiv N^{-1} \sum_{i \in S} g(x_i, \hat{\theta}^I).$$

It is straightforward to derive the following results:

Theorem 27 *Suppose that Assumptions 1, 2, 25,and 26 holds and* $\hat{\theta}^I$ *satisfies Assumption 3. Then,* $\hat{\mu}^{DS}$ *satisfies the following properties:*

$$\hat{\mu}^{DS} \xrightarrow{p} \mu \quad \text{and} \quad \sqrt{N}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, V_g + r^{-1} M_g^T \Omega^{-1} M_g) \quad \text{as} \quad N \to \infty. \tag{10}$$

Theorem 27 is particularly useful for estimating the population mean μ in a cost effective manner, which can be done by choosing r appropriately. To show this, we make the following assumption:

Assumption 28 *The ratio of the cost of observing only* x_i *to that of observing both* x_i *and* Y_i *for a given subject* i *is* $\kappa (< 1)$.

Hereafter, we normalize the cost of observing both x_i and Y_i to be unity and ignore the fact that N and N^I are integers. The following theorem is relevant if there is a binding budget constraint C (i.e., the budget is just enough to collect C observations of both x_i and Y_i):

Theorem 29 *Given the binding budget constraint* C*, the variance of* $\hat{\mu}^{SD}$ *is minimized when* N *and* r *satisfy:*

$$\begin{aligned} (N, r) &= \arg\min_{(\nu,\rho)} [V_g + \rho^{-1} M_g^T \Omega^{-1} M_g]/\nu \quad \text{s.t.} \quad [\rho + \kappa(1-\rho)]\nu = C \\ &= \left(C / \left[\rho^* + \kappa(1-\rho^*)\right], \rho^*\right), \end{aligned} \tag{11}$$

where $\rho^* = \min\left(1, \sqrt{\kappa(1-\kappa)^{-1} M_g^T \Omega^{-1} M_g V_g^{-1}}\right)$ *and the minimized variance is given by:*

$$\begin{cases} [V_g + M_g^T \Omega^{-1} M_g]/C & \text{if} \quad \rho^* = 1 \\ \left[\sqrt{V_g \kappa} + \sqrt{M_g^T \Omega^{-1} M_g (1-\kappa)}\right]^2 / C & \text{otherwise.} \end{cases}$$

In other cases, we may want to minimize the cost of data collection for a given accuracy. In this case, the dual version of Theorem 29 is relevant.

Corollary 30 *Suppose that the maximum acceptable variance of* $\hat{\mu}^{SD}$ *is given by* $\bar{A}$*. Then, the cost of achieving this accuracy is minimized when* N *and* r *satisfy the following equation:*

$$\begin{aligned} (N, r) &= \arg\min_{\nu,\rho} [\rho + \kappa(1-\rho)]\nu \quad \text{s.t.} \quad [V_g + M_g^T \Omega^{-1} M_g/\rho]/\nu = \bar{A} \\ &= ([V_g + M_g^T \Omega^{-1} M_g/\rho^*]/\bar{A}, \rho^*), \end{aligned}$$

where $\rho^* = \min\left(1, \sqrt{\kappa(1-\kappa)^{-1} M_g^T \Omega^{-1} M_g V_g^{-1}}\right)$ *and the minimized cost is given by*

$$\begin{cases} [V_g + M_g^T \Omega^{-1} M_g]/\bar{A} & \text{if} \quad \rho^* = 1 \\ \left[\sqrt{V_g \kappa} + \sqrt{M_g^T \Omega^{-1} M_g (1-\kappa)}\right]^2 / \bar{A} & \text{otherwise.} \end{cases}$$

Note that when $k \to 0$, we have: $N \to \infty$, $r \to 0$, and $N^I = rN \to C$. In other words, in the limit where the covariates x_i can be collected at no cost ($k \to 0$), it is cost-effective to collect x_i for an infinite number of subjects (or the entire population) without spending any resources (i.e., $N \to \infty$) and spend all budget on collecting the outcome variable Y_i (i.e., $N^I \to C$). Under this scenario, the variance will be entirely driven by the model error; with $N \to \infty$, the sample variance component will tend to zero.

Note also that under these simplifying assumptions concerning the cost function, the ratio of the minimized variance under double-sampling relative to the variance under single-sampling coincides with the ratio of the minimized budget under double-sampling relative to the budget under single-sampling:

$$\zeta(\kappa, \alpha) = \left(\sqrt{\alpha\kappa} + \sqrt{(1-\alpha)(1-\kappa)}\right)^2, \tag{12}$$

where $\alpha = V_g/(V_g + M_g^T \Omega^{-1} M_g)$, which is the proportion of the variance due to sampling error to the total variance of the double-sampling of the single sample estimator.

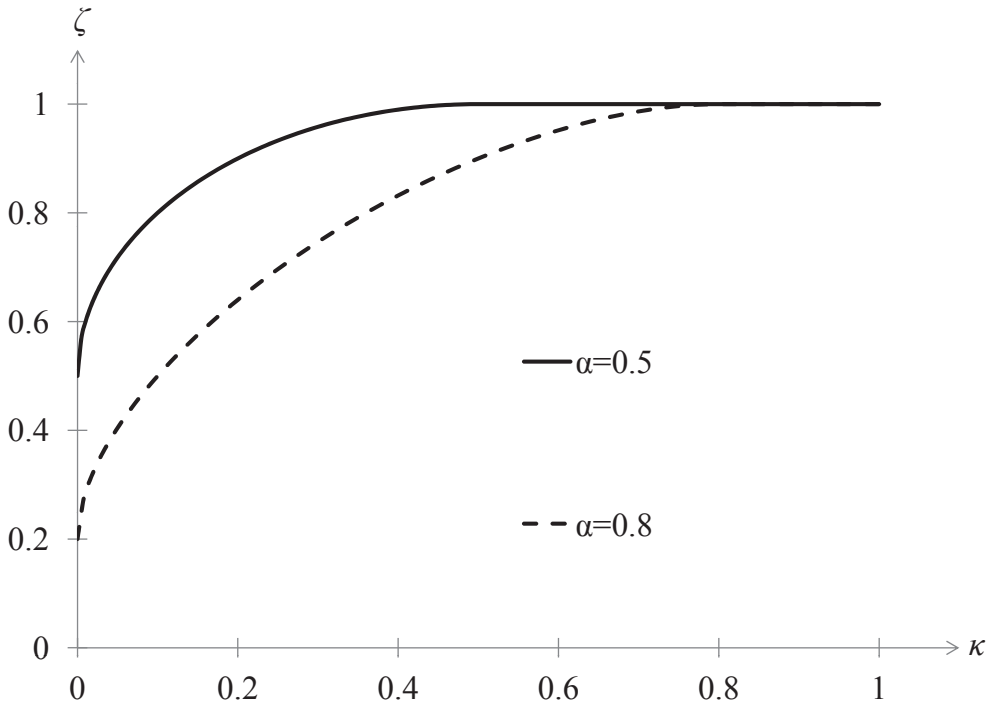

Figure 1: $\zeta(\kappa, \alpha)$ as a function of κ for: $\alpha = 0.5$ and $\alpha = 0.8$.

Figure 1 plots the variance ratio (or budget ratio) as a function of κ for $\alpha = 0.5$ and $\alpha = 0.8$. The figure shows that the benefits of double-sampling are larger when the model error component is small relative to the original sampling error component. The figure also confirms that there is a threshold value for κ above which there will be no gains in using double-sampling relative to single-sampling.

6 Extensions

In the discussion so far, we maintained Assumption 1 to keep the presentation simple. However, this assumption can be relaxed in at least two important ways.

First, we will consider cluster sampling which is a common feature of surveys. This can be easily accommodated in Theorem 4. Since we essentially only rely on Assumptions 2 and 3 for the proof of Theorem 4, all we need to do is to compute the estimates of θ, Ω, V_g, and M_g in a clustering-consistent manner.[7]

[7] See Appendix A for further discussion on this.

Second, Assumption 26 implies that Y_i is missing completely at random. Because we are concerned about cost-effectively estimating the population average by choosing not to observe Y_i for some subjects, this is a reasonable assumption for our main purpose. However, Assumption 26 may be too strong in other cases for which our prediction estimator is potentially applicable. One example is the case of a censored outcome variable.

Example 31 *Consider Example 17 again. This time, assume that Y_i is missing when the (unobservable) state variable y_i satisfies $y_i \leq 0$. If θ is estimated only with Sample 1 (i.e., subjects for which both x_i and Y_i are observed), $\hat{\theta}$ will be biased and thus $\hat{\mu}$ will also be biased.*

Assumption 1 implies that Samples 1 and 2 discussed in Section 5 are generated from the same underlying population. In some cases, the validity of this assumption is not clear, especially when the two samples are taken from two different data sources. In such cases, one can test the validity of the prediction estimator using the following statistic:

Theorem 32 *Suppose that the estimator $\hat{\theta}^I$ of θ is estimated only with S^I and satisfies Assumption 3. Further, let $\bar{Y}^I \equiv \sum_{i \in S^I} Y_i / N^I$ be the sample average of the outcome variable for S^{II} and $\tilde{\mu}^{II} \equiv \sum_{i \in S^{II}} g(x_i, \hat{\theta}^1)/N^{II}$ be the prediction estimator for Sample 2, respectively. Then, under the null hypothesis that Assumptions 1, 2, and 26 hold, the test statistic Z defined below asymptotically follows a standard normal distribution:*

$$Z \equiv \frac{\bar{Y}^I - \tilde{\mu}^{II}}{\sqrt{((1+r^{-1})V_g + \text{var}[\epsilon_i] + r^{-1} M_g^T \Omega^{-1} M_g)/N}}.$$

Note that the rejection of the null hypothesis is consistent with various possibilities. For example, the distribution of x_i may be different between Sample 1 and Sample 2, which can be tested separately. It is also possible that θ is different in the underlying populations of two samples because, for example, the timing of sampling is different. The test described in Theorem 32 alone cannot distinguish these and other possibilities.

7 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. We make a case for using a prediction estimator to obtain more efficient estimates of the population mean by extending the results of Matloff (1981). In an application to cost-effective double sampling, we show how prediction estimators may be adopted to maximize statistical precision [minimize financial costs] under a budget constraint [statistical precision constraint]. The approach is particularly useful when the outcome variable is relatively more expensive to observe than its covariates.

There have been a number of cases in which predictions are used *ex post* to estimate the outcome variable of interest. For example, Elbers et al. (2003) have put forward an approach where the prediction of household consumption per capita is made for each census record, which in turn is used to compute poverty statistics for small areas. Their approach is useful because household consumption is expensive to observe and thus typically observed only for a small set of households in the population of interest. A similar approach can be used to estimate the prevalence of stunting and underweight of children for small areas (Fujii, 2010). There are also a number of recent case studies in which a consumption survey is combined with secondary surveys (without a consumption component) to supplement existing poverty estimates (Stifel and Christiaensen, 2007; Douidich et al., 2013).

Our study shows how prediction methods may be fruitfully adopted even *ex ante* by designing surveys appropriately. A recent study by Ahmed et al. (2013) on Bangladesh shows that poverty can be estimated with reasonable accuracy with a relatively small sample (e.g., 64 Primary Sampling Units (PSUs) with 10 households per PSU) of the outcome variable and its covariates when combined with a larger sample of the covariates only (612 PSUs). Their cost estimates suggest that, compared with a full scale consumption survey (complete observations of both outcome variable and its covariates for 612 PSUs), the manpower cost for data collection could by cut by more than 90 percent without severely undermining the accuracy of the national poverty estimate. If our approach is adopted, the trade-off between financial costs and statistical precision can be explicitly optimized in a situation like theirs.

To what extent we can successfully bring down the costs while maintaining statistical pre-

cision is intrinsically an empirical question. In general, our prediction estimator is most useful when there are covariates that are both inexpensively observable (i.e., low value of κ) and offer a good model for prediction (i.e., low value of $[M_g^T \Omega^{-1} M_g] V_g^{-1}$). It should be apparent that prior estimates of these parameters are required to calculate the optimal sample sizes under double sampling. When these are not available, data collection can be done in two stages. That is, one can collect a small sample with both Y_i and x_i in the first stage, and use this sample to compute the parameters needed to determine the optimal double sampling structure. In practice, this exercise can be done at the time of a pilot survey, in which case there may be practically no additional logistical costs. While the estimate of ρ obtained from the first-stage sample is likely to be noisy, this does not bias the estimate of μ. Moreover, the resulting estimates are likely to be more accurate than the estimates created with only one sample under the same budget constraint. Therefore, the cost-effective estimation approach proposed in Section 5 is likely to be a helpful alternative that will benefit various institutions, including governments, hospitals, laboratories, and factories, that are financially constrained or wish to minimize the cost of estimation while achieving a given level of accuracy.

References

Ahmed, F., C. Dorji, S. Takamatsu, and N. Yoshida (2013) 'Conducting a hybrid survey to improve the reliability and frequency of poverty statistics in Bangladesh.' mimeo, World Bank

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

Cameron, A.C., and P.K. Trivedi (2005) *Microeconometrics: Methods and Applications* (Cambridge University Press)

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

Douidich, M., A. Ezzrari, R. van der Weide, and P. Verme (2013) 'Estimating quarterly poverty rates using labor force surveys: A primer.' World Bank Policy Research Working Paper 6466, World Bank

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

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

Hansen, L.P. (1982) 'Large sample properties of generalized method of moments estimators.' *Econometrica* 50(4), 1029–1054

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

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

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

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

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

A Random-effects model

It is not uncommon that errors between subjects are correlated , which calls for a relaxation of Assumption 1. A popular model that accommodates correlated errors is the random-effects model. To allow for random effects in each cluster, we replace the index for each subject i by the combination of "cluster" c and "household" h. The disturbance term u, therefore, has the following structure:

$$u_{ch} = \eta_c + e_{ch},$$

where η_c and e_{ch} are, respectively, the cluster-specific and household-specific random-effects terms that satisfy $E[\eta_c] = E[e_{ch}] = 0$. We define $\sigma_\eta^2 \equiv \text{var}[\eta_c]$ and $\sigma_e^2 \equiv \text{var}[e_{ch}]$ and denote the size of cluster c in the sample by $k_c (< \infty)$ (i.e., there are k_c households in cluster c). Therefore, the total number of households is equal to $N = \sum_c k_c$.

In what follows, we make the following assumption instead of Assumptions 1 and 2:

Assumption 33 *The variables (x_{ch}, e_{ch}) are independently and identically distributed across ch. Further, η_c and k_c are independently and identically distributed across c. The variables x_{ch}, e_{ch}, η_c, and k_c are independent with each other.*

Under Assumption 33, the variance of the household disturbance term u_{ch} is $\text{var}[u_{ch}] = \sigma_\eta^2 + \sigma_e^2 (= \sigma_u^2)$. Note that u_{ch} is not independent across households because of the cluster-specific random-effects term η_c. The correlation of u_{ch} in a given cluster c is given by: $\gamma_1 = \sigma_\eta^2 / \sigma_u^2$.

While the OLS estimator of θ is still consistent under Assumption 33, it is in general inefficient. Therefore, we consider the case when θ is estimated by a (feasible) Generalized-Least-Squares (GLS) estimator $\hat{\theta}^{GLS}$. To facilitate the discussion, we further make two additional assumptions:

Assumption 34 *The number of clusters t tends to infinity as the total number of households N tends to infinity.*

We also assume that consistent estimators of the variance parameters are available.

Assumption 35 *For each t, we can compute $\hat{\sigma}_\eta^2$ and $\hat{\sigma}_e^2$ with the sample. Further, the following asymptotic properties are satisfied as $t \to \infty$:*

$$\begin{aligned} \hat{\sigma}_\varepsilon^2 &\xrightarrow{p} \sigma_\varepsilon^2 \\ \hat{\sigma}_u^2 &\xrightarrow{p} \sigma_u^2. \end{aligned}$$

In the discussion below, it is convenient to introduce matrix notation. Plugging $\tilde{u} \equiv u/\sigma_u$ into eq. (7), and writing the result in matrix form, we obtain the following equation:

$$\mathbf{y} = \mathbf{X}\beta + \sigma_u \tilde{\mathbf{U}} \tag{13}$$

where $\mathbf{y}$, $\mathbf{X}$, and $\tilde{\mathbf{U}}$ are, respectively, the matrix version of y_{ch}, x_{ch}, and $\tilde{u}_{ch}$ formed by stacking all the households in the sample vertically. For example, $\mathbf{y}$ is defined as follows:

$$\mathbf{y} \equiv (y_{11}, \cdots, y_{1k_1}, y_{21}, \cdots, y_{2k_2}, \cdots, y_{t1} \cdots, y_{tk_t})^T.$$

Now, define $H = \sigma_u^2 E[\tilde{\mathbf{U}}\tilde{\mathbf{U}}^T]$. Due to the cluster-specific random effects, H is a block-diagonal matrix with the number of blocks equal to t. The c-th block solves: $B_c \equiv \sigma_\eta^2 1_{k_c} 1_{k_c}^T + \sigma_e^2 I_{k_c}$, where 1_k and I_k are a column k-vector of ones and a $k \times k$-identity matrix, respectively. With these notations, we obtain the following lemma.

Lemma 36 *Let $\hat{\beta}^{GLS} \equiv (X^T \hat{H}^{-1} X)^{-1} X^T \hat{H}^{-1} y$ denote the feasible GLS estimator for β, where $\hat{H}$ is a consistent estimator for H. Furthermore, for $\hat{u} \equiv y - X\hat{\beta}^{GLS}$, let $\hat{\sigma}_u^{GLS} \equiv \left[\hat{u}^T \hat{u}/(N-1)\right]^{1/2}$ be the corresponding estimator for σ_u. Further, define $\hat{\theta}^{GLS} \equiv ((\hat{\beta}_1^{GLS} - z)/\hat{\sigma}^{GLS}, \hat{\beta}_2^{GLS}/\hat{\sigma}^{GLS}, \ldots, \hat{\beta}_K^{GLS}/\hat{\sigma}^{GLS})$, $\Sigma_{xx} \equiv E[x_{ch} x_{ch}^T]$, $\gamma_c \equiv k_c \sigma_\eta^2/(k_c \sigma_\eta^2 + \sigma_e^2)$, and $\gamma_s = (1-\gamma_c)(k_c - 1)/k_c$. Then, under Assumptions 33, 34, and 35, $\hat{\theta}^{GLS}$ satisfies the following properties as $t \to \infty$:*

$$\begin{cases} \hat{\theta}^{GLS} \xrightarrow{p} \theta \\ \sqrt{N}(\hat{\theta}^{GLS} - \theta) \xrightarrow{d} \mathcal{N}\left(0, (1-\gamma_1)\left[(1 - E[\gamma_c])\Sigma_{xx} + E[\gamma_s]\text{var}[x_{ch}]\right]^{-1}\right), \end{cases}$$

where the expectations $E[\gamma_c]$ and $E[\gamma_s]$ are taken over k_c.

Let $\hat{\mu}^{GLS}$ denote the estimator for μ that is based on $\hat{\theta}^{GLS}$. Because Lemma 36 shows that $\hat{\theta}^{GLS}$ satisfies Assumption 3, the asymptotic variance of $\hat{\mu}^{GLS}$ has the same form as in the model under Assumption 1. The only difference lies in the fact that the asymptotic variance matrix Ω^{-1} must be adjusted to account for clustering.

B Proofs

Proof of Theorem 4 By an exact first-order Taylor expansion of $\hat{\theta}$ around θ, the Law of Large Numbers, and Assumption 3, we have:

$$\hat{\mu}(\hat{\theta}) = \frac{1}{N}\sum_i g(x_i, \hat{\theta}) + \frac{1}{N}\sum_i \frac{\partial g(x_i, \theta^+)}{\partial \theta^T}(\hat{\theta} - \theta) \xrightarrow{p} \mu, \tag{14}$$

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

By the Central Limit Theorem and Assumptions 2 and 3, we have:

$$\begin{aligned}\sqrt{N}(\hat{\mu} - \mu) &= \frac{1}{\sqrt{N}}\sum_i g(x_i, \hat{\theta}) - \mu + \frac{1}{N}\sum_i \frac{\partial g(x_i, \theta^+)}{\partial \theta^T}\sqrt{N}(\hat{\theta} - \theta) \\ &\xrightarrow{p} \mathcal{N}(0, V_g + M_g^T \Omega^{-1} M_g),\end{aligned}$$

which completes the proof. □

Proof of Theorem 7 Because this theorem is well known, we only provide a sketch of proof that is relevant to the rest of the paper. Taking the first order condition of the minimization problem in eq. (2) and dividing by two, we have:

$$\left[\frac{1}{N}\sum_i \frac{\partial m^T(\hat{\theta}, x_i, y_i)}{\partial \theta}\right] \hat{W}_N \left[\frac{1}{N}\sum_i m(\hat{\theta}, x_i, y_i)\right] = 0.$$

Taking an exact first-order Taylor expansion around $\hat{\theta} = \theta$, there exist θ^{++} between θ and $\hat{\theta}$ such that:

$$\left[\frac{1}{N}\sum_i \frac{\partial m^T(\hat{\theta}, x_i, y_i)}{\partial \theta}\right] \hat{W}_N \left[\frac{1}{N}\sum_i m(\theta, x_i, y_i) + \frac{1}{N}\sum_i \frac{\partial m(\theta^{++}, x_i, y_i)}{\partial \theta^T}(\hat{\theta} - \theta)\right] = 0.$$

Solving for θ and applying the Law of Large Numbers, we obtain:

$$\begin{aligned}\hat{\theta} &= \theta - \left[\left[\frac{1}{N}\sum_i \frac{\partial m^T(\hat{\theta}, x_i, y_i)}{\partial \theta}\right] \hat{W}_N \left[\frac{1}{N}\sum_i \frac{\partial m(\theta^{++}, x_i, y_i)}{\partial \theta^T}\right]\right]^{-1} \times \\ &\quad \left[\frac{1}{N}\sum_i \frac{\partial m^T(\hat{\theta}, x_i, y_i)}{\partial \theta}\right] \hat{W}_N \left[\frac{1}{N}\sum_i m(\theta, x_i, y_i)\right] \qquad (15) \\ &\xrightarrow{p} \theta.\end{aligned}$$

This in turn implies $\theta^{++} \xrightarrow{p} \theta$. By this, $\hat{W}_N \xrightarrow{p} V_m^{-1}$, and the Central Limit Theorem, we have:

$$\begin{aligned}\sqrt{N}(\hat{\theta} - \theta) &= -\left[\left[\frac{1}{N}\sum_i \frac{\partial m^T(\hat{\theta}, x_i, y_i)}{\partial \theta}\right] \hat{W}_N \left[\frac{1}{N}\sum_i \frac{\partial m(\theta^{++}, x_i, y_i)}{\partial \theta^T}\right]\right]^{-1} \times \\ &\quad \left[\frac{1}{N}\sum_i \frac{\partial m^T(\hat{\theta}, x_i, y_i)}{\partial \theta}\right] \hat{W}_N \left[\frac{1}{\sqrt{N}}\sum_i m(\theta, x_i, y_i)\right] \qquad (16) \\ &\xrightarrow{d} \mathcal{N}(0, [M_m^T V_m^{-1} M_m]^{-1}).\end{aligned}$$

□

Proof of Lemma 9 By eq. (3), we obtain eq. (4) since:

$$\text{avar}[\bar{Y} - \hat{\mu}] = \text{avar}[\bar{Y}] + \text{avar}[\hat{\mu}] - 2\text{acov}[\bar{Y}, \hat{\mu}] = \text{avar}[\bar{Y}] - \text{avar}[\hat{\mu}] \geq 0.$$

The equation above is satisfied with equality if and only if $\text{avar}[\bar{Y} - \hat{\mu}] = 0$, which in turn is equivalent to $\hat{\mu} = \bar{Y}$ almost surely because $\hat{\mu}$ and $\bar{Y}$ are both consistent estimator of μ. It is clear that $\check{\mu}$ is consistent and its asymptotic variance is:

$$\text{avar}[\check{\mu}] = \alpha^2 \text{avar}[\bar{Y}] + (1-\alpha)^2 \text{avar}[\hat{\mu}] + 2\alpha(1-\alpha)\text{acov}[\bar{Y}, \hat{\mu}].$$

If $\hat{\mu} = \bar{Y}$ holds almost surely, the asymptotic variance of $\check{\mu}$ is constant and therefore trivially minimized when $\alpha = 0$ (or any other value of α). Thus, we assume below that $\hat{\mu} \neq \bar{Y}$ almost surely.

By taking the first-order condition with respect to α, it can be seen that $\text{avar}[\check{\mu}]$ is minimized

when:

$$\alpha = \frac{\text{avar}[\hat{\mu}] - \text{acov}[\bar{Y}, \hat{\mu}]}{\text{avar}[\bar{Y}] + \text{avar}[\hat{\mu}] - 2\text{acov}[\bar{Y}, \hat{\mu}]}.$$

Hence, if eq. (3) holds, avar$[\check{\mu}]$ is minimized when $\alpha = 0$, in which case we have $\check{\mu} = \hat{\mu}$. □

Proof of Theorem 10 By Theorems 4 and 7, we have

$$\text{avar}[\hat{\mu}] = V_g + M_g^T [M_m^T V_m^{-1} M_m]^{-1} M_g. \tag{17}$$

Let us define $g_i \equiv g(x_i, \theta)$, $g_i^+ \equiv g(x_i, \theta^+)$, $m_i^{++} \equiv m(\theta^{++}, x_i, Y_i)$, and $\hat{m}_i \equiv m(\theta^{++}, x_i, Y_i)$. Then, using eqs. (14) and (15), we obtain:

$$\begin{aligned}
& N(\bar{Y} - \mu)(\hat{\mu} - \mu) \\
= \; & N \left(\frac{1}{N} \sum_i (g_i - \mu) + \epsilon_i \right) \left(\left[\frac{1}{N} \sum_i (g_i - \mu) \right] - \left[\frac{1}{N} \sum_i \frac{\partial g_i^+}{\partial \theta^T} \right] \times \right. \\
& \left. \left[\left[\frac{1}{N} \sum_i \frac{\partial \hat{m}_i^T}{\partial \theta} \right] \hat{W}_N \left[\frac{1}{N} \sum_i \frac{\partial m_i^{++}}{\partial \theta^T} \right] \right]^{-1} \left[\frac{1}{N} \sum_i \frac{\partial \hat{m}_i^T}{\partial \theta} \right] \hat{W}_N \left[\frac{1}{N} \sum_i m_i \right] \right) \\
\xrightarrow{p} \; & V_g - M_g^T [M_m^T V_m^{-1} M_m]^{-1} M_m^T V^{-1} \text{cov}[m_i, \epsilon_i] (= \text{acov}[\bar{Y}, \hat{\mu}]).
\end{aligned} \tag{18}$$

By the law of total covariance and $E_u[m_i | x_i] = 0$, we have:

$$\begin{aligned}
\text{cov}[m_i, \epsilon_i] &= \text{cov}_x[E_u[m_i|x_i], E_u[\epsilon_i|x_i]] + E_x[\text{cov}_u[m_i, \epsilon_i|x_i]] \\
&= E_x[\text{cov}_u[m_i, Y_i|x_i]] \\
&= E_x[E_u[m_i Y_i|x_i] - E_u[Y_i|x_i] E_u[m_i|x_i]] \\
&= E[m_i Y_i].
\end{aligned} \tag{19}$$

Therefore, by eqs. (17), (18), and (19), eq. (3) holds if and only if we have:

$$M_g^T [M_m^T V_m^{-1} M_m]^{-1} M_g = -M_g^T [M_m^T V_m^{-1} M_m]^{-1} M_m^T V_m^{-1} E[m_i Y_i].$$

For this to hold regardless of the distribution of x, it is necessary and sufficient to have eq. (5). In the exactly-identified case, premultiplying eq. (5) by $M_m^{-T} V_m$, we obtain eq. (6). □

Proof of Corollary 12 Notice that this is an exactly-identified case. Solving the first order condition, we obtain $N^{-1}\sum_i m_i = 0$. Hence, Assumption 6 holds. Further, we have:

$$E[m_i Y_i] = E_x\left[E_u\left[Y_i(Y_i - g_i)\,\frac{\partial g_i}{\partial \theta}\,\middle|\, x_i\right]\right] = \text{var}[\epsilon_i^2]E_x\left[\frac{\partial g_i}{\partial \theta}\right] = \text{var}[\epsilon_i^2]M_g. \tag{20}$$

By the definitions of M_m and V_m, we have:

$$M_m = E_x\left[E_u\left[-\frac{\partial g_i}{\partial \theta}\frac{\partial g_i}{\partial \theta^T} + (Y_i - g_i)\frac{\partial g_i}{\partial \theta \partial \theta^T}\right]\right] = -E_x\left[\frac{\partial g_i}{\partial \theta}\frac{\partial g_i}{\partial \theta^T}\right] \tag{21}$$

$$V_m = E_x[\text{var}_u[m_i]] = E_x\left[\frac{\partial g_i}{\partial \theta}\frac{\partial g_i}{\partial \theta^T}\right]\text{var}[\epsilon^2]. \tag{22}$$

By these, we can verify that eq. (6) holds. Therefore, eq. (3) holds by Theorem 10. □

Proof of Corollary 15 Notice that this is also an exactly-identified case. Assumption 6 trivially holds because $N^{-1}\sum_i m_i = 0$ by Assumption 13. Further, under the standard ML regularity conditions, we have:

$$V_m = E[m_i m_i^T] = E\left[\frac{\partial l_i}{\partial \theta}\frac{\partial l_i}{\partial \theta^T}\right] = -E\left[\frac{\partial^2 l_i}{\partial \theta \partial \theta^T}\right] = -E\left[\frac{\partial^2 m}{\partial \theta^T}\right] = -M_m.$$

Using this and noting M_m is a symmetric matrix, eq. (6) reduces to $E[m_i Y_i] = M_g$. This holds because:

$$\begin{aligned} E[m_i Y_i] &= E\left[\frac{f_\epsilon'(Y_i - g(x_i,\theta)|x_i)}{f_\epsilon(Y_i - g(x_i,\theta)|x_i)}\cdot\left(-\frac{\partial g(x_i,\theta)}{\partial \theta}\right)\cdot(g(x_i,\theta)+\epsilon_i)\right] \\ &= E_x\left[-E_u\left[\frac{f_\epsilon'(\epsilon_i|x_i)}{f_\epsilon(\epsilon_i|x_i)}\right]\frac{\partial g(x_i,\theta)}{\partial \theta}g(x_i,\theta) - E_u\left[\frac{f_\epsilon'(\epsilon_i|x_i)}{f_\epsilon(\epsilon_i|x_i)}\epsilon_i\,\middle|\, x_i\right]\frac{\partial g(x_i,\theta)}{\partial \theta}\right] \\ &= E_x\left[\frac{\partial g(x_i,\theta)}{\partial \theta}\right] = M_g, \end{aligned}$$

where we have used:

$$\begin{aligned} E_u\left[\frac{f_\epsilon'(\epsilon_i|x_i)}{f_\epsilon(\epsilon_i|x_i)}\right] &= \int_{-\infty}^{\infty}\frac{f_\epsilon'(\epsilon_i|x_i)}{f_\epsilon(\epsilon_i|x_i)}f_\epsilon(\epsilon_i|x_i)\mathrm{d}\epsilon_i = [f_\epsilon(\epsilon_i|x_i)]_{\epsilon_i=-\infty}^{\epsilon_i=\infty} = 0 \\ E_u\left[\frac{f_\epsilon'(\epsilon_i|x_i)}{f_\epsilon(\epsilon_i|x_i)}\epsilon_i\right] &= \int_{-\infty}^{\infty} f_\epsilon'(\epsilon_i|x_i)\epsilon_i\mathrm{d}\epsilon_i = [f_\epsilon(\epsilon_i|x_i)\epsilon_i - F_\epsilon(\epsilon_i|x_i)]_{\epsilon_i=-\infty}^{\epsilon_i=\infty} = -1. \end{aligned}$$

By Theorem 10, eq. (3) holds. □

Proof of Theorem 19 First, notice that this is also an exactly identified case. Note also that the following relationship holds:

$$y_i > z \quad \Leftrightarrow \quad x_i\beta - z + u_i > 0 \quad \Leftrightarrow \quad \tilde{u}_i > -x_i^T\tilde{\beta} = -x_i^T\theta \equiv B_i.$$

Thus, we have $g(x_i, \theta) = E[\text{Ind}(\tilde{u}_i > B_i)] = 1 - \tilde{F}(B_i)$. The individual log-likelihood function is: $l_i = Y_i \ln[1 - \tilde{F}(B_i)] + (1 - Y_i) \ln \tilde{F}(B_i)$. Hence, solving the first order condition, we obtain:

$$m_i = \frac{\partial l_i}{\partial \theta} = \left[\frac{-Y_i}{1 - \tilde{F}(B_i)} + \frac{1 - Y_i}{\tilde{F}(B_i)}\right] \frac{\partial B_i}{\partial \theta} = \frac{\tilde{F}(B_i) + Y_i - 1}{\tilde{F}(B_i)(1 - \tilde{F}(B_i))} \tilde{f}(B_i) x_i. \tag{23}$$

Because Y_i is binary, $Y_i^2 = Y_i$. Using this, we have:

$$E_u[Y_i m_i | x_i] = \frac{E_u[Y_i \tilde{F}(B_i) + Y_i^2 - Y_i | x_i]}{\tilde{F}(B_i)(1 - \tilde{F}(B_i))} \tilde{f}(B_i) x_i = \frac{E_u[Y_i | x_i]}{1 - \tilde{F}(B_i)} \tilde{f}(B_i) x_i = \tilde{f}(B_i) x_i.$$

Taking expectations over x yields $E[Y_i m_i] = M_g$. Furthermore, we have:

$$\begin{aligned} M_m &= -E\left[\left(\frac{Y_i}{(1 - \tilde{F}(B_i))^2} + \frac{1 - Y_i}{\tilde{F}^2(B_i)}\right) \tilde{f}^2(B_i) \frac{\partial B_i}{\partial \theta} \frac{\partial B_i}{\partial \theta^T}\right] \\ &\quad + E\left[\left(\frac{-Y_i}{1 - \tilde{F}(B_i)} + \frac{1 - Y_i}{\tilde{F}(B_i)}\right) \tilde{f}'(B_i) x_i \frac{\partial B_i}{\partial \theta^T}\right] \\ &= -V_m. \end{aligned}$$

Because V_m is symmetric, eq. (6) holds. By Theorem 10, eq. (3) holds. □

Proof of Theorem 20 By eq. (23) and Assumption 13, $\hat{\theta}$ satisfies:

$$\frac{1}{N} \sum_i \left(\tilde{F}(-x_i^T\hat{\theta}) + Y_i - 1\right) \left(\frac{\tilde{f}(-x_i^T\hat{\theta})}{\tilde{F}(-x_i^T\hat{\theta})(1 - \tilde{F}(-x_i^T\hat{\theta}))}\right) x_i = 0. \tag{24}$$

Because eq. (24) holds for each component of x_i, it also holds for the first component, or the constant term. Thus, letting $\hat{B}_i \equiv -x_i^T\hat{\theta}$, the following equation must hold to have $\hat{\mu} = \bar{Y}$

almost surely:

$$\frac{1}{N}\sum_i \left(\tilde{F}(-x_i^T\hat{\theta}) + Y_i - 1\right)\left(\frac{\tilde{f}(\hat{B}_i)}{\tilde{F}(\hat{B}_i)(1-\tilde{F}(\hat{B}_i))}\right) = 0. \tag{25}$$

For this to hold for any distribution of x_i, the following equation must hold for all i:

$$\frac{\tilde{f}(\hat{B}_i)}{\tilde{F}(\hat{B}_i)(1-\tilde{F}(\hat{B}_i))} = 1. \tag{26}$$

Rearranging terms, we have:

$$\hat{B}_i = \int \left[\frac{1}{\tilde{F}(\hat{B}_i)} + \frac{1}{1-\tilde{F}(\hat{B}_i)}\right] \mathrm{d}\tilde{F} = \ln \tilde{F}(\hat{B}_i) - \ln(1-\tilde{F}(\hat{B}_i)) + c,$$

where c is a constant. Solving for $\tilde{F}(\hat{B}_i)$, we obtain:

$$\tilde{F}(\hat{B}_i) = [1 + \mathrm{e}^{-(\hat{B}_i - c)}]^{-1}. \tag{27}$$

The mean of $\hat{u}_i$ in this case is:

$$E[\tilde{u}_i] = \int_{-\infty}^{\infty} \tilde{u}\tilde{f}(\tilde{u})\mathrm{d}\tilde{u} = \int_0^1 (\ln \tilde{F} - \ln(1-\tilde{F}) + c)\mathrm{d}\tilde{F} = c. \tag{28}$$

Because $E[\tilde{u}_i] = 0$, we must have $c = 0$. Plugging this into eq. (27), we have eq. (8). Conversely, if eq. (8) holds, eq. (26) holds. Hence, eq. (24) reduces to $\bar{Y} = \hat{\mu}$. □

Proof of Theorem 22 To distinguish the moments for different prediction estimators, we use the superscripts OLS, MLB, and MLC. Letting $C_i \equiv (y_i - x_i^T\beta)/\sigma_u$, and solving the first

order conditions yields:

$$\begin{aligned} m_i^{OLS} &= \begin{bmatrix} -(Y_i - x_i^T\beta)x_i \\ \sigma_u^2 - (y_i - x_i^T\beta)^2 \end{bmatrix} \\ m_i^{MLB} &= \frac{\tilde{F}(B_i) + Y_i - 1}{\tilde{F}(B_i)(1-\tilde{F}(B_i))}\tilde{f}(B_i)x_i \\ m_i^{MLC} &= -\frac{\tilde{f}'(C_i)}{\tilde{f}(C_i)}\begin{bmatrix} x_i/\sigma_u \\ C_i/\sigma_u \end{bmatrix} - \begin{bmatrix} 0_{J\times 1} \\ 1/\sigma_u \end{bmatrix} \end{aligned}$$

To calculate the relevant moments below, the following formulae are useful:

$$\begin{aligned} E_u\left[\frac{\tilde{f}(\tilde{u})\tilde{f}''(\tilde{u})}{\tilde{f}^2(\tilde{u})}\right] &= \int_{-\infty}^{\infty}\frac{\tilde{f}(\tilde{u})\tilde{f}''(\tilde{u})}{\tilde{f}^2(\tilde{u})}\tilde{f}(\tilde{u})\mathrm{d}\tilde{u} = \int_{-\infty}^{\infty}\tilde{f}''(\tilde{u})\mathrm{d}\tilde{u} = \left[\tilde{f}'(\tilde{u})\right]_{\tilde{u}=-\infty}^{\tilde{u}=\infty} = 0 \\ E_u\left[\frac{\tilde{f}(\tilde{u})\tilde{f}''(\tilde{u})}{\tilde{f}^2(\tilde{u})}\tilde{u}\right] &= \int_{-\infty}^{\infty}\tilde{f}''(\tilde{u})\tilde{u}\mathrm{d}\tilde{u} = \left[\tilde{f}'(\tilde{u})u - f(\tilde{u})\right]_{\tilde{u}=-\infty}^{\tilde{u}=\infty} = 0 \\ E_u\left[\frac{\tilde{f}(\tilde{u})\tilde{f}''(\tilde{u})}{\tilde{f}^2(\tilde{u})}\tilde{u}^2\right] &= \int_{-\infty}^{\infty}\tilde{f}''(\tilde{u})\tilde{u}^2\mathrm{d}\tilde{u} = \left[\tilde{f}'(\tilde{u})\tilde{u}^2 - 2\tilde{u}\tilde{f}(\tilde{u}) + 2\tilde{F}(\tilde{u})\right]_{\tilde{u}=-\infty}^{\tilde{u}=\infty} = 2 \\ E_u\left[\frac{\tilde{f}'(\tilde{u})}{\tilde{f}(u)}\right] &= \int_{-\infty}^{\infty}\frac{\tilde{f}'(\tilde{u})}{\tilde{f}(\tilde{u})}\tilde{f}(\tilde{u})\mathrm{d}\tilde{u} = \int_{\infty}^{\infty}\tilde{f}'(\tilde{u})\mathrm{d}\tilde{u} = \left[\tilde{f}(\tilde{u})\right]_{\tilde{u}=-\infty}^{\tilde{u}=\infty} = 0 \\ E_u\left[\frac{\tilde{f}'(\tilde{u})}{\tilde{f}(\tilde{u})}u\right] &= \int_{-\infty}^{\infty}\tilde{f}'(\tilde{u})\tilde{u}\mathrm{d}\tilde{u} = \left[\tilde{f}(\tilde{u})\tilde{u} - F(\tilde{u})\right]_{\tilde{u}=-\infty}^{\tilde{u}=\infty} = -1 \\ E_u\left[\frac{Y_i\tilde{f}'(\tilde{u})}{\tilde{f}(\tilde{u})}\right] &= \int_{B_i}^{\infty}\tilde{f}'(\tilde{u})\mathrm{d}\tilde{u} = \left[\tilde{f}(\tilde{u})\right]_{\tilde{u}=B_i}^{\tilde{u}=\infty} = \tilde{f}(B_i) \\ E_u\left[\frac{Y_i\tilde{f}'(\tilde{u})}{\tilde{f}(\tilde{u})}\tilde{u}\right] &= \int_{B_i}^{\infty}\tilde{f}'(\tilde{u})\tilde{u}\mathrm{d}\tilde{u} = \left[\tilde{f}(\tilde{u})\tilde{u} - F(\tilde{u})\right]_{\tilde{u}=B_i}^{\tilde{u}=\infty} = \tilde{F}(B_i) - B_i\tilde{f}(B_i) - 1. \end{aligned}$$

By Assumption 21, we also have:

$$E_u\left[\left(\frac{\tilde{f}'(\tilde{u})}{\tilde{f}(\tilde{u})}\right)^2\tilde{u}\right] = 0.$$

Using these and $\tilde{u}_i = C_i$, we have:

$$
\begin{aligned}
M_m^{OLS} &= \begin{bmatrix} \Sigma_{xx} & 0 \\ 0 & 2\sigma_u \end{bmatrix} \\
V_m^{OLS} &= E\begin{bmatrix} (y_i - x_i^T)^2 x_i x_i^T & -(\sigma_u^2 - (y_i - x_i^T\beta)^2)(y_i - x_i^T\beta)x_i \\ -(\sigma_u^2 - (y_i - x_i^T\beta)^2)(y_i - x_i^T\beta)x_i^T & (\sigma_u^2 - (y_i - x_i^T\beta)^2)^2 \end{bmatrix} \\
&= \begin{bmatrix} \sigma_u^2\Sigma_{xx} & 0_{1\times J} \\ 0_{J\times 1} & \nu_u - \sigma_u^2 \end{bmatrix} \\
M_m^{MLB} &= -E\left[\left[\frac{\tilde{F}(1-\tilde{F}) - (\tilde{F} + Y_i - 1)(1-2\tilde{F})}{\tilde{F}^2(1-\tilde{F})^2} f^2 + \frac{\tilde{F} + Y_i - 1}{\tilde{F}(1-\tilde{F})} f'\right] x_i x_i^T\right] \\
&= -E_x\left[\frac{\tilde{f}^2(B_i)}{\tilde{F}(B_i)(1-\tilde{F}(B_i))} x_i x_i^T\right] = -V_m^{MLB} \\
M_m^{MLC} &= E\left[\frac{\tilde{f}\tilde{f}'' - (\tilde{f}')^2}{\tilde{f}^2}\begin{bmatrix} \frac{x_i x_i^T}{\sigma_u^2} & \frac{\tilde{u}_i x_i}{\sigma_u^2} \\ \frac{\tilde{u}_i x_i^T}{\sigma_u^2} & \frac{\tilde{u}_i^2}{\sigma_u^2} \end{bmatrix} + \frac{\tilde{f}'}{\tilde{f}}\begin{bmatrix} 0_{J\times J} & \frac{x_i}{\sigma_u^2} \\ \frac{x_i^T}{\sigma_u^2} & \frac{2u_i}{\sigma_u^2} \end{bmatrix} + \begin{bmatrix} 0_{J\times J} & 0_{J\times 1} \\ 0_{1\times J} & \frac{1}{\sigma_u^2} \end{bmatrix}\right] \\
&= \frac{1}{\sigma_u^2}\begin{bmatrix} -E_u\left[\left(\frac{\tilde{f}'(\tilde{u}_i)}{\tilde{f}(\tilde{u}_i)}\right)^2\right] E_x\left[x_i x_i^T\right] & 0_{J\times 1} \\ 0_{1\times J} & E_u\left[2 - \left(\frac{\tilde{f}'(\tilde{u}_i)\tilde{u}_i}{\tilde{f}(\tilde{u}_i)}\right)^2 - 2 + 1\right] \end{bmatrix} \\
&= \frac{1}{\sigma_u^2}\begin{bmatrix} -E_u\left[\left(\frac{\tilde{f}'(\tilde{u}_i)}{\tilde{f}(\tilde{u}_i)}\right)^2\right] E_x\left[x_i x_i^T\right] & 0_{J\times 1} \\ 0_{1\times J} & -E_u\left[\left(\frac{\tilde{f}'(\tilde{u}_i)\tilde{u}_i}{\tilde{f}(\tilde{u}_i)}\right)^2 - 1\right] \end{bmatrix} = -V_m^{MLC}.
\end{aligned}
$$

Also, note the following:

$$g^{OLS}(x_i^T, \theta) = g^{MLC}(x_i^T, \theta) = 1 - \tilde{F}(B_i) = g^{MLB}(x_i^T, \theta).$$

Therefore, we have:

$$
\begin{aligned}
M_g^{OLS} &= \begin{bmatrix} \Sigma_{fx}/\sigma_u \\ \Sigma_{fB}/\sigma_u \end{bmatrix} \\
M_g^{MLB} &= \Sigma_{fx} \\
M_g^{MLC} &= \begin{bmatrix} \Sigma_{fx}/\sigma_u \\ 0 \end{bmatrix}.
\end{aligned}
$$

Using these, we obtain:

$$\begin{aligned}
\text{avar}[\hat{\mu}^{OLS}] &= V_g + \Sigma_{fx}^T \Sigma_{xx}^{-1} \Sigma_{fx} + \frac{(\nu_u - \sigma_u^4)\Sigma_{fB}^2}{4\sigma_u^2} \\
\text{avar}[\hat{\mu}^{MLB}] &= V_g + \Sigma_{fx}^T E_x^{-1} \left[\frac{\tilde{f}^2(B_i)}{\tilde{F}(B_i)(1-\tilde{F}(B_i))} x_i x_i^T \right] \Sigma_{fx} \\
\text{avar}[\hat{\mu}^{MLC}] &= V_g + E_u^{-1} \left[\left(\frac{\tilde{f}'(\tilde{u}_i)}{\tilde{f}(\tilde{u}_i)} \right)^2 \right] \Sigma_{fx}^T \Sigma_{xx}^{-1} \Sigma_{fx}.
\end{aligned}$$

For the asymptotic covariance between $\hat{\mu}^a$ and $\hat{\mu}^b$ with $a, b \in \mathcal{E}$, the following follows from eq. (16):

$$N(\hat{\mu}^a - \mu)(\hat{\mu}^b - \mu) \xrightarrow{p} V_g + P^a M_m^{a,b} (P^b)^T (= \text{acov}[\hat{\mu}^a, \hat{\mu}^b]), \tag{29}$$

where $P \equiv M_g^T [M_m^T V_m^{-1} M_m]^{-1} M_m V_m^{-1}$ for each of the superscripts a and b, and $M_m^{a,b} \equiv E[m_i^a (m_i^b)^T]$.

$$\begin{aligned}
P^{OLS} &= \frac{1}{\sigma_u} \left[\Sigma_{fx}^T \Sigma_{xx}^{-1} \quad \frac{\Sigma_{fB}}{2\sigma_u^3} \right] \\
P^{MLB} &= -\Sigma_{fx}^T E_x^{-1} \left[\frac{\tilde{f}^2(B_i)}{\tilde{F}(B_i)(1-\tilde{F}(B_i))} x_i x_i^T \right] \\
P^{MLC} &= -\sigma_u \Sigma_{fx}^T E_u^{-1} \left[\left(\frac{\tilde{f}'(\tilde{u}_i)}{\tilde{f}(\tilde{u}_i)} \right)^2 \right] \left[\Sigma_{xx}^{-1} \quad 0_{J\times 1} \right].
\end{aligned}$$

Further, we have:

$$
\begin{aligned}
M_m^{OLS,MLC} &= E\left[\begin{array}{cc} \frac{\tilde{f}'(\tilde{u}_i)\tilde{u}_i}{\tilde{f}(\tilde{u}_i)}\frac{x_i x_i^T}{\sigma_u} & \left(\frac{\tilde{f}'(\tilde{u}_i)\tilde{u}_i^2}{\tilde{f}(\tilde{u}_i)\sigma_u}-\tilde{u}_i\right)\frac{x_i}{\sigma_u} \\ \frac{-(\sigma_u^2-\tilde{u}_i^2)\tilde{f}'(\tilde{u}_i)x_i^T}{\tilde{f}(\tilde{u}_i)} & -\frac{\sigma_u^2-\tilde{u}_i^2}{\sigma_u}-\frac{(\sigma_u^2-\tilde{u}_i^2)u_i\tilde{f}'(\tilde{u}_i)}{\sigma_u^2\tilde{f}(\tilde{u}_i)} \end{array}\right] \\
&= \left[\begin{array}{cc} -\Sigma_{xx} & 0_{J\times 1} \\ 0_{1\times J} & -4 \end{array}\right] \\
M_m^{MLB,MLC} &= -E\left[\frac{(F(B_i)+Y_i-1)f(B_i)}{F(B_i)(1-F(B_i))}\frac{f'(\tilde{u}_i)}{f(\tilde{u}_i)}\left[\frac{x_i x_i^T}{\sigma_u}, \quad \frac{\tilde{u}_i x_i}{\sigma_u}\right]\right] \\
&\quad -E\left[0_{1\times J}, \quad \frac{m_i^{MLB}}{\sigma_u}\right] \\
&= E_x\left[\frac{\tilde{f}(B_i)}{F(B_i)(1-\tilde{F}(B_i))}\left[\tilde{f}(B_i)\frac{x_i x_i^T}{\sigma_u}, \quad (\tilde{F}(B_i)-B_i\tilde{f}(B_i))-1)\frac{x_i}{\sigma_u}\right]\right] \\
&\quad -E_x\left[\frac{\tilde{f}(B_i)}{\tilde{F}(B_i)}\left[0_{J\times J}, \quad -\frac{x_i}{\sigma_u}\right]\right] \\
&= \frac{1}{\sigma_u}E_x\left[\frac{\tilde{f}^2(B_i)}{\tilde{F}(B_i)(1-\tilde{F}(B_i))}\left[x_i x_i^T, \quad x_i B_i\right]\right].
\end{aligned}
$$

Plugging these moments in eq. (29), we obtain:

$$
\text{acov}[\hat{\mu}^{MLB},\hat{\mu}^{MLC}] = \text{acov}[\hat{\mu}^{OLS},\hat{\mu}^{MLC}] = V_g + E_u^{-1}\left[\left(\frac{\tilde{f}'(\tilde{u}_i)}{\tilde{f}(\tilde{u}_i)}\right)^2\right]\Sigma_{fx}^T E_x^{-1}[x_i x_i^T]\Sigma_{fx},
$$

which completes the proof. □

Proof of Corollary 23 Because $\text{avar}[\hat{\mu}^{MLC}] = \text{acov}[\hat{\mu}^{OLS},\hat{\mu}^{MLC}] = \text{acov}[\hat{\mu}^{MLB},\hat{\mu}^{MLC}]$, the same argument as Lemma 9 is applicable to the comparison between $\hat{\mu}^{OLS}$ and $\hat{\mu}^{MLC}$ and between $\hat{\mu}^{MLB}$ and $\hat{\mu}^{MLC}$. □

Proof of Theorem 27 By an exact first-order Taylor expansion of $\hat{\theta}^I$ around θ, the Law of Large Numbers, and Assumption 3, we have:

$$
\hat{\mu}^{DS} = \frac{1}{N}\sum_i g(x_i,\hat{\theta}) + \frac{1}{N}\sum_i \frac{\partial g(x_i,\theta^+)}{\partial\theta^T}(\hat{\theta}^I-\theta) \xrightarrow{p} \mu, \tag{30}
$$

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

By the Central Limit Theorem and Assumption 3, we obtain:

$$\begin{aligned}\sqrt{N}(\hat{\mu}^{DS}-\mu) &= \frac{1}{\sqrt{N}}\sum_i g(x_i,\hat{\theta})-\mu+\sqrt{\frac{N}{N^I}}\frac{1}{N}\sum_i \frac{\partial g(x_i,\theta^+)}{\partial\theta^T}\sqrt{N^I}(\hat{\theta}^I-\theta)\\ &\xrightarrow{p} \mathcal{N}(0, V_g + r^{-1}M_g^T\Omega^{-1}M_g),\end{aligned}$$

which completes the proof. □

Proof of Theorem 29 Plugging the constraint into the objective function from eq. (11), the objective function is seen to solve:

$$[V_g k + M_g^T\Omega^{-1}M_g(1-k) + V_g(1-k)\rho + M_g^T\Omega^{-1}M_g k/\rho]/C.$$

This function is convex with respect to ρ. The theorem follows from this and $\rho \in (0,1]$. □

Proof of Corollary 30 The proof is similar to the proof of Theorem 29. □

Proof of Theorem 32 It is straightforward to show that:

$$\sqrt{N^I}(\bar{Y}^I-\mu) \xrightarrow{d} \mathcal{N}(0, V_g + \text{var}[\epsilon_i]). \tag{31}$$

Because $\bar{Y}^I$ and $\tilde{\mu}^{II}$ are independent under Assumption 26, Theorem 32 follows from eq. (31) and Theorem 4. □

Proof of Lemma 36 Under Assumptions 33, 34, and 35, $\hat{\beta}^{GLS}$ is consistent and asymptotically normal distributed as $t \to \infty$:

$$\begin{aligned}\hat{\beta}^{GLS} &\xrightarrow{p} \beta\\ \sqrt{N}(\hat{\beta}^{GLS}-\beta) &\xrightarrow{d} \mathcal{N}\left(0, (X^TH^{-1}X)^{-1}\right).\end{aligned}$$

We are interested in $\hat{\theta}^{GLS} = ((\hat{\beta}_1^{GLS}-z)/\hat{\sigma}^{GLS}, \hat{\beta}_2^{GLS}/\hat{\sigma}^{GLS}, \ldots, \hat{\beta}_K^{GLS}/\hat{\sigma}^{GLS})$. The consistency of $\hat{\theta}^{GLS}$ immediately follows from the consistency of $\hat{\beta}^{GLS}$ and $\hat{\theta}^{GLS}$. What remains to be established is the asymptotic variance and normality of $\hat{\theta}^{GLS}$.

Since the constant z merely introduces a deterministic shift in the intercept it does not affect

the asymptotic variance and normality. This means that our interest is in asymptotic distribution of $\hat{\beta}^{GLS}/\hat{\sigma}^{GLS}$. An application of Slutsky's theorem yields:

$$\sqrt{N}(\hat{\theta}^{GLS} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \Omega_{GLS}^{-1}\right) \quad \text{as} \quad t \to \infty,$$

where Ω_{GLS}^{-1} satisfies:

$$(X^T H^{-1} X/N)^{-1}/\sigma_u^2 = (\sigma_u^2 X^T H^{-1} X/N)^{-1} \xrightarrow{p} \Omega_{GLS}^{-1}.$$

To obtain the analytic expression of Ω_{GLS}^{-1}, let us first evaluate the matrix inverse of H. As H is a block-diagonal matrix, its inverse too is a block-diagonal matrix with blocks B_c^{-1}. The inverse of B_c is:

$$B_c^{-1} = \frac{1}{\sigma_e^2}\left(I_{k_c} - \frac{\sigma_\eta^2}{\sigma_e^2 + k_c\sigma_\eta^2} 1_{k_c} 1_{k_c}^T\right).$$

Substituting this into $X^T H^{-1} X$ yields:

$$\begin{aligned} X^T H^{-1} X &= \sum_c X_c^T B_c^{-1} X_c \\ &= \frac{1}{\sigma_e^2}\sum_c X_c^T \left(I_{k_c} - \frac{\sigma_\eta^2}{\sigma_e^2 + k_c\sigma_\eta^2} 1_{k_c} 1_{k_c}^T\right) X_c \\ &= \frac{1}{\sigma_e^2}\left[\sum_c X_c^T X_c - \gamma_c k_c \bar{x}_c \bar{x}_c^T\right], \end{aligned}$$

where $X_c \equiv (x_{c1}^T, \cdots, x_{ck_c}^T)^T$ denotes the matrix of x_{ch} stacked for cluster c and $\bar{x}_c \equiv X_c^T 1_{k_c}/k_c$ is the cluster average of x_{ch}.

Therefore, Ω^{GLS} is given by:

$$\begin{aligned} \Omega^{GLS} &= \sigma_u^2 (X^T H^{-1} X)/N \\ &= \frac{\sigma_u^2}{\sigma_e^2}\frac{1}{N}\left[\sum_c X_c^T X_c - \gamma_c k_c \bar{x}_c \bar{x}_c^T\right] \\ &= \frac{\sigma_u^2}{\sigma_e^2}\frac{1}{N}\left[\sum_c \sum_h (1-\gamma_c) x_{ch} x_{ch}^T + \frac{(1-\gamma_c)(k_c - 1)}{k_c} svar_c[x_{ch}]\right] \\ &\xrightarrow{p} (1-\gamma_1)^{-1}\left[(1 - E[\gamma_c])\Sigma_{xx} + E[\gamma_s]\text{var}[x_{ch}]\right], \end{aligned}$$

by the independence between x_c and k_c and Law of Large Numbers, which completes the proof. □

C Further discussion on Example 17

To clearly show the violation of eq. (5), we explicitly compute each moment. Let $\theta \equiv (\beta^T, \sigma_u)^T$ and denote the probability density function and the cumulative distribution function for the standard normal distribution by ϕ and Φ, respectively. Then, we have:

$$\begin{aligned} g(x,\theta) &= E[y_i \text{Ind}(y_i > 0)|x_i] \\ &= E[(x_i^T\beta + \sigma_u\tilde{u}_i)\text{Ind}(\tilde{u}_i > -x_i^T\beta/\sigma_u)|x_i] \\ &= \int_{-x_i^T\beta/\sigma_u}^{\infty} (x_i^T\beta + \sigma_u\tilde{u}_i)\frac{\phi(\tilde{u})}{\sigma_u}\text{d}\tilde{u} \\ &= A_i(\Phi(A_i) - 1) + \phi(A_i), \end{aligned}$$

where $A_i \equiv -x_i^T\beta/\sigma_u$.

$$M_g = E_x \begin{bmatrix} \partial g_i/\partial\beta \\ \partial g_i/\partial\sigma_u \end{bmatrix} = E_x \begin{bmatrix} (1-\Phi(A_i))x/\sigma_u \\ -(1-\Phi(A_i))A_i/\sigma_u \end{bmatrix}$$

For the Tobit model, the individual log-likelihood function l_i can be written as:

$$l_i = \text{Ind}(Y_i = 0)\ln\Phi(A_i) + \text{Ind}(Y_i > 0)\ln\left(\frac{1}{\sigma_u}\phi\left(\frac{Y_i - x_i^T\beta}{\sigma_u}\right)\right).$$

Then, the moment function is:

$$m_i = \begin{bmatrix} \frac{\partial l_i}{\partial\beta} \\ \frac{\partial l_i}{\partial\sigma_u} \end{bmatrix} = \begin{bmatrix} \left[-\text{Ind}(Y_i = 0)\frac{1}{\sigma_u}\frac{\phi(A_i)}{\Phi(A_i)} + \text{Ind}(Y_i > 0)\frac{Y_i - x_i^T\beta}{\sigma_u^2}\right]x_i \\ \text{Ind}(Y_i = 0)\frac{x_i^T\beta}{\sigma_u^2}\frac{\phi(A_i)}{\Phi(A_i)} + \text{Ind}(Y_i > 0)\left[-\frac{1}{\sigma_u} + \frac{(Y_i - x_i^T\beta)^2}{\sigma_u^3}\right] \end{bmatrix}.$$

The following relationships are useful:

$$\begin{aligned}
E[Y_i > 0] &= E[\tilde{u} > A_i] = \int_{A_i}^{\infty} \phi(\tilde{u}) \mathrm{d}\tilde{u} = 1 - \Phi(A_i) \\
E[\tilde{u}|Y_i > 0] &= \int_{A_i}^{\infty} \tilde{u}\phi(\tilde{u}) \mathrm{d}\tilde{u} = -[\phi(\tilde{u})]_{\tilde{u}=A_i}^{\tilde{u}=\infty} = \phi(A_i) \\
E[\tilde{u}^2|Y_i > 0] &= \int_{A_i}^{\infty} \tilde{u}^2\phi(\tilde{u}) \mathrm{d}\tilde{u} = [\Phi(\tilde{u}) - \tilde{u}\phi(\tilde{u})]_{\tilde{u}=A_i}^{\tilde{u}=\infty} = A_i\phi(A_i) + 1 - \Phi(A_i) \\
E[\tilde{u}^3|Y_i > 0] &= \int_{A_i}^{\infty} \tilde{u}^3\phi(\tilde{u}) \mathrm{d}\tilde{u} = -[(\tilde{u}^2 + 2)\phi(\tilde{u})]_{\tilde{u}=A_i}^{\tilde{u}=\infty} = (A_i^2 + 2)\phi(A_i) \\
E[\tilde{u}^4|Y_i > 0] &= \int_{A_i}^{\infty} \tilde{u}^4\phi(\tilde{u}) \mathrm{d}\tilde{u} = [3\Phi(\tilde{u}) - \phi(\tilde{u})(\tilde{u}^3 + 3\tilde{u})]_{\tilde{u}=A_i}^{\tilde{u}=\infty} \\
&= 3(1 - \Phi(A_i)) + \phi(A_i)(A_i^3 + 3A_i)
\end{aligned}$$

Therefore, letting $C_i \equiv (y_i - x_i^T\beta)/\sigma_u (= \tilde{u})$ and taking expectations, we have:

$$\begin{aligned}
& M_m = E\left[\frac{\partial m}{\partial \theta}\right] \\
= \; & E\left[\mathrm{Ind}(Y_i = 0)\left[\frac{\zeta_i}{\sigma_u^2}\begin{bmatrix} 0_{J\times J} & x_i \\ x_i^T & 2A_i \end{bmatrix} - \frac{A_i\zeta_i + \zeta_i^2}{\sigma_u^2}\begin{bmatrix} x_i x_i^T & x_i A_i \\ A x_i^T & A_i^2 \end{bmatrix}\right]\right] \\
& + E\left[\mathrm{Ind}(Y_i > 0)\frac{1}{\sigma_u^2}\begin{bmatrix} -x_i x_i^T & -2C_i x_i \\ -2C_i & 1 - 3C_i^2 \end{bmatrix}\right] \\
= \; & \frac{1}{\sigma_u^2}E_x\begin{bmatrix} [1 - (A_i\zeta_i + \zeta_i^2 + 1)\Phi(A_i)]x_i x_i^T & -x_i\upsilon_i\phi(A_i)) \\ -\upsilon_i\phi(A_i)x_i^T & -\upsilon_i A_i\phi(A_i) + 2(1 - \Phi(A_i)) \end{bmatrix},
\end{aligned}$$

where $\zeta_i \equiv \phi(A_i)/\Phi(A_i)$ and $\upsilon_i \equiv [(A_i + \zeta_i)A_i + 1]$. Also, we have:

$$
\begin{aligned}
& V_m \\
= \; & E[m_i m_i^T] \\
= \; & E\left[\text{Ind}(Y_i = 0)\frac{\zeta_i^2}{\sigma_u^2}\begin{bmatrix} x_i x_i^T & x_i A_i \\ A_i x_i^T & A_i^2 \end{bmatrix}\right] + \\
& E\left[\text{Ind}(Y_i > 0)\frac{1}{\sigma_u^2}\begin{bmatrix} C_i^2 x_i x_i^T & (-C_i + C_i^3)x_i \\ (-C_i + C_i^3)x_i^T & (C_i^2 - 1)^2 \end{bmatrix}\right] \\
= \; & E_x\left[\frac{\zeta_i^2\Phi(A_i)}{\sigma_u^2}\begin{bmatrix} x_i x_i^T & x_i A_i \\ A_i x_i^T & A_i^2 \end{bmatrix}\right] + \\
& E_x\left[\frac{1}{\sigma_u^2}\begin{bmatrix} (A_i\phi(A_i) - \Phi(A_i) + 1)x_i x_i^T & (A_i^2 + 1)\phi(A_i)x_i \\ (A_i^2 + 1)\phi(A_i)x_i^T & 2(1 - \Phi(A_i)) + (A_i^3 + A_i)\phi(A_i) \end{bmatrix}\right] \\
= \; & \frac{1}{\sigma_u^2}E_x\begin{bmatrix} ((\zeta_i + A_i)\phi(A_i) + 1 - \Phi(A_i))x_i x_i^T & (A_i\Phi(A_i) + \upsilon_i)x_i \\ (A_i\Phi(A_i) + \upsilon_i)x_i^T & 2(1 - \Phi(A_i)) + A_i(\zeta_i\phi(A_i)A_i + \upsilon_i), \end{bmatrix}
\end{aligned}
$$

where $\upsilon_i \equiv (A_i^2 + 1)\phi(A_i)$.

$$
\begin{aligned}
E[m_i Y_i] &= E_x\left[E_u\left[\begin{bmatrix} [\text{Ind}(\tilde{u}_i > A_i) \cdot [-A_i\tilde{u}_i + \tilde{u}_i^2]]\, x_i \\ \text{Ind}(\tilde{u}_i > A_i) \cdot [-A_i + \tilde{u}_i]\,[-1 + \tilde{u}_i^2] \end{bmatrix} \middle| x_i\right]\right] \\
&= \frac{1}{\sigma_u}E_x\begin{bmatrix} [1 - (A_i\phi(A_i) + \Phi(A_i) - A_i\phi(A_i))]x_i \\ -(\phi(A_i) + A_i^2\phi(A_i) - (A_i^2 + 2)\phi(A_i)) \end{bmatrix} \\
&= E_x\begin{bmatrix} (1 - \Phi(A_i))x_i/\sigma_u \\ \phi(A_i)/\sigma_u \end{bmatrix}.
\end{aligned}
$$

Using the moments calculated above, we can verify that eq. (5) do not hold for the Tobit model.