**Data Collection, Weighting, and Modeling Techniques to Estimate Unbiased Population Parameters**

**Dale Robbennolt**
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712, USA
Email: dar4836@utexas.edu

**Ram M. Pendyala**
Arizona State University
School of Sustainable Engineering and the Built Environment
660 S. College Avenue, Tempe, AZ 85287, USA
Email: ram.pendyala@asu.edu

**Chandra R. Bhat (corresponding author)**
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712, USA
Tel: +1-512-471-4535; Email: bhat@mail.utexas.edu

**ABSTRACT**
Empirical research studies regularly encounter sampling-related challenges that can impact the validity and reliability of model estimation results. This paper presents a comprehensive examination of the implications of nonrandom sampling for estimator bias and appropriate modeling techniques to achieve unbiased results. Through theoretical and simulation-backed support, we underscore the importance of adopting appropriate sampling and estimation methods in two broad scenarios. First, we demonstrate that achieving range variation in exogenous variables, rather than strict population representativeness, is crucial for estimating individual-level causal relationships when sampling is based only on observed exogenous variables. Second, we investigate the efficacy of weighting approaches when sampling is endogenous and use a joint modeling approach to accommodate unobserved self-selection effects where traditional weighting approaches prove inadequate. Our proposed approach accommodates unobserved correlations and successfully recovers true population parameters when the joint distribution of exogenous variables in the population is known. The methodology also shows improved performance compared to existing methods even when only the population marginal distribution of exogenous variables is available. Notably, our simulation experiments extend beyond the conventional linear regression framework to include binary outcomes, providing crucial insights for nonlinear choice modeling applications. The findings underscore the importance of carefully considering sampling mechanisms and their implications for model estimation, while offering practical guidance for researchers facing various sampling-related challenges in empirical studies.

**Keywords:** Sample Selection, Selection Bias, Weighting, Survey Methods, Nonresponse, Joint Modeling

## 1. INTRODUCTION

Empirical research studies across multiple fields, including the transportation field, employ data from large surveys for their analysis. In doing so, studies must address such sampling-related issues as non-response, missing data, unequal sampling, and other survey biases (Hudson et al., 2004; Couper, 2017; Alhassan et al., 2024). The voluntary nature of most surveys means that, in many empirical applications, data are not randomly selected from the population. Instead, researchers only observe the responses of those who choose to respond to the survey, potentially resulting in sample selection biases (see, for example, Wittwer et al., 2024). In this context, there has been widespread debate about the ways that sampling considerations impact modeling results as well as the best approaches to achieve unbiased and consistent results (Winship and Radbill, 1994; Elwert and Winship, 2014; Solon et al., 2015). At the same time, unequal sampling does not always result in estimation biases. Instead, specific sampling techniques may result in more efficient model estimation (more certainty in model results) without any loss of consistency. Specifically, in the transportation sector, it is critical to develop appropriate survey sampling techniques that capture sufficient data from population groups that have small shares in the population (see, for example, Liévanos et al., 2019). This often necessitates the use of sampling approaches that do not capture truly representative samples but instead emphasize good range and coverage of exogenous variables.

Beyond the sample selection mechanisms themselves, a variety of modeling approaches have been proposed to accommodate the aforementioned selection biases. Specifically, sampling weights have long been considered essential when undertaking descriptive statistical analysis (such as determining population averages) on data with unequal sampling probabilities (Kish and Frankel, 1974; Pfeffermann, 1993). However, there has been much more debate about the appropriate circumstances to use sampling weights for causal effects modeling (see Solon et al., 2015; Bollen et al., 2016). The question of whether survey weights should be applied in different contexts has been discussed in a wide range of fields, such as statistics (Bollen et al., 2016; F. Wang et al., 2023), economics (Nguyen and Murphy, 2015; Gluschenko, 2018), sociology (Winship and Radbill, 1994; Becker and Ismail, 2016), epidemiology and medicine (Frohlich et al., 2001; Tchetgen et al., 2012; Howe et al., 2016; Avery et al., 2019), and transportation (Pendyala et al., 1991; Thill and Horowitz, 1997; Boto-García, 2023). There is general consensus that, if individuals have nearly equal sampling selection probabilities given their values of exogenous variables, then both weighted and unweighted estimators are consistent, and the lower variance of the unweighted estimator is preferred. But, when the probability of selection differs significantly among individuals due to a selection mechanism that is endogenous (that is, the probability of selection is not completely explainable based on exogenous variables), using sampling weights (representing the inverse probability of sample selection) can yield consistent estimates of population parameters, while unweighted estimators are generally inconsistent (see Manski and McFadden, 1981; Hausman and Wise, 1981; Wooldridge, 1999, 2001; Gelman, 2007; Solon et al., 2015; F. Wang et al., 2023). For instance, see the two sampling mechanisms considered for the same population shown in Figure 1, in the context of a simple linear regression. The graph on the left shows an exogenous sampling procedure (that is, the sample includes only those for whom the independent variable, $x$, is greater than or equal to 4), demonstrating that selection based on the exogenous variable does not worsen the estimated relationship between the variables even without the use of sampling weights. However, the endogenous selection procedure shown in the graph on the right demonstrates that sampling based on the endogenous outcome

(that is, the sample includes only those for whom the dependent variable, $y$, is greater than or equal to 3) biases the unweighted estimation results.

A critical issue, however, is that the true probability of selection is generally unknown in cases of nonresponse (Gelman, 2007). In these cases, weights are not based on the true probability of selection. Instead, they are estimated using post-data collection comparisons with population statistics to match the proportion of respondents in each demographic group with their population proportions in an external independent control (such as census data) (Biemer and Christ, 2008; Gary et al., 2023). The basic idea is that, by employing such weighting, one essentially gets back to the case of an equal probability sample with exogenous sampling, which takes care of any endogeneity in selection into the sample. However, unobserved factors may also play a significant role in response decisions (and thus, sampling probabilities), and such unobserved factors may also be correlated with the main outcome of interest. Such situations cannot be addressed through post-data collection weights, which rely on the assumption that selection is based solely on observed characteristics (Wooldridge, 2007; Brewer and Carlson, 2024). For instance, those with intrinsic existing knowledge or intrinsic interest in transportation technologies ("intrinsic" here means over and beyond the knowledge/interest in transportation technologies that can be captured by the exogenous variables collected) may be more likely to respond to a survey about electric vehicles, while those who have less intrinsic interest may be less likely to respond. Here, developing sampling weights based on other observed variables (such as demographics) as they relate to knowledge/interest in transportation technologies, or based on independent controls (such as census data), would not be adequate for consistent parameter estimation.

Further, in situations of truncation on the dependent outcome of interest (such as the sample shown on the right side of Figure 1), it is unclear how sampling weights could be developed at all. In these cases, whether for descriptive statistical analysis or for model estimation, the probability for selection given some values of the outcome is zero, and weights cannot be developed to accommodate these sampling mechanisms. Instead, model-based approaches offer mechanisms to account for truncation by conditioning on additional variables that may underly the truncation mechanism, as well as accommodating unobserved effects. Relatedly, while descriptive statistics are often considered separately from model-based approaches, since weights are needed even when sampling is based only on exogenous variables using standard formulas for descriptive statistics (see Solon et al., 2015), these same statistics can be calculated using model-based approaches that condition on exogenous variables. For example, consider a stratified sampling scheme that oversamples low-income individuals to get a sample (but still retains a good spread of income values), which is then used to estimate the average commute distance for individuals in the population. Given the general relationship between income and commute distance (that is, lower income individuals commute shorter distances; see Bhat, 2015; Bogomolov et al., 2021) we would expect that the standard unweighted sample mean would underestimate the average commute distance. To accommodate this sampling bias using a weight-based approach, the population mean (or the population variance or any other population parameter) could be estimated using sampling weights developed based on the probability of selection into the sample given a respondent's income. Alternatively, assuming the relationship between the commute distance and income is linear, the same population mean (or the population variance or other population parameters) could be predicted by estimating a linear regression of commute distance on income (at the individual level and using the exogenous-sampling based stratified sample) and then applying the overall population distribution of incomes along with the estimated coefficients to calculate predicted

commute distances.[1] That is, calculating the mean and variance of the predicted individual-level outcomes for the population data (from the estimated relationship using the stratified sample) would produce unbiased population estimates (in the case of variance, adding the appropriate error variance based on the residuals in the sample would also be needed for this approach). In fact, more generally, the use of sampling weights based on exogenous variables yields identical results to conditioning on all the variables used for weighting (and their interactions) in the model-based approach. This is because the exogenous sampling effectively renders the stratified sample-based regression estimator unbiased for the population regression (as depicted on the left side graph of Figure 1). But, while either approach (the weighting or the model-based approach) yields an unbiased result when sampling is based only on exogenous variables (as in the example above), the traditional weight-based approach cannot accommodate unobserved self-selection effects, while the model-based approach offers the flexibility to accommodate these unobserved selection effects as well, as discussed in more detail in Section 4. The same holds for the estimation of descriptive statistics based on other types of dependent variables, such as the use of a discrete choice modeling approach to predict population shares in each of many discrete categories (as we will empirically demonstrate in Sections 4.1 and 4.2).

Motivated by the discussion above, in this paper, we consider the ways that appropriate sampling strategies and modeling techniques can be used to improve estimation results when the collection of a representative sample is unnecessary or impractical. We focus on two broad types of sampling techniques, each of which carry different implications for the biases present in the resulting sample and require different modeling approaches. These approaches are visually depicted in Figure 2 (the annotations by the boxes and equations near the top of the figure refer to the sampling definitions used in the next section). First, *exogenous sampling* (shown on the left side of the figure) includes all cases where the probability of an individual being included in the sample is equal to the probability of being included conditional only on the values of the exogenous variables. Exogenous sampling includes many stratified sampling techniques as well as simple random sampling. It also includes cases of non-response and missing data, but when the non-response and "missingness" is completely determined by the exogenous variables. Second, *endogenous sampling* (shown on the right side of the figure) includes situations when sampling is dependent on unobserved variables (alone or in addition to observed variables) that impact or are correlated with the outcome of interest. Endogenous sampling encompasses cases of nonresponse and missing data, when unobserved characteristics influence an individual's decision of whether to respond to the survey. Since these variables are not observed, they cannot be accounted for by traditional weighting methods, which rely on the assumption that selection is based on observed characteristics (see Brewer and Carlson, 2024). Finally, choice-based sampling (where sampling is directly dependent on the outcome, or alternative that is selected), is another form of an endogenous sampling approach, because the probability of an individual being included in the sample is not equal to the probability of being included conditional on exogenous variables (Manski and Lerman, 1977). However, it is distinct from other forms of endogenous sampling discussed here because the relationship between selection probability and the outcome is direct

[1] Of course, non-linear relationships may exist between commute distance and income, but as long as the relationship is correctly specified in this instance through an appropriate non-linear regression relationship, using weights or using the non-linear relationship (applied to the overall population distribution) would essentially provide the same population-level results. Of course, this requires that the individual-level relationship be specified correctly, which is anyway the hallmark of any data analysis exercise. In the rest of this paper, we will assume that the analyst has done all due diligence in developing a good model specification.

and observed rather than being dependent on unobserved variables. Thus, in contrast to the form of endogenous sampling discussed above (and discussed in more detail in Section 4) where sampling is based on unobserved variables, weights can be developed to represent the inverse probability of selection and essentially gets back to the case of an equal probability sample with exogenous sampling.

In this research, through theoretical and simulation-backed support, we underscore the importance of adopting appropriate sampling and estimation methods in each of these situations. We contribute to the existing literature in several ways. First, we explicate, rigorously and comprehensively, why the unweighted approach is to be preferred over the weighted approach in the case of exogenous sampling in discrete choice models. We then use simulations to demonstrate that range variation in exogenous variables needs to be the key in survey designs (not necessarily population representativeness) to estimate individual-level causal relationships. Second, we demonstrate that weighting approaches are unable to accommodate endogenous selection, when sampling is based on unobserved variables (see also Brewer and Carlson, 2024). Instead, we propose a joint modeling approach to accommodate these unobserved sampling biases. We show that this approach of jointly modeling sample selection along with the outcome of interest can accommodate unobserved correlations and recover the true population parameters when the joint distribution of exogenous variables in the population is known. We also demonstrate that this method can be used to improve upon existing methods that do not account for endogenous selection even when only the population marginal distribution of exogenous variables is known. This would be the case, for instance, when using an address-based sampling frame to recruit participants (such that not all exogenous variable data is known in advance for all participants) and using the population marginal distribution of exogenous variables (such as Census data for the target population) to accommodate endogenous selection (in this case, nonresponse from some portion of the targeted random sample). Finally, while most existing simulation studies exclusively consider the case of linear regression (including Brewer and Carlson, 2024), we use binary outcomes in each of the simulations. This is a critical extension given the widespread use of nonlinear choice models and the fact that some techniques to deal with sample selection in the case of linear regression (such as the use of a Heckman two-stage formulation) do not extend to the nonlinear setting (see Greene, 2003; Galimard et al., 2018). Our analysis should be of interest to all empirical researchers working in the area of survey research and associated data modeling.

## 2. SAMPLE SELECTION PROPERTIES

Let W represent the population of interest and let $w_q$ be a random draw from the population (where every individual has the same chance of being selected). If the analyst uses ("considers") say $Q$ random draws from the population, that corresponds to the case of a random sample of size $Q$. Now consider the case of estimation of a binary choice model with this random sample (extension to the case of a multinomial choice model is straightforward and does not present any additional complications). Assume the usual underlying utility structure for each alternative, such that the latent propensity $y_q^*$ of individual $q$ (corresponding to a draw from the population) selecting the first alternative may be written as the difference between the utilities of the two alternatives:

$$y_q^* = U_{q0} - U_{q1} = \boldsymbol{\beta}'\mathbf{x}_q + \varepsilon_q \tag{1}$$

where $\mathbf{x}_q$ is an $(A\times 1)$ vector of exogenous variables (including a constant), $\boldsymbol{\beta}$ is an $(A\times 1)$ column vector of corresponding coefficients to be estimated, and $\varepsilon_q$ is a random normal error term. Then, each individual is assumed to select alternative $y_q = m_q$ ($m_q \in \{0,1\}$) if $\theta^{m_q-1} < y_q^* < \theta^{m_q}$, for $\theta^{-1} = -\infty$, $\theta^0 = 0$, and $\theta^1 = \infty$. Of course, all that is observed by the researcher is the actual chosen binary outcome $m_q$ rather than the underlying propensity. Then, according to the usual binary probit model, the likelihood function for the individual can be written as:

$$L_q(\boldsymbol{\beta}, m_q) = \Pr(y_q = m_q) = \int_{D_{qr}} f(r \mid \boldsymbol{\beta}'\mathbf{x}_q, 1)dr \tag{2}$$

where the integration domain $D_{qr} = \{r : \theta^{m_q-1} < r < \theta^{m_q}\}$ is simply the region of $y_q^*$ truncated by the appropriate upper and lower thresholds. $f(r \mid \boldsymbol{\beta}'\mathbf{x}_q, 1)$ is the univariate normal density function with a mean of $\boldsymbol{\beta}'\mathbf{x}_q$ and a variance of one. Collect the elements $y_q$ and $m_q$ into vectors **y** and **m**, respectively. Then, the likelihood function of the sample (assuming independence across individuals) is given by:

$$L(\boldsymbol{\beta}, \mathbf{m}) = \prod_{q=1}^{Q} L_q(\boldsymbol{\beta}, m_q) \tag{3}$$

Under usual regularity conditions needed for likelihood objects, the logarithm of the likelihood function can be maximized through solving the corresponding first-order (score) equations (see Molenberghs and Verbeke, 2005, p. 191). These score equations can be shown to be unbiased since they are linear combinations of individual likelihood score functions based on the choice probabilities for each individual. That is, the estimation of the maximum likelihood model is achieved by solving the score equations given by:

$$\boldsymbol{s}(\boldsymbol{\beta}, \mathbf{m}) = \nabla \log L(\boldsymbol{\beta}, \mathbf{m}) = \sum_{q=1}^{Q} \boldsymbol{s}_q(\boldsymbol{\beta}, m_q) = \mathbf{0}, \tag{4}$$

where $\boldsymbol{s}_q(\boldsymbol{\beta}, m_q) = \dfrac{\partial \log L_q(\boldsymbol{\beta}, m_q)}{\partial \boldsymbol{\beta}}$. Then, to prove unbiasedness, we can show that

$$\begin{aligned}
E\left[\boldsymbol{s}(\boldsymbol{\beta}, \mathbf{m})\right] &= E\left[\sum_{q=1}^{Q} \boldsymbol{s}_q(\boldsymbol{\beta}, m_q)\right] = \sum_{q=1}^{Q} E\left[\boldsymbol{s}_q(\boldsymbol{\beta}, m_q)\right] = \sum_{q=1}^{Q} E\left[\frac{\partial \log L_q(\boldsymbol{\beta}, m_q)}{\partial \boldsymbol{\beta}}\right] \\
&= \sum_{q=1}^{Q} \sum_{m_q \in \{0,1\}} \frac{\partial \log L_q(\boldsymbol{\beta}, m_q)}{\partial \boldsymbol{\beta}} P(y_q = m_q) = \sum_{q=1}^{Q} \sum_{m_q \in \{0,1\}} \frac{\partial \log L_q(\boldsymbol{\beta}, m_q)}{\partial \boldsymbol{\beta}} L_q(\boldsymbol{\beta}, m_q) \\
&= \sum_{q=1}^{Q} \sum_{m_q \in \{0,1\}} \frac{1}{L_q(\boldsymbol{\beta}, m_q)} \frac{\partial L_q(\boldsymbol{\beta}, m_q)}{\partial \boldsymbol{\beta}} L_q(\boldsymbol{\beta}, m_q) = \sum_{q=1}^{Q} \sum_{m_q \in \{0,1\}} \frac{\partial L_q(\boldsymbol{\beta}, m_q)}{\partial \boldsymbol{\beta}} \\
&= \sum_{q=1}^{Q} \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{m_q \in \{0,1\}} L_q(\boldsymbol{\beta}, m_q) == \sum_{q=1}^{Q} \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{m_q \in \{0,1\}} P(y_q = m_q) = \sum_{q=1}^{Q} \frac{\partial}{\partial \boldsymbol{\beta}}(1) = \sum_{q=1}^{Q} \mathbf{0} = \mathbf{0}.
\end{aligned} \tag{5}$$

Since the expectation of the score equations are all equal to zero, the estimator is unbiased. For the asymptotic properties, let $\boldsymbol{\beta}_0$ be the true unknown parameter vector in the population. Then, each random draw ( $w_q$ ) from the population amounts to drawing a value $\boldsymbol{s}_q(\boldsymbol{\beta}_0, m_q)$ from a distribution in the population with a zero mean and variance:

$$\mathbf{J} = VAR[\boldsymbol{s}_q(\boldsymbol{\beta}_0, m_q)] = E\left[\left(\frac{\partial \log L_q(\boldsymbol{\beta}_0, m_q)}{\partial \boldsymbol{\beta}_0}\right)\left(\frac{\partial \log L_q(\boldsymbol{\beta}_0, m_q)}{\partial \boldsymbol{\beta}_0'}\right)\right] \tag{6}$$

Then, using the Central Limit Theorem, we get:

$$\frac{1}{\sqrt{Q}}\boldsymbol{s}(\boldsymbol{\beta}_0, \mathbf{m}) \xrightarrow{d} MVN_A(\mathbf{0}, \mathbf{J}) \tag{7}$$

where $MVN_A(\cdot)$ is the multivariate normal distribution function with $A$ dimensions and $\boldsymbol{s}(\boldsymbol{\beta}_0, \mathbf{m}) = \sum_{q=1}^{Q} \boldsymbol{s}_q(\boldsymbol{\beta}_0, m_q)$ (see also, Yi et al., 2011; Bhat, 2014). Then, for the estimator $\hat{\boldsymbol{\beta}}$, we can write a Taylor Series expansion of the score function for the estimator around the score function for the true population parameters $\boldsymbol{\beta}_0$ to get

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = [-\nabla \boldsymbol{s}(\boldsymbol{\beta}_0, \mathbf{m})]^{-1} \boldsymbol{s}(\boldsymbol{\beta}_0, \mathbf{m}) . \tag{8}$$

Using the Law of Large Numbers, we also know that $\frac{1}{Q}\nabla \boldsymbol{s}(\boldsymbol{\beta}_0, \mathbf{m})$ converges to the population expected value since it is a sample mean of $\nabla \boldsymbol{s}_q(\boldsymbol{\beta}_0, m_q)$. So,

$$-\frac{1}{Q}\nabla \boldsymbol{s}(\boldsymbol{\beta}_0, \mathbf{m}) \xrightarrow{d} E[-\frac{1}{Q}\nabla \boldsymbol{s}(\boldsymbol{\beta}_0, \mathbf{m})] = \mathbf{H} = E\left[\frac{-\partial^2 \log L_q(\boldsymbol{\beta}_0, m_q)}{\partial \boldsymbol{\beta}_0 \partial \boldsymbol{\beta}_0'}\right]. \tag{9}$$

Combining Equation (6) through Equation (9), and applying Slutsky's Theorem, we get the asymptotic distribution:

$$\sqrt{Q}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} MVN_A(\mathbf{0}, \mathbf{H}^{-1}\mathbf{J}\mathbf{H}^{-1}), \tag{10}$$

where the quantity $\mathbf{H}\mathbf{J}^{-1}\mathbf{H}$ represents the well-known Godambe (1960) information matrix. Additionally, in the case of random sampling, the asymptotic variance simplifies to

$$\mathbf{I}(\boldsymbol{\beta}_0)^{-1} = -\mathbf{J}^{-1} = \mathbf{H}^{-1} = E\left[\frac{-\partial^2 \log L_q(\boldsymbol{\beta}_0, m_q)}{\partial \boldsymbol{\beta}_0 \partial \boldsymbol{\beta}_0'}\right]^{-1} \tag{11}$$

where $\mathbf{I}(\boldsymbol{\beta}_0)$ represents the Fisher (1922) information matrix for the sample. Thus, the estimator $\hat{\boldsymbol{\beta}}$ converges in probability to $\boldsymbol{\beta}_0$ as $Q \to \infty$ leading to the consistency of the estimator.

Finally, while descriptive statistics in this case of random sampling can be computed directly from the sample using standard formulas, the model estimates themselves can be used to calculate unbiased descriptive statistics (in this case, for the binary probit model considered, the population shares predicted to select each alternative for the outcome). Specifically, the overall

probability that some randomly selected individual from the population would select alternative $y=1$ is simply the average over $\Pr(y=1)=\frac{1}{Q}\sum_{q=1}^{Q}L_q(\boldsymbol{\beta},1)$.

The analyst, however, may not always "consider" each random draw to include for analysis. Let $\omega_q$ be an indicator of whether the analyst "considers" draw $w_q$ or not ($\omega_q=1$ if the analyst "considers" the draw and $\omega_q=0$ otherwise) and let the actual observed sample (the subset of the random sample $Q$ where $\omega_q=1$) be $\tilde{Q}$. Then, the case of non-random, but exogenous, sampling corresponds to the assumption that the probability of selection into the actual observed sample ($P(\omega_q=1)$) does not depend on the outcome, conditional on the exogenous variables:

$$P(\omega_q=1\,|\,\mathbf{x}_q,y_q)=P(\omega_q=1\,|\,\mathbf{x}_q). \tag{12}$$

Estimation issues for this case are discussed in Section 3. Next, the second case of endogenous sampling corresponds to situations where selection is informative of the outcome, even after conditioning on all the observed exogenous variables $\mathbf{x}_q$:

$$P(\omega_q=1\,|\,\mathbf{x}_q,y_q)\neq P(\omega_q=1\,|\,\mathbf{x}_q). \tag{13}$$

Estimation issues for this case are discussed in Section 4.

## 3. EXOGENOUS SELECTION

Two estimators can be considered in the situation of exogenous sampling. First, the usual unweighted estimator for a sample of decision-makers is simply determined using the product of the individual likelihoods for those individuals in the observed sample $\tilde{Q}$. Equivalently, the likelihood function for the sample can be written as a function of all the individuals in the underlying random sample $Q$ as:

$$L_u(\boldsymbol{\beta},\mathbf{m})=\prod_{q=1}^{Q}L_q(\boldsymbol{\beta},m_q)^{\omega_q} \tag{14}$$

where the product is over all individuals in the underlying random sample of $Q$ draws, but the selection indicator $\omega_q$ means that only those individuals considered by the researcher influence the likelihood function. We can then show that the unweighted estimator is unbiased under exogenous sampling because the expected value of the score function (label the score function for the unweighted estimator as $\boldsymbol{s}^u(\boldsymbol{\beta},\mathbf{m})$) again converges to zero. Using the Law of Iterated Expectations, we can show the following:

$$\begin{aligned}
&E\left[\boldsymbol{s}^u(\boldsymbol{\beta},\mathbf{m})\right]=E\left[\sum_{q=1}^{Q}\omega_q\boldsymbol{s}_q(\boldsymbol{\beta},m_q)\right]=\sum_{q=1}^{Q}E\left[\omega_q\boldsymbol{s}_q(\boldsymbol{\beta},m_q)\right]\\
&=\sum_{q=1}^{Q}E\left[E\left\{\omega_q\boldsymbol{s}_q(\boldsymbol{\beta},m_q)\,|\,\mathbf{x}_q\right\}\right]=\sum_{q=1}^{Q}E\left[E\left\{\omega_q\,|\,\mathbf{x}_q\right\}E\left\{\boldsymbol{s}_q(\boldsymbol{\beta},m_q)\,|\,\mathbf{x}_q\right\}\right]\\
&=\sum_{q=1}^{Q}E\left[P(\omega_q=1\,|\,\mathbf{x}_q)E\left\{\boldsymbol{s}_q(\boldsymbol{\beta},m_q)\,|\,\mathbf{x}_q\right\}\right]=\sum_{q=1}^{Q}E\left[P(\omega_q=1\,|\,\mathbf{x}_q)\mathbf{0}\right]=\mathbf{0}
\end{aligned} \tag{15}$$

which holds because of the assumption that the probability of selection does not depend on the outcome, conditional on the exogenous variables ($P(\omega_q = 1 | \mathbf{x}_q, y_q) = P(\omega_q = 1 | \mathbf{x}_q)$) and the expected value of the score function conditional on exogenous variables ($E\left\{\boldsymbol{s}_q(\boldsymbol{\beta}, m_q) | \mathbf{x}_q\right\}$) is equal to zero (following the same steps shown in Equation (5)). Similarly, following the logic from the previous section, the asymptotic properties of the unweighted estimator can be derived. For the unweighted estimator, we get the asymptotic distribution

$$\sqrt{Q}(\hat{\boldsymbol{\beta}}_u - \boldsymbol{\beta}_0) \xrightarrow{d} MVN_A(\mathbf{0}, \mathbf{H}_u^{-1}\mathbf{J}_u\mathbf{H}_u^{-1}) \tag{16}$$

where $\mathbf{H}_u = E\left[\omega_q \frac{-\partial^2 \log L_q(\boldsymbol{\beta}_0, m_q)}{\partial \boldsymbol{\beta}_0 \partial \boldsymbol{\beta}_0'}\right]$ and $\mathbf{J}_u = E\left[\omega_q \left(\frac{\partial \log L_q(\boldsymbol{\beta}_0, m_q)}{\partial \boldsymbol{\beta}_0}\right)\left(\frac{\partial \log L_q(\boldsymbol{\beta}_0, m_q)}{\partial \boldsymbol{\beta}_0'}\right)\right]$.

Further, the law of iterated expectations can be applied again to simplify this expression for the asymptotic variance. Specifically, we can condition on $\mathbf{x}_q$ in the expressions for both $\mathbf{H}_u$ and $\mathbf{J}_u$. Then, since $E\left[\frac{-\partial^2 \log L_q(\boldsymbol{\beta}_0, m_q)}{\partial \boldsymbol{\beta}_0 \partial \boldsymbol{\beta}_0'} | \mathbf{x}_q\right] = E\left[\left(\frac{\partial \log L_q(\boldsymbol{\beta}_0, m_q)}{\partial \boldsymbol{\beta}_0}\right)\left(\frac{\partial \log L_q(\boldsymbol{\beta}_0, m_q)}{\partial \boldsymbol{\beta}_0'}\right) | \mathbf{x}_q\right]$, it holds that $\mathbf{H}_u^{-1} = -\mathbf{J}_u$ and the asymptotic variance simplifies to a weighted version of the Fisher Information, without the need to compute the full robust "sandwich" variance estimator (Greene, 2018; Wooldridge, 2002):

$$\mathbf{H}_u^{-1}\mathbf{J}_u\mathbf{H}_u^{-1} = \mathbf{I}_u(\boldsymbol{\beta}_0)^{-1} = -\mathbf{J}_u^{-1} = \mathbf{H}_u^{-1} = E\left[\omega_q \frac{-\partial^2 \log L_q(\boldsymbol{\beta}_0, m_q)}{\partial \boldsymbol{\beta}_0 \partial \boldsymbol{\beta}_0'}\right]^{-1}. \tag{17}$$

Since the unweighted estimator still converges to the true population value, the estimator is consistent. However, the variance of the estimator is increased relative to that of the estimator with the larger random sample, as the size of the actual observed sample is reduced (because not all individuals from the underlying random sample are observed).

An alternative to the unweighted estimator is the inverse probability weighed estimator, which includes a weight that is defined to be equal to the inverse of the probability of selection. Define the weight as $v_q = 1 / P(\omega_q = 1 | \mathbf{x}_q)$, which depends on an individual's values of exogenous variables $\mathbf{x}_q$. For a stratified sampling technique these weights can be estimated based on the selection probabilities in the sampling design, while techniques such as raking or iterative proportional fitting (IPF) can be used to generate weights based on the population marginal distributions of exogenous variables when sampling probabilities are unknown (if we continue to assume that sampling is based only on these exogenous variables, then these techniques approximate inverse probability weights; see Biemer and Christ, 2008). Then, the weighted likelihood function for a sample of decision-makers is given by:

$$L_w(\boldsymbol{\beta}, \mathbf{m}) = \prod_{q=1}^{Q} L_q(\boldsymbol{\beta}, m_q)^{\omega_q v_q} \tag{18}$$

where the estimator is again based on the same underlying sample of random draws ($w_q$), but only individuals with $\omega_q = 1$ contribute to the overall log-likelihood function, with their contributions dependent on their weight $v_q$. Following the same logic as the proof for the unweighted estimator,

the weighted estimator can be shown to be unbiased because the expected value of the score function (label the score function for the weighted estimator as $\boldsymbol{s}^{w}(\boldsymbol{\beta},\mathbf{m})$) again converges to zero:

$$
\begin{aligned}
&E\left[\boldsymbol{s}^{w}(\boldsymbol{\beta},\mathbf{m})\right] = E\left[\sum_{q=1}^{Q}\omega_q v_q \boldsymbol{s}_q(\boldsymbol{\beta},m_q)\right] = \sum_{q=1}^{Q} E\left[\omega_q v_q \boldsymbol{s}_q(\boldsymbol{\beta},m_q)\right] \\
&= \sum_{q=1}^{Q} E\left[E\left\{\omega_q v_q \boldsymbol{s}_q(\boldsymbol{\beta},m_q)\,|\,\mathbf{x}_q\right\}\right] = \sum_{q=1}^{Q} E\left[E\left\{\omega_q v_q\,|\,\mathbf{x}_q\right\}E\left\{\boldsymbol{s}_q(\boldsymbol{\beta},m_q)\,|\,\mathbf{x}_q\right\}\right] \\
&= \sum_{q=1}^{Q} E\left[P(\omega_q=1\,|\,\mathbf{x}_q)v_q E\left\{\boldsymbol{s}_q(\boldsymbol{\beta},m_q)\,|\,\mathbf{x}_q\right\}\right] = \sum_{q=1}^{Q} E\left[\mathbf{0}\right] = \mathbf{0}.
\end{aligned}
\tag{19}
$$

The above holds because of the definition of the sampling weight as $v_q = 1/P(\omega_q=1\,|\,\mathbf{x}_q)$, and the assumption that both weights and selection are based only on exogenous variables $P(\omega_q=1\,|\,\mathbf{x}_q,y_q) = P(\omega_q=1\,|\,\mathbf{x}_q)$ (see Equation (5) for proof that the individual score functions converge to zero). Therefore, in the case of exogenous sampling, both the unweighted and weighted estimators are unbiased. Further, the asymptotic distribution of the weighted estimator is given by:

$$
\sqrt{Q}(\hat{\boldsymbol{\beta}}_w - \boldsymbol{\beta}_0) \xrightarrow{d} MVN_A(\mathbf{0},\mathbf{H}_w^{-1}\mathbf{J}_w\mathbf{H}_w^{-1}),
\tag{20}
$$

where $\mathbf{H}_w = E\left[\omega_q v_q \dfrac{-\partial^2 \log L_q(\boldsymbol{\beta}_0,m_q)}{\partial\boldsymbol{\beta}_0\partial\boldsymbol{\beta}_0'}\right]$ and $\mathbf{J}_w = E\left[\omega_q v_q^2\left(\dfrac{\partial \log L_q(\boldsymbol{\beta}_0,m_q)}{\partial\boldsymbol{\beta}_0}\right)\left(\dfrac{\partial \log L_q(\boldsymbol{\beta}_0,m_q)}{\partial\boldsymbol{\beta}_0'}\right)\right]$.

In the case of the weighted estimator, the full robust "sandwich" variance estimator is needed because there is no cancellation between the $\mathbf{H}_w$ and $\mathbf{J}_w$ matrices (see Murphy and Topel, 2002; Wooldridge, 2007). Therefore, both the weighted and unweighted estimators are also consistent under exogenous sampling and converge to the true population values. However, by comparing the asymptotic variances of the two estimators, we can show that the unweighted estimator is more efficient than the weighted estimator, making it a more appropriate choice for small-sample estimation. Specifically, if the difference $\mathbf{H}_w^{-1}\mathbf{J}_w\mathbf{H}_w^{-1} - \mathbf{H}_u^{-1}\mathbf{J}_u\mathbf{H}_u^{-1}$ is positive semi-definite then the asymptotic variance of the unweighted estimator will be less than that of the weighted estimator. This will hold if:

$$
\begin{aligned}
\mathbf{H}_u\mathbf{J}_u^{-1}\mathbf{H}_u - \mathbf{H}_w\mathbf{J}_w^{-1}\mathbf{H}_w &= E\left[\omega_q \frac{\partial^2 \log L_q(\boldsymbol{\beta}_0,m_q)}{\partial\boldsymbol{\beta}_0\partial\boldsymbol{\beta}_0'}\right] \\
&\quad - E\left[\frac{\partial^2 \log L_q(\boldsymbol{\beta}_0,m_q)}{\partial\boldsymbol{\beta}_0\partial\boldsymbol{\beta}_0'}\right]E\left[\frac{1}{\omega_q}\frac{\partial^2 \log L_q(\boldsymbol{\beta}_0,m_q)}{\partial\boldsymbol{\beta}_0\partial\boldsymbol{\beta}_0'}\right]^{-1}E\left[\frac{\partial^2 \log L_q(\boldsymbol{\beta}_0,m_q)}{\partial\boldsymbol{\beta}_0\partial\boldsymbol{\beta}_0'}\right]
\end{aligned}
\tag{21}
$$

is positive semi-definite. Then, since the expression in Equation (21) can be rewritten in the form

$$
E\left[\mathbf{A}'\mathbf{A}\right] - E\left[\mathbf{A}'\mathbf{B}\right]E\left[\mathbf{B}'\mathbf{B}\right]^{-1}E\left[\mathbf{B}'\mathbf{A}\right]
\tag{22}
$$

where $\mathbf{A} = \left( \omega_q \frac{\partial^2 \log L_q(\boldsymbol{\beta}_0, m_q)}{\partial \boldsymbol{\beta}_0 \partial \boldsymbol{\beta}_0'} \right)^{1/2}$ and $\mathbf{B} = \left( \frac{1}{\omega_q} \frac{\partial^2 \log L_q(\boldsymbol{\beta}_0, m_q)}{\partial \boldsymbol{\beta}_0 \partial \boldsymbol{\beta}_0'} \right)^{1/2}$, it follows from the Cauchy-Schwarz Inequality that the expression is positive semi-definite (Tripathi, 1999). Therefore, since both estimators are unbiased and consistent, and the unweighted estimator is more efficient, the unweighted estimator is preferred under exogenous sampling.

Beyond the choice of a specific estimator, the asymptotic properties of the unweighted estimator reveal implications from a data collection standpoint. While the collection of a representative sample is often a stated goal of many large-scale survey efforts, this is not necessary for the estimation of a causal effects model. Based on the result that exogenous sampling techniques produce consistent and unbiased results without the need for weighting, many exogenous sampling techniques can be considered as alternatives to random selection. Comparing the efficiency of the estimators, random selection is only the best sampling approach when it results in a more efficient estimator than other exogenous selection schemes.

Using the asymptotic variance of the unweighted estimator, we can assess the impact of specific exogenous sampling schemes on estimator efficiency. Specifically, the asymptotic variance of the unweighted estimator can be rewritten as:

$$\mathbf{H}_u^{-1}\mathbf{J}_u\mathbf{H}_u^{-1} = E\left[ \omega_q \frac{\partial^2 \log L_q(\boldsymbol{\beta}_0, m_q)}{\partial \boldsymbol{\beta}_0 \partial \boldsymbol{\beta}_0'} \right]^{-1} = \left\{ \int_{\mathbf{x}_q} g(\mathbf{x}_q) P(\omega_q = 1 \mid \mathbf{x}_q) \mathbf{I}(\boldsymbol{\beta}_0 \mid \mathbf{x}_q) d\mathbf{x}_q \right\}^{-1} \tag{23}$$

where $g(\mathbf{x}_q)$ represents the distribution of exogenous variables present in the population and $\mathbf{I}(\boldsymbol{\beta}_0 \mid \mathbf{x}_q) = E\left[ \frac{\partial^2 \log L_q(\boldsymbol{\beta}_0, m_q)}{\partial \boldsymbol{\beta}_0 \partial \boldsymbol{\beta}_0'} \mid \mathbf{x}_q \right]$ is the expected Fisher information matrix for an individual with an exogenous variable vector $\mathbf{x}_q$. Using this expression, it is apparent that the variance of the estimator is directly dependent on the relationship between the distribution of $\mathbf{x}_q$ in the sample and the Fisher information. Notably, the distribution of $\mathbf{x}_q$ influencing the variance of the estimator (accounting for both $g(\mathbf{x}_q)$ and $P(\omega_q = 1 \mid \mathbf{x}_q)$) is that of the sample, not the population, implying that researchers can select the best exogenous sampling scheme to minimize the asymptotic variance of the unweighted estimator. While the effects of the distribution of exogenous variables on the variance of the estimator depend on the specific form of the likelihood function, we can show that maximizing the variance of the exogenous distribution in the sample minimizes the variance of the estimator in the case of the binary probit model of interest here. Specifically, the Fisher information of an individual for the binary probit model is given by:

$$\mathbf{I}(\boldsymbol{\beta}_0 \mid \mathbf{x}_q) = \frac{f(0 \mid \boldsymbol{\beta}_0'\mathbf{x}_q, 1)}{\int_{D_{qr}} f(r \mid \boldsymbol{\beta}_0'\mathbf{x}_q, 1) dr \left( 1 - \int_{D_{qr}} f(r \mid \boldsymbol{\beta}_0'\mathbf{x}_q, 1) dr \right)} \mathbf{x}_q \mathbf{x}_q' \tag{24}$$

where $f(r \mid \boldsymbol{\beta}'\mathbf{x}_q, 1)$ is again the univariate normal density function with a mean of $\boldsymbol{\beta}'\mathbf{x}_q$ and a variance of one and the expression does not depend on the chosen outcome $m_q$ (see also,

Demidenko, 2001). The initial ratio in this expression takes its maximum value when $\boldsymbol{\beta}_0'\mathbf{x}_q = 0$, suggesting that choice occasions where both alternatives are well represented in terms of choice provide more information. Although an intuitive result, this has important implications for stated preference survey design, indicating that designing choice scenarios with realistic alternatives can help improve estimator efficiency (see also, Terawaki et al., 2003; Rose and Bliemer, 2013). More generally, however, the sum over the outer product $\mathbf{x}_q\mathbf{x}_q'$, in Equation (24), indicates that a greater variance in the distribution for $\mathbf{x}_q$ present in the sample (as a combination of the underlying distribution in the population $g(\mathbf{x}_q)$ and the sampling probabilities $P(\omega_q = 1 \mid \mathbf{x}_q)$) also serves to increase the total information in the sample, and thereby reduces the variance of the estimator. Thus, selecting a specific sampling scheme that provides greater exogenous variation will result in an improved estimator efficiency compared with simply collecting a random representative sample (assuming that these samples are of the same size). Therefore, when estimating individual-level causal relationships is the main objective, one need not get too fixated about representativeness; rather, good range and coverage of the exogenous variables is important.

For a simple textbook example of this issue using a linear regression, see Figure 3. In the figure, two samples are drawn from the same population with different distributions of exogenous variables (values of the exogenous variable $x$ are drawn from a normal distribution with mean 3 and standard deviation 1 for the "high variance sample" and from a normal distribution with mean 3 and standard deviation 0.2 for the "low variance sample"). Based on the discussion above, both sampling techniques are consistent (with either sampling technique, generating a larger sample will lead each estimated linear regression to converge towards the black population line). However, for this fixed number of draws, the sample with the greater variance of the exogenous variable performs better than the one with the smaller variance. This issue is further explored in the following simulation in the context of the binary probit model.

### 3.1 Exogenous Selection Simulation Design

To compare and evaluate the effects of sampling different exogenous populations at different rates, we undertake a simulation exercise based on the binary probit choice system described in the previous section. The design of this simulation is shown visually in Figure 4. For the simulation, four "populations" of 100,000 individuals each are generated with values of two exogenous dummy (binary) variables (drawn from underlying standard normal distributions with specific correlations between the two underlying continuous variables of $\sigma = \{0.00, 0.25, 0.50, 0.75\}$; see the top section of Figure 4 labeled "Generate Population Exogenous Data"). We consider populations with different levels of correlation between the exogenous variables because the presence of correlations (which are often present in realistic empirical applications) leads to higher standard errors and may mask or amplify sampling biases. Then, appropriate thresholds are applied such that there is a 5% chance that $x^1 = 1$ and 50% chance that $x^2 = 1$. For the outcome (see the second section of Figure 4 labeled "Generate Outcome Data"), each individual's latent propensity $y_q^*$ is calculated based on Equation (1) using the coefficients $\beta_0 = -0.75$ (for the constant), $\beta_1 = 0.50$ (corresponding to $x^1$), and $\beta_2 = -0.50$ (corresponding to $x^2$) along with a random normal error $\varepsilon_q$ drawn from an independent standard normal distribution. Finally, the outcome $y_q$ for each individual is determined using the appropriate thresholds.

Then, given each fixed population, a binary probit model is used to generate a set of true coefficient values using the full population data. Next, five sampling strategies are used to generate

datasets of 500 individuals each (see the third section of the figure labeled "Sample Data") within each of the four populations (each population refers to individuals with a given correlation value). First, samples are drawn such that the percentage of cases with $x^1 = 1$ in the sample matches that of the population (5%), essentially ensuring that all individuals in the population have the same probability of selection. Second, four additional sets of samples involve oversampling individuals with $x^1 = 1$ such that they appear at a higher rate in each sample than they do in the population (and thus individuals with $x^1 = 0$ have a lower probability of selection). Individuals with $x^1 = 1$ make up 7.5%, 10%, 12.5%, and 15% of the sample, respectively, for the final four sets of samples. 1,000 samples (each of 500 individuals) are generated from each population with each of the aforementioned five sampling strategies.

For each sample, an unweighted binary probit model is run (see the fourth section of the figure labeled "Run Binary Probit Models"), and the results are stored to evaluate the performance of each strategy. For each coefficient, the following results are reported: (1) the mean coefficient value across the 1,000 runs, (2) the average percentage error (APE) compared with the coefficient value calculated for the population, (3) the standard deviation of the coefficient values calculated across the 1,000 runs, and (4) the mean standard error (this is done for each set of 1,000 samples, shown in the bottom section of Figure 4 labeled "Evaluate Performance"). Finally, in addition to the model estimates themselves, the estimates are used to predict the share of the population selecting $y = 1$ (this is again done using each sampling method for each of the four populations, although this step is not shown in the figure). To do so, the estimated parameters are applied to calculate the probability of each individual in the population (not the estimation sample) predicted to select the outcome $y = 1$. The average across these individual probabilities then represents the predicted population share selecting the outcome $y = 1$.

### 3.2 Exogenous Selection Simulation Results

Table 1 presents an overall summary of the performance of the unweighted estimator under each sampling procedure. In the table, each column-panel represents a population with the specified correlation between the two exogenous variables, while each row-panel indicates a set of samples selected from that population to include the specified proportion of individuals with $x^1 = 1$. Within each panel, the performance metrics are shown for each of the three coefficients. The mean coefficient values consistently converge to the true population values (shown at the top of each column), confirming the consistency of the unweighted estimator under each of these exogenous sampling mechanisms. However, the mean standard errors, which represent the efficiency of the estimator, are significantly influenced by the sampling procedure. For the population with no correlation between the exogenous variables, the standard error for $\beta_1$ (corresponding to the exogenous variable with highly unequal proportions in the population) is reduced by more than 40% (from a standard error of 0.28 to a standard error of 0.17) when its representation in the sample is increased from the population level of 5% to a level of 15% in the sample. This dramatic improvement in estimator efficiency highlights the benefits of considering the distribution of exogenous variables in the sample, as relatively small changes in representation in the sample may allow statistical inferences to be made about small population segments that may not be sufficiently represented in a random sample.

The effects discussed above are even stronger when correlations are present between the exogenous variables, as the standard errors increase as correlations are introduced (in each row the mean standard error increases moving to the right). Similar results can be observed by comparing

the average percentage errors (APEs) of the coefficients compared with the true population values. The APEs are reduced significantly when individuals with $x^1 = 1$ are oversampled compared to when they are sampled to maintain their highly skewed population proportions, demonstrating the importance of considering these impacts for small sample estimation.

The results of this simulation confirm that maintaining a representative sample of the population should not always be the goal when sampling. In fact, in all four of the populations, the sample that is selected to be representative of the population yields the worst results in our simulations. Instead, ensuring that there is enough variation in the exogenous variables in the collected sample, and particularly that there is sufficient representation from minority population groups, is critical to ensuring the accuracy of small sample estimation results. In practice, these results mean that (a) samples that are not representative of the population in terms of the distribution of exogenous variables should not be a concern for researchers if the goal is the estimation of a causal effects model, (b) researchers can intentionally target surveys to underrepresented population groups to ensure that there is sufficient exogenous variation and reduce the necessary sample size to estimate effects for these groups, and (c) these strategies become particularly important when researchers have interest in untangling the effects of multiple correlated exogenous variables that may otherwise pose greater efficiency challenges.

In addition to the model estimates presented in Table 1, Table 2 presents the population share predictions using the estimated coefficients from each model. In the table, each column again represents a population with the specified correlation between the two exogenous variables, while each row-panel indicates a set of samples selected from that population to include the specified proportion of individuals with $x^1 = 1$. Within each row-panel four values are shown. First, the predicted population share is calculated for each sample based on the application of the estimated parameters (using each sampling strategy for each sample) to calculate the probability that each individual in the population (not the estimation sample) is predicted to select the outcome $y = 1$. The average across these individual probabilities then represents the predicted population share for the sample (and the average predicted population share across the 1,000 samples is reported in the table). Second, the APE shown below the predicted population share is the average percentage error of the predicted population share compared with the true population share (shown at the top of the table). Third, the in-sample share is simply the average (across the 1,000 samples) of the proportion of individuals in the sample who selected the outcome $y = 1$. Finally, the last value is the average percentage error between the in-sample share and the true population share.

As mentioned in Section 1, the theoretical results apply to descriptive statistics as well as model coefficients themselves. Since the population shares can be predicted using the estimated coefficients applied to exogenous variable data from the overall population, the predicted shares will be unbiased for all sampling approaches because selection variables are controlled for in the model. This is evidenced in Table 1 in that the population share predictions (see the first row within each row-panel in the table) yield unbiased results regardless of the extent of oversampling or the correlation between the exogenous variables. As also may be observed in the table, the in-sample share is an unbiased estimate of the true population value when the sample is selected to maintain the population proportions of exogenous variables (see the first row-panel of Table 2). As before, this initial case does not include any selection bias, so the basic in-sample share effectively predicts the true population values. However, as oversampling is introduced in the lower rows, the in-sample shares (or equivalently, the model estimates applied *within the sample* to make predication) become increasingly poor estimates of the true population shares, particularly in the population without any correlation between the exogenous variables. The implication is that, for model

estimation of individual relationships with exogenous sampling, the sampling strategy should focus on obtaining a good range of the exogenous variables (not necessarily representativeness), though to make any population share predictions or to compute average treatment effects of variables in the population on the outcome of interest, the estimated relationship needs to be applied to a sample that reasonably represents the population distribution of exogenous variables.

## 4. ENDOGENOUS SELECTION

In contrast to the assumptions of the previous sections, most researchers do not have full control over sample selection and may not observe all variables relevant to selection. Specifically, while researchers can work toward administering surveys to a representative sample of the population based on known exogenous variables, unobserved self-selection effects are still likely to impact responses. For instance, online opinion panels provide platforms to get high response rates and allow some control over representation of respondents in terms of a subset of observed control variables (sociodemographic variables), but opinion panel respondents have been shown to have different attitudes and characteristics than the broader population (Wang et al., 2023). Therefore, while much of the existing literature relies on the assumption that selection is based on observed characteristics, it is increasingly important to develop methods that accommodate unobserved selection effects.

As before, we begin by considering the properties of the unweighted estimator, which can be defined in the same way as Equation (14). However, since the sample selection is informative of the outcome, even after conditioning on exogenous variables, the unweighted estimator will no longer be unbiased. To show this, we will define a $(B\times 1)$ vector of variables $\mathbf{z}_q$ that impact selection, are correlated with the outcome $y_q$, but are not observed by the researcher. Then, as defined in Section 2, we assume that selection is informative of the outcome, even after conditioning on all the observed exogenous variables $\mathbf{x}_q$ ($P(\omega_q = 1 \mid \mathbf{x}_q, y_q) \neq P(\omega_q = 1 \mid \mathbf{x}_q)$), but we can assume that the probability of selection does not depend on the outcome after also conditioning on $\mathbf{z}_q$ ($P(\omega_q = 1 \mid \mathbf{x}_q, \mathbf{z}_q, y_q) = P(\omega_q = 1 \mid \mathbf{x}_q, \mathbf{z}_q)$). Following the same steps as above, we can show that

$$
\begin{aligned}
&E\left[\boldsymbol{s}^u(\boldsymbol{\beta},\mathbf{m})\right] = E\left[\sum_{q=1}^{Q}\omega_q \boldsymbol{s}_q(\boldsymbol{\beta},m_q)\right] = \sum_{q=1}^{Q}E\left[\omega_q \boldsymbol{s}_q(\boldsymbol{\beta},m_q)\right] \\
&= \sum_{q=1}^{Q}E\left[E\left\{\omega_q \boldsymbol{s}_q(\boldsymbol{\beta},m_q)\mid \mathbf{x}_q,\mathbf{z}_q\right\}\right] = \sum_{q=1}^{Q}E\left[E\left\{\omega_q v_q \mid \mathbf{x}_q,\mathbf{z}_q\right\}E\left\{\boldsymbol{s}_q(\boldsymbol{\beta},m_q)\mid \mathbf{x}_q,\mathbf{z}_q\right\}\right] \qquad (25)\\
&= \sum_{q=1}^{Q}E\left[P(\omega_q = 1\mid \mathbf{x}_q,\mathbf{z}_q)E\left\{\boldsymbol{s}_q(\boldsymbol{\beta},m_q)\mid \mathbf{x}_q,\mathbf{z}_q\right\}\right].
\end{aligned}
$$

However, we cannot show that $E\left\{\boldsymbol{s}_q(\boldsymbol{\beta},m_q)\mid \mathbf{x}_q,\mathbf{z}_q\right\} = \mathbf{0}$ because the underlying probability of the random outcome is only conditional on the explanatory variables $\mathbf{x}_q$, not the endogenous selection variables $\mathbf{z}_q$ which are unobserved. Therefore, conditioning the expectation of the score function on $\mathbf{z}_q$ introduces a bias such that $E\left[\boldsymbol{s}^u(\boldsymbol{\beta},\mathbf{m})\right] \neq \mathbf{0}$.

For the asymptotic properties, while draws of $w_q$ in the underlying random sample of individuals in $Q$ yield $\boldsymbol{s}_q(\boldsymbol{\beta}, m_q)$ which are normally distributed with mean zero, this same result does not hold for the unweighted estimator under endogenous sampling. Instead, since $E\left[\boldsymbol{s}^u(\boldsymbol{\beta}, \mathbf{m})\right] \neq \mathbf{0}$ (since sampling in informative on the outcome, as shown above) the limiting distribution of $\boldsymbol{s}_q{}^u(\boldsymbol{\beta}_0, m_q)$ does not have a mean of zero. Instead, as $Q \to \infty$, the unweighted estimator will converge to a value ($\boldsymbol{\beta}^*$) different from the true population value based on the selection probabilities given the endogenous selection variable $\boldsymbol{\beta}^*(P(\omega_q = 1 \mid \mathbf{x}_q, \mathbf{z}_q)) \neq \boldsymbol{\beta}_0$.

Weighting approaches would seem to offer a solution since, as discussed in the previous section, inverse probability weights cancel with the expected value of the selection indicator. Thus, for weights defined as $v_q = 1/P(\omega_q = 1 \mid \mathbf{x}_q, \mathbf{z}_q)$ to represent the true probability of selection, the weighted estimator remains unbiased:

$$
\begin{aligned}
&E\left[\boldsymbol{s}^w(\boldsymbol{\beta}, \mathbf{m})\right] = E\left[\sum_{q=1}^{Q} \omega_q v_q \boldsymbol{s}_q(\boldsymbol{\beta}, m_q)\right] = \sum_{q=1}^{Q} E\left[\omega_q v_q \boldsymbol{s}_q(\boldsymbol{\beta}, m_q)\right] \\
&= \sum_{q=1}^{Q} E\left[E\left\{\omega_q v_q \boldsymbol{s}_q(\boldsymbol{\beta}, m_q) \mid \mathbf{x}_q, \mathbf{z}_q\right\}\right] = \sum_{q=1}^{Q} E\left[E\left\{\omega_q v_q \mid \mathbf{x}_q, \mathbf{z}_q\right\} E\left\{\boldsymbol{s}_q(\boldsymbol{\beta}, m_q) \mid \mathbf{x}_q, \mathbf{z}_q\right\}\right] \\
&= \sum_{q=1}^{Q} E\left[P(\omega_q = 1 \mid \mathbf{x}_q, \mathbf{z}_q) v_q E\left\{\boldsymbol{s}_q(\boldsymbol{\beta}, m_q) \mid \mathbf{x}_q, \mathbf{z}_q\right\}\right] = \sum_{q=1}^{Q} E\left[E\left\{\boldsymbol{s}_q(\boldsymbol{\beta}, m_q) \mid \mathbf{x}_q, \mathbf{z}_q\right\}\right] = \\
&\sum_{q=1}^{Q} E\left[\boldsymbol{s}_q(\boldsymbol{\beta}, m_q)\right] = \sum_{q=1}^{Q} \mathbf{0} = \mathbf{0}.
\end{aligned}
\tag{26}
$$

As shown above, the weighted estimator remains unbiased when the weights can be appropriately calculated to represent the true probability of selection, even in cases when the unweighted estimator is biased. The asymptotic properties, likewise, follow from the steps shown in the previous section, with the asymptotic distribution of the weighted estimator identical to that of Equation (20), where $\mathbf{H}_w$ and $\mathbf{J}_w$ likewise take the same form, with the updated weights $v_q = 1/P(\omega_q = 1 \mid \mathbf{x}_q, \mathbf{z}_q)$ to accommodate endogenous selection.

This result indicates that when weights can be developed to represent the true inverse probability of selection, the weighted estimator can be used to estimate unbiased parameters (Manski and Lerman, 1977; Cosslett, 1981; Wooldridge, 1999; Solon et al., 2015). For instance, as mentioned in Section 1, choice-based sampling is a specific form of endogenous sampling method where sampling is based directly on the outcome $y_q$ such that specific alternatives are over- or under-sampled. In this case, if the population proportions of individuals selecting each alternative are known, then weights can be developed to accommodate this sampling approach, and the weighted estimator will yield unbiased results. For instance, a destination choice model based on a sampling strategy that recruits individuals at specific destinations would be a choice-based sample, and appropriate weights could be developed if the population proportions of individuals selecting each destination were known. In this scenario, the weighted estimator could be used to estimate the model and achieve unbiased results (in this case, the population shares would be needed in advance to develop weights for model estimation rather than being predicted

using the estimated model parameters, but these results could still be applied for additional unbiased predictions, such as calculating treatment effects).

However, as several authors have discussed, weights used in practice are not always based on the true inverse probability of selection. Rather, in the more general case, selection may be based on observed and unobserved variables rather than directly on observed choice values, and weights are often estimated using techniques such as raking or iterative proportional fitting (IPF) that rely only on the observed exogenous variables and their respective population marginal proportions (Gelman, 2007, 2023). Thus, when sampling is also based on unobserved variables (as is generally almost always likely to be the case, or at least it is highly questionable practice if the analyst were to *a priori* preclude such a possibility), these weights will not represent the true inverse probability of selection $v_q = 1/P(\omega_q = 1 | \mathbf{x}_q, \mathbf{z}_q)$, but instead only represent the probability of selection conditional on the observed exogenous variables $\tilde{v}_q = 1/P(\omega_q = 1 | \mathbf{x}_q)$. Such a weighting procedure, therefore, will not accommodate the bias in the estimation because the product $P(\omega_q = 1 | \mathbf{x}_q, \mathbf{z}_q)\tilde{v}_q$ does not equal one and the steps in Equation (26) will not hold. Since, under this sampling design, we cannot show that $E\left[\boldsymbol{s}^w(\boldsymbol{\beta}, \mathbf{m})\right] = \mathbf{0}$, the weighted estimator (along with the unweighted estimator) will yield biased results. Thus, while weighting approaches are effective in situations where selection is based only on observed variables, they are unable to accommodate self-selection effects resulting from participant non-response (see also Brewer and Carlson, 2024).

Since neither the unweighted or weighted approach yields unbiased results, an alternative approach involves using a second selection equation to control for endogenous selection (Heckman, 1979). For instance, in the case of a linear outcome, Heckman applies a two-stage estimation approach, where the selection indicator $\omega_q$ is assumed to be determined using a binary probit framework in the first stage, given by:

$$s_q^* = \boldsymbol{\gamma}' \boldsymbol{d}_q + \eta_q, \tag{27}$$

where $s_q^*$ is the propensity of a respondent to answer a survey if it is presented to them, $\boldsymbol{d}_q$ is an $(H \times 1)$ vector of exogenous variables impacting selection, $\boldsymbol{\gamma}$ is an $(H \times 1)$ column vector of corresponding coefficients to be estimated, and $\eta$ is a standard random normal error term. Then, each individual is assumed to select alternative $s_q = \omega_q$ $(\omega_q \in \{0,1\})$ if $\theta^{\omega_q - 1} < s_q^* < \theta^{\omega_q}$, for $\theta^{-1} = -\infty$, $\theta^0 = 0$, and $\theta^1 = \infty$. As before, $\omega_q$ is the actual binary indicator that determines whether an individual appears in the sample, while the underlying propensities do not appear to the researcher. Collect the elements of $s_q$ and $\omega_q$ into vectors $\boldsymbol{s}$ and, $\boldsymbol{\omega}$ respectively. Of course, this estimator relies on the ability to estimate the first-stage selection equation, implying that exogenous variable data must be available for both those individuals with an observed outcome, and those who are unobserved. Thus, the Heckman approach is more appropriate for censored outcomes than completely missing data due to nonresponse (for the moment, we will maintain this assumption, that we observe exogenous variable data from an underlying random sample of individuals while censoring the outcome endogenously).

In the case of a linear outcome variable, Heckman's two-step estimator can be used to estimate the parameters $\boldsymbol{\gamma}$ from the selection model (in the first stage), and then the parameters $\boldsymbol{\beta}$ (in the second stage) from an adjusted form of the main outcome model with a correction term

based on the inverse Mills ratio (Heckman, 1979). Under the assumption that the errors are jointly normal, this procedure provides a consistent estimator of $\boldsymbol{\beta}$.

However, while the original form of the Heckman two-stage procedure for sample selection has been extended to additional distributional assumptions (Newey, 2009; Liu and Yu, 2022), two additional issues remain with this formulation. First, the two-stage approach relies on the assumption that the vector of variables in the selection model ($\boldsymbol{d}_q$) includes an instrument that is observed by the researcher, strongly influences selection, and is irrelevant in the main outcome equation. Although this exclusion restriction is not strictly required for model identification, identification without this type of instrument relies entirely on the distributional assumptions and the nonlinear relationship between the two estimation equations (see Wolfolds and Siegel, 2019). Thus, when these conditions on the instrument are not met, the Heckman two-stage approach performs poorly (Puhani, 2000; Wolfolds and Siegel, 2019). Second, and more importantly, while this approach works for linear models, the inverse Mills ratio correction term does not extend to the case of nominal outcomes (Greene, 2018; Galimard et al., 2018). As Dubin and Rivers (1989) describe, the general model conceptualization used in the Heckman procedure has straightforward analogies to nonlinear models (including binary outcomes), where the model system can be set up with a binary probit selection equation and a binary probit outcome equation. However, the two-stage approach using the inverse Mills ratio as a correction factor cannot be applied to accommodate the correlations between the error terms in nonlinear outcome models.

Therefore, rather than using a two-stage approach, we use a joint modeling approach that models sample selection and the main outcome in a single stage, allowing for unobserved correlations between the two error terms ($\varepsilon_q$ and $\eta_q$) in the model. This approach maintains the structure of the nominal main outcome decision (as shown in Equation (1)) as well as the structure of the binary probit formulation (as shown in Equation (27)) for sample selection, but allows for direct correlations between the two error terms in a single-stage estimation. Lee (1979) proposed the use of a joint likelihood function of this type, in the case of a switching regression with linear outcomes and a binary selection variable as well as extending the framework to a polychotomous discrete choice selection variable with a continuous outcome (using a copula approach to tie extreme value error terms of the multinomial selection equation to the normal distribution of the continuous outcome equation; see Lee, 1983). The copula approach (for an extended discussion of this approach see Bhat and Eluru, 2009) relaxes the normality assumption on the univariate error terms, while providing a multivariate functional form for the joint distribution of random variables. However, while these approaches consider continuous outcome equations, the same joint approach may be applied to the case of sample selection with a limited dependent main outcome.

Specifically, in our current context of a binary main outcome, consider $\varepsilon_q$ and $\eta_q$ to be jointly normally distributed with a mean vector of zeros and a correlation matrix given by $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. The correlation term $\rho$ captures the error correlations among the underlying latent propensity for the main outcome and sample selection, accommodating the presence of unobserved self-selection effects that influence the outcome. This accounts for the variables $\mathbf{z}_q$ that both impact selection and are informative of the main outcome. Finally, while the parameters of the model system may be identified in this bivariate normal distribution case even without the need for an exclusion restriction when there is at least one common continuous exogenous variable (but not a binary variable) in both the selection and outcome equations, only weak identification is

suggested. In general, an exclusion restriction is still generally a necessary and sufficient condition for identification, and is absolutely necessary for point identification in cases with more flexible error distributions (see Han and Lee, 2019 and Bhat, 2024).

Next, define a vector $\boldsymbol{\delta}$ that collects the parameters to be estimated $\boldsymbol{\delta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}', \rho)'$. Then, using this joint modeling approach, the likelihood for each individual can be written as:

$$L_q(\boldsymbol{\delta}; m_q, \omega_q) = \Pr(y_q = m_q, s_q = \omega_q) = \int_{D_r} f_2(r \mid \boldsymbol{\mu}_q, \boldsymbol{\Sigma}) dr \,, \tag{28}$$

where the integration domain $D_r = \{r : \boldsymbol{\theta}^{low} < r < \boldsymbol{\theta}^{high}\}$ is now the multivariate region of $y_q^*$ and $s_q^*$ truncated by the respective upper and lower thresholds. $f_2(r \mid \boldsymbol{\mu}_q, \Sigma)$ is the bivariate normal density function with a mean of $\boldsymbol{\mu}_q = \begin{bmatrix} \boldsymbol{\beta}'\mathbf{x}_q \\ \boldsymbol{\gamma}'\boldsymbol{d}_q \end{bmatrix}$ and a covariance matrix given by $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. For now, we assume that the exogenous variables for all individuals in the underlying sample of size $Q$ are observed, but the outcome $m_q$ is only observed for individuals with $\omega_q = 1$. Then, this likelihood function applies to all members of the underlying random sample. However, for individuals with unobserved outcomes ($\omega_q = 0$), we can simply marginalize over the main outcome $m_q$, so that only the selection outcome $\omega_q$ is available for estimation, giving the univariate likelihood:

$$L_q(\boldsymbol{\delta}; \omega_q) = \Pr(s_q = \omega_q) = \int_{D_r} f_1(r \mid \boldsymbol{\gamma}'\boldsymbol{d}_q, 1) dr \,. \tag{29}$$

Since unobserved error correlations between the sample selection model and the main outcomes are estimated using the joint approach through the matrix $\boldsymbol{\Sigma}$, the unweighted estimator can be used, giving the likelihood function for the sample:

$$L(\boldsymbol{\delta}; \mathbf{m}, \boldsymbol{\omega}) = \prod_{q=1}^{Q} L_q(\boldsymbol{\delta}; m_q, \omega_q)^{\omega_q} L_q(\boldsymbol{\delta}; \omega_q)^{1-\omega_q} \,. \tag{30}$$

Under the usual regularity conditions needed for likelihood objects, the logarithm of the joint likelihood function can again be maximized through solving the corresponding first-order (score) equations, now defined as:

$$\boldsymbol{s}(\boldsymbol{\delta}; \mathbf{m}, \boldsymbol{\omega}) = \nabla \log L(\boldsymbol{\delta}; \mathbf{m}, \boldsymbol{\omega}) = \sum_{q=1}^{Q} \boldsymbol{s}_q(\boldsymbol{\delta}; m_q, \omega_q) = \mathbf{0}, \tag{31}$$

where $\boldsymbol{s}_q(\boldsymbol{\delta}; m_q, \omega_q) = \omega_q \dfrac{\partial \log L_q(\boldsymbol{\delta}; m_q, \omega_q)}{\partial \boldsymbol{\delta}} + (1 - \omega_q) \dfrac{\partial \log L_q(\boldsymbol{\delta}; \omega_q)}{\partial \boldsymbol{\delta}}$. The first component of the score function corresponds to those individuals with an observed outcome (where the joint likelihood function is applied), while the second component corresponds to those individuals with no observed outcome (where the marginalized likelihood function is applied). Then, to show that the estimator is unbiased we can show that

$$
\begin{aligned}
E\left[\boldsymbol{s}(\boldsymbol{\delta};\mathbf{m},\boldsymbol{\omega})\right] &= E\left[\sum_{q=1}^{Q}\boldsymbol{s}_q(\boldsymbol{\delta};m_q,\omega_q)\right] = \sum_{q=1}^{Q}E\left[\omega_q\frac{\partial\log L_q(\boldsymbol{\delta};m_q,\omega_q)}{\partial\boldsymbol{\delta}}+(1-\omega_q)\frac{\partial\log L_q(\boldsymbol{\delta};\omega_q)}{\partial\boldsymbol{\delta}}\right] \\
&= \sum_{q=1}^{Q}E\left[\omega_q\frac{\partial\log L_q(\boldsymbol{\delta};m_q,\omega_q)}{\partial\boldsymbol{\delta}}\right]+E\left[(1-\omega_q)\frac{\partial\log L_q(\boldsymbol{\delta};\omega_q)}{\partial\boldsymbol{\delta}}\right] \\
&= \sum_{q=1}^{Q}E\left[\omega_q\frac{\partial\log L_q(\boldsymbol{\delta};\omega_q)}{\partial\boldsymbol{\delta}}\right]+E\left[\omega_q\frac{\partial\log L_q(\boldsymbol{\delta};m_q\mid\omega_q)}{\partial\boldsymbol{\delta}}\right]+E\left[(1-\omega_q)\frac{\partial\log L_q(\boldsymbol{\delta};\omega_q)}{\partial\boldsymbol{\delta}}\right] \\
&= \sum_{q=1}^{Q}E\left[\frac{\partial\log L_q(\boldsymbol{\delta};\omega_q)}{\partial\boldsymbol{\delta}}\right]+P(\omega_q=1)E\left[\frac{\partial\log L_q(\boldsymbol{\delta};m_q\mid\omega_q)}{\partial\boldsymbol{\delta}}\mid\omega_q=1\right]
\end{aligned}
\tag{32}
$$

where the ability to split the joint likelihood function into the product of a conditional likelihood and marginal likelihood relies on the appropriate specification of the correlation term $\rho$ estimated using the data from those with both outcomes. Then, we can show that each of these expected value terms equal zero to prove that the joint estimator is unbiased:

$$
\begin{aligned}
E\left[\boldsymbol{s}(\boldsymbol{\delta};\mathbf{m},\boldsymbol{\omega})\right] &= \sum_{q=1}^{Q}E\left[\frac{\partial\log L_q(\boldsymbol{\delta};\omega_q)}{\partial\boldsymbol{\delta}}\right]+P(\omega_q=1)E\left[\frac{\partial\log L_q(\boldsymbol{\delta};m_q\mid\omega_q)}{\partial\boldsymbol{\delta}}\mid\omega_q=1\right] \\
&= \sum_{q=1}^{Q}\sum_{\omega_q\in\{0,1\}}\frac{\partial\log L_q(\boldsymbol{\delta};\omega_q)}{\partial\boldsymbol{\delta}}P(s_q=\omega_q)+P(\omega_q=1)\sum_{m_q\in\{0,1\}}\frac{\partial\log L_q(\boldsymbol{\delta};m_q\mid\omega_q=1)}{\partial\boldsymbol{\delta}}P(y_q=m_q\mid\omega_q=1) \\
&= \sum_{q=1}^{Q}\sum_{\omega_q\in\{0,1\}}\frac{\partial L_q(\boldsymbol{\delta};\omega_q)}{\partial\boldsymbol{\delta}}+P(\omega_q=1)\sum_{m_q\in\{0,1\}}\frac{\partial L_q(\boldsymbol{\delta};m_q\mid\omega_q=1)}{\partial\boldsymbol{\delta}} \\
&= \sum_{q=1}^{Q}\frac{\partial}{\partial\boldsymbol{\delta}}\sum_{\omega_q\in\{0,1\}}L_q(\boldsymbol{\delta};\omega_q)+P(\omega_q=1)\frac{\partial}{\partial\boldsymbol{\delta}}\sum_{m_q\in\{0,1\}}L_q(\boldsymbol{\delta};m_q\mid\omega_q=1) \\
&= \sum_{q=1}^{Q}\frac{\partial}{\partial\boldsymbol{\delta}}(1)+P(\omega_q=1)\frac{\partial}{\partial\boldsymbol{\delta}}(1)=\mathbf{0}.
\end{aligned}
\tag{33}
$$

Thus, when the exogenous variables for the entire random sample of $Q$ individuals are known, this estimator returns unbiased estimates of both selection and the outcome of interest, even when marginalizing over nonrandom selection for the main outcome.

Finally, this approach, as described above, relies on the availability of exogenous variable data from all individuals in the underlying random sample, even those without an observed outcome. This is a reasonable assumption for some applications, such as accounting for attrition in panel data models or when demographics are known for all potential respondents in the sampling frame. In many cases, however, the exogenous variables are unknown for individuals who choose not to participate, and many existing approaches do not account for these scenarios (Galimard et al., 2018; Brewer and Carlson, 2024). We address these scenarios where the exogenous data is unavailable for nonrespondents using an additional step to generate an augmented sample based on the observed characteristics of individuals in the sample, the size of the sampling frame, and the known population marginal distribution of exogenous variables. In such situations, an IPF process can be used to generate an augmented sample that includes the real data for all individuals

in the observed sample $\tilde{Q}$ and exogenous variable data for a hypothetical set of individuals who are assumed to be unwilling to respond (Choupani and Mamdoohi, 2016; Durán-Heras et al., 2018). To do so, we first use IPF to generate a hypothetical augmented sample that is representative (in terms of the exogenous variables) of the real population $W$ and matches the size of the sampling frame $Q$, using the population marginal distributions of exogenous variables and multi-way distribution table of exogenous variables in the sample $\tilde{Q}$ (and assign each generated individual a value of $\omega_q = 0$). Then, each individual in the real sample $\tilde{Q}$ replaces one individual from the generated augmented sample that has a corresponding set of exogenous variables, taking on the value $\omega_q = 1$ and their real observed main outcome $m_q$. Essentially, this process recovers a final augmented sample that is representative in terms of the population marginal distribution of the exogenous variables, with the unobserved segment of the sample being representative of nonrespondents. Then, jointly modeling the main outcome (available for only those in the observed sample) with sample selection (using the generated individuals as nonrespondents) will result in an improved estimate of the true parameters.

To understand the properties of the estimator using this method, we can use the same process as shown in Equation (32). However, while there is no change to the estimator for those individuals with $\omega_q = 1$, we must consider the possibility that the generated individuals with $\omega_q = 0$ are no longer representative of the population (and therefore the distribution of the outcome $s_q$ given $\mathbf{x}_q$ may no longer match that of the population). Specifically, the expected value of the score function is now

$$\begin{aligned} E\left[\boldsymbol{s}(\boldsymbol{\delta};\mathbf{m},\boldsymbol{\omega})\right] &= E\left[\sum_{q=1}^{Q}\boldsymbol{s}_q(\boldsymbol{\delta};m_q,\omega_q)\right] = \sum_{q=1}^{Q}E\left[\omega_q\frac{\partial\log L_q(\boldsymbol{\delta};m_q,\omega_q)}{\partial\boldsymbol{\delta}} + (1-\omega_q)\frac{\partial\log \hat{L}_q(\boldsymbol{\delta};\omega_q)}{\partial\boldsymbol{\delta}}\right] \\ &= \sum_{q=1}^{Q}E\left[\omega_q\frac{\partial\log L_q(\boldsymbol{\delta};m_q,\omega_q)}{\partial\boldsymbol{\delta}}\right] + E\left[(1-\omega_q)\frac{\partial\log \hat{L}_q(\boldsymbol{\delta};\omega_q)}{\partial\boldsymbol{\delta}}\right] \\ &= \sum_{q=1}^{Q}E\left[\omega_q\frac{\partial\log L_q(\boldsymbol{\delta};\omega_q)}{\partial\boldsymbol{\delta}} + (1-\omega_q)\frac{\partial\log \hat{L}_q(\boldsymbol{\delta};\omega_q)}{\partial\boldsymbol{\delta}}\right] + P(\omega_q=1)E\left[\frac{\partial\log L_q(\boldsymbol{\delta};m_q\mid\omega_q)}{\partial\boldsymbol{\delta}}\mid\omega_q=1\right] \end{aligned} \tag{34}$$

where $\hat{L}_q(\boldsymbol{\delta};\omega_q)$ is the marginal likelihood of selection for each unobserved individual. The only difference between this term and the original likelihood of selection $L_q(\boldsymbol{\delta};\omega_q)$ is that we now have estimated values $\hat{\mathbf{x}}_q$ that replace the true values $\mathbf{x}_q$ for these unobserved individuals. If $\hat{L}_q(\boldsymbol{\delta};\omega_q) = L_q(\boldsymbol{\delta};\omega_q)$ then the remaining steps of Equations (32) and (33) hold, and this estimator is unbiased. This will be true if the distribution of $\hat{\mathbf{x}}_q$ matches the distribution of $\mathbf{x}_q$ for nonrespondents in the population (because then, the likelihood of selection given a value of $\mathbf{x}_q$ is known). However, using the IPF algorithm only guarantees convergence to the population marginal distribution, while maintaining the contingency table least distinguishable from the sample, rather than the population joint distribution of the exogenous variables (Ireland and Kullback, 1968; Beckman et al., 1996). Therefore, while this procedure yields better results than ignoring the unobserved selection effects (which amounts to an assumption that the likelihood of selection given $\mathbf{x}_q$ is always one), it is not guaranteed to be unbiased. The following simulation examines the characteristics of each of these estimators under endogenous selection,

demonstrating that even though this method is not unbiased, it results in significant improvements over independent models if the exogenous variables of nonrespondents are unknown.

### 4.1 Endogenous Selection Simulation Design

To evaluate the effects of weighting approaches and the joint modeling approach under endogenous selection, we undertake a second simulation exercise based on the same binary probit model. For this simulation, a new population of 100,000 individuals is generated using the same approach as the previous simulation (for the purpose of this simulation we consider only a single population, see the top row of Figure 5 labeled "Generate Population Exogenous Data"). In this case, however, thresholds are now applied to the underlying (uncorrelated) standard normal variables for the exogenous variables $\mathbf{x}_q$ such that there is a 40% chance that $x^1 = 1$ and 50% chance that $x^2 = 1$. In addition, for the selection model, the vector of explanatory variables $\boldsymbol{d}_q$ is assumed to include both $x^1$ and $x^2$, as well as a third exogenous variable $r^1$ which is drawn from an independent standard normal distribution (and included as a continuous variable). The variable $r^1$ satisfies the exclusion restriction as it is informative of selection and included in the selection equation but is not informative of the main outcome. Values of the random error term $\varepsilon_q$ for the main outcome and $\eta_q$ for the selection model were drawn from a bivariate random normal distribution with correlation $\rho = 0.5$. Finally, the latent propensity $y_q^*$ is calculated based on Equation (1) using the same coefficients used in the previous simulation ($\beta_0 = -0.75$, $\beta_1 = 0.50$, and $\beta_2 = -0.50$) and the latent propensity of selection $s_q^*$ is calculated based on Equation (27) using the coefficients $\gamma_0 = -0.50$ (for the constant), $\gamma_1 = 1.00$ (corresponding to $x^1$), $\gamma_2 = -0.50$ (corresponding to $x^2$) and $\gamma_3 = 0.50$ (corresponding to $r^1$). Appropriate thresholds are applied to get the binary outcome $y_q$ and binary selection indicator $s_q$ (see the second row of Figure 5 labeled "Generate Outcome Data").

Next, 1,000 independent samples are drawn from the population. For each sample, 5,000 individuals are selected at random from the entire population (including those with $s_q = 0$ and $s_q = 1$). This approach amounts to offering the survey to a random sample of 5,000 individuals in the population, some of whom will choose to respond based on their value of $s_q$ (using the coefficients defined above, this results in approximately 2,000 individuals, or 40%, choosing to respond to the survey), and all of these individuals become part of the underlying random sample $Q$ (see the third section of Figure 5 labeled "Select Underlying Random Sample"). Using these samples of 5,000 individuals, four models are considered (see the fourth section of Figure 5 labeled "Run Models"). First, an unweighted binary probit model is run using only those individuals from the sample with $s_q = 1$ (those with an observed outcome; approximately 2,000 individuals in each sample). Second, a weighted binary probit model is run using only those individuals with $s_q = 1$ and weights generated based on the respondent values of $x^1$, $x^2$ and $r^1$ (weights cannot be generated based on the endogenous selection variable in this case, because it is unobserved by the researcher). Third, a joint binary probit model for the main outcome and sample selection is run using all 5,000 individuals but suppressing the outcome $y_q$ for those individuals with $s_q = 0$.

Finally, the last model is another joint model for the main outcome and selection, but where no data (exogenous or endogenous) is observed from individuals with $s_q = 0$. Instead, an IPF method is used to generate new exogenous data for individuals with $s_q = 0$ based on the distribution of respondents with $s_q = 1$ and the known population marginal distributions of $x^1$, $x^2$ and $r^1$ (see the far right side of Figure 5, where the section labeled "Run Models" is split into (a) generate hypothetical augmented sample using IPF and (b) run joint binary probit model using augmented sample).

As before, the results from each model are stored to evaluate the performance of each modeling strategy (see the final row of Figure 5 labeled "Evaluate Performance"). The same metrics are used to evaluate the performance of these models under the endogenous selection scenario as used in the previous simulation. Further, the estimates are again used to predict the share of the population selecting $y = 1$. In this case, the estimated likelihood functions for the unweighted binary probit model, the weighted binary probit model, and the joint binary probit model estimated with the known sampling frame are applied to the exogenous variable data from the underlying random sample of 5,000 individuals selected in each draw (regardless of whether each individual chose to respond to the survey) to calculate the probability of each individual selecting the outcome $y = 1$ in the underlying random sample (not the estimation sample). For the joint binary probit model estimated with the unknown sampling frame, the estimated likelihood function is applied to the augmented sample generated using IPF for estimation (we continue to assume at this stage that the exogenous variable data for the underlying random sample is unavailable) to calculate the probability of each individual selecting the outcome $y = 1$ in this augmented sample (not the estimation sample). The average across the likelihood predictions for these 5,000 individuals (either the underlying random sample or augmented sample) is calculated for each model to represent the predicted population share selecting the outcome $y = 1$.

### 4.2 Endogenous Selection Simulation Results

Table 3 presents the performance results of the four models in this simulation. In the table, each row-panel represents a modeling approach while each column represents the values for each of the variables in the model (including the main outcome equation, the selection model, and the correlation term). As seen in the table, neither the unweighted nor weighted models (see the first two row-panels in Table 3) are consistent when estimated independently using only the sample of observed individuals. In this case, since the selection variable is unobserved, it is not conditioned for in the model and the weights that are based only on the exogenous variables are unable to accommodate the selection bias, so both models are biased. This has important implications, confirming that weighting is not an appropriate strategy to address sample selection biases when they are due to unobserved self-selection effects.

In contrast, when the sampling frame is assumed to be known, the parameters can be consistently estimated with the joint binary probit model shown in the third row-panel. As expected, this joint model consistently estimates the main outcome, as well as the selection model parameters and correlation. Thus, accounting for the correlation between selection and the main outcome effectively eliminates the sample selection bias and allows the model to be estimated consistently. This approach does, however, rely on the exogenous variable data from the entire underlying random sample of individuals, even when the outcome is unobserved. Therefore, the fourth row-panel gives the results if the sampling frame is assumed to be unknown, generating an augmented population based on the known population marginals to replace the unknown

nonresponding individuals. Although the performance of the joint model in this case is worse than when the sampling frame is known, the estimates are significantly superior to those of the independent models (and much of the reduction in performance is in the selection equation rather than for the main outcome). Thus, even if not all bias is removed, the joint modeling approach still represents a significant improvement over an assumption of sampling on observables.

From a modeling perspective, these results highlight the importance of considering self-selection effects when unobserved factors may influence survey response, as significant biases occur based on the unobserved self-selection effects. From a sampling perspective, these results indicate that researchers can accommodate self-selection when sufficient information is known about the underlying population or an underlying representative sample. One implication is that researchers can accommodate situations when individuals choose not to respond to specific survey questions based on unobserved factors, as long as unobserved factors don't influence response rates to the survey overall. This means that, particularly when asking about sensitive data, allowing participants to choose not to answer specific questions may be preferable to forcing responses and having participants abandon the survey entirely (as this still allows for the collection of the exogenous variable data for these participants). However, even when unobserved endogenous factors influence sample selection and no data is observed for those who respond, researchers can use the IPF population generation techniques described above to improve estimation relative to ignoring these sample selection biases. Also, since consistency can be proven when the joint distribution of exogenous variables in the population is known, the collection of population joint distributions of common exogenous variables would be beneficial.

In addition to the model estimates themselves, the share predictions of the individuals in the population selecting $y=1$ based on the results of each model are shown in Table 4. In the table, the true population share is shown in the first row, followed by the average in-sample share (that is, the proportion of individuals, in the underlying random sample of 5,000 individuals and with $s_q=1$, who selected the outcome $y=1$, averaged across the 1,000 samples) in the second row. The following four rows show the share predictions using each of the four models under consideration. These shares are predicted for each sample based on the application of the estimated parameters (for each modeling approach and each of the 1,000 samples) to calculate the probability of each individual in the underlying random sample (or augmented sample in the case where underlying exogenous variable data is assumed to be unobserved) predicted to select the outcome $y=1$. The average across these individual probabilities for each modeling approach then represents the respective predicted population share for the sample (and the average predicted population share for each modeling approach across the 1,000 samples is reported in the table). Then, for each of these prediction mechanisms, the average percentage error between the prediction and the true population share (shown at the top of the table) is also shown in the final column.

As may be observed in the table, the in-sample share is an extremely poor prediction of the true population share as the sampling mechanism is biased. Model-based estimates using either a weighted or unweighted approach do not accommodate these sampling biases. In fact, both of these approaches overestimate the population share by nearly as much as the in-sample share. Notably, the poor performance of these models implies that calculating descriptive statistics based on weights can be a poor reflection of population statistics, aligning with the recent findings of Brewer and Carlson (2024) for the case of a linear regression, who found that weighting schemes can actually worsen predictions when endogenous selection effects are present. On the other hand, the share predictions based on the model results from the joint modeling approaches both yield

significantly improved predictions of the population share. As expected, the model with the known sampling frame performs slightly better than the model with the generated sampling frame, but both models result in unbiased predictions of the population shares.

## 5. CONCLUSIONS

The current research has several practical implications for researchers in terms of survey design and dissemination as well as for modeling practices. We show that weighting approaches can adversely affect model efficiency under exogenous sampling, and unweighted estimators should be used in these cases. Further, given that exogenous sampling strategies are consistent, researchers should work to improve the efficiency of their estimates in the survey dissemination process by intentionally sampling to ensure that there is sufficient exogenous variation in the sample. These results demonstrate the close connection between sampling considerations and modeling approaches, highlighting the need to carefully integrate these processes and consider the modeling impacts of various sampling approaches.

In the case of endogenous sampling, we show that selection on unobserved variables cannot be addressed using weighting approaches, which are only effective when sampling is based on observed variables. In particular, our results caution against assuming a priori that sample selection is based solely on observed variables, because such an assumption can lead even descriptive statistics based on weights to be poor reflections of population statistics. In this context, our investigation strongly advocates for the use of sample selection models that accommodate unobserved self-selection effects. We show that jointly modeling sample selection with the main outcome can accommodate unobserved correlation effects and lead to unbiased results, when the exogenous variables of an underlying representative sample are known. Finally, our proposed method of generating an augmented exogenous population using IPF when this underlying sample is not observed also results in significantly improved results in simulation efforts both in terms of model coefficients and prediction of population statistics.

Overall, while this paper provides detailed theoretical and simulation-based support for our findings, there are several avenues for additional research. First, in terms of intentional sampling techniques, more research is needed to further quantify the efficiency gains of sampling techniques that move away from representative sampling to improve the variation of variables in the modeling effort. Additionally, although we provide intuition as well as simulation to examine improvements under endogenous sampling using our proposed joint modeling approach with an IPF augmented sample, consistency is not guaranteed. Further work should explore more robust population generation techniques with provable asymptotic properties to develop consistent estimators for samples with unobserved endogenous selection.

## ACKNOWLEDGEMENTS

This research was partially supported by the U.S. Department of Transportation through the Center for Understanding Future Travel Behavior and Demand (TBD) (Grant No. 69A3552344815 and No. 69A3552348320). The author is grateful to Lisa Macias for help in formatting this document.

## REFERENCES

Alhassan, V.O., Yu, F., Dimas Valle, J.R., Magassy, T.B., Batur, I., Salon, D., Bhat, C.R., Pendyala, R.M., 2024. Investigating the Influence of Alternative Survey Participant Recruitment Strategies on Measurement and Inference of Mobility Patterns. Technical paper, School of Sustainable Engineering and the Built Environment, Arizona State University.

Avery, L., Rotondi, N., McKnight, C., Firestone, M., Smylie, J., Rotondi, M., 2019. Unweighted Regression Models Perform Better Than Weighted Regression Techniques for Respondent-Driven Sampling Data: Results from a Simulation Study. BMC Medical Research Methodology 19, 202. https://doi.org/10.1186/s12874-019-0842-5

Becker, J.-M., Ismail, I.R., 2016. Accounting for Sampling Weights in PLS Path Modeling: Simulations and Empirical Examples. European Management Journal 34(6), 606–617. https://doi.org/10.1016/j.emj.2016.06.009

Beckman, R.J., Baggerly, K.A., McKay, M.D., 1996. Creating Synthetic Baseline Populations. Transportation Research Part A 30(6), 415–429. https://doi.org/10.1016/0965-8564(96)00004-3

Bhat, C.R., 2024. Transformation-Based Flexible Error Structures for Choice Modeling. Journal of Choice Modelling 53, 100522. https://doi.org/10.1016/j.jocm.2024.100522

Bhat, C.R., 2015. A Comprehensive Dwelling Unit Choice Model Accommodating Psychological Constructs Within a Search Strategy for Consideration Set Formation. Transportation Research Part B 79, 161–188. https://doi.org/10.1016/j.trb.2015.05.021

Bhat, C.R., 2014. The Composite Marginal Likelihood (CML) Inference Approach with Applications to Discrete and Mixed Dependent Variable Models. Foundations and Trends in Econometrics 7(1), 1–117. https://doi.org/10.1561/0800000022

Bhat, C.R., Eluru, N., 2009. A Copula-Based Approach to Accommodate Residential Self-Selection Effects in Travel Behavior Modeling. Transportation Research Part B 43(7), 749–765. https://doi.org/10.1016/j.trb.2009.02.001

Biemer, P.P., Christ, S., 2008. Weighting Survey Data, in: International Handbook of Survey Methodology. Routledge.

Bogomolov, Y., He, M., Khulbe, D., Sobolevsky, S., 2021. Impact of Income on Urban Commute Across Major Cities in Us. Procedia Computer Science, 10th International Young Scientists Conference in Computational Science, YSC2021 193, 325–332. https://doi.org/10.1016/j.procs.2021.10.033

Bollen, K.A., Biemer, P.P., Karr, A.F., Tueller, S., Berzofsky, M.E., 2016. Are Survey Weights Needed? A Review of Diagnostic Tests in Regression Analysis. Annual Review of Statistics and Its Application 3, 375–392. https://doi.org/10.1146/annurev-statistics-011516-012958

Boto-García, D., 2023. Good Results Come to Those Who Weight: On the Importance of Sampling Weights in Empirical Research Using Survey Data. Current Issues in Tourism 27, 268–287. https://doi.org/10.1080/13683500.2023.2178394

Brewer, D., Carlson, A., 2024. Addressing Sample Selection Bias for Machine Learning Methods. Journal of Applied Econometrics 39, 383–400. https://doi.org/10.1002/jae.3029

Choupani, A.-A., Mamdoohi, A.R., 2016. Population Synthesis Using Iterative Proportional Fitting (IPF): A Review and Future Research. Transportation Research Procedia, International Conference on Transportation Planning and Implementation Methodologies for Developing Countries (12th TPMDC) Selected Proceedings, IIT Bombay, Mumbai, India, 10-12 December 2014 17, 223–233. https://doi.org/10.1016/j.trpro.2016.11.078

Cosslett, S.R., 1981. Maximum Likelihood Estimator for Choice-Based Samples. Econometrica 49, 1289–1316. https://doi.org/10.2307/1912755

Couper, M.P., 2017. New Developments in Survey Data Collection. Annual Review of Sociology 43, 121–145. https://doi.org/10.1146/annurev-soc-060116-053613

Demidenko, E., 2001. Computational Aspects of Probit Model. Mathematical Communications 6, 233–247.
Dubin, J.A., Rivers, D., 1989. Selection Bias in Linear Regression, Logit and Probit Models. Sociological Methods & Research 18, 360–390. https://doi.org/10.1177/0049124189018002006
Durán-Heras, A., García-Gutiérrez, I., Castilla-Alcalá, G., 2018. Comparison of Iterative Proportional Fitting and Simulated Annealing as Synthetic Population Generation Techniques: Importance of the Rounding Method. Computers, Environment and Urban Systems 68, 78–88. https://doi.org/10.1016/j.compenvurbsys.2017.11.001
Elwert, F., Winship, C., 2014. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. Annual Review of Sociology 40, 31–53. https://doi.org/10.1146/annurev-soc-071913-043455
Fisher, R.A., 1922. On the Mathematical Foundations of Theoretical Statistics. Philosophical Transactions of the Royal Society of London, Containing Papers of a Mathematical or Physical Character 222, 309–368. https://doi.org/10.1098/rsta.1922.0009
Frohlich, N., Carriere, K.C., Potvin, L., Black, C., 2001. Assessing Socioeconomic Effects on Different Sized Populations: To Weight or Not to Weight? Journal of Epidemiology & Community Health 55, 913–920. https://doi.org/10.1136/jech.55.12.913
Galimard, J.-E., Chevret, S., Curis, E., Resche-Rigon, M., 2018. Heckman Imputation Models for Binary or Continuous MNAR Outcomes and MR Predictors. BMC Medical Research Methodology 18, 90. https://doi.org/10.1186/s12874-018-0547-1
Gary, S., Lenhard, W., Lenhard, A., Herzberg, D., 2023. A Tutorial on Automatic Post-Stratification and Weighting in Conventional and Regression-Based Norming of Psychometric Tests. Behavior Research Methods 56, 4632–4642. https://doi.org/10.3758/s13428-023-02207-0
Gelman, A., 2023. Unifying Design-Based and Model-Based Sampling Inference by Estimating a Joint Population Distribution for Weights and Outcomes. Presented at the Joint Statistical Meetings, American Statstical Association, Toronto.
Gelman, A., 2007. Struggles with Survey Weighting and Regression Modeling. Statistical Science 22(2), 153–164.
Gluschenko, K., 2018. Measuring Regional Inequality: To Weight or Not to Weight? Spatial Economic Analysis 13(1), 36–59. https://doi.org/10.1080/17421772.2017.1343491
Godambe, V.P., 1960. An Optimum Property of Regular Maximum Likelihood Estimation. The Annals of Mathematical Statistics 31(4), 1208–1211.
Greene, W., 2018. Econometric Analysis, 8th ed. Pearson Education.
Han, S., Lee, S., 2019. Estimation in a Generalization of Bivariate Probit Models with Dummy Endogenous Regressors. Journal of Applied Econometrics 34, 994–1015. https://doi.org/10.1002/jae.2727
Hausman, J.A., Wise, D.A., 1981. Stratification on Endogenous Variables and Estimation: The Gary Income Maintenance Experiement, in: Manski, C.F., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge, MA, pp. 365–391.
Heckman, J.J., 1979. Sample Selection Bias as a Specification Error. Econometrica 47(1), 153–161. https://doi.org/10.2307/1912352

Howe, C.J., Cole, S.R., Lau, B., Napravnik, S., Eron, J.J.J., 2016. Selection Bias Due to Loss to Follow Up in Cohort Studies. Epidemiology 27(1), 91–97. https://doi.org/10.1097/EDE.0000000000000409

Hudson, D., Seah, L.-H., Hite, D., Haab, T., 2004. Telephone Presurveys, Self-Selection, and Non-Response Bias to Mail and Internet Surveys in Economic Research. Applied Economics Letters 11(4), 237–240. https://doi.org/10.1080/13504850410001674876

Ireland, C.T., Kullback, S., 1968. Contingency Tables with Given Marginals. Biometrika 55(1), 179–188. https://doi.org/10.1093/biomet/55.1.179

Kish, L., Frankel, M.R., 1974. Inference from Complex Samples. Journal of the Royal Statistical Society: Series B (Methodological) 36(1), 1–22. https://doi.org/10.1111/j.2517-6161.1974.tb00981.x

Lee, L.-F., 1983. Generalized Econometric Models with Selectivity. Econometrica 51(2), 507–512. https://doi.org/10.2307/1912003

Lee, L.-F., 1979. Identification and Estimation in Binary Choice Models with Limited (Censored) Dependent Variables. Econometrica 47(4), 977–996. https://doi.org/10.2307/1914142

Liévanos, R.S., Lubitow, A., McGee, J.A., 2019. Misrecognition in a Sustainability Capital: Race, Representation, and Transportation Survey Response Rates in the Portland Metropolitan Area. Sustainability 11(16), 4336. https://doi.org/10.3390/su11164336

Liu, R., Yu, Z., 2022. Sample Selection Models with Monotone Control Functions. Journal of Econometrics 226(2), 321–342. https://doi.org/10.1016/j.jeconom.2021.01.010

Manski, C.F., Lerman, S.R., 1977. The Estimation of Choice Probabilities from Choice Based Samples. Econometrica 45(8), 1977–1988. https://doi.org/10.2307/1914121

Manski, C.F., McFadden, D., 1981. Alternative Estimators and Sample Designs for Discrete Choice Analysis, in: Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge, MA, pp. 2–50.

Molenberghs, G., Verbeke, G., 2005. Models for Discrete Longitudinal Data, Springer Series in Statistics. Springer, New York ; London.

Murphy, K.M., Topel, R.H., 2002. Estimation and Inference in Two-Step Econometric Models. Journal of Business & Economic Statistics 20(1), 88–97. https://doi.org/10.1198/073500102753410417

Newey, W.K., 2009. Two-Step Series Estimation of Sample Selection Models. The Econometrics Journal 12, S217–S229. https://doi.org/10.1111/j.1368-423X.2008.00263.x

Nguyen, N.D., Murphy, P., 2015. To Weight or Not To Weight? A Statistical Analysis of How Weights Affect the Reliability of the Quarterly National Household Survey for Immigration Research in Ireland. The Economic and Social Review 46(4), 567–603.

Pendyala, R.M., Goulias, K.G., Kitamura, R., 1991. Impact of Telecommuting on Spatial and Temporal Patterns of Household Travel. Transportation 18, 383–409. https://doi.org/10.1007/BF00186566

Pfeffermann, D., 1993. The Role of Sampling Weights When Modeling Survey Data. International Statistical Review 61(2), 317–337. https://doi.org/10.2307/1403631

Puhani, P., 2000. The Heckman Correction for Sample Selection and Its Critique. Journal of Economic Surveys 14(1), 53–68. https://doi.org/10.1111/1467-6419.00104

Rose, J.M., Bliemer, M.C.J., 2013. Sample Size Requirements for Stated Choice Experiments. Transportation 40, 1021–1041. https://doi.org/10.1007/s11116-013-9451-z

Solon, G., Haider, S.J., Wooldridge, J.M., 2015. What Are We Weighting For? Journal of Human Resources 50(2), 301–316. https://doi.org/10.3368/jhr.50.2.301
Tchetgen, E.J.T., Glymour, M.M., Shpitser, I., Weuve, J., 2012. Rejoinder: To Weight or Not to Weight? On the Relation Between Inverse-Probability Weighting and Principal Stratification for Truncation by Death. Epidemiology 23(1), 132–137.
Terawaki, T., Kuriyama, K., Yoshida, K., 2003. The Importance of Excluding Unrealistic Alternatives in Choice Experiment Designs. Discussion Paper No. 03002, College of Economics, Ritsumeikan University.
Thill, J.-C., Horowitz, J.L., 1997. Modelling Non-Work Destination Choices with Choice Sets Defined by Travel-Time Constraints, in: Fischer, M.M., Getis, A. (Eds.), Recent Developments in Spatial Analysis. Springer, Berlin, Heidelberg, pp. 186–208. https://doi.org/10.1007/978-3-662-03499-6_10
Tripathi, G., 1999. A Matrix Extension of the Cauchy-Schwarz Inequality. Economics Letters 63(1), 1–3. https://doi.org/10.1016/S0165-1765(99)00014-2
Wang, F., Wang, H., Yan, J., 2023. Diagnostic Tests for the Necessity of Weight in Regression With Survey Data. International Statistical Review 91(1), 55–71. https://doi.org/10.1111/insr.12509
Wang, X., Shaw, F.A., Mokhtarian, P.L., Circella, G., Watkins, K.E., 2023. Combining Disparate Surveys Across Time to Study Satisfaction with Life: The Effects of Study Context, Sampling Method, and Transport Attributes. Transportation 50, 513–543. https://doi.org/10.1007/s11116-021-10252-x
Winship, C., Radbill, L., 1994. Sampling Weights and Regression Analysis. Sociological Methods & Research 23(2), 230–257. https://doi.org/10.1177/0049124194023002004
Wittwer, R., Hubrich, S., Gerike, R., 2024. New Evidence on Nonresponse in Household Travel Surveys. Transportation Research Procedia, 12th International Conference on Transport Survey Methods 76, 233–245. https://doi.org/10.1016/j.trpro.2023.12.051
Wolfolds, S.E., Siegel, J., 2019. Misaccounting for Endogeneity: The Peril of Relying on the Heckman Two-Step Method Without a Valid Instrument. Strategic Management Journal 40(3), 432–462. https://doi.org/10.1002/smj.2995
Wooldridge, J.M., 2007. Inverse Probability Weighted Estimation for General Missing Data Problems. Journal of Econometrics 141(2), 1281–1301. https://doi.org/10.1016/j.jeconom.2007.02.002
Wooldridge, J.M., 2002. Inverse Probability Weighted M-Estimators for Sample Selection, Attrition, and Stratification. Portuguese Economic Journal 1, 117–139. https://doi.org/10.1007/s10258-002-0008-x
Wooldridge, J.M., 2001. Asymptotic Properties of Weighted M-Estimators for Standard Stratified Samples. Econometric Theory 17(2), 451–470. https://doi.org/10.1017/S0266466601172075
Wooldridge, J.M., 1999. Asymptotic Properties of Weighted M-Estimators for Variable Probability Samples. Econometrica 67(6), 1385–1406. https://doi.org/10.1111/1468-0262.00083
Yi, G.Y., Zeng, L., Cook, R.J., 2011. A Robust Pairwise Likelihood Method for Incomplete Longitudinal Binary Data Arising in Clusters. Canadian Journal of Statistics 39(1), 34–51. https://doi.org/10.1002/cjs.10089

**Table 1: Exogenous Sampling Simulation Results**

| | | Correlation: 0.00 | | | Correlation 0.25 | | | Correlation 0.50 | | | Correlation 0.75 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_0$ | $\beta_1$ | $\beta_2$ |
| **True Population Values** | | **-0.75** | **0.50** | **-0.49** | **-0.75** | **0.50** | **-0.49** | **-0.75** | **0.51** | **-0.50** | **-0.75** | **0.51** | **-0.51** |
| **Maintaining Population Proportions of X1 (5%)** | **Mean Coefficient Value** | **-0.75** | **0.48** | **-0.50** | **-0.75** | **0.48** | **-0.50** | **-0.75** | **0.48** | **-0.50** | **-0.75** | **0.48** | **-0.51** |
| | APE with True Value | 0.47 | 3.76 | 0.76 | 0.50 | 4.74 | 0.73 | 0.55 | 5.27 | 0.73 | 0.54 | 5.92 | 0.69 |
| | Standard Deviation of Coefficient Value | 0.09 | 0.29 | 0.14 | 0.09 | 0.31 | 0.14 | 0.09 | 0.32 | 0.14 | 0.09 | 0.34 | 0.15 |
| | Mean Standard Error | 0.09 | 0.28 | 0.14 | 0.09 | 0.65 | 0.14 | 0.09 | 1.17 | 0.14 | 0.09 | 2.91 | 0.14 |
| **Oversampling of X1 (7.5%)** | **Mean Coefficient Value** | **-0.75** | **0.49** | **-0.50** | **-0.75** | **0.49** | **-0.50** | **-0.75** | **0.49** | **-0.50** | **-0.75** | **0.49** | **-0.51** |
| | APE with True Value | 0.44 | 2.32 | 0.97 | 0.40 | 2.74 | 1.15 | 0.32 | 2.69 | 1.46 | 0.38 | 3.05 | 1.16 |
| | Standard Deviation of Coefficient Value | 0.09 | 0.24 | 0.14 | 0.09 | 0.25 | 0.14 | 0.09 | 0.26 | 0.14 | 0.09 | 0.27 | 0.14 |
| | Mean Standard Error | 0.09 | 0.23 | 0.14 | 0.09 | 0.24 | 0.14 | 0.09 | 0.25 | 0.14 | 0.09 | 0.56 | 0.15 |
| **Oversampling of X1 (10%)** | **Mean Coefficient Value** | **-0.75** | **0.50** | **-0.50** | **-0.75** | **0.50** | **-0.50** | **-0.75** | **0.51** | **-0.51** | **-0.75** | **0.51** | **-0.51** |
| | APE with True Value | 0.26 | 0.60 | 1.23 | 0.39 | 0.61 | 1.04 | 0.23 | 0.01 | 1.82 | 0.44 | 0.28 | 1.18 |
| | Standard Deviation of Coefficient Value | 0.09 | 0.20 | 0.14 | 0.09 | 0.21 | 0.14 | 0.09 | 0.22 | 0.15 | 0.09 | 0.23 | 0.15 |
| | Mean Standard Error | 0.09 | 0.20 | 0.14 | 0.09 | 0.21 | 0.14 | 0.09 | 0.22 | 0.14 | 0.09 | 0.23 | 0.15 |
| **Oversampling of X1 (12.5%)** | **Mean Coefficient Value** | **-0.75** | **0.50** | **-0.50** | **-0.75** | **0.50** | **-0.50** | **-0.75** | **0.51** | **-0.50** | **-0.75** | **0.51** | **-0.51** |
| | APE with True Value | 0.40 | 0.56 | 1.27 | 0.51 | 0.56 | 1.02 | 0.38 | 0.02 | 1.74 | 0.60 | 0.43 | 0.83 |
| | Standard Deviation of Coefficient Value | 0.09 | 0.19 | 0.14 | 0.09 | 0.19 | 0.14 | 0.09 | 0.20 | 0.14 | 0.09 | 0.21 | 0.15 |
| | Mean Standard Error | 0.09 | 0.19 | 0.14 | 0.09 | 0.19 | 0.14 | 0.09 | 0.20 | 0.14 | 0.09 | 0.21 | 0.15 |
| **Oversampling of X1 (15%)** | **Mean Coefficient Value** | **-0.75** | **0.50** | **-0.50** | **-0.75** | **0.50** | **-0.50** | **-0.75** | **0.51** | **-0.51** | **-0.75** | **0.51** | **-0.51** |
| | APE with True Value | 0.25 | 0.25 | 1.52 | 0.35 | 0.32 | 1.16 | 0.01 | 0.39 | 2.59 | 0.29 | 0.15 | 1.34 |
| | Standard Deviation of Coefficient Value | 0.09 | 0.17 | 0.14 | 0.09 | 0.18 | 0.14 | 0.09 | 0.19 | 0.14 | 0.09 | 0.20 | 0.15 |
| | Mean Standard Error | 0.09 | 0.17 | 0.14 | 0.09 | 0.18 | 0.14 | 0.09 | 0.19 | 0.14 | 0.09 | 0.20 | 0.15 |

**Table 2: Share Prediction from Exogenous Sampling Simulation**

| | | Correlation: 0.00 | Correlation 0.25 | Correlation 0.50 | Correlation 0.75 |
|---|---|---|---|---|---|
| **True Population Share** | | **0.174** | **0.174** | **0.173** | **0.173** |
| **Maintaining Population Proportions of X1 (5%)** | **Predicted Population Share** | **0.174** | **0.173** | **0.173** | **0.172** |
| | APE | 0.097 | 0.109 | 0.125 | 0.100 |
| | In-Sample Share | 0.174 | 0.173 | 0.172 | 0.172 |
| | In-Sample Share APE | 0.110 | 0.124 | 0.145 | 0.101 |
| **Oversampling of X1 (7.5%)** | **Predicted Population Share** | **0.174** | **0.173** | **0.173** | **0.172** |
| | APE | 0.127 | 0.109 | 0.072 | 0.052 |
| | In-Sample Share | 0.177 | 0.176 | 0.175 | 0.174 |
| | In-Sample Share APE | 1.920 | 1.411 | 1.055 | 0.807 |
| **Oversampling of X1 (10%)** | **Predicted Population Share** | **0.174** | **0.174** | **0.173** | **0.172** |
| | APE | 0.051 | 0.008 | 0.030 | 0.023 |
| | In-Sample Share | 0.181 | 0.179 | 0.177 | 0.176 |
| | In-Sample Share APE | 4.342 | 3.238 | 2.426 | 1.816 |
| **Oversampling of X1 (12.5%)** | **Predicted Population Share** | **0.174** | **0.173** | **0.173** | **0.172** |
| | APE | 0.120 | 0.142 | 0.127 | 0.115 |
| | In-Sample Share | 0.185 | 0.182 | 0.179 | 0.177 |
| | In-Sample Share APE | 6.236 | 4.646 | 3.402 | 2.554 |
| **Oversampling of X1 (15%)** | **Predicted Population Share** | **0.174** | **0.174** | **0.173** | **0.173** |
| | APE | 0.026 | 0.014 | 0.029 | 0.071 |
| | In-Sample Share | 0.189 | 0.185 | 0.181 | 0.179 |
| | In-Sample Share APE | 8.582 | 6.456 | 4.760 | 3.647 |

**Table 3: Endogenous Sampling Simulation Results**

| | | Outcome of Interest | | | Selection Model | | | | Correlation |
|---|---|---|---|---|---|---|---|---|---|
| | | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\gamma_0$ | $\gamma_1$ | $\gamma_2$ | $\gamma_3$ | $\rho$ |
| **True Population Values** | | **-0.77** | **0.51** | **-0.49** | **-0.52** | **1.00** | **-0.48** | **0.50** | **0.50** |
| **Unweighted Binary Probit** | **Mean Coefficient Value** | **-0.30** | **0.28** | **-0.40** | -- | -- | -- | -- | -- |
| | APE with True Value | 60.57 | 44.83 | 18.86 | -- | -- | -- | -- | -- |
| | Standard Deviation of Coefficient Value | 0.05 | 0.06 | 0.06 | -- | -- | -- | -- | -- |
| | Mean Standard Error | 0.05 | 0.06 | 0.06 | -- | -- | -- | -- | -- |
| **Weighted Binary Probit** | **Mean Coefficient Value** | **-0.30** | **0.28** | **-0.40** | -- | -- | -- | -- | -- |
| | APE with True Value | 60.58 | 44.83 | 18.84 | -- | -- | -- | -- | -- |
| | Standard Deviation of Coefficient Value | 0.05 | 0.06 | 0.06 | -- | -- | -- | -- | -- |
| | Mean Standard Error | 0.05 | 0.06 | 0.06 | -- | -- | -- | -- | -- |
| **Joint Binary Probit (Known Sampling Frame)** | **Mean Coefficient Value** | **-0.79** | **0.53** | **-0.50** | **-0.51** | **1.00** | **-0.49** | **0.50** | **0.52** |
| | APE with True Value | 2.70 | 3.86 | 1.74 | 0.43 | 0.15 | 0.59 | 0.08 | 3.31 |
| | Standard Deviation of Coefficient Value | 0.08 | 0.06 | 0.06 | 0.03 | 0.04 | 0.04 | 0.02 | 0.08 |
| | Mean Standard Error | 0.09 | 0.07 | 0.06 | 0.03 | 0.04 | 0.04 | 0.02 | 0.08 |
| **Joint Binary Probit (Unknown Sampling Frame)** | **Mean Coefficient Value** | **-0.75** | **0.51** | **-0.49** | **-0.51** | **1.03** | **-0.51** | **0.57** | **0.47** |
| | APE with True Value | 3.08 | 0.04 | 1.26 | 0.87 | 2.48 | 5.70 | 13.30 | 5.39 |
| | Standard Deviation of Coefficient Value | 0.08 | 0.07 | 0.06 | 0.04 | 0.06 | 0.06 | 0.04 | 0.08 |
| | Mean Standard Error | 0.09 | 0.07 | 0.06 | 0.03 | 0.04 | 0.04 | 0.02 | 0.08 |

**Table 4: Share Prediction from Endogenous Sampling Simulation**

| | | Share | APE |
|---|---|---|---|
| **True Population Share** | | **0.221** | -- |
| **Average In-Sample Share** | | 0.387 | 74.769 |
| **Predicted Share Using** | **Unweighted Binary Probit** | 0.352 | 58.950 |
| | **Weighted Binary Probit** | 0.383 | 72.827 |
| | **Joint Binary Probit (Known Sampling Frame)** | 0.218 | 1.594 |
| | **Joint Binary Probit (Unknown Sampling Frame)** | 0.228 | 2.870 |

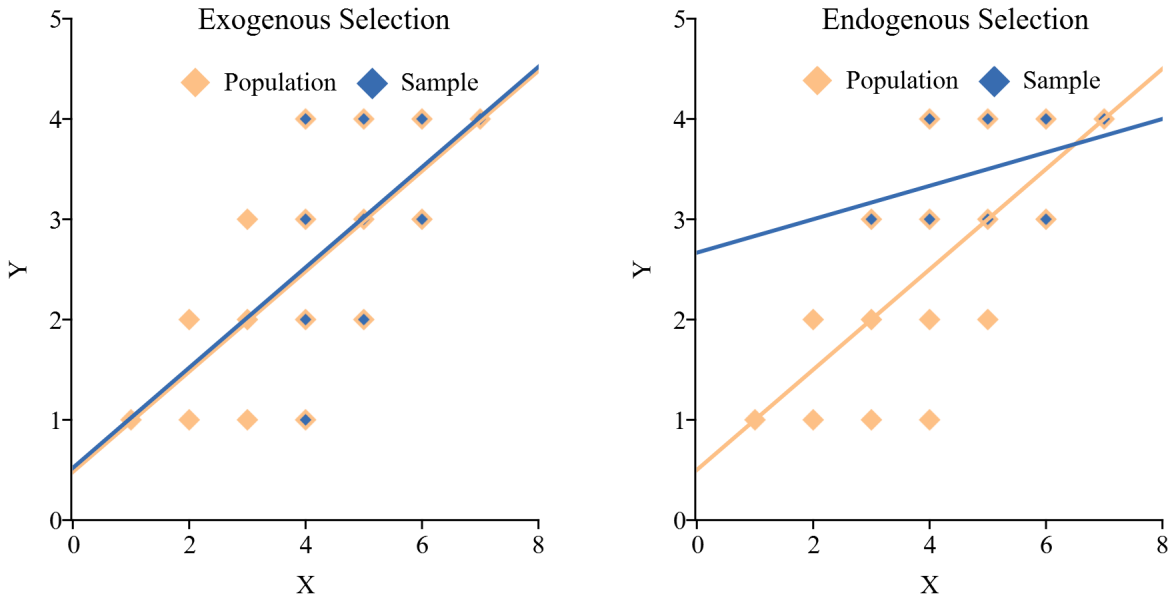


**Figure 1: Exogenous and Endogenous Selection Demonstration**

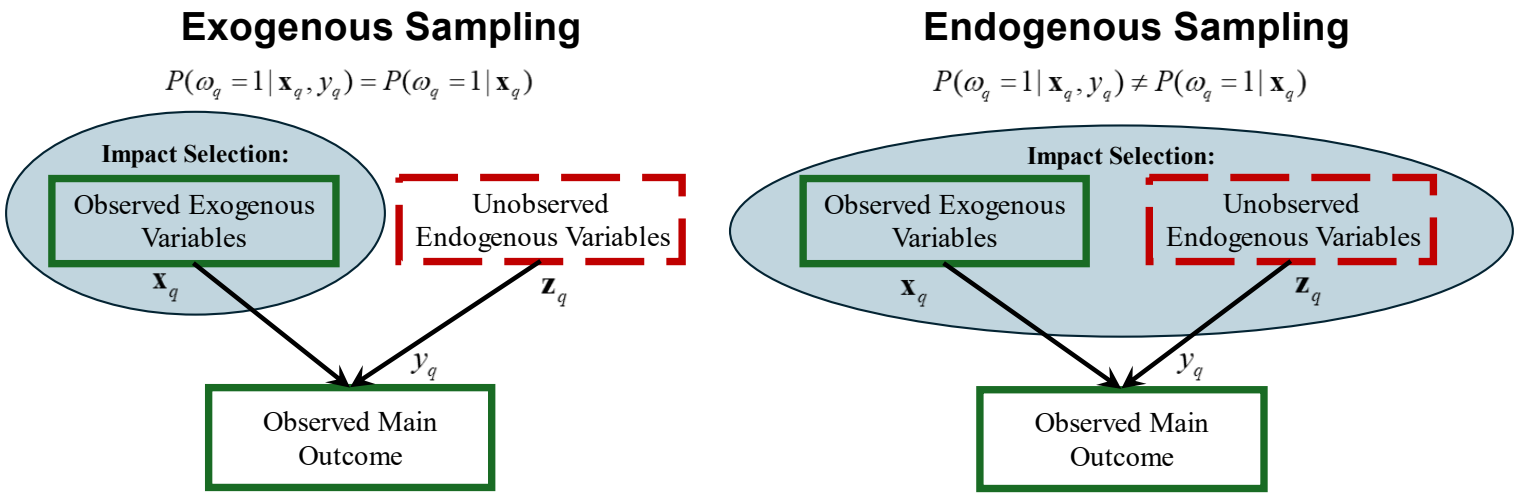


**Figure 2: Sampling Approaches**

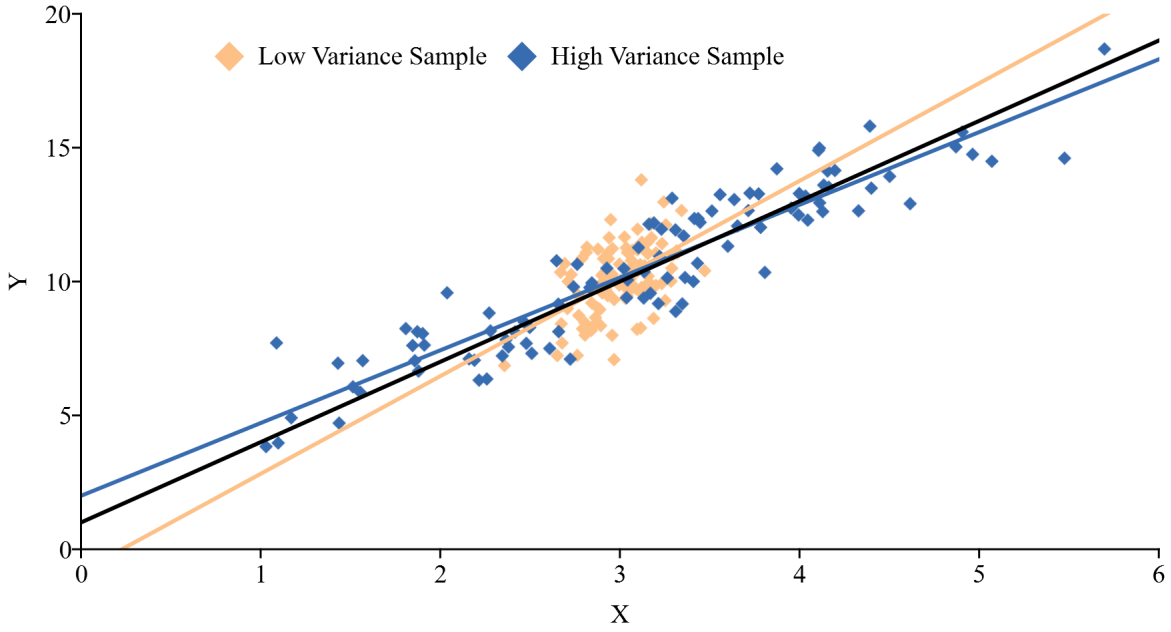


**Figure 3: Effects of Exogenous Variation on Estimator Performance**

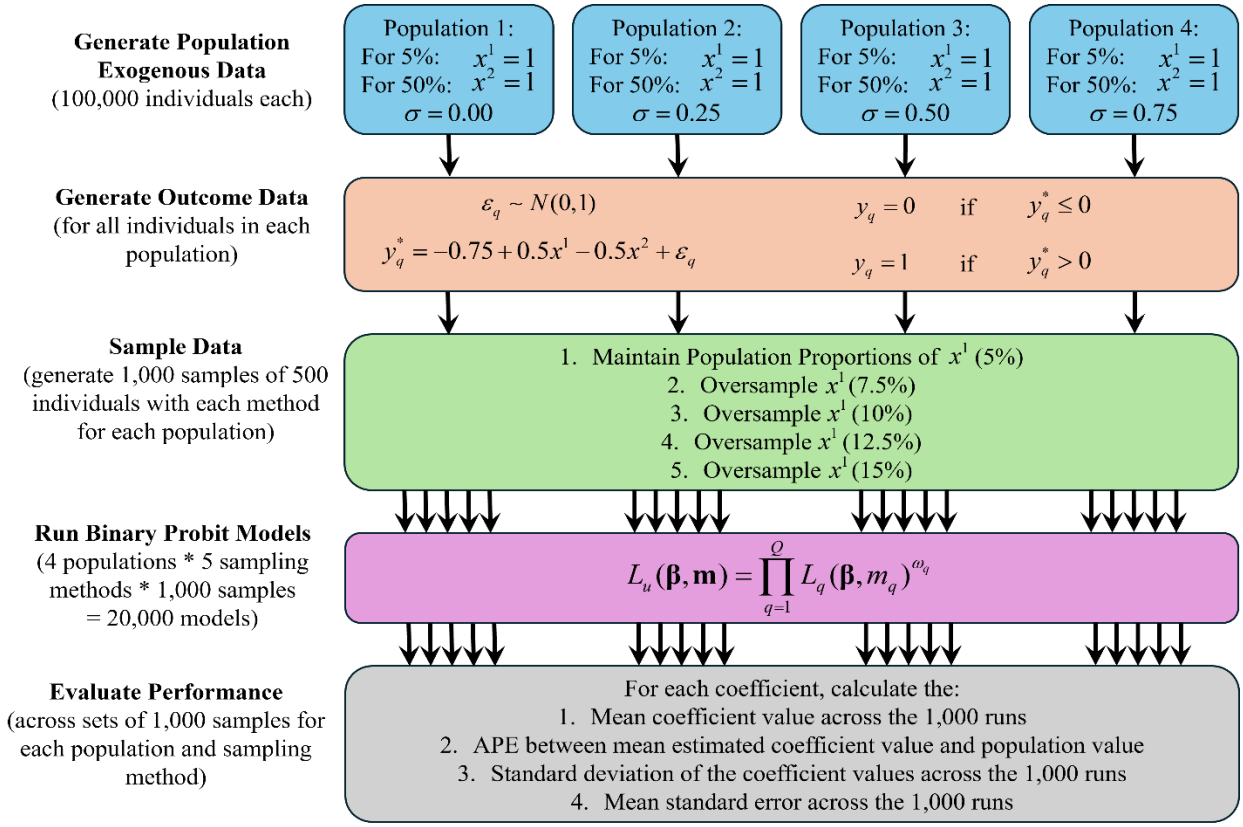


**Figure 4: Design of Exogenous Sampling Simulation**

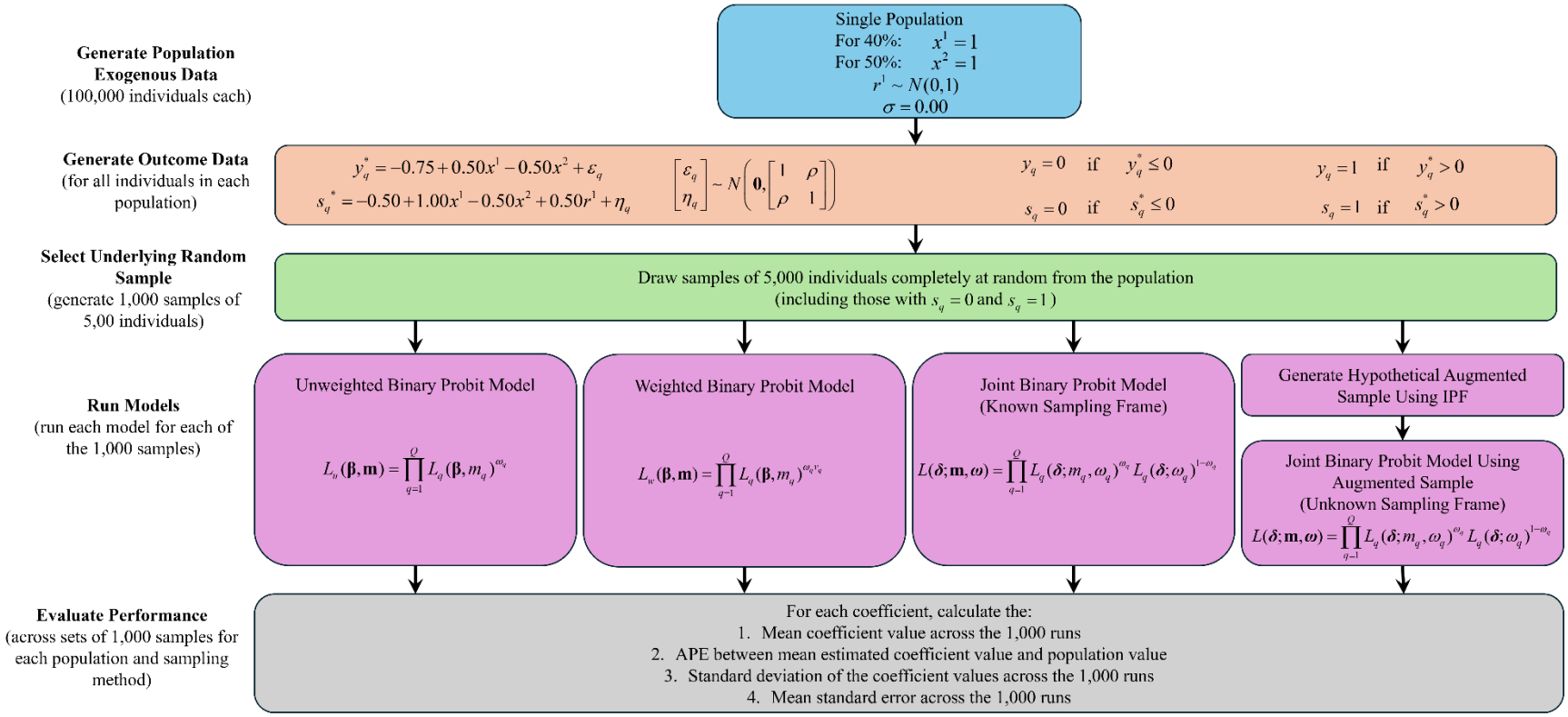


**Figure 5: Design of Endogenous Sampling Simulation**