NBER WORKING PAPER SERIES

# SIMPLIFYING BIAS CORRECTION FOR SELECTIVE SAMPLING: A UNIFIED DISTRIBUTION-FREE APPROACH TO HANDLING ENDOGENOUSLY SELECTED SAMPLES

Yi Qian
Hui Xie

Working Paper 28801
http://www.nber.org/papers/w28801

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2021

We thank the grant support from Social Sciences and Humanities Research Council of Canada [grant 435-2018-0519], Natural Sciences and Engineering Research Council of Canada [grant RGPIN-2018-04313] and US National Institute of Health [grants R01CA178061]. We thank Dr. K. Sudhir, our colleagues, and the review team of Marketing Science for invaluable comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Simplifying Bias Correction for Selective Sampling: A Unified Distribution-Free Approach to Handling Endogenously Selected Samples
Yi Qian and Hui Xie
NBER Working Paper No. 28801
May 2021
JEL No. C01,C1,C5,C8

## ABSTRACT

Unlike random sampling, selective sampling draws units based on the outcome values, such as over-sampling rare events in choice outcomes and extreme activities on continuous and count outcomes. Despite high cost effectiveness for marketing research, such endogenously selected samples must be carefully analyzed to avoid selection bias. We introduce a unified and efficient approach based on semiparametric odds ratio (SOR) models applicable for categorical, continuous and count response data collected using selective sampling. Unlike extant sampling-adjusting methods and Heckman-type selection models, the proposed approach requires neither modeling selection mechanisms nor imposing parametric distributional assumptions on the response variables, eliminating both sources of mis-specification bias. Using this approach, one can quantify and test for the relationships among variables as if samples had been collected via random sampling, simplifying bias correction of endogenously selected samples. We evaluate and illustrate the method using extensive simulation studies and two real data examples: endogenously stratified sampling for linear/nonlinear regressions to identify drivers of the share-of-wallet outcome for cigarettes smokers, and using truncated and on-site samples for count data models of store shopping demand. The evaluation shows that selective sampling followed by applying the SOR approach reduces required sample size by more than 70% compared with random sampling, and that in a wide range of selective sampling scenarios SOR offers novel solutions outperforming extant methods for selective samples with opportunities to make better managerial decisions.

Yi Qian
Sauder School of Business
University of British Columbia
2053 Main Mall
Vancouver, BC V6T 1Z2
CANADA
and NBER
yi.qian@sauder.ubc.ca

Hui Xie
Department of Biostatistics
School of Public Health
University of Illinois at Chicago
huixie@uic.edu

## 1. Introduction

Nonrandom samples resulting from selective sampling are prevalent in marketing research, given the nature of much of the primary and secondary data produced in the field. Unlike random sampling in which each unit in the population has equal probability of being sampled, selective sampling has a unit sampling probability that differs from its frequency in the population. When the sampling scheme involves the dependent variable $Y$ (*a.k.a.* response-dependent sampling), the resulting sample forms an "endogenously selected sample", the analysis of which requires special handling of selection bias associated with such samples and is the focus of this paper. For exposition simplicity, henceforth we use the term "selective sampling" to refer to a sampling scheme involving $Y$ and producing endogenously selected samples.

In many situations, marketing researchers *purposefully* conduct such selective sampling in data collection to garner its high cost-effectiveness. The idea is to obtain a sample over-represented by more informative units. In practice, it is often necessary to adopt selective sampling and account for it in the analysis to more effectively infer the population density/mass function of interest $f_\theta(Y|X)$ in marketing research, where $X$ contains a set of independent variables and $\theta$ is a vector of parameters. Examples of response-dependent sampling abound in marketing research and below are four running examples considered and dealt with in this work.

*Example 1: Selective Sampling for Binary Choice Models.* In database marketing, when a choice outcome (e.g., consumer churn, web ad click) is rare (Donkers *et al.*, 2003, Kamakura *et al.*, 2005), random sampling can be inefficient or even impossible as it can require a very large sample or a very long time to achieve enough events of interest for meaningful analysis. Even if large consumer databases may already have enough events, relevant explanatory variables may be lacking (Qian and Xie, 2014, 2015). Separate efforts (e.g., surveys or measurements) are often required to collect data on these variables, which are feasible only in a sample of consumers. Donkers *et al.* (2003) showed that, for binary logit choice models studying the determinants of the rare churn outcome, oversampling defecting customers can reduce required survey sample size by more than 50%.

*Example 2: Selective Sampling for Continuous Outcome Models.* Selective sampling can also be applied to continuous outcomes (Feldt, 1961, Hausman and Wise, 1981, Cosslett, 1993, 2013).

One can enrich the sample by oversampling extreme or infrequently occurring values of a continuous behavioral outcome. For implementation convenience, sampling is often based on intervals of $Y$. We illustrate such an endogenous stratified sampling of linear/nonlinear mean regressions to identify factors affecting a smoker's consumption share of the main cigarette brand. Over-sampling extreme values of the consumption share outcome combined with a stratification of the highly skewed regressor (price) reduces the required sample size by $\approx 70\%$ as compared with random sampling.

*Example 3: Selective (truncated and on-site) Sampling for Count Models.* Besides used for increasing sample information, selective sampling could be a *byproduct* of database construction. This case study uses a firm's shopper database to construct a count model for the number of shopping visits ($Y$) made by store shoppers for consumer profiling and targeting purpose. To be included in the database, a consumer must made at least one shopping visit. Thus, the sample is truncated with no zero count. This is an extreme case of selective sampling, whereby the subpopulation of store non-users with $Y = 0$ is not sampled at all. An alternative way to estimate such a count model is to analyze mall intercept surveys that interview shoppers about the number of shopping visits (Clow and James, 2014). With such on-site sampling, besides the truncation of those non-users, among those who visit the store, consumers who visit the store frequently are more likely to be sampled than consumers who visit the site only occasionally. By oversampling these more active consumers, such on-site samples can contain a lot more information than random samples.

*Example 4: Selection on $(Y, X, \theta)$.* Endogenously selected samples can also occur out of the control of researchers when data owners selecting observations (e.g., to preserve privacy) are different from researchers, or as a result of the behaviors of the units being sampled (i.e., self-selection). An example for the latter one is nonresponse to surveys. Selection mechanisms can depend on $(Y, X)$ or $(Y, X, \theta)$ and frequently be unobserved to researchers.

### Methods to Adjust for Selective Sampling

Despite the substantial benefits of selective sampling noted above, care must be taken when analyzing resulting nonrandom samples to avoid selection bias (e.g., Breslow and Day (1980), Hausman and Wise (1981), Donkers *et al.* (2003), Feinberg *et al.* (2012)). To understand the bias, let $G$ be the indicator variable for a population unit to be in the selected sample and $p_\psi(G|Y, X, \theta)$ denote the selection mechanism where $\psi$ collects parameters distinct from $\theta$. When $p_\psi(G|Y, X, \theta)$ depends on $Y$, the selection mechanism corresponds to nonignorable missingness in Rubin's missing data framework (Little and Rubin, 2020)[1] in that $p_\psi(G|Y, X, \theta)$ generally cannot be ignored when estimating $\theta$ with the selected sample.[2] Naive methods that analyze endogenously selective samples as random samples without accounting for selective sampling can yield significant selection bias.

Existing sampling-adjustment methods require specification/restriction of the selection mechanism $P_\psi(G|Y, X, \theta)$ and/or imposing outcome distributional assumptions on $f_\theta(Y|X)$ (see Tables 1 and 2 (excl. the last columns) for typical modeling assumptions and restrictions). When sampling is under the control of researchers with known sampling weights or known forms of sampling schemes (Table 1 and Examples 1-3 above), the main concern is that mis-specifying the outcome error distributions can yield biased results and erroneous conclusions. To address the concern, Cosslett (2013) developed an adjustment method with unspecified error distributions for endogenously stratified regressions. However, the method is applicable only to the case of two strata defined by whether $Y$ exceeds a known cutoff value. When units not in the selected sample are excluded because some of their $X$s are unobserved (e.g., Example 1 above), missing-covariate methods (Qian and Xie, 2011) can be applied to model and impute these missing $X$ values. However, in order to include these data into analysis to increase efficiency, an additional covariate model is required (Table 2) which may not be desirable or even feasible in many selective sampling applications.

Heckman-type selection models (Heckman (1979), *a.k.a.* Heckit) are the classical approach to adjusting for self-selection bias when selective sampling mechanisms are not under the control of and unobserved to researchers (Example 4 above and see also Ying *et al.* (2006), Wachtel and Otter

[1] See Little and Rubin (2020) Chpt.15 for models that handle non-ignorable missingness with arbitrary patterns of missingness. Little and Rubin (2020) note that these models require modeling missing-data mechanism if the mechanism is not completely known, and model non-identifiability and sensitivity to model specifications are serious challenges in these models. As shown later, by exploiting special structures of selective sampling problems, one can robustly identify key model parameters without modeling missing-data mechanisms and outcome error distributions.

[2] In contrast, sampling based only on $X$ yields exogenously selected samples, for which valid analysis can ignore selection rules.

(2013), Feinberg *et al.* (2016), Tian and Feinberg (2020)). The Heckit procedures have restricted forms for outcome distributions and selection rules (Table 2). When these forms are specified correctly the procedures can work well to correct for self-selection bias. The key concern with Heckit is its brittleness in that Heckit depends exquisitely on untestable modeling assumptions (e.g., normal error distribution, exclusion restriction) and is sensitive to mis-specification bias (e.g., Puhani (2000)). For the reason, alternative approaches propose using these selection models as a device for sensitivity analysis (Little and Rubin, 2020, Xie and Qian, 2012, Yuan *et al.*, 2020). A particular relevant issue is the rigidity of the linear probit selection assumed in Heckit to handle the wide variety of selection rules in selective sampling problems (see Section 3.2).

Overall, despite considerable progress made in addressing selection bias, the existing methods are specific to particular sampling schemes, or imposing strong assumptions on outcome error distributions and/or selection mechanisms. As compared with handling random samples, they often require extra modeling efforts and different (sometimes very complex) estimation algorithms to correct for selection bias, hindering the use of endogenously selected samples. There is a considerable need for simple and broadly applicable sampling-adjustment methods with reduced modeling assumptions/efforts and optimal statistical properties.

To overcome these challenges in analyzing endogenous selected samples, we introduce a unified bias-correction approach that requires neither specifying the outcome error distributions nor the selection mechanisms, thereby eliminating these sources of mis-specification bias altogether for a wide variety of selective sampling schemes. The proposed approach corrects selection bias as if the samples had been collected via random sampling and requires a minimum amount of additional modeling and estimation efforts (Tables 1 and 2). It provides efficient maximum likelihood estimation of selective samples based on a semiparametric odds ratio (SOR) modeling framework (Chen, 2007, Chen *et al.*, 2015). In general, all estimates of regression model parameters, such as those in the commonly-used generalized linear models (GLM) (McCullagh and Nelder, 1989), are biased without correcting for the response-dependent sampling. An exception is the logit model for binary outcomes (Donkers *et al.*, 2003), in which the regression coefficients are unaffected by response-dependent sampling. The SOR sampling-adjustment method provides a unified framework that covers as a special case this practically important result of Donkers *et al.* (2003) and extends it to polytomous, continuous and count outcomes with a wide variety of selective sampling schemes. An important merit of SOR is that the population association parameters as captured by regression coefficients in SOR for polytomous, continuous and count outcomes continue to be unaffected by response-dependent sampling, and consistent estimation of these population association parameters using these nonrandom samples requires neither modeling selection mechanisms nor specifying the outcome error distributions. While nesting GLMs as special cases, the regression coefficients in SOR are closely linked to familiar association measures in GLMs: they are the regression coefficients (e.g., in logit and Poisson regressions) themselves or standardized by the dispersion parameter when present (e.g., in a normal linear regression) in a population GLM model.

We further extend the SOR approach in a number of important ways relevant to marketing and other fields alike. *First*, we derive new SOR models with regression coefficients mapped to parameters in the (nested) parametric models with unobserved consumer heterogeneity and outside GLMs. When the population follows no parametric distribution, the regression coefficients in SOR can be interpreted as OR association parameters, or alternatively we show how to map SOR models to nonlinear regression models that relax not only the outcome distributional assumptions, but also the linearity assumption in the mean structure of linear regression models for continuous outcomes. In this general case, we derive simple relationships mapping the OR parameters to other parameters of interest, such as marginal mean effects, to aid the interpretation of the SOR results.

*Second*, we extend the SOR to handle endogenous samples selected on $(Y, X)$ with and without the knowledge of sampling weights as well as selection on $(Y, X, \theta)$. These extensions render SOR applicable for a wider range of selective sampling, including sampling and self-selection depending on $(Y, X, \theta)$. As compared with Heckit, the SOR approach requires specification of neither the outcome distribution nor the functional form of response-dependent selection. These semiparametric features make the SOR approach attractive to bias correction in a broad range of selective sampling problems (Sections 2.4 and 3.2). Our extension also shows that, in stark contrast to selection on the outcome considered in the Chen (2007) and Chen *et al.* (2015), when sampling weights depend on both $Y$

and $X$, knowledge of sampling weights[3] can lead to a substantial gain in the estimation precision and power to select important explanatory variables and a significant reduction in the sample size required, compared to cases in which the weights are unknown. A further benefit of knowing sampling weights is that, by incorporating them into estimation as offsets to correct for selective sampling, one can recover the entire distribution to examine various aspects of the distribution, such as regression means, percentiles and marginal mean effects. To facilitate inference about these quantities for marketing research, we develop a novel bootstrap procedure that, unlike regular bootstrap procedures for random samples, can properly account for selective sampling schemes when computing the exact standard errors of complex functions of model parameter estimates.

*Finally*, we investigate the use of the unified framework in a broad range of marketing applications, including endogenous stratified sampling, truncated samples, on-site sampling and sample selection problems, which yields a novel set of principled, robust, and amenable solutions to the selective sampling problems. To our knowledge, this is the first time in the literature to treat such a broad set of selective sampling problems under one unified modeling framework.

Next, Section 2 describes the SOR models that extend the logit model for binary choice outcomes to more general outcome types, and develops estimation methods for selective samples. Section 3 evaluates the performance of these methods using synthetic data in the context of Examples 2 and 4, and studies factors affecting the efficiency of selective sampling. Section 4 applies SOR to Examples 3. These analyses show a large reduction (up to 90%) in required sample size with selective sampling for continuous and count outcomes compared with random sampling as well as the flexibility and robustness of the SOR. The analysis of the two real-life data sets also reveals that, because of the mis-specifications of the underlying distributions for the continuous and count outcomes, the existing sampling-adjustment methods yield biased estimates and erroneous identification of outcome determinants that misinform managerial decisions. In contrast, SOR can adapt itself to the underlying distributions and protect the analysis of selective samples from such

[3] In practice, access to large consumer information databases, census data and study screening data permits obtaining sampling weights (Donkers *et al.*, 2003). Thus, the extension is especially relevant to marketing researchers given the pervasiveness of stratifying on explanatory variables.

bias. We end with a discussion in Section 5.

## 2. Methodology

### 2.1 *Odds ratio models*

The aim is to learn about the population distribution, $f_\theta(Y|X)$, where $X = (X_1, \cdots, X_K)$ contains $K$ explanatory variables, $f_\theta(Y|X)$ denotes either a probability density function (PDF) when $Y$ is continuous, or a probability mass function (PMF) when $Y$ is categorical or count. For a categorical outcome $Y$, the odds ratio (OR) is frequently used to measure the association between Y and X (Breslow and Day, 1980). The well-known Chi-squared tests for contingency table analysis and multivariable logistic regressions are frequently used to estimate OR and test for the presence, direction, and strength of associations via ORs. Below we describe the SOR model (Chen, 2007) for use with endogenously selected samples. The model nests as a special case the logit model for categorical outcomes and extends to continuous and count outcomes. The SOR model has been adapted for handling missing data and data combinations in marketing and business analytics (Qian and Xie, 2011, 2014, 2015) and been discussed by Feit and Bradlow (2018). Let $(y_0, x_0)$ be a fixed point in the sample space of $(Y, X)$. We define the OR for the two variables as

$$\eta(y, y_0; x, x_0) \quad = \quad \frac{f(y|x)f(y_0|x_0)}{f(y|x_0)f(y_0|x)}. \tag{1}$$

The OR parameter $\eta(y, y_0; x, x_0)$ is the ratio of the odds of observing $y$ relative to observing $y_0$ when $X$ varies from $x_0$ to $x$. For notational simplicity, we write $\eta(y, y_0; x, x_0)$ as $\eta(y; x)$ henceforth. The OR can be used to measure the strength of the associations among general types of variables. In particular, $\eta(y; x) \equiv 1$ for all values of $Y$ and $X$ means no association between $Y$ and $X$. To better understand the OR for continuous variables, consider a simple linear regression model where $X$ consists of a single variable and

$$y \quad = \quad \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

Plugging the conditional normal density functions $f(y|x), f(y_0|x_0), f(y|x_0), f(y_0|x)$ into Eqn (1), it can be readily verified that the OR takes the following log-bilinear form:

$$\ln \eta(y; x) = \ln \eta(y, y_0; x, x_0) = \frac{\beta_1}{\sigma^2}(x - x_0)(y - y_0).$$

Thus for this normal linear model, the OR $\eta(y;x)$ is closely related to the familiar regression coefficient parameters $\beta_1$. In particular, $\eta(y;x) \equiv 1$ is equivalent to $\beta_1 = 0$.

For modeling and estimation in general settings, consider the following SOR modeling approach, which will be used to handle endogenously selected samples. Multiplying both sides of Eqn (1) by the denominator of the right-hand side and then integrating both sides with respect to $y$, we obtain

$$f(y|x) \quad = \quad \frac{\eta(y;x)f(y|x_0)}{\int \eta(y;x)f(y|x_0)dy}. \tag{2}$$

The right-hand side of the above equation re-expresses $f(y|x)$ as a function of two components. The first component is the OR, $\eta(y;x)$, which describes the association between $Y$ and $X$, and the second component $f(y|x_0)$ is a marginal-like density function for $Y$ at a fixed point $X = x_0$, which we call a baseline function. As illustrated above for the simple linear regression model, the OR parameter $\eta(y;x)$ is closely related to other familiar ways of describing associations, such as the regression coefficients in a linear model. Motivated by the above simple linear regression model, we consider the following log-bilinear model for the OR function for $x = (x_1, \cdots, x_K)$:

$$\ln \eta(y;x) \quad = \quad \sum_{k=1}^{K} \gamma_k (x_k - x_{k0})(y - y_0), \tag{3}$$

where $\gamma = (\gamma_1, \cdots, \gamma_K)$ is a set of parameters in the log-odds-ratio function, and $\gamma_k$ captures the association between $Y$ and $X_k$ when holding other variables in $X$ constant. The selection of the reference point $(x_0, y_0)$ can be arbitrary, but a value of $(x_0, y_0)$ remote from the observed data points may lead to computational instability in model estimation. The log-bilinear form of OR function is closely related to the GLMs (McCullagh and Nelder, 1989), commonly used to analyze marketing responses. To see this, consider the GLM with the following density function:

$$f_{\beta,\tau}(y|x) = \exp\left\{ \frac{y\Psi(\beta, x) - b(\Psi(\beta, x))}{a(\tau)} + c(y, \tau) \right\},$$

where $\Psi$ is the canonical parameter; functions $b(\cdot)$ and $c(\cdot,\cdot)$ determine a distribution in the exponential family; and $a(\tau) = \tau/w$, with dispersion parameter $\tau$ and known weight $w$. The GLMs include Gaussian, logistic, Poisson, and Gamma regressions as special cases. For GLMs with canonical link functions and $g(E(Y|x)) = \beta_0 + \sum_{k=1}^{K} \beta_k x_k$, the OR function is

$$\ln \eta(y;x) = \sum_{k=1}^{K} \frac{\beta_k}{a(\tau)} (x_k - x_{k0})(y - y_0).$$

Therefore, the parameters in this log-bilinear OR function are a re-parametrization of parameters in GLMs with $\gamma_k = \frac{\beta_k}{a(\tau)}$. In the log-bilinear form, $\gamma_k = 0$ is equivalent to $\beta_k = 0$. The simple linear regression model described earlier is a special case of GLM with $a(\tau) = \sigma^2$. Higher-order terms of the explanatory variables and the interaction terms between them can be added in the above log-bilinear form of OR function (Eqn (3)) and tested using likelihood-based test statistics. Alternative forms of OR functions can be motivated by parametric models outside of classical GLMs. See Section 4 for an example motivated by the negative binomial distribution. With sufficiently flexible functional forms the OR function can approximate any arbitrary and complex forms of associations among variables. Using model selection measures such as the likelihood-ratio, AIC or BIC statistics, the flexible OR function specification permits selection of models within and outside of the family of GLMs under one unified modeling framework.

The second component $f(y|x_0)$ is the baseline density or mass function for $Y$ at the fixed reference point $X = x_0$ and behaves like a marginal distribution. For modeling robustness, one can employ a nonparametric empirical modeling approach. Specifically, let $(u_1, ..., u_L)$ be the uniquely observed values in the dataset for this variable. A nonparametric model assigns probability mass $p = (p_1, ..., p_L)$ to $f(u_1|x_0), \cdots, f(u_L|x_0)$ with a constraint $\sum_{l=1}^{L} p_l = 1$. To relax the constraint, we re-parameterize $p$ as $\lambda = (\lambda_1, ..., \lambda_L)$, such that $\lambda_l = \ln(p_l/p_L)$ and $p_l = \exp(\lambda_l)/\sum_{l'=1}^{L} \exp(\lambda_{l'})$, for $l = 1, ..., L$. The resulting SOR model nests the parametric GLMs as special cases by eschewing the parametric assumptions in the baseline distribution in GLMs. Given a sample of $n$ independent units, $(x_i, y_i), i = 1, \cdots, n$, the likelihood for the SOR model is

$$L(\gamma, \lambda) \propto \prod_{i=1}^{n} f_{\gamma,\lambda}(y_i|x_{i1}, \cdots, x_{iK}) \;=\; \prod_{i=1}^{n} \frac{\eta_\gamma\{y_i; x_{i1}, \cdots, x_{iK}\} f_\lambda(y_i|x_{10}, \cdots, x_{K0})}{\sum_{l=1}^{L} \eta_\gamma\{u_l; x_{i1}, \cdots, x_{iK}\} f_\lambda(u_l|x_{10}, \cdots, x_{K0})}. \quad (4)$$

The maximum likelihood estimates of the model parameters, $(\hat{\gamma}, \hat{\lambda})$, can be obtained using an algorithm for function optimization, such as the quasi-Newton algorithm, with details and an illustrating example given in Online Appendix I.1. In Online Appendix I.2, we describe a Bayesian estimation of the SOR model, which can be useful in marketing applications. After model estimation, the predictive distribution for the outcome conditional on $X$ can be obtained as a multinomial distribution on the set of observed unique values $(u_1, ..., u_L)$ with

$$Pr(Y = u_l|x) \;\propto\; \exp(\hat{\lambda}_l + \ln(\eta_{\hat{\gamma}}\{u_l; x_{i1}, \cdots, x_{iK}\})), \qquad l = 1, \cdots, L, \quad (5)$$

where the right-hand side is a product of the OR function $\eta_\gamma\{u_l; x_{i1}, \cdots, x_{iK}\}$ and the exponentiated parameter $\lambda_l$, the $l$th intercept parameter in $(\lambda_1, \cdots, \lambda_{L-1})$ with no constraint among them in this multiplicative model. Furthermore, the OR function and the intercept parameters in SOR are variational independent, and thus can be modeled separately (Chen, 2007). The SOR model thus has the unique features of the multiplicative intercept (MI) model for categorical data (Scott and Wild, 1997), such as logistic models, and extends them to other outcome types, which simplifies the analysis of endogenously selected samples as shown next.[4]

### 2.2 *Selection on outcome*

One advantage of the above SOR model is that the OR parameter remains invariant to response-dependent sampling. To demonstrate this, we first consider a selection mechanism, in which the sampling probability depends on the outcome only with $p_\psi(G = 1|y)$ denoting the conditional probability of being sampled, given that $Y = y$. Let $f(Y = y|x, G = 1)$ denote the PDF/PMF of $Y$ given $x$ in the selected sample, and the purpose is to learn about the conditional distribution of $Y$ given $x$ in the population, i.e., $f_\theta(Y = y|x)$. Using Bayes' theorem,

$$f(Y = y|x, G = 1) \quad = \quad \frac{f(Y = y, x, G = 1)}{f(x, G = 1)} = \frac{p_\psi(G = 1|y) f_\theta(Y = y|x)}{\int p_\psi(G = 1|u) f_\theta(u|x) du}. \tag{6}$$

We note two important points regarding the above distribution of $Y|X$ in the selected samples. First, the marginal distribution of $X$ does not appear in $f(Y = y|x, G = 1)$ because it cancels out in both the numerator and denominator of the rightmost side of the equation. This means that we do not need to model the covariate distribution, and inference is valid for arbitrary distributions of $X$. This can be an important modeling and computational advantage. Second, using the OR model expression in Eqn (2) for $f_\theta(y|x)$, we have

$$f(Y = y|x, G = 1) \quad = \quad \frac{\eta_\gamma(y; x) f'(y|x_0)}{\int \eta_\gamma(u; x) f'(u|x_0) du}, \text{ where } f'(y|x_0) = \Big[p_\psi(G = 1|y) f_\lambda(y|x_0)\Big]. \tag{7}$$

As seen above, $f(Y = y|x, G = 1)$, which can be directly learned from the selective sample, has an OR expression in the form of Eqn (2). In this expression, the OR function is $\eta_\gamma(y; x)$, which

[4] The MI model does not require that $\eta_\gamma\{u_l; x_{i1}, \cdots, x_{iK}\}$ in (5) take certain log-bilinear forms. Furthermore, not all GLMs are multiplicative intercept models. For example, normal regression models are not MI models.

is exactly the same as the OR function in the population conditional distribution of $Y$ given $x$, demonstrating that the OR is invariant to the sampling scheme and not subject to selection bias.[5] This implies that the population OR parameters can be consistently estimated using the selective sample as if it was a random sample. One can identify and evaluate the association of explanatory variables with the outcome via OR parameters without the need to model selection mechanism.

On the other hand, the baseline distribution $f'(y|x_0)$ for the selective sample is $p_\psi(G = 1|y)f_\gamma(y|x_0)$, i.e., the population baseline distribution $f_\gamma(y|x_0)$ modified multiplicatively by the sampling weight function $p_\psi(G = 1|y)$. Thus, without correcting for endogenous selectivity, the estimates of the baseline distribution using the selective sample will be biased for the population baseline distribution. Below, we describe two estimation procedures based on maximization of the conditional likelihood in Eqn (7), depending on whether one has supplemental data on sampling weights. For the MI models such as ours, they yield fully-efficient MLEs of the population OR parameters, regardless of whether or not sampling weights are known as well as of the baseline distribution function when known sampling weights are included as offsets to correct for selective sampling (Scott and Wild, 1997, Chen *et al.*, 2015).

### *2.2.1 Estimation without data on sampling weights (SOR)*

Estimation proceeds as if the sample was a random sample, without the need to correct for selective sampling. The MLEs of the OR parameters $\gamma$ can be obtained jointly with $\lambda'$, the parameters in the baseline distribution $f'_{\lambda'}(y|x_0) = p_\psi(G = 1|y)f_\lambda(y|x_0)$ using the algorithm described in Online Appendix I. The baseline distribution $f'_{\lambda'}(y|x_0)$ for the selective sample depends on the unknown sampling weights and may behave differently from any standard theoretical distributions, even if the population baseline distribution $f_\lambda(y|x_0)$ can be assumed to follow a simple parametric model. As shown later in simulation studies, an incorrect specification of $f'_{\lambda'}(y|x_0)$ via a parametric model can lead to significant bias in the estimation of OR parameter $\gamma$. By contrast, the SOR model employs a nonparametric model for $f'_{\lambda'}(y|x_0)$. Thus, the SOR estimation requires no prior

[5] Theoretically, this invariance property is a natural result of the SOR model having the multiplicative intercept model form in Eqn (5), from which it is readily seen that sampling weights modify the intercept terms multiplicatively but do not affect OR functions.

knowledge about the potential complex form of the baseline distribution function in the selective sample and can automatically generate suitable distributions that match data.

### *2.2.2 Estimation using supplemental data on sampling weights (SOR-FI)*

When supplemental data on sampling weights are available as described in the Introduction, one can make use of the supplemental information to recover $f_\lambda(y|x_0)$, the population baseline distribution function. One can fit the model to obtain the MLEs of $(\gamma, \lambda)$ using a program written for fitting the SOR model, provided that the program allows for including sampling weights as offsets in the estimation of the baseline distribution. For illustration purposes, below we provide three examples of selection on outcomes with known sampling weights.

**Case I. Stratified sampling on a categorical outcome** For a categorical outcome with $L-$levels, i.e., $Y \in (y_1, \cdots, y_L)$, a selective sampling first selects a category with probabilities $n_l/n$, where $n_l$ is the sample size allocated to category $l$ and $n$ is the total number of units in the selective sample. Then, the observation is drawn randomly from the subpopulation with $y_i$ in the selected category. With SOR models, Equations (6) and (7) for this example now become

$$Pr(Y = y_l|x, G = 1) \quad = \quad \frac{w_l f_\theta(y_l|x)}{\sum_{l'=1}^{L} w_{l'} f_\theta(y_{l'}|x)} = \frac{\eta_\gamma(y_l; x)e^{\lambda_l + \ln(w_l/w_L)}}{\sum_{l'=1}^{L} \eta_\gamma(y_{l'}; x)e^{\lambda_{l'} + \ln(w_{l'}/w_L)}}, \tag{8}$$

where $w_l = p_\psi(G = 1|Y = y_l) = \frac{n_l/n}{N_l/N}$ denotes sampling weights, and $N$ and $N_l$ are the supplemental information on the entire population size and the population total for the response category $y_l$, respectively. That is, the sampling weight is the ratio of the probability that a sampling unit belongs to category $l$ in the sample to the probability that it is in category $l$ in the population.

**An example for Case I: Choice-based sampling on a binary response (Example 1 continued)** For a binary outcome $Y \in (y_1 = 1, y_2 = 0)$, consider the population OR model:

$$Pr(Y = y_l|x) \quad = \quad \frac{\exp(\lambda_l + \ln \eta_\gamma(y_l, y_0; x, x_0))}{1 + \exp(\lambda_1 + \ln \eta_\gamma(y_1, y_0; x, x_0))},$$

where $l \in (1, 2)$, $y_0 = y_2 = 0$ and the intercept parameters in the baseline distribution $\lambda = (\lambda_1, \lambda_2)$ have $\lambda_1 = \ln(Pr(Y = 1|x_0)/Pr(Y = 0|x_0))$ and $\lambda_2 = 0$. When the OR function $\eta_\gamma(y_l, y_0; x, x_0)$ takes a log-bilinear form, i.e., $\ln \eta_\gamma(y_l, y_0; x, x_0) = \gamma(x - x_0)y_l$, the population OR model is a logistic model in which $\gamma$ is the regression coefficient in the logistic model. With selection on $Y$, the conditional likelihood in Eqn (8) for the selected sample becomes

$$Pr(Y = y_l|x, G = 1) \quad = \quad \frac{\exp(\lambda'_l + \ln \eta_\gamma(y_l, y_0; x, x_0))}{1 + \exp(\lambda'_1 + \ln \eta_\gamma(y_1, y_0; x, x_0))},$$

where $\lambda'_l = (\lambda_l + \ln(w_l/w_2))$, $w_l (l = 1, 2)$ denotes the sampling weight for each level of the binary choice outcome $y_l$. It is clear that the above conditional likelihood in the selected sample is a likelihood from an SOR model with the same OR function but a different intercept parameter $\lambda'_1 = \lambda_1 + \ln(w_1/w_2)$. Consequently, the population OR function remains invariant to selective sampling and can be identified in the endogenously selected sample as if it were a random sample. It is worth noting that, for a log-bilinear OR function, the above conditional likelihood reduces to that considered in Donkers *et al.* (2003), with a logit model for binary choices and offsets $\{\ln(w_l/w_L)\}, l = 1, 2$. A more general OR function takes the following form:

$$\eta_\gamma(y_l, y_0; x, x_0)) \quad = \quad \frac{\Gamma(y_l + \exp(\alpha + \gamma x))\Gamma(y_0 + \exp(\alpha + \gamma x_0))}{\Gamma(y_0 + \exp(\alpha + \gamma x))\Gamma(y_l + \exp(\alpha + \gamma x_0))}, \tag{9}$$

where $\Gamma(\dot{})$ is the gamma function. Using the property of gamma function, $\Gamma(1+u) = u\Gamma(u)$, one can show for a binary outcome $Y$ the above OR function reduces to $\ln \eta_\gamma(y_l, y_0; x, x_0) = \gamma(x - x_0)y_l$, the OR in the binary logistic model. The OR function in Eqn (9) can be used for a binomial outcome $(n, p_i)$ with over-dispersion, such as the number of months visiting a store in a year. In this case a common modeling option is the beta-binomial distribution assuming $p_i$ follows a beta-distribution with parameters $(a, b)$, $a = \exp(\beta_0 + \beta_1 + \gamma x)$ and $b = \exp(\beta_0)$. Using Eqn (1), one can show that the beta-binomial distribution has the OR function in Eqn (9) with $\alpha = \beta_0 + \beta_1$. Thus, the SOR nests the beta-binomial distribution as a special case because SOR replaces the parametric baseline function in the beta-binomial distribution by a nonparametric baseline function. An illustration of this point in the similar setting of negative binomial for count outcomes is provided in Section 4.

**Case II. Stratified sampling on a continuous outcome.** This sampling scheme first divides the sample space of $Y$ into $M$ mutually exclusive intervals: $\Delta_1, \cdots, \Delta_M$, and selects an interval with probabilities $n_m/n$ and then an observation is selected randomly with equal probabilities from the subpopulation with $y_i$ in the selected interval. Equations (6) and (7) in this case now become:

$$f(Y = y|x, G = 1) \quad = \quad \frac{\sum\limits_{m:y\in\Delta_m} w_m f_\theta(y|x)}{\sum\limits_{l=1}^{L} \sum\limits_{m:u_l\in\Delta_m} w_m f_\theta(u_l|x)} = \frac{\sum\limits_{m:y\in\Delta_m} \sum\limits_{l=1}^{L} 1_{y=u_l} \eta_\gamma(y; x) e^{\lambda_l + \ln(w_m/w_M)}}{\sum\limits_{l=1}^{L} \sum\limits_{m:u_l\in\Delta_m} \eta_\gamma(u_l; x) e^{\lambda_l + \ln(w_m/w_M)}}, \tag{10}$$

where the sampling weight is $w_m = p_\psi(G = 1|y \in \Delta_m) = \frac{n_m/n}{N_m/N}$, and $n_m$ and $N_m$ are the sample size allocated to the interval $\Delta_m$ in the selective sample and the population total in the interval $\Delta_m$ in the population, respectively.

**An example for Case II: Stratified sampling on a normal outcome (Example 2 continued)** Hausman and Wise (1981) considered the above stratified sampling and assumed a normal error distribution for $Y$ as $y_i = x_i\beta + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$. In the simple case of stratified sampling with two strata ($y \leq \alpha$ and $y > \alpha$) and a known strata cutoff value $\alpha$, they showed that

$$E(Y|x, G = 1) \quad = \quad x\beta - \sigma\frac{(1-P)\phi(\frac{\alpha - X\beta}{\sigma})}{(1-P)\Phi(\frac{\alpha - X\beta}{\sigma}) + P}, \quad P = w_2/w_1,$$

where $w_1$ and $w_2$ are the sampling weights for strata 1 and 2, respectively, and $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and distribution function of standard normal, respectively. Thus, the OLS estimates are biased when $P \neq 1$ (i.e., endogenous selective sampling) with bias being a nonlinear and complex function of departure from random sampling, as captured by $P$, cutoff point $\alpha$, and unknown model parameters. To correct for the bias, Hausman and Wise (1981) developed an MLE procedure using a normal regression model for $f_\theta(y|x)$ in the conditional likelihood given in Eqn (6). The bias-correction approach of Hausman and Wise (1981) requires the strata cut-off points be known to data analysts and sampling weights depend on the outcome $Y$ only. As shown in Section 2.2.1, the SOR approach does not require known strata cut-off points (e.g., the value $\alpha$ above).

In certain situations, some strata may have a sampling weight of zero, which creates truncated data. Examples are the demand for a college class or a concert, when the class/concert opening requires a minimum number of students/audience. When the demand falls below this lower limit, the event gets canceled and will not appear in the sample, resulting in a truncated sample. Since truncation is a special case of endogenously stratified sampling, it follows that OLS can also create bias in the truncated samples. In practice, researchers may not know the sample truncation point or even are unaware of the existence of sample truncation problems. The SOR has an advantage of not requiring these knowledge to correct selection bias associated with truncated samples. In the following sections, we will compare SOR with the sampling adjustment methods assuming normal error distributions in handling endogenously stratified samples and truncated samples.

**Case III. Selective sampling directly on a continuous/count outcome.** Both the above selective sampling schemes constitute stratified sampling, whereby selection probabilities depend on $Y$ via a finite number of strata. Case III considers a more flexible case of selection directly on $Y$. One example is on-site sampling mentioned in the Introduction and described further below.

**An example for Case III: On-site sampling with a normal and count outcome (Example 3 continued)** In on-site sampling, the sampling weight $p_\psi(G = 1|y) \propto y$. Shaw (1988) derived the MLE based on the conditional likelihood, assuming $Y$ follows either a normal or a Poisson regression. For a normal outcome, the conditional likelihood is

$$f(y|x, G = 1) \quad = \quad \frac{yf(y|x)}{\int_0^\infty uf(u|x)du}, y > 0. \tag{11}$$

Shaw (1988) showed that OLS estimates of the demand function using on-site samples are biased and proposed an MLE procedure to correct for the bias. For $Y$ following a Poisson regression, the conditional likelihood for the on-site samples becomes

$$f(y|x, G = 1) \quad = \quad \frac{yf(y|x)}{\sum_{u=1}^\infty uf(u|x)du} = \exp(\lambda_i)\lambda_i^{y_i-1}/(y_i - 1)!, \quad y \geq 1. \tag{12}$$

This implies a simple solution to on-site samples with Poisson distribution: one can simply subtract 1 from observed counts in an on-site sample and analyze the resulting data as a random sample using Poisson regression. These correction procedures critically depend on the outcome distributional assumptions. We compare these approaches with SOR for handling on-site samples in Section 4.2.

### 2.3 *Selection on both y and x*

We show above that the OR association parameters in SOR are invariant and unaffected by arbitrary (and possibly unknown) forms of selection on $Y$ only (Selection Type A). We next consider the more general case of selection on both $Y$ and $X$, for which using Bayes' theorem, we have

$$f(Y = y|x, G = 1) \quad = \quad \frac{p_\psi(G = 1|y, x)f_\theta(Y = y|x)}{\int p_\psi(G = 1|u, x)f_\theta(u|x)du} = \frac{\eta_\gamma(y; x)\Big[p_\psi(G = 1|y, x)f_\lambda(y|x_0)\Big]}{\int \eta_\gamma(u; x)\Big[p_\psi(G = 1|u, x)f_\lambda(u|x_0)\Big]du} \tag{13}$$

When sampling weight $p_\psi(G = 1|y, x)$ is known (Selection Type B), it can be added as an offset term when fitting the SOR to the endogenous selected sample to recover both the OR function and baseline function.[6] When sampling weight is of an arbitrary and unknown form, then in general neither the population baseline function nor the OR function is identifiable. An exception is when selection depends on a combination of $Y$ with the strata of the independent variables $X$ (Selection Type C); for this important class of sampling schemes, the OR association parameters in SOR are invariant and can be identified even if the sampling weight is unknown, as shown below.

Specifically, we extend the SOR approach to response-dependent sampling with selection following $p_\psi(G = 1|y, x, s) = p_\psi(G = 1|y, s)$ and $s$ being the strata variable defined by the explanatory variables. Marketing researchers often stratify their samples on independent variables so the response-dependent sampling schemes including such stratification is highly relevant. Because samples from different strata are drawn independently, we consider the conditional likelihood below:

$$f(Y = y|x, S = s, G = 1) \quad = \quad \frac{f(Y = y, x, s, G = 1)}{f(x, s, G = 1)} = \frac{p_\psi(G = 1|y, s)f(Y = y|x, s)}{\int p_\psi(G = 1|u, s)f(u|x, s)du}.$$

Since $Y$ is independent of $s$, given $X = x$, we thus have $f(Y = y|x, s) = f_\theta(y|x)$. When we apply the OR model for $f_\theta(y|x)$, we obtain

$$f(Y = y|x, S = s, G = 1) \quad = \quad \frac{\eta_\gamma(y; x)\Big[p_\psi(G = 1|y, s)f_\lambda(y|x_0)\Big]}{\int \eta_\gamma(u; x)\Big[p_\psi(G = 1|u, s)f_\lambda(u|x_0)\Big]du}. \tag{14}$$

Thus $f(Y = y|x, S = s, G = 1)$ has an SOR model expression with the same OR function $\eta_\gamma(y; x)$ as $f_\theta(Y|x)$, and a strata-specific baseline function $f'_{\lambda'_s}(y|x_0, s) = \Big[p_\psi(G = 1|y, s)f_\lambda(y|x_0)\Big]$. We may consider a more general model that permits not only heterogeneous baseline functions but also heterogeneous OR functions across strata, as follows

$$f(Y = y|x, S = s, G = 1) \quad = \quad \frac{\eta_\gamma(y; x|S = s)\Big[p_\psi(G = 1|y, s)f_\lambda(y|x_0)\Big]}{\int \eta_\gamma(u; x|S = s)\Big[p_\psi(G = 1|u, s)f_\lambda(u|x_0)\Big]du}, \tag{15}$$

where the OR function $\eta_\gamma(y; x|S = s)$ varies by strata $s$. For instance, when a log-bilinear form of OR function in Eqn (3) is used, we can have strata-specific log-odds-ratio parameter $\gamma_k(s)$, which can be used to incorporate functional and preference heterogeneity across stratum in the framework.

[6] An example is Example 1 in which known sampling weights depend on the churn outcome and a covariate (duration of consumers with the company) (Donkers *et al.*, 2003). Even when sampling weight is known, SOR has the advantage of not relying on outcome distributional assumptions.

#### *2.3.1 Estimation without data on sampling weights (SOR)*

When the sampling weight function $p_\psi(G = 1|y, s)$ is unknown, the baseline function $f'_{\lambda'_s}(y|x_0, s) = \Big[p_\psi(G = 1|y, s)f_\lambda(y|x_0)\Big]$ varies across the strata in $S$. Therefore, when estimating the model with unknown sampling weights, we need to assume a different baseline distribution function for each stratum separately. One can fit the SOR model with stratum-specific nonparametric baseline distribution functions on the stratified selective samples as if this were a (stratified) random sample and obtain the estimates of common OR parameters $\hat{\gamma}$ as well as strata-specific $\hat{\lambda}'_s$. When the OR function is also allowed to vary freely across strata (i.e., Eqn (15)), this is equivalent to performing SOR estimation of $\hat{\gamma}_s$ as well as $\hat{\lambda}'_s$ for each stratum $s$ separately. It is unnecessary to correct the OR estimates $\hat{\gamma}$ or $\hat{\gamma}_s$ for selective sampling in either case.

#### *2.3.2 Estimation with supplemental data on sampling weights (SOR-FI)*

When the sampling weights are known, we can use this supplemental information to obtain a more efficient full-information (FI) estimation of model parameters. For MI model such as ours, maximization of the conditional likelihood in Eqn (14) yields consistent but not necessarily fully efficient estimates when sampling weights are known. Fully efficient estimates are obtained when a full set of dummies for the strata $S$ are included in the model, or via an iterative procedure in the absence of the full set of dummy variables for strata $S$ (Scott and Wild, 1997). As shown in Eqn (14), when the sampling weight $p_\psi(G = 1|y, s)$ is known, the logarithm of these sampling weights can be taken as the offsets when estimating the parameters $\lambda$ in the baseline distribution function. Thus, when selection is on both $Y$ and $X$ and sampling weights are known, information can be pooled across strata to estimate the common baseline distribution function $f_\lambda(y|x_0)$. As a result, unlike selection on outcome $Y$ only, where knowledge of sampling weights does not affect the OR parameter estimation, selection on both $Y$ and $X$ can have substantial gain in the efficiency of OR estimates $\hat{\gamma}$ when sampling weights are known compared to when sampling weights are unknown. This point will be further elaborated in the simulation study and real-life data below.

### 2.4 *Selection on $Y, X, \theta$*

In this case, sampling weight $p_\psi(G = 1|Y, X, \theta)$ is unknown because of unknown $\theta$ (Example 4). When $p_\psi(G = 1|Y, X, \theta)$ is of arbitrary form, neither the population OR nor the baseline distribution is identifiable. An exception is when selection depends on $(Y, X, \theta)$ as follows (Selection Type D):

$$p_{\psi}(G = 1|y, x, \theta, s) \quad = \quad h_{1,\psi}(y, s, \theta)h_{2,\psi}(x, s, \theta), \tag{16}$$

where $s$ is the strata variable defined by the explanatory variables $X$; both $h_{1,\psi}(\cdot)$ and $h_{2,\psi}(\cdot)$ are unspecified functions so long as their product gives a probability. Following similar derivation for Eqn (14), we obtain the following conditional likelihood

$$f(Y = y|x, s, G = 1) \quad = \quad \frac{\eta_{\gamma}(y; x)\Big[h_{1,\psi}(y, s, \theta)f_{\lambda}(y|x_0)\Big]}{\int \eta_{\gamma}(u; x)\Big[h_{1,\psi}(u, s, \theta)f_{\lambda}(u|x_0)\Big]du}, \tag{17}$$

where $h_{2,\psi}(x, s, \theta)$ appears in both numerator and denominator and cancels out. It is clear that $f(Y = y|x, s, G = 1)$ has the same OR, $\eta_{\gamma}(y; x)$, as $f_{\theta}(Y|X)$ does and a strata-specific baseline function, meaning that the population OR $\eta_{\gamma}(y; x)$ is invariant to selective sampling and can be identified using the algorithm in Section 2.3.1.

Selection Type D in Eqn (16) is known to hold in many selective sampling schemes for which researchers have at least some control or partial knowledge, including endogenous stratified sampling, truncation and on-site sampling. The SOR nests as special cases the existing parametric sampling-adjustment methods (Table 1) designed specifically for these situations by eschewing outcome error distributional assumptions. Selection Types A and C described in the previous subsection are special cases of Type D by setting $h_{2,\psi}(x, s, \theta) = 1$ and additionally $h_{1,\psi}(y, s, \theta) = h_{1,\psi}(y)$ for Type A and $h_{1,\psi}(y, s, \theta) = h_{1,\psi}(y, s)$ for Type C. By further permitting selection to depend on unknown $\theta$ and leaving unspecified the functional forms of $h_{1,\psi}(\cdot)$ and $h_{2,\psi}(\cdot)$, Type D can permit a much wider range of sampling schemes unobserved to researchers, as compared with selection on $(Y, X)$. For example, $h_{2,\psi}(x, s, \theta)$ permits selection on the conditional mean of $Y$, $E_{\theta}(Y|x)$, which is a function of $(x, s, \theta)$. This captures a form of conformation biases leading to certain types of people more likely to be sampled due to their anticipated responses. Similarly, $h_{1,\psi}(y, s, \theta)$ permits selection on $Y$, $S$ and $\theta$, such as trimming samples spaces of $Y$ and $X$ based on $\theta$. Note that Type D in Eqn (16) means that *given* $s$, the impacts of $Y$ and $X$ on selection are functionally multiplicative but in the entire population, selection can depend on $Y$ and $X$ in a functionally non-multiplicative way.

### 2.5 *Estimation of Standard Errors*

The literature has shown that, for MI models such as ours, the regular standard errors of the OR estimates computed from the inverse information matrix of the above conditional likelihood functions are correct (Scott and Wild, 1997). One exception is that for selection on both $Y$ and $X$ with supplemental data on sampling weights, and when the model does not include a complete set of separate constant terms for each stratum in $S$, these regular standard errors can be used as conservative estimates, but the true standard errors will be smaller[7]. Furthermore, the standard errors for the parameters $\lambda$ in the baseline distribution need to be corrected. Scott and Wild (1997) derived a formula to obtain corrected standard errors in these situations for case-control data. An alternative means of obtaining the estimates of standard errors that can be applied to our setting is based on bootstrap samples. For this purpose, we develop a bootstrap procedure that properly accounts for selective sampling schemes with details described in Online Appendix II. One benefit of the bootstrap method is that it is straightforward to obtain the standard errors of complex functions of these model parameters, such as means, percentiles, and correlations coefficients, etc.

### 2.6 *Practical Usage of SOR*

Regression coefficients (i.e., the $\gamma$ OR parameter) in SOR regression can be used to quantify and test for the conditional association between each independent variable and $Y$, given all the other independent variables. These coefficients in SOR remain unaffected by the response-dependent sampling schemes (Selection Types A, B, C and D) described in Sections 2.3 and 2.4. Thus, one practical usage of SOR is a tool to identify/select important explanatory variables with selective samples. Furthermore, these SOR regression coefficients produce results identical or closely related to standard analysis results when population distribution follows the GLMs (Section 2.1). To give some examples, these SOR regression coefficients are the logistic (Poisson) regression coefficients when $Y$ given $X$ follows a logistic (Poisson) regression; they are the linear regression coefficients divided by the error variance when $Y$ given $X$ follows a normal regression model. The SOR model reveals which parameters in these standard models are invariant to response-dependent sampling and identifiable without knowing or modeling functional forms of these selection schemes. Notably,

[7] The regular ones can be up to three times of the true ones (see notes of Tables 1 and 2).

the intercept parameters in these standard models are not invariant to selective sampling.

When population does not follow any parametric distribution, one can still use SOR for variable identification/selection. In this case it becomes a moot point to situate the OR association parameters within the parametric modeling framework as above. To aid the interpretation and use of the SOR, we connect these OR parameters with other quantifies of interest as follows. When sampling weight is known and the entire distribution is recovered, summary statistics of the predictive distribution (Eqn (5)) useful for continuous and count outcomes, such as mean and tail probabilities, can be obtained for consumer profiling and targeting. In particular, the regression mean is

$$E(Y|x) \quad = \quad \frac{\sum_{l=1}^{L} u_l \exp(\lambda_l + \ln(\eta_\gamma\{u_l; x_{i1}, \cdots, x_{iK}\}))}{\sum_{l=1}^{L} \exp(\lambda_l + \ln(\eta_\gamma\{u_l; x_{i1}, \cdots, x_{iK}\}))}. \tag{18}$$

Estimation and testing of covariate effects on the outcome mean can be done as follows. With the log-bilinear OR function in Eqn (3), the marginal effect of a continuous covariate $x_k$, which informs the average change in $Y$ for one-unit change in $X_k$ holding other covariates constant, is

$$\frac{\partial E(Y|x)}{\partial x_k} \quad = \quad \gamma_k \sigma^2_{Y|x}, \tag{19}$$

where $\sigma^2_{Y|x} = E(Y^2|x) - E^2(Y|x)$, and the expectation is taken with respect to the multinomial distribution in Eqn (5). The equation above connects the OR parameter $\gamma_k$ with the marginal mean effect of $x_k$, and is known to hold for the special case of GLMs. This connection has several desirable properties enhancing the interpretation and use of OR parameters: (1) testing the marginal effect of $X_k$ on the mean of $Y$ is zero is equivalent to testing $\gamma_k = 0$; (2) the effect of a covariate on OR is in the same direction as its marginal effect on mean; (3) the ratio of the marginal effects for two covariates ($X_k$ and $X_j$) is the same as the ratio of two OR parameters ($\frac{\gamma_k}{\gamma_j}$); (4) the marginal effect of $X_k$ is not restricted to be constant when $X_k$ varies, permitting a flexible data-driven nonlinear mean function. An important implication of the last point is that SOR relaxes not only the distributional assumptions, but also the linearity-in-mean assumption imposed in linear regression models. More flexible relationships between OR parameters and marginal mean effects can be obtained using higher-order and interaction terms or alternative forms of OR functions.

## 3. Simulation Studies

### 3.1 *Study I: Endogenously Stratified Sampling*

In this section, we use synthetic data to demonstrate the benefits of selective sampling and evaluate the proposed method's performance for analyzing selective samples. The simulation emulates a setting in which the aim is to assess the relationship between the dependent variable $Y$ (total purchase expenditure or the count of shopping visits) and two independent variables $X_1$ and $X_2$, where variables ($Y$, $X_1$ and $X_2$) can have either symmetric or skewed distributions. Data were simulated from the following four regression models: (1) Normal regression: $Y|x \sim N(\mu_{Y|x}, \sigma^2)$, (2) Truncated normal regression: $Y|x \sim TN(\mu_{Y|x}, \sigma^2, \text{lb} = 0)$, (3) Poisson regression: $Y|x \sim \texttt{Poisson}(\mu_{Y|x})$, and (4) Zero-truncated Poisson regression: $Y|x \sim \texttt{TPoisson}(\mu_{Y|x}, \text{lb=1})$, where $\mu_{Y|x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ in (1) and (2) and $\ln \mu_{Y|x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ in (3) and (4), and $Y \geq \text{lb}$ in (2) and (4). The independent variable $X_1$ is a binary variable with an uneven distribution: $\Pr(X_1 = 1)$=0.15 and $X_2$ follows a standard normal. We set $\beta_0 = 0, \beta_1 = \beta_2 = 0.1, \sigma = 1$.

All four population regression models for $f_\theta(Y|x)$ can be re-expressed as SOR models with $\ln \eta_\gamma(y; x) = \gamma_1(y - y_0)(x_1 - x_{10}) + \gamma_2(y - y_0)(x_2 - x_{20})$, where the log-odds-ratio parameter $\gamma_j = \beta_j/\sigma^2 = 0.1, j = 1, 2$ for models (1) and (2) and $\gamma_j = \beta_j = 0.1, j = 1, 2$ for models (3) and (4), respectively, and the baseline distribution $f(y|x_0)$ is normal, half-normal, Poisson, and zero-truncated Poisson for models (1) to (4), respectively. Both $\gamma_j$ and $\beta_j$ can be used equivalently for testing the relationship between $X_j$ and $Y$. The purpose here is to uses these models to demonstrate that the SOR approach performs as well as extant sampling adjustment methods when their assumed familiar distributions (i.e., normal and Poisson) are correct, and performs better than them when there are departures (i.e., boundedness/truncation) from these standard distributions. In fact, the baseline distribution in SOR is unspecified and can permit arbitrary forms of departures from familiar normal and Poisson distributions. The use of truncated distributions here should be viewed as one particular instance of departures from normal and Poisson distributions.

We generate a population of 1,000,000 consumers from each of the four models, mimicking large consumer databases owned by firms. We consider the following schemes of sampling 500 consumers.

**Sampling Schemes**

- `RS`: The benchmark sampling approach is the simple random sampling of 500 consumers.

- ES-Y: We conducted an endogenously selective (ES) sampling that over-samples consumers with extreme purchase outcome values. We allocated $p = 25\%$ of sample size to be a random sample of the population and the remaining sample size as an augmented sample that drew randomly from consumers with $Y$ values above the $\alpha$=90th percentile or below the $(100 - \alpha)$=10th percentile.

- ES-YX: Sample selection depends on a combination of both $Y$ and a stratification of $X$ values. We divided sample size equally between $X_1 = 0$ and $X_1 = 1$. This will distribute $X_1$ more evenly in the selected sample, which, as shown below, can further significantly increase its estimation efficiency. Then, within each stratum, a selection based on $Y$ only as described in ES-Y is conducted.

**Estimation Method** For each generated sample, we applied the following estimation methods.

- Regressions without adjusting for endogenous selection. These are OLS (truncated normal regression (TNREG)) for the (truncated) continuous outcome and Poisson (truncated Poisson (TPREG)) regression for the (truncated) count outcome without adjusting for endogenous selection. We expect that they perform well in random sampling but yield bias with endogenous sampling.

- Parametric sampling adjustment methods. These extant sampling adjustment methods assume parametric outcome distributions and are denoted as HW and HW-FI (Hausman and Wise, 1981) for ES-Y and COSS-FI (Cosslett, 1993) for ES-YX, where "FI" means that data on sampling weights are available and used.[8] These are benchmark adjustment methods for comparison.

- POR. This estimation fits an odds ratio model with a parametric baseline distribution. This is the same method as described in Sections 2.2.1 and 2.3.1, except that it uses parametric baseline distribution functions: a normal (Poisson) distribution for the continuous (count) outcome. Thus, the parametric model for baseline distribution is correctly specified for random samples in models (1) and (3) but mis-specified for other models and for endogenous samples where the baseline distribution in the sample depends on sampling weights. We use this model to demonstrate the potential estimation bias associated with the mis-specification in the baseline distribution.

- SOR. This estimation fits an odds ratio model with the nonparametric baseline distribution

[8] Hausman and Wise (1981) considered only a normal error distribution, so we extend their approach to Poisson regression here. Cosslett (1993) only described the method requiring sampling weights, which is the COSS-FI.

described in Sections 2.2.1 and 2.3.1. No data on sampling weights are required or used.

- SOR-FI: This is the same as the SOR above but uses supplemental data on sampling weights. This is the full-information estimation approach described in Sections 2.2.2 and 2.3.2.

**Result** Conclusions for $Y$ being continuous and count are broadly similar. For brevity, we summarize the main findings for modes (1) and (2) below. Appendix III provides a more detailed description of findings, results for count data (Models (3) and (4)) and further results on models with more complex functional forms and small sample sizes.

Robustness of SOR for correcting selection bias. Fig 1 plots the means ± SD of the estimates of $\beta_1$ and $\gamma_1$ over repeated samples. In random sampling (RS), all estimation methods (except HW for truncated normal data) perform equally well and can recover the true parameter value with no bias and the same estimation variability (Fig 1). They also have the coverage rate of 95% confidence interval equal to nominal rate and the same power to reject the null hypothesis of no association between $X$ and $Y$ (Tables 1 and 2 in Appendix III). Although SOR has substantially more parameters than the parametric methods and POR, no estimation efficiency is lost.

As expected, methods without sampling adjustment (OLS/TNREG) yield significant bias in estimating $\beta_1$ with ES-Y and ES-YX (Fig 1), demonstrating selection bias associated with endogenously selected samples. The parametric sampling adjustment methods (HW, HW-FI and COSS-FI), the benchmark adjustment methods, can correct for selection bias when the outcome distributional assumptions are met (Normal in Fig 1). However, the parameter estimates have significant bias when the outcome follows truncated normal (red lines in Fig 1): the red dots for means of their estimates under selective sampling (ES-Y and ES-YX) differ significantly from the true value and also from the means of estimates under random sampling. This indicates that the performance of these parametric sampling adjustment methods depends critically on the distributional assumptions, and when these assumptions are violated, these methods cannot recover either the true parameter values or the estimates from random samples. By contrast, the SOR can estimate the true regression coefficient value without bias across sampling schemes and outcome distributions.

To further demonstrate the importance of properly modeling outcome distributions in handling endogenously selected samples, consider POR, which is the same as SOR except that the baseline function $f(y|x_0)$ is specified parametrically. POR appears to perform reasonably well when $Y$ follows a conditional normal, even though its normal baseline distribution is mis-specified for endogenous samples (`ES-Y` and `ES-YX`). This may mean that the log-odds-ratio parameter $\gamma$ is more robust to sampling schemes than the regression coefficient $\beta$ as an association measure. However, the estimates of association parameters from POR suffer from significant bias in endogenously selected samples (`ES-Y` and `ES-YX`) when $Y$ follows a skewed Poisson distribution (Fig 1 in Appendix III), indicating the significant advantage of SOR over POR for handling selective samples.

Estimation efficiency of SOR. SOR-FI has the same variability of estimates (i.e., the same length of blue lines in Fig 1) as HW-FI and COSS-FI do under the normal error distribution. This means the SOR method performs as efficient as these parametric sampling adjustment methods.

Efficiency gain in selective sampling schemes. SOR estimates have substantially smaller SDs in `ES-Y` and `ES-YX` as compared to `RS` (Fig 1), indicating substantial efficiency gain from selective sampling. For `ES-Y` sampling, the reduction in sample size to achieve the same level of estimation precision, compared with `RS`, is 60—70% (Appendix III Tables 1 and 2 column "Eff."). Furthermore, knowledge of sampling weights does not improve the estimation of association parameters.

By contrast, for `ES-YX` sampling, knowledge of sampling weights and their use in estimation as the SOR-FI does can significantly improve the efficiency of the association parameter estimation: SDs of SOR-FI estimates are substantially smaller than those of SOR estimates (Fig 1 `ES-YX`). The power to detect the effect of $X_1$ on $Y$ is increased from almost no power (6%—14%) with random sampling to almost full power (90%—100%) with `ES-YX` (Appendix III Tables 1 and 2), a striking difference indicating great gain with selective sampling. The increase in power to detect relevant independent variables increases the percentages of correct models selected across all simulated datasets from below 10% in `RS` to 90% in most cases in `ES-YX` (Column "CMS" in Appendix III Tables 1 and 2). `ES-YX` also increases the reduction in required sample size, compared with `RS`, to 90—94%, a significant improvement on the `ES-Y` (Appendix III Tables 1 and 2 column "Eff.").

Factors to consider in practice. In selective sampling, one can vary the following factors as noted in the "Sampling Schemes": (1) the value of $\alpha$ (used to define strata of extreme $Y$ values based on above $\alpha th$ and below $(100-\alpha)th$ percentiles of $Y$); (2) the value of $p$ (the proportion of sample size allocated to a random sample), and (3) whether to stratify on unevenly distributed covariates (i.e. ES-YX .vs. ES-Y ). Fig 2 investigates the effects of these factors on sampling efficiency and magnitude of selection bias by varying these factors in the above simulation set-up. The left panel of Fig 2 plots the sample information averaged over 500 simulated data, where sample information is defined as the ratio of sampling size required for RS divided by that required by ES-YX and ES-Y, separately, to have the same estimation precision.[9] The right panel of Fig 2 plots the percentage bias of the estimates, where percentage bias is defined as $\frac{1}{2}\sum_{k=1}^{2}\frac{|E(\hat{\beta}_k)-\beta_k|}{\beta_k}, k = 1, 2$.

Below are some major findings: (1) Sample information increases with more extreme strata cutoff point $\alpha$, and the larger proportion $(1-p)$ of sample size allocated to nonrandom samples in selective sampling; efficiency gain is smallest when $\alpha = 50th$ percentile of Y or $1-p=0$; (2) ES-YX can obtain substantially more sample information than both ES-Y and RS; (4) By increasing $(1-p)$, the proportion of sample sizes oversampling extreme values of $Y$, ES-YX can obtain more than 12 times of sample information than RS, and ∼6 times more information than exogenously stratified sampling (i.e., ES-YX with $p = 1$); this indicates the substantially additional gain associated with oversampling extreme $Y$ values, even after stratifying on the unevenly distributed covariate $X_1$; (5) Efficiency gain is largest when samples are limited to have only extreme $Y$ values ($p = 0$);[10] (6) Factors increasing the sampling efficiency also increase the selection bias of OLS[11]whereas SOR realizes the efficiency gain of selective sampling without incurring selection bias.

3.2 *Simulation Study II: Selective Sampling Resulting in Missing Outcomes*

The above simulation considered endogenously stratified sampling that over-sampled extreme outcome values for cost-effectiveness purposes, and showed that when sampling weights are known,

[9] Estimation precision is quantified using the D-error measure $|\Sigma|^{1/K}$(Arora and Huber, 2001), where $\Sigma$ is the covariance matrix of the parameter estimates, and $K$ is the number of covariates.

[10] One strength of the SOR approach is that the OR in a local region is unaffected by trimming data in other regions. Thus there is a less need of complete data coverage for consistently estimating OR than for estimating conventional mean regression parameters which are affected by trimming sample space of Y. To ensure data to cover the region of interest, however, we do recommend setting $p$ and the sampling ranges of $X$ and $Y$ properly.

[11] An exception is the stratification on unevenly distributed covariates. As shown in Fig 1, the selection biases for OLS in ES-Y and ES-YX are in opposite direction and may not be comparable.

SOR-FI analyzes selective samples as if it were a random sample except incorporating sampling weights as offsets.[12]In this section, we consider more flexible sampling rules with unknown sampling weights (e.g., Example 4). A useful result is that the SOR can identify and consistently estimate OR association parameters without modeling the functional form of response-dependent selection.

In this study, the $Y$ value in a random sample is selected for observation according to selection mechanisms depicted in Fig. 3. Thus, the selection indicator $G$ is the non-missingness indicator for $Y$. One example of these selection rules is data confidentiality, whereby extreme $Y$ values (too high or two low) have reduced or no chance to be shared with others and are designated missing by data owners. Another example is nonresponse to sensitive questions in surveys. For example, nonresponse to a question on expense is more likely when the expense is too high or too low. In these settings, selection may depend directly on $Y$ taking infinite possible values, rather than on a finite number of known strata defined by $Y$. We assume that data analysts do not know the forms of these response-selection rules. Thus, the extant methods for stratified sampling considered in the above section are not applicable here, because these methods require selection depend on a finite number of strata with known cutoff points of $Y$ values for determining strata. It is unclear how they may be generalized to the sampling schemes considered here with unknown sampling rules and unknown outcome distributions. By contrast, the SOR method is more general and can be applied to these selection rules. As described in Section 2.2.1, there is no need to specify *a priori* the form of selection rules in SOR, which means that selection on $Y$ can be of arbitrary form. An alternative and classical approach for handling missingness in $Y$ is the Heckit procedure (Heckman, 1979), which takes the following form for a normal linear outcome:

$$Y_i = X_i\beta + \epsilon_{1i}, \qquad G_i^* = Z_i\delta + \epsilon_{2i}, \quad (\epsilon_{1i}, \epsilon_{2i}) \sim \text{BVN}(0, \Sigma),$$

where BVN stands for bivariate normal and the selection indicator $G$ is determined by the latent variable $G^*$ as G=1 if $G^* > 0$ and G=0 if otherwise. The model estimation can be performed through maximum likelihood, denoted as FIML, or a simpler two-step procedure. As seen above, one important limitation of the Heckit selection model for use with selective sampling is that the

[12]Appendix IV demonstrates the use of SOR for endogenously stratified sampling with the real-life data from Example 2. The conclusions are broadly similar to those reached in simulation study I.

BVN error model imposes a linear probit selection rule on the outcome $Y$, and the estimation results depend sensitively on the assumed selection form. As demonstrated below, departures from the linear probit selection rules can lead to significant estimation bias.

In this simulation study, we generate a random sample of 10,000 observations from the normal linear model, as specified in model (1) in Section 3.1 above. For each sample, we apply each of the four selection rules (Selection on $Y$ or $(Y, \theta)$) in the top row of Fig. 3 to create endogenously selected samples. We then apply three methods to analyze each sample: the OLS, Heckit, and SOR. We include variables $X_1$ and $X_2$ in both stages of Heckit. Unlike OLS, both Heckit and SOR adjust for nonignorable missingness in the sample selection problem; unlike Heckit, however, SOR analyzes the endogenously selected samples as if they were random samples without the need to model the selection mechanism. Results over 500 simulated datasets are shown in Table 3. As expected, the OLS has large bias that can go in either direction (the ratio of bias to true value goes from -86% to 300%). Heckit can correct for bias for the linear probit selection rule (Selection Rule 1 in Table 3) and reduce the bias to less than 10%. However, there is a considerable variation in estimates. When the selection rules differ from linear probit, the estimates from Heckit have significant bias and can be even worse than the OLS. This demonstrates the exquisite sensitivity of the Heckit procedure to the assumed selection rules. By contrast, SOR does not require modeling selection mechanisms and performs very well across all selection rules with no bias and substantially smaller variability in estimates. We then further considered the `ES-YX` selection (Selection Type C described in Section 2.3) that depends on a combination of $Y$ and the strata on $X_2$ as well as `ES-YX`$\theta$ selection (Selection Type D described in Section 2.4) depending additionally on $\theta$ as depicted on the bottom row of Fig. 3. Again, we assume that data analysts do not know the exact functional form of selection on $Y$ within each strata, although they know selection stratified on $X_2$. The findings under the `ES-YX` are broadly similar to those under the `ES-Y` (Table 3), except that Heckit also shows significant bias in Selection Rule 5. This is due to selection's dependence on the interaction of $Y$ and $X$ in Selection Rule 5, whereas in Selection Rule 1, selection depends on $Y$ only. Note that the selection rules 2 and 5 permit selection to depend on $Y$, $X$ and $\beta$ (population parameters). The dependence on $\beta$ can occur when people performing selection have the population data and know $\beta$. Table 7 in

Online Appendix III reports results with sparse data (sample size=100 with about 50% missingness in $Y$). The main conclusions remain the same.

Besides imposing the restrictive assumption of linear probit selection, another limitation of using Heckit for selective sampling is that it requires knowing the $X$ values for all observations in the random sample to estimate the selection rules and thus cannot be applied to selective sampling with $X$ and $Y$ simultaneously observed or missing. In the next section, we compare different methods for handling truncated and on-site samples, for which Heckit is not applicable.

## 4. Application: Truncated and On-site Samples of Shoppers' Store Visits

Count data are ubiquitous in marketing, economics, social and health research. Examples are prescription counts, website visits and store visits. Our sample comes from a retail store's shoppers database and contains 3092 customers who made at least one shopping visit to the store during the first year of store opening. The purpose is to construct a count model of the store shopping visits made by the shoppers to identify relevant demographic, social-economic, geographic profiles (as listed in Table 4), and to conduct consumer profiling and targeting to boost sales[13]. Such studies often use either administrative databases generated during routine operations in firms, hospitals, organizations and government, or surveys of consumers. These surveys or administrative databases frequently include only individuals engaging in the activity of interest. A sample truncation issue arises if the interest is in the general population rather than the individuals participating in the activity exclusively. Even if one is interested in the latter individuals only, modeling has to account for the bounded distributional feature of no zero counts in the data (Shonkwiler and Englin, 2009). Besides truncation, the sampling probabilities in on-site samples further depend on the *intensity* of the activity of interest —the number of store visits. Similar examples abound in marketing and management, for instance, the demand for a recreational site (Shaw, 1988, Englin and Shonkwiler, 1995), household demand studies sampled in supermarkets, the number of physician visits sampled in doctors' offices, and transportation uses sampled on transportation modes. Below we revisit Example 3 and demonstrate the use of the SOR for truncated and on-site samples.

[13] See Wachtel and Otter (2013) for an example of consumer targeting using demographic variables.

### 4.1 *Truncated Samples*

Similar to the truncation in continuous outcomes (Case II in Section 2.2), fitting truncated count data with regular count data models ignoring truncation can yield significant bias in parameter estimation. To account for truncation, truncated count data models (Shaw, 1988, Englin and Shonkwiler, 1995) can be used to provide consistent model estimation with the following PMF:

$$f_c(Y_i|x_i) = f_\theta(Y_i|Y_i > c, x_i) \quad = \quad \frac{f_\theta(Y_i|x_i)}{1-F(c)}, \quad y = c+1, \cdots, \text{ where } F(c) = \sum_{k=0}^{c} f_\theta(Y_i = k|x_i), \tag{20}$$

where $f_\theta(Y_i|x_i)$ is the PMF for the un-truncated count distribution. When $f_\theta(Y_i|x_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$ and $c = 0$, we obtain the PMF for the zero-truncated Poisson regression as

$$f_c(Y_i|x_i) = f_\theta(Y_i|Y_i > 0, x_i) \quad = \quad \frac{\lambda_i^{y_i}}{(e^{\lambda_i}-1)y_i!}, \tag{21}$$

where $\lambda_i = \exp(\beta_0 + \sum_{k=1}^{K} \beta_k x_{i,k})$. One can generalize the truncated Poisson regression to permit over-dispersion using the truncated negative binomial (NB) regression, in which

$$f_\theta(Y_i = y|x_i) = \frac{\Gamma(y+\frac{1}{\alpha})}{\Gamma(y+1)\Gamma(\frac{1}{\alpha})}(\alpha\lambda_i)^y \left[1+\alpha\lambda_i\right]^{-(y+1/\alpha)}, \tag{22}$$

where $\alpha \geq 0$ is a nuisance parameter to be estimated along with $\beta$ with $\alpha = 0$ giving Poisson regression. The NB regression can be derived as a random-intercept Poisson regression with its mean as $\exp(\beta_0 + \sum_{k=1}^{K} \beta_k x_{i,k} + b_{0i})$, where $\exp(b_{0i})$ follows a Gamma distribution with mean 1 and a dispersion parameter of $\alpha$. Thus the NB extends a Poisson regression with only observed heterogeneity to include both observed and unobserved heterogeneity in count means.

As shown in Eqn (20), the mean structure in truncated samples involves $F(c)$. Thus correct identification of mean structure using truncated samples depends more critically on the correct specification of distributional forms that yield correct $F(c)$. Indeed, parametric truncated count models are more sensitive to departures from the imposed distributional forms, and when mis-specified can yield inconsistent estimates even for mean parameters (Englin and Shonkwiler, 1995), not to mention count probabilities and other distributional aspects. Thus, the importance of not relying on distributional assumptions becomes more salient for truncated data. We demonstrate below that the SOR provides a framework to generate distribution-free models nesting these parametric truncated count models as special cases and avoiding the reliance on distributional assumptions.

It is readily seen the zero-truncated Poisson in Eqn (21) with a log-linear function for $\lambda_i$ has an log-bilinear OR function and a parametric zero-truncated Poisson baseline PMF, and thus is a special case of SOR. The NB and truncated NB regression have the following OR function

$$\ln \eta(y; x) = \sum_{k=1}^{K} \beta_k (y - y_0)(x_k - x_{0k}) -$$

$$(y - y_0) \left\{ \ln \left[ 1 + \exp \left( \ln \alpha + \beta_0 + \sum_{k=1}^{K} \beta_k x_k \right) \right] - \ln \left[ 1 + \exp \left( \ln \alpha + \beta_0 + \sum_{k=1}^{K} \beta_k x_{0k} \right) \right] \right\} . \quad (23)$$

Inside the bracket of Eqn (23) is the difference of two functions of the form, $\ln(1 + \exp(u))$, that is called the softplus function in the machine learning literature. We thus name this OR function as log-linear-softplus (LLS). The NB has a parametric baseline PMF as

$$f(y|x_0) \quad = \quad \frac{\Gamma(y + \frac{1}{\alpha})}{\Gamma(y+1)\Gamma(\frac{1}{\alpha})} \left[ \exp \left( \ln \alpha + \beta_0 + \sum_{k=1}^{K} \beta_k x_{0k} \right) \right]^{y} \left[ 1 + \exp \left( \ln \alpha + \beta_0 + \sum_{k=1}^{K} \beta_k x_{0k} \right) \right]^{-(y+1/\alpha)} \quad (24),$$

which is an NB distribution with mean $\exp\left(\beta_0 + \sum_{k=1}^{K} \beta_k x_{0k}\right)$ and the dispersion parameter $\alpha$. The baseline PM for truncated NB is the truncated version of the above PMF. These decompositions suggest a way to nest these parametric count models within SOR by replacing the parametric baseline PMFs with nonparametric PMFs. One can test the distributional assumptions of the parametric count regressions by model selection statistics, such as likelihood ratio test or AIC, to compare the parametric count models and the SOR with the same OR functions. Should the parametric count model be rejected, SOR provides an alternative regression model to use.[14]

Table 4 summarizes the model fitting results from six store visit demand models: Poisson and zero-truncated Poisson (TP) regression, NB and zero-truncated NB (TNB), SOR with log-bilinear (SOR$^{LLL}$) and SOR with log-linear-softplus (SOR$^{LLS}$) OR function. Fig. 4 presents the observed and predicted counts from the six models with the two SOR models generating the same fitted values. It shows that the sample contained no consumers with zero store visits (i.e.,

[14] Model selection is needed for NB because unlike the GLMs whose parameters in the OR and baseline functions are variation independent (Chen, 2007), and thus the baseline function provides no additional information about OR parameters, the form of the baseline function in the NB model can provide additional information about OR parameters. Consequently unlike for GLMs, it is possible to gain estimation efficiency of the OR parameters for NB compared with SOR.

sample truncation), indicating the need to use truncated regressions. Furthermore, there are large discrepancies between the observed and predicted counts for both Poisson and TP regressions, an indication of over-dispersion of count outcome. The use of NB and TNB reduces the discrepancies (Fig. 4) and improves model fitting with greater likelihood and smaller AIC (Table 4). Among all four parametric models, TNB performs best, indicating the importance of accounting for both over-dispersion and truncation. However, there are still significant and systematic mismatches between observed and predicted frequencies for TNB. In contrast, Fig. 4 shows that SOR provides excellent fit to the marginal counts of store visits. Table 4 also shows that the two SOR models have the same likelihood that are substantially greater than all the other four parametric count data models. According to AIC, SOR with the log-bilinear OR function is the best model.

Besides incorrectly estimating the count probabilities (Fig. 4), these parametric (truncated) count models can lead to incorrect identification of relevant consumer profiling variables. For the latter issue, one simple approach to correcting for over-dispersion/incorrect random-effect distribution is to use empirical sandwich standard errors of parameter estimates (Cameron and Trivedi, 2009). The robust standard errors are reported as $\text{SE}_R$ with the default model-based standard errors as $\text{SE}_D$ in the parenthesis in Table 4, which shows large differences between them. The $t_D$ statistics using model-based $\text{SE}_D$ for Poisson and TP indicate that all variables except `Age` are significant at the 0.05 level, whereas $t$ statistics using robust $\text{SE}_R$ show significance only for `DTS`, $\texttt{DTS}^2$ and `Married`. The model-based $\text{SE}_D$ for NB and TNB also have substantially downward bias compared with the robust $\text{SE}_R$, in some cases by more than 50%, although the downward bias is smaller than Poisson and TP. The downward bias in model-based $\text{SE}_D$ also causes TNB to incorrectly identify `Kids` as a significant predictor. The estimation results from SOR show broad similarity in $t$ statistics with those using the robust standard errors from the four parametric count regression models, and find statistical significance only for `DTS`, $\texttt{DTS}^2$ and `Married`. This indicates that the SOR's performance is on-par with existing robust procedures for truncated data, yet SOR has the advantages of correctly estimating the count probabilities (Fig. 4) and providing a full data distribution needed for other marketing decision problems and for being used in Bayes' theorem to handle general selective sampling schemes. Overall, the analysis suggests that SOR provides inference of truncated samples that is robust to distributional assumptions.

### 4.2 *On-site Sampling*

To demonstrate the merits of SOR for on-site sampling problems, we consider the following scenario. Suppose that we need to conduct an intercept interview of consumers in the store to obtain data on those independent variables. In such on-site sampling situations, the probability of being sampled is proportional to the outcome, the number of store visits, which creates an endogenously selected sample. One main reason to use on-site sampling is that a random sampling of the general population is unlikely to contain enough people who have made shopping visits to the store. We conduct three experiments. In the first experiment, we generate an on-site sample of 3092 customers from the population TP model estimated in Table 4 with sampling probability proportional to the number of store visits. We then apply five estimation methods: TP, TP adjusting for on-site sampling (TP-A) (Eqn (12)), TNB, TNB adjusting for on-site sampling (TNB-A), and the SOR. The TNB-A uses the following PMF for on-site samples (Englin and Shonkwiler, 1995)[15]

$$f_\theta(Y=y|x,G=1)=\frac{y_i\Gamma(y+\frac{1}{\alpha})}{\Gamma(y+1)\Gamma(\frac{1}{\alpha})}\alpha^y\lambda_i^{y-1}\left[1+\alpha\lambda_i\right]^{-(y+1/\alpha)}. \tag{25}$$

Note that the parametric PMFs in Eqns (12) and (25) are also special cases of SOR in that their baseline distributions take parametric forms. We repeated the sampling and estimation 500 times, with results over all samples reported in the top panel of Table 5. In the second experiment, we follow the same procedure, except that the on-site samples were generated from the population TNB model estimated in Table 4 and the SOR with the log-linear-softplus OR function was used in SOR estimation. In the third experiment, on-site samples were generated from the population SOR model with log-bilinear form estimated in Table 4, with results reported in the bottom panel of Table 5. As a comparison, random samples of the same sample size were also drawn from the population models and estimated in the three experiments.

As shown in the top panel of Table 5, TP and TNB (TP-A and TNB-A) can recover the population parameters with random sampling (on-site sampling) when the population follows a truncated Poisson distribution. TP and TNB (TP-A and TNB-A) produce, however, biased estimates for

[15] Englin and Shonkwiler (1995) used a slightly different parameterization replacing $\alpha$ with $\alpha_0/\lambda_i$, which fitted their data better but fitted our data worse than the parameterization used here.

on-site sampling (random sampling). This indicates these parametric procedures depend on the correctness of sampling weights. When the population departs from the assumed distributional forms, the estimates assuming incorrect parametric distributional forms (TP and TP-A in the middle panel of Table 5, as well as TP, TP-A, TNB and TNB-A in the bottom panel of Table 5) are subject to bias regardless of sampling schemes. Furthermore, when distributional forms are misspecified, parametric sample adjustment methods (TP-A and TNB-A) in on-site samples do not recover the estimates of their counterparts (TP and TNB) in random samples either. These results indicate that the parametric methods depend critically on the outcome distributional assumptions. Only SOR performs well in all scenarios as shown in Table 5, indicating its robustness to both outcome distributional assumptions and sampling weights. Although TNB-A performs better than TP-A, its performance depends on the assumed forms of random effects and other distributional assumptions imposed in NB; also the estimation algorithm of TNB-A loses the simplicity of that of TP-A. By contrast, SOR does not need to impose these distributional assumptions and excels in simplicity, in that the algorithm used to analyze random samples can be straightforwardly used for on-site samples. Finally, the D-error measure in Table 5 shows that on-site sampling can substantially increase the estimation efficiency. In the bottom panel of Table 5, D-error is reduced from 0.663 for random sampling to 0.082 for on-site sampling, a reduction of 90% in sample size required for the same estimation precision. This is due to over-sampling of rare and informative large counts in the on-site samples. For the reason stated in Footnote 14, we do find that unlike when the population follows GLMs (Poisson here) for which the estimation efficiency is about the same for correct parametric models and SOR models (row "$D \times 1000$" in top panel of Table 5), estimation efficiency is higher (D-errors are smaller) in correct parametric models than SOR when the population follows truncated NB (row "$D \times 1000$" in the middle panel of Table 5), although they all provide consistent results in this case. In all cases, AIC is capable of choosing the correct and most efficient (smallest D-errors) method for use with on-site samples (Table 5).

### 4.3 *Managerial Implications*

The estimated demand functions can be used for profiling and targeting store shoppers. Unlike Shonkwiler and Englin (2009) whose method can be used for inferring behavior of the (unconditional) average shopper, our approach is more flexible and permits conditioning on drivers of shopping trips. To illustrate, we examine the effects of $\texttt{DTS}$ (distance to store) on the expected shopping frequency and proportion of returning shoppers in the population of store shoppers, holding other covariates constant at their population mean values. This analysis contributes to understand the impact of shopping distance on demand from the shoppers of the focal store or to take the finding to a similar store. Table 6 compares the results using the demand functions estimated from different sample adjustment methods. For the truncated sample, we use the estimated PMFs for TP, TNB and SOR$^{LLL}$ in Table 4, among which SOR$^{LLL}$ fits data best with the smallest AIC value. The expected shopping visits for shoppers at the $95th, 50th$ and $5th$ percentiles of $\texttt{DTS}$ (denoted as $E(Y_{\texttt{DTS95}})$, $E(Y_{\texttt{DTS50}})$, $E(Y_{\texttt{DTS5}})$) are all close across the three models (Table 6 under "Truncated Sample"), indicating that incorrect distributional assumptions in TP and TNB have less effect on these quantities estimated using this truncated sample. Significant differences, however, exist in the estimated proportions of returning shoppers ($Pr(Y > 1)$) across these models (Table 6), indicating bias associated with TP and TNB. This finding is consistent with the pattern observed in Fig. 4.

For on-site samples, we use the demand functions estimated using TP, TP-A, TNB, TNB-A and SOR in the bottom panel of Table 5 (i.e., with population distribution following SOR that fits observed data best). Both TP and TNB do not adjust for on-site sampling, and they estimate the quantify of shopping trips for the sample of shoppers *interviewed* at the store, instead of the *population* of shoppers. As expected, the predicted average shopping visits from TP and TNB have large upward bias, indicating strong selection bias associated with on-site samples. TP-A and TNB-A attempt to correct for the selection bias. TNB-A performs better than TP-A but still has significant upward bias, e.g., $E(Y_{\texttt{DTS5}})$ estimated as 4.08 versus a true value of 2.58 and $Pr(Y_{\texttt{DTS5}} > 1)$ estimated as 0.62 versus a true value of 0.38 with both estimates having about 60% bias (Table 6). We note the bias may be in another direction for a different application. These biases are due to incorrect distributional assumptions imposed in these parametric approaches, and can lead store managers to mistakenly overemphasize the importance of shoppers closer to the store and over-weight them in managerial decisions. Overall this analysis demonstrates how ignoring selective sampling and mis-specifying population distributions can cause significant bias and sub-optimal managerial decisions, and the usefulness of SOR for guarding against these biases.

## 5. Discussion

In this paper, we introduced a unified framework based on SOR that extends the efficient approach of Donkers *et al.* (2003) for endogenous stratified sampling of binary choice outcomes to broader types of outcomes (polytomous, continuous and count) and selective sampling schemes, including endogenously stratified sampling, on-site sampling, truncated samples, and sample selection problems. Our study using synthetic and real-life examples shows endogenous selective sampling can substantially improve sampling efficiency and reduce the cost associated with marketing research, and sometimes is the only feasible means of addressing the question of interest given the nature of available datasets and resources. The results also show that selection on both the outcome and a strata on unevenly distributed explanatory variables can further substantially increase estimation efficiency when the sampling weights are known and included as offsets to a program written for fitting the SOR to random samples. Other design factors ($\alpha$ and $p$ determining the extremeness and amount of informative data points) in endogenously stratified sampling are shown to have systematic and substantial effects on the sampling efficiency (Fig 2). Depending on choices made on factors affecting sampling efficiency, the gain using selective sampling can be more than 10 times of that using random sampling, which can translate into large savings in financial and time costs.

Despite these substantial benefits, our evaluation shows that the selection bias associated with selective sampling also increases simultaneously, demanding bias correction to realize the benefits of selective sampling. Two thorny issues in correcting for the selection bias are the specifications of outcome distributions and selection mechanisms. In practice, researchers rarely have complete knowledge to specify them confidently. Selective sampling can hide or artificially create complex distributional features of the outcomes. There may only be partial and limited knowledge on selection rules. Unfortunately, as shown in the analysis of synthetic and real-life data, mis-specifications of these functional forms can yield large bias, sometimes even larger than not correcting for selectivity, as well as sub-optimal marketing decisions. Notably, the marginal effect of the price variable is substantially underestimated and not selected as a significant variable in Example 2, and the importance of consumers closer to the store is mistakenly over-estimated in Example 3 because of mis-specifications of outcome distributions in the parametric sample adjustment methods.

Thus, the proposed SOR approach's capability of "Killing Two Birds with One Stone" is very desirable, in that it requires neither modeling selection mechanisms nor relying on parametric distributional assumptions about marketing response variables, thereby guarding against the mis-specification bias associated with these parametric sample adjustment methods. Specifically, when the population distribution follows GLMs, nonparametric modeling of the baseline distribution in SOR leads to no or little efficiency loss and performs almost as well as existing parametric sampling adjustment methods in estimating the population association parameters of interests. When the distributional assumptions are violated in the parametric sample adjustment methods, using them to adjust for selective sampling can yield significant bias, whereas the SOR approach remains valid. Thus SOR should be preferred to sampling adjustment methods assuming GLMs for the population distribution. When the population follows a parametric distribution outside GLMs (e.g., a consumer demand model using a truncated negative binomial regression including both observed and unobserved consumer heterogeneity), it is possible to use a correctly-specified parametric sample adjustment method to gain higher efficiency than SOR. In this case, SOR can be useful for testing the adequacy of distributional assumptions imposed in the parametric method, by nesting the parametric approach as a special case. Should the parametric approach be rejected, SOR provides a better alternative for analyzing selected samples. As shown in Example 3, model selection statistics such as AIC can be used to select best-performing sample adjustment method(s).

Requiring no modeling of selection mechanisms is another advantage of SOR. When sampling probabilities are unknown, the outcome error distributional form in the selective sample depends on the unknown sampling weights and may depart from standard distributions in any unknown and arbitrary way. Unlike random sampling, these distributional departures can be difficult to check in endogenously selected samples, as selection can artificially create or hide the departures. The fact that a *priori* knowledge of sampling weights or error distributional forms is not required and the ability to automatically generate a suitable outcome distribution, an important feature of the SOR approach, can thus significantly increase the robustness and applicability of selective sampling.

The property of the OR association parameters in SOR being invariant to and free of bias from response-dependent sampling holds much more broadly than the logit models for binary outcomes considered in Donkers *et al.* (2003) and has substantially wider applicability. These OR parameters are regression coefficients or standardized ones in the presence of the dispersion parameter in the popular GLMs. When population follows these standard distributions, SOR directly estimates these familiar population parameters even if response-dependent sampling alters the distributions in selected samples away from these standard distributions in an unknown fashion. When the population does not follow GLMs or any parametric models and no standard model describes the population adequately, the OR parameters in SOR can still be used as population association parameters for identifying drivers of the outcomes. Simple connections between OR parameters and marginal mean effects and other quantifies of interest are derived and can aid the interpretation of SOR results in this more general case. Our empirical analysis of real-life data demonstrates that in the population or a random sample the inferences based on SOR are identical or correspond closely to the inference using other distribution-free procedures, such as OLS or parametric regressions with robust standard errors. In comparison with these other distribution-free procedures, advantages of the SOR approach include its ability to properly recover the entire distribution of interest, to provide a full data distribution to use in Bayes' theorem for handling selective sampling, and to relax the assumption of linearity in the mean function of the OLS regression. In addition to the unique property of the OR parameters being invariant to response-dependent sampling, these desirable features make SOR well suited to handle selective sampling.

This study points toward some avenues for future research. It would be useful to extend the approach to selective sampling with panel data to reduce the significantly more costs associated with conducting panel studies. In this work we assume all the regressors in $X$ are exogenous. Future work can extend the method to account for both selective sampling and regressor endogeneity. Throughout the paper we assume that all the data in the endogenously selected sample are fully observed. In practice, there could be missing values in $Y$ or $X$ (i.e., item missingness), *given* the units are already included in the selected sample. Oftentimes it can be plausible to assume the mechanism of such item missingness to be ignorable (Little and Rubin, 2020). Future research can develop methods that combine missing data methods assuming ignorability with the sampling-adjustment method proposed here for efficient use of such data.

**Acknowledgment**

We would like to thank the Editor, the Associate Editor and the reviewers for many constructive comments that substantially improved the manuscript. We would like to acknowledge the support from Social Sciences and Humanities Research Council of Canada [grant 435-2018-0519], Natural Sciences and Engineering Research Council of Canada [grant RGPIN-2018-04313] and US National Institute of Health [grants R01CA178061].

## References

Arora, N. and Huber, J. (2001). Improving parameter estimates and model prediction by aggregate customization in choice experiments. *Journal of Consumer Research*, **28**, 273–83.

Breslow, N. and Day, N. (1980). *Statistical Methods in Cancer Research*. International Agency for Research on Cancer, Lyon, France.

Cameron, A. and Trivedi, P. (2009). *Microeconometrics using stata*. Stata Press, College Station.

Chen, H. (2007). A semiparametric odds ratio model for measuring association. *Biometrics*, **63**, 413–421.

Chen, H., Rader, D., and Li, M. (2015). Likelihood inferences on semiparametric odds ratio model. *Journal of the American Statistical Association*, **110**, 1125–1135.

Clow, K. and James, K. (2014). *Essentials of marketing research: Putting research into practice*. SAGE Publications, Thousands Oaks, CA.

Cosslett, S. (1993). Estimation from endogenously stratified samples. In C. R. G.S. Maddala and H. Vinod, editors, *Handbook of Statistics*. Elsevier Science Publishers, Amsterdam.

Cosslett, S. (2013). Efficient semiparametric estimation for endogenously stratified regression via smoothed likelihood. *Journal of Econometrics*, **177**, 116–129.

Donkers, B., Franses, P., and Verhoef, P. (2003). Selective sampling for binary choice models. *Journal of Marketing Research*, **XL**, 492–97.

Englin, J. and Shonkwiler, J. (1995). Estimating social welfare using count data models: An application to long-run recreation demand under conditions of endogenous stratification and truncation. *Review of Economics & Statistics*, **77(1)**, 104–112.

Feinberg, F., Kinnear, T., and Taylor, J. (2012). *Modern Marketing Research: Concepts, Methods and Cases*. Thomson Academic Publishing, Mason, OH.

Feinberg, F., Salisbury, L., and Ying, Y. (2016). When random assignment is not enough: Accounting for item selectivity in experimental research. *Marketing Science*, **35**, 976–994.

Feit, E. and Bradlow, E. (2018). Fusion modeling. In K. M. Homberg C. and V. A., editors, *Handbook of Marketing Research*. Springer, New York.

Feldt, L. (1961). The use of extreme groups to test for the presence of a relationship. *Psychometrika*, **26**, 307–316.

Hausman, J. and Wise, D. (1981). Stratification on endogenous variables and estimation: the gary income maintenance experiment. In C. Manski and D. E. McFadden, editors, *Structural Analysis of Discrete Data with Econometric Applications*, pages 365–391. MIT Press, Cambridge, MA.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, **47**, 153–61.

Kamakura, W., Mela, C., Ansari, A., Bodapati, A., Fader, P., Iyengar, R., Naik, P., Neslin, S., Sun, B., Verhoef, P., Wedel, M., and Wilcox, R. (2005). Choice models and customer relationship management. *Marketing Letters*, **16**, 279–291.

Little, R. and Rubin, D. (2020). *Statistical Analysis with Missing Values, 3rd Ed.* John Wiley.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, New York.

Puhani, P. (2000). The heckman correction for sample selection and its critique. *Journal of Econometric Surveys*, **14**, 53–68.

Qian, Y. and Xie, H. (2011). No customer left behind: A distribution-free bayesian approach to accounting for missing xs in marketing models. *Marketing Science*, **30**, 717–736.

Qian, Y. and Xie, H. (2014). Which brand purchasers are lost to counterfeiters? an application of new data fusion approaches. *Marketing Science*, **33**, 437–448.

Qian, Y. and Xie, H. (2015). Driving more effective data-driven innovations: Enhancing the utility of secure catabases. *Management Science*, **61**, 520–541.

Scott, A. and Wild, C. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, **84(1**, 57–71.

Shaw, D. (1988). On-site samples' regression: Problems of non-negative integers, truncation and endogenous stratification. *Journal of Econometrics*, **37**, 211–223.

Shonkwiler, J. and Englin, J. (2009). Approximating the distribution of recreational visits from on-site survey data. *Journal of Environmental Management*, **90**, 1850–1853.

Tian, L. and Feinberg, F. (2020). Optimizing price menus for duration discounts: A subscription selectivity field experiment. *Marketing Science*, **39**, 1033–1201.

Wachtel, S. and Otter, T. (2013). Successive sample selection and its relevance for management decisions. *Marketing Science*, **32**, 170–185.

Xie, H. and Qian, Y. (2012). Measuring the impact of nonignorability in panel data with non-monotone nonresponse. *Journal of Applied Econometrics*, **27**, 129–159.

Ying, Y., Feinberg, F., and Wedel, M. (2006). Leveraging missing ratings to improve online recommendation systems. *Journal of Marketing Research*, **43**, 355–365.

Yuan, C., Hedeker, D., Mermelstein, R., and Xie, H. (2020). A tractable method to account for high-dimensional nonignorable missing data in intensive longitudinal data. *Statistics in Medicine*, **39**, 2589–2605.

**Table 1**

*Comparison of Estimation Methods for Endogenously Selective Sampling.*

| | DFV (2003) | HW (1981) | Cosslett (1993) | Cosslet (2013) | Shaw (1988), ES (1995) | SOR |
|---|---|---|---|---|---|---|
| Outcome | | | | | | |
| Type | Binary | Continuous | Discrete & Continuous | Continuous | Discrete & Continuous | Discrete & Continuous |
| Model | Binary Regression | Normal Regression | Parametric Regression | Distribution-free Linear Regression | Normal, Poisson, Negative Binomial Regression | Distribution-free SOR regression |
| Sampling Scheme | Stratified Sampling (with known SW) | Stratified Sampling | Stratified Sampling (with known SW) | Stratified Sampling with two strata | On-site Sampling (with known SW) | Stratified Sampling,<br>On-site sampling,<br>general selective sampling |
| Factors Affecting Selection | $Y, X$ | $Y$ | $Y, X$ | $Y$<br>$(Y > c$ .vs. $Y < c)$ | $Y$ | $Y, X, \theta$ |
| Estimation | Analyzed as random samples with known SWs as offsets | Separate estimation algorithms from those for random samples are needed and sometimes can be difficult to implement/generalize | | | | Analyzed as random samples (SWs, if known, are used as offsets) |

Stratified Sampling: selection depends on $Y$ and $X$ via a finite number of known strata defined on $Y$ and $X$; On-site sampling: sampling weights proportional to $Y$ values; General selective sampling: sampling weights depend directly on $Y$ or a combination of $Y, X$ and $\theta$ with known or unknown sampling weights. SW: Sampling Weights. GLMs: Generalized Linear Models, including normal regression, binomial regression, Poisson regression and Gamma regression as special cases. SOR: Semiparametric Odds-Ratio model, including GLMs as special cases. DFV (2003): Donkers, Franses, and Verhoef (2003); HW (1981): Hausman and Wise (1981); ES (1995): Englin and Shonkwiler (1995).

**Table 2**

*Comparison with Missing Data Methods for Endogenously Selective Sampling.*

| | Qian and Xie (2011) | Heckit | SOR |
|---|---|---|---|
| Outcome Model | Parametric | Parametric | Distribution-Free SOR |
| Covariate Modeling Required | Yes | No | No |
| Selection Modeling Required | No[a] | Yes (linear probit model) | No |
| Factors Affecting Selection | $Y, X$ | $Y, X, \theta$ | $Y, X, \theta$ |

[a]: The method of Qian and Xie (2011) was designed for addressing missing-covariate issues in a random sample and considered the case of ignorable missingness for which no selection modeling is needed. The method can be extended to permit nonignorable missingness which will require modeling missing data selection mechanisms.

Table 3: Comparison of Methods for Dealing with Missing $Y$ Values Due to Selective Sampling.

| | $X_1$ (true coefficient value=0.1) | | | $X_2$(true coefficient value=0.1) | | |
|---|---|---|---|---|---|---|
| Selection Rules | OLS | Heckit-FIML | SOR | OLS | Heckit-FIML | SOR |
| ES-Y or ES-Y$\theta$ | | | | | | |
| 1. Linear Probit Selection | 0.067 (0.029) | 0.092 (0.743) | 0.101 (0.043) | 0.069 (0.012) | 0.090 (0.151) | 0.101 (0.019) |
| 2. Quadratic Probit Selection | 0.150 (0.053) | -2.104 (16.603) | 0.101 (0.037) | 0.152 (0.019) | -0.110 (1.229) | 0.099 (0.013) |
| 3. Extreme Values Selected | 0.304 (0.103) | 66.325 (487.235) | 0.099 (0.033) | 0.320 (0.041) | -1.017 (26.40) | 0.100 (0.014) |
| 4. Middle Value Selected | 0.014 (0.015) | -17.80 (295.87) | 0.097 (0.101) | 0.014 (0.006) | -0.65 (3.254) | 0.100 (0.040) |
| ES-YX or ES-YX$\theta$ | | | | | | |
| 5. Linear Probit Selection | 0.079 (0.036) | -7.02 (115.91) | 0.099 (0.048) | -0.039 (0.013) | -0.307 (1.100) | 0.101 (0.030) |
| 6. Quadratic Probit Selection | 0.095 (0.040) | 0.090 (0.042) | 0.100 (0.044) | -0.197(0.017) | -0.150 (0.165) | 0.099 (0.024) |
| 7. Extreme Values Selected | 0.252 (0.081) | 0.230 (0.232) | 0.100 (0.032) | 0.160 (0.035) | -0.034 (0.620) | 0.101 (0.016) |
| 8. Middle Value Selected | 0.016 (0.016) | 0.011 (0.030) | 0.098 (0.094) | 0.069 (0.013) | 0.034 (0.079) | 0.100 (0.057) |

1. $\Phi(y)$; 2. $\Phi(-1+10\beta_1 y+10\beta_2 y^2)$; 3. and 4.: 0/1 function of percentiles of $Y$. 5. $\Phi(10\beta_1 y-20\beta_2 y*I(x_2>0))$; 6: $\Phi(-1+y+y^2$ $+2*I(x_2>0)-2*y*I(x_2>0)-2*y^2*I(x_2>0))$; 7. and 8.: 0/1 functions of $Y$ percentiles and $I(X_2>0)$. Also see Fig 3.

[hp]

**Table 4**

*Demand Function Estimation Using the Truncated Retail Store Shopper Visits Data.*

| | Poisson | | Truncated Poisson (TP) | | Negative Binomial | | Truncated Negative Binomial (TNB) | | SOR$^{LLL}$ | | SOR$^{LLS}$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Est. | SE$_R$ (SE$_D$)<br>t (t$_D$) | Est. | SE$_R$ (SE$_D$)<br>t (t$_D$) | Est. | SE$_R$ (SE$_D$)<br>t (t$_D$) | Est. | SE$_R$ (SE$_D$)<br>t (t$_D$) | Est. | S.E.<br>t | Est. | S.E.<br>t |
| `Income` | 0.028 | 0.045 (0.013) | 0.043 | 0.070 (0.017) | 0.026 | 0.040 (0.018) | 0.049 | 0.075 (0.040) | 0.013 | 0.017 | 0.015 | 0.022 |
| | | 0.63 (**2.12**) | | 0.62 (**2.57**) | | 0.64 (1.44) | | 0.65 (1.21) | | 0.74 | | 0.70 |
| `DTS` | -0.283 | 0.057 (0.020) | -0.515 | 0.101 (0.028) | -0.270 | 0.048 (0.026) | -0.609 | 0.096 (0.055) | -0.318 | 0.047 | -0.368 | 0.156 |
| | | **-4.96** (**-14.13**) | | **-5.09** (**-18.36**) | | **-5.59** (**-10.56**) | | **-6.36** (**-11.12**) | | **-6.77** | | **-2.36** |
| `DTS2` | 0.078 | 0.021 (0.009) | 0.138 | 0.036 (0.012) | 0.073 | 0.018 (0.011) | 0.152 | 0.035 (0.025) | 0.073 | 0.021 | 0.086 | 0.042 |
| | | **3.75** (**9.14**) | | **3.87** (**11.94**) | | **4.07** (**6.49**) | | **4.30** (**6.19**) | | **3.50** | | **2.05** |
| `Age` | -0.003 | 0.028 (0.014) | -0.006 | 0.048 (0.019) | -0.003 | 0.026 (0.019) | -0.002 | 0.058 (0.043) | -0.002 | 0.023 | -0.003 | 0.028 |
| | | -0.11 (-0.21) | | -0.13 (-0.34) | | -0.10 (-0.14) | | -0.04 (-0.05) | | -0.10 | | -0.10 |
| `Married` | 0.057 | 0.021 (0.014) | 0.107 | 0.040 (0.020) | 0.055 | 0.021 (0.018) | 0.133 | 0.055 (0.040) | 0.070 | 0.033 | 0.084 | 0.042 |
| | | **2.74** (**4.05**) | | **2.69** (**5.43**) | | **2.67** (**3.04**) | | **2.40** (**3.35**) | | **2.15** | | **2.01** |
| `Kids` | -0.033 | 0.029 (0.014) | -0.057 | 0.049 (0.019) | -0.034 | 0.028 (0.018) | -0.095 | 0.064 (0.040) | -0.023 | 0.024 | -0.029 | 0.034 |
| | | -1.15 (**-2.36**) | | -1.15 (**-3.06**) | | -1.24 (-1.86) | | -1.47 (**-2.34**) | | -0.98 | | -0.85 |
| `Const.` | 0.583 | 0.026 (0.016) | 0.224 | 0.060 (0.024) | 0.590 | 0.026 (0.021) | -21.313 | 0.045 (41.66) | 0.557 | 0.025 | 0.557 | 0.035 |
| | | 22.30 (35.54) | | 3.72 (9.27) | | 22.82 (28.15) | | -475.69 (-0.51) | | 22.28 | | 15.91 |
| $\ln \alpha$ | | | | | -1.021 | 0.144 (0.048) | 21.968 | 0.035 (41.66) | | | -2.071 | 3.085 |
| | | | | | | -7.09 (-21.27) | | 627.7 (0.53) | | | | -0.67 |
| -logLik | 6328.9 | | 5659.8 | | 5546.3 | | 3819.9 | | **3667.5** | | **3667.5** | |
| AIC | 12672 | | 11334 | | 11109 | | 7656 | | **7399** | | 7401 | |

Note: `Income`: Household income; `DTS`: Distance to Store; `DTS2`= `DTS`∗`DTS`; `Age`: Age of the customer; `Married`: Marriage status of the customer; `Kids`: Number of kids less than ≤ 18 years old at home. SE$_R$: Robust (empirical) sandwich standard error estimate; SE$_D$: Default model-based standard error estimate. t: Wald t statistics using robust standard error or using distribution-free SOR; t$_D$: Wald t statistics using default standard error based on parametric distributions. SOR$^{LLL}$: SOR model using log-bilinear OR function; SOR$^{LLS}$: SOR model using log-linear-softplus OR function. Bolded t values means statistical significance at the 0.05 level. Greatest likelihood and smallest AIC are bolded.

Table 5: Demand Function Estimation Using Random Samples and On-site Samples of the Shopper Visit Data.

| | Population | Random Sampling | | | | | On-site Sampling | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Trunc. Pois | TP | TP-A | TNB | TNB-A | SOR$^{LLL}$ | TP | TP-A | TNB | TNB-A | SOR$^{LLL}$ |
| `Income` | 0.043 | 0.044 | 0.053 | 0.043 | 0.053 | 0.044 | 0.035 | 0.043 | 0.036 | 0.044 | 0.043 |
| | | (0.018) | (0.022) | (0.018) | (0.022) | (0.018) | (0.012) | (0.015) | (0.012) | (0.038) | (0.015) |
| `DTS` | -0.515 | -0.515 | -0.606 | -0.516 | -0.610 | -0.516 | -0.425 | -0.516 | -0.426 | -0.516 | -0.517 |
| | | (0.028) | (0.033) | (0.028) | (0.033) | (0.030) | (0.020) | (0.024) | (0.020) | (0.025) | (0.026) |
| `DTS2` | 0.138 | 0.138 | 0.163 | 0.137 | 0.163 | 0.138 | 0.113 | 0.138 | 0.113 | 0.137 | 0.138 |
| | | (0.011) | (0.013) | (0.011) | (0.013) | (0.012) | (0.009) | (0.010) | (0.009) | (0.010) | (0.011) |
| `Age` | -0.006 | -0.006 | -0.007 | -0.006 | -0.007 | -0.006 | -0.005 | -0.006 | -0.006 | -0.007 | -0.006 |
| | | (0.018) | (0.021) | (0.019) | (0.022) | (0.019) | (0.013) | (0.016) | (0.014) | (0.016) | (0.016) |
| `Married` | 0.107 | 0.106 | 0.127 | 0.107 | 0.126 | 0.107 | 0.089 | 0.108 | 0.089 | 0.107 | 0.108 |
| | | (0.019) | (0.022) | (0.019) | (0.022) | (0.019) | (0.014) | (0.017) | (0.014) | (0.017) | (0.017) |
| `Kids` | -0.057 | -0.057 | -0.068 | -0.058 | -0.068 | -0.057 | -0.047 | -0.057 | -0.047 | -0.057 | -0.057 |
| | | (0.018) | (0.021) | (0.018) | (0.021) | (0.019) | (0.013) | (0.016) | (0.013) | (0.016) | (0.016) |
| `Const.` | 0.224 | 0.222 | -0.281 | 0.222 | -0.281 | 0.222 | 0.650 | 0.223 | 0.650 | 0.224 | 0.224 |
| | | (0.024) | (0.028) | (0.024) | (0.028) | (0.024) | (0.017) | (0.020) | (0.017) | (0.020) | (0.020) |
| -logLik | | 3960.1 | 3983.2 | 3960.1 | 3983.5 | **3955.9** | 4710.2 | 4677.1 | 4710.2 | 4677.1 | **4672.4** |
| AIC | | **7934** | 7980 | 7936 | 7983 | 7939 | 9432 | **9366** | 9434 | 9368 | 9374 |
| $D \times 1000$ | | 0.285 | − − − | 0.287 | − − − | 0.297 | − − − | 0.214 | − − − | 0.214 | 0.223 |
| | Trunc. NB | TP | TP-A | TNB | TNB-A | SOR$^{LLS}$ | TP | TP-A | TNB | TNB-A | SOR$^{LLS}$ |
| `Income` | 0.049 | 0.033 | 0.040 | 0.049 | 0.040 | 0.049 | 0.040 | 0.049 | 0.060 | 0.049 | 0.049 |
| | | (0.030) | (0.036) | (0.040) | (0.035) | (0.048) | (0.020) | (0.025) | (0.029) | (0.024) | (0.028) |
| `DTS` | -0.609 | -0.446 | -0.526 | -0.603 | -0.510 | -0.605 | -0.500 | -0.607 | -0.731 | -0.606 | -0.605 |
| | | (0.045) | (0.052) | (0.056) | (0.049) | (0.223) | (0.029) | (0.035) | (0.040) | (0.034) | (0.134) |
| `DTS2` | 0.152 | 0.108 | 0.128 | 0.155 | 0.122 | 0.150 | 0.120 | 0.150 | 0.182 | 0.150 | 0.150 |
| | | (0.019) | (0.023) | (0.027) | (0.022) | (0.067) | (0.013) | (0.016) | (0.019) | (0.016) | (0.037) |
| `Age` | -0.002 | -0.005 | -0.005 | -0.002 | -0.001 | -0.002 | -0.001 | -0.001 | -0.003 | -0.002 | -0.002 |
| | | (0.031) | (0.037) | (0.044) | (0.033) | (0.042) | (0.020) | (0.025) | (0.029) | (0.024) | (0.024) |
| `Married` | 0.133 | 0095 | 0.113 | 0.130 | 0.114 | 0.136 | 0.107 | 0.134 | 0.160 | 0.133 | 0.132 |
| | | (0.030) | (0.036) | (0.041) | (0.033) | (0.058) | (0.020) | (0.024) | (0.028) | (0.023) | (0.036) |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| `Kids` | -0.095 | -0.066 | -0.079 | -0.098 | -0.078 | -0.093 | -0.075 | -0.092 | -0.112 | -0.093 | -0.092 |
| | | (0.030) | (0.036) | (0.043) | (0.035) | (0.056) | (0.020) | (0.025) | (0.029) | (0.024) | (0.033) |
| `Const.` | -21.313 | 0.261 | -0.237 | -21.312 | -21.756 | -21.312 | 1.004 | 0.654 | -20.790 | -21.313 | -21.538 |
| | | (0.032) | (0.037) | (0.021) | (0.018) | (0.024) | (0.024) | (0.029) | (0.019) | (0.014) | (0.015) |
| $\ln\alpha$ | 21.968 | – – – | – – – | 21.971 | 21.524 | 21.733 | – – – | – – – | 22.490 | 21.967 | 22.008 |
| | | – – – | – – – | (0.023) | (0.018) | (0.781) | – – – | – – – | (0.019) | (0.014) | (0.373) |
| -logLik | | 4675.1 | 4976.3 | 4053.2 | 4119.0 | **4041.1** | 7322.4 | 7821.5 | 6414.1 | 6234.0 | **6215.8** |
| AIC | | 9364 | 9967 | **8122** | 8254 | 8130 | 14659 | 15657 | 12844 | **12484** | 12497 |
| $D \times 1000$ | | – – – | – – – | 1.266 | – – – | 2.231 | – – – | – – – | – – – | 0.460 | 0.757 |
| SOR | | TP | TP-A | TNB | TNB-A | SOR$^{LLL}$ | TP | TP-A | TNB | TNB-A | SOR$^{LLL}$ |
| `Income` | 0.013 | 0.046 | 0.055 | 0.052 | 0.048 | 0.015 | 0.102 | 0.115 | 0.098 | 0.089 | 0.013 |
| | | (0.058) | (0.070) | (0.066) | (0.060) | (0.022) | (0.038) | (0.043) | (0.040) | (0.033) | (0.005) |
| `DTS` | -0.318 | -0.519 | -0.612 | -0.611 | -0.554 | -0.322 | -1.091 | -1.270 | -1.041 | -0.920 | -0.319 |
| | | (0.100) | (0.118) | (0.093) | (0.088) | (0.049) | (0.079) | (0.086) | (0.059) | (0.052) | (0.022) |
| `DTS2` | 0.073 | 0.136 | 0.161 | 0.159 | 0.144 | 0.070 | 0.305 | 0.353 | 0.293 | 0.260 | 0.072 |
| | | (0.036) | (0.043) | (0.039) | (0.035) | (0.022) | (0.024) | (0.026) | (0.030) | (0.022) | (0.007) |
| `Age` | -0.002 | -0.006 | -0.007 | -0.006 | -0.007 | -0.003 | -0.012 | -0.015 | -0.005 | -0.005 | -0.003 |
| | | (0.053) | (0.063) | (0.057) | (0.052) | (0.025) | (0.038) | (0.043) | (0.040) | (0.033) | (0.006) |
| `Married` | 0.070 | 0.109 | 0.129 | 0.127 | 0.117 | 0.074 | 0.217 | 0.253 | 0.198 | 0.184 | 0.072 |
| | | (0.043) | (0.050) | (0.055) | (0.050) | (0.033) | (0.028) | (0.032) | (0.040) | (0.032) | (0.015) |
| `Kids` | -0.023 | -0.057 | -0.068 | -0.060 | -0.057 | -0.025 | -0.137 | -0.157 | -0.110 | -0.104 | -0.023 |
| | | (0.051) | (0.061) | (0.060) | (0.054) | (0.026) | (0.036) | (0.042) | (0.039) | (0.032) | (0.007) |
| `Const.` | 0.557 | 0.218 | -0.286 | -21.323 | -21.771 | 0.556 | 1.129 | 0.823 | -20.539 | -21.122 | 0.557 |
| | | (0.054) | (0.062) | (0.030) | (0.026) | (0.024) | (0.044) | (0.051) | (0.025) | (0.018) | (0.014) |
| -logLik | | 5635.9 | 6167.2 | 3810.1 | 4103.8 | **3647.1** | 16425.7 | 17641.6 | 7264.0 | 7585.4 | **7195.3** |
| AIC | | 11286 | 12348 | 7636 | 8224 | **7353** | 32865 | 32597 | 14544 | 15187 | **14457** |
| $D * 1000$ | | – – – | – – – | – – – | – – – | 0.663 | – – – | – – – | – – – | – – – | 0.082 |

Note: the table reports the averages of estimates and standard errors in the parenthesis over all 500 samples drawn under random sampling and on-site sampling. TP (TNB): truncated Poisson (negative binomial) regression; TP-A (TNB-A): truncated Poisson (negative binomial) regression adjusting for on-site sampling; SOR$^{LLL}$ and SOR$^{LLS}$: see note under Table 4; D*1000: D-error measure multiplied by 1000. Greatest likelihood and smallest AIC are bolded.

**Table 6**

*Expected Shopping Frequencies of Shoppers and Proportions of Returning Shoppers by Shopping Distance.*

| | Truncated Sample | | | On-site Sample | | | | |
|---|---|---|---|---|---|---|---|---|
| | TP | TNB | SOR | TP | TP-A | TNB | TNB-A | SOR |
| $E(Y_{dist95})$ | 1.43 | 1.46 | 1.52 | 1.70 | 1.41 | 2.36 | 1.57 | 1.52 |
| $E(Y_{dist50})$ | 1.95 | 1.98 | 1.89 | 4.83 | 3.88 | 5.10 | 2.56 | 1.89 |
| $E(Y_{dist5})$ | 2.65 | 2.53 | 2.58 | 12.68 | 11.74 | 9.71 | 4.08 | 2.58 |
| $Pr(Y_{dist95} > 1)$ | 0.33 | 0.29 | 0.26 | 0.48 | 0.32 | 0.49 | 0.32 | 0.26 |
| $Pr(Y_{dist50} > 1)$ | 0.58 | 0.43 | 0.32 | 0.96 | 0.91 | 0.65 | 0.51 | 0.32 |
| $Pr(Y_{dist5} > 1)$ | 0.76 | 0.51 | 0.38 | 1.00 | 1.00 | 0.73 | 0.62 | 0.38 |

Note: For shoppers $Y$ takes integer values of larger than or equal to 1.

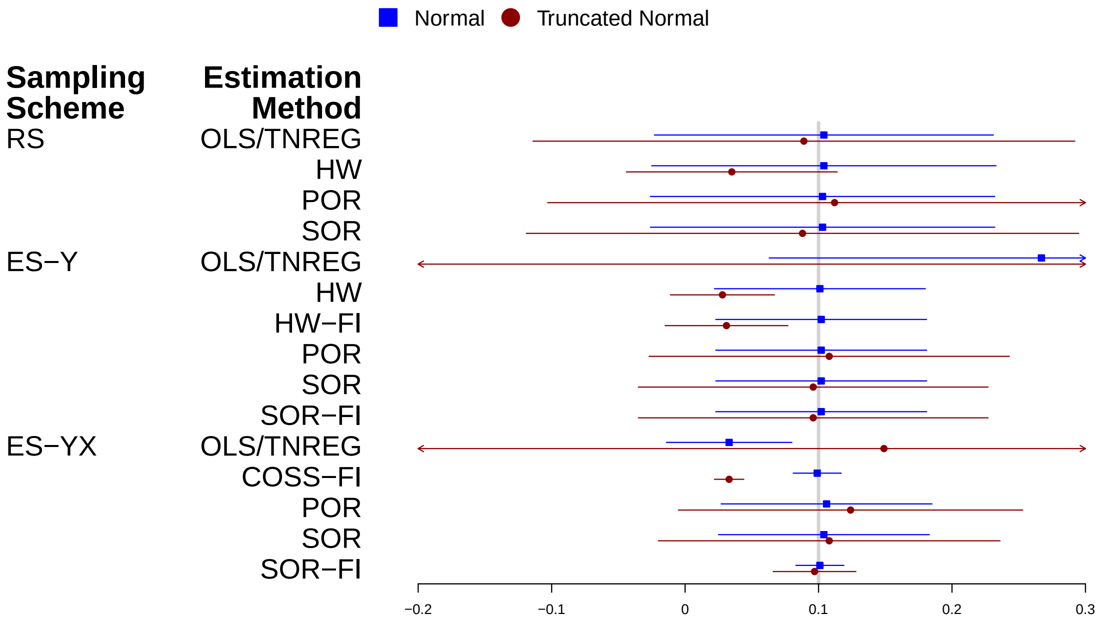

**Figure 1.** Mean ± SD of regression coefficient estimates of $X_1$ over 500 repeated samples. Arrows indicate intervals beyond the limits of x-axis. The mean of estimates for TNREG in ES-Y is 1.136.

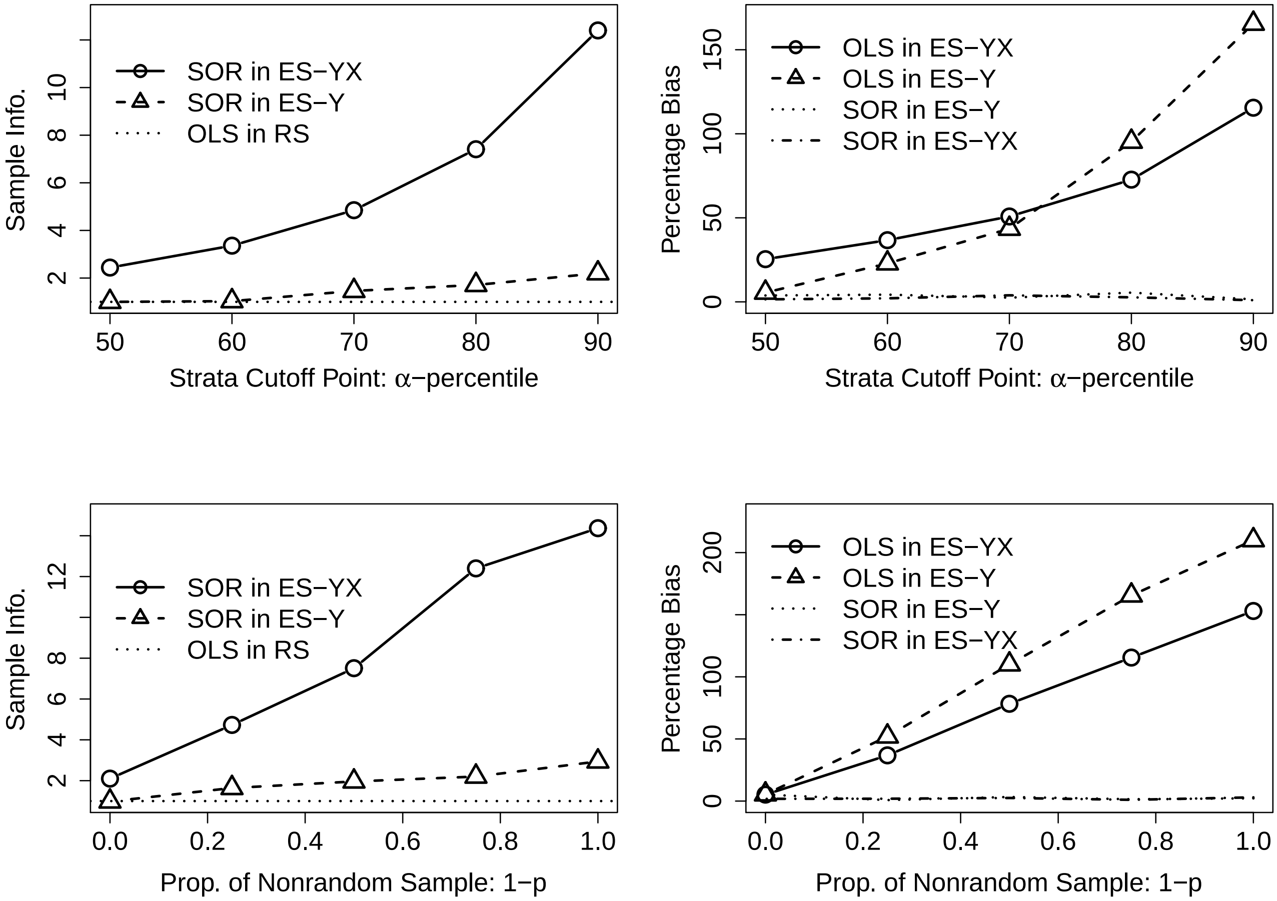

**Figure 2.** Factors affecting efficiency and selection bias of endogenous sampling schemes.

[h]

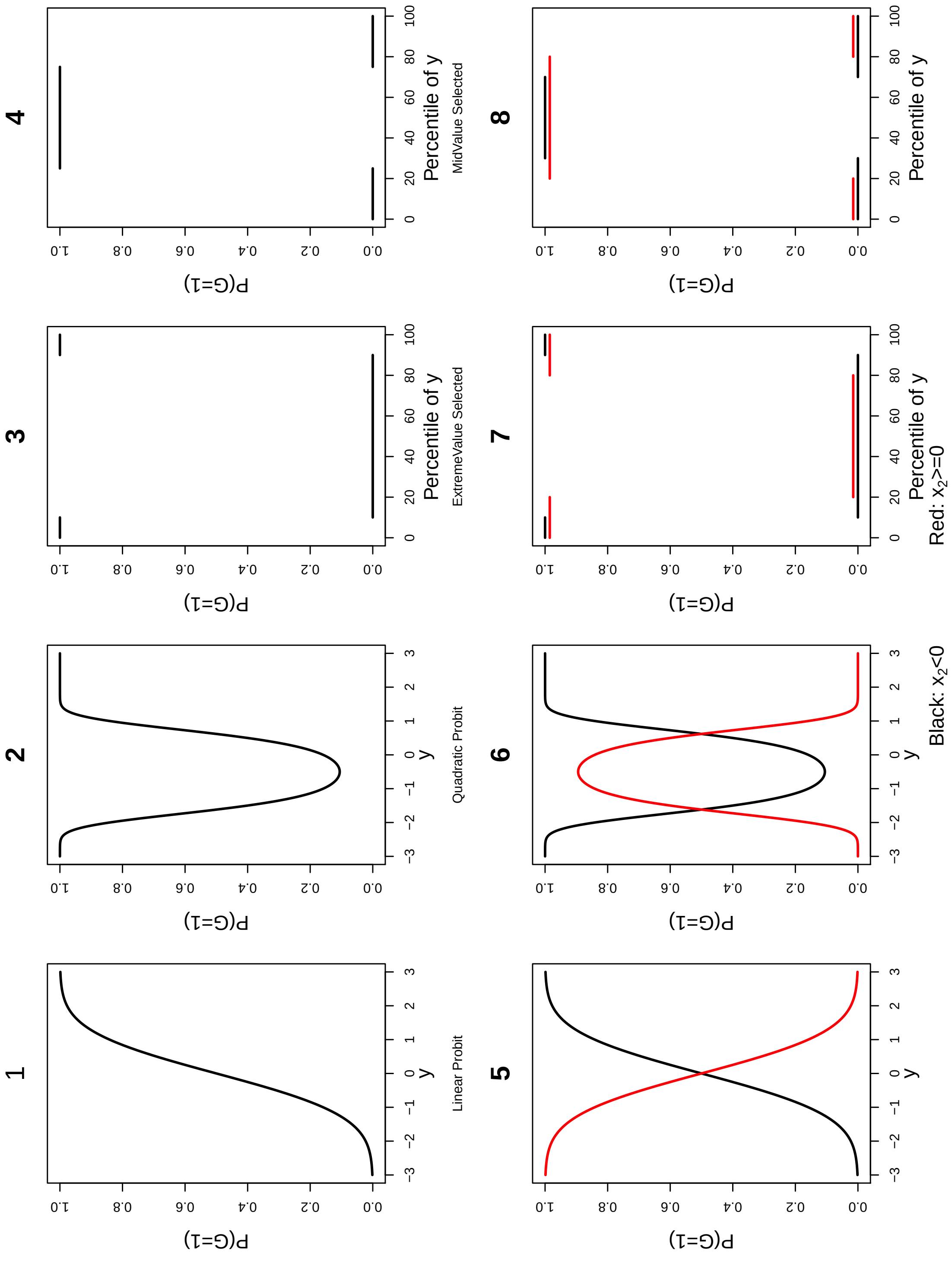

**Figure 3.** Selection Rules for $Y$ Being Missing.

[h]

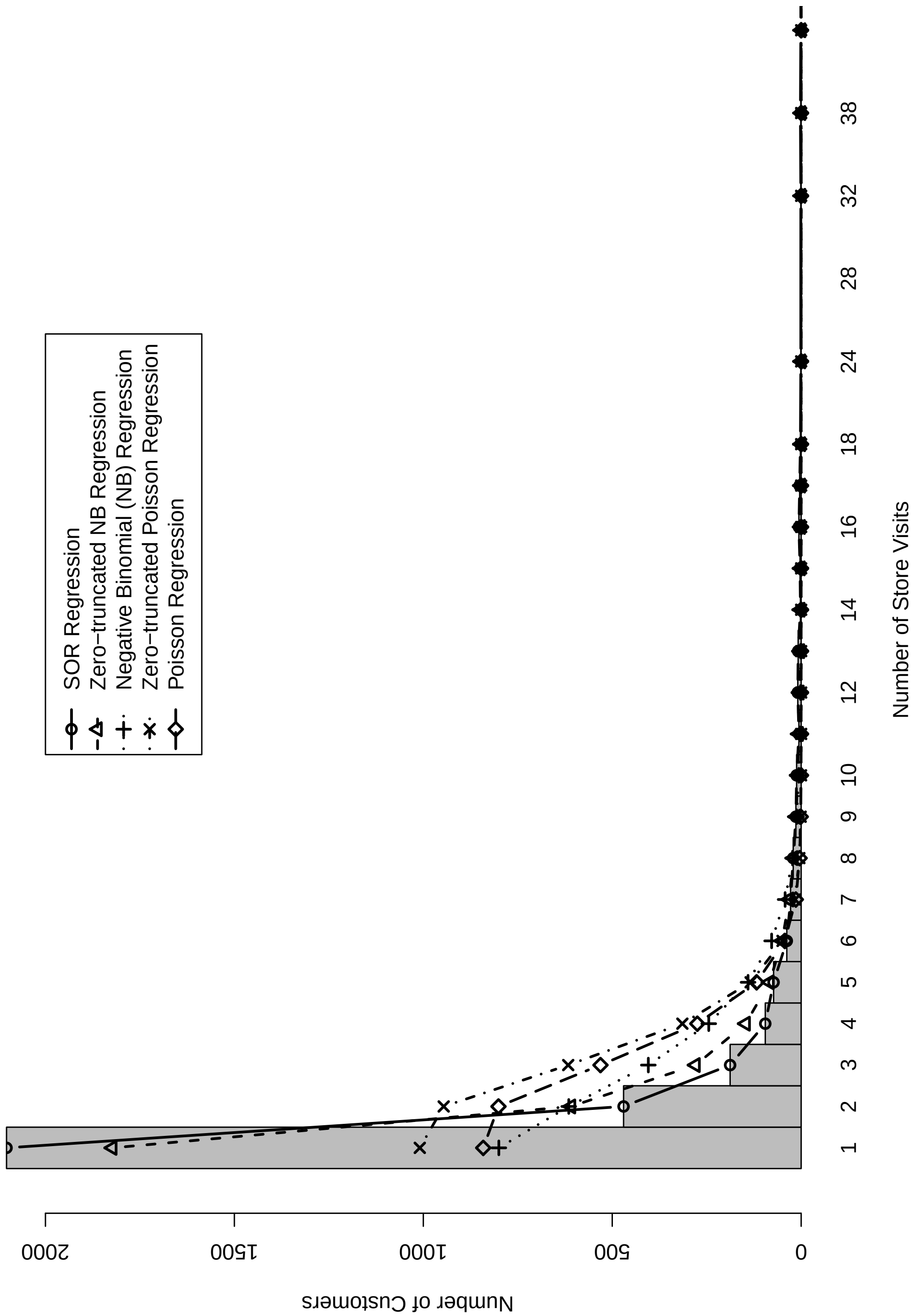

**Figure 4.** Comparison of Model Fittings. The bars represents the observed numbers of customers. The five curves represent the predicted numbers of customers.