Web Appendix to

"Simplifying Bias Correction for Selective Sampling: A Unified Distribution-Free Approach to Handling Endogenously Selected Samples"

Online Appendix I: Estimation Algorithm

I.1: MLE of the SOR Approach to Selective Sampling.

One merit of the SOR approach to selective sampling is that it analyzes endogenously selected samples as if it were a random sample. For a random sample of n independent units, $(x_i, y_i), i = 1, \cdots, n$, the likelihood $L(\gamma, \lambda)$ is given in Eqn (4) in the main paper with the resulting log likelihood as

$$\begin{aligned} LL(\gamma,\lambda) &= \ln L(\gamma,\lambda) = \sum_{i=1}^{n} LL_i(\gamma,\lambda) \\ &= \sum_{i=1}^{n} \ln \left\{ \frac{\eta_\gamma\{y_i; x_{i1},\cdots,x_{iK}\} f_\lambda(y_i|x_{10},\cdots,x_{K0})}{\sum_{l=1}^{L} \eta_\gamma\{u_l; x_{i1},\cdots,x_{iK}\} f_\lambda(u_l|x_{10},\cdots,x_{K0})} \right\} \\ &= \sum_{i=1}^{n} \left\{ \ln \eta_\gamma(y_i; x_{i1},\cdots,x_{iK}) + \sum_{l=1}^{L} 1_{(y_i=u_l)}\lambda_l - \ln \left[\sum_{l=1}^{L} \eta_\gamma\{u_l; x_{i1},\cdots,x_{iK}\} \exp(\lambda_l) \right] \right\}, \end{aligned}$$

where $(u_1, ..., u_L)$ are the uniquely observed Y values in the dataset, and a nonparametric model for the baseline function $f_\lambda(y_i|x_{10}, \cdots, x_{K0})$ assigns probability mass $p = (p_1, ..., p_L)$ to $f(u_1|x_{10}, \cdots, x_{K0})$, $\cdots$, $f(u_L|x_{10}, \cdots, x_{K0})$ with the constraint $\sum_{l=1}^{L} p_l = 1$. To relax the constraint, we reparameterize p as $\lambda = (\lambda_1, ..., \lambda_L)$, such that $\lambda_l = \ln(p_l/p_L)$ and thus $p_l = \exp(\lambda_l)/\sum_{l=1}^{L} \exp(\lambda_l)$, for $l = 1, ..., L$.

Let $(\hat\gamma, \hat\lambda)$ denote the MLE of (γ, λ), i.e., $(\hat\gamma, \hat\lambda) = \text{argmax}_{\gamma,\lambda} LL(\gamma, \lambda)$. The MLE of (γ, λ) can be obtained through a Newton-Raphson (NR) type algorithm, which iteratively searches for the parameter values that maximize the model likelihood. The NR type algorithm also requires evaluating the score functions which can be straightforwardly evaluated as follows,

$$\frac{\partial \ln LL_i(\gamma,\lambda)}{\partial \lambda_l} = 1_{(y_i=u_l)} - \frac{\eta_\gamma(u_l; x_{i1},\cdots,x_{iK}) \exp(\lambda_l)}{\sum_{l'=1}^{L} \eta_\gamma(u_{l'}; x_{i1},\cdots,x_{iK}) \exp(\lambda_{l'})},$$

$$\frac{\partial \ln LL_i(\gamma,\lambda)}{\partial \gamma} = \frac{\partial \ln \eta_\gamma(y_i; x_{i1},\cdots,x_{iK})}{\partial \gamma} - \sum_{l=1}^{L} \frac{\eta_\gamma(u_l; x_{i1},\cdots,x_{iK}) \exp(\lambda_l)}{\sum_{l'=1}^{L} \eta_\gamma(u_{l'}; x_{i1},\cdots,x_{iK}) \exp(\lambda_{l'})} \frac{\partial \ln \eta_\gamma(u_l; x_{i1},\cdots,x_{iK})}{\partial \gamma}.$$

The term $\frac{\partial \ln \eta_\gamma(y_i;x_{i1},\cdots,x_{iK})}{\partial \gamma}$ depends on the form of the OR function. When a log-bilinear form is adopted for the OR function, we have

$$\frac{\partial \ln \eta_\gamma(y_i; x_{i1},\cdots,x_{iK})}{\partial \gamma} = (x_i - x_0)(y_i - y_0),$$

where $(x_i - x_0) = (x_{1i} - x_{10}, \cdots, x_{ki} - x_{k0})'$. When the odds ratio function contains interaction and higher-order terms of independent variables, these score functions can be derived similarly. More general forms of odds ratio functions can also be used. The exact form of $\frac{\partial \ln \eta_\gamma(y_i;x_{i1},\cdots,x_{iK})}{\partial \gamma}$ in the score functions will thus depend on the specific functional forms of the odds ratio function. We then feed the evaluated model likelihood value and score function values into the Quasi-Newton routine

UMING in Fortran IMSL library to obtain the MLE. Because none of these evaluations involves integration and only summations over a finite number of points are required, the Quasi-Newton algorithm performs well and converges very quickly.

We illustrate the estimation algorithm using the retail store visit data. To fit the odds ratio model, we specify the odds ratio function and baseline function separately. We first specify a log-bilinear odds ratio function $\ln \eta_\gamma(\texttt{Visit}; \texttt{Income}, \texttt{DTS}, \texttt{DTS}^2, \texttt{Age}, \texttt{Married}, \texttt{Kids})$ as follows

$$\gamma_1(\texttt{Visit} - \texttt{Visit}_0)(\texttt{Income} - \texttt{Income}_0) + \gamma_2(\texttt{Visit} - \texttt{Visit}_0)(\texttt{DTS} - \texttt{DTS}_0) + \gamma_3(\texttt{Visit} - \texttt{Visit}_0)(\texttt{DTS}^2 - \texttt{DTS}_0^2)$$
$$+\gamma_4(\texttt{Visit} - \texttt{Visit}_0)(\texttt{Age} - \texttt{Age}_0) + \gamma_5(\texttt{Visit} - \texttt{Visit}_0)(\texttt{Married} - \texttt{Married}_0)$$
$$+\gamma_6(\texttt{Visit} - \texttt{visit}_0)(\texttt{Kids} - \texttt{Kids}_0),$$

The baseline function $f(\texttt{Visit}; \texttt{Income}_0, \texttt{DTS}_0, \texttt{DTS}_0^2, \texttt{Age}_0, \texttt{Married}_0, \texttt{Kids}_0)$ is specified to be a non-parametric function that assigns a probability mass on the unique values of `Visit` in the data. In retail store visit data set, there are 27 unique values of retail store visits as follows

$$[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20, 24, 26, 28, 31, 32, 35, 38, 131]\,.$$

Thus the probability mass assigned to the lth unique value is $p_l = \exp(\lambda_l)/\sum_{l'=1}^{L} \exp(\lambda_{l'})$, for $l = 1, ..., L$, where $L = 27$ and λ_L is fixed at zero for model identification purpose. We then form the model likelihood and score functions based on the above specifications of odds ratio function and baseline function and apply the NR estimation algorithm as described above. At the convergence, the MLEs of log-odds-ratio parameters γ are reported in Table 4. Note that the choice of reference point $(\texttt{Income}_0, \texttt{DTS}_0, \texttt{Age}_0, \texttt{Married}_0, \texttt{Kids}_0)$ does not affect the estimates of γ parameters. For example, setting these values at the maximums, minimums or means of these variables all yield identical log-odds-ratio estimates. At convergence, the MLEs of $\hat{\lambda}$ in the baseline function are

$$[22.8, 21.3, 20.4, 19.7, 19.4, 18.7, 18.3, 18.0, 17.5, 17.3, 16.6, 16.9, 16.7, 15.3, 15.6, 16.2, 15.8,$$
$$15.0, 14.1, 13.8, 14.3, 13.4, 13.1, 13.0, 12.7, 12.4, 0]\,.$$

In our analysis, the reference point $(\texttt{Income}_0, \texttt{DTS}_0, \texttt{DTS}_0^2, \texttt{Age}_0, \texttt{Married}_0, \texttt{Kids}_0)$ is chosen at the mean of these variables. Thus the baseline function $f(\texttt{Visit}; \texttt{Income}_0, \texttt{DTS}_0, \texttt{DTS}_0^2, \texttt{Age}_0, \texttt{Married}_0, \texttt{Kids}_0)$ gives the probability mass function of store visit for a consumer with average values of profiling variables.

As noted in Section 2, the above estimation algorithm can also be used to obtain MLEs of (γ, λ') in selective sampling when sampling weight is unknown. When sampling weights are known, they can be simply added as offset terms to the above algorithm to obtain MLEs of (γ, λ).

I.2: Bayesian Estimation in the SOR approach to Selective Sampling

In this section, we outline a Bayesian estimation of the SOR model for selective sampling problems. The Bayesian approach can be useful for a broad range of marketing applications. First, because the MLE approach described above relies on a relatively large sample for its optimal asymptotic properties, it may not perform as well in small samples. Furthermore, in many extended marketing applications, there is a need to account for consumer heterogeneity via a high-dimensional of individual consumer parameters (e.g, in panel data). In these cases, a Bayesian estimation approach can have advantages. Below we outline a Bayesian SOR approach to selective sampling problems that can be extended for use with such marketing applications. We describe here an MCMC algorithm to draw samples from the posterior distribution of the odds ratio model parameters (γ, λ).

The posterior density function for (γ, λ) is given as follows:

$$f(\gamma, \lambda | y; x_1, \cdots, x_K) \quad \propto \quad \prod_{i=1}^{n} \left\{ \frac{\eta_\gamma\{y_i; x_{i1}, \cdots, x_{iK}\} f_\lambda(y_i | x_{10}, \cdots, x_{K0})}{\sum_{l=1}^{L} \eta_\gamma\{u_l; x_{i1}, \cdots, x_{iK}\} f_\lambda(u_l | x_{10}, \cdots, x_{K0})} \right\} \pi(\gamma, \lambda), \qquad (1)$$

Because λ can be viewed as the scale parameters and γ can be viewed as the location parameters, we assign an independent prior, $\pi(\gamma, \lambda) = \pi_1(\lambda)\pi_2(\gamma)$, where $\gamma \sim MVN(\mu_\gamma, \nu_\gamma I_{n_\gamma})$ and $\lambda \sim MVN(\mu_\lambda, \nu_\lambda I_{n_\lambda})$, μ_γ and μ_λ are vectors of prior means with length n_γ and n_λ, respectively; n_γ and n_λ are the length of parameters γ and λ, respectively; I_{n_γ} and I_{n_λ} are identity matrices of size $n_\gamma \times n_\gamma$ and $n_\lambda \times n_\lambda$, respectively. One can assign μ_γ and μ_λ to be vectors of zeros and the large values for variances, ν_γ, ν_λ, which lead to diffuse and non-informative priors relative to the data.

The posterior distribution of (γ, λ) as specified in Equation (1) generally does not have a closed form. We adopt the Hybrid Monte Carlo (HMC) method to sample for this posterior distribution. The HMC method is adopted in Qian and Xie (2011) for handling missing covariates in marketing models. For completeness, we describe the HMC sampler below. The HMC sampler uses the idea of Molecular Dynamic (MD) to propose new draws, which is followed by a Metropolis acceptance-rejection method to sample from a target distribution. As compared with random-walk Metropolis-Hasting sampling algorithm, the HMC sampler exploits the local dynamics of the target distribution. Thus the HMC sampler suppresses the randomness of making proposal draws in the RW-MH algorithm. As a result, the HMC can substantially increase the acceptance rate to the level of around 70% for a Markov chain while maintaining a fast mixing of the chain. To sample from the posterior distribution in Equation (1), the HMC augments the parameter $\theta = (\gamma, \lambda)$ with a vector of invented momentum variables p, which has the same dimension as θ, and defines the following Hamiltonian function:

$$H(\theta, p) = U(\theta) + \varphi(p),$$

where $U(\theta) = -\ln f(\gamma, \lambda | y; x_1, \cdots, x_K)$, $\varphi(p) = -\frac{1}{2} p^T p$. The HMC algorithm consists of the following steps to generate a new draw:

- Initialize the system with (θ^{old}, p^{old}).
  Let θ^{old} be the current value of θ, and let the initial value $\theta^0 = \theta^{old}$. Generate p' from a standard Gaussian distribution and then assign to the system an initial momentum: $p^0 = p' - \frac{\delta}{2} U'(\theta^0)$, where $U'(\theta^0)$ is the derivative of $U(\cdot)$ with respect to its argument and δ is a user-specified stepsize.

- Generate a new state (θ^{new}, p^{new}) in the phase space.
  Starting from the initial phase space (θ^0, p^0), an approximate MD algorithm, called leap-frog algorithm, is run L steps to generate a new state (θ^L, p^L) in the phase space, where

  $$\theta^l = \theta^{l-1} + \delta p^{l-1}, \qquad p^l = p^{l-1} - \delta^l U'(\theta^l),$$

  $l = 1, ..., L$, $\delta^l = \delta$ for $l < L$ and $\delta^L = \frac{\delta}{2}$; $U'(\theta^l)$ is the derivative of $U(\theta)$ with respect to θ evaluated at θ^l, and the derivative is

  $$\frac{\partial U(\theta)}{\partial \theta} \quad = \quad -\sum_i \frac{\partial LL_i(\gamma, \lambda)}{\partial \theta} - \frac{\partial \ln \pi(\theta)}{\partial \theta}, \qquad (2)$$

  where $\frac{\partial LL_i(\theta)}{\partial \theta}$ is the score function given in Appendix I.1 and $\frac{\partial \ln \pi(\theta)}{\partial \theta}$ is given below,

  $$\frac{\partial \ln \pi(\theta)}{\partial \lambda_l} = -\frac{\lambda_l - \mu_{\lambda_l}}{\nu_\lambda}, \qquad \frac{\partial \ln \pi(\theta)}{\partial \gamma_v} = -\frac{\gamma_v - \mu_{\gamma_v}}{\nu_\gamma},$$

and μ_{λ_l} is the corresponding element for λ_l in μ_λ, and μ_{γ_v} is the corresponding element for γ_v in μ_γ. Note that the larger ν_λ and ν_γ, the smaller the contribution of the priors in the updating. With sufficiently large values of $\nu_\lambda = \nu_\gamma$, the proirs becomes diffuse and non-informative relative to data.

- Perform accept-rejection of new draw.
  At the end of the leap-frog steps, let the candidate draw $(\theta^{prop}, p^{prop}) = (\theta^L, p^L)$. The algorithm accepts the candidate draw according to the following probability:
  $$min(1, \exp\left\{-H(\theta^{prop}, p^{prop}) + H(\theta^{old}, p^{old})\right\}).$$
  If the candidate draw is accepted, θ^{prop} becomes the new draw; otherwise, θ^{old} becomes the new draw.

Online Appendix II: Bootstrap Procedure

It is important to note that, unlike typical bootstrap procedures using random samples, the bootstrap in our cases must properly take into account selective sampling schemes to reproduce population. Reproducing true population is feasible here because the bootstrap procedure is only required when sampling weights are known. When sampling weights are unknown, only the population odds ratio parameters can be identified without imposing additional assumptions, and thus only standard errors for the odds ratio parameters are needed. In these situations, the regular standard errors are the same as the true standard errors, and bootstrapping is not required. Given supplemental data on sampling weights, we can regenerate the population and thus obtain bootstrap samples. Consider the general case of selection on both Y and X; the bootstrap procedure proceeds as follows:

1. To regenerate the population, we re-sample N_s times with replacement from the set of explanatory values, $x_i, i = 1, \cdots, x_{n_s}$ in $\texttt{stratum} = s$ in the endogenously selected sample, where N_s is the population total in the $\texttt{stratum} = s$ given in the supplemental data. Repeating the step for all strata will generate a simulated population with the same distribution of the covariates X and strata S as in the original population. For each simulated unit, we generate Y from an odds ratio model $f_{\lambda,\gamma}(Y|x)$ where (λ, γ) are replaced with their estimates $(\hat{\lambda}, \hat{\gamma})$ estimated using the endogenously selected sample.

2. Repeatedly draw new endogenously selected samples from the simulated population, using the same sampling scheme and sample size as used to obtain the observed selective sample. Denote these data sets as $D^j, j = 1, \cdots, J$, where J is the pre-specified number of bootstrap samples.

3. Perform the MLE estimation with known sampling weights on the selective sample D^j and denote the resulting estimates as $(\hat{\lambda}^j, \hat{\gamma}^j)$.

4. Repeat Step 3 for $j = 1, \cdots, J$ to obtain a set of J bootstrap estimates $(\hat{\lambda}^j, \hat{\gamma}^j), j = 1, \cdots, J$. Standard error of an estimator is computed as the standard deviation of the estimates across the J bootstrap samples.

Online Appendix III: Full Results from Simulation Studies

Simulation Study I: Endogenously Stratified Sampling

In this Appendix, we provide a full description of the simulation study I, including the motivation for conducting endogenously stratified results, a descriptions of all simulation results, including further results on more complex functional forms and small sample sizes.

The simulation emulates a setting in which the aim is to assess the relationship between the dependent variable Y 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 are simulated from the following four regression models: (1) Normal regression: $Y|x \sim N(\mu_{Y|x}, \sigma^2)$, (2) Truncated normal regression:$Y|x \sim TN(\mu_{Y|x}, \sigma^2, \mathrm{lb} = 0)$, (3) Poisson regression: $Y|x \sim \texttt{Poisson}(\mu_{Y|x})$, and (4) Zero-truncated Poisson regression: $Y|x \sim \texttt{TPoisson}(\mu_{Y|x}, \mathrm{lb{=}1})$, 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 $Y \geq \mathrm{lb}$ 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 $\beta_0 = 0, \beta_1 = \beta_2 = 0.1, \sigma = 1$. These regression models mimic database marketing applications, where Y is the total purchase expenditure (a continuous outcome) for (1) and (2) or the purchase frequency (a count outcome) in stores during the past year for (3) and (4), respectively, and x contains predictors of these purchase outcomes, e.g., a binary variable of having a long relationship with the firm (X_1) and the household income (X_2).

All four population regression models for $f_\theta(Y|x)$ can be re-expressed as odds ratio 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. 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.[1]

We generate a population of 1,000,000 consumers from each of the four models, mimicking large consumer databases possessed by firms. Although the population data on the outcome Y are available to researchers, it is necessary to survey on a sample of the consumer population to obtain complete information on X in order to assess the relationship between X and Y. We consider the sampling schemes as described in Section 3.1 in the main text with a sample size of 500 for all sampling schemes. For each generated sample of 500 surveyed units, we applied the following estimation methods to obtain the effect estimates of X on Y.

Sampling Schemes See the main text.

Estimation Method

- Parametric regressions without sampling adjustment. On each generated survey sample, we estimate the population regression models without adjusting for endogenous selection. This

[1]Furthermore, in all four models, odds ratio parameters capture the association between X and Y parsimoniously, even for truncated outcomes. Alternative approaches for truncated outcomes, such as linear regression with nonparametric mean functions to permit the effects of boundedness/truncation, can be hard to interpret, difficult to adjust for selective sampling, and do not guarantee boundedness in model predictions. Some of these reasons are why parametric truncated regressions are preferred to these alternative approaches for truncated outcomes. Unlike parametric truncated regressions, SOR requires no parametric distributional assumptions and performs better, as shown in Section 4.

analysis is for comparison purposes. We expect that the estimation method will perform well in random sampling but will be subject to selection bias with endogenous selection.

- Parametric sampling adjustment methods. We apply extant sampling adjustment methods assuming parametric outcome distributions, denoted as HW and HW-FI (Hauseman and Wise 1981) for `ES-Y` and COSS-FI (Cosslett 1993) for `ES-YX`, where "FI" denotes that supplement data on sampling weights are available and have been used.[2] These are benchmark adjustment methods for comparison purposes. We expect that these methods will perform well and can correct for selection bias when outcome distributional assumptions are met (i.e, models (1) and (3)) but can yield bias otherwise (i.e, models (2) and (4)). The purpose here is to show that the performance of these parametric adjustment methods depends critically on the outcome distributional assumptions. A further evaluation in Section 4.2 uses the adjustment method assuming parametric truncated distributions as the benchmark and the conclusions remain the same.

- POR. This estimation fits an odds ratio model with a parametric baseline distribution to each generated survey sample. No supplemental data on sampling weights are required or used. 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 distribution for the continuous outcome and a Poisson distribution for the 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 unknown 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 to each generated survey sample. No data on sampling weights are required or used. This is the method described in Sections 2.2.1 and 2.3.1.

- 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 Results over 500 repeated samples are summarized in Tables 1 and 2. We compare the performances of different approaches in estimating and testing the values of effect parameters of X on Y. In odds ratio models, these parameters are the log-odds-ratio parameter γ_1 and γ_2. These log-odds-ratio parameters are also the same as β_1 and β_2 in the population regression models, whose estimates are reported for OLS, Poisson regression (PReg), Truncated regression (TNReg and TPReg) and parametric sampling adjustment methods (HW, HW-FI and COSS-FI). In random sampling (`RS`), all estimation methods perform equally well and can recover the true parameter without bias, with the same efficiency, with 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. Interestingly, although the SOR model has substantially more parameters than those parametric approaches and POR, no efficiency is lost in estimating the odds ratio parameters.

As expected, estimation methods without sampling adjustment (OLS, PReg, TNReg and TPReg) yield significant bias in estimating the regression parameters β_1 and β_2 with `ES-Y` and `ES-YX`, demonstrating the 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

[2] Hauseman 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 used here for comparison.

Table 1 and Poisson in Table 2). However, the parameter estimates have significant bias when the outcomes are truncated normal (TN in Table 1) and truncated Poisson (TPois in Table 2): the means of estimates under selective sampling (`ES-Y` and `ES-YX`) differ significantly from the true parameter values 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 estimation can consistently estimate the odds ratio parameters without bias across sampling schemes and outcome distributions. In fact, baseline distribution function $f(y|x_0)$ is left completely unspecified. It is also highly efficient and performs as well as when the correct parametric sampling adjustment method is used to analyze endogenous selective samples and performs better than them when data violates these distributional assumptions. To further demonstrate the importance of properly accounting for outcome distributions in handling endogenously selected samples, consider the performance of POR, which is the same as SOR except that the baseline function $f(y|x_0)$ is specified parametrically. The POR method appears to perform reasonably well when Y is a conditional normal in population, even though the normal baseline distribution is mis-specified in endogenous samples (`ES-Y` and `ES-YX`). This may mean that the log-odds-ratio parameter γ is more robust to sampling schemes than the regression coefficient β as an association measure. However, the POR method performs poorly in endogenously selected samples (`ES-Y` and `ES-YX`) when the outcome Y is a skewed Poisson distribution in population. This demonstrates that when the baseline distribution is mis-specified, significant bias may be associated with the estimation of odds ratio parameters, indicating the significant advantage of SOR over POR for handling selective samples.

We next examine the efficiency gain of selective sampling over random sampling. To summarize efficiency gain, we examine the D-error measure. The D-error measure is defined as $|\Sigma|^{1/K}$, where Σ is the variance-covariance matrix of the parameter estimates, and K is the number of explanatory variables. A smaller D-error value means greater efficiency with an $\alpha\%$ reduction in D-error corresponding to an $\alpha\%$ smaller sample required to achieve the same level of estimation precision. For `ES-Y` sampling, the reduction in sample size to achieve the same level of estimation precision, compared with random sampling, is in the range of 60—70% (Tables 1 and 2 column "Eff.")[3]. The gain in the power to detect the association between Y and X is also substantial at the same sample size (Tables 1 and 2 column "Power"). Furthermore, knowledge of sampling weights does not improve the estimation of the odds ratio parameters.

By contrast, for `ES-YX` sampling where the selection depends on both outcome Y and explanatory variable X, knowledge of sampling weights and their use in estimation as the SOR-FI does can improve the estimation efficiency by two- to threefold times and the power to detect the effect of relevant explanatory variables from little power (∼20—30%) to full power (100%) with the same sample size, and reduce the required sample size by 50—70% for the same estimation efficiency (Tables 1 and 2 comparing rows "SOR" and "SOR-FI" under column "Eff.").[4] This occurs because, when known sampling weights are not used, a different baseline distribution needs to be estimated

[3] A smaller standard error value means greater efficiency with an $\alpha\%$ reduction in the variance of the estimate corresponding to an $\alpha\%$ reduction in sample size with the same level of estimation precision.

[4] In `ES-YX` scheme, the levels of the binary variable X_1 are used as strata in sampling (i.e., $X_1 = s$). As shown in Eqn (14) in the main paper, X_1 appears in both odds ratio function and sampling weight function. In SOR estimation, these known sampling weights are not used and become unknown quantities. Thus, the effect of X_1 in odds ratio function is collinear with its effect on the sampling weight, which causes a large variability in the estimate of γ_1 and a huge loss of estimation efficiency in SOR. To give SOR a comparison advantage relative to SOR-FI, for SOR we used X_2 for stratified sampling with strata $X_2 <= 0$ and $X_2 > 0$. By contrast, SOR-FI includes known sampling weights in the estimation, avoids this collinearity issue, and permits accurate effect estimation of X_1 when sampling stratifies on X_1. This represents an extreme case of the efficiency gain of SOR-FI relative to SOR in `ES-YX`.

separately for each stratum. This can negatively affect the estimation precision of the odds ratio parameter such that SOR in `ES-YX` has a lower efficiency than in `ES-Y`. Therefore, it appears that when sampling weights are unknown, `ES-Y` is preferred to `ES-YX`. When sampling weights are known, SOR-FI can make use of this information and pool information across strata to estimate a common baseline distribution, which improves the precision of the odds ratio parameters estimation. The result shows that selection on both Y and X followed by SOR-FI can significantly improve the efficiency of the odds ratio parameter estimation: the reduction in sample size, compared with random sampling, can be further improved to the range of 90—94%, a significant improvement on the `ES-Y` sampling scheme. Furthermore, 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`, a striking difference indicating great gain with selective sampling. The increase in power to detect relevant independent variables also leads to improvement in model selection. The percentages in correct models selected using the backward stepwise model selection procedure across all simulated datasets are substantially increased from below 10% in random sampling to 90% in most cases in `ES-YX` (Column "CMS" in Tables 1 and 2).

Further Results We have conducted additional evaluations when the outcome model has more complex functional forms. Tables 5 and 6 in the Online Appendix II report the simulation results when the models (1)-(4) include interaction terms between X_1 and X_2. Section 4.2 reports results over repeated samples when the outcome model includes quadratic terms of independent variables. The conclusions from these additional analyses have broadly the same conclusions and support the superiority of the SOR method for handling selective samples.

As a maximum likelihood approach, the SOR is asymptotically efficient. As reported in Tables 3 and 4, for simulation results when the sample size is reduced to 100, we can see that SOR has excellent finite sample performance. It performs as well as the parametric methods (HW, HW-FI, COSS-FI, and POR) when their distributional assumptions are met: the efficiency loss is negligible even for sample sizes as small as 100.

[bht!]

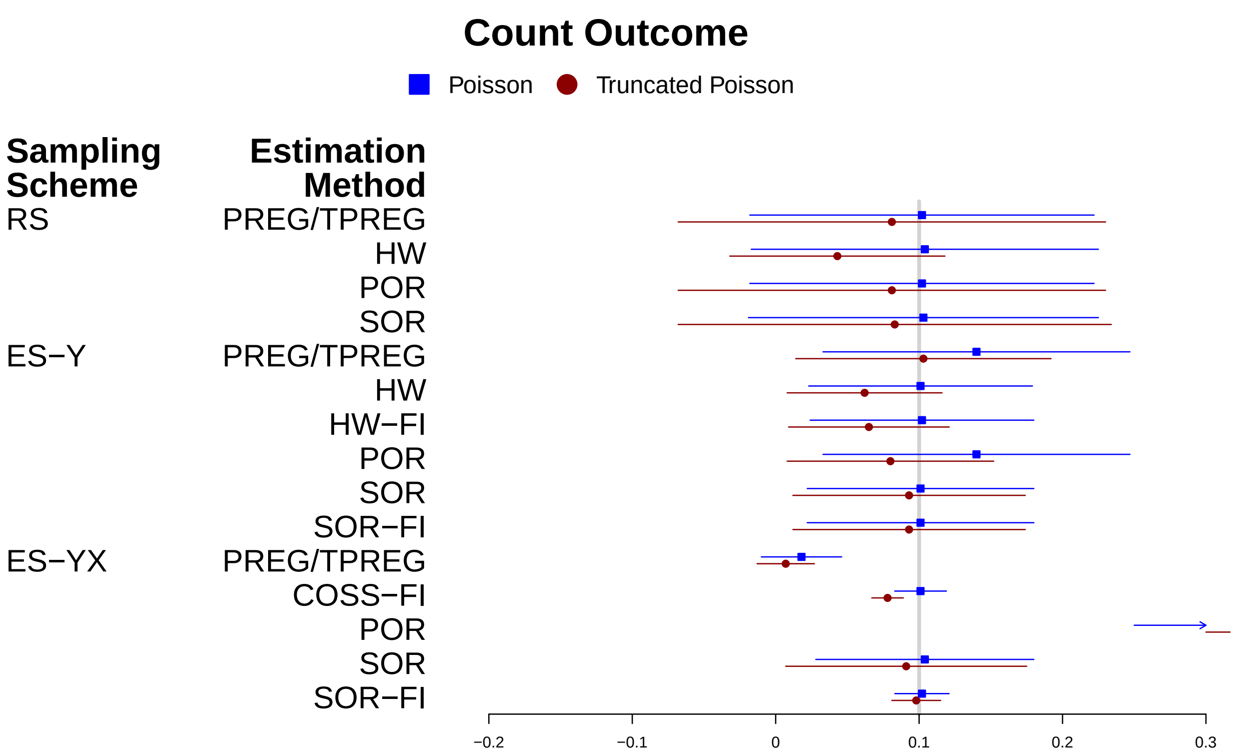

Figure 1. Mean $\pm$ SD of estimates of the regression coefficient of X_1 over 500 repeated samples. Arrows indicate intervals extending out of the limits of x-axis.

Table 1: Performance Comparison for Estimation and Testing of Effect Parameters for continuous outcome. n=500

Y	Sampling Scheme	Est. Method	X_1 (true coeff. value=0.1) Est. (SD) (Emp.SD)	C.R.	Power	X_2 (true coeff. value=0.1) Est. (SD) (Emp.SD)	C.R.	Power	Eff.	CMS
Normal	RS	OLS	0.104 (0.125) (0.127)	95%	14%	0.099 (0.045) (0.046)	95%	60%	0%	5%
		HW	0.104 (0.126) (0.129)	95%	14%	0.099 (0.046) (0.046)	95%	59%	0%	5%
		POR	0.103 (0.127) (0.129)	95%	14%	0.099 (0.046) (0.046)	95%	59%	0%	5%
		SOR	0.103 (0.127) (0.129)	95%	14%	0.099 (0.046) (0.047)	95%	60%	0%	5%
	ES-Y	OLS	0.267 (0.205) (0.204)	88%	−−	0.264 (0.073) (0.071)	38%	−−	−−	−−
		HW	0.101 (0.078) (0.079)	95%	26%	0.100 (0.028) (0.029)	95%	94%	61%	22%
		HW-FI	0.102 (0.078) (0.079)	95%	27%	0.101 (0.028) (0.029)	95%	95%	61%	22%
		POR	0.102 (0.078) (0.079)	95%	−−	0.102 (0.028) (0.029)	94%	−−	−−	−−
		SOR	0.102 (0.078) (0.079)	95%	27%	0.101 (0.028) (0.029)	94%	95%	60%	22%
		SOR-FI	0.102 (0.078) (0.079)	95%	27%	0.101 (0.028) (0.029)	94%	95%	61%	22%
	ES-YX	OLS	0.033 (0.145) (0.047)	100%	−−	0.264 (0.072) (0.069)	37%	−−	−−	−−
		COSS-FI	0.099 (0.019) (0.018)	95%	100%	0.101 (0.028) (0.028)	94%	95%	91%	95%
		POR	0.106 (0.078) (0.079)	94%	−−	0.104 (0.046) (0.048)	95%	−−	−−	−−
		SOR	0.104 (0.078) (0.079)	96%	27%	0.099 (0.046) (0.049)	94%	61%	36%	18%
		SOR-FI	0.101 (0.019) (0.018) (0.055)*	95%	100%	0.100 (0.028) (0.028) (0.028)*	94%	95%	91%	95%
TN	RS	TNReg	0.089 (0.205) (0.203)	96%	6%	0.098 (0.075) (0.076)	95%	28%	0%	2%
		HW	0.035 (0.081) (0.079)	87%	−−	0.038 (0.029) (0.030)	42%	−−	−−	−−
		POR	0.112 (0.210) (0.215)	95%	−−	0.097 (0.075) (0.074)	95%	−−	−−	−−
		SOR	0.088 (0.207) (0.207)	95%	6%	0.097 (0.075) (0.075)	95%	28%	0%	2%
	ES-Y	TNReg	1.136 (2.355) (1.775)	100%	−−	1.412 (1.267) (0.619)	100%	−−	−−	−−

continued on next page

Table 1: *continued*

Y	Sampling Scheme	Est. Method	Est. (SD) (Emp.SD)	C.R.	Power	Est. (SD) (Emp.SD)	C.R.	Power	Eff.	CMS
		HW	0.028 (0.038) (0.039)	55%	−−	0.029 (0.014) (0.014)	0%	−−	−−	−−
		HW-FI	0.031 (0.045) (0.046)	69%	−−	0.032 (0.016) (0.016)	1%	−−	−−	−−
		POR	0.108 (0.131) (0.135)	95%	−−	0.102 (0.047) (0.049)	96%	−−	−−	−−
		SOR	0.096 (0.130) (0.131)	95%	13%	0.102 (0.048) (0.048)	95%	58%	61%	8%
		SOR-FI	0.096 (0.130) (0.131)	95%	13%	0.102 (0.048) (0.048)	95%	58%	61%	8%
	ES-YX	TNReg	0.149 (1.486) (0.509)	100%	−−	1.450 (1.334) (0.651)	100%	−−	−−	−−
		COSS-FI	0.033 (0.010) (0.011)	0%	−−	0.034 (0.017) (0.017)	3%	−−	−−	−−
		POR	0.124 (0.131) (0.129)	95%	−−	0.108 (0.079) (0.081)	95%	−−	−−	−−
		SOR	0.108 (0.129) (0.128)	95%	14%	0.106 (0.079) (0.081)	95%	27%	35%	6%
		SOR-FI	0.097 (0.030) (0.031) (0.094)*	93%	90%	0.100 (0.047) (0.048) (0.048)*	94%	59%	92%	50%

Note: "Est. (SD)": averages of estimates and standard errors over all samples. "Emp.SD.": standard deviation of estimates over all samples. "C.R.": coverage rate. "Power": power to reject hypothesis of the parameter being zero. "Eff.": reduction in sample size relative to the simple random sampling. OLS: ordinary least square. TNReg: truncated normal regression. HW and HW-FI: parametric adjustment methods by Uausman and Wise (1981). COSS-FI: parametric adjustment methods by Cosslett (1993). POR: odds ratio model with parametric baseline distribution. SOR: odds ratio model with nonparametric baseline distribution and unknown sampling weights. SOR-FI: odds ratio model with nonparametric baseline distribution and supplemental data on sampling weights. RS: random sampling. ES-Y: endogenous selection on Y only. ES-YX: endogenous selection on Y and X. *: regular standard errors computed using the inverse information matrix.

Table 2: Performance Comparison for Estimation and Testing of Effect Parameters for count outcome. n=500

			X_1 (true coeff. value=0.1)			X_2 (true coeff. value=0.1)				
Y	Sampling Scheme	Est. Method	Est. (SD) (Emp.SD)	C.R.	Power	Est. (SD) (Emp.SD)	C.R.	Power	Eff.	CMS
Poisson	RS	PReg	0.102 (0.121) (0.120)	95%	14%	0.099 (0.045) (0.045)	94%	61%	0%	10%
		HW	0.104 (0.121) (0.121)	95%	14%	0.100 (0.045) (0.046)	94%	61%	0%	10%
		POR	0.102 (0.121) (0.120)	95%	14%	0.099 (0.045) (0.045)	94%	61%	0%	10%
		SOR	0.103 (0.122) (0.122)	95%	14%	0.101 (0.045) (0.046)	95%	61%	0%	10%
	ES-Y	PReg	0.140 (0.093) (0.107)	88%	−−	0.143 (0.035) (0.043)	68%	−−	−−	−−
		HW	0.101 (0.080) (0.078)	95%	25%	0.098 (0.030) (0.031)	95%	93%	55%	17%
		HW-FI	0.102 (0.080) (0.078)	95%	26%	0.101 (0.030) (0.031)	95%	93%	56%	18%
		POR	0.140 (0.093) (0.107)	88%	−−	0.143 (0.035) (0.043)	68%	−−	−−	−−
		SOR	0.101 (0.080) (0.079)	96%	25%	0.098 (0.030) (0.030)	95%	93%	55%	18%
		SOR-FI	0.101 (0.080) (0.079)	96%	25%	0.098 (0.030) (0.030)	95%	93%	55%	18%
	ES-YX	PReg	0.018 (0.071) (0.028)	98%	−−	0.146 (0.035) (0.041)	72%	−−	−−	−−
		COSS-FI	0.101 (0.018) (0.018)	95%	100%	0.099 (0.029) (0.029)	95%	94%	90%	90%
		POR	0.318 (0.074) (0.068)	14%	−−	0.152 (0.093) (0.104)	87%	−−	−−	−−
		SOR	0.104 (0.080) (0.076)	96%	28%	0.102 (0.049) (0.050)	95%	54%	27%	10%
		SOR-FI	0.103 (0.018) (0.019) (0.058)*	93%	100%	0.100 (0.029) (0.029) (0.030)*	95%	93%	91%	91%
TPois	RS	TPReg	0.081 (0.148) (0.149)	96%	9%	0.100 (0.055) (0.054)	96%	43%	0%	5%
		HW	0.043 (0.105) (0.075)	99%	−−	0.050 (0.038) (0.026)	82%	−−	−−	−−
		POR	0.081 (0.148) (0.149)	95%	−−	0.100 (0.055) (0.054)	95%	−−	−−	−−
		SOR	0.083 (0.150) (0.151)	96%	9%	0.101 (0.055) (0.055)	96%	43%	0%	5%
	ES-Y	TPReg	0.103 (0.085)	95%	−−	0.115 (0.032)	91%	−−	−−	−−

continued on next page

Table 2: *continued*

Y	Sampling Scheme	Est. Method	Est. (SD) (Emp.SD)	C.R.	Power	Est. (SD) (Emp.SD)	C.R.	Power	Eff.	CMS
			(0.089)			(0.0.035)				
		HW	0.062 (0.066) (0.054)	97%	−−	0.066 (0.024) (0.020)	75%	−−	−−	−−
		HW-FI	0.065 (0.068) (0.056)	97%	−−	0.067 (0.025) (0.021)	77%	−−	−−	−−
		POR	0.080 (0.074) (0.072)	95%	−−	0.088 (0.028) (0.025)	94%	−−	−−	−−
		SOR	0.093 (0.081) (0.081)	95%	21%	0.099 (0.031) (0.031)	94%	90%	69%	17%
		SOR-FI	0.093 (0.081) (0.081)	95%	21%	0.099 (0.031) (0.031)	94%	90%	69%	17%
	ES-YX	TPReg	0.007 (0.065) (0.020)	95%	−−	0.120 (0.033) (0.035)	89%	−−	−−	−−
		COSS-FI	0.078 (0.012) (0.011)	76%	−−	0.070 (0.021) (0.021)	54%	−−	−−	−−
		POR	0.853 (0.057) (0.041)	0%	−−	0.073 (0.075) (0.071)	96%	−−	−−	−−
		SOR	0.091 (0.081) (0.084)	95%	19%	0.099 (0.051) (0.052)	96%	52%	49%	9%
		SOR-FI	0.098 (0.018) (0.017) (0.060)*	94%	100%	0.102 (0.030) (0.031) (0.031)*	93%	90%	94%	90%

PReg: Poisson regression; TPReg: truncated Poisson regression. Additionally please see notes under Table 1.

Table 3: Performance Comparison for Estimation and Testing of Effect Parameters for continuous outcome. n=100

Y	Sampling Scheme	Est. Method	X_1 (true coeff. value=0.1) Est. (SD) (Emp.SD)	C.R.	Power	X_2 (true coeff. value=0.1) Est. (SD) (Emp.SD)	C.R.	Power	Eff.
Normal	RS	OLS	0.109 (0.285) (0.286)	94%	7%	0.098 (0.102) (0.101)	95%	16%	0%
		HW	0.109 (0.290) (0.304)	95%	7%	0.099 (0.104) (0.110)	96%	15%	0%
		POR	0.114 (0.300) (0.301)	95%	7%	0.104 (0.107) (0.104)	95%	15%	0%
		SOR	0.114 (0.301) (0.305)	95%	7%	0.105 (0.107) (0.109)	95%	15%	0%
	ES-Y	OLS	0.268 (0.466) (0.494)	92%	−−	0.280 (0.164) (0.169)	80%	−−	−−
		HW	0.109 (0.183) (0.201)	94%	9%	0.109 (0.065) (0.069)	94%	34%	59%
		HW-FI	0.108 (0.183) (0.200)	94%	9%	0.109 (0.064) (0.069)	94%	34%	60%
		POR	0.106 (0.181) (0.198)	93%	−−	0.111 (0.065) (0.069)	94%	−−	−−
		SOR	0.109 (0.184) (0.205)	94%	9%	0.110 (0.065) (0.069)	94%	34%	58%
		SOR-FI	0.109 (0.184) (0.205)	94%	9%	0.110 (0.065) (0.069)	94%	34%	58%
	ES-YX	OLS	0.020 (0.330) (0.105)	100%	−−	0.269 (0.166) (0.161)	82%	−−	−−
		COSS-FI	0.098 (0.048) (0.042)	96%	47%	0.104 (0.066) (0.065)	96%	32%	90%
		POR	0.117 (0.181) (0.184)	91%	−−	0.110 (0.108) (0.113)	96%	−−	−−
		SOR	0.104 (0.184) (0.188)	96%	9%	0.105 (0.108) (0.113)	95%	16%	33%
		SOR-FI	0.099 (0.050) (0.044)	97%	41%	0.100 (0.065) (0.065)	94%	32%	88%
TN	RS	TNReg	0.046 (0.491) (0.509)	97%	3%	0.099 (0.180) (0.179)	97%	9%	0%
		HW	0.033 (0.190) (0.207)	92%	−−	0.052 (0.067) (0.068)	84%	−−	−−
		POR	0.089 (0.493) (0.541)	95%	−−	0.098 (0.174) (0.175)	96%	−−	−−
		SOR	0.051 (0.495) (0.558)	96%	5%	0.098 (0.175) (0.174)	96%	10%	0%
	ES-Y	TNReg	0.454 (5.205) (3.567)	100%	−−	1.053 (2.088) (1.100)	100%	−−	−−
		HW	0.031 (0.089)	86%	−−	0.029 (0.031)	34%	−−	−−

continued on next page

Table 3: *continued*

Y	Sampling Scheme	Method Method	Est. (SD) (Emp.SD)	C.R.	Power	Est. (SD) (Emp.SD)	C.R.	Power	Eff.
			(0.094)			(0.032)			
		HW-FI	0.039 (0.105) (0.113)	88%	−−	0.034 (0.037) (0.037)	53%	−−	−−
		POR	0.104 (0.307) (0.319)	95%	−−	0.103 (0.110) (0.112)	96%	−−	−−
		SOR	0.092 (0.307) (0.321)	95%	5%	0.102 (0.109) (0.112)	96%	15%	70%
		SOR-FI	0.092 (0.307) (0.321)	95%	5%	0.102 (0.109) (0.112)	96%	15%	70%
	ES-YX	TNReg	0.133 (2.993) (0.983)	100%	−−	1.178 (2.415) (1.229)	100%	−−	−−
		COSS-FI	0.033 (0.029) (0.025)	36%	−−	0.036 (0.041) (0.038)	61%	−−	−−
		POR	0.121 (0.306) (0.314)	95%	−−	0.113 (0.182) (0.174)	96%	−−	−−
		SOR	0.110 (0.305) (0.313)	96%	6%	0.114 (0.183) (0.175)	96%	9%	49%
		SOR-FI	0.097 (0.076) (0.074)	95%	26%	0.104 (0.112) (0.111)	95%	14%	93%

Please see notes under Table 1

Table 4: Performance Comparison for Estimation and Testing of Effect Parameters for count outcome. n=100

			X_1 (true coeff. value=0.1)			X_2 (true coeff. value=0.1)			
Y	Sampling Scheme	Est. Method	Est. (SD) (Emp.SD)	C.R.	Power	Est. (SD) (Emp.SD)	C.R.	Power	Eff.
Poisson	RS	PReg	0.043 (0.284) (0.300)	96%	6%	0.101 (0.102) (0.102)	94%	14%	0%
		HW	0.045 (0.290) (0.310)	96%	6%	0.103 (0.104) (0.106)	94%	14%	0%
		POR	0.043 (0.285) (0.300)	96%	6%	0.101 (0.102) (0.102)	94%	14%	0%
		SOR	0.050 (0.295) (0.320)	95%	6%	0.105 (0.106) (0.108)	94%	14%	0%
	ES-Y	PReg	0.135 (0.218) (0.277)	89%	−−	0.150 (0.080) (0.094)	85%	−−	−−
		HW	0.109 (0.189) (0.202)	95%	9%	0.105 (0.068) (0.069)	95%	32%	55%
		HW-FI	0.109 (0.188) (0.200)	95%	10%	0.105 (0.068) (0.068)	95%	32%	56%
		POR	0.135 (0.218) (0.277)	89%	−−	0.150 (0.080) (0.094)	85%	−−	−−
		SOR	0.111 (0.191) (0.205)	95%	9%	0.107 (0.069) (0.070)	95%	32%	54%
		SOR-FI	0.111 (0.191) (0.205)	95%	9%	0.107 (0.069) (0.070)	95%	32%	54%
	ES-YX	PReg	0.009 (0.159) (0.065)	100%	−−	0.157 (0.080) (0.096)	84%	−−	−−
		COSS-FI	0.096 (0.046) (0.045)	96%	66%	0.109 (0.068) (0.068)	93%	31%	90%
		POR	0.112 (0.220) (0.279)	88%	−−	0.147 (0.135) (0.162)	87%	−−	−−
		SOR	0.095 (0.190) (0.203)	95%	8%	0.105 (0.115) (0.118)	95%	14%	24%
		SOR-FI	0.096 (0.047) (0.045)	96%	67%	0.109 (0.071) (0.070)	95%	28%	89%
TPois	RS	TPReg	-0.015 (2.306) (0.809)	97%	4%	0.102 (0.125) (0.123)	94%	12%	0%
		HW	0.019 (0.246) (0.141)	100%	−−	0.050 (0.088) (0.052)	97%	−−	−−
		POR	0.019 (0.227) (0.168)	100%	−−	0.044 (0.081) (0.062)	96%	−−	−−
		SOR	-0.006 (1.576) (0.735)	97%	4%	0.108 (0.131) (0.134)	95%	11%	0%
	ES-Y	TPReg	0.105 (0.194) (0.200)	94%	−−	0.122 (0.074) (0.083)	92%	−−	−−

continued on next page

Table 4: *continued*

Y	Sampling Scheme	Method Method	Est. (SD) (Emp.SD)	C.R.	Power	Est. (SD) (Emp.SD)	C.R.	Power	Eff.
		HW	0.069 (0.151) (0.120)	99%	−−	0.070 (0.057) (0.048)	94%	−−	−−
		HW-FI	0.075 (0.156) (0.128)	98%	−−	0.072 (0.057) (0.050)	94%	−−	−−
		POR	0.083 (0.169) (0.149)	98%	−−	0.089 (0.063) (0.061)	96%	−−	−−
		SOR	0.106 (0.188) (0.186)	96%	9%	0.107 (0.071) (0.075)	96%	34%	95%
		SOR-FI	0.106 (0.188) (0.186)	96%	9%	0.107 (0.071) (0.075)	96%	34%	95%
	ES-YX	TPReg	0.007 (0.146) (0.049)	100%	−−	0.122 (0.073) (0.082)	91%	−−	−−
		COSS-FI	0.083 (0.035) (0.029)	90%	−−	0.072 (0.051) (0.049)	91%	−−	−−
		POR	0.093 (0.171) (0.159)	97%	−−	0.087(0.106) (0.099)	96%	−−	−−
		SOR	0.107 (0.191) (0.195)	95%	8%	0.107 (0.117) (0.121)	95%	52%	92%
		SOR-FI	0.100 (0.048) (0.042)	96%	55%	0.104 (0.076) (0.072)	95%	30%	98%

Please see notes under Table 2

Table 5: Performance Comparison for Estimation and Testing of Effect Parameters for continuous outcome. n=500

			X_1 (true coeff. value=0.1)			X_2 (true coeff. value=0.1)			$X_1 * X_2$ (true coeff. value=0.1)			
Y	Sampling Scheme	Est. Method	Est. (SD) (Emp.SD)	C.R.	Power	Est. (SD) (Emp.SD)	C.R.	Power	Est. (SD) (Emp.SD)	C.R.	Power	Eff.
Normal	RS	OLS	0.106 (0.126) (0.126)	95%	14%	0.100 (0.049) (0.047)	95%	56%	0.097 (0.128) (0.126)	96%	11%	0%
		HW	0.106 (0.127) (0.128)	95%	14%	0.100 (0.049) (0.048)	95%	55%	0.096 (0.129) (0.127)	95%	10%	0%
		POR	0.107 (0.128) (0.128)	95%	14%	0.100 (0.049) (0.048)	95%	56%	0.098 (0.129) (0.127)	95%	10%	0%
		SOR	0.106 (0.127) (0.128)	95%	14%	0.099 (0.049) (0.048)	95%	56%	0.098 (0.129) (0.127)	95%	10%	0%
	ES-Y	OLS	0.255 (0.203) (0.207)	87%	−−	0.269 (0.079) (0.075)	40%	−−	0.218 (0.199) (0.189)	92%	−−	−−
		HW	0.102 (0.080) (0.083)	94%	25%	0.102 (0.031) (0.030)	95%	92%	0.100 (0.083) (0.087)	94%	20%	74%
		HW-FI	0.102 (0.080) (0.084)	94%	26%	0.102 (0.030) (0.030)	95%	92%	0.101 (0.083) (0.087)	94%	20%	74%
		POR	0.099 (0.078) (0.080)	94%	−−	0.106 (0.031) (0.031)	96%	−−	0.082 (0.077) (0.073)	95%	−−	−−
		SOR	0.102 (0.080) (0.082)	94%	25%	0.102 (0.030) (0.030)	95%	92%	0.101 (0.084) (0.088)	95%	20%	73%
		SOR-FI	0.102 (0.080) (0.082)	94%	25%	0.102 (0.030) (0.030)	95%	92%	0.101 (0.084) (0.088)	95%	20%	73%
	ES-YX	OLS	0.026 (0.145) (0.054)	100%	−−	0.274 (0.101) (0.103)	57%	−−	0.225 (0.142) (0.130)	86%	−−	−−
		COSS-FI	0.099 (0.023) (0.021)	95%	100%	0.103 (0.039) (0.036)	96%	80%	0.097 (0.056) (0.056)	94%	42%	95%
		POR	0.101 (0.078) (0.077)	95%	−−	0.103 (0.048) (0.046)	96%	−−	0.091 (0.078) (0.071)	97%	−−	−−
		SOR	0.103 (0.079) (0.079)	95%	25%	0.102 (0.048) (0.046)	96%	58%	0.103 (0.083) (0.080)	95%	22%	58%
		SOR-FI	0.099 (0.022)	95%	100%	0.102 (0.039)	96%	80%	0.098 (0.056)	94%	42%	95%

continued on next page

Table 5: *continued*

Y	Sampling Scheme	Method Method	Est. (SD) (Emp.SD)	C.R.	Power	Est. (SD) (Emp.SD)	C.R.	Power	Est. (SD) (Emp.SD)	C.R.	Power	Eff.
			(0.021)			(0.037)			(0.058)			
TN	RS	TNReg	0.089 (0.209) (0.212)	95%	9%	0.111 (0.082) (0.082)	96%	28%	0.096 (0.209) (0.202)	97%	7%	0%
		HW	0.043 (0.081) (0.082)	90%	−−	0.043 (0.031) (0.032)	54%	−−	0.041 (0.082) (0.082)	89%	−−	−−
		POR	0.116 (0.210) (0.217)	95%	−−	0.112 (0.081) (0.079)	95%	−−	0.112 (0.213) (0.212)	95%	−−	−−
		SOR	0.088 (0.209) (0.212)	95%	9%	0.112 (0.081) (0.080)	95%	29%	0.100 (0.210) (0.206)	96%	7%	0%
	ES-Y	TNReg	0.952 (2.173) (1.749)	100%	−−	1.387 (1.186) (0.610)	98%	−−	0.909 (1.990) (1.571)	100%	−−	−−
		HW	0.033 (0.039) (0.041)	59%	−−	0.031 (0.015) (0.015)	0%	−−	0.034 (0.040) (0.043)	58%	−−	−−
		HW-FI	0.041 (0.045) (0.048)	72%	−−	0.036 (0.017) (0.017)	5%	−−	0.041 (0.047) (0.051)	71%	−−	−−
		POR	0.111 (0.132) (0.136)	93%	−−	0.107 (0.052) (0.052)	97%	−−	0.103 (0.132) (0.133)	94%	−−	−−
		SOR	0.101 (0.132) (0.136)	94%	12%	0.105 (0.052) (0.052)	96%	52%	0.103 (0.136) (0.141)	95%	12%	75%
		SOR-FI	0.101 (0.132) (0.136)	94%	12%	0.105 (0.052) (0.052)	96%	52%	0.103 (0.136) (0.141)	95%	12%	75%
	ES-YX	TNReg	-0.151 (1.271) (0.521)	100%	−−	1.289 (1.221) (0.797)	99%	−−	1.094 (1.476) (1.127)	100%	−−	−−
		COSS-FI	0.039 (0.012) (0.012)	1%	−−	0.036 (0.023) (0.023)	20%	−−	0.035 (0.033) (0.031)	49%	−−	−−
		POR	0.115 (0.132) (0.132)	94%	−−	0.102 (0.081) (0.079)	97%	−−	0.092 (0.133) (0.129)	96%	−−	−−
		SOR	0.104 (0.131) (0.131)	95%	14%	0.103 (0.081) (0.079)	96%	25%	0.095 (0.133) (0.132)	96%	10%	59%
		SOR-FI	0.103 (0.033)	96%	90%	0.105 (0.067)	97%	38%	0.098 (0.091)	96%	24%	96%

continued on next page

Table 5: *continued*

Y	Sampling Scheme	Method Method	Est. (SD) (Emp.SD)	C.R.	Power	Est. (SD) (Emp.SD)	C.R.	Power	Est. (SD) (Emp.SD)	C.R.	Power	Eff.
			(0.034)			(0.067)			(0.088)			

Please see notes under Table 1

Table 6: Performance Comparison for Estimation and Testing of Effect Parameters for count outcome. n=500

Y	Sampling Scheme	Est. Method	X_1 (true coeff. value=0.1) Est. (SD) (Emp.SD)	C.R.	Power	X_2 (true coeff. value=0.1) Est. (SD) (Emp.SD)	C.R.	Power	$X_1 * X_2$ (true coeff. value=0.1) Est. (SD) (Emp.SD)	C.R.	Power	Eff.
Poisson	RS	PReg	0.095 (0.124) (0.124)	96%	15%	0.100 (0.049) (0.049)	95%	52%	0.109 (0.122) (0.126)	94%	16%	0%
		HW	0.096 (0.125) (0.126)	96%	14%	0.102 (0.049) (0.051)	93%	52%	0.110 (0.123) (0.126)	94%	17%	0%
		POR	0.095 (0.124) (0.124)	96%	15%	0.100 (0.049) (0.050)	95%	52%	0.109 (0.122) (0.126)	94%	16%	0%
		SOR	0.098 (0.125) (0.128)	96%	14%	0.102 (0.049) (0.051)	95%	51%	0.110 (0.124) (0.127)	95%	16%	0%
	ES-Y	PReg	0.120 (0.098) (0.128)	84%	−−	0.144 (0.039) (0.044)	74%	−−	0.100 (0.093) (0.099)	94%	−−	−−
		HW	0.098 (0.082) (0.088)	96%	25%	0.097 (0.033) (0.031)	96%	87%	0.097 (0.085) (0.080)	97%	20%	69%
		HW-FI	0.097 (0.082) (0.088)	96%	25%	0.098 (0.033) (0.031)	97%	87%	0.098 (0.084) (0.080)	97%	20%	69%
		POR	0.120 (0.098) (0.128)	84%	−−	0.144 (0.039) (0.044)	74%	−−	0.100 (0.093) (0.099)	94%	−−	−−
		SOR	0.098 (0.083) (0.088)	96%	25%	0.098 (0.033) (0.031)	96%	87%	0.097 (0.085) (0.080)	97%	20%	68%
		SOR-FI	0.098 (0.083) (0.088)	96%	25%	0.097 (0.033) (0.031)	96%	87%	0.097 (0.085) (0.080)	97%	20%	68%
	ES-YX	PReg	-0.030 (0.074) (0.040)	66%	−−	0.148 (0.050) (0.061)	76%	−−	0.126 (0.070) (0.082)	88%	−−	−−
		COSS-FI	0.098 (0.025) (0.023)	96%	95%	0.103 (0.042) (0.044)	94%	67%	0.097 (0.060) (0.062)	95%	40%	88%
		POR	0.308 (0.074) (0.071)	21%	−−	0.139 (0.095) (0.104)	86%	−−	0.143 (0.061) (0.076)	79%	−−	−−
		SOR	0.103 (0.081) (0.079)	95%	26%	0.099 (0.051) (0.052)	94%	56%	0.097 (0.082) (0.080)	97%	22%	58%
		SOR-FI	0.098 (0.025)	96%	95%	0.103 (0.043)	93%	67%	0.098 (0.061)	94%	40%	88%

continued on next page

Table 6: *continued*

Y	Sampling Scheme	Method Method	Est. (SD) (Emp.SD)	C.R.	Power	Est. (SD) (Emp.SD)	C.R.	Power	Est. (SD) (Emp.SD)	C.R.	Power	Eff.
			(0.023)			(0.044)			(0.064)			
TPois	RS	TPReg	0.080 (0.152)	93%	11%	0.103 (0.060)	93%	42%	0.096 (0.148)	94%	12%	0%
			(0.166)			(0.063)			(0.154)			
		HW	0.047 (0.106)	98%	−−	0.051 (0.042)	86%	−−	0.049 (0.103)	98%	−−	−−
			(0.079)			(0.031)			(0.075)			
		POR	0.045 (0.099)	98%	−−	0.043 (0.039)	76%	−−	0.048 (0.099)	98%	−−	−−
			(0.079)			(0.031)			(0.075)			
		SOR	0.082 (0.153)	93%	11%	0.104 (0.061)	93%	41%	0.097 (0.150)	95%	12%	0%
			(0.161)			(0.064)			(0.158)			
	ES-Y	TPReg	0.106 (0.090)	92%	−−	0.127 (0.036)	87%	−−	0.064 (0.082)	93%	−−	−−
			(0.100)			(0.039)			(0.084)			
		HW	0.070 (0.068)	96%	−−	0.070 (0.027)	85%	−−	0.055 (0.066)	94%	−−	−−
			(0.057)			(0.022)			(0.054)			
		HW-FI	0.077 (0.069)	96%	−−	0.072 (0.028)	85%	−−	0.041 (0.069)	95%	−−	−−
			(0.061)			(0.023)			(0.059)			
		POR	0.084 (0.077)	96%	−−	0.093 (0.031)	97%	−−	0.057 (0.072)	94%	−−	−−
			(0.073)			(0.029)			(0.063)			
		SOR	0.105 (0.084)	94%	25%	0.104 (0.034)	95%	89%	0.095 (0.084)	95%	18%	81%
			(0.088)			(0.034)			(0.087)			
		SOR-FI	0.105 (0.084)	94%	25%	0.105 (0.034)	95%	89%	0.095 (0.084)	95%	18%	81%
			(0.088)			(0.034)			(0.087)			
	ES-YX	TPReg	-0.032 (0.069)	54%	−−	0.124 (0.046)	91%	−−	0.098 (0.064)	94%	−−	−−
			(0.034)			(0.049)			(0.067)			
		COSS-FI	0.094 (0.018)	91%	−−	0.073 (0.029)	83%	−−	0.062 (0.040)	84%	−−	−−
			(0.016)			(0.029)			(0.041)			
		POR	0.835 (0.058)	0%	−−	0.086 (0.075)	96%	−−	0.090 (0.048)	97%	−−	−−
			(0.041)			(0.070)			(0.044)			
		SOR	0.104 (0.083)	95%	25%	0.103 (0.052)	96%	51%	0.095 (0.084)	95%	18%	71%
			(0.084)			(0.052)			(0.080)			
		SOR-FI	0.103 (0.030)	96%	90%	0.105 (0.043)	94%	69%	0.095 (0.061)	95%	32%	97%

continued on next page

Table 6: *continued*

Y	Sampling Scheme	Method Method	Est. (SD) (Emp.SD)	C.R.	Power	Est. (SD) (Emp.SD)	C.R.	Power	Est. (SD) (Emp.SD)	C.R.	Power	Eff.
			(0.025)			(0.044)			(0.061)			

Please see notes under Table 2

Table 7: Comparison of Methods for Dealing with Missing Y Values and Sparse Data (n=100) due to Selective Sampling. Sparse data arises due to three reasons here: a small sample size of complete data (n=100) compounded by a significant amount of missingness and the uneven distribution of the binary covariate X_1 ($P(X_1 = 1) = 0.15$).

	Missing	X_1 (true coeff. value=0.1)			X_2 (true coeff. value=0.1)		
Selection Rules	Prop	OLS	Heckit-FIML	SOR	OLS	Heckit-FIML	SOR
ES-Y							
1. Linear Probit Selection	50%	0.052 (0.328)	0.064 (4.072)	0.074 (0.559)	0.073 (0.123)	0.087 (0.369)	0.117 (0.205)
2. Quadratic Probit Selection	59%	0.086 (0.590)	1.845 (33.516)	0.334 (3.889)	0.156 (0.200)	0.151 (1.076)	0.115 (0.153)
3. Extreme Values Selected	50%	0.190 (0.554)	-36.243 (243.237)	0.112 (0.382)	0.167 (0.186)	-3.656 (21.735)	0.097 (0.111)
4. Middle Value Selected	50%	-0.007 (0.165)	-17.02 (518.819)	-0.081 (1.336)	0.015 (0.050)	1.395 (24.249)	0.123 (0.381)
ES-YX							
5. Linear Probit Selection	52%	0.094 (0.405)	-1.133 (44.68)	0.356 (5.059)	-0.403 (0.139)	-0.102 (4.993)	0.109 (0.394)
6. Quadratic Probit Selection	51%	0.087 (0.424)	-0.103 (10.829)	0.173 (0.646)	-0.188 (0.165)	0.069 (1.565)	0.128 (0.350)
7. Extreme Values Selected	50%	0.175 (0.547)	11.606 (279.459)	0.125 (0.394)	0.142 (0.221)	-0.402 (12.877)	0.094 (0.207)
8. Middle Value Selected	50%	0.022 (0.198)	-9.233 (123.569)	0.136 (1.485)	0.063 (0.146)	1.202 (25.073)	0.128 (0.788)

The selection rules for $P(G = 1)$ are as follows:
1. $\Phi(y)$; 2. $\Phi(-1 + 10\beta_1 y + 10\beta_2 y^2)$; 3. =1 for above top or below bottom 25th percentiles of y values and =0 otherwise; 4. =0 for above top or below bottom 25th percentiles of y values and =1 otherwise;
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: =1 for above top or below bottom 20th (30th) percentiles of y values when $x_2 <= 0 (x_2 > 0)$ and =0 otherwise; 8: =0 for above top or below bottom 30th (20th) percentiles of y values when $x_2 <= 0 (x_2 > 0)$ and =1 otherwise.

Online Appendix IV: Example 2: Selective Sampling in Cigarettes Survey Data

In this section, we apply the proposed methodology for endogenously selected samples to real-life data concerning cigarette consumers in a developing country, which is among the leading markets in cigarettes consumption globally. Counterfeits have been a pervasive issue in different industries in fast-growing emerging markets and developing countries (Sudhir *et al.* 2015, Fink *et al.* 2016), and understanding the impacts of and developing strategies against counterfeits can be useful from both policy-makers' and practitioners' perspectives (Qian 2007, Qian 2008, Qian 2014a, Qian 2014b, Qian *et al.* 2015). The data considered here are large-scale survey data collected by a major company in cigarettes sales and marketing in the country on the smoking and consumption behaviors of cigarette consumers. This application focuses on the role of experiencing counterfeits in smokers' consumption of authentic branded cigarette products. Given the high prevalence of counterfeit cigarettes and the impact they can have on authentic products, the in-person survey included the following question: "List the name(s) of cigarette brands for which you purchased/encountered counterfeits in the past year." The survey also asked the smokers to state the name of the authentic cigarette brand consumed most currently, the price per pack (`Price`), and the consumption share (`Share`) of the authentic brand. We then created a binary variable `Buyfake` which is 1 if the smoker purchased/encountered counterfeits for the brand consumed most at the time of being surveyed. The purpose here is to evaluate whether purchasing/encountering counterfeits of an authentic brand in the recent past related to the consumption share of the authentic branded product. It was also of interest to learn whether the consumption share of the authentic brand product depended on its price and the age of the smoker (`Age`). Table 8 provides the definitions and summary statistics of these variables for a total of 47,971 smokers.

The consumption share of the main authentic cigarettes product can be considered as a share-of-wallet (SOW) outcome, a construct often used to measure consumer loyalty and frequently of interest to marketing researchers. Given the difficulty to gather all external data, firms/researchers often have to rely on survey data to measure the SOW outcomes. For example, Baumann *et al.* (2004) analyzed survey data to study determinants of a consumer' (self-reported) SOW at main bank using multiple regressions that relate consumer demographics and service characteristics of main bank to the SOW outcome. We consider a similar model set up in our example. One data limitation is that the main outcome variable `Share` is a self-reported measure rather than actual behavioral outcome. However, the proposed methodology is general and can be applied to the case when data on actual behavioral SOW outcome are available, as explained below.

To illustrate the method for the endogenously selected samples, we consider the following scenario. Suppose that the survey did not ask the question regarding the counterfeit, and thus `Buyfake` was unobserved for all smokers. However, the company could afford to conduct another round of survey to obtain the `Buyfake` values on a sub-sample of the 47,971 smokers. Similar sampling needs can arise in other database marketing situations. Consider that a firm maintains a large consumer database, from which the share of the most purchased sub-brand of this firm can be calculated and reflect actual behaviors. However, whether or not the consumer purchased/encountered the counterfeits for the most consumed sub-brand is not captured in the database and must be asked via a survey on a subsample of the consumer population in the database. In such situations, when the existing databases do not capture all key variables required for the study at hand, a sampling of the existing database is often required to collect the information on these key variables when it is impossible to collect the data on the entire population. We demonstrate below that endogenous sampling schemes followed by the SOR methods can substantially reduce the sample size required to achieve the same estimation precision.

As a gold standard for comparison purposes, the estimation results listed under the column

"Population" in Table 9 are the regression coefficient estimates from three estimation methods (OLS, HW-FI and SOR) using all 47,971 smokers, who can be considered as the finite population here. Although these gold standard population parameters are unknown in practice, they are presented here to facilitate the performance evaluation of different sampling schemes and estimation methods. For ease of interpretation, all continuous variables (`Share`, `Price`, and `Age`) are standardized to have mean zero and variance of one. The SOR model uses a log-bilinear odds ratio function $\ln \eta_\gamma(\texttt{Share}; \texttt{Buyfake}, \texttt{Price}, \texttt{Age})$ of the following specified form:

$$[\gamma_1(\texttt{Buyfake} - \texttt{Buyfake}_0) + \gamma_2(\texttt{Price} - \texttt{Price}_0) + \gamma_3(\texttt{Age} - \texttt{Age}_0)]\,(\texttt{Share} - \texttt{Share}_0),$$

which captures the dependence of the response variable `Share` on the set of explanatory variables (`Buyfake`, `Price`, and `Age`). The γ odds ratio parameters, when exponentiated, represent the fold changes in the odds of having one standard deviation (SD) increase in `Share` with one SD increase in a continuous explanatory variable or a one-unit increase in a binary explanatory variable. These odds ratio association parameters can be used to capture and test for the relationship between each of these covariates with the outcome when holding other covariates constant. A zero value of a parameter estimate thus indicates no association between the outcome and the corresponding explanatory variable. We employ the nonparametric model for the baseline distribution function for `Share`, $f_\lambda(\texttt{Share}|\texttt{Buyfake}_0, \texttt{Price}_0, \texttt{Age}_0)$. The intercept parameter for `const.` is calculated as $\int u\ f_\lambda(u|\texttt{Buyfake}_0, \texttt{Price}_0, \texttt{Age}_0)du = \sum_{l=1}^L u_l \frac{\exp(\lambda_l)}{\sum_{l'=1}^L \exp(\lambda_l')}$, where $u_l, l = 1, \cdots, L$ is the set of uniquely observed values of `Share` in the data and $\lambda_l, l = 1, \cdots, L$ is the set of parameters in the baseline distribution function.[5] The fixed points $(\texttt{Buyfake}_0, \texttt{Price}_0, \texttt{Age}_0)$ can be chosen arbitrarily and do not change the estimates of γ. We chose them to be zeros so that the intercept parameter has the same meaning as those in OLS and HW. Both OLS and HW estimate the regression parameters in the linear regression model: $\texttt{Share} = \beta_0 + \beta_1\texttt{Buyfake} + \beta_2\texttt{Price} + \beta_3\texttt{Age} + \epsilon$. Although OLS assumes no particular error distributional form and HW assumes a normal error distribution, their estimates using population data are identical (Column "Population" in Table 9).

When Y given X follows a normal distribution, the log odds ratio parameter γ is the regression coefficient β divided by the residual variance in the linear regression model. More generally, when the normal error distributional assumption is not satisfied as it is the case here, the SOR model maps to the nonlinear regression model that not only relaxes distributional assumptions but also relaxes the assumption of linear mean structure of the linear regression models. Specifically, the above SOR model maps to the following regression mean model

$$E(\texttt{Share}|x) \quad = \quad m(\gamma_1(\texttt{Buyfake} - \texttt{Buyfake}_0) + \gamma_2(\texttt{Price} - \texttt{Price}_0)) = m(\gamma^T(x - x_0)),$$

where the regression mean relates to the covariates x via the single linear predictor $\gamma^T(x - x_0)$ as

$$m(\gamma^T(x - x_0)) \quad = \quad \frac{\sum_{l=1}^L u_l \exp(\lambda_l + (u_l - \texttt{Share}_0)\gamma^T(x - x_0))}{\sum_{l=1}^L \exp(\lambda_l + (u_l - \texttt{Share}_0)\gamma^T(x - x_0))},$$

which is not restricted to have linear functional forms for regression mean and, unlike linear regression models, can permit both constant and non-constant marginal effects for covariates. In this case, we can interpret the γ estimates as OR association parameters. Alternatively, one can rely on the approach described in Section 2.6 "Practical Usage of SOR" to map the OR parameters in SOR models to the marginal effects on the mean outcome that inform the effects of unit change in X on the unit change in Y to aid the interpretation of the SOR model for continuous outcomes. As

[5] The standard error of the intercept estimate is obtained using the bootstrap algorithm described in Section 2.4.

shown below, we computed the marginal effects of the three covariates on the mean consumption share outcomes and reported average marginal effects of these covariates in Table 9.

Based on the SOR estimation results for all 47,971 consumers, the log-odds-ratio coefficient parameter for `Buyfake` is -0.066, which means the odds of seeing one SD increase in the share of the most consumed brand is reduced by 6.6% when experiencing counterfeits compared with not experiencing counterfeits. Thus, experience of counterfeits related negatively to the brand's consumption share. When consumers encounter more counterfeits of an authentic brand, they may want to reduce their consumption of the brand to avoid bad tastes and health hazards associated with counterfeit cigarettes, as revealed by some smokers' comments on counterfeit cigarettes. To interpret the effect of `Buyfake` on the mean of the outcome `Share` , we convert the log-odds-ratio parameter to marginal effect estimate using Eqn (19) in the main paper. As `Buyfake` is a binary variable, we consider the following marginal effect of a binary variable

$$\Delta(w) \quad = \quad E(Y|\texttt{Buyfake} = 1, w) - E(Y|\texttt{Buyfake} = 0, w),$$

where w denotes the vector of independent variables except the binary regressor `Buyfake`, and $E(Y|\texttt{Buyfake} = 1, w)$ and $E(Y|\texttt{Buyfake} = 1, w)$ denote the means of Y using the Eqn (18), estimated using the MLEs of SOR model parameters. Thus $\Delta(w)$ denotes the difference in the mean of Y when the binary variable `Buyfake` is changed from 0 to 1, when holding the other regressors fixed at the value of w. The marginal effect $\Delta(w)$ varies with the value of w. One can report the marginal effect for particular value(s) of w of interest. Alternatively one can compute the average marginal effect (AME) as follows

$$\bar{\Delta} \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\Delta(w_i),$$

where N is the sample size and w_i is the observed values of W in the population. Our calculation shows that $\bar{\Delta} = -0.0662$ (Table 9). This means that the share of the most consumed brand is associated with a reduction of 0.0662 in the unit of its SD, when experiencing counterfeits compared with not experience counterfeits. Given that the SD of `Share` is 0.18, this amounts to an expected reduction of 1.1 percentage point (=0.0662*0.18) in the outcome `Share`.

We can compute the marginal effects of continuous regressors on the mean of Y using Eqn (19). Unless $\sigma_{Y|x}$ does not depend on x, the marginal effect $\frac{\partial E(Y|x)}{\partial x_k}$ varies with the value of x. In the absence of a particular value x of interest, one can compute the average marginal effect (AME) for a continuous covariate as

$$\overline{\frac{\partial E(Y|x)}{\partial x_k}} \quad = \quad \frac{\gamma_k}{N}\sum_{i=1}^{N}\sigma_{Y|x_i}, \tag{3}$$

where x_i is the vector of regressors for consumer $i, i = 1, \cdots, N$. Using this approach, we connect the log-odds-ratio parameters of continuous regressors to their marginal effects on mean. The log-odds-ratio coefficient parameter for `Price` is -0.020, which means that price also relates negatively to the brand's consumption share. The AME is -0.0195 (Table 9), which means that with one SD increase in price, the share of most consumed product is reduced on average by 0.0195 in the unit of its SD. The variable `Age` has a log-odds-ratio coefficient parameter of 0.061, which means older smokers tend to have a larger consumption share of the most-consumed cigarettes product, likely because the brand consumption habit had been stable among older smokers while younger smokers were experimenting with different brands. Its AME is 0.060 (Table 9), which means with one SD increase in the age, the share of most consumed product is increased on average by 0.06 in the unit of the SD of Share. Model fitting to the population data also shows statistical significance of all three explanatory variables at 0.001 level (Table 9).

Next, we compare the performances of different sampling schemes and estimation methods in recovering these population relationships. For comparison purposes, we first consider a simple random sampling (RS) that randomly draws 2000 consumers from whom the Buyfake values are revealed. We then consider an endogenously sampling scheme (ES-Y) that also draws 2000 consumers with sampling weights depending on the outcome variable Share. The selective sampling scheme over-samples consumers with extreme values of the dependent variable Share. Specifically, 40% of sample size is allocated to each of the two groups above the upper and below the lower 10th percentiles of Share. The remaining 20% is allocated to the third group with a Share value between the upper and lower 10th percentiles. Within each of the three groups, we randomly sample units with the same weight. Finally, we consider an endogenous sampling scheme with sampling weights depending on both the outcome variable and explanatory variables (ES-YX). Specifically, we consider a stratification on the highly skewed variable Price. We first stratify the population of 47,971 smokers at the mean of Price and allocate 1000 sample size to each of the two strata, and within each stratum we apply the ES-Y sampling scheme described above. For each generated sample, we apply three estimation methods: OLS, sampling adjustment methods assuming normal outcome error distributions (i.e, HW-FI for ES-Y and COSS-FI for ES-YX), and SOR-FI. We repeat these sampling and analysis procedures 500 times, with results summarized in Table 9.

The estimation results obtained using random sampling are reported under the column "RS" and show that all three estimation methods are unbiased for estimating population AME reported in "Population" but with considerably larger standard errors.[6] Consequently, all explanatory variables except Age are nonsignficant at the 0.05 level based on the average estimates and standard errors in Table 9. However, under selective sampling schemes (ES-Y) and (ES-YX), only the SOR can recover the population AME well; both the OLS and the adjustment using normal outcome distribution (HW-FI and COSS-FI) generate estimates that have significant bias and do not recover either the population parameter values or the estimates under random sampling (Table 9 Columns ES-Y and ES-YX). To explain the different performances of the parametric adjustment method and the SOR under selective sampling, one can refer to the goodness-of-fit comparison in Table 10, which compares the observed frequencies for intervals of Share and predicted frequencies using normal linear regression and SOR estimated on the population of 47,491 consumers. To compute the predicted frequency for an interval, for an observation in the sample we first compute the predicted probability of falling into this interval for a new observation with the same covariate values as this observation using model parameter estimates, and then sum the predicted probabilities over all observations in the sample to obtain the predicted frequency for this interval. Large discrepancies are evident between observed and predicted frequencies under normal linear regression, whereas the SOR provides excellent goodness-of-fit, matching data very well. Marketing applications often require more than just accurate mean regression parameters; frequently, researchers need accurate and robust estimation of the entire distributions, such as tail probabilities, for prediction, scoring, and targeting purposes. The results in Table 10 demonstrate an important advantage of SOR over OLS/normal linear regression in providing accurate estimation of the entire distribution. This merit is also important for handling selective sampling problems; when considered a missing data problem, it becomes clear that a method that can automatically generate suitable distributions for variables subject to missingness, such as SOR, performs better in imputing missing values and recovering parameter values than methods based on standard but restrictive distributional assumptions.

We next compare the performances of different sampling algorithms using SOR for estimation.

[6] In these OLS and normal linear models, the regression coefficients are equal to AMEs.

Using the endogenous selective sampling scheme (`ES-Y`), the standard errors of parameters estimates are substantially smaller for all three explanatory variables with `Age` significant at the 0.001 level and with `Buyfake` significant at the 0.1 level on average. The analysis of the `ES-Y` sample using SOR is simple and is conducted as if the data were a random sample without the need to correct for selective sampling. This is justified in Section 2.2, which shows that the odds ratio parameters are invariant to sampling schemes and supported empirically by the simulation studies. Furthermore, the estimation does not require knowledge of sampling weights and remains the same regardless of whether or not the sampling weights are known with the same estimation efficiency of D-error$\times$1000=0.48 (Table 9 under column `ES-Y`). We can further improve the estimation efficiency by conducting additional selection on X. In this example, the variable `Price` is a highly-skewed variable, and stratifying the variable in selective sampling and incorporating sampling weights to the analysis of the `ES-YX` sample significantly increases the estimation efficiency, with all three explanatory variables showing statistically significant results at the 0.05 level on average (Table 9 Column "ES-YX" under "Adjustment with SOR-FI"). The analysis is also straightforward and amounts to adding the sampling weights as offsets to the program written for the estimation of the `ES-Y` sample (Section 2.3.2). As Table 9 shows, the efficiency gain for `ES-Y` .vs. `RS` is 0.96/0.48=2.0, which means a reduction of 50% in sample size without compromising the estimation precision. The efficiency gain for `ES-YX` .vs. `RS` is 0.96/0.34=2.8, which means a reduction of $\sim$70% in sample size without compromising the estimation precision. It is important to note that unlike `ES-Y` where SOR and SOR-FI have the same D-error value, when sampling weights are not incorporated into the estimation of `ES-YX` samples, D-error is increased from 0.34 (SOR-FI) to 0.52 (SOR) (Table 9 Column "ES-YX"), erasing all the gain of `ES-YX`. This demonstrates the importance of incorporating the known sampling weights into analysis to realize the efficiency gain of `ES-YX`. The efficiency gain of including known sampling weights in the estimation would have been even larger if we were to stratify on the binary variable `Buyfake`; in this case, the effects of `Buyfake` on the outcome and on sampling weights become collinear if known sampling weights are not included in the estimation, which can cause large variability in its outcome effect estimate.

Table 9 also reports the computational time for estimating each dataset for each method under different sampling schemes. These computational times are obtained on a desktop computer with a 2.7 GHz Intel Xeon processor and 4GM memory. We found that the SOR is computationally feasible for analyzing the population data with 47,491 consumers (52.5 seconds), and it requires less time than HW-FI when handling selective samples. The reason is that SOR replaces the integration $\int p_\psi(G=1|u)f_\theta(u|x)du$ in the denominator of the conditional likelihood Eqn (6) with summations over the finite number of observed data points, whereas HW-FI and COSS-FI need to evaluate this integration numerically.

Table 8: Definition and Summary Statistics of Variables in Counterfeit Cigarettes Survey Data.

Variable	Definition	Mean	SD
`Share`	The fraction of the most-consumed cigarettes brand	0.78	0.18
`Buyfake`	Bought/encountered counterfeits for the most-consumed cigarettes brand in the past year	0.18	0.38
`Price`	The price of the most-consumed cigarettes brand (RMB/pack)	7.3	7.7
`Age`	Age of the smoker	46.2	13.0

References for Online Appendix

Baumann, C., Burton, S., and Elliott, G. (2004) "Determinants of customer loyalty and share of wallet in retail banking," *Journal of Financial Services Marketing,* **9**, 231–248.

Qian, Y. (2007) "Do national patent laws stimulate domestic innovation in a global patenting environment? a cross-country analysis of pharmaceutical patent protection, 1978-2002.," *Review of Economics and Statistics,* **89**, 436–453.

Qian, Y. (2014a) "Brand management and strategies against counterfeits," *Journal of Economics and Management Strategy,* **23**, 317-343.

Qian, Y. (2014b) "Counterfeiters: Foes or friends," *Management Science,* **60**, 2381–2400.

Fink, C., Maskus, K., and Qian, Y. (2016), "The economic effects of counterfeiting and piracy: A review and implications for developing countries." *World Bank Research Observer,* **31**: 1–28.

Sudhir, K., Priester, J., Shum, M., and et al. (2015). "Research opportunities in emerging markets: an inter-disciplinary perspective from marketing, economics, and psychology." *Customer Needs and Solutions,* **2**: 264–276.

Table 9: Comparison of Consumption Share Estimation Using Different Sampling Schemes and Estimation Methods in Counterfeit Cigarettes Survey Data.

Share	Population	RS	ES-Y	ES-YX
	OLS		OLS with No Adjustment	
Buyfake	-0.066 (0.012)***	-0.063 (0.058)	-0.164 (0.084)*	-0.157(0.082)*
Price	-0.020 (0.005)***	-0.022 (0.024)	-0.037 (0.029)	-0.014 (0.025)
Age	0.060 (0.005)***	0.059 (0.023)**	0.108 (0.031) ***	0.106 (0.032)***
Const.	0.012 (0.005)***	0.011 (0.025)	-0.211 (0.034)***	-0.197(0.035)***
Time in seconds	0.06	0.01	0.01	0.01
	HW-FI		Adjustment with HW-FI and COSS-FI	
Buyfake	-0.066 (0.012)***	-0.063 (0.058)	-0.053 (0.028)*	-0.072(0.035)**
Price	-0.020 (0.005)***	-0.022 (0.024)	-0.013 (0.010)	-0.013 (0.011)
Age	0.061 (0.005)***	0.059 (0.023)**	0.033 (0.010) ***	0.046 (0.013)***
Const.	0.012 (0.005)***	0.011 (0.025)	-0.179 (0.011)***	-0.062(0.015)***
Time in seconds	0.06	0.01	9.28	11.75
	SOR		Adjustment with SOR-FI	
Buyfake	-0.066 (0.012)***	-0.062 (0.058)	-0.068 (0.042)*	-0.068(0.034)**
Price	-0.020 (0.005)***	-0.021 (0.022)	-0.020 (0.015)	-0.020 (0.010)**
Age	0.060 (0.004)***	0.059 (0.023)**	0.059 (0.015) ***	0.058 (0.016)***
Const.	0.012 (0.005)***	0.012 (0.025)	0.014 (0.023)	0.015 (0.020)
Time in Seconds	52.5	0.42	0.41	0.50
N	47971	2000	2000	2000
D*1000	0.039	0.96	0.48 (0.48$^{\$}$)	0.34 (0.52$^{\$}$)

Note: the table reports the estimate and standard error in the parenthesis under column "Population" and the averages of estimates and standard errors in the parenthesis over all 500 samples drawn under "RS", "ES-Y" and "ES-YX". '*', '**', and '***' represent statistical significance at 0.1, 0.05, and 0.001 levels, respectively. The estimates for SOR are the average marginal effects (AME) on the mean consumption share outcome. N: Sample size, D*1000: D-error measure multiplied by 1000. $^{\$}$: D*1000 for SOR when known sampling weighs are not included in the estimation. Population: the finite population of all smokers. RS: Random Sampling. ES-Y: Selective Sampling on the outcome only. ES-YX: Selective Sampling on both the outcome and explanatory variables.

Table 10: Comparison of Observed and Predicted Data Frequency in Counterfeit Cigarettes Survey Data.

`Share`	Frequency			Cumulative Frequency		
	Observed	OLS	SOR	Observed	OLS	SOR
> 100	0.000	0.109	0.000	1.000	1.000	1.000
$(95, 100]$	0.235	0.061	0.235	1.000	0.891	1.000
$(90, 95]$	0.025	0.080	0.025	0.765	0.829	0.765
$(85, 90]$	0.108	0.096	0.108	0.740	0.749	0.740
$(80, 85]$	0.018	0.107	0.018	0.633	0.653	0.633
$(75, 80]$	0.210	0.111	0.210	0.614	0.546	0.614
$(70, 75]$	0.013	0.106	0.013	0.405	0.435	0.405
$(65, 70]$	0.135	0.094	0.135	0.391	0.329	0.391
$(60, 65]$	0.008	0.077	0.008	0.256	0.235	0.256
$(55, 60]$	0.119	0.058	0.119	0.248	0.159	0.248
$(50, 55]$	0.006	0.041	0.006	0.129	0.100	0.129
$(45, 50]$	0.086	0.026	0.086	0.123	0.060	0.123
$(40, 45]$	0.003	0.016	0.003	0.037	0.033	0.037
$(35, 40]$	0.022	0.009	0.022	0.034	0.017	0.034
$(30, 35]$	0.001	0.005	0.001	0.012	0.008	0.012
≤ 30	0.011	0.003	0.011	0.011	0.003	0.011