Research Article

Brian Erard*

Modeling Qualitative Outcomes by Supplementing Participant Data with General Population Data: A New and More Versatile Approach

https://doi.org/10.1515/jem-2021-0004

Received January 18, 2021; accepted November 6, 2021; published online December 2, 2021

Abstract: Although one often has detailed information about participants in a program, the lack of comparable information on non-participants precludes standard qualitative choice estimation. This challenge can be overcome by incorporating a supplementary sample of covariate values from the general population. This paper presents new estimators based on this sampling strategy, which perform comparably to the best existing supplementary sampling estimators. The key advantage of the new estimators is that they readily incorporate sample weights, so that they can be applied to Census surveys and other supplementary data sources that have been generated using complex sample designs. This substantially widens the range of problems that can be addressed under a supplementary sampling estimation framework. The potential for improving precision by incorporating imperfect knowledge of the population prevalence rate is also explored.

Keywords: qualitative response, discrete choice, choice-based sampling, supplementary sampling, contaminated controls

JEL Classification: C13, C25, C35

1 Introduction

Often providers of a program or service have detailed information about their clients, but only very limited information about potential clients. Likewise, ecologists frequently have extensive knowledge regarding habitats where a given animal or plant species is known to be present, but they lack comparable information on habitats where they are certain not to be present. In epidemiology, comprehensive information is routinely collected about patients who have been diagnosed with a given disease; however, commensurate information may not be available for individuals who are known to be free of the disease. While it may be highly beneficial to learn about the determinants of participation (in a program or service) or presence (in a habitat or of a disease), the lack of a comparable sample of observations on subjects that are not participants (or that are non-present) precludes the application of standard qualitative response models, such as logit or probit.

***Corresponding author: Brian Erard**, B. Erard & Associates, LLC, 2350 Swaps Ct., Reston, VA 20191-2630, USA,
E-mail: Brian@BrianErard.com

In fact, though, if a *supplementary* random sample can be drawn from the general population of interest, it is feasible to estimate conditional response probabilities. Importantly, this supplementary sample need not include information on whether the subjects are participants or non-participants, present or not present. Rather, it only must include measures of the relevant covariates, comparable to those collected from the *primary* sample (of subjects that are participants or that are present). This sampling scheme, involving a primary sample consisting exclusively of participants and a supplementary sample that includes both participants and non-participants, has been assigned various names in the literature, including "use-availability sampling", "supplementary sampling", "case control sampling with contaminated controls", "presence pseudo-absence sampling", and "presence-background sampling".[1]

The existing literature on qualitative response estimation under this sampling scheme (see, for example, Cosslett 1981, and Lancaster and Imbens 1996) has focused on developing efficient estimators for the case where the primary and supplementary samples are each unstratified random samples from their respective underlying populations. Unfortunately, however, generalization of these estimators to more complex sampling schemes can be challenging and requires detailed knowledge of the designs of both samples. This precludes the application of these estimators when the requisite sample design information is not available. For instance, applications based on supplementary samples from large Census data sources, such as the Current Population Survey and the American Community Survey are ruled out, because the stratification criteria and other design details are not publicly disclosed.[2]

Our new estimators are derived using an approach that is very similar to that used in the earlier work by Cosslett (1981) and Lancaster and Imbens (1996). The key difference is that our derivation relies on the empirical distribution of the covariates in the supplementary sample, whereas theirs relies on the empirical distribution in the combined (primary and supplementary) sample. This new approach leads to estimators that are based on moments, computed separately from each sample, that involve the absolute probability of participation rather than the conditional probability of selection into one of the samples. Consequently, these new estimators are readily adapted to more complex sampling designs. In particular, one can separately construct weighted versions of the moments within each sample using the available sample weights.

The remainder of this paper is organized as follows. To motivate our estimators for the empirically relevant case where the distribution of the covariates is unknown, Section 2 begins by outlining a consistent estimation method for the case where the distribution is known. Section 3 then demonstrates how a very similar approach can be implemented when the covariate distribution is unknown, leading to estimators that are relatively easy to apply even under fairly complex sampling designs. A method for generating more precise estimates when uncertain information about the prevalence rate is available is provided in Section 4. Section 5 offers some considerations for applying the new estimators, and Section 6 concludes. Proofs for the consistency of the new estimators are provided in Appendix A. Appendix B presents some Monte Carlo simulation results, which show that the new use-availability sampling estimators perform comparably to the best existing estimators in small unstratified samples.

2 Known Covariate Distribution

Following the notation of Lancaster and Imbens (1996), let y be a binary indicator equal to 1 (for participation/presence) or 0 (for non-participation/non-presence), and let x represent a vector of discrete and/or continuous covariates with cumulative distribution function $F(x)$. We assume that the conditional probability

1 Discussions of applications of use-availability sampling in various fields include Breslow (1996) [epidemiology]; Keating and Cherry (2004), Royle et al. (2012), and Phillips and Elith (2013) [ecology]; Erard et al. (2020) [tax compliance]; and Rosenman, Goates, and Hill (2012) [substance abuse prevention programs].

2 Even if all sample design information were publicly available, it would be extremely difficult to generalize these existing models to account for the complexity of the designs.

of participation given x follows a known parametric form, $\Pr(y = 1|x; \beta) = P(x; \beta)$, where β is an unknown parameter vector we desire to estimate. Finally, we define the prevalence rate q (the marginal probability that y equals 1) as:[3]

$$q = \int P(x; \beta)\ \mathrm{d}F(x). \tag{1}$$

2.1 Identification

Suppose we have a random sample of size N_1 from the subpopulation of cases with y equal to one. The conditional distribution of x given $y = 1$ is defined as:

$$g(x|y = 1) = \frac{P(x; \beta) f(x)}{q}, \tag{2}$$

where $f(x)$ is the probability density (mass) function associated with x, which is assumed not to depend on β. If $f(x)$ is known, it follows from Eq. (2) that the function $P(x; \beta)/q$ is nonparametrically identified under such a sampling scheme. In many instances, one will be able to measure (at least to some degree of confidence) the value of q. For instance, one may have a reasonably reliable estimate of the take-up rate for a particular government program or the prevalence rate for a given disease. If q is known, then $P(x; \beta)$ is also nonparametrically identified.

When q is unknown, the relative conditional response probability $[P(x_1; \beta)/P(x_2; \beta)]$ continues to be nonparametrically identified. However, identification of β in this case relies on the functional form that has been assigned to $P(x; \beta)$. For certain specifications, it is not possible to separately identify all of the elements of β. For instance, under a linear probability model, $\frac{P(x;\beta_0,\beta_1)}{q} = \left(\frac{\beta_0}{q}\right) + \left(\frac{\beta_1}{q}\right)' x$, so that only the ratio of each element of β to q is identified. Ecological models of resource selection often rely on an exponential (log-linear) probability model. Under this specification, $\ln\left(\frac{P(x;\beta_0,\beta_1)}{q}\right) = (\beta_0 - \ln q) + \beta_1' x$. In this case, each of the slope coefficients of the conditional response probability is identified, but the intercept is not.[4]

Fortunately, the above two cases are exceptional. As discussed by Solymos and Lele (2016), all of the elements of β are identified under most parametric specifications of the conditional response probability, including the logit, probit, arctan, and complementary log–log models, so long as the specification includes at least one continuous covariate. Nonetheless, although formal identification can easily be achieved by relying on commonly used parametric specifications, one will tend to have less confidence in the quality of estimates of absolute probabilities than estimates of relative probabilities when the prevalence rate is unknown.

2.2 Estimation

If the joint distribution of the covariates $F(x)$ is known, consistent estimation of the conditional response probability parameters is relatively straightforward. Consider first the case where the prevalence rate q is unknown. From Eq. (2), the conditional probability of the covariates x given participation is equal to $P(x; \beta) f(x)/q$. So the likelihood function for a sample of N_1 participants can be specified as:

$$L = \left(\sum_{n=1}^{N_1} \ln\left(P(x_n; \beta)\right) + \ln\left(f(x_n)\right) \right) - N_1 \ln(q) \tag{3}$$

3 When x is discrete, the corresponding formula is: $q = \sum p(x) f(x)$, where $f(x)$ is the probability mass function.

4 Under pure choice-based sampling (which is referred to as a "case-control sampling" by epidemiologists and ecologists), the function $\left(\frac{P(x;\beta)}{1-P(x;\beta)}\right)\left(\frac{1-q}{q}\right)$ is identified rather than $\left(\frac{P(x;\beta)}{q}\right)$. As a consequence, the intercept of the logit specification is not identified under a pure choice-based model when the prevalence rate is unknown, whereas it is the intercept of the exponential probability specification that is not identified under a supplementary sampling design.

where, from Eq. (1), q is constrained to equal $\int P(x;\beta)\ \mathrm{d}F(x)$. After substituting this expression for q into the objective function, an estimate of β can therefore be obtained by maximizing the concentrated likelihood function:[5]

$$\max_{\beta}\left(\sum_{n=1}^{N_1}\ln\left(P(x_n;\beta)\right)\right)-N_1\ln\left(\int P(x;\beta)\ \mathrm{d}F(x)\right). \tag{4}$$

An estimate ($\tilde{q}$) of the prevalence rate can be obtained, if desired, using the formula $\tilde{q}=\int P\left(x;\tilde{\beta}\right)\ \mathrm{d}F(x)$, where $\tilde{\beta}$ represents the estimated value of β.

If the prevalence rate is known, one can instead estimate β by solving the following constrained maximum likelihood estimation problem:

$$\max_{\beta}\sum_{n=1}^{N_1}\ln\left(P(x_n;\beta)\right)\quad s.t.\quad q=\int P(x;\beta)\ \mathrm{d}F(x)\,. \tag{5}$$

As shown in Appendix A, the consistency of these estimators can be verified using the same lemmas employed by Manski and McFadden (1981) to establish the consistency of their choice-based sampling estimators. Rather remarkably, then, if one actually knew the covariate distribution, it would be possible to estimate the conditional probability of participation using a sample that consists entirely of participants.

3 Estimation under a Use-Availability Design

In most practical applications, the joint distribution of the covariates is unknown. However, we can overcome our ignorance of the covariate distribution by incorporating a supplementary sample of covariate values from the overall population. Under the baseline use-availability design discussed in this paper, a primary sample of N_1 observations is randomly drawn from the subpopulation of participants, and a separate supplementary sample of N_0 observations is randomly drawn from the general population. However, estimation under more complex designs, such as stratified sampling within one or both samples, is also explored.

As noted by Lancaster and Imbens (1996), the supplementary sample under a use-availability design would permit identification of $f(x)$, while the primary sample would permit identification of $P(x;\beta)f(x)/q$. Consequently, the function $P(x;\beta)/q$ continues to be non-parametrically identified under this approach, as does the vector β of conditional response probability parameters when the prevalence rate is known.

3.1 Estimation when $F(x)$ and q are Both Unknown

To motivate our new estimator for the case where both the covariate distribution and the prevalence rate are unknown, we begin by deriving the estimator that was independently developed for this case by Cosslett (1981) and Lancaster and Imbens (1996). In our presentation, we explain why it can be challenging to implement a generalized version of their estimator that is appropriate under more complex sampling designs, and we discuss how our new estimator avoids this difficulty.

3.1.1 The Cosslett-Lancaster-Imbens Estimator

Following the approach of Lancaster and Imbens (1996), this estimator is initially derived based on the assumption that the covariate x is discrete with K known points of support. The resulting estimator is then shown to be more generally applicable to cases involving continuous covariates. Whereas the share of

5 The term $\ln\left(f\left(x_n\right)\right)$ has been excluded from Eq. (4) since $f(x)$ is assumed not to depend on β.

observations in the primary sample (N_1/N) is treated as fixed under our estimation framework, Lancaster and Imbens assume more generally that this share is randomly determined through N trials of a Bernoulli process with parameter h.[6] Under this process, h represents the probability that an observation is drawn from the subpopulation of participants for the primary sample, while $(1-h)$ represents the probability that the observation is drawn from the general population of participants and non-participants for the supplementary sample.

Building on Eq. (2) above, the probability $g(x_k)$ that the discrete covariate x is equal to x_k in the combined sample can be expressed as:

$$g(x_k) = \left[h\left(\frac{P(x_k;\beta)}{q}\right) + (1-h)\right] f(x_k), \quad k = 1, \ldots, K, \tag{6}$$

where $f(x_k)$ represents the unconditional probability that x is equal to x_k. The conditional probability that an observation from the combined sample with value $x = x_k$ belongs to the primary sample may therefore be expressed as $R(x_k;\beta,q,h) = \left(\frac{hP(x_k;\beta)/q}{hP(x_k;\beta)/q+1-h}\right)$. This leads to the following constrained maximum likelihood problem:

$$\underset{\beta,q,h,g(x_1),\ldots,g(x_K)}{\text{Max}} L = \sum_{k=1}^{K} N_{1k} \ln\left(R(x_k;\beta,q,h)\right) + N_{0k} \ln\left(1 - R(x_k;\beta,q,h)\right) + N_k \ln\left(g(x_k)\right) \tag{7}$$
$$s.t. \sum_{k=1}^{K} g(x_k) = 1 \quad \text{and} \quad \sum_{k=1}^{K} R(x_k;\beta,q,h)\, g(x_k) = h.$$

In this expression, N_{1k} (N_{0k}) represents the number of observations in the primary (supplementary) sample with covariate value $x = x_k$ and $N_k = N_{0k} + N_{1k}$. If one substitutes the empirical mass function of the covariate within the combined sample $\left(\tilde{g}(x_k) = \frac{N_k}{N}; k = 1, \ldots, K\right)$ in place of the theoretical mass function and optimizes over the remaining parameters, Lancaster and Imbens have demonstrated that one obtains $\tilde{h} = N_1/N$ as a consistent estimator of h, while consistent estimators of β and q may then be obtained from the much simpler unconstrained optimization problem:

$$\max_{\beta,q} L = \sum_{k=1}^{K} N_{1k} \ln\left(R\left(x_k;\beta,q,\tilde{h}\right)\right) + N_{0k} \ln\left(1 - R\left(x_k;\beta,q,\tilde{h}\right)\right) \tag{8}$$

Equivalently, this optimization problem may be expressed as:

$$\max_{\beta,q} L = \sum_{n=1}^{N} s_n \ln\left(R\left(x_n;\beta,q,\tilde{h}\right)\right) + (1-s_n) \ln\left(1 - R\left(x_n;\beta,q,\tilde{h}\right)\right), \tag{9}$$

where s_n is an indicator equal to 1 for observations from the primary sample and 0 for those from the supplementary sample.[7] The first-order conditions for this problem are:

$$\sum_{n=1}^{N} \frac{P'_\beta(x_n;\beta)}{P(x_n;\beta)} \left(s_n - R\left(x_n;\beta,q,\tilde{h}\right)\right) = 0 \tag{10}$$

$$-\frac{1}{q}\left(N_1 - \sum_{n=1}^{N} R\left(x_n;\beta,q,\tilde{h}\right)\right) = 0, \tag{11}$$

---

6 This generalization plays an important role in the Lancaster-Imbens estimator for the case of a known prevalence rate. For that estimator, the restriction $\sum_{n=1}^{N} R(x_n;\beta,q,h) = Nh$ plays an analogous role to our restriction $\sum_{j=1}^{N_0} P(x_j;\beta) = N_0 q$. In our estimation framework, the estimator of h (N_1/N) is completely independent of the estimators for the other parameters, so it is more convenient to condition the analysis on the observed values of N_1 and N_0.

7 Cosslett (1981) derives the same estimator based on the maximization of the following pseudo-likelihood function: $L = \sum_{n=1}^{N} s_n \ln\left(\lambda P(x_n;\beta)\right) - \ln\left(\lambda P(x_n;\beta) + \frac{N_0}{N}\right)$ over β and λ, where the weight factor λ is related to the prevalence rate through the constraint $\left(\lambda q + \frac{N_0}{N}\right) = 1$.

where $P_\beta(x;\beta) = \frac{\partial P(x;\beta)}{\partial \beta}$. As noted by Lancaster and Imbens, this formulation of the sample moment conditions does not require knowledge of the points of support for x or otherwise rely on the assumption of a discrete covariate distribution. Consequently, the scores of the likelihood function can also be used to consistently estimate the parameters under the more general case involving continuous covariates.

Observe that the above estimators of β and q rely on the computation of $R\left(x_n;\beta,q,\tilde{h}\right)$ for each observation in the combined sample. Under a more general sampling scheme, this expression for the conditional probability that an observation belongs to the primary sample will depend on the designs for each of the samples. Consider, for example, the relatively simple modification where simple random sampling is replaced by stratified random sampling only in the case of the supplementary sample. Suppose that a normalized set of sample weights is provided with the supplementary sample: $w_{0b} = \left(\frac{T_b}{N_{0b}}\right)\left(\frac{N_0}{T}\right),\ b = 1,\ldots,B$, where B represents the number of strata, T is the population size, and T_b and N_{0b} respectively represent the population and supplementary sample counts of observations within stratum b.[8] Under this scenario, the generalized Cosslett-Lancaster-Imbens estimators of β and q would be obtained through the following optimization problem:

$$\max_{\beta,q} L = \sum_{b=1}^{B}\sum_{n=1}^{N_b} s_{nb}\,\ln\left(R\left(x_{nb};\beta,q_b(q),\tilde{h}_b\right)\right) + (1-s_{nb})\ln\left(1-R\left(x_{nb};\beta,q_b\left(q\right),\tilde{h}_b\right)\right), \tag{12}$$

where $R\left(x_{nb};\beta,q_b(q),\tilde{h}_b\right) = \left(\frac{\tilde{h}_b P(x_{nb};\beta)/q_b}{\tilde{h}_b P(x_{nb};\beta)/q_b+1-\tilde{h}_b}\right)$, $\tilde{h}_b = \frac{N_{1b}}{N_b}$, $q_b = \left(\frac{N_{1b}}{N_1}\right)\left(\frac{N_0}{w_{0b}N_{0b}}\right)q$, N_{1b} represents the number of primary sample observations in stratum b, and N_b represents the total number of observations in stratum b in the combined sample. To implement this generalized estimator, one would need more information than just the sample weights. Since computation of the stratum-specific terms in Eq. (12) requires the observations from the two samples to be assigned to a common set of strata, the stratification criteria for the supplementary sample would need to be available. Furthermore, the underlying stratifiers would need to be present in both of the samples.

3.1.2 A New Estimator for the Case of an Unknown Prevalence Rate

The derivation of our new estimator for the case of an unknown prevalence rate also begins with the scenario of a discrete covariate distribution with K known points of support. However, it relies on the empirical distribution of x within the supplementary sample alone rather than its distribution within the combined sample. Specifically, rather than estimating $g(x_k)$ using $\tilde{g}(x_k) = \frac{N_k}{N}$ and specifying the likelihood function in terms of the conditional probability of sample assignment, we estimate $f(x_k)$ using $\tilde{f}(x_k) = \frac{N_{0k}}{N_0}$ and rely on an analogue of the concentrated likelihood function specified in Eq. (4):

$$L_{\text{qunk}} = \left(\sum_{k=1}^{K} N_{1k}\,\ln\left(P\left(x_k;\beta\right)\right)\right) - N_1\,\ln\left(\sum_{k=1}^{K} P\left(x_k;\beta\right)\tilde{f}(x_k)\right). \tag{13}$$

Since $\tilde{f}(x_k)$ is a consistent estimator of $f(x_k)$, $\left(\sum_{k=1}^{K} P\left(x_k;\beta\right)\tilde{f}(x_k)\right)$ is a consistent estimation formula for q, so that maximization of this pseudo-likelihood function over β yields consistent estimates of the response probability parameters. Equivalently, this optimization problem may be expressed as:

$$\tilde{\beta}_{\text{qunk}} = \arg\max_{\beta}\left(\sum_{i=1}^{N_1}\ln\left(P\left(x_i;\beta\right)\right)\right) - N_1\,\ln\left(\frac{\sum_{j=1}^{N_0} P\left(x_j;\beta\right)}{N_0}\right). \tag{14}$$

8 This formula produces normalized weights with a mean value of one within the supplementary sample in order to ensure compatibility with the implicit sample weight of one for observations within the primary sample.

The first-order conditions for this estimator are:

$$\sum_{n=1}^{N} s_n \frac{P'_\beta(x_n;\beta)}{P(x_n;\beta)} - \left(\frac{N_1}{N_0 \hat{q}(\beta)}\right)(1-s_n)P'_\beta(x_n;\beta) = 0, \tag{15}$$

where $\hat{q}(\beta) = \sum_{j=1}^{N_0} P\left(x_j;\beta\right)/N_0$. The prevalence rate can be estimated by: $\tilde{q}_{\text{qunk}} = \hat{q}\left(\tilde{\beta}_{\text{qunk}}\right)$. Thus, the new estimators of β and q jointly satisfy the following sample moment conditions:[9]

$$\sum_{n=1}^{N} s_n \frac{P'_\beta(x_n;\beta)}{P(x_n;\beta)} - \frac{N_1}{N_0 q}(1-s_n)P'_\beta(x_n;\beta) = 0 \tag{16}$$

$$\sum_{n=1}^{N} (1-s_n)(q - P(x_n;\beta)) = 0. \tag{17}$$

As with the sample moments associated with the Cosslett–Lancaster–Imbens estimators, these moments do not depend on the points of support for x, but rather only on the realized values of the observations. In fact, they remain valid even when x is not discrete. The consistency of these new estimators in the case of a more general covariate distribution is established in Appendix A. Intuitively, the analog formula for the prevalence rate converges to the true formula as the supplementary sample size increases, so that the pseudo-likelihood function in Eq. (14) converges to the true likelihood function in Eq. (4).[10] As discussed below in Section 3.3, the standard errors of these parameter estimates can be derived from a GMM framework based on the sample moment conditions presented in Eqs. (16) and (17).

A key advantage of the new estimators of β and q is that the likelihood function and its scores involve only sample-specific terms. Consequently, it is straightforward to generalize these estimators to account for more complex sampling schemes using only the provided sample weights. For instance, suppose that each of the samples is exogenously stratified with respective sample weights of w_1 and w_0, which are normalized to sum to N_1 and N_0, respectively. Then the generalized pseudo-likelihood function is: $L_w = \left(\sum_{i=1}^{N_1} w_{1i} \ln\left(\mathrm{P}\left(x_i;\beta\right)\right)\right) - N_1 \ln(q)$. Maximization of this function over β and q subject to $N_0 q = \sum_{j=1}^{N_0} w_{0j}\mathrm{P}\left(x_j;\beta\right)$ yields the generalized first-order conditions:

$$\sum_{i=1}^{N_1} w_{1i} \frac{P'_\beta(x_i;\beta)}{P(x_i;\beta)} - \frac{N_1}{N_0 q}\sum_{j=1}^{N_0} w_{0j} P'_\beta\left(x_j;\beta\right) = 0, \tag{18}$$

$$N_0 q - \sum_{j=1}^{N_0} w_{0j} P\left(x_j;\beta\right) = 0. \tag{19}$$

3.2 Estimation when $F(x)$ is Unknown and q is Known

When the prevalence rate is known, Lancaster and Imbens (1996) again begin by deriving an estimator for the case where x is discrete with K known points of support, relying on the empirical probabilities $\tilde{g}(x_k) = \frac{N_k}{N}$. As with the unknown q case, this leads to moment conditions involving the conditional probability of selection into the primary sample $R(x;\beta,q,h)$ for each observation in the combined sample. Consequently, generalization of the estimator for application with more complex sampling schemes is again challenging and requires detailed knowledge of the designs for each sample.[11]

---

9 These moment conditions also can be obtained directly by maximizing the unconcentrated likelihood function $L = \left(\sum_{n=1}^{N} s_n \ln P(x_n;\beta)\right) - N_1 \ln(q)$ subject to the constraint $N_0 q = \sum_{n=1}^{N}(1-s_n)\,P(x_n;\beta)$.

10 See Lele and Keim (2006) for a related simulation-based approach to estimation when the prevalence rate is unknown.

11 One existing qualitative response model estimator for the case of a known prevalence rate (Steinberg and Cardell 1992) can be readily implemented using the sample weights under a use-availability sample design with exogenous stratification. However, this estimator has been shown to be quite inefficient and is subject to convergence problems in small samples (see Lancaster and Imbens 1996).

When the prevalence rate is known, our new estimator for the case when x is discrete is based on an analog of the optimization problem described in Eq. (5); specifically, it involves maximizing $\sum_{k=1}^{K} N_{1k} \ln(P(x_k;\beta))$ subject to $q = \sum_{k=1}^{K} P(x_k;\beta)\tilde{f}(x_k)$. This estimator ($\tilde{\beta}_{\text{qkn}}$) can be expressed in an alternative way as:

$$\tilde{\beta}_{\text{qkn}} = \underset{\beta}{\text{argmax}} \sum_{i=1}^{N_1} \ln(P(x_i;\beta)) \quad s.t. \quad N_0 q = \sum_{j=1}^{N_0} P(x_j;\beta). \tag{20}$$

The Lagrangian for the optimization problem in Eq. (20) is:

$$\mathcal{L}(\beta,\mu) = \sum_{i=1}^{N_1} \ln(P(x_i;\beta)) + \mu\left(N_0 q - \sum_{j=1}^{N_0} P(x_j;\beta)\right), \tag{21}$$

and the first-order conditions are:

$$\frac{\partial \mathcal{L}}{\partial \beta} = \sum_{i=1}^{N_1} \frac{P'_\beta(x_i;\beta)}{P(x_i;\beta)} - \mu \sum_{j=1}^{N_0} P'_\beta(x_j;\beta) = 0. \tag{22}$$

$$\frac{\partial \mathcal{L}}{\partial \mu} = N_0 q - \sum_{j=1}^{N_0} P(x_j;\beta) = 0. \tag{23}$$

Once again, the sample moment conditions do not require knowledge of the points of support, and they remain valid even when x is not discrete. A proof for the consistency of this estimator is provided in Appendix A.

It is desirable to have a consistent estimate of β to use as an initial value in the search for a solution to the above optimization problem. It can be shown that the limit value for the Lagrange multiplier μ in Eq. (21) is equal to $N_1/(N_0 q)$. Similar to the approach used by Manski and McFadden (1981) to develop an initial consistent estimator for the standard choice-based sampling problem, one can consistently estimate β by replacing μ with its limit value and maximizing the following pseudo-likelihood function:

$$L = \sum_{n=1}^{N} s_n \ln(P(x_n;\beta)) - \frac{N_1}{N_0 q}(1-s_n) P(x_n;\beta). \tag{24}$$

The first-order conditions for this optimization problem are equivalent to those provided in Eq. (16).

We refer to our estimation methodology for the known prevalence rate case as "calibrated qualitative response estimation", because the estimator is obtained by calibrating the response probabilities so that their average value within the supplementary sample is equal to the population prevalence rate q. Following standard terminology for the classic qualitative response framework, we refer to our model as a "calibrated probit" when $P(x;\beta)$ is cumulative standard normal, and as a "calibrated logit" when $P(x;\beta)$ is cumulative standard logistic.

The estimator $\tilde{\beta}_{\text{qkn}}$ is calibrated to ensure that the average predicted probability of participation in the supplementary sample is consistent with the prevalence rate, even in small samples. To solve the constrained optimization problem for this estimator, one can rely on readily available algorithms, such as the maxLik package in R, the nonlinear optimization routines in SAS@/IML@, or the CML application in GAUSS@.

As discussed in Section 3.3, the covariance matrix for $\tilde{\beta}_{\text{qkn}}$ can be obtained using a GMM framework. Since the terms of Eq. (21) through (23) are sample-specific, it is again straightforward to generalize this estimator to account for more complex sampling schemes by separately applying the relevant set of sample weights to each term. The generalized optimization problem is:

$$\mathcal{L}(\beta,\mu) = \sum_{i=1}^{N_1} w_{1i} \ln(P(x_i;\beta)) + \mu\left(N_0 q - \sum_{j=1}^{N_0} w_{0j} P(x_j;\beta)\right), \tag{25}$$

and the associated first-order conditions are:

$$\frac{\partial \mathcal{L}}{\partial \beta} = \sum_{i=1}^{N_1} w_{1i} \frac{P'_\beta(x_i;\beta)}{P(x_i;\beta)} - \mu \sum_{j=1}^{N_0} w_{0j} P'_\beta(x_j;\beta) = 0. \tag{26}$$

$$\frac{\partial \mathcal{L}}{\partial \mu} = N_0 q - \sum_{j=1}^{N_0} w_{0j} P(x_j;\beta) = 0. \tag{27}$$

3.3 GMM Estimation Framework

Although the conditional response probability parameters can be estimated for both the unknown and known prevalence rate cases using the pseudo-maximum likelihood framework described previously, the usual estimate of the covariance matrix of the parameter estimates based on the information matrix will not be valid. Intuitively, the replacement of the exact formula for $q\left(\int P(x;\beta)\,\mathrm{d}F(x)\right)$ with its sample analog $\left(\sum_{j=1}^{N_0} P(x_j;\beta)/N_0\right)$ reduces the precision of the estimators to some degree. Following Lancaster and Imbens (1996), we recast our estimation problems using the GMM framework, relying on moment conditions derived from the scores of our pseudo-likelihood functions. Consider the population moment conditions $E(g_1(x;\theta)) = 0$ and $E(g_2(x;\theta)) = 0$, where:

$$g_1(x;\theta) = s\frac{P'_\beta(x;\beta)}{P(x;\beta)} - (1-s)\left(\frac{N_1}{N_0 q}\right) P'_\beta(x;\beta), \tag{28}$$

$$g_2(x;\theta) = (1-s)(q - P(x;\beta)). \tag{29}$$

The validity of these moment conditions can be verified by applying the following formula for $E(s|x)$:

$$E(s|x) = P(s=1|x) = \frac{\frac{N_1}{N} P(x;\beta)/q}{\frac{N_1}{N} P(x;\beta)/q + \frac{N_0}{N}}. \tag{30}$$

Taking the conditional expectation of $g_1(x;\theta)$ in Eq. (28) and substituting the formula for $E(s|x)$ from Eq. (30) yields:

$$E(g_1(x;\theta)|x) = \frac{\frac{N_1}{N} P'_\beta(x;\beta)/q}{\frac{N_1}{N} P(x;\beta)/q + \frac{N_0}{N}} - \left(\frac{N_1}{N_0 q}\right) \frac{\frac{N_0}{N} P'_\beta(x;\beta)}{\frac{N_1}{N} P(x;\beta)/q + \frac{N_0}{N}}, \tag{31}$$

which does in fact equal zero. It is easily verified that the second moment condition is also satisfied.

Let $\theta = \begin{pmatrix} \beta \\ q \end{pmatrix}$ for the case where the prevalence rate is unknown and $\theta = \beta$ for the case where it is known. Thus, the model is exactly identified in the former case and over-identified in the latter. Let $g(x;\theta)$ represent the vector $\begin{bmatrix} g_1(x;\theta) \\ g_2(x;\theta) \end{bmatrix}$ and define the associated vector of sample moment conditions as: $g_N(x;\theta) = \begin{bmatrix} g_{N_1}(x;\theta) \\ g_{N_2}(x;\theta) \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{n=1}^{N} g_1(x_n;\theta) \\ \frac{1}{N}\sum_{n=1}^{N} g_2(x_n;\theta) \end{bmatrix}$. In the exactly identified case involving an unknown prevalence rate, the GMM estimator based on the above moment conditions will be identical to the pseudo-maximum likelihood

estimator $\tilde{\theta}_{\text{qunk}} = \begin{bmatrix} \tilde{\beta}_{\text{qunk}} \\ \tilde{q}_{\text{unk}} \end{bmatrix}$. Since the moment conditions are valid, the asymptotic covariance matrix for this estimator can be estimated, subject to the usual regularity conditions,[12] using the standard GMM formula:

$$V\left[\sqrt{N}\left(\tilde{\theta}-\theta\right)\right] \cong \left(G_{N,qunk}\left(x;\tilde{\theta}_{\text{qunk}}\right)' \tilde{S}^{-1}_{N,\text{qunk}} G_{N,\text{qunk}}\left(x;\tilde{\theta}_{\text{qunk}}\right)\right)^{-1}, \tag{32}$$

where $\tilde{S}_{N,\text{qunk}} = \frac{1}{N}\sum_{n=1}^{N} g\left(x_n;\tilde{\theta}_{\text{qunk}}\right) g\left(x_n;\tilde{\theta}_{\text{qunk}}\right)'$ and $G_{N,\text{qunk}}\left(x;\tilde{\theta}_{\text{qunk}}\right) = \begin{bmatrix} \frac{\partial g_{N_1}}{\partial \beta} & \frac{\partial g_{N_1}}{\partial q} \\ \frac{\partial g_{N_2}}{\partial \beta} & \frac{\partial g_{N_2}}{\partial q} \end{bmatrix}$ evaluated at $\theta = \tilde{\theta}_{\text{qunk}}$. Define the elements of the matrix $S_N = \frac{1}{N}\sum_{n=1}^{N} g\left(x_n;\theta\right) g\left(x_n;\theta\right)'$ as $\begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}$, where:

$$\begin{aligned}
s_{11} &= \frac{1}{N}\sum_{i=1}^{N_1} \frac{P'_\beta\left(x_i;\beta\right) P_\beta\left(x_i;\beta\right)}{P\left(x_i;\beta\right)^2} + \frac{1}{N}\left(\frac{N_1}{N_0 q}\right)^2 \sum_{j=1}^{N_0} P'_\beta\left(x_j;\beta\right) P_\beta\left(x_j;\beta\right) \\
s_{12} &= -\frac{1}{N}\left(\frac{N_1}{N_0 q}\right) \sum_{j=1}^{N_0} \left(q - P\left(x_j;\beta\right)\right) P'_\beta\left(x_j;\beta\right) \\
s_{21} &= -\frac{1}{N}\left(\frac{N_1}{N_0 q}\right) \sum_{j=1}^{N_0} \left(q - P\left(x_j;\beta\right)\right) P_\beta\left(x_j;\beta\right) \\
s_{22} &= \frac{1}{N}\sum_{j=1}^{N_0} \left(q - P\left(x_j;\beta\right)\right)^2.
\end{aligned} \tag{33}$$

Then $\tilde{S}_{N,\text{qunk}} = S_N$ evaluated at $\theta = \tilde{\theta}_{\text{qunk}}$. The elements of matrix $G_{N,\text{qunk}}\left(x;\theta\right)$ are defined as follows:

$$\begin{aligned}
\frac{\partial g_{N_1}}{\partial \beta} &= \frac{1}{N}\sum_{i=1}^{N_1} \frac{1}{P\left(x_i;\beta\right)}\left(\frac{\partial P\left(x_i;\beta\right)}{\partial \beta \partial \beta'} - \frac{P'_\beta\left(x_i;\beta\right) P_\beta\left(x_i;\beta\right)}{P\left(x_i;\beta\right)}\right) - \frac{1}{N}\left(\frac{N_1}{N_0 q}\right) \sum_{j=1}^{N_0} \frac{\partial P\left(x_j;\beta\right)}{\partial \beta \partial \beta'} \\
\frac{\partial g_{N_1}}{\partial q} &= \frac{1}{N}\left(\frac{N_1}{N_0 q^2}\right) \sum_{j=1}^{N_0} P'_\beta\left(x_j;\beta\right) \\
\frac{\partial g_{N_2}}{\partial \beta} &= -\frac{1}{N}\sum_{j=1}^{N_0} P_\beta\left(x_j;\beta\right) \\
\frac{\partial g_{N_2}}{\partial q} &= \frac{N_0}{N}.
\end{aligned} \tag{34}$$

For the case of a known prevalence rate, one can estimate β using the constrained pseudo-maximum likelihood estimator $\tilde{\beta}_{\text{qkn}}$ defined in Eq. (20) and rely on the following GMM formula for its estimated covariance matrix:

$$V\left[\sqrt{N}\left(\tilde{\beta}_{\text{qkn}}-\theta\right)\right] \cong \left(G_{N,\text{qkn}}\left(x;\tilde{\beta}_{\text{qkn}}\right)' \tilde{S}^{-1}_{N,\text{qkn}} G_{N,\text{qkn}}\left(x;\tilde{\beta}_{\text{qkn}}\right)\right)^{-1}, \tag{35}$$

where $\tilde{S}_{N,\text{qkn}} = \frac{1}{N}\sum_{n=1}^{N} g\left(x_n;\tilde{\beta}_{\text{qkn}}\right) g\left(x_n;\tilde{\beta}_{\text{qkn}}\right)'$ and $G_{N,\text{qkn}}\left(x;\tilde{\beta}_{\text{qkn}}\right) = \begin{bmatrix} \frac{\partial g_{N_1}}{\partial \beta} \\ \frac{\partial g_{N_2}}{\partial \beta} \end{bmatrix}$, evaluated at $\beta = \tilde{\beta}_{\text{qkn}}$.

12 These include, for example, rank, order, and uniqueness conditions for identification as well as conditions to ensure that the sample moments converge in probability to their expectation and have a finite asymptotic covariance matrix. See Hansen (1982) and Newey and McFadden (1994) for detailed discussions of the regularity conditions.

Alternatively, one can use the related GMM estimator of β that solves the following optimization problem:

$$\min_{\beta} g_N(x;\beta)' \tilde{S}_N^{-1} g_N(x;\beta), \tag{36}$$

where $\tilde{S}_N = \frac{1}{N}\sum_{n=1}^{N} g(x_n;\tilde{\beta})\, g(x_n;\tilde{\beta})'$ is an estimate of the asymptotic covariance matrix of $\sqrt{N} g_N(x;\beta)$ based on the initial consistent estimator $(\tilde{\beta})$ of β, obtained via unconstrained maximization of the pseudo-likelihood function defined in Eq. (24). The covariance matrix for this alternative estimator ($\tilde{\beta}_{\text{gmm}}$) may be estimated as:

$$V\left[\sqrt{N}(\tilde{\beta}_{\text{gmm}} - \theta)\right] \cong \left(G_N(x;\tilde{\beta}_{\text{gmm}})' \tilde{S}_{N,\text{gmm}}^{-1} G_N(x;\tilde{\beta}_{\text{gmm}})\right)^{-1}, \tag{37}$$

where $\tilde{S}_{N,\text{gmm}} = \frac{1}{N}\sum_{n=1}^{N} g(x_n;\tilde{\beta}_{\text{gmm}})\, g(x_n;\tilde{\beta}_{\text{gmm}})'$ and $G_N(x;\tilde{\beta}_{\text{gmm}}) = \begin{bmatrix} \frac{\partial g_{N1}}{\partial \beta} \\ \frac{\partial g_{N2}}{\partial \beta} \end{bmatrix}$, evaluated at $\beta = \tilde{\beta}_{\text{gmm}}$.

The two estimators for the case where the prevalence rate is known, $\tilde{\beta}_{\text{qkn}}$ and $\tilde{\beta}_{\text{gmm}}$, are asymptotically equivalent. Intuitively, as the sample size increases, the Lagrange multiplier μ in the first-order condition presented in Eq. (22) will converge to $\left(\frac{N_1}{N_0 q}\right)$; consequently, the two estimators will converge to the same estimator as the first-order conditions and the moment conditions become aligned and the solution to the moment conditions becomes more exact. For a formal proof of the asymptotic equivalence between a constrained discrete choice maximum likelihood estimator under choice-based sampling and a GMM estimator that relies on the score and constraint conditions from this problem as moments, refer to Imbens (1992).

The above formulas for the estimators and their covariance matrices are readily generalized to account for more complex sampling schemes by simply weighting the sample moment conditions:

$$g_{N_1,w}(x;\theta) = \frac{1}{N}\sum_{n=1}^{N} w_n \left(s_n \frac{P'_\beta(x_n;\beta)}{P(x_n;\beta)} - (1-s_n)\left(\frac{N_1}{N_0 q}\right) P'_\beta(x_n;\beta) \right) \tag{38}$$

$$g_{N_2,w}(x;\theta) = \frac{1}{N}\sum_{n=1}^{N} w_n (1-s_n)\left(q - P(x_n;\beta)\right), \tag{39}$$

where w_n represents the sample weight for observation n.

4 Incorporating Imperfect Knowledge About *q*

Existing research on choice-based sampling estimators for qualitative response problems has focused on the polar opposite cases of a known and an unknown prevalence rate. In practical applications, however, knowledge of the prevalence rate will often fall somewhere between these extremes. For instance, an estimate of the prevalence rate might be available based on a relatively small representative sample of participants and nonparticipants. Alternatively, a working estimate might be available based on informal projections, experience, and/or judgment.

Imbens and Lancaster (1994) have proposed a general approach for incorporating stochastic information regarding population parameters within a GMM framework through the introduction of additional moment conditions.[13] The uncertainty surrounding the true parameter values is then accounted for by assigning less weight to these conditions than would be assigned if the knowledge were certain. Below, we propose an adaptation of the Imbens–Lancaster methodology that permits the incorporation of imperfect information about the prevalence rate into our estimation framework.

13 I am grateful to the reviewer for suggesting that I explore the applicability of the Imbens-Lancaster methodology to this problem.

The GMM estimator for the case of a known prevalence rate ($\tilde{\beta}_{\text{gmm}}$) was derived in Section 3 by applying the sample moment conditions based on Eqs. (28) and (29) after setting q equal to q^*. When the value of a population parameter, such as the prevalence rate, is not known but is instead estimated based on an independent sample, Imbens and Lancaster propose incorporating this knowledge via a new moment condition: $E(m(x;\theta)) = 0$. The general formula for the covariance matrix of this constrained GMM estimator $\left(\hat{\theta}_{\text{gmm}}\right)$ is $V\left[\sqrt{N}\left(\hat{\theta}_{\text{gmm}} - \theta\right)\right] = \left(\Omega^{-1} + \Gamma_m' \Delta_m^{-1} \Gamma_m\right)^{-1}$, where Ω represents the asymptotic covariance matrix of the unconstrained estimator of θ, $\Gamma_m = \partial m(x;\theta)/\partial\theta'$, and $\Delta_m = E\left(m(x;\theta)\, m(x;\theta)'\right)$. The term $\Gamma_m' \Delta_m^{-1} \Gamma_m$ is referred to as the "gain in precision" due to the imposition of the constraint.

When the prevalence rate is estimated based on an independent sample of size M, one can account for this information by introducing an additional moment condition, $E\left(g_3(x;\beta)\right) = 0$, where

$$g_3(x;\beta) = (1-s)\left(q^* - P(x;\beta)\right). \tag{40}$$

So, whereas $g_3(x;\beta)$ replaces $g_2(x;\theta)$ in Eq. (29) and q^* replaces q in Eq. (28) when the prevalence rate is known to equal q^* with certainty, $E\left(g_3(x;\beta)\right) = 0$ is introduced as an additional moment condition when q^* instead represents an independent estimate of the prevalence rate.[14] One then obtains an estimator $\hat{\theta}_{\text{gmm}} = \begin{pmatrix} \hat{\beta}_{\text{gmm}} \\ \hat{q}_{\text{gmm}} \end{pmatrix}$ by solving the following GMM optimization problem:

$$\min_{\theta} g_N(x;\theta)' \, W_N g_N(x;\theta), \tag{41}$$

where the vector of sample moment conditions is now defined as $g_N(x;\theta) = \begin{bmatrix} g_{N_1}(x;\theta) \\ g_{N_2}(x;\theta) \\ g_{N_3}(x;\theta) \end{bmatrix}$. The new weight matrix is specified as $W_N = \begin{bmatrix} \tilde{S}_N^{-1} & 0 \\ 0 & \omega(\tilde{\beta}) \end{bmatrix}$. The term $\tilde{S}_N^{-1}$ is equal to the inverse of the value of S_N, as defined in Eq. (33), evaluated using a consistent initial estimator of β ($\tilde{\beta}$) and the independently estimated value of the prevalence rate. The term $\omega(\tilde{\beta})$ represents the weight assigned to the new sample moment condition based on a consistent estimate of β. Under the assumption that (M/N) converges to a constant as N grows large, Imbens and Lancaster show that the optimal value for $\omega(\beta)$ is the inverse of $\left(\frac{N_0}{M}\Delta_{g_3}\right)$, where $\Delta_{g_3} = E\left(g_3(x;\beta)\, g_3(x;\beta)'\right)$.[15] A feasible estimate of the optimal weight $\left(\omega(\tilde{\beta})\right)$ is used for estimation, which is computed by replacing Δ_{g_3} with the following estimate based on the supplementary sample: $\tilde{\Delta}_{g_3} = \sum_{j=1}^{N_0}\left(q^* - P(x_j;\tilde{\beta})\right)^2 / N$. The asymptotic covariance matrix of the resulting constrained estimator $\left(\hat{\theta}_{\text{gmm}} = \begin{pmatrix} \hat{\beta}_{\text{gmm}} \\ \hat{q}_{\text{gmm}} \end{pmatrix}\right)$ will be equal to:

$$V\left[\sqrt{N}\left(\tilde{\theta}_{\text{gmm}} - \theta\right)\right] = \left(\Omega^{-1} + \begin{bmatrix} \Gamma_{g_3}' \Delta_{g_3} \Gamma_{g_3} & 0 \\ 0 & 0 \end{bmatrix}\right)^{-1}, \tag{42}$$

where $\Omega^{-1} = V\left[\sqrt{N}\left(\tilde{\theta}_{\text{qunk}} - \theta\right)\right]^{-1}$ is the inverse of the asymptotic covariance of the unconstrained estimator $\tilde{\theta}_{\text{qunk}}$ and $\Gamma_{g_3} = -\frac{1}{N}\sum_{j=1}^{N_0} P_\beta(x_j;\beta)$. The expression $\Gamma_{g_3}' \Delta_{g_3} \Gamma_{g_3}$ reflects the gain in precision of $\hat{\beta}_{\text{gmm}}$ over $\tilde{\beta}_{\text{qunk}}$.

---

14 This sample need not include covariates relevant to the participation decision, so long as participation status is recorded. The researcher does not require direct access to the sample so long as the sample proportion of participants and sample size are available.

15 When the size of the independent sample used to estimate the prevalence rate (M) is small (large) relative to the size of the supplementary estimation sample (N_0), less (more) weight will be placed on the moment condition to reflect the difference in precision of an estimate of q based on a sample of size M rather than a supplementary sample of size N_0.

The estimated covariance matrix of the constrained GMM estimator is obtained from the following equation:

$$V\left[\sqrt{N}\left(\hat{\theta}_{\text{gmm}}-\theta\right)\right] \cong \left(G_N\left(x;\hat{\theta}_{\text{gmm}}\right)' \hat{W}_{N,\text{gmm}} G_N\left(x;\hat{\theta}_{\text{gmm}}\right)\right)^{-1}, \tag{43}$$

where $\hat{W}_{N,\text{gmm}} = \begin{bmatrix} S_N^{-1} & 0 \\ 0 & \omega(\theta) \end{bmatrix}$ and $G_N\left(x;\hat{\theta}_{\text{gmm}}\right) = \begin{bmatrix} \frac{\partial g_{N1}}{\partial \beta} & \frac{\partial g_{N1}}{\partial q} \\ \frac{\partial g_{N2}}{\partial \beta} & \frac{\partial g_{N2}}{\partial q} \\ \frac{\partial g_{N2}}{\partial \beta} & 0 \end{bmatrix}$, both evaluated at $\hat{\theta}_{\text{gmm}}$. The elements of $G_N(x;\theta)$ are defined in Eq. (34).

If q^* is not based on an independent representative sample, one can specify $\omega(\theta)$ based on one's perceived level of confidence in the estimate. At one extreme, when the value of this weight is set close to zero, the value of the constrained estimate $\left(\hat{\theta}_{\text{gmm}}\right)$ will approach the value of the unconstrained estimate $\left(\tilde{\theta}_{\text{qunk}}\right)$. Towards the other extreme, when the value of the weight is increased to a very high value, the value of the constrained estimate will approach the value of the estimate associated with a fully known prevalence rate $\left(\tilde{\theta}_{\text{gmm}}\right)$.

5 Considerations for Applying the New Estimators

As emphasized in Section 1, use-availability sampling is a viable strategy for estimating conditional response probabilities in a variety of circumstances where there is an abundance of information on participants, but a relative dearth of information on non-participants. In practice, however, there have been very few empirical applications of this strategy to estimate the drivers of participation (or presence) when standard qualitative choice estimation techniques are not feasible. Hopefully, the new estimators introduced in this paper, which readily incorporate sample weights from relatively complex sampling designs, will generate renewed interest in use-availability sampling strategies for qualitative choice analysis.

5.1 Representativeness and Compatibility of Data Sources

An important consideration when selecting data sources for implementing a use-availability model of participation is that they need to be representative of the underlying populations of interest. The supplementary sample must be representative of the overall population of participants and non-participants, while the primary sample must be representative of the sub-population of participants. In the case where stratified or clustered samples are available, sample weights must be available to make the samples representative, and these weights must be applied when estimating the models.

Many social benefit programs are means-tested. When selecting a supplementary sample to study the drivers of participation in such programs, it is important to restrict the sample to individuals who are eligible for the benefit. It is therefore important for the supplementary data source to include reliable measures of the income concepts, family characteristics, and other relevant variables underlying the program eligibility criteria.

Another consideration is that the relevant covariates for the analysis need to be present in both the primary and supplementary samples, and they must be comparably measured across the two data sources. It is essential that the differences in covariate values across data sets reflect the underlying population differences rather than conceptual discrepancies or systematic measurement errors.

5.2 Precision

For a fixed overall sample size, use-availability estimates will be less precise than standard qualitative choice model estimates. However, in the case a low prevalence rate, the supplementary sample under a

use-availability design will consist largely of non-participants, while the primary sample will consist entirely of participants, similar to the overall estimation sample used for a traditional qualitative choice analysis. In such cases, the relative difference in precision will be more modest. In the case of a higher prevalence rate, this disadvantage will be more substantial, especially in the absence of any knowledge regarding the prevalence rate. Nevertheless, the use-availability approach serves as a viable estimation strategy in a variety of circumstances where standard qualitative choice model estimation is not feasible. Moreover, in some cases, the use-availability estimation strategy will permit estimation based on a much larger overall sample, thereby overcoming its disadvantage with regard to estimator precision. For instance, Erard et al. (2020) were able to employ a large primary sample of income tax filers from IRS records and a large sample of filers and nonfilers from the Current Population Survey to obtain reasonably precise calibrated probit estimates of the drivers of filing compliance, despite prevalence rates (filing rates) well in excess of 90 percent.

6 Conclusions

Frequently, researchers have access to detailed information on the relevant characteristics of participants in a program. However, the lack of comparable information on non-participants precludes the application of standard qualitative response models to examine the drivers of participation. A feasible alternative approach is to supplement the data on participants with a representative sample of observations from the general population of participants and nonparticipants. This paper presents some new qualitative choice estimators for a such a sampling scheme that rival the performance of the best existing estimators. An important advantage of these new estimators is that they are more readily generalized for application under relatively complex use-availability designs; all they require is a set of sample weights to make the samples representative of their underlying populations. These new estimators therefore significantly broaden the scope of potential data sources that can be used to estimate conditional response probabilities.

In the case where the prevalence rate is unknown, both the new and existing estimators are less precise and are subject to periodic convergence problems, particularly when q is fairly close to either of its boundaries (0 or 1). These problems can be alleviated either by using a larger estimation sample or by incorporating imperfect knowledge regarding the prevalence rate using the methodology described in this paper.

Acknowledgements: Research support provided by the Internal Revenue Service under contracts TIRNO-10-D-00021-D0004, TIRNO-14-P-00157, and TIRNO-15-P-00172 is gratefully acknowledged. The views expressed in this paper are my own and do not necessarily reflect the opinions of the IRS. I thank Stephen Cosslett, Subhash Lele, and the anonymous reviewers for their extremely helpful comments and suggestions. I am also grateful to John Guyton, Patrick Langetieg, Mark Payne, and Alan Plumley for helping me to refine my methodology as we worked on applying the approach to understand the determinants of taxpayer filing behavior.

Appendix A: Consistency of Maximum Likelihood Estimators

The consistency of the new quasi-maximum likelihood estimators presented in this paper follows directly from the proofs of consistency provided by Manski and McFadden (1981, pp. 36–45) for the qualitative choice estimators they constructed for application with a purely choice-based sampling design. We maintain their assumption that certain regularity conditions are satisfied (Assumptions 1.1–1.5, p. 12) and we rely on the same lemmas provided in that study (pp. 37–38):

Lemma 1.1. *Let* $f_N(x,\phi)$, $N = 1,\ldots,\infty$, *be a sequence of measurable functions on a measurable space X and for each* $x \in X$, *a continuous function for* $\phi \in \Phi$, Φ *being compact. Then there exists a sequence of measurable functions* $\phi_N(x)$, $N = 1,\ldots\infty$, *such that* $f_N(x,\phi_N(x)) = \sup_{\phi\in\Phi} f_N(x,\phi)$ *for all* $x \in X$ *and* $N = 1,\ldots,\infty$.

Furthermore, if for almost every $x \in X$, $f_N(x, \phi)$ converges to $f(\phi)$ uniformly for all $\phi \in \Phi$, and if $f(\phi)$ has a unique maximum at $\phi^* \in \Phi$, then ϕ_n converges to ϕ^* for almost every $x \in X$.

Lemma 1.2. *Let μ be a probability measure over a Euclidean space S, let Φ be a compact subset of a Euclidean space, and let gs, ϕ be a continuous function of ϕ for each $s \in S$ and a measurable function of s for each $\phi \in \Phi$. Assume also that $|g(s, \phi)| \leq \alpha(s)$ for all s, ϕ, and some μ–integrable α. For any sequence $x = s_1 s_2, \ldots$, let $f_N(x, \phi) = \sum_{n=1}^{N} g(s_n, \phi)/N$, and let X be the set of all sequences x. If sequences x are drawn as random samples from S, then for almost every realized such sequence, as $N \to \infty$, $f_N(x, \phi) \to E(g(x, \phi)) \equiv f(\phi)$ uniformly for all $\phi \in \Phi$.*

Lemma 1.3 . *Let $g(s, \phi)$ be a real valued function over a space S x Φ such that g is integrable with respect to a measure μ over S and $g(s, \phi) \geq 0$, all $s \in S$, $\phi \in \Phi$. Let ϕ^* be an element of Φ such that $g(s, \phi^*) > 0$ for almost every $s \in S$ and $\int_S (g(s, \phi^*) - g(s, \phi))\, d\mu \geq 0$, all $\phi \in \Phi$. Then the expression $f(\phi) = \int_S g(s, \phi^*) \ln(g(s, \phi))\, d\mu$ attains its maximum at $\phi = \phi^*$. The maximum is unique if, for every $\phi \in \Phi$ such that $\phi \neq \phi^*$, there exists an $S_\phi \subset S$ such that*

$$\int_{S_\phi} g(s, \phi)\, d\mu \neq \int_{S_\phi} g(s, \phi^*)\, d\mu.$$

A.1 Estimators for a known covariate distribution

For the estimator described in Eq. (4) for the case of an unknown prevalence rate, let $f_N(x; \beta) = \frac{1}{N_1} \sum_{n=1}^{N_1} \ln\left(\frac{P(x_n;\beta)}{\int P(z;\beta)\, \mathrm{d}F(z)}\right)$ for a random primary sample $x_p = x_1, x_2, \ldots, x_{N_1}$, and allow β^* to represent the true value of β. It follows from Lemma 1.2 that, as $N_1 \to \infty$, $f_N(x; \beta) \xrightarrow{\text{a.s.}} E\left[\ln\left(\frac{P(x;\beta)}{\int P(z;\beta)\, \mathrm{d}F(z)}\right)\right] = \frac{P(x;\beta^*)}{\int P(z;\beta^*)\, \mathrm{d}F(z)} \ln\left(\frac{P(x;\beta)}{\int P(z;\beta)\, \mathrm{d}F(z)}\right) = f(\beta)$ uniformly in β. From Lemma 1.3, it further follows that $f(\beta)$ is uniquely maximized at $\beta = \beta^*$. Finally, since $f_N(x; \beta)$ converges to $f(\beta)$ uniformly in β and has a unique optimum at β^*, the estimator $\tilde{\beta}$ based on Eq. (4) is consistent in accordance with Lemma 1.1. The consistency of $\tilde{q} = \int P(x; \tilde{\beta})\, \mathrm{d}F(x)$ follows directly from the consistency of $\tilde{\beta}$. Since the estimator defined in Eq. (5) for the case of a known prevalence rate is a constrained version of the estimator $\tilde{\beta}$, its consistency is assured by the consistency of $\tilde{\beta}$.

A.2 Estimators for an Unknown Covariate Distribution

For the estimator described in Eq. (14) for an unknown prevalence rate, let $f_N(x; \beta) = \frac{1}{N_1} \sum_{n=1}^{N_1} \ln\left(\frac{P(x_n;\beta)}{\frac{1}{N_0}\sum_{m=1}^{N_0} P(x_m;\beta)}\right)$ for the randomly selected primary ($x_\mathrm{p} = x_1, x_2, \ldots, x_{N_1}$) and supplementary ($x_\mathrm{s} = x_1, x_2, \ldots, x_{N_0}$) samples. This function may be re-expressed as: $f_n(x; \beta) = \frac{1}{N_1}\left(\sum_{n=1}^{N_1} \ln\left(\frac{P(x_n;\beta)}{\int P(z;\beta)\, \mathrm{d}F(z)}\right)\right) + \ln\left(\frac{\int P(z;\beta)\, \mathrm{d}F(z)}{\frac{1}{N_0}\sum_{m=1}^{N_0} P(x_m;\beta)}\right)$. By Lemma 1.2, the first term of this function converges to $\frac{P(x;\beta^*)}{\int P(z;\beta^*)\, \mathrm{d}F(z)} \ln\left(\frac{P(x;\beta)}{\int P(z;\beta)\, \mathrm{d}F(z)}\right) = f(\beta)$ uniformly in β as $N_1 \to \infty$. In addition, the second term converges to zero as $N_0 \to \infty$; the denominator of the expression in parentheses is a consistent estimator of the numerator (i.e., the prevalence rate), so that the natural log of the expression tends to zero as the supplementary sample size increases. Therefore, function $f_n(x; \beta)$ converges uniformly to $f(\beta)$ as the overall sample size ($N = N_0 + N_1$) increases and the ratio (N_1/N) is held fixed. Lemma 1.3 further implies that $f(\beta)$ is uniquely maximized at $\beta = \beta^*$. Since $f_N(x; \beta)$ converges to $f(\beta)$ uniformly in β and has a unique optimum at β^*, it follows from Lemma 1.1 that the estimator $\tilde{\beta}_{\text{qunk}}$ based on Eq. (14) is consistent. Since $\sum_{j=1}^{N_0} P(x_j; \beta)/N_0$ converges to q and $\tilde{\beta}_{\text{qunk}}$ converges to β, it follows that $\tilde{q}_{\text{unk}} = \sum_{j=1}^{N_0} P(x_j; \tilde{\beta}_{\text{qunk}})/N_0$ is a consistent estimator of q. The estimator $\tilde{\beta}_{\text{qknown}}$ presented in Eq. (20) for the

case of a known prevalence rate is a constrained version of the estimator $\tilde{\beta}_{\text{qunk}}$. Therefore, its consistency is assured by the consistency of $\tilde{\beta}_{\text{qunk}}$.

Appendix B: Monte Carlo Analysis

We have undertaken a Monte Carlo analysis to compare the small sample performance of our new estimators against the Cosslett (1981) and Lancaster and Imbens (1996) estimators. In the simulations, the conditional probability of participation is described by a logistic function of two independent standard normal regressors and an intercept. The coefficients of the two regressors are fixed at one, while the intercept is varied to achieve alternative approximate values of the prevalence rate q, including 0.125, 0.25, 0.50, 0.75, and 0.875. The Bernoulli sampling scheme described in Section 3.1.1 is used to draw the primary and supplementary samples. For each experiment, 1000 replications are completed and the following statistics are reported for each parameter: mean and median parameter estimate; mean asymptotic standard deviation (ASD) of the parameter estimate; standard deviation of the parameter estimates (SSD); and the median absolute deviation from the median (MAD) parameter estimate. For each replication, the combined sample size (N) is set to 600, and the Bernoulli parameter (h) is fixed at 0.50.

B.1 Known prevalence rate

For the known prevalence rate case, we compare the small sample performance of the calibrated logit estimator $\left(\tilde{\beta}_{\text{qkn}}\right)$ defined in Eq. (20) and the associated GMM estimator obtained based on the optimization problem described in Eq. (36) against the performances of two benchmark estimators. The first of these benchmarks is motivated by the choice-based estimation framework of Cosslett (1981), and it is obtained as the solution to the following optimization problem:

$$\max_{\beta} \min_{\lambda_1} \sum_{n=1}^{N} \left(s_n \ln\left(P(x_n;\beta)\right) - \ln\left(\lambda_1 P(x_n;\beta) + 1 - \lambda_1 q\right)\right), \tag{44}$$

where λ_1 is a weight factor that is estimated jointly with β. We refer to this estimator as the "Cosslett" estimator in our Monte Carlo simulations. Observe that the solution for this estimator is at a saddle point of the objective function in Eq. (44).

The second benchmark estimator is the one proposed by Lancaster and Imbens (1996) for the case of a known prevalence rate. This estimator is obtained by applying GMM estimation based on the following three moment conditions:

$$E\left(\frac{P'_{\beta}(x;\beta)}{P(x;\beta)}\left(s - R(x;\beta,q,h)\right)\right) = 0 \tag{45}$$

$$-\frac{1}{q}E\left(s - R(x;\beta,q,h)\right) = 0 \tag{46}$$

$$E\left(h - R(x;\beta,q,h)\right) = 0, \tag{47}$$

where $R(x;\beta,q,h)$ was previously defined in Section 3.1.1.

The Monte Carlo simulation results are summarized in Table 1. All of the estimators perform similarly. For prevalence rates below 75 percent, the estimators show little sign of bias. At higher prevalence rates, a modest degree of upward bias is present in each of the intercept estimates, but the slope estimates remain essentially unbiased. The precision of each of the estimators deteriorates as the prevalence rate increases. Presumably, this is because the contrast between the primary and supplementary samples is less sharp at high prevalence rates, owing to the high shares of participants in both samples. Overall, the estimators all perform well and exhibit comparable levels of precision.

Table 1: Monte Carlo simulation results, prevalence rate known.

	Cosslett			Lancaster–Imbens			Calibrated logit			GMM alternative		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
						$q = 0.125, N = 600, h = 0.50$						
Actual	−2.574	1.00	1.00	−2.574	1.00	1.00	−2.574	1.00	1.00	−2.574	1.00	1.00
Mean	−2.688	1.00	0.99	−2.570	1.00	1.00	−2.574	1.02	1.01	−2.574	1.02	1.01
Median	−2.685	1.00	0.99	−2.566	1.00	1.00	−2.572	1.01	1.01	−2.571	1.01	1.01
ASD	0.081	0.13	0.13	0.080	0.13	0.13	0.086	0.15	0.15	0.085	0.15	0.15
SSD	0.118	0.13	0.14	0.085	0.13	0.15	0.087	0.15	0.16	0.087	0.15	0.16
Mad	0.080	0.09	0.10	0.057	0.09	0.10	0.059	0.10	0.11	0.059	0.10	0.11
						$q = 0.25, N = 600, h = 0.50$						
Actual	−1.492	1.00	1.00	−1.492	1.00	1.00	−1.492	1.00	1.00	−1.492	1.00	1.00
Mean	−1.530	1.01	1.01	−1.490	1.00	1.00	−1.488	1.02	1.02	−1.489	1.02	1.02
Median	−1.522	1.00	1.00	−1.487	1.00	1.00	−1.486	1.01	1.01	−1.486	1.01	1.01
ASD	0.065	0.16	0.16	0.064	0.15	0.15	0.067	0.17	0.17	0.066	0.17	0.17
SSD	0.087	0.16	0.16	0.065	0.17	0.17	0.067	0.17	0.18	0.066	0.17	0.18
Mad	0.055	0.10	0.12	0.045	0.11	0.12	0.045	0.11	0.12	0.045	0.11	0.12
						$q = 0.50, N = 600, h = 0.50$						
Actual	0.00	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00
Mean	0.017	1.02	1.03	0.013	1.00	1.01	0.021	1.01	1.02	0.018	1.01	1.02
Median	0.007	1.01	1.02	−0.000	0.99	1.00	0.006	1.00	1.01	0.004	1.00	1.01
ASD	0.081	0.22	0.22	0.079	0.21	0.21	0.087	0.23	0.23	0.087	0.23	0.23
SSD	0.084	0.22	0.23	0.084	0.23	0.23	0.096	0.23	0.24	0.091	0.23	0.23
Mad	0.052	0.15	0.15	0.050	0.15	0.15	0.056	0.15	0.15	0.055	0.15	0.15
						$q = 0.75, N = 600, h = 0.50$						
Actual	1.492	1.00	1.00	1.492	1.00	1.00	1.492	1.00	1.00	1.492	1.00	1.00
Mean	1.591	1.05	1.07	1.568	1.01	1.03	1.583	1.03	1.06	−1.568	1.02	1.04
Median	1.549	1.03	1.05	1.534	1.00	1.02	1.536	1.01	1.04	−1.523	0.99	1.02
ASD	0.271	0.37	0.37	0.255	0.36	0.36	0.268	0.37	0.37	0.265	0.36	0.37
SSD	0.281	0.39	0.38	0.288	0.42	0.41	0.284	0.39	0.38	0.270	0.39	0.38
Mad	0.170	0.25	0.25	0.180	0.28	0.26	0.170	0.26	0.25	0.164	0.26	0.25
						$q = 0.875, N = 600, h = 0.50$						
Actual	2.574	1.00	1.00	2.574	1.00	1.00	2.574	1.00	1.00	2.574	1.00	1.00
Mean	2.870	1.04	1.07	2.790	0.96	0.98	2.853	1.01	1.04	−2.810	1.00	1.03
Median	2.733	1.03	1.05	2.642	0.94	0.94	2.716	1.03	1.03	−2.653	0.99	0.99
ASD	0.722	0.68	0.72	0.598	0.78	0.83	0.612	0.59	0.61	0.613	0.61	0.64
SSD	0.633	0.65	0.68	0.717	0.70	0.74	0.606	0.68	0.71	0.694	0.65	0.70
Mad	0.338	0.42	0.40	0.434	0.49	0.46	0.345	0.42	0.41	0.347	0.42	0.42

B.2 Unknown Prevalence Rate

For the case of an unknown prevalence rate, the simulations compare the small sample performances of our new estimator of β and q based on Eq. (14) and the Cosslett–Lancaster–Imbens estimator described in Section 3.1.1. The results are summarized in Table 2. Both estimators are subject to periodic convergence problems in small samples. We report the performance statistics for a given estimator based on the subset of replications that are free from convergence issues. The number of replications for which an estimator has failed to converge is reported as "#Failures".

Overall, our pseudo-maximum likelihood estimator performs comparably to the Cosslett–Lancaster–Imbens estimator in terms of mean and median performance as well as precision. As with the

Table 2: Monte Carlo simulation results, prevalence rate unknown.

	Cosslett–Lancaster–Imbens				Pseudo-MLE			
	β_0	β_1	β_2	q	β_0	β_1	β_2	q
				$q = 0.125$, $N = 600$, $h = 0.50$				
Actual	−2.574	1.00	1.00	0.125	−2.574	1.00	1.00	0.125
Mean	−2.667	1.04	1.04	0.14	−2.505	1.08	1.08	0.15
Median	−2.525	1.03	1.03	0.14	−2.423	1.08	1.08	0.15
ASD	1.460	0.20	0.20	0.09	4.633	0.82	0.82	0.36
SSD	0.819	0.19	0.18	0.07	0.653	0.20	0.18	0.07
Mad	0.452	0.13	0.12	0.05	0.384	0.12	0.12	0.05
#Failures		99				297		
				$q = 0.25$, $N = 600$, $h = 0.50$				
Actual	−1.492	1.00	1.00	0.25	−1.492	1.00	1.00	0.25
Mean	−1.602	1.06	1.05	0.25	−1.558	1.07	1.06	0.25
Median	−1.474	1.04	1.04	0.25	−1.441	0.05	1.05	0.26
ASD	1.104	0.24	0.24	0.11	4.00	0.90	0.89	0.48
SSD	0.735	0.24	0.24	0.10	0.667	0.24	0.23	0.09
Mad	0.389	0.16	0.16	0.06	0.374	0.16	0.15	0.06
#Failures		26				123		
				$q = 0.50$, $N = 600$, $h = 0.50$				
Actual	0.000	1.00	1.00	0.50	0.000	1.00	1.00	0.50
Mean	0.032	1.09	1.09	0.48	0.002	1.08	1.09	0.48
Median	0.041	1.04	1.04	0.50	0.048	1.04	1.04	0.50
ASD	0.934	0.37	0.37	0.15	3.295	1.01	1.02	0.56
SSD	0.846	0.39	0.42	0.13	0.839	0.38	0.40	0.13
Mad	0.463	0.22	0.21	0.08	0.451	0.22	0.21	0.08
#Failures		13				53		
				$q = 0.75$, $N = 600$, $h = 0.50$				
Actual	1.492	1.00	1.00	0.75	1.492	1.00	1.00	0.75
Mean	1.951	1.28	1.25	0.72	1.832	1.24	1.21	0.71
Median	1.558	1.06	1.09	0.75	1.560	1.06	1.08	0.74
ASD	2.478	0.88	0.87	0.20	5.185	1.53	1.53	0.75
SSD	1.991	0.88	0.85	0.15	1.801	0.82	0.74	0.15
Mad	0.782	0.36	0.37	0.08	0.789	0.38	0.38	0.08
#Failures		86				89		
				$q = 0.875$, $N = 600$, $h = 0.50$				
Actual	2.574	1.00	1.00	0.875	2.574	1.00	1.00	0.875
Mean	3.339	1.33	1.35	0.81	3.496	1.38	1.39	0.81
Median	2.801	1.13	1.08	0.87	2.867	1.13	1.09	0.87
ASD	3.889	1.16	1.20	0.26	7.146	2.01	1.96	0.79
SSD	2.859	1.11	1.25	0.16	3.126	1.26	1.29	0.17
Mad	1.391	0.55	0.55	0.06	1.456	0.58	0.57	0.07
#Failures		244				191		

estimators for the case of a known prevalence rate, the current estimators show a deterioration in precision as the prevalence rate increases, along with an upwardly biased intercept estimate when the prevalence rate is relatively high ($q = 0.75$ and $q = 0.875$). However, this upward bias extends to the slope coefficients as well for the case of an unknown prevalence rate. Overall, precision suffers when the prevalence rate is unknown, especially in the case of the intercept estimate.

Lancaster and Imbens (1996) have reported that their estimator has periodic convergence issues in small samples, particularly when the true value of q is close to zero. This problem extends to our estimator. As noted by Lancaster and Imbens, when q is close to zero, supplementary sampling is close to pure choice-based sampling, and the choice-based sampling estimator of the intercept in a logit model is not identified when q is unknown. Our simulation results indicate that convergence problems are also prevalent when the true value of q is relatively high ($q = 0.75$ and $q = 0.875$). Convergence problems can be alleviated by employing larger estimation samples or by incorporating imperfect knowledge regarding the prevalence rate using the methodology described in Section 4.

References

Breslow, N. E. 1996. "Statistics in Epidemiology: The Case-Control Study." *Journal of the American Statistical Association* 91 (433): 14–28.

Cosslett, S. R. 1981. "Efficient Estimation of Discrete Choice Models." In *Structural Analysis of Discrete Data with Econometric Applications*, edited by C. Manski, and D. McFadden, 51–111. Cambridge: MIT Press.

Erard, B., P. Langetieg, M. Payne, and A. Plumley. 2020. "Flying under the Radar: Ghosts and the Income Tax." *CESifo Economic Studies* 66 (3): 185–97.

Hansen, L. P. 1982. "Large Sample Properties of Generalized Method of Moment Estimators." *Econometrica* 50 (4): 1029–54.

Imbens, G. W. 1992. "An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-Based Sampling." *Econometrica* 60 (5): 1187–214.

Imbens, G. W., and T. Lancaster. 1994. "Combining Micro and Macro Data in Microeconometric Models." *The Review of Economic Studies* 61 (4): 655–80.

Keating, K. A., and S. Cherry. 2004. "Use and Interpretation of Logistic Regression in Habitat Selection Studies." *Journal of Wildlife Management* 68 (4): 774–89.

Lancaster, T., and G. Imbens. 1996. "Case Controlled Studies with Contaminated Controls." *Journal of Econometrics* 71: 145–60.

Lele, S. R., and J. L. Keim. 2006. "Weighted Distributions and Estimation of Resource Selection Probability Functions." *Ecology* 87 (12): 3021–8.

Manski, C. F., and D. McFadden. 1981. "Alternative Estimators and Sample Designs for Discrete Choice Analysis." In *Structural Analysis of Discrete Data with Econometric Applications*, edited by C. Manski, and D. McFadden, 2–49. Cambridge: MIT Press.

Newey, W. K., and D. McFadden. 1994. "Large Sample Estimation and Hypothesis Testing." In *Handbook of Econometrics*, Vol. 4, edited by R. F. Engle, and D. McFadden, 2111–245. Amsterdam: Elsevier.

Phillips, S. J., and J. Elith. 2013. "On Estimating Probability of Presence from Use-Availability or Presence-Background Data." *Ecology* 94 (6): 1409–19.

Rosenman, R., S. Goates, and L. Hill. 2012. "Participation in Universal Prevention Programs." *Applied Economics* 44 (2): 219–28.

Royle, J. A., R. B. Chandler, C. Yackulic, and J. D. Nichols. 2012. "Likelihood Analysis of Species Occurrence Probability from Presence-Only Data for Modelling Species Distributions." *Methods in Ecology and Evolution* 3: 545–54.

Solymos, P., and S. R. Lele. 2016. "Revisiting Resource Selection Probability Functions and Single-Visit Methods: Clarifications and Extensions." *Methods in Ecology and Evolution* 7 (2): 196–205.

Steinberg, D., and N. S. Cardell. 1992. "Estimating Logistic Regression Models when the Dependent Variable has no Variance." *Communications in Statistics - Theory and Methods* 21 (2): 423–50.