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

Erard, Brian

B. Erard Associates, LLC

24 June 2017

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

Brian Erard
B. Erard & Associates, LLC
Reston, VA

Revised April 25, 2020

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. New estimators are introduced that exploit the parameter restrictions implied by the relationship between the marginal and conditional response probabilities in the supplementary sample. An important advantage of these estimators over the existing alternatives is that they can be applied to exogenously stratified samples even when the underlying stratification criteria are unknown. The ability of these new estimators to readily incorporate sample weights make them applicable to a much wider range of data sources. The new estimators are also easily generalized to address polychotomous outcomes.

Key words: Qualitative response, Discrete choice, Choice-based sampling, Supplementary sampling, Contaminated controls

JEL Classification: C13; C25; C35

* CONTACT Brian Erard, Brian@BrianErard.com, 2350 Swaps Ct., Reston, VA 20191-2630.

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.

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".[2]

The existing literature on qualitative response estimation under this sampling scheme (Cosslett, 1981; 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. This paper shows that the extension of the methods developed in these studies to permit estimation using exogenously stratified random samples requires detailed knowledge of the sample design. In many cases, however, such information is not available. Rather, only the sample weights are made public. For instance, researchers may be interested in using a general survey of the overall population as a supplementary data source. In the U.S., a few examples include the Current Population Survey (CPS), the Survey of Income and Program Participation (SIPP), and the American Community Survey (ACS).[3] Often, however, such data sources are derived using complex sampling designs and specific design details, such as the stratification criteria, are not made available to the public. As a consequence, it is not feasible to apply the estimation approaches developed by these authors to such data sources.[4]

In this paper, we present some new estimators that can be applied to stratified samples even when the underlying details of the sampling design are not available; all that is required are the sample weights, which are routinely available. We develop separate estimators for the cases

[2] 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. (2016) [tax compliance]; and Rosenman, Goates, and Hill (2012) [substance abuse prevention programs].

[3] If eligibility for a program or service is limited, one may be able to restrict the supplementary sample to include only those survey respondents who are eligible, providing that eligibility can be deduced from the survey information that has been collected. For instance, the CPS has detailed income information that can be useful in assessing eligibility for means-tested programs and services.

[4] Even if the full details of the sampling design were made available to researchers, it would be difficult to adapt the Cosslett and Lancaster-Imbens estimators for application with the complex sampling designs employed in many national surveys, which typically involve multi-stage sampling, clustering, and post-stratification adjustment.

in which the prevalence rate (i.e., the overall take-up rate in the case of a program, the percentage of habitats occupied by a species in the case of wildlife presence, or the share of the population that is infected in the case of a disease), is and is not known. These new estimators are derived using an approach similar to that used in earlier work by Cosslett (1981) and Lancaster and Imbens (1996). The key difference is that the derivation relies on the empirical distribution of the covariates in the supplementary sample rather than their empirical distribution in the combined (primary and supplementary) sample.

We perform Monte Carlo simulations involving unstratified random samples to compare the small sample performance of our new estimators against other existing estimators. We find that the performance of our estimators for both the known and unknown prevalence rate cases rivals that of the best existing estimators (Cosslett, 1981, and Lancaster and Imbens, 1996). We further show that our new estimators are easily generalized to address polychotomous response problems. As an illustration of this generalization, we estimate a multinomial logit specification of voting behavior using stratified primary and supplementary data samples that were respectively drawn from the CPS and the ACS.

2. Known covariate distribution

Using the notation of Lancaster and Imbens (1996), let y be a binary response variable equal to 1 (for participation/presence) or 0 (for non-participation/non-presence) and let x represent a vector of attributes with cumulative distribution function $F(x)$. We assume that the conditional probability that y is equal to 1 given x follows a known parametric form,

$\Pr(y = 1|x;\beta) = P(\mathrm{x};\beta)$, 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 in the population) as:

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

2.1 *Identification*

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

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

where $f(x)$ represents the joint marginal p.d.f. of x $\left[f(x) = \frac{dF(x)}{dx}\right]$. If $f(x)$ is known, it follows from Equation (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 probability $P(x;\beta)/P(y;\beta)$ continues to be nonparametrically identified. However, identification of β in this case depends on the parametric specification of the conditional response probability. For certain specifications, it is not possible to separately identify all of the elements of β. For instance, under a linear probability model, $\frac{P(x;\beta_0,\beta_1)}{q} = \left(\frac{\beta_0}{q}\right) + \left(\frac{\beta_1}{q}\right)' x$. Therefore, 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, $\ln\left(\frac{P(x;\beta_0,\beta_1)}{q}\right) = (\beta_0 - lnq) + \beta_1{}'x$. In this case, each of the slope coefficients of the conditional response probability is identified, but the intercept is not.[5]

[5] 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

Fortunately, the above two cases are exceptional. As discussed by Solymos and Lele (2016), all of the elements of β are identified under most qualitative choice specifications, including the logit, probit, arctan, and complementary log-log models, so long as the specification includes at least one continuous covariate. Nevertheless, the above examples involving the linear and log-linear probability models do raise concerns about the possible consequences of relying on assumed functional forms to identify certain parameters of the conditional response probability function when the prevalence rate is unknown. 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.

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.[6]

Consider first the case where the prevalence rate q is unknown. In this case, one can estimate β by solving the following unconstrained maximum likelihood estimation problem:

---

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.

[6] Consistency of the estimators we present for this case follows from the proofs provided by Manski and McFadden (1981) for the estimators they have reviewed for a pure choice-based sampling design. Under the assumptions they present on pp. 12-13, the consistency of the estimator we present in Equation (3) for the case where the prevalence rate is unknown follows their proof for Estimator 1.16 on pp. 38-39, with their expression for $g_N(i,z,\phi)$ replaced by $\ln P(z;\beta) - ln(\int P(x;\beta)dF(x))$ and their expression for $f(\phi)$ replaced by $\left(\frac{P(z;\beta^*)f(z)}{q}\right) ln\left(\frac{P(z;\beta)f(z)}{\int P(x;\beta)dF(x)}\right) + K$, where β^* represents the true value of β. As they note on p. 38, consistency of the unconstrained version of an estimator ensures the consistency of a constrained version of the estimator. If therefore follows that our constrained maximum likelihood estimator in Equation (4) for the case of a known prevalence rate is also consistent.

$$\max_{\beta} \left(\sum_{i=1}^{N_1} \ln(P(x_i;\beta)) \right) - N_1 ln\left(\int P(x;\beta) dF(x) \right). \quad (3)$$

Observe that the objective function in Equation (3) is a concentrated likelihood function obtained by substituting the expression in Equation (1) for the unknown value of q. An estimate ($\tilde{q}$) of the prevalence rate can be obtained, if desired, using the formula $\tilde{q} = \int P(x;\tilde{\beta}) dF(x)$, 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:

$$\max_{\beta} \sum_{i=1}^{N_1} \ln(P(x_i;\beta)) \text{ s.t. } q = \int P(x;\beta) dF(x). \quad (4)$$

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. Identification under a use-availability design

Unfortunately, the joint distribution of the covariates will not generally be known in practice. However, we can overcome our ignorance of the covariate distribution by incorporating a supplementary sample of covariate values from the overall population into the analysis.

Under this use-availability design, one would draw a primary random sample of covariate values from the subpopulation of participants and a separate supplementary random sample of covariate values from the general population. Assume, for now, that both the primary and supplementary samples are simple random samples. In Section 7 we will generalize our approach to account for exogenous stratification of one or both samples.

As noted by Lancaster and Imbens (1996), the supplementary sample under this design would permit identification of $f(x)$, while the primary sample would permit identification of $P(x;\beta)f(x)/q$. Thus, by implementing a use-availability design, the function $P(x;\beta)/q$ would continue to be non-parametrically identified. As noted previously for the case of a known covariate distribution, however, one would need to make parametric assumptions in order to separately identify the elements of β and q if the prevalence rate is unknown.

4. Estimation when $F(x)$ and q are both unknown

Development of our new estimators of β follows the approach introduced by Imbens (1992) and later employed by Lancaster and Imbens (1996). Under this approach, we begin by constructing an estimator for the case in which x is discrete. We then demonstrate that our estimator can be expressed in a way that not only requires no knowledge of the points of support for x, but which remains valid even when x is continuous. However, whereas Lancaster and Imbens (1996) and Cosslett (1981) develop their estimates based on the empirical probability distribution of x in the combined sample, we rely instead on the empirical probability distribution of x in the supplementary sample. As we shall see below in Section 7, this greatly simplifies estimation in cases where the primary and/or supplementary sample have been generated using a stratified sampling design. In particular, implementation of our estimators requires only application of the sample weights, whereas the Cosslett and Lancaster and Imbens estimators require detailed knowledge of the sampling design.

4.1 *Derivation of estimator*

If x is discrete with K known points of support, one can consistently estimate the probability (λ_k) that x is equal to x_k from a supplementary sample using the empirical

probability $\tilde{\lambda}_k = \frac{N_{0k}}{N_0},\ k = 1, \ldots, K$, where N_{0k} represents the number of observations in the supplementary sample with covariate value $x = x_k$.[7] The empirical probability $\tilde{\lambda}_k$ represents the unconstrained maximum likelihood estimate of λ_k based on the supplementary sample.

If the prevalence rate is unknown, one can estimate the conditional response probability parameters by maximizing: $L_{qunknown} = \left(\sum_{k=1}^{K} N_{1k} ln\big(P(x_k;\beta)\big)\right) - N_1 ln\left(\sum_{k=1}^{K} \tilde{\lambda}_k P(x_k;\beta)\right)$ over β, where N_{1k} represents the number of observations in the primary sample of participants with covariate value $x = x_k$. Equivalently, this optimization problem may be expressed as:

$$\tilde{\beta}_{qunknown} = \underset{\beta}{\operatorname{argmax}} \left(\sum_{l=1}^{N_1} ln\big(P(x_i;\beta)\big) \right) - N_1 ln\left(\frac{\sum_{j=1}^{N_0} P\big(x_j;\beta\big)}{N_0} \right). \tag{5}$$

The first order conditions for this estimator are:

$$\sum_{i=1}^{N_1} \frac{P'_\beta(x_i;\beta)}{P(x_i;\beta)} - \frac{N_1}{N_0 \tilde{q}(\beta)} \left(\sum_{j=1}^{N_0} P'_\beta\big(x_j;\beta\big) \right) = 0, \tag{6}$$

where $P_\beta(x;\beta) = \frac{\partial P(x;\beta)}{\partial \beta}$ and $\tilde{q}(\beta) = \sum_{j=1}^{N_0} P\big(x_j;\beta\big)/N_0$. Observe that these moments do not require knowledge of the points of support for x and that they remain valid even when x is not discrete.

The above estimator for β can be obtained using a standard maximum likelihood estimation routine.[8] However, the usual estimates of the standard errors (based on the information matrix) will not be valid, owing to the reliance on a sample analog $[\tilde{q}(\beta)]$ of the population relationship between q and β. Intuitively, the reliance on an approximate relationship

[7] Whereas our approach focuses on the unconditional probability (λ_k) of x_k and estimates it based on the supplementary sample moment (N_{0k}/N_0), the Lancaster and Imbens (1996) approach focuses on the conditional probability (π_k) that an observation with value x_k is included in the combined sample and estimates this probability using the combined sample moment $(N_{0k} + N_{1k})/(N_0 + N_1)$.

[8] See Lele and Keim (2006) for a related simulation-based approach to estimation in this case.

between β and q rather than the exact relationship tends to reduce the precision of our estimator to some degree. We rely on insights from generalized method of moments (GMM) theory to develop a covariance matrix estimator that properly accounts for this effect.

4.2 *GMM approach*

Following the approach taken by Lancaster and Imbens (1996) for their estimator, we recast the above problem using the GMM framework. Consider the following moments:

$$g_1(x;\theta) = s\frac{P'_\beta(x;\beta)}{P(x;\beta)} - (1-s)\frac{N_1}{N_0 q}P'_\beta(x;\beta). \tag{7}$$

$$g_2(x;\theta) = (1-s)\big(q - P(x;\beta)\big), \tag{8}$$

where $\theta = \begin{pmatrix}\beta \\ q\end{pmatrix}$ and s is a 1/0 indicator that identifies observations from the primary sample in the combined primary and supplementary sample. The moment $g_1(x;\theta)$ is the single observation score from the pseudo-log-likelihood function defined in Equation (5), while $g_2(x;\theta)$ reflects the relationship between marginal q and conditional $P(x;\beta)$. These moments have an expected value of zero when evaluated at the true value of θ.

Let $g(x;\theta)$ represent the vector $\begin{bmatrix}g_1(x;\theta) \\ g_2(x;\theta)\end{bmatrix}$, $N = (N_0 + N_1)$ represent the size of the combined primary and supplementary sample, and $g_N(x;\theta) = \frac{1}{N}\sum_{n=1}^{N} g(x_n;\theta)$ represent the $(H+1)$ x 1 vector of sample moment conditions. Based on our estimator $\tilde{\beta}_{qunk}$, we can rely on $\tilde{q}(\tilde{\beta}_{qunk})$ to estimate q, so that $\tilde{\theta} = \begin{pmatrix}\tilde{\beta}_{qunknown}, \\ \tilde{q}(\tilde{\beta}_{qunknown})\end{pmatrix}$. Then asymptotic covariance of our estimators can be estimated as:

$$V\left[\sqrt{N}\left(\tilde{\theta}-\theta\right)\right] \cong G_N\left(x;\tilde{\theta}\right)'\tilde{S}_N G_N\left(x;\tilde{\theta}\right), \tag{9}$$

where $\tilde{S}_N = \left[\frac{1}{N}\sum_{n=1}^{N} g(x_n;\theta)\, g(x_n;\theta)'\right]^{-1}$ and $G_N\left(x;\tilde{\theta}\right) = \frac{\partial g_N(x;\theta)}{\partial \theta'}\Big|_{\tilde{\theta}}$.

Alternatively, one can apply GMM estimation to develop asymptotically equivalent estimators of β and q:

$$\min_{\beta,q} g_N\,(x;\theta)' W_N g_N(x;\theta), \tag{10}$$

where $W_N = \frac{1}{N}\sum_{n=1}^{N} g\left(x_n;\tilde{\theta}\right) g\left(x_n;\tilde{\theta}\right)'$ is an estimate of the asymptotic covariance matrix of $\sqrt{N} g_N(x;\theta)$ based on $\tilde{\theta}$, a consistent initial estimate of θ.

5. Estimation when F(x) is unknown and q is known

Suppose that the prevalence rate is known. Returning to the example above in Section 4.1 where x is discrete with K known points of support, one could consistently estimate β in this case by maximizing the likelihood function, $L_{qknown} = \sum_{k=1}^{K} N_{1k} ln[P(x_k;\beta)]$, subject to the analog of the constraint on β that is imposed by prevalence rate from Equation (1): $q = \sum_{k=1}^{K} \tilde{\lambda}_k\, P(x_k;\beta)$, where N_{1k} represents the number of observations in the primary sample of participants with covariate value $x = x_k$, N_{0k} represents the corresponding number of participants in the supplementary sample, and $\tilde{\lambda}_k = \frac{N_{0k}}{N_0}$.

This estimator ($\tilde{\beta}_{qknown}$) can be expressed in an alternative way as the solution to:

$$\tilde{\beta}_{qknown} = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{N_1} ln\left(P(x_i;\beta)\right) \;\; s.t.\; q = \frac{\sum_{j=1}^{N_0} P\left(x_j;\beta\right)}{N_0}. \tag{11}$$

The Lagrangian for the optimization problem in Equation (11) is:

$$\mathcal{L}(\beta,\mu)=\sum_{i=1}^{N_1} ln\big(P(x_i;\beta)\big)\ +\mu\left(N_0 q-\sum_{j=1}^{N_0} P\big(x_j;\beta\big)\right). \tag{12}$$

The first-order conditions are:

$$\frac{\partial\mathcal{L}}{\partial\beta}=\sum_{i=1}^{N_1}\frac{P'_\beta(x_i;\beta)}{P(x_i;\beta)}-\mu\sum_{j=1}^{N_0} P'_\beta\big(x_j;\beta\big)=0. \tag{13}$$

$$\frac{\partial\mathcal{L}}{\partial\mu}=N_0 q-\sum_{j=1}^{N_0} P\big(x_j;\beta\big)=0. \tag{14}$$

Observe that these moments do not require knowledge of the points of support and that they remain valid even when x is not discrete.

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 Equation (13) 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 in Equation (12) and maximizing the following pseudo-likelihood function:

$$L=\sum_{i=1}^{N} s_i ln\big(P(x_i;\beta)\big)-\frac{N_1}{N_0 q}(1-s_i)P(x_i;\beta), \tag{15}$$

where s is a 1/0 indicator that identifies observations from the primary sample in the combined primary and supplementary sample.

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 $\tilde{\beta}_{qknown}$ 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, or the nonlinear optimization routines in SAS$^{@}$/IML$^{@}$, and the CML application in GAUSS$^{@}$,.

5.1 *GMM framework*

Although the conditional response probability parameters can be estimated using a constrained maximum likelihood algorithm, the usual estimate of the covariance matrix of the parameter estimates from such an algorithm will not be valid. Again, this is because we have replaced the exact formula for q $\left(\int P(x;\beta)dF(x)\right)$ with its sample analog $\left(\sum_{j=1}^{N_0} P\left(x_j;\beta\right)/N_0\right)$. We rely on insights from generalized method of moments (GMM) theory to develop a covariance matrix estimator that properly accounts for this substitution.

Consider the following moments:

$$g_1(x;\beta) = s\frac{P'_\beta(x;\beta)}{P(x;\beta)} - (1-s)\frac{N_1}{N_0 q}P'_\beta(x;\beta). \tag{16}$$

$$g_2(x;\beta) = (1-s)\left(q - P(x;\beta)\right). \tag{17}$$

The moment $g_1(x;\beta)$ is the single observation score from the pseudo-log-likelihood function defined in Equation (15), while $g_2(x;\beta)$ reflects the relationship between marginal q and conditional $P(x;\beta)$. These moments have an expected value of zero when evaluated at the true value of β. An estimate of asymptotic covariance matrix for our estimator $\tilde{\beta}_{qknown}$ can be

obtained by evaluating the standard formula for the GMM estimator of the covariance matrix based on these moment conditions at $\tilde{\beta}_{qknown}$. Alternatively, one can directly apply GMM estimation to the above moment conditions to obtain an asymptotically equivalent estimator of β. Depending on one's preference, then, one can rely either on maximum likelihood estimation or unconstrained GMM estimation.

6. Monte Carlo analysis

We have undertaken a Monte Carlo analysis to compare the small sample performance of our estimators against other existing estimators. In our simulations, we employ a logit specification for the conditional probability of participation involving 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. We perform 1,000 replications for each experiment.

A standard use-availability design is employed that includes a supplementary random sample of $N_0 = 400$ participants and non-participants and a primary random sample of $N_1 = 400 * q$ participants. We also experiment with a larger supplementary sample of $N_0 = 1{,}600$ participants and non-participants.

6.1 *Known prevalence rate*

For the known prevalence rate case, we compare the small sample performance of our calibrated logit estimator $\left(\tilde{\beta}_{qknown}\right)$, defined in Equation (11), against several alternative estimators from the existing literature on supplementary sampling. Below, we briefly describe these alternative estimators, which are explored in more detail in Appendix A. We then present our findings.

Cosslett (1981) developed a generalized framework for estimating discrete choice models using choice-based samples. Through a straightforward extension of his estimation methodology for the case of a known prevalence rate, a supplementary sampling estimator for the response parameters (β) can be obtained as the solution to the following optimization problem:

$$\max_{\beta} \min_{\lambda_1} \sum_{i=1}^{N} s_i ln(P(x_i;\beta)) - l\,n(\lambda_1 P(x_i;\beta) + 1 - \lambda_1 q), \tag{18}$$

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 Equation (18).

The Lancaster-Imbens (1996) estimator is obtained by applying GMM estimation to the following three moment conditions:

$$g_1(x;\beta,h) = \frac{P'_\beta(x;\beta)}{P(x;\beta)}\big(s - R(x;\beta,q,h)\big). \tag{19}$$

$$g_2(x;\beta,h) = -\frac{1}{q}\big(s - R(x;\beta,q,h)\big). \tag{20}$$

$$g_3(x;\beta,h) = h - R(x;\beta,q,h), \tag{21}$$

where $R(x;\beta,q,h) = \frac{h\left(\frac{P(x;\beta)}{q}\right)}{\left[h\left(\frac{P(x;\beta)}{q}\right)+(1-h)\right]}$ represents the conditional probability of selection into the primary sample. In this model, parameter h, which represents the Bernoulli probability that a sample observation is drawn from the subpopulation of participants, is estimated jointly with β.This contrasts with the approach taken by Cosslett (1981) as well as our current approach wherein the values of N_0 and N_1 are treated as fixed.

The Steinberg-Cardell (1992) estimator is the solution to the following optimization problem:

$$\max_{\beta} \sum_{i=1}^{N} s_i \left(\frac{N_0 q}{N_1}\right) ln\left(\frac{P(x_i;\beta)}{1-P(x_i;\beta)}\right) + (1-s_i) ln\big(1-P(x_i;\beta)\big). \quad (22)$$

The Monte Carlo simulation results are summarized in Table 1. For each case, we report the mean and median estimates, the mean asymptotic standard deviation of the estimates (ASD), the standard deviation of the estimates (SSD) over the 1,000 replications, and the mean absolute deviation from the median estimates (MAD) over the 1,000 replications. In some of the simulations, certain estimators are subject to convergence problems. For such estimators, we perform our tabulations based on the subset of replications that are free from convergence problems. The number of replications where an estimator has failed to converge is reported as "#Failures".

When q is relatively low, the estimators all perform similarly, with the exception of the Steinberg-Cardell estimator. Even when the prevalence rate is small, this estimator is relatively inefficient in comparison with the other estimators. As the prevalence rate rises, the calibrated logit estimator continues to perform comparably to the Cosslett and Lancaster-Imbens estimators. On the other hand, the Steinberg-Cardell estimator suffers not only from relatively high standard errors, it also is subject to periodic convergence problems.

In the final case presented in Table 1, we explore the performance of the various estimators when the prevalence rate is high ($q = 0.875$), but a larger estimation sample is employed. In particular, we quadruple the sample size (from $N_0 = 400$ and $N_1 = 350$ to $N_0 = 1{,}600$ and $N_1 = 1{,}400$). The application of a larger estimation sample largely eliminates

the convergence problems associated with the Steinberg-Cardell estimator. As well, the precision of all of the estimators is substantially improved.

We have also performed some Monte Carlo simulations for our alternative GMM-based estimator. Our GMM estimator based on the moment conditions in Equations (16) and (17) produces very similar results to our calibrated logit estimator, even in small samples.

6.2 *Unknown prevalence rate*

Alternative estimators for the case of an unknown prevalence rate have been proposed by Cosslett (1981) and Lancaster and Imbens (1996). We show in Appendix A that these two estimators are actually the same. The Cosslett-Lancaster-Imbens estimator is the solution to the following optimization problem:

$$\max_{\beta} \max_{q\in(0,1)} \sum_{i=1}^{N} s_i ln\left(\frac{N_1}{Nq}P(x_i;\beta)\right) - ln\left(\frac{N_1}{Nq}P(x_i;\beta) + \frac{N_0}{N}\right). \quad (23)$$

We have undertaken some Monte Carlo simulations to compare the small sample performance of our pseudo-maximum likelihood estimator based on Equation (5) and the Cosslett-Lancaster-Imbens estimator based on Equation (23) for the unknown prevalence rate case. The results are summarized in Table 2. For each case, we report the mean and median estimates, the standard deviation of the estimates (SSD), and the mean absolute deviation from the median estimates (MAD) over the 1,000 replications. We also present the mean asymptotic standard deviation of the estimates based on the pseudo-likelihood function (LSD). In the case of the Cosslett-Lancaster-Imbens estimator, we derive the standard error estimates using the inverse of the information matrix. Lancaster and Imbens (1996) have shown that these standard error estimates are consistent for the coefficients (but not for q). For our pseudo-maximum likelihood model, we rely on the Huber-White standard errors for our LSD estimates. The LSD estimate of

the standard error for our pseudo-maximum likelihood estimate of q is computed using the delta method. In large samples, these estimates will tend to be somewhat too small, because they do not account for our reliance on a sample analogue of the true relationship between marginal q and conditional $P(x;\beta)$. We compare our LSD estimates to the GMM-based standard error estimates (GSD), which do account for this relationship.

In small samples, both estimators are subject to periodic convergence problems. We base our performance measures for a given estimator on the subset of replications that are free from such problems. The number of replications for which an estimator has failed to converge is reported as "#Failures".

Comparing findings for the cases involving a known and an unknown prevalence rate, it is clear that precision suffers when q is unknown. The discrepancy in performance across these two cases is especially pronounced when q is relatively large ($q = .75$ and $q = .875$). In addition, the discrepancy is much larger for the intercept than for the slope coefficients.

Overall, our pseudo-maximum likelihood estimator performs quite comparably to the Cosslett-Lancaster-Imbens estimator in terms of mean and median performance as well as precision. 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. We find that the estimated covariance matrices for both supplementary sampling estimators tend to become ill-conditioned at solutions involving estimated values of q close to zero, and the standard error of the intercept estimate becomes very large in such cases.

Our simulation results indicate that convergence problems are also prevalent when the true value of q is relatively high ($q = 0.75$ and $q = .875$). One source of such problems is that, despite the high actual prevalence rate, some replications result in an estimated prevalence rate that is actually close to zero. Another source of convergence problems when q is relatively high involves estimates of the prevalence rate that are very close to the upper bound of one. Typically in such cases, the average predicted conditional probability of participation approaches one within the primary sample, while the average predicted probability is just slightly smaller within the supplementary sample.

In the final case presented in Table 2, we explore the performance of the estimators when the prevalence rate is high ($q = 0.875$), but a larger estimation sample is employed. This not only leads to substantial improvements in precision, it also greatly reduces the incidence of convergence problems. In general, then, when the prevalence rate is not known, it is especially beneficial to employ a reasonably large overall sample in estimation.

We have also performed Monte Carlo simulations using our GMM estimator based on the moment conditions in Equations (7) and (8). The results indicate that this estimator and our pseudo-maximum likelihood estimator for the case of an unknown prevalence rate produce very similar estimates, even in small samples.

7. Exogenously stratified samples

The results from the Monte Carlo simulations indicate that our new estimators rival the performance of the existing estimators developed by Cosslett and Lancaster-Imbens for the case in which the primary and supplementary data sources are simple random samples from their respective populations. The advantage of our new estimators is that they can also be applied in situations where the Cosslett and Lancaster-Imbens estimators are not feasible.

In Appendix B, we show how each of the existing supplementary sampling estimators can be generalized to accommodate exogenous stratification of the primary and/or supplementary samples. With the sole exception of the relatively inefficient Steinberg-Cardell estimator, however, implementation of these generalized estimators would require one to allocate observations from both the primary and supplementary samples to a common set of sampling strata (or substrata in the likely case that the stratum definitions differ across the two samples). Unfortunately, the requisite information is not always available to do so. For instance, the U.S. Census Bureau does not publicly disclose its stratification criteria for national surveys such as the CPS, SIPP, and ACS.[9] Therefore, if public-use data from one of these surveys were used as a supplementary sample from the general population, researchers would not know how to construct comparable strata for members of the primary sample, which would preclude application of these estimation methods.[10]

An important advantage of our new supplementary sampling estimators is that they can be implemented even when the stratification criteria are unavailable; all that is needed are the sample weights. Let the sample weights for the primary data source be represented by w_1 and

[9] Under a fairly simple stratified random sampling design, it may be possible to deduce the stratification criteria (at least approximately) by analyzing the characteristics of each subsample of observations with a common value for the sample weight (assuming that the relevant stratifying variables are present in the data sample). However, such an approach is not feasible for more complex survey designs. For instance, Census surveys often involve multi-stage sampling, clustering, post-stratification adjustment, and imputation. As a consequence, the final sample weight often varies among observations within the same initial stratum. Even when the sampling criteria for the supplementary sample can be deduced, it is only feasible to evaluate the relative sampling weights if the stratifying variables are also present in the primary sample. In cases where both the primary and supplementary data sources are stratified, one would further need to divide the existing strata for the two data samples into sub-strata that are comparable across the two samples. In such cases, the presence of sparse or empty sub-strata would complicate estimation.

[10] Even if the specific survey design criteria were made known for such surveys, it would be very difficult to adapt the Cosslett and Lancaster-Imbens estimators to account for the complexity of these survey designs. Although Appendix B shows how to apply these estimators under a relatively simple stratified random sampling process, accounting for more complex designs involving multi-stage sampling and clustering would be much more challenging.

those for the supplementary data source by w_0. We assume that these weights have been normalized so that they sum to the size of their respective samples, N_1 and N_0. If either of the samples is not stratified, the weight for each observation in that sample would be set equal to one.

Our generalized constrained pseudo-maximum likelihood estimator for the case of a known prevalence rate ($\tilde{\beta}_{W,qknown}$) is constructed by incorporating the relevant sample weights into the objective function and constraint of the optimization problem described by Equation (11):

$$\tilde{\beta}_{W,qknown} = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{N_1} w_{1i} ln\big(P(x_i;\beta)\big) \;\; s.t. \; q = \frac{\sum_{j=1}^{N_0} w_{0j} P\big(x_j;\beta\big)}{N_0}. \tag{24}$$

When the prevalence rate is unknown, the objective function in Equation (5) is easily generalized to account for stratified sampling as follows:

$$\tilde{\beta}_{W,qunknown} = \max_{\beta} \left(\sum_{l=1}^{N_1} w_{1i} ln\big(P(x_i;\beta)\big) \right) - N_1 ln\left(\frac{\sum_{j=1}^{N_0} w_{0i} P\big(x_j;\beta\big)}{N_0} \right).^{11} \tag{25}$$

The GMM versions of our estimators can also be generalized by appropriately weighting the moment conditions; these weighted moment conditions can also be employed to estimate the covariance matrices for the above estimators.

8. Polychotomous response models

Our estimation approach readily generalizes to account for more than two outcomes. For instance, suppose there are M+1 possible outcomes indexed by the values $y = 0, 1, \cdots, M$. Define

[11] If the available sample weights for the primary and supplementary samples sum to their respective population totals, then the prevalence rate actually will be known since it can be computed as the ratio of the sum of the sample weights for the primary sample to the sum of the sample weights for the supplementary sample. However, if only normalized sample weights are available (which instead sum to the respective sample sizes), it will not be possible to deduce the prevalence rate from such weights.

the outcome probabilities as: $\Pr(y = m|x;\beta) = P(m|x;\beta),\ m = 0, 1, \ldots, M$, where $P(m|x;\beta)$ has a known parametric form. This framework is sufficiently general to include both ordinal and multinomial response models.

Let the outcome $y = 0$ represent non-participation and let the remaining M outcomes represent alternative forms of participation. Suppose one has a random participant-only sample of size N_1 that includes observations with outcomes 1 through M. Sampling among these participants may be choice-based, meaning that the sampled number of observations (N_{1m}) for a given participation outcome m may not be representative of the incidence of this outcome within the participant population. In addition, suppose one has a supplementary random sample of size N_0 from the general population that includes observations on all types of participants as well as non-participants.

Define q_m as the prevalence rate for outcome $m,\ m = 1, \cdots, M$. Assuming these prevalence rates are known, our calibrated qualitative response estimator for the binary response case described in Equation (11) may be adapted to account for polychotomous responses as follows:

$$\tilde{\beta}_{P,qknown} = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{N_1} ln\big(P(y_i|x_i;\beta)\big)\ \ s.t.\ q_m = \frac{\sum_{j=1}^{N_0} P\big(m|x_j;\beta\big)}{N_0},\ m = 1 \cdots, M. \tag{26}$$

Thus, the generalized form of our calibrated qualitative response estimator involves M constraints, one for each outcome in the primary sample. To estimate the covariance matrix of $\tilde{\beta}_{P,qknown}$, one can rely on the GMM covariance matrix formula associated with the following moment conditions:

$$g_0(x;\beta) = \frac{P'_\beta(y|x;\beta)}{P(y|x;\beta)} - (1-s)\sum_{m=1}^{M} \frac{N_{1m}}{N_0 q_m} P'_\beta(m|x;\beta). \tag{27}$$

$$g_m(x;\beta) = (1-s)(q_m - P(m|x;\beta)),\ \ m = 1,\dots,M. \tag{28}$$

Alternatively, one can derive an asymptotically equivalent estimator of β by applying GMM estimation to these moment conditions.

If the prevalence rates are unknown, the optimization problem defined in Equation (5) may be generalized to:

$$\tilde{\beta}_{P,qunknown} = \underset{\beta}{\text{argmax}} \sum_{i=1}^{N_1} ln\big(P(y_i|x_i;\beta)\big) - \sum_{m=1}^{M} N_{1m} ln\left(\frac{\sum_{j=1}^{N_0} P\big(m|x_j;\beta\big)}{N_0}\right). \tag{29}$$

The parameters q_m can then be estimated using the analogue estimator: $\tilde{q}_m = \frac{\sum_{j=1}^{N_0} P\big(m|x_j;\tilde{\beta}_{P,qunk}\big)}{N_0}$.

For estimation of the covariance matrix, one can rely on the GMM covariance matrix formula based on the moment conditions in Equations (27) and (28), where these conditions are now taken as a function of the unknown parameters q_m as well as β. Alternatively, one can derive asymptotically equivalent estimators of β and q by applying GMM estimation to these moment conditions.

To extend the above estimators to account for stratified random sampling on exogenous factors, one simply applies the appropriate primary and supplementary sample weights to the terms in equations (26) through (29).

9. An example using stratified data samples

Burden et al. (2014) estimate the determinants of voting behavior using data from the Current Population Survey (CPS) for 2004 and 2008 using both binary and multinomial logit specifications. In this section, we focus on their analysis for 2008. We begin by estimating similar specifications to those used in their study based on the same 2008 CPS data sample. We

then compare the results against alternative estimates derived from a use-based sample design involving a voter-only subsample from the CPS and a supplementary sample from the overall voting-eligible population from the American Community Survey (ACS). These alternative estimates include results based on our calibrated binary and multinomial logit models, the Steinberg-Cardell binary logit estimator as well as a multinomial logit generalization of their estimator that we have derived in Appendix B, and our pseudo-MLE binary and multinomial logit estimators for the case involving an unknown prevalence rate.[12]

The binary logit specification employed by Burden et al. distinguished voters and non-voters. The multinomial logit specification distinguished among the following modes of voting: (1) election-day voting; (2) early voting in person; and (3) early voting by mail. Both specifications relied on the following explanatory variables:

Early: Dummy for residence in a state that permits early voting.
EDR: Dummy for residence in a state that permits one to both register and vote on Election Day.
Early*SDR: Dummy for residence in an early voting state that permits same-day registration.
Early*EDR: Interaction between Early Voting and EDR.
Early*EDR*SDR: Interaction between Early Voting, SDR, and EDR.
30-Day Reg. Close: Dummy for residence in a state that requires voters to be registered 30 days in advance of an election.
ID Requirement: Dummy for residence in a state that requires voters to show some form of identification.
Education: Indicator for educational attainment (4 values ranging from less than high school diploma to Bachelor's degree or higher).
African American: Dummy for self-identified race of Black only or Black in combination with another race.
Hispanic: Dummy for self-identified race of Hispanic origin.
Naturalized Citizen: Dummy for naturalized citizenship.
Married: Dummy for married.
Female: Dummy for female.
Age: Age in years.
Age 18–24: Dummy for age between 18 and 24.
Age 75 plus: Dummy for age 75 or over.

[12]The authors have kindly posted the Stata code they used in their analysis at https://electionadmin.wisc.edu/BCMM_AJPS_CPSanalysis.zip. This code greatly facilitated the replication of their original results.

Competitiveness: A poll-based index of campaign competitiveness (a higher value indicates a more competitive campaign).
South: Dummy for residence in a southern state.
North Dakota: Dummy for residence in North Dakota (which does not require voter registration).
Oregon: Dummy for residence in Oregon (a "vote-by-mail" state).
Washington: Dummy for residence in Washington state (a "vote-by-mail" state).
Self-Reported Vote: Dummy equal to one if voting status was self-reported and zero if reported by another family member.
Natural. 10+ Years: Dummy for naturalized citizen who entered the U.S. prior to 1998.
Residence 1 Year: Dummy for tenure of at least one year at current residence.
Income: Indicator for total family income (16 values ranging from less than $5,000 to $150,000 and over).

The estimation sample was restricted to individuals who appeared eligible to vote (age 18 or over and a U.S. citizen) and who did not reside in the District of Columbia.

In order to apply the calibrated qualitative response methodology, it is essential to have comparably defined and measured variables in the primary and supplementary data sources. As this example demonstrates, this requirement imposes limitations on the set of explanatory variables that one can use in an analysis. In particular, the last four variables listed above do not satisfy this requirement. Although a comparable family income concept can be constructed from the ACS data, it turns out that the CPS family income measure is missing for approximately 20 percent of the voting-eligible sample, including a disproportionate share of lower-income households.[13] Consequently, the (weighted) subsample with non-missing information is not representative of the overall population and therefore cannot be validly compared against the (weighted) ACS sample. A similar missing data problem exists with regard to tenure at the

[13] Based on a comparison of the ACS (which has complete income information) and the CPS, it appears that a disproportionate share of the missing responses in the CPS is attributable to lower income households. Burden et al. restrict their analysis to the portion of their CPS sample with non-missing income information. This restriction might be justified if it can be assumed that willingness to provide income information on the CPS survey is uncorrelated with voting behavior. However, the validity of this assumption is uncertain. Note that even if this assumption were valid, it would not be feasible to include the income measure as a regressor in the calibrated qualitative response model.

current address.[14] In addition, two of the variables used in the Burden et al. analysis (Naturalized 10+ Years and Self-Reported Vote) cannot be constructed from the ACS data.[15]

For purposes of illustration and comparison of methodologies, we have therefore estimated specifications that exclude these four variables from the analysis. Tables 3 and 4, respectively, compare the standard binary and multinomial logit estimates based on the CPS to the corresponding estimates of our alternative models based on a supplementary sampling scheme that includes the subsample of voters in the CPS as our primary sample and a 10 percent random subsample of voting-eligible individuals in the ACS as our supplementary sample. Both the CPS and the ACS rely on stratified sampling designs, so we incorporate the publicly available sample weights from both surveys in our analysis as discussed in Section 7.[16]

Overall, our calibrated binary and multinomial logit estimates are qualitatively quite similar to the standard binary and multinomial logit results.[17] Differences in the relative magnitudes of certain coefficients across methods are largely attributable to moderate differences in the weighted mean values of the underlying regressors (such as the dummies for marital status,

[14] This Information is missing for approximately 13 percent of the CPS voter-eligible sample. The authors set the Residence 1 Year dummy equal to one when this information was missing, which resulted in an unknown number of instances of misclassified residential tenure status. Such an approach introduces bias into the binary and multivariate logit findings. Moreover, it invalidates the comparison against ACS data employed under our calibrated qualitative response approach.

[15] The Naturalized 10+ Years dummy is based on information concerning the date of entry to the U.S. The ACS inquires about the date of naturalization but not the date of entry (which typically occurs many years earlier).

[16] In the case of the standard binary and multinomial logit specifications based on the CPS data sample, we have followed the authors in performing an unweighted analysis, followed by the computation of cluster-robust standard errors by state.

[17] For our calibrated binary logit model, we have relied on the weighted mean value of the voting indicator in the CPS sample, inclusive of those observations with missing income information, as our measure of the prevalence rate. Similarly, for the calibrated multinomial logit model, we have relied on the weighted mean values of reported shares of individuals voting on election day in person, voting early in person, and voting early by mail (inclusive of observations with missing income information) as our measures of the prevalence rates for these three different voting methods.

age range, and residence in certain states with different voting requirements) across the two data sources.

Overall, the Steinberg-Cardell coefficient estimates are also qualitatively quite similar to the logit estimates for the binary response case. However, some of the multinomial response coefficient estimates based on the Steinberg-Cardell approach deviate fairly substantially from the corresponding multinomial logit and calibrated multinomial logit estimates.

Our pseudo-maximum likelihood estimates based on unknown prevalence rates are very similar to our calibrated binary and multinomial logit results based on specified values for the prevalence rates. In addition, the pseudo-maximum likelihood estimates of the prevalence rates are reasonably close to measures computed using the weighted CPS statistics. Overall, our combined estimation sample is quite large (273,933). Although we do lose some precision when we do not specify a prevalence rate in estimation, the large overall size of the combined sample (273,933) ensures that we are still able to obtain reasonably precise estimates of the conditional response probability parameters.

10. Summary and conclusion

Frequently, researchers have access to detailed information on the relevant characteristics of participants in a program, patients suffering from a disease, or habitats where a species is known to be present. However, their lack of comparable information about households that do not participate in the program, individuals who are free of the disease, or habitats where the species is not present precludes the application of standard qualitative response models to analyze the determinants of the outcome under investigation.

If the joint probability distribution of the underlying covariates were known, we have demonstrated how a constrained maximum likelihood procedure could be used to estimate the

parameters of the conditional response probability distribution based solely on an available sample of participants. This approach exploits the parameter restrictions implied by the relationship between the marginal and conditional probabilities of participation: $q = \int P(x;\beta)\, dF(x)$, where q is the marginal probability of participation (i.e., the prevalence rate), $P(x;\beta)$ is the conditional probability of participation, and $F(x)$ is the joint distribution function of the covariates. In practice, however, this approach is not generally feasible to implement, because $F(x)$ is unknown.

To overcome this problem, we have shown that one can replace the unknown relationship between the marginal and conditional response probability distributions with its analogue based on a supplementary sample of size N_0 from the general population: $\tilde{q} = \frac{1}{N_0}\sum_{i=1}^{N_0} P(x_i;\beta)$. Using this analogue relationship, we have derived some feasible new constrained and unconstrained pseudo-maximum likelihood estimators of the parameters of the conditional response probability distribution. Following Lancaster and Imbens (1996), we show how our optimization problem can be recast under a GMM framework. This leads to some asymptotically equivalent estimators as well as a straightforward way to obtain appropriate standard errors for our pseudo-maximum likelihood estimators. We also demonstrate that our framework is readily generalized to accommodate polychotomous responses.

We have conducted some Monte Carlo simulations to compare the small sample performance of our new estimators against that of existing estimation approaches, including Cosslett (1981), Lancaster and Imbens (1996), and Steinberg and Cardell (1992). Our Monte Carlo simulations reveal several insights. When the prevalence rate is known, our calibrated qualitative response estimator rivals the performance of the best existing estimators (Lancaster-Imbens and Cosslett) in small samples. The Steinberg-Cardell estimator exhibits less precision in

our Monte Carlo simulations, and it is also subject to convergence issues, particularly when the sample size is small and q is relatively large.

When the prevalence rate is unknown, our pseudo-maximum likelihood estimator performs comparably to the Cosslett-Lancaster-Imbens estimator. Our Monte Carlo simulations reveal that both estimators are relatively imprecise in small samples and are subject to periodic convergence problems, particularly when q is fairly close to either of its boundaries (0 or 1). Both of these problems are alleviated by using a larger estimation sample. However, owing to the reliance on specific parametric assumptions to identify the conditional response probability parameters when the prevalence rate is unknown, one will tend to have greater confidence in estimates of relative, rather than absolute, probabilities.

An important advantage of our new estimators over those proposed by Cosslett and Lancaster-Imbens is that the latter estimators require detailed knowledge of the sampling criteria when the primary and/or supplementary sample is exogenously stratified. This precludes their application when the relevant sampling criteria have not been made publicly available, such as when the supplementary sample has been drawn from a Census survey. In contrast, our estimators require knowledge only of the sample weights, which are routinely available. The new estimators therefore significantly broaden the scope of potential data sources that can be used in estimation of qualitative response probabilities. With these new estimators, for example, one can rely on publicly available data from national surveys, such as the CPS, ACS, and SIPP, as supplementary data sources for estimation.

References

Breslow, N.E., 1996, Statistics in epidemiology: The case-control study. Journal of the American Statistical Association 91:433, 14-28.

Burden, B.C., D.T. Canon, K.R. Mayer, and D.P. Moynihan, 2014, Election laws, mobilization, and turnout: The unanticipated consequences of election reform. American Journal of Political Science 58:1, 95-109.

Cosslett, S.R.,1981, Efficient estimation of discrete choice models, in: C. Manski and D. McFadden, (Eds.), Structural analysis of discrete data with econometric applications. MIT Press, Cambridge, pp. 51-111.

Erard, B., J. Guyton, P. Langetieg, M. Payne, and A. Plumley, 2016, What drives income tax filing compliance? IRS Research Bulletin, Publication 1500. Internal Revenue Service, Washington, DC, pp. 32-37.

Imbens, G.W. , 1992, An efficient method of moments estimator for discrete choice models with choice-based sampling. Econometrica 60:5, 1187-1214.

Keating, K.A. and S. Cherry, 2004, Use and interpretation of logistic regression in habitat selection studies. Journal of Wildlife Management 68:4, 774-789.

Lancaster, T. and G. Imbens, 1996, Case controlled studies with contaminated controls. Journal of Econometrics 71, 145-160.

Lele, S.R., 2009, A new method for estimation of resource selection probability function. Journal of Wildlife Management 73:1, 122-127.

Lele, S.R. and J.L. Keim, 2006, Weighted distributions and estimation of resource selection probability functions. Ecology 87:12, 3021-3028.

Manski, C.F. and D. McFadden, 1981, Alternative estimators and sample designs for discrete choice analysis, in: C. Manski and D. McFadden, (Eds.), Structural analysis of discrete data with econometric applications. MIT Press, Cambridge, pp. 2-49.

Phillips, S.J. and J. Elith, 2013, On estimating probability of presence from use-availability or presence-background data. Ecology 94:6, 1409-1419.

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, 545-554.

Solymos, P. and S.R. Lele, 2016, Revisiting resource section 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. Communication in Statistics –Theory and Methods 21:2, 423-450.

Ward, G., T. Hastie, S. Barry, J. Elith, and J.R. Leathwick, 2009, Presence-only data and the EM algorithm. Biometrics 65, 554-563.

Appendix A: Existing alternative estimators

This appendix provides a discussion of existing alternative estimators of the conditional response probability parameters under a supplementary sampling scheme.

Estimators based on the Cosslett framework for a known prevalence rate

In his seminal study of discrete choice estimation under choice-based sampling, Cosslett (1981) derives a generalized framework for asymptotically efficient estimation. Although he extends his framework to consider the case of supplementary sampling when the prevalence rate is unknown, he does not derive a corresponding supplementary sampling estimator for the situation involving a known prevalence rate. We employ Cosslett's estimation framework to derive an estimator for this situation below.

The first step is to consider the optimization problem under a specified functional form for the covariate distribution:

$$\max_{\beta} \sum_{i=1}^{N} s_i ln(P(x_i;\beta)) + lnf(x_i) \;\; s.t. \;\; q = \int P(x;\beta)\, f(x)\, dx\,. \tag{30}$$

Under Cosslett's approach, one replaces the covariate density $f(x)$ in Equation (30) with an empirical density characterized by a weight factor w_i:

$$\max_{\beta, w_1, w_2, \cdots, w_N} \sum_{i=1}^{N} s_i ln(P(x_i;\beta)) + ln(w_i) \;\; s.t. \;\; q = \sum_{i=1}^{N} P(x_i;\beta)\, w_i \text{ and } \sum_{i=1}^{N} w_i = 1. \tag{31}$$

The first-order condition for w_i implies: $\frac{1}{w_i} = \lambda_1 P(x_i;\beta) + \lambda_0$,[18] where λ_1 and λ_0 are the Lagrange multipliers associated with the two constraints in Equation (31). Substitution of this result into Equation (31) yields the dual optimization problem:

$$\max_{\beta} \min_{\lambda_0, \lambda_1} \sum_{i=1}^{N} s_i ln(P(x_i;\beta)) - l\,n(\lambda_1 P(x_i;\beta) + \lambda_0) - N(1 - \lambda_1 q - \lambda_0). \quad (32)$$

Observe that, whereas the original optimization problem involved *maximization* over the weights w_i, the dual optimization problem involves *minimization* over the Lagrange multipliers. The optimization problem in Equation (32) is equivalent to the following problem:

$$\max_{\beta} \min_{\lambda_0, \lambda_1} \sum_{i=1}^{N} s_i ln(P(x_i;\beta)) - l\,n(\lambda_1 P(x_i;\beta) + \lambda_0) \;\; s.t. \;\; \lambda_1 q + \lambda_0 = 1. \quad (33)$$

Further simplification is possible by substituting the above constraint on the multipliers into the objective function:

$$\max_{\beta} \min_{\lambda_1} \sum_{i=1}^{N} s_i ln(P(x_i;\beta)) - l\,n(\lambda_1 P(x_i;\beta) + 1 - \lambda_1 q). \quad (34)$$

A less efficient but simpler feasible estimator of β can be obtained by substituting the limit values for λ_0 and λ_1 (N_0/N and N_1/Nq, respectively) into Equation (32):

$$\max_{\beta} \sum_{i=1}^{N} s_i ln(P(x_i;\beta)) - ln\left(\frac{N_1}{Nq} P(x_i;\beta) + \frac{N_0}{N}\right). \quad (35)$$

This simplified estimator was proposed by Lancaster and Imbens (1996, p. 153) as a feasible means to obtain an initial consistent estimate for use in solving the GMM estimation problem associated with their estimator.

[18] Note that the weights w_i must be positive, which implies that $(\lambda_1 P(x_i;\beta) + \lambda_0)$ must also be positive.

Lancaster-Imbens estimator for a known prevalence rate

Lancaster and Imbens (1996) develop a GMM approach to the estimation of response probabilities using a supplementary sampling scheme. In their formulation of the problem, the primary and supplementary samples are drawn using a sequence of Bernoulli trials with unknown parameter h. They begin by considering a case involving discrete covariate values that have a finite number of points of support, characterized by the p.d.f. $f(x;\lambda)$. The likelihood function for this problem may be expressed as:

$$L = \sum_{i=1}^{N} (s_i ln\, P(x_i;\beta) + f(x_i;\lambda)) - N_1 lnh - N_0 \ln(1-h) - N_1 lnq. \quad (36)$$

Lancaster and Imbens then reparametrize this likelihood function in terms of the sampling distribution of the covariates: $g(x;\lambda) = [(h/q)P(x;\beta) + (1-h)]f(x;\lambda)$. The reformulated likelihood function is specified in terms of β, q, h, and π:

$$L^R = \sum_{i=1}^{N} s_i ln\, R(x_i;\beta,q,h) + (1-s_i)\ln\big(1 - R(x_i;\beta,q,h)\big) + \ln g(x_i;\, \pi), \quad (37)$$

where $R(x;\beta,q,h) = \frac{(h/q)P(x;\beta)}{(h/q)P(x;\beta)+(1-h)}$ and the value of π at the k^{th} point of support (x^k) is equal to $\pi_k = [(h/q)P(x^k;\beta) + (1-h)]\lambda_k$.

Whereas maximization of the original likelihood function is subject to the restriction $q = \int P(x;\beta)dF(x;\lambda)$, maximization of the reformulated likelihood function is subject to the restriction $h = \int R(x;\beta)dG(x;\pi)$.[19] Rather than pursue a constrained maximum likelihood estimation strategy, Lancaster and Imbens derive their estimator by applying GMM estimation based on the following three moment conditions:

[19] Ward et al. (2009) develop an expectation-maximization (EM) algorithm that solves for the constrained maximum likelihood solution under a logistic specification for the conditional response probability distribution.

$$g_1(x;\beta,h) = \frac{P'_\beta(x;\beta)}{P(x;\beta)}\big(s - R(x;\beta,q,h)\big). \tag{38}$$

$$g_2(x;\beta,h) = -\frac{1}{q}\big(s - R(x;\beta,q,h)\big). \tag{39}$$

$$g_3(x;\beta,h) = h - R(x;\beta,q,h). \tag{40}$$

The third moment condition is the sample analogue of the restriction $h = \int R(x;\beta)dG(x;\pi)$, while the first two conditions represent the single observation scores of the likelihood function in Equation (37) for β and h, respectively. Observe that these three moment conditions do not require knowledge of the points of support for x, and they remain valid even when x is continuous.

Steinberg-Cardell estimator for a known prevalence rate

The Steinberg-Cardell (1992) estimator is motivated by the estimator that one might use under the hypothetical scenario where the primary sample includes all participants in the population and the supplementary sample includes all participants and non-participants in the population. Even if the participants and non-participants in the supplementary sample could not be distinguished, one could effectively estimate a binary choice model by solving the following optimization problem:

$$\max_{\beta} \sum_{i=1}^{T} s_i ln\, P(x_i;\beta) \; + ln\big(1 - P(x_i;\beta)\big) - s_i\, ln\big(1 - P(x_i;\beta)\big), \tag{41}$$

where T represents the population size. Under the standard binary choice framework, the likelihood function accumulates the values of $lnP(x_i;\beta)$ across all participants and the values of $ln\big(1 - P(x_i;\beta)\big)$ across all non-participants. The former tally is achieved by the first term in

Equation (41), while the latter is achieved by the combination of the second and third terms. Rearranging terms, the optimization problem in Equation (41) can equivalently be expressed as:

$$\max_{\beta} \sum_{i=1}^{T} s_i ln\left(\frac{P(x_i;\beta)}{1-P(x_i;\beta)}\right) + ln\big(1-P(x_i;\beta)\big). \tag{42}$$

Now consider our supplementary sampling design under which a simple random sample of size N_1 is drawn from the overall subpopulation of participants and a simple random sample of size N_0 is drawn from the overall population of participants and non-participants. The Steinberg-Cardell estimator approximates the optimization problem in Equation (42) by scaling up the sample probabilities by the inverse of the sampling rates:

$$\max_{\beta} \sum_{i=1}^{N} s_i \left(\frac{qT}{N_1}\right) ln\left(\frac{P(x_i;\beta)}{1-P(x_i;\beta)}\right) + (1-s_i)\left(\frac{T}{N_0}\right) ln\big(1-P(x_i;\beta)\big). \tag{43}$$

Cosslett-Lancaster-Imbens estimator for an unknown prevalence rate

Cosslett (1981) has derived an alternative supplementary sampling estimator for the case of an unknown prevalence rate based on maximization of the following pseudo-likelihood function:

$$L = \sum_{i=1}^{N} s_i ln(\lambda P(x_i;\beta)) - ln\left(\lambda P(x_i;\beta) + \frac{N_0}{N}\right). \tag{44}$$

The above expression is maximized jointly over β and λ. If desired, an estimate of the prevalence rate can be obtained from the estimated value of λ by applying the normalization condition: $\left(\lambda q + \frac{N_0}{N}\right) = 1$.[20]

[20] See pp. 71-73 of Cosslett (1981) for a discussion of how to estimate prevalence rates by applying scale factors based on the relevant normalization condition for a problem. Although Cosslett imposed the restriction $\lambda > 0$ for

Alternatively, one can use this condition to re-specify the optimization problem directly in terms of β and q:

$$\max_{\beta} \max_{q\in(0,1)} \sum_{i=1}^{N} s_i ln\left(\frac{N_1}{Nq}P(x_i;\beta)\right) - ln\left(\frac{N_1}{Nq}P(x_i;\beta) + \frac{N_0}{N}\right). \quad (45)$$

This is, in fact, the same as the optimization problem that Lancaster and Imbens (1996) have derived for the case involving an unknown prevalence rate.[21]

Appendix B: Exogenously Stratified Samples

Each of the existing supplementary sampling estimators can be generalized to accommodate exogenous stratification of the primary and/or supplementary samples. Assume, for simplicity, that the observations in each sample belong to one of *B* commonly defined strata. Let the sample weights be represented by w_{1b} for stratum *b* of the primary sample and w_{0b} for stratum *b* of the supplementary sample. Assume these weights have been normalized so that they sum to the sizes of their respective samples, N_1 and N_0. In particular, let $w_{1b} = \left(\frac{T_{1b}}{N_{1b}}\right)\left(\frac{N_1}{T_1}\right)$ and $w_{0b} = \left(\frac{T_b}{N_{0b}}\right)\left(\frac{N_0}{T}\right)$, where *N* represents sample totals, *T* represents the population totals, and subscripts are used to identify subtotals associated with a specific sample stratum.[22]

The Cosslett estimator in Equation (34) for the case in which the prevalence rate is known can be generalized to accommodate this exogenously stratified sampling design as follows:

---

the maximization of the likelihood function in Equation (44), one would actually need to impose the stronger restriction $\lambda > \frac{N_1}{N}$ to insure that the estimated prevalence rate is less than one.

[21] Lele (2009) has introduced a data-cloning algorithm as an alternative to standard maximum likelihood estimation routines for this problem.

[22] If sampling strata differ across the two samples, *b* is meant to index a common set of substrata that have been constructed so that, within each sample, each member of a given substratum has a common weight.

$$\max_{\beta} \min_{\lambda_{11},\lambda_{12},\ldots,\lambda_{1B}} \sum_{b=1}^{B}\sum_{i=1}^{N_b} s_i ln(P(x_{ib};\beta)) - l\,n(\lambda_{1b}P(x_{ib};\beta) + 1 - \lambda_{1b}q_b), \quad (46)$$

where B represents the number of strata, N_b represents the combined sample size of stratum b, and λ_{1b} is the stratum-specific multiplier. The prevalence rate within stratum b (q_b) can be computed from the overall prevalence rate (q), the shares of the overall primary and supplementary samples belonging to the stratum, and the stratum-specific weights according to the formula: $q_b = \left(\frac{w_{1b}}{w_{0b}}\right)\left(\frac{N_{1b}/N_1}{N_{0b}/N_0}\right) q = \frac{T_{1b}}{T_b}$.

The Lancaster-Imbens estimator for the case of a known prevalence rate is based on the conditional probability $R(x_i;\beta,q,h)$ that an observation i was selected into the primary sample, where $R(x_i;\beta,q,h) = \frac{(h/q)P(x_i;\beta)}{(h/q)P(x_i;\beta)+(1-h)}$. Under a stratified sampling design, this probability will depend on the stratum to which the observation belongs. For an observation i from stratum b, the stratum-specific probability of the observation being drawn from the primary sample is computed as: $R(x_{ib};\beta,q_b,h_b) = \frac{(h_b/q_b)P(x_{ib};\beta)}{(h_b/q_b)P(x_{ib};\beta)+(1-h_b)}$. Therefore, to accommodate exogenous stratification, the terms q, h, and $R(x;\beta,q,h)$ in moment equations (19) through (21) would need to be made stratum specific [q_b, h_b, and $R(x_{ib};\beta,q_b,h_b)$, $b = 1,\cdots,B$].

When the prevalence rate is unknown, the generalized Cosslett-Lancaster-Imbens pseudo-likelihood function for exogenously stratified samples can be expressed as:

$$L = \sum_{b=1}^{B}\sum_{i=1}^{N_b} s_{ib} ln\left(\frac{1}{w_{1b}}\left(\frac{N_1}{Nq}\right)P(x_{ib};\beta)\right) - ln\left(\frac{1}{w_{1b}}\left(\frac{N_1}{Nq}\right)P(x_{ib};\beta) + \frac{1}{w_{0b}}\left(\frac{N_0}{N}\right)\right). \quad (47)$$

Observe that the generalized version of each of the above estimators requires knowledge of which specific stratum ($b = 1,\cdots,B$) any given observation from either sample has been

assigned.[23] In the case of the generalized Cosslett estimator for a known prevalence rate, such knowledge is necessary both to compute the stratum-specific prevalence rates (q_b) and to associate the common stratum members from each sample with their specific multiplier (λ_{1b}). In the case of the generalized Lancaster-Imbens estimator for a known prevalence rate, one needs to know the stratification assignments in order to compute the stratum-specific prevalence rates and to associate the common stratum members from each sample with their stratum-specific Bernoulli parameter h_b and stratum-specific conditional probability of assignment to the primary sample ($R(x_{ib}; \beta, q_b, h_b)$). In the case of the generalized Cosslett-Lancaster-Imbens estimator for an unknown prevalence rate, one needs to know which members from the two samples belong to a common stratum so that one can identify the corresponding weights associated with that stratum in the two samples; observe that the second expression in Equation (47) requires knowledge of both w_{0b} and w_{1b} for each member of stratum b in the combined sample.

Thus, even when the sampling strata are comparably defined in the primary and supplementary data samples, one needs to know more than just the supplied values of the sample weights in order to apply these generalized estimators. In particular, one needs to be able to identify which members from the two samples belong to a common stratum. In practice, of course, it is reasonable to expect that the sampling strata will be defined differently in the two samples. For instance, one might have a simple random primary sample and a stratified random supplementary sample. In such cases, one would need to divide one or both samples into

[23] Provided that each stratum within a sample is associated with a unique weight, knowledge of the sample weights would be sufficient to distinguish the strata *within* the sample. However, estimation of the generalized models requires aligning observations from the same stratum *across* the two samples so that they can be associated with the same stratum-specific parameters. Consequently, knowledge of the sample weights alone would not be sufficient to estimate these models even in the unrealistic case where the two samples have commonly defined strata.

substrata that are comparably defined for the two samples. To do so would require even more detailed knowledge of the sampling designs. In particular, one would need to know the specific criteria underlying the stratum assignments. Furthermore, both data samples would need to contain the variables associated with these sampling criteria in order to divide the existing strata within each sample into common sets of substrata.[24]

Unfortunately, the requisite information about the sampling criteria may not be available in practice, in which case these estimators cannot be applied. For instance, one might want to rely on data from a Census Survey, such as the CPS, ACS, or SIPP, as one's supplemental sample from the general population. For such surveys, the stratification criteria are not publicly disclosed. Moreover, even if such information were available, it would be difficult to generalize these estimators to account for the complex multi-stage stratified sampling designs underlying such surveys.

Steinberg and Cardell (1992) have proposed an extension of their estimation framework for the case of a known prevalence rate to accommodate exogenously stratified primary and/or supplementary samples. The generalized Steinberg-Cardell estimator is obtained as the solution to the following optimization problem:

$$\max_{\beta} \sum_{i=1}^{N} w_{1i} s_i \left(\frac{N_0 q}{N_1}\right) ln\left(\frac{P(x_i;\beta)}{1-P(x_i;\beta)}\right) + w_{0i}(1-s_i) ln\big(1-P(x_i;\beta)\big). \tag{48}$$

Like our new estimators, the generalized Steinberg-Cardell estimator requires knowledge only of the sample weights. However, the Steinberg-Cardell estimator has been shown to be relatively inefficient in the simulations presented in Section 6 as well as in prior studies (e.g., Lancaster and Imbens, 1996).

[24] A further complication of such an approach is the possibility of sparse or empty substrata.

Steinberg and Cardell (1992) restrict their attention to binary choice estimation. However, we demonstrate below that their approach can be extended to accommodate multinomial response problems. Using the notation presented in Section 8, let $P(m|x;\beta)$ represent a parametric specification of the probability of outcome m for $m = 0, \cdots, M$. Denote the prevalence rate associated with this outcome as q_m, and allow s_m to serve as a 1/0 indicator of presence of an observation with this outcome in the participant-only sample. A generalization of the Steinberg-Cardell estimator for this problem is obtained as the solution to the following estimation problem:

$$\max_{\beta} \left[\left(\sum_{i=1}^{N_1} \sum_{m=1}^{M} w_{1i} s_{mi} \left(\frac{N_0 q_m}{N_1} \right) ln \left(\frac{P(m|x;\beta)}{P(0|x;\beta)} \right) \right) - \sum_{j=1}^{N_0} w_{0i} ln\big(P(0|x;\beta)\big) \right]. \tag{49}$$

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 and Subhash Lele for their very 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.

Table 1: Monte Carlo Simulation Results, Prevalence Rate Known

	Steinberg-Cardell			Lancaster-Imbens			Calibrated Logit			Cosslett		
	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_0 = 400, N_1 = 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	-2.64	1.05	1.04	-2.58	1.00	1.00	-2.61	1.03	1.02	-2.61	1.03	1.03
Median	-2.59	1.01	1.01	-2.56	0.99	0.98	-2.58	1.01	1.01	-2.59	1.01	1.01
ASD	0.30	0.32	0.32	0.18	0.23	0.24	0.20	0.25	0.25	0.20	0.25	0.25
SSD	0.26	0.31	0.30	0.21	0.26	0.26	0.20	0.26	0.25	0.21	0.26	0.25
Mad	0.19	0.24	0.23	0.16	0.21	0.20	0.16	0.20	0.19	0.16	0.20	0.19
#Failures	0			0			0			0		
q = 0.25, N_0 = 400, N_1 = 100												
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	-1.53	1.04	1.05	-1.50	1.00	1.01	-1.51	1.02	1.03	-1.51	1.03	1.04
Median	-1.50	0.99	1.00	-1.49	0.99	0.99	-1.50	1.00	1.02	-1.50	1.01	1.01
ASD	0.21	0.32	0.32	0.10	0.22	0.22	0.11	0.23	0.23	0.11	0.23	0.23
SSD	0.15	0.31	0.30	0.11	0.22	0.23	0.11	0.22	0.23	0.11	0.22	0.23
Mad	0.11	0.23	0.23	0.08	0.17	0.18	0.08	0.17	0.18	0.08	0.17	0.18
#Failures	0			0			0			0		
q = 0.50, N_0 = 400, N_1 = 200												
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	0.02	1.10	1.08	0.01	1.02	1.01	0.01	1.03	1.02	0.01	1.03	1.02
Median	0.01	1.05	1.02	0.00	1.01	1.00	0.01	1.03	1.01	0.01	1.03	1.01
ASD	0.28	0.48	0.47	0.07	0.23	0.23	0.08	0.24	0.24	0.07	0.24	0.24
SSD	0.09	0.42	0.41	0.07	0.25	0.23	0.07	0.25	0.23	0.07	0.25	0.23
Mad	0.06	0.30	0.29	0.05	0.20	0.19	0.06	0.20	0.18	0.05	0.19	0.18
#Failures	2			0			0			0		
q = 0.75, N_0 = 400, N_1 = 300												
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	1.71	1.16	1.19	1.55	1.01	1.02	1.56	1.04	1.05	1.57	1.05	1.06
Median	1.53	1.01	1.01	1.52	1.01	1.02	1.54	1.03	1.03	1.54	1.03	1.04
ASD	1.33	1.20	1.21	0.23	0.35	0.36	0.24	0.34	0.35	0.24	0.35	0.35
SSD	0.58	0.76	0.75	0.26	0.38	0.38	0.24	0.34	0.36	0.25	0.35	0.36
Mad	0.38	0.54	0.53	0.19	0.30	0.30	0.18	0.27	0.28	0.18	0.27	0.28
#Failures	30			0			0			0		
q = 0.875, N_0 = 400, N_1 = 350												
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	2.96	1.02	1.10	2.75	0.97	1.01	2.81	1.02	1.06	2.82	1.04	1.08
Median	2.63	0.88	0.95	2.65	0.94	1.00	2.72	1.03	1.07	2.72	1.02	1.08
ASD	3.90	2.25	2.44	0.50	0.60	0.62	0.54	0.55	0.55	0.61	0.63	0.63
SSD	1.02	0.94	1.02	0.63	0.63	0.63	0.55	0.61	0.61	0.64	0.60	0.63
Mad	0.70	0.70	0.74	0.46	0.49	0.50	0.41	0.45	0.47	0.44	0.44	0.47
#Failures	181			0			0			0		
q = 0.875, N_0 =1,600, N_1 = 1,400												
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	2.96	1.02	1.10	2.75	0.97	1.01	2.81	1.02	1.06	2.80	1.04	1.08
Median	2.63	0.88	0.95	2.65	0.94	1.00	2.72	1.03	1.07	2.72	1.03	1.07
ASD	3.90	2.25	2.44	0.50	0.60	0.62	0.54	0.55	0.55	0.59	0.61	0.62
SSD	1.02	0.94	1.02	0.63	0.63	0.63	0.55	0.61	0.61	0.63	0.58	0.61
Mad	0.70	0.70	0.74	0.46	0.49	0.50	0.41	0.45	0.47	0.44	0.43	0.46
#Failures	181			0			0			16		

Table 2: Monte Carlo Simulation Results, Prevalence Rate Unknown

	q Known						q Unknown							
	Lancaster-Imbens			Calibrated Logit			Cosslett-Lancaster-Imbens				Pseudo-MLE			
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	q	β_0	β_1	β_2	q
q = 0.125, N_0 = 400, N_1 = 50														
Actual	-2.574	1.00	1.00	-2.574	1.00	1.00	-2.574	1.00	1.00	0.125	-2.574	1.00	1.00	0.125
Mean	-2.58	1.00	1.00	-2.61	1.03	1.02	-2.42	1.26	1.25	0.18	-2.47	1.24	1.24	0.17
Median	-2.56	0.99	0.98	-2.58	1.01	1.01	-2.40	1.15	1.17	0.16	-2.43	1.16	1.17	0.16
GSD	0.18	0.23	0.24	0.20	0.25	0.25	1.30	0.49	0.49	0.12	2.93	1.16	1.13	0.38
LSD							1.33	0.47	0.46	0.12	1.53	0.47	0.45	0.11
SSD	0.21	0.26	0.26	0.20	0.26	0.25	0.93	0.60	0.57	0.10	0.94	0.52	0.51	0.10
Mad	0.16	0.21	0.20	0.16	0.20	0.19	0.68	0.34	0.34	0.08	0.68	0.32	0.32	0.08
#Failures	0			0			288				297			
q = 0.25, N_0 = 400, N_1 = 100														
Actual	-1.492	1.00	1.00	-1.492	1.00	1.00	-1.492	1.00	1.00	0.125	-1.492	1.00	1.00	0.125
Mean	-1.50	1.00	1.01	-1.51	1.02	1.03	-1.45	1.14	1.16	0.27	-1.49	1.12	1.15	0.27
Median	-1.49	0.99	0.99	-1.50	1.00	1.02	-1.41	1.09	1.09	0.27	-1.44	1.08	1.09	0.27
GSD	0.10	0.22	0.22	0.11	0.23	0.23	0.95	0.39	0.40	0.13	2.69	0.98	1.00	0.42
LSD							0.91	0.37	0.37	0.13	0.91	0.35	0.35	0.11
SSD	0.11	0.22	0.23	0.11	0.22	0.23	0.76	0.38	0.40	0.11	0.79	0.37	0.38	0.11
Mad	0.08	0.17	0.18	0.08	0.17	0.18	0.58	0.28	0.29	0.09	0.59	0.27	0.28	0.09
#Failures	0			0			138				136			
q = 0.5, N_0 = 400, N_1 = 200														
Actual	0.00	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.50	0.00	1.00	1.00	0.50
Mean	0.01	1.02	1.01	0.01	1.03	1.02	0.09	1.13	1.12	0.50	0.02	1.11	1.10	0.49
Median	0.00	1.01	1.00	0.01	1.03	1.01	0.04	1.04	1.05	0.50	0.01	1.03	1.04	0.50
GSD	0.07	0.23	0.23	0.08	0.24	0.24	0.96	0.46	0.44	0.15	2.63	1.03	1.00	0.45
LSD							0.83	0.40	0.40	0.14	0.85	0.39	0.38	0.12
SSD	0.07	0.25	0.23	0.07	0.25	0.23	0.89	0.47	0.45	0.13	0.89	0.45	0.43	0.14

	q Known						q Unknown							
	Lancaster-Imbens			Calibrated Logit			Cosslett-Lancaster-Imbens				Pseudo-MLE			
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	q	β_0	β_1	β_2	q
Mad	0.05	0.20	0.19	0.06	0.20	0.18	0.65	0.32	0.32	0.11	0.65	0.31	0.31	0.11
#Failures	0			0			56				57			
q = 0.75, N_0 = 400, N_1 = 300														
Actual	1.492	1.00	1.00	1.492	1.00	1.00	1.492	1.00	1.00	0.75	1.492	1.00	1.00	0.75
Mean	1.55	1.01	1.02	1.56	1.04	1.05	2.10	1.30	1.33	0.72	1.91	1.25	1.25	0.72
Median	1.52	1.01	1.02	1.54	1.03	1.03	1.62	1.06	1.10	0.75	1.58	1.05	1.07	0.75
GSD	0.23	0.35	0.36	0.24	0.34	0.35	2.34	0.92	0.93	0.17	3.26	1.22	1.22	0.36
LSD							1.80	0.70	0.75	0.15	1.51	0.61	0.61	0.12
SSD	0.26	0.38	0.38	0.24	0.34	0.36	2.77	1.07	1.34	0.15	2.30	0.95	1.09	0.15
Mad	0.19	0.30	0.30	0.18	0.27	0.28	1.39	0.59	0.59	0.11	1.26	0.55	0.52	0.11
#Failures	0			0			67				61			
q = 0.875, N_0 = 400, N_1 = 350														
Actual	2.574	1.00	1.00	2.574	1.00	1.00	2.574	1.00	1.00	0.875	2.574	1.00	1.00	0.875
Mean	2.75	0.97	1.01	2.81	1.02	1.06	4.40	1.51	1.73	0.83	4.31	1.48	1.68	0.83
Median	2.65	0.94	1.00	2.72	1.03	1.07	2.86	1.11	1.08	0.88	2.89	1.10	1.08	0.88
GSD	0.50	0.60	0.62	0.54	0.55	0.55	4.61	1.40	1.63	0.19	4.26	1.44	1.48	0.26
LSD							5.38	1.59	2.02	0.17	2.93	0.99	1.04	0.13
SSD	0.60	0.62	0.63	0.55	0.61	0.61	7.44	2.15	3.25	0.14	7.30	2.23	3.11	0.15
Mad	0.45	0.48	0.50	0.41	0.45	0.47	2.86	1.02	1.16	0.09	2.83	1.01	1.14	0.10
#Failures	1			0			220				181			
q = 0.875, N_0 = 1,600, N_1 = 1,400														
Actual	2.574	1.00	1.00	2.574	1.00	1.00	2.574	1.00	1.00	0.875	2.574	1.00	1.00	0.875
Mean	2.59	0.98	0.98	2.61	1.01	1.00	2.81	1.09	1.08	0.86	2.75	1.07	1.07	0.86
Median	2.57	0.99	0.99	2.59	1.00	1.00	2.62	1.00	1.00	0.87	2.58	1.01	1.00	0.87
GSD	0.22	0.25	0.25	0.23	0.25	0.25	1.26	0.46	0.46	0.08	2.05	0.71	0.70	0.13
LSD							1.08	0.41	0.41	0.07	0.96	0.36	0.36	0.70
SSD	0.24	0.28	0.27	0.23	0.24	0.25	1.79	0.60	0.64	0.07	1.76	0.62	0.60	0.08
Mad	0.19	0.21	0.21	0.17	0.19	0.19	0.92	0.35	0.35	0.05	0.91	0.35	0.34	0.05
#Failures	0			0			15				10			

Table 3: Standard Logit vs. Supplementary Sampling Estimators of the Decision to Vote

Variable	Original Specification		Restricted Specification							
	Standard Logit		Standard Logit		Calibrated Logit		Steinberg-Cardell		Pseudo-MLE *q* unknown	
	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.
Early	-0.1845	-3.32	-0.1283	-2.18	-0.1083	-2.75	-0.1270	-2.82	-0.1108	-2.63
EDR	0.1870	2.07	0.2392	3.31	0.2745	3.65	0.3295	3.59	0.2825	3.31
Early*SDR	0.0037	0.08	0.0004	0.01	0.0336	0.71	0.0568	1.06	0.0328	0.67
Early*EDR	-0.0723	-0.57	0.0283	0.25	0.0218	0.17	0.0863	0.56	0.0198	0.15
Early*EDR*SDR	0.1292	1.58	0.2033	2.68	0.1778	2.31	0.2763	2.99	0.1807	2.22
30-Day Reg. Close	-0.1220	-2.51	-0.1048	-2.46	-0.0581	-1.54	-0.0628	-1.47	-0.0596	-1.50
ID Requirement	0.0036	0.06	-0.0090	-0.16	-0.0042	-0.10	-0.4393	- 0.09	-0.6029	-0.13
Education	0.6002	28.64	0.6277	31.93	0.7074	41.17	0.6893	32.67	0.7322	5.91
African American	0.7181	11.83	0.4030	7.09	0.6192	11.34	0.4960	8.36	0.6429	4.84
Hispanic	-0.0489	-0.48	-0.1068	-1.00	0.0600	1.11	0.0153	0.27	0.0650	1.06
Naturalized Citizen	-1.0275	-5.88	-0.5793	-8.31	-0.5242	-8.34	-0.6899	-11.03	-0.5319	-7.30
Married	0.4258	18.04	0.4619	19.06	0.8235	24.01	0.8329	21.46	0.8515	6.03
Female	0.1489	8.26	0.1693	12.08	0.2353	7.57	0.2291	6.45	0.2424	5.21
Age	0.0254	21.29	0.0237	21.89	0.0248	17.98	0.0236	14.58	0.0256	5.92
Age 18–24	0.4257	11.37	0.2141	6.23	0.3308	6.14	0.2718	4.60	0.3455	3.82
Age 75 plus	-0.1085	-2.03	-0.2443	-6.12	-0.3448	-4.95	-0.3703	-4.40	-0.3564	-3.96
Competitiveness	0.0119	4.33	0.0095	3.86	0.0121	5.22	0.0117	4.46	0.0126	4.17
South	-0.0760	-1.25	-0.0457	-0.87	-0.1154	-2.68	-0.0710	-1.44	-0.1205	-2.34
North Dakota	-0.3501	-4.28	-0.2542	-3.23	-0.2570	-1.16	-0.3112	-1.18	-0.2579	-1.11
Oregon	0.1872	4.01	0.0912	1.62	0.2453	1.89	0.3755	2.19	0.2467	1.84
Washington	-0.0204	-0.34	0.0305	0.51	0.0814	0.69	0.1634	1.15	0.0818	0.67
Self-Reported Vote	0.8231	28.51								
Natural. 10+ Years	0.4565	2.76								
Residence 1 Year	0.2681	7.58								
Income	0.0828	25.57								
Constant	-4.9878	-19.83	-3.4479	-14.49	-4.2386	-19.72	-4.0733	-16.38	-4.3398	-8.34
Estimated value of *q*									0.6484	11.12
CPS-based value of *q*	0.6362									
# Overall Sample	73,333		91,161		274,172		274,172		274,172	
# Partic. Sample	50,362		59,090		59,090		59,090		59,090	
# Suppl. Sample					215,082		215,082		215,082	

Table 4: Standard Multinomial Logit vs. Supplementary Sampling Estimators of the Decision to Vote

Vote on Election Day in Person

Variable	Original Specification		Restricted Specification							
	Standard MNL		Standard MNL		Calibrated MNL		Steinberg-Cardell		Pseudo-MLE *q* Unknown	
	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.
Early	-0.5173	-5.07	-0.4576	-4.14	-0.4294	-10.92	-0.5589	-24.66	-0.3653	-2.18
EDR	0.1368	1.52	0.1906	2.21	0.2148	2.89	0.1083	1.58	0.2194	1.93
Early*SDR	-0.3858	-3.94	-0.3932	-3.64	-0.3541	-7.30	-0.5767	-24.62	-0.2634	-1.41
Early*EDR	-0.3898	-2.38	-0.2845	-1.82	-0.3011	-2.36	-0.4730	-2.75	-0.2240	-1.08
Early*EDR*SDR	-0.1721	-1.69	-0.0928	-0.91	-0.1036	-1.34	-0.3125	-5.09	-0.0374	-0.30
30-Day Reg. Close	-0.1394	-1.68	-0.1231	-1.55	-0.0959	-2.51	-0.0718	-3.35	-0.0987	-2.10
ID Requirement	-0.0749	-0.81	-0.0895	-1.03	-0.0888	-2.03	-0.1382	-4.99	-0.0657	-1.09
Education	0.5522	24.37	0.5724	26.95	0.6470	37.01	0.2820	30.96	0.6985	3.47
African American	0.6633	9.57	0.3625	5.38	0.5597	10.18	0.1912	8.00	0.6056	3.01
Hispanic	-0.0501	-0.39	-0.1046	-0.77	0.0676	1.21	-0.0060	-0.22	0.0695	0.84
Naturalized Citizen	-1.0241	-5.53	-0.5721	-7.12	-0.5060	-7.90	-0.3619	-11.69	-0.5230	-6.07
Married	0.4624	16.67	0.4903	16.08	0.8455	24.19	0.4658	24.68	0.8792	3.28
Female	0.1166	6.65	0.1440	10.10	0.2058	6.52	0.0924	5.27	0.2202	3.34
Age	0.0189	12.39	0.0178	12.50	0.0186	13.28	0.0033	4.36	0.0216	4.27
Age 18–24	0.2708	6.36	0.0605	1.48	0.1877	3.40	-0.0941	-3.03	0.2483	2.80
Age 75 plus	-0.1958	-3.22	-0.3312	-7.68	-0.4067	-5.67	-0.2721	-6.38	-0.3987	-2.60
Competitiveness	0.0068	1.40	0.0051	1.09	0.0097	4.12	0.0024	1.78	0.1108	3.46
South	-0.2331	-1.95	-0.2056	-1.84	-0.2776	-6.40	-0.2882	-13.05	-0.2638	-2.07
North Dakota	-0.2383	-1.92	-0.1588	-1.38	-0.1888	-0.85	-0.0736	-0.19	-0.2158	-0.90
Oregon	-1.9307	-23.43	-1.9537	-19.93	-1.6540	-10.21	-2.4138	-21.03	-1.3481	-2.46
Washington	-1.5068	-17.79	-1.4311	-16.34	-1.2843	-9.33	-1.8263	-27.11	-1.0219	-2.20
Self-Reported Vote	0.8387	27.56								
Natural. 10+ Years	0.4540	2.66								
Residence 1 Year	0.3311	8.77								
Income	0.0770	19.10								
Constant	-4.0707	-9.40	-2.9532	-6.20	-3.5269	-16.06	-1.6888	-13.76	-3.9669	-8.85
Estimated value of *q*									0.4384	3.96
CPS-based value of *q*	0.4455									
# Overall Sample	73,333		91,161		273,933		273,933		273,933	
# Election Day	36,027		42,468		42,468		42,468		42,468	
# Early in Person	6,518		7,473		7,473		7,473		7,473	
# Early by Mail	7,667		8,910		8,910		8,910		8,910	
# Supp. Sample					215,082		215,082		215,082	

Vote Early in Person

Variable	Original Specification		Restricted Specification							
	Standard MNL		Standard MNL		Calibrated MNL		Steinberg-Cardell		Pseudo-MLE *q* Unknown	
	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.
Early	1.6829	4.13	1.7419	4.15	1.8617	29.16	1.6248	30.62	1.9331	11.52
EDR	-0.1572	-0.32	-0.1359	-0.27	0.4320	2.83	0.2166	0.93	0.4392	2.25
Early*SDR	1.5449	3.05	1.5848	3.06	1.8319	25.89	1.4648	28.14	1.9296	10.45
Early*EDR	1.6352	3.47	1.7582	3.65	2.0800	13.13	1.7810	5.74	2.1590	9.73
Early*EDR*SDR	1.8385	3.92	1.9231	4.07	2.2332	22.00	1.9063	18.26	2.3007	17.26
30-Day Reg. Close	0.2923	1.29	0.2945	1.26	0.4199	8.81	0.4102	13.14	0.4188	7.31
ID Requirement	-0.4379	-1.12	-0.3991	-1.02	-0.3684	-6.23	-0.4124	-8.07	-0.3439	-4.49
Education	0.7469	20.50	0.8114	22.07	0.8968	38.90	0.4621	35.93	0.9529	4.87
African American	1.1944	9.54	0.8339	7.80	1.1334	17.16	0.6482	22.33	1.1909	6.18
Hispanic	-0.0047	-0.03	-0.0413	-0.22	0.1637	2.19	0.1226	2.98	0.1629	1.63
Naturalized Citizen	-1.0959	-3.61	-0.7775	-6.12	-0.7515	-8.23	-0.5184	-10.25	-0.7637	-6.32
Married	0.4207	8.34	0.4805	12.46	0.8676	19.19	0.4302	16.54	0.9034	3.44
Female	0.2119	5.72	0.2072	7.56	0.2903	7.05	0.1597	6.70	0.3061	4.30
Age	0.0380	14.07	0.0345	13.30	0.0363	19.94	0.0177	17.47	0.0394	7.73
Age 18–24	0.5669	5.65	0.3272	3.44	0.4198	5.01	0.0770	1.59	0.4871	4.27
Age 75 plus	-0.3148	-3.96	-0.4794	-6.95	-0.6119	-6.61	-0.4349	-7.61	-0.6009	-3.64
Competitiveness	0.0422	2.01	0.0363	1.72	0.0371	11.43	0.0278	10.73	0.3848	9.30
South	1.0992	4.09	1.1374	4.07	1.2982	23.37	1.2318	31.71	1.3129	10.43
North Dakota	-0.0975	-0.31	0.0152	0.05	0.0638	0.26	0.2182	0.38	0.0438	0.16
Oregon	-0.9134	-2.04	-1.0924	-2.39	-0.2739	-0.57	-1.1670	-2.90	-0.0091	-0.01
Washington	-0.7455	-1.76	-0.7733	-1.80	-0.2162	-0.49	-0.9737	-3.11	0.0520	0.08
Self-Reported Vote	0.8745	20.42								
Natural. 10+ Years	0.3071	1.04								
Residence 1 Year	0.0659	1.09								
Income	0.1066	12.15								
Constant	-12.7293	-8.01	-10.8073	-6.64	-11.9484	-38.76	-10.9415	-47.89	-12.2765	-17.16
Estimated value of *q*									0.1029	8.89
CPS-based value of *q*	0.0911									
# Overall Sample	73,333		91,161		273,933		273,933		273,933	
# Election Day	36,027		42,468		42,468		42,468		42,468	
# Early in Person	6,518		7,473		7,473		7,473		7,473	
# Early by Mail	7,667		8,910		8,910		8,910		8,910	
# Supp. Sample					215,082		215,082		215,082	

Vote Early by Mail

Variable	Original Specification		Restricted Specification							
	Standard MNL		Standard MNL		Calibrated MNL		Steinberg-Cardell		Pseudo-MLE *q* Unknown	
	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.
Early	0.6610	1.72	0.6854	1.76	0.6341	11.67	0.4749	12.47	0.7022	4.29
EDR	-0.0728	-0.14	-0.0241	-0.05	-0.1503	-1.42	-0.2575	-1.86	-0.1497	-1.03
Early*SDR	1.5043	3.27	1.4634	3.07	1.5338	23.21	1.2309	30.30	1.6298	8.87
Early*EDR	1.1422	2.79	1.1938	2.87	1.3157	8.46	1.0582	4.78	1.3952	6.47
Early*EDR*SDR	1.2403	3.52	1.2763	3.59	1.1877	12.38	0.9369	9.72	1.2502	9.97
30-Day Reg. Close	-0.5417	-1.66	-0.4977	-1.51	-0.6267	-11.66	-0.5876	-15.49	-0.6329	-10.27
ID Requirement	0.8577	2.71	0.8069	2.63	1.0427	16.95	-0.9419	17.53	1.0710	13.75
Education	0.7245	22.38	0.7658	26.52	0.8615	38.15	0.4199	32.26	0.9183	4.70
African American	0.3408	2.52	-0.0272	-0.23	0.1493	1.91	-0.2523	-5.82	0.2067	1.06
Hispanic	-0.0909	-0.74	-0.2001	-1.74	-0.1276	-1.66	-0.1994	-4.40	-0.1236	-1.29
Naturalized Citizen	-0.9739	-4.18	-0.5166	-4.92	-0.4301	-5.29	-0.2707	-6.25	-0.4327	-3.89
Married	0.2802	6.05	0.3513	8.78	0.7246	16.50	0.3199	12.31	0.7571	2.86
Female	0.2638	10.09	0.2803	11.91	0.3640	9.03	0.2192	9.01	0.3818	5.52
Age	0.0511	16.95	0.0477	15.66	0.0510	27.06	0.0276	25.17	5.4712	10.88
Age 18–24	1.2234	8.06	1.0245	6.59	1.1358	13.77	0.6122	12.04	1.2215	11.10
Age 75 plus	0.1292	1.69	-0.0007	-0.01	-0.0654	-0.76	0.0293	0.58	-0.0554	-0.34
Competitiveness	0.0131	1.11	0.0107	0.89	0.0055	1.87	-0.0038	-1.56	0.0712	1.92
South	-0.8552	-2.36	-0.8136	-2.31	-0.9299	-16.07	-0.9242	27.60	-0.9226	-7.32
North Dakota	-1.1583	-3.50	-0.9977	-2.99	-1.0984	-4.59	-0.8634	-1.58	-1.1334	-4.35
Oregon	3.2773	10.37	3.0915	9.65	3.3895	22.45	2.3700	34.74	3.7038	7.20
Washington	2.0571	5.46	2.1073	5.51	2.0251	15.00	1.3197	24.11	2.2957	5.15
Self-Reported Vote	0.7550	19.95								
Natural. 10+ Years	0.5019	1.96								
Residence 1 Year	0.1443	2.44								
Income	0.0980	13.46								
Constant	-9.9828	-9.61	-8.3723	-7.68	-8.7091	-31.10	-7.3882	-36.28	-9.0592	-13.06
Estimated value of *q*									0.1146	8.37
CPS-based value of *q*	0.0986									
# Overall Sample	73,333		91,161		273,933		273,933		273,933	
# Election Day	36,027		42,468		42,468		42,468		42,468	
# Early in Person	6,518		7,473		7,473		7,473		7,473	
# Early by Mail	7,667		8,910		8,910		8,910		8,910	
# Supp. Sample					215,082		215,082		215,082	