ELSEVIER

Journal of Econometrics 71 (1996) 145–160

JOURNAL OF Econometrics

Case-control studies with contaminated controls

Tony Lancaster*,[a], Guido Imbens[b]

[a]*Department of Economics, Brown University, Providence, RI 02912, USA*
[b]*Department of Economics, Harvard University, Cambridge, MA 02138, USA*

(Received August 1992; final version received July 1994)

Abstract

This paper considers inference about a parametric binary choice model when the data consist of two distinct samples. The first is a random sample from the people who made choice 1, say, with all relevant covariates completely observed. The second is a random sample from the whole population *with only the covariates observed.* This is called a contaminated sampling scheme. An example might be where we have a random sample of female labor force participants and their covariate values and a second random sample of working age women, with covariates, whose participant status is unknown. We consider the cases in which the fraction of the population making choice 1 is known and that in which it is not. For both cases we give semiparametrically efficient procedures for estimating the choice model parameters.

Key words: Endogenous sampling; Choice-based sampling; Binary choice; Discrete choice
JEL classification: C13; C14; C25; C21

1. Introduction

There is a significant body of literature in statistics and econometrics dealing with discrete response models under various types of nonrandom sampling. Such sampling schemes might reduce the cost of the study, particularly if one of

* Corresponding author.

We are grateful for financial support to the National Science Foundation under grant SES 9122477, to CentER, KUB, Tilburg, for hospitality and to Steve Pischke and a referee for useful comments.

SSDI 0304407694016 98 Y

the responses is rare. A leading case is case-control, retrospective, choice-based, or response-based sampling. In the simplest example the researcher has two samples, one containing observations with response $y = 1$ (the cases) and the second containing observations with response $y = 0$ (the controls). In both samples we observe the attributes x for all observations. When the model for the conditional probabilities of the choices given the covariates is of logit form it has long been known that the investigator can proceed as though the data were obtained by random sampling so far as estimation of the covariate coefficients is concerned; see for example Prentice and Pyke (1979). For the general case Manski and Lerman (1977) proposed a weighted maximum likelihood estimator. Cosslett (1981) and Imbens (1992) proposed efficient solutions to the general estimation problem.

A case that has not received as much attention, and one that is not covered by the general sampling schemes in Hsieh, Manski, and McFadden (1985) and Imbens (1992), is that where the second sample is a random sample *from the whole population* with only the attributes or covariate values, and not the responses, observed. The second sample, that formed the control group in case-control sampling, now consists of an unknown mixture of cases and controls. Such a situation might occur if the researcher obtains a sample of observations with a particular response or disease and wishes, possibly for reasons of economy, to compare them with a random sample from a very different source in which the particular response was not measured. We describe this set-up as one of contaminated controls, following the usage of Heckman and Robb (1984). Neither sample in itself identifies the parameters of the conditional response probability, but the combination of cases and contaminated controls might do so.

This paper deals with efficient estimation of parametric discrete choice models using samples of this type. In Section 2 we discuss identifiability of choice models under contaminated sampling and point out that the choice model is nonparametrically identified if the marginal probabilities of the choices are known to the investigator. In Section 3 we give an efficient generalized method of moments (GMM) estimator for the case in which the marginal probabilities are unknown. The estimator is identical to a constrained maximum likelihood estimator when the covariates have a multinomial distribution with known support. In Section 4 we give an efficient GMM estimator for the case in which the marginal probabilities are known. This estimator is asymptotically equivalent to a constrained maximum likelihood estimator when the covariates are multinomial. The estimator proposed in Section 3 achieves the semiparametric efficiency bound as defined by Chamberlain (1987) or Begun et al. (1983). The problem is semiparametric because of the appearance in the likelihood of the unknown population covariate distribution.

In Section 5 we discuss the case in which the choice model is logit and the marginal probabilities are known. This case has been considered by Steinberg

and Cardell (1992) who have given a consistent estimator of the logit parameters. Section 6 reports a small Monte Carlo study of the estimators.

2. The model and its identifiability

Let y be a binary random response variable, equal to 0 or 1, and x a vector of attributes. In the population the distribution function of x is $F(x)$ which is unknown. We will assume that the conditional probability of $y = 1$ given x in the population is equal to $\Pr(y = 1 \mid x) = P(x; \beta)$, where $P(\cdot;\cdot)$ is a known function and β an unknown parameter. Finally, we define q to be the marginal probability of choice 1 in the population, $q = \int P(x; \beta)\, \mathrm{d}F(x)$.

The sampling scheme is that two independent random samples of sizes N_1 and N_0 are available. The first is drawn from the subset of the population who made choice 1 and the covariate is observed; the second is drawn from the whole population with only the covariate observed. We let s denote a binary stratum indicator, taking the value 1 if an observation is drawn from the subpopulation who made choice 1, and 0 if it was drawn from the whole population.

An observation from stratum 1 has probability $p(x \mid y = 1) = P(x)\, f(x)/q$; an observation from stratum 0 has probability $f(x)$. If we knew these probabilities, we could determine the function $P(x)/q$ for all values of x with positive probability. This function is therefore nonparametrically identified. It follows that the relative probabilities $P(x)/P(x_0)$ are identified. This contrasts with standard case-control sampling which identifies the relative odds, $P(x)/(1 - P(x)) \div P(x_0)/(1 - P(x_0))$.

If q is also known, then clearly $P(x)$ is identifiable. Alternatively, if the parametric form of $P(x; \beta)$ is known, then β can generally be deduced from knowledge of the function $P(x)/q$ for a sufficiently large set of values of x. In this case $P(x)$ is parametrically identifiable. In this paper we shall consider parametric models for $P(x)$ with and without prior knowledge of q. When q is known, $P(x)$ is parametrically overidentified.

3. Efficient estimation

In this section we will propose an estimator for the parameters of the conditional choice probability function $P(x; \beta)$. This function $P(x)$ will be assumed known up to a finite parameter vector β and there is no prior knowledge of the marginal probability q. In Section 4 we shall show how to take account of prior information such as knowledge of q.

To derive this estimator we will assume initially that the regressors x have a discrete distribution with unknown probabilities λ_l on $L + 1$ known points of

support, x^l. This allows us to use standard maximum likelihood theory, and to derive an efficient estimator for that case. This estimator does not depend on either the number or the location of points of support of the covariate distribution that do not appear in the sample. We then show that this estimator is asymptotically semiparametrically efficient.

It is convenient, first of all, to enlarge the model. We do this by supposing that the sample sizes were determined by a sequence of Bernoulli trials with unknown parameter h which is functionally independent of the other parameters β and λ. Thus the data is provided by repeatedly conducting such trials; if a success occurs, we randomly sample from the subpopulation who made choice 1; if a failure, we randomly sample from the whole population. This procedure is repeated N times. The population is assumed sufficiently large that the probability of overlap between the sampled individuals is zero. A consequence of this enlargement is that the sample now constitutes N independently and identically distribution realisations from the joint distribution of stratum and covariate $g(s, x) = (hPf/q)^s((1-h)f)^{1-s}$. The quantity h will be treated as an unknown parameter. Its maximum likelihood estimator will be the sample fraction of observations from stratum 1, N_1/N. As long as h is functionally independent of β, N_1/N is ancillary and the asymptotic distribution of the ML estimator of β is independent of that of h.

If $N = N_1 + N_0$ is the total number of observations, the log-likelihood is

$$L(\beta, h, \lambda) = \sum_{n=1}^{N} \left[s_n \log[P_n(\beta) f_n(\lambda)/q(\beta, \lambda)] + (1 - s_n) \log f_n(\lambda) \right] + N_1 \log h + N_0 \log(1-h), \tag{3.1}$$

where $f_n(\lambda) = f(x_n; \lambda)$ and $P_n(\beta) = P(x_n; \beta)$. Since L involves β and λ in a rather awkward way because of the term in q, it is convenient to reparametrize. The following transformation changes the log-likelihood into the form that would arise under a random sampling scheme in which there exists a conditional distribution and a marginal distribution each depending on distinct sets of parameters.

Define

$$R_1(x; \beta, q, h) = \frac{(h/q)P(x;\ \beta)}{(h/q)P(x;\ \beta) + 1 - h}, \quad R_0 = 1 - R_1,$$

$$g(x) = [(h/q)P(x;\beta) + 1 - h] f(x). \tag{3.2}$$

R_1 is the conditional probability that an observation comes from stratum 1 given the covariate and the sampling scheme. The distribution $g(x)$, which is also multinomial with parameters $\pi_l = [(h/q)P(x^l;\beta) + 1 - h]\lambda_l$ on the same point of support as $f(x)$, is the covariate distribution induced by the sampling

scheme. Then L may be rewritten as

$$L(\beta, q, h, \pi) = \sum_{n=1}^{N} [s_n \log R_{1n}(\beta, q, h) + (1 - s_n) \log R_{0n}(\beta, q, h)] + \sum_{n=1}^{N} \log g_n(\pi) = L_1(\beta, q, h) + L_2(\pi). \tag{3.3}$$

Formally, we can regard (3.3) as the log-likelihood corresponding to a random sample from a population in which the covariate has a multinomial distribution with probabilities π and the conditional probability of choice 1 is $R_1(\beta, q, h)$. But (3.3), which is just a rewriting of (3.1), appears to have one more parameter than (3.1). This is false, of course, because the parameters β, q, h, and π are subject to the constraint that $q = \int P(x;\beta)\, \mathrm{d}F(x;\lambda)$ which may be rewritten in terms of the new parametrization as

$$h = \int R_1(x;\beta, q, h)\, \mathrm{d}G(x;\pi). \tag{3.4}$$

So to maximize this expression we must, in principle, take account of the constraint (3.4). But in fact the values of β, q, h, and π which maximize the log-likelihood without imposing the constraint do, in fact, satisfy the constraint. This implies that to compute maximum likelihood estimates of the choice model parameters all we need to do is to maximize the first component of (3.3). This is just a random sampling binary choice log-likelihood with choice probabilities given by (3.2). We now show that this is so by examining the unconstrained maximum likelihood estimators of β, q, h, and π.

Let a hat denote an estimator which maximizes L without imposing the restriction (3.4). Then $\hat{\pi}_l = n_l/N$ for all l, where n_l is the sample number of observations which have covariate value x^l. At this solution for π the constraint (3.4) becomes

$$h = N^{-1} \sum_{n=1}^{N} R_{1n}(\beta, q, h). \tag{3.5}$$

Next consider the β, q, and h likelihood equations from L_1:

$$\frac{\partial L_1}{\partial \beta} = \sum_{n=1}^{N} p'_{\beta n}(s_n - R_{1n}(\beta, q, h))/P_n = 0, \tag{3.6}$$

$$\frac{\partial L_1}{\partial q} = -(1/q) \sum_{n=1}^{N} (s_n - R_{1n}(\beta, q, h)) = 0, \tag{3.7}$$

$$\frac{\partial L_1}{\partial h} = (1/\bar{h}) \sum_{n=1}^{N} (s_n - R_{1n}(\beta, q, h)) = 0. \tag{3.8}$$

Here $p_{\beta n} = \partial P_n/\partial\beta$ of order $1 \times K$, where K is the dimension of β and $\bar{h} = h(1-h)$.

Let $\hat{\beta}$ and $\hat{q}$ solve (3.6) and (3.7) with $h = \hat{h} = N_1/N$. Then $\hat{\beta}$, $\hat{q}$, and $\hat{h}$ solve (3.6), (3.7), and (3.8), and they also satisfy the constraint (3.5) which may be written $N^{-1}\sum(s_n - R_{1n}(\hat{\beta}, \hat{q}, \hat{h})) = 0$. Hence the constrained ML estimator of β and q can be found by maximising $L_1(\beta, q, \hat{h})$ with respect to variation in β and q. Since L_1 is just a random sampling binary choice log-likelihood, this is an essentially simple computation.

The above derivation gives $\hat{\beta}$ as a constrained ML estimator after a parameter transformation. It may also be given a generalized method of moments (GMM) interpretation.[1] Consider the generalized moments

$$\begin{aligned}
&\psi_1(\beta, q, h, s, x) = p'_\beta(x;\beta)(s - R_1(x;\beta, q, h))/P(x;\beta),\\
&\psi_2(\beta, q, h, s, x) = -(1/q)(s - R_1(x;\beta, q, h)),\\
&\psi_3(\beta, h, q, s, x) = q - P(x;\beta)/[(h/q)P(x;\beta) + 1 - h] \propto h - R_1(x;\beta, q, h).
\end{aligned} \tag{3.9}$$

The moments ψ_1, and ψ_2 are the single observation scores for β and q from the log-likelihood L_1, (3.3). In the form $q - P/[(h/q)P + 1 - h]$ the moment ψ_3 is just the definitional relation between marginal, q, and conditional, $P(x)$, choice probabilities after allowing for the fact that the covariate distribution induced by the sampling scheme is not $f(x)$, but $g(x) = f(x)[(h/q)P + 1 - h]$. In the form $h - R_1$ the moment ψ_3 is the single-observation version of the constraint (3.4). These moments have mean zero at the true parameter point. Equating their sample analogues to zero gives $\hat{\beta}$, $\hat{q}$ and $\hat{h}$ which are then GMM estimates. Thus the asymptotic distribution of the estimator may be found equivalently from GMM theory or from constrained ML theory. The former is rather simpler since we do not have to consider the estimation of π. Moreover note that these are valid moments whether the distribution of x is discrete or continuous so they do not hinge on the assumption of a discrete covariate with known support.

Theorem 1. Let $\delta = (\beta, q, h)$ *and* $\psi = (\psi_1, \psi_2, \psi_3)$, *where* $\psi_3 = h - R_1(x;\beta, q, h)$. *Under regularity conditions, the solution,* $\hat{\delta}$ *to* $\sum_{n=1}^{N}\psi_n(\hat{\delta}) = 0$ *is a consistent estimator for* δ^* *and* $\sqrt{N}(\hat{\delta} - \delta^*) \to \mathcal{N}(0, V)$ *with*

$$V = \Gamma^{-1}\Delta(\Gamma')^{-1}, \quad \Delta = \mathscr{E}[\psi(\delta)\cdot\psi(\delta)']_{\delta=\delta^*}, \quad \Gamma = \mathscr{E}\left[\frac{\partial\psi}{\partial\delta\delta'}\right]_{\delta=\delta^*},$$

where $\mathscr{E}[.]$ *denotes expectation taken over the distribution induced by the*

[1] See Hansen (1982) and Manski (1988).

sampling scheme, $g(s, x) = (hPf/q)^s((1-h)f)^{1-s}$, *and an asterisk denotes the true value. The above covariance matrix is the semiparametric efficiency bound of Chamberlain* (1987) *or Begun, Hall, Huang, and Wellner* (1984).

Proof. See Appendix.

An explicit form for the asymptotic covariance matrix of $\hat{\beta}$ and $\hat{q}$ is as follows. Let

$$\Delta_{11} = \mathscr{E}(p'_\beta \bar{R} p_\beta / P^2), \quad \Delta_{12} = -(1/q)\mathscr{E}(p'_\beta \bar{R}/P),$$
$$\Delta_{22} = (1/q^2)\mathscr{E}(\bar{R}), \qquad \Delta_{33} = \bar{h} - \mathscr{E}(\bar{R}), \tag{3.10}$$

which are the nonzero elements of Δ. Here $\bar{R} = R_1(1 - R_1)$ and the expectation is with respect to $g(x)$, defined in (3.2). Then the limiting covariance matrix of $\hat{\beta}, \hat{q}$ is

$$V(\hat{\beta}, \hat{q}) = \Delta_1^{-1} - \begin{pmatrix} 0 & 0 \\ 0 & q^2/\bar{h} \end{pmatrix} \quad \text{where} \quad \Delta_1 = \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{pmatrix}.$$

The variance of $\hat{h}$ is $\bar{h}$ and it is distributed independently of $\hat{\beta}$ and $\hat{q}$.

We see that the covariance matrix of β can be found from the upper left submatrix of Δ_1^{-1} which is the inverse information matrix for β and q from L_1. This means that (a) an efficient estimate of β can be found by maximizing the binary choice log-likelihood, L_1, with respect to β and q with h replaced by N_1/N, and (b) the standard inverse information matrix estimate of the $\hat{\beta}$ and $\hat{q}$ covariance matrix will give the correct standard errors for $\hat{\beta}$ (though not for $\hat{q}$).

4. Efficient estimation with known q

Suppose that extra sample information provides the numerical value of the marginal choice probability, q^*. One way of proceding is to maximize the log-likelihood (3.3) subject to the constraint provided by knowledge of q^*. The log-likelihood becomes

$$\begin{aligned} L(\beta, h, \pi) &= \sum_{n=1}^{N} [s_n \log R_{1n}(\beta, q^*, h) + (1 - s_n) \log R_{0n}(\beta, q^*, h)] \\ &\quad + \sum_{n=1}^{N} \log g_n(\pi) \\ &= L_1(\beta, h) + L_2(\pi). \end{aligned} \tag{4.1}$$

The constraint relating β, h, and π is $q^* = \int P(x;\beta)\,\mathrm{d}F(x;\lambda)$, which is equivalent to

$$h = \int R_1(x;\beta, h)\,\mathrm{d}G(x;\pi). \tag{4.2}$$

Here, $R_1 = (h/q^*)P/[(h/q^*)P + 1 - h]$. The ML estimator of β, h, and π maximizes (4.1) subject to (4.2). Unlike the case in which q was unknown, it is no longer true that the unconstrained ML estimator satisfies the constraint, so this simplification no longer applies. But a constrained optimization can be avoided if we adopt a Generalized Method of Moments approach.

Consider the moments ψ with q replaced q^*. These are

$$\begin{aligned}
&\psi_1(\beta, q^*, h, s, x) = p'_\beta(x;\beta)(s - R_1(x;\beta, q^*, h))/P(x;\beta),\\
&\psi_2(\beta, q^*, h, s, x) = -(1/q^*)(s - R_1(x;\beta, q^*, h)),\\
&\psi_3(\beta, q^*, h, s, x) = h - R_1(x;\beta, q^*, h).
\end{aligned} \tag{4.3}$$

The covariance matrix of these moments is Δ whose elements were given in (3.10). Then:

Theorem 2. Let $\psi_n = \psi(\beta, q^*, h, s_n, x_n)$, $\delta_1 = (\beta, h)$, $\Delta = \mathscr{E}(\psi\psi')$, *and* $\Gamma_1 = \mathscr{E}(\partial\psi/\partial\delta)$, *where* Γ_1 *is a submatrix of the* Γ *of Theorem 1 – the column corresponding to q has been deleted. Finally, let* $\hat{\delta}_1$ *minimize*

$$\sum_{n=1}^{N} \psi_n(\delta_1)\Delta^{-1}\psi_n(\delta_1).$$

Then $\sqrt{N}(\hat{\delta}_1 - \delta_1^*) \to \mathcal{N}(0, V_1)$, *where*

$$V_1 = [\Gamma_1'\Delta^{-1}\Gamma_1]^{-1}.$$

This covariance matrix is the same as that of the estimator of β and h which maximizes (4.1) subject to (4.2), under the usual regularity conditions. Thus the GMM estimator is asymptotically equivalent to the ML estimator and is efficient when the covariate is discrete with known points of support. Semiparametric efficiency of $\hat{\beta}$ can be proved using the arguments in Imbens (1992, Thm. 2).

Notice the simplicity of the GMM procedure. It avoids estimation of the covariate distribution; it avoids a constrained optimization problem; and it is a procedure that can be applied without any restrictive assumption about the covariate distribution.

An explicit form for the asymptotic covariance matrix of $\hat{\beta}$ is[2]

$$V(\hat{\beta}) = \Delta_{11}^{-1} - \Delta_{11}^{-1}\Delta_{12}[\Delta_{21}\Delta_{11}^{-1}\Delta_{21} + (\bar{h}/q^2) - \Delta_{22}]^{-1}\Delta_{21}\Delta_{11}^{-1}. \tag{4.4}$$

The corresponding expression when q is not known is found from (3.11) to be

$$V(\hat{\beta}) = \Delta_{11}^{-1} - \Delta_{11}^{-1}\Delta_{12}[\Delta_{21}\Delta_{11}^{-1}\Delta_{21} - \Delta_{22}]^{-1}\Delta_{21}\Delta_{11}^{-1}. \tag{4.5}$$

---

[2] $\sqrt{N}(\hat{h} - h)$ is distributed independently of $\hat{\beta}$ with variance $\bar{h}$.

The feasible form of the estimator will require an initial consistent estimate of δ in order to estimate the covariance matrix Δ. This might be provided by the estimator which solves

$$\sum_{n=1}^{N} \psi_1(\beta, q^*, \hat{h}, s_n, x_n) = 0. \tag{4.6}$$

This uses only the first moment, which is the score from the conditional likelihood of s given x with h replaced by N_1/N. It is similar to Manski and McFaddens' (1981) conditional maximum likelihood estimator in the standard case-control or choice-based sampling set up. The asymptotic covariance matrix of this estimator is

$$V(\hat{\beta}_{\mathrm{CML}}) = \Delta_{11}^{-1} - \Delta_{11}^{-1}\Delta_{12}[\bar{h}/q^2]^{-1}\Delta_{21}\Delta_{11}^{-1}. \tag{4.7}$$

This estimator is distributed independently of $\hat{h}$. Its inefficiency is revealed by comparison with (4.4) since $\Delta_{22} - \Delta_{21}\Delta_{11}^{-1}\Delta_{12}$ is nonnegative definite.

5. The logit case

The logit model for P is of interest since it is widely used and there are known simplifications under this model in standard case-control sampling. The model is

$$P(x;\beta) = 1/(1 + \exp\{\beta_0 + \beta_1' x\}).$$

Under standard case-control sampling the conditional probability of choice 1 given the covariate *and* the sampling scheme is

$$R_1(x;\beta) = 1/(1 + \exp\{\beta_0 + \log[q(1-h)/h(1-q)] + \beta_1' x\}),$$

which is the original logit model with intercept displaced. This is the reason why under standard case-control sampling with a logit model an investigator can proceed *as if* the data had been obtained by random sampling so far as inference about the covariate effects is concerned. But in the present application the conditional probability of stratum 1 given the covariate and the sampling scheme is

$$R_1(x;\beta) = 1/(1 + [q(1-h)/h] + \exp\{\beta_0 + \log[q(1-h)/h] + \beta_1' x\}).$$

This is not a logit model. Thus it would be incorrect for an investigator to proceed to make inferences about covariate effects as if the data originated in random sampling.

Steinberg and Cardell (1992) have suggested an estimator for the logit model when q is known. They propose choosing β to maximize

$$L_{\mathrm{SC}} = \sum_{n=1}^{N} (1 - s_n)\log(1 - P_n(\beta)) + \omega s_n \log[P_n(\beta)/(1 - P_n(\beta))]. \tag{5.1}$$

Here $\omega = q^*(1 - \hat{h})/\hat{h}$. In this section we shall give an interpretation of the Steinberg and Cardell (SC) estimator and comment on its properties.[3] In the next Section we report some Monte Carlo comparisons of this estimator and the efficient procedure.

Consider a two-stage estimation procedure. In the first stage, a nonparametric estimate of the population joint distribution of choice and covariate is constructed. In the second stage, an estimate of β is formed by minimizing the Kullback–Leibler (KL) distance between the nonparametric estimate and a proposed parametric (logit) model. Let $(y, x) = f(x)\, P(x)^y [1 - P(x)]^{1-y}$, the population joint distribution of choice and covariate. The second stage therefore minimizes

$$C = \sum_{y,x} \hat{u}(y, x) \log[\hat{u}(y, x)/u(y, x; \beta)], \tag{5.2}$$

where $\hat{u}$ is the nonparametric estimate and $u(y, x; \beta)$ is the parametric model with a logit form for $P(x)$ depending on the parameter β. Dropping terms from (5.2) which do not involve β it may be written

$$\begin{aligned} C &= \sum_l \hat{P}_l \hat{f}_l \log P_l(\beta) + (1 - \hat{P}_l)\, \hat{f}_l \log(1 - P_l(\beta)) \\ &= \sum_l \hat{f}_l \log(1 - P_l(\beta)) + \hat{P}_l \hat{f}_l \log[P_l(\beta)/(1 - P_l(\beta))]. \end{aligned} \tag{5.3}$$

In this expression, $f_l = f(x^l)$, $P_l = P(x^l)$, and a caret indicates the nonparametric estimate.

Now consider nonparametric estimation of P and f. The log-likelihood (3.3) with $g(x)$ multinomial leads us to such estimates. The ML estimate of $g(x)$ is $\hat{\pi}_l = n_l/N$. The nonparametric estimate of $R_1(x^l) = \hat{R}_{1l}$ is n_{1l}/n_l. Here n_l is the number of observations having covariate value x^l and n_{1l} is the number of observations having covariate x^l and originating from stratum 1. n_{0l} is similarly defined. Note that these estimates do satisfy the constraint (3.4) or (3.5) when $\hat{h} = N_1/N$, so they do in fact maximize the constrained log-likelihood. They do not, of course, make any use of the fact that q is known.

The definitions (3.2) and the definition of $\omega = q^*(1 - \hat{h})/\hat{h}$ enable us to go from estimates of R_1 and g to estimates of f and Pf which are

$$\hat{P}_l \hat{f}_l = \omega n_{1l}/N_0, \qquad \hat{f}_l = n_{0l}/N_0. \tag{5.4}$$

---

[3] Steinberg and Cardell actually study a slightly different case where the population is finite, and the two samples, one containing observations with $y = 1$ and one randomly from the whole population, may partially overlap. The model we study can be viewed as a limit of their framework where the size of the population goes to infinity. They also gave a quite different justification for their estimator than the one which follows.

Note that the nonparametric estimator of $f(x)$ is the sample distribution from stratum 0, the random sample.

Inserting these estimates into the KL measure, (5.3), gives

$$C = N_0^{-1} \sum_l n_{0l} \log(1 - P_l(\beta) + \omega n_{1l} \log(P_l(\beta/(1 - P_l(\beta))$$

$$= N_0^{-1} \sum_{n=1}^{N} (1 - s_n) \log(1 - P_n(\beta)) + \omega s_n \log[P_n(\beta)/(1 - P_n(\beta))]. \quad (5.5)$$

This is proportional to the Steinberg and Cardell criterion function, (5.1).

While the preceding argument is formally correct, it suffers from the difficulty that the implicit 'nonparametric ML' estimate of P may lie outside the interval zero to one. This is obvious from the relation between R_1 and P given in (3.2) where, even though $\hat{R}_1$ is a proper probability, there is no guarantee that $\hat{P}$ is. This suggests that the Steinberg–Cardell estimator may behave poorly in small samples, even though when P is logit the criterion function (5.1) is globally concave.

It is interesting the look at the form of the Steinberg–Cardell estimator in more detail, as it explains some of the finding of the Monte Carlo study. Suppose that x is a scalar random variable, taking on two values, 0 and 1. Also, assume that β_0 is known. The first-order condition for maximization of L_{SC} is

$$L_{SC}^{\beta} = \sum_{n=1}^{N} x_n[\omega s_n - (1 - s_n)P_n] = 0.$$

Since x is binary this becomes

$$L_{SC}^{\beta} = \omega S - P(N_1 - S) = 0, \quad (5.6)$$

where S is the number of the N_1 observations from stratum 1 having covariate value 1 and $P = 1/(1 + \exp\{\beta_0 + \beta_1\})$. Conditional on $x = 1$, S is Binomial $(N_1, R_1(1; \beta, q, h))$.

Eq. (5.6) will have a finite solution for β_1 if and only if $\omega S \leqslant N_1 - S$, an event of probability less than 1. As a particular example suppose that $\beta_1^* = 0$ and that equal numbers of observations come from each stratum, $h = 0.5$. Then, using the Normal approximation to the Binomial, we find

$$\Pr(\text{no finite solution for } \hat{\beta}_1) = 1 - \Phi\left(\sqrt{N_1}\frac{1 - q}{1 + q}\right).$$

Some values of this probability are given in Table 1.

Under these circumstances the efficient GMM estimator can be expected to perform much better. The third moment compares q to the average value of $P(x; \beta)/(hP(x; \beta)/q + 1 - h)$. In this case with β_1 close to 0, this moment has very little variance and gives an almost exact restriction on β_1. This information is

Table 1
Probabilities of no solution

N	N_1	q	Probability
100	50	0.90	0.355
400	200	0.90	0.228
1000	500	0.90	0.120
100	50	0.80	0.216
400	200	0.80	0.058
1000	500	0.80	0.0065

not used by the Steinberg–Cardell estimator. This is of course no proof that the GMM estimator will in fact perform better in practice. It relies on a first round of consistent estimates to get an estimate of the optimal weight matrix. The choice of the first-round weight matrix does not matter asymptotically, but there is no guarantee that the first-round estimator will actually converge. In practice, however, we had no difficulty in obtaining convergence for the GMM estimator using prior knowledge of q.

We did have occasional convergence problems with the GMM estimator *for the unknown q case*, especially where the true value of q was close to 0. This is not surprising because when q is close to 0 the sampling is close to pure choice-based sampling in which one subsample is chosen from those who make choice 0 and one from those who make choice 1. But under pure choice-based sampling, with a logit model, the intercept is not identified when q is unknown. While all parameters are formally identified in our Monte Carlo experiment, with $q = 0.2$ we may be close enough to the nonidentified case that convergence problems occur.

In the Monte Carlo experiment x was choosen to have a bivariate normal distribution with zero means, unit variance, and zero correlation. Three sets of parameter values were used: $(\beta_0, \beta_1, \beta_2)$ equal to $(0, 1, 1)$, $(0, 2, 0.5)$, and $(-1.89, 1, 1)$. The implied values for q were 0.5, 0.5, and 0.2. h was fixed at 0.5. The number of observations was in all simulations equal to 400. The number of replications was equal to 200 for each experiment. We report the averages of the 200 estimates (mean), the average of the asymptotic standard devations (asd), the standard deviation of the 200 replications (ssd), the median, and the median of the absolute deviation from the median (mad). The results are reported in Tables 2 to 4.

The SC estimator performed significantly worse than the efficient GMM estimator proposed in this paper. In fact, in the first and third set of simulations 9 and 27 of the replications did not lead to convergence. The GMM estimator without knowledge of q did not converge for 11, 2, and 19 of the simulations.

Table 2
Design I: $\beta_0 = 0.0$, $\beta_1 = 1.0$, $\beta_2 = 1.0$, $q = 0.5$, $h = 0.5$

	GMM (unknown q)			GMM (known q)			SC		
Failure to converge	11			0			9		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Mean	0.06	1.18	1.18	0.02	1.04	1.04	0.10	1.20	1.22
Asd	33.98	14.43	11.41	0.10	0.26	0.26	0.42	0.77	0.77
Ssd	0.98	0.48	0.49	0.10	0.29	0.27	0.34	0.64	0.61
Med	0.06	1.09	1.11	0.01	1.01	1.02	0.04	1.05	1.10
Mad	0.66	0.32	0.30	0.06	0.19	0.15	0.20	0.30	0.31

Table 3
Design II: $\beta_0 = 0.0$, $\beta_1 = 2.0$, $\beta_2 = 0.5$, $q = 0.5$, $h = 0.5$

	GMM (unknown q)			GMM (known q)			SC		
Failure to converge	2			0			27		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Mean	−0.01	2.15	0.52	0.00	2.03	0.50	0.08	2.58	0.70
Asd	17.00	16.35	5.02	0.13	0.38	0.25	0.92	5.15	2.10
Ssd	0.81	0.66	0.31	0.13	0.38	0.26	0.46	1.98	0.93
Med	−0.01	2.04	0.46	−0.01	1.98	0.49	0.00	2.10	0.51
Mad	0.46	0.33	0.17	0.08	0.26	0.17	0.23	0.62	0.24

Table 4
Design III: $\beta_0 = -1.89$, $\beta_1 = 1.0$, $\beta_2 = 1.000$, $q = 0.2$, $h = 0.5$

	GMM (unknown q)			GMM (known q)			SC		
Failure to converge	19			0			0		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Mean	−1.89	1.12	1.10	−1.87	1.04	1.03	−1.92	1.09	1.06
Asd	133.68	61.76	61.90	0.09	0.18	0.18	0.20	0.32	0.32
Ssd	0.75	0.31	0.26	0.10	0.20	0.18	0.20	0.36	0.36
Med	−1.77	1.09	1.09	−1.87	1.04	1.03	−1.92	1.03	1.01
Mad	0.39	0.16	0.18	0.06	0.12	0.13	0.12	0.20	0.17

There were no problems with convergence of the GMM estimator with known q. The standard errors for the unknown q GMM estimator and the Steinberg–Cardell estimator reflect the convergence problems: they are markedly different from what one would expect given normality and given the median deviation from the mean. The finite-sample properties of the known q GMM estimator seem satisfactory and reflect its theoretical asymptotic superiority to the Steinberg–Cardell estimator when the model is correctly specified.

6. Summary and conclusions

We have given computationally simple and asymptotically efficient estimators in the contaminated sampling problem. When the marginal choice probability, q, is unknown the estimator maximizes a binary-choice log-likelihood and, if the covariate distribution is multinomial with known support, it is interpretatable as a constrained maximum likelihood estimator. When the marginal choice probability is known the estimator solves a generalized method of moments problem. When the covariate distribution is multinomial with known support, the estimator is asymptotically equivalent to a constrained maximum likelihood estimator. We also gave explicit forms for the asymptotic covariance matrices in both cases as well as for a conditional likelihood estimator applicable when q is known. Additional a priori information can be readily incorporated into the GMM procedure as long as it is expressible as a moment condition. Imbens and Lancaster (1992) gives further examples of this.

We have also discussed the logit model as a special case and compared numerically the properties of the estimators proposed in this paper with an alternative method suggested by Steinberg and Cardell (1992) which is applicable when q is known. When q is known, the efficient generalized method of moments estimator exhibited satisfactory performance. The estimator of Steinberg and Cardell failed to exist in a significant fraction of simulations, as did the efficient GMM procedure in the absence of knowledge of the marginal choice probability.

Appendix: Outline of proofs of Theorems 1 and 2

Consistency and asymptotic normality of the GMM estimators, both when q is known and when it is unknown, can be proved in a generalized method of moments framework as described by Hansen (1982) and Manski (1988). For instance, Theorems 2.1 and 3.1 in Hansen (1982) prove consistency and asymptotic normality for generalized method of moments estimators. Conditions that ensure that the regularity conditions for these theorems are satisfied are: (i) compactness of sample and parameter spaces (with true parameters interior to

the parameter space), (ii) continuity of $P(x;\beta)$ and its derivative with respect to β, (iii) uniqueness of the solution to $\mathscr{E}(\psi(\delta)) = 0$, and (iv) full rank of Δ and Γ.

The estimator of Theorem 1 was derived initially for the case in which x has a discrete distribution with known finite support. The estimator was shown to be a maximum likelihood estimator in that case and therefore achieves the Cramer–Rao bound for regular estimators. This result can be extended to the continuous regressor case using the approach to semiparametric efficiency bounds of Begun, Hall, Huang, and Wellner (1984).

From (3.1) the log-density of a single observation is

$$\log g(s, x) = s \log P(x;\beta) - s \log q + s \log h + (1 - s)\log(1 - h) + \log f(x). \tag{A.1}$$

Consider a parametric submodel in which the unknown density $f(.)$ is parametrized by η. In this submodel the scores for β and η are

$$S_\beta = s(p'_\beta/P - q_\beta/q), \qquad S_\eta = -s(q_\eta/q) + f_\eta/f. \tag{A.2}$$

The tangent set, $\mathscr{T}$,[4] is of the form

$$d(x) - s(\mathscr{E}(\mathrm{d}(x))/h)$$

where $d(x)$ is unrestricted apart from the requirement that $\int d(x)\,\mathrm{d}F(x) = 0$. The efficient score for β is

$$S^* = (s - R(x;\beta)(p'_\beta/P + \delta/q),$$

where

$$\delta = -q\frac{\mathscr{E}(p'_\beta \bar{R}/P)}{\mathscr{E}(\bar{R})} = -\Delta_{12}\Delta_{22}^{-1}.$$

The inverse of the covariance matrix of S^* is the variance of the GMM estimator described in Theorem 1.

The claim in Theorem 2 that the GMM estimator is asymptotically equivalent to the constrained ML estimator when the covariate distribution is discrete with known support is established by direct calculation using classical results on the covariance matrix of the constrained maximum likelihood estimator. The general form of this result is given in Lemma 1 of Imbens (1992). The extension to the continuous covariate case can be based on Theorem 2 of that paper.

[4] See Newey (1990).

References

Begun, J.M., W.J. Hall, W.M. Huang, and J.A. Wellner, 1983, Information and asymptotic efficiency in parametric–nonparametric models, Annals of Statistics 11, 432–452.

Breslow, N.E. and N. Day, 1980, Statistical methods in cancer research, 1: The analysis of case-control studies (IARC, Lyon).

Chamberlain, G., 1987, Asymptotic efficiency in estimation with conditional moment restrictions, Journal of Econometrics 34, 305–334.

Cosslett, S.R., 1981, Efficient estimation of discrete choice models, in: C.F. Manski and D. McFadden, eds., Structural analysis of discrete data with econometric applications (MIT Press, Cambridge, MA).

Cox, D. and D. Hinkley, 1974, Theoretical statistics (Chapman and Hall, London).

Hansen, L.P., 1982, Large sample properties of generalized method of moment estimators, Econometrica 50, 1029–1054.

Heckman, J.J. and Robb, 1984, Alternative methods for evaluating the impact of interventions , in: J.J. Heckman and B. Singer, eds., Longitudinal analysis of labor market data (Cambridge University Press, Cambridge).

Hsieh, D.A., C.F. Manski, and D. McFadden, 1985, Estimation of response probabilities from augmented retrospective observations, Journal of the American Statistical Association 80, 651–662.

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

Imbens, G.W. and T. Lancaster, 1991, Efficient estimation and stratified sampling, Journal of Econometrics, forthcoming.

Imbens, G.W. and T. Lancaster, 1994, Combining micro and macro data in microeconometric models, Review of Economic Studies 61, 655–680.

Manski, C.F., 1988, Analog estimation methods in econometrics (Chapman and Hall, New York, NY).

Manski, C.F. and S.R. Lerman, 1977, The estimation of choice probabilities from choice-based samples, Econometrica 45, 1977–1988.

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

Newey, W.K., 1990, Semiparametric efficiency bounds, Journal of Applied Econometrics 5, 99–135.

Prentice, R.L. and R. Pyke, 1979, Logistic disease incidence models and case-control studies, Biometrika 66, 403–411.

Steinberg, D. and N.S. Cardell, 1992, Estimating logistic regression models when the dependent variable has no variance, Communications in Statistics – Theory and Methods 21, 423–450.