Endogenous Sampling and Matching Method in Duration Models

Takeshi Amemiya and Xinghua Yu

Endogenous sampling with matching (also called "mixed sampling") occurs when the statistician samples from the non-right-censored subset at a predetermined proportion and matches on one or more exogenous variables when sampling from the right-censored subset. This is widely applied in the duration analysis of firm failures, loan defaults, insurer insolvencies, and so on, due to the low frequency of observing non-right-censored samples (bankrupt, default, and insolvent observations in respective examples). However, the common practice of using estimation procedures intended for random sampling or for the qualitative response model will yield either an inconsistent or inefficient estimator. This paper proposes a consistent and efficient estimator and investigates its asymptotic properties. In addition, this paper evaluates the magnitude of asymptotic bias when the model is estimated as if it were a random sample or an endogenous sample without matching. This paper also compares the relative efficiency of other commonly used estimators and provides a general guideline for optimally choosing sample designs. The Monte Carlo study with a simple example shows that random sampling yields an estimator of poor finite sample properties when the population is extremely unbalanced in terms of default and non-default cases while endogenous sampling and mixed sampling are robust in this situation.

Keywords: Duration models; Endogenous sampling with matching; Maximum likelihood estimator; Manski-Lerman estimator; Asymptotic distribution

JEL Classification: C13, C24, C41

Takeshi Amemiya: Edward Ames Edmonds Professor of Economics, Stanford University, Department of Economics (E-mail: amemiya@stanford.edu)

Xinghua Yu: Stanford University, Department of Economics (E-mail: xhyu@stanford.edu)

This paper was prepared in part while Takeshi Amemiya and Xinghua Yu were at the Institute for Monetary and Economic Studies (IMES) of the Bank of Japan (BOJ) as a visiting scholar and as a visitor, respectively. The authors are very grateful to the anonymous referees for their valuable comments and suggestions. Any remaining errors are our own. Views expressed in this paper are those of the authors and do not necessarily reflect the official views of the BOJ.

I. Introduction

Endogenous sampling is a sample design in which the statistician stratifies the population based on endogenous variables, such as choices or alternatives in discrete choice probability models, and then selects samples at different rates from the different strata. In financial and labor economic research, one frequently must analyze data that measure the time until the occurrence of certain event, such as default and unemployment. Due to the restriction of the observation window, the event may not occur to some observations during the study period. Such observations are called "right-censored." In duration analysis, endogenous sampling refers to the design in which the population is divided into two subsets (non-right-censored and right-censored) and the statistician samples only from one subset or from both subsets at a predetermined ratio.

Endogenous sampling is widely used in the duration analysis of such phenomena as firm failures, loan defaults, and insurer insolvencies, because in these areas default cases are rarely observed, relative to non-default cases, while they are the most interesting to the researchers. In addition, many studies augment the endogenous sampling by exogenous sampling, which is referred to as "matching." First, a random sample is drawn from the default subsets; a second sample is then drawn from the non-default subsets in such a way that the distributions of some exogenous variables are matched for the two samples; and finally, the combined sample is used for estimation. Unfortunately, these empirical applications have either used standard estimation procedures intended for random sampling or used *ad hoc* modified estimation procedures without investigating their statistical properties. Thus, they have failed to consider the full implications of endogenous sampling, as well as those of the matching procedure. Since the sample is no longer representative of the population, without proper adjustment, any statistical inference about the population is biased. The seriousness of this problem calls for a rigorous treatment for non-random sampling in duration analysis.

The importance and necessity of such treatment is best illustrated by a series of examples:

(1) Lane, Looney, and Wansley (1986) analyze bank failures in the United States from 1979 to 1983 to identify the factors that increase the default risk of commercial banks and to construct a model to predict the default probability, based on the banks' characteristics. They match each failed bank with one or more non-failed banks, based on geographic location and four other criteria. A Cox proportional hazards model is estimated using the PHGLM procedure provided by SAS software.

(2) Luoma and Laitinen (1991) study Finnish company failures using empirical data, consisting of "36 Finnish failed limited companies and their non-failed mates." They do not describe how the non-failed "mates" were chosen. A Cox proportional hazards model is estimated using the BMDP statistical software.

(3) Kim *et al.* (1995) investigate property-liability and life insurer insolvencies in the United States. They select the default and non-default samples in an equal proportion in the framework of proportional hazard models. In their

duration analysis, they use the weighted maximum likelihood estimation method originally proposed by Manski and Lerman (1977) in the qualitative response model.

(4) Lee and Urrutia (1996) study the insolvency problem in the U.S. property-liability insurance industry in the 1980s. They choose insolvent insurers based on data availability. An equal number of matching solvent insurers are then selected, based on state domicile and total admitted assets. Lee and Urrutia (1996) find that their procedure creates a choice-based sample and claim to have "appropriately" corrected the over-sampling problem by making the adjustments following Palepu (1986) and BarNiv (1990), who address the problem in the context of a discrete choice probability models.

The common feature of the above examples is that all of these papers apply duration models with non-random sampling schemes, although they differ in sample design details, estimation methods, and, as will be clarified later, in the appropriateness of their estimation methods.

The properties of endogenous sampling have been investigated in various models, most notably in qualitative response models, as summarized in Amemiya (1985, section 9.5). However, the first paper that considered this problem in duration models, at least as far as we know, is Amemiya (2001), who derives the asymptotic properties of the endogenous sampling maximum likelihood estimator (ESMLE) in a duration model, in which defaults and non-defaults are sampled in a certain proportion. A counterintuitive finding for the case of a scalar parameter is that the optimal sampling proportion for the duration model is always zero or one, never in between. In other words, depending on the functional assumptions, it is optimal to use only the default sample or the non-default sample. Amemiya (2001) proves that the random sampling maximum likelihood estimator (RSMLE) is inconsistent under an endogenous sampling scheme. Furthermore, Amemiya (2001) compares the two estimators with regard to their respectively favorable conditions, showing that in certain cases the ESMLE can be more efficient than the RSMLE. One weakness of the ESMLE is the necessity of estimating the starting-time distribution of the spells. With regard to this problem, Amemiya (2001) proposes a conditional ESMLE and investigates the effects of estimating the starting-time distribution from a separate sample.

However, Amemiya (2001) deals only with the case of endogenous sampling without matching and does not address the frequently used method of mixing endogenous sampling with the matching procedure in duration analysis. The wide application of such sample designs in many empirical studies has heightened the need to provide a consistent estimator with its statistical properties fully characterized. Furthermore, it would be interesting to directly compare Amemiya's (2001) ESMLE with the Manski-Lerman weighted maximum likelihood estimator (WMLE) in the context of duration analysis, which was applied in Kim *et al.* (1995) as a solution to the over-sampling problem generated by their sample design.

In this paper, we focus on the maximum likelihood estimator under the mixed sampling scheme described above (see Section III for a rigorous definition). We aim to answer two questions: first, how to properly estimate a duration model with

non-random samples; second, whether it is advantageous to apply non-random sampling to duration models. The next section compares the statistical properties of the Manski-Lerman WMLE with those of the ESMLE within the context of the duration model. Section III derives the asymptotic distribution of the mixed sampling maximum likelihood estimator (MSMLE). Section IV investigates the relative efficiencies of the RSMLE, ESMLE, and MSMLE. Using a simple example, we show that none of the estimators unambiguously dominates the others. However, the ESMLE and MSMLE could outperform the RSMLE when the population is extremely unbalanced in terms of the frequency of defaults and non-defaults. A Monte Carlo study confirms this statement for small sample sizes. Finally, Section V summarizes the findings and offers a general guideline for sampling in empirical duration analysis.

II. The Manski-Lerman WMLE in Duration Models

A. Asymptotic Properties of the WMLE

The references cited above used the WMLE, originally proposed by Manski and Lerman (1977) for the qualitative response model, without questioning its validity. Although the estimator can be shown to be consistent for the duration model as well, it cannot be recommended for the duration analysis for two reasons. First, the Manski-Lerman estimator has the intrinsic assumption that the true probability of choice (for the qualitative response model) or of default (for the duration model) is known. This assumption may be justified for the qualitative response model where, for example, the true proportion of people riding a train in the entire region can be estimated reasonably well. However, for most duration model applications, this assumption is inappropriate. For example, the true proportion of firms that default cannot be accurately estimated, given the small sample size that one could possibly obtain. Furthermore, even if the true probability of default is known, it is equally easy and more efficient to maximize the true likelihood function than the weighted log likelihood, as will be demonstrated below.

Following Amemiya's (2001) notation, we assume that the duration data are obtained from the following data generating process. A spell, defined as individual duration of stay in one state, starts in an interval (a, b), and the starting time X is distributed according to the density $h(x)$. The duration T is distributed according to the density $f(t|\beta)$ and the distribution function $F(t|\beta)$, where β is a parameter vector. For simplicity, we assume that X and T are independent. Let D represent the indicator function of whether a spell is a default ($t < b - x$) or a non-default ($t \geq b - x$), and $P_1(P_0)$ be the probability of default (non-default), which the WMLE assumes to be known to the statistician. Prior knowledge of the causal structure is assumed to allow the statistician to specify $f(t|\bullet)$ up to a parameter vector β, contained in a subset B of a finite-dimensional Euclidean space. The goal is to estimate β from a sample generated by the following design: default samples are selected with probability λ_1 and non-defaults with probability $\lambda_0 (= 1 - \lambda_1)$. The WMLE defines weights as $W_j = P_j/\lambda_j$ for $j = 0, 1$ and maximizes the following weighted likelihood:

$$S = \sum_1 W_1 \ln f(x, t|D=1) + \sum_0 W_0 \ln f(x|D=0)$$

$$= \sum_1 W_1 \ln\left[\frac{h(x)f(t)}{P_1}\right] + \sum_0 W_0 \ln\left[\frac{h(x)[1-F(b-x)]}{P_0}\right], \tag{1}$$

subject to

$$P(D=1) = \int_a^b h(x)F(b-x)dx = P_1, \tag{2}$$

or equivalently

$$P(D=0) = \int_a^b h(x)[1-F(b-x)]dx = P_0, \tag{3}$$

where $\sum_1$ and $\sum_0$ mean summarizing over the default and non-default samples, respectively.

One can show the consistency of the WMLE in the same way as to prove consistency of the WMLE in the qualitative response model (see Amemiya [1985, section 9.5.2]). To derive the asymptotic distribution, it is convenient to rewrite the constraint in the form of

$$\beta = g(\alpha), \tag{4}$$

where α is a $(k-1)$ vector and k is the dimension of β.

By Amemiya's (1985) Theorem 4.1.3, we have

$$\sqrt{N}(\hat{\alpha}_{WMLE} - \alpha) \to N(0, AV(\hat{\alpha}_{WMLE})), \tag{5}$$

$$AV(\hat{\alpha}_{WMLE}) = (A+B)^{-1}\left[\frac{P_1}{\lambda_1}A + \frac{P_0}{\lambda_0}B\right](A+B)^{-1}, \tag{6}$$

where

$$A = \int_a^b \int_0^{b-x} \frac{h(x)}{f(t)} \frac{\partial f(t)}{\partial \alpha} \frac{\partial f(t)}{\partial \alpha'} dt\, dx, \tag{7}$$

$$B = \int_a^b \frac{h(x)}{1-F(b-x)} \frac{\partial(1-F(b-x))}{\partial \alpha} \frac{\partial(1-F(b-x))}{\partial \alpha'} dx. \tag{8}$$

Therefore, using a Taylor series approximation

$$\hat{\beta}_{WMLE} - \beta \cong G(\hat{\alpha}_{WMLE} - \alpha), \tag{9}$$

where $G = \partial\beta/\partial\alpha'$, we can derive the asymptotic distribution of the WMLE as

$$\sqrt{N}\,(\hat{\beta}_{WMLE} - \beta) \to N(0, AV(\hat{\beta}_{WMLE})), \tag{10}$$

$$AV(\hat{\beta}_{WMLE}) = G\,(A+B)^{-1}\left[\frac{P_1}{\lambda_1}A + \frac{P_0}{\lambda_0}B\right](A+B)^{-1}G'. \tag{11}$$

B. Comparing the WMLE with the ESMLE, Assuming That P_1 (P_0) Is Known

To compare the WMLE with the ESMLE, we have to make the very restrictive assumption that the probabilities of default and non-default are known, which is associated with the WMLE.

Imposing this assumption on the derivation in Amemiya (2001), we have

$$\sqrt{N}\,(\hat{\beta}_{ESMLE} - \beta) \to N(0, AV(\hat{\beta}_{ESMLE})), \tag{12}$$

$$AV(\hat{\beta}_{ESMLE}) = G\left[\frac{\lambda_1}{P_1}A + \frac{\lambda_0}{P_0}B\right]^{-1}G'. \tag{13}$$

We can show through simple algebra[1] that

$$AV(\hat{\beta}_{ESMLE}) \le AV(\hat{\beta}_{WMLE}). \tag{14}$$

When α is a scalar, it is also possible to compare the two estimators under their respective optimal (in terms of asymptotic efficiency) sampling designs, which are

$$\lambda_1^*(WMLE) = \frac{1}{1 + \sqrt{\frac{P_0 B}{P_1 A}}}, \tag{15}$$

$$\lambda_1^*(ESMLE) = 1_{\{\frac{A}{P_1} > \frac{B}{P_0}\}}. \tag{16}$$

Then, the asymptotic variances under the two sample designs are, respectively:

$$AV^*(\hat{\beta}_{WMLE}) = G\,(A+B)^{-1}\left[\sqrt{P_1 A} + \sqrt{P_0 B}\right]^2(A+B)^{-1}G', \tag{17}$$

$$AV^*(\hat{\beta}_{ESMLE}) = G\left[\frac{P_1}{A}1_{\{\frac{A}{P_1} > \frac{B}{P_0}\}} + \frac{P_0}{B}1_{\{\frac{A}{P_1} \le \frac{B}{P_0}\}}\right]G'. \tag{18}$$

It can be easily shown that under their respective optimal sampling designs, the ESMLE dominates the WMLE in terms of asymptotic efficiency.[2]

1. $AV(\hat{\beta}_{ESMLE}) \le AV(\hat{\beta}_{WMLE}) \Leftrightarrow [A + (P_0\lambda_1/P_1\lambda_0)B]^{-1} \le [(P_0\lambda_1/P_1\lambda_0)B]^{-1}$, which is true since both A and B are non-negative definite. See Appendix 1 for details.
2. If $A/P_1 > B/P_0$, then $AV^*(\hat{\beta}_{ESMLE}) \le AV^*(\hat{\beta}_{WMLE}) \Leftrightarrow \sqrt{P_1 A} + \sqrt{P_0 B} \ge (A+B)\sqrt{P_1/A} \Leftrightarrow A/P_1 \ge B/P_0$; if $A/P_1 \le B/P_0$, then $AV^*(\hat{\beta}_{ESMLE}) \le AV^*(\hat{\beta}_{WMLE}) \Leftrightarrow \sqrt{P_1 A} + \sqrt{P_0 B} \ge (A+B)\sqrt{P_0/B} \Leftrightarrow A/P_1 \le B/P_0$.

III. Matching in Duration Models

As we mentioned above, researchers usually use samples that are not only based on endogenous variables, but are also *matched* on one or more exogenous variables. For this purpose, we introduce a new variable. Assume that matching is based on covariate Z, which is distributed according to density $g(z)$. Conditional on covariate Z, the duration T is distributed according to the density $f(t|z)$ and the distribution function $F(t|z)$. We further assume that X and T are independent conditional on Z.[3]

We assume that the matching sampling scheme is designed as follows: for fixed N_1 and N_0, we first randomly sample N_1 defaults; then we compute the empirical distribution of Z among the default samples, denoted by $\hat{g}_1(z)$. N_0 non-defaults are then drawn, such that the empirical distribution of Z among these non-default samples are the same as $\hat{g}_1(z)$. Let $N = N_1 + N_0$ be the total number of observations under such a matching sampling scheme, $\lambda_1 = N_1/N$ and $\lambda_0 = 1 - \lambda_1 = N_0/N$, all of which are predetermined by the statistician.

A. Likelihood Function

The likelihood of a single observation in the subsample of defaults ($t_i < b - x_i$) is given by

$$f(x_i, t_i, z_i|D_i = 1) = \frac{h(x_i)f(t_i|z_i)g(z_i)}{P(D_i = 1)}, \tag{19}$$

where

$$P(D = 1) = P(T < b - X) = \int_Z \int_a^b h(x)F(b - x|z)g(z)dx\,dz. \tag{20}$$

The likelihood of a single observation in the subsample of non-defaults with attribute z ($t_i < b - x_i$, $z_i = z$) is given by

$$f(x_i|z, D_i = 0) = \frac{h(x_i)[1 - F(b - x_i|z)]g(z)}{P(D_i = 0)g(z|D_i = 0)}$$

$$= \frac{h(x_i)[1 - F(b - x_i|z)]}{P(D_i = 0|z)}, \tag{21}$$

where

$$P(D = 0|z) = P(T \geq b - X|z) = \int_a^b h(x)[1 - F(b - x|z)]dx. \tag{22}$$

Let $\hat{g}_1(z) = \hat{g}(z|D = 1)$ be the empirical conditional distribution of z among the subsample of defaults. The endogenous sampling process with matching implies that

3. Although this assumption is usually made in duration analysis, it does involve a loss of generality. The relaxation of this assumption is a matter for future research.

we will sample with probability λ_1 from the stratum of defaults, and with probability $[\lambda_0 \hat{g}_1(z)]$ from the stratum of non-defaults with attribute z. Therefore, assuming the parameter of interest, β, only characterizes f, but not h or g, the likelihood for such a generated sample is

$$L_M(\beta; x, t, z, D) = \prod_1 \lambda_1 f(x_i, t_i, z_i | D_i = 1) \prod_0 \lambda_0 \hat{g}_1(z) [f(x_i | z, D_i = 0)], \tag{23}$$

where $\prod_1$ and $\prod_0$ mean taking the product over the default and non-default samples, respectively.

Ignoring the terms that do not depend on β, we have the log likelihood of the sample as follows:

$$\begin{aligned} \ln L_M = & \sum_1 \ln f(t_i | z_i) - N_1 \ln P(D = 1) + \sum_0 \ln[1 - F(b - x_i | z_i)] \\ & - \sum_0 \ln P(D = 0 | z_i). \end{aligned} \tag{24}$$

B. Consistency of the MSMLE $\hat{\beta}$

The consistency of MSMLE $\hat{\beta}$ follows from Theorem 2.2.1 (the generalized Amemiya conditions) of Goto (1993) under certain regularity conditions (see Appendix 2). Here, we only verify that $(1/N)(\partial \ln L_M / \partial \beta) \to 0$ in probability.

$$\begin{aligned} \frac{1}{N} \frac{\partial \ln L_M}{\partial \beta} = & \frac{1}{N} \sum D_i \frac{1}{f(t_i | z_i)} \frac{\partial f(t_i | z_i)}{\partial \beta} \\ & - \lambda_1 \frac{1}{P(D = 1)} \frac{\partial P(D = 1)}{\partial \beta} \\ & + \frac{1}{N} \sum (1 - D_i) \frac{1}{1 - F(b - x_i | z_i)} \frac{\partial [1 - F(b - x_i | z_i)]}{\partial \beta} \\ & - \frac{1}{N} \sum (1 - D_i) \frac{1}{P(D = 0 | z_i)} \frac{\partial P(D = 0 | z_i)}{\partial \beta}, \end{aligned} \tag{25}$$

$$\begin{aligned} & \operatorname{plim} \frac{1}{N} \sum D_i \frac{1}{f(t_i | z_i)} \frac{\partial f(t_i | z_i)}{\partial \beta} \\ & = \lambda_1 E \left[\frac{1}{f(t | z)} \frac{\partial f(t | z)}{\partial \beta} \middle| D = 1 \right] \end{aligned}$$

$$= \lambda_1 \frac{1}{P(D=1)} \int_Z \int_a^b \int_0^{b-x} \frac{1}{f(t|z)} \frac{\partial f(t|z)}{\partial \beta} h(x) f(t|z) g(z) dt\, dx\, dz$$

$$= \lambda_1 \frac{1}{P(D=1)} \frac{\partial P(D=1)}{\partial \beta}, \tag{26}$$

$$\text{plim} \frac{1}{N} \sum (1 - D_i) \frac{1}{1 - F(b - x_i|z_i)} \frac{\partial [1 - F(b - x_i|z_i)]}{\partial \beta}$$

$$= \lambda_0 E \left[\frac{1}{1 - F(b - x|z)} \frac{\partial [1 - F(b - x|z)]}{\partial \beta} \middle| D = 0 \right]$$

$$= \lambda_0 \int_Z \frac{\int_a^b \frac{1}{1 - F(b - x|z)} \frac{\partial [1 - F(b - x|z)]}{\partial \beta} h(x) [1 - F(b - x|z)] dx}{P(D = 0|z)} g_1(z) dz$$

$$= \lambda_0 \int_Z \frac{1}{P(D=0|z)} \frac{\partial P(D=0|z)}{\partial \beta} g_1(z) dz, \tag{27}$$

where

$$g_1(z) = g(z|D=1) = \text{plim} \hat{g}_1(z),^4 \tag{28}$$

$$\text{plim} \frac{1}{N} \sum (1 - D_i) \frac{1}{P(D=0|z_i)} \frac{\partial P(D=0|z_i)}{\partial \beta}$$

$$= \lambda_0 \int_Z \frac{1}{P(D=0|z_i)} \frac{\partial P(D=0|z_i)}{\partial \beta} g_1(z) dz. \tag{29}$$

Thus, the consistency of the MSMLE $\hat{\beta}$ follows from equations (26), (27), and (29).

C. Asymptotic Variance of the MSMLE $\hat{\beta}$

We derive the asymptotic variance using the following formula:

$$AV \left[\sqrt{N} (\hat{\beta} - \beta) \right]^{-1} = \lim E \left[\frac{1}{N} \frac{\partial \ln L}{\partial \beta} \frac{\partial \ln L}{\partial \beta'} \right]. \tag{30}$$

Rearranging the terms on the right-hand side of equation (25) and multiplying them by $\sqrt{N}$, we have

4. We implicitly assume that the empirical distribution of Z among the default samples can be consistently estimated, that is, $\text{plim} \hat{g}_1(z) = g_1(z)$. This assumption is true only if we have a large enough number of defaults in our constructed sample.

$$\frac{1}{\sqrt{N}}\frac{\partial \ln L_M}{\partial \beta}$$

$$= \frac{1}{\sqrt{N}}\sum D_i\left(\frac{1}{f(t_i|z_i)}\frac{\partial f(t_i|z_i)}{\partial \beta} - \frac{1}{P(D=1)}\frac{\partial P(D=1)}{\partial \beta}\right)$$

$$+ \frac{1}{\sqrt{N}}\sum (1-D_i)\left(\frac{1}{1-F(b-x_i|z_i)}\frac{\partial [1-F(b-x_i|z_i)]}{\partial \beta}\right.$$

$$\left. - \frac{1}{P(D=0|z_i)}\frac{\partial P(D=0|z_i)}{\partial \beta}\right). \tag{31}$$

Since

$$E\left[\begin{array}{l}\left(\dfrac{1}{f(t|z)}\dfrac{\partial f(t|z)}{\partial \beta} - \dfrac{1}{P(D=1)}\dfrac{\partial P(D=1)}{\partial \beta}\right) \\ \times\left(\dfrac{1}{f(t|z)}\dfrac{\partial f(t|z)}{\partial \beta'} - \dfrac{1}{P(D=1)}\dfrac{\partial P(D=1)}{\partial \beta'}\right)\end{array}\middle| D=1\right]$$

$$= \frac{\int_Z\int_a^b\int_0^{b-x}\left[\begin{array}{l}\left(\dfrac{1}{f(t|z)}\dfrac{\partial f(t|z)}{\partial \beta} - \dfrac{1}{P(D=1)}\dfrac{\partial P(D=1)}{\partial \beta}\right) \\ \times\left(\dfrac{1}{f(t|z)}\dfrac{\partial f(t|z)}{\partial \beta'} - \dfrac{1}{P(D=1)}\dfrac{\partial P(D=1)}{\partial \beta'}\right)\end{array}\right] h(x)f(t|z)g(z)dt\,dx\,dz}{P(D=1)}$$

$$= \frac{\int_Z\int_a^b\int_0^{b-x}\frac{h(x)g(z)}{f(t|z)}\frac{\partial f(t|z)}{\partial \beta}\frac{\partial f(t|z)}{\partial \beta'}dt\,dx\,dz - \frac{1}{P(D=1)}\frac{\partial P(D=1)}{\partial \beta}\frac{\partial P(D=1)}{\partial \beta'}}{P(D=1)},$$

(32)

and

$$\lim E\left[\begin{array}{l}\left(\dfrac{1}{1-F(b-x|z)}\dfrac{\partial [1-F(b-x|z)]}{\partial \beta} - \dfrac{1}{P(D=0|z)}\dfrac{\partial P(D=0|z)}{\partial \beta}\right) \\ \times\left(\dfrac{1}{1-F(b-x|z)}\dfrac{\partial [1-F(b-x|z)]}{\partial \beta'} - \dfrac{1}{P(D=0|z)}\dfrac{\partial P(D=0|z)}{\partial \beta'}\right)\end{array}\middle| D=0\right]$$

$$= \int_Z \frac{\int_a^b \left[\left(\frac{1}{1-F(b-x|z)} \frac{\partial[1-F(b-x|z)]}{\partial\beta} - \frac{1}{P(D=0|z)} \frac{\partial P(D=0|z)}{\partial\beta} \right) \times \left(\frac{1}{1-F(b-x|z)} \frac{\partial[1-F(b-x|z)]}{\partial\beta'} - \frac{1}{P(D=0|z)} \frac{\partial P(D=0|z)}{\partial\beta'} \right) \right] h(x)[1-F(b-x|z)]dx}{P(D=0|z)} g_1(z)dz$$

$$= \int_Z \frac{\int_a^b \frac{h(x)}{1-F(b-x|z)} \frac{\partial[1-F(b-x|z)]}{\partial\beta} \frac{\partial[1-F(b-x|z)]}{\partial\beta'} dx - \frac{1}{P(D=0|z)} \frac{\partial P(D=0|z)}{\partial\beta} \frac{\partial P(D=0|z)}{\partial\beta'}}{P(D=0|z)} g_1(z)dz,$$

(33)

the asymptotic variance of the MSMLE $\hat{\beta}$ is given by

$$AV(MSMLE)^{-1} = \lim E\left[\frac{1}{N} \frac{\partial \ln L_M}{\partial\beta} \frac{\partial \ln L_M}{\partial\beta'} \right]$$

$$= \lambda_1 \frac{\int_Z \int_a^b \int_0^{b-x} \frac{h(x)g(z)}{f(t|z)} \frac{\partial f(t|z)}{\partial\beta} \frac{\partial f(t|z)}{\partial\beta'} dt\,dx\,dz - \frac{1}{P(D=1)} \frac{\partial P(D=1)}{\partial\beta} \frac{\partial P(D=1)}{\partial\beta'}}{P(D=1)}$$

$$+ \lambda_0 \int_Z \frac{\int_a^b \frac{h(x)}{1-F(b-x|z)} \frac{\partial[1-F(b-x|z)]}{\partial\beta} \frac{\partial[1-F(b-x|z)]}{\partial\beta'} dx - \frac{1}{P(D=0|z)} \frac{\partial P(D=0|z)}{\partial\beta} \frac{\partial P(D=0|z)}{\partial\beta'}}{P(D=0|z)} g_1(z)dz.$$

(34)

D. Estimation of Mixed Sample When $g(z)$ Is Unknown

The above discussion is based on the assumption that the marginal distribution $g(z)$ is known to the statistician. However, in many empirical contexts such prior knowledge is not likely to be available. This section covers the case when $g(z)$ is estimated from a separate random sample.

Kiefer and Wolfowitz (1956) show that the empirical distribution is the maximum likelihood estimator of an unknown distribution function. Therefore, suppose the sample from which $g(z)$ is estimated is of size K, the log likelihood function of β will be modified as

$$\ln \tilde{L}_M = \sum_1 \ln f(t_i|z_i) - N_1 \ln \hat{P}(D=1) + \sum_0 \ln[1-F(b-x_i|z_i)] - \sum_0 \ln P(D=0|z_i), \quad (35)$$

where

$$\hat{P}(D=1)=\frac{1}{K}\sum_{k=1}^{K}\int_a^b h(x)F(b-x|z_k)dx. \tag{36}$$

Note that

$$\frac{1}{\sqrt{N}}\frac{\partial \ln \tilde{L}_M}{\partial \beta}$$

$$=\frac{1}{\sqrt{N}}\frac{\partial \ln L_M}{\partial \beta}-\lambda_1\sqrt{N}\left[\frac{1}{\hat{P}(D=1)}\frac{\partial \hat{P}(D=1)}{\partial \beta}-\frac{1}{P(D=1)}\frac{\partial P(D=1)}{\partial \beta}\right]$$

$$\overset{LD}{=}\frac{1}{\sqrt{N}}\frac{\partial \ln L_M}{\partial \beta}-\lambda_1\sqrt{N}\left[\frac{\frac{\partial \hat{P}(D=1)}{\partial \beta}-\frac{\partial P(D=1)}{\partial \beta}}{P(D=1)}-\frac{\frac{\partial P(D=1)}{\partial \beta}(\hat{P}(D=1)-P(D=1))}{P(D=1)^2}\right],$$

where $\overset{LD}{=}$ denotes equivalency in the limit distribution.

The consistency follows from Theorem 2.3.1 in Goto (1993) under certain regularity conditions (see Appendix 2). Since this is a two-step estimator, the asymptotic efficiency might be affected by the fact that $g(z)$ is estimated. In fact, if we define

$$\Omega=\lim E\left[\frac{1}{N}\frac{\partial \ln L_M}{\partial \beta}\frac{\partial \ln L_M}{\partial \beta'}\right],$$

$$W_1(z)=\int_a^b h(x)F(b-x|z)dx,$$

$$W_2(z)=\int_a^b h(x)\frac{\partial F(b-x|z)}{\partial \beta}dx,$$

then the asymptotic variance of $\tilde{\beta}$ can be defined as

$$AV\left[\sqrt{N}(\tilde{\beta}-\beta)\right]$$

$$=\Omega^{-1}\left[\Omega+\frac{N}{K}\frac{Var(W_2(z))}{P(D=1)^2}+\frac{N}{K}\frac{Var(W_1(z))}{P(D=1)^4}\frac{\partial P(D=1)}{\partial \beta}\frac{\partial P(D=1)}{\partial \beta'}\right.$$

$$\left.-2\frac{N}{K}\frac{1}{P(D=1)^3}\frac{\partial P(D=1)}{\partial \beta}Cov(W_1(z),W_2(z)')\right]\Omega^{-1}. \tag{37}$$

Equation (37) shows that for the two-step estimator $\tilde{\beta}$ to have the proper asymptotic property, K must increase to infinity at least as fast as N.

IV. Comparison of the Maximum Likelihood Estimators under Different Sample Designs

A. Relative Asymptotic Efficiency

The question of the relative efficiency of different sample designs is natural to raise in an investigation such as ours. However, the nonlinear structure precluded much progress in solving this problem. As shown below, the relative efficiency depends on the choice of λ's and functional forms of the densities, as well as prior knowledge of the parameters to be estimated. An explicitly Bayesian approach to the design problem might be a possible solution. This paper, however, will not undertake that task. Instead, we limit ourselves to a general discussion of the optimal design problem and illustrate the result with a simple example.

To compare the asymptotic efficiency of the MSMLE with that of endogenous sampling without matching, we introduce covariate z into Amemiya's (2001) endogenous sampling model and derive the log likelihood and the asymptotic distribution of ESMLE without matching as[5,6]

$$\ln L_E = \sum_1 \ln f(t_i|z_i) - N_1 \ln P(D=1) + \sum_0 \ln[1-F(b-x_i|z_i)] - N_0 \ln P(D=0), \tag{38}$$

$$\sqrt{N}(\hat{\beta}_{ESMLE} - \beta) \to N(0, AV(\hat{\beta}_{ESMLE})), \tag{39}$$

$$AV(ESMLE)^{-1} = \frac{\lambda_1}{P_1}\left[\int_Z\int_a^b\int_0^{b-x} \frac{hg}{f}\frac{\partial f}{\partial \beta}\frac{\partial f}{\partial \beta'}dt\,dx\,dz - \frac{1}{P_1}\frac{\partial P_1}{\partial \beta}\frac{\partial P_1}{\partial \beta'}\right] + \frac{\lambda_0}{P_0}\left[\int_Z\int_a^b \frac{hg}{1-F}\frac{\partial(1-F)}{\partial \beta}\frac{\partial(1-F)}{\partial \beta'}dx\,dz - \frac{1}{P_0}\frac{\partial P_0}{\partial \beta}\frac{\partial P_0}{\partial \beta'}\right], \tag{40}$$

where

$$P_0 = P(D=0) = \int_Z\int_a^b h(x)[1-F(b-x|z)]g(z)dx\,dz. \tag{41}$$

Similarly, by introducing covariate z into the derivation of equation (21) in Amemiya (2001), we have

$$\ln L_R = \sum_1 \ln f(t_i|z_i) + \sum_0 \ln[1-F(b-x_i|z_i)], \tag{42}$$

5. We use the following simplified notation from now on: $P_1 = P(D=1)$, $P_0 = P(D=0)$, which are defined in equations (20) and (41), respectively. We also suppress the arguments of density and distribution functions.
6. Please refer to Amemiya (2001) for the derivation.

$$\sqrt{N}(\hat{\beta}_{RSMLE} - \beta) \to N(0, AV(\hat{\beta}_{RSMLE})), \tag{43}$$

$$AV(RSMLE)^{-1}$$

$$= \int_Z \int_a^b \int_0^{b-x} \frac{hg}{f} \frac{\partial f}{\partial \beta} \frac{\partial f}{\partial \beta'} dt\, dx\, dz + \int_Z \int_a^b \frac{hg}{1-F} \frac{\partial (1-F)}{\partial \beta} \frac{\partial (1-F)}{\partial \beta'} dx\, dz. \tag{44}$$

Comparing equations (34), (40), and (44), we see that the relative efficiency depends on the choice of λ's and the functional form of the densities, as well as β, the parameters to be estimated. Notice that when covariate z has no prediction power in terms of default probability, that is, if $P(D = 0|z) = P(D = 0)$, matching on z will not add additional efficiency to what an endogenous sample can achieve, in other words, in that special case $AV(MSMLE) = AV(ESMLE)$. In general, however, none of the estimators unambiguously dominates the others. To illustrate this point, we consider a very simple example:

Example 1: *Assume*

$$h(x) = 1, \quad 0 \le x \le 1, \tag{45}$$

$$f(t|z) = e^{\beta z} e^{-e^{\beta z} t}, \quad t \ge 0, \tag{46}$$

$$z = \begin{cases} 1 & \textit{with probability } p \\ 0 & \textit{with probability } (1-p). \end{cases} \tag{47}$$

Substituting the above assumptions into equations (34), (40), and (44) yields

$$AV(ESMLE)^{-1}$$

$$= \frac{\lambda_1 p}{P_1}[1 - 3e^{-\beta} + 3e^{-\beta - e^{\beta}} + e^{\beta - e^{\beta}} + 2e^{-e^{\beta}}] - \frac{\lambda_1 p^2}{P_1^2}[e^{-\beta} - e^{-\beta - e^{\beta}} - e^{-e^{\beta}}]^2$$

$$+ \frac{\lambda_0 p}{P_0}[-e^{\beta - e^{\beta}} - 2e^{-e^{\beta}} - 2e^{-\beta - e^{\beta}} + 2e^{-\beta}] - \frac{\lambda_0 p^2}{P_0^2}[e^{-\beta} - e^{-\beta - e^{\beta}} - e^{-e^{\beta}}]^2$$

$$= \frac{\lambda_1 p}{P_1}[1 - 3e^{-\beta} + 3e^{-\beta - e^{\beta}} + e^{\beta - e^{\beta}} + 2e^{-e^{\beta}}]$$

$$+ \frac{\lambda_0 p}{P_0}[-e^{\beta - e^{\beta}} - 2e^{-e^{\beta}} - 2e^{-\beta - e^{\beta}} + 2e^{-\beta}]$$

$$-\left[\frac{\lambda_1 p^2}{P_1^2}+\frac{\lambda_0 p^2}{P_0^2}\right][e^{-\beta}-e^{-\beta-e^{\beta}}-e^{-e^{\beta}}]^2, \tag{48}$$

$$\begin{aligned}
&AV(RSMLE)^{-1}\\
&=p[1-3e^{-\beta}+3e^{-\beta-e^{\beta}}+e^{\beta-e^{\beta}}+2e^{-e^{\beta}}]\\
&\quad+p[-e^{\beta-e^{\beta}}-2e^{-e^{\beta}}-2e^{-\beta-e^{\beta}}+2e^{-\beta}]\\
&=p[1-e^{-\beta}+e^{-\beta-e^{\beta}}],
\end{aligned} \tag{49}$$

$$\begin{aligned}
&AV(MSMLE)^{-1}\\
&=\frac{\lambda_1 p}{P_1}[1-3e^{-\beta}+3e^{-\beta-e^{\beta}}+e^{\beta-e^{\beta}}+2e^{-e^{\beta}}]-\frac{\lambda_1 p^2}{P_1^2}[e^{-\beta}-e^{-\beta-e^{\beta}}-e^{-e^{\beta}}]^2\\
&\quad+\frac{\lambda_0 p(1-e^{-\beta}+e^{-\beta-e^{\beta}})}{P_1}\frac{-e^{\beta-e^{\beta}}-2e^{-e^{\beta}}-2e^{-\beta-e^{\beta}}+2e^{-\beta}}{-e^{-\beta-e^{\beta}}+e^{-\beta}}\\
&\quad-\frac{\lambda_0 p(1-e^{-\beta}+e^{-\beta-e^{\beta}})}{P_1}\left[\frac{e^{-\beta-e^{\beta}}+e^{-e^{\beta}}-e^{-\beta}}{-e^{-\beta-e^{\beta}}+e^{-\beta}}\right]^2.
\end{aligned} \tag{50}$$

P_1 and P_0 in the above equations can be calculated as

$$\begin{aligned}
P_1&=p\int_0^1\int_0^{1-x}e^{\beta}e^{-e^{\beta}t}dt\,dx+(1-p)\int_0^1\int_0^{1-x}e^{-t}dt\,dx\\
&=p(1-e^{-\beta}+e^{-\beta-e^{\beta}})+(1-p)e^{-1},
\end{aligned} \tag{51}$$

$$P_0=p(e^{-\beta}-e^{-\beta-e^{\beta}})+(1-p)(1-e^{-1}). \tag{52}$$

Table 1 illustrates the relative efficiency of the three classes of estimators under different assumptions of the true values of β and p. The examples are listed in ascendant order of the true probability of default $P(D=1)$, which is a function of β and p. For a sample of size N, the asymptotic variance for any estimator $\hat{\beta}$ can be calculated as $1/N$ divided by the corresponding value displayed in the table. As can be seen, none of the estimators unambiguously dominates the others. For each pair of β and p, we observe a rather large variation of the asymptotic variances of different estimators, which indicates that the sample design plays an important role in duration analysis. Despite the inconclusiveness of relative efficiency, the table seems to reveal a very intuitive pattern: when the population is extremely unbalanced in terms of the ratio of defaults and non-defaults, that is, when $P(D=1)$ takes extremely small or

Table 1 Inverse of Asymptotic Variance (Scaled by 1/*N*): Selected Examples

β	p	$P(D=1)$	Inverse of asymptotic variance					
			RSMLE	*ESMLE*		*MSMLE*		
				$\lambda_1 = 0$	$\lambda_1 = 1$	$\lambda_1 = 0.3$	$\lambda_1 = 0.5$	$\lambda_1 = 0.7$
–3.0	0.8	0.0932	0.0196	0.0003	0.1607*	0.0482	0.0804	0.1125
–2.0	0.7	0.1557	0.0453	0.0019	0.1888*	0.0570	0.0946	0.1323
0.5	0.4	0.4248	0.2040*	0.1506	0.1399	0.1088	0.1177	0.1266
1.0	0.3	0.4544	0.1969*	0.1834	0.1516	0.1793	0.1714	0.1635
2.3	0.9	0.8466	0.8098	0.8273	0.7326	0.8863*	0.8424	0.7985
2.7	0.95	0.9046	0.8862	0.8903	0.8335	0.9358*	0.9066	0.8773

Note: Asterisks denote the most asymptotically efficient estimator for each pair of β and p.

large values, over-sampling from the less frequent subsample would usually result in a significant efficiency gain, while random sampling is preferable when the population is relatively balanced.

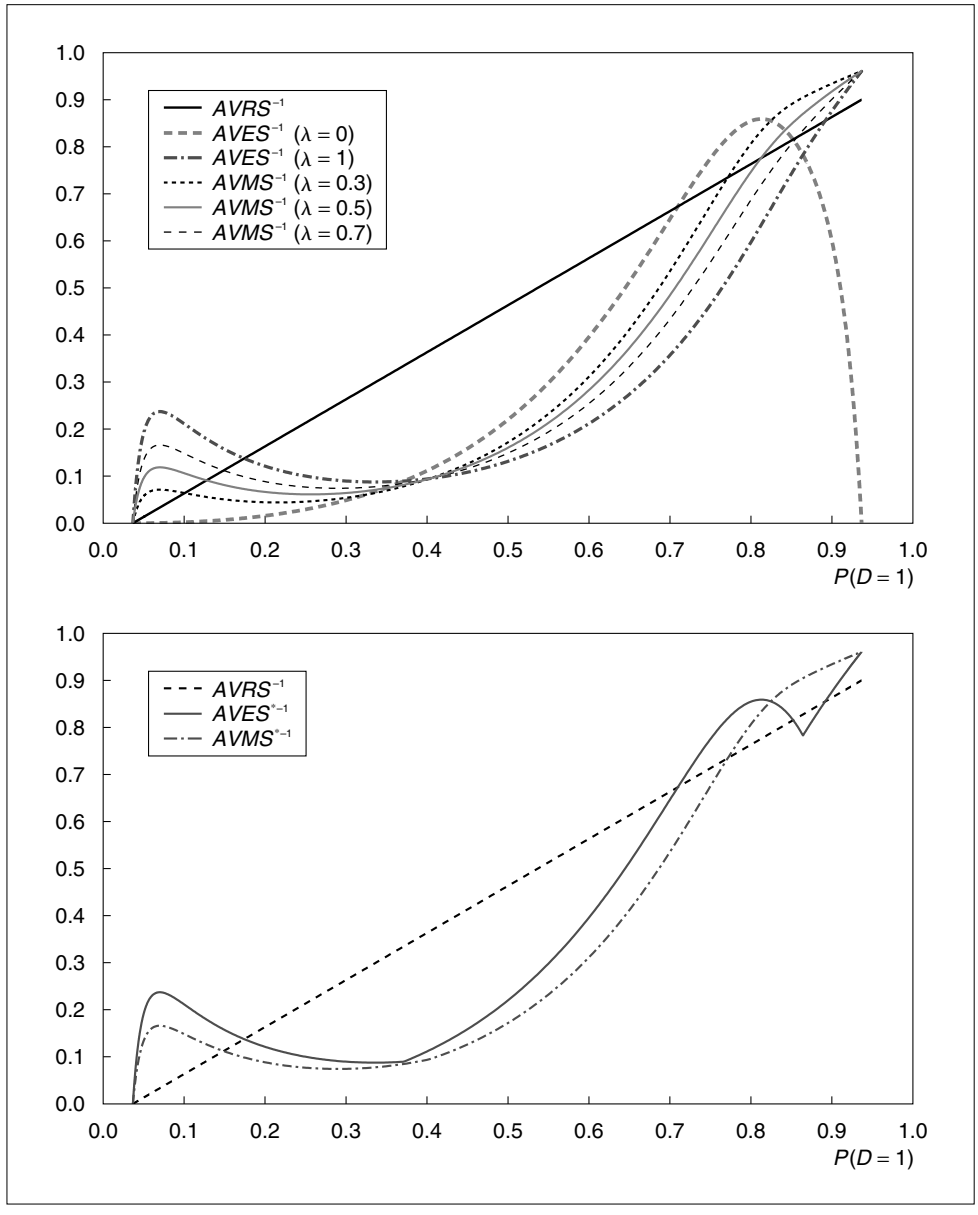
This phenomenon is further illustrated in Figure 1, which was obtained by fixing $p = 0.9$ and varying β from −10 to 10. The horizontal axis displays the variation of $P(D = 1)$ as a result of changes in the value of β. The upper panel shows inverted asymptotic variances (scaled by $1/N$) for all six estimators, while the lower panel displays those of the most efficient estimators within each class. As can be seen, the curves for endogenous sampling, with or without matching, reach their local maxima at the two tails of $P(D = 1)$. The same pattern persists when we change the value of p, except that as p becomes smaller, the middle range in which the RSMLE dominates the ESMLE and the MSMLE shrinks dramatically, and the range on the right tail, in which the MSMLE outperforms the ESMLE, increases considerably. Both Table 1 and Figure 1 suggest a general guideline for sampling in empirical duration analysis: it might be optimal to consider endogenous sampling, with or without matching, when the population is extremely unbalanced, whereas the random sampling design is usually a better choice when there is no significant difference in the observed frequency of defaults and non-defaults.

The above discussion, based on Example 1, only applies to the case when β is a scalar parameter. It is noted in Amemiya (2001) that the optimal choice of λ_1 in the endogenous sampling design is either one or zero. However, this conclusion is true only for the case of the scalar parameter and cannot be generalized to the vector parameters case. One of the possible criteria in terms of relative efficiency in the vector case can be defined over a linear combination of the parameter vector, that is

$$\min_{\lambda_1} AV\left[\sqrt{N}\,(c'(\hat{\beta} - \beta))\right], \tag{53}$$

where $c'c = 1$. Notice that the asymptotic variances of both $\sqrt{N}\,(\hat{\beta}_{ESMLE} - \beta)$ and $\sqrt{N}\,(\hat{\beta}_{MSMLE} - \beta)$ take the form of $[\lambda_1 A + (1 - \lambda_1)B]^{-1}$, for appropriate A's and B's (see equations [40] and [34]). Therefore, we can apply the following discussion to both sample designs:

Figure 1 Inverse of Asymptotic Variance (Scaled by 1/*N*)



$$
\begin{aligned}
AV\left[\sqrt{N}\,(c'(\hat{\beta}-\beta))\right] &= c'[\lambda_1 A + (1-\lambda_1)B]^{-1}c \\
&= c'B^{-1/2}[I + \lambda_1 B^{-1/2}(A-B)B^{-1/2}]^{-1}B^{-1/2}c \\
&= g'[I + \lambda_1 D]^{-1}g \\
&= \sum_{k=1}^{K} \frac{g_k^2}{1+\lambda_1 d_k}, \qquad (54)
\end{aligned}
$$

where K is the dimension of the vector β; $\{d_k\}_{k=1}^{K}$ are the eigenvalues of the matrix $[B^{-1/2}(A-B)B^{-1/2}]$; $D = diag[d_1, \ldots, d_k] = H'[B^{-1/2}(A-B)B^{-1/2}]H$; $g = H'B^{-1/2}c$.

Equation (54) shows that optimal choice of λ_1 depends on the characteristics of the variance-covariance matrix, which in turn depends on the distributions and the true parameters. We can easily offer examples where the optimal λ_1 is an interior value.

B. Asymptotic Bias

Amemiya (2001) demonstrates that the RSMLE is inconsistent under the endogenous sampling scheme without matching. This statement is also valid for the case of the mixed sampling scheme. When the data are collected through a mixed sampling scheme, the first-order condition of $\ln L_R$ does not have zero mean. Instead,

$$E^M\left[\frac{1}{N}\frac{\partial \ln L_R}{\partial \beta}\right] = \lambda_1 \frac{1}{P(D=1)}\frac{\partial P(D=1)}{\partial \beta} + \lambda_0 \int_Z \frac{1}{P(D=0|z)}\frac{\partial P(D=0|z)}{\partial \beta} g_1(z)dz, \tag{55}$$

which is not zero, in general.[7]

Not surprisingly, the ESMLE is also inconsistent under the mixed sampling scheme. Its inconsistency can be illustrated by investigating the expectation of the first-order condition of $\ln L_E$ with respect to the true likelihood under the mixed sampling:

$$E^M\left[\frac{1}{N}\frac{\partial \ln L_E}{\partial \beta}\right] = \lambda_0 \int_Z \frac{1}{P(D=0|z)}\frac{\partial P(D=0|z)}{\partial \beta} g_1(z)dz - \lambda_0 \frac{1}{P(D=0)}\frac{\partial P(D=0)}{\partial \beta}, \tag{56}$$

which, in general, is not zero either.

Since some of the examples cited at the beginning of the paper ignored the problem of the mixed sampling scheme and estimated the model either by conventional means (as if the data were from a random sample), or by a modified method for the qualitative response model (as if the data were from an endogenous sample without matching), it is interesting to study the magnitude of the bias when using the wrong estimators. For this purpose, we again consider the simple exponential example given in the above section.

Tables 2 and 3 present the results of asymptotic bias when the data obtained by mixed sampling are estimated as if they were obtained by random or endogenous sampling without matching, using the maximum likelihood estimator. Let γ be the probability limit of the biased estimator, and β be the probability limit of the MSMLE (which is equal to the true value of the parameter, denoted as β^*). For any given case, the value in each cell is calculated as $\gamma - \beta$. Given the true value β^*, γ is defined by solving

7. E^M means that the expectation is taken with respect to the true likelihood under mixed sampling.

Table 2 Asymptotic Bias for Mixed Sample Estimated as a Random Sample

β	p	$P(D=1)$	Asymptotic bias		
			$\lambda_1 = 0.3$	$\lambda_1 = 0.5$	$\lambda_1 = 0.7$
–3.0	0.8	0.0932	2.6029	3.1906	3.6105
–2.0	0.7	0.1557	1.6173	2.2051	2.6249
0.5	0.4	0.4248	–0.6065	–0.0238	0.3902
1.0	0.3	0.4544	–0.9011	–0.3257	0.0798
2.3	0.9	0.8466	–1.1697	–0.6356	–0.2754
2.7	0.95	0.9046	–1.1821	–0.6565	–0.3049

Table 3 Asymptotic Bias for Mixed Sample Estimated as an Endogenous Sample without Matching

β	p	$P(D=1)$	Asymptotic bias		
			$\lambda_1 = 0.3$	$\lambda_1 = 0.5$	$\lambda_1 = 0.7$
–3.0	0.8	0.0932	0.2886	0.1081	0.0442
–2.0	0.7	0.1557	0.7473	0.1997	0.0774
0.5	0.4	0.4248	–0.3702	–0.2600	–0.1603
1.0	0.3	0.4544	–0.6614	–0.4890	–0.3182
2.3	0.9	0.8466	–0.2793	–0.2138	–0.1373
2.7	0.95	0.9046	–0.2214	–0.6139	–0.5374

$$E^*\left[\frac{1}{N}\frac{\partial \ln L(\gamma)}{\partial \gamma}\right]=0.$$

When the mixed sample is estimated as if it were random, we obtain an explicit function of γ as

$$\gamma=\beta^*+\ln\left[\frac{\lambda(1-e^{e^{\beta^*}})(1-e^{e^{\beta^*}}+e^{e^{\beta^*}+\beta^*})}{(1+\lambda)(1-2e^{e^{\beta^*}}+e^{2e^{\beta^*}})+(1-\lambda)e^{e^{\beta^*}+2\beta^*}+e^{\beta^*}-e^{2e^{\beta^*}+2\beta^*}}\right]. \quad (57)$$

In the case when the mixed sample is estimated as if it were endogenous without matching, we have to resort to numerical approximation. The direction of the bias is not certain. However, based on the examples displayed in the table, the bias is generally upward when defaults are rarely observed in the population and downward when the population is relatively balanced or concentrated over the non-defaults. Such bias exists when the mixing nature of the sample is either ignored or partially treated.

C. Finite Sample Properties

We have explored the asymptotic properties of the three estimators in the previous two subsections. In this subsection, we will focus on how sample size affects the performance of different estimators using the Monte Carlo method.

The data generating process (DGP) is as follows: for the RSMLE, we assume that the data are generated randomly as in Example 1; for the ESMLE, we specify a probability λ_1, and sample defaults and non-defaults according to this pre-specified

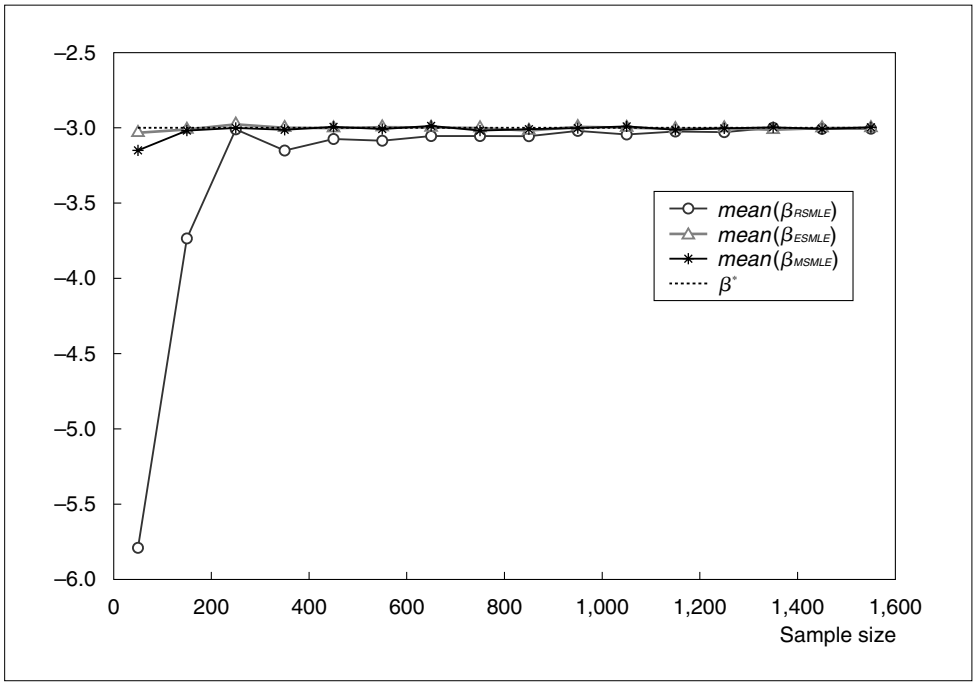
probability; for the MSMLE, we first specify the number of defaults N_1 and randomly sample N_1 default cases, then we calculate the empirical distribution of the matched variable z among the defaults; using this empirical distribution, we finally sample non-defaults. The DGP is repeated for different sample sizes. The likelihood functions and the first-order conditions of each estimator are presented in Appendix 3.

To illustrate the finite sample properties of the three estimators, we consider both a case of an extremely unbalanced population in terms of the frequency of defaults and non-defaults and a case where the population distribution is relatively balanced.

Case 1: *First of all, we consider a case in which the dataset is extremely unbalanced. In particular, we assume* $\beta^* = -3$ *and* $p = 0.8$ *in Example 1, which implies that the probability of default in the population* $P(D = 1)$ *is as low as 0.09. When applying endogenous sampling and mixed sampling, we assume the probability of sample default case* λ_1 *to be 0.7.*

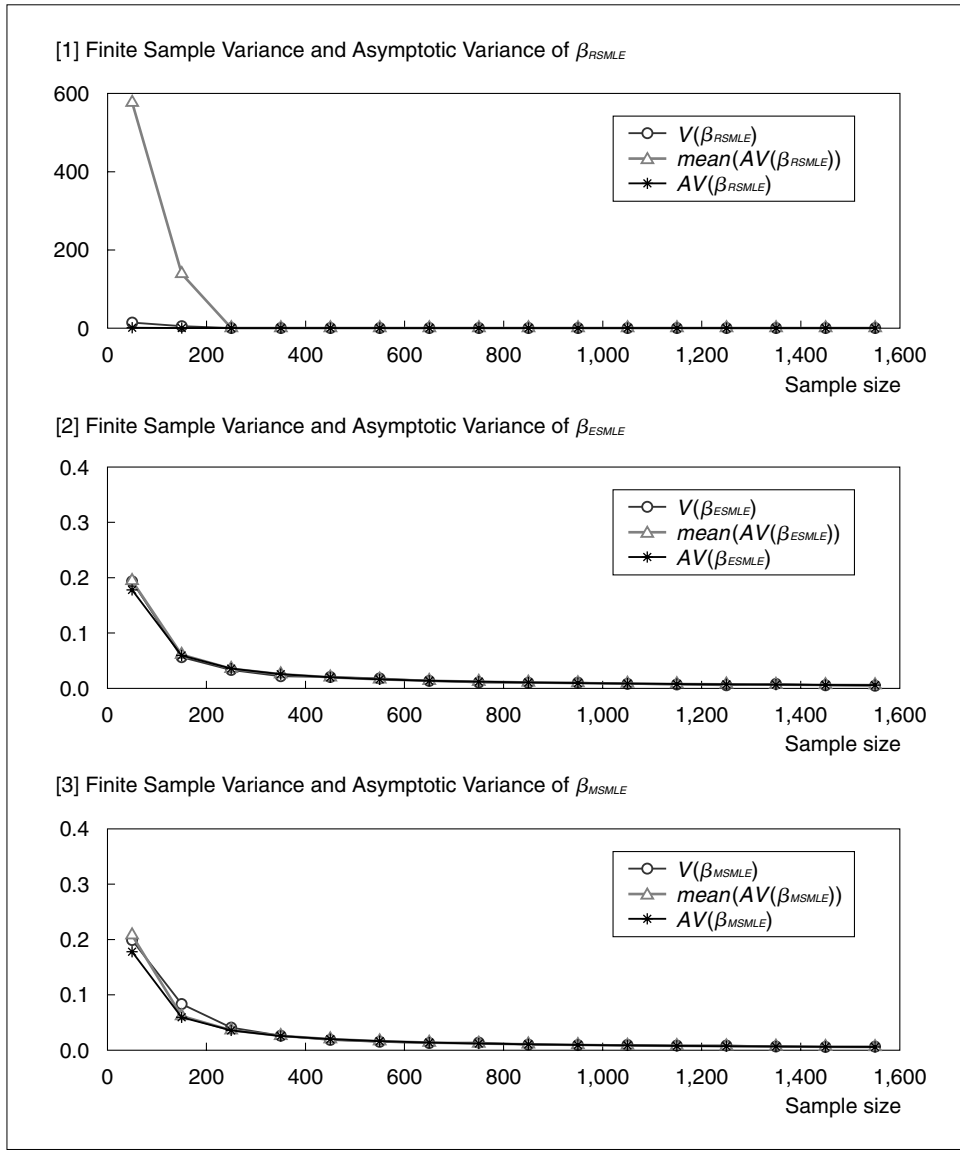
Figure 2 shows the three estimators with a sample size varying from 50 to 1,550. For each fixed sample size, we simulate 100 times and report the mean of the estimators. We experiment with a larger number of replications and the shape of the convergence is the same. As can be seen, the RSMLE converges much more slowly when the sample size is small, while both the ESMLE and MSMLE are very robust to changes in sample size. Figure 3 reports the empirical variance of the estimators based on the

Figure 2 Comparison of Three Estimators with Varying Sample Sizes (Unbalanced Population)



Monte Carlo simulations $\{V(\beta_k)\}$; the mean of the estimated asymptotic variances $\{mean(AV(\beta_k))\}$; and the asymptotic variances $\{AV(\beta_k)\}$ for the three different estimators $\{k = RSMLE, ESMLE, MSMLE\}$. The same pattern appears: when the sample size is small, the RSMLE's estimated variance is much larger than its asymptotic variance. In contrast, both the ESMLE and MSMLE's finite sample variances are good proximates of their asymptotic variances even with a very small sample size. Notice that to cover the full range of estimated asymptotic variance for the RSMLE, the upper panel is drawn on a much larger scale than the other panels. If we used a much finer scale as with the other panels, we would also observe a large gap between the average empirical variance for the RSMLE and its asymptotic variance for small sample sizes.

Figure 3 Monte Carlo Variance and Asymptotic Variance (Unbalanced Population)



The gap disappears as the sample size grows. For both the ESMLE and MSMLE, the empirical variances approximate the asymptotic variances very well, even for small sample sizes.

Case 2: *Next, we consider the case in which the default probability in the population is relatively balanced. In particular, we simulate the case of* $\beta^* = 1$ *and* $p = 0.3$*, which implies that the probability of default in the population* $P(D = 1)$ *is about 0.45. Again, we assume that* $\lambda_1 = 0.7$ *when applying endogenous sampling and mixed sampling.*

Figure 4 reports the mean of estimated β_{RSMLE}, β_{ESMLE}, and β_{MSMLE} from 100 simulations for each given sample size. Figure 5 presents the empirical variance of the estimators based on the Monte Carlo simulations $\{V(\beta_k)\}$; the mean of the estimated asymptotic variances $\{mean(AV(\beta_k))\}$; and the asymptotic variances $\{AV(\beta_k)\}$ for the three different estimators $\{k = RSMLE, ESMLE, MSMLE\}$. The two figures show that all three estimators are quite robust to the variation of sample size, although the RSMLE slightly outperforms the MSMLE and ESMLE in terms of small sample properties when the population is relatively balanced.

However, the above Monte Carlo study is based on an extremely simple example with only one coefficient to estimate. For a model with more coefficients to estimate, it is reasonable to believe that a non-random sampling scheme has an advantage over random sampling, given that most empirical studies deal with datasets that are extremely unbalanced.

Figure 4 Comparison of Three Estimators with Varying Sample Sizes (Balanced Population)

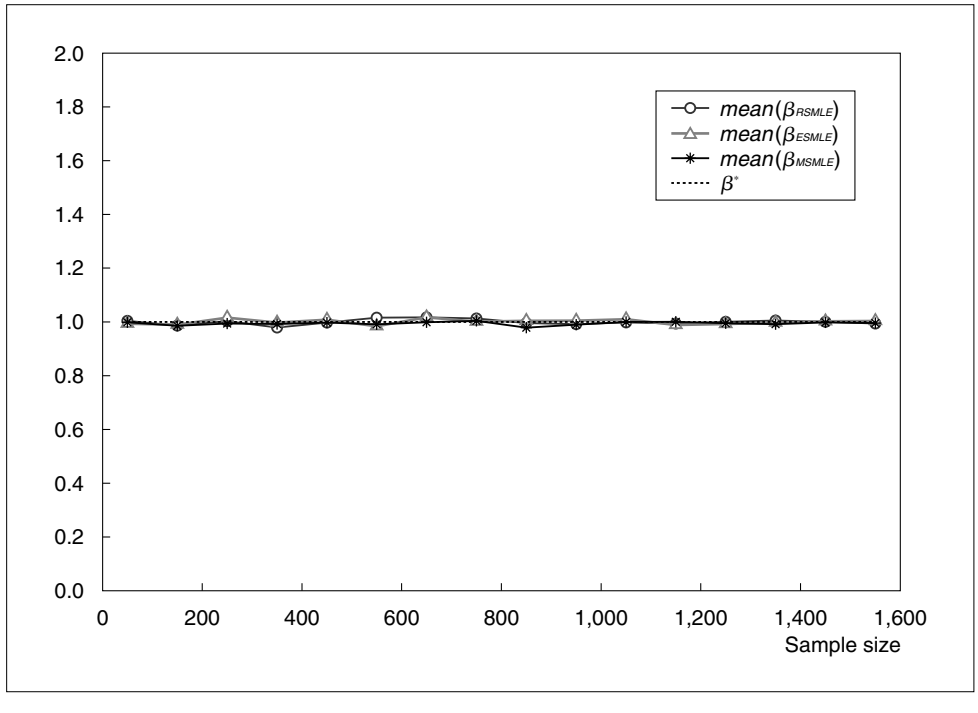
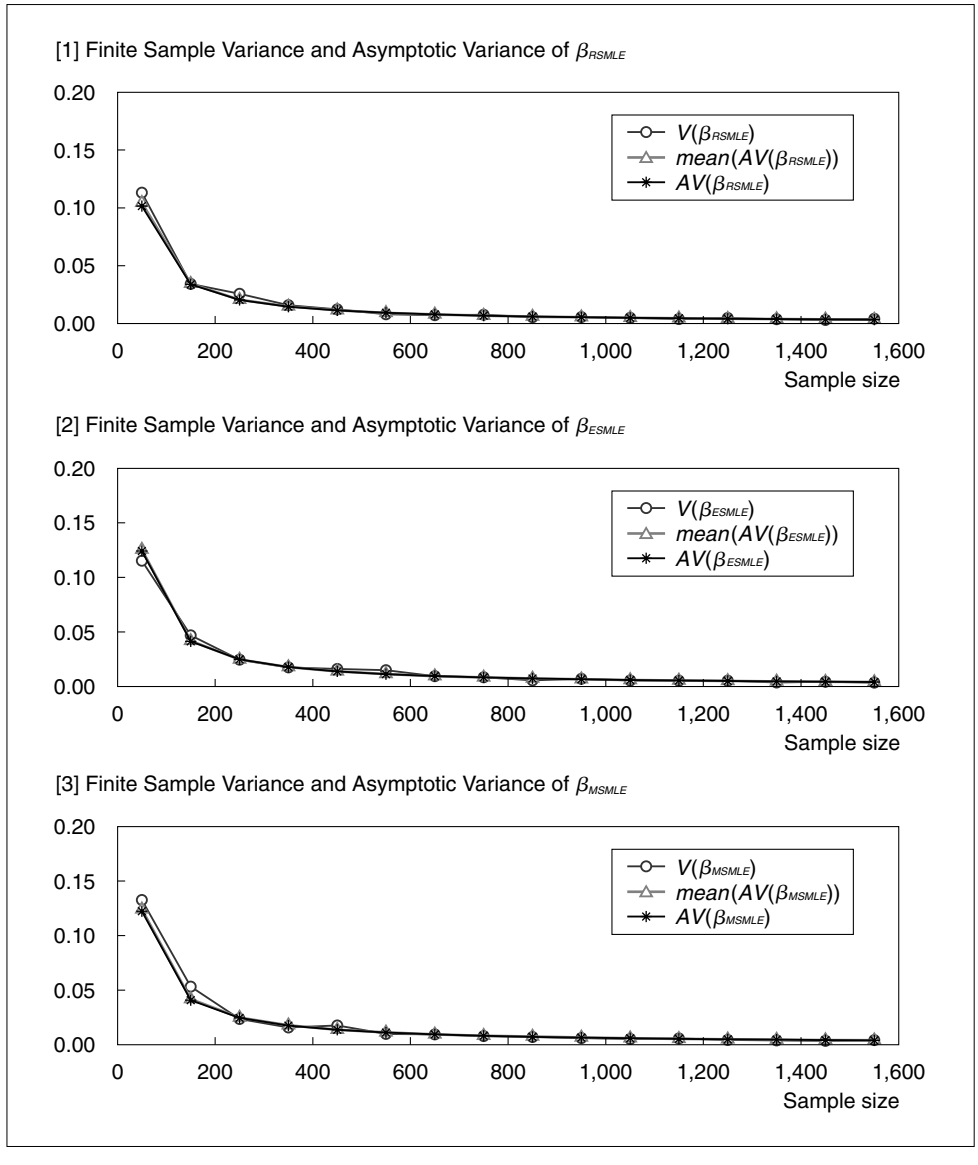


Figure 5 Monte Carlo Variance and Asymptotic Variance (Balanced Population)



In this Monte Carlo study, we fix λ_1 at 0.7. Using the same DGP, Appendix 4 illustrates the effect of varying λ_1 on the performance of the ESMLE and MSMLE.

V. Conclusion

Due to the low frequency of observing a right-censored (default) sample, many empirical duration analyses apply a mixed sampling procedure to collect data, that is, to over-sample from the right-censored (default) subset and match on one or more exogenous variables when sampling from the non-right-censored (non-default) subset. However,

the common practice of using estimation procedures intended for random sampling or for the qualitative response model under choice-based sampling will yield either inconsistent or inefficient estimators. This paper has established the fact that the parameters of a duration model can be estimated consistently under a mixed sampling procedure using the MSMLE and derived its asymptotic properties. In addition, this paper studies the relative efficiency of the RSMLE, ESMLE, and MSMLE through a simple example, which suggests a general guideline for sampling design in duration analyses: to over-sample default cases will usually gain efficiency when such cases are rarely observed; on the other hand, random sampling is generally better when the population is relatively balanced in terms of the frequency of default and non-default cases. In the case of vector parameters, the optimal choice of sampling proportions can have an interior solution. Since many empirical studies cited in the references tend to ignore the mixed sampling property, this paper also evaluates the asymptotic bias when the model is estimated as if it were a random sample or an endogenous sample without matching. We observe a large bias in certain examples, which indicates the importance of taking the sample designs into consideration when estimating duration models.

Despite the wide usage of mixed sampling in duration studies, we are not aware of any rigorous theoretical treatment of the problem yet. It is hoped that the present paper has bridged this gap in the literature and will improve the quality of empirical studies using duration models.

This paper, however, has several restrictive assumptions, and therefore, raises some future topics for research. For example, we are not yet in a position to treat cases where the matching procedure is based on more than one exogenous variable, although our conclusion can be directly applied to the case in which the matching procedure is based on linear combinations or other functions of multiple exogenous variables. In addition, we completely ignore the problem of heterogeneity. In future research, we will relax these restrictions and generalize the results to a more realistic level.

APPENDIX 1: PROOF OF EQUATION (14)

We copy equations (11) and (13) here for convenience:

$$AV(\hat{\beta}_{WMLE}) = G(A+B)^{-1}\left[\frac{P_1}{\lambda_1}A + \frac{P_0}{\lambda_0}B\right](A+B)^{-1}G', \tag{A.1}$$

$$AV(\hat{\beta}_{ESMLE}) = G\left[\frac{\lambda_1}{P_1}A + \frac{\lambda_0}{P_0}B\right]^{-1}G'. \tag{A.2}$$

To show $AV(\hat{\beta}_{ESMLE}) \leq AV(\hat{\beta}_{WMLE})$ is equivalent to showing

$$\frac{\lambda_1}{P_1}A + \frac{\lambda_0}{P_0}B \geq (A+B)\left[\frac{P_1}{\lambda_1}A + \frac{P_0}{\lambda_0}B\right]^{-1}(A+B). \tag{A.3}$$

Let $h = \lambda_1 P_0 / \lambda_0 P_1$, we can rewrite the above inequality as

$$A + \frac{1}{h}B \geq (A+B)[A+hB]^{-1}(A+B)$$

$$= A + (2-h)B + (1-h)^2 B[A+hB]^{-1}B. \tag{A.4}$$

Rearranging the above inequality yields

$$\frac{1}{h}B \geq B[A+hB]^{-1}B$$

$$\Leftrightarrow (hB)^{-1} \geq [A+hB]^{-1}. \tag{A.5}$$

Equation (14) follows from the fact that both A and B are non-negative definite.

APPENDIX 2: CONSISTENCY OF THE MSMLE

The consistency of the MSMLE when $g(z)$ is known and unknown follows from the two theorems in Goto (1993).

THEOREM 2.2.1 (The generalized Amemiya conditions) *The maximum likelihood estimator of γ is consistent if the density $f(x|\gamma)$ of i.i.d. random variables $\{x_i\}$ satisfies the following four assumptions:*

ASSUMPTION 1 (Parameter space): *The closure $\bar{\Gamma}$ of the parameter space Γ is a compact metric space.*

ASSUMPTION 2 (Continuity): *$f(x|\gamma)$ is a measurable function of x in a Euclidean space for all $\gamma \in \bar{\Gamma}$ and continuous in γ for almost all x.*

ASSUMPTION 3 (Identifiability): *If* $\gamma \neq \gamma_0$ *(the true value), then there exists some set A such that* $\int_A f(x|\gamma)dx \neq \int_A f(x|\gamma_0)dx$.

ASSUMPTION 4 (Integrability): $E \sup_{\gamma \in \bar{\Gamma}} |\ln f(x|\gamma)| < \infty$, *where the expectation is taken under the true value* γ_0.

THEOREM 2.3.1 *The maximum likelihood estimator of* γ *is consistent if the density* $f(x|\gamma)$ *of i.i.d. random variables* $\{x_i\}$ *satisfies the following four assumptions:*

ASSUMPTION 1 (Parameter space): *The parameter space is* $\Gamma = \Theta \times M$ *(i.e., the product of a subset* Θ *of* R^K *and a subset M of* H^S*, the product space of S identical H's, where H denotes the space of uniformly bounded nondecreasing measurable functions) with the metric defined as* $d(\gamma_1, \gamma_2) = \sum_{i=1}^{K} |\theta_1^i - \theta_2^i| + \sum_{i-1}^{S} \int_Z |H_1^i(z) - H_2^i(z)| dz$.

ASSUMPTION 2 (Continuity): $f(x|\gamma)$ *is a measurable function of x in a Euclidean space for all* $\gamma \in \bar{\Gamma}$ *and continuous in* γ *for almost all x, where* $\bar{\Gamma} = \bar{\Theta} \times \bar{M}$ *is the completion of* $\Gamma = \Theta \times M$. *In addition, if* Θ *is not bounded,* $\lim_{|\theta| \to \infty} f(x|\theta, H) = 0$ *for almost all x's and all H's.*

ASSUMPTION 3 (Identifiability): *If* $\gamma \neq \gamma_0$ *(the true value), then there exists some set A, such that* $\int_A f(x|\gamma)dx \neq \int_A f(x|\gamma_0)dx$.

ASSUMPTION 4 (Integrability): *For the true value* γ_0, $E|\ln f(x|\gamma_0)| < \infty$. *For any* $\gamma \in \bar{\Gamma}$, *there exist* $\rho > 0$, *such that* $E[\sup_{d(\gamma', \gamma) \leq \rho} \ln f(x|\gamma')]^+ < \infty$. *In addition, if* Θ *is not bounded, there exists* $\rho > 0$, *such that* $E[\sup_{d(\gamma', 0) \leq \rho} \ln f(x|\gamma')]^+ < \infty$, *where all the expectations above are taken under the true value* γ_0.

APPENDIX 3: LIKELIHOOD FUNCTIONS FOR THE MONTE CARLO STUDY

A. RSMLE

Log-likelihood function:

$$l(\beta; D, X, T, Z, p) = \sum_i [D(\beta z - e^{\beta z} t) - (1 - D)e^{\beta z}(1 - x)] + constant. \tag{A.6}$$

First-order condition:

$$\sum_i [D(z - e^{\beta z} t z) - (1 - D)z e^{\beta z}(1 - x)] = 0. \tag{A.7}$$

B. ESMLE
Log-likelihood function:

$$l(\beta; D, X, T, Z, p) = \sum_i [D(\beta z - e^{\beta z} t - \ln P_1) - (1-D)(e^{\beta z}(1-x) + \ln P_0)] + constant. \quad \text{(A.8)}$$

First-order condition:

$$\sum_i [D(z - e^{\beta z} tz) - (1-D) z e^{\beta z}(1-x)] - \left(\frac{N_1}{P_1} - \frac{N_0}{P_0}\right)\frac{\partial P_1}{\partial \beta} = 0. \quad \text{(A.9)}$$

C. MSMLE
Log-likelihood function:

$$l(\beta; D, X, T, Z, p) = \sum_i [D(\beta z - e^{\beta z} t - \ln P_1) + (1-D)(\beta z - e^{\beta z}(1-x) - \ln(1 - e^{-e^{\beta z}}))] + constant. \quad \text{(A.10)}$$

First-order condition:

$$\sum_i \left[z - e^{\beta z} z \left(Dt + (1-D)\left(1 - x + \frac{e^{-e^{\beta z}}}{1 - e^{-e^{\beta z}}}\right)\right)\right] - \frac{N_1}{P_1}\frac{\partial P_1}{\partial \beta} = 0. \quad \text{(A.11)}$$

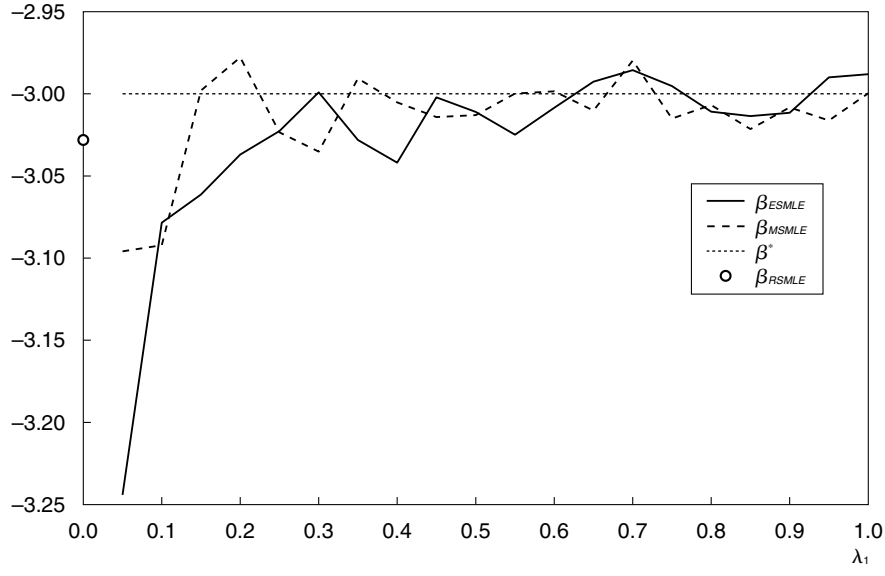
APPENDIX 4: EFFECTS OF VARYING λ_1 ON THE ESMLE AND MSMLE

CASE 1. *Unbalance Population:* $\beta = -3$, $p = 0.8$, $P(D = 1) = 0.09$.

CASE 2. *Balance Population:* $\beta = 1$, $p = 0.3$, $P(D = 1) = 0.45$.

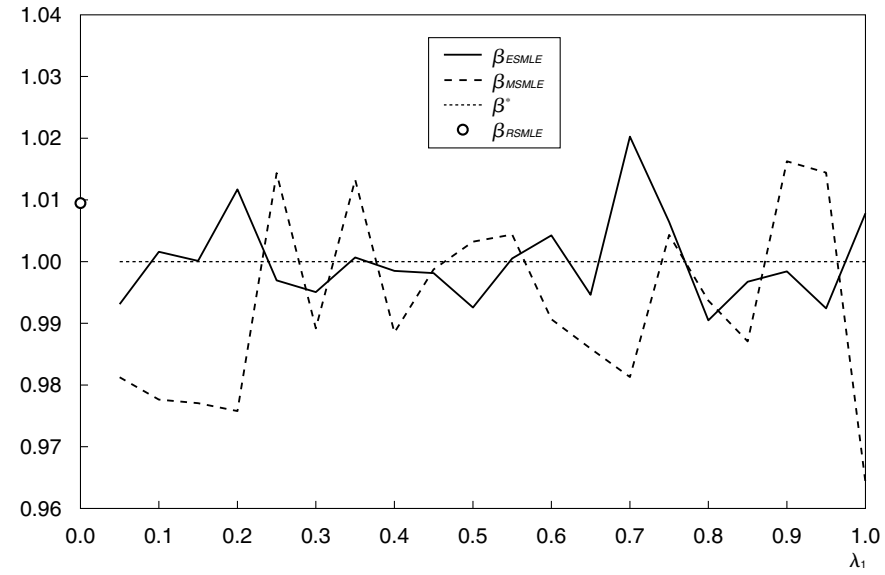
In both of the cases, we fix the sample size at 500 and estimate β_{ESMLE} and β_{MSMLE} from 100 simulations. Appendix Figures 1 and 2 report the mean of estimated β_{ESMLE} and β_{MSMLE} as λ_1 varies, for the unbalanced and balanced cases, respectively. Appendix Figures 3 and 4 present the empirical variance of the estimators based on the Monte Carlo simulations $\{V(\beta_k)\}$ and the asymptotic variances $\{AV(\beta_k)\}$ for the two estimators $\{k = ESMLE, MSMLE\}$. The values of β_{RSMLE} and $AV(\beta_{RSMLE})$ are circled on the vertical axes for comparison. Appendix Figures 1 and 3 show that when $P(D = 1)$ is extremely small, performance of β_{ESMLE} and β_{MSMLE} improves as λ_1 increases, both in terms of small sample properties and in terms of asymptotic efficiency. Appendix Figures 2 and 4 shows that the opposite is true when the population is relatively balanced.

Appendix Figure 1 β_{ESMLE} and β_{MSMLE} with Varying λ_1 (Unbalanced Population)



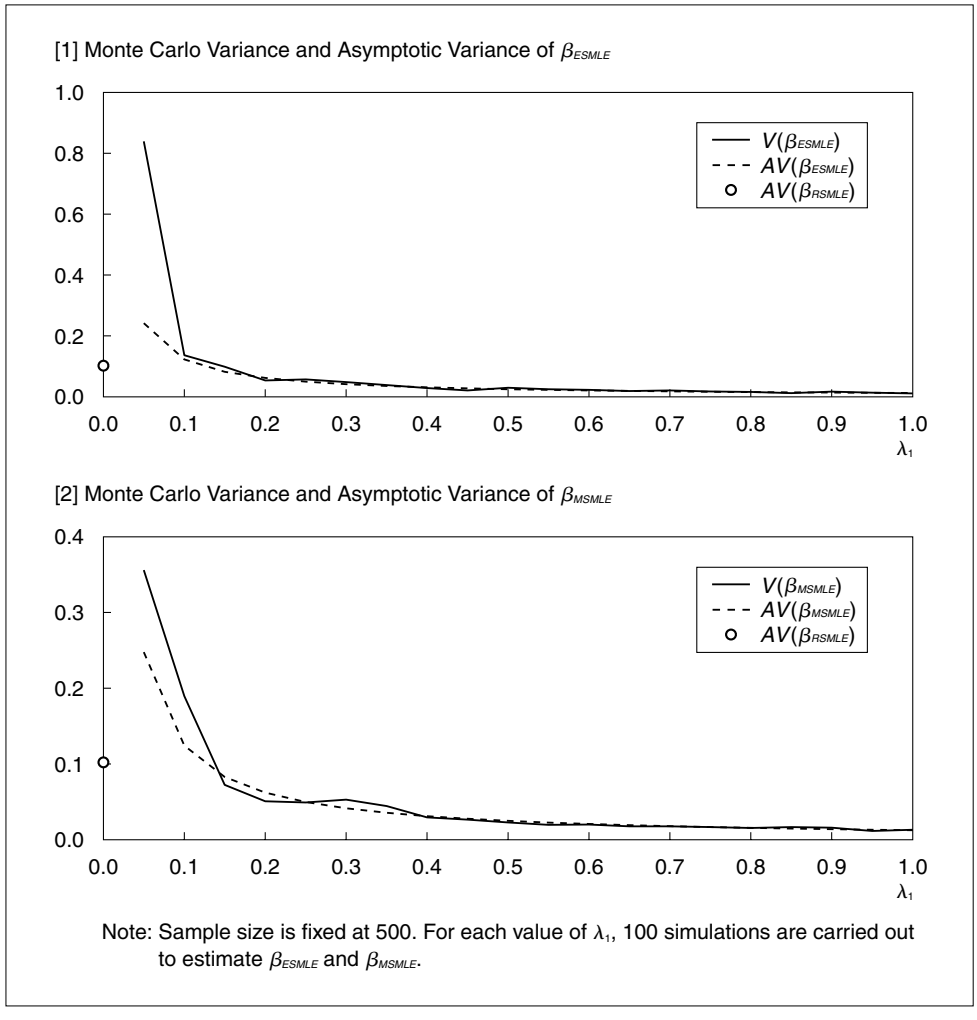
Note: Sample size is fixed at 500. For each value of λ_1, 100 simulations are carried out to estimate β_{ESMLE} and β_{MSMLE}.

Appendix Figure 2 β_{ESMLE} and β_{MSMLE} with Varying λ_1 (Balanced Population)



Note: Sample size is fixed at 500. For each value of λ_1, 100 simulations are carried out to estimate β_{ESMLE} and β_{MSMLE}.

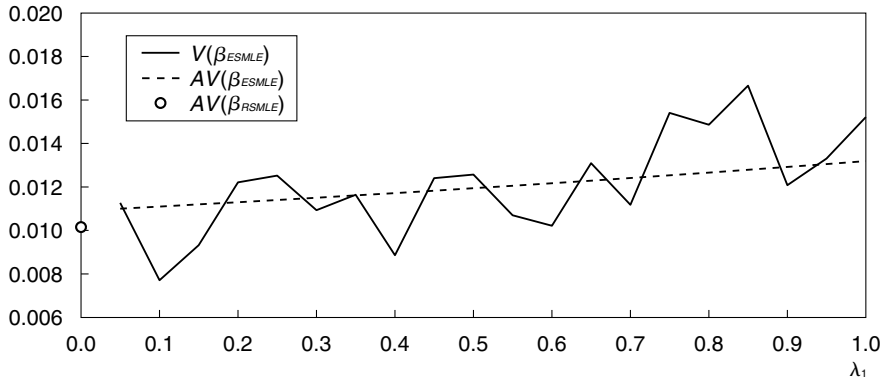
Appendix Figure 3 Monte Carlo Variance and Asymptotic Variance (Unbalanced Population)



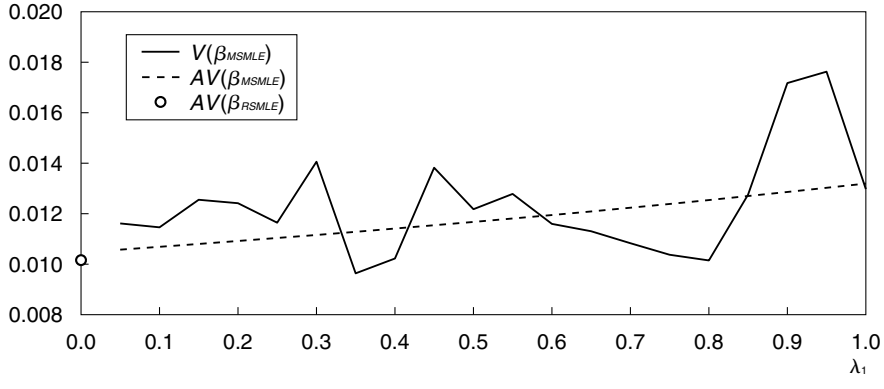
Note: Sample size is fixed at 500. For each value of λ_1, 100 simulations are carried out to estimate β_{ESMLE} and β_{MSMLE}.

Appendix Figure 4 Monte Carlo Variance and Asymptotic Variance (Balanced Population)

[1] Monte Carlo Variance and Asymptotic Variance of β_{ESMLE}



[2] Monte Carlo Variance and Asymptotic Variance of β_{MSMLE}



Note: Sample size is fixed at 500. For each value of λ_1, 100 simulations are carried out to estimate β_{ESMLE} and β_{MSMLE}.

References

Amemiya, T., *Advanced Econometrics*, Cambridge, Massachusetts: Harvard University Press, 1985.

———, "Endogenous Sampling in Duration Models," *Monetary and Economic Studies*, 19 (3), Institute for Monetary and Economic Studies, Bank of Japan, 2001, pp. 77–96.

BarNiv, R., "Accounting Procedures, Market Data, Cash Flow Measures and Insolvency Classification: The Case of the Insurance Industry," *Accounting Review*, 65 (3), 1990, pp. 578–604.

Goto, F., "Consistency and Efficiency of Semiparametric Estimators in Left-Censored Duration Models," Ph.D. thesis, Stanford University, 1993.

Kiefer, J., and J. Wolfowitz, "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *Annals of Mathematical Statistics*, 27, 1956, pp. 887–906.

Kim, Y.-D., D. R. Anderson, T. L. Amburgey, and J. C. Hickman, "The Use of Event History Analysis to Examine Insurer Insolvencies," *Journal of Risk and Insurance*, 62, 1995, pp. 94–110.

Lane, W. R., S. W. Looney, and J. W. Wansley, "An Application of the Cox Proportional Hazards Model to Bank Failure," *Journal of Banking and Finance*, 10, 1986, pp. 511–531.

Lee, S. H., and J. L. Urrutia, "Analysis and Prediction of Insolvency in the Property-Liability Insurance Industry: A Comparison of Logit and Hazard Models," *Journal of Risk and Insurance*, 63, 1996, pp. 121–130.

Luoma, M., and E. K. Laitinen, "Survival Analysis as a Tool for Company Failure Prediction," *Omega (International Journal of Management Science)*, 19, 1991, pp. 673–678.

Manski, C. F., and S. R. Lerman, "The Estimation of Choice Probabilities from Choice-Based Samples," *Econometrica*, 45, 1977, pp. 1977–1988.

Palepu, K., "Predicting Takeover Targets: A Methodological and Empirical Analysis," *Journal of Accounting and Economics*, 8 (1), 1986, pp. 3–35.