

**Maximum Score Estimation of Binary Choice Panel Data Models
with Sample Selection: an Application to Navy Officer Retention**

by Walter J. Mayer, Tanja Blackstone, Baoping Shang, and N. Keith Womer

School of Business Administration
University of Mississippi
University, MS 38677
(wmayer@bus.olemiss.edu)

*This research was supported by Office of Naval Research grant N00014-98-1-0768. We would like to thank Dong Chen, Meijun Qian, and Hong Xie for their valuable research assistance.

Abstract

Predictions of job retention are often made from maximum likelihood estimates that rely on strong but convenient assumptions for the distribution of unobserved random variables. If the distribution is misspecified, the predictions can be misleading. We propose a new approach that relies on less restrictive assumptions than previous studies. The approach is an extension of Manski's (1985, 1987) maximum score estimator to a binary choice panel data model with repeated sample selection. The estimator is shown to be consistent under assumptions of conditional symmetry and independence. The estimator also allows more general forms of autocorrelation, and is more robust to measurement errors than maximum likelihood. For a retention model of U.S. Navy officers, maximum score estimates are computed and compared to maximum likelihood estimates. The maximum score and maximum likelihood predicted probabilities of retention are also compared.

To design compensation policies and control turnover costs, policymakers and personnel managers need accurate predictions of job retention. Often, predictions are based on models that rely on strong, but convenient assumptions for the distributions of unobserved random variables. The latter are usually assumed to be normally distributed with a simple autocorrelation structure. If the distribution is misspecified, predictions of retention can be misleading. Unfortunately, we seldom, if ever, possess enough prior knowledge to justify such assumptions. This paper applies a new approach to retention data that relies on less restrictive assumptions than previous studies. The approach is an extension of Manski's (1985,1987) maximum score estimator to a binary choice panel data model with repeated sample selection.

We developed the approach to analyze a sample of officers in the Navy's nuclear submarine and surface communities for the period 1978 to 1994. The decision to stay or leave the Navy differs from the usual employment situation where the employee can leave at any time. The option to leave the Navy is only available at specific points of time. At each of those points, we observe the decision. Consequently, we model retention as a sequence of decisions rather than just a duration. The net value of staying in the Navy for the i -th officer at decision point t , y^*_{it} , is assumed to take the form:

$$y^*_{it} = x_{it}\beta_t + c_i + u_{it} \quad i=1, \dots, N; t=1, \dots, T_i \quad (1)$$

where x_{it} is a vector of observed explanatory variables, β_t is a parameter vector; the unobserved disturbance, $c_i + u_{it}$, consists of two components: c_i is a permanent, individual-specific component that does not change across decision points, while u_{it} is a transitory component. Let $y_{it} = I(y^*_{it} > 0)$, where $I(\cdot)$ is the indicator function. Thus, $y_{it}=1$ if the i -th officer stays, and $y_{it}=0$ if leaves. Sample data consist of observations on

(y_{it}, x_{it}) . Individuals are tracked only as long as they are in the Navy; therefore, the data are censored: y_{it} is observed only when $y_{it-1} = 1$. More formally, we have:

Assumption 1

- (a) $\{(y_{it}, x_{it}); i=1, \dots, N\}$ is an i.i.d. sequence for each $t=1, \dots, T_i$.
- (b) For all i and t , (y_{it}, x_{it}) is observed only if $y_{it-1} = \dots = y_{i1} = 1$.

The problem is to estimate β_i and the choice probabilities. Under Assumption 1, the latter are given by:

$$Q_{ij}^L = P(\text{officer } i \text{ leaves in period } J \leq T_i \mid x_{iJ}, \dots, x_{i1})$$

$$= P(y_{iJ}=0, y_{iJ-1}=1, \dots, y_{i1}=1 \mid x_{iJ}, \dots, x_{i1})$$

$$Q_{ij}^S = P(\text{officer } i \text{ leaves in period } J > T_i \mid x_{iJ}, \dots, x_{i1})$$

$$= P(y_{iT_i}=1, y_{iT_i-1}=1, \dots, y_{i1}=1 \mid x_{iT_i}, \dots, x_{i1})$$

If the choice probabilities are correctly specified up to an unknown parameter vector, then maximum likelihood (ML) is (given regularity conditions) $n^{1/2}$ -consistent (for fixed T_i), asymptotically normal and efficient. An early approach is the probit MLE proposed by Butler and Moffitt (1982), and used by Black, Moffitt, and Warner (1990) to analyze retention data on civilian workers in the U.S. government sector. The Butler-Moffitt approach assumes:

Assumption 2

- (a) For all i and t : c_i and u_{it} are normally distributed conditional on (x_{it}, \dots, x_{i1}) with means zero, and variances σ_c^2 and σ_u^2 , and covariance zero.
- (b) When $i \neq j$ or $t \neq s$, $E(u_{it}u_{js}) = E(c_i c_j) = 0$.

Butler and Moffitt (1982) is computationally convenient, but consistency for the estimator depends on assumptions that cannot be justified by a priori considerations. For one, Assumption 2(b) implies that the correlation between disturbances $c_i + u_{it}$ for any two decision points is the same regardless of how far apart the decisions. This is obviously unrealistic. If, for example, u_{it} reflects tastes that are gradually changing over time, then one would expect the decision to stay in the Navy at a given decision point to be most correlated with the most recent decisions.

When Butler and Moffitt (1982) was published, the computational costs of evaluating high dimensional integral expressions precluded autocorrelation in u_{it} . With the development of methods such as the Geweke-Hajivassiliou-Keane (GHK) simulator and simulated MLE, however, this is no longer the case. Assumption 2(b) can now be relaxed by specifying an autoregressive process for u_{it} , and using simulated MLE. (Surveys of simulation methods for panel probit and other models can be found in Hajivassiliou (1993), Keane (1993), and Hajivassiliou and Ruud (1994).

While methods are available for relaxing the equicorrelation assumption, none have been proposed for the assumption of normally distributed errors (Assumption 2a). To develop such a method, our starting point is Manski (1987). Manski (1987) proposes a

maximum score (MSCORE) estimator for binary choice panel data model that is consistent under weak distributional assumptions.¹ The score function is based on the transformation $(y_{it} - y_{it-1})I(y_{it} \neq y_{it-1})$, which “differences out” the individual effect, c_i , in (1). Consequently, the consistency theorem (Manski 1987, p.361) places no restrictions on the distribution of c_i . Manski’s (1987) assumptions, however, do not allow the sample selection specified in Assumption 1(b) above, and Manski’s (1987) estimator is in fact inconsistent under Assumption 1(b).

In Section 2, we show that consistent MSCORE estimation is possible under Assumption 1(b) if the score function is based on y_{it} instead of $(y_{it} - y_{it-1})I(y_{it} \neq y_{it-1})$. This form of MSCORE allows sample selection, but requires restrictions on c_i . Propositions 1 and 2 in Section 2 establish that the MSCORE estimator based on y_{it} is consistent under “random effects” assumptions of conditional symmetry and independence. This is in contrast to the weaker assumptions of median independence in Manski (1985,1987). The need for stronger assumptions here arises from the sample selection.

An obvious advantage of the proposed MSCORE estimator over maximum likelihood is that it does not require prior knowledge of normality or any other finite-dimensional parametric family. In Section 2, we show that MSCORE has also has two other less obvious advantages. First, it does not require the researcher to parameterize the autocorrelation function for u_{it} and thus allows more flexible dynamics for unobserved tastes. Second, MSCORE is more robust to measurement errors than maximum likelihood. More specifically, we show that if the observations at the t -th decision point: $(y_{it}, x_{it}); i=1, \dots, N$, are not measured with error, then the

MSCORE estimator of β_i is consistent even if the other T-1 decision points contain measurement errors. In other words, when MSCORE is used, the adverse effect of measurement errors is confined to the decision points that contain them. Section 2 also discusses specification testing and prediction.

Semiparametric methods such as MSCORE have not been used previously to estimate models of retention decisions. Consequently, it is not known whether they would produce results substantially different than traditional ML methods. Section 3 reports ML and MSCORE estimates of models of retention for a sample of 8042 officers. Some findings are robust across the ML and MSCORE specifications. In both, for example, the “annualized cost of leaving” variable, ACOL is positively and significantly related to the probability that the officer stays in the Navy. The ACOL variable was originally proposed by Warner and Goldberg (1984), and is widely used in applied retention studies. Other findings, however, are sensitive to the estimation method. For example, several variables are statistically significant only for the ML estimates, the ML and MSCORE predicted probabilities of retention differ by as much as 30 percent. Using a goodness-of-fit measure adapted from Horowitz (1993), we also find that MSCORE provides a better description of the data at some decision points. Traditionally, researchers have relied on ML to predict retention. Since ML imposes stronger assumptions than MSCORE, these results suggest that ML predictions should be viewed cautiously.

2. Maximum Score Approach

2.1 Estimator

Let $Y_{it} = \text{sgn}(y_{it}^*)$, and $z_{it} = I(y_{it-1} = \dots = y_{i1} = 1)$. Therefore,

$$E(Y_{it}|x_{it}, z_{it}=1) = \{P(y_{it}=1, z_{it}=1|x_{it}) - P(y_{it}=0, z_{it}=1|x_{it})\} / P(z_{it}=1|x_{it}) \quad (2)$$

Identification of β_t requires:

$$x_{it}\beta_t > 0 \Leftrightarrow E(Y_{it}|x_{it}, z_{it}=1) > 0 \text{ and } x_{it}\beta_t < 0 \Leftrightarrow E(Y_{it}|x_{it}, z_{it}=1) < 0 \quad (3)$$

The inequalities in (3) are analogous to Manski (1985, p.315), and provide the basis for a MSCORE estimator of β_t under “random effects” assumptions given below.

Obviously, $E(Y_{it}|x_{it}, z_{it}=1)$ depends on the joint distribution of the current and past disturbances. As we will see, when this distribution is restricted by the next assumption, β_t can be identified through restrictions on the conditional distributions of c_i and u_{it} .

Assumption 3

$u_{it} = g_{it}(v_{it-1}, v_{it-2}, \dots) + v_{it}$, where g_{it} is a measurable function, and

(i) v_{it} is conditionally independent of $(v_{it-1}, v_{it-2}, \dots)$ given $(\mathbf{x}_{it}, c_i + g_{it}(v_{it-1}, v_{it-2}, \dots))$,

where $\mathbf{x}_{it} = (x_{it}, \dots, x_{i1})$,

(ii) v_{it} is conditionally independent of $c_i + g_{it}(v_{it-1}, v_{it-2}, \dots)$ given \mathbf{x}_{it} .

Assumption 3 allows most types of autocorrelation considered in practice such as ARMA processes. In what follows, we let $\varepsilon_{it} = g_{it}(v_{it-1}, v_{it-2} \dots)$, $\eta_{it} = c_i + \varepsilon_{it}$, and $\boldsymbol{\eta}_{it} = (\eta_{it}, \dots, \eta_{i1})$.

Assumption 3 implies that the decision to stay in the Navy in period t is *conditionally* independent of previous decisions given $(\mathbf{x}_{it}, \boldsymbol{\eta}_{it})$. Therefore, we can write:

$$P(y_{it}=1, z_{it}=1 | \mathbf{x}_{it}, \boldsymbol{\eta}_{it}) - P(y_{it}=0, z_{it}=1 | \mathbf{x}_{it}, \boldsymbol{\eta}_{it}) = D_{it}(\eta_{it}) P(z_{it}=1 | \mathbf{x}_{it}, \boldsymbol{\eta}_{it}), \text{ where}$$

$$D_{it}(\eta_{it}) = P(v_{it} > -x_{it}\beta_t - \eta_{it} | \mathbf{x}_{it}, \boldsymbol{\eta}_{it}) - P(v_{it} \leq -x_{it}\beta_t - \eta_{it} | \mathbf{x}_{it}, \boldsymbol{\eta}_{it}).$$

Conditional on $(\mathbf{x}_{it}, \boldsymbol{\eta}_{it})$, $D_{it}(\eta_{it})$ is the difference between the “stay” and “leave” probabilities for the i th officer in period t , while $P(z_{it}=1 | \mathbf{x}_{it}, \boldsymbol{\eta}_{it})$ is the probability that the officer stays in the Navy at least t periods. By iterated expectations, we have:

$$E(Y_{it} | \mathbf{x}_{it}, z_{it}=1) = \int \{ \int D_{it}(\eta_{it}) P(z_{it}=1 | \mathbf{x}_{it}, \boldsymbol{\eta}_{it}) dF(\boldsymbol{\eta}_{it} | \mathbf{x}_{it},) \} dF(\mathbf{x}_{it-1} | \mathbf{x}_{it}) / P(z_{it}=1 | \mathbf{x}_{it}) \quad (4)$$

where $F(\boldsymbol{\eta}_{it} | \mathbf{x}_{it},)$ denotes the distribution function of $\boldsymbol{\eta}_{it}$ conditional on \mathbf{x}_{it} .

Consequently, under Assumption 3, $E(Y_{it} | \mathbf{x}_{it}, z_{it}=1)$ can be viewed as a weighted average of the difference $D_{it}(\eta_{it})$, where the “weights” depend on the conditional distributions of v_{it} given $(\mathbf{x}_{it}, \boldsymbol{\eta}_{it})$ and $\boldsymbol{\eta}_{it}$ given \mathbf{x}_{it} .

The next task is to derive sufficient conditions for (3) from (4). We first note that median independence, which suffices for the standard cross-section maximum score (Manski 1985), does not suffice in the present problem because of the sample selection. To see this, consider the term in brackets in (4) which can be written as:

$$\int_A D_{it}(\eta_{it})P(z_{it}=1 | \mathbf{x}_{it}, \boldsymbol{\eta}_{it})dF(\boldsymbol{\eta}_{it} | \mathbf{x}_{it},) + \int_{A^c} D_{it}(\eta_{it}) P(z_{it}=1 | \mathbf{x}_{it}, \boldsymbol{\eta}_{it})dF(\boldsymbol{\eta}_{it} | \mathbf{x}_{it},) \quad (5)$$

where $A = \{ \eta_{it} \in (-x_{it}\beta_t, \infty) \}$, and $A^c = \{ \eta_{it} \in (-\infty, -x_{it}\beta_t) \}$. Suppose $x_{it}\beta_t > 0$.

Clearly, $MED(v_{it} | \mathbf{x}_{it}, \boldsymbol{\eta}_{it}) = 0$ implies $D_{it}(\eta_{it}) > 0$ on A and $D_{it}(\eta_{it}) < 0$ on A^c . Median independence, however, restricts neither the “weights” $P(z_{it}=1 | \mathbf{x}_{it}, \boldsymbol{\eta}_{it}) dF(\boldsymbol{\eta}_{it} | \mathbf{x}_{it},)$ nor the magnitudes of $D_{it}(\eta_{it})$. Consequently, it follows from (4) and (5) that the sign of $E(Y_{it} | \mathbf{x}_{it}, z_{it}=1)$ is ambiguous.

To restrict the weights and magnitudes of $D_{it}(\eta_{it})$, we assume that $P(z_{it}=1 | \mathbf{x}_{it}, \boldsymbol{\eta}_{it})$ is a nondecreasing function of η_{it} , and that the conditional distributions of v_{it} given $(\mathbf{x}_{it}, \boldsymbol{\eta}_{it})$ and $\boldsymbol{\eta}_{it}$ given \mathbf{x}_{it} are each symmetric around zero. The assumption on $P(z_{it}=1 | \mathbf{x}_{it}, \boldsymbol{\eta}_{it})$ merely requires that “effects” that increase the probability of staying in the t -th period (that is, large values of η_{it}) do not decrease the probability of staying at least t periods (given \mathbf{x}_{it}). Given $x_{it}\beta_t > 0$, conditional symmetry implies: (a) $D_{it}(\eta_{it}) > 0$ on A , and $D_{it}(\eta_{it}) < 0$ on A^c ; and (b) negative values of $D_{it}(\eta_{it})$ on A^c are offset by positive values of $D_{it}(\eta_{it})$ on a *proper* subset of A , namely $(x_{it}\beta_t, \infty)$, of the same magnitude. As a result, provided that $P(\eta_{it} \in (-x_{it}\beta_t, x_{it}\beta_t)) > 0$, the first term in

(5) dominates the second term and, thus, $E(Y_{it}|x_{it}, z_{it}=1) > 0$. A similar argument establishes $x_{it}\beta_t < 0 \Leftrightarrow E(Y_{it}|x_{it}, z_{it}=1) < 0$.

Assumption 4 and Proposition 1 below formalize the above assumptions and argument.

Assumption 4

- a. For all i and t , and almost every \mathbf{x}_{it} : $\boldsymbol{\eta}_{it}$ has a positive density conditional on \mathbf{x}_{it} that is *component wise* symmetric around zero; that is, $f_{it}(\boldsymbol{\eta}_{it}) = f_{it}(-\boldsymbol{\eta}_{it})$.
- b. For all i and t , and almost every $(\mathbf{x}_{it}, \boldsymbol{\eta}_{it})$: v_{it} has a positive density conditional on $(\mathbf{x}_{it}, \boldsymbol{\eta}_{it})$ that is symmetric around zero.
- c. $P(z_{it}=1 | \mathbf{x}_{it}, \boldsymbol{\eta}_{it})$ is a nondecreasing function of $\boldsymbol{\eta}_{it}$.

Assumptions 4(a) and 4(b) are a “random effects” specification that nests Assumption 2. Assumptions 4(a) and 4(b) hold if the individual component, c_i , and the transitory components, ε_{it} and v_{it} , are each conditionally symmetric around zero. Conditional symmetry, in turn, implies that these components are each uncorrelated with the regressors, x_{it} .

Proposition 1

Under Assumptions 1, 3, and 4:

$$x_{it}\beta_t > 0 \Leftrightarrow E(Y_{it}|x_{it}, z_{it}=1) > 0 \text{ and } x_{it}\beta_t < 0 \Leftrightarrow E(Y_{it}|x_{it}, z_{it}=1) < 0.$$

Proof: See the Appendix A.

Proposition 1 implies that the function $E(Y_{it}|x_{it}, z_{it}=1)I(x_{it}b_t > 0)$ has a maximum at $b_t = \beta_t$ and, therefore, suggests the use of the score function:

$$S_{NT}(\mathbf{b}) = \sum_{t=1}^T s_{Nt}(\mathbf{b}_t) \quad (6)$$

where $\mathbf{b} = (\mathbf{b}_1, \dots, \mathbf{b}_T)$ and $s_{Nt}(\mathbf{b}_t) = N^{-1} \sum_{i=1}^N Y_{it} I(\mathbf{x}_{it} \mathbf{b}_t \geq 0)$.

With the following additional, standard assumption (see Manski 1985, Assumption 2), Proposition 2 below shows that $S(\mathbf{b}) \equiv E[S_{NT}(\mathbf{b})]$ has a unique maximum at the normalized vector $\beta^* \equiv \beta / \|\beta\|$.

Assumption 5

For all i and t :

- (a) The support of \mathbf{x}_{it} is not contained in any proper linear subspace of k -dimensional Euclidean space.
- (b) There is at least one component of \mathbf{x}_{it} that has a nonzero coefficient, and has everywhere positive density conditional on the other components of \mathbf{x}_{it} .

Proposition 2

Let Assumptions 1, 3, 4 and 5 hold, $\beta_t^* \equiv \beta_t / \|\beta_t\|$ and $s_t(\mathbf{b}_t) = E[s_{Nt}(\mathbf{b}_t)]$.

Then $s_t(\beta_t^*) > s_t(\mathbf{b}_t)$ for all \mathbf{b}_t such that $\mathbf{b}_t \neq \beta_t^*$.

Proof: See Appendix B.

Under the assumptions of Proposition 2, uniform convergence of $S_{NT}(\mathbf{b})$ to $S(\mathbf{b})$ and continuity of $S(\mathbf{b})$ can be established using arguments analogous to those in Manski (1985). Consequently, given Proposition 2, consistency for any global maximizer of (6) follows from an application of Manski (1985).

2.2 Comparison of MSCORE and MLE

One difference between MSCORE and ML is that when no cross-equation restrictions are imposed on the coefficients, the MSCORE estimator of β_t uses only the sample observations of the t -th decision point, (Y_{it}, X_{it}) . This can be easily seen from (6): the MSCORE estimator of β_t requires only maximization of $s_{Ni}(b_t)$. Consequently, the MSCORE estimator of β_t will be robust to some forms of misspecification such as measurement errors at the other decision points. In particular, the MSCORE estimator of β_t will be consistent provided that for the previous decision points we have

$$y^*_{is} = x_{is}\beta_s + c_i + u_{is} \quad (s=1, \dots, t-1)$$

for *some* x_{is} (not necessarily observable), and c_i , and u_{is} that satisfy Assumptions 3 and 4. This robustness comes at a cost of course, since the estimator of β_t does not use all of the observed data.

The MSCORE approach can produce consistent estimates from single cross-sections and, thus, essentially turns a panel data problem into a cross-section problem. ML, by contrast, requires panel data to correct for the sample selectivity: Applying standard cross-sectional ML to any period except the first does not produce consistent estimates. As a result, ML is less robust to measurement errors. This difference arises because ML depends on a stronger restriction of the form $E(y_{it}|x_{it}, z_{it}=1) = m_{it}(x_{it}\beta_t)$, where m_{it} is a known function. Except for the special case of no cross-equation correlation (that is, $c_i \equiv 0$ and $E(u_{it}u_{js}) = 0$ for all $t \neq s$), the function m_{it}

depends on *all* the data up to period t (see 2). Therefore, misspecifications at any of the previous decision points can adversely affect the ML estimator of β_t .

Another way to see this is to consider the score for the simple cross-sectional probit. Let $L_t(\beta)$ denote the “log-likelihood” for cross-sectional probit, and $\Phi(x_{it}\beta_t)$ and $\phi(x_{it}\beta_t)$ denote the corresponding standard normal CDF and density. Consider

$$E(\partial L_t(\beta) / \partial \mathbf{b} | x_{it}, z_{it}=1) = \Sigma E(y_{it} | x_{it}, z_{it}=1) \phi(x_{it}\beta_t) / \Phi(x_{it}\beta_t) \\ - \Sigma(1 - E(y_{it} | x_{it}, z_{it}=1)) \phi(x_{it}\beta_t) / (1 - \Phi(x_{it}\beta_t))$$

Since sample selectivity implies $E(y_{it} | x_{it}, z_{it}=1) \neq \Phi(x_{it}\beta_t)$ and, thus, in general $E(\partial L_t(\beta) / \partial \mathbf{b} | x_{it}, z_{it}=1) \neq 0$, the cross-sectional probit is inconsistent. In other words, cross-sectional probit breaks down because it depends on correct specification of $E(y_{it} | x_{it}, z_{it}=1)$ up to an unknown finite-dimensional parameter vector. In contrast, consistency for MSCORE depends on a weaker restriction, $\text{sgn}(E(Y_{it} | x_{it}, z_{it}=1)) = \text{sgn}(x_{it}\beta_t)$, whose sample analog can be constructed from cross-sectional data.

2.3 Prediction

One objective of retention models is to predict the effects on individual stay-leave decisions of changes in earnings and other factors. Consequently, it is of interest to compare the predictions implied by the ML and MSCORE estimates. We consider the problem of predicting the stay/leave decision in period J conditional on decisions up to period $J-1$. The relevant choice probability is:

$$Q_{ij} = P(y_{ij}=1 | z_{ij}=1, x_{ij}, \dots, x_{i1}) \quad (7)$$

Under Assumption 2a, (7) is estimated by approximating the multivariate normal CDF with the GHK simulator and replacing $\beta \equiv (\beta_1, \dots, \beta_J)$ with the MLE.

Under the semiparametric assumptions, we combine the MSCORE estimate of β , $\hat{\beta}$, with kernel regression. Let k be a kernel function, and γ_n be a bandwidth.

We assume:

Assumption 6

a. For all i and almost every (x_{ij}, \dots, x_{i1}) :

$$f(u_{i1}, \dots, u_{ij}, c_i | x_{ij}, \dots, x_{i1}) = f(u_{i1}, \dots, u_{ij}, c_i | x_{ij}\beta_J, \dots, x_{i1}\beta_1)$$

b. $\sup_{\tau} f_{\tau} < \infty$, where f_{τ} is the density of $\tau_i \equiv (x_{ij}\beta_J, \dots, x_{i1}\beta_1)$.

c. $\int |k(v)| dv < \infty$, $\int k(v) dv = 1$, $k(\cdot)$ has continuous and bounded first order derivatives.

d. $\lim \gamma_n = 0$, and $\lim n\gamma_n^J = \infty$.

It is possible to estimate (7) completely nonparametrically by computing the kernel estimate of the conditional mean of y_{ij} conditional on $(z_{ij}=1, x_{ij}, \dots, x_{i1})$. A well-known problem, however, is that the precision of nonparametric estimates deteriorates rapidly as the dimension of the conditioning vector increases. Since $(z_{ij}=1, x_{ij}, \dots, x_{i1})$ is relatively large (especially when $J > 1$), completely nonparametric estimates of (7) would probably not be very precise. To reduce the dimensionality of the conditioning vector, we impose an index restriction (Assumption 6a) that implies:

$$Q_{ij} = P(y_{ij}=1|z_{ij}=1, x_{ij}\beta_j, \dots, x_{i1}\beta_1) \quad (8)$$

Assumptions 6b-d are standard regularity conditions for kernel regression. For the bandwidth in our empirical application, we use $\gamma_n \equiv hn^{-1/(J+1)}$, where the value for h was chosen by least-squares cross-validation.

Given a kernel function k, define:

$$K_{imn}^x(\beta) = \prod_{i=1}^J k((x_{ij}\beta_j - x_{mj}\beta_j)/\gamma_n)$$

$$K_{imn}^y = \prod_{i=1}^{J-1} k((1 - y_{mj})/\gamma_n)$$

A semiparametric estimate of (8) is:

$$\hat{Q}_{ij}(\hat{\beta}) = \frac{\sum_{m \neq i} y_{mi} K_{imn}^y K_{imn}^x(\hat{\beta})}{\sum_{m \neq i} y_{mi} K_{imn}^y K_{imn}^x(\hat{\beta})} \quad (9)$$

Assuming $\text{plim } \hat{\beta} = \beta^*$, (9) is a consistent estimator under Assumption 6:

Proposition 3

Let Assumptions 1, 3, 4 and 6 hold. Then for any τ_i such that $f_{\tau} > 0$,

$$\text{plim } \hat{Q}_{ij}(\hat{\beta}) = \beta^*$$

Proof: See Appendix B.

2.3 Specification Testing

If the assumptions underlying the MSCORE estimator hold, then $\text{sgn}(E(Y_{it}|x_{it}, z_{it}=1)) = \text{sgn}(x_{it}\beta_t)$ for all i and t . Consequently, the MSCORE specification can be tested using a uniform confidence band for $E(Y_{it}|x_{it}, z_{it}=1)$. If the specification is correct, then for positive values of the MSCORE estimate, $\hat{x}_{it}\hat{\beta}_t$, the upper bound of the confidence band should be positive; for negative values of $\hat{x}_{it}\hat{\beta}_t$, the lower bound of the band should be negative. Similar procedures have been used by Horowitz (1993) and Charlier et al. (1995).

The confidence band is constructed using the kernel regression estimator of $E(Y_{it}|x_{it}, z_{it}=1)$. Since $E(Y_{it}|x_{it}, z_{it}=1) = E(Y_{it}|x_{it}\beta_t, z_{it}=1)$ under Assumption 6a, the band takes the form:

$$\hat{g}(x_{it}\hat{\beta}_t) = \frac{\sum_{j=1}^n Y_{jt} k((x_{it}\hat{\beta}_t - x_{jt}\hat{\beta}_t)/\gamma_n)}{\sum_{j=1}^n k((x_{it}\hat{\beta}_t - x_{jt}\hat{\beta}_t)/\gamma_n)}$$

where k and $\gamma_n \equiv hn^{-\delta}$ ($\delta > 0$) are the kernel and bandwidth. Assume:

Assumption 7

- (i) $E(Y_{it}|x_{it}\beta_t, z_{it}=1)$ and the density of $x_{it}\beta_t$, $f(x_{it}\beta_t)$, are continuous, uniformly bounded and have continuous and uniformly bounded first and second derivatives.
- (ii) $k(\cdot)$ is such that $\partial g(\cdot)/\partial \beta = O_p(1)$, $\int k(z)dz=0$, and $|\int z^2 k(z)dz| < \infty$.

The confidence band is based on:

Proposition 4

Let $G(x_{it}\beta_t) = E(Y_{it}|x_{it}\beta_t)$, $\sigma_t^2 = E((Y_{it} - G(x_{it}\beta_t))^2 | x_{it}\beta_t)$, and

$$\hat{\psi} = n\gamma_n \sum (Y_{jt} - \hat{g}(x_{it}\beta_t))^2 k((x_{it}\beta_t - \hat{x}_{jt}\beta_t) / \gamma_n) / \{ \sum k((x_{it}\beta_t - \hat{x}_{jt}\beta_t) / \gamma_n) \}^2$$

Suppose Assumptions 1,3,4,5,6(a) and 7 hold and $\delta > 1/3$. Then:

$$(i) \quad (n\gamma_n)^{1/2} (\hat{g}(x_{it}\beta_t) - G(x_{it}\beta_t)) \Rightarrow N(0, [\sigma_t^2 / f(x_{it}\beta_t)] \int z^2 k(z) dz)$$

$$(ii) \quad \text{plim } \hat{\psi} = [\sigma_t^2 / f(x_{it}\beta_t)] \int z^2 k(z) dz.$$

Proof: See Appendix B.

Given Proposition 4, an asymptotic $(1-\alpha)100\%$ confidence interval for $E(Y_{it}|x_{it}\beta_t, z_{it}=1)$ is $\hat{g}(x_{it}\beta_t) \pm Z_{\alpha/2} \hat{\psi}^{1/2}$. To compute the interval in the next section, we specified k as the standard normal density, and $\gamma_n = 0.25n^{-3/8}$, and chose $h=0.25$ by cross-validation.

3. Application to Officer Retention

3.1 Data

ML and MSCORE estimators were applied to data on the retention decisions of officers in the Navy's nuclear submarine and surface communities. The data were provided by the Office of Naval Research and assembled from the Navy's Officer Master File. It consist of observations on 8,042 unrestricted line officers for 10 decision points during the fiscal years 1978 to 1994.

Table 1 lists the explanatory variables in x_{it} . Human capital theory implies that the net value of staying in the Navy at decision point t , y_{it}^* , depends on the present value of earnings from alternative career paths over the remaining working life of the officer, costs of moving to the civilian sector, and on individual tastes for military life. Consequently, x_{it} includes: a measure of the difference between expected military earnings and alternative civilian earnings, ACOL (annualized cost of leaving); a measure of demand conditions in the civilian sector, UE (the national unemployment rate); and personal characteristics potentially related to job preference: NONWHT, LOS, ACAD, ROTC, and DEP.

The ACOL variable has become standard in retention studies (see Warner and Goldberg 1984, Black, Moffitt, and Warner 1991, and Nakada, Mackin, and Mackie 1998) and is based on the so called Annualized Cost of Leaving model. Since detailed treatments can be found in the above papers, we will provide only a brief description here. The ACOL variable used in our models was constructed by the Office of Naval Research. Let:

W_{Nm} = discounted expected military wage stream for the next N years.

W_{oc} = discounted civilian wage stream if leave the military immediately.

W_{Nc} = discounted expected civilian wage stream if leave the military after N more years.

r = personal discount rate.

The annualized pay difference is given by:

$$A_N = (W_{Nm} + W_{Nc} - W_{oc}) / \sum_{i=1}^N (1+r)^{-i}$$

The ACOL variable is defined as the maximum of A_N over values of N . The expected military wage stream, W_{Nm} , is calculated using historical pay tables. The values used for the civilian wage streams, W_{Nc} and W_{oc} , are based on predictions from an experience-earnings regression for a sample of civilian workers in engineering and managerial occupations. Black et al. (1991) and others have argued that ACOL is a reasonable approximation, and computationally much simpler than alternative approaches used by Gotz and McCall (1984) and Daula and Moffitt (1995), for example.

Table 2 presents sample means by decision point. As one would expect, the stay rate generally increases with the number of decision points: At decision point 1 (DP1) 72% stay, while at DP10 97% stay. The sample average of ACOL also increase with the decision point. Across all ten decision points most officers are white (97% at DP1) and have graduated from the Naval Academy or a ROTC program (65% at DP1). About 48% have dependents at DP1, while 86% have dependents at DP 10.

3.2 Estimation Results

Tables 3 and 4 report the estimates of the retention model. Simulated maximum likelihood estimates (SMLE) were computed using the Geweke-Hajivassiliou-Keane simulator to simulate the likelihood function. The simulated likelihood function was maximized by an optimization routine based on the Powell algorithm. Because gradient algorithms can converge to sub-optimal local maxima, the solution found by the Powell algorithm was checked using simulated annealing (Goffe, Ferrier, and Rogers 1994)

and genetic-algorithm searches (Dorsey and Mayer 1995 and Ostermark 1999).

The SMLE assume that the individual and transitory components of the disturbance are normally distributed and an AR(1) process: $u_{it} = \rho_A u_{it-1} + v_{it}$. In addition to the unrestricted SMLE, SML/ ML estimates were computed under restrictions of “no heterogeneity” (denoted $SMLE(\sigma_c^2 = 0)$), (“no autocorrelation” (denoted $MLE(\rho_A = 0)$) and “no autocorrelation or heterogeneity” (denoted $MLE(\rho_A = \sigma_c^2 = 0)$).

The MSCORE estimates, which require maximization of a discontinuous step function (see 6), were computed using a genetic-algorithm search. The numbers in parentheses are asymptotic standard errors. In contrast to ML, convenient expressions are not available for MSCORE and bootstrapping must be used (see Manski and Thompson 1986). It should be noted, however, that the MSCORE estimator does not satisfy the assumptions underlying the standard theories of the bootstrap (see Horowitz 1999). No theoretical justification is presently available, but the bootstrap has proven useful in other MSCORE applications. For example, the Monte Carlo results of Horowitz (1992) suggest that the bootstrap is more accurate in finite samples than approximations of first-order asymptotic theory.

First consider the “pooled” SML and ML estimates in the first three rows of Table 3. A positive (negative) coefficient implies the probability to stay increases (decreases) with the associated explanatory variable. ACOL measures the financial incentive to stay in the Navy as the difference between expected military and civilian earnings. Consistent with economic theory and estimates reported in previous studies, the coefficient of ACOL is positive and significant for all three sets of estimates. The coefficients of UE and LOS are also all positive. The positive coefficients for UE implies

that the probability to stay in the Navy is counter cyclical which agrees with Black, Moffitt, and Warner (1990), and other studies. A possible explanation for the positive coefficients of LOS is that individuals with a greater taste for military life initially choose longer obligations. In terms of statistical significance, however, the estimates for UE and LOS are sensitive to the restrictions: The SMLE are insignificant at conventional levels, while both $\text{MLE}(\rho_A = \sigma_c^2 = 0)$ and $\text{MLE}(\rho_A = 0)$ are significant.

Table 3 also reports the simple cross-sectional probit estimates. The simple probit estimates for decision point 1 (DP1) are obviously unaffected by sample selection. Those for DP2-5, by contrast, are subject to selection biases. Such biases are present if the errors, $c_i + u_{it}$, are correlated across the decision points, or in other words, if $\rho_A \neq 0$ or $\sigma_c^2 \neq 0$. Consequently, the presence of selection bias can be tested using a likelihood ratio test of the null: $\rho_A = \sigma_c^2 = 0$. The computed value of the likelihood ratio statistic is about 142 which greatly exceeds the critical value at the one percent level. In view of this, comparisons of $\text{MLE}(\rho_A = \sigma_c^2 = 0)$ and the probits for DP2-5 to the probit for DP1, SMLE, $\text{SMLE}(\sigma_c^2 = 0)$, and $\text{MLE}(\rho_A = 0)$ give some indication of the nature of the selection biases. One notable difference is that with the exception of $\text{MLE}(\rho_A = 0)$ and the simple probit for DP1, UE and LOS are statistically significant only in the models that assume no selection bias.

A common finding in the literature is that retention rates and job tenure are positively correlated. The retention rates in Table 2 generally increase across the ten decision points and, consequently, such correlation also characterizes the present data set. One explanation offered by many studies is the so called heterogeneity hypothesis: The retention rate of a given cohort increases over time as individuals are sorted out by

permanent differences in preferences and other unobserved factors such as skills. These differences are captured by the individual effect c_i in equation (1). Since none of estimates of σ_c^2 are significant, we find no support for the heterogeneity hypothesis. An alternative explanation for the positive correlation between retention and job tenure is that individuals acquire a greater taste for military life over time. In terms of our model, this suggests positive autocorrelation in the transitory error, u_{it} . Consistent with this explanation, the autocorrelation parameter, ρ_A , is positive and significant at the one percent level across all three sets of estimates.

As we have emphasized, the SML and ML specifications rely heavily on the assumption of normally distributed errors and on the parametric model specified for the autocorrelation. Moreover, the SML/ML estimates can be contaminated by misspecifications at any decision points. The MSCORE estimator, in contrast, is consistent under non-normal disturbances, allows more flexible forms of correlation, and for the unpooled version, effects of misspecification are confined to misspecified decision points. Therefore, the robustness of the SML/ML results can be examined by comparing the SML/ML and MSCORE estimates in Tables 3 and 4.

With the exception of UE and LOS, the signs of MSCORE, and SML/ML estimates agree for the pooled models. Both the MSCORE and SML/ML estimates for ACOL are positive and significant. One notable difference between MSCORE and SML/ML is that the standard errors of the MSCORE estimates are larger. This difference is consistent with asymptotic theory since SML/ML converge at faster rates than MSCORE. In contrast to SML/ML, the MSCORE estimates for LOS, UE, NWHT, ACD, RTC and DEP are insignificant. Table 4 also reports the unpooled MSCORE

estimates for the first five decision points. Unlike the simple probits in Table 3, consistency for the unpooled MSCORE estimates is not adversely affected by the sample selection. In terms of signs, about half of the unpooled MSCORE estimates conflict with the pooled MSCORE estimates. One possible explanation for this discrepancy between the unpooled and pooled estimates is the large difference in sample sizes coupled with the slow convergence rate of MSCORE. Another possibility is misspecification.

To test the MSCORE specification, we used the procedure described in section 2.3 based on 95% confidence bands for $E(Y_{it}|x_{it}\beta_t, z_{it}=1)$. Recall the basic idea: The MSCORE assumptions imply $\text{sgn}(E(Y_{it}|x_{it}, z_{it}=1)) = \text{sgn}(x_{it}\beta_t)$ and, therefore, if the specification is correct, then the upper bound of the band should be positive when the estimate of $x_{it}\beta_t$ is positive, and the lower band should be negative when the estimate of $x_{it}\beta_t$ is negative. We computed 95% confidence bands and carried out the procedure for both the pooled and unpooled MSCORE estimates. The results offer limited support for the MSCORE specification. On the one hand, the confidence band constructed with the pooled estimates satisfies the above condition across all observations and, consequently, does not reject the MSCORE specification. On the other hand, when the unpooled estimates are used, the condition is violated at some decision points for a few observations (10 at decision point 2, 2 at decision point 7, and 1 observation at decision points 3, 5, 8 and 9). Therefore, in contrast to the pooled estimates, the unpooled estimates reject the MSCORE specification at decision points 2, 3, 5, 7, 8 and 9. Although misspecification cannot be ruled out, one possible explanation for the discrepancy is sample size. MSCORE estimators are known to converge very slowly. The pooled MSCORE estimator is based on over 22,000 observations, while the

unpooled estimators are based on considerably less (an average of 1815 observations). Rather than misspecification, the discrepancy in the results might merely reflect a more accurate estimate of $E(Y_{it}|x_{it}\beta_t, z_{it}=1)$ by the pooled MSCORE.

We next compare the predicted probabilities of retention obtained from the SML and MSCORE estimates that were described in section 2.2. Tables 5 and 6 report the SML and MSCORE predictions of

$$Q_{ij} = P(y_{ij}=1 | y_{i,j-1}=1, \dots, y_{i1}=1, x_{ij}, \dots, x_{i1})$$

for $J=1,2$. The SML predictions are obtained by replacing the unknown parameters of the multivariate normal CDF (approximated with the GHK simulator) with SML estimates. Unlike SML, the MSCORE predictions do not impose a parametric functional form on Q_{ij} . The MSCORE predictions combine the MSCORE estimates with kernel regression.

Table 5 reports SML and MSCORE predicted probabilities at the first decision point. The predictions were made for the 12 possible values of (NONWHT, ACAD, ROTC, DEP) with ACOL, LOS and UE set equal to their sample means. The differences between the SML and MSCORE predictions vary from about 1 percent to 31 percent. The average of the absolute value of the differences of the predictions is 9.37 percent. For NONWHT=DEP=1, however, the average absolute difference is 17.36 percent. Table 6 reports similar results for the second decision point, with the SML predictions smaller than the MSCORE predictions in all except one case.

Predictions of retention appear sensitive to whether one adopts the SML or MSCORE specification. To see which specification best describes the data, we compared fits using a method adapted from Horowitz (1993, pp.61-62). For each decision point, the observations were divided into 10 cells obtained by dividing the range of ACOL into 10 intervals of equal length. The within-cell average SML and MSCORE predicted probabilities of staying in the Navy were then compared to the observed fractions of officers choosing to stay. If the specification is correct, then the average predicted probability should be close to the observed fraction in the cell.

Tables 7-9 report the results for decision points 1-3. The bottom half of each table reports the absolute value of the difference between the predicted probability and the observed fraction for each cell. As the tables reveal, the best specification in terms of this goodness-of-fit measure depends on the decision point. For decision point 1 (see Table 7), SML fits the data better than MSCORE in 7 out of 10 cells that represent 97 percent of the observations. On the other hand, MSCORE does better at decision points 2 and 3. For decision point 2, MSCORE fits better than SML in 7 out of 10 cells (97 percent of the observations). For decision point 3, MSCORE fits better than SML in only 4 out of 10 cells, but the 4 cells represent over 78 percent of the observations.

Appendix A

Proof of Proposition 1

Recall that $u_{it} = \varepsilon_{it} + v_{it}$, and $\eta_{it} = c_i + \varepsilon_{it}$. We have

$$E(Y_{it}|x_{it}, z_{it}=1) = \int \left\{ \int D_{it}(\eta_{it}) h_{it}(\eta_{it}) dF(\boldsymbol{\eta}_{it} | \mathbf{x}_{it},) \right\} dF(\mathbf{x}_{it-1} | x_{it}) / P(z_{it}=1) \quad (10)$$

where

$$D_{it}(\eta_{it}) = P(v_{it} > -x_{it}\beta_t - \eta_{it} \mid \mathbf{x}_{it}, \eta_{it}, \boldsymbol{\eta}_{it-1}) - P(v_{it} \leq -x_{it}\beta_t - \eta_{it} \mid \mathbf{x}_{it}, \eta_{it}, \boldsymbol{\eta}_{it-1}),$$

$$h_{it}(\eta_{it}) = P(v_{i1} > x_{i1}\beta_1 - \eta_{i1}, \dots, v_{it-1} > -x_{it-1}\beta_{t-1} - \eta_{it-1} \mid \mathbf{x}_{it}, \eta_{it}, \boldsymbol{\eta}_{it-1})$$

Case 1: $x_{it}\beta_t > 0 \Leftrightarrow E(Y_{it} \mid \mathbf{x}_{it}, z_{it}=1) > 0$.

Under Assumption 4a, the term in the brackets in (10) can be written as:

$$\begin{aligned} & \int \int_{A1} D_{it}(\eta_{it}) h_{it}(\eta_{it}) f_{it}(\eta_{it}) d\eta_{it} d\boldsymbol{\eta}_{it-1} + \int \int_{A2} D_{it}(\eta_{it}) h_{it}(\eta_{it}) f_{it}(\eta_{it}) d\eta_{it} d\boldsymbol{\eta}_{it-1} \\ & + \int \int_{A3} D_{it}(\eta_{it}) h_{it}(\eta_{it}) f_{it}(\eta_{it}) d\eta_{it} d\boldsymbol{\eta}_{it-1} \end{aligned} \quad (11)$$

where $f_{it}(\eta_{it})$ denotes the density of $\boldsymbol{\eta}_{it}$ conditional on \mathbf{x}_{it} , $A1 = \{\eta_{it}: x_{it}\beta_t \leq \eta_{it}\}$, $A2 = \{\eta_{it}: -x_{it}\beta_t < \eta_{it} < x_{it}\beta_t\}$, and $A3 = \{\eta_{it}: -x_{it}\beta_t \geq \eta_{it}\}$. Assumptions 4a and 4b imply that the first two terms in (11) are positive. Therefore, it suffices to show that

$$\int_{A1} D_{it}(\eta_{it}) h_{it}(\eta_{it}) f_{it}(\eta_{it}) d\eta_{it} \geq \int_{A3} D_{it}(\eta_{it}) h_{it}(\eta_{it}) f_{it}(\eta_{it}) d\eta_{it} \quad (12)$$

By Assumption 3 (which implies that $(v_{it-1}, \dots, v_{i1})$ and η_{it} are independent conditionally on \mathbf{x}_{it}):

$$h_{it}(\eta_{it}) = h_{it}(-\eta_{it}) \text{ for all } \eta_{it} \quad (13)$$

Now consider

$$\begin{aligned}
D_{it}(\eta_{it}) + D_{it}(-\eta_{it}) &= P(v_{it} > -x_{it}\beta_t - \eta_{it} \mid \mathbf{x}_{it}, \eta_{it}, \boldsymbol{\eta}_{it-1}) - P(v_{it} \leq -x_{it}\beta_t + \eta_{it} \mid \mathbf{x}_{it}, -\eta_{it}, \boldsymbol{\eta}_{it-1}) \\
&\quad + P(v_{it} > -x_{it}\beta_t + \eta_{it} \mid \mathbf{x}_{it}, -\eta_{it}, \boldsymbol{\eta}_{it-1}) - P(v_{it} > -x_{it}\beta_t - \eta_{it} \mid \mathbf{x}_{it}, \eta_{it}, \boldsymbol{\eta}_{it-1}) \\
&> 0 \text{ for all } \eta_{it} \text{ in } A1
\end{aligned} \tag{14}$$

where the inequality in (14) follows from Assumption 3 (which implies that v_{it} and η_{it} are independent conditionally on \mathbf{x}_{it}) (DETAILS?) and Assumption 4b. By Assumption 4a, $f_{it}(\eta_{it}) = f_{it}(-\eta_{it})$. Therefore, from (13) and (14), we have

$$\int_{A1} D_{it}(\eta_{it}) h_{it}(\eta_{it}) f_{it}(\eta_{it}) d\eta_{it} \geq - \int_{A1} D_{it}(-\eta_{it}) h_{it}(-\eta_{it}) f_{it}(-\eta_{it}) d\eta_{it} \tag{15}$$

Substituting $\xi = -\eta_{it}$ on the right hand side of (15) yields (12).

Proof of Proposition 2

Let $X = \{x_{it} : \text{sgn}(x_{it}b_t) \neq \text{sgn}(x_{it}\beta_t)\}$ and choose any b_t such that $b_t/\|b_t\| \neq \beta_t^*$. By Proposition 1, we have

$$s_t(\beta_t^*) - s_t(b_t^*) = \int_X E(Y_{it} \mid x_{it}) \text{sgn}(x_{it}\beta_t) dF_{it} = \int_X |E(Y_{it} \mid x_{it})| dF_{it} \tag{16}$$

Assumption 5 implies $P(x_{it} \in X) > 0$ (see Manski (1985, p.317) for example), and Assumption 5 and Proposition 1 imply $|E(Y_{it} \mid x_{it})| \neq 0$ for almost every x_{it} . It follows that the right hand side of (16) is positive.

Proof of Proposition 3

By the Mean Value Theorem under Assumption 6c, and Proposition 2, we have:

$$\hat{Q}_{ij}(\hat{\beta}) = \hat{Q}_{ij}(\beta) + (\hat{\beta} - \beta) \partial \hat{Q}_{ij}(\beta) / \partial \beta = \hat{Q}_{ij}(\beta) + o_p(1)$$

where β is a vector lying on the segment connecting $\hat{\beta}$ and β . Let x_{i1}, \dots, x_{ij} be such that $f_z > 0$. Under Assumptions 6a-d, Bierens (1994, Theorem 10.1.1) implies $\text{plim } \hat{Q}_{ij}(\hat{\beta}) = Q_{ij}$.

Proof of Proposition 4

Under Assumption 7(ii), $\partial g / \partial \beta = O_p(1)$. Since the MSCORE estimator converges at rate $n^{-1/3}$ (see Kim and Pollard 1988), $\delta > 1/3$ implies $(n\gamma_n)^{1/2}(\hat{\beta}_t - \beta_t) = o_p(1)$. Therefore, a Taylor series expansion yields

$$(n\gamma_n)^{1/2}(g(x_{it}\hat{\beta}_t) - g(x_{it}\beta_t)) = (n\gamma_n)^{1/2}(\hat{\beta}_t - \beta_t) \partial g(x_{it}\beta_t) / \partial \beta = o_p(1) \quad (17)$$

It follows from (17) that

$$(n\gamma_n)^{1/2}(g(x_{it}\hat{\beta}_t) - G(x_{it}\beta_t)) = (n\gamma_n)^{1/2}(g(x_{it}\beta_t) - G(x_{it}\beta_t)) + o_p(1) \quad (18)$$

The conclusion of part (i) of Proposition 4 follows by applying Bierens (1994, Theorem 10.2.1) to the right hand side of (18); part (ii) follows from Bierens (1994, Theorem 10.2.8).

Notes

1. Charlier, Melenberg and Van Soest (1995) apply the idea of Horowitz (1992) and propose a smoothed version of Manski's (1987) estimator. Other extensions of the maximum score method have been proposed by Mayer and Dorsey (1998) and Mayer (1999).

Table 1

Explanatory Variables

<u>Variable</u>	<u>Definition</u>
ACOL	Annualized cost of leaving Navy.
UE	National average unemployment rate divided by the full sample average (6.5).
NONWHT	Equals 1 if nonwhite, 0 if white.
LOS	Length of service at decision point.

ACAD	Equals 1 if Academy is source of commission, 0 otherwise.
ROTC	Equals 1 if ROTC is source of commission, 0 otherwise.
DEP	Equals 1 if officer has any Dependents, 0 otherwise.

Table 2

Number of Observations and Sample Means by Decision Point

	DP1	DP2	DP3	DP4	DP5	DP6	DP7	DP8	DP9	DP10
N	8027	4806	3455	2324	1483	940	602	355	191	103
Means										
y	0.717	0.878	0.806	0.834	0.849	0.906	0.929	0.947	0.979	0.971
ACOL	1.032	1.106	1.202	1.307	1.418	1.558	1.687	1.822	1.890	2.037
UE	1.045	1.046	1.046	1.036	1.013	0.986	0.974	0.977	0.969	0.951
LOS	4.702	5.983	7.484	8.806	10.08	11.26	12.19	13.01	13.50	14.17
NWHT	0.030	0.029	0.026	0.023	0.016	0.018	0.022	0.014	0.011	0.010
ACD	0.373	0.397	0.434	0.445	0.453	0.451	0.457	0.465	0.461	0.437

RTC	0.275	0.275	0.282	0.271	0.260	0.269	0.269	0.293	0.283	0.262
DEP	0.478	0.574	0.661	0.718	0.763	0.795	0.811	0.811	0.817	0.854

Table 3

Maximum Likelihood Estimates

DP INT ACOL UE LOS NWHT ACD RTC DEP
 1-10 (pooled models)

SMLE: est. $\rho_A=0.482(14.05)$, est. $\sigma_c^2=0.62E-06(0.00)$, *log-likelihood* = -10463.94

-0.320	0.698	0.044	0.011	-0.227	0.148	0.066	0.063
(3.556)	(10.04)	(0.722)	(0.917)	(3.197)	(5.248)	(2.229)	(2.739)

SMLE($\sigma_c^2=0$): est. $\rho_A=0.481(16.55)$, *log-likelihood* = -10463.95

-0.311	0.696	0.037	0.011	-0.278	0.148	0.065	0.062
(3.490)	(10.20)	(0.608)	(1.193)	(3.900)	(5.361)	(2.235)	(2.768)

MLE($\rho_A=0$) est. $\sigma_c^2=0.043(1.44)$ *log-likelihood* = -10516.48

-0.365	0.550	0.163	0.050	-0.228	0.119	0.042	0.064
(4.056)	(9.167)	(3.260)	(5.225)	(3.801)	(3.975)	(1.444)	(3.296)

MLE($\rho_A=\sigma_c^2=0$). *log-likelihood* = -10534.98

-0.386	0.507	0.178	0.611	-0.213	0.104	0.032	0.060
(4.483)	(8.074)	(3.138)	(9.211)	(3.383)	(4.374)	(1.262)	(2.904)

Cross-sectional Probit

1	0.497	0.724	0.228	-0.182	-0.249	0.152	0.147	0.161
	(3.079)	(8.116)	(0.283)	(6.767)	(2.741)	(3.755)	(3.485)	(5.331)
2	-0.196	0.769	0.206	0.002	-0.204	0.602	0.338	0.081
	(0.853)	(5.200)	(1.532)	(0.079)	(1.324)	(9.925)	(5.579)	(0.165)
3	-0.421	-0.400	0.366	0.207	-0.091	-0.079	-0.139	-0.079
	(1.457)	(1.922)	(2.361)	(8.109)	(0.563)	(1.254)	(2.077)	(1.474)
4	-1.289	-0.489	0.781	0.261	-0.187	-0.019	-0.173	-0.109
	(3.409)	(1.628)	(3.705)	(7.255)	(0.874)	(0.229)	(2.028)	(1.516)
5	-0.652	-0.959	0.275	0.290	1.087	-0.083	-0.071	-0.079
	(1.445)	(2.680)	(0.913)	(5.805)	(2.151)	(0.810)	(0.642)	(0.818)

Table 4

Maximum Score Estimates

<u>DP</u>	<u>INT</u>	<u>ACOL</u>	<u>UE</u>	<u>LOS</u>	<u>NWHT</u>	<u>ACD</u>	<u>RTC</u>	<u>DEP</u>
-----------	------------	-------------	-----------	------------	-------------	------------	------------	------------

1-10 (pooled)

-0.028	0.720	-0.084	-0.061	-0.334	0.251	0.018	0.542
(0.20)	(2.57)	(0.31)	(0.86)	(1.19)	(1.14)	(0.08)	(0.73)

(cross-sections)

1	0.075	0.207	0.714	-0.033	-0.630	-0.158	-0.109	-0.084
	(0.41)	(0.48)	(0.99)	(0.47)	(1.26)	(0.56)	(0.49)	(0.14)
2	0.115	-0.152	-0.046	0.026	0.146	0.161	0.233	0.926
	(0.38)	(0.19)	(0.20)	(0.16)	(0.69)	(0.89)	(0.67)	(1.20)
3	0.002	-0.783	0.338	0.135	0.000	-0.001	0.001	0.504
	(0.00)	(2.11)	(1.12)	(1.68)	(0.00)	(0.00)	(0.00)	(1.07)
4	-0.047	-0.768	0.606	0.106	-0.030	0.101	-0.006	-0.137
	(0.36)	(1.67)	(1.08)	(1.33)	(0.11)	(0.50)	(0.03)	(0.36)
5	0.186	0.479	-0.180	-0.050	0.192	0.359	0.421	0.598
	(0.34)	(0.73)	(0.41)	(0.38)	(0.66)	(0.65)	(1.11)	(0.96)

Table 5

Values of the explanatory variables				Predicted $\Pr(y_1=1 x_1)$	
<u>NONWHT</u>	<u>ACAD</u>	<u>ROTC</u>	<u>DEP</u>	<u>ML</u>	<u>MSCORE</u>
0	0	0	0	0.7361	0.7115
1	0	0	0	0.6382	0.7593
0	1	0	0	0.7822	0.7346
0	0	1	0	0.7571	0.7201

0	0	0	1	0.7562	0.6952
1	1	0	0	0.6921	0.7068
1	0	1	0	0.6625	0.7352
1	0	0	1	0.6615	0.8452
0	1	0	1	0.8002	0.6849
0	0	1	1	0.7764	0.6664
1	1	0	1	0.7139	0.7422
1	0	1	1	0.6852	0.9940

Table 6

Values of the explanatory variables Predicted $\Pr(y_2=1 | y_1=1, x_1, x_2)$

<u>NONWHT</u>	<u>ACAD</u>	<u>ROTC</u>	<u>DEP</u>	<u>ML</u>	<u>MSCORE</u>
0	0	0	0	0.8148	0.8461
1	0	0	0	0.7515	0.7642
0	1	0	0	0.8447	0.9232
0	0	1	0	0.8283	0.8679
0	0	0	1	0.8278	0.7640
1	1	0	0	0.7863	0.9735

1	0	1	0	0.7671	0.8518
1	0	0	1	0.7665	0.6474
0	1	0	1	0.8565	0.9217
0	0	1	1	0.8409	0.8637
1	1	0	1	0.8004	0.9995
1	0	1	1	0.7818	0.9999

Table 7

Average Observed and Predicted Probabilities: Decision Point 1

Cell	Avg.ACOL	Count	Observed	ML	MSCORE
1	0.462	12	0.7500	0.5811	0.7116
2	0.570	182	0.5329	0.6151	0.6834
3	0.702	596	0.6611	0.6445	0.7049
4	0.832	968	0.6818	0.6823	0.7143
5	0.930	869	0.6869	0.6957	0.7181
6	1.102	3570	0.7342	0.7385	0.7151
7	1.181	1643	0.7480	0.7578	0.7320
8	1.359	67	0.7341	0.7123	0.7167
9	1.432	117	0.7948	0.7329	0.7502
10	1.536	3	1.0000	0.7926	0.6891

Cell	ABS(Observed-ML)	ABS(Observed-MSCORE)
1	0.1011	0.0384
2	0.0822	0.1505
3	0.0166	0.0438
4	0.0005	0.0325
5	0.0085	0.0312
6	0.0043	0.0191
7	0.0098	0.0160
8	0.0218	0.0174
9	0.0619	0.0446
10	0.2074	0.3109

Table 8

Average Observed and Predicted Probabilities: Decision Point 2

Cell	Avg.ACOL	Count	Observed	ML	MSCORE
1	0.564	58	0.6897	0.7292	0.7788
2	0.710	265	0.8075	0.7628	0.8084
3	0.893	748	0.8717	0.7978	0.8684
4	1.085	1125	0.8667	0.8208	0.8886
5	1.199	2320	0.8953	0.8406	0.8879
6	1.401	200	0.8950	0.8593	0.8753
7	1.556	77	0.8831	0.8736	0.8943
8	1.751	8	1.0000	0.9315	0.9391
9	1.923	3	1.0000	0.9497	0.9546
10	1.536	3	1.0000	0.9735	0.9101

Cell	ABS(Observed-ML)	ABS(Observed-MSCORE)
1	0.0395	0.0891
2	0.0447	0.0009
3	0.0739	0.0033
4	0.0459	0.0219
5	0.0547	0.0074
6	0.0357	0.0197
7	0.0095	0.0112
8	0.0685	0.0609
9	0.0503	0.0454
10	0.0265	0.0899

Table 9

Average Observed and Predicted Probabilities: Decision Point 3

Cell	Avg.ACOL	Count	Observed	ML	MSCORE
1	0.668	20	0.8000	0.7613	0.8547
2	0.842	161	0.8385	0.8090	0.8426
3	0.972	459	0.8126	0.8244	0.8251
4	1.169	1254	0.7815	0.8468	0.7951
5	1.281	1222	0.7995	0.8683	0.7839
6	1.474	234	0.8932	0.8852	0.8356
7	1.636	71	0.8732	0.9075	0.8526
8	1.854	21	1.0000	0.9426	0.9227
9	1.988	8	1.0000	0.9476	0.9390
10	2.204	5	1.0000	0.9667	0.8767

Cell	ABS(Observed-ML)	ABS(Observed-MSCORE)
1	0.0387	0.0541
2	0.0295	0.0041
3	0.0118	0.0125
4	0.0653	0.0136
5	0.0688	0.0156
6	0.0080	0.0576
7	0.0343	0.0206
8	0.0574	0.0773
9	0.0524	0.0610
10	0.0333	0.1233

References

- Bierens, H.J., (1994), *Topics in Advanced Econometrics*, Cambridge, MA: Cambridge University Press.
- Black, M., Moffitt, R. and Warner, J.T. (1990), "The Dynamics of Job Separation: The Case of Federal Employees," *Journal of Applied Econometrics*, 5, 245-262.
- Butler, J.S., and Moffitt, R. (1982), "A Computationally Efficient Quadrature Procedure for the One-Factor Multinomial Probit Model," *Econometrica* , 50, 761-764.

- Charlier, E., Melenberg, B. and Van Soest, A.H.O. (1995), "A Smoothed Maximum Score Estimator for Binary Choice Panel Data Model with an Application to Labour Force Participation," *Statistica Neerlandica*, 49-3, 324-342.
- Daula, T. and Moffitt, R. (1995), "Estimating Dynamic Models of Quit Behavior: The Case of Military Reenlistment," *Journal of Labor Economics*, 13-3, 499-523.
- Dorsey, R.E. and Mayer, W.J. (1995), "Genetic Algorithms for Estimation Problems with Multiple Optima, Nondifferentiability and Irregular Features," *Journal of Business and Economic Statistics*, 13, 53-66.
- Goffe, W.L., Ferrier, and Rogers, J. (1994), "Global Optimization of Statistical Functions with Simulated Annealing," *Journal of Econometrics*, 60, 65-99.
- Gotz, G. and McCall, J.A. (1984), *A dynamic Retention Model for Air Force Officers: Theory and Estimates*. Santa Monica, CA: RAND.
- Hajivassiliou, V.A. (1993), "Simulation Estimation Methods for Limited Dependent Variable Models," in *Handbook of Statistics Vol. 11*, G.S. Maddala, C.R. Rao, and H.D. Vinod eds. New York: North-Holland.
- Hajivassiliou, V.A. and Ruud, P.A. (1994), "Classical Estimation Methods for LDV Using Simulation," in *Handbook of Econometrics Vol. IV*, R.F. Engle and McFadden eds. Amsterdam: Elsevier Science B.V.
- Horowitz, J.L., (1992), "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60, 505-531.
- Horowitz, J.L., (1993), "Semiparametric Estimation of a Work-Trip Mode Choice Model," *Journal of Econometrics*, 58, 49-70.
- Horowitz, J.L., (1998), *Semiparametric Methods*. New York: Springer.

- Horowitz, J.L., (1999), "The Bootstrap," in *Handbook of Econometrics Vol. 5*,
Forthcoming.
- Keane, M.P., (1993), "Simulation Estimation for Panel Data Models with Limited
Dependent Variables," in *Handbook of Statistics Vol. 11*, G.S. Maddala, C.R. Rao,
and H.D. Vinod eds. New York: North-Holland.
- Manski, C.F., (1985), "Semiparametric Analysis of Discrete Response: Asymptotic
Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27, 313-333.
- Manski, C.F., (1987), "Semiparametric Analysis of Random Effects Linear Models
from Binary Panel Data," *Econometrica*, 55-2, 357-362.
- Manski, C.F. and Thompson, T.S., (1986), "Operational Characteristics of Maximum
Score Estimation," *Journal of Econometrics*, 32, 65-108.
- Mayer, W.J. and Dorsey, R.E., (1998), "Maximum Score Estimation of
Disequilibrium Models and the Role of Anticipatory Price-Setting," *Journal of
Econometrics*, 87, 1-24.
- Mayer, W.J., (1999), "An Extension of the Maximum Score Estimator for
Disequilibrium Models," *Economics Letters*, 64, 143-149.
- Nakada, M.K., Mackin, P.C., and Mackie, C.D., (1998), "Nuclear Officer Retention:
MSR and Beyond," report for Navy Personnel Research and Development Center.
- Ostermark, R. (1999), "Solving Irregular Econometric and Mathematical
Optimization Problems with Genetic Hybrid Algorithm," *Computational Economics*
13, 103-115.

Warner, J.T. and Goldberg, M.S., (1984), "The Influence of Non-pecuniary Factors on Labor Supply: The Case of Navy Enlisted Personnel," *Review of Economics and Statistics*, 26-35.