WI

SSRI Working Series 85-11

Social Systems Research Institute
University of Wisconsin - Madison

MAXIMUM LIKELIHOOD ESTIMATION OF DISCRETE CONTROL PROCESSES

John Rust

8511
(Revision of 8407)

November 1985

The theory of statistical inference for stochastic processes, which began with the monograph by Ulf Grenander (1950), is now a well-developed field (cf. Basawa and Prakasa Rao, 1980). The theory of stochastic control, which began with the work of Richard Bellman (1957), is now also well-developed (cf. Gihman and Skorohod, 1979). For reasons which are unclear, there has been little interface between these fields, and only recently has work begun on a theory of statistical inference for controlled stochastic processes, i.e. stochastic processes which arise as solutions to well-defined optimization problems. In certain fields such as economics, observed time series data can be interpreted as realizations of controlled stochastic processes of the general form $\{i_t, x_t\}$ where i_t is a variable representing the action taken by an agent at time t, and x_t is the observed state of the agent at time t. The goal of statistical inference for these data is not simply to infer the form of the stochastic process governing the historical evolution of $\{i_t, x_t\}$, but to go deeper and attempt to infer the ultimate determinant of this stochastic process, namely, the mathematical objective function of the agent. This type of structural statistical inference is required in order to test the hypothesis that the observed data is in fact generated by an agent solving the specified stochastic control problem. If indeed such an hypothesis is supported by the data, then structural inference is also required in order to perform policy experiments which forecast how the stochastic process governing $\{i_t, x_t\}$ changes when certain parameters of the agent's objective function are changed. While the existing literature on estimation on stochastic processes may permit us to consistently estimate the form of the historical stochastic process governing $\{i_t, x_t\}$, it is of limited use for forecasting the effects of policy changes which alter the agent's objective function. An

alteration in the agent's objective function induces a corresponding shift in the solution to the stochastic control problem, implying that the stochastic process governing $\{i_t, x_t\}$ after the policy change is generally not equal to the historical stochastic process governing $\{i_t, x_t\}$ before the policy change. Marschak (1953), and Lucas (1976) have shown that the existing non-structural or reduced-form statistical models can produce dramatic inference and forecasting errors under commonly analyzed policy experiments.

To make the above somewhat abstract discussion more concrete, consider the following example which we analyze and estimate in Rust (1986b). Our data $\{i_{t\ell}, x_{t\ell}\}$, $t=1,\ldots,T_\ell$, $\ell=1,\ldots,L$, consists of monthly observations on the mileage $x_{t\ell}$ of each bus ℓ in the fleet of the Madison Metropolitan Bus Company. The agent is Harold Zurcher, maintenance manager at Madison Metro, who decides each month whether or not to replace the engine on bus ℓ with a rebuilt engine: $i_{t\ell} = 1$ versus $i_{t\ell} = 0$. Our hypothesis is that Harold Zurcher follows an engine replacement strategy which minimizes the expected discounted costs of operating each bus over its lifetime. The statistical problem is to use the data $\{i_{t\ell}, x_{t\ell}\}$ to infer the unknown parameter vector $(\beta, \theta_1, \theta_2, \theta_3)$ where β is Harold's intertemporal discount factor, θ_2 is the cost of a replacement engine, θ_3 is a vector of parameters describing the stochastic evolution of the state variables $\{x_{t\ell}\}$, and θ_1 is a vector of parameters specifying the functional form of the operating cost function, $c(x_{t\ell}, \theta_1)$, which equals the monthly operating and maintenance cost of each bus ℓ as a function of the accumulated mileage since last replacement $x_{t\ell}$. Structural estimation of this model is required because 1) we want to test the hypothesis that Harold Zurcher's behavior is in fact consistent with this simple model of optimal replacement, and 2) we want to forecast the effect of

certain policy changes on the timing of investment in replacement engines. For example, we might want to study the impact on the frequency of engine replacement of an increase in the cost of a replacement engine θ_2, or a change in bus utilization intensity (represented by an appropriate change in θ_3). Since engine replacement costs and utilization rates have not changed much in the past, existing reduced form methods of inference which attempt to directly measure engine costs or utilization rates and include them as explanatory variables in the model are likely to yield imprecise and unreliable forecasts.

In this paper we define a class of controlled stochastic processes, discrete control processes, which are explicitly derived as solutions to infinite horizon markovian decision problems. Our definition provides a general method for directly incorporating unobservable state variables (state variables which are observed by the agent but not by the statistician) into the solution of the stochastic control problem so as to produce an internally consistent statistical model. Most importantly, we present a nested fixed point algorithm which computes maximum likelihood estimates of the structural parameters of discrete control processes. We prove that the maximum likelihood estimator is consistent and asymptotically normally distributed, and derive explicit formulae for the gradient of the likelihood function and the asymptotic covariance matrix. To our knowledge, this nested fixed point algorithm allows us to formulate and estimate structural parameters of a wide class of controlled stochastic processes for which there were no previously known estimation methods.

To put our results in perspective, a brief summary the existing literature on estimation of controlled stochastic processes is in order. The literature dichotomizes according to whether the time variable t is discrete or con-

tinuous, and more importantly, according to whether the control variable i_t is discrete or continuous. If i_t can take any value in some convex subset of a Euclidean space we call $\{i_t,x_t\}$ a continuous control process, otherwise if i_t is restricted to lie in a countable set we call $\{i_t,x_t\}$ a discrete control process (the intermediate case where certain components of i_t are discrete and others are continuous has not been analyzed). In an important contribution, Hansen and Singleton (1982) have developed a practical, internally consistent technique for estimating structural parameters of a fairly general class of discrete-time, continuous control processes. Their method uses the generalized method of moments technique (Hansen (1982), Manski (1982)) to estimate first order necessary conditions of the agent's stochastic control problem (stochastic Euler equations), avoiding the need for an explicit solution of the optimal decision rule and analytic formulae for the probability distribution of the controlled stochastic process governing $\{i_t,x_t\}$. The Hansen-Singleton method depends critically on the assumption that the agent's control variable i_t is continuous so that first order necessary conditions can be derived by the usual variational methods. In many circumstances, such as in our bus engine replacement problem, the agent's control variable will be discrete, ruling out the use of the Hansen-Singleton method. A further limitation is that the method relies on the assumption that all variables entering the agent's objective function are observed by both the agent and the statistician; "this latter qualification does rule out some models in which the implied Euler equations involves unobservable forcing variables" (Hansen-Singleton (1982), p. 1271). Garber and King (1985) show that this "exclusion restriction" is necessary for parameter identification. The presence of unobservables will ordinarily lead to inconsistent parameter estima-

tes. Replacing unobservables by instruments will not work because the Euler equations are generally nonlinear functions of the state variables.

The difficulties created by allowing for discrete control variables i_t and unobservable state variables ϵ_t are twofold. First, discrete stochastic control problems rarely possess closed-form solutions or convenient first order necessary conditions amenable to estimation. Instead the solution is almost always given only implicitly from the solution to the fundamental equation of dynamic programming, Bellman's equation. The second problem is that the optimal control i_t will generally be a function of all the state variables in the model. This implies that if the agent's state variables consist of the vector (x_t,ϵ_t) but the statistician observes only x_t, then we obtain a statistical model of the general form

$$(1) \qquad i_t = f(x_t,\epsilon_t,\theta)$$

which is generally a highly nonlinear, non-separable function of the observed explanatory variables x_t and the unobserved "error terms" ϵ_t. These considerations lead us to search for a statistical model in which 1) the dependent variable i_t is finite valued, 2) the data $\{i_t,x_t\}$ are serially dependent, 3) the functional form f of the statistical model does not have an a priori known closed-form solution, 4) the unobservables appear in a possibly nonlinear and non-separable fashion, and 5) the stochastic process governing the unobservables $\{\epsilon_t\}$ is possibly serially dependent.

Despite this formidable list of requirements, pioneering efforts by Heckman and Coleman (1983), Miller (1984), Pakes (1985), and Wolpin (1984) have yielded successful estimation methods for several special classes of models. Heckman and Coleman produced a method for estimating structural para-

meters of a class of continuous-time, infinite-horizon markovian models of employment transitions. Miller developed an infinite horizon multi-armed bandit model of occupation choice, solved by numerically computing Gitten's (1978) "dynamic allocation indices". Pakes developed a finite horizon optimal stopping model of patent renewal behavior of European firms and evaluated the likelihood function for non-renewal of patents by computing the distribution of times at which realizations of a simulated stochastic process of patent returns crossed a parametrically determined optimal stopping barrier. Wolpin developed a finite horizon model of Malaysian womens' decisions about the number and timing of births over their fertile period. The method involved numerically solving a woman's dynamic programming problem by backward induction starting with the last year of the woman's fertile period. Although each of these papers have produced ingenious methods for interfacing stochastic control theory and statistical estimation theory, each method is application specific: none of the methods can claim to offer a general estimation method for a wide class of discrete control processes. In particular, many specific structural estimation problems (such as the bus engine replacement problem) fall outside the domain of any existing method.

In section 2 we define a general class of discrete control processes and prove that their solution is obtained by computing the fixed point EV_θ to a particular differentiable contraction mapping. We derive a general formula for the conditional probability $P(i_t|x_t,\theta)$ of choosing alternative i_t given the observed state variable x_t, which forms the basis of the sample likelihood function. For specific distributional assumptions about the unobserved state variables ϵ_t, we show that $P(i_t|x_t,\theta)$ takes very simple forms. For example, if the conditional distribution of ϵ_t given x_t is assumed to be multivariate extreme value, then $P(i_t|x_t,\theta)$ has the well-known conditional logit form

$$(2) \qquad P(i_t|x_t,\theta) = \frac{\exp\{u(x_t,i_t,\theta_1)+\beta EV_\theta(x_t,i_t)\}}{\sum_{j\in C(x_t)} \exp\{u(x_t,j,\theta_1)+\beta EV_\theta(x_t,j)\}}$$

where $u(x_t,j,\theta_1)$ is the current period utility of being in observed state x_t and choosing alternative j, and $\beta EV_\theta(x_t,j)$ is the discounted expected future utility from time t+1 onward. In section 3 we derive the likelihood function for a panel of data $\{i_{t\ell},x_{t\ell}\}$ $t=1,\ldots,T_\ell$, $\ell=1,\ldots,L$ and present a nested fixed point algorithm which computes maximum likelihood estimates of the structural parameters of discrete control processes. The idea behind this algorithm is quite simple. Except for the presence of the unknown value function EV_θ we have a standard nonlinear maximum likelihood problem. The nested fixed point algorithm, therefore, consists of an "inner" algorithm which computes the fixed point EV_θ corresponding to the current value of θ, and an "outer" hill-climbing algorithm which searches over values of θ in order to maximize the likelihood function. In section 4 we show that EV_θ is in fact a fixed point to a differentiable contraction mapping. This differentiability property enables us to use the very efficient Newton-Kantorovich algorithm to compute EV_θ, and allows us to derive closed-form solutions for its θ-derivatives. This, in turn, allows us in section 5 to derive closed-form solutions for the gradients of the likelihood function, and prove the consistency, asymptotic normality and asymptotic efficiency of the maximum likelihood estimator. Thus, except for the necessity of computing a contraction mapping fixed point at each evaluation of the likelihood function, our proposed maximum likelihood estimator can be computed by standard optimization algorithms such as BHHH (Berndt, Hall, Hall and Hausman (1974)) or Newton's method.

2. Solution of the Stochastic Control Problem: Infinite Horizon Case

To simplify notation, we will state our results for an infinite horizon, stationary Markovian decision problem. By allowing our state variables to be elements of a general complete, separable metric space, however, our arguments implicitly handle the finite horizon nonstationary case as well. In addition, via a well-known transformation of state variables, (see for example, Bertsekas (1976)), our framework implicitly handles Bayesian control of Markov processes where agents have imperfect information about some of the state variables and learn optimally by sequentially updating their beliefs via Bayes rule. We need the following notation:

$C(x_t)$	Choice set; a finite set of allowable values of the control variable i_t when state variable is x_t.
$\epsilon_t = \{\epsilon_t(i) \mid i \epsilon C(x_t)\}$	A $\#C(x_t)$ dimensional vector of state variables observed by agent but not by the statistician. $\epsilon_t(i)$ is interpreted as an unobserved component of utility of alternative i in time period t.
$x_t = \{x_t(1), \ldots, x_t(M)\}$	M-dimensional vector of state variables observed by the agent and statistician.
$u(x_t, i, \theta_1) + \epsilon_t(i)$	Realized single period reward or utility when alternative i is selected and when the state variable is (x_t, ϵ_t). θ_1 is a vector of unknown parameters to be estimated.
$p(x_{t+1}, \epsilon_{t+1} \mid x_t, \epsilon_t, i, \theta_2, \theta_3)$	Markov transition density for state variable (x_t, ϵ_t) when alternative i is selected and when the variable is (x_t, ϵ_t). θ_1 is a vector of unknown parameters to be estimated.
$\theta \equiv (\beta, \theta_1, \theta_2, \theta_3)$	The complete $(1+K_1+K_2+K_3)$ vector of parameters to be estimated, where $\beta \epsilon (0,1)$ is the agent's intertemporal discount factor.

Given the stochastic evolution of the state variables (x_t, ϵ_t), the agent must choose a sequence of decision-rules or <u>controls</u> $f_t(x_t, \epsilon_t, \theta)$ to maximize expected discounted utility over an infinite horizon. Define the <u>value function</u> V_θ by

$$(3) \qquad V_\theta(x_t, \epsilon_t) = \sup_{\pi} E\{ \sum_{j=t}^{\infty} \beta^{(j-t)} [u(x_j, f_j, \theta_1) + \epsilon_j(f_j)] | x_t, \epsilon_t, \theta_2, \theta_3 \}$$

where $\pi = \{f_t, f_{t+1}, f_{t+2}, \ldots\}$, and $f(x_t, \epsilon_t) \epsilon C(x_t)$ for all t, x_t and ϵ_t, and where the expectation is taken with respect to the <u>controlled</u> stochastic process $\{x_t, \epsilon_t\}$ whose probability distribution is defined on cylinder sets from π and the transition probability for $\{x_t, \epsilon_t\}$ by

$$(4) \quad dp\{x_t, \epsilon_t, x_{t+1}, \epsilon_{t+1}, \ldots x_{t+n}, \epsilon_{t+n}\} = \left[\prod_{i=t}^{t+n-1} p(x_{i+1}, \epsilon_{i+1} | x_i, \epsilon_i, f_i(x_i, \epsilon_i), \theta_2, \theta_3) \right] \Omega(x_t, \epsilon_t, \theta)$$

where $\Omega(x_t, \epsilon_t, \theta)$ is an initial probability density for (x_t, ϵ_t). We have implicitly assumed that the controls f_t are 1) nonstochastic, and 2) depend only on the current state of the process (x_t, ϵ_t). By the markovian structure of the decision problem, these assumptions involve no loss of generality, as we show below.

Problem (3) is known as an infinite-horizon discounted markovian decision problem. It differs from the standard formulation presented by Denardo (1967), Blackwell (1968), and Bertsekas and Shreve (1978), due to the fact the utility function $[u(x_t, i_t, \theta_1) + \epsilon_t(i_t)]$ will generally not be uniformly bounded in (x_t, ϵ_t) under usual statistical assumptions about the distribution of unobservables. The unboundedness of the utility function could potentially lead to unbounded values of the objective function (3) and nonexistence of an optimal policy π.

We need to find assumptions sufficient to guarantee the existence of a stationary optimal policy $\pi = (f,f,f,\ldots)$ where f is given by

(5) $\quad i_t = f(x_t, \epsilon_t, \theta)$

and is interpreted as a decision rule specifying the agent's optimal decision i_t when the state variables are given by (x_t, ϵ_t). First we present some more notation. We will require that the number of elements in each choice set C(x) to be uniformly bounded in x, so that we have $C(x) \subseteq C = \{1,\ldots,N\}$ where N is an upper bound on the number of elements in each C(x). Define the state space of the controlled process by $S = \{(x,\epsilon) | x\epsilon\Delta, \epsilon\epsilon R^N\}$ where Δ is a Borel subset of a complete separable metric space. Typically Δ will be a Borel subset of R^M, as indicated by the notation $x_t = \{x_t(1),\ldots,x_t(M)\}$. Strictly speaking $\epsilon_t \epsilon R^{\#C(x_t)}$, however since the dimensionality of ϵ_t changes with the number of elements in $C(x_t)$, we imbed the state space for ϵ_t in the common space R^N and zero out the unnecessary components in order to obtain an ϵ_t vector with at most $\#C(x_t)$ positive elements. Let λ be Lesbesque measure on R^N and let μ be a Borel measure on Δ, the latter which need not be nonatomic. We define a measure ν on S in the usual way by $\nu = \mu \times \lambda$. Our first assumption is needed in order to guarantee the existence of probability densities which will permit us to perform maximum likelihood estimation.

(A1) For each $i\epsilon C(x_t)$ and $(x_t,\epsilon_t)\epsilon S$, the conditional probability distribution of $(x_{t+1}, \epsilon_{t+1})$ given (x_t,ϵ_t,i) is regular and has a Radon-Nikodym density $p(x_{t+1},\epsilon_{t+1}|x_t,\epsilon_t,i,\theta_2,\theta_3)$ with respect to the measure ν on S.

The remaining assumptions guarantee the existence of a stationary optimal policy π to the stochastic control problem (3). Throughout the remainder of

the paper we will use the shorthand notation $Ef(x,\epsilon,i)$ to denote the conditional expectation of the function $f:S\to R$ with respect to the conditional probability density $p(x_{t+1},\epsilon_{i+1}|x_t,\epsilon_t,i,\theta_2,\theta_3)$: $Ef(x,\epsilon,i) \equiv \int_S f(y,\eta)p(y,\eta|x,\epsilon,i,\theta_2,\theta_3)\mu(dy)x\lambda(d\eta)$.

(A2) $0 < \beta < 1$

(A3) $C(x) \subseteq C = \{1,\ldots,N\}$ for all $x\epsilon\Delta$.

(A4) For each $i\epsilon C(x)$, $u(x,i,\theta_1)$ is upper semi-continuous at x for each $x\epsilon\Delta$ and we have

$$R(x,\epsilon) \equiv \sum_{j=0}^{\infty} \beta^j r_j(x,\epsilon) < +\infty \qquad (x,\epsilon)\epsilon S$$

where $r_0(x,\epsilon) = \max_{i\epsilon C(x)} |u(x,i,\theta_1) + \epsilon(i)|$

$$r_{j+1}(x,\epsilon) = \max_{i\epsilon C(x)} Er_j(x,\epsilon,i)$$

(A5) For each $i\epsilon C(x)$, $Eh(x,\epsilon,i)$ is continuous at each point $(x,\epsilon)\epsilon S$ for all Borel measurable functions $h:S\to R$ satisfying

$|h(x,\epsilon)| \leq |R(x,\epsilon)|+1 \qquad (x,\epsilon)\epsilon S$

Theorem 1 Under assumptions (A1),...,(A5) a stationary optimal policy $\pi^* = (f,f,f,\ldots)$ exists for some Borel measurable function $f:S\to C$. The decision rule f is nonstochastic, markovian and is determined from Bellman's equation

$$(6) \qquad V_\theta(x,\epsilon) = \max_{i\epsilon C(x)} [u(x,i,\theta_1) + \epsilon(i) + \beta EV_\theta(x,\epsilon,i)]$$

by the identity

$$(7) \qquad f(x,\epsilon,\theta) \equiv \underset{i\epsilon C(x)}{\text{argmax}} \left[u(x,i,\theta_1) + \epsilon(i) + \beta EV_\theta(x,\epsilon,i)\right]$$

Theorem 1 is a specialization of Theorem 2.1 of Bhattacharya and Majumdar (1985) who extend the basic results of stochastic dynamic programming to allow for unbounded rewards. Note that although the value function $V_\theta(x,\epsilon)$ is finite for each $(x,\epsilon)\epsilon S$, it is generally not uniformly bounded in (x,ϵ). As a result, we cannot apply the standard results of Blackwell (1968) and Denardo (1967) who use the uniform boundedness property to show that V_θ is a fixed point to a contraction mapping on the Banach space B of all uniformly bounded upper semicontinuous functions from S to R. This is unfortunate since the main numerical method for solving stochastic control problems consists of computing the value function by solving the associated fixed point problem. Lippman (1975) has provided alternative conditions under which Theorem 1 holds and V_θ is the unique fixed point to a contraction mapping on the Banach space B_w of all bounded, upper semicontinuous functions under a weighted supremum norm defined by

$$(8) \qquad ||h||_w = \sup_{(x,\epsilon)\epsilon S} |h(x,\epsilon)|/w(x,\epsilon)$$

where w is a specified weight function which satisfies $w \geq 1$. However even adopting Lippman's approach, there are two difficulties which hamper direct statistical implementation of the model $i_t = f(x_t,\epsilon_t,\theta)$ given by the solution to (6) and (7). First, standard distributional assumptions for unobservables imply that ϵ_t will be continuously distributed on R^N with unbounded support. However, this raises serious dimensionality problems since the optimal stationary policy f will ordinarily be computed by solving for the fixed point

V_θ from Bellman's equation. Even taking a rough grid approximation to the true continuous distribution of ϵ_t, the dimensionality of the resulting finite approximation will still be too large to be computationally tractable. Secondly, since ϵ_t appears nonlinearly in the unknown function EV_θ, we face the additional problem of integrating out over the ϵ_t distribution to obtain conditional choice probabilities. Since EV_θ is an unknown function, this will require the dual task of integrating V_θ with respect to a finite grid approximation of the density $p(x_{t+1},\epsilon_{t+1}|x_t, \epsilon_t, i, \theta_2, \theta_3)$ to obtain EV_θ, and then numerically integrating Bellman's equation (6) to obtain the conditional choice probability $P(i_t|x_t,\theta)$ needed to form the likelihood function. The following assumption enables us to circumvent these problems.

(A6) For each $(x_t, \epsilon_t)\epsilon S$, the markov transition density factors as

$$(9) \quad p(x_{t+1},\epsilon_{t+1}|x_t,\epsilon_t,i,\theta_2,\theta_3) = q(\epsilon_{t+1}|x_{t+1},\theta_2)p(x_{t+1}|x_t,i,\theta_3), \; i\epsilon C(x_t)$$

Assumption (A6) involves two restrictions. First, x_{t+1} is a sufficient statistic for ϵ_{t+1}, which implies that any statistical dependence between ϵ_t and ϵ_{t+1} is transmitted entirely through the vector x_{t+1}. Second, the probability density for x_{t+1} depends only on x_t and not ϵ_t. Intuitively, the $\{\epsilon_t\}$ process can be regarded as noise superimposed on the underlying $\{x_t\}$ process, since in each period t, ϵ_t is drawn according to the density $q(\epsilon_t|x_t,\theta_2)$ given the realized value of x_t. The pattern of dependence implied by assumption (A1) is displayed graphically in diagram 1 below.

Actually (A6) is a stronger assumption than is necessary. All our results will go through under the weaker assumption

(A6') For each $(x_t, \epsilon_t, i_t) \in S \times C$ the markov transition density $p(x_{t+1}, \epsilon_{t+1}|x_t, \epsilon_t, i_t, \theta_2, \theta_3)$ is independent of ϵ_t.

This implies that p admits the more general factorization

$$(10) \quad p(x_{t+1},\epsilon_{t+1}|x_t,\epsilon_t,i_t,\theta_2,\theta_3) = q(\epsilon_{t+1}|x_{t+1},x_t,i_t,\theta_2)p(x_{t+1}|x_t,i_t,\theta_3) \quad i_t \in C(x_t)$$

In this case (x_{t+1},x_t,i_t) is a sufficient statistic for ϵ_{t+1}, allowing more complicated forms of dependence than is allowed under (A6). We adopt (A6) throughout the rest of the paper only to simplify notation.

The payoff to adopting (A6) or (A6') is twofold. First, we will show that (A6) implies that EV_θ is not a function of ϵ_t, so that required choice probabilities will not require integration over the unknown function EV_θ. Second, we will show that EV_θ is a fixed point of a separate contraction mapping on the space $\Gamma = \{(x,i)|x \in \Delta, i \in C(x)\}$, eliminating the need to compute the fixed point V_θ on the much larger space S and avoiding the numerical integration required to obtain EV_θ from V_θ.

Before proving these claims, we need yet some more notation and a mild regularity assumption. Let g be a measurable, real-valued function from Γ to R. Define the norm of g, $||g||_\infty$ in the usual manner by $||g||_\infty = \sup_{(x,i)\in\Gamma}|g(x,i)|$. It follows that the set B of all measurable, real-valued and $||\cdot||_\infty$-bounded functions on Γ is a Banach space. Given a vector $r(x) = \{r(x,i)|i \in C(x)\} \in R^N$, define the function $G(r(x)|x,\theta_2)$ by

$$(11) \quad G(r(x)|x,\theta_2) \equiv \int_\epsilon \{\max_{i\in C(x)} [r(x,i) + \epsilon(i)]\}q(\epsilon|x,\theta_2)\lambda(d\epsilon)$$

$G(r(x)|x,\theta_2)$ is simply the conditional expectation of the maximum of $[r(x,i) + \epsilon(i)]$, $i \in C(x)$. McFadden (1981) calls G a social surplus function. The social surplus function G has an important property, apparently first noted by

Williams (1977) and Daly and Zachary (1979), which is a key to our subsequent results.

Theorem 2 Suppose the density $q(\epsilon|x,\theta_2)$ has finite first moments. Then for any vector $r(x) \equiv \{r(x,i)|i\epsilon C(x)\}$, the social surplus function $G(r(x)|x,\theta_2)$ defined in (11) has the following properties:

A. G is a positively linear homogeneous, convex function of r(x).

B. G has the additivity property $G(r(x)+\alpha|x,\theta_2) = \alpha + G(r(x)|x,\theta_2)$ for any constant α, where $r(x)+\alpha \equiv \{r(x,i)+\alpha|i\epsilon C(x)\}$.

C. Let G_i denote the partial derivative of G with respect to r(x,i), and let $P(i|x,\theta_2)$ denote the conditional probability that $r(x,i)+\epsilon(i)$ is the largest: $P(i|x,\theta_2) \equiv \int_\epsilon I\{r(x,i)+\epsilon(i) = \max_{j\epsilon C(x)} [r(x,j)+\epsilon(j)]\}q(d\epsilon|x,\theta_2)$. Then we have

(12) $P(i|x,\theta_2) = G_i(r(x)|x,\theta_2)$ and P is a differentiable function of r(x).

The proof of Theorem 2 is given in McFadden (1981). The key result is equation (12) which has the following intuitive explanation. The social surplus function can be regarded as the expected value of the maximum utility of choosing alternatives $i\epsilon C(x)$ for a population of consumers indexed by ϵ. If we increase the utility r(x,i) of the i^{th} alternative by 1 unit, how much does social utility go up? The amount is simply 1 times the fraction of the population choosing alternative i, or $P(i|x,\theta_2)$. Before stating our main result we need to make one final regularity assumption.

(A7) $u\epsilon B$, and for each $r\epsilon B$, $EG\epsilon B$ where $G(r(x)|x,\theta_2)$ is given by (12) and EG is defined by

(13) $EG(x,i) \equiv \int_y G(r(y)|y,\theta_2)p(dy|x,i,\theta_3)$

(Notice that in order to simplify notation in (13) we have left the dependence of EG on the vector r implicit). Assumption (A7) is a regularity condition which will be satisfied for most choices of densities $q(\epsilon|x,\theta_2)$. A sufficient condition for (A7) to hold is that $q(\epsilon|x,\theta_2)$ have finite first moments which are uniformly bounded in x. An immediate consequence of Assumption (A7) is given in Lemma 1 below.

Lemma 1 Under assumptions (A1), ..., (A7), $EV_\theta \epsilon B$.

Proof By assumption (A7) it follows that for any $x_t \epsilon \Delta$ and any measurable function f,

$$(14) \qquad E\{[u(x_t,f(x_t,\epsilon_t),\theta_1) + \epsilon_t(f(x_t,\epsilon_t))]|x_t\} \leq G(||u||_\infty|x_t,\theta_2)$$

It follows from (3) that for each $(x_t,\epsilon_t)\epsilon S$

$$(15) \qquad V_\theta(x_t,\epsilon_t) \leq \sup_\pi E\left\{\sum_{j=t}^{\infty} \beta^{(j-t)} G(||u||_\infty|x_j,\epsilon_j)|x_t,\epsilon_t,\theta_2,\theta_3\right\}$$

From (15) it follows that

$$(16) \qquad ||EV_\theta||_\infty \leq ||EG||_\infty/(1-\beta)$$

Since $||EG||_\infty < \infty$ by (A7) it follows that $EV_\theta \epsilon B$. Q.E.D.

We are now ready to state the main result of this paper.

Theorem 3 Let $P(i|x,\theta)$ denote the conditional probability that the agent chooses alternative i given that he is in observed state x. Then under assumptions (A1),...,(A7), $P(i|x,\theta)$ given by

$$(17) \qquad P(i|x,\theta) = G_i(r(x,\theta)|x,\theta_2)$$

where $r(x,\theta) \equiv \{r(x,i,\theta)|i\epsilon C(x)\}$ is given by

$$(18) \qquad r(x,i,\theta) = u(x,i,\theta_1) + \beta EV_\theta(x,i),\ i\epsilon C(x)$$

and where the function $EV_\theta \epsilon B$ is the unique fixed point to the contraction mapping $T_\theta: B \to B$ defined by

$$(19) \qquad T_\theta(EV_\theta)(x,i) = \int_y G([u(y,\theta_1) + \beta EV_\theta(y)]|y,\ \theta_2)p(dy|x,i,\theta_3).$$

Proof First we show that (A6) implies that the conditional expectation $EV_\theta(x,\epsilon,i)$ does not depend on ϵ, and so can be written as $EV_\theta(x,i)$. By (A6) and Fubini's theorem we have

$$(20) \qquad EV_\theta(x,\epsilon,i) = \int_S V_\theta(y,\eta)p(\eta,y|x,\epsilon,i,\theta_2,\theta_3)\mu(dy)x\lambda(d\eta)$$

$$= \int_y \{\int_\eta V_\theta(y,\eta)q(\eta|y,\theta_2)\lambda(d\eta)\}p(dy|x,i,\theta_3)\mu(dy)$$

$$= EV_\theta(x,i)$$

which is clearly independent of ϵ. Substituting the formula for $V_\theta(y,\eta)$ given by Bellman's equation (6) into equation (20) we obtain the fixed point equation for EV_θ

$$(21) \qquad EV_\theta(x,i) = \int_y \int_\epsilon \{\max_{i\epsilon C(y)} [u(y,i,\theta_1)+\epsilon(i)+\beta EV_\theta(y,i)]\}q(d\epsilon|y,\theta_2)p(dy|x,i,\theta_3)$$

$$= \int_y G([u(y,\theta_1) + \beta EV_\theta(y)]|y,\theta_2)p(dy|x,i,\theta_3).$$

By (A7) and Lemma 1, equation (19) defines a nonlinear operator $T_\theta: B \to B$. To show that T_θ is a contraction mapping, notice that for each y and for each $g,h \epsilon B$ we have

$$(22)\qquad \max_{i\in C(y)} [u(y,i,\theta_1)+\epsilon(i)+\beta g(y,i)] - \max_{i\in C(y)} [u(y,i,\theta_1)+\epsilon(i)+\beta h(y,i)]$$

$$\leq \max_{i\in C(y)} \beta|g(y,i)-h(y,i)|$$

It follows immediately from (22) that

$$(23)\qquad ||T_\theta(g) - T_\theta(h)||_\infty \leq \beta||g-h||_\infty$$

so that T_θ is a contraction mapping and EV_θ is the unique fixed point to T_θ in B. The formula for the conditional choice probabilities $P(i|x,\theta)$ given in (17) follows from the fact that $r(\theta,x) = u(x,\theta_1) + \beta EV_\theta(x)$ does not depend on ϵ and equation (12) of Theorem 2.

Q.E.D.

The significance of Theorem 3 is that the conditional choice probabilities of the stochastic control problem (3) can be computed using the same formulas used in the static case with the addition of expected discounted future utility $\beta EV_\theta(x,i)$ to the usual static utility term $u(x,i,\theta_1)$. Notice that the general static model of discrete choice arises as a special case of this model when $p(\cdot|x,i,\theta_3)$ is independent of i. In that case the expected utilities $EV_\theta(x,i)$ are also independent of i which implies (by Theorem 2) that G_i is a function of $\{u(x,j,\theta)|j\epsilon C(x)\}$ alone, so that $P(i|x,\theta) = G_i(u(x,\theta)|x,\theta_2)$ can be interpreted at the usual static choice probability. The intuition behind this result is clear; when $p(\cdot|x,i,\theta_3)$ is independent of i, current choices do not affect the evolution of the state variables $\{x_t,\epsilon_t\}$ and so have no future consequences. Therefore, it is optimal to behave myopically and choose the alternative i_t which maximizes the single period utility $u(x_t,i,\theta_1) + \epsilon_t(i)$. When current choices do have future consequences, the

term $\beta EV_\theta(x,i)$ provides the appropriate valuation of the future consequences of each action and must be added to the current utility in order to correctly describe the optimal behavior of the agent.

By choosing specific functional forms for $q(\epsilon|y,\theta_2)$ we can obtain concrete formulas for the choice probability $P(i|x,\theta)$ and the contraction mapping T_θ. For example, if $q(\epsilon|y,\theta_2)$ is given by a multivariate extreme value distribution with cdf $Q(\epsilon|y,\theta_2)$ given by

$$(24) \qquad Q(\epsilon|y,\theta_2) = \prod_{i\epsilon C(y)} \exp\{-\exp\{-\epsilon(i) - \theta_2\}\}$$

where $\theta_2 = \gamma \cong .577216$, then $P(i|x,\theta)$ is given by the well-known multinomial logit formula

$$(25) \qquad P(i|x,\theta) = \frac{\exp\{u(x,i,\theta_1)+\beta EV_\theta(x,i)\}}{\sum_{j\epsilon C(x)} \exp\{u(x,j,\theta_1)+\beta EV_\theta(x,j)\}}$$

and EV_θ is given by the unique solution to the functional equation

$$(26) \qquad EV_\theta(x,i) = \int_y \log\left\{\sum_{j\epsilon C(y)} \exp[u(y,j,\theta_1) + \beta EV_\theta(y,j)]\right\} p(dy|x,i,\theta_3)$$

Similarly, if $q(\epsilon|y,\theta_2)$ is a multivariate normal density $P(i|x,\theta)$ will take the form of a probit function. The drawback of the Gaussian family of distributions is that they are not closed under the operation of maximization (this property characterizes the family of extreme value distributions), so that numerical integration is required to evaluate $P(i|x,\theta)$. Since maximum likelihood estimation will require numerical computation of the fixed point EV_θ of each evaluation of the likelihood function, it may make sense to economize computer time by using distributions like the multivariate extreme

value (24) or McFadden's generalized extreme value distribution (GEV) (McFadden, (1981)) which yield closed-form nested logit expressions for $P(i|x,\theta)$. If computer costs decline sufficiently, more general distributions for $q(\epsilon|y,\theta_2)$ could be used in order to allow for more general patterns of dependence across alternatives, thus avoiding the well-known IIA property of the multinomial logit model (see Domencich and McFadden (1975)).

3. Derivation of the Maximum Likelihood Estimator

Suppose we have a panel of L individuals. For each individual we have T_ℓ periods of data. Thus our data consists of $(i_1^\ell, \ldots, i_{T_\ell}^\ell, x_1^\ell, \ldots, x_{T_\ell}^\ell)$ $\ell = 1, \ldots, L$. What is the likelihood function for our sample of data? The following lemmas will be useful in deriving the likelihood function.

Lemma 2 Under assumptions (A1),...,(A7), the controlled process $\{x_t,\epsilon_t\}$ is jointly markovian.

Proof

$$(27)\quad dp\{x_{t+1},\epsilon_{t+1}|x_t,\ldots,x_1,\epsilon_t,\ldots,\epsilon_1\} = q(\epsilon_{t+1}|x_{t+1},\theta_2)\, p(x_{t+1}|x_t,f(x_t,\epsilon_t,\theta),\theta_3)$$

which depends only on (x_t,ϵ_t).

Q.E.D.

Lemma 3 Under assumptions (A1),...,(A7), the controlled process $\{x_t,i_t\}$ is jointly markovian.

Proof

$$(28)\quad dp\{x_{t+1},i_{t+1}|x_t,\ldots,x_1,i_t,\ldots,i_1\}$$

$$= \int_\eta I\{f(x_{t+1},\eta,\theta) = i_{i+1}\}q(d\eta|x_{t+1},\theta_2)p(x_{t+1}|x_t,i_t,\theta_3)$$

which depends only on (x_t, i_t).

Q.E.D.

Lemma 4 Let $dp\{x_{t+1}, i_{t+1} | x_t, i_t\}$ denote the conditional probability density of (x_{t+1}, i_{t+1}) given (x_t, i_t). Under assumptions (A1),...,(A7) we have

$$(29) \qquad dp\{x_{t+1}, i_{t+1} | x_t, i_t\} = P(i_{t+1} | x_{t+1}, \theta) p(x_{t+1} | x_t, i_t, \theta_3)$$

Proof The conditional probability $dp\{x_{t+1}, i_{t+1} | x_t, i_t\}$ can be decomposed as

$$(30) \qquad dp\{x_{t+1}, i_{t+1} | x_t, i_t\} = dp\{i_{t+1} | x_{t+1}, x_t, i_t\} p(x_{t+1} | x_t, i_t, \theta_3).$$

However, by Bellman's equation (6) and assumption (A6), it is obvious that at time t+1 the state variable x_{t+1} is a sufficient statistic for computing the probability of i_{t+1}. Thus

$$(31) \qquad dp\{i_{t+1} | x_{t+1}, x_t, i_t\} = P(i_{t+1} | x_{t+1}, \theta)$$

where $P(i_{t+1} | x_{t+1}, \theta)$ is given by (17).

Q.E.D.

Suppose we are given initial conditions (x_o, i_o). We now wish to compute the joint likelihood of the observations $(x_1,...,x_T, i_1,...,i_T)$ given (x_o, i_o), which we denote by $L^f(x_1,...,x_T, i_1,...,i_T | x_o, i_o, \theta)$.

Theorem 4 Under assumptions (A1), ..., (A7) we have

$$(32) \qquad L^f(x_1,...,x_T, i_1,...,i_T | x_o, i_o, \theta) = \prod_{t=1}^{T} P(i_t | x_t, \theta) p(x_t | x_{t-1}, i_{t-1}, \theta_3)$$

Proof It is always possible to decompose L^f as a product of conditional likelihoods as follows:

(33) $L^f(x_1,\ldots,x_T,i_1,\ldots,i_T|i_o,x_o,\theta) =$

$L^f(x_1,i_1|x_o,i_o,\theta)L^f(x_2,i_2|x_1,x_o,i_1,i_o,\theta)\ \ldots.$

$L^f(x_T,i_T|x_{T-1},\ldots,x_o,i_{T-1},\ldots,i_o,\theta)$

By Lemmas 3 and 4 we have

(34) $L^f(x_t,i_t|x_{t-1},\ldots,x_o,i_{t-1},\ldots,i_o,\theta) = P(i_t|x_t,\theta)p(x_t|x_{t-1},i_{t-1},\theta_3)$

Substituting (34) into (33) yields our result.

Q.E.D.

Under appropriate regularity conditions, maximization of the likelihood L^f given in (32) provides a consistent, asymptotically normal, asymptotically efficient estimate of θ^*. In addition to L^f, we define two partial-likelihood functions L^1, L^2 which provide alternative methods for estimating θ.

(35) $L^1(x_1,\ldots,x_T,i_1,\ldots,i_T|x_o,i_o,\theta_3) \equiv \prod_{t=1}^{T} p(x_t|x_{t-1},i_{t-1},\theta_3)$

(36) $L^2(x_1,\ldots,x_T,i_1,\ldots,i_T|x_o,\theta) \equiv \prod_{t=0}^{T} P(i_t|x_t,\theta)$

Neither L^1 nor L^2 correspond to the true conditional likelihoods

(37) $L_1^c(x_1,\ldots,x_T|i_1,\ldots,i_T,i_o,x_o,\theta)$

and

(38) $L_2^c(i_1,\ldots,i_T|x_1,\ldots,x_T,i_o,x_o,\theta)$

obtained by dividing the joint likelihood L^f by the appropriate marginals.

Nevertheless, both L^1 and L^2 provide simpler partial-likelihood estimators which are consistent, and asymptotically normally distributed. Note that using L^1, one can only identify θ_3. This drawback is compensated by the fact that no internal calculation of EV_θ is required in order to estimate θ_3. Thus, L^1 may provide an easier means of obtaining initial consistent estimates of θ_3 to be used as starting values in L^2 or L^f. Let the maximum likelihood estimator corresponding to the likelihoods L^f, L^1, L^2 be denoted by $\hat{\theta}^f$, $\hat{\theta}^1$, $\hat{\theta}^2$, respectively.

In the case where the initial condition is stochastic L^f should be regarded as a conditional likelihood function. A more efficient estimator than $\hat{\theta}^f$ can be obtained by maximizing L^f multiplied by the probability density $\Omega(x_o,i_o,\theta)$ for the initial condition (x_o,i_o). There are two cases to consider: (1) the process has been operating for a long period of time and the joint distributions of (x_t,i_t) and (x_t,ϵ_t) have converged to the ergodic distribution, or (2) the process has only recently started up so (x_t,ϵ_t) or (x_t,i_t) cannot be regarded as having been drawn from the ergodic distribution. In the latter case, unless we know the initial condition of the process there is no way we can compute the probability distribution of (x_o,i_o) so that we have to be content with the conditional likelihood function (32). In the former case we can compute the unconditional probability of (x_o,i_o) as follows. First compute the ergodic distribution of the controlled process $\{x_t,i_t\}$. Under appropriate regularity conditions given in Appendix 3, this distribution is given by the unique solution $\Omega(x,i,\theta)$ to the following functional equation:

$$(39) \qquad \Omega(x,i,\theta) = \int_y \int_j P(i|x,\theta)p(x|y,j,\theta_3)\Omega(dy,dj,\theta)$$

In the case where $p(x|y,j,\theta_3)$ has finite support, (39) reduces to a matrix equation of which $\Omega(x,i,\theta)$ is the unit eigenvector. Using the computed value of Ω, we obtain the unconditional full information likelihood $L^*(x_o,\ldots,x_T,i_o,\ldots,i_T,\theta)$ as the product

$$(40) \qquad L^*(x_o,\ldots,x_T,i_o,\ldots,i_T,\theta) = \Omega(x_o,i_o,\theta) \prod_{t=1}^{T} P(i_t|x_t,\theta)p(x_t|x_{t-1},i_{t-1},\theta_3)$$

Denote the maximum likelihood estimator corresponding to L^* by $\hat{\theta}^*$.

If panel time lengths are relatively long the extra information contributed by $\Omega(x_o,i_o,\theta)$ is likely to add little to the efficiency of the estimator $\hat{\theta}^*$ since, intuitively, the contribution of the first term of the likelihood becomes negligible as $T_\ell \to \infty$. Note that under the assumption that all agents in the sample have the same parameter θ (i.e. no heterogeneity), there is no problem of inconsistency in parameter estimates created by conditioning on (x_o,i_o) (for a discussion of this problem of "initial conditions" see Heckman (1981)). Therefore, as a general practice we recommend using the likelihood function L^f which, while not the true full information maximum likelihood equation when (x_o,i_o) is stochastic, is a full information maximum likelihood equation <u>conditional</u> on (x_o,i_o).

With the exception of likelihood function L^1 given in (35), none of the likelihood functions derived in this section have an a priori known, or closed functional form. Each of the likelihood functions L^2 and L^f involve the conditional choice probability $P(i_t|x_t,\theta)$, which, by Theorem 3, depends on the unknown value function EV_θ. Since EV_θ is a fixed point to the contraction mapping T_θ defined in (19), our results suggest the following <u>nested fixed point</u> algorithm: an "inner" fixed point algorithm computes the fixed point EV_{θ_t} corresponding to the current parameter estimate θ_t, and an "outer" hill-climbing algorithm searches over alternative values of θ_t

to maximize the likelihood function. Given the computational burden involved in repeated fixed point calculations for each successive estimate θ_t, it seems clear that one should design an outer algorithm which finds the maximum using the smallest number of function evaluations. Perhaps the most important strategy for designing an efficient outer maximization algorithm, is to find successively stronger conditions under which the likelihood function is 1) continuous, 2) continuously differentiable, and 3) twice continuously differentiable in θ. Differentiability conditions 2) and 3) permit use of more efficient gradient optimization methods, with condition 2) permitting use of the BHHH (1974) optimization algorithm, and condition 3) permitting use of Newton's method. These conditions also determine the asymptotic distribution of the estimator, since condition 1) is used to prove consistency, and conditions 2) and 3) are used to prove asymptotic normality. Since the fundamental objects $u(x_t,i_t,\theta_1)$, $q(\epsilon_t|x_t,\theta_2)$ and $p(x_{t+1}|x_t,i_t,\theta_3)$ have a priori known functional forms, they can be chosen to be continuously differentiable. Thus, the only place where differentiability is at question is in the conditional choice probability $P(i_t|x_t,\theta)$. By Theorem 3, $P(i|x,\theta) = G_i([u(x,\theta_1)+\beta EV_\theta(x)]|x,\theta_2)$ is continuously differentiable function of $[u(x,\theta_1)+\beta EV_\theta(x)]$. Therefore the question of differentiability reduces to the question of finding sufficient conditions under which the mapping $\theta \rightarrow EV_\theta$ is a smooth mapping from $R^{1+K_1+K_2+K_2}$ into B. We turn to this problem in section 4.

4. Differentiability of the Expected Value Function.

Since EV_θ is a fixed point of the contraction mapping T_θ, EV_θ is an implicit function of θ. In this section we provide sufficient conditions for EV_θ to be twice continuously differentiable in θ. These results will be used in Section 5 to prove that the likelihood equations are twice continuously differentiable which (along with other regularity conditions) will imply that the maximum likelihood estimator is asymptotically normally distributed. Our approach is to provide successively stronger conditions which imply that EV_θ is 1) continuous, 2) continuously differentiable, and 3) twice continuously differentiable. Throughout, we will let B denote the Banach space of all measurable $||\cdot||_\infty$-bounded real valued functions on Γ, where $||\cdot||_\infty$ is the supremum norm.

(A8) u is a continuous function of θ_1.

(A9) For any $g \in B$, EG is a continuous function of (θ_2, θ_3) and g where

$$EG(x,i) = \int_y G(g(y)|y,\theta_2)p(dy|x,i,\theta_3).$$

Assumptions (A8) and (A9) are continuity assumptions on u, and q and p, respectively. For fixed θ_3 the dominated convergence theorem (Pratt, (1960)) implies that a sufficient condition for (A9) to hold is that $q(\epsilon|y,\theta_2)$ is continuous in θ_2 uniformly for $y \in \Delta$ for all ϵ in R^N except on sets A(y) of Lesbesgue measure zero. Similarly, by Scheffe's Theorem (Billingsley, (1979)), a sufficient condition for EG to be continuous in θ_3 is that $p(\cdot|x,i,\theta_3)$ is continuous in θ_3 uniformly for $(x,i) \in \Gamma$ and for all $y \in \Delta$ except on sets A(x,i) of μ measure zero.

Theorem 5 Under assumptions (A1),...,(A9), EV_θ is a continuous function of θ.

Proof The result follows from Theorem 3 of Kantorovich and Aikilov ((1982), p. 476), provided we can show that T_θ is continuous in θ, i.e. for each $g \in B$, $T_{\hat{\theta}}(g) \to T_\theta(g)$ as $\hat{\theta} \to \theta$. By formula (19) we can write

$$(41) \quad (T_{\hat{\theta}}-T_\theta)(g)(x,i) =$$

$$\int_y [G([u(y,\hat{\theta}_1)+\hat{\beta}g(y)]|y,\hat{\theta}_2) - G([u(y,\theta_1)+\beta g(y)]|y,\hat{\theta}_2)]p(dy|x,i,\hat{\theta}_3)$$

$$+ \int_y G([u(y,\theta_1)+\beta g(y)]|y,\theta_2)[p(dy|x,i,\hat{\theta}_3)-p(dy|x,i,\theta_3)]$$

$$+ \int_y [G([u(y,\theta_1)+\beta g(y)]|y,\hat{\theta}_2) - G([u(y,\theta_1)+\beta g(y)]|y,\theta_2)]p(dy|x,i,\hat{\theta}_3)$$

Since G satisfies

$$(42) \quad |G([u(y,\hat{\theta}_1)+\hat{\beta}g(y)]|y,\theta_2) - G([u(y,\theta_1)+\beta g(y)]|y,\theta_2)|$$

$$\leq ||u(\theta_1)-u(\hat{\theta}_1)||_\infty + |\beta-\hat{\beta}|\ ||g||_\infty ,$$

it follows from (A8) and (A9) that the first and third terms in (41) converge to zero uniformly on Γ. By assumption (A9) the second term converges to zero uniformly on Γ. Thus T_θ is continuous in θ which implies that EV_θ is also continuous in θ.

Q.E.D.

The following assumptions are sufficient to guarantee that EV_θ is a continuously differentiable function of θ. In what follows, let $L(R^k,B)$ denote the Banach space of linear operators from R^k to B.

(A10) $\partial u/\partial\theta_1 \in L(R^{K_1}, B)$ and is a continuous function of θ_1.

(A11) For any $r\in B$, $\partial G(r(y)|y,\theta_2)/\partial\theta_2$ exists and is dominated by a $p(\cdot|x,i,\theta_3)$ integrable function $g(y)$ for all y except on sets $A(x,i)$ of $p(\cdot|x,i,\theta_3)$ measure zero.

(A12) For each $r\in B$, $\partial EG/\partial\theta_2 \in L(R^{K_2}, B)$ and is a continuous function of θ and r.

(A13) For each $r\in B$, $\partial EG/\partial\theta_3 \in L(R^{K_3}, B)$ and is a continuous function of θ and r.

By the Lesbesgue dominated convergence theorem for derivatives (Billingsley (1979), Pratt (1960)), a sufficient condition for (A13) to hold is that $\partial p(y|x,i,\theta_3)/\partial\theta_3$ exists, is bounded, and is continuous in θ_3, uniformly for $(x,i)\in\Gamma$, except for y in sets $A(x,i)$ of μ measure zero. When this holds, we can write

$$(43)\qquad \partial EG/\partial\theta_3(x,i) = \int_y G(r(y)|y,\theta_2)\,[\partial p(y|x,i,\theta_3)/\partial\theta_3]\mu(dy)$$

For example, if for each $(x,i)\in\Gamma$ $p(\cdot|x,i,\theta_3)$ has finite support, then (A13) will hold provided for each $y\in\Delta$ and each $(x,i)\in\Gamma$, $p(y|x,i,\theta_3)$ is continuously differentiable in θ_3. Similarly (A11) and dominated converage imply that $\partial EG/\partial\theta_2(x,i) = \int_y \{\partial G(r(y)|y,\theta_2)/\partial\theta_2\}p(dy|x,i,\theta_3)$.

Theorem 5 Under Assumptions (A1),...,(A13), EV_θ is continuously differentiable in θ in a neighborhood of each point $\theta^*\in R^{\dim(\theta)} = R^{(1+K_1+K_2+K_3)}$

Proof (EV_θ,θ) is a zero of the nonlinear operator Λ defined on $B \times R^{\dim(\theta)}$ defined by

$$(44) \qquad 0 = \Lambda(EV_\theta,\theta) \equiv (I-T_\theta)(EV_\theta)$$

where 0 is the zero element in B and I is the identity operator on B. By the implicit function theorem for Banach spaces (Kantorovich and Aikilov (1982), Theorems 1 and 3, pp. 518-520), EV_θ will be continuously differentiable in θ in a neighborhood of any point θ^* satisfying (44) provided the partial (Gateaux) derivatives of Λ with respect to EV and θ exist and are continuous, and provided the partial derivative of Λ with respect to EV has a continuous inverse. Since I is the identity operator which is independent of θ, we first verify that $T_\theta(EV)$ has derivatives in EV and θ which exist and are continuous. We have

$$(45) \qquad \partial/\partial\theta_1[T_\theta(EV)](x,i) = \int_y \{ \sum_{j\in C(y)} \partial u(y,j,\theta_1)/\partial\theta_1 P(j|y,\theta)\} p(dy|x,i,\theta_3)$$

$$(46) \qquad \partial/\partial\theta_2[T_\theta(EV)](x,i) = \int_y \{\partial G([u(y,\theta_1) + \beta EV(y)]|y,\theta_2)/\partial\theta_2\} p(dy|x,i,\theta_3)$$

$$(47) \qquad \partial/\partial\theta_3[T_\theta(EV)](x,i) = \partial/\partial\theta_3 \{\int_y G([u(y,\theta_1) + \beta EV(y)]|y,\theta_2) p(dy|x,i,\theta_3)\}$$

$$(48) \qquad \partial/\partial\beta[T_\theta(EV)](x,i) = \int_y \{ \sum_{j\in C(y)} EV(y,j)\ P(j|y,\theta)\} p(dy|x,i,\theta_3)$$

The derivative in formula (47) exists and is continuous by (A13). Assumption (A11) and the dominated convergence theorem justify interchanging the operations of differentiation and integration to yield formula (46) which is continuous in θ by assumption (A12). Similarly (A10) and dominated convergence justifies interchanging differentiation and integration to compute

$\partial/\partial\theta_1[T_\theta(EV)]$ and $\partial/\partial\beta[T_\theta(EV)]$. By Theorem 5 and Theorem 2, $G_i([u(y,\theta_1) + \beta EV]|y,\theta_2) = P(i|y,\theta)$ and is a Lesbesgue almost everywhere continuous function of $u(y,\theta_1)$ and θ_2. Using an argument similar to the proof of Theorem 4, it is easy to show that (45) and (48) are continuous in θ. These formulas define elements of a Banach space since $\partial u/\partial\theta_1 \epsilon L(R^{K_1},B)$ and $EV\epsilon B$.

Let $T'_\theta(EV_\theta)$ denote the partial Gateaux derivative of T_θ with respect to EV evaluated at the point EV_θ. Since $EV\epsilon B$, it follows that $T'_\theta(EV_\theta)$ is a linear operator on B. Let $T'_\theta(EV_\theta)(m)$ denote the value of $T'_\theta(EV_\theta)$ evaluated at the point $m\epsilon B$, defined formally by

$$(49) \qquad T'_\theta(EV_\theta)(m) = \lim_{t\to 0} [T_\theta(EV_\theta+tm) - T_\theta(EV_\theta)]/t$$

where $t\epsilon R$. Then, by Theorem 2 and the dominated convergence theorem we have

$$(50) \qquad T'_\theta(EV_\theta)(m)(x,i) = \beta \int_y \{\sum_{i\epsilon C(y)} m(y,i)P(i|y,\theta)\}p(dy|x,i,\theta_3)$$

It is evident from (52) that $T'_\theta(EV_\theta)$ satisfies

$$(51) \qquad ||T'_\theta(EV_\theta)(m)||_\infty \leq \beta||m||_\infty$$

which implies that $T'(EV_\theta)$ is a continuous linear operator with norm $||T'_\theta(EV_\theta)||_\infty \leq \beta < 1$. By the Banach inverse theorem (Kantorovich and Aikilov (1982), Theorem 3, p. 154), it follows that $\partial\Lambda(EV,\theta)/\partial EV$ is a continuous linear operator and has continuous inverse given by

$$(52) \qquad [\partial\Lambda(EV,\theta)/\partial EV]^{-1} = [I - T'_\theta(EV)]^{-1} = \sum_{t=0}^{\infty} [T'_\theta(EV)]^t$$

Thus, all the conditions of the implicit function theorem are satisfied

so that EV_θ is a continuously differentiable function of θ in some neighborhood of θ^*.

Q.E.D.

Corollary 1 Under assumptions (A1),...,(A13), EV_θ is a continuously differentiable function for all θ in a compact set Θ.

Proof. From Theorem 6 we know that EV_θ is continuously differentiable in some neighborhood N_θ about each point θ in Θ. These neighborhoods form an open cover of Θ, so that by compactness there exists a finite number of points $\theta_1,\ldots,\theta_r$ whose associated neighborhoods N_{θ_i} cover Θ. Suppose $\theta \epsilon N_{\theta_i} \cap N_{\theta_j}$ for $j \neq i$. If $EV_\theta(\theta_i)$ is the implicit function of θ defined on N_{θ_i}, then by the uniqueness of the fixed point EV_θ of T_θ, $EV_\theta(\theta_i) = EV_\theta(\theta_j)$ on $N_{\theta_i} \cap N_{\theta_j}$, and hence EV_θ is uniquely defined and continuously differentiable over all of Θ.

Q.E.D.

Corollary 2 Under assumptions (A1),...,(A13), the derivative of EV_θ with respect to θ is given by

$$\partial EV_\theta/\partial\theta = \Big[\sum_{t=0}^{\infty} [T'_\theta(EV_\theta)]^t\Big] [\partial T_\theta(EV_\theta)/\partial\theta] = [I - T'_\theta(EV_\theta)]^{-1}[\partial T_\theta(EV_\theta)/\partial\theta] \tag{53}$$

where $\partial T_\theta(EV_\theta)/\partial\theta$ is given by (45),...,(48) and $T'_\theta(EV_\theta)$ is given by (50).

Proof. By the implicit function theorem we have

$$0 = \Lambda(EV_\theta, \theta) \tag{54}$$

for all θ in Θ. Differentiating the identity (54) and making use of the chain rule yields (53).

Q.E.D.

Corollary 3 Under assumptions (A1),...,(A13), $\partial EV_\theta/\partial\theta \epsilon L(R^{\dim(\theta)},B)$.

Proof. Using assumptions (A11),...,(A13) it is easy to verify that for each j, $\partial T_\theta(EV_\theta)/\theta_j \epsilon B$. Since the linear operator $[I-T'_\theta]^{-1}$ is continuous with norm not exceeding $1/(1-\beta)$, it follows immediately from (53) that $\partial EV_\theta/\partial\theta_j \epsilon B$, j=1, ..., dim($\theta$).

Q.E.D.

By formula (53) it appears that we can formally differentiate EV_θ a second time, using the product rule to obtain

$$(55)\quad \partial EV_\theta/\partial\theta\partial\theta' = \left[\sum_{t=0}^{\infty} t[T'_\theta(EV_\theta)]^{t-1}\right]\Big[[\partial T'_\theta(EV_\theta)/\partial\theta][\partial T_\theta(EV_\theta)/\partial\theta'] + T'_\theta(EV_\theta)(\partial EV_\theta/\partial\theta')\Big] + \left[\sum_{t=0}^{\infty}[T'_\theta(EV_\theta)]^t\right][\partial T_\theta(EV_\theta)/\partial\theta\partial\theta']$$

The first term consists of the product of dim(θ) linear operators on B times dim(θ) elements of B, yielding a matrix of $[\dim(\theta)]^2$ elements of B. The second term consists of a single linear operator on B acting on $[\dim(\theta)]^2$ elements of B, again yielding a matrix of $[\dim(\theta)]^2$ elements of B. We now present assumptions sufficient to guarantee that this formal differentiation is valid.

(A14) $\partial^2 u/\partial\theta_1\partial\theta_1' \epsilon L(R^{K_1^2}, B)$ and is a continuous function of θ_1 .

(A15) For any $r\epsilon B$, $j,k\epsilon C(y)$, $G_{jk}(r(y)|y,\theta_2)$ exists and is dominated by a $p(\cdot|x,i,\theta_3)$ integrable function for all $y\epsilon\Delta$ except on sets $A(x,i)$ of $p(\cdot|x,i,\theta_3)$ measure zero.

(A16) For any $g,r\epsilon B$, $\int_y \{ \sum_{j,k\epsilon C(y)} \sum g(y,j)g(y,k)G_{jk}(r(y)|y,\theta_2)\}p(dy|x,i,\theta_3)\epsilon B$ and is a continuous function of θ and r.

(A17) For any $r\epsilon B$, $\partial^2 G(r(y)|y,\theta_2)/\partial\theta_2\partial\theta_2'$ exists and is dominated by a $p(\cdot|y,i,\theta_3)$ integrable function for all y except on sets $A(x,i)$ of $p(\cdot|x,i,\theta_3)$ measure zero.

(A18) For any $r\epsilon B$, $E\{\partial^2 G/\partial\theta_2\partial\theta_2'\} \epsilon L(R^{K_2^2}, B)$ and is a continuous function of θ and r.

(A19) For any $r\epsilon B$, $\partial^2[\int_y G(r(y)|y,\theta_2)p(dy|x,i,\theta_3)]/\partial\theta_3\partial\theta_3' \epsilon L(R^{K_3^2}, B)$ and is a continuous function of θ and r.

(A20) For any $j\epsilon C(y)$, $\partial G_j(r(y)|y,\theta_2)/\partial\theta_2$ exists and is dominated by a $p(\cdot|x,i,\theta_3)$ integrable function for all y except on sets $A(x,i)$ of $p(\cdot|x,i,\theta_3)$ measure zero.

(A21) For any $r,g\epsilon B$, $\int_y \{ \sum_{j\epsilon C(y)} g(y,j)\ \partial G_j(r(y)|y,\theta_2)/\partial\theta_2\}p(dy|x,i,\theta_3)\epsilon L(R^{K_2}, B)$ and is a continuous function of θ and r.

Theorem 6 Under assumptions (A1),...,(A21), the second derivative of EV_θ given by formula (55) exists and is an element of $L(R^{\dim(\theta)^2}, B)$ and is a continuous function of θ.

Proof. In order to compute $\partial^2 EV_\theta/\partial\theta\partial\theta'$ by applying the product rule to formula (55), we must verify that each of the terms is a differentiable function of θ. Assumptions (A15),...,(A20) and the dominated convergence theorem imply that $\partial^2 T_\theta(EV_\theta)/\partial\theta\partial\theta'$ exists, is an element of $L(R^{\dim(\theta)^2},B)$ and is a continuous function of θ. Explicit formulas for these derivatives are given in Appendix 1, which should make clear exactly where each of the assumptions (A15),...,(A21) are used to prove existence and continuity of these derivatives. The linear operator $[I-T'_\theta(EV_\theta)]^{-1}$ is a continuous function of θ since from formula (50) it is evident that $T'_\theta(EV_\theta)$ is continuous in θ with norm uniformly bounded by β, and thus the dominated convergence theorem implies that $\sum_{t=0}^{\infty} [T'_\theta(EV_\theta)]^t = [I-T'_\theta(EV_\theta)]^{-1}$ is a continuous function of θ. Since both $[I-T'_\theta(EV_\theta)]^{-1}$ and $[T_\theta(EV_\theta)]/\partial\theta\partial\theta'$ are continuous functions of θ, so is their product, which is the second sum and in equation (55). To complete the proof we must verify that $\partial/\partial\theta[I-T'_\theta(EV_\theta)]^{-1}$ exists, is a bounded linear operator in $L(R^{\dim(\theta)},B)$ and is a continuous function of θ. First we show that the operator $\partial T'_\theta(EV_\theta)/\partial\theta'$ is a bounded linear operator on $L(R^{\dim(\theta)},B)$ and is a continuous function of θ. We have

(56) $$\partial T'_\theta(EV_\theta)(m)/\partial\theta_1(x,i) =$$

$$\beta \int_y \{ \sum_{j,k\epsilon C(y)} \sum m(y,j)\, \partial r(\theta,y,k)/\partial\theta_1' \, G_{jk}(r(\theta,y)|y,\theta_2)\} p(dy|x,i,\theta_3)$$

(57) $$\partial T'_\theta(EV_\theta)(m)/\partial\theta_2(x,i) =$$

$$\beta \int_y \{ \sum_{j,k\epsilon C(y)} \sum m(y,j)\, \partial r(\theta,y,k)/\partial\theta_2 G_{jk}(r(\theta,y)|y,\theta_2)\} p(dy|x,i,\theta_3)$$

$$+ \beta \int_y \{ \sum_{j\epsilon C(y)} m(y,j)\, \partial G_j(r(\theta,y)|y,\theta_2)/\partial\theta_2\} p(dy|x,i,\theta_3)$$

$$(58)\quad \partial T'_\theta(EV_\theta)(m)/\partial\theta_3(x,i) =$$

$$\beta \int_y \{ \sum_{j,k\epsilon C(y)} \sum m(y,j)\partial r(\theta,y,k)/\partial\theta_3 G_{jk}(r(\theta,y)|y,\theta_2)\} p(dy|x,i,\theta_3)$$

$$+ \beta\, \partial/\partial\theta_3 \int_y \{ \sum_{j\epsilon C(y)} m(y,j)P(j|y,\theta)\} p(dy|x,i,\theta_3)$$

$$(59)\quad \partial T'_\theta(EV_\theta)(m)/\partial\beta(x,i) = \int_y \{ \sum_{j\epsilon C(y)} m(y,j)P(j|y,\theta)\} p(dy|x,i,\theta_3)$$

$$+ \beta \int_y \{ \sum_{j,k\epsilon C(y)} \sum m(y,j)\partial r(\theta,y,k)/\partial\beta\, G_{jk}(r(\theta,y)|y,\theta_2)\} p(dy|x,i,\theta_3)$$

Using assumptions (A16),...,(A21), and Corollary 3 of Theorem 6, it is evident that $\partial T'_\theta(EV_\theta)/\partial\theta$ is a bounded linear operator on $L(R^{\dim(\theta)},B)$ and is a continuous function of θ. It remains to show that the operator $L_\theta = [\sum_{t=0}^{\infty} t[T'_\theta(EV_\theta)]^{t-1}]$ is a bounded linear operator on B and a continuous function of θ. To show that L_θ exists and is bounded, note that for each N, $L_\theta^N = [\sum_{t=0}^{N} t[T'_\theta(EV_\theta)]^{t-1}]$ is a bounded linear operator with norm not exceeding $1/(1-\beta)^2$. Furthermore, for each $g\epsilon B$, $L^N(g)$ is a fast Cauchy sequence in B, i.e. $\sum_{N=0}^{\infty} ||L_\theta^N(g)-L_\theta^{N+1}(g)||_\infty = \sum_{N=0}^{\infty} (N+1)||T'(EV_\theta;g)||^N < +\infty$. By the Banach-Steinhaus Theorem (Kantorovich and Aikilov, (1982), p. 203), it follows that $L_\theta^N \to L_\theta$ where L_θ is a linear operator with norm not exceeding $1/(1-\beta)^2$. Since $||t[T'(EV_\theta)]^{t-1}||_\infty \le t\beta^{t-1}$ with $\sum_{t=0}^{\infty} \beta^{t-1} = 1/(1-\beta)^2$, the dominated convergence theorem permits us to interchange summation and differentiation to compute $\partial/\partial\theta[I-T_\theta(EV_\theta)]^{-1}$ as $[\sum_{t=0}^{\infty} t[T'(EV_\theta)]^{t-1}]\,[\partial T'(EV_\theta)/\partial\theta']$ which verifies that it is valid to apply the product rule for differentiation to $\partial EV_\theta/\partial\theta$ given in (53) to obtain the formula for $\partial^2 EV_\theta/\partial\theta\partial\theta'$ given in (55). Furthermore, since $T'_\theta(EV_\theta)$ is continuous in θ, the dominated convergence

theorem also implies that L_θ is continuous in θ. In conclusion, we have shown that the second derivatives $\partial^2 EV_\theta/\partial\theta\partial\theta'$ exist, are continuous in θ, and are given by formula (55).

Q.E.D.

Using the results obtained in Theorem 6, Corollary 2, we can how describe in more detail the nested fixed point algorithm suggested in Section 3. Recall that the algorithm consists of two components, an outer "hill-climbing" algorithm and an "inner" fixed point algorithm. The outer algorithm searches over values of θ in order to maximize the likelihood functions L^f or L^2. Each time a new estimate θ_t is generated by the outer algorithm, we must invoke the inner algorithm in order to compute the corresponding fixed point EV_{θ_t}. Theorem 5 established the Gateaux differentiability of the operator T_θ, which implies that the fixed point EV_θ can be computed using the quadratically convergent Newton-Kantorovich algorithm. A by-product of this method and Corollary 2 is that analytic derivatives $\partial EV_\theta/\partial\theta$ can be computed with marginal additional effort since the computation of $\partial EV_\theta/\partial\theta$ requires the inverse operator $(I-T'_\theta)^{-1}$ which is already computed as part of the Newton-Kantorovich iterations. If the values of the likelihood function can be computed with about the same effort as its derivatives, it seems clear that one wants an outer hill climbing algorithm which minimizes the number of function evaluations needed to get within any pre-specified tolerance of a local maximum. Ordinarily this would mandate use of Newton's method, however, ordinary Newton's method requires <u>second</u> derivatives of the log-likelihood function which in turn requires the second derivatives $\partial^2 EV_\theta/\partial\theta\partial\theta'$. In general the reduction in the number of likelihood function evaluations gained by using Newton's method is outweighed by the considerably greater effort required to compute $\partial^2 EV_\theta\partial\theta\partial\theta'$ at each function evaluation. An alternative, asymptotically equivalent

algorithm is based on the well-known identity $-E\{\partial^2 \log L(\theta)/\partial\theta\partial\theta'\} = E\{\partial \log L(\theta)/\partial\theta \partial \log L(\theta)/\partial\theta'\}$. The BHHH algorithm (Berndt, Hall, Hall and Hausman (1974)) uses this identity to yield an alternative Newton-like algorithm which requires only first derivatives of the likelihood function. This algorithm substitutes the sample average of $-\partial \log L(\theta)/\partial\theta \partial \log L(\theta)/\partial\theta'$ for the true Hessian matrix of the log-likelihood function. If the sample size is large and the model is correctly specified, such a substitution should have negligible impact in the rate of convergence of the outer hill-climbing algorithm, while at the same time avoiding the computational burden of calculating $\partial^2 EV_\theta/\partial\theta\partial\theta'$. The resulting nested fixed point algorithm adopts BHHH method as the outer hill-climbing algorithm and is summarized in Diagram 2.

5. Asymptotic Properties of the Maximum Likelihood Estimator

In principle, we must consider three separate cases in order to derive the asymptotic distribution of the maximum likelihood estimator. In case one, we have a fixed number T_ℓ periods of observation for each individual, but the number of individuals L tends to infinity. In case two, we have a fixed number of individuals L, but the number of time periods T_ℓ tends to infinity. In the final case, both T_ℓ and L tend to infinity. Currently, most panel data sets have only a limited number of time periods T_ℓ so that approximating the finite sample distribution of the maximum likelihood estimator by its asymptotic distribution as $T_\ell \to \infty$ is not likely to yield accurate results. Furthermore, calculation of the asymptotic distribution of the maximum likelihood estimator in cases two and three is complicated by serial dependence in the $\{x_t, i_t\}$ process. Since $\{x_t, i_t\}$ is jointly markov by lemma 2 of section 3, one can show that the log likelihood forms a zero mean

martingale, so that a martingale strong law of large numbers must be applied to prove consistency, and a martingale central limit theorem must be used to prove asymptotic normality of the maximum likelihood estimator. Application of these results appears to require continuous third derivatives of the likelihood (see Billingsley (1961)). In the interest of space, we have decided to handle the complications arising in cases two and three in a sequel to this paper. In case one, the assumption of independence between individuals in a cross section seems tenable. With the additional assumptions that 1) the observation period T_ℓ is the same for all individuals and equal to a finite constant T, and, 2) the initial distribution of (x_o, i_o) is given by the ergodic distribution $\Omega(x_o, i_o, \theta)$, it follows that the likelihood equations are independent and <u>identically</u> distributed among a cohort of individuals. Thus enables us to use the simpler i.i.d. strong law of large numbers and the Lindeberg-Levy Central Limit theorem to prove consistency and asymptotic normality of the estimators $\hat{\theta}^f, \hat{\theta}^1, \hat{\theta}^2$. Use of these theorems will only require boundedness and continuity of the second derivatives of the log-likelihood equation. Actually, Huber (1967) has shown that asymptotic normality obtains with a weaker Lipschitz condition on the first derivatives of the log-likelihood function. The cost of this extra generality is exacted in the form of more stringent and less intuitive conditions on moments of the score function. Therefore, we have opted for the standard approach using continuous second derivatives. Our treatment of the i.i.d. maximum likelihood case is entirely standard and follows previous treatments, e.g. White (1982). In the following discussion, let $\hat{\theta}_L^i$ denote the estimator obtained by maximizing the likelihood function L^i derived from a sample of L individuals, i = f,1,2. The following assumption guarantees that these estimators are well-defined random vectors.

(A22) The parameter space Θ is a compact subset of $R^{(1+K_1+K_2+K_3)}$.

Theorem 8 Under assumptions (A1),...,(A10) and (A22), the maximum likelihood estimator $\hat{\theta}_i$, i = f,1,2 is a measurable function of the random variables $(x_1^{\ell},\ldots,x_T^{\ell},i_1^{\ell},\ldots,i_T^{\ell})$, ℓ=1,...,L.

Proof Follows immediately from Theorem 2.1 of White (1982).

(A23) For all $\theta\epsilon\Theta$ and initial distributions Ω, $|\log p(x_{t+1}|x_t,i_t,\theta_3)|$ has finite expectation with respect to (x_{t+1},x_t,i_t).

Assumption (A23) is a standard regularity assumption necessary to guarantee consistency. The next assumption (A24) is the standard identification condition necessary to insure consistency of the maximum likelihood estimator.

(A24) For any initial distribution Ω, $E\{\log P(i|x,\theta)\}$ and $E\{\log p(y|x,i,\theta_3)\}$ have unique maxima at $\theta = \theta^*$ and $\theta_3 = \theta_3^*$, respectively.

Assumption (A24) can be deduced from the well-known "information inequality" provided the sets A and B have nonzero probability, where $A = \{(x,i)\epsilon\Gamma | P(i|x,\theta) \neq P(i|x,\theta^*)\}$ and $B = \{(y,x,i)\epsilon\Delta x\Gamma | p(y|x,i,\theta_3) \neq p(y|x,i,\theta_3^*)\}$. Thus, it seems likely that we could prove assumption (A24) from more basic conditions on the underlying functions u,p, and q. However, since even in simple models it is difficult to determine a priori whether the identification condition holds, and since in practice lack of identification will show up in the form of a singular or near-singular hessian matrix, in the interests of space we simply

assume that (A24) holds. The following theorem proves consistency of the estimators $\hat{\theta}_L^i$, i=1,2,f.

Theorem 9 Under assumptions (A1),...,(A10) and (A21),...,(A24), as $L \to \infty$ $\hat{\theta}_L^i$ converges to θ^* with probability one, i = f,1,2.

Proof The result for $\hat{\theta}_L^f$ follows immediately from Theorem 2.2 of White (1982). The consistency of $\hat{\theta}^1$ and $\hat{\theta}^2$ follows since L^1 and L^2 are partial-likelihood functions, and the Cox's consistency theorem for "partial-likelihood" estimators, Cox (1975).

Q.E.D.

The next set of assumptions guarantee the boundedness and continuity of the score function and its derivative, and are used to prove the asymptotic normality of the maximum likelihood estimator. To motivate these conditions it is helpful to write out explicit formulas for the score function. Let $L^f(\theta) = \prod_{t=1}^{T} P(i_t|x_t,\theta)p(x_t|x_{t-1},i_{t-1},\theta_3)$ be the likelihood for a single individual. Then

$$(60)\quad \partial \log L^f(\theta)/\partial\theta_1 = \sum_{t=1}^{T}\Big\{\sum_{j\in C(x_t)} [\partial r(\theta,x_t,j)/\partial\theta_1\, G_{i_t j}(r(\theta,x_t)|x_t,\theta_2)/P(i_t|x_t,\theta)\Big\}$$

$$(61)\quad \partial \log L^f(\theta)/\partial\theta_2 = \sum_{t=1}^{T}\Big\{\sum_{j\in C(x_t)} \beta\partial EV_\theta(x_t,j)/\partial\theta_2\, G_{i_t j}(r(\theta,x_t)|\theta_2)/P(i_t|x_t,\theta)\Big\}$$

$$+ \sum_{t=1}^{T}\{\partial G_{i_t}(r(\theta,x_t)|x_t,\theta_2)/\partial\theta_2\}/P(i_t|x_t,\theta)$$

$$(62)\quad \partial \log L^f(\theta)/\partial\theta_3 = \sum_{t=1}^{T}\Big\{\sum_{j\in C(x_t)} \beta\partial EV_\theta(x_t,j)/\partial\theta_3\, G_{i_t j}(r(\theta,x_t)|x_t,\theta_2)/P(i_t|x_t,\theta)\Big\}$$

$$+ \sum_{t=1}^{T}\{\partial \log\, p(x_t|x_{t-1},i_{t-1},\theta_3)/\partial\theta_3\}$$

$$(63)\quad \partial \log L^f(\theta)/\partial\beta = \sum_{t=1}^{T} \{ \sum_{j\epsilon C(x_t)} [EV_\theta(x_t,j)+\beta\partial EV_\theta(x_t,j)/\partial\beta]G_{i_t j}(r(\theta,x_t)|x_t,\theta_2)/P(i_t|x_t,\theta)\}$$

where $r(\theta,x_t,i_t) = u(x_t,i_t,\theta_1) + \beta EV_\theta(x_t,i_t)$.

Assumptions (A1),...,(A14) and (A23), guarantee that the score function and its expectation are bounded and continuous in θ. The remaining conditions we need are ones to guarantee that the second derivatives of the log likelihood have finite expectation and are continuous functions of θ. The second derivatives of the log likelihood function are presented in Appendix 2. From these formulas it is apparent that the following additional assumptions are required.

(A25) $\partial^2 \log p(x_{t+1}|x_t,i_t,\theta_3)/\partial\theta_3\partial\theta_3'$ is a continuous function of θ for $\mu x\mu x\kappa$ almost all $(x_{t+1},x_t,i_t)\epsilon\Delta x\Gamma$, where κ is counting measure on R.

(A26) $|\partial^2 \log p(x_{t+1}|x_t,i_t,\theta_3)/\partial\theta_3\partial\theta_3'|$ and $|[\partial \log p(x_{t+1}|x_t,i_t,\theta_3)/\partial\theta_3][\partial \log p(x_{t+1}|i_t,\theta_3)/\partial\theta_3']|$ have finite expectation in (x_{t+1},x_t,i_t) for all $\theta\epsilon$ and all initial distributions Ω.

(A27) For each $r\epsilon B$, μ almost all $x\epsilon\Delta$, and all $i,j,k\epsilon C(x)$, the third partial derivatives $G_{ijk}(r(x)|x,\theta_2)$ exist and are continuous in $r(x)$ and θ_2. Furthermore, for any $g,h\epsilon B$

$$\int_y \{\sum_{i,j,k\epsilon C(y)} \sum \sum g(y,j)h(y,k)G_{ijk}(r(y)|y,\theta_2)\}p(dy|x,i,\theta_3)\epsilon B.$$

(A28) For each $r\epsilon B$, μ almost all $x\epsilon\Delta$, and all $j,k\epsilon C(x)$, the derivative $\partial G_{jk}(r(x)|x,\theta_2)/\partial\theta_2$ exists and is continuous in $r(x)$ and θ_2. Furthermore, the function $\int_y \{ \sum_{j,k\epsilon C(y)} \sum g(y,k)\partial G_{jk}(r(y)|y,\theta_2)/\partial\theta_2\}p(dy|x,i,\theta_3)$ is an element of B, for any $g\epsilon B$.

(A29) For each $r \epsilon B$, μ almost all $x \epsilon \Delta$ and all $j \epsilon C(x)$, the derivative $\partial^2 G_j(r(x)|x,\theta_2)/\partial\theta_2\partial\theta_2'$ exists and is continuous in $r(x)$ and θ_2. Furthermore,

$$\int_y \{ \sum_{j \epsilon C(y)} \partial^2 G_j (r(x)|x,\theta_2)/\partial\theta_2\partial\theta_2' \} p(dy|x,i,\theta_3) \epsilon L(R^{K_2^2}, B).$$

Assumptions (A25) and (A26) are standard assumptions necessary to allow us to apply the Lindeberg-Levy central limit theorem to prove asymptotic normality of $\hat{\theta}_L^i$, i=f,1,2. Assumptions (A25) and (A26) guarantee that the expectation of the hessian and outer product of the first derivatives of $\log p(y|x,i,\theta_3)$ are finite for all $\theta \epsilon \Theta$. These assumptions are exactly the ones used by White (1982) to prove asymptotic normality in the general i.i.d. case. Assumptions (A28), (A29), (A30) are equivalent to differentiability assumptions on the choice probabilities $P(i|x,\theta)$ which guarantee continuity and integrability of $\partial^2 \log P(i|x,\theta)/\partial\theta\partial\theta'$. For example, continuity of $G_{ijk}(r(x)|x,\theta_2)$ is equivalent to assuming that the choice probability $P(i|x,\theta)$ is twice continuously differentiable in $r(x,j)$ and $r(x,k)$. Similarly, continuity of $\partial^2 G_i(r(x)|x,\theta_2)/\partial\theta_2\partial\theta_2'$ guarantees that the choice probability $P(i|x,\theta)$ is twice continuously differentiable in the parameters θ_2 of the distribution of unobservables, $Q(\epsilon|x,\theta_2)$. For example, in the case of gaussian ϵ_t's with unrestricted covariance matrix, (A30) requires the probit choice probabilities to be twice continuously differentiable in the covariance matrix parameters. Note that assumptions (A28), (A29),(A30) could also be stated directly in terms of assumptions on the distribution function $Q(\epsilon|x,\theta_2)$. For example, a sufficient condition for (A28) to hold is that $Q(\epsilon|x,\theta_2)$ is three times continuously differentiable in ϵ, with the integrals

$$\int_{-\infty}^{\infty} Q_{ijk}(\iota(x)\epsilon - r(x)]|x,\theta_2)d\epsilon$$

uniformly bounded for $x\epsilon\Delta$ for each $r \epsilon B$.

To complete the list of assumptions define the [dim(θ) x dim(θ)] matrices $J_T^i(\theta)$ and $A_T^i(\theta)$ by

$$(64) \qquad H_T^1(\theta) = \sum_{t=1}^{T} E\{\partial^2 \log P(i_t|x_t,\theta)/\partial\theta\partial\theta'\}$$

$$(65) \qquad H_T^2(\theta) = \sum_{t=1}^{T} E\{\partial^2 \log p(x_t|x_{t-1},i_{t-1},\theta_3)/\partial\theta\partial\theta'\}$$

$$(66) \qquad H_T^f(\theta) = H_T^1(\theta) + H_T^2(\theta)$$

$$(67) \qquad A^1(\theta) = E\{[\sum_{t=1}^{T} \partial \log P(i_t|x_t,\theta)/\partial\theta][\sum_{t=1}^{T} \partial \log P(i_t|x_t,\theta)/\partial\theta']\}$$

$$(68) \qquad A_T^2(\theta) = E\{[\sum_{t=1}^{T} \partial \log p(x_t|x_{t-1},i_{t-1},\theta_3)/\partial\theta][\sum_{t=1}^{T} \partial \log p(x_t|x_{t-1},i_{t-1},\theta_3)/\partial\theta')]\}$$

$$(69) \qquad A_T^f(\theta) = E\{[\sum_{t=1}^{T} \partial \log P(i_t|x_t,\theta)p(x_t|x_{t-1},i_{t-1},\theta_3)/\partial\theta]$$

$$\cdot [\sum_{t=1}^{T} \partial \log P(i_t|x_t,\theta)p(x_t|x_{t-1},i_{t-1},\theta_3)/\partial\theta']\}$$

where the expectation in (64),...,(69) are taken with respect to the density $[\prod_{t=1}^{T} P(i_t|x_t,\theta)p(x_t|x_{t-1},i_{t-1},\theta_3)]\Omega_\theta(x_o,i_o)$, and where Ω is the density of the ergodic distribution of (x_o,i_o) given by the solution to equation (39).

Note that $-H_T^f(\theta)$ is simply the information matrix of the full information maximum likelihood estimator $\hat{\theta}^f$. Formulas (64), (65), (66) clearly display the information value of adding extra time periods to the panel. The matrix

$A_T^i(\theta)$ is the expectation of the outer product of the first derivatives of $\log L_T^i$, i = f,1,2. When $[H^i(\theta)]^{-1}$ exists, define the matrix $J_T^i(\theta)$ by

(70) $$J_T^i(\theta) = [H_T^i(\theta)]^{-1}[A_T^i(\theta)][H_T^i(\theta)]^{-1} \quad i=f,1,2$$

The following assumption guarantees the existence of the required inverses of $H_T^i(\theta)$ and $A_T^i(\theta)$.

(A30) θ^* is interior to Θ, $A_T^i(\theta^*)$ is nonsingular, and θ^* is a regular point of $H_T^i(\theta^*)$, i=f,1,2.

Theorem 10 Under Assumptions (A1),...,(A30), $H_T^i(\theta^*)$ is negative definite, i=f,1,2.

The proof is a direct application of Theorem 3.1 of White (1982). In the absence of problems of non-identification, the regularity of $H_T^i(\theta)$ is a generic property. This follows from the theory of Morse functions. Specifically, a Morse function is a twice continuously differentiable function which has a nonsingular hessian at each critical point. Clearly, if we knew that the score function was a Morse function, we could do without the assumption that θ^* is a regular point of $H_T^i(\theta)$. The Morse Lemma (Guilleman and Pollock, (1974)) tells us (roughly speaking) that almost every C^2 function is a Morse function. Furthermore, as a practical matter, singularity in $H_T^i(\theta)$ will be detected in the course of finding the optimum of the likelihood function.

Theorem 11 Under assumptions (A1),...,(A30),

$$\sqrt{L}\ (\hat{\theta}_L^i - \theta^*) \xrightarrow{D} N(0, J_T^i(\theta^*)) \quad i=f,1,2$$

Theorem 10 is the main result of this paper, proving the asymptotic normality of the maximum likelihood estimators $\hat{\theta}^i$, i=f,1,2. A consistent estimate of the matrix $J_T^i(\theta^*)$ can be obtained by evaluating the sample analogs of $H_T^i(\theta)$ and $A_T^i(\theta)$ at $\hat{\theta}_L^i$. Note that the asymptotic covariance matrix $J_T^i(\theta^*)$ is not automatically guaranteed to equal the inverse of the information matrix $-H_T^i(\theta^*)$. To obtain that result we need two final assumptions.

(A31) For $\mu \times \kappa$ almost all $(x,i)\epsilon\Gamma$, the minimal support of $p(\cdot|x,i,\theta_3)$ does not depend on θ.

(A32) There exists an integrable function $h(x_{t+1},x_t,i_t)$ which for almost all (x_{t+1},x_t,i_t) (relative to the product measure $\mu x\mu x\kappa$ on $\Delta x\Gamma$) satisfies

$$(71) \qquad h(x_{t+1},x_t,i_t) \geq p(x_{t+1}|x_t,i_t,\theta_3)$$

$$(72) \qquad \int_{x_{t+1}} \int_{x_t} \int_{i_t} h(dx_{t+1},dx_t,di_t) < +\infty$$

Note that if μ has finite support, the right hand side of (71) will be dominated by h = 1 so (A32) will be satisfied. In empirical applications discrete valued x variables will always be used in order to compute the required fixed point of T_θ.

Theorem 12 Under Assumptions (A1),...,(A32), $A_T^i(\theta^*) = -H_T^i(\theta^*)$ and

$$\sqrt{L}\,(\hat{\theta}_L^i - \theta^*) \xrightarrow{D} N(0, -[H_T^i(\theta^*)]^{-1}),\ i=f,1,2$$

The proofs of Theorems 11 and 12 are straightforward applications of the general results of White (1982), but since our assumptions aren't perfectly matched to theirs, some discussion is in order. White requires that the mini-

mal support of the likelihood function not to depend on θ. By Assumption (A1) the minimal support of $P(i|x,\theta)$ equals the set $C(x)$ which is independent of θ. By (A31) the minimal support of $p(\cdot|x,i,\theta_3)$ is independent of θ which implies in turn that the likelihoods L^i, i=f,1,2 have support which are independent of θ. Finally, White places a domination condition on the derivative of the product of the score and the likelihood in order to prove that $A^f(\theta_T^*) = -H^f(\theta_T^*)$. Given that we have proved the uniform boundedness of the score and its derivative, it follows that it is sufficient to bound the likelihood uniformly in θ. This is precisely condition (A32).

Theorem 13 Under assumptions (A1),...,(A30), the random variables

$$\partial \log P(i_t|x_t,\theta)p(x_t|x_{t-1},i_{t-1},\theta_3)/\partial\theta \text{ and}$$

$$\sum_{s=1}^{w} \partial \log P(i_s|x_s,\theta)p(x_s|x_{s-1},i_{s-1},\theta_3)/\partial\theta$$

are uncorrelated for $w < t$, which implies

$$(73) \quad A_T^f(\theta) = \sum_{t=1}^{T} E\{[\partial \log P(i_t|x_t,\theta)p(x_t|x_{t-1},i_{t-1},\theta_3)/\partial\theta] \cdot [\partial \log P(i_t|x_t,\theta)p(x_t|x_{t-1},i_{t-1},\theta_3)/\partial\theta']\}$$

Proof Since the controlled process $\{x_t,i_t\}$ is jointly markovian by lemma 3, it follows that the derivative of the log-likelihood function forms a zero-mean martingale (see Billingsley, (1961)) which immediately implies (73).

Q.E.D.

Theorem 14 Under assumptions (A1),...,(A30), and assuming that $Q(x_o,i_o,\theta)$ is the unique ergodic density, we have

$$(74)\quad A_T^f(\theta) = TE\{[\partial \log P(i_1|x_1,\theta)p(x_1|x_o,i_o,\theta_3)/\partial\theta] \cdot [\partial \log P(i_1|x_1,\theta)p(x_o,i_o,\theta_3)/\partial\theta']\}$$

$$(75)\quad H_T^f(\theta) = TE\{\partial^2 \log P(i_1|x_1,\theta)/\partial\theta\partial\theta'\} + TE\{\partial^2 \log p(x_1|x_o,i_o,\theta_3)/\partial\theta\partial\theta'\}$$

where the expectation is taken with respect to the density

$$(76)\quad \pi(x_1,i_1,x_o,i_o,\theta) = P(i_1|x_1,\theta)p(x_1|x_o,i_o,\theta_3)\Omega(x_o,i_o,\theta)$$

Proof. Since the ergodic density Ω is self-reproducing by formula (39), it follows that for all t the joint density of (x_t,i_t) is given by $\Omega(x_t,i_t,\theta)$, and the joint density of $(x_t,i_t,x_{t-1},i_{t-1})$ is given by $\pi(x_t,i_t,x_{t-1},i_{t-1},\theta)$ defined in formula (76). This immediately implies that formulas (66) and (69) for $A_T^f(\theta)$ and $H_T^f(\theta)$ reduce to formulas (74) and (75), respectively.

Q.E.D.

It follows from Theorem 14 that increasing the number of time periods T decreases the asymptotic variance of $\hat{\theta}_L^f$ at rate 1/T. The value of adding the initial probability $\Omega(x_o,i_o,\theta)$ to obtain the full-information likelihood function L_T^* given in formula (40) is to increase information by amount $-E\{\partial^2\Omega(x_o,i_o,\theta)/\partial\theta\partial\theta'\}$, which equals $E\{[\partial\Omega(x_o,i_o,\theta)/\partial\theta][\partial\Omega(x_o,i_o,\theta)/\partial\theta']\}$ under an appropriate analog of Assumption (A32). We have not been able to prove that $\Omega(x_o,i_o,\theta)$ is twice continuously differentiable in θ, so we have focused our results on the estimators $\hat{\theta}_T^i$, i = f,1,2. One can see that if T is relatively large, the effect of adding log Ω to log L^f will have relatively small impact on the asymptotic standard errors of the estimator. Given the added difficulty of computing Ω and $\partial\Omega/\partial\theta$, the researcher might be

justified in concluding that the incremental value of this information is low in comparison to its computation cost.

We should point out that the presence of alternative estimators for θ^* provides the basis for Hausman-type specification tests (Hausman, (1978)). One particularly easy test statistic to compute is $L(\hat{\theta}_3^f-\hat{\theta}_3^2)\{[\hat{J}_T^f(\hat{\theta}_3^f)]^{-1}-[\hat{J}_T^z(\hat{\theta}_3^z)]\}^{-1}(\hat{\theta}_3^f-\hat{\theta}_3^2)'$ which is asymptotically chi-square with K_3 degrees of freedom, where $\hat{\theta}_3^f$ and $\hat{\theta}_3^2$ are the full information and partial-likelihood estimates of θ_3^* and $[\hat{J}_T^i(\hat{\theta}_3^i)]^{-1}$ is a consistent estimate of the (θ_3,θ_3) block of the asymptotic covariance matrix of $\hat{\theta}_3^i$ i=f,2. Under the null hypothesis of no model misspecification, both $\hat{\theta}_3^f$ and $\hat{\theta}_3^z$ are consistent estimates of θ_3^*, with $\hat{\theta}_3^f$ efficient in the class of all CUAN estimators which condition on (x_o,i_o). It follows that $\sqrt{L}(\hat{\theta}_3^f-\theta_3^*)$ and $\sqrt{L}(\hat{\theta}_3^z-\theta_3^*)$ are asymptotically uncorrelated, so that the covariance matrix of their difference is consistently estimated by $\{[\hat{J}_T^f(\hat{\theta}_3^f)]^{-1} - [\hat{J}_T^2(\hat{\theta}_3^z)^{-1}]\}$. Inverting this latter expression and pre- and post multiplying by $\sqrt{L}(\hat{\theta}_3^f-\hat{\theta}_3^2)$ yields the Hausman test statistic which is asymptotically chi-square with degrees of freedom K_3. The idea behind the test is that under misspecification, there is no reason to assume that $\hat{\theta}_3^f$ and $\hat{\theta}_3^z$ will converge to the same value. Large discrepancies in these estimators then provide evidence of model misspecification. The test statistic is relatively easy to compute since the estimator $\hat{\theta}_3^z$ is obtained from the likelihood L^z which doesn't require internal computation of the optimal policy $f(x_t,\epsilon_t,\theta)$.

Although parameter estimates $\hat{\theta}$ are the end result of maximum likelihood estimation, in many cases interest focuses not on the parameters themselves but on the form of the estimated value function $V_\theta(x,i) \equiv [u(x,i,\hat{\theta}_1)+\hat{\beta}EV_\theta(x,i)]$. Treating the estimated value function V_θ as a B-valued random element, we

would like to compute its asymptotic distribution directly in order, for example, to compute confidence bands which tell us how precisely we estimate the unknown value function V_{θ^*}. Recent progress on the theory of probability distributions in linear spaces (Vakhania, (1981)) enables us to prove the following result.

Theorem 15 Let $\hat{\theta}_N$ be a consistent estimator of θ^* with $\sqrt{N}(\hat{\theta}_N-\theta^*) \xrightarrow{D} N(0,\Sigma)$.

Then 1) the value function $V_{\hat{\theta}}$ is a B-valued random element,

2) $V_{\hat{\theta}}$ is a consistent estimator of V_{θ^*}, and

3) $\sqrt{N}(V_{\hat{\theta}_N}-V_{\theta^*})$ converges weakly to a Gaussian random element on B with expectation 0 and covariance operator $[\partial V_{\theta^*}/\partial\theta][\Sigma][V_{\theta^*}/\partial\theta']\epsilon L(B,L(B,B))$.

Proof First we show that $V_{\hat{\theta}}$ is a B-valued random element for any random variable $\hat{\theta}$. Let $L \epsilon B^*$, the dual space of B. Since $LoV_\theta: \rightarrow R$ is a continuous mapping from Θ to R, it follows immediately that for any random variable $\hat{\theta}$, $LoV_{\hat{\theta}}$ is a real random variable. By Lemma 2.2.2 of Taylor (1978), $LoV_{\hat{\theta}}$ is a random variable for each $L \epsilon B^*$ iff $V_{\hat{\theta}}$ is a B-valued random element, establishing 1). Since $V_\theta: \Theta \rightarrow B$ is a continuously differentiable mapping from Θ to B, we have

$$(77)\quad ||V_{\hat{\theta}_N}-V_{\theta^*}-\partial V_{\theta^*}/\partial\theta(\hat{\theta}_N-\theta^*)||_\infty \leq ||\hat{\theta}_N-\theta^*|| \sup_{0<\alpha<1} ||\partial V_{\theta^*}/\partial\theta-\partial V_{\theta^*}/\partial\theta-\partial V_{\theta^*+\alpha(\hat{\theta}_N-\theta^*)}/\partial\theta||_\infty$$

by the mean value theorem for Banach spaces (Kantorovich and Aikilov (1982), p. 500). Since with probability 1, $\hat{\theta}_N \rightarrow \theta^*$, (77) implies that with probability 1 $V_{\hat{\theta}_N} \rightarrow V_{\theta^*}$. Furthermore, we have

$$(78)\quad ||\sqrt{N}(V_{\hat{\theta}_N}-V_{\theta^*})-\partial V_{\theta^*}/\partial\theta(\sqrt{N}(\hat{\theta}_N-\theta^*))||_\infty$$

$$\leq \sqrt{N}||\hat{\theta}_N-\theta^*|| \sup_{0<\alpha<1} ||\partial V_{\theta^*+\alpha(\hat{\theta}_N-\theta^*)}/\partial\theta-\partial V_{\theta^*}/\partial\theta||_\infty$$

Since $||\sqrt{N}(\hat{\theta}_N-\theta^*)||$ is $O_p(1)$, continuity of $\partial V_\theta/\partial\theta$ in θ implies that the right hand side of (78) converges to 0 with probability 1. Thus, the asymptotic distribution of $\sqrt{N}(V_{\hat{\theta}_N}-V_{\theta^*})$ equals the asymptotic distribution of $\sqrt{N}\ \partial V_{\theta^*}/\partial\theta(\hat{\theta}_N-\theta^*)$. By the Banach-space version of Prohorov's Theorem (Araujo and Giné (1980)) we have that $\sqrt{N}\ \partial V_{\theta^*}/\partial\theta(\hat{\theta}_N-\theta^*)$ converges weakly to a B-valued random element iff $L[\sqrt{N}\ \partial V_{\theta^*}/\partial\theta(\hat{\theta}_N-\theta^*)]$ converges weakly to a random variable for each $L\epsilon B^*$, provided the sequence of random elements $\{\sqrt{N}\ \partial V_{\theta^*}/\partial\theta(\hat{\theta}_N-\theta^*)\}$ is flatly concentrated, (i.e., has all its probability mass concentrated on a finite dimensional subspace of B). Since $L \equiv L[\partial V_{\theta^*}/\partial\theta]\epsilon L(R^{\dim(\theta)},R)$, it follows that $L[\sqrt{N}\ \partial V_{\theta^*}/\partial\theta(\hat{\theta}_N-\theta^*)] \xrightarrow{D} N(0,L\Sigma L')$, for every $L \epsilon B^*$. By the definition of a Gaussian random element, it follows immediately that $\sqrt{N}\ \partial V_{\theta^*}/\partial\theta(\hat{\theta}_N-\theta^*)$ converges weakly to a Gaussian random element on B with mean 0 and covariance operator $[\partial V_{\theta^*}/\partial\theta][\Sigma]\partial[V_{\theta^*}/\partial\theta']\epsilon L(B,L(B,B))$.

Q.E.D.

We conclude this section by noting that our rather extensive list of assumptions does not lead to a vacuous class of estimators; the extreme value distribution given in (20) satisfies all the required conditions for consistency and asymptotic normality. Thus, if, the functions u and p satisfy their own regularity conditions (A2), (A3), (A4), (A10), (A13), (A14), (A19), (A26) and (A27), then $Q(\epsilon|x,\theta_2)$ given by (24) satisfies all the remaining conditions necessary to prove consistency and asymptotic normality of the maximum likelihood estimator. These latter conditions, given by assumptions (A7), (A11), (A12), (A15), (A16), (A17), (A18), (A21), (A24), (A28), (A29), and (A30), can be easily verified using the following special property of the extreme-value distribution (24)

$$(77)\qquad G_{ij}([u(x,\theta_1)+\beta EV_\theta(x)]|x,\theta_2) = \begin{cases} P(i|x,\theta)[1-P(i|x,\theta)] & \text{if } i=j \\ -P(i|x,\theta)P(j|x,\theta) & \text{if } i\neq j \end{cases}$$

This property also enables us to obtain very simple formulas for the derivatives of $\log P(i|x,\theta)$.

$$(78) \qquad \partial \log P(i|x,\theta)/\partial\theta_1 = \partial r(\theta,x,i)/\partial\theta_1 - \sum_{j\epsilon C(x)} \partial r(\theta,x,j)/\partial\theta_1 P(j|x,\theta)$$

$$(79) \qquad \partial \log P(i|x,\theta)/\partial\theta_3 = \partial r(\theta,x,i)/\partial\theta_3 - \sum_{j\epsilon C(x)} \partial r(\theta,x,j)/\partial\theta_3 P(j|x,\theta)$$

$$(80) \qquad \partial \log P(i|x,\theta)/\partial\beta = \partial r(\theta,x,i)/\partial\beta - \sum_{j\epsilon C(x)} \partial r(\theta,x,j)/\partial\beta P(j|x,\theta)$$

From formulas (78), (79), (80) we can see directly that $E\{\partial \log P(i|x,\theta)/\partial\theta\} = 0$. The simplicity yielded by the extreme value distribution makes it a natural starting point for empirical implementation.

BIBLIOGRAPHY

Araujo, A., Giné, E. (1980), The Central Limit Theorem for Real and Banach Valued Random Variables, Wiley.

Basawa, I.V., Prakasa Rao, B.L.S. (1980), Statistical Inference for Stochastic Processes, Academic Press.

Berndt, E., Hall, B., Hall, R., Hausman, T. (1974), "Estimation and Inference in Nonlinear Structural Models," Annals of Economic and Social Measurement 3/4, 653-665.

Bertsekas, D. (1976), Dynamic Programming and Stochastic Control, Academic Press.

Bertsekas. D., Shreve, S. (1978), Stochastic Optimal Control: the Discrete Time Case, Academic Press.

Bhattacharya, R. N., Majumdar, M. (1984), "Dynamic Programming for Discounted and Long-run Average Rewards," SSRI Working Paper 8416.

Billingsley, P. (1961), Statistical Inference for Markov Processes, University of Chicago Press.

Billingsley, P. (1979), Probability and Measure, Wiley.

Blackwell, D. (1968), "Discounted Dynamic Programming," Annals of Mathematical Statistics, 226-235.

Bokar, V., Varaiya, P. (1982), "Identification and Adaptive Control of Markov Chains" SIAM Journal of Control, 10-4, 470-495.

Chamberlain, G. (1984), "Panel Data" in Handbook of Econometrics, ed. by Zvi Griliches, North Holland.

Coleman, T. S. (1983), "A Dynamic Model of Labor Supply Under Uncertainty," manuscript, University of Chicago.

Cosslett, S. (1983), "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model," Econometrica, 51, 765-782.

Cox, D. R. (1975), "Partial Likelihood," Biometrika 62-2, 269-276.

Daly, A., Zachary, S. (1979), Improved Multiple Choice Models in D. Henscher, Q. Dalvi (eds.), Identifying and Measuring the Determinants of Mode Choice, Teakfield, London.

Denardo, E. (1967), "Contraction Mapping in the Theory Underlying Dynamic Programming," SIAM Review, 165-177.

Domencich, T. A., McFadden, D. (1975), Urban Travel Demand, North-Holland.

Dunford, N., Schwartz, J. (1957), Linear Operators, Wiley.

Flinn, C., Heckman, J. (1982), "New Methods for Analyzing Structural Models of Labor Force Dynamics," Journal of Econometrics, 18, 115-168.

Garber, P. M., King, R. G. (1983), "Deep Structural Excavation? A Critique of Euler Equation Methods," manuscript.

Gihman, I. I., Skorohod, A. V. (1979), Controlled Stochastic Processes, Springer-Verlag.

Gill, P., Murray, W., Wright, M. (1981), Practical Optimization, Academic Press.

Grenander, U. (1950), "Stochastic Processes and Statistical Inference" Ark. Mat. 1, 195-277.

Grenander, U. (1981), Abstract Inference, Wiley.

Guilleman, V., Pollock, A. (1974), Differential Topology, Prentice Hall.

Futia, C. (1982), "Invariant Distributions and the Limiting Behavior of Markovian Economic Models" Econometrica, 50-4, 1029-1054,

Hansen, L. (1982), "Large Sample Properties of Generalized Method of Moment Estimators," Econometrica, 50-4, 1029-1054.

Hansen, L., Singleton, K. (1982), "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," Econometrica, 50-5, 1269-1286.

Hausman, J. (1978), "Specification Tests in Econometrics" Econometrica, 46, 1251-1272.

Heckman, J. (1981), "Statistical Models for Discrete Panel Data," in Structural Analysis of Discrete Data, ed. by C. Manski, and D. McFadden, M.I.T. Press.

Heckman, J. (1981), "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete-Time, Discrete-Data Stochastic Processes," in Structural Analysis of Discrete Data, ed. by C. Manski, and D. McFadden, M.I.T. Press.

Howard, R. (1971), Dynamic Probabilistic Systems Volume I: Markov Models Wiley.

Huber, P. (1967), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions" in Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability. Berkeley, University of California Press.

Kantorovich, L., Akilov, G. (1982), Functional Analysis, Pergamon Press.

Karlin, S., Taylor, H. (1975), A First Course in Stochastic Processes, Academic Press.

Lippman, S. (1975), "On Dynamic Programming with Unbounded Rewards," Management Science, 1225-1233.

Lucas, R. E. (1976), "Econometric Policy Evaluation: A Critique" in Brunner, K., Meltzer, A. K. (eds.) The Phillips Curve and Labor Markets, Carnegie-Rochester Conference on Public Policy 1, North Holland.

Mannering, F., Winston, C. (1982), "Dynamic Models of Household Vechicle Ownership and Utilization: An Empirical Analysis," M.I.T. Working Paper.

Manski, C. (1982), "Closest Empirical Distribution Estimation," Econometrica, 51, 305-320.

Manski, C. (1984), "Recent Work on Estimation of Econometric Models Under Weak Assumptions," Les Cahiers Du Seminaire d'Econométrie 26, July 1984.

Manski, C. (1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice," Journal of Econometrics 3, 205-228.

Manski, C., McFadden, D. (1981), Structural Analysis of Discrete Data with Econometric Applications, M.I.T. Press.

Marschak, J. (1953), "Economic Measurements for Policy and Prediction" in Hood, W. C., Koopmans, T. C. (eds.) Studies in Econometric Method, Wiley.

McFadden, D. (1981), "Econometric Models of Probabilistic Choice," in Structural Analysis of Discrete Data, edited by C. Manski and D. McFadden, M.I.T. Press.

McFadden, D. (1982), "Econometric Analysis of Qualitative Response Models," manuscript, M.I.T.

McFadden, D. (1973), "Conditional Logit Analysis of Qualitative Choice Behavior," in Frontiers of Econometrics, ed. by P. Zarembka, Academic Press.

McFadden, D., Newey, W. (1983), "Asymptotic Properties of Nonlinear Estimators," 14:383 Lecture Notes, M.I.T.

Miller, R. (1984), "Job Matching and Occupational Choice," Journal of Political Economy 92-6, 1086-1120.

Ortega, J., Rheinboldt, W. (1970), Iterative Solution of Nonlinear Equations in Several Variables, Academic Press.

Pakes, A. (1985), "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," forthcoming, Econometrica.

Pratt, J. (1960), "On Interchanging Limits and Integrals," Annals of Mathematical Statistics, 74-77.

Rust, J. (1986a), "Nested Fixed Point Optimization Algorithms," manuscript.

Rust, J. (1986b), "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," manuscript.

Sargent, T. J. (1981), "Interpreting Economic Time Series," Journal of Political Economy 89-2, 213-248.

Sargent, T. J. (1978), "Estimation of Dynamic Labor Demand Schedules Under Rational Expectations" 86-6, 1009-1044.

Taylor, R. L. (1978), Stochastic Convergence of Weighted Sums in Linear Spaces, Springer-Verlag.

Vakhania, N. N. (1981), Probability Distributions on Linear Spaces, North Holland.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," Econometrica, 50-1, 1-26.

Whittle, P. (1982), Optimization Over Time: Dynamic Programming and Stochastic Control, Wiley.

Williams, H. (1977), "On the Formation of Travel Demand Models and Economic Evaluation Measures of User Benefit," Environment Planning A-9, 285-344.

Wolpin, K. (1984), "An Estimable Dynamic Stochastic Model of Fertility and Child Mortality," Journal of Political Economy, 92-6, 1086-1120.

Appendix 1. Second Derivatives of $T_\theta(EV)$.

In this appendix we derive formulas for $\partial^2[T_\theta(EV)]/\partial\theta\partial\theta'$ which are used to verify that EV_θ is twice continuously differentiable in θ, and are also used to compute the hessian and asymptotic covariance matrix of the maximum likelihood estimator. Throughout we define $r(\theta,y) \equiv \{r(\theta,y,j) | j\epsilon C(y)\}$ with $r(\theta,y,j) \equiv [u(y,j,\theta_1) + \beta EV_\theta(y,j)]$. $G_{ij}(r(\theta,x)|x,\theta_2)$ denotes the second mixed partial derivative of $G(r(\theta,x)|x,\theta_2)$ with respect to $r(\theta,x,i)$ and $r(\theta,x,j)$. By the results of McFadden (1981), $G_{ij}(r(\theta,x)|x,\theta_2)$ is continuous in $r(\theta,x)$ for each x and θ_2. Assumptions (A15), (A17) and (A20) and the dominated convergence theorem justify interchanging differentiation and integration to obtain the following formulas for $\partial^2 EV_\theta/\partial\theta\partial\theta'$.

$$(1)\quad \partial^2 T_\theta(EV_\theta)/\partial\theta_1\partial\theta_1'(x,i) = \int_y \{\sum_{j\epsilon C(y)} \partial^2 u(y,j,\theta_1)/\partial\theta_1\partial\theta_1' P(j|y,\theta)\}p(dy|x,i,\theta_3)$$

$$+ \int_y \{\sum_{j,k\epsilon C(y)} \partial u(y,j,\theta_1)/\partial\theta_1\ \partial r(\theta,y,k)/\partial\theta_1' G_{jk}(r(\theta,y)|y,\theta_2)\}p(dy|x,i,\theta_3)$$

$$(2)\quad \partial^2 T_\theta(EV_\theta)/\partial\theta_1\partial\theta_2'(x,i) = \int_y \{\sum_{j\epsilon C(y)} \partial u(y,j,\theta_1)/\partial\theta_1\partial G_j(r(\theta,y)|y,\theta_2)/\partial\theta_2'\}p(dy|x,i,\theta_3)$$

$$+ \int_y \{\sum_{j,k\epsilon C(y)} \partial u(y,j,\theta_1)/\partial\theta_1\partial r(\theta,y,k)/\partial\theta_2' G_{jk}(r(\theta,y)|y,\theta_2)\}p(dy|x,i,\theta_3)$$

$$(3)\quad \partial^2 T_\theta(EV_\theta)/\partial\theta_1\partial\theta_3'(x,i) =$$

$$\int_y \{\sum\sum_{j,k\epsilon C(y)} \partial u(y,j,\theta_1)/\partial\theta_1\partial r(\theta,y,k)/\partial\theta_3' G_{jk}(r(\theta,y)|y,\theta_2)\}p(dy|x,i,\theta_3)$$

$$\partial[\int_y \{\sum_{j\epsilon C(y)} \partial u(y,j,\theta_1)/\partial\theta_1 P(j|y,\theta)\}p(dy|x,i,\theta_3)]/\partial\theta_3'$$

$$(4)\quad \partial^2 T_\theta(EV_\theta)/\partial\theta_1\partial\beta(x,i) =$$

$$\int_y \{ \sum_{j,k\epsilon C(y)} \sum \partial u(y,j,\theta_1)/\partial\theta_1 \; \partial r(\theta,y,k)/\partial\beta G_{jk}(r(\theta,y)|y,\theta)\} p(dy|x,i,\theta_3)$$

$$(5)\quad \partial^2 T_\theta(EV_\theta)/\partial\theta_2\partial\theta_2'(x,i) = \int_y \{ \sum_{j\epsilon C(y)} \partial r(\theta,y,j)/\partial\theta_2\partial G_j(r(\theta,y)|y,\theta_2)/\partial\theta_2')\} p(dy|x,i,\theta_3)$$

$$+ \int_y \{\partial^2 G(r(\theta,y)|y,\theta_2)/\partial\theta_2\partial\theta_2'\} p(dy|x,i,\theta_3)$$

$$(6)\quad \partial^2 T_\theta(EV_\theta)/\partial\theta_2\partial\theta_3'(x,i) = \int_y \{ \sum_{j\epsilon C(y)} \partial r(\theta,y,j)/\partial\theta_3'\partial G_j(r(\theta,y)|y,\theta_2)/\partial\theta_2\} p(dy|x,i,\theta_3)$$

$$+ \partial[\int_y \{\partial G(r(\theta,y)|y,\theta_2)/\partial\theta_2\} p(dy|x,i,\theta_3)]/\partial\theta_3'$$

$$(7)\quad \partial^2 T_\theta(EV_\theta)/\partial\theta_2\partial\beta(x,i) = \int_y \{ \sum_{j\epsilon C(y)} \partial r(\theta,y,j)/\partial\beta \; \partial G_j(r(\theta,y)|y,\theta_2)/\partial\theta_2\} p(dy|x,i,\theta_3)$$

$$(8)\quad \partial^2 T_\theta(EV_\theta)/\partial\theta_3\partial\theta_3'(x,i) = \partial^2[\int_y G(r(\theta,y)|y,\theta_2) p(dy|x,i,\theta_3)]/\partial\theta_3\partial\theta_3'$$

$$+ \partial[\int_y \{ \sum_{j\epsilon C(y)} \partial r(\theta,y,j)/\partial\theta_3 P(j|y,\theta)\} p(dy|x,i,\theta_3)]/\partial\theta_3'$$

$$(9)\quad \partial^2 T_\theta(EV_\theta)/\partial\theta_3\partial\beta(x,i) = \partial [\int_y \{ \sum_{j\epsilon C(y)} \partial r(\theta,y,j)/\partial\beta P(j|y,\theta)\} p(dy|x,i,\theta_3)]/\partial\theta_3$$

$$(10)\quad \partial^2 T_\theta(EV_\theta)/\partial\beta\partial\beta(x,i) = \int_y \{ \sum_{j\epsilon C(y)} \partial EV_\theta(y,j)/\partial\beta \; P(j|y,\theta)\} p(dy|x,i,\theta_3)$$

$$+ \int_y \{ \sum_{j,k\epsilon C(y)} \sum [EV_\theta(y,j)\partial r(\theta,y,k)/\partial\beta] G_{jk}(r(\theta,y)|y,\theta_2)\} p(dy|x,i,\theta_3)$$

Appendix 2. Second Derivatives of $L^f(\theta)$

In this appendix we derive formulas for $\partial^2 \log L^f(\theta)/\partial\theta\partial\theta'$. In order to simplify notation, we simply compute $\partial^2[\log\{P(i_t|x_t,\theta)P(x_t|x_{t-1},i_{t-1},\theta_3)\}/\partial\theta\partial\theta'$, since the second derivative of $\log L^f(\theta)$ is simply the sum of these terms, i.e.

$$(1)\quad \partial^2[\log L^f(\theta)]/\partial\theta\partial\theta' = \sum_{t=1}^{T} \partial^2[\log\{P(i_t|x_t,\theta)p(x_t|x_{t-1},i_{t-1},\theta_3)]/\partial\theta\partial\theta'$$

Furthermore, we omit the "t" subscript on the x and i variables to further simplify notation. It is important to note that the zero mean martingale property of the score function (Theorem 13) implies that the only terms surviving after taking expectations of these second derivatives are the expectations of the bracketed terms containing the product of the scores. As in appendix 1, define $r(\theta,x,j) \equiv [u(x,j,\theta_1) + \beta EV_\theta(x,j)]$.

$$(2)\quad \partial^2 \log P(i|x,\theta)/\partial\theta_1\partial\theta_1' = -[\partial \log P(i|x,\theta)/\partial\theta_1][\partial \log P(i|x,\theta)/\partial\theta_1']$$

$$+ \sum_{j\in C(x)} \partial^2 r(\theta,x,j)/\partial\theta_1\partial\theta_1' G_{ij}(r(\theta,x)|x,\theta_2)/P(i|x,\theta)$$

$$+ \sum_{j,k\in C(x)} \sum \partial r(\theta,x,j)/\partial\theta_1 \; \partial r(\theta,x,k)/\partial\theta_1' G_{ijk}(r(\theta,x)|x,\theta_2)/P(i|x,\theta)$$

$$(3)\quad \partial^2 \log P(i|x,\theta)/\partial\theta_1\partial\theta_2' = -[\partial \log P(i|x,\theta)/\partial\theta_1][\partial \log P(i|x,\theta)/\partial\theta_2']$$

$$+ \sum_{j\in C(x)} [\partial^2 r(\theta,x,j)/\partial\theta_1\partial\theta_2']/P(i|x,\theta)$$

$$+ \sum_{j,k\in C(x)} \sum \partial r(\theta,x,j)/\partial\theta_1 \partial r(\theta,x,k)/\partial\theta_2' \partial G_{ijk}(r(\theta,x)|x,\theta_2)/P(i|x,\theta)$$

$$+ \sum_{j\in C(x)} [\partial r(\theta,x,j)/\partial\theta_1 \partial G_{ij}(r(\theta,x)|x,\theta_2)/\partial\theta_2']/P(i|x,\theta)$$

$$(4)\quad \partial^2 \log\{P(i|x,\theta)\}/\partial\theta_1\partial\theta_3' = -[\partial \log P(i|x,\theta)/\partial\theta_1][\partial \log P(i|x,\theta)/\partial\theta_3']$$

$$+ \sum_{j\in C(x)} \partial^2 r(\theta,x,j)/\partial\theta_1\partial\theta_3' G_{ij}(r(\theta,x)|x,\theta_2)/P(i|x,\theta)$$

$$+ \sum_{j,k\in C(x)} \sum \partial r(\theta,x,j)/\partial\theta_1 \; \partial r(\theta,x,k)/\partial\theta_3' G_{ijk}(r(\theta,x)|x,\theta_2)/P(i|x,\theta)$$

$$(5)\quad \partial^2 \log P(i|x,\theta)/\partial\theta_1\partial\beta = -[\partial \log P(i|x,\theta)/\partial\theta_1][\partial \log P(i|x,\theta)/\partial\beta]$$

$$+ \sum_{j\in C(x)} \partial^2 r(\theta,x,j)/\partial\theta_1\partial\beta \; G_{ij}(r(\theta,x)|x,\theta_2)/P(i|x,\theta)$$

$$+ \sum_{j,k\in C(x)} \sum \partial r(\theta,x,j)/\partial\theta_1 \; \partial r(\theta,x,k)/\partial\beta \; G_{ijk}(r(\theta,x)|x,\theta_2)/P(i|x,\theta)$$

$$(6)\quad \partial^2 \log P(i|x,\theta)/\partial\theta_2\partial\theta_2' = -[\partial \log P(i|x,\theta)/\partial\theta_2][\partial \log P(i|x,\theta)/\partial\theta_2']$$

$$+ \sum_{j\in C(x)} \partial^2 r(\theta,x,j)/\partial\theta_2\partial\theta_2' G_{ij}(r(\theta,x)|x,\theta_2)/P(i|x,\theta)$$

$$+ \sum_{j,k\in C(x)} \sum \partial r(\theta,x,j)/\partial\theta_1 \partial r(\theta,x,k)/\partial\theta_2' \; G_{ijk}(r(\theta,x)|x,\theta_2)/P(i|x,\theta)$$

$$+ [\partial^2 G_i(r(\theta,x)|x,\theta_2)/\partial\theta_2\partial\theta_2']/P(i|x,\theta)$$

$$(7)\quad \partial^2 \log P(i|x,\theta)/\partial\theta_2\partial\theta_3' = -[\partial \log P(i|x,\theta)/\partial\theta_2]\;[\partial \log P(i|x,\theta)/\partial\theta_3']$$

$$+ \sum_{j\in C(x)} \partial^2 r(\theta,x,j)/\partial\theta_2\partial\theta_3' G_{ij}(r(\theta,x)|x,\theta_2)/P(i|x,\theta)$$

$$+ \sum_{j,k\in C(x)} \sum \partial r(\theta,x,j)/\partial\theta_2 \partial r(\theta,x,k)/\partial\theta_3' G_{ijk}(r(\theta,x)|x,\theta_2)/P(i|x,\theta)$$

$$+ \sum_{j\in C(x)} [\partial G_{ij}(r(\theta,x)|x,\theta_2)/\partial\theta_2 \; \partial r(\theta,x,j)/\partial\theta_3']/P(i|x,\theta)$$

$$(8)\quad \partial^2 \log P(i|x,\theta)/\partial\theta_2\partial\beta = -\,[\partial \log P(i|x,\theta)/\partial\theta_2][\partial \log P(i|x,\theta)/\partial\beta]$$

$$+\sum_{j\in C(x)} \partial^2 r(\theta,x,j)/\partial\theta_2\partial\beta\; G_{ij}(r(\theta,x)|x,\theta_2)/P(i|x,\theta)$$

$$+\sum_{j,k\in C(x)} \partial r(\theta,x,j)/\partial\theta_2\; \partial r(\theta,x,k)/\partial\beta\; G_{ijk}(r(\theta,x)|x,\theta_2)/P(i|x,\theta)$$

$$+\sum_{j\in C(x)} [\partial G_{ij}(r(\theta,x)|x,\theta_2)/\partial\theta_2\partial r(\theta,x,j)/\partial\beta]/P(i|x,\theta)$$

$$(9)\quad \partial^2 \log\{P(i|x,\theta)p(x|y,\ell,\theta_3)\}/\partial\theta_3\partial\theta_3' = -\,[\partial \log P(i|x,\theta)/\partial\theta_3][\partial \log P(i|x,\theta)/\partial\theta_3']$$

$$+\sum_{j\in C(x)} \partial^2 r(\theta,x,j)/\partial\theta_3\partial\theta_3' G_{ij}(r(\theta,x)|x,\theta_2)/P(i|x,\theta)$$

$$+\sum_{j,k\in C(x)} \partial r(\theta,x,j)/\partial\theta_3 \partial r(\theta,x,k)/\partial\theta_3'\; G_{ijk}(r(\theta,x)|x,\theta_2)/P(i|x,\theta)$$

$$+\,\partial^2 \log p(x|y,\ell,\theta_3)/\partial\theta_3\partial\theta_3'$$

$$(10)\quad \partial^2 \log P(i|x,\theta)/\partial\theta_3\partial\beta = -\,[\partial \log P(i|x,\theta)/\partial\theta_3]\;[\partial \log P(i|x,\theta)/\partial\beta]$$

$$+\sum_{j\in C(x)} \partial^2 r(\theta,x,j)/\partial\theta_3\partial\beta G_{ij}(r(\theta,x)|x,\theta_2)/P(i|x,\theta)$$

$$+\sum_{j,k\in C(x)} \partial r(\theta,x,j)/\partial\theta_3\; \partial r(\theta,x,k)/\partial\beta\; G_{ijk}(r(\theta,x)|x,\theta_2)/P(i|x,\theta)$$

$$(11)\quad \partial^2 \log P(i|x,\theta)/\partial\beta\partial\beta = -\,[\partial \log P(i|x,\theta)/\partial\beta]^2$$

$$+\sum_{j\in C(x)} \partial^2 r(\theta,x,j)/\partial\beta\partial\beta\; G_{ij}(r(\theta,x)|x,\theta_2)/P(i|x,\theta)$$

$$+\sum_{j,k\in C(x)} \partial r(\theta,x,j)/\partial\beta\; \partial r(\theta,x,k)/\partial\beta\; G_{ijk}(r(\theta,x)|x,\theta_2)/P(i|x,\theta)$$

Appendix 3. Existence of an Invariant Distribution Ω.

In this section we provide sufficient conditions for the existence of a unique invariant distribution Ω which is the solution to the functional equation

$$(1) \qquad \Omega(x_o,i_o,\theta) = \int_y \sum_{i\in C(y)} P(i_o|x_o,\theta)p(x_o|y,i,\theta_3)\Omega(dy,di,\theta)$$

where $\Omega(x_o,i_o,\theta)$ is interpreted as an invariant probability density corresponding to the joint controlled stochastic process $\{x_t,i_t\}$. Recall the following notation.

Δ = A Borel subset of R^K.

Γ = $\{(x,i)|x\in\Delta,\ i\in C(x)\}$

B = Banach space of all measurable, real valued functions on Γ under norm $||\cdot||_\infty$

B_c = closed subspace of B consisting of all real valued continuous functions on Γ.

A linear operator E_θ on B is defined by

$$(2) \qquad E_\theta g(x,i) = \int\{\sum_{y\ i\in C(y)} g(y,i)P(i|y,\theta)\}p(dy|x,i,\theta_3)$$

It is obvious that E_θ is a continuous linear operator with norm equal to 1. The following assumptions guarantee the existence of an invariant distribution Ω.

(A32) For all $\theta\in\Theta$, E_θ is a linear operator on B_c, i.e., $E_\theta: B_c \to B_c$.

(A33) For all $\theta\in\Theta$ and for each $g\in B_c$, the sequence $\frac{1}{T}\sum_{t=0}^{T-1} E_\theta^t g$ converges to an element of B_c as $T \to \infty$.

(A34) For all $\theta\epsilon\Theta$, for all $\epsilon > 0$, there is a compact set $F_\epsilon \subseteq \Delta$ such that for all $(x_o,i_o)\epsilon\Gamma$ we have

(3) $$\int_{F_\epsilon} p(dy|x_o,i_o,\theta_3) \geq 1-\epsilon.$$

Theorem Under assumptions (A32),...,(A34) there exists at least one invariant distribution Ω which is a solution to equation (1). Furthermore, given any initial density $\lambda(x_o,i_o)$ the sequence $\frac{1}{T}\sum_{t=o}^{T-1}(E_\theta^*)^t\lambda$ converges to a unique invariant density where the adjoint operator E_θ^* is given by

(4) $$E_\theta^*\lambda(x,i) = \int_y \sum_{i\epsilon C(y)} P(i|x,\theta)p(x|y,j,\theta_3)\lambda(dy,dj)$$

For the proof of this Theorem, see Theorems 2.9 and 2.10 of Futia (1982). The following condition guarantees uniqueness of the invariant distribution Ω, and is the analog of the requirement in finite state markov chains that all states "communicate".

(A35) For all $\theta\epsilon\Theta$, there is a point $x_o\epsilon\Delta$ such that for every neighborhood N_{x_o} and for all $x\epsilon\Delta$, $i\epsilon C(x)$, we have

(5) $$\int_{N_{x_o}} p(dy|x,i,\theta_3) > 0.$$

Theorem Under assumptions (A32),...,(A35) there exists a unique invariant distribution Ω which is the solution to (1).

Proof. Assumption (A35) implies that the transition probability $P(i|y,\theta)p(y|x,i,\theta_3)$ satisfies the "uniqueness criterion" given in Futia (1982), so that Theorem 2.12 of that paper guarantees the existence of a unique invariant distribution Ω.