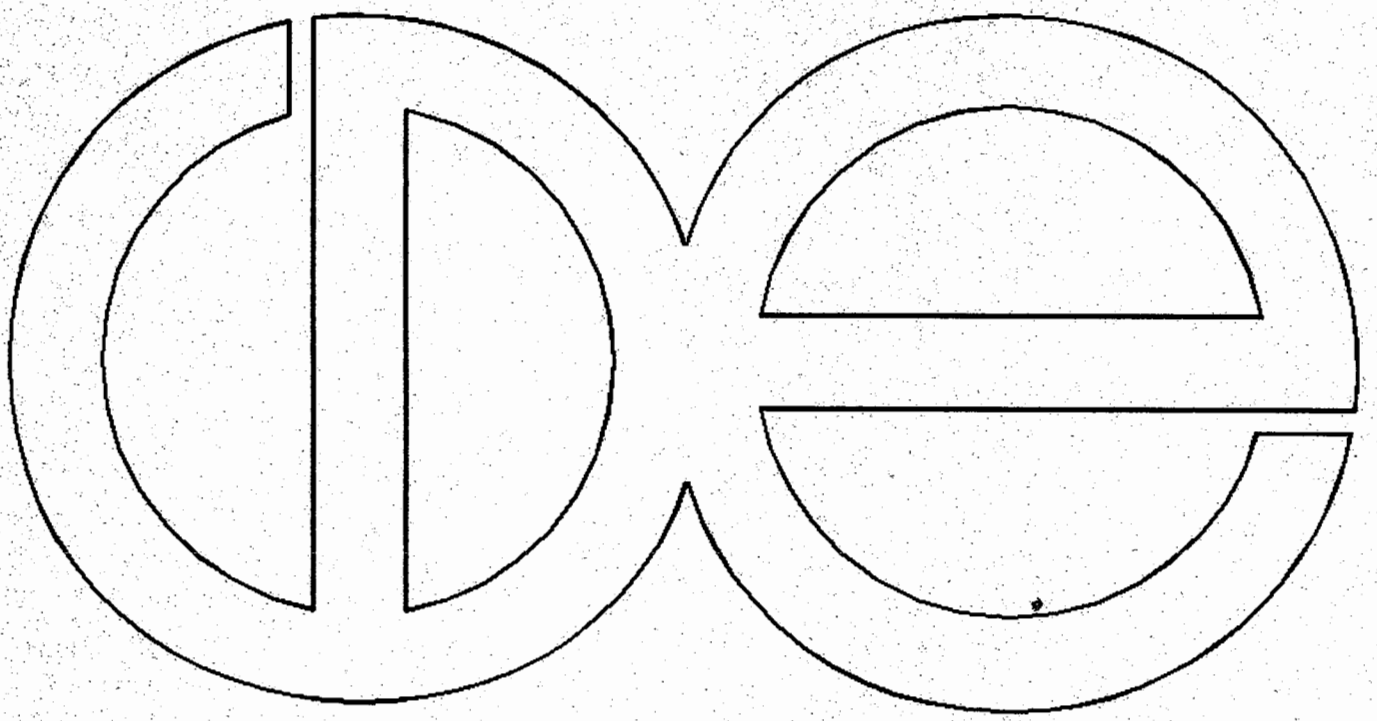
Center for Demography and Ecology

University of Wisconsin-Madison

ESTIMATING TWO-SIDED LOGIT MODELS

John Allen Logan

CDE Working Paper No. 96-07



Estimating Two-Sided Logit Models

Revised
July 31, 1996

John Allen Logan
Department of Scoiology
University of Wisconsin
Madison, Wisconsin 53706
logan@ssc.wisc.edu
phone: (608) 262-0995
fax: (608) 265-5389

Revised and expanded version of a paper presented at the World Congress of Sociology, July 1994, Bielefeld, Germany.

Estimating Two-Sided Logit Models

Abstract

Logan (1996a) introduced both a new micro-behavioral model of employment opportunity and choice, and a multivariate statistical method based on the micro-behavioral model. This article considers the connection between the behavioral model and the statistical method, two-sided logit (TSL), in more detail than the original article, discussing issues of parameter identification, model constraints, data reduction, and practical estimation contingencies. The article also presents the EM algorithm used previously and an accelerated EM algorithm (Jamshidian and Jennrich 1993) which turns out to be seven times faster on a sample problem. Strategies for further technical development of TSL methods are also considered.

Estimating Two-Sided Logit Models

Introduction

Logan (1996a) introduced a new, multivariate method for studying the determinants of labor market outcomes, the two-sided logit or TSL model. Unlike many standard sociological methods, the TSL model is based on an explicit, behavioral model. This behavioral model represents labor market outcomes as the result of choices made by workers and employers, all of which choices are constrained by other choices made by other workers and other employers. Logan (1996b) showed that the TSL model is closely related to formal, two-sided matching games, while Logan (1996c) compared TSL and multivariate log-linear models, demonstrating that the latter have important relative disadvantages when applied to labor market outcomes, as in occupational mobility models.

Estimating TSL models requires more care than many familiar techniques. This article examines the estimation of TSL models in greater detail than earlier articles. Issues of identification, estimation algorithms, and the uniqueness of the estimates obtained from algorithms are addressed.

Identification of the model's behavioral parameters depends upon the type of data available to the researcher; some behavioral parameters of the model as described in Logan (1996a) would be identifiable only in very rich data sets, although the most important parameters can be identified in very ordinary data sets, such as the General Social Survey (GSS). Identification also depends on a series of decisions

regarding details of specification and constraint which are driven primarily by data limitations.

Two practical estimation algorithms will be described for the TSL model, an EM algorithm and an accelerated EM algorithm (AEM) which improves on the basic EM performance by a factor of 7 in the estimations to be presented. Whether other methods might also be practical for this problem will also be considered.

Finally, the possibility of obtaining non-global, local maxima with either of the two algorithms will be discussed. Strategies to reduce the possibility of accepting less than optimal solutions include ascertaining that any obtained set is at least a local maximum, exploring different starting values, and insuring that the absolute fit of the model to the data is acceptable.

The next section of the paper motivates the subsequent development by describing the behavioral model first presented in Logan (1996a).

The TSL Behavioral Model

Considered as a voluntary outcome, obtaining a job has two necessary components: a willing employer and a willing worker. The TSL behavioral model is made up of separate, micro-level models of these two components. One such submodel describes the decisions of employers whether to offer jobs to workers, while the other submodel describes the decisions of workers whether to accept offered jobs.

The model of the employer's decision is comprised of two linear utility functions. For a given employer, j, the utility of hiring a worker, i, is defined as

$$U_j(i) = \boldsymbol{\beta}_j^* \mathbf{x}_i^* + m_j + \varepsilon_{1ij} \ , \tag{1}$$

while the utility of not hiring i is

$$U_j(\neg i) = b_j + s_j + \varepsilon_{0ij} \ . \tag{2}$$

Here $\boldsymbol{\beta}_j^*$ is a row vector of employer j's preferences for relevant characteristics of employees, and $\mathbf{x}_i^*$ is a column vector of i's measured values on those characteristics. Three scalar values in (1) and (2) are introduced to represent influences on the employer's utilities which are not dependent on the characteristics of individual i: m_j represents market effects on the utility of hires for the employer in general; b_j is the baseline utility the employer would experience without an additional hire; and s_j is a strategic increment to the threshold which the employer might set to achieve a better stable match. The terms ε_{1ij} and ε_{0ij} are random components representing any other factors which influence j's utilities for hiring i. The decision rule each employer follows is to evaluate each worker in turn, and to make an offer to each one for which $U_j(i) > U_j(\neg i)$.

On the other side of the market, a submodel representing the decisions of workers whether to accept particular jobs consists of a single utility function which is simultaneously evaluated for each job a worker finds available. For worker i

the utility of a job offered by employer j is:

$$V_i(j) = \alpha_i w_{ij} + \upsilon_{ij} \quad . \tag{3}$$

Here w_{ij} contains characteristics of the job offered by the employer j, α_i is a vector of worker i's preferences for those characteristics, and υ_{ij} is a random term representing additional influences on i's utility. The worker is also assumed to evaluate (3) for a non-employment (or "unemployment") alternative, represented by $j = 0$; in this evaluation, the characteristics in w_{i0} are those i would obtain without a job. The worker evaluates (3) for all available jobs, and for unemployment, and then selects the single alternative which gives the highest utility.

As discussed in Logan (1996b), equations (1) through (3), together with the above-mentioned rules governing job offering and acceptance behavior, describe the basic elements of a two-sided matching game between workers and employers. The particular game is a random variant of the "college admissions" game of the formal game theory literature (Roth and Sotomayor 1990). Because the deterministic game has been analyzed extensively, and because the deterministic results are transferable to the random matching game, it is known that at least one stable matching of employers and workers exists such that no worker-employer pair who are not matched to each other can improve their utilities by abandoning any current partners and establishing a new match together (Logan 1996b). It has also been shown that a simple, random process

of information flow is sufficient to move the overall system toward a set of stable matches (Roth and Vande Vate 1990; Logan 1996b).

Estimation of the parameters of the TSL model is done under the assumption that the matches observed in empirical data are stable matches, at current levels of demand and supply. It is important to note that this stability assumption does not presume that either employers or workers are good optimizers in the economic sense: their preferences for employment partners can be based on immanent as well as instrumental values (Hechter 1994), and can include any mix of norms, habits, traditions and value-rational motives, regardless of market values (Logan 1996b).

The parameters of primary interest in the TSL model are the α_i and β_j vectors of preference coefficients. In particular, the elements of β_j constitute *rules of access* to positions since they determine which individuals will be preferred to others as hiring decisions are made (Logan 1996c). Elements of α_i are also rules of access of a different sort, determining which workers can be hired by which employers. The scalars m_j, b_j, and s_j have a more heuristic role, describing other considerations going into the employers' decisions while not necessarily being the subjects of estimation.

Identification

Practical estimation of the TSL model's parameters

depends both on the nature of available data and on simplifying assumptions imposed on the basic model. Distributional assumptions, parametric constraints and prior data reduction are all important to estimation.

Distributional assumptions

A basic assumption concerns the distributions of the random terms in (1), (2) and (3). Each of these terms is intended to represent unknown influences on utilities, attributes which could in principle be measured but are not available to the researcher. Since it is reasonable to suppose that the unknown influences in any employer's utility of hiring could also be influences in the decisions of other employers, or perhaps even in the decisions of workers to accept jobs, a prima facie case exists for allowing disturbance correlations across employers and/or workers: such correlations should be induced by the presence of shared components in the disturbances.

However, there are arguments against including such correlations in the specification. The simply pragmatic argument is that such correlations would greatly increase the difficulty of estimation, and that it is sensible to begin with a simpler model. The issue has also been addressed in the one-sided, discrete choice context, where multinomial probit models, which allow intercorrelations of disturbances, are theoretically preferable to polytomous conditional logit models, which do not. Horowitz (1991) argues against the

multinomial probit's correlated errors and in favor of better measurement and specification of observable factors which applied research has shown to be important. Among other drawbacks, he points out that correlated error structures lead to problems of comparing or transferring coefficient estimates between data sets, which better specification of observables would obviate. For these and other reasons discussed by Horowitz, the first identifying assumption imposed on the TSL model is that the errors in (1), (2) and (3) are mutually independent.

Once mutual independence is assumed, it is appropriate to choose a distributional form for the disturbances. The assumption that many unobservables are represented in the disturbances suggests the normal distribution because of the central limit theorem. However, with independent disturbances it is well known that choosing the type I extreme value, or Gumbel, distribution leads to essentially equivalent results and easier estimation (Maddala 1983). Assuming then that the random components ε_{1ij} and ε_{0ij} of (1) and (2) have independent, standard Gumbel distributions, the probability that employer j will offer a job to individual i can be shown to be[1]

$$\Pr(o_{ij}=1) = \frac{\exp(\boldsymbol{\beta}_j \mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}_j \mathbf{x}_i)}, j>0, \qquad (4a)$$

where o_{ij} is a dummy variable indicating an offer is made, $\mathbf{x}_i = \left(1, \mathbf{x}_i^{*T}\right)^T$ is the original vector of individual characteristics augmented by an entry of unity in the first

position, and $\beta_j = (\beta_{j0}, \beta_j^*)$ is the original vector of employer j's preferences, augmented by an intercept parameter in its first position. This last term, called a demand intercept for short, is mathematically equal to the net effect of the j-subscripted scalars in (1) and (2):

$$\beta_{j0} = m_j - b_j - s_j \ .$$

Since unemployment is always available, the probability of an "offer" of unemployment, represented by $j = 0$, is set to 1, without regard to the worker's characteristics:

$$\Pr(o_{i0} = 1) = 1. \tag{4b}$$

Regarding the workers' submodel, assuming that the disturbances in (3) are independently distributed across job alternatives according to standard Gumbel distributions implies this probability of selecting a particular job j, given a set of offers O_k:

$$\Pr(A_{ij}|O_k) \begin{cases} = \dfrac{\exp(\alpha_i w_{ij})}{\sum_{h \in O_k} \exp(\alpha_i w_{ih})}, j \in O_k \\ = 0, j \notin O_k \end{cases} \tag{5}$$

This is the polytomous conditional logit model of the discrete choice literature, with the restriction that choice can occur only from among the offered jobs in the set.[2] For notational convenience, the set of offers is represented by the subscripts of the employers making the offers. Thus the set contains the numeral 1 if and only if employer 1 offers a job, and so on. To facilitate later notation, all offering sets also contain the numeral 0, representing the constant

availability of unemployment to all workers. There are $R = 2^J$ distinct, possible offering sets when J employers (not counting unemployment) make separate decisions.[3]

The implication of the assumed independence of disturbances across employers is that each acts independently in evaluating workers, conditional on the observed characteristics in $\mathbf{x}_i$. For this reason, the probability that worker i will obtain any given offering set O_k, which will be denoted as the event S_{ik}, is found from the multiplication rule for (conditionally) independent events, with reference to formula (4) above:

$$\Pr(S_{ik}) = \prod_{m \in O_k} \Pr(o_{im} = 1) \prod_{n \in \bar{O}_k} \Pr(o_{in} = 0) \quad . \tag{6}$$

Here $\bar{O}_k$ is the complement of the set O_k. Note that because the probabilities of offers depend on the personal characteristics of workers (in equation (4a)), workers will differ widely in their probabilities of receiving particular offering sets of greater or lesser desirability.

The independence of disturbances among formulas (1), (2), and (3) further implies that the probability that worker i will accept a job from employer j is:

$$\begin{aligned} \Pr(A_{ij}) &= \sum_{k=1}^{R} \Pr(A_{ij}|S_{ik}) \Pr(S_{ik}) \\ &= \sum_{k=1}^{R} \Pr(A_{ij}|S_{ik}) \prod_{m \in O_k} \Pr(o_{im} = 1) \prod_{n \in \bar{O}_k} \Pr(o_{in} = 0) \\ &= \sum_{k: j \in O_k} \frac{\exp(\boldsymbol{\alpha}_i \mathbf{w}_{ij})}{\sum_{h \in O_k} \exp(\boldsymbol{\alpha}_i \mathbf{w}_{ih})} \prod_{\substack{m \in O_k \\ m > 0}} \frac{\exp(\boldsymbol{\beta}_m \mathbf{x}_i)}{1 + \exp(\boldsymbol{\beta}_m \mathbf{x}_i)} \prod_{\substack{n \in \bar{O}_k \\ n > 0}} \frac{1}{1 + \exp(\boldsymbol{\beta}_n \mathbf{x}_i)} \end{aligned} \tag{7}$$

Equations (4) and (5) would form the foundation of a

practical estimation method if the availability of all jobs for each sample member could be observed. In such an ideal case, (4) could be used (with additional parametric restrictions) to estimate the preferences of employers, and (5) could be used (again with additional parametric restrictions) to estimate workers' preferences, in the knowledge of the opportunities each had found available. Because obtaining data on the actual availability of each of a large number of jobs in a labor market of any size is clearly impractical, equation (7), which integrates over the distribution of unobserved offers, must be the starting point for practical estimation. In principle (7) requires observations on the characteristics of all jobs in the system, but not on which offers have actually been made to which workers in the sample. The fact that this requirement is still excessive means that further constraints are needed.

Parametric constraints

Even assuming for the moment that the characteristics of all jobs in the system could be known, (7) would still not be directly estimable. An obvious difficulty is the presence of the subscript on α_i, which implies a different set of preference coefficients for each sampled worker. This difficulty can be remedied either by dropping the subscript altogether, implying an assumption of shared preferences across all sample members, or by estimating different vectors for different groups of workers, the so-called market

segmentation strategy (Ben-Akiva and Lerman 1985: 64). In the estimations below, the subscript is simply dropped, but the other alternative is just as practical.

A further problem with (7) is that the vector of job characteristics offered to i, that is $\mathbf{w}_{ij}$, is i-subscripted, implying that different characteristics are offered to different individuals by the same employer. Collecting data on such differences would be impractical for the same reason that collecting data on the presence of offers is impractical, whether or not differences might be important, so that this subscript must also be dropped. The implication is that employers have fixed characteristics of the jobs they offer, and exercise their judgment only in deciding which workers should get which offers.

Similarly to the problem implied by the subscript on α_i, the subscript on the $\boldsymbol{\beta}_j$ vectors means that each employer is assumed to have a unique structure of preference coefficients, which would be impossible to estimate without repeated observations on individual employers. Here either all employers can be specified as sharing the same preferences, by dropping the subscript, or employers of different types can be given shared preference vectors. The latter strategy will be adopted below, with employers of different types of workers having differentiated preferences.

Aside from these simplifications, it should also be remarked that the demand scalars of equations (1) and (2), namely, m_j, b_j, and s_j, no longer appear in (7). Only their

net effect, represented in the demand intercept β_{j0}, can be estimated from this formula. This limitation is not intrinsic to the model, but is an implicit restriction imposed on the parameterization. Measuring the factors which determine the levels of the three scalars is not possible with common data sets, so only the net effect of the scalars is retained. If appropriate data were actually available, the scalars could be reformulated as linear regression functions of the observables. The primary usefulness of distinguishing the three scalars in the absence of such data comes in considering an employer-optimal matching strategy which figures in the game-theoretic analysis of the model (Logan 1996b).

It is also noteworthy that the model contains no α intercept terms in either the utility function of workers (3) or the conditional logit submodel derived from it, (5). This is neither an oversight nor a drawback. Formulations like (3) are in principle complete descriptions of the factors affecting choice among alternatives, since all relevant characteristics of alternatives can in principle be listed in the $\mathbf{w}_{ij}$ vectors. Such formulations are called abstract mode models in the one-sided choice literature, and their desirable properties gave them a certain following in applied studies even in the face of evidence that introducing intercepts produced better fits (Amemiya 1981). In the TSL model, the effects of hypothetical j-specific intercepts in the α coefficient vector would not be easily distinguishable

from the existing j-specific intercepts in the β_j coefficient vectors, though there is in principle a difference in the form of the mathematical effects on the likelihood. When the model is estimated as specified, without any j-specific α intercepts, the β_j intercepts will therefore tend to pick up whatever α intercept effects may be present. This would be problematic if there were substantive interest in the magnitudes of the β_j intercepts, but as explained above, these demand intercepts represent the net influence of three separate scalar effects, and therefore their magnitudes are of no particular interest. The role of the demand intercepts from an estimation point of view is simply to insulate the preference coefficient estimates from demand effects, so it is of no concern that they may also be insulating the coefficient estimates from residual differences in the attractiveness of job types, the role which α intercepts would normally play.

All of the constraints discussed in this subsection are broadly consistent with the aims of sociological research. As a generalizing science, sociology is properly interested in classes of workers and employers, which justifies the treatments of α and β as shared preferences characterizing such classes. Dropping the i subscript from w_{ij} is consistent with a characteristically sociological, as opposed to economic, position that many jobs have relatively fixed characteristics. Estimating only the net effects of the demand scalars, in the form of unconstrained demand

intercepts, is a way of regarding levels of demand as exogenous, rather than attempting to explain them within the model; this frees the model from a problem which would otherwise be extremely difficult to solve. Unfortunately there is a remaining problem related to data availability which requires a solution not really in accord with sociological goals, though perhaps not greatly in conflict with them.

Data reduction

The final problem standing in the way of estimation has to do with the types of data which can reasonably be collected on the two sets of observable characteristics, those of workers, $\mathbf{x}_i$, and those of employers, $\mathbf{w}_j$. It is easy to obtain the necessary $\mathbf{x}_i$ observations, which are simple characteristics of the sample members as found in a typical survey of a general adult population. The $\mathbf{w}_j$ observations required for the model as defined so far, however, are the characteristics of *all* the jobs in the (perhaps local) labor market which might be offered to *any* of the individuals in the sample of workers. This is an unrealistic requirement. All that is typically available in a sample survey are the characteristics of jobs actually held by sample members. The relevant consideration is how these available characteristics can best be used in place of the unobtainable requirement implied by the model.

A random sample of workers provides the basis for a

random sample of the characteristics of filled jobs.[4] If it is reasonable to assume that the characteristics of filled jobs are roughly similar to the characteristics of all jobs at any given time,[5] then it is also reasonable to take the distribution of job characteristics across the sample members as representative of the distribution of characteristics among jobs which might have been available to each sample member. Thus the inability to obtain the characteristics of all jobs potentially available is offset to some extent by the ability to estimate roughly the distribution of the characteristics of these jobs from the data at hand.

Two uses of this information seem reasonable. First, it is possible to use the behavioral model and the observed distribution of job characteristics to estimate relevant features of the choice situation which is most likely to have confronted each worker, given the characteristics of the job he/she is observed to hold. Such an individualized-expectations approach is relatively complicated, but seems feasible and is under development. The second approach is to use the available data in a substantially simplified interpretation of the model, the method used in Logan (1996a), which will now be described.

In the simplified interpretation of the model, each worker is thought of as making choices among types of jobs, or occupations, rather than individual jobs. Types of jobs can be characterized by their mean attributes, which are easily approximated by the mean characteristics of jobs of

each type in the available sample. In this approach each worker's own job is measured only as a job type so that his/her obtained job type is attributed the mean characteristics of all jobs in the type. The alternatives which form the elements of the offering sets O_k are then the job types, rather than jobs per se. Since there are only a small number of job types compared with jobs as such, the available data are sufficient to estimate the characteristics of all alternatives which may have been offered to each worker. Equation (7), with its subscripts altered according to the previous section, still stands as the appropriate formula for the model, but with the understanding that subscripts j, h, m, and n now refer to job types rather than jobs:

$$\Pr(A_{ij}) = \sum_{k:j \in O_k} \frac{\exp(\alpha w_j)}{\sum_{h \in O_k} \exp(\alpha w_h)} \prod_{\substack{m \in O_k \\ m>0}} \frac{\exp(\beta_m x_i)}{1+\exp(\beta_m x_i)} \prod_{\substack{n \in \bar{O}_k \\ n>0}} \frac{1}{1+\exp(\beta_n x_i)} \qquad (8)$$

Because the subscripts m and n appear on β preference vectors, it is implied that the preferences of employers are being estimated separately by type as well. That is, the recasting of jobs into types of jobs leads naturally to the segmentation of employer preferences by the types of jobs employers are offering. (Earlier it was said that segmentation of employer preferences would be desirable, without specifying what principle should determine the segmentation.)

A micro-simulation study (Logan 1996c) using the TSL

individual-level behavioral model showed good results when estimating the TSL model on job type characteristics, rather than on individual job characteristics. Substituting job type data did result in downwardly-biased estimates of α, the workers' preferences, but did not appreciably affect the estimates of employers' preferences, β_j. Note that good estimates of the β_j would typically be much more important for sociological studies, since it is these coefficients which imply rules of access to employment opportunities. The simulation also showed that both the α and β_j estimates using either job type data or individual job data were *insensitive* to a shift in the overall levels of demand in the simulated system, a key property of TSL estimates which is not shared by log-linear and other common models (see Logan 1996c). This demand insensitivity argues for the particular suitability of TSL models for comparisons of opportunity across locations or time periods which may differ in overall levels of demand.

Thus simulation suggests that the TSL model, as constrained in this section, provides a valid means of inference regarding the determinants of employment opportunity, even when mean occupational characteristics are substituted for the job characteristics actually determining outcomes. This substitution is not without costs, however. One drawback is that the number of variables which may appear in the $\mathbf{w}_j$ vectors is limited to J - 1, where J now stands for the number of categories or job types being used rather than the number of jobs. This limitation should be removed when

and if the method directly using characteristics of each worker's own job, mentioned above, proves practical. In the meantime, there is some comfort to be found in the fact that it is the effects on the other side of the model, the β_j, which are of more sociological interest, and that there is no important limitation on the number of these effects which can be included. Experimentation suggests that so long as the w_j vector contains at least one reasonable overall measure of the desirability of jobs, such as mean status, the estimates of the β_j coefficients are not dramatically affected by variations in the specification of w_j. So it seems reasonable to estimate models with detailed specifications of β_j effects but only limited specifications of α, at least when broad categories of jobs are in use.

Other possible constraints

It is tempting to constrain the TSL model even further, in order to simplify estimation. One possibility would be to assume that workers all share a common and strict order of preferences across jobs or job types. The special case of TSL which arises under this assumption is known as the sequential logit model (Amemiya 1981). Another is to assume that jobs of all types are freely available, which leads to the one-sided polytomous conditional logit model (Ben-Akiva and Lerman 1985). But the simulation in Logan (1996c) shows that both of these special cases perform poorly when the underlying micro-level behavioral model is TSL: estimates of coefficients are

inaccurate, and are inappropriately sensitive to demand shifts.

A third, plausible possibility, suggested by reviewers, is to assume an ordering (or, alternatively, a partial ordering) of occupations or job types such that a worker who is given access to a higher occupation is presumed to have access to all lower ones. Thus, for example, a worker able to work as a professional would be considered to have free access to sales jobs. However, a main purpose of the TSL model is to find evidence for different determinants of opportunity in different types of jobs. Making the assumption that access to higher jobs entails access to lower jobs would imply that the β_j preference coefficients for all job types must be identical. For if the coefficients were not identical then the rank ordering of candidates would differ across employers hiring in different types of jobs, making it possible for some workers to qualify for higher jobs without qualifying for lower ones; only equality of coefficients across all β_j could assure this would not happen. Actual application of TSL does in fact show believable differences in estimated β_j

coefficients among employers of different job types, in contrast to the implication of the suggested simplification.[6]

Thus there are reasons not to simplify the TSL model further, even though simplification would ease estimation.

Two Practical Estimation Algorithms

To ease exposition, estimation of the TSL model will be discussed in terms of workers and employers, rather than workers and sets of employers offering types of jobs; the algorithmic considerations are not affected by the reduction of the data from job characteristics to job type characteristics. The simplified form of the model in equation (8) will be the object of estimation.

An EM Estimation Algorithm

An often robust method for finding maximum likelihood estimates is the EM algorithm described by Dempster, Laird and Rubin (1977). The EM algorithm works by expressing a likelihood as a function of unobserved "complete data," which should be defined in such a way that a simpler likelihood maximization could be performed if the complete data were observed. In the estimation, the unobserved complete data are replaced by their expectations, given the observed, incomplete data (the E step), and the simpler maximization is then performed (the M step). The results of the maximization are used to obtain new expected values of the complete data, and the process is iterated to convergence. Assuming the simpler maximization is well-behaved, the EM algorithm is

inherently numerically stable, though it is often slow to converge. It can also often be implemented using appropriate weights with standard estimation programs, as in the present case. Dempster et al. prove that the algorithm increases the likelihood at each non-stationary point of the function, while Wu (1983) gives regularity conditions, met by the TSL model, which insure that the algorithm converges to a stationary point of the likelihood. The latter result does not preclude convergence to saddle points, which means that additional post-convergence checks are needed, as described below.

In the case of TSL, if the offers made by employers to workers could be observed, estimation of all the model parameters would become straightforward. Given the offers, the α parameters could be found from one polytomous conditional logit estimation, and the β parameters from J binary conditional logit estimations. This observation suggests an EM algorithm in which the offers of employers, together with the observed matches, are considered as the complete data. The estimation becomes a finite mixture distribution problem, in which the probability distribution over the offering sets for each individual is the mixing distribution for the observed job choices.

The following development of the appropriate EM algorithm is very closely based on the finite mixture discussion in Dempster et al. (1977, pp. 15-16), and the complementary development in Everitt and Hand (1981, pp. 8-

11). The former discussion is especially helpful for understanding the details of implementation.

New notation is required to show the correspondence with Dempster et al.'s presentation of the EM algorithm. The observed, incomplete data are the jobs accepted by each worker i, $i = 1, 2, \ldots n$, represented as n observations $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$. Each element y_i of $\mathbf{y}$ is an integer recording the index number of the job accepted by a worker; if worker i accepted job j, then $y_i = j$. (Unemployment is indicated by $y_i = 0$.)

The offering sets O_k comprise $R = 2^J$ possible states, only one of which actually corresponds to each worker's choice. Which of these sets was actually available for each worker is unknown to the researcher, making the states the *unobserved data*. These data are represented by the set $\mathbf{z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$, whose elements are R-length row vectors $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iR})$ each of which contains all zero entries except for a single entry of unity indicating which of the states (i.e., which set of offers) is associated with $\mathbf{y}_i$. For example, $\mathbf{z}_1 = (0\ 0\ 1\ 0\ \ldots\ 0)$ would indicate that the unobserved state for case number 1 is set number 3, that is, the one containing a job offer only from employer 2 (using the binary translation rule for offering set numbers given in footnote 3). The complete data are then defined to be $\mathbf{c} = \{\mathbf{y}, \mathbf{z}\}$.

The $\mathbf{z}_i$ are independently drawn from densities related to formula (6) above:

$$v(\mathbf{r}_k \mid \beta_1, \beta_2, \ldots, \beta_J, \mathbf{x}_i) = \Pr(S_{ik})$$

defined for R-length row vectors $\mathbf{r}_1 = (1, 0, \ldots, 0)$, $\mathbf{r}_2 = (0, 1, \ldots, 0)$, $\ldots$, $\mathbf{r}_R = (0, 0, \ldots, 1)$. Conditionally on $\mathbf{z}_i$, the observed $\mathbf{y}_i$ are independently drawn from densities related to formula (5) above:

$$u(\mathbf{y}_i \mid \mathbf{r}_k, \alpha, \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_J) = \Pr(A_{ij} \mid S_{ik}),$$

where it is understood that $j = y_i$.

To make the correspondence with Dempster et al. clearer, define $\phi = \{\beta_1, \beta_2, \ldots, \beta_J, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \alpha, \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_J\}$. Then, following their equations 4.3.1-4.3.2, define $\mathbf{U}(\mathbf{y}_i \mid \phi)$ to be a length-R column vector with elements $\log u(\mathbf{y}_i \mid \mathbf{r}_k, \phi) = \log \Pr(A_{ij} \mid S_{ik})$, $k = 1, 2, \ldots, R$; and $\mathbf{V}(\phi)$ to be a length-R column vector with elements $\log v(\mathbf{r}_k \mid \phi) = \log \Pr(S_{ik})$, $k = 1, 2, \ldots, R$. The log-likelihood of the complete data--that is, assuming knowledge of $\mathbf{z}$ and $\mathbf{y}$--then becomes:

$$\log f(\mathbf{c} \mid \phi) = \sum_{i=1}^{n} \mathbf{z}_i \mathbf{U}(\mathbf{y}_i \mid \phi) + \sum_{i=1}^{n} \mathbf{z}_i \mathbf{V}(\phi) \qquad (9)$$

Since formula (9) represents the complete data, it is true that only a single element of $\mathbf{z}_i$ is non-zero. That is, only one of the elements z_{ik}, $k = 1, 2, \ldots R$, is unity, while the rest are zeroes, so that formula (9) can also being expressed as

$$\log f(\mathbf{c} \mid \phi) = \sum_{i=1}^{n} \sum_{k=1}^{R} z_{ik} \log \Pr(A_{ij} \mid S_{ik}) + \sum_{i=1}^{n} \sum_{k=1}^{R} z_{ik} \log \Pr(S_{ik}) \; ; \qquad (10)$$

the non-zero z_{ik} serve as switches to pick out the

appropriate elements of $\mathbf{U}(\mathbf{y}_i|\boldsymbol{\phi})$ and $\mathbf{V}(\boldsymbol{\phi})$ for each case. Of course (10) does not imply knowledge by the researcher of which of the z_{ik} is non-zero for any particular case.

Since knowledge of the z_{ik} values is lacking, equation (10) cannot be evaluated directly. Instead, in the E-step of the EM algorithm, the z_{ik} are estimated as the conditional probabilities, given the observed outcome, that i experienced each of the R states (cf. Everitt and Hand 1981, pp. 9-10):

$$z_{ik}^{*} = \Pr^{*}(S_{ik} \mid A_{ij}) = \frac{\Pr^{*}(A_{ij} \mid S_{ik}) \Pr^{*}(S_{ik})}{\Pr^{*}(A_{ij})} \qquad (11)$$

The right-side quantities in this equation are found in formulas (5), (6) and (7) above; here the asterisks indicate that the formulas are to be evaluated using the current parameter estimates at each step. Each z_{ik}^{*} value is non-zero (because of the logistic form of the constituent probabilities), with $\sum_k z_{ik} = 1$.

The M step of the EM algorithm maximizes the complete data log-likelihood (10) with respect to α and β_j using the estimated z_{ik}^{*} values in place of the z_{ik}. Since the two terms on the right of the complete-data log-likelihood (10) do not depend on any shared parameters, they can be maximized separately, a possibility Dempster et al. emphasize. The M step then becomes a set of separate conditional logit maximizations, where the z_{ik}^{*} serve as weights in each. The first term on the right of (10) is a weighted polytomous conditional logit for the selection of an offer, given

knowledge of the offering set. Each worker i appears in the estimation as many times as there are offering sets which could have given rise to his/her observed choice, and is weighted by the appropriate value of z_{ik}^{*} at each appearance.

The second term on the right of (10) is the product of J weighted binary conditional logit models for the offers of the J employers. Since the employer submodels do not share parameters, the necessary maximization of the term can be performed as J separate, weighted logit estimations, using the same weights z_{ik}^{*} as before. Each logit estimation is done on dummy variables indicating whether or not the particular employer's offer was part of each of the possible sets. Each offering set which could have led to the observed choice enters into the estimation, weighted by its corresponding z_{ik}^{*}.

The overall algorithm works by first calculating the z_{ik}^{*} for all possible offering sets for all workers (the E step), and then performing polytomous and binary conditional logit maximizations to obtain updates of the parameter estimates, using the z_{ik}^{*} as weights across the sets (the M step). Because of the global concavity and simple form of the likelihood, Newton's algorithm is the method of choice for conditional logit models (Greene 1993: 667, 670). Details of the Newton algorithm and the first and second derivatives of the conditional logit likelihood necessary to apply it can be found in standard sources (e.g., Greene 1993, Maddala 1983). Once updated parameter estimates are obtained, a new cycle is begun by calculating new values of the z_{ik}^{*}.

Dempster et al. (1977: 10) remark that computation time can sometimes be saved by simply increasing the objective function in the M step, rather than maximizing it. Surprisingly, in the case of the TSL model repeated experiments have consistently found that performing only a single Newton step at each M step produces convergence in the same, or very close to the same, number of EM steps as does performing complete Newton maximizations at each M step. In practice, therefore, there is no need to do more than a single Newton algorithm step at each M step for this model.

Convergence of the EM algorithm itself is determined by observing both the likelihood and the parameter values. Convergence in the likelihood is detected when the largest of the relative gradients of the likelihood with respect to each of the parameters has fallen below a specified tolerance.[7] Convergence in the parameters occurs when the largest relative change in successive estimates of the parameters falls below tolerance (see Dennis and Schnabel 1983: 160).

Experience shows that very poorly specified models can converge by the likelihood criterion while parameter estimates do not converge, or converge very slowly. This is especially likely to occur in over-parameterized models (especially tabular models estimated on grouped data), where identification breaks down. In such cases the likelihood values are still informative even though the parameters are poorly identified.

AEM: An Accelerated EM Algorithm

Though EM, as a generally stable and easily implemented method, is the algorithm of choice for many problems with large numbers of parameters, it is often very slow to converge. Of various acceleration methods proposed for EM, Jamshidian and Jennrich's (1993) generalized conjugate gradient method was chosen for the TSL problem because of the high level of acceleration reported by its authors, namely, improvements in speed by factors of 2.5 to 92.0.

Linear conjugate gradient methods (Gill, Murray and Wright 1981; Press, Teukolsky, Vetterling, and Flannery 1992) choose successive search directions in the parameter space which preserve the maximizations achieved in previous search directions: that is, at each step, the gradient of the likelihood function remains perpendicular to the previous search directions, so that no movement along those directions will increase the function. EM does not have this property. However, Jennrich and Jamshidian (1993) observe that each EM step may be considered an approximate generalized gradient of the likelihood function, with respect to an appropriate norm. Using the EM steps as generalized gradients, a conjugate gradient algorithm is defined which achieves orthogonal search steps in the appropriate metric, which changes at each iteration. The result is called the accelerated EM, or AEM, algorithm.

In addition to the calculations required for EM, the AEM algorithm requires the gradient vector of first derivatives

of the log-likelihood for the full model. With reference to equation (8), the full model log-likelihood is:

$$L = \sum_{i=1}^{n} \sum_{j=0}^{J} y_{ij} \ln \Pr(A_{ij})$$

where y_{ij} is a dummy variable defined to take the value 1 if worker i has obtained a job in category j, and 0 otherwise. The first derivatives, suppressing the i subscripts, are[8]

$$\frac{\partial L}{\partial \alpha_g} = \sum_{i=1}^{n} \sum_{j=0}^{J} y_j \sum_{k=1}^{R} \Pr(S_k \mid A_j) \left[w_{jg} - \sum_{h=0}^{J} w_{hg} \Pr(A_h \mid S_k) \right]$$

$$\frac{\partial L}{\partial \beta_{fg}} = \sum_{i=1}^{n} x_g \sum_{j=0}^{J} y_j \sum_{k=1}^{R} \Pr(S_k \mid A_j) (\omega_{fk} - p_f) \quad ,$$

where

$$\Pr(S_k \mid A_j) = \frac{\Pr(A_j \mid S_k) \Pr(S_k)}{\Pr(A_j)} \quad ,$$

$$\omega_{fk} \begin{cases} = 1 \text{ if } f \in O_k \\ = 0 \text{ otherwise} \end{cases} \quad , \text{ and}$$

$$p_f = \Pr(o_f = 1) .$$

The AEM algorithm also requires programming a simple line search routine given by Jamshidian and Jennrich.

Jamshidian and Jennrich recommend starting the AEM algorithm only after successive EM steps produce changes in the log-likelihood of less than 0.5. My experience indicates that this rule allows the AEM algorithm to converge in the large majority of cases, but that sometimes divergence is observed instead. In such cases the solution is to change the

triggering criterion to a smaller value, say half as large. This expedient, which may need to be repeated, is not known to be completely failproof. However, if AEM were to continue to fail even at very small criteria, it would be simple to fall back on EM by setting the criterion to 0.

Implementation of the algorithms

Though one attractive feature of EM algorithms is that the maximization step can often be performed with standard statistical packages, this seems less attractive with the TSL model. Many standard conditional logit routines require preparation of data in case-by-alternative record format: that is, one record for each alternative available to each case. In the TSL model there must be separate, iteratively-reweighted sets of case-alternative records for each possible offering set available to each worker, with the numbers and identities of the records in each set varying across workers according to the observed occupational outcomes. Preparing the records itself requires a special program, but the worse drawback is that the number of records becomes very large, which slows the estimation very seriously.

Because of such drawbacks of adapting canned programs, a stand-alone FORTRAN program has been developed for TSL estimation. This program requires no special data preparation, using only a single record per case, as found in most standard data sets. It includes a recoding command for easily redefining occupational categories; this allows

flexibility in exploratory work, taking advantage of the faster convergence available with smaller numbers of categories. It also has other useful features such as occupational category and variable labels, and reports of descriptive statistics by category. The program relies on public-domain LINPACK routines for numerical programming operations (Dongarra, Moler, Bunch and Stewart 1979).[9]

The stand-alone program calculates asymptotic standard error estimates for all parameters using analytical second derivatives rather than numerical techniques. The second derivatives of the model are:

$$\frac{\partial^2 L}{\partial \alpha_g \partial \alpha_h} = \sum_{i=1}^{n} \sum_{j=0}^{J} y_j \sum_{k=1}^{R} \Pr(S_k \mid A_j) \left[T_{jkg} U_{jkh} - \sum_{f=0}^{J} w_{fg} \Pr(A_f \mid S_k) T_{fkh} \right] ,$$

$$\frac{\partial^2 L}{\partial \alpha_g \partial \beta_{fh}} = \sum_{i=1}^{n} x_h \sum_{j=0}^{J} y_j \sum_{k=1}^{R} \Pr(S_k \mid A_j) (\omega_{fk} - p_f) U_{jkg} , \qquad (12)$$

$$\frac{\partial^2 L}{\partial \beta_{dh} \partial \beta_{fg}} = \sum_{i=1}^{n} x_g x_h \sum_{j=0}^{J} y_j \sum_{k=1}^{R} \Pr(S_k \mid A_j) [V_{jdk} (\omega_{fk} - p_f) - \delta_{df} p_f (1 - p_f)] ,$$

where

$$T_{jkg} = w_{jg} - \sum_{m=0}^{J} w_{mg} \Pr(A_m \mid S_k) ,$$

$$U_{jkg} = T_{jkg} - \sum_{d=1}^{R} \Pr(S_d \mid A_j) T_{jdg} ,$$

$$V_{jdk} = \omega_{dk} - p_d - \sum_{b=1}^{R} \Pr(S_b \mid A_j) (\omega_{db} - p_d) , \text{ and}$$

$$\delta_{df} \begin{cases} = 1 \text{ if } d = f \\ = 0 \text{ otherwise.} \end{cases}$$

Though EM does not require second derivatives and does not produce standard error estimates as a by-product, available numerical methods for estimating standard errors without second derivatives can take several times longer than the EM estimation itself (e.g., Meng and Rubin 1991), and were therefore rejected as inappropriate for production work.

Relative performance

Table 1 compares the performance of the EM and AEM algorithms when re-estimating the preferred model of Logan (1996a). Coefficient estimates and relative and absolute goodness of fit measures can be found in that paper. The data are four independent samples from the General Social Survey, representing females and males surveyed in two time periods, 1972-80 and 1982-90. The sample sizes, shown in the table, average 2,679. The number of parameters estimated in each model is 22.

The first panel of table 1 presents statistics for the convergence of the algorithms to 3 and 6 digits of accuracy in the 1972-80 female data, when using zero start values for all parameters. An estimate was considered accurate to 3(6) digits when all estimated parameters, rounded to 3(6) significant digits, agreed with similarly rounded values obtained from runs to several more digits of accuracy. As the first line of the panel shows, the EM algorithm took 594 iterations to reach 3 digit accuracy, consuming 28.9 minutes of CPU time.[10] The AEM algorithm began taking accelerated steps after 37 initial EM iterations, when the change in the

log-likelihood between steps first dropped to less than 0.5, as is reported in the column labeled "Initial EM Iterations." After the accelerated steps began, AEM ran 29 more iterations, terminating after a total execution time, including the EM startup, of 6.0 CPU minutes. The relative performance ratio in the last column is the EM total CPU time divided by the AEM total, 4.8 in this case. AEM, that is, was 4.8 times faster than EM for this problem. For 6 digits of accuracy, AEM converged in 8.0 minutes versus 57.4 minutes for EM, or 7.2 times as fast.

The second panel of table 1 shows convergence results for the 1982-90 female data. As mentioned in Logan (1996a), these data do not produce convergence to the MLE's from zero start values, but instead converge to a local, non-global maximum. The performance shown in table 1 reflects instead the use of the 1972-80 female estimates as starting values for the 1982-90 data, and the convergence then is to the values reported in Logan (1996a). The problem of local maxima will be discussed in the next section. In this data, the AEM algorithm was 8.2(7.5) times faster than EM to 3(6) digits of accuracy. In the 1972-80 male data, shown in the third panel, AEM was 4.5(6.1) times faster than EM, while in the 1982-90 male data, it was 9.0(8.8) times faster. On average, AEM was 6.6 times as fast as EM to 3 digits of accuracy and 7.4 times as fast to 6 digits, or 7.0 times as fast overall.

Local Maxima

Table 2 presents two sets of estimates obtained from the 1982-90 female data. The first set results when zero starting values are used in either algorithm (the AEM steps commencing from initial EM iterations based on zero starting values). The second set results when the 1972-80 female estimates are used as starting values. Since the (doubled) log-likelihood of the first set, -9035.70, is lower in magnitude than that of the second, -8975.10, it is clear the values in the first set are not the MLE's, while those in the second set may be.

The results shown in panel A of table 2 are the only example of convergence to a local, non-global maximum which I have knowingly encountered with real data. Though such problems may in fact be rare, it is well worth considering when they are likely to occur and how they may be detected.

In the context of estimating four parallel models in independent samples, as was done for Logan (1996a), it was instantly apparent that the first set of results was suspect. The large positive intercept and the large negative coefficient for education in the Sales/Clerical category were dissimilar to the patterns seen in the 1972-80 female data and in both sets of male data. On closer inspection, the standard errors for the intercept and for the nonwhite coefficient in the same category were seen to be unusually large, not only compared with other samples, but also in comparison with other categories in the same model. These seem to be good patterns to look for in general. Restarting

the algorithm with the 1972-80 female estimates as starting values at once produced the second set of estimates, with a higher likelihood value.

This experience suggests another strategy for detecting local maxima, which is purposely to estimate parallel models on similar sets of data. When, unlike the case here, there is no substantive interest in comparing parallel models on separate sets of data, it may be worthwhile to divide a data set arbitrarily into smaller pieces to see if divergent results ensue. In fact, in the 1982-90 female data, estimating the same model on two halves of the data set, with zero starting values, gives one set of results similar to each of the solutions seen in table 2.

If results are available from similar data sets, it would seem always to be wise to use these as starting values, in addition to using zero starting values. It is commonly suggested that arbitrarily different starting values be tried in any problem in which local maxima may occur, but devising a good set of such starting values is not trivial when there are 20 or more parameters. It is also reasonable to try estimates obtained from special cases of the TSL model as starting values. Appropriate special cases might be the conditional logit model using mean occupational characteristics, the sequential logit model based on worker characteristics, or the logit analog of the Abowd and Farber model discussed above, depending on which simplification of the model the researcher believes may be closest to the

actual parametric situation (see Logan 1996a). The first of these models can be estimated quickly with GLIM (e.g., Lindsey 1995), the second with LIMDEP (Greene 1989), and the last with the available TSL program. It may happen that starting values obtained from these special cases will produce numerical overflows in the estimation program, in which case they should be adjusted appropriately to relieve the problem. Reducing the absolute values of demand intercepts or of the α coefficients on important w_j variables would usually be appropriate remedies.

Cox (1970: 89), in the context of simpler binary models, observed, "It seems likely on general grounds that multiple maxima will not arise unless there are either very limited data or gross discrepancies with the model." Though the data used here are not very limited in quantity, there is evidence that the data set for which multiple maxima are observed is the most poorly fitting of the four models. Three of the five Hosmer-Lemeshow absolute goodness of fit statistics, $\hat{C}$, shown for each of the two sets of results in table 2 are extremely high compared to their expectations of 8.0 and their nominal chi-square critical values of 15.51 at $p = 0.05$ (Hosmer and Lemeshow 1989:140-5).[11] Examination of these statistics (which are printed by the available TSL estimation program) may help to identify models particularly likely to have converged to local maxima.

Of course, a model with poor absolute fit may not be worth interpreting in the first place, so that the appearance

of local maxima *only* in such instances might not really be problematic. For example, the $\hat{C}$ statistics reported in Logan (1996a, table 2) for these estimations show very acceptable fits for both male samples, but also a single, large outlying $\hat{C}$ value for the other female sample. Even though a search produced no other maximum for the second female sample, the appearance of the outlying $\hat{C}$ value is in itself enough to discount the estimates in that model. The discussion in Logan (1996a) therefore qualifies its interpretation of both female models' estimates, while the male estimates are interpreted more seriously because of their good absolute fits.[12] Researchers who are reluctant to interpret models with poor absolute fits seem likely thereby to reduce their chances of interpreting estimates from local, non-global maxima, if Cox's observation pertains.

It should also be mentioned that it is possible in principle for the algorithms considered in this paper to reach saddle points, rather than local or global maxima. If this occurred using the available software with the calculation of analytical standard errors requested, the program would detect the consequent non-negative-definiteness of the matrix of second derivatives and issue an error message. If other software were used, or if standard errors were not being calculated, slightly perturbing the estimates and restarting the algorithm should cause movement away from the saddle point.

None of the above should obscure the point that the most

useful starting values of all are zeroes for all parameters. Almost always zeroes produce reasonable fits to which all workable alternative starting values also lead, and do so with relatively little extra cost in execution time. Exceptions have occurred in very poorly specified models, but such models generally do not produce alternative solutions with alternative starting values; instead they diverge from all attempted starting values. The other known exception is the local, non-global maximum in table 2.

Until more experience is obtained, the single known appearance of a local, non-global maximum in a very poorly fitting model is not especially worrisome, though it is important to watch for signs of this problem, to explore alternative starting values for any final model or unusual result, and to exercise caution regarding models with poor absolute fits.

Other Estimation Algorithms

EM has proved to be a reliable method for TSL estimation, while acceleration with Jamshidian and Jennrich's AEM has provided the same reliability (overlooking the occasional triggering point adjustment) and much greater speed. Nonetheless, the appearance of 2^J terms in the likelihood means that execution times approximately double with each additional job type or occupation being fitted, so that more efficient algorithms would be very valuable.

From a numerical programming perspective, TSL is a

smooth, unconstrained optimization problem. That is, the objective function, the log-likelihood corresponding to (8), is a continuous function of continuous variables, which to an optimization algorithm are the α and β_j vectors rather than the data vectors, which are taken as given constants. The discreteness of the choice described in the underlying behavioral model has no effect on the continuity of (8), considered as a function of α and β_j. Furthermore, because the coefficients and data enter (8) exclusively in the form of logits (binary and polytomous), there are no finite parameter values, positive or negative, which are prohibited by the structure of the model or which could be prohibited by any configuration of the observed data. Equation (8) always comes out to a value between 0 and 1, for finite parameters and data; the problem is unconstrained.

As a smooth, unconstrained optimization problem, TSL in principle is a candidate for Newton or quasi-Newton estimation. However, the computational complexity of the model tends to move it fairly quickly outside the realm of small-to-medium size problems for which the simpler forms of these methods are optimal. The analytical second derivatives required for the Newton algorithm per se are extremely time consuming to compute, as examination of (12) should suggest.

It is possible that Newton-type algorithms designed to take advantage of the particular structure of the model could achieve relatively high speed. It is also possible that quasi-Newton methods which avoid calculation of analytical

second derivatives might be effective. However, my own experience with such an algorithm, the Davidon-Fletcher-Powell routine provided in LIMDEP, proved disappointing because of extreme instability. In addition, Swait and Ben-Akiva had similar problems with their simpler but mathematically similar independent availability logit model, even when using the highly-regarded quasi-Newton algorithm of Dennis and Schnabel (1983). The problems were enough to make them suggest that they might jeopardize "the practical usefulness of probabilistic choice set models in general" (Swait and Ben-Akiva 1986: 82).[13] Because of the complexity of the TSL model, it seems likely that successful implementation of a useful Newton-type algorithm would require substantial numerical programming expertise.

A completely different approach to the problem would be to employ one of the many recently-popular algorithms which use random sampling to approximate features of the likelihood. Monte Carlo EM, for example, would sample from the distribution of unobserved offering sets rather than exhaustively evaluate the distribution to form the expectation (see Tanner 1993, Diebolt and Ip 1996). Markov chain Monte Carlo would sample values in both the parameter space and the distribution of unobserved offering sets to estimate the entire model (see Gilks, Richardson and Spiegelhalter, eds., 1996). Such algorithms have proved highly effective in very complex models, and might perform well for TSL even with relatively large values of J.

Discussion

The basic TSL behavioral model described above is very simple: employers choose the best workers they can get, workers choose the best jobs they can get, and the intersecting preferences of workers and employers create constraints on the choices of all. There are relatively rich interpretative and theoretical advantages associated with the basic model, which could not be addressed in detail here. For example, the preference coefficients of the model can be interpreted as measures of *relative opportunity*, where opportunity is given a mathematical definition (Logan 1996a). In addition, the β parameters can be interpreted as *rules of access* to positions, along the lines of a nonmathematical suggestion by Hauser (1978), and it can then also be shown both that TSL estimation of these rules should be insensitive to changes in levels of demand for workers, and that a class of extended log-linear models which includes most applied mobility table models should *not* be insensitive to changes in levels of demand, contrary to the common belief (Logan 1996c). Finally, as mentioned earlier, there is a direct connection between the TSL behavioral model and the very rich theoretical literature on formal, two-sided matching games (Logan 1996b, Roth and Sotomayor 1990).

The problem which this paper has addressed in detail is how best to move from such an attractive micro-level model to a useful tool for empirically estimating the determinants of opportunity and the rules of access to positions. The guiding

principle throughout has been pragmatic: the availability of certain types of data must force certain simplifications and even compromises in the model to achieve identification and estimability. Even with appropriate accommodations to the realities of data gathering, TSL estimation remains a relatively lengthy procedure which is severely limited in the numbers of occupational categories which can be distinguished.[14] However, it is also an estimation method which can be applied immediately to many existing, generic sets of data, such as the GSS, using available software.

The most important substantive consideration in applying the TSL estimation methods developed here is whether and to what degree the desirable properties of the micro-behavioral model carry over into desirable properties of the obtained estimates. The focus, as explained earlier, is on the α and β preference coefficients, rather than on the variety of effects which enter the demand intercepts. Short of mathematically deriving a new and practical estimation method directly from the micro-level model (on which work is in progress), the best way of addressing the issue is by simulation.

The micro-simulation reported in Logan (1996c) gives encouraging results: estimates using the occupational category method developed in this paper are good for β, downwardly biased but strongly significant for α, and, very importantly, insensitive to a simulated shift in demand. Since the estimates for β are good, accurate inferences can

be made about the determinants of relative opportunity in the simulated micro-behavioral system. Since the estimates are insensitive to demand shifts in the micro-behavioral system, there is reason to think the estimates of relative opportunity are suitable for cross-national and cross-temporal comparisons. Additional simulation studies varying the parameters of the micro-behavioral system in new ways should lead to new information about the robustness of these results.

A combination of micro-simulation with exploration of alternative specifications and constraints within the general TSL behavioral framework seems likely to be a fruitful mode of research for future improvements in TSL estimation. The very fact that TSL estimation is connected to a behavioral model which allows simulation seems extremely important both as an aid to further model development and as a brake on any tendency toward abstract specifications which lack clear behavioral referents.

REFERENCES

Abowd, John M., and Henry S. Farber. 1982. "Job Queues and the Union Status of Workers." *Industrial and Labor Relations Review* 35(3): 354-367.

Amemiya, Takeshi. 1981. "Qualitative Response Models: A Survey." *Journal of Economic Literature* XIX:1483-1536.

Ben-Akiva, Moshe, and Steven R. Lerman. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA: MIT Press.

Cox, D.R. 1970. *The Analysis of Binary Data*. London: Methuen.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." (With Discussion). *Journal of the Royal Statistical Society, Series B (Methodological)* 39:1-38.

Dennis, J.E., and R.B. Schnabel. 1983. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs, NJ: Prentice-Hall.

Diebolt, Jean, and Eddie H.S. Ip. 1996. "Stochastic EM: Method and Application." Pp. 259-273 in *Markov Chain Monte Carlo in Practice*, edited by W.R. Gilks, S. Richardson, and N.G. Speigelhalter. New York: Chapman and Hall.

Dongarra, J.J., C.B. Moler, J.R. Bunch, and G.W. Stewart. 1979. *LINPACK User's Guide*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Everitt, B.S., and D.J. Hand. 1981. *Finite Mixture*

Distributions. New York, London: Chapman and Hall.

Gilks, W.R., S. Richardson, and D.J. Spiegelhalter (eds.). 1996. *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Gill, Philip E., Walter Murray, and Margaret H. Wright. 1981. *Practical Optimization*. London; New York: Academic.

Greene, William H. 1989. *LIMDEP Version 5.1*. New York: Econometric Software, Inc.

_____. 1993. *Econometric Analysis*. 2nd Edition. New York: Macmillan.

Hauser, Robert M. 1978. "A Structural Model of the Mobility Table." *Social Forces* 56:919-953.

Hechter, Michael. 1994. "The Role of Values in Rational Choice Theory." *Rationality and Society* 6:318-333.

Horowitz, Joel L. 1991. "Reconsidering the Multinomial Probit Model." *Transportation Research B* 25B:433-438.

Hosmer, David W., and Stanley Lemeshow. 1989. *Applied Logistic Regression*. New York: Wiley.

Jamshidian, Mortaza, and Robert J. Jennrich. 1993. "Conjugate Gradient Acceleration of the EM Algorithm." *Journal of the American Statistical Association* 88(421):221-228.

Lindsey, James K. 1995. *Modelling Frequency and Count Data*. New York: Oxford University Press.

Logan, John Allen. 1996a. "Opportunity and Choice in Socially Structured Labor Markets." *American Journal of Sociology*

102(1; July).

_____. 1996b. "Rational Choice and the TSL Model of Occupational Opportunity." *Rationality and Society* 8(2;May): 207-230.

_____. 1996c. "Rules of Access and Shifts in Demand: A Comparison of Log-Linear and Two-Sided Logit Models." *Social Science Research* 25 (June):174-199.

Maddala, G.S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Meng, Xiao-Li, and Donald B. Rubin. 1991. "Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm." *Journal of the American Statistical Association* 86(416):899-909.

Press, William H., Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 1992. *Numerical Recipes in FORTRAN; The Art of Scientific Computing*. 2nd ed. Cambridge; New York: Cambridge University Press.

Pudney, Stephen. 1989. *Modelling Individual Choice*. Oxford: Basil Blackwell.

Roth, Alvin E., and Marilda A. Oliveira Sotomayor. 1990. *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Cambridge: Cambridge University Press.

Roth, Alvin E., and John H. Vande Vate. 1990. "Random Paths to Stability in Two-Sided Matching." *Econometrica* 58:1475-80.

Swait, Joffre, and Moshe Ben-Akiva. 1986. "Constraints on Individual Travel Behavior in a Brazilian City." *Transportation Research Record* 1085:75-85.

Tanner, Martin A. 1993. *Tools for Statistical Inference; Methods for the Exploration of Posterior Distributions and Likelihood Functions.* 2nd. ed. New York: Springer-Verlag.

Wu, C.F.J. 1983. "On the Convergence Properties of the EM Algorithm." *Annals of Statistics* 11:95-103.

NOTES

[1] The standard Gumbel distribution has distribution function $\exp(-e^{-x})$, mode 0, mean 0.57722, and variance $\pi^2/6$. The derivation of (3) from equations similar to (1) and (2) is a standard result in the discrete choice literature; see, e.g., Ben-Akiva and Lerman (1985), Greene (1993), Maddala (1983), or Pudney (1989).

[2] See Ben-Akiva and Lerman (1985) or Pudney (1989) for the derivation of the multinomial conditional logit model from these distributional assumptions.

[3] Though the particular ordering of the offering sets is unimportant for what follows, a lexicographic ordering is adopted for specificity. Numbering the sets from 1 through 2^J, the contents of each set can be determined by expressing the set number, minus 1, in binary notation, and reading off the nonzero elements from right to left. For example, for set $k = 12$, expressing $k - 1 = 11$ in binary form as 1011 and reading off the nonzero elements from right to left shows that the numerals 1, 2 and 4 are contained in the set, in addition to the always-present numeral 0 indicating the availability of unemployment. Under this scheme, the possible offering sets in order are {0}, {0,1}, {0,2}, {0,1,2}, {0,3}, {0,1,3}, {0,2,3}, {0,1,2,3}, {0,4}, ..., {0,1,2,...,2^J}.

[4] In principle, using a sample of workers to estimate the distribution of job characteristics requires adjustments for multiple job holding, which will not be considered further

here.

[5] Possibly apart from the characteristics of chronically hard to fill jobs in the lower depths of the job market.

[6] See Logan (1996a) for a discussion of the rank ordering implied by β coefficients, for empirical results showing different estimates across job types, and for a derivation of the model implied by the suggested simplification, which is a logit analog of the partial observability probit model of Abowd and Farber (1982).

[7] See Dennis and Schnabel, p. 160, eq. 7.2.5. The gradient of the EM likelihood itself, required for the relative gradient calculation, appears in the next section of this article.

[8] All derivative formulas reported in this paper have been verified numerically using Ridders's method of polynomial extrapolation (see Press, et al., 1992, pp. 182-183).

[9] The EM algorithm, with analytical standard error calculations, as used in Logan (1996a), is available in a public-release program, TSLogit, v.1.0, with a user's guide. The program is written in DEC Fortran.

[10] Reported computer times are total CPU times (system plus user times) for the complete estimation until convergence, including the negligible setup routines and the initial EM iterations prior to the start of the AEM steps. The computer was a 133 MHz DEC 3000/400 workstation, running the OSF/1, V3.2, UNIX operating system. All computer times and ratios of times were calculated with more significant digits than are

shown here.

[11] The Hosmer-Lemeshow statistic, $\hat{C}$, allows a chi-square test comparing the proportions of outcomes actually observed in each decile of risk with the proportions which are expected according to a model. For each occupational outcome, deciles of risk are formed across the observations by collecting the first ten percent of cases which have the lowest predicted probabilities of accepting a job within the occupation, followed by the ten percent with the next lowest probabilities, and so on. The test statistic should be distributed approximately as $\chi^2(8)$ when calculated as just described. I am following the spirit of Hosmer and Lemeshow's own recommendation for assessing the fit of the multinomial logit model, which is to calculate separate binary model test statistics for each outcome category, even though neither theoretical nor simulation justification is given for this recommendation. I believe the results should be approximately correct here as well.

[12] Logan (1996a) should probably have been even more circumspect overall in interpreting the female estimates. It seems reasonable to attribute the poor fit among females to a lack of relevant data particularly appropriate to their labor force participation, such as the presence of children at home.

[13] The independent availability logit model is conceptually distinct from TSL, and does not involve two sets of actors.

It and TSL were developed independently. See Logan (1996a).

[14] My experience extends to 7 categories, plus unemployment, with perhaps 10 being a reasonable goal using the AEM algorithm on problems similar to the ones reported here, at least for researchers who are patient.

Table 1. EM versus Accelerated EM (AEM) Convergence Performance.

Sample Convergence	EM Algorithm Total Iterations	EM Algorithm Total CPU Minutes	AEM Algorithm Initial EM Iterations	AEM Algorithm AEM Iterations	AEM Algorithm Total CPU Minutes	Relative Performance
Females, 1972-80, n = 2,632						
to 3 digits	594	28.9	37	29	6.0	4.8
to 6 digits	1190	57.4	37	43	8.0	7.2
Females, 1982-90, n = 3,283*						
to 3 digits	905	47.3	10	34	5.7	8.2
to 6 digits	1132	59.1	10	48	7.8	7.5
Males, 1972-80, n = 2,149						
to 3 digits	445	13.5	24	25	3.0	4.5
to 6 digits	952	28.8	24	44	4.7	6.1
Males, 1982-90, n = 2,651						
to 3 digits	1326	49.4	22	43	5.5	9.0
to 6 digits	2356	87.7	22	85	10.0	8.8

*Performance using female, 1972-80, estimates as start values; all others used zero start values.
Note: All AEM algorithm runs took initial EM iterations until the change in log-likelihood values between successive steps fell below 0.5.

Table 2. Different Estimates from Different Start Values, 1982-90 Female Data.

A. Estimates from Zero Start Values. (2*Log-Likelihood = -9035.70.)

Workers' Preferences	
Prestige	.030 (.003)
Autonomy	.086 (.013)

Firms' Preferences	Prof	Mgmt	Clerical/ Service	Mfg. Blue	Other Blue
Intercept	-1.064 (.113)	-1.106 (.092)	3.339 (.600)	-1.253 (.308)	-2.249 (.126)
Education	.862 (.054)	.281 (.034)	-.847 (.155)	-.807 (.133)	-.098 (.047)
Age	.004 (.015)	.038 (.013)	-.070 (.049)	.027 (.031)	-.011 (.020)
Nonwhite	-.398 (.229)	-.851 (.216)	2.990 (2.475)	1.627 (.521)	-.353 (.313)
$\hat{C}$	42.139	29.767	76.982	12.000	7.533

B. Estimates Using 1972-80 Female Estimates as Start Values. (2*Log-likelihood = -8975.10)

Workers' Preferences	
Prestige	.059 (.007)
Autonomy	.136 (.013)

Firms' Preferences	Prof	Mgmt	Clerical/ Service	Mfg. Blue	Other Blue
Intercept	-1.563 (.116)	-1.477 (.092)	1.669 (.308)	-.679 (.362)	-2.401 (.128)
Education	.791 (.048)	.311 (.031)	.370 (.084)	-.691 (.154)	-.023 (.047)
Age	.005 (.013)	.039 (.012)	.039 (.020)	.053 (.040)	-.005 (.020)
Nonwhite	-.422 (.193)	-.864 (.205)	-.144 (.237)	1.656 (.674)	-.401 (.305)
$\hat{C}$	25.533	24.618	51.438	11.739	9.034

Note: Standard errors in parentheses.