Simulation Appendix

In each of three periods, individuals make binary education decisions $d_t \in \{0,1\}$. The decisions are sequential: if an individual chooses $d_t = 0$, then for all $t' > t$, $d_{t'} = 0$. The $0-1$ variable C indicates whether an individual "completes college"; C equals 1 if and only if $d_t = 1$ for all t. After each decision has been made, individuals receive realizations on state variables that affect future decisions. An example of such a process occurs when an individual: (1) decides to take the SAT and receives a realization of the SAT score; (2) decides to apply to college and receives a realization on financial aid; and (3) decides whether to attend college and receives an earnings realization.

After the final decision is made, individuals receive earnings, Y, which follow

$$\ln(Y) = \left[\begin{array}{ccc} X & C & Type \end{array}\right]\gamma + \eta, \tag{1}$$

where X is a vector of individual characteristics known to the individual at all stages and $Type$ is the individual's unobserved (to the econometrician) type, which takes on one of K values. The component of earnings which is unknown to the individual until after all educational decisions are made is η. We assume that η is distributed $N(0, \sigma_Y^2)$, and is independent of X, C, and $Type$. The vector of coefficients to be estimated is then γ.

Each of the actions needed to set C equal to one is costly, with the costs varying by individual. At $t = 3$, the expected present value of lifetime utility from choosing $d_3 = 1$ conditional on both d_1 and d_2 equalling one is given by

$$v_3(d_3 = 1) = [\begin{array}{ccc} Z_3 & Type & E_3(Y|C=1) - E_3(Y|C=0) \end{array}]\alpha_3 + \epsilon_3. \tag{2}$$

Z_3 is a vector of individual characteristics that may overlap with X. Note further that some of the elements of Z_3 may be known only at the time of the third period decision; when making earlier decisions, individuals may face uncertainty over these variables. The unobserved (to the econometrician) preference for choosing $d_3 = 1$ is given by ϵ_3, which is assumed to follow a logistic distribution. As is standard in dynamic discrete choice models, this unobserved preference is unknown to the individual before period three. The utility of choosing $d_3 = 0$ is normalized to zero, so that individuals choose $d_3 = 1$ whenever the above expression is positive.

Conditional on $d_{t-1} = 1$, the value of setting $d_t = 1$ for $t \in \{1,2\}$ is given by

$$v_t(d_t = 1) = [\begin{array}{ccc} Z_t & Type & E_t(V_{t+1}|d_t = 1) \end{array}]\alpha_t + \epsilon_t, \tag{3}$$

where as before: Z_t is a vector of individual characteristics; ϵ_t is an unobserved logistically-distributed preference shifter; and utility is normalized so that $d_t = 1$ whenever $v_t(d_t = 1)$ is positive. As is standard, we assume that the preference shifters (the ϵ's) are independent across time—embedding the problem in a mixture model is useful in part because it allows for persistent taste differences. Note that evaluating equation (3) requires the individual to take expectations over both future preferences and future state variables. For the simulations, we assume that there is one random state variable in each period. We assume further that each of these state variables is distributed $N(0, \sigma_s^2)$ and is independent of any information the individual might have in preceding periods.

Using the distributional assumptions for unobserved preferences and integrating over uncertain future states, Rust (1987) showed that equation (3) can be rewritten as

$$v_t(d_t = 1) = [\ Z_t \quad Type \quad \int \ln(\exp[\overline{v}_{t+1}(d_{t+1} = 1|Z_{t+1})] + 1)\pi_t(Z_{t+1})dZ_{t+1}\ ]\alpha_t + \epsilon_t,$$

where $\pi_t(\cdot)$ is the pdf of the state vector Z_{t+1}, and $\overline{v}_{t+1}(\cdot)$ is the value of "choosing college" net of preferences, i.e., with $\epsilon_{t+1} = 0$. Discretizing the state space allows us to calculate $\overline{v}_t(d_t = 1)$, the expected lifetime utility from college across realizations of Z_{t+1} (when $\epsilon_t = 0$). In the simulations, we assume that one element apiece of Z_2 and Z_3 must be discretized in this fashion.

The probability of choosing choosing college at time t is then

$$P_t(d_t = 1|X, Z, Type; \gamma, \alpha) = \frac{\exp(\overline{v}_t(d_t = 1))}{\exp(\overline{v}_t(d_t = 1)) + 1},$$

taking the familiar logit form, with any nonlinear terms appearing inside the parentheses. For each individual, the likelihood can be written as

$$\mathcal{L}(\gamma, \alpha) = \prod_{t=1}^{3} P_t(d_t|X, Z, Type; \gamma, \alpha) f(Y|X, C, Type; \gamma),$$

where f is the pdf of the outcome Y.

Note that if an individual's *Type* were observed, his log-likelihood would factor as

$$L(\gamma, \alpha) = L_1(\gamma) + L_2(\gamma, \alpha_3) + L_3(\gamma, \alpha_3, \alpha_2) + L_4(\gamma, \alpha_3, \alpha_2, \alpha_1),$$

where L_1 refers to the log-likelihood contribution of earnings, Y, and L_2 through L_4 are the log-likelihood contributions from the preceding periods. Estimation can proceed sequentially. Consistent (though not efficient) estimates of γ can be obtained from maximizing the sample average of L_1 alone. Taking these estimates as the true values, consistent estimates of α_3

can be obtained from maximizing the sample average of L_2. With estimation proceeding sequentially, L_3 and L_4 are then used to obtain estimates of α_2 and α_1, respectively.

When an individual's $Type$ is unobserved, his log-likelihood becomes

$$L(\gamma, \alpha, p) = \ln \left(\sum_{k=1}^{K} p_k \mathcal{L}_{1k}(\gamma) \mathcal{L}_{2k}(\gamma, \alpha_3) \mathcal{L}_{3k}(\gamma, \alpha_3, \alpha_2) \mathcal{L}_{4k}(\gamma, \alpha_3, \alpha_2, \alpha_1) \right),$$

where p_k is the unconditional probability of being the kth type. Although L is not additively separable, at the maximization step of the ESM algorithm the log-likelihood contribution for an individual is given by

$$\begin{aligned} L(\gamma, \alpha, p) = \sum_{k=1}^{K} & Pr(k|Y, X, C, Z; \gamma, \alpha, p) \\ & \times \left[L_{1k}(\gamma) + L_{2k}(\gamma, \alpha_3) + L_{3k}(\gamma, \alpha_3, \alpha_2) + L_{4k}(\gamma, \alpha_3, \alpha_2, \alpha_1) \right], \end{aligned} \tag{4}$$

where $Pr(k|Y, X, C, Z; \gamma, \alpha, p)$ is the conditional probability that the individual is of type k. Since the conditional probabilities are taken as given in the maximization step, the sample log-likelihood based on equation (4) can be maximized sequentially. One can then estimate the model by iterating between maximizing equation (4) sequentially, updating the conditional probabilities, and updating the unconditional probabilities (the p's) with the averages of the conditional probabilities.

With the description of the model complete, we now give the values of the parameters used to generate the data. We begin with equation (1), the log-earnings equation. The error on earnings is drawn from a $N(0, 1)$ distribution. The premium for attending college, γ_C, is set to 0.2. There are two types, with the earnings premium for the second type set at 0.5. The vector X contains nine variables, each drawn from a $N(0, 1)$ distribution, and a constant. The last three variables are set to zero if the individual did not attend college. The coefficients on the nine variables all equal 0.1, while the coefficient on the constant is 0.

Next consider utility in period one, as given by equation (3). Besides the constant term, the vector Z_1 has two elements, each drawn from a $N(0, 1)$ distribution. The coefficients (elements of α) for these two variables are both set to 1, as is the coefficient on expected future utility. The coefficient on the constant term is set at 0.4, with the utility premium for being the second type set at 0.6. The variance of ϵ_1 is 1.

In the second and third period utility functions, the coefficients on the constant term and expected future utility, the variance of the ϵ's, and the utility premium for being the second type are all the same as in period one. Z_1, Z_2 and Z_3 share the same first element as well, with a common coefficient of 1. The second variable in Z_2 is a $N(0, 1)$ variable unknown to the individual until after the first period decision. The second variable in Z_3 is the sum of the second variable in Z_2 and a $N(0,1)$ shock. The coefficient on both of these variables is 1.

For each artificial data set, we use the estimates we get under the assumption there is no selection problem as the starting values for mixture estimation. To get the starting values for the constant and the type premium in the earnings equation, we use the estimated constant from the selection-unadjusted equation and its standard deviation. Subtracting a quarter of the standard deviation from the selection-unadjusted constant gives the starting value of the selection-adjusted constant, while the earnings premium for being the second type is initialized at half of the standard deviation. The starting value of the utility premium for being the second type is set to zero.