Lecture 5. Heterogeneity, Mixture Distributions and the Expectation Maximization Algorithm

Ciprian Domnisoru

Aalto University

Heterogeneity in econometrics

- The “representative consumer” and “representative firm” have been shown to lack empirical support.
- Initial microdata in economics were cross sections, and unobservables were treated as independent preference shocks.
- With the advent of panel data, distinguishing between **duration dependence** and **permanent unobserved heterogeneity** became important for formulating policy recommendations.
- Early work used parametric distributions for the permanent unobservables, but one remaining question was whether it was possible to distinguish heterogeneity from state dependence without using parametric assumptions.
- The nonparametric method of Heckman and Singer (1984) shows the distribution of unobservables can be approximated by low dimensional finite mixtures of types.

Review: mixed logit and alternatives

$U_{ij} = X_{ij}'\beta_i + \epsilon_{ij}$

- You can specify $\beta_i \mid Z_i \sim N(Z_i'\gamma, \Sigma)$
  - Evaluating the likelihood can be difficult with a large number of choices: the random coefficients have to be integrated out: computationally intensive.

- Assume the existence of a finite number of types of individuals, $\beta_i \in \{b_0, b_1, ..., b_K\}$ with $P(\beta_i = b_k \mid Z_i) = p_k$ or $P(\beta_i = b_k \mid Z_i) = \frac{exp(Z_i'\gamma_k)}{1+\sum_{l=1}^{K} exp(Z_i'\gamma_l)}$
  - Use the EM algorithm (Dempster, Laird, and Rubin, 1977)

Mixed Logit

- Mixed logit is a highly flexible model.[1]
- McFadden and Train (2000) show that mixed logit can approximate any RUM.
- 3 features of mixed logit:

1. allows for random taste variation.
2. unrestricted substitution patterns.
3. correlation of unobservables over time.

[1] Note: terminology is unfortunately unsettled. Here I use "mixed logit" as in Train; Cameron and Trivedi also call this model a random parameters logit (section 15.7)

Mixed Logit

- A mixed logit is any model where choice probabilities are given by

$$p_j = \int \frac{\exp(V_j(\beta))}{\sum_k \exp(V_k(\beta))} f(\beta) d\beta \tag{1}$$

- weighted average of the logit formula evaluated at different values of β, with the weights given by the density $f(\beta)$. Weighted average of several functions is called a mixed function.
- Concrete example:

$$V_{ji} = x_{ji}\beta_i + \epsilon_{ji}$$

  where $i =$consumer and $j =$choice and

$$\beta_i \sim N(b, W = \sum{}_{\beta})$$

- Research estimates parameters b and W along with parameters β.

Review: Unobserved Heterogeneity in Duration Models

- Duration models are very sensitive to the presence of unobserved heterogeneity, even if it is uncorrelated with the explanatory variables (unlike OLS).
- We care whether hazards are actually functions of spell length, a property called **duration dependence**, or whether the observed time variation in the aggregate hazard is just a function of unobserved heterogeneity.
- We will use an approach similar to what we introduced in the mixed logit. Random effects techniques (in biostatistics random effects models are sometimes referred to as models of **frailty**)

Review: Unobserved Heterogeneity in Duration Models

- Consider an example: half the population has λ_1=1, the other half λ_2=2.
- At the start $(t = 0)$, the average hazard is 1.5. But at t=1, the survival rate for each type z=1,2 is $exp(-\lambda_z)$. At that point, the average hazard is the weighted average between the survivors of each type:
$$\bar{\lambda}(1)= \frac{0.5exp(-1)\cdot 1+0.5exp(-2)\cdot 2}{0.5exp(-1)+0.5exp(02)} \approx 1.25$$
- The process appears to have negative duration dependence (it appears for example that finding a job is more difficult the longer the unemployment spell), but it is actually a consequence of the unobserved heterogeneity in the hazard function.

Review: Unobserved Heterogeneity in Duration Models

- Parametric solutions: multiplicative unobservable in the hazard function:

$$\lambda(t|\boldsymbol{x}) = v \cdot \lambda_0(t, \boldsymbol{\alpha})\phi(\mathbf{x},\beta) \tag{2}$$

- Need to integrate over the distribution of v. Common choice for v is $\Gamma(a, b)$
- Semiparametric approach (Heckman and Singer, 1984): nonparametric distribution for the unobserved heterogeneity. Discrete distribution: h(v)=π_k, for $v = \eta_k$, k=1...K, where the points of support η_k and their probabilities π_k are estimated.

Example: Discrete time proportional hazard models with mixture models

- Suppose there are individuals i = 1,...,N, who each enter a state (e.g. illness) at time t = 0 and are observed for j time periods, at which point each person either remains in the state (censored duration data) or dies.
- Without unobserved heterogeneity, the discrete hazard rate in period t (based on the Prentice-Gloeckler (1978) model) is

  $$h_t = 1 - exp(-exp(b_0 + X_{it} * b))$$

  where b0 is an intercept and the linear index function, $X_{it} * b$, incorporates the impact of covariates X_{it}.
- The contribution to the sample likelihood for a subject with a spell length of j periods is

  $$S(j) * (h_j/(1 - h_j))^c$$

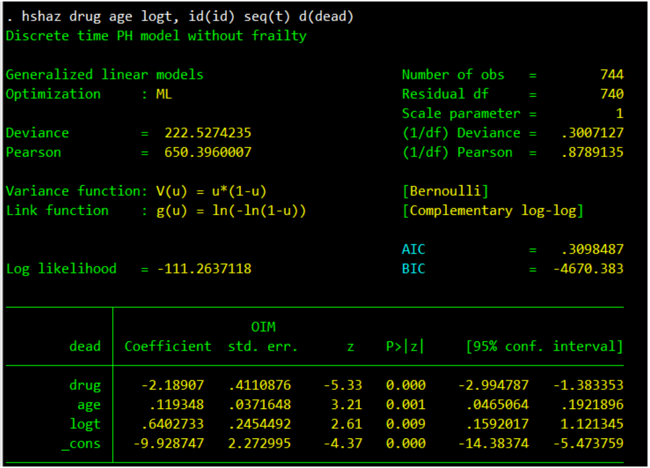
  where $S(j)$ is the probability of remaining in the state j periods, i.e. the survivor function, and c is a censoring indicator, equal to one for a completed spell and zero otherwise.

Example: Discrete time proportional hazard models with mixture models

- Suppose now that each individual belongs to one of a number of different types, and membership of each class is unobserved. This is parameterized by allowing the intercept term in the hazard function to differ across types.
- Thus, for a model with types $z = 1, ..., Z$, the hazard function for an individual belonging to type z is: $h_{zt} = 1 - exp(-exp(m_z + b0 + X_{it} * b))$ and the probability of belonging to type z is p_z.
- The m_z characterize the discrete points of support of a multinomial distribution ('mass points'), with m_1 normalized to equal zero and $p_1 = 1 - \sum_{z=2}^{Z} p_z$. The z th mass point equals $m_z + b_0$.
- The contribution to the sample likelihood of a subject with observed duration j is:

$$\mathrm{L} = \sum_{z=1}^{Z}[p_z * S_z(j) * (h_{zj}/(1 - h_{zj}))^c]$$

Example: **hshaz** command



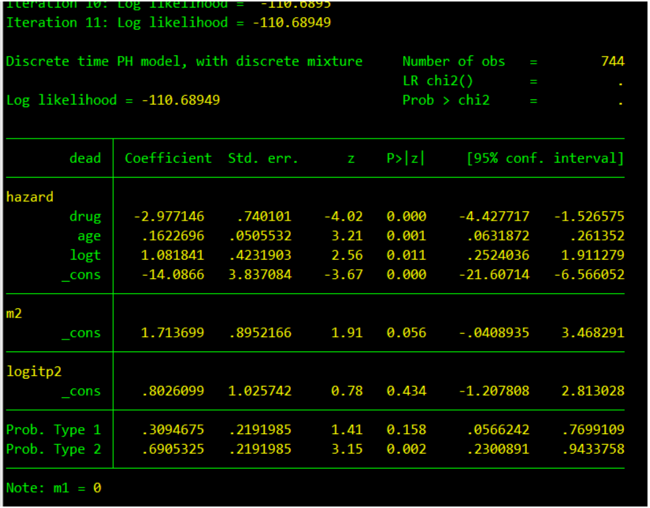
```
. hshaz drug age logt, id(id) seq(t) d(dead)
Discrete time PH model without frailty

Generalized linear models                         Number of obs     =        744
Optimization     : ML                             Residual df       =        740
                                                  Scale parameter   =          1
Deviance         =  222.5274235                   (1/df) Deviance   =   .3007127
Pearson          =  650.3960007                   (1/df) Pearson    =   .8789135

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(-ln(1-u))            [Complementary log-log]

                                                  AIC               =   .3098487
Log likelihood   = -111.2637118                   BIC               =  -4670.383

------------------------------------------------------------------------------
             |                 OIM
        dead | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
        drug |   -2.18907   .4110876    -5.33   0.000    -2.994787   -1.383353
         age |    .119348   .0371648     3.21   0.001     .0465064    .1921896
        logt |   .6402733   .2454492     2.61   0.009     .1592017    1.121345
       _cons |  -9.928747   2.272995    -4.37   0.000    -14.38374   -5.473759
------------------------------------------------------------------------------
```



```
Iteration 10:  Log likelihood = -110.6895
Iteration 11:  Log likelihood = -110.68949

Discrete time PH model, with discrete mixture    Number of obs    =        744
                                                 LR chi2()        =          .
Log likelihood = -110.68949                      Prob > chi2      =          .

------------------------------------------------------------------------------
        dead | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
hazard       |
        drug |  -2.977146    .740101    -4.02   0.000    -4.427717   -1.526575
         age |   .1622696   .0505532     3.21   0.001     .0631872     .261352
        logt |   1.081841   .4231903     2.56   0.011     .2524036    1.911279
       _cons |   -14.0866   3.837084    -3.67   0.000    -21.60714   -6.566052
-------------+----------------------------------------------------------------
m2           |
       _cons |   1.713699   .8952166     1.91   0.056    -.0408935    3.468291
-------------+----------------------------------------------------------------
logitp2      |
       _cons |   .8026099   1.025742     0.78   0.434    -1.207808    2.813028
-------------+----------------------------------------------------------------
Prob. Type 1 |   .3094675   .2191985     1.41   0.158     .0566242    .7699109
Prob. Type 2 |   .6905325   .2191985     3.15   0.002     .2300891    .9433758
------------------------------------------------------------------------------
Note: m1 = 0
```

EM algorithm

In general, Expectation-maximization (EM) algorithms are procedures for maximizing a log likelihood function when standard procedures are numerically difficult or infeasible.

- The procedure was introduced by Dempster, Laird, and Rubin (1977) as a way of handling missing data. However, it is applicable far more generally and has been used successfully in many fields of statistics.
- In our application, the missing information consists of the type (or class) share probabilities.

EM algorithm: Motivational example

- Two-component mixture model:

$$W_i = (B_i, Y_{1,i}, Y_{2,i})' \textit{ for } i = 1, ..., n. \tag{3}$$

Suppose that

$$\{W_i\} \equiv i.i.d. \tag{4}$$

and suppose that

$$Y_{1,i} \sim N(\mu_1, \sigma_1^2),\, Y_{2,i} \sim N(\mu_2, \sigma_2^2),$$

$$B_i = \begin{cases} 1 & \textit{with prob } p \\ 0 & \textit{with prob } 1 - p \end{cases}$$

Moreover, let

$$Y_i = (1 - B_i)Y_{1,i} + B_i Y_{2,i} \tag{5}$$

and assume that B_i, $Y_{1,i}$, and $Y_{2,i}$ are mutually independent and that we only observe Y_i and not B_i, $Y_{1,i}$, and $Y_{2,i}$ separately

Probability density function

The pdf of Y_i is given by

$$g_Y(Y) = (1 - p)\phi_{\theta_1}(y) + p\phi_{\theta_2}(y), \tag{6}$$

where for s = 1, 2

$$\theta_s = (\mu_s, \sigma_s^2)', \tag{7}$$

Log-likelihood function

$$\theta = \begin{pmatrix} p \\ \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} p \\ \mu_1 \\ \sigma_1^2 \\ \mu_2 \\ \sigma_2^2 \end{pmatrix} \ and \ Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}. \tag{8}$$

- Log-likelihood Function:

$$l(\theta, Y) = \sum_{i=1}^{n} \ln\Big\{(1-p)\phi_{\theta_1}(y) + p\phi_{\theta_2}(y)\Big\}. \tag{9}$$

EM Algorithm (1/2)

To understand the idea behind the expectation-maximization algorithm, consider the case where we could observe the values of B_i (the complete data case). Then, the problem would be a lot easier, since if $B_i = 1$; then, Y_i comes from model 2; otherwise, it comes from model 1. Hence, if the values of B_i are observable, then the probability density function of the data would take the form

$$\begin{aligned} & f(B_i, Y_i|\theta) \\ & = f(Y_i|B_i, \theta_1, \theta_2) f(B_i|p) \\ & = \left\{ \left[\phi_{\theta_1}(y_i) \right]^{(1-B_i)} \left[\phi_{\theta_2}(y_i) \right]^{B_i} \right\} \left\{ [1-p]^{(1-B_i)} p^{B_i} \right\} \end{aligned} \tag{10}$$

The likelihood function can be written down as:
$L_0(\theta, B, Y) = \prod_{i=1}^{n} \phi_{\theta_1}(y_i)^{(1-B_i)} \phi_{\theta_2}(y_i)^{B_i} [1-p]^{(1-B_i)} p^{B_i}$

EM algorithm (2/2)

Since in reality the values of B_i are typically unknown, the EM algorithm proceeds in an iterative manner, substituting for each Bi with its expected value

$$\begin{aligned}
\gamma_i(\theta) &= E[B_i|\theta, Y] \\
&= E[B_i|\theta, Y_i] \quad \text{(by independence)} \\
&= Pr(B_i = 1|\theta, Y_i) \\
&= \frac{f(B_i = 1 \cap Y_i)|\theta)}{f(Y_i|\theta)} \\
&= \frac{f(B_i = 1|p)f(Y_i|B_i = 1, \theta)}{f(Y_i, \theta)} \\
&= \frac{p\phi_{\theta_2}(y_i)}{(1-p)\phi_{\theta_1}(y_i) + p\phi_{\theta_2}(y_i)}. \qquad (11)
\end{aligned}$$

- The quantity $\gamma_i(\theta) = E[B_i|\theta, Y]$ is often called the **responsibility** of model 2 for observation i.

...With conditional expectations

Taking conditional expectation of the log-likelihood function, we get

$$\begin{aligned}
&E\left[\ell_0\left(\theta,B,Y\right)\middle|\,\hat{\theta},\,Y\right] \\
&= \sum_{i=1}^{n}\left\{\left(1-E\left[B_i\middle|\,\hat{\theta},Y_i\right]\right)\ln\phi_{\theta_1}(y_1)+E\left[B_i\,\middle|\,\hat{\theta},Y_i\,\right]\ln\phi_{\theta_2}(y_i)\right\} \\
&\quad+\sum_{i=1}^{n}\left\{\left(1-E\left[B_i|\hat{\theta},Y_i\right]\right)\ln\left(1-p\right)+E\left[B_i\,|\hat{\theta},Y_i\right]\ln p\right\} \\
&= \sum_{i=1}^{n}\left\{\left(1-\gamma_i\left(\hat{\theta}\right)\right)\ln\phi_{\theta_1}(y_i)+\gamma_i\left(\hat{\theta}\right)\ln\phi_{\theta_2}\left(y_i\right)\right\} \\
&\quad+\sum_{i=1}^{n}\left\{\left(1-\gamma_i\left(\hat{\theta}\right)\right)\ln\left(1-p\right)+\gamma_i\left(\hat{\theta}\right)\ln p\right\} \qquad (12)
\end{aligned}$$

EM algorithm

- The EM algorithm iterates back and forth between an expectation step and a maximization step.
- Under the expectation step, we do a soft assignment of each observation to each model, i.e., the current estimates of the parameters are used to assign responsibilities according to the relative density (under each model) of the sample points.
- Under the maximization step these responsibities are used to construct a weighted log-likelihood, which we then maximize to update our estimates of the parameters.

Step by step (1/2)

- More precisely, the EM algorithm goes as follows:
  - **Step 1:** Take initial estimates of the parameters $\hat{\mu}_{1,0}$, $\hat{\sigma}^2_{1,0}$, $\hat{\mu}^2_{2,0}$, $\hat{\sigma}^2_{2,0}$, $\hat{p}_0$ (to be specified below)
  - **Step 2:** (Expectation Step) In the k^{th} step, compute the responsibilities

$$\hat{\gamma}_{i,k} = \frac{\hat{p}_{k-1}\phi_{\hat{\theta}_{2,k-1}}(y_i)}{(1-\hat{p}_{k-1})\,\phi_{\hat{\theta}_{1,k-1}}(y_i) + \hat{p}_{k-1}\phi_{\hat{\theta}_{2,k-1}}(y_i)} \; \textit{for } i = 1, \ldots, n; \tag{13}$$

where

$$\hat{\theta}_{1,k-1} = \left(\hat{\mu}_{1,k-1}, \hat{\sigma}^2_{1,k-1}\right)' \text{ and } \hat{\theta}_{2,k-1} = \left(\hat{\mu}_{2,k-1}, \hat{\sigma}^2_{2,k-1}\right). \tag{14}$$

Step by step (2/2)

- **Step 3:** (Maximization Step) Compute the weighted means and variances for the $(k+1)^{th}$ step as

$$\hat{\mu}_{1,k} = \frac{\sum_{i=1}^{n} \left(1 - \hat{\gamma}_{i,k}\right) y_i}{\sum_{i=1}^{n} \left(1 - \hat{\gamma}_{i,k}\right)}, \hat{\mu}_{2,k} = \frac{\sum_{i=1}^{n} \hat{\gamma}_{i,k} y_i}{\sum_{i=1}^{n} \hat{\gamma}_{i,k}}, \hat{\sigma}^2_{1,k} = ..., \hat{\sigma}^2_{1,k} = .. \tag{15}$$

- Also compute the mixing probabilities as

$$\hat{p}_{k+1} = \frac{1}{n} \sum_{i=1}^{n} \hat{\gamma}_{i,k}. \tag{16}$$

- **Step 4:** Iterate steps 2 and 3 until convergence.

Remark

- Initial estimates $\hat{\mu}_{1,0}$ and $\hat{\mu}_{2,0}$ could be made by simply choosing two of the y_i's. $\hat{\sigma}^2_{1,0}$ and $\hat{\sigma}^2_{2,0}$ could both be set equal to the overall sample variance

  $$\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 \tag{17}$$

  and the initial mixing proportion $\hat{p}_0$ can be set to 0.5.

General Formulation

- The more general EM algorithm goes as follows:
- **Step 1:** Take initial estimate of the parameter vector $\hat{\theta}^{(0)}$
  Step 2: (Expected Step) In the k^{th} step, compute the responsibilities

  $$Q\left(\theta^{'}\ ,\hat{\theta}\ ^{(k-1)}\right) = E\left[\ell_0\left(\theta^{'},W\right) Y,|\hat{\theta}\ ^{(k-1)}\right]. \tag{18}$$

  as a function of the dummy argument $\theta^{'}$.
  Step 3: (Maximization Step) Determine $\hat{\theta}^{(k)}$ as

  $$\hat{\theta}\ ^{(k)} = \arg\ \max_{\theta'}\ Q\left(\theta^{'}\ ,\hat{\theta}\ ^{(k-1)}\right) \tag{19}$$

  Step 4: Iterate 2 and 3 until convergence.