

Ec241a

Econometrics

Spring 2004

UC Berkeley Department of Economics

MAXIMUM LIKELIHOOD ESTIMATION IV:
UNOBSERVED HETEROGENEITY, MIXTURE MODELS
AND THE EM ALGORITHM (LANCASTER, 1979)

In the Lancaster paper the Weibull extension of the exponential model fits the data considerably better. This can be confirmed by any of the tests considered. Everything suggests that as an individual is unemployed for longer, the hazard rate goes down. In other words, the chances of finding a job during the next day fall over the duration of the unemployment spell. Specifically, Lancaster estimates α to be -0.23, with a standard error of 0.09.

Lancaster then notices an interesting phenomenon. As he increases the number of covariates, the estimated Weibull shape coefficient increases from a negative number to something closer to zero, the value corresponding to the exponential model. Might it be that if he had even more covariates than he actually includes in his most general model that the estimated Weibull parameter is actually equal to zero? In other words, could it be true that the durations really do have exponential distributions but that the unobserved covariates are making it look like there is negative duration dependence? This is a question of parameter heterogeneity or unobserved heterogeneity. Recall that the model is

$$h(y|x, \beta, \alpha) = (\alpha + 1)y^\alpha \exp(x'\beta).$$

Consider the following example. Suppose there are two types of people in the population, “movers” and “stayers”, movers making up a fraction p of the population and stayers the remaining $1 - p$. Movers, when they find themselves unemployed, have a constant hazard of $\lambda_m = 2$. Stayers are not so good at finding a job and when they find themselves unemployed their hazard rate is constant at a lower rate of $\lambda_s = 1$. Let us see what happens with the population hazard if we treat the population as a homogenous one. The cdf of the durations

Table 1: ESTIMATES OF WEIBULL PARAMETER FROM LANCASTER STUDY

$\hat{\alpha}$	regressors
-0.33	none
-0.26	log age
-0.232	log age, log unemployment rate
-0.227	log age, log unemployment rate, log replacement ratio

for the movers is

$$F_m(y) = 1 - \exp(-2y),$$

and for the stayers:

$$F_s(y) = 1 - \exp(-y).$$

With movers making up a fraction p of the population, the probability of Y less than y is

$$\begin{aligned} F(y) &= Pr(Y \leq y) = E[Pr(Y \leq y | \text{type})] \\ &= Pr(\text{mover}) \cdot Pr(Y \leq y | \text{mover}) + Pr(\text{stayer}) \cdot Pr(Y \leq y | \text{stayer}) \\ &= p \cdot (1 - \exp(-2y)) + (1 - p) \cdot (1 - \exp(-y)) \\ &= 1 - p \cdot \exp(-2y) - (1 - p) \cdot \exp(-y). \end{aligned}$$

The survivor function is

$$S(y) = 1 - F(y) = p \cdot \exp(-2y) + (1 - p) \cdot \exp(-y).$$

The hazard function is equal to minus the derivative of the logarithm of the survivor function:

$$h(y) = -\frac{\partial \ln S}{\partial y}(y) = p(y) \cdot 2 + (1 - p(y)) \cdot 1,$$

where the fraction of movers in the subpopulation of people with durations exceeding y is

$$p(y) = \frac{p \cdot \exp(-2y)}{p \cdot \exp(-2y) + (1-p) \cdot \exp(-y)}.$$

At $y = 0$, the hazard equals $p \cdot 2 + (1-p) \cdot 1$, the simple average of the two hazard functions. As $y \rightarrow \infty$, it declines monotonically to 1, the lower of the two hazards. If you were to estimate a Weibull model on data from such a population, the Weibull parameter would be negative, suggesting erroneously that there is negative duration dependence.

Based on this argument, Lancaster worries about the potential presence of unobserved heterogeneity. To test for this there are a number of options. One can specify a parametric model allowing the parameter β_0 , the intercept in the $x'\beta$ part to have some distribution. Lancaster assumes that β_0 has the following form, $\beta_0 = \beta_{00} + v_i$, where v_i has a Gamma distribution with mean 1 and variance σ^2 , independent of the covariates X_i . Formally,

$$f(y|x, \alpha, \beta, \sigma^2) = \int (\alpha + 1)t^\alpha v \exp(x'\beta) \exp(-y^{\alpha+1} v \exp(x'\beta)) f(v|\sigma^2) dv.$$

One can then test the hypothesis of no unobserved heterogeneity by doing any of the tests we discussed before, Likelihood Ratio, Wald, Lagrange Multiplier, or Hausman–Wu, with the null hypothesis $H_0 : \sigma^2 = 0$ against the alternative $H_a : \sigma^2 > 0$.

Let us look at this in a slightly simpler context. Suppose Y_i , for $i = 1, \dots, N$, is assumed to have an exponential distribution with parameter λ :

$$f(y|\lambda) = \lambda \exp(-y\lambda), \quad y > 0.$$

Under the alternative model Y_i has an exponential with a unit-specific value of λ , say λ_i , which has some unknown distribution $g(\lambda|\mu, \sigma^2)$ with mean μ and variance σ^2 :

$$f(y|\mu, \sigma^2) = \int_{\lambda} \lambda \exp(-y\lambda) g(\lambda|\mu, \sigma^2) d\lambda.$$

We wish to test the hypothesis that the variance of λ is equal to zero, in other words, whether there is any heterogeneity in the coefficient or not. This seems difficult to do for general unknown $g(\cdot)$, so let us assume a particular specification for $g(\cdot)$. To separate out the issue

of heterogeneity from the issue of the average hazard, let us model the density of y as:

$$f(y|\lambda_0, \sigma^2) = \nu\lambda_0 \exp(-y\lambda_0\nu)g(\nu|\sigma^2)d\nu,$$

where ν is assume to have mean one and variance σ^2 , without loss of generality, and we still wish to test the hypothesis that the variance of ν , σ^2 , is equal to zero.

Now let us parametrize the distribution of ν :

$$Pr(\nu = 1 - \sqrt{\eta}) = Pr(\nu = 1 + \sqrt{\eta}) = 1/2,$$

for $0 \leq \eta < 1$. The mean of ν is one for all values of η , and the variance is η . Hence our null hypothesis now corresponds to

$$H_0 : \eta = 0,$$

against the alternative

$$H_1 : \eta \neq 0.$$

Let us do a score test for this hypothesis, to avoid having to estimate the model under the alternative. The density function is

$$f(y|\lambda_0, \eta) = \frac{1}{2}\lambda_0(1 + \sqrt{\eta}) \exp(-y\lambda_0(1 + \sqrt{\eta})) + \frac{1}{2}\lambda_0(1 - \sqrt{\eta}) \exp(-y\lambda_0(1 - \sqrt{\eta})).$$

The part of the score function corresponding to η is

$$\begin{aligned} \mathcal{S}(y, \lambda_0, \eta) = & \left[\frac{1}{4}\lambda_0\eta^{-1/2} \exp(-y\lambda_0(1 + \sqrt{\eta})) - \frac{1}{4}\lambda_0^2 y(1 + \eta^{1/2})\eta^{-1/2} \exp(-y\lambda_0(1 + \sqrt{\eta})) \right. \\ & \left. - \frac{1}{4}\lambda_0\eta^{-1/2} \exp(-y\lambda_0(1 - \sqrt{\eta})) + \frac{1}{4}\lambda_0^2 y(1 - \eta^{1/2})\eta^{-1/2} \exp(-y\lambda_0(1 - \sqrt{\eta})) \right] \\ & / \left[\frac{1}{2}\lambda_0(1 + \sqrt{\eta}) \exp(-y\lambda_0(1 + \sqrt{\eta})) + \frac{1}{2}\lambda_0(1 - \sqrt{\eta}) \exp(-y\lambda_0(1 - \sqrt{\eta})) \right]. \end{aligned}$$

At $\eta = 0$ the denominator of the score function is

$$\lambda_0 \exp(-y\lambda_0).$$

The numerator is messier because of the factors $\eta^{-1/2}$ which would amount in the limit (as $N \rightarrow \infty$) to dividing by zero. Write the numerator as

$$\left[\frac{1}{4} \lambda_0 \exp(-y \lambda_0 (1 + \sqrt{\eta})) - \frac{1}{4} \lambda_0^2 y (1 + \eta^{1/2}) \exp(-y \lambda_0 (1 + \sqrt{\eta})) \right. \\ \left. - \frac{1}{4} \lambda_0 \exp(-y \lambda_0 (1 - \sqrt{\eta})) + \frac{1}{4} \lambda_0^2 y (1 - \eta^{1/2}) \exp(-y \lambda_0 (1 - \sqrt{\eta})) \right] / \eta^{1/2}.$$

Here both numerator and denominator of the score function go to zero as η goes to zero. We apply l'Hospital's rule of differentiating both numerator and denominator with respect to η . Evaluating this at $\eta = 0$ this leads for the numerator to

$$-2y \lambda_0^2 \exp(-y \lambda_0) + y^2 \lambda_0^3 \exp(-y \lambda_0).$$

Dividing by the denominator $\lambda_0 \exp(-y \lambda_0)$ gives us the limiting score function

$$\mathcal{S}_\eta(y, \lambda_0, \eta \rightarrow 0) = -2y \lambda_0 + y^2 \lambda_0^2.$$

To calculate the score for λ given $\eta = 0$, we first evaluate the log of the density at $\eta = 0$ to get:

$$f(y|\lambda, \eta = 0) = \frac{1}{2} \lambda (1 + \sqrt{\eta}) \exp(-y \lambda (1 + \sqrt{\eta})) + \frac{1}{2} \lambda (1 - \sqrt{\eta}) \exp(-y \lambda (1 - \sqrt{\eta})) \Big|_{\eta=0} \\ = \lambda \exp(-y \lambda).$$

The score at $\eta = 0$ is

$$\mathcal{S}_\lambda(y, \lambda_0, \eta = 0) = \frac{1}{\lambda} - y.$$

Hence estimating the information matrix is straightforward: estimate λ under the null hypothesis as $\hat{\lambda}_r = 1/\bar{y}$, estimate the information matrix using the outerproduct of the first derivatives (i.e., using $\hat{\mathcal{I}}_4$), and average the score for η at the restricted mle for λ . Using the most popular for for the LM test in this setting, we would get

$$LM = \iota'_N \mathcal{S} (\mathcal{S}' \mathcal{S})^{-1} \mathcal{S}' \iota_N,$$

where \mathcal{S} is the $N \times 2$ matrix with i th row equal to $(1/\hat{\lambda} - y_i, -2y_i \hat{\lambda} + y_i^2 \hat{\lambda}^2)$.

Let us return to the Lancaster paper and consider estimation of the more general model where the intercept in the hazard has a nondegenerate distribution. For this type of mixture model the previous algorithms discussed for maximization of the likelihood function work very poorly. A much more powerful approach is based on an algorithm called the Expectation–Maximization Algorithm (EM algorithm, Dempster, Laird, Rubin, 1974; Tanner, 1996). It alternates between calculation the expectation of missing components (in this case the mixture component) and maximizing the likelihood function as if the missing components are observed.

The big advantage of the algorithm is its stability, and widespread applicability. The disadvantage is that it can be very slow compared to algorithms that directly maximize the likelihood function.

Example

We illustrate the EM algorithm here for two point mixture of exponential distribution. Suppose the density of y is

$$f(y|\lambda_0, \lambda_1, p) = (1 - p)\lambda_0 \exp(-y\lambda_0) + p\lambda_1 \exp(-y\lambda_1).$$

We can think of this as the conditional density of y given ν being equal to

$$f(y|\nu) = \lambda_0^{1-\nu} \lambda_1^\nu \exp(-y\lambda_0^{1-\nu} \lambda_1^\nu),$$

with $\nu \in \{0, 1\}$, and $Pr(\nu = 1) = p$. If we observed ν we would just split the sample by the value of ν and estimate

$$\begin{aligned} \hat{p} &= \sum \nu / N, \\ \hat{\lambda}_0 &= \sum_{i=1}^N (1 - \nu_i) / \sum_{i=1}^N (1 - \nu_i) y_i, \\ \hat{\lambda}_1 &= \sum_{i=1}^N \nu_i / \sum_{i=1}^N \nu_i y_i. \end{aligned}$$

This is the M–step of the EM algorithm.

The E step of the EM algorithm calculates the expectation of ν_i given the current estimates of the parameters:

$$\hat{\nu}_i = \frac{p\lambda_1 \exp(-y_i\lambda_1)}{p\lambda_1 \exp(-y_i\lambda_1) + (1-p)\lambda_0 \exp(-y_i\lambda_0)}.$$

Given the estimates of ν_i estimates of p , λ_0 and λ_1 are obtained. Then ν_i is re-estimated, and the algorithm iterates between the E and M steps till the whole thing converges. \square

Let us consider a second example.

Example

Let $Y + 1, \dots, Y_N$ have an exponential distribution with hazard rate λ . Let D_1, \dots, D_N be a censoring indicator at C . So, we observe D_i and $T_i = \min(Y_i, C)$ for $i = 1, \dots, N$. If we actually observed all the failure times, the maximum likelihood estimator for λ would be $1/\bar{y}$. However, for the censored observations, we do not observe Y_i . Nevertheless, we can estimate it:

$$E[Y|D = 1, T] = C + 1/\lambda.$$

So the E step is:

$$\hat{y}_i = (1 - d_i) \cdot t_i + d_i(t_i + 1/\lambda).$$

The M-step is

$$\lambda = 1 / \left(\frac{1}{N} \sum_{i=1}^N \hat{y}_i \right).$$

The EM algorithm then iterates between the two steps till things converge. In this case of course we could find the maximum likelihood estimator directly, but this approach still works in cases where we could not do that (e.g., with covariates). \square

Heckman and Singer (1984) worry about the sensitivity of the estimates of the coefficients of the hazard function to the specification of the distribution of the unobserved heterogeneity. They consider a Weibull model

$$h(y|x, \beta, \alpha) = (\alpha + 1) \cdot y^\alpha \exp(x'\beta),$$

with heterogeneity in the intercept of the hazard. Using a real data set they find that depending on the specification of the distribution of the heterogeneity one can get widely varying estimates. See the attached table. Since then simulation studies have shown that if the Weibull part of the specification is correct, the specification of the distribution of the unobserved heterogeneity is not all that important.

REFERENCES

DEMPSTER, A., N. LAIRD, AND D. RUBIN, (1974), "Maximum Likelihood from Incomplete Data via the EM Algorithm", (with discussion), *Journal of the Royal Statistical Society*, Series B, Vol. 39, 1-38.