OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 57, 1 (1995)
0305-9049

PRACTITIONER'S CORNER

Easy Estimation Methods for Discrete-Time Duration Models

Stephen P. Jenkins†

I. INTRODUCTION

For empirical analysis of duration data, discrete-time models have several advantages over continuous-time models. For example, one can straightforwardly estimate — without writing special programs — discrete-time models combining both time-varying covariates and flexible specifications of duration dependence.[1] The recent introduction of duration analysis modules into widely-available econometric software packages has reduced these relative advantages, but these modules remain limited. In particular, no module allows modification of the likelihood function to take account of the type of sampling scheme used.[2] This paper shows how the '"serious but occasional" applied econometrician' (MacKie-Mason, 1992, p. 165) can estimate discrete-time duration models taking account of some empirically-important sampling schemes and do so using readily-available packages.

Although inferences about some underlying population are the usual goal, our data are often derived, not from a random sample of that population, but from a sample of people flowing into or out of a particular state (flow sampling) or from a sample of those occupying a state at a given time (stock sampling). If we do not control for this — e.g. by continuing to use existing

†This research was supported by a Nuffield Foundation Social Science Research Fellowship. I thank John Ermisch, Peter Lambert, John Micklewright, Chris Orme and Robert Wright for helpful discussions.

[1] Multiple occurrences of tied failure times are also easily handled.

[2] LIMDEP, STATA, and BMDP all have modules for continuous-time parametric log-linear models with time-varying covariates. STATA and BMDP have modules for the continuous-time semi-parametric Cox models with time-varying covariates. (LIMDEP's Cox module does not allow time-varying covariates.) LIMDEP also has a module for a Weibull model with unobserved heterogeneity (but not time-varying covariates as well), and a module for the discrete-time semi-parametric Han-Hausman ordered-logit model (with time-varying covariates). SABRE estimates discrete-time semi-parametric logistic hazard models with unobserved heterogeneity; more about this later.

package estimation modules — estimates may be contaminated by a form of sample selection bias.[3]

To illustrate my case, I focus first on the sampling scheme in which respondents are randomly selected from a stock at one date and then interviewed after some time interval has elapsed. This scheme, labelled stock sampling with observation over a fixed interval by Lancaster (1990, p. 193), is common. For example, Lancaster (1979) models unemployment durations for a sample of the unemployment stock drawn at a date during 1973 who were interviewed some five weeks later. Jenkins (1993) models social assistance benefit durations for a sample drawn on 1 April 1989 of the stock of lone mothers receiving benefit who were then interviewed during the summer of 1989. Notice that the 'average' spell length in these stock samples is likely to be longer than the 'average' spell length amongst the population we are interested in — an illustration of the sample selection bias problem.

The paper's results generalize to other sampling schemes where the likelihood function requires conditioning on survival to a date between sample selection and interview. For example, Narendrenathan and Stewart (1993) use a sample of men registering as unemployed in 1978–79 who were interviewed approximately six, 16, and 52 months later. The nature of their income data led Narendrenathan and Stewart to model unemployment durations conditional on remaining unemployed for more than four weeks.

The empirically useful 'trick' I wish to communicate is that the discrete-time duration model based on data derived from sampling schemes of the type described can be estimated as a regression model for a binary dependent variable, and hence estimated using widely-available packages. Lancaster (1979) pioneered the analysis of stock-sampled data observed over a fixed interval using a Weibull model incorporating unobserved heterogeneity, but his model requires there to be no time-varying covariates and it cannot be estimated with standard duration analysis package modules. I build on an approach set out by Allison (1982) in a paper which deserves to be better known amongst economists. Allison explains the 'trick' in the context of random samples but does not discuss the likelihood function modifications required for other types of sampling scheme. Hence the principal contribution of this paper is to show that the approach he espouses may be extended.

In Section II I set out the discrete-time hazard model and show how the complex sequence likelihood function appropriate for the sampling schemes described above may be rewritten as a 'standard' (non-sequence) likelihood, and hence easily estimated. Section III summarizes the results and the three steps (data reorganization, selection, estimation) which are required to estimate the model using widely-available software packages. Some illustrative program code is provided in the Appendix.

[3] Chesher and Lancaster (1983) and Lancaster (1990) emphasize the difference between duration distributions derived from population samples, flow samples and stock samples.

II. REWRITING 'SEQUENCE' LIKELIHOOD FUNCTIONS IN EASILY-ESTIMABLE FORMS

To exposit the model I shall suppose we are interested in modelling the length of Income Support (IS) spells for lone mothers and have available a random stock sample of lone mothers receiving IS which was drawn on 1 April 1989 (as in Jenkins, 1993). Index the lone mothers in the sample by $i = 1, \ldots, n$, and describe the passage of calendar time (months, say) in terms of the set of positive integers. Month $t = 1$ is the earliest month in which there was a respondent receiving IS (and is before the sample selection month). Let $t = \tau$ index the sample selection month — April 1989 in the example — and so by construction each respondent receives IS at $t = \tau$.[4]

Each respondent is then interviewed some months later, where the length of the interval between the drawing of the sample and the interview is exogenously fixed. Some respondents remain IS recipients throughout the interval between sample selection and interview (contributing censored IS duration data), while some respondents leave IS during the interval (contributing completed IS duration data). Define $\delta_i = 1$ for those with completed spells and $\delta_i = 0$ for those still receiving IS when interviewed.[5] Let $t = \tau + s_i$ index the month in which the IS spell finishes if $\delta_i = 1$ and index the interview month if $\delta_i = 0$. Each respondent i thus contributes s_i months of IS spell data from the interval between sample selection and interview.

The distribution of durations is modelled via the probabilities of ending a spell at each value of t: there is a one-to-one relationship between these probabilities — 'hazard rates' — and the probabilities of having completed spell durations of different lengths — summarized by the 'survivor function'.[6]

More precisely, the discrete-time hazard rate h_{it} is

$$h_{it} = \text{prob}(T_i = t \mid T_i \geq t; \quad X_{it}) \tag{1}$$

where X_{it} is a vector of regressor variables ('covariates') which may vary with time. T_i is a discrete random variable representing the time at which the end of the spell occurs; it is T_i's distribution which is of primary interest. It can be shown that

$$\text{prob}(T_i = t) = h_{it} \cdot \prod_{k=1}^{t-1} (1 - h_{ik}) = [h_{it}/(1 - h_{it})] \cdot \prod_{k=1}^{t} (1 - h_{ik}). \tag{2}$$

[4] I assume that the 'month' is the natural unit of time in which to measure benefit spells — whereas the 'week' is in fact nearer the truth in this example — and that each explanatory variable is constant within each month. Alternatively, view the discrete-time model as an approximation to some underlying continuous-time model. Both assumptions raise the issue of whether the grouping of time intervals introduces serious aggregation bias. My view is that it usually does not: see, for example, the reassuring results from the extensive analysis of Bergström and Eden (1992). In my own work too, corresponding discrete- and continuous-time duration models have always provided similar estimates and implications (Jenkins, 1990).

[5] This definition assumes that the data are censored immediately before the end of the interval between selection and interview.

[6] See Kalbfleisch and Prentice (1980) or Lancaster (1990) for further details.

and

$$\text{prob}(T_i > t) = \prod_{k=1}^{t} (1 - h_{ik}). \tag{3}$$

To motivate the derivation of the sample likelihood, let us consider two illustrative cases.

Consider first a respondent, i, who is still on IS when interviewed in July 1989 ($\delta_i = 0$). The probability of her remaining on IS between the start of her spell and July 1989 (i.e. between $t = 1$ and $t = \tau + 3$) is

$$(1 - h_{i,\text{July}})(1 - h_{i,\text{June}})(1 - h_{i,\text{May}})(1 - h_{i,\text{April}})(1 - h_{i,\text{March}})(\ldots)(1 - h_{i,1}). \tag{4}$$

If the data had been derived from a random population sample, this unconditional survivor probability would also represent i's contribution to the sample likelihood. With stock sampling and observation over a fixed interval, the likelihood must be modified.

What we require is an expression for the probability of i remaining on IS between the start of her spell and July 1989, *conditional on not having left IS before the end of March 1989* (the condition which made her eligible for selection in the sample). Using the definition of a conditional probability, this is given by

$$\frac{(1 - h_{i,\text{July}})(1 - h_{i,\text{June}})(1 - h_{i,\text{May}})(1 - h_{i,\text{April}})(1 - h_{i,\text{March}})(\ldots)(1 - h_{i,1})}{(1 - h_{i,\text{March}})(\ldots)(1 - h_{i,1})}$$

$$= (1 - h_{i,\text{July}})(1 - h_{i,\text{June}})(1 - h_{i,\text{May}})(1 - h_{i,\text{April}}). \tag{5}$$

Similarly, the probability that a respondent, j, completes her IS spell in July 1989, *conditional on not having left IS before the end of March 1989* ($\delta_j = 1$) is

$$\frac{(h_{j,\text{July}})(1 - h_{j,\text{June}})(1 - h_{j,\text{May}})(1 - h_{j,\text{April}})(1 - h_{j,\text{March}})(\ldots)(1 - h_{j,1})}{(1 - h_{j,\text{March}})(\ldots)(1 - h_{j,1})}$$

$$= (h_{j,\text{July}})(1 - h_{j,\text{June}})(1 - h_{j,\text{May}})(1 - h_{j,\text{April}}). \tag{6}$$

In both examples, the conditioning of the survivor probability is handled remarkably simply, via a 'cancelling' of terms. The conditional survivor probability, and hence likelihood contribution, depends only on hazard rates and data for the months at risk between sample selection and interview.[7]

This 'cancelling' result also applies to other types of sampling scheme where derivation of the likelihood function requires conditioning on survival

[7] For an alternative approach to estimation using data from stock samples, see Nickell (1979). The advantage of Nickell's method is that it is potentially more efficient than the method described here since it utilizes the data on spell months prior to the sample selection date; its disadvantages are its strong, potentially debatable, assumptions about inflows, and a very complicated sequence likelihood to program and maximize. For stock samples with individuals observed only once, Nickell's method is the only one available.

to a date between the sample selection date and the interview. For example, in the Narendrenathan and Stewart (1993) case the likelihood contribution of each unemployed man depends on data for the weeks between spell week 5 and the interview. Moreover, the methods outlined can easily handle samples where s_i is unknown ('left censoring') if one is prepared to assume that there is no duration dependence in the hazard rate: inspect (5) and (6).[8]

Let us summarize the 'cancelling' result for these sampling schemes more formally. The conditional probability of observing the event history of someone with an uncompleted spell at interview is

$$\text{prob}(T_i > t + s_i \mid T_i > \tau - 1) = \prod_{t=\tau}^{\tau+s_i} (1 - h_{it}), \tag{7}$$

and the conditional probability of observing the event history of someone completing a spell between the 'conditioning' month and interview is

$$\text{prob}(T_i = t + s_i \mid T_i > \tau - 1) = [h_{i\tau+s_i}/(1 - h_{i\tau+s_i})] \prod_{t=\tau}^{\tau+s_i} (1 - h_{it}). \tag{8}$$

Hence the likelihood of observing the event history data for the whole sample is

$$\mathcal{L} = \prod_{i=1}^{n} \left[[h_{i\tau+s_i}/(1 - h_{i\tau+s_i})] \prod_{t=\tau}^{\tau+s_i} (1 - h_{it}) \right]^{\delta_i} \left[\prod_{t=\tau}^{\tau+s_i} (1 - h_{it}) \right]^{1-\delta_i} \tag{9}$$

and the corresponding log-likelihood function is

$$\log \mathcal{L} = \sum_{i=1}^{n} \delta_i \cdot \log[h_{i\tau+s_i}/(1 - h_{i\tau+s_i})] + \sum_{i=1}^{n} \sum_{t=\tau}^{\tau+s_i} \log(1 - h_{it}). \tag{10}$$

Equation (9) is an example of a 'sequence' likelihood function — note the product of terms for each t — and is very difficult to maximize directly without advanced programming skills.

To derive the easy estimation method, I define a variable $y_{it} = 1$ if $t = \tau + s_i$ and $\delta_i = 1$, and $y_{it} = 0$ otherwise. For stayers, $y_{it} = 0$ for all spell months; for exiters, $y_{it} = 0$ for all spell months except the exist month. Using this indicator variable, the log-likelihood function can be rewritten as

$$\log \mathcal{L} = \sum_{i=1}^{n} \sum_{t=\tau}^{\tau+s_i} y_{it} \cdot \log[h_{it}/(1 - h_{it})] + \sum_{i=1}^{n} \sum_{t=\tau}^{\tau+s_i} \log(1 - h_{it}) \tag{11}$$

[8] My thanks to John Micklewright for pointing this out. A similar property holds for continuous-time duration models: see Tuma and Hannan (1984, p 131).

which has the same form as the 'standard' log-likelihood function for regression analysis of a binary variable, in this case y_{it}, and where the unit of analysis is now the spell month.[9]

In sum, the complex sequence likelihood can be transformed into a 'standard' likelihood for a data set which is differently organized. Creation of the re-organized data set, plus creation of time-varying covariates, is a simple matter using readily-available software (more on this in the Appendix).

All that is required to complete the specification of the likelihood, and hence to estimate the model, is an expression for the hazard rate, plus a software package which can estimate regression models for binary dependent variables.

One specification for the hazard rate is the *complementary log-log* one

$$h_{it} = 1 - \exp\{-\exp[\theta(t) + \beta' X_{it}]\} \Leftrightarrow \log[-\log(1 - h_{it})] = \theta(t) + \beta' X_{it}. \quad (12)$$

This has the property that the resulting model is the discrete-time counterpart of an underlying continuous-time proportional hazards model (Prentice and Gloeckler, 1978).[10] On the other hand, there is no particular reason why, with economic duration data, hazards should be proportional. Widely-available software packages to estimate the dichotomous variable regression model in the complementary log-log case include GLIM and STATA.

A commonly-used non-proportional hazard specification for the hazard function is the *logistic* one:

$$h_{it} = 1/\{1 + \exp[-\theta(t) - \beta' X_{it}]\} \Leftrightarrow \log[h_{it}/(1 - h_{it})] = \theta(t) + \beta' X_{it}. \quad (13)$$

In this case, the model likelihood has exactly the same form as that for a standard binary logit regression model (applied to the reorganized data set), and so can be estimated with a very large number of software packages. So too could a *probit* specification.

The logistic model turns out to be very similar to the complementary log-log one in most empirical applications. The reason is that the logistic model converges to a proportional hazard model as the hazard rate becomes increasingly small, and the rate is indeed sufficiently small in most applications.[11] This fact, combined with the wide availablity of logit regression programs makes the logistic hazard model especially useful.

[9] Each respondent contributes as many observations to this reorganized data set as she has months at risk of exiting, i.e. s_i.

[10] A proportional hazards model is one for which absolute differences in covariates imply proportionate differences in hazard rates, i.e. multiplicative scaling of the baseline hazard function $\theta(t)$: see Kalbfleisch and Prentice (1980) or Lancaster (1990) for further details. If $\theta(t)$ is summarized by set of constants differing for each t, the discrete-time duration model is fully semi-parametric, just like the continuous-time Cox model though, by contrast, estimates of the baseline hazard are derived directly as part of the estimation procedure in the discrete-time model (another advantage).

[11] See Bergström and Eden (1992) for some recent evidence about this.

What is there is unobserved heterogeneity? Let us generalize the hazard rate specification (13) to

$$\log[h_{it}/(1-h_{it})] = \theta(t) + \beta' X_{it} + \varepsilon_i \qquad (14)$$

where ε_i is an unobserved individual-specific error term with zero mean, uncorrelated with the Xs. The standard way of estimating such a model is to assume the ε_i follow some parametric distribution and to integrate them out of the likelihood. The complication for feasible estimation in the current context is that, if the i.i.d. assumptions about the ε_i are assumed to hold in the population, the assumptions need not hold in the stock sample (Chesher and Lancaster, 1983). Tractable estimation is possible, however, if the assumptions about ε_i are made 'with respect to our sample and not with respect to the population' (Narendrenathan and Stewart, 1993, p. 370). Although one cannot rewrite the sequence likelihood as a standard likelihood function in this case, this modified logistic model can nonetheless still be straightforwardly estimated using the SABRE package (Barry *et al.*, 1990), assuming the ε_i are Normally distributed. Again one uses the reorganized data set in which spells at risk are the unit of analysis.

III. SUMMARY OF RESULTS AND IMPLEMENTATION

Recent literature has emphasized that estimation of duration models should be modified to take account of the sampling scheme generating the data used. However the estimation methods required cannot be implemented using the duration modules in currently available packages and the advanced programming skills consequently required are a barrier to implementing the methods. This paper aims to help break down such barriers by showing how an important class of models can be estimated easily using widely-available, non-specialist, software packages.

The paper has focused on discrete-time models and the case where data are derived from respondents randomly selected from a stock at one date and then interviewed after some time interval has elapsed. The results generalize to other sampling schemes where the likelihood function requires conditioning on survival to a date between sample selection and interview.

There are three steps involved in implementing the proposed estimation strategy.

(1) *Data Step.* Reorganize the data set, converting the unit of analysis from the 'respondent' to the 'spell month at risk of event occurrence'. At the same time, create the time-varying covariates, if required.
(2) *Selection Step.* Select the spells corresponding to the time periods between sample selection and interview date.
(3) *Estimation Step.* Estimate the model using the reorganized data set.

In the Appendix, I elaborate on how to implement the Data and Selection Steps with some widely-available packages.

The same estimation strategy can of course be used when the sample likelihood does not have to be modified to take account of the sampling scheme. The only difference is that the Selection Step is omitted.[12]

University of Essex

Date of Receipt of Final Manuscript: December 1993

REFERENCES

Allison, P. D. (1982). 'Discrete-time methods for the analysis of event histories', in Leinhardt, S. (ed.), *Sociological Methodology 1982*, Jossey-Bass Publishers, San Francisco, pp. 61–97.

Barry, J., Francis, B. and Davies, R. (1990). *SABRE: Software for the Analysis of Binary Recurrent Events. A guide to users*, Centre for Applied Statistics, University of Lancaster, Lancaster.

Bergström, R. and Edin, P.-A. (1992). 'Time aggregation and the distributional shape of unemployment duration', *Journal of Applied Econometrics*, Vol. 7, pp. 5–30.

Chesher, A. and Lancaster, T. (1983). 'The estimation of models of labor market behavior', *Review of Economic Studies*, Vol. 50, pp. 609–24.

Han, A. and Hausman, J. A. (1990). 'Flexible parametric estimation of duration and competing risk models', *Journal of Applied Econometrics*, Vol. 5, pp. 1–28.

Jenkins, S. P. (1990). 'The length of lone mothers' spells on Supplementary Benefit/Income Support', *Report to the Department of Social Security.*

Jenkins, S. P. (1993). 'How long do lone mothers remain on Income Support?, mimeo.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*, Wiley, New York.

Lancaster, T. (1979). 'Econometric methods for the duration of unemployment', *Econometrica*, Vol. 47, pp. 939–56.

Lancaster, T. (1990). *The Econometric Analysis of Transition Data*, Cambridge University Press, Cambridge.

MacKie-Mason, J. K. (1992). 'Econometric software: a user's view', *Journal of Economic Perspectives*, Vol. 6, pp. 165–88.

Meyer, B. D. (1990). 'Unemployment insurance and unemployment spells', *Econometrica*, Vol. 58, pp. 757–82.

Narendrenathan, W. and Stewart, M. B. (1993). 'How does the benefit effect vary as unemployment spells lengthen?', *Journal of Applied Econometrics*, Vol. 8, pp. 361–81.

Nickell, S. (1979). 'Estimating the probability of leaving unemployment', *Econometrica*, Vol. 47, 1249–66.

Prentice, R. L. and Gloeckler, L. A. (1978). 'Regression analysis of grouped survival data with application to breast cancer data', *Biometrics*, Vol. 34, pp. 57–67.

Tuma, N. B. and Hannan, M. T. (1984). *Social Dynamics: Models and Methods*, Academic Press, Orlando, FL.

[12] Hence in (11) the summations would be over all t (Allison, 1982). Notice, for example, that the rather daunting-looking likelihood used by Meyer (1990, equation 5) in his study of unemployment durations can be rewritten as a complementary log-log version of equation (11), and hence estimated straightforwardly using the method proposed here.

APPENDIX: IMPLEMENTATION OF THE DATA AND SELECTION STEPS

To elaborate the data and selection steps cited in section III, I shall assume that, in the original data set, the unit of analysis is the respondent, and that for each respondent there are counterparts to the following variables: ID (respondent unique identifier); DUR (observed spell length); CENSORED (censoring indicator = 1 if duration censored, = 0 otherwise); START (date when spell began, in integer form); <FIXED COV> (a vector of covariates, assumed fixed throughout the spell).

In SPSS, use the INPUT PROGRAM, LEAVE, and LOOP commands to reorganize the data set in the required way. The LEAVE command ensures that each spell month observation for a given respondent in the reorganized data set contains the respondent's fixed covariates from the original data. The LOOP command creates an integer variable for each respondent, call it T, which indexes each different spell month. Within the loop, one can simultaneously generate time-varying covariates using T and START — not only covariates varying with duration *per se*, but also variables varying with calendar time. (Sometimes it is easier to do this after the Selection step.) The Selection step itself is simple: use SELECT.

To implement the Estimation step using the new data set with other programs, either WRITE the reorganized data set out from SPSS in ASCII format, or EXPORT it as a SPSS portable file and use a utility such as STAT/TRANSFER to convert this to a format readable by the other programs.

The DATA and Selection Steps can also be straightforwardly implemented using programs such as SAS and STATA. For example, in STATA, reorganize the data set using the one command: expand DUR. Generate the index T with the following commands: sort ID, ge x = ID = = ID[_n − 1], quietly by ID: ge T = 1 + sum(x). Then generate the time-varying covariates. Implement the Selection step using keep and/or drop.

The following is some illustrative SPSS program code for the Data and Selection steps which should be modified according to context.

```
********************  DATA  STEP  ************************************
INPUT PROGRAM
DATA LIST FILE = <name of input data file (file in ascii format) >
  <variable list, including ID, DUR, CENSORED, START, <FIXED COV> >
LEAVE <variable list, including ID, DUR, CENSORED, START, <FIXED COV> >
LOOP T = 1 TO DUR
*****************  CREATE YIT  ********************
IF (T < DUR) YIT = 0
IF (T = DUR AND CENSORED = 1) YIT = 0
IF (T = DUR AND CENSORED = 0) YIT = 1
**********  CREATE TIME-VARYING COVARIATES ********
** DUMMY VARIABLES TO ESTIMATE SEMI-PARAMETRIC BASELINE HAZARD **
COMPUTE THETA2 = 0
IF (T = 2) THETA2 = 1
COMPUTE THETA3 = 0
IF (T = 3) THETA3 = 1
. . etc
** ALTERNATIVELY, FOR A PARAMETRIC BASELINE HAZARD, E.G. 'WEIBULL' .    **
COMPUTE LOGT = LN(T)
** VARIABLES VARYING WITH CALENDAR-TIME **
IF ( (START + T - 1 = ...) AND ( ... ) ) TVC1 = ....
... etc
** VARIABLES VARYING WITH DURATION **
IF ( (T = ...) AND ( ... ) ) TVC2 = ....
. . etc
*************************************************************************
END CASE
END LOOP
END INPUT PROGRAM
...
********************  SELECTION STEP  ********************************
SELECT IF ( (START + T - 1 GE   . ) AND ( ... ) )
... etc
```

Competing risks models can also be estimated using the methods espoused in this paper, because the competing risk conditional likelihood can be partitioned into a sum of terms corresponding to single cause-specific likelihoods. See W. Narendranathan and M. B. Stewart (1993). 'Modelling the probability of leaving unemployment: Competing risk models with flexible base-line hazards', *Applied Statistics*, Vol. 42, pp. 63–83.