Estimation of discrete time (grouped duration data) proportional hazards models: pgmhaz

Stephen P. Jenkins <stephenj@essex.ac.uk>
ESRC Research Centre on Micro-Social Change
University of Essex, Colchester CO4 3SQ, U.K.

Hazard rate models are widely used to model duration data in a wide range of disciplines, from bio-statistics to economics. The aim of this insert is to supplement the portfolio of duration data analysis tools which Stata provides.

pgmhaz estimates, by maximum likelihood, two discrete time (grouped duration data) proportional hazards regression models, one of which incorporates a gamma mixture distribution to summarize unobserved individual heterogeneity (or 'frailty'). Covariates may include regressor variables summarizing observed differences between persons (either fixed or time-varying), and variables summarizing the duration dependence of the hazard rate. With suitable definition of covariates, models with a fully non-parametric specification for duration dependence may be estimated; so too may parametric specifications. **pgmhaz** thus provides a useful complement to **cox** and **st stcox**, **weibull** and **st stweib**.

The two models estimated are: (1) the Prentice-Gloeckler (1978) model; and (2) the Prentice-Gloeckler (1978) model incorporating a gamma mixture distribution to summarize unobserved individual heterogeneity, as proposed by Meyer (1990). These are referred to as Model 1 and Model 2 respectively below. The versions of the Prentice-Gloeckler-Meyer hazard models estimated are as described by Stewart (1996), and my exposition of the models draws heavily on his paper.

The models

Suppose there are individuals $i = 1,...,N$, who each enter a state (e.g. unemployment) at time $t = 0$. The (instantaneous) hazard rate function for person i at time $t > 0$ is assumed to take the proportional hazards form

(1) $\lambda_{it} = \lambda_0(t).\exp(X_{it}'\beta)$

where $\lambda_0(t)$ is the baseline hazard function which may take a parametric or non-parametric form (see below); exp(.) is the exponential function; X_{it} is a vector of covariates summarizing observed differences between individuals at t; and β is a vector of parameters to be estimated. The associated continuous time survivor function is

(2) $$S(t;X_{it}) = \exp\left[-\int_0^t \lambda(\tau;X_{it})d\tau\right] = \exp\left\{-\exp[X_{it}'\beta + \log(H_t)]\right\}$$

where $H_t = \int_0^t \lambda_0(\tau)d\tau$ is the integrated baseline hazard at t.

The underlying continuous durations are only observed in disjoint time intervals $[0=a_0,a_1),[a_1,a_2),[a_2,a_3),...,[a_{k-1},a_k=\infty)$. (Alternatively durations are intrinsically discrete.) Covariates may vary between time intervals but are assumed to be constant within each of them.

The probability of exit in the j-th interval for person i is

(3) $\text{prob}\{T \in [a_{j-1},a_j)\} = S(a_{j-1};X_{it}) - S(a_j;X_{it})$

and the survivor function at the start of the jth interval is

(4) $\text{prob}\{T \geq a_{j-1}\} = S(a_{j-1};X_{it})$.

The hazard of exit in the jth interval is thus given by

(5) $h_j(X_{it}) \equiv \text{prob}\{T \in [a_{j-1},a_j) | T \geq a_{j-1}\} = 1 - [S(a_j;X_{it})/S(a_{j-1};X_{it})]$.

Given the proportional hazards assumption, the survivor function has the same form as (2). This can be conveniently re-written in the discrete case as:

(6) $S(a_j;X_{it}) = \exp[-\exp(X_{it}'\beta + \delta_j)]$ where $\delta_j = \log(H_{it})$ for $j = 1,\ldots,k$.

To simplify, all intervals are now assumed to be of unit length (e.g. a week, or a month), so the recorded duration for each person i corresponds to the interval $[t_i\text{-}1,t_i)$. Persons are also recorded as either having left the state during the interval, or as still remaining in the state. The former group, contributing completed spell data, are identified using with censoring indicator $c_i = 1$. For the latter group, contributing right-censored spell data, $c_i = 0$. Observe that the number of intervals comprising a censored spell is defined here to include the last interval within which the person is observed.

With these assumptions, the log-likelihood can be written

(7) $$\log L(\beta,\delta) = \sum_{i=1}^{n}\left\{c_i \log[S(t_i-1;X_{it}) - S(t_i;X_{it})] - (1-c_i)\log S(t_i;X_{it})\right\}$$

with $S(.)$ as in (6). This expression can be rewritten in terms of the hazard function as:

(8) $$\log L = \sum_{i=1}^{n}\left\{c_i \log\left\{h_{ti}(X_{iti})\prod_{s=1}^{t_i-1}[1-h_s(X_{is})]\right\} + (1-c_i)\log\left\{\prod_{s=1}^{t_i}[1-h_s(X_{is})]\right\}\right\}$$

where the discrete time hazard in the jth interval is

(9) $h_j(X_{ij}) = 1 - \exp[-\exp(X_{ij}'\beta + \gamma_j)]$ with $\gamma_j = \log\int_{a_{j-1}}^{a_j}\lambda_0(\tau)d\tau$.

This specification allows for a fully non-parametric baseline hazard with a separate parameter for each duration interval (depending on the duration data sample—see below), in which case the γ_j can be interpreted as the logarithm of the integral of the baseline hazard over the relevant interval. Alternatively, the γ_j may be described by some semi-parametric or parametric function, call it $\theta(j)$.

If one defines an indicator variable $y_{it} = 1$ if person i exits the state during the interval $[t\text{-}1,t]$, $y_{it} = 0$ otherwise, then the log-likelihood in (8) can be re-written in sequential binary response form:

(10) $$\log L = \sum_{i=1}^{n}\sum_{j=1}^{t_i}\left\{y_{ij}\log h_j(X_{ij}) + (1-y_{ij})\log[1-h_j(X_{ij})]\right\}$$.

This is the version of the Model 1 log-likelihood which is estimated by **pgmhaz**.

Model 2 incorporates a Gamma distributed random variable to describe unobserved (or omitted) heterogeneity between individuals. (For a discussion of and comparison with other mixed proportional hazards models, see Stewart, 1996.)

The instantaneous hazard rate is now specified as (cf. (1)):

(11) $\lambda_{it} = \lambda_0(t).\varepsilon_i.\exp(X_{it}'\beta) = \lambda_0(t).\exp[X_{it}'\beta + \log(\varepsilon_i)]$

where ε_i is a Gamma distributed random variate with unit mean and variance $\sigma^2 \equiv v$, and the discrete-time hazard function corresponding to (11) is now

(12) $h_j(X_{ij}) = 1 - \exp\{-\exp[X_{ij}'\beta + \gamma_j + \log(\varepsilon_i)]\}$.

Conveniently, the survivor function for the augmented model has a closed form expression (see Meyer 1990 for details), and thence so too does the log-likelihood function.

The Model 2 log-likelihood function is:

(13) $$\log L = \sum_{i=1}^{N} \log\{(1-c_i).A_i + c_i.B_i\}$$

where

$$A_i = \left[1 + v \sum_{j=1}^{t_i} \exp\left[X_{ij}'\beta + \theta(j)\right]\right]^{-(1/v)}, \text{ and}$$

$$B_i = \left[1 + v \sum_{j=1}^{t_i-1} \exp\left[X_{ij}'\beta + \theta(j)\right]\right]^{-(1/v)} - A_i, \text{ if } t_i > 1, \text{ or}$$

$$= 1 - A_i, \text{ if } t_i = 1,$$

where $\theta(j)$ is a function describing duration dependence in the hazard rate, including the non-parametric baseline hazard specification. The functional form for $\theta(j)$ is chosen by the user and specified by defining appropriate covariates—see below. Model 1's log-likelihood function is the limiting case as $v \to 0$.

For suitably organised data, the log-likelihood function for Model 1 is the same as the log-likelihood for a generalized linear model of the binomial family with complementary log-log link: see Allison (1982) or Jenkins (1995). Model 1 is estimated by ML using Stata's **glm** command. Model 2 is estimated using Stata's **ml deriv0** command, with starting values for β taken from Model 1's estimates. Given the potential fragility of models incorporating unobserved heterogeneity, estimates for both models are always reported.

Syntax

The syntax of **pgmhaz** is

pgmhaz *covariates* [**if** exp] [**in** *range*], **id(***idvar***) dead(***deadvar***) seq(***seqvar***)**
[**lnvar0(#) eform level(#) nolog trace nocons**]

Options

lnvar0(#**)** specifies the value for ln(v) which is used as the starting value in the maximization. The default is -1 (i.e. $v \approx 0.37$).

eform reports coefficients transformed to relative risk format, i.e. exp(β) rather than β. Standard errors and confidence intervals are similarly transformed. **eform** may be specified at estimation or when redisplaying results.

level(#**)** specifies the significance level, in percent, for confidence intervals of the parameters.

nolog suppresses the iteration logs.

trace reports the current value of the estimated parameters of Model 2 at each iteration. See [R] maximize.

nocons specifies no intercept term in the index function $X_{ij}'\beta$.

Saved results include the global macros set by **ml post**
plus S_1 Model 2 log-likelihood value at maximum, and
S_2 Model 1 log-likelihood value at maximum.
Access to estimated coefficients and standard errors is available in the usual way: see [U] 20.5 Accessing coefficients and [R] matrix get.

Data organization and mandatory variables

The data set must be organised before estimation so that, for each person, there are as many data rows as there are time intervals at risk of the event occuring for each person. Given the definitions above, this means t_i rows for each person $i = 1,...,N$. In effect an unbalanced panel data set-up is required. This data organisation is closely related to that required for estimation of Cox regression models with time-varying covariates. **expand** is useful for putting the data in this form: see [R] expand, and the example below. Also see the 'data step' discussion in Jenkins (1995).

Three variables must be defined by the user:

id(*idvar***)** specifies the variable uniquely identifying each person, i.

seq(*seqvar***)** is the variable uniquely identifying each time period at risk for each person. For each i, *seqvar* is the integer sequence 1,2,..., t_i.

dead(*deadvar***)** summarizes censoring status during each time interval at risk, and corresponds to the indicator variable y_{it} described earlier. If $c_i = 0$, *deadvar* = 0 for all $j = 1,2,...,t_i$; if $c_i = 1$, *deadvar* = 0 for all $j = 1,2,...,t_i$-1, and *deadvar* = 1 for $j = t_i$.

Examples of how to construct these variables are given below.

Example

This illustration uses the Cancer data set (**cancer.dta**), supplied with Stata, and described in the Stata version 5.0 Reference Manual P-Z, p. 257. The data provides information about survival times for 48 participants in a cancer drug trial. Of the 48 people, 28 receive the experimental drug treatment (**drug** = 1) and 20 receive the control treatment (**drug** = 0). The participants range in age from 47 to 67 years. We wish to analyse time until death, measured in months. The variable **studytim** records either the month of their death or the last month that they were known to be alive (the maximum value in the data is 39). The persons known to have died have variable **died** = 1 (contributing completed duration data); those still alive have **died** = 0 (contributing censored duration data).

First we **use** the data and recode **drug** so that it matches the Manual example:

```
. use cancer
(Patient Survival in Drug Trial)
. replace drug = 0 if drug == 1
```

```
(20 real changes made)
. replace drug = 1 if drug > 1
(28 real changes made)
```

To run **pgmhaz** we must re-organise the data set and create the mandatory variables. To understand what is going on, look at how the data for the first four people is currently organised, and compare this with their data in the re-organized data set later on.

```
. ge id = _n  /* create unique person identifier */
. list id studytim died drug age in 1/4
            id  studytim      died      drug       age
  1.         1         1         1         0        61
  2.         2         1         1         0        65
  3.         3         2         1         0        59
  4.         4         3         1         0        52
```

Now expand the data set so that there's one data row per person per month at risk of dying, and create *seqvar* and *dead*:

```
. expand studytim
(696 observations created)
. sort id
. quietly by id: ge seqvar = _n
. quietly by id: ge dead = died & _n==_N
```

Compare this data format with the earlier one, taking the same four persons:

```
. list id studytim seqvar died dead age if id <= 4
            id  studytim     seqvar      died       dead       age
  1.         1         1          1         1          1        61
  2.         2         1          1         1          1        65
  3.         3         2          1         1          0        59
  4.         3         2          2         1          1        59
  5.         4         3          1         1          0        52
  6.         4         3          2         1          0        52
  7.         4         3          3         1          1        52
```

At this stage, with the data re-organised into person-month form, it would be straightforward to generate time-varying covariates. None are available in **cancer.dta** however. The illustrative estimations use the fixed covariates **drug** and **age**, capturing observed heterogeneity, and use the gamma mixing distribution to capture unobserved heterogeneity.

The first models estimated using **pgmhaz** assume duration dependence in the hazard rate is summarised by a parametric 'Weibull' specification. This is achieved by including a covariate defined as the logarithm of *seqvar*. (If the estimated coefficient on this regressor is greater than zero, the hazard increases monotonically; if less than zero, it decreases monotonically.) The Model 1 estimates are precisely those which would be produced by the command

```
. gen logd = ln(seqvar)
. glm deadvar logd drug age, f(b) l(c)
```

except that in **pgmhaz** I have added output giving log-likelihood values. [Incidentally, the logistic hazard counterpart to this proportional hazards model could have been estimated with **logit** applied to the same re-organised data set (Allison 1982, Jenkins 1995).]

```
. pgmhaz logd drug age, id(id) s(seqvar) d(dead)

(1) PGM hazard model without unobserved heterogeneity

Iteration 1 : deviance =  298.3504
Iteration 2 : deviance =  237.2426
Iteration 3 : deviance =  224.1963
Iteration 4 : deviance =  222.5673
Iteration 5 : deviance =  222.5275
Iteration 6 : deviance =  222.5274
Iteration 7 : deviance =  222.5274
```

```
Residual df  =       740                                 No. of obs =       744
Pearson X2   =  650.3937                                 Deviance   =  222.5274
Dispersion   =  .8789105                                 Dispersion =  .3007127

Bernoulli distribution, cloglog link
------------------------------------------------------------------------------
    dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    logd |   .6402733   .2448109      2.615   0.009       .1604526    1.120094
    drug |   -2.18907   .4125618     -5.306   0.000      -2.997676   -1.380463
     age |    .119348   .0369335      3.231   0.001       .0469596    .1917364
   _cons |  -9.928747   2.262543     -4.388   0.000      -14.36325   -5.494243
------------------------------------------------------------------------------
Log likelihood (-0.5*Deviance) = -111.26371
   Cf. log likelihood for intercept-only model (Model 0) = -128.86467
   Chi-squared statistic for Model (1) vs. Model (0) = 35.201924
   Prob. > chi2(3) = 1.104e-07

(2) PGM hazard model with gamma distributed unobserved heterogeneity

Iteration 0:  Log Likelihood = -112.22135
Iteration 1:  Log Likelihood = -111.09624
Iteration 2:  Log Likelihood = -111.08967
Iteration 3:  Log Likelihood = -111.08965
Iteration 4:  Log Likelihood = -111.08965

PGM hazard model with gamma heterogeneity           Number of obs    =     744
                                                    Model chi2(3)    =       .
                                                    Prob > chi2      =       .
Log Likelihood =   -111.0896470

------------------------------------------------------------------------------
    dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
hazard   |
    logd |   .8664734   .4787207      1.810   0.070       -.071802    1.804749
    drug |  -2.578879   .8275313     -3.116   0.002      -4.200811   -.9569476
     age |    .141193   .0569466      2.479   0.013       .0295798    .2528062
   _cons |  -11.29142   3.510489     -3.216   0.001      -18.17185   -4.410984
---------+--------------------------------------------------------------------
ln_varg  |
   _cons |  -1.247006   1.845572     -0.676   0.499      -4.864262    2.370249
------------------------------------------------------------------------------
Gamma variance, exp(ln_varg) = .28736375; Std. Err. = .53035058; z = .54183734

Likelihood ratio statistic for testing models (1) vs (2) = .34812954
Prob. test statistic > chi2(1) = .55517386
```

Comparing **pgmhaz** Model 1 and Model 2 estimates, we see that the duration dependence parameter is larger in the latter. Moreover the coefficients in Model 2 on **drug** and **age** are slightly larger in absolute value. These differences are not unexpected: not accounting for unobserved heterogeneity induces an under-estimate of the extent to which the hazard rate increases with duration (or an over-estimate of the decline), and attentuates the magnitude of the impact of covariates on the hazard rate (see Lancaster 1990, chapter 4).

The size of the variance of the gamma mixture distribution relative to its standard error suggests, however, that unobserved heterogeneity is not significant in this data set. A likelihood ratio test of Model 2 versus Model 1 also suggests the same conclusion. Users should be aware though that standard likelihood ratio tests cannot, strictly speaking, be used to choose between Models 1 and 2, because the former is not a nested version of the latter.

The discrete time ‘Weibull’ estimates can be compared with estimates of a continuous time Weibull model derived using **stweib**:

```
. stset seqvar dead, id(id)

note:  making entry-time variable t0
       (within id, t0 will be 0 for the 1st observation and the
       lagged value of exit time seqvar thereafter)
```

```
   data set name:  cancer.dta
              id:  id
      entry time:  t0
       exit time:  seqvar
  failure/censor:  dead

. stweib drug age, nohr

<output omitted>

Weibull regression -- entry time t0
log relative hazard form

No. of subjects =           48                  Log likelihood =  -42.931336
No. of failures =           31                  chi2(2)        =       35.39
Time at risk    =          744                  Prob > chi2    =      0.0000

------------------------------------------------------------------------------
  seqvar |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    drug |  -2.197157    .408785     -5.375   0.000      -2.998361   -1.395953
     age |   .1202128   .0371591      3.235   0.001       .0473823    .1930433
   _cons |  -10.58395   2.326241     -4.550   0.000       -15.1433   -6.024599
---------+--------------------------------------------------------------------
    ln p |   .5203303   .1389099      3.746   0.000       .2480718    .7925887
       p |   1.682583                                     1.281552    2.209108
     1/p |   .5943242                                     .4526714    .7803039
------------------------------------------------------------------------------
```

As it happens the coefficient estimates are very similar to corresponding estimates in the discrete time 'Weibull' model without unobserved heterogeneity. The duration dependence parameters are similar too: compare 1-p = 0.683 with the coefficient on logd, 0.640.

We should be wary about drawing conclusions about duration dependence from parametric models like the 'Weibull' which tightly constrain the general shape of the baseline hazard function shape, when in fact it may be non-monotonic. Moreover it is well-known that conclusions about the significance of unobserved heterogeneity are more reliably drawn if a flexible specification for the baseline hazard has been used (for a recent discussion, see Dolton and van der Klaauw, 1995).

Let us therefore compare some models which allow for more flexibility in the shape of the baseline hazard function. One obvious reference point is the (continuous time) Cox model, estimates for which are are reported in Reference Manual P-Z, p. 257. These are reproduced with the commands:

```
. stcox drug age, nohr baseh(coxbaseh)

<output omitted>

Cox regression -- entry time t0

No. of subjects =           48                  Log likelihood =  -83.323546
No. of failures =           31                  chi2(2)        =       33.18
Time at risk    =          744                  Prob > chi2    =      0.0000

------------------------------------------------------------------------------
  seqvar |
    dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    drug |  -2.254965   .4548338     -4.958   0.000      -3.146423   -1.363507
     age |   .1136186   .0372848      3.047   0.002       .0405416    .1866955
------------------------------------------------------------------------------
```

The Cox baseline hazard function is graphed on p. 279 of Stata 4.0 Reference Manual Volume Two, and the figure suggests that the hazard increases non-monotonically with duration. (The picture can be reproduced with the command `gr coxbaseh studytim, xlab ylab`.) The estimates of the baseline hazard estimate are as follows, and there is clear evidence of non-monotonicity:

```
.
. sort seqvar
. list seqvar coxbaseh if coxbaseh~=.
```

```
        seqvar    coxbaseh
 16.         1   .00013425
 46.         1   .00013425
 70.         2   .00007571
112.         3   .00007924
149.         4   .00018067
151.         4   .00018067
196.         5    .0002276
201.         5    .0002276
253.         6   .00026013
255.         6   .00026013
301.         7   .00013608
306.         8   .00044866
307.         8   .00044866
335.         8   .00044866
380.        10    .0001891
404.        11   .00041499
429.        11   .00041499
433.        12   .00056331
435.        12   .00056331
476.        13   .00035089
526.        15   .00038132
532.        16   .00043044
559.        17   .00047959
639.        22   .00149966
641.        22   .00149966
652.        23   .00249846
662.        23   .00249846
672.        24   .00168347
680.        25   .00189012
705.        28   .00228993
734.        33   .00437874
```

Let us now compare the Cox model estimates with various discrete time proportional hazard model specifications. One example of a flexible parametric specification for the baseline hazard function is a polynomial in duration. One could

```
. gen seqvar_2 = seqvar^2
. gen seqvar_3 = seqvar^3
```

and include **seqvar**, **seqvar_2**, and **seqvar_3**, as covariates instead of **logd** in order to specify a cubic baseline hazard function.

pgmhaz also allows the estimation of fully non-parametric specifications for the baseline hazard (analogously to the Cox model). The interval-specific baseline hazard can only be identified for those duration intervals during which events (‘deaths’) occur, i.e. values of **seqvar** for which there are observations (person-months) with **dead** = 1. If there are duration intervals for which this is not true, then either one must refine the grouping on the duration dimension—the piece-wise constant model is an example of this—or one must drop the relevant person months from the estimation. (Cf. the discussion of identification of the logit model in Reference Manual K-M, p. 371-375.)

To estimate the non-parametric baseline model, first one has to create binary dummy variables corresponding to each duration interval. It is the user’s responsibility to do this and also to check identifiability. This is straightforward. We can create duration-specific dummy variables, one for each spell month at risk (the maximum number is 39 here), with the following command:

```
. quietly for 1-39, ltype(numeric): ge byte d@ = seqvar== @
```

Next we check identifiability of the baseline hazard at each duration interval with:

```
. tab seqvar dead
           | dead
    seqvar |         0         1 |     Total
-----------+----------------------+----------
         1 |        46         2 |        48
```

```
          2 |        45          1 |        46
          3 |        44          1 |        45
          4 |        42          2 |        44
          5 |        40          2 |        42
          6 |        38          2 |        40
          7 |        36          1 |        37
          8 |        33          3 |        36
          9 |        32          0 |        32
         10 |        30          1 |        31
         11 |        27          2 |        29
         12 |        24          2 |        26
         13 |        23          1 |        24
         14 |        23          0 |        23
         15 |        22          1 |        23
         16 |        20          1 |        21
         17 |        19          1 |        20
         18 |        18          0 |        18
         19 |        18          0 |        18
         20 |        16          0 |        16
         21 |        15          0 |        15
         22 |        13          2 |        15
         23 |        11          2 |        13
         24 |        10          1 |        11
         25 |         9          1 |        10
         26 |         8          0 |         8
         27 |         8          0 |         8
         28 |         7          1 |         8
         29 |         6          0 |         6
         30 |         6          0 |         6
         31 |         6          0 |         6
         32 |         6          0 |         6
         33 |         3          1 |         4
         34 |         3          0 |         3
         35 |         2          0 |         2
         36 |         1          0 |         1
         37 |         1          0 |         1
         38 |         1          0 |         1
         39 |         1          0 |         1
------------+----------------------+----------
      Total |       713         31 |       744
```

There are no deaths during months 9, 14, 18-21, 26, 27, 29-32, 34-39, and so a month-specific hazard rate cannot be estimated for these intervals.

The non-parametric baseline model is estimated by including all the relevant duration dummies, excluding observations to ensure identifiability (if necessary), and excluding the intercept using the **nocons** option. (An alternative estimation strategy would be to include the intercept term and exclude one of the interval-specific duration dummy variables.)

```
. pgmhaz d1-d8 d10-d13 d15-d17 d22-d25 d28 d33 drug age  /*
>         */ if (seqvar>=1 & seqvar<=8) | (seqvar>=10 & seqvar<=13) /*
>         */ | (seqvar>=15 & seqvar<=17) | (seqvar>=22 & seqvar<=25) /*
>         */ | seqvar==28 | seqvar==33 , /*
>         */ i(id) s(seqvar) d(dead) nocons

(1) PGM hazard model without unobserved heterogeneity

Iteration 1 : deviance =  255.2345
Iteration 2 : deviance =  206.7409
Iteration 3 : deviance =  195.2580
Iteration 4 : deviance =  193.6490
Iteration 5 : deviance =  193.5947
Iteration 6 : deviance =  193.5944
Iteration 7 : deviance =  193.5943
Iteration 8 : deviance =  193.5943

Residual df  =       550                              No. of obs =       573
Pearson X2   =  590.1132                              Deviance   =  193.5943
Dispersion   =  1.072933                              Dispersion =  .3519897

Bernoulli distribution, cloglog link
------------------------------------------------------------------------------
        dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
          d1 |  -9.321505   2.325432     -4.009   0.000      -13.87927   -4.763742
          d2 |  -9.888197   2.408603     -4.105   0.000      -14.60897   -5.167421
          d3 |  -9.841984   2.411291     -4.082   0.000      -14.56803   -5.115941
          d4 |  -9.008131   2.296365     -3.923   0.000      -13.50892   -4.507338
```

```
      d5 |  -8.758806   2.240112     -3.910   0.000      -13.14934   -4.368267
      d6 |  -8.617634   2.211921     -3.896   0.000      -12.95292    -4.28235
      d7 |  -9.269882    2.31374     -4.006   0.000      -13.80473   -4.735034
      d8 |  -8.075462   2.152716     -3.751   0.000      -12.29471   -3.856216
     d10 |  -8.931846   2.322521     -3.846   0.000       -13.4839   -4.379788
     d11 |  -8.144971   2.201254     -3.700   0.000      -12.45935   -3.830592
     d12 |  -7.819553   2.202429     -3.550   0.000      -12.13624   -3.502872
     d13 |   -8.27514   2.282109     -3.626   0.000      -12.74799   -3.802288
     d15 |  -8.190081   2.265841     -3.615   0.000      -12.63105   -3.749113
     d16 |  -8.068544   2.291659     -3.521   0.000      -12.56011   -3.576975
     d17 |  -7.959319   2.257287     -3.526   0.000      -12.38352   -3.535118
     d22 |  -6.799641   2.161635     -3.146   0.002      -11.03637   -2.562914
     d23 |  -6.231227    2.12435     -2.933   0.003      -10.39488   -2.067578
     d24 |  -6.597669   2.287659     -2.884   0.004       -11.0814    -2.11394
     d25 |  -6.481679   2.285573     -2.836   0.005      -10.96132   -2.002038
     d28 |  -6.293319   2.302273     -2.734   0.006      -10.80569   -1.780946
     d33 |  -5.654198   2.350609     -2.405   0.016      -10.26131    -1.04709
    drug |   -2.45515   .4668781     -5.259   0.000      -3.370214   -1.540086
     age |   .1208959    .037461      3.227   0.001       .0474738     .194318
------------------------------------------------------------------------------
Log likelihood (-0.5*Deviance) = -96.797174
   Cf. log likelihood for intercept-only model (Model 0) = -120.56974
   Chi-squared statistic for Model (1) vs. Model (0) = 47.545131
   Prob. > chi2(22) = .0012454

(2) PGM hazard model with gamma distributed unobserved heterogeneity

Iteration 0:  Log Likelihood =   -97.7371
(nonconcave function encountered)
Iteration 1:  Log Likelihood = -97.178695
Iteration 2:  Log Likelihood = -96.672111
Iteration 3:  Log Likelihood =  -96.66698
Iteration 4:  Log Likelihood = -96.666692
Iteration 5:  Log Likelihood = -96.666692

PGM hazard model with gamma heterogeneity           Number of obs    =     573
                                                    Model chi2(23)   =       .
                                                    Prob > chi2      =       .
Log Likelihood =    -96.6666923

------------------------------------------------------------------------------
        dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
hazard       |
          d1 |  -10.80181   3.943248     -2.739   0.006      -18.53043   -3.073185
          d2 |  -11.30775   3.896149     -2.902   0.004      -18.94406   -3.671436
          d3 |  -11.24518   3.875214     -2.902   0.004      -18.84046   -3.649898
          d4 |  -10.35178   3.703692     -2.795   0.005      -17.61088   -3.092676
          d5 |  -9.991745   3.504439     -2.851   0.004      -16.86032    -3.12317
          d6 |   -9.78455   3.385348     -2.890   0.004      -16.41971    -3.14939
          d7 |  -10.42769   3.437018     -3.034   0.002      -17.16412   -3.691257
          d8 |  -9.193994   3.290664     -2.794   0.005      -15.64358   -2.744411
         d10 |  -10.01127   3.344376     -2.993   0.003      -16.56612   -3.456411
         d11 |  -9.191331   3.235603     -2.841   0.005        -15.533   -2.849666
         d12 |  -8.820771   3.177368     -2.776   0.006       -15.0483   -2.593245
         d13 |  -9.261921   3.224425     -2.872   0.004      -15.58168   -2.942165
         d15 |  -9.145481   3.191924     -2.865   0.004      -15.40154   -2.889424
         d16 |  -8.978681   3.134165     -2.865   0.004      -15.12153    -2.83583
         d17 |  -8.819529   3.071867     -2.871   0.004      -14.84028   -2.798779
         d22 |  -7.626859   2.946399     -2.589   0.010      -13.40169   -1.852023
         d23 |  -7.076276   2.946105     -2.402   0.016      -12.85053   -1.302017
         d24 |  -7.436352    3.02082     -2.462   0.014      -13.35705   -1.515653
         d25 |  -7.276834    2.97176     -2.449   0.014      -13.10138   -1.452292
         d28 |  -7.038744   2.933167     -2.400   0.016      -12.78765   -1.289843
         d33 |  -6.276739   2.859726     -2.195   0.028       -11.8817   -.6717789
        drug |  -2.863101   .9693994     -2.953   0.003      -4.763088   -.9631126
         age |   .1468383   .0667209      2.201   0.028       .0160677    .2776089
-------------+----------------------------------------------------------------
ln_varg      |
       _cons |  -1.114756   2.041279     -0.546   0.585      -5.115589    2.886076
------------------------------------------------------------------------------
Gamma variance, exp(ln_varg) = .32799525; Std. Err. = .66952971; z = .48988902

Likelihood ratio statistic for testing models (1) vs (2) = .26096361
Prob. test statistic > chi2(1) = .60945891
```

The results suggest that unobserved heterogeneity is not significant in this context, so our preferred specification is Model 1. Parameter estimates for this model correspond quite closely to the Cox ones. In

particular, there is a close match in the pattern of variation of the baseline hazard with duration: compare the duration dummy variable coefficient estimates with the Cox model estimates listed earlier). The coefficients on **drug** and **age** are each somewhat larger in absolute value in the discrete time model compared to the Cox model.

The earlier tablulation showed that, even for the months in which there were deaths, the number of deaths was relatively small. Some additional grouping of duration intervals might therefore be considered desirable. A model with a piece-wise constant baseline hazard is an example of a compromise model which allows some non-parametric flexibility in the duration dependence specification, but may help estimation precision when there are few spell endings per duration interval. To specify a baseline hazard which is constant within six month intervals but allowed to vary between these, one would simply:

```
. ge dur1 = d1+d2+d3+d4+d5+d6
. ge dur2 = d7+d8+d9+d10+d11+d12
. ge dur3 = d13+d14+d15+d16+d17+d18
. ge dur4 = d19+d20+d21+d22+d23+d24
. ge dur5 = d25+d26+d27+d28+d29+d30
. ge dur6 = d31+d32+d33+d34+d35+d36+d37+d38+d39
```

and use the command

```
. pgmhaz dur1-dur6 drug age, i(id) s(seqvar) d(dead) nocons
```

Estimates of Models 1 and 2 for this case (not shown here) indicate that unobserved heterogeneity is not significant and there is evidence of a non-monotonic increase in the baseline hazard with duration. The coefficients on **age** and **drug** are similar to those estimated by both the Cox model and the discrete time proportional hazards model with non-parametric baseline hazard**.**

Computational and other issues

pgmhaz can be slow, or rather estimation of Model 2 can be. This is partly because the maximization procedure uses numerical derivatives, and also partly because re-organized data sets can be relatively ‘large’. Models with fully non-parametric baseline hazard function specifications also take significantly longer to estimate than models with parsimonious parametric baseline specifications. Using a Pentium P-120 PC with 32MB RAM, running Stata 5.0 for Windows 3.11 for Workgroups, the ‘Weibull’ **pgmhaz** model took about one minute to run, and the non-parametric baseline model about seven minutes. Using a different dataset from **cancer.dta**, one with 7410 person-months, a model with one covariate and thirteen duration dummy variables took about 30 minutes to complete.

The log-likelihood function for Model 2 is not globally concave, but convergence is usually achieved without problems. If there are maximization difficulties, users may find the **trace** option useful for diagnosing problems. Setting different starting values for the logarithm of the gamma variance with the **lnvar0(**#**)** option may also be helpful.

A warning. Because of the particular ordered sequence person-month structure of the data, the **if** option should be used with great care (and the **in** option should probably never be used). An **if** expression which refers to all the data rows for each person will be handled correctly: e.g. selection of an estimation sub-sample according to values of a fixed covariate. Do not select cases using an expression referring to a duration-varying variable or the results may be unpredictable. One exception to this rule arises when some observations need to be excluded to ensure identifiability of a model with a non-parametric baseline hazard function—as illustrated earlier. (In this context, there is one situation I am aware of in which the program will be incorrect if this strategy is

followed. This is when the data contain a person contributing $s > 1$ intervals to the analysis who 'dies' in the s-th interval, and there are no 'deaths' observed for any person in the sample during any of the duration intervals before s. This situation is likely to be rare.)

Acknowledgements

This work forms part of the scientific programme of the ESRC Research Centre on Micro-Social Change, supported by the UK Economic and Social Research Council and the University of Essex. I owe many thanks to Espen Bratberg, Andy Dickerson, John Ermisch, Gene Fisher, Andy Henley, Wilbert van der Klaauw, and most especially Bill Sribney and Mark Stewart, for comments and advice.

References

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

Dolton, P. and W. van der Klaauw (1995), 'Leaving teaching in the UK: a duration analysis', Economic Journal, 105(429), 431-444.

Jenkins, S.P. (1995), 'Easy estimation methods for discrete-time duration models', Oxford Bulletin of Economics and Statistics, 57(1), 129-138.

Lancaster, T. (1990), The Econometric Analysis of Transition Data, Econometric Society Monograph No. 17, Cambridge University Press, Cambridge.

Meyer, B.D. (1990), 'Unemployment insurance and unemployment spells' Econometrica, 58(4), 757-782.

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

Stewart, M.B. (1996), 'Heterogeneity specification in unemployment duration models', unpublished paper, Department of Economics, University of Warwick, Coventry, UK.