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 \( i = 1, \ldots, N \), 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

\[
\lambda_{it} = \lambda_0(t).\exp(X_{it}'\beta)
\]

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 \( \beta \) is a vector of parameters to be estimated. The associated continuous time survivor function is

\[
S(t;X_i) = \exp\left(-\int_0^t \lambda_0(\tau) d\tau\right) = \exp\left(-\exp[X_i'\beta + \log(H_i)]\right)
\]

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 \([0=a_0,a_1],[a_1,a_2],[a_2,a_3], \ldots, [a_{k-1},a_k=\infty] \). (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

\[
\text{prob}\{T \in [a_{j-1},a_j]\} = S(a_{j-1};X_{it}) - S(a_j;X_{it})
\]
and the survivor function at the start of the \( j \)th interval is

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

The hazard of exit in the \( j \)th interval is thus given by

\[
h_j(X_{it}) = \text{prob}\{T \in [a_{j-1}, a_j) \mid T \geq a_{j-1}\} = 1 - \frac{S(a_j; X_{it})}{S(a_{j-1}; X_{it})}.
\]

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:

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

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 \([t_{i-1}, t_i)\). 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

\[
\log L(\beta, \delta) = \sum_{i=1}^{n} \left\{ c_i \log[S(t_{i-1}; X_0) - S(t_i; X_0)] - (1-c_i) \log S(t_i; X_0) \right\}
\]

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

\[
\log L = \sum_{i=1}^{n} \left\{ c_i \log \left[ h(x_{it}) \prod_{s=1}^{t_i-1} [1-h(x_{is})] \right] + (1-c_i) \log \left[ \prod_{s=1}^{t_i} [1-h(x_{is})] \right] \right\}
\]

where the discrete time hazard in the \( j \)th interval is

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

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 \( \gamma_j \) can be interpreted as the logarithm of the integral of the baseline hazard over the relevant interval. Alternatively, the \( \gamma_j \) 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 \([t_{i-1}, t_i]\), \( y_{it} = 0 \) otherwise, then the log-likelihood in (8) can be re-written in sequential binary response form:

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

This is the version of the Model 1 log-likelihood which is estimated by \texttt{pgmhz}.
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)):

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

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

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

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:

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

where

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

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

\[ = 1 - A_i, \quad \text{if } t_i = 1, \]

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 \texttt{glm} command. Model 2 is estimated using Stata’s \texttt{ml deriv0} command, with starting values for \( \beta \) 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 \texttt{pgmhaz} is

\texttt{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. $\exp(\beta)$ rather than $\beta$. 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 organized before estimation so that, for each person, there are as many data rows as there are time intervals at risk of the event occurring for each person. Given the definitions above, this means $t_i$ rows for each person $i = 1, \ldots, N$. 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 $j = 1,2,\ldots,t_i$; if $c_i = 1$, deadvar = 0 for all $j = 1,2,\ldots,t_i-1$, 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
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
  +---------------------+
<table>
<thead>
<tr>
<th>id</th>
<th>studytim</th>
<th>died</th>
<th>drug</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
  +---------------------+
```

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
  +---------------------+---------------------+-------+-------+---+
<table>
<thead>
<tr>
<th>id</th>
<th>studytim</th>
<th>seqvar</th>
<th>died</th>
<th>dead</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>61</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>65</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>59</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>52</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>52</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>52</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>52</td>
</tr>
</tbody>
</table>
  +---------------------+---------------------+-------+-------+---+
```

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
bernoulli distribution, cloglog link

|     | Coef. | Std. Err. |      z | P>|z| |  [95% Conf. Interval] |
|-----|-------|-----------|-------|-----|----------------------|
| dead |       |           |       |     |                      |
| logd | 0.6403 | 0.2448    | 2.615 | 0.009 | 0.1605 - 1.1201     |
| drug | -2.1909 | 0.4126   | -5.306 | 0.000 | -2.9977 - -1.3805    |
| age  | 0.1193 | 0.0369    | 3.231 | 0.001 | 0.0470 - 0.1917     |
| _cons| -9.9287 | 2.2625   | -4.388 | 0.000 | -14.3633 - -5.4943    |

Log likelihood (-0.5*Deviance) = -111.26371
Cf. log likelihood for intercept-only model (Model 0) = -128.86467
Prob. > chi2(3) = 1.104e-07

(2) PGM hazard model with gamma distributed unobserved heterogeneity

|     | Coef. | Std. Err. |      z | P>|z| |  [95% Conf. Interval] |
|-----|-------|-----------|-------|-----|----------------------|
| hazard |       |           |       |     |                      |
| logd | 0.8664 | 0.4787    | 1.810 | 0.070 | -0.0718 - 1.8047    |
| drug | -2.5789 | 0.8275   | -3.116 | 0.002 | -4.2008 - -0.9569    |
| age  | 0.1411 | 0.0569    | 2.479 | 0.013 | 0.0295 - 0.2528     |
| _cons| -11.29 | 3.5105    | -3.216 | 0.001 | -18.1718 - -4.4109    |

ln var |       |           |       |     |                      |
| _cons | -1.2470 | 1.8456   | -0.676 | 0.499 | -4.8642 - 2.3703    |

Gamma variance, exp(ln var) = 0.28736375; Std. Err. = 0.53035058; z = 0.54183734

Comparing pghaz 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 attenuates 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)
Weibull regression -- entry time t0
log relative hazard form

No. of subjects = 48
No. of failures = 31
Time at risk = 744

| 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 -15.14330 -6.024599 |

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.

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~=

. stcox drug age, nohr baseh(coxbaseh)
Cox regression -- entry time t0
log relative hazard form

No. of subjects = 48
No. of failures = 31
Time at risk = 744

| seqvar | 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  |
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
```

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

```
. pmghaz 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

<p>| dead | Coef. | Std. Err. |    z  | P&gt;|z| |   [95% Conf. Interval] |
|------|-------|-----------|-------|------|------------------------|
| d1   | -9.321958 | 2.328835 | -4.009 | 0.000 | -13.87925 -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 |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>d5</td>
<td>-8.758806</td>
<td>2.240112</td>
<td>-3.910</td>
<td>0.000</td>
<td>-13.14934 -4.368267</td>
</tr>
<tr>
<td>d6</td>
<td>-8.617634</td>
<td>2.211921</td>
<td>-3.896</td>
<td>0.000</td>
<td>-12.95292 -4.28235</td>
</tr>
<tr>
<td>d7</td>
<td>-9.269882</td>
<td>2.31374</td>
<td>-4.006</td>
<td>0.000</td>
<td>-13.80473 -4.735034</td>
</tr>
<tr>
<td>d8</td>
<td>-8.075462</td>
<td>2.152716</td>
<td>-3.751</td>
<td>0.000</td>
<td>-12.29471 -3.856216</td>
</tr>
<tr>
<td>d10</td>
<td>-8.931846</td>
<td>2.322521</td>
<td>-3.846</td>
<td>0.000</td>
<td>-13.4839 -4.379788</td>
</tr>
<tr>
<td>d11</td>
<td>-8.144971</td>
<td>2.201254</td>
<td>-3.700</td>
<td>0.000</td>
<td>-12.45935 -3.830592</td>
</tr>
<tr>
<td>d12</td>
<td>-7.819553</td>
<td>2.282109</td>
<td>-3.626</td>
<td>0.000</td>
<td>-12.74799 -3.802288</td>
</tr>
<tr>
<td>d15</td>
<td>-8.190081</td>
<td>2.265841</td>
<td>-3.615</td>
<td>0.000</td>
<td>-12.63105 -3.749113</td>
</tr>
<tr>
<td>d16</td>
<td>-8.06844</td>
<td>2.291659</td>
<td>-3.521</td>
<td>0.000</td>
<td>-12.56011 -3.576975</td>
</tr>
<tr>
<td>d17</td>
<td>-7.959319</td>
<td>2.257287</td>
<td>-3.526</td>
<td>0.000</td>
<td>-12.38352 -3.535118</td>
</tr>
<tr>
<td>d22</td>
<td>-6.799641</td>
<td>2.161635</td>
<td>-3.146</td>
<td>0.002</td>
<td>-11.03637 -2.562914</td>
</tr>
<tr>
<td>d23</td>
<td>-6.231227</td>
<td>2.12435</td>
<td>-2.933</td>
<td>0.003</td>
<td>-10.39488 -2.067578</td>
</tr>
<tr>
<td>d24</td>
<td>-6.597669</td>
<td>2.287659</td>
<td>-2.884</td>
<td>0.004</td>
<td>-11.0814 -2.11394</td>
</tr>
<tr>
<td>d25</td>
<td>-6.481679</td>
<td>2.285573</td>
<td>-2.836</td>
<td>0.005</td>
<td>-10.96132 -2.002038</td>
</tr>
<tr>
<td>d28</td>
<td>-6.293319</td>
<td>2.302273</td>
<td>-2.734</td>
<td>0.006</td>
<td>-10.80569 -1.780946</td>
</tr>
<tr>
<td>d33</td>
<td>-5.654198</td>
<td>2.350609</td>
<td>-2.405</td>
<td>0.016</td>
<td>-10.26131 -1.04709</td>
</tr>
<tr>
<td>drug</td>
<td>-2.45515</td>
<td>.4668781</td>
<td>-5.259</td>
<td>0.000</td>
<td>-3.370214 -1.540086</td>
</tr>
</tbody>
</table>

---

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

---

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 -3.260836 |
d3 | -11.24518 | 3.875214 | -2.902 | 0.004 | -18.84046 -3.650178 |
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 -3.249358 |
d7 | -10.42769 | 3.437018 | -3.034 | 0.002 | -17.16412 -3.299578 |
d8 | -9.193994 | 3.290664 | -2.794 | 0.005 | -15.64358 -2.744411 |
d10 | -10.01127 | 3.344439 | -2.993 | 0.003 | -16.56612 -3.456411 |
d11 | -9.19331 | 3.235603 | -2.841 | 0.005 | -15.533 -2.849666 |
d12 | -8.820771 | 3.177368 | -2.776 | 0.006 | -15.0483 -2.529345 |
d13 | -9.261921 | 3.224425 | -2.872 | 0.004 | -15.58168 -2.542165 |
d15 | -8.796861 | 3.134165 | -2.865 | 0.004 | -15.12533 -2.535835 |
d16 | -8.819529 | 3.071867 | -2.871 | 0.004 | -15.0483 -2.584752 |
d17 | -9.261921 | 3.224425 | -2.872 | 0.004 | -15.58168 -2.542165 |
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 -1.515635 |
d25 | -7.276834 | 2.97176 | -2.449 | 0.014 | -13.10138 -1.652929 |
d28 | -7.037443 | 2.931367 | -2.400 | 0.016 | -12.78765 -1.289843 |
d33 | -6.276739 | 2.859726 | -2.195 | 0.028 | -11.8817 -6.717789 |
drug | -2.45515 | 0.4668781 | -5.259 | 0.000 | -3.370214 -1.540086 |
age | .1468383 | 0.0667209 | 2.201 | 0.028 | .0160677 .2776089 |

ln_varg | cons | -1.114756 | 2.044127 | -0.546 | 0.585 | -5.115589 | 2.886076 |

---

Gamma variance, exp(ln_varg) = .32799525; Std. Err. = .66952971; z = .49883902

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 tabulation 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:

```stata
    . 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

```stata
    . 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

histories’, in Sociological Methodology 1982 (S. Leinhardt, ed.),
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.
survival data with application to breast cancer data’ Biometrics, 34,
57-67.
models’, unpublished paper, Department of Economics, University of
Warwick, Coventry, UK.