UNIVERSITY OF ESSEX
INSTITUTE FOR SOCIAL AND ECONOMIC RESEARCH
Professor Stephen P. Jenkins <stephenj@essex.ac.uk>

Essex Summer School course 'Survival Analysis'
and
EC968. Part II: Introduction to the analysis of spell duration data

Lesson 7. Unobserved heterogeneity ('frailty')

Contents

1 Aim

The aim of this Lesson is to show how to estimate models incorporating unobserved heterogeneity (or 'frailty' as biostatisticians label it).

2 Introduction

The models considered in this Lesson are a generalisation of those that we considered in Lessons 5 and 6.

2.1 Model specification

For the continuous time parametric models that we estimated in Lesson 5, we now write the hazard rate for each observation as

$$\theta_v(t, X) \equiv \theta(t, X \mid v) = \theta(t, X).v$$

where $\theta(t, X)$ is the hazard function considered earlier (and assuming an absence of time-varying covariates for now). Thus unobserved differences between observations are introduced via a multiplicative scaling factor, v. This is a random variable taking on positive values, with the mean normalised to one (for identification reasons) and finite variance σ^2. A crucial assumption in these models is that v is distributed independently of X and t.

It can be shown that the frailty survivor function is related to the non-frailty one by the relationship:

$$S_v(t, X) \equiv S(t, X \mid v) = [S(t, X)]^v.$$

Thus unobserved differences also imply a scaling of the non-frailty survivor function.

Observe that, for proportional hazards models, the frailty hazard rate may be written as

$$\theta_v(t, X) \equiv \theta(t, X \mid \beta, v) = \theta_0(t).\exp(\beta'X).v = \theta_0(t).\exp(\beta'X + u)$$

or

$$\ln[\theta_v(t, X)] = \ln[\theta_0(t)] + \beta'X + u$$

where $\theta_0(t)$ is the baseline hazard function and the ‘error’ term $u \equiv \ln(v)$ which is random variable with a mean of zero.

The random variable v, or equivalently u, may be interpreted in several ways. The most common one is that it summarises the impact of ‘omitted variables’ on the hazard rate – whether the missing regressors are intrinsically unobservable or simply unobserved in the data set to hand. Alternative interpretations can be offered in terms of errors of measurement in recorded regressors or recorded survival times (see Lancaster 1990, Chapter 4).

In the discrete time proportional hazards model, the model specification follows directly from above. The standard cloglog model generalises to:

$$\text{cloglog}[p(t, X \mid \beta, v)] = D(t) + \beta'X + u$$

where $D(t)$ characterises the baseline hazard function. The logistic hazard regression model is typically generalised in an analogous way:

$$\text{logit}[p(t, X \mid \beta, v)] = D(t) + \beta'X + e$$

where the ‘error’ term e is a random variable with mean zero and finite variance. These are random intercept models where randomness is characterized using some parametric distribution (see below).

To estimate these models requires expressions for survival and density functions that do not condition on the unobserved effects for, since each individual v is unobserved, how could one write down the likelihood contribution for each observation? The way forward to specify a distribution for the v, where the distribution is characterised in terms of parameters (that can be estimated), and the unconditional survivor function is written in terms of this. This is known as ‘integrating out’ the unobserved effect. Referring to the example above, one works with survivor function $S(t, X \mid \beta, \sigma^2)$ rather than $S(t, X \mid \beta, v)$, and similarly for the density function.

In principle, any continuous distribution with positive support, mean one and finite variance, is a suitable candidate to represent the frailty distribution. For tractability reasons, however, the choice of distribution is typically limited to those that provide a closed form expression for the frailty survivor function.

For *continuous* time models the Gamma and Inverse Gaussian distributions have been the two that have been most commonly used. For the Gamma mixture model, the survivor function is given by

$$S(t, X \mid \beta, V) = (\ 1 - V.\ln[S(t)]\ )^{-1/V}$$

where $V \equiv \sigma^2$ and the non-frailty survivor function is $S(t)$. For the Weibull model, $\ln[S(t)] = -\lambda t^{\alpha}$ where $\lambda = \exp(\beta' X)$, and so in this case,

$$S(t, X \mid \beta, V) = (\ 1 + V.\lambda t^{\alpha}\ )^{-1/V}$$

and the median duration (integrating over the distribution of the frailty v) is $[(2^V - 1)/(V\lambda)]^{1/\alpha}$. One could derive predicted survival probabilities assuming that v took on a particular value, for example the mean $v = 1$. In this case, the formula for the median is the same as the standard non-frailty median formula, but of course evaluated using the parameters estimated from the model with unobserved heterogeneity.

For the Inverse Gaussian mixture model, the survivor function is given by

$$S(t, X \mid \beta, V) = \exp[\ (1/V)(1 - \{1 - 2V.\ln[S(t)]\}^{1/2})\ ].$$

For the Weibull model, the median duration (averaging over the frailty v) is $[\{\ [1 + V.\ln(2)]^2 - 1\}/(2V\lambda)]^{1/\alpha}$. Estimates of the median conditioning on particular frailty v values can also be derived, as for the Gamma model.

For the *discrete time* PH model, the Gamma distribution has been the most popular distribution. For cloglog and logistic models, it also straightforward to assume a Normal (Gaussian) distribution for u and e, respectively. (In these latter two cases, closed form expressions are not available; numerical quadrature techniques are used for the integrating out.) Prediction of survivor functions is rather more complicated than for the continuous time case (as Lesson 6 showed), as the survivor function is a product of the complements of the period-specific hazard rates. And, with frailty, these hazard rates also depend on an unobserved individual error term. The most common empirical practice has been to calculate survivor functions conditioning on a particular error term value – using the estimates of covariate coefficients from the frailty model but setting the error term equal to its mean.

There is also a literature, following the pioneering work of Heckman and Singer in the 1980s, which eschewed parametric forms for the frailty distribution. Instead non-parametric characterisations were proposed, by which an arbitrary distribution was fitted using a set of parameters representing a set of 'mass points' along the distribution's support together with the probabilities of a subject being at each mass point. Consider for illustrative purposes a discrete time proportional hazards model. When there is no frailty, the discrete hazard rate in period t is

$$h_t = 1 - \exp(-\exp(\beta_0 + \beta' X_{it}))$$

where β_0 is an intercept and the linear index function, $\beta' X_{it}$, incorporates the impact of covariates X_{it}. Suppose now that each individual belongs to one of a number of different types, and membership of each class is unobserved. This is parameterized by allowing the intercept term in the hazard function to differ across types. For example, for a model with types $z = 1, ..., Z$, the hazard function for an individual belonging to type z is:

$$hz_t = 1 - \exp(-\exp(m_z + \beta_0 + \beta' X_{it}))$$

and the probability of belonging to type z is p_z. The m_z characterize the discrete points of support of a multinomial distribution ('mass points'), with m_1 normalized to equal zero and $p_1 = 1 - \sum_{z = 2,\ldots,Z} (p_z)$. Mass point z equals $m_z + \beta_0$. This is a random intercept model where randomness is characterized using a discrete distribution.

We focus mostly on parametric representations of unobserved heterogeneity in this lesson. The particular models considered are listed in the overview provided by Section 2.3 below.

2.2 The implications of unobserved heterogeneity

What happens to parameter estimates if one (mistakenly) ignores unobserved heterogeneity? The theoretical literature has suggested several results, typically derived with reference to a continuous time PH model:

- The non-frailty model will over-estimate the degree of negative duration dependence in the (true) baseline hazard, and under-estimate the degree of positive duration dependence. (This is a selection effect. In the negative duration dependence case, observations with high v values fail faster, other things equal, so the survivors at any given survival time are increasingly composed of observations with relatively low v values and thence lower hazard rates.)
- The proportionate effect of a given regressor on the hazard rate is no longer constant and independent of survival time (in the non-frailty PH model, the proportionate effect for regressor X_k is the fixed amount β_k).
- The presence of unobserved heterogeneity attentuates the proportionate response of the hazard to variation in each regressor at any survival time. In short the estimate of a positive (negative) β_k derived from the (wrong) no-frailty model will underestimate (overrestimate) the 'true' estimate. (Lancaster, 1990, chapter 4, proves this for the case when v follows a Gamma distribution.)

The empirical literature has generally confirmed these results. There has also been discussion of the magnitude of the effects and how 'serious' the biases are in practice. Verdicts have been contingent on the choice of shape of the non-frailty hazard function and the choice of the distribution for the unobserved heterogeneity. The results from several recent papers have suggested that if a fully flexible specification for the baseline hazard function is used, then the magnitude of the biases in the non-frailty model (relative to the 'true' model) are diminished.

In sum, the literature to date provides a number of important results and guidelines, but conclusions about the empirical relevance of unobserved heterogeneity are likely to differ from application to application. Moreover, frailty models can be relatively 'fragile' in the statistical sense – they can be relatively hard to fit especially if the frailty variance is close to zero.

2.3 Frailty models available in Stata – overview

For continuous time models, Stata estimates frailty generalisations of all the non-frailty parametric models that were cited in Lesson 5: Exponential, Weibull, Gompertz, Log-logistic, Lognormal, Gamma. As we shall see below, estimation is – in principle – as straightforward as adding a **frailty(.)** option to one's **streg** command, and choosing between Gamma and Inverse Gaussian representations of the frailty. I say 'in principle' because the frailty models can sometimes be difficult to fit.

Several discrete time survival models with frailty can be estimated in Stata. The discrete time PH (cloglog) model with Gamma heterogeneity can be estimated using my program **pgmhaz8**. This can be downloaded for free from the SSC-IDEAS archive using the command **ssc install pgmhaz8**. The cloglog model with Normal distributed errors can be estimated using Stata's **xtcloglog** command and the logistic model with Normal distributed errors can be estimated using **xtlogit**. (Note that a Normal distribution for *u* corresponds to a lognormal distribution for *v*.)

Illustrations of the programs cited are provided below. Section 3 considers continuous time models and Section 4 discrete time models.

Some discrete time models with Heckman and Singer-type non-parametric representations of frailty can be estimated using my program **hshaz** (for proportional hazard models, obtained via **ssc install hshaz**), or Sophia Rabe-Hesketh's program **gllamm** (obtained via **ssc install gllamm**). Split-population (or 'cure') survival models can also be interpreted as models with a particular type of mover-stayer unobserved heterogeneity. See e.g. my program for discrete time data, **spsurv.** This is downloadable using **ssc install spsurv.**

Two types of frailty are currently incorporated in the Stata **streg** programs. The first (and default) is known as observation level frailty – there is one value of *v* for each record in the data. This is what we focus on here. If there is multiple record data, e.g. because of episode splitting to incorporate time-varying covariates, then each separate record counts as an observation. This treatment of frailty may not be what you want. For example you may wish to have a single value of *v* that is common to a group of observations (e.g. the same individual or the same family). This is known as *shared frailty*. To incorporate this you need to specify the **shared** option in addition to the **frailty(.)** one. If there is only a single record per person (such as when there are no time-varying covariates), or there is non-informative episode splitting, then the two approaches are equivalent.

The **stcox** command for Cox PL regression includes an option for estimating models with shared frailty, assuming a Gamma mixture. In the illustrative data that we use here, there is only a single record per person (see previous paragraph), and so the command is not applicable.

The discrete time frailty models cited above all incorporate shared frailty rather than observation level frailty.

3 Continuous time parametric models

Let us now illustrate the frailty models available via **streg**. The discussion here draws heavily on that in the Stata 7 Reference Manual Volume 3, Q–St, pp. 359–363, and uses the same data set, the breast cancer data (bc.dta). These (hypothetical) data refer to survival times for 80 women with breast cancer. Covariates summarise patient's age, whether she smokes, and average weekly calorific intake over the course of the study.

We shall first estimate a Weibull model without frailty but using all the covariates. (The data were in fact created by simulation from a Weibull distribution.)

```
. use bc.dta, clear
.
. stset t, f(dead)

     failure event:  dead ~= 0 & dead ~= .
obs. time interval:  (0, t]
 exit on or before:  failure

------------------------------------------------------------------------------
       80  total obs.
        0  exclusions
------------------------------------------------------------------------------
       80  obs. remaining, representing
       58  failures in single record/single failure data
  1257.07  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =        35

. streg age smoking dietfat, d(weib) nohr nolog

         failure _d:  dead
   analysis time _t:  t


Weibull regression -- log relative-hazard form

No. of subjects =           80                     Number of obs   =        80
No. of failures =           58
Time at risk    =      1257.07
                                                   LR chi2(3)      =    250.96
Log likelihood  =   -13.352142                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .559197   .0563239     9.93   0.000     .4488042    .6695899
     smoking |   1.649311   .3276501     5.03   0.000     1.007128    2.291493
     dietfat |   2.222411   .2404553     9.24   0.000     1.751128    2.693695
       _cons |  -45.97988   4.634153    -9.92   0.000    -55.06265   -36.89711
-------------+----------------------------------------------------------------
       /ln_p |   1.431728   .0978872    14.63   0.000     1.239872    1.623583
-------------+----------------------------------------------------------------
           p |   4.185925   .4097485                      3.455172    5.071228
         1/p |   .2388958   .0233848                      .1971909    .2894212
------------------------------------------------------------------------------
```

We can see that higher hazard rates – shorter survival times – are positively associated with age, smoking, and dietary fat at conventional levels of statistical significance. Let us now drop the dietfat variable with the aim of introducing unobserved heterogeneity. We will use this next model as the reference non-frailty model:

```
. streg age smoking , d(weib) nohr nolog

         failure _d:  dead
   analysis time _t:  t


Weibull regression -- log relative-hazard form

No. of subjects =           80                     Number of obs   =        80
No. of failures =           58
Time at risk    =      1257.07
                                                   LR chi2(2)      =    118.82
Log likelihood  =   -79.419727                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1644213   .0149837    10.97   0.000     .1350538    .1937888
     smoking |   .9056537   .3061656     2.96   0.003     .3055801    1.505727
       _cons |  -11.20242   .9989083   -11.21   0.000    -13.16024   -9.244594
-------------+----------------------------------------------------------------
       /ln_p |   .3633523   .0955797     3.80   0.000     .1760195    .5506852
-------------+----------------------------------------------------------------
           p |   1.438142   .1374573                      1.192461    1.734441
         1/p |   .6953414   .0664605                      .5765546    .8386016
------------------------------------------------------------------------------
```

Let us also predict median survival times. First we predict the median for the person with characteristics equal to the sample mean values on all covariates, and then we predict the median for each person in the sample. The first prediction has to be done ‘manually’ but the second can be done directly using **predict**.

```
. predict xb, xb

. su xb

    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+-----------------------------------------------------
          xb |      80   -3.834933     1.93139  -6.927465  -.1026448

. di "Pred. Median [at sample mean X] = "  (ln(2)*exp(-r(mean)))^(1/e(aux_p))
Pred. Median [at sample mean X] = 11.153305

. * median duration for each person in sample
. * NB Stata allows you to generate these directly:
. predict mediand, time
(option median time assumed; predicted median time)

. su mediand, de

                     predicted median _t
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .8323699       .8323699
 5%     1.109575       1.046216
10%     1.707746       1.046216       Obs                  80
25%     3.789622       1.046216       Sum of Wgt.          80

50%      11.5727                      Mean           23.37674
                        Largest       Std. Dev.      26.22382
75%     34.23132       85.43642
90%     67.97328       95.78455       Variance       687.6885
95%     80.82134       95.78455       Skewness       1.359739
99%     95.78455       95.78455       Kurtosis       3.853934

. drop xb mediand
```

So the median duration for the person with mean characteristics is 11.2, and the median among the sample as a whole is 11.6.

Now we allow for unobserved heterogeneity, first assuming a Gamma mixture distribution and then an Inverse Gaussian one. We shall also look at predicted median distributions.

```
. streg age smoking , d(weib) nohr nolog frailty(gamma)

         failure _d:  dead
   analysis time _t:  t


Weibull regression -- log relative-hazard form
                      Gamma frailty

No. of subjects =           80                  Number of obs   =         80
No. of failures =           58
Time at risk    =      1257.07
                                                LR chi2(2)      =     135.75
Log likelihood  =   -68.135804                  Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .3893002   .0934984     4.16   0.000     .2060467    .5725537
     smoking |   1.025521   .5225054     1.96   0.050     .0014291    2.049613
       _cons |   -23.8082   5.204923    -4.57   0.000    -34.00966   -13.60674
-------------+----------------------------------------------------------------
       /ln_p |   1.087761    .222261     4.89   0.000     .6521376    1.523385
     /ln_the |   .3307466   .5250758     0.63   0.529     -.698383    1.359876
-------------+----------------------------------------------------------------
           p |   2.967622   .6595867                       1.91964    4.587727
         1/p |   .3369701   .0748953                      .2179729     .520931
       theta |   1.392007   .7309092                      .4973889    3.895711
------------------------------------------------------------------------------
Likelihood ratio test of theta=0: chibar2(01) =    22.57 Prob>=chibar2 = 0.000
```

The ‘theta’ value reported in the output is the estimate of the frailty distribution variance. Note that the frailty model is preferred to the reference non-frailty model according to the relevant likelihood ratio test. The test is a ‘boundary’ test that takes account of the fact that the null distribution is not the usual chi-squared(d.f. = 1) but is rather a 50:50 mixture of a chi-squared(d.f. = 0) variate (which is a point mass at zero) and chi-squared(d.f. = 1) – hence the reference to ‘chibar2(01)’ in the output. Click on the blue ‘chibar2(01)’ in the output window for an explanation, or see Gutierrez *et al.* (2001) for more details (Gutierrez, R.G., Carter, S., and Drukker, D., ‘On boundary-value likelihood-ratio tests’, insert sg160, Stata Technical Bulletin, STB-60, StataCorp, College Station TX.) The *p*-value = 0.000 in this case.

The frailty has expected effects on model parameters. The estimated coefficients on the regressors age and smoking are larger in magnitude that the corresponding coefficients in the reference model. Also the Weibull distribution shape parameter *p* is larger in the frailty models than in the reference model – the baseline hazard slopes upwards to a greater extent. Observe too that with frailty present, an exponentiated coefficient is simply that, losing its interpretation in terms of a hazard ratio (a proportional change in the hazard for a one unit change in the relevant covariate).

Now let us consider the predictions of median duration and compare them with those of the reference non-frailty model. Observe that because there are no time-varying covariates, observation-level and shared frailty models are equivalent. The former is the default.

```
. predict xb, xb
(option unconditional assumed)

. su xb

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------
          xb |      80   -6.685866   4.461222   -13.6864   1.353928

. di "Pred. Median [at sample mean X] = " ((2^e(theta) - 1)/
(e(theta)*exp(r(mean))))^(1/e(aux_p))
Pred. Median [at sample mean X] = 10.023875

. * median duration for each person in sample
. * NB Stata allows you to generate these directly:
. predict mediand, time
(option unconditional assumed)
(option median time assumed; predicted median time)

. su mediand, de

                        predicted _t
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .6675103       .6675103
 5%      .867764       .8271137
10%     1.107688       .8271137       Obs                  80
25%     3.071027        .867764       Sum of Wgt.          80

50%     10.94738                      Mean           24.43518
                        Largest       Std. Dev.      29.05888
75%     37.13202       93.01468
90%      71.5497       106.0531       Variance       844.4185
95%     87.29696       106.0531       Skewness       1.378743
99%     106.0531       106.0531       Kurtosis       3.908608
```

The median for the person with mean characteristics is now predicted to be 10.02. This is smaller than the corresponding value for the non-frailty model – as expected, perhaps, given the change in the coefficients and increase in duration dependence parameter.

We can also see what the median is for the case in which we condition on frailty $v = 1$ (the mean value):

```
. predict mediand2, time alpha1
(option median time assumed; predicted median time)

. su mediand2, de

                     predicted median _t
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .5600455       .5600455
 5%     .7280598       .6939539
10%     .9293576       .6939539       Obs                  80
25%     2.576612       .7280598       Sum of Wgt.          80

50%     9.184921                      Mean           20.50128
                        Largest       Std. Dev.      24.38059
75%     31.15401       78.03993
90%     60.03067       88.97925       Variance       594.4133
95%     73.24273       88.97925       Skewness       1.378743
99%     88.97925       88.97925       Kurtosis       3.908608

.
. drop mediand mediand2 xb
```

It turns out that the median is a little bit smaller than before, at all corresponding points of the sample distribution.

Now let us repeat the analysis for the Inverse Gaussian frailty model.

```
. streg age smoking , d(weib) nohr nolog frailty(invgauss)

         failure _d:  dead
   analysis time _t:  t


Weibull regression -- log relative-hazard form
                      Inverse-Gaussian frailty

No. of subjects =           80                  Number of obs   =        80
No. of failures =           58
Time at risk    =      1257.07
                                                LR chi2(2)      =    125.44
Log likelihood  =   -73.838578                  Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .2500841   .0360754     6.93   0.000      .1793777    .3207906
     smoking |   1.066574   .4311907     2.47   0.013      .2214561    1.911692
       _cons |  -16.01035   2.097075    -7.63   0.000     -20.12054   -11.90015
-------------+----------------------------------------------------------------
       /ln_p |   .7173904   .1434382     5.00   0.000      .4362567    .9985241
     /ln_the |   .2374778   .8568064     0.28   0.782     -1.441832    1.916788
-------------+----------------------------------------------------------------
           p |   2.049079   .2939162                       1.546906    2.714273
         1/p |   .4880241   .0700013                       .3684228    .6464518
       theta |   1.268047   1.086471                       .2364941    6.799082
------------------------------------------------------------------------------
Likelihood ratio test of theta=0: chibar2(01) =    11.16 Prob>=chibar2 = 0.000
```

It turns out that frailty is again statistically significant, and again the parameters of the model change in the expected direction. The improvement in log-likelihood relative to the no-frailty model is largest for the Gamma mixture model, but choice between the Gamma and Inverse Gaussian specifications is complicated by the fact that the two models are non-nested.

How do model predictions differ between the two frailty specifications? We explore this using **predict** after the estimation command.

```
. predict xb, xb
(option unconditional assumed)

. su xb

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------
          xb |      80   -4.893819   2.902277  -9.508158   .5614455

.  di  "Pred.  Median  [at  sample  mean  X]  =  "   (  ((1+e(theta)*ln(2))^2  -
1)/(2*e(theta)*exp(r(mean))) )^(1/e(
> aux_p))
Pred. Median [at sample mean X] = 10.883087

. * median duration for each person in sample
. * NB Stata allows you to generate these directly:
. predict mediand, time
(option unconditional assumed)
(option median time assumed; predicted median time)

. su mediand, de

                        predicted _t
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .7595033       .7595033
 5%     1.032402       .9694791
10%     1.454846       .9694791       Obs                  80
25%     3.711818       .9694791       Sum of Wgt.          80

50%     11.31624                      Mean           24.43678
                        Largest       Std. Dev.      28.26338
75%     36.10971       91.56709
90%     71.73492       103.4532       Variance       798.8188
95%     86.30687       103.4532       Skewness       1.382187
99%     103.4532       103.4532       Kurtosis       3.930592

. drop xb mediand
```

The predicted median for the person with ‘average’ characteristics is now 10.9.

What happens if we re-estimate the models but now re-introduce dietfat as a regressor? Here are the estimates from the model with Gamma frailty.

```
. streg age smoking dietfat, d(weib) nolog frailty(gamma) nohr

         failure _d:  dead
   analysis time _t:  t


Weibull regression -- log relative-hazard form
                      Gamma frailty

No. of subjects =           80                  Number of obs   =         80
No. of failures =           58
Time at risk    =      1257.07
                                                LR chi2(3)      =     245.32
Log likelihood  =   -13.352142                  Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .5592066   .0563201     9.93   0.000     .4488212    .6695919
     smoking |   1.649354   .3276412     5.03   0.000     1.007188    2.291519
     dietfat |   2.222451   .2404404     9.24   0.000     1.751196    2.693706
       _cons |  -45.98067   4.633832    -9.92   0.000    -55.06281   -36.89852
-------------+----------------------------------------------------------------
       /ln_p |   1.431747   .0978783    14.63   0.000     1.239909    1.623585
     /ln_the |  -15.92927   6628.419    -0.00   0.998    -13007.39    12975.53
-------------+----------------------------------------------------------------
           p |   4.186005   .4097189                      3.455299    5.071237
         1/p |   .2388912   .0233823                      .1971906    .2894106
       theta |   1.21e-07   .0008006                             0           .
------------------------------------------------------------------------------
Likelihood ratio test of theta=0: chibar2(01) =     0.00 Prob>=chibar2 = 1.000
```

Here are the estimates from the model with Inverse Gaussian frailty.

```
. streg age smoking dietfat, d(weib) nolog frailty(invg) nohr

         failure _d:  dead
   analysis time _t:  t


Weibull regression -- log relative-hazard form
                      Inverse-Gaussian frailty

No. of subjects =           80                  Number of obs   =         80
No. of failures =           58
Time at risk    =      1257.07
                                                LR chi2(3)      =     246.41
Log likelihood  =   -13.352142                  Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .5592044   .0563229     9.93   0.000     .4488136    .6695952
     smoking |   1.649341   .3276494     5.03   0.000      1.00716    2.291522
     dietfat |   2.222442   .2404515     9.24   0.000     1.751166    2.693718
       _cons |  -45.98049   4.634064    -9.92   0.000    -55.06309   -36.89789
-------------+----------------------------------------------------------------
       /ln_p |   1.431742   .0978845    14.63   0.000     1.239892    1.623592
     /ln_the |    -14.424   2866.444    -0.01   0.996    -5632.551    5603.703
-------------+----------------------------------------------------------------
           p |   4.185987   .4097431                      3.455242    5.071276
         1/p |   .2388923   .0233838                       .197189    .2894154
       theta |   5.44e-07   .0015598                             0           .
------------------------------------------------------------------------------
Likelihood ratio test of theta=0: chibar2(01) =     0.00 Prob>=chibar2 = 1.000
```

For both models, we now see that there is negligible unobserved heterogeneity – observe the near-zero frailty variances, and the *p*-values for the likelihood ratio test equal to one. The coefficients on the covariates are almost exactly the same as those in the corresponding model without unobserved heterogeneity that we estimated at the very start.

4 Discrete time models

We will now repeat our analysis of the same data but use discrete time models instead. To facilitate comparability with the Weibull models of Section 3, we shall use proportional hazards models with a log(time) specification for duration dependence.

One initial minor complication is that survival times in bc.dta are non-integer. We therefore need to create a new survival time variable (‘td’ below, rather than ‘t’), in addition to doing the usual episode-splitting to form a data set organised in person-month form. We use the **ceil()** function to round up the survival times to the nearest integer.

```
use bc, clear

. * convert survival times to discrete integers
. ge td = ceil(t)  // has same effect as: ge td = round(t+.49,1)

. su t td

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------
           t |      80    15.71337   13.59278        .33         35
          td |      80     16.0875   13.40243          1         35
.
. sort t
.
. ge id = _n
```

Now we do the episode-splitting to derive the data organised in person-month form, and create the appropriate spell month identifier. Then we create the variable to summarize duration dependence in the discrete hazard, log(time) in this case.

```
. expand td /* expand on td (not t) since time intervals indexed on td */
(1207 observations created)

. sort id

. by id: ge newt = _n

. by id: ge died = dead==1 & _n==_N

. ge logt = ln(newt)
```

Let us begin with the model including all covariates (but no frailty) and then the reference model from which the dietfat regressor has been omitted:

```
. cloglog died logt age smoking dietfat, nolog

Complementary log-log regression                  Number of obs     =       1287
                                                  Zero outcomes     =       1229
                                                  Nonzero outcomes  =         58

                                                  LR chi2(4)        =     244.53
Log likelihood = -114.18864                       Prob > chi2       =     0.0000

------------------------------------------------------------------------------
        died |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        logt |   3.594979   .4918103     7.31   0.000     2.631048    4.558909
         age |   .5449406    .060167     9.06   0.000     .4270153    .6628658
     smoking |   1.638842   .3753424     4.37   0.000     .9031841    2.374499
     dietfat |   2.149625   .2561884     8.39   0.000     1.647505    2.651745
       _cons |  -44.72472     4.9638    -9.01   0.000    -54.45359   -34.99585
------------------------------------------------------------------------------

. cloglog died logt age smoking, nolog

Complementary log-log regression                  Number of obs     =       1287
                                                  Zero outcomes     =       1229
                                                  Nonzero outcomes  =         58

                                                  LR chi2(3)        =     126.56
Log likelihood = -173.17144                       Prob > chi2       =     0.0000

------------------------------------------------------------------------------
        died |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        logt |   .5328684   .1743352     3.06   0.002     .1911776    .8745592
         age |   .1647403    .015838    10.40   0.000     .1336985    .1957821
     smoking |   .8752191   .3078178     2.84   0.004     .2719074    1.478531
       _cons |  -11.10694   1.042009   -10.66   0.000    -13.14924   -9.064643
------------------------------------------------------------------------------
```

The estimates are similar to the corresponding Weibull model estimates (as expected).

Now what if we suppose that the frailty term *u* is Normally distributed? We use the **xtcloglog** command. **xt** stands for 'cross-section time series' or 'panel data' estimator – we are using a panel data estimator to estimate a survival analysis model. The mandatory **i(.)** option is used to identify the observations with distinct values for the heterogeneity term. The 'sigma_u' reported is the standard deviation of the heterogeneity variance. The reported 'rho' is the ratio of the heterogeneity variance to one plus the heterogeneity variance. So if the hypothesis that rho is zero cannot be rejected, then frailty is unimportant.

```
xtcloglog died logt age smoking , nolog i(id)

Random-effects complementary log-log            Number of obs      =      1287
Group variable (i) : id                         Number of groups   =        80

Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =      16.1
                                                               max =        35

                                                Wald chi2(3)       =     21.88
Log likelihood  = -164.02912                    Prob > chi2        =    0.0001

------------------------------------------------------------------------------
        died |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        logt |   3.344268   .9695804     3.45   0.001     1.443925    5.244611
         age |   .5170562   .1222468     4.23   0.000     .2774569    .7566554
     smoking |   1.587545   .6984732     2.27   0.023     .2185631    2.956528
       _cons |  -32.61654   7.741936    -4.21   0.000    -47.79045   -17.44262
-------------+----------------------------------------------------------------
    /lnsig2u |   1.743016   .5665486                      .6326016    2.853431
-------------+----------------------------------------------------------------
     sigma_u |   2.390514    .677171                      1.372043    4.164997
         rho |   .8510698     .07181                      .6530791    .9454958
------------------------------------------------------------------------------
Likelihood ratio test of rho=0: chibar2(01) =    18.28 Prob >= chibar2 = 0.000
```

The likelihood ratio test suggests statistically significant frailty. The frailty has expected effects on model parameters. The estimated coefficients on the regressors age and smoking are larger in magnitude that the corresponding coefficients in the reference model. Also the duration dependence is much larger in the frailty models than in the reference model – the baseline hazard slopes upwards to a greater extent.

What happens if we re-estimate our model but instead assume Gamma-distributed unobserved heterogeneity? To investigate this, we use **pgmhaz8**. Its syntax differs from that for the other programs (it was modelled on the syntax for the Cox model in Stata 5, i.e. in a pre-**stset** era), but is relatively straightforward. One specifies the covariates after the program name, and then options are used to specify identifiers for each person (the **id(.)** option), the spell interval identifier for each person (**seq(.)**, which is an integer sequence from 1 to the total survival time for that person) and the censoring variable (the **dead(.)** option). See **help pgmhaz8** for further details. The program first reports the non-frailty cloglog model (estimated using **glm**), and then the Gamma frailty model. (Observe that the non-frailty estimates correspond to those reported earlier. **pgmhaz8** has an option to turn off the reporting of these.)

```
. pgmhaz8 logt age smoking , id(id) seq(newt) dead(died) nolog
PGM hazard model without gamma frailty

Generalized linear models                          No. of obs      =       1287
Optimization     : ML: Newton-Raphson              Residual df     =       1283
                                                   Scale parameter =          1
Deviance         =  346.3428729                    (1/df) Deviance =  .2699477
Pearson          =  827.8333299                    (1/df) Pearson  =  .6452325

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

Log likelihood   = -173.1714365                    AIC             =  .2753247
BIC              =  -8840.02592

------------------------------------------------------------------------------
        died |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        logt |   .5328684   .1743352     3.06   0.002     .1911776    .8745592
         age |   .1647403    .015838    10.40   0.000     .1336985    .1957821
     smoking |   .8752191   .3078178     2.84   0.004     .2719074    1.478531
       _cons |  -11.10694   1.042009   -10.66   0.000    -13.14924   -9.064643
------------------------------------------------------------------------------

PGM hazard model with gamma frailty               Number of obs   =       1287
                                                  LR chi2()       =          .
Log likelihood = -162.55447                       Prob > chi2     =          .

------------------------------------------------------------------------------
        died |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
hazard       |
        logt |   2.742215   .9879201     2.78   0.006     .8059276    4.678503
         age |   .4369005   .1197364     3.65   0.000     .2022216    .6715795
     smoking |   .9551981   .5663482     1.69   0.092     -.154824     2.06522
       _cons |  -26.43864   6.772158    -3.90   0.000    -39.71182   -13.16545
-------------+----------------------------------------------------------------
ln_varg      |
       _cons |   .5409866   .5287288     1.02   0.306    -.4953028    1.577276
-------------+----------------------------------------------------------------
  Gamma var. |   1.717701   .9081978     1.89   0.059     .6093864    4.841749
------------------------------------------------------------------------------
LR test of Gamma var. = 0: chibar2(01) =   21.2339  Prob.>=chibar2 =  2.0e-06
```

The *p*-value for the likelihood ratio test is virtually zero, indicating statistically significant frailty – which is entirely consistent with what we found for the corresponding continuous time model. We see again too that regressor coefficients are larger in magnitude and the degree of positive duration dependence in the hazard is larger.

If we add dietfat back into the model as a regressor, then the frailty appears to become negligible. Observe that I omitted the **nolog** option this time, so the iteration log for the frailty model is reported. There are obviously some problems with convergence (note the ‘not concave’ messages) and, although the model did finally converge satisfactorily, the frailty variance is tiny.

```
. pgmhaz8 logt age smoking dietfat, id(id) seq(newt) dead(died) nolog

. pgmhaz8 logt age smoking dietfat, id(id) seq(newt) dead(died) iter(25)
PGM hazard model without gamma frailty

Generalized linear models                          No. of obs      =      1287
Optimization     : ML: Newton-Raphson              Residual df     =      1282
                                                   Scale parameter =         1
Deviance         =    228.37728                    (1/df) Deviance =  .1781414
Pearson          =  679.2477092                    (1/df) Pearson  =  .5298344

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

Log likelihood   =   -114.18864                    AIC             =  .1852193
BIC              = -8950.831444

------------------------------------------------------------------------------
        died |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        logt |   3.594979   .4918101     7.31   0.000     2.631049    4.558909
         age |   .5449406    .060167     9.06   0.000     .4270154    .6628657
     smoking |   1.638842   .3753423     4.37   0.000     .9031842    2.374499
     dietfat |   2.149625   .2561882     8.39   0.000     1.647506    2.651745
       _cons |  -44.72472   4.963796    -9.01   0.000    -54.45358   -34.99586
------------------------------------------------------------------------------

Iteration 0:   log likelihood = -116.47328
Iteration 1:   log likelihood = -114.39827
Iteration 2:   log likelihood = -114.29907
Iteration 3:   log likelihood = -114.22669
Iteration 4:   log likelihood = -114.20848
Iteration 5:   log likelihood = -114.19356
Iteration 6:   log likelihood =  -114.1906
Iteration 7:   log likelihood =  -114.1895
Iteration 8:   log likelihood = -114.18888
Iteration 9:   log likelihood = -114.18871
Iteration 10:  log likelihood = -114.18865
Iteration 11:  log likelihood = -114.18864  (not concave)
numerical derivatives are approximate
nearby values are missing
Iteration 12:  log likelihood = -114.18864  (not concave)
Iteration 13:  log likelihood = -114.18864  (not concave)
Iteration 14:  log likelihood = -114.18864

PGM hazard model with gamma frailty               Number of obs   =       1287
                                                  LR chi2()       =          .
Log likelihood = -114.18864                       Prob > chi2     =          .

------------------------------------------------------------------------------
        died |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
hazard       |
        logt |   3.595206   .1925172    18.67   0.000     3.217879    3.972532
         age |   .5449706   .0197679    27.57   0.000     .5062262    .5837149
     smoking |   1.638924   .3370305     4.86   0.000     .9783566    2.299492
     dietfat |   2.149746   .1156383    18.59   0.000       1.9231    2.376393
       _cons |  -44.72717   1.595761   -28.03   0.000    -47.85481   -41.59954
-------------+----------------------------------------------------------------
ln_varg      |
       _cons |  -13.22453   1390.536    -0.01   0.992    -2738.625    2712.176
-------------+----------------------------------------------------------------
  Gamma var. |   1.81e-06    .002511     0.00   0.999            0           .
------------------------------------------------------------------------------
LR test of Gamma var. = 0: chibar2(01) =  -8.2e-07  Prob.>=chibar2 =       .5
```

What is probably happening is that the program is trying to find ever-smaller values of the variance: a value of the log of the gamma variance of –13.22 implies a value for the variance that is very very close to zero! But the model is programmed with the gamma variance constrained to be positive – hence convergence problems. This type of result is quite common when estimating frailty models.

To reiterate, although our first applications of **pgmhaz8** converged relatively quickly and without difficulty, this is often not the case in practice with real-world data sets. With ‘large’ data sets, the model can be slow to converge, because the program uses numerical derivatives. Moreover the likelihood surface is not globally concave and non-concavities may sometimes be reported at the final iteration, or the maximisation may sometimes get stuck on a ‘flat’ part and fail to converge with an error message:

```
pgmhaz8_ll does not compute a continuous nonconstant function
could not calculate numerical derivatives
r(430);
```

If such situations arise, users are recommended to use the **trace** option to **pgmhaz8** – e.g. check whether the estimate of ln(Gamma variance) is heading off towards minus infinity – and to experiment with different starting values for the Gamma variance using the **lnvar(.)** option.

xtcloglog and **xtlogit** may also take a long time to run with expanded person-month data sets if they are ‘large’.

4.1 Prediction for discrete-time frailty models

There are no built-in commands for predicting medians and so on for the discrete-time frailty models, as there were for the frailty models estimated using **streg**. However predictions can be derived straightforwardly, by adapting the strategy that was illustrated in Lesson 6 for discrete-time models without unobserved heterogeneity. For the discrete-time frailty models, predictions are derived assuming that the frailty term is set equal to its mean value.

The idea is to first derive predicted hazard rates for persons with given characteristics, and thence the implied survivor functions. I start with the **pgmhaz8** model estimated earlier, with age and smoking as the covariates, and consider predictions for three persons: (a) someone with the sample mean values of each regressor (age = 35, smoking = 0.23); (b) someone aged 30 and a non-smoker; and (c) someone aged 50 and a smoker.

Recall that the cloglog hazard rate with frailty has specification:

$$p(t) \; = 1 - \exp[-\exp(\, z(t)\,)]$$

where $z(t) = D(t) + X\beta + u$. $D(t)$ is the baseline hazard function and $X\beta$ includes an intercept term. For predictions, we will take the case with $u = 0$, and we will consider the case with $D(t) = (q-1)\log(t)$, as earlier.

What we have to do is, first, generate values of $z(t)$ for each of our three persons, using the coefficient estimates left behind after **pgmhaz8** Model 2, and then, second, the implied hazard rate and survivor functions. At the first step, one uses Stata’s **matrix** operations to extract the required coefficients (one doesn’t need the estimate of the frailty variance). Then one generates the values of $z(t)$ = β_0 + β_1*age + β_2*smoking + c.log(t) for the given values of age and smoking. Observe the result used to get $z(t)$ when age and smoking are at sample mean values: (the mean of a linear combination equals the linear combination of means). So, **matrix score** is used to calculate β_0 + β_1*age + β_2*smoking for all cases, and then **summarise** is used to derive the mean of this. The **if newt == 1** qualifier is used to ensure the mean is taken over individual cases, rather than over time intervals.

The following code illustrates the strategy as a whole:

```
. * First re-do earlier model
. pgmhaz8 logt age smoking , id(id) seq(newt) dead(died) nolog
< output omitted >

. * Now generate predicted probabilities
. mat b = e(b)

. mat list b

b[1,5]
       hazard:    hazard:    hazard:    hazard:   ln_varg:
         logt        age    smoking      _cons      _cons
y1  2.7422156  .43690055  .95519812  -26.43864  .54098674

. scalar n = colsof(b) - 1

. scalar list n
         n =          4

.
. mat b = b[1,2..n]  /* exclude coeffs on dur dep var(s), and lnvarg est */

. mat list b

b[1,3]
       hazard:    hazard:    hazard:
          age    smoking      _cons
y1  .43690055  .95519812  -26.43864

. mat score xb = b

. sum xb if newt == 1

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------
          xb |      80   -7.278995   4.992778  -15.07923   1.604393

. ge z0 = r(mean) + _b[logt]*logt

. ge z1 = _b[_cons] + _b[age]*45 + _b[smoking]*1 + _b[logt]*logt

. ge z2 = _b[_cons] + _b[age]*30 + _b[smoking]*0 + _b[logt]*logt

.
. sort id newt

. by id: gen p0 = 1 - exp(-exp(z0))

. lab var p0 "Predicted h(t) at mean of covariates"

. by id: gen p1 = 1 - exp(-exp(z1))

. lab var p1 "Predicted h(t),age=45,smoking"

. by id: gen p2 = 1 - exp(-exp(z2))

. lab var p2 "Predicted h(t),age=30,non-smoking"

.
. by id: ge s0 =  exp(sum(ln(1-p0)))

. lab var s0 "Predicted S(t) at mean of covariates"

. by id: ge s1 =  exp(sum(ln(1-p1)))

. lab var s1 "Predicted S(t),age=45,smoking"

. by id: ge s2 =  exp(sum(ln(1-p2)))
. lab var s2 "Predicted S(t),age=30,non-smoking"
```

Having generated the predicted hazard and survivor probabilities, we can graph them:

```
. . twoway (connect p0 newt , sort  msymbol(t) ) ///
>         (connect p1 newt, sort msymbol(o) ) ///
>         (connect p2 newt, sort msymbol(x) )  ///
>         , title("Predicted discrete hazard rates from -pgmhaz8-") ///
>         saving(pgmh1, replace) ytitle("p(t)")
(file pgmh1.gph saved)
```

```
. twoway (connect s0 newt , sort  msymbol(t) )   ///
>         (connect s1 newt, sort msymbol(o) )   ///
>         (connect s2 newt , sort  msymbol(x) )  ///
>         , title("Predicted survivor functions from -pgmhaz8-") ///
>         saving(pgms1, replace) ytitle("S(t)")
(file pgms1.gph saved)
```

Figure 8.1. Predicted hazard rates from -pgmhaz8-

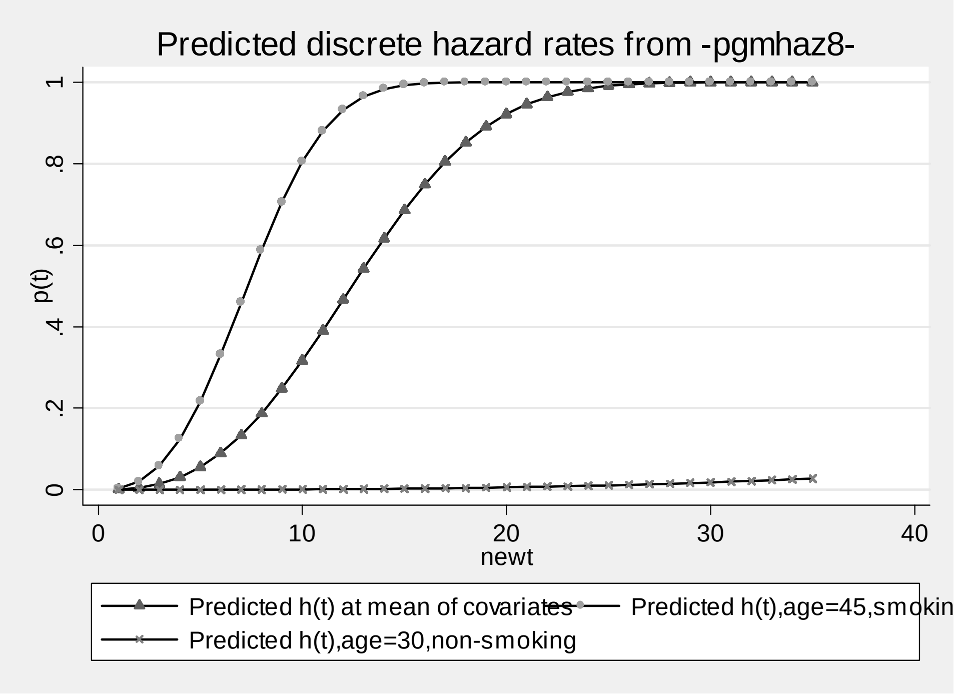
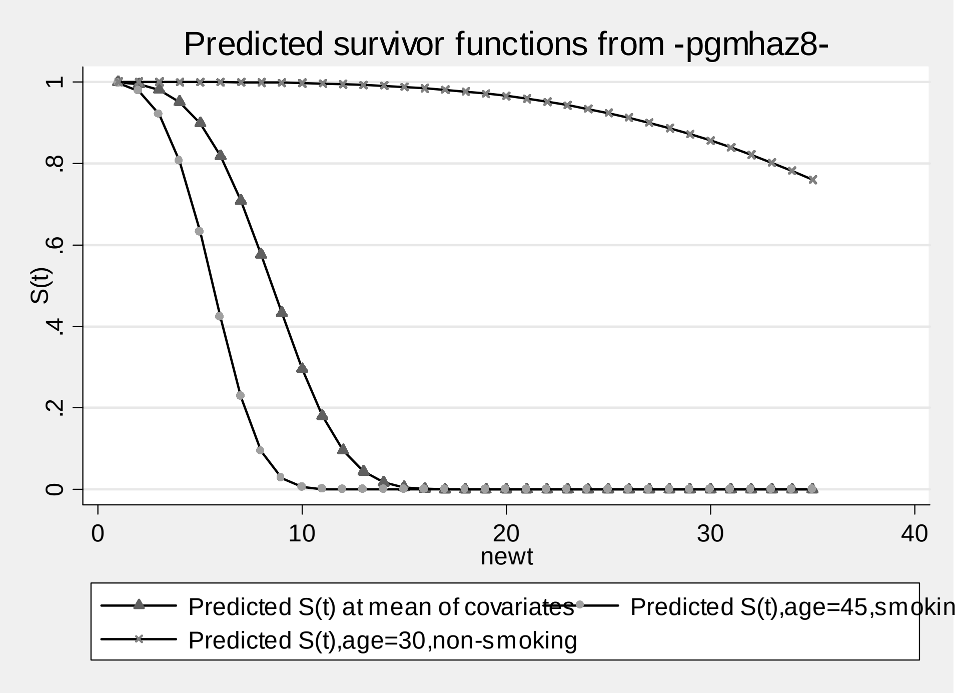


Figure 8.2. Predicted survivor functions from -pgmhaz8-



The median duration for the person with sample mean values is about 10, but clearly very much longer for the 30 year old non-smoker (out of sample range). The median duration for the 45 year-old smoker is shorter – about six – and the survival probability in this case is zero by the tenth interval. Observe from the first graph how the hazard rate for this person rises very quickly.

Finally, we do a bit of tidying up:

```
. drop z0 z1 z2 xb p0 p1 p2 s0 s1 s2
```

This code only produces predictions for survival times ranging between the minimum and the maximum in the sample. If you want predictions for longer survival times, then you have to generate extra sample observations – just as we did in Lesson 6.

Observe that the same code could be used to generate predictions after estimating the proportional hazards or logistic models with Normal heterogeneity (the order in which the regressors are listed has to be the same as before for this to work without change). The commands would be:

```
. xtcloglog died logt age smoking , i(id)  nolog
```

```
. xtlogit died logt age smoking , i(id)  nolog
```

Note that, for these models, one could predict hazard rates in this model for each person within the sample using their built-in **predict** commands:

```
. predict p, pu0
```

Observe the ‘pu0’ option. This ensures that the prediction is generated while conditioning on frailty being set equal to its mean value. You could then generate the survivor function predictions with

```
. bysort id (newt): ge s =  exp(sum(ln(1-p)))
```

The within-sample predictions could be listed and so on, as we did in Lesson 6.

5 Exercise 7.1

(i) Repeat the estimation of frailty and non-frailty models using the bc.dta for a Log-logistic hazard model rather than a Weibull one. In what way do the results change?

(ii) For any one of the three persons considered in the discrete-time model predictions, compare the predicted median survival times from the proportional hazards models with and without Normal heterogeneity (i.e. **xtcloglog** and **cloglog** with age and smoking as the regressors). Which model predicts the shorter median? Explain the result you find, referring to the parameter estimates from the two models.

(iii) Repeat the estimation of Normal frailty and non-frailty models using the breast cancer data (bc.dta) for a discrete time logistic model rather than a discrete time proportional hazards model. Use **logit** and **xtlogit** rather than **cloglog** and **xtcloglog**. Compare model predictions too for the three persons considered in the text.

(iv) Verify that uninformative episode-splitting does not affect the estimates of the continuous time frailty models. Use the Weibull model to do this and the following commands:

```
use bc, clear
ge id = _n
stset t, f(dead) id(id)
streg age smoking , d(weib) nohr nolog frailty(gamma)
stsplit time, every(1)
stset
streg age smoking , d(weib) nohr nolog frailty(gamma)
```

(v) Compare the results of the various frailty and non-frailty models, discrete and continuous, using a different data. I suggest the cancer data set (cancer.dta) rather than bc.dta.

(vi) Using bc.dta, we compared the estimates derived using

```
pgmhaz8 logt age smoking , id(id) seq(newt) dead(died) nolog
xtcloglog died logt age smoking , nolog i(id)
```

These assume a parametric continuous distribution for the frailty mixture distribution. Re-estimate the model assuming a two mass point discrete mixture using the command

```
hshaz logt age smoking , id(id) seq(newt) dead(died) nolog
```

(By default, **hshaz** assumes that there are two mass points; the number can be changed using the **nmp()** option.) Comment on the estimates and compare them with those for the earlier models. Derive and compare the predicted hazard and survivor functions for persons from Type 1 and Type 2 (setting the other covariates at their mean values).