**7. Unobserved heterogeneity (‘frailty’)**

1. **Aim**

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

1. **Introduction**

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

* 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

θ*v*(*t*,*X*) ≡ θ(*t*,*X*|*v*)=θ(*t*,*X*).*v*

where θ(*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 σ2. 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:  
Thus unobserved differences also imply a scaling of the non-frailty survivor function.

*Sv*(*t*,*X*) ≡ *S*(*t*,*X*|*v*) = [*S*(*t*,*X*)]*v*.

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

θ*v*(*t*, *X*) ≡ θ(*t*, *X* | β, *v*) = θ0(*t*).exp(β′*X*).*v* = θ0(*t*).exp(β′*X* + *u*)

ln[θ*v*(*t*, *X*)] = ln[θ0(*t*)] + β′*X* + *u*

where θ0(*t*) is the baseline hazard function and the ‘error’ term *u* ≡ 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

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

cloglog[*p*(*t*, *X* | β, *v*)] = *D*(*t*) + β′*X* + *u*

where *D*(*t*) characterises the baseline hazard function. The logistic hazard regression

model is typically generalised in an analogous way: logit[*p*(*t*, *X* | β, *v*)] = *D*(*t*) + β′*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* | β, σ2) rather than *S*(*t*, *X* | β, *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 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.

* 1. **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 *Xk* 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.

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.

* 1. **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*.)

1. **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)

streg age smoking dietfat, d(weib) nohr nolog

Failure \_d: dead

Analysis time \_t: t

Weibull PH regression

No. of subjects = 80 Number of obs = 80

No. of failures = 58

Time at risk = 1,257.07

LR chi2(3) = 250.96

Log likelihood = -13.352142 Prob > chi2 = 0.0000

------------------------------------------------------------------------------

\_t | Coefficient 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 PH regression

No. of subjects = 80 Number of obs = 80

No. of failures = 58

Time at risk = 1,257.07

LR chi2(2) = 118.82

Log likelihood = -79.419727 Prob > chi2 = 0.0000

------------------------------------------------------------------------------

\_t | Coefficient 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

------------------------------------------------------------------------------

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 PH regression

Gamma frailty

No. of subjects = 80 Number of obs = 80

No. of failures = 58

Time at risk = 1,257.07

LR chi2(2) = 135.75

Log likelihood = -68.135804 Prob > chi2 = 0.0000

------------------------------------------------------------------------------

\_t | Coefficient 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

/lntheta | .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

------------------------------------------------------------------------------

LR 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

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.

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 PH regression

Gamma frailty

No. of subjects = 80 Number of obs = 80

No. of failures = 58

Time at risk = 1,257.07

LR chi2(3) = 245.32

Log likelihood = -13.352142 Prob > chi2 = 0.0000

------------------------------------------------------------------------------

\_t | Coefficient Std. err. z P>|z| [95% conf. interval]

-------------+----------------------------------------------------------------

age | .5592066 .05632 9.93 0.000 .4488214 .6695918

smoking | 1.649354 .327641 5.03 0.000 1.007189 2.291518

dietfat | 2.222451 .2404402 9.24 0.000 1.751197 2.693705

\_cons | -45.98067 4.633826 -9.92 0.000 -55.0628 -36.89853

-------------+----------------------------------------------------------------

/ln\_p | 1.431747 .0978781 14.63 0.000 1.239909 1.623584

/lntheta | -15.92571 6058.646 -0.00 0.998 -11890.65 11858.8

-------------+----------------------------------------------------------------

p | 4.186005 .4097183 3.4553 5.071235

1/p | .2388912 .0233822 .1971906 .2894105

theta | 1.21e-07 .0007344 0 .

------------------------------------------------------------------------------

LR test of theta=0: chibar2(01) = 0.00 Prob >= chibar2 = 1.000

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.