ON FRAILTY MODELS IN STATA

Roberto G. Gutierrez

Stata Corporation

OUTLINE

I. Introduction

   A. Basic concepts of survival analysis

   B. Frailty

   C. Frailty vs. shared frailty

II. Example – Breast Cancer Data

III. Parametric Frailty Models

   A. The unconditional survival function

   B. Example

   C. Comparing the gamma and inverse–Gaussian.

IV. Parametric Shared Frailty Models

   A. Some calculations

   B. Example

   C. Some fun comparisons of frailty vs. shared frailty.

V. Frailty and Cox Regression

VI. Conclusions

INTRODUCTION

Basic Concepts of Survival Analysis

T – Response, time to failure

x^t – row vector of covariates

$[T|x^t\beta]$ – some density $f(t)$

Survival analysis characterized by censoring and truncation

Much more convenient to think in terms of survival function $S(t) = P(T > t)$ and hazard function $h(t) = f(t)/S(t)$, i.e. instantaneous probability of failure given survival up to t.

Response is actually the triple (t_0, t, d) where subject observed from $(t_0, t]$ and either failed $(d = 1)$ or was censored $(d = 0)$. The covariates are assumed constant over $(t_0, t]$.

Effect of x^t can either be parameterized as proportional hazards (PH) or accelerated failure time (AFT).

PH assumes

$$h(t_i) = h_0(t_i)\exp(x_i^t\beta)$$

for some baseline hazard $h_0(t)$.

AFT takes

$$S(t_i) = S_0\{\exp(-x_i^t\beta)t_i\}$$

for some baseline survival function $S_0(t)$.

Parametric survival models assume some function form for $h_0(t)$, and hence for $S_0(t)$.

Parametric families supported by Stata (`streg`) are the exponential, Weibull, Gompertz, lognormal, log-logistic, and generalized gamma.

For example, Weibull PH formulation takes $h_0(t) = pt^{p-1}$, and requires the additional estimation of the shape p.

Cox regression is a PH model that makes no assumption about the functional form of $h_0(t)$.

Frailty models

Parametric specification plus covariates can only go so far in explaining the variability in observed time to failure. Excess unexplained variability is known as overdispersion.

Overdispersion is caused either by misspecification or omitted covariates. As such, current model cannot adequately account for why subjects with shorter times to failures are more "frail" than others.

A frailty model attempts to measure this overdispersion by modeling it as resulting from a latent multiplicative effect on the hazard function, i.e. the hazard becomes

$$h(t|\alpha) = \alpha h(t)$$

where $h(t)$ is a hazard function from a model we may have considered previously.

From a PH perspective, it is easy to see how α may correspond to an omitted covariate (or set of covariates).

$$h(t_i|\alpha_i) = \alpha_i h(t_i) = \alpha_i h_0(t_i) \exp(x_i^t \beta)$$

Same goes for AFT models, just harder to see since the frailty enters multiplicatively on the hazard.

Frailty vs. Shared Frailty

Distinction is critical to success in using Stata's `streg, frailty() [shared()]` .

For the jth observation in the ith group, a frailty model treats

$$h(t_{ij}|\alpha_{ij}) = \alpha_{ij}h(t_{ij})$$

while a shared frailty model has

$$h(t_{ij}|\alpha_i) = \alpha_i h(t_{ij}),$$

i.e., the frailty is shared among the group.

"Group" may represent a family, for example, or simply a single subject for which multiple episodes are observed.

Thinking in terms of omitted variables, a frailty model could be used when you think you lack measurements that vary within the group, or a shared frailty model when you have a latent common group effect.

If considering the analogy to Stata's `poisson` command, a frailty model would be equivalent to `nbreg` while a shared frailty model is analogous to `xtpois`.

Even when you have a single record per subject, the above still represent different models, and hence may give different results.

EXAMPLE – BREAST CANCER DATA

We'll consider this data in one form or another throughout.

80 subjects, time $t = 0$ corresponds to date of diagnosis. Analysis time in years until death or censoring. Covariates are age at diagnosis, smoking status (0/1), and weekly calories from fat in diet ($\times 10^3$).

Subjects observed over two-year intervals where dietary fat remeasured over each interval.

```
. list id _t0 _t _d age smoking dietfat if id==35
```

	id	_t0	_t	_d	age	smoking	diet~t
255.	35	0	2	0	48	0	4.227
256.	35	2	4	0	48	0	4.334
257.	35	4	6	0	48	0	4.239
258.	35	6	8	0	48	0	4.514
259.	35	8	10	0	48	0	4.389
260.	35	10	11.03	1	48	0	4.324

Data generated so that time to failure given the covariates is Weibull.

Omitting a covariate here and there creates "unexplained" heterogeneity which we can capture via a frailty model.

```
. streg age smoking dietfat, dist(weib) nolog

Weibull regression -- log relative-hazard form

No. of subjects =           80                     Number of obs   =       671
No. of failures =           58
Time at risk    =      1257.07
                                                   LR chi2(3)      =    248.31
Log likelihood  =   -14.675006                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.710954    .090628    10.14   0.000     1.542236    1.898129
     smoking |    5.57421   1.831668     5.23   0.000     2.927393    10.61416
     dietfat |   7.977746   1.751895     9.46   0.000     5.187502     12.2688
-------------+----------------------------------------------------------------
       /ln_p |   1.405362   .0968303    14.51   0.000     1.215578    1.595146
-------------+----------------------------------------------------------------
           p |   4.077004   .3947774                      3.372244    4.929049
         1/p |   .2452782   .0237504                      .2028789    .2965384
------------------------------------------------------------------------------
```

PARAMETRIC FRAILTY MODELS

AVAILABILITY: Stata 7

The unconditional survival function

Suppressing the index, recall that

$$h(t|\alpha) = \alpha h(t)$$

for $h(t)$ corresponding to any of our six parametric models.

This implies that the conditional survival function is

$$S(t|\alpha) = \exp\{-\int_0^t h(u|\alpha)du\} = \{S(t)\}^\alpha$$

where, again, $S(t)$ is a survival function to which we are accustomed.

Since α is unobservable we require the unconditional survival function.

For purposes of identifiability, assume the distribution of α has positive support with mean one and variance θ. Problem then reduces to estimating the additional frailty variance θ.

Unconditional survival function is then given by

$$S_\theta(t) = \int_0^\infty \{S(t)\}^\alpha g(\alpha)d\alpha$$

where $g(\alpha)$ is the pdf of α.

We currently offer two choices for $g(\alpha)$.

(1) Gamma$(1/\theta, \theta)$ for which

$$g(\alpha) = \frac{\alpha^{1/\theta - 1} \exp(-\alpha/\theta)}{\Gamma(1/\theta)\theta^{1/\theta}}$$

$$S_\theta(t) = [1 - \theta \ln\{S(t)\}]^{-1/\theta}$$

(2) Inverse–Gaussian$(1, 1/\theta)$ for which

$$g(\alpha) = \left(2\pi\theta\alpha^3\right)^{-1/2} \exp\left\{-\frac{1}{2\theta}\left(\alpha - 2 + \frac{1}{\alpha}\right)\right\}$$

$$S_\theta(t) = \exp\left\{\frac{1}{\theta}\left(1 - [1 - 2\theta \ln\{S(t)\}]^{1/2}\right)\right\}$$

Log–normal distributed α is a possibility, but this would require quadrature.

Using L'Hopital's rule, one can show that $\lim_{\theta \to 0} S_\theta(t) = S(t)$ in either case.

Example

Applying this to our data, we purposely omit the covariate `dietfat` from our model to get some heterogeneity.

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

Weibull regression -- log relative-hazard form
                     Gamma frailty

No. of subjects =           80                     Number of obs   =       671
No. of failures =           58
Time at risk    =      1257.07
                                                   LR chi2(2)      =    135.75
Log likelihood  =   -68.135804                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.475948   .1379987     4.16   0.000     1.228811    1.772788
     smoking |   2.788548   1.457031     1.96   0.050      1.00143    7.764894
-------------+----------------------------------------------------------------
       /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
```

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

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

No. of subjects =           80                     Number of obs   =       671
No. of failures =           58
Time at risk    =      1257.07
                                                   LR chi2(2)      =    125.44
Log likelihood  =   -73.838578                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.284133   .0463256     6.93   0.000     1.196473    1.378217
     smoking |   2.905409   1.252785     2.47   0.013     1.247892    6.764528
-------------+----------------------------------------------------------------
       /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
```

"chibar2" is a result of testing on the boundary. The LR test compares Weibull frailty model to the standard Weibull.

Hazard ratios now have an interpretation that is conditional on the frailty. Unconditionally, hazard ratios are only valid at time 0.

Parameter estimates for AFT models have the same interpretation, either serving to accelerate or decelerate time.

Note the similarity in $\hat{\theta}$ for both models.

Let's now add `dietfat` back in and watch the frailty disappear.

```
. streg age smoking dietfat, dist(weib) frailty(invgauss) nolog

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

No. of subjects =           80                     Number of obs   =       671
No. of failures =           58
Time at risk    =      1257.07
                                                   LR chi2(3)      =    243.77
Log likelihood  =   -14.675007                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.710977   .0906212    10.14   0.000      1.54227    1.898137
     smoking |   5.574535   1.831704     5.23   0.000     2.927638    10.61451
     dietfat |   7.978179    1.75185     9.46   0.000     5.187961    12.26905
-------------+----------------------------------------------------------------
       /ln_p |    1.40539   .0968185    14.52   0.000     1.215629     1.59515
     /ln_the |  -14.73854   1798.306    -0.01   0.993    -3539.353    3509.876
-------------+----------------------------------------------------------------
           p |   4.077115   .3947401                      3.372414     4.92907
         1/p |   .2452715   .0237468                       .202878    .2965235
       theta |   3.97e-07   .0007145                            0           .
------------------------------------------------------------------------------
Likelihood ratio test of theta=0: chibar2(01) =     0.00 Prob>=chibar2 = 1.000
```

Comparing the gamma and inverse–Gaussian

As dissimilar as the frailty survival functions $S_\theta(t)$ appear for the gamma vs. inverse–Gaussian, the associated hazard functions do look a lot alike.

For the gamma,

$$h_\theta(t) = h(t)[1 - \theta \ln\{S(t)\}]^{-1}$$

For the inverse–Gaussian,

$$h_\theta(t) = h(t)[1 - 2\theta \ln\{S(t)\}]^{-1/2}$$

The above equations do, however, highlight an important difference between the two frailty distributions.

Consider two individuals with common frailty. Conditional on the frailty, there respective hazards are proportional with $h^{(2)}(t)/h^{(1)}(t) = c$, say.

Marginally, however, for gamma frailties the hazard ratio $h_\theta^{(2)}(t)/h_\theta^{(1)}(t) = c$ at $t = 0$, but diminishes with time so that

$$\lim_{t\to\infty} \frac{h_\theta^{(2)}(t)}{h_\theta^{(1)}(t)} = 1$$

This is known as the frailty effect, or attenuation due to frailty.

For the inverse-Gaussian, $h_\theta^{(2)}(t)/h_\theta^{(1)}(t) = c$ at $t = 0$ also, however

$$\lim_{t\to\infty} \frac{h_\theta^{(2)}(t)}{h_\theta^{(1)}(t)} = c^{1/2}$$

and so the effect does not completely diminish with time.

Question: Is there a frailty distribution which would allow $h_\theta(t)$ to retain its proportional hazards interpretation?

Answer: Yes. The positive stable distribution. For some $\delta < 1$,

$$g_\delta(\alpha) = \frac{1}{\pi\alpha} \sum_{i=1}^{\infty} \frac{\Gamma(k\delta+1)}{k!}(-\alpha^{-\delta})^k \sin(\delta k\pi)$$

For this frailty distribution

$$\frac{h_\theta^{(2)}(t)}{h_\theta^{(1)}(t)} = c^\delta$$

and so you get a diminished effect, but this is constant over time.

Positive stable family currently not available in Stata, but we're looking to add it.

PARAMETRIC SHARED FRAILTY MODELS

AVAILABILITY: Future ado update to Stata 7.

Some calculations

Recall, for the jth observation in the ith group, a shared frailty model treats

$$h(t_{ij}|\alpha_i) = \alpha_i h(t_{ij})$$

for $i = 1, ..., G$ and $j = 1, ..., n_i$.

Contribution to the likelihood function for a subject who was observed from $(t_{0ij}, t_{ij}]$ is

$$L(t_{ij}|t_{0ij}, \alpha_i) = \left\{\frac{S(t_{ij})}{S(t_{0ij})}\right\}^{\alpha_i} \{\alpha_i h(t_{ij})\}^{d_{ij}}$$

Contribution to the likelihood for the ith group is

$$L(i\text{th group}|\alpha_i) = \alpha_i^{D_i} \prod_{j=1}^{n_i} \left[\left\{\frac{S(t_{ij})}{S(t_{0ij})}\right\}^{\alpha_i} \{h(t_{ij})\}^{d_{ij}}\right]$$

where $D_i = \Sigma_{j=1}^{n_i} d_{ij}$ is the number of deaths in the group.

Unconditionally,

$$L(i\text{th group}) = \int_0^\infty \alpha_i^{D_i} \prod_{j=1}^{n_i} \left[\left\{\frac{S(t_{ij})}{S(t_{0ij})}\right\}^{\alpha_i} \{h(t_{ij})\}^{d_{ij}}\right] g(\alpha_i) d\alpha_i$$

and we are free to choose $g(\alpha_i)$ as before, i.e. gamma or inverse–Gaussian.

Example

Recall, our breast cancer data has multiple records per subject. Let's now leave out `age` to introduce group–level heterogeneity.

```
. streg smoking dietfat, dist(weib) frailty(gamma) nolog

Weibull regression -- log relative-hazard form
                     Gamma frailty

No. of subjects =           80                     Number of obs   =       671
No. of failures =           58
Time at risk    =      1257.07
                                                   LR chi2(2)      =     11.88
Log likelihood  =   -130.06979                     Prob > chi2     =    0.0026

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     smoking |   9.765663   17.58528     1.27   0.206      .286366    333.0289
     dietfat |   5.418364   7.253305     1.26   0.207     .3930114    74.70181
-------------+----------------------------------------------------------------
       /ln_p |   1.281455   .6605948     1.94   0.052    -.0132872    2.576197
     /ln_the |   2.325339   .8010715     2.90   0.004     .7552672     3.89541
-------------+----------------------------------------------------------------
           p |   3.601876    2.37938                      .9868007    13.14704
         1/p |   .2776331    .183403                      .0760627    1.013376
       theta |   10.23014   8.195076                       2.12818     49.1762
------------------------------------------------------------------------------
Likelihood ratio test of theta=0: chibar2(01) =     9.87 Prob>=chibar2 = 0.001
```

Is this really what we want? Probably not.

Let's try this instead:

```
. streg smoking dietfat, dist(weib) frailty(gamma) shared(id) nolog

        failure _d:  dead
  analysis time _t:  t
                id:  id


Weibull regression -- log relative-hazard form
                      Gamma frailty

No. of subjects =           80                     Number of obs   =       671
No. of failures =           58
Time at risk    =      1257.07
                                                   LR chi2(2)      =     11.05
Log likelihood  =   -130.48938                     Prob > chi2     =    0.0040

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     smoking |   5.376692   7.068356     1.28   0.201     .4087904    70.71794
     dietfat |    3.00329   1.869374     1.77   0.077     .8866941    10.17234
-------------+----------------------------------------------------------------
       /ln_p |   .9551898   .4955395     1.93   0.054    -.0160498    1.926429
     /ln_the |   1.923936   .6585433     2.92   0.003     .6332148    3.214657
-------------+----------------------------------------------------------------
           p |   2.599164   1.287988                      .9840783    6.864954
         1/p |   .3847391   .1906534                      .1456674    1.016179
       theta |   6.847858   4.509611                      1.883657    24.89475
------------------------------------------------------------------------------
Likelihood ratio test of theta=0: chibar2(01) =     9.04 Prob>=chibar2 = 0.001
```

Here we know which model is more appropriate, but in practice ask yourself: Do I want observation–level frailty or do I want to impose a grouping constraint on the frailties?

Question: How do we handle `predict`? Do we

(a) Go the `xt` route and give everyone $\alpha = 1$.

(b) Use $\hat{\theta}$ from a shared frailty model and revert to the non-shared forms for $S_\theta(t)$, $h_\theta(t)$, etc.

Some fun comparisons of frailty vs. shared frailty

Comparison I: single record per subject, full time span

Let's drop `dietfat` from out data so that we can collapse our multiple records per subject into single records.

```
. drop dietfat

. stjoin
(option censored(0) assumed)
(591 obs. eliminated)

. list id _t0 _t _d age smoking in 20/30

            id         _t0          _t        _d       age    smoking
 20.        20           0        1.55         1        62          1
 21.        21           0       14.97         1        36          1
 22.        22           0          35         0        29          1
 23.        23           0       13.28         1        41          1
 24.        24           0        1.62         1        53          0
 25.        25           0        1.89         1        59          0
 26.        26           0   26.540001         1        43          0
 27.        27           0       10.86         1        41          0
 28.        28           0   .55000001         1        60          1
 29.        29           0       34.23         1        27          0
 30.        30           0        5.04         1        52          0
```

Surely for these data the frailty and shared frailty models should agree if we specify `shared(id)`, and in fact they do.

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

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 | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.475948   .1379987     4.16   0.000     1.228811    1.772788
     smoking |   2.788548   1.457031     1.96   0.050      1.00143    7.764894
-------------+----------------------------------------------------------------
       /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


. streg age smoking, dist(weib) frailty(gamma) shared(id) nolog

No. of subjects =           80                     Number of obs   =        80
No. of failures =           58
Time at risk    =      1257.07
                                                   LR chi2(2)      =    135.75
Log likelihood  =   -68.135803                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.475948    .137998     4.16   0.000     1.228812    1.772787
     smoking |   2.788548   1.457032     1.96   0.050      1.00143    7.764895
-------------+----------------------------------------------------------------
       /ln_p |   1.087762   .2222597     4.89   0.000     .6521405    1.523383
     /ln_the |   .3307477   .5250732     0.63   0.529    -.6983769    1.359872
-------------+----------------------------------------------------------------
           p |   2.967624   .6595832                      1.919645    4.587717
         1/p |     .33697   .0748948                     .2179733    .5209295
       theta |   1.392009   .7309064                      .497392    3.895696
------------------------------------------------------------------------------
Likelihood ratio test of theta=0: chibar2(01) =    22.57 Prob>=chibar2 = 0.000
```

Comparison II: Non–informative episode splitting

```
. stsplit cat, at(5(5)35)
(205 observations (episodes) created)

. list id _t0 _t _d age smoking if (id==24) | (id==35)

            id         _t0          _t        _d       age    smoking
 73.        24           0        1.62         1        53          0
110.        35           0           5         0        48          0
111.        35           5          10         0        48          0
112.        35          10       11.03         1        48          0
```

By "non–informative" we mean that none of our covariates vary between episodes. Recall, we have dropped `dietfat`.

In this case, again we do not expect to see any difference, and in fact, we don't.

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

No. of subjects =           80                     Number of obs   =       285
No. of failures =           58
Time at risk    =      1257.07
                                                   LR chi2(2)      =    135.75
Log likelihood  =   -68.135804                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.475948   .1379987     4.16   0.000     1.228811    1.772788
     smoking |   2.788548   1.457031     1.96   0.050      1.00143    7.764894
-------------+----------------------------------------------------------------
       /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


. streg age smoking, dist(weib) frailty(gamma) shared(id) nolog

No. of subjects =           80                     Number of obs   =       285
No. of failures =           58
Time at risk    =      1257.07
                                                   LR chi2(2)      =    135.75
Log likelihood  =   -68.135803                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.475947   .1379978     4.16   0.000     1.228812    1.772786
     smoking |   2.788547    1.45703     1.96   0.050     1.001431    7.764889
-------------+----------------------------------------------------------------
       /ln_p |   1.087761   .2222597     4.89   0.000     .6521399    1.523382
     /ln_the |   .3307461   .5250734     0.63   0.529    -.6983788    1.359871
-------------+----------------------------------------------------------------
           p |   2.967622   .6595826                      1.919644    4.587714
         1/p |   .3369702   .0748949                      .2179735    .5209298
       theta |   1.392006   .7309054                       .497391     3.89569
------------------------------------------------------------------------------
Likelihood ratio test of theta=0: chibar2(01) =    22.57 Prob>=chibar2 = 0.000
```

Comparison III: single record per subject, left–truncation

```
. drop cat

. stjoin
(option censored(0) assumed)
(205 obs. eliminated)

. stsplit cat, at(2)
(67 observations (episodes) created)

. drop if _t0==0
(80 observations deleted)

. list id _t0 _t _d age smoking in 20/30

            id         _t0          _t        _d       age    smoking
 20.        23           2       13.28         1        41          1
 21.        26           2   26.540001         1        43          0
 22.        27           2       10.86         1        41          0
 23.        29           2       34.23         1        27          0
 24.        30           2        5.04         1        52          0
 25.        31           2   4.4099998         1        53          0
 26.        32           2   3.3399999         1        52          0
 27.        33           2          35         0        34          0
 28.        34           2          35         0        28          0
 29.        35           2       11.03         1        48          0
 30.        36           2          35         0        39          0
```

Here we will see a difference in model estimations, even though we are running a shared frailty model on groups all of size 1. Why?

In general, if you have time gaps and/or informative episode splitting you are running different models with different assumptions.

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

No. of subjects =           67                     Number of obs   =        67
No. of failures =           45
Time at risk    =      1110.37
                                                   LR chi2(2)      =    101.89
Log likelihood  =   -37.963915                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.741891   .3157656     3.06   0.002     1.221005     2.48499
     smoking |   3.058021   2.526777     1.35   0.176     .6054955    15.44437
-------------+----------------------------------------------------------------
       /ln_p |   1.481744   .3425452     4.33   0.000     .8103676     2.15312
     /ln_the |   .7252144   .5674721     1.28   0.201    -.3870105    1.837439
-------------+----------------------------------------------------------------
           p |   4.400613   1.507409                      2.248734    8.611685
         1/p |   .2272411   .0778403                      .1161213    .4446946
       theta |   2.065174   1.171929                       .679084    6.280435
------------------------------------------------------------------------------
Likelihood ratio test of theta=0: chibar2(01) =    25.87 Prob>=chibar2 = 0.000


. streg age smoking, dist(weib) frailty(gamma) shared(id) nolog

No. of subjects =           67                     Number of obs   =        67
No. of failures =           45
Time at risk    =      1110.37
                                                   LR chi2(2)      =     97.04
Log likelihood  =    -40.38613                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.534071    .155921     4.21   0.000     1.256986    1.872235
     smoking |   3.057159   1.992844     1.71   0.086     .8520254    10.96942
-------------+----------------------------------------------------------------
       /ln_p |   1.305881   .2681155     4.87   0.000      .780384    1.831377
     /ln_the |   .5454618   .4723077     1.15   0.248    -.3802442    1.471168
-------------+----------------------------------------------------------------
           p |   3.690938   .9895978                      2.18231    6.242479
         1/p |   .2709338   .0726416                      .1601928      .45823
       theta |   1.725405    .814922                      .6836944    4.354317
------------------------------------------------------------------------------
Likelihood ratio test of theta=0: chibar2(01) =    21.02 Prob>=chibar2 = 0.000
```

FRAILTY AND COX REGRESSION

AVAILABILITY: future

Frailty models for Cox regression are essential to making our frailty package "complete".

Consider gamma distributed frailties. For ith group, the joint distribution of the shared frailty and the data is

$$\begin{aligned} f(\alpha_i, \mathbf{t}, \mathbf{d}) &= g(\alpha_i) f(\mathbf{t}, \mathbf{d} | \alpha_i) \\ &= g(\alpha_i) \alpha_i^{D_i} \prod_{j=1}^{n_i} \left[h_0(t_{ij}) \exp(x_{ij}^t \beta) \right]^{d_{ij}} \{ S_0(t_{ij}) \}^{\alpha_i \exp(x_{ij}^t \beta)} \end{aligned}$$

$h_0()$ is a nuisance parameter, just like in standard Cox regression.

Can show that distribution of α_i given the observed data is also a gamma, but with different shape and scale. In particular the E–step of an EM algorithm would only require

$$E(\alpha_i | \mathbf{t}, \mathbf{d}) = \frac{1/\theta + D_i}{1/\theta - \Sigma_{i=1}^{n_i} \ln\{S_0(t_{ij})\} \exp(x_{ij}^t \beta)} \equiv \frac{A_i}{C_i}$$

and

$$E\{\ln(\alpha_i)|\mathbf{t}, \mathbf{d}\} = \Psi(A_i) - \ln(C_i),$$

where $\Psi()$ is the digamma function.

The M–Step of EM would then consist of fitting (for a current $\hat{\theta}$) a Cox regression with A_i/C_i as an offset to obtain $\widehat{\beta}$ and an estimate of the baseline survival function $S_0()$.

Using the updated $\widehat{\beta}$ and baseline survival function, we can update $\hat{\theta}$ using the conditional distribution of α_i given the data, which depends on the quantities obtained from Cox.

This EM algorithm is slow to converge, but there exist modifications to make it faster.

CONCLUSIONS

Parametric frailty models offer a generalization of our current models for those who wish to account for unobservable heterogeneity.

There are two types: frailty and shared frailty.

Results can vary according to the choice of frailty distribution, so it is important to offer some variety here.

Frailty for Cox regression is coming.

REFERENCES

Hougaard, P. (1984). Life table methods for heterogeneous populations: distributions describing the heterogeneity. *Biometrika*, **71**, 75–83.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, **73**, 387–96.

Hougaard, P. (1995). Frailty models for survival data. *Lifetime data analysis*, **1**, 255–273.

Kalbfleish, J. D. & R. L. Prentice (1980). *The statistical analysis of failure time data.* New York: John Wiley.

Klein, J. P. & M. L. Moeschberger (1997). *Survival analysis: Techniques for censored and truncated data.* New York: Springer.

Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica*, **47**, 939–56.

Sahu, S. K., D. K. Dey, H. Aslanidou, & D. Sinha. (1997) A Weibull regression model with gamma frailties for multivariate survival data. *Lifetime data analysis*, **3**, 123–137.

Therneau, T.M. & P. M. Grambsch. Penalized cox models and frailty. Working manuscript.