



AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

Parametric frailty and shared frailty survival models

Roberto G. Gutierrez
Stata Corporation

Abstract. Frailty models are the survival data analog to regression models, which account for heterogeneity and random effects. A frailty is a latent multiplicative effect on the hazard function and is assumed to have unit mean and variance θ , which is estimated along with the other model parameters. A frailty model is an heterogeneity model where the frailties are assumed to be individual- or spell-specific. A shared frailty model is a random effects model where the frailties are common (or shared) among groups of individuals or spells and are randomly distributed across groups. Parametric frailty models were made available in Stata with the release of Stata 7, while parametric shared frailty models were made available in a recent series of updates. This article serves as a primer to those fitting parametric frailty models in Stata via the `streg` command. Frailty models are compared to shared frailty models, and both are shown to be equivalent in certain situations. The user-specified form of the distribution of the frailties (whether gamma or inverse Gaussian) is shown to subtly affect the interpretation of the results. Methods for obtaining predictions that are either conditional or unconditional on the frailty are discussed. An example that analyzes the time to recurrence of infection after catheter insertion in kidney patients is studied.

Keywords: st0006, parametric survival analysis, frailty, random effects, overdispersion, heterogeneity

1 Introduction

Parametric survival models are regression models in which the distribution of the response is chosen to be consistent with what one would see if the response is time-to-failure. In particular, the distribution of the response should have positive support. Examples of such distributions are the exponential, Weibull, log-normal, log-logistic, Gompertz, and the generalized gamma, among others. Survival models also differ from standard regression models in their ability to account for censoring and truncation.

For purposes of interpretability, the distribution of time-to-failure is oftentimes characterized by the hazard function, which is the ratio of the probability density function to one minus the cumulative density function. Hazard functions also provide a convenient means to adjust for regressors, either by assuming that the covariates serve to multiplicatively shift the hazard function (proportional hazards) or by assuming that the covariates serve to accelerate or decelerate the effect of time (accelerated failure time).

A frailty model is a generalization of a survival regression model. In addition to the observed regressors, a frailty model also accounts for the presence of a latent multiplicative effect on the hazard function. This effect, or *frailty*, is not directly estimated from the data, but instead is assumed to have unit mean and finite variance, which is estimated. In cases where the frailty is greater than one, subjects experience an increased hazard (or risk) of failure and are said to be more frail than their cohorts. In this way, frailty models can provide a useful alternative to a standard survival model when the standard model fails to adequately account for all the variability in the observed failure times.

Frailty models come in two flavors. The *frailty model* is used with univariate data and is used to model heterogeneity among individuals, analogous to how negative binomial regression generalizes Poisson regression, since the negative binomial may be derived from a Poisson model by introducing a latent gamma-distributed effect. The *shared frailty model* is used with multivariate survival data where the unobserved frailty is shared among groups of individuals, and thus a shared frailty model may be thought of as a random effects model for survival data. In this way, a shared frailty model would be analogous to how a Poisson random effects model (estimated via `xtpois`) generalizes the standard Poisson model.

Frailty models for univariate data have long been used to account for heterogeneous times-to-failure. The term *frailty* was first suggested by Vaupel et al. (1979) in the context of mortality studies, and Lancaster (1979) incorporated the frailty concept into a study of duration of unemployment. Hougaard (1984) discusses the ramifications of the assumed distribution of the frailty, whether gamma or inverse Gaussian. In particular, for the gamma frailty distribution the relative variability of the frailties among survivors is shown to remain constant, whereas for the inverse Gaussian the relative variability decreases with age, making the population of survivors more homogeneous with the passage of time. Hougaard (1986b) considers the positive stable family of distributions for the frailty, not available for estimation in current Stata. The positive stable family is appealing under this family because if individual hazards are proportional, then so are the population hazards. In general, however, the relative shapes of the individual and population (unconditional on the frailty) hazard functions can differ extensively due to the *frailty effect* — the more frail individuals tend to fail earlier, leaving a more homogeneous population of survivors.

Shared frailty models for multivariate data have the added appeal that the frailty can be used to model intragroup correlation. Early considerations of these models can be found in Clayton (1978) and Clayton and Cuzick (1985), and much of the development in this area stems from the extension of methods used to measure correlation in bivariate survival data with arbitrary individual hazard functions (including Cox models). Hougaard (1986a) examines the shared frailty model with Weibull individual hazards, and Whitmore and Lee (1991) look at the inverse Gaussian shared frailty model with constant individual hazards (exponential time-to-failure). Sahu et al. (1997) estimate Bayesian shared frailty models using Gibbs sampling.

Section 2 discusses the univariate (or non-shared) frailty model, and derives the likelihood used to estimate model parameters. A case study involving times to infection after catheter insertion in kidney patients is introduced, and methods for obtaining predictions are given, as well as a note on the interpretation of hazard ratios in the presence of frailty effects. Section 3 mirrors the material covered in Section 2, with the discussion now focused on the multivariate, or shared, frailty model. The kidney data is re-examined in this section. Section 4 describes situations for which both the non-shared and the shared frailty model prove equivalent.

2 Frailty models

For a random time-to-failure, T , we define the probability density function of T as $f(t)$ and the cumulative distribution function as $F(t) = P(T \leq t)$. Two other functions that prove useful in this context are the survivor function $S(t) = P(T > t) = 1 - F(t)$, and the hazard function $h(t) = f(t)/S(t)$, which can be interpreted as the instantaneous rate of failure given survival up until time t .

Consider a parametric survival model characterized by its hazard function, $h(t)$. Implicit in the definitions of all these functions are the effects of any covariates, whether we parameterize the model as having proportional hazards (PH) with respect to changes in covariate values, or accelerated failure time (AFT) due to the covariates. For example, in a Weibull PH regression, the hazard function at time t for individual i with covariate vector \mathbf{x}_i is

$$h_i(t) = \exp(\mathbf{x}_i\beta)pt^{p-1} \quad (1)$$

In general, we use the notation $h_i(t) = h(t|\mathbf{x}_i)$. The shape parameter p and regression coefficients β are estimated from the data. [R] **streg** contains a list of those forms of $h(t)$ currently available in Stata.

A frailty model in the univariate case introduces an unobservable multiplicative effect α on the hazard, so that conditional on the frailty

$$h(t|\alpha) = \alpha h(t) \quad (2)$$

where α is some random positive quantity assumed to have mean one (for purposes of model identifiability) and variance θ . Those individuals who possess $\alpha > 1$ are said to be more frail for reasons left unexplained by the covariates and will have an increased risk of failure. Conversely, those individuals with $\alpha < 1$ are less frail and will tend to survive longer all else being equal (i.e., given a certain covariate pattern). Since α is a multiplicative effect, it is easy to see from (1) how one can think of a frailty as representing the cumulative effect of one or more omitted covariates.

Given the relationship between the hazard and survival functions, it can be shown that the individual survival function conditional on the frailty is $S(t|\alpha) = \{S(t)\}^\alpha$, where $S(t)$ is the survival function from a standard survival model and may include ancillary parameters and covariate effects. The population survival function is then

obtained by integrating out the unobservable α . If α has probability density function $g(\alpha)$, then the population or unconditional survival function is given by

$$S_\theta(t) = \int_0^\infty \{S(t)\}^\alpha g(\alpha) d\alpha \quad (3)$$

where we use the subscript θ to emphasize the dependence on the frailty variance θ .

When α is distributed as gamma with mean one and variance θ ,

$$g(\alpha) = \frac{\alpha^{1/\theta-1} \exp(-\alpha/\theta)}{\Gamma(1/\theta)\theta^{1/\theta}} \quad (4)$$

and (3) becomes

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

When α follows an inverse Gaussian distribution,

$$g(\alpha) = \left(\frac{1}{2\pi\theta\alpha^3}\right)^{1/2} \exp\left\{-\frac{1}{2\theta}\left(\alpha - 2 + \frac{1}{\alpha}\right)\right\} \quad (5)$$

and (3) becomes

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

2.1 Derivation of the likelihood

The unconditional probability of survival past time t is given by $S_\theta(t)$, and we can treat this as we would any other survival function. We can take one minus this function to get the cumulative distribution function (unconditional) of t , and then differentiate to get the probability density function of t . Since $S(t)$ may include covariate effects, $S_\theta(t)$ can also be used to measure the effects of the covariates, although the interpretation requires some special consideration (See Section 2.5). The point is that in the univariate case, a frailty model is really just a survival model with an additional parameter θ . If fitting a Weibull model, we would estimate p and β — with a Weibull–gamma model, we now estimate p , β , and θ . When $\theta = 0$, $S_\theta(t)$ reduces to the standard survival function $S(t)$.

The relationship between the survival function and hazard function still holds unconditional on α , and thus we can obtain the population hazard function using

$$h_\theta(t) = -\frac{d}{dt} S_\theta(t) \{S_\theta(t)\}^{-1}$$

In fact, one can show equivalently that $h_\theta(t) = h(t)E(\alpha|T > t)$. That is, the unconditional hazard is the average hazard over the survivors at any given time (Hougaard 1995).

As with the standard survival models estimated via Stata's `streg` command, the response is organized as the triple (t_{0i}, t_i, d_i) for $i = 1, \dots, n$ with the i th observation corresponding to the time span $(t_{0i}, t_i]$, with either failure occurring at time t_i ($d_i = 1$) or the failure time being right-censored at time t_i ($d_i = 0$). The log likelihood is formed as a combination of the failures and the censored observations in the usual way as

$$\begin{aligned} \ln L &= \ln \prod_{i=1}^n \frac{\{S_{\theta_i}(t_i)\}^{1-d_i} \{f_{\theta_i}(t_i)\}^{d_i}}{S_{\theta_i}(t_{0i})} \\ &= \sum_{i=1}^n [\ln\{S_{\theta_i}(t_i)\} - \ln\{S_{\theta_i}(t_{0i})\} + d_i h_{\theta_i}(t_i)] \end{aligned}$$

where we use the subscript i notation so that, for example, $h_{\theta_i}(t) = h_{\theta}(t|\mathbf{x}_i)$.

2.2 Kidney data

We consider a dataset given in McGilchrist and Aisbett (1991) that examines the recurrence times to infection at point of catheter insertion for kidney patients using portable dialysis equipment. The data consist of 38 patients with two recurrence times given for each, along with each patient's age and gender (0==male, 1==female).

The recurrence time measures the time between catheter insertion and infection, which occurs where the catheter is inserted. When infection occurs, the catheter is removed and the infection is treated, and then, after a pre-determined period of time, the catheter is reinserted. When the catheter is removed for reasons other than infection, the time to infection is treated as censored. As such, the second recurrence time may also be censored if the follow-up period terminated prior to infection.

```
. use kidney, clear
. list in 1/10, nodisplay
      time1   fail1   time2   fail2   age   gender
  1.     16      1      8      1     28      0
  2.     13      0     23      1     48      1
  3.     22      1     28      1     32      0
  4.    318      1    447      1    31.5     1
  5.     30      1     12      1     10      0
  6.     24      1    245      1    16.5     1
  7.      9      1      7      1     51      0
  8.     30      1    511      1    55.5     1
  9.     53      1    196      1     69      1
 10.    154      1     15      1    51.5     0
```

For the purposes of our discussion of the univariate model with heterogeneity, we'll confine ourselves to considering the first recurrence time for each patient. However, in subsequent sections on shared frailty models we do consider both recurrence times.

We fit a Weibull regression model with gamma-distributed heterogeneity using the `frailty(gamma)` option to `streg`.

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	-.0097803	.0144447	-0.68	0.498	-.0380913	.0185307
gender	1.260102	.5566201	2.26	0.024	.1691466	2.351057
_cons	3.653898	.7309858	5.00	0.000	2.221192	5.086604
/ln_sig	.1257008	.2199912	0.57	0.568	-.305474	.5568757
/ln_the	-1.965189	2.298354	-0.86	0.393	-6.46988	2.539502
sigma	1.133943	.2494575			.736774	1.745211
theta	.1401294	.3220669			.0015494	12.67336

Likelihood ratio test of theta=0: $\text{chibar2}(01) = 0.21$ Prob>=chibar2 = 0.324

The above analysis could be redone using the `frailty(invgauss)` option instead of `frailty(gamma)` throughout — thus assuming an inverse Gaussian distribution for the frailties — with similar results.

The log-normal homogeneous model does result in a slightly higher likelihood, yet it is still hard to say which of the two alternate explanations best represent our data. We could further explore this issue by considering a more flexible individual hazard function like a piecewise exponential, and in fact we do this later when we revisit these data in Section 3.

2.3 Obtaining predictions

Predictions are obtained in the usual way using `predict`, and after fitting a univariate frailty model the predictions are automatically adjusted for the estimate of the frailty variance, θ , and the assumed frailty distribution. A wide array of predictions are available, including median times-to-failure, the hazard function, the survival function, martingale residuals, and Cox–Snell residuals to name a few. For a complete list, see [R] `streg`. In order to predict the population hazard and survival function, one could type, for example

```
. predict h, hazard
. predict S, surv
```

which generates the new variables `h` and `S` that respectively contain estimates of $h_\theta(t)$ and $S_\theta(t)$.

We can predict the population hazard functions for both models considered in the previous section using `predict`, and graph them for comparison. Before we do that, however, for plotting purposes we append some dataset containing a grid of values of `_t`. For the appended data, we set the values of `gender` and `age` to baseline. Note that since we have already `stset` the data prior to appending the grid, the estimation sample will remain the same.


```
. set obs 438
obs was 38, now 438
. replace _t = _n - 39 in 39/1
(400 real changes made)
. replace gender = 0 in 39/1
(400 real changes made)
. replace age = 0 in 39/1
(400 real changes made)
```

We can now re-estimate the two models, predict $h_{\theta}(t)$ for both, and graph.

```
. quietly streg age gender, dist(weibull) frailty(gamma) time nohr
. predict h_weib in 39/1, hazard
(38 missing values generated)
. quietly streg age gender, dist(lnormal) frailty(gamma)
. predict h_lnorm in 39/1, hazard
(39 missing values generated)
. graph h_weib h_lnorm _t in 39/1, c(l1[-]) s(ii) sort /*
> */ l1title("Population hazard") /*
> */ key1(c(1) "Weibull") key2(c(1[-]) "Log-Normal")
```

and this produces the graph given in Figure 1.

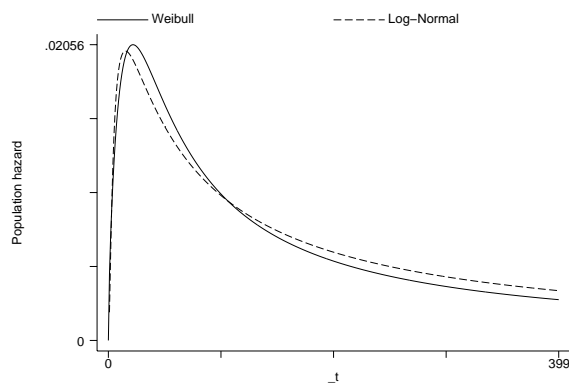


Figure 1: Estimated population hazard functions

Both models give very similar population hazards, even though for the Weibull model the shape is more a result of the frailty. For the Weibull, we estimate the individual hazard $h(t)$ as monotone increasing ($\hat{p} = 1.64$ which is greater than one), yet the population hazard $h_{\theta}(t)$ starts decreasing past a certain point. For the log-normal model, since the frailty variance is estimated to be near zero, the individual hazard would closely resemble the estimate of $h_{\theta}(t)$ given in Figure 1.

2.4 Obtaining conditional predictions

The previous section serves to point out that `predict` used after a univariate frailty model will give population (unconditional) predictions. That is, the predictions are adjusted for the estimated frailty variance θ and for the assumed frailty distribution. Should we want predictions that are instead conditional on arbitrary α , we would have to start from the hazard function given in (2) and work our way through to how α enters into the prediction, given the functional form of the individual hazard, parameter estimates, and covariate values.

Of particular interest is the case where $\alpha = 1$, the mean frailty. When $\alpha = 1$, population or frailty-adjusted quantities such as $h_\theta(t)$ reduce to their respective individual quantities such as $h(t)$. Predictions that condition on $\alpha = 1$ are obtained via the option `alpha1` to `predict`.

For example, in order to compare the individual hazards for the Weibull/gamma versus the Log-normal/gamma models, we repeat the previous example, this time utilizing the `alpha1` option:

```
. quietly streg age gender, dist(weibull) frailty(gamma) time nohr
. predict ih_weib in 39/1, hazard alpha1
(38 missing values generated)
. quietly streg age gender, dist(lnormal) frailty(gamma)
. predict ih_lnorm in 39/1, hazard alpha1
(39 missing values generated)
. graph ih_weib ih_lnorm _t in 39/239, c(11[-]) s(ii) sort /*
> */ l1title("Individual hazard") /*
> */ key1(c(1) "Weibull") key2(c(1[-]) "Log-Normal")
```

The above generates the graph shown in Figure 2, and demonstrates the frailty effect in the Weibull model.

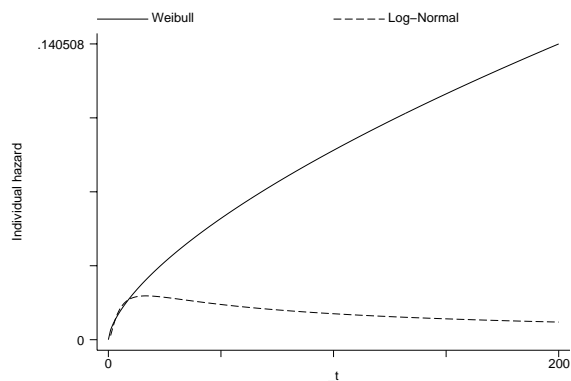


Figure 2: Estimated individual hazard functions

This technique may be used not only to obtain the predicted individual hazard, but to obtain any prediction conditional on $\alpha = 1$. In general, since α is a scalar one may use the relationship in (2) to take a prediction that is conditional on $\alpha = 1$ and transform it into a prediction conditional on any α . For example, from (2) we get $S(t|\alpha) = \{S(t)\}^\alpha$, and so if we wanted to predict the survival function conditional on $\alpha = 0.5$, say, we would simply type

```
. predict S, surv alpha1
. replace S = S^0.5
```

Note: In general, `predict, surv` calculates $S(t)/S(t_0)$, and so if you are using `predict, surv` to obtain $S(t)$ over a plotting grid of values of `_t`, be sure to set `_t0` equal to zero over that grid.

2.5 A note on hazard ratios

When we analyzed the kidney data in Section 2.2, we fit a Weibull/gamma model in the accelerated failure time metric (AFT). We chose this parameterization to ease the comparison with the log-normal model which is parameterized as AFT. However, we could just have easily estimated the Weibull model in the proportional hazards (PH) metric, since both parameterizations are available in `streg`.

```
. use kidney, clear
. stset time1 fail1
  (output omitted)
. streg age gender, dist(weibull) frailty(gamma) nolog
      failure _d:  fail1
      analysis time _t:  time1

Weibull regression -- log relative-hazard form
      Gamma frailty

No. of subjects =          38                Number of obs =          38
No. of failures =          27
Time at risk   =         3630

Log likelihood =  -48.410833                LR chi2(2) =          4.78
                                                Prob > chi2 =          0.0917
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age	1.018152	.0251684	0.73	0.467	.9699991	1.068696
gender	.1109124	.1306617	-1.87	0.062	.0110208	1.116217
/ln_p	.4965477	.31294	1.59	0.113	-.1168033	1.109899
/ln_the	.3610336	.7067754	0.51	0.609	-1.024221	1.746288
p	1.643039	.5141726			.8897602	3.034051
1/p	.6086282	.1904641			.3295923	1.123898
theta	1.434812	1.01409			.3590761	5.733281

```
Likelihood ratio test of theta=0:  chibar2(01) =          4.20 Prob>=chibar2 = 0.020
```

The results are the same as before, except for the regression coefficients, which are now parameterized in the PH metric and displayed (by default) as hazard ratios.

In standard models, a hazard ratio is interpreted as a proportional shift in the hazard function due to a unit change in the associated covariate. When using a frailty model, however, reported “hazard ratios” carry this usual interpretation only if comparing two hazards conditional on a given α . For example, one would interpret the hazard ratio reported for the covariate **gender** as “all other things equal (including the frailty α), the hazard function for females is 0.11 times the hazard for males”.

Unconditionally, however, the proportionality of hazards does not generally hold. When comparing two population hazards, for most distributions of the frailty the reported hazard ratios are only true hazard ratios at time $t = 0$. The exception to this rule is when the frailties follow a distribution from the positive stable family, in which case the hazard ratios remain constant over time; see Hougaard (1986b).

However, for those distributions of frailty currently in Stata, namely the gamma and inverse Gaussian, with the passage of time the effect of the covariates on the population hazard will diminish in favor of the frailty effect. In other words, with the passage of time, gender (for instance) becomes less of a factor and frailty becomes more of a factor in determining the chance of remaining infection free.

It is on this point where we can see a subtle, yet important, difference between the gamma and inverse Gaussian distributions. For gamma frailties, the relationship between individual and population hazards is

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

If we have an individual hazard $h^*(t) = ch(t)$ for some constant c , then $S^*(t) = \{S(t)\}^c$ and the ratio of the population hazards

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

equals c at $t = 0$ but will tend towards 1 as t goes to infinity.

For inverse Gaussian frailties, the relationship between the individual and population hazards is remarkably similar to that for gamma frailties

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

However, in this case an individual hazard ratio of c translates to a population hazard ratio equal to c at $t = 0$ tending towards \sqrt{c} as t goes to infinity. In other words, when the frailty is distributed as inverse Gaussian the effect of the covariates diminishes with time, but never completely vanishes as it does when frailties follow a gamma distribution.

If estimating in the AFT metric, however, the interpretation of regression coefficients remains the same, conditionally and unconditionally. Recall that in a standard (non-frailty) regression the AFT parameterization assumes that

$$S(t) = S^0\{\exp(\mathbf{x}\beta)t\}$$

for some baseline survival function $S^0(t) = S(t|\mathbf{x} = \mathbf{0})$, vector of covariates \mathbf{x} , and regression coefficients β . That is, under this metric the effect of the covariates is to either accelerate or decelerate the effect of time on the probability of survival. Examination of (3) shows that this relationship still holds in a frailty model. That is,

$$S_\theta(t) = S_\theta^0\{\exp(\mathbf{x}\beta)t\}$$

with $S_\theta^0(t) = S_\theta(t|\mathbf{x} = \mathbf{0})$.

3 Shared frailty models

A natural extension of the univariate frailty model would be a multivariate survival model where individuals are allowed to share the same frailty value. Sharing a frailty value also generates dependence between those individuals who share frailties, whereas conditional on the frailty those individuals are independent.

For data consisting of n groups with the i th group comprised of n_i individuals ($i = 1, \dots, n$), (2) generalizes to

$$h_{ij}(t|\alpha_i) = \alpha_i h_{ij}(t)$$

for $j = 1, \dots, n_i$ with $h_{ij}(t) = h(t|\mathbf{x}_{ij})$. That is, for any member of the i th group, the standard hazard function is now multiplied by the shared frailty α_i . For example, in the case of Weibull PH regression, the conditional hazard for an individual is given by

$$h_{ij}(t|\alpha_i) = \alpha_i h_{ij}(t) = \alpha_i \exp(\mathbf{x}_{ij}\beta) p t^{p-1}$$

and the conditional survival function is

$$S_{ij}(t|\alpha_i) = \{S_{ij}(t)\}^{\alpha_i} = \exp\{-\alpha_i \exp(\mathbf{x}_{ij}\beta) t^p\}$$

3.1 Derivation of the likelihood

The likelihood of the observed data is obtained by calculating the group-level conditional likelihoods and integrating out the frailty. Suppose we have data for $i = 1, \dots, n$ groups with $j = 1, \dots, n_i$ observations per group consisting of the trivariate response $(t_{0ij}, t_{ij}, d_{ij})$, which indicate the start time, end time, and failure/censoring for the j th individual from the i th group.

Given α_i , the contribution to the likelihood of the ij th individual is thus

$$\begin{aligned} L_{ij}(\alpha_i) &= \frac{S_{ij}(t_{ij}|\alpha_i)}{S_{ij}(t_{0ij}|\alpha_i)} \{h_{ij}(t_{ij}|\alpha_i)\}^{d_{ij}} \\ &= \left\{ \frac{S_{ij}(t_{ij})}{S_{ij}(t_{0ij})} \right\}^{\alpha_i} \{\alpha_i h_{ij}(t_{ij})\}^{d_{ij}} \end{aligned}$$

and if we define $D_i = \sum_{j=1}^{n_i} d_{ij}$, the likelihood of the i th group is

$$L_i(\alpha_i) = \alpha_i^{D_i} \prod_{j=1}^{n_i} \left\{ \frac{S_{ij}(t_{ij})}{S_{ij}(t_{0ij})} \right\}^{\alpha_i} \{h_{ij}(t_{ij})\}^{d_{ij}}$$

Unconditionally, we can obtain the likelihood of the i th group by integrating out α_i ,

$$L_i = \int_0^\infty L_i(\alpha_i) g(\alpha_i) d\alpha_i$$

where $g()$ is the probability density function of the frailty, given by (4) in the case of gamma, and by (5) for the inverse Gaussian. When the frailties follow a gamma distribution, L_i can be expressed compactly as

$$L_i = \left[\prod_{j=1}^{n_i} \{h_{ij}(t_{ij})\}^{d_{ij}} \right] \frac{\Gamma(1/\theta + D_i) \theta^{D_i}}{\Gamma(1/\theta)} \left\{ 1 - \theta \sum_{j=1}^{n_i} \ln \frac{S_{ij}(t_{ij})}{S_{ij}(t_{0ij})} \right\}^{-1/\theta - D_i}$$

When the frailties follow an inverse Gaussian distribution, the expression for L_i is quite complex, and involves integer half orders of the modified Bessel function of the third kind (commonly known as the BesselK function).

Given the unconditional group likelihoods, we can estimate the regression parameters and frailty variance θ by maximizing the overall log-likelihood $\ln L = \sum_{i=1}^n \ln L_i$.

3.2 Kidney data revisited

Let's once again consider the kidney data from Section 2.2. This time, however, we'll consider both recurrence times given for each patient. The data are in wide form, and so before we `stset` our dataset, we need to `reshape` it into long form.

```
. use kidney, clear
. gen patient = _n
. reshape long time fail, i(patient)
  (output omitted)
. list patient time fail age gender in 1/10
```

	patient	time	fail	age	gender
1.	1	16	1	28	0
2.	1	8	1	28	0
3.	2	13	0	48	1
4.	2	23	1	48	1
5.	3	22	1	32	0
6.	3	28	1	32	0
7.	4	318	1	31.5	1
8.	4	447	1	31.5	1
9.	5	30	1	10	0
10.	5	12	1	10	0

```
. stset time fail
  (output omitted)
```


_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	-.0067002	.0098278	-0.68	0.495	-.0259623	.0125619
gender	1.422234	.3343544	4.25	0.000	.7669111	2.077556
_cons	3.331102	.4911662	6.78	0.000	2.368434	4.29377
/ln_sig	.0670784	.1187767	0.56	0.572	-.1657197	.2998764
/ln_the	-1.823592	.995803	-1.83	0.067	-3.77533	.1281455
sigma	1.069379	.1270173			.8472837	1.349692
theta	.1614447	.1607671			.0229295	1.136718

Likelihood ratio test of theta=0: $\chi^2(01) = 1.57$ Prob>= $\chi^2 = 0.105$

we find the frailty to be insignificant at the 10% level. As in Section 2.2, it is of interest to determine which model better fits the data, the Weibull model with significant frailty effect, or the log-normal model with insignificant frailty.

In order to answer that question, we follow the lead of Sahu et al. (1997), who also fit via Gibbs sampling a piecewise exponential model (piecewise constant hazard) with shared frailty. This requires that we first `stsplit` our records so that we can generate the appropriate indicator variables that specify the pieces. However, before we do that we need to `re-stset` our dataset and assign each record a unique identifier variable.

```
. gen obs = _n
. stset time fail, id(obs)
  (output omitted)
. stsplit my_t, at(1(1)562) /* 562 is the max(_t) */
(7348 observations (episodes) created)
```

Note that we still plan on sharing frailties at the patient level, and the variable `obs` used above merely uniquely identifies the individual observations (distinct recurrences of infection) for purposes of splitting the records. Once we split the records, we generate the indicator variables, which comprise a grid of ten intervals spanning $[0, 60)$, $[60, 120)$, \dots , $[480, 540)$, $[540, \infty)$.

```
. forvalues k = 1/9 {
2.     gen in_`k' = ((`k'-1)*60 <= my_t) & (my_t < `k'*60)
3. }
```

and the indicators are constructed so that the baseline interval (that for which all the indicators equal zero) is the interval $[540, \infty)$.

We are now ready to fit the piecewise exponential/gamma shared frailty model.

```
. streg age gender in_*, dist(exp) frailty(gamma) shared(patient) nolog nohr time
      failure _d: fail
      analysis time _t: time
      id: obs
```



```

Exponential regression -- accelerated failure-time form
                        Gamma shared frailty
No. of subjects =      76                Number of obs =      7424
No. of failures =      58
Time at risk   =      7424
Log likelihood = -95.241838
LR chi2(11)    =      20.90
Prob > chi2    =      0.0344

```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	-.0060301	.0119445	-0.50	0.614	-.0294409	.0173807
gender	1.416426	.4525747	3.13	0.002	.5293963	2.303456
in_1	2.566041	1.350038	1.90	0.057	-.0799854	5.212067
in_2	3.229574	1.339194	2.41	0.016	.6048016	5.854347
in_3	2.366064	1.29691	1.82	0.068	-.1758328	4.907961
in_4	2.1349	1.305189	1.64	0.102	-.4232243	4.693024
in_5	2.44805	1.373428	1.78	0.075	-.2438189	5.139919
in_6	2.788362	1.52622	1.83	0.068	-.2029743	5.779697
in_7	2.659777	1.518699	1.75	0.080	-.3168189	5.636373
in_8	2.319784	1.501516	1.54	0.122	-.6231325	5.262701
in_9	1.233349	1.296951	0.95	0.342	-1.308628	3.775326
_cons	1.057467	1.488013	0.71	0.477	-1.858986	3.973919
/ln_the	-.8007126	.5527555	-1.45	0.147	-1.884094	.2826683
theta	.4490089	.2481921			.1519667	1.326665

```

Likelihood ratio test of theta=0: chibar2(01) =      6.62 Prob>=chibar2 = 0.005

```

The estimated frailty variance using maximum likelihood, $\hat{\theta} = 0.449$, again compares favorably with the Bayesian estimate of the related model of Sahu et al. (1997), for whom $\hat{\theta} = 0.499$. After assuming a more flexible individual hazard, the results seem to favor the Weibull model over the log-normal, and we now have stronger evidence of the existence of latent patient frailty.

3.3 Obtaining conditional predictions

When one specifies the option `shared()` with `streg`, `frailty()`, the default predictions obtained from `predict` are conditional on $\alpha_i = 1$ for all $i = 1, \dots, n$. This is different from when `shared()` is not specified (see Section 2.3), in which case `predict` will give unconditional predictions that are averaged over the distribution of α .

The main reason for this change in behavior in `predict` when going from a univariate frailty model to a shared frailty model has to do with the interpretation of “population” functions like $h_{\theta}(t)$ and $S_{\theta}(t)$. In a non-shared frailty model the interpretation is relatively straightforward — the function $S_{\theta}(t)$ is the new generalized survival function with all the usual interpretations and from which all the likelihood calculations are derived.

In a shared frailty model, however, the function $S_{\theta}(t)$ from (3) is not always a true survival function in the strictest sense. For example, consider a hypothetical population of groups where the relative size of each group is correlated with the frailty for

that group. In this case the use of $S_\theta(t)$ to estimate population dynamics such as the proportion of the population that survives past a given time t would lead to misleading results. One can get around this problem by interpreting $S_\theta(t)$ as giving the probability of survival, say, of a randomly chosen individual from a randomly chosen group, or by assuming (and carefully checking the assumption) that group size is orthogonal to frailty. In any case, there are difficulties with having a statement such as

```
. predict S, surv
```

which implies that one is getting the population “survival” function, when in fact this might not be the case. Of course, we may still be interested in obtaining an estimate of $S_\theta(t)$ if we are willing to make the appropriate assumptions as to its interpretation, and this is easily accomplished as shown in Section 3.4.

For now, however, we can use `predict` to obtain the estimated individual hazard functions (given $\alpha_i = 1$) for the three shared frailty models we have already fit.

```
. use kidney, clear
. gen patient = _n
. reshape long time fail, i(patient)
  (output omitted)
. gen obs = _n
. stset time fail, id(obs)
  (output omitted)
. stssplit my_t, at(1(1)562)      /* _N == 7424, now */
(7348 observations (episodes) created)
. set obs 7824                    /* add 400 observations for graphing */
obs was 7424, now 7824
. replace _t = _n - 7424 in 7425/1 /* generate graphing grid */
(400 real changes made)
. replace my_t = _t in 7425/1     /* indicators for exp graph */
(400 real changes made)
. replace gender = 0 in 7425/1   /* baseline for graphs */
(400 real changes made)
. replace age = 0 in 7425/1
(400 real changes made)
. forvalues k = 1/9 {
  2.     gen in_`k' = (('k'-1)*60 <= my_t) & (my_t < 'k'*60)
  3. }
. quietly streg age gender in_*, dist(exp) fr(gamma) sh(patient) nohr time
. predict h_exp in 7425/1, hazard
(7424 missing values generated)
. quietly streg age gender, dist(weibull) fr(gamma) sh(patient) nohr time
. predict h_weib in 7425/1, hazard
(7424 missing values generated)
. quietly streg age gender, dist(lnormal) fr(gamma) sh(patient)
. predict h_lnorm in 7425/1, hazard
(7424 missing values generated)
```

```

. graph h_exp h_weib h_lnorm my_t in 7425/1, c(11[-]11[.]) s(iii) /*
>      */ sort l1title("Individual hazard") /*
>      */ key1(c(1) "Exponential") key2(c(1[-]) "Weibull") /*
>      */ key3(c(1[.]) "Log-Normal")

```

produces the graph given in Figure 3. The graph shows that there is little agreement among any of the individual hazards past time $t = 100$.

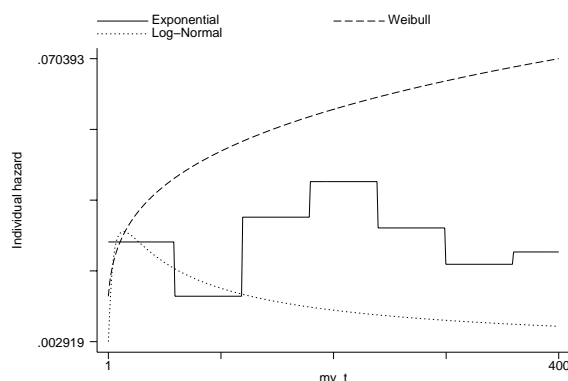


Figure 3: Estimated individual hazard functions

3.4 Obtaining unconditional predictions

For our kidney data example, all “groups” are of size two, and we can envision a population of patients for which two infection recurrence times are measured for each. Thus, in this case the group size is orthogonal to the frailty and population functions such as $h_{\theta}(t)$ and $S_{\theta}(t)$ do carry their usual interpretations. For instance, $S_{\theta}(t)$ could be used to evaluate the proportion of infection-free catheter insertions in the population, adjusted for covariate values.

Overriding the default prediction for the shared frailty case (i.e., that conditional on $\alpha_i = 1$) is done by specifying the `unconditional` option to `predict`. Continuing our previous example, we now predict the population hazards for our three competing shared frailty models and compare.

```

. quietly streg age gender in_*, dist(exp) fr(gamma) sh(patient) nohr time
. predict ph_exp in 7425/1, hazard unconditional
(7424 missing values generated)
. quietly streg age gender, dist(weibull) fr(gamma) sh(patient) nohr time
. predict ph_weib in 7425/1, hazard unconditional
(7424 missing values generated)
. quietly streg age gender, dist(lnormal) fr(gamma) sh(patient)
. predict ph_lnorm in 7425/1, hazard unconditional
(7424 missing values generated)

```

```
. graph ph_exp ph_weib ph_lnorm my_t in 7425/1, c(11[-]1[.]) s(iii) /*
>      */ sort l1title("Population hazard") /*
>      */ key1(c(1) "Exponential") key2(c(1[-]) "Weibull") /*
>      */ key3(c(1[.]) "Log-Normal")
```

produces the graph in Figure 4. Note the similarity in shape of the estimated population hazard functions, regardless of whether the shape is derived from the shape of the individual hazard or from the frailty effect.

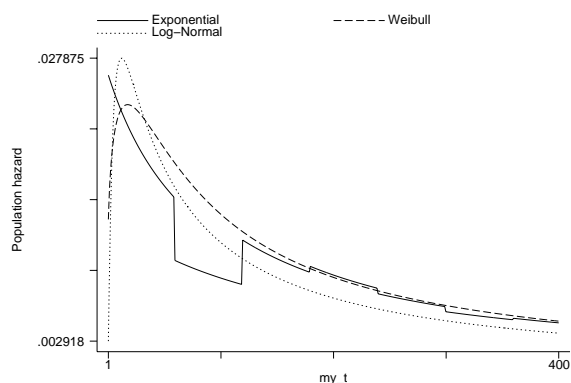


Figure 4: Estimated population hazard functions

Finally, note that one may, rather than rely on defaults, be specific and use `predict, alpha1` to obtain conditional predictions and `predict, unconditional` for unconditional predictions, regardless of which type of frailty model (shared or univariate) is estimated.

4 When is a frailty model equivalent to a shared frailty model?

For certain situations, a frailty model is equivalent to a shared frailty model. For such situations, opting for a shared frailty model either proves to be redundant or merely a consequence of how the data are organized rather than representing a substantive assumption as to the source of the frailty. By contrast, there also exist cases where choosing to share frailties may appear redundant but instead represent a theoretically distinct departure from a univariate non-shared frailty model.

In order to illustrate these cases, we return to the kidney data, where we only consider the first recurrence time for each of the 38 patients as we did in Section 2.2.

```
. use kidney, clear
. gen patient = _n
```

```
. list patient time1 fail1 age gender in 1/10
```

	patient	time1	fail1	age	gender
1.	1	16	1	28	0
2.	2	13	0	48	1
3.	3	22	1	32	0
4.	4	318	1	31.5	1
5.	5	30	1	10	0
6.	6	24	1	16.5	1
7.	7	9	1	51	0
8.	8	30	1	55.5	1
9.	9	53	1	69	1
10.	10	154	1	51.5	0

4.1 Singleton groups of complete observations

We define a *complete observation* to be a single record representing a time span that begins at time $t = 0$ and ends either in failure or right-censoring, in other words, the most simple form of survival data. For the kidney data given above, the entry time $t = 0$ is assumed for each patient since we do not specify otherwise when we `stset` the data.

```
. stset time1 fail1, id(patient)
(output omitted)
. list patient _t0 _t age gender in 1/5
```

	patient	_t0	_t	age	gender
1.	1	0	16	28	0
2.	2	0	13	48	1
3.	3	0	22	32	0
4.	4	0	318	31.5	1
5.	5	0	30	10	0

Since we have only one observation per patient, specifying `id(patient)` when we `stset` our dataset is for most purposes not necessary, but we do so anyway because we plan on `stsplitting` the dataset later.

In this case, when fitting a frailty model whether we choose to share the frailties on the patient level does not matter, and we would get identical results for either model. In order to save space, we demonstrate this fact by only displaying the log likelihoods for each model, but in fact all the parameter estimates, standard errors, etc. would also be identical.

```
. quietly streg age gender, dist(weibull) frailty(invgauss)
. display e(ll)
-49.253638
. quietly streg age gender, dist(weibull) frailty(invgauss) shared(patient)
. display e(ll)
-49.253638
```

The above seems obvious given that we are sharing frailties on groups each consisting of one observation, but as we illustrate below, this is only true because we have singleton groups *and* each singleton observation is complete.

4.2 Singleton groups with left-truncation

Left-truncated observations are those for which subjects are not observed beginning at $t = 0$ but instead initially come under observation at some time $t_0 > 0$. For example, suppose we now assume that our 38 patients came under observation at time $t = 1$; i.e., by the time we initially observe each patient, he has already been infection free up to time $t = 1$. Otherwise we would not have observed him. We represent this assumption by specifying an entry time equal to 1 when we `stset` our dataset,

```
. stset time1 fail1, id(patient) enter(time 1)
(output omitted)
. list patient _t0 _t age gender in 1/5
```

	patient	_t0	_t	age	gender
1.	1	1	16	28	0
2.	2	1	13	48	1
3.	3	1	22	32	0
4.	4	1	318	31.5	1
5.	5	1	30	10	0

and observe that the only difference is that now `_t0==1` for all observations.

In cases where some or all observations are not complete but are instead left-truncated, sharing frailties at the patient level will result in a different estimation, even though the sharing still takes place across singleton groups.

```
. quietly streg age gender, dist(weibull) frailty(invgauss)
. display e(11)
-49.075185
. quietly streg age gender, dist(weibull) frailty(invgauss) shared(patient)
. display e(11)
-49.063301
```

In this case, since the observations are not complete, whether we share frailties at the patient level is relevant to what we assume about the model. If we do not share frailties, we assume that each patient has a spell-specific frailty for the unobserved period $(0, 1]$ and a different frailty for the period afterwards. For example, we can think of the frailty as representing a latent time-dependent covariate with unique values over each period. However, when we do share frailties at the patient level, for each patient the frailties for the two periods (unobserved and observed) are assumed equal. In this case, the frailty may represent a latent covariate that is unique to each patient, yet constant over time. That the period $(0, 1]$ is unobserved doesn't matter since although we do not observe the patient over this period, we still condition on the patient's survival to the end of that period when evaluating the likelihood.

4.3 Uninformative episode splitting

Stata's `st` family of commands are well-suited for dealing with multiple records per individual, with each record representing the time span $(t_0, t]$. Using multiple records is necessary to deal with time-varying covariates, left truncation, and gaps in which subjects are left unobserved. In cases where multiple records are utilized for any or all of these purposes, the splitting of episodes for a given individual is said to be *informative*. Informative episode splitting, for reasons described previously assuming that frailties are shared over these episodes, yields different parameter estimates than those obtained by not sharing frailties and by instead assuming that the frailties are episode-specific.

However, sometimes multiple records exist for individuals as a result of how the data were collected or organized, and the splitting of records does not reflect time-varying covariates, left truncation, or observation gaps. The splitting of records in this case is *uninformative*, and the records could be collapsed into single complete observations without any loss or change of information. In these situations, sharing frailties on the groups formed by having what are really complete observations that have been split uninformatively will produce the same results as if the frailties had not been assumed shared.

To illustrate, let's `stsplit` the complete observations in our kidney data to form a dataset containing multiple records for each patient.

```
. stset time1 fail1, id(patient)
(output omitted)
. stsplit my_t, at(5(5)562)
(702 observations (episodes) created)
. list patient _t0 _t age gender if patient <= 3
```

	patient	_t0	_t	age	gender
1.	1	0	5	28	0
2.	1	5	10	28	0
3.	1	10	15	28	0
4.	1	15	16	28	0
5.	2	0	5	48	1
6.	2	5	10	48	1
7.	2	10	13	48	1
8.	3	0	5	32	0
9.	3	5	10	32	0
10.	3	10	15	32	0
11.	3	15	20	32	0
12.	3	20	22	32	0

Since the splitting was done arbitrarily beginning with a dataset comprised of complete observations, the splitting of records is uninformative with no loss, gain, or change in the information contained in the complete observations. As a result, a model with frailties shared over the multiple records for each patient is equivalent to one with frailties shared on the singleton groups of complete observations, and this in turn is equivalent to a model where the frailties are not shared.

```
. quietly streg age gender, dist(weibull) frailty(invgauss)
. display e(11)
-49.253638
. quietly streg age gender, dist(weibull) frailty(invgauss) shared(patient)
. display e(11)
-49.253638
```

5 References

- Clayton, D. 1978. A model for association in bivariate life tables and its applications in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 65: 141–151.
- Clayton, D. and J. Cuzick. 1985. Multivariate generalisations of the proportional hazards model (with discussion). *Journal of the Royal Statistical Society, Series A* 148: 82–117.
- Hougaard, P. 1984. Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika* 71: 75–83.
- . 1986a. A class of multivariate failure time distributions. *Biometrika* 73: 671–8.
- . 1986b. Survival models for heterogeneous populations derived from stable distributions. *Biometrika* 73: 397–96.
- . 1995. Frailty models for survival data. *Lifetime Data Analysis* 1: 255–273.
- Lancaster, T. 1979. Econometric methods for the duration of unemployment. *Econometrica* 47: 939–56.
- McGilchrist, C. A. and C. W. Aisbett. 1991. Regression with frailty in survival analysis. *Biometrics* 47: 461–466.
- Sahu, S. K., D. K. Dey, H. Aslanidou, and D. Sinha. 1997. A Weibull regression model with gamma frailties for multivariate survival data. *Lifetime Data Analysis* 3: 123–137.
- Vaupel, J. W., K. Manton, and E. Stallard. 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 16: 439–54.
- Whitmore, G. A. and M.-L. T. Lee. 1991. A multivariate survival distribution generated by an inverse Gaussian mixture of exponentials. *Technometrics* 33: 39–50.

About the Author

Roberto G. Gutierrez is Director of Statistics at Stata Corporation.