Statistics Project

A Dissertation on Frailty Models - Survival Analysis

Under the Guidance of Dr. Sumanta Adhya

Kaustav Pal

STSPCOR18M

M.Sc Statistics, SEM IV

July, 2022

ACKNOWLEDGEMENT

Place: Barasat

Date: 22-July-2022

First, I wish to express my sincere gratitude to my project supervisor, Dr. Sumanta Adhya, for his insightful comments and helpful information, practical advice and unceasing ideas that have helped me at all times in my Project. His immense knowledge, profound experience and professional expertise in Survival Analysis and Statistics have enabled me to complete this project successfully. Without his support and guidance, this project would not have been possible.

I would like to express my sincere gratitude to several individuals and West Bengal State University for supporting me throughout my project.

I am ensuring that this project is finished by me.

Kaustav Pal
M.Sc Statistics
SEM-IV
West Bengal State University

Contents:

Abstract

The idea of frailty provides an easier way to introduce random effects, association and unobserved heterogeneity into models for survival data. Frailty models are getting more and more popular to account for overdispersion and clustering of survival data. When, somehow we know the baseline hazard, the parametric estimation approach can be used advantageously. A baseline hazard represents the hazard when all the independent variables are equal to 0.

Introduction

The notion of frailty makes it easier to introduce random effects into statistical models for survival data. In layman's terms, a frailty is an "unobserved random proportionality factor that modifies the hazard function of an individual, or of related individuals". Essentially, the frailty concept dates back to work of Greenwood and Yule (1920) on "accident proneness". The term frailty itself was coined by Vaupel et al. (1979) in univariate survival models and the model was hugely promoted by its application to multivariate survival data in a seminal paper by Clayton (1978) (without using the notion "frailty") on chronic disease incidence in families.

Survival Analysis

Generally, survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs.

By time, we mean years, months, weeks, or days from the beginning of follow-up of an individual until an event occurs; alternatively, time can refer to the age of an individual when an event occurs.

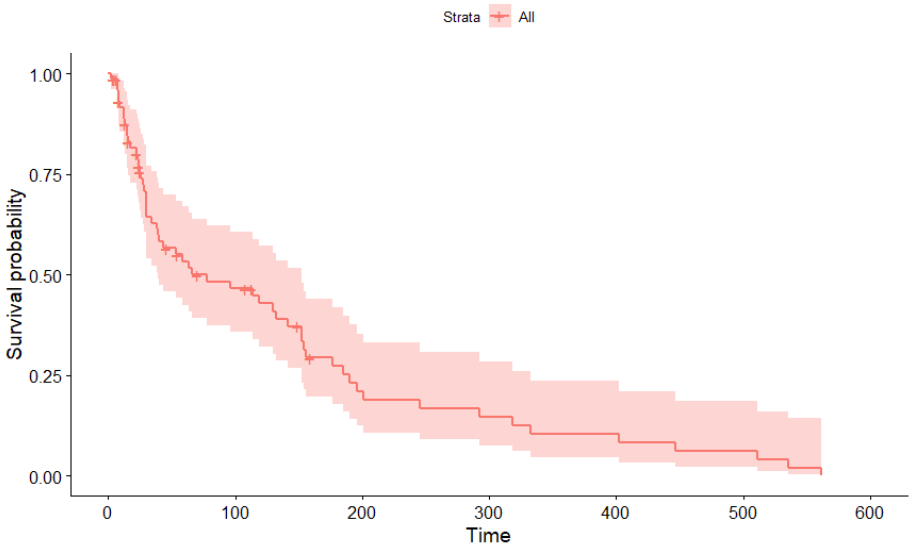
By event, we mean death, disease incidence, relapse from remission, recovery (e.g., return to work) or any designated experience of interest that may happen to an individual.

Survival Functions

The survival function is a function that gives the probability that a patient, device, or other object of interest will survive past a certain time.

We can define it as *T*, given *T* is a continuous RV with CDF *F(t)* on the interval *[0,∞)*. The survival or reliability function then will be:

$$S(t) = P(T > t) = \int_t^{\infty} f(u)du = 1 - F(t)$$



This is the survival plot for the data on the recurrence times to infection, at the point of insertion of the catheter, for kidney patients

Assumptions of Survival Analysis:

One of the basic assumptions for this is the **homogeneity** of the subjects of the population, which is somewhat *questionable*.

As in reality, the individuals differ substantially from each other in susceptibility to various diseases and morbidity status. No matter how many covariates we want to add, it will never be a complete one, especially in the case of human studies.

Standard Survival models also assume independence between the survival times. Frailty models provide an useful extension of the standard survival models by introducing a

Random Effect (Frailty) when the survival data are correlated. Frailty models can be used in the survival analysis to represent random effects or unexplained heterogeneity between individuals or groups.

Multivariate or cluster failure time data are commonly encountered in the survival analysis, and finding an appropriate method to model the correlation among the observations is a very important issue. Frailty models provide an appropriate method to model the correlation among the multivariate data.

Hazard Function

The **hazard function** (also called the *force of mortality*, *instantaneous failure rate*, *instantaneous death rate*, or *age-specific failure rate*) is a way to model data distribution in survival analysis. The most common use of the function is to **model a participant's chance of death as a function of their age**. However, it can be used to model any other time-dependent event of interest. In short, it is the risk of hazard at time ***t***.
More specifically, the hazard function models **which periods have the highest or lowest chances of an event**. The function is defined as the instantaneous risk that the event of interest happens, within a very narrow time frame. (Note: If you're familiar with calculus, you may recognize that this instantaneous measurement is the derivative at a certain point).

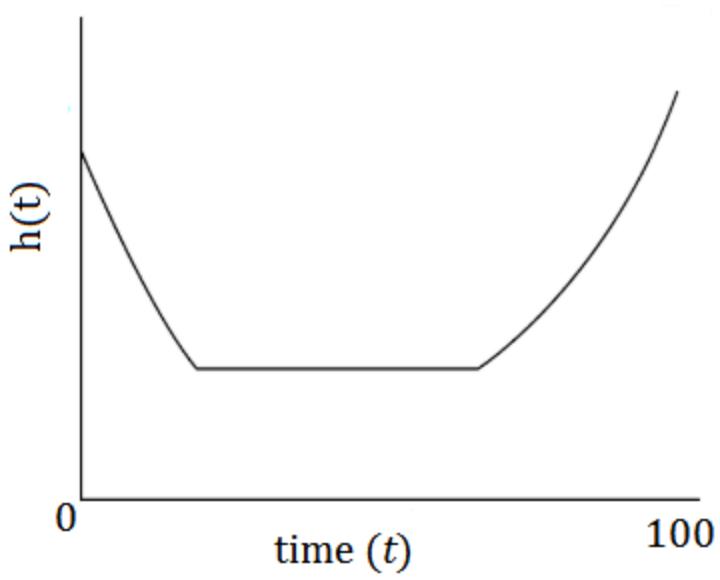
Hazard functions and survival functions are alternatives to traditional probability density functions (PDFs). They are better suited than PDFs for modeling the types of data found in survival analysis.

The hazard function formula is:

$$h_T(t) = \frac{f_T(t)}{S_T(t)}$$

$f_T(t)$ = The pdf of Survival time T

S_T = The Survivor Function



This is an example of a Hazard Function Graph

Censoring of Data

Most Survival Analysis must consider one key analytical problem called censoring. In essence censoring occurs when we have some information about individual survival time, but we don't know the survival time exactly.

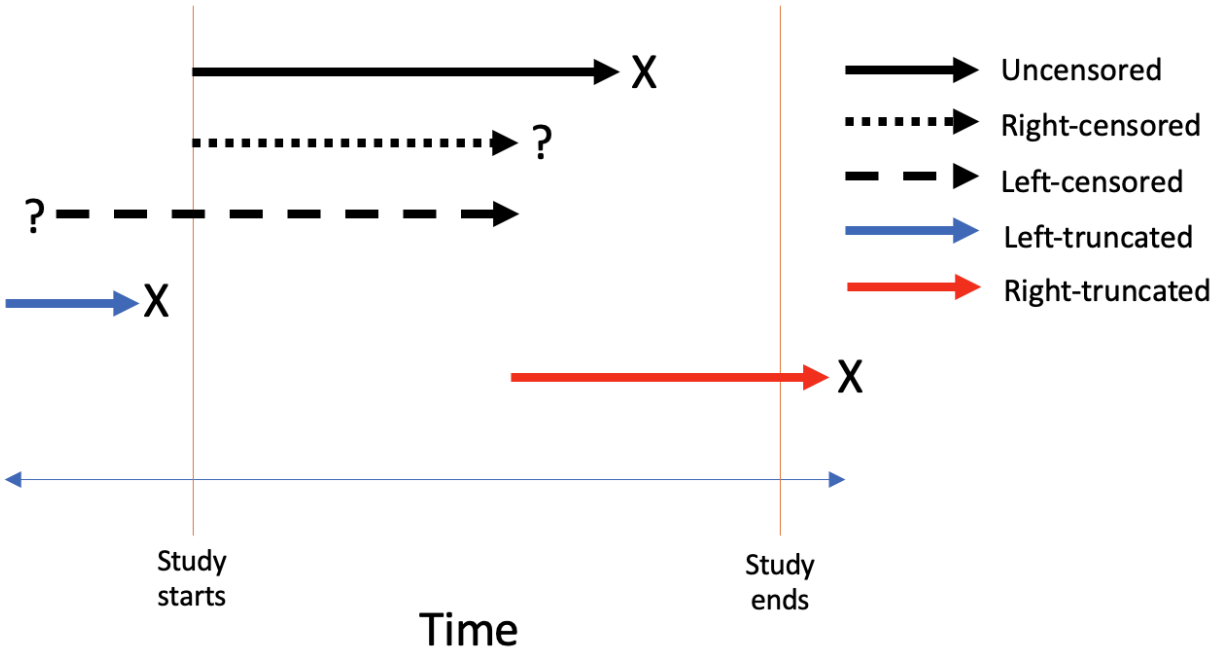
As a simple example of censoring, consider leukemia patients followed until they go out of remission, shown here as X. If for a given patient, the study ends while the patient is still in remission (i.e., doesn't get the event), then that patient's survival time is considered censored. We know that, for this person, the survival time is at least as long as the period that the person has been followed, but if the person goes out of remission after the study ends, we do not know the complete survival time.

There are generally three reasons why censoring may occur:

(1) a person does not experience the event before the study ends;

(2) a person is lost to follow-up during the study period;

(3) a person withdraws from the study because of death (if death is not the event of interest) or some other reason (e.g., adverse drug reaction or other competing risk)



This is an example of different types of Censoring in a diagrammatic representation

Conditional and Variations

The hazard function is a conditional failure rate, in that it is **conditional a person has actually survived until time *t***. In other words, the function at year 10 only applies to those who were actually alive in year 10; it doesn't count those who died in previous periods.

There are other variations on the function, other than as a conditional rate. The Kaplan Meier (KM) method uses rates, has no upper limit, and is preferred for clinical trials (Fink & Brown, 2006). Conversely, with the *actuarial method*, the hazard function is a proportion, with values between 0 and 1.

Cox Proportional Hazard Models

Cox Proportional Hazard Models are used in Survival Analysis to determine the influence of different variables on survival time.

The C-P Hazard Model is defined by the hazard function ***h(t)***

$$h(t) = h_0(t)e^{x'\beta}$$

Or we can expand it to:

$$h(t) = h_0(t) \times exp(b_1 x_1 + b_2 x_2 + \ldots + b_p x_p)$$

Where,

- ***t*** represents the survival time
- ***h(t)*** is the hazard function determined by a set of ***p*** covariates ***(x_1,x_2,...,x_p)***
- the coefficients ***(b_1,b_2,...,b_p)*** measure the impact (i.e., the effect size) of covariates
- the term ***h_0*** is called the **baseline hazard**. It corresponds to the value of the hazard if all the x_i are equal to zero (the quantity ***exp(0)*** equals 1). The ***'t'*** in ***h(t)*** reminds us that the hazard may vary over time.

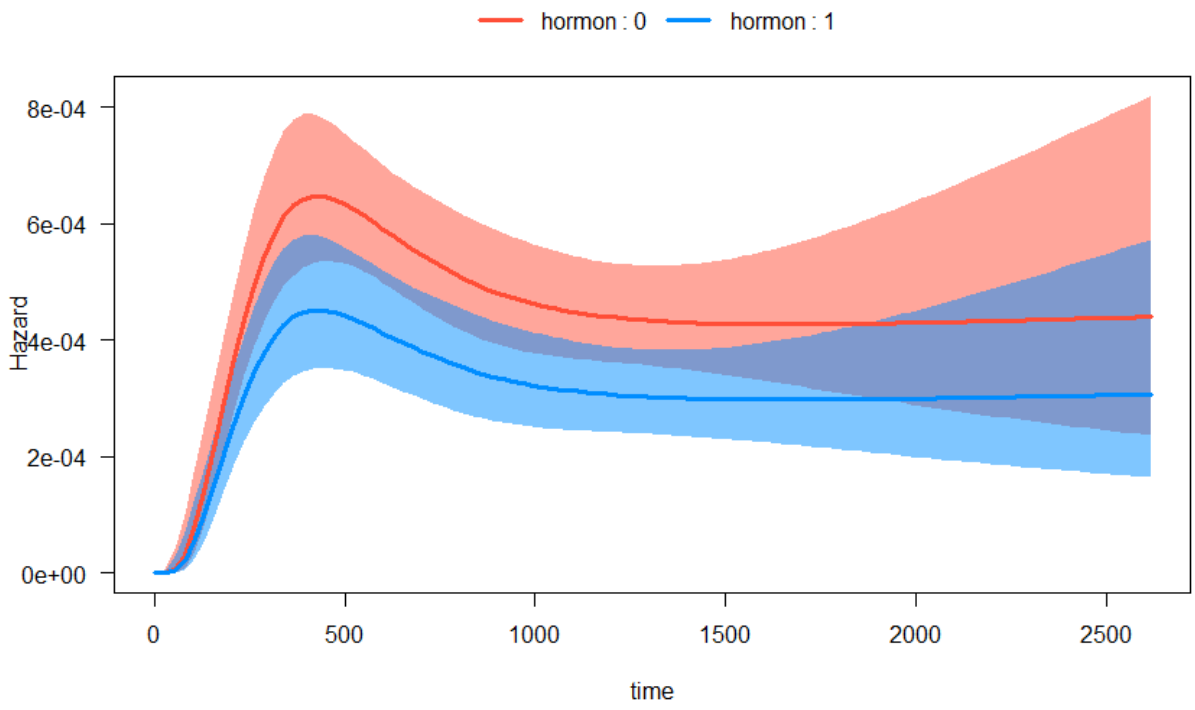
The Cox model can be written as a multiple linear regression of the logarithm of the hazard on the variables ***x_i***, with the baseline hazard being an 'intercept' term that varies with time.

The quantities ***exp(b_i)*** are called hazard ratios (HR). A value of ***b_i*** greater than ***zero***, or equivalently a hazard ratio greater than ***one***, indicates that as the value of the i^{th} covariate increases, the event hazard increases and thus the length of survival decreases.

Put another way, a hazard ratio above 1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival.

In summary,

- HR = 1: No effect
- HR < 1: Reduction in the hazard
- HR > 1: Increase in Hazard



This is the hazard ratio plot for two hormones from the breast cancer dataset

Frailty Models

A frailty model is a random effects model for time variables, where the random effect (the frailty) has a multiplicative effect on the hazard. It can be used for univariate (independent) failure times, i.e. to describe the influence of unobserved covariates in a proportional hazards model. More interesting, however, is to consider multivariate (dependent) failure times generated as conditionally independent times given the frailty.

The motivation for studying Frailty Models are the neglected covariates.

A few notable points about frailty:

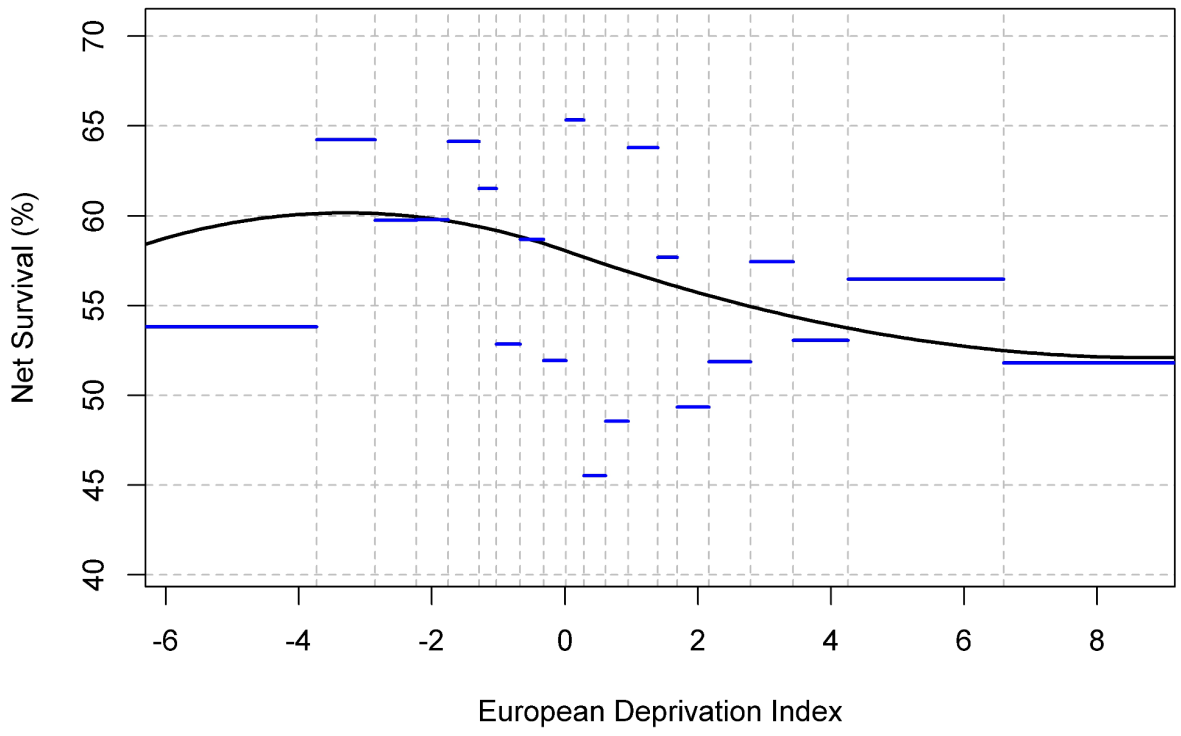
- This effect, or frailty, is not directly estimated from the data, but instead is assumed to have unit mean and finite variance, which is to be estimated.
- This shared frailty parameter is due to some genetic factors which are unobserved in an individual and are common and shared by both paired organs in humans or twins in a family.
- This shared frailty parameter is also responsible for the dependence between the

two components.

Clayton (1978) first used unobserved random covariates in multivariate survival models on chronic disease incidence in families without using the notion “frailty”. The term frailty itself was introduced by Vapuel et al. (1979) in univariate survival models.

Some consequences of ignoring frailty might be:

- If the frailty is significant in the model and is ignored, then the model will be inappropriate and the decision based on such models will be misleading.
- If frailty is ignored, the estimates of regression coefficients will underestimate or overestimate and AIC, BIC, DIC will be more as compared to the model with frailty.
- The decision based on the model without frailty will go wrong.
- In general terms, we let the heterogeneity go into the error term. This will lead to an increase in the variability of the response as compared to the case, when the frailty is included.



This is a diagrammatic example of a joint frailty model to estimate the recurrence process and the disease-specific mortality process without needing the cause of death

We are going to dig deeper into the idea of Frailty Models as we carry on further.

Frailty Models as an extension of Cox Proportional Hazard Model

- A frailty is an extension of the Cox proportional hazard model. In addition to the observed regressors, a frailty model also accounts for the presence of a latent multiplicative effect on the hazard function.
- A frailty is an unobserved random proportionality factor that modifies the hazard function of an individual, or of related individuals.
- This effect, or frailty, is not directly estimated from the data, but instead assumed to have unit mean and finite variance which is to be estimated.
- This shared frailty parameter is due to some genetic factors or environmental factors which are unobserved in an individual and are common and shared by both paired organs in humans or twins in a family.
- This shared frailty parameter is also responsible for the dependence between two components.

Cox PH Model:

$$h(t) = h_0(t)e^{x'\beta}$$

Unobserved covariate, (U):

$$h(t) = h_0(t)e^{x'\beta+u'\beta^*}$$

$$h(t|z) = zh_0(t)e^{x'\beta}$$

Where,

$$z = e^{u'\beta^*}$$

Conditional Survival Function:

$$S(t|z) = exp(-zH_0(t)e^{x'\beta})$$

Where $H_0(t)$ is the integrated hazard.

Unconditional Survival Function:

$$S(t) = E_z[e^{-zH_0(t)e^{x'\beta}}] = L_z[H_0(t)e^{x'\beta}]$$

Shared Gamma Frailty Models

Let us consider an RV Z follows a Gamma distribution which has an expected value of **1**. Under restriction, the density function and the Laplace transformation of a gamma distribution reduces to,

$$f_z(z) = \begin{cases} \frac{\frac{1}{\theta}^{\frac{1}{\theta}}}{\Gamma\frac{1}{\theta}} z^{\frac{1}{\theta}-1} e^{\frac{-z}{\theta}}; & z>0, \theta>0 \\ 0; & o.w \end{cases} \quad \ldots.(1)$$

And,

$$L_z(s) = (1+\theta s)^{-\frac{1}{\theta}}$$

with variance of Z as θ.

For the two component system:

$$S(t_1, t_2) = L_z[(H_0(t_1) + H_0(t_2))e^{x'\beta}]$$

Where T_1 and T_2 are independent given the frailty, Z. When the frailty variable is integrated out, T_1 and T_2 will become dependent because the frailty is common to both components and it is called the Shared Frailty Model.

Unconditional bivariate survival function for the j[th] individual at the time $t_{1j}>0$ and $t_{2j}>0$ as,

$$S(t_{ij}, t2j) = [1 + \theta\eta((H_{01}(t_{1j}) + H_{02}(t_{2j})))]^{-\frac{1}{\theta}} \quad \text{....(2)}$$

Where, $H_{01}(t_{1j})$ and $H_{02}(t_{2j})$ are cumulative baseline hazard functions of the lifetime random variables T_{1j} and T_{2j} respectively. The bivariate distribution in the presence of covariates, when the frailty variable is degenerate is given by,

$$S_{WOF}(t_{1j}, t_{2j}) = exp[-\eta((H_{01}(t_{1j}) + H_{02}(t_{2j})))] \quad \text{.....(3)}$$

According to different assumptions on the baseline distributions, we get different shared gamma frailty models.

If frailty acts additively on hazard rate function, in such situations, an univariate frailty model is derived as given below. Let a continuous RV T be a lifetime of an individual and the RV Z be a frailty variable. The conditional hazard function for a given frailty variable, $Z=z$ at time $t>0$ is,

$$h(t|z) = h_0(t) + e^{x\beta + U\beta_U}$$

$$h(t|z) = h_0(t) + ze^{x\beta}, z > 0, -\infty < U < \infty \quad \text{....(4)}$$

Where $z=exp(U\beta_U)$, $h_0(t)$ is a baseline hazard function at time $t>0$, X is a row vector of covariates, and β is a column vector of regression coefficients.

The cumulative hazard rate function is given by:

$$H(t|z) = H_0(t) + zte^{X\beta} \quad \text{.....(5)}$$

The conditional survival function for given frailty at time $t>0$ is,

$$S(t|z) = exp[-\int_0^t h(x|z)dx] = exp[H_0(t) + zte^{X\beta}] \quad \text{.....(6)}$$

Where $H_0(t)$ is the cumulative baseline hazard function at time $t>0$.

Integrating over the range of frailty variable Z having density $f(z)$, we get the marginal survival function as,

$$S(t) = \int_0^\infty S(t|z)f(z)dz$$

$$S(t) = \int_0^\infty exp[-(H_0(t) + zte^{\mathbf{X}\beta})]f(z)dz$$

$$S(t) = exp[-H_0(t)]L_Z(te^{\mathbf{X}\beta}) \quad \text{.....(7)}$$

For the Gamma Distribution, the Kendall's τ, which measures the association between any two event times from the same cluster in the multivariate case, can be computed as:

$$\tau = \frac{\theta}{\theta+2} \in (0,1)$$

General Shared Frailty Model

Suppose n individuals are observed for the study and let a bivariate RV (T_{1j},T_{2j}) represent the first and the second survival times of the j^{th} individual $(j=1,2,3...n)$. Also, suppose that there are k observed covariates collected in a vector $X_j=(X_{1j},X_{2j},...,X_{kj})$ for the jth individual where X_{aj} $(a=1,2...k)$ represents the value of the a^{th} observed covariate for the j^{th} individual. Here, we assume that the first and the second survival times for each individual share the same value of the covariates. Let Z_j be shared frailty for the j^{th} individual. Assuming that the frailties are acting additively on the baseline function and both the survival times of individuals are conditionally independent for giving frailty, the conditional hazard function for the j^{th} individual at the i^{th} $(i=1,2)$ survival time $T_{ij}>0$ for given frailty $Z_j=z_j$ has the form,

$$h(t_{ij}|z_j, \mathbf{X_j}) = h_0(t_{ij}) + z_j e^{\mathbf{X_j}\beta} \quad \text{.....(8)}$$

Where $h_0(t_{ij})$ is the baseline hazard at time $t_{ij}>0$ and β is a vactor of order k, of the regression coefficients. The conditional cumulative hazard function for the j^{th} individual at the i^{th} survival time $t_{ij}>0$ for a given faulty $Z_j=z_j$ is,

$$h(t_{ij}|z_j, \mathbf{X_j}) = H_0(t_{ij}) + z_j t_{ij} \eta_j \quad \ldots(9)$$

Where $\eta_{ij}=exp(X_j\beta)$ and $H_0(t_{ij})$ is the cumulative baseline hazard function at time $t_{ij}>0$. The conditional survival function for the jth individual at the ith survival time the $t_{ij}>0$ for a given frailty $Z_j=z_j$ is,

$$S(t_{ij}|z_j, X_j) = exp[-H(t_{ij}|z_j, X_j)]$$

$$S(t_{ij}|z_j, X_j) = exp[-(H_0(t_{ij}) + z_j t_{ij} \eta_j)] \quad \ldots(10)$$

Under the assumption of independence, the bivariate conditional survival function for a given frailty, $Z_j=z_j$ at time $t_{1j}>0$ and $t_{2j}>0$ is,

$$S(t_{1j}, t_{2j}|z_j, X_j) = S(t_{1j}|z_j, \mathbf{X_j}) S(t_{2j}|z_j, \mathbf{X_j})$$

$$S(t_{1j}, t_{2j}|z_j, X_j) = exp[-(H_{01}(t_{1j}) + H_{02} + t_{2j}) + z_j(t_{1j} + t_{2j}\eta_j)] \quad \ldots(11)$$

The unconditional bivariate survival function at time $t_{1j}>0$ and $t_{2j}>0$ can be obtained by integrating over the frailty variable Z_j having the probability function $f_Z(Z_j)$, for the j^{th} individual.

$$S(t_{1j}, t_{2j}|\mathbf{X_j}) = \int_{Zj} S(t_{1j}, t_{2j}|z_j) f_Z(z_j) dz_j$$

$$S(t_{1j}, t_{2j}|\mathbf{X_j}) = \int_{Zj} e^{-[(H_{01}(t_{1j}) + H_{02}(t_{2j})) + z_j(t_{1j} + 1_{2j})\eta_j]} f_z(z_j) dz_j$$

$$S(t_{1j}, t_{2j}|\mathbf{X_j}) = e^{-(H_{01}(t_{1j}) + H_{02}(t_{2j}))} L_{Z_j}[(t_{1j} + t_{2j})\eta_j] \quad \ldots(12)$$

The equations (2) and (12) provide us the survival of t_{1j} and t_{2j} with frailty. One is for the multiplicative structure and another is for the additive structure. While equation (3) provides us survival time without frailty.

Replacing the Laplace transformation in equation (12), we get the unconditional

bivariate survival function for the jth individual at the time $t_{1j}>0$ and $t_{2j}>0$ as,

$$S(t_{1j}, t_{2j}) = e^{-(H_{01}(t_{1j})+H_{02}(t_{2j}))}[1+\theta\eta_j(t_{1j}+t_{2j})]^{-\frac{1}{\theta}} \quad \ldots(13)$$

Where $H_{01}(t_{1j})$ and $H_{02}(t_{2j})$ are cumulative baseline hazard functions of the lifetime random variables T_{1j} and T_{2j} respectively. The Bivariate distribution in the presence of covariates, when the frailty variable is degenerate is given by,

$$S(t_{1j}, t_{2j}) = e^{-[(H_{01}(t_{1j})+H_{02}(t_{2j}))+\eta_j(t_{1j}+t_{2j})]} \quad \ldots(14)$$

According to different assumptions on the baseline distributions, we get different shared gamma frailty models.

Generalized log-logistic distribution

If a continuous RV T follows Generalized log-logistic distribution then the distribution function, cumulative hazard rate function and the hazard rate are as follows:

Distribution Function:

$$F(t) = 1-(1+\lambda t^{\gamma})^{-\alpha} \quad \ldots(15)$$

Hazard Function:

$$h(t) = \alpha\frac{\lambda\gamma t^{\gamma-1}}{1+\lambda t^{\gamma}} \quad \ldots(16)$$

Cumulative Hazard Function:

$$H(t) = \frac{\alpha}{n(1+\lambda t^{\gamma})} \quad \ldots(17)$$

Generalized Weibull Distribution

If a continuous RV T follows Generalized Weibull Distribution then the distribution function, cumulative hazard rate function and the hazard rate are respectively:

$$F(t) = (1 - e^{-(\lambda t)^{\gamma}})^{\alpha}, t > 0, \alpha > 0, \lambda > 0, \gamma > 0 \quad \ldots(18)$$

$$H(t) = -ln[1 - (1 - e^{-(\lambda t)^{\gamma}})^{\alpha}] \quad \ldots(19)$$

$$h(t) = \frac{\lambda^{\gamma}\gamma\alpha(t)^{\gamma-1}e^{-(\lambda t)^{\gamma}}(1-e^{-(\lambda t)^{\gamma}})^{\alpha-1}}{1-(1-e^{-(\lambda t)^{\gamma}})^{\alpha}} \quad \ldots(20)$$

Proposed Multiplicative Models

Substituting cumulative hazard function for the generalized log-logistic and the generalized Weibull baseline distribution in equation (2) and equation (3), we get unconditional bivariate survival functions at time $t_{1j}>0$ and $t_{2j}>0$ as,

Baseline: Generalized Log-Logistic, Frailty: Gamma

$$S(t_{1j}, t2j) = [1 + \theta\eta_j(\alpha_1 \ln(1 + \lambda_1 t_{1j}^{\gamma_1}) + \alpha_2 \ln(1 + \lambda_2 t_{2j}^{\gamma_2}))]^{-\frac{1}{\theta}}$$

Without Frailty:

$$S_{WOF}(t_{1j}, t2j) = exp[-\eta_j(\alpha_1(\ln(1 + \lambda_1 t_{1j}^{\gamma_1}) + \alpha_2 \ln(1 + \lambda_2 t_{2j}^{\gamma_2})))]$$

Baseline: Generalized Weibull, Frailty: Gamma

$$S_(t_{1j}, t2j) = [1 - \theta\eta_j(\ln(1 - (1 - e^{-\lambda_1 t_{1j}^{\gamma_1}})^{\alpha_1}] + \ln[1 - (1 - e^{-\lambda_2 t_{2j}^{\gamma_2}})^{\alpha_2}])]^{-\frac{1}{\theta}}$$

Without Frailty:

$$S_{WOF}(t_{1j}, t2j) = exp[\eta_j(\ln(1 - (1 - e^{-\lambda_1 t_{1j}^{\gamma_1}})^{\alpha_1} + \ln(1 - (1 - e^{-\lambda_2 t_{2j}^{\gamma_2}})^{\alpha_2}]$$

From now on, we mention equation (10), (11), (12) and (13) as Model 1,2,3 and 4 respectively.

Copula Representation

To every bivariate distribution function ***F(t₁,t₂)*** with absolute marginal distribution functions ***F(t₁)*** and ***F(t₂)***, corresponds a unique function

$$C : [0,1] \times [0,1] \rightarrow [0,1]$$

Is called a Copula such that:

$$F(t_1, t_2) = C(F(t_1), F(t_2)) \quad \text{for} \quad (t_1, t_2) \in (0,\infty) \times (0,\infty)$$

For a given Copula ***C***, there exists a unique survival Copula ***C'***, such that

$$C'(u,v) = u + v - 1 + C(1-u, 1-v)$$

For details, Nelsen (2006) can be checked. For eq (10) and (12), we get survival copula,

$$C'(u,v) = [u^{-\theta} + v^{-\theta} - 1]^{-\frac{1}{\theta}}$$

Where $u = S_{T_1}(.)$ and $v = S_{T_2}(.)$

For Model 1

$$S_{T_i}(t-i) = [1 + \theta\eta(\alpha \ln(1 + \lambda_i t_i^{\gamma_i})]^{-\frac{1}{\theta}} \quad \text{i=1,2}$$

For Model 3

$$S_{T_i}(t-i) = [1 - \theta\eta(1 - (1 - e^{-\lambda_i t_i^{\gamma_i}})^{\alpha_i})]^{-\frac{1}{\theta}} \quad \text{i=1,2}$$

Literature Review

Frailty models are an extension to the **Cox-proportional** hazards model (Cox, 1972), the foremost popular statistical model in survival analysis. Generally, the assumption of a homogenous population to be studied is inferred in most clinical applications of survival

analysis. This implies that each one individual within the study sample is in theory subjected to the same risk (e.g., risk of death or risk of disease relapse). On several occasions, we aren't able to assume the study population to be homogeneous and are compelled to consider it as a heterogeneous sample, i.e. a mixture of individuals with different hazards. For instance, in many cases it's impossible to measure all relevant covariates related to the disease of interest, sometimes thanks to economical reasons, sometimes the importance of some covariates is still unknown. The frailty model approach is a statistical modeling concept which aims to account for heterogeneity, caused by unmeasured covariates. In statistical terms, a frailty is a random effect for time-to-event data, where the random effect (the frailty) has a multiplicative effect on the baseline hazard function. One can distinguish two broad classes of frailty models:

1. models with an univariate survival time as endpoint and
2. models which describe multivariate survival endpoints (e.g; competing risks, recurrence of events within the same individual, occurrence of a disease in relatives).

In the first case, an univariate (independent) lifetime is used to explain the influence of unobserved covariates in a proportional hazards model (heterogeneity). The variability of survival data is split into an element that depends on risk factors, and is therefore theoretically predictable, and a component that's initially unpredictable, even when all relevant information is known. A separation of these two sources of variability has the advantage that heterogeneity can explain some unexpected results or give an alternate interpretation of some results, for instance, crossing-over effects or convergence of hazard functions of two different treatment arms (see Manton and Stallard (1981)) or leveling-off effects - which means the decline in the increase of mortality rates, which could result in a hazard function at old ages parallel to the x-axis (Aalen and Tretli (1999)). More interesting, however, is the second case when multivariate survival times are considered. There one aims to account for the dependence in clustered event times, for instance in the lifetimes of patients in study centers in a multi-center clinical trial, caused by center-specific conditions (see Andersen et al. (1999)). A natural way to model dependence of clustered event times is through the introduction of a cluster-specific random effect - the frailty. This random effect explains the dependence in the sense that had we known the frailty, the events would be independent. In other words, the lifetimes are conditional independent, given the frailty. This approach may be used for survival times of related individuals like family members or recurrent observations on the same

person. Different extensions of univariate frailty models to multivariate models are possible and will be considered below.

We want to discuss the key ideas of univariate frailty models by an illustrative example from Aalen and Tretli (1999). The authors analyzed the incidence of testis cancer by means of a frailty model based on data from the Norwegian Cancer Registry collected during 1953-93. The incidence of testicular cancer is greatest among younger men, and then declines from a certain age. The frailty is considered to be established by birth, and due to a mixture of genetic and environmental effects. The idea of the frailty model is that a subgroup of men is particularly vulnerable to testicular cancer. This would explain why testis cancer is primarily a disease of young men. As time goes by the members of the frail group acquire the disease, and at some age this group is more or less exhausted. Then the incidence, computed on the basis of all men at a particular age, will necessarily decline.

Univariate frailty models

The standard situation of the application of survival methods in clinical research projects assumes that a homogeneous population is investigated when subject under different conditions (e.g. experimental treatment and standard treatment). The suitable survival model then assumes that the survival data of the different patients are independent from one another and that each patient's individual survival time distribution is the same (independent and identically distributed failure times).

This basic presumption implies a homogeneous population. However, in the field of clinical trials one observes in many most practical situations that patients differ substantially. The effect of a drug, a treatment or the influence of various explanatory variables may differ greatly between subgroups of patients. To account for such unobserved heterogeneity in the study population Vaupel et al. (1979) introduced univariate frailty models into survival analysis. The key idea is, that individuals possess different frailties, and that those patients who are most frail will die earlier than the others. Consequently, systematic selection of robust individuals (that means patients with low frailty) takes place. When mortality rates are estimated, one may be interested in how these rates change over time or age. Quite often it is observed that the hazard function (or mortality rate) rises at the beginning, reaches a maximum, and then declines (unimodal intensity) or levels-off at a constant value. The longer the patient lives after manifestation of the disease, the more improved are his or her chances of survival. It is

likely that unimodal intensities are often a result of a selection process acting in a heterogeneous population and do not reflect individual mortality. The population intensity may start to decline simply because the high-risk individuals have already died out. The hazard rate of a given individual might well continue to increase. If protective factors or risk factors are known, those could be included in the analysis by using the proportional hazards model, which is of the form

$$\mu(t, \mathbf{X}) = \mu_0(t) exp(\beta^T \mathbf{X})$$

Where $\mu_0(t)$ denotes the baseline hazard function, assumed to be unique for all individuals in the study population. $\underline{X}$ is a vector of observed covariates and β the respective vector of regression parameters to be estimated. The mathematical convenience of this model is based on the separation of the effects of aging in the baseline hazard $\mu_0(t)$ from the effects of covariates in the parametric term $exp(\beta^T X)$.

There are two main reasons why it is often impossible to include all important factors on the individual level into the analysis. Sometimes there are too many covariates to be considered in the model, in other cases the researcher does not know or is not able to measure all the relevant covariates. In both cases, there are two sources of variability in survival data: variability accounted for by measurable risk factors, which is thus theoretically predictable, and heterogeneity caused by unknown covariates, which is thus theoretically unpredictable, even if knowing all the relevant information. There are advantages to separating these two sources of variability since heterogeneity in contrast to variability can explain some "unexpected" results or can provide an alternative explanation of some results. Consider for example, non-proportional hazards or decreasing hazards when unexpected extra variability prevails.

In a proportional hazards model, neglect of a subset of the important covariates leads to biased estimates of both regression coefficients and the hazard rate. The reason for such bias lies in the fact that the time-dependent hazard rate results in changes in the composition of the study population over time with respect to the covariates.

If there are two groups of patients in a clinical trial where some individuals experience a higher risk of failure, then the remaining individuals at risk tend to form a more or less selected group with a lower risk. An estimate of the individual hazard rate, without taking into account the unobserved frailty, would therefore underestimate the true hazard function and the extent of underestimation would increase as time progresses.

The univariate frailty model extents the Cox model such that the hazard of an individual

depends in addition on an unobservable random variable ***Z***, which acts multiplicatively on the baseline hazard function ***μ***:

$$\mu(t, \mathbf{Z}, \mathbf{X}) = \mathbf{Z}\mu_0(t)exp(\beta^T \mathbf{X}) \quad \ldots(1)$$

Again, $\mu_0(t)$ is the baseline hazard function, the vector of regression coefficients, X is the vector of observed covariates. and ***Z*** now is the frailty variable. The frailty ***Z*** is a random variable varying over the population which lowers ***(Z<1)*** or increases ***(Z>1)*** the individual risk. Frailty corresponds to the notions of liability or susceptibility in different settings (Falconer, 1967). The most important point here is that frailty is unobservable. The respective survival function ***S***, describing the fraction of surviving individuals in the study population, is given by

$$S(t|\mathbf{Z}, \mathbf{X}) = exp[-\mathbf{Z}\int_0^1 \mu_0(s)ds \cdot exp(\beta^T \mathbf{X})] \quad \ldots(2)$$

S(t|Z,X) may be interpreted as the fraction of individuals surviving the time ***t*** after the beginning of follow-up given the vector of observable covariates ***X*** and frailty ***Z***. Note, that relations (1) and (2) describe the same model using different notions. Up to now, the model has been described at the level of individuals. However, this individual model is not observable. Consequently, it is necessary to consider the model at the population level. The survival function of the total population is the mean of the individual survival functions (2). It can be viewed as the survival function of a randomly drawn individual, and corresponds to that which is actually observed. It is important to note that the observed hazard function will not be similar to the individual hazard rate. What may be observed in the population is the net result for a number of individuals with different ***Z***. The population hazard rate may have a completely different shape than the individual hazard rate as shown in the following picture:

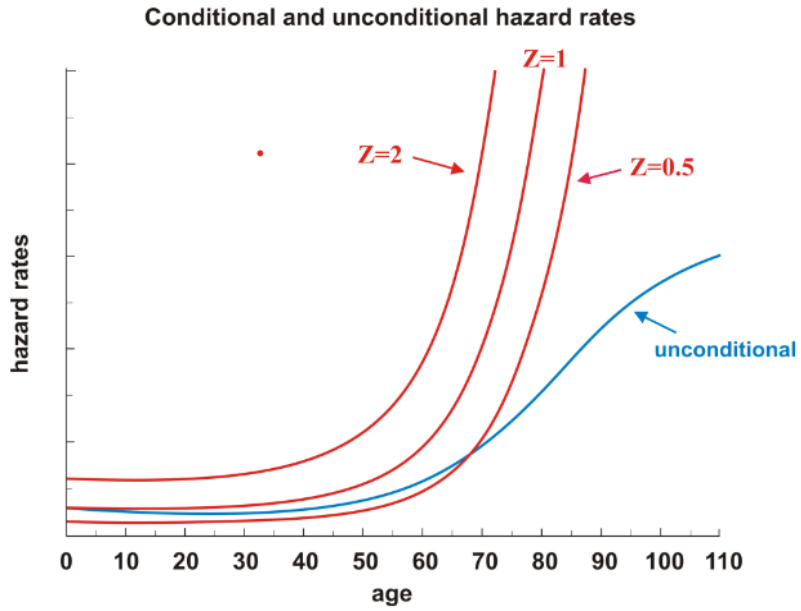


Figure 1: Conditional and unconditional hazard rates in a simulated data set of human mortality. The red lines denote the conditional (individual) hazard rates for individuals with frailty 0.5, 1 and 2, respectively. The blue line denotes the unconditional (population) hazard rate.

One important problem in the area of frailty models is the choice of the frailty distribution. The frailty distributions most often applied are the gamma distribution (Clayton 1978; Vaupel et al. 1979), the positive stable distribution (Hougaard 1986b), a three-parameter distribution (PVF) (Hougaard 1986a), the compound poisson distribution (Aalen 1988, 1992) and the log-normal distribution (McGilchrist and Aisbett, 1991).

We will further study and see how gamma distribution works as a frailty model with the data used by McGilchrist and Aisbett, 1991 in their work, later on.

Univariate frailty models are widely applied. A few examples which can be consulted for more details are listed here. Aalen and Tretli (1999) applied the compound Poisson distribution to testicular cancer data already introduced above. The idea of the model is that a subgroup of men is particularly susceptible to testicular cancer, which results in selection over time. Another example is the malignant melanoma data set including records of patients who had radical surgery for malignant melanoma (skin cancer) at the University Hospital of Odense in Denmark. Hougaard (2000) compared the traditional Cox regression model with a gamma frailty and PVF frailty model, respectively, to these data.

The third example deals with the time from insertion of a catheter into dialysis patients until it has to be removed due to infection. A subset of the complete data, including the first two infection times of 38 patients, was published by McGilchrist and Aisbett (1991). To account for heterogeneity within the data, Hougaard (2000) used a univariate gamma frailty model. This is what we are going to do further study in.

Henderson and Oman (1999) tried to quantify the bias which may occur in estimated covariate effects, and fitted marginal distributions when frailty effects are present in survival data but the latter are ignored in a misspecified proportional hazards analysis.

Congdon (1995) investigated the influence of different frailty distributions (gamma, inverse Gaussian, stable, binary) on total and cause-specific mortality from the London area (1988-1990).

Multivariate frailty models

A second important application of frailty models is in the field of multivariate survival data. Such kind of data occurs for example if lifetimes (or times of onset of a disease) of relatives (twins, parent-child) or recurrent events like infections in the same individual are considered. In such cases independence between the clustered survival times can not be assumed. Multivariate models are able to account for the presence of dependence between these event times. A commonly used and very general approach is to specify independence among observed data items conditional on a set of unobserved or latent variables (Hougaard, 2000). The dependence structure in the multivariate model arises from a latent variable in the conditional models for multiple observed survival times, for example let $S(t_1|Z,X_1)$ and $S(t_2|Z,X_2)$ be the conditional survival functions of two related individuals with different vectors of observed covariates X_1 and X_2, respectively, (see (2)). Averaging over an assumed distribution for the latent variables (e.g., using a gamma, log-normal, stable distribution) then induces a multivariate model for the observed data. In the case of paired observations, the two-dimensional survival function is of the form,

$$S(t_1, t_2) = \int_0^\infty S(t_1|z, X_1)S(t_2|z, X_2)g(z)dz$$

where g denotes the density of the frailty Z. In the case of twins, $S(t_1,t_2)$ denotes the fraction of twin pairs with twin 1 surviving t_1 and twin 2 surviving t_2.

Frailty models for multivariate survival data are derived under conditional

independence assumption by specifying latent variables that act multiplicatively on the baseline hazard.

The shared frailty model

The shared frailty model is relevant to event times of related individuals, similar organs and repeated measurements. Individuals in a cluster are assumed to share the same frailty ***Z***, which is why this model is called the shared frailty model. It was introduced by **Clayton (1978)** and extensively studied in **Hougaard (2000)**. The survival times are assumed to be conditional independent with respect to the shared (common) frailty. For ease of presentation we will consider the case of groups with pairs of individuals (bivariate failure times, e.g. event times of twins or parent - child). Extensions to multivariate data are straightforward. Conditional on the frailty ***Z***, the hazard function of an individual in a pair is of the form $Z\mu_0(t)exp(\beta^T X)$, where the value of ***Z*** is common to both individuals in the pair, and thus is the cause for dependence between survival times within pairs. Independence of the survival times within a pair corresponds to a degenerate frailty distribution $(Z=1, \sigma^2=0)$. In all other cases with $\sigma^2>0$ the dependence is positive by construction of the model. Conditional on Z, the bivariate survival function is given as,

$$S(t_1, t_2|\mathbf{Z}) = S_1(t_1)^z S_2(t_2)^z$$

In most applications it is assumed that the frailty distribution (i.e. the distribution of the random variable ***Z***) is a gamma distribution with mean 1 and variance σ^2. Averaging the conditional survival function produces under this assumption survival functions of the form

$$S(t_1, t_2) = (S_1(t_1)^{-\sigma^2} S_2(t_2)^{-\sigma^2} - 1)^{\frac{1}{\sigma^2}}$$

Shared frailty explains correlation between subjects within clusters. However, it does have some limitations. Firstly, it forces the unobserved factors to be the same within the cluster, which may not always reflect reality. For example, at times it may be inappropriate to assume that all partners in a cluster share all their unobserved risk factors. Secondly, the dependence between survival times within the cluster is based on marginal distributions of survival times. However, when covariates are present in a proportional hazards model with gamma distributed frailty the dependence parameter

and the population heterogeneity are confounded (Clayton and Cuzick, 1985). This implies that the joint distribution can be identified from the marginal distributions (Hougaard, 1986a). Thirdly, in most cases, a one-dimensional frailty can only induce positive association within the cluster. However, there are some situations in which the survival times for subjects within the same cluster are negatively associated. For example, in the Stanford Heart Transplantation Study, generally the longer an individual must wait for an available heart, the shorter he or she is likely to survive after the transplantation. Therefore, the waiting time and the survival time afterwards may be negatively associated.

To avoid the above mentioned limitations of shared frailty models **correlated frailty models** were developed.

The correlated frailty model

Originally, correlated frailty models were developed for the analysis of bivariate failure time data, in which two associated random variables are used to characterize the frailty effect for each pair. For example, one random variable is assigned for partner 1 and one for partner 2 so that they would no longer be constrained to have a common frailty. These two variables are associated and have a joint distribution. Knowing one of them does not necessarily imply knowing the other. There is no more a restriction on the type of correlation. These two variables can also be negatively associated, which would induce a negative association between survival times. Assuming **gamma distributed frailties**, Yashin and Iachine (1995) used the correlated gamma- frailty model resulting in a bivariate survival distribution of the form

$$S(t_1, t_2) = \frac{S_1(t_1)^{1-\rho} S_2(t_2)^{1-\rho}}{(S_1(t_1)^{-\sigma^2} + S_2(t_2)^{-\sigma^2} - 1)^{\frac{\rho}{\sigma^2}}}$$

Examples of the use of multivariate frailty models are various and emphasize the importance of this family of statistical models for survival data.

1. a shared log-normal frailty model for the catheter infection data mentioned above used by McGilchrist and Aisbett (1991),
2. a shared frailty model with gamma and log-normal distributed frailty, applied to the recurrence of breast cancer by dos Santos et al. (1995),
3. a shared positive stable frailty model, applied by Manatunga and Oakes (1999) to

the data from the Diabetic Retinopathy Study, which examined the effectiveness of laser photo- coagulation in delaying the onset of blindness in patients with diabetic retinopathy. Positive stable frailty allows for proportional hazards both in the marginal and the conditional model.

4. a study of Andersen et al. (1999), who tested for center effects in multi-centre survival studies by means of a frailty model with unspecified frailty distribution,
5. a correlated gamma-frailty model, applied by Pickles et al. (1994) to age of onset of puberty and antisocial behavior in British twins,
6. a correlated gamma-frailty model by Yashin and Iachine (1995) and Yashin et al. (1995) to analyze mortality in Danish twins
7. a correlated gamma-frailty model by Wienke et al. (2001) and Zdravkovic et al. (2002) to analyze genetic factors involved in mortality due to coronary heart disease in Danish and Swedish twins, respectively,
8. an extension of the correlated gamma-frailty model by Wienke et. al (2002a) used to model death due to coronary heart disease in Danish twins,
9. different versions of the correlated gamma-frailty model applied by Zahl (1997) to cause- specific cancer mortality in Norway to model the excess hazard.

Tools used

We are going to use the R programming language to analyze the data, model fitting and estimation.

DATA

Each observation corresponds to a kidney, the variable **Patient** is the patient's code. The time from insertion of the catheter to infection or censoring is stored in **Time** while **Status** is 1 when infection has occurred and 0 for censored observations. They may also have been removed for reasons other than infection as well.

Three covariates are present: **Age**, **Sex** (1 for male and 2 for female) and **Disease** types. The **Frailty** is the prediction from the original paper which fits a semi-parametric lognormal frailty model.

We do make some necessary changes later in the data as per our needs.

Patient	Time	Status	Age	Sex	Disease	Frailty
1	8	1	28	1	3	2.3
1	16	1	28	1	3	2.3
2	23	1	48	2	0	1.9
2	13	0	48	2	0	1.9
3	22	1	32	1	3	1.2
3	28	1	32	1	3	1.2
4	447	1	31	2	3	0.5
4	318	1	32	2	3	0.5
5	30	1	10	1	3	1.5
5	12	1	10	1	3	1.5
6	24	1	16	2	3	1.1

Model Building and Analysis

We use the parfm library. The parfm package provides a wide range of frailty models in R.

```
library("parfm")
```

```
## Loading required package: survival
```

```
## Loading required package: optimx
```

We input the data from the directory and check whether the data is showing properly or not.

We further make some necessary changes to the data. The description of the data is provided previously under the "Data" section.

```
##   Patient Time Status Age Sex Disease Frailty
## 1       1    8      1  28   1       3     2.3
## 2       1   16      1  28   1       3     2.3
## 3       2   23      1  48   2       0     1.9
## 4       2   13      0  48   2       0     1.9
## 5       3   22      1  32   1       3     1.2
## 6       3   28      1  32   1       3     1.2
```

We model the hazard of infection as a function of the patient's age, sex and disease. Trivially, kidneys from the same patient cannot be considered independent. Hence, we use a shared frailty model with cluster size of 2 corresponding to patients.

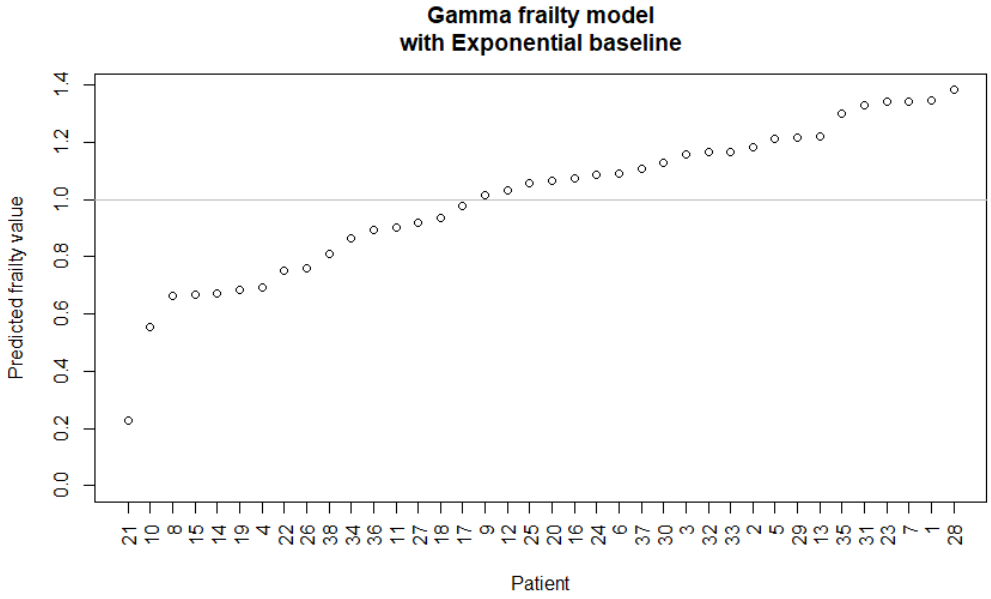
Model Estimation: We fit a model with exponential baseline hazard and gamma frailty distribution.

```
## Frailty distribution: gamma
## Baseline hazard distribution: Exponential
## Loglikelihood: -332.868
##
##         ESTIMATE SE    p-val
## theta    0.289   0.156
## lambda   0.037   0.027
## Sex     -1.461   0.398 <.001 ***
## Age      0.000   0.012 0.971
## Disease -0.135   0.153 0.378
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Kendall's Tau: 0.126
```

According to this model, we can see that **sex** has a significant effect on the hazard of infection while **age** and *desease occured** does not. Conditional to the patient's frailty and the age, the hazard of infection for a female at any time t is estimated to be exp(-1.461) $\approx$ 0.232 times that of a male, with Wald Confidence Interval as:

```
##   low    up
## 0.106 0.506
```

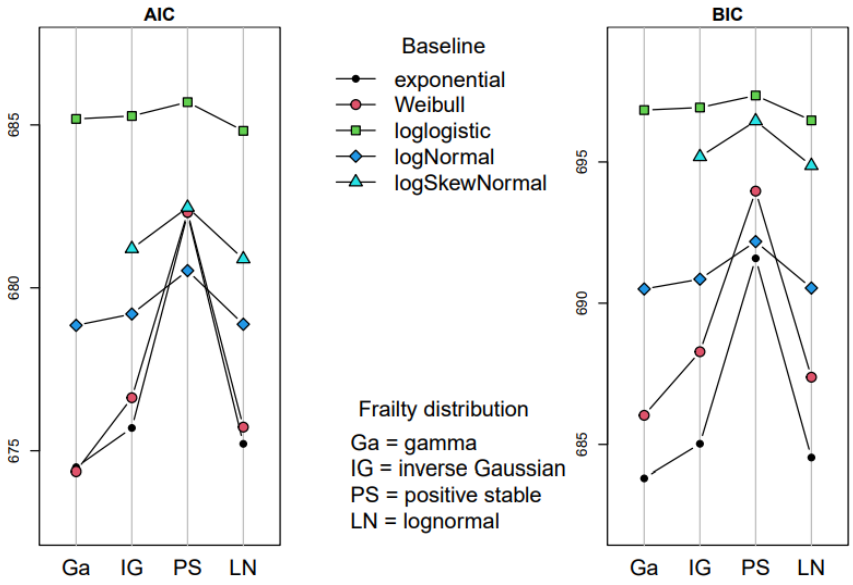
We can obtain the prediction of frailties and further plot them in a graph.



Comparing different models: It might be useful to obtain AIC and BIC values for a series of candidate models for baseline function and the results can further be plotted.

```
##
##
## ### - Parametric frailty models - ###
## Progress status:
##   'ok' = converged
##   'nc' = not converged
##
##                Frailty
## Baseline            gamma   invGau  posSta  lognor
## exponential.........ok......ok......ok......ok....
## Weibull.............ok......ok......ok......ok....
## loglogistic.........ok......ok......ok......ok....
## lognormal...........ok......ok......ok......ok....
## logskewnormal.......nc......ok......ok......ok....
```

```
##  AIC:              gamma  ingau possta lognor
##    exponential       674    676    682    675
##    weibull           674    677    682    676
##    loglogistic       685    685    686    685
##    lognormal         679    679    681    679
##    logskewnormal     ----   681    682    681
##
##  BIC:              gamma  ingau possta lognor
##    exponential       684    685    692    685
##    weibull           686    688    694    687
##    loglogistic       697    697    697    696
##    lognormal         691    691    692    691
##    logskewnormal     ----   695    696    695
```



It seems like the exponential baseline function is a good candidate.

We can compare the model with Inverse Gaussian frailties.

```
##
## Frailty distribution: inverse Gaussian
## Baseline hazard distribution: Exponential
## Loglikelihood: -333.368
##
##          ESTIMATE SE    p-val
## theta     0.350   0.249
## lambda    0.035   0.025
## Sex      -1.296   0.370 <.001 ***
```

```
## Age     -0.001   0.012 0.955
## Disease -0.151   0.152 0.318
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Kendall's Tau: 0.119
```

So, we see that the conclusions drawn in this case are somewhat similar as this model also hints at **sex** being the only significant effect.

```
## Frailty distribution: positive stable
## Baseline hazard distribution: Exponential
## Loglikelihood: -335.47
##
##          ESTIMATE SE    p-val
## nu        0.096   0.085
## lambda    0.023   0.016
## Sex      -0.954   0.336 0.005 **
## Age      -0.001   0.011 0.913
## Disease -0.169   0.136 0.214
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Kendall's Tau: 0.096
```

No difference in this case as well.

Lastly, we compare with a semi-parametric model. As an example, we fit a semi-parametric model with gamma frailties via the Cox-Proportional Hazard Model.

```
## Call:
## coxph(formula = Surv(Time, Status) ~ Sex + Age + Disease + frailty(Patient,
##     distribution = "gamma", eps = 1e-11), data = df, outer.max = 15)
##
##                               coef se(coef)      se2    Chisq   DF       p
## Sex                        -1.56576  0.45827  0.35051 11.67352  1.0 0.00063
## Age                         0.00055  0.01293  0.00942  0.00181  1.0 0.96609
## Disease                    -0.14711  0.16967  0.12523  0.75184  1.0 0.38589
## frailty(Patient, distribu                             21.73953 12.2 0.04461
##
## Iterations: 11 outer, 52 Newton-Raphson
##      Variance of random effect= 0.398   I-likelihood = -181.3
## Degrees of freedom for terms=  0.6  0.5  0.5 12.2
## Likelihood ratio test=46.7  on 13.9 df, p=2e-05
## n= 76, number of events= 58
```

Estimates of regression parameters are very similar to those of the Exponential-Gamma model, while the frailty variance is significantly different. This can arguably be due to the way in which the baseline hazard is treated in a parametric and a semi-parametric model.

RESULTS

We find that Sex is a significant effect on the infection of kidney. This was seen in various models with different frailty distributions the exponential baseline hazard. We also checked a semi-parametric Cox-Proportional Hazard Model, and found a similar result.

Appendix: R-Code

#Libraries needed

library("parfm")

#Inputting the data

df<-read.csv("data.csv", sep = " ")

head(df)

```
df$Sex<-df$Sex-1

#Model Estimation (Gamma Frailty)

mod <- parfm(Surv(Time, Status) ~ Sex + Age + Disease, cluster="Patient", data=df,
dist="exponential", frailty="gamma")

mod

#Confidence Interval

ci.parfm(mod, level=0.05)["Sex",]

#Frailty Prediction and Plotting

u<-predict(mod)

plot(u,sort = "i")

#Comparison for AIC and BIC

df.parfm <- select.parfm(Surv(Time, Status) ~ Sex + Age + Disease + Time + Status,
cluster="Patient", data=df, dist=c("exponential", "weibull", "loglogistic",
"lognormal", "logskewnormal"), frailty=c("gamma", "ingau", "possta",
"lognormal"))

df.parfm

plot(df.parfm)


#Inverse Gaussian Frailty

parfm(Surv(Time, Status) ~ Sex + Age + Disease, cluster="Patient", data=df,
dist="exponential", frailty="ingau")

#Positive Stable Frailty

parfm(Surv(Time, Status) ~ Sex + Age + Disease + Time, cluster="Patient", data=df,
```

dist="exponential", frailty="possta", iniFpar=0.1)

#Semi-parametric Model

coxph(Surv(Time, Status) ~ Sex + Age + Disease + frailty(Patient, distribution="gamma", eps=1e-11), outer.max=15, data=df)

REFERENCES

1. Frailty Models, Andreas Wienke, MPIDR WORKING PAPER WP 2003-032 SEPTEMBER 2003, Max Planck Institute for Demographic Research
2. Survival Analysis - A Self Learning Text, Third Edition by David G. Kleinbaum and Mitchel Klein
3. Parfm: Parametric Frailty Models in R, Package vignette, V. 1.4 (January 25th, 2017), Marco Munda (Arlenda), Federico Rotolo (Gustave Roussy), Catherine Legrand (Universite´e catholique de Louvain)
4. McGilchrist and Aisbett (1991), Regression with Frailty in Survival Analysis. Biometrics 47, 461-466
5. Yashin, A.I., Vaupel, J.W., Iachine, I.A. (1995) Correlated Individual Frailty: An Advantageous Approach to Survival Analysis of Bivariate data. Mathematical Population Studies 5, 145 - 159
6. Clayton, D.G. (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika 65, 141-151
7. Aurélien Belot, Virginie Rondeau, Laurent Remontet, Roch Giorgi (2014), A joint frailty model to estimate the recurrence process and the disease-specific mortality process without needing the cause of death, John Wiley & Sons, Ltd
8. Roger B. Nelse, An Introduction to Copulas, 2006