

APPLICATION OF SHARED GAMMA AND INVERSE-GAUSSIAN FRAILTY MODELS TO CANCER DATA.

A. A. ABIODUN

(Received 5 October 2007; Revision Accepted 23 November 2007)

ABSTRACT

Shared Gamma and Inverse-Gaussian Frailty models are used to analyze the survival times of patients who are clustered according to cancer/tumor types under Parametric Proportional Hazard framework. The result of the analysis shows that patients within the same cluster share some common unobserved heterogeneity which proves the analysis of such data under the assumption of independence inappropriate. This is an indication that conclusion based only on the observed covariates, ignoring the characteristic unobservable covariates in the data can be misleading. However, no evidence is strong enough for preference of either Gamma or Inverse Gaussian Frailty.

KEY WORDS: Failure time, Shared Frailty, Baseline hazard.

INTRODUCTION

Clustered also known as Multivariate or Correlated failure times are frequently encountered when there is grouping of individuals into clusters. The Cox (1972) Proportional hazards model which is a popular model in assessing covariate effects on failure time, based on the assumption of independence is inappropriate under this setting due to the complication caused by the dependence of related failure times within the clusters. For example, in animal experiments, due to common genetic and environmental factors, the failure times of animals within the same litter may be correlated. Also in multicentre clinical trials, outcomes for participants within the same centre, sharing the same environment are likely to be correlated. In the presence of such correlation, standard analytical approaches such as the Cox proportional model can lead to confidence intervals with incorrect coverage probabilities (Huster, Brookmayer and Self, 1989). Thus there is the need to account for such intracluster dependence. Two classes of model approaches that are often used in analyzing clustered failure time data are the frailty model and the marginal model. In frailty model, association of failure times of subjects within a group (cluster) is usually focused. Marginal model considers the population –average covariate effects and leaves the association among subjects within clusters as a nuisance parameter. This paper focuses on the frailty model. Frailty simply means "Proneness" or "Propensity" to fail or die. By definition, frailty is an unobserved heterogeneity that modifies the hazard rate of an individual. Frailty could be univariate, which is often incorporated into a model on the basis of idea that individuals under study possess some heterogeneous unobserved characteristics (frailties) and that those patients who are frailer will die earlier than the others. This form of frailty was considered by Vaupel et al (1979) in analyzing survival time data. The other form of frailty, the one considered in this study is the "Shared frailty". This form of frailty model was introduced by Clayton (1978). The shared frailty model is relevant to event times of related individuals, similar organs and repeated measurements. Individuals in the same group or cluster are assumed to share some common unobserved heterogeneity (frailty), hence the name shared frailty model (Wienke, 2003). In statistical term, frailties are random effects that act multiplicatively on the baseline hazard, thereby modifying the hazard rate of all individuals in the study population or members of each cluster.

Cancer has been variously defined in the literature as uncontrolled growth of abnormal cell on or within the body. Research on cancer has become a global interest. WHO (2003) reports that cancer accounts for 7.1 million deaths annually, which is about 12% of global total. Report also shows that the number of new cancer cases annually is estimated to rise to 15 million by 2020, with more than half of all cancer cases occurring in developing countries. The challenges of cancer however, lie in their complexities. The many different tumor types each with distinct sites, present different clinical behaviours (Eschenbach and Collins, 2005). The crude percentage of death due to cancer within each cluster can only serve as a rough index if effects of covariates are not taken into account. Even then, analysis done based on observed covariates alone may not be adequate. For example, some types of tumor may have higher mortality rate just because the patients are older, whereas there could be more to that due to some unobservable heterogeneity. This heterogeneity may arise in part from the fact that tumors are complex organ systems that are shaped by cellular biological context. It is therefore imperative in cancer research to note the characteristic heterogeneity due to various types of cancers and tumors as the case may be. This study, therefore aims at analyzing data collected on cancer patients when they are clustered according to cancer or tumor types. The data were collected on 215 cancer patients from University of Ilorin Teaching Hospital between year 1999 and 2002. The failure time is the time between patient's admission and the time when he/she died or censored. Censored patient here refers to patient who dropped out alive between the entry date and time of data collection and patient who is still alive as at the time of data collection. Two covariates, sex and age at the time of admission were also collected on each patient. In section 2, we introduce some basic notations and then specify the models. We also give the parametric specification of the baseline hazards and analyze a dataset on cancer patients. In section 3, we present and discuss the results of the analysis. Some concluding remarks are given in section 4.

A. A. Abiodun, Department of Statistics, University of Ilorin, Ilorin, Nigeria

2. MODELING AND ESTIMATION

2.1 Notations and Model Specification

Suppose that we observe censored time-to-event data from a study with J clusters and n_j subjects from the j th cluster

($j=1, \dots, J$), such that the total sample size is $N = \sum_{j=1}^J n_j$. Let X_{ij} and C_{ij} denote the latent failure time and censoring

time for subject i in cluster j respectively

($j = 1, \dots, J; i = 1, \dots, n_j$)

Then the observed failure time is given by

$$T_{ij} = \min(X_{ij}, C_{ij}). \quad (2.1)$$

$$\text{Let } \delta_{ij} = \begin{cases} 1 & \text{if } T_{ij} = X_{ij} \\ 0 & \text{if } T_{ij} = C_{ij} \end{cases}$$

If we also let Z_{ij} denote the covariate vector for subject i in cluster j , then the observed data structure is $(t_{ij}, \delta_{ij}, Z_{ij})$.

The proportional hazard (PH) model, extending from Cox (1972) is given by

$$\lambda(t_{ij} | Z_{ij}) = \lambda_0(t) \exp(\beta^T Z_{ij}), \quad (2.2)$$

where $\lambda_0(t)$ is the baseline hazard function (for which a functional form may be assumed or may be left unspecified as for Cox PH model) and β is a vector of regression coefficients. The model given in (2.2) is based on the assumption of independent failure times of subjects regardless of the cluster he/she belongs to. However, if the independence of failure times within the clusters is of interest, then $\lambda_0(t)$ in (2.2) has to be modified by incorporating a frailty term w . Thus for the i th subject in the j th cluster with shared frailty w_j , the Proportional hazard model is given as

$$\lambda(t_{ij} | Z_{ij}, w_j) = w_j \lambda_0(t) \exp(\beta^T Z_{ij}), \quad (2.3)$$

where w_j is a frailty term shared among all members of cluster j . Clearly, (w_1, w_2, \dots, w_J) are independently and identically distributed samples from some distributions with positive supports. Common distributions for frailty include the one parameter gamma distribution (Clayton, 1978), the Inverse Gaussian distribution (Hougaard, 1986b), the positive stable distribution (Hougaard, 1986a) and the Lognormal distribution (McGilchrist and Aisbett, 1991). However, for tractability reason, the choice of distribution is limited to those that provide a close form expression for the frailty survivor function. For continuous failure time model, the Gamma and the Inverse-Gaussian distributions have been the two most commonly used. In fact, it is believed by some (Hougaard, 1995) that all models should contain frailties. Hendersen and Oman (1999) showed that when frailty is present but ignored in a Cox model, the regression coefficients are biased towards zero. However, when censoring is present, the bias is reduced. Literatures have shown that the presence of frailty attenuates the proportionate response of the hazard to variation in each regressor and that the estimate of a positive (negative) effect derived from the wrong, no-frailty model, will underestimate (overestimate) the true estimate.

Suppose we consider the modified baseline hazard in (2.3) as

$$\begin{aligned} \lambda_{(ij)}(t) &= \lambda(t | w_j) \\ &= w_j \lambda_0(t) \end{aligned} \quad (2.4)$$

which is essentially a proportional hazard model interpreted as the conditional hazard function given frailty w_j , shared by all members of cluster j . The corresponding conditional survival function is

$$\begin{aligned} S_{(ij)}(t) &= S(t | w_j) \\ &= \exp\left(-\int \lambda_0(s | w_j) ds\right) \\ &= S_0(t)^{w_j} \end{aligned} \quad (2.5)$$

which represents the probability of being alive at time t given the shared frailty w_j .

However, estimation requires expression for survival function and hazard function that are not conditional on the frailty term and this is obtained by integrating out the frailty term

The unconditional survival function is thus given (suppressing the subscript) as the integrated survival with respect to the frailty

$$S(t) = \int_0^{\infty} S_0(t)^w g(w) dw, \quad (2.6)$$

where $g(w)$ is the density function of the frailty variable w
The unconditional hazard function can then be obtained as follows
If we let $\Lambda(t) = -\log S(t)$

$$\begin{aligned} &= -\log \int_0^{\infty} S_0(t)^w g(w) dw \\ &= -\log \int_0^{\infty} S_0(t)^w g(w) dw \end{aligned} \quad (2.7)$$

from (2.5) and (2.6) respectively, then the unconditional hazard function is given as

$$\lambda(t) = \frac{\int_0^{\infty} \frac{d}{dt} S_0(t)^w g(w) dw}{\int_0^{\infty} S_0(t)^w g(w) dw} = \frac{\int_0^{\infty} w \lambda_0(t) S_0(t)^w g(w) dw}{\int_0^{\infty} S_0(t)^w g(w) dw} \quad (2.8)$$

For purposes of identifiability, we assume that the random variable W has the mean and variance given respectively, as

$$E(W) = 1 \text{ and } \text{Var}(W) = \theta \quad (2.9)$$

Thus if we specify Gamma distribution for W with $\text{Gamma}(1/\theta, \theta)$ as

$$g(w) = \frac{w^{1/\theta-1} \exp(-w/\theta)}{\Gamma(1/\theta) \theta^{1/\theta}} \quad (2.10)$$

from (2.8) and (2.10) the modified baseline hazard function with Gamma frailty becomes

$$\lambda_{G,t}(t) = \frac{\lambda_0(t)}{1 + \theta \Lambda_0(t)} \quad (2.11)$$

And if we specify Inverse-Gaussian distribution for w , with Inverse-Gaussian $(1, 1/\theta)$ as

$$g(w) = (2\pi\theta w^3)^{-1/2} \exp\left\{-\frac{1}{2\theta}\left(w - 2 + \frac{1}{w}\right)\right\} \quad (2.12)$$

then the modified baseline hazard function with Inverse-Gaussian frailty is given by

$$\lambda_{IG,t}(t) = \frac{\lambda_0(t)}{(1 + 2\theta \Lambda_0(t))^{1/2}} \quad (2.13)$$

where

$$\Lambda_0(t) = -\log S_0(t)$$

2.2. Parametric specification of baseline hazard ($\lambda_0(t)$)

The parametric Proportional hazard model requires that functional form of the baseline hazard $\lambda_0(t)$ be specified. Weibull and Gompertz distributions have been two popular parametric distributions for increasing baseline hazards, which is characteristic of cancer data. Increasing hazard occurs in cases when patients have low positive response to treatment. (Lee, 1980).

The baseline density function for Weibull with parameter γ and p is given by

$$f_0(t) = \gamma p (\gamma t)^{p-1} \exp(-(\gamma t)^p) \quad (2.14)$$

with the baseline hazard

$$\lambda_0(t) = \gamma p (\gamma t)^{p-1} \quad (2.15)$$

where γ and p are the scale and shape parameters respectively.

The baseline density function for Gompertz distribution with parameter γ and p is given by

$$f_0(t) = \exp\left[(p + \gamma t) - \frac{1}{\gamma} (e^{p + \gamma t} - e^p) \right] \quad (2.16)$$

with the corresponding baseline hazard function given by

$$\lambda_0(t) = \exp(p + \gamma t) \quad (2.17)$$

2.3. Data Analysis

We analyze the cancer data described in section 1 using the methods described above. Two covariates are used in the analysis. These are age on admission and sex which is coded 1 for male and 0 for female patients. Patients are grouped into nine clusters according to cancer/tumor types/sites, which include carcinoma, leukemia, lymphoma, melanoma, sarcoma, rectum, lung, liver and stomach. Prostate and breast cancers are not included because they are gender related and may possibly introduce gender bias into the analysis.

The major interest of the study is to examine how frailty models compare with the model without frailty (reference model), and also investigate whether there are differences between Gamma and Inverse Gaussian frailty models. We thus carry out our parameter estimation under five models stated below.

Model I: The reference model of no frailty and Weibull baseline hazard.

Model II: The reference model of no frailty and Gompertz baseline hazard.

Model III: The model with Gamma frailty and Weibull baseline hazard.

Model IV: The model with Gamma frailty and Gompertz baseline hazard.

Model V: The model with Inverse-Gaussian frailty and Weibull baseline hazard.

Model VI: The model with Inverse-Gaussian frailty and Gompertz baseline hazard.

All analyses have been done with STATA, Version 8.0.

3. RESULTS

Table 1 shows the effect of each covariate upon the hazard and their associated standard errors under Weibull and Gompertz baseline specifications for the reference model (no frailty) and with shared Gamma and Inverse-Gaussian frailties. Both the reference model and those with frailties adequately fit the data as seen from the likelihood ratio tests.

The frailty is seen to have an increasing effect on the parameter values, in accordance with the submission of Hendersen and Oman (1999). There is no significant difference in the hazard rates between male and female patients (though slightly higher for male) in all the models. Effect of age is found to be moderately significant in all the models.

The θ value is the estimate of frailty distribution variance which is a measure of degree of heterogeneity among the clusters and of course the degree of association within the cluster. Clearly, there is evidence of heterogeneity among the clusters indicated by the significant difference of θ from zero under the two frailty specifications. This means that the cancer patients within the same cluster are correlated.

Table 1: Proportional Hazards Models under Weibull and Gompertz baselines with No Frailty and with Shared Gamma and Inverse-Gaussian Frailties

WEIBULL BASELINE									
Covariates	No Frailty			Gamma Frailty			Inverse Gaussian Frailty		
	Coef.	Std. Err.	P-Value	Coef.	Std. Err.	P-Value	Coef.	Std. Err.	P-Value
Sex	0.0656	0.131	0.160	0.0667	0.134	0.180	0.0668	0.134	0.181
Age	0.0343	0.003	0.039	0.0554	0.004	0.026	0.0540	0.005	0.033
LR Test	14.64 (0.0019)			18.07(0.0012)			17.54(0.0015)		
Test for $\theta = 0$:				3.75(0.0001)			20.69(0.0001)		
GOMPERTZ BASELINE									
Covariates	No Frailty			Gamma Frailty			Inverse-Gaussian Frailty		
	Coef.	Std. Err.	P-Value	Coef.	Std. Err.	P-Value	Coef.	Std. Err.	P-Value
Sex	0.0654	0.121	0.183	0.0657	0.134	0.225	0.0658	0.134	0.224
Age	0.0341	0.003	0.048	0.0536	0.004	0.037	0.0533	0.004	0.041
LR Test	14.59 (0.0025)			17.82(0.0019)			16.20(0.0021)		
Test for $\theta=0$:				22.08(0.0001)			18.95(0.0001)		

Note: Values in the parentheses are the P-values of the tests

For model with Gompertz baseline, the results are similar to those of Weibull baseline. The reference model, the Gamma and Inverse-Gaussian frailty models all fit the data adequately well. The frailties are also statistically significant and the parameters of the models change in the expected direction. As for the choice between Gamma and Inverse-Gaussian frailty specifications, literatures have shown that the choice is complicated by the fact that the two models are non-nested.

Simulation studies

To compare Gamma and Inverse Gaussian frailties under Weibull and Gompertz models, simulation studies are conducted. Data generations are done as follows. In each sample there are $J=10$ clusters with equal cluster size $n_j=10$ so that $n=100$ observations are generated in each sample. The survival times t_{ij} , $i=1, \dots, n_j$, $j=1, \dots, J$, are generated from model $\lambda(t_{ij}|z_{ij}, w_j) = w_j \lambda_{ij}(t) \exp(\beta' z_{ij})$ where $w_j \sim \text{Gamma}(1/3, 3)$. The baseline distribution is taken to be Weibull with scale and shape parameters 1 and 2 respectively, z_{ij} is Bernoulli with success probability 0.5 and $\beta = 0.5$. $N=100$ samples are generated and analyzed using models I to VI. The results are given in table 2. It can

be seen that parameter estimates are far from the true values when analyses are done using Weibull or Gompertz baselines without frailty. When frailty is incorporated, no remarkable differences are seen between gamma frailty and inverse Gaussian models when analyses are done with either Weibull and Gompertz baselines.

Table 2. Results of simulations to compare Gamma and Inverse Gaussian Frailty Models under Weibull and Gompertz baselines

Model	Weibull Baseline	Gompertz Baseline
	Coeff(std error)	Coeff(std error)
No Frailty	0.449(0.434)	0.424(0.433)
Gamma Frailty	0.501(0.458)	0.499(0.455)
Inverse Gaussian Frailty	0.503(0.459)	0.497(0.455)

It is seen that parameter estimates for Gamma and Inverse Gaussian frailty models are 0.501(0.458) and 0.503(0.459) respectively under Weibull baseline, while for Gompertz baseline, they are 0.499(0.455) and 0.497(0.454) respectively.

4. DISCUSSION AND CONCLUSION

We have demonstrated how frailty model can be used under Parametric proportional hazard framework. Whereas the model that ignores the unobserved heterogeneity in the cancer data appears to fit the data well, simulation results clearly reveal a clear difference between the situation when frailty is incorporated and when it is not. This shows that conclusion based only on the observed covariates, ignoring the characteristic unobservable covariates in the data can be misleading. A possible extension of this paper is to examine how these models behave under correlated time-varying covariates.

REFERENCES

- Clayton, D., 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.
- Cox, D.R., 1972. Regression Models and Life Tables (with discussions). *Journal of the Royal Statistical Society, B*, 34, 187-220.
- Eschenbach, A. and Collins, F., 2005. Toward a Comprehensive Genomic Analysis of Cancer. Workshop Executive Summary. The Ritz Carlton Washington, DC.
- Henderson, R. and Oman, P., 1999. Effect of Frailty on Marginal Regression Estimates in Survival Analysis. *Journal of the Royal Statistical Society, B*, 61(2), 367-379.
- Hougaard, P., 1986a. Survival Models for Heterogeneous Population derived from stable distributions. *Biometrika*, 73, 671-678.
- Hougaard, P., 1986b. A class of Multivariate failure time distributions. *Biometrika*, 73, 671-78.
- Hougaard, P., 1995. Frailty models for survival data. *Lifetime Data Analysis*, 1, 255-273.
- Huster, W. J., Brookmeyer, R. and Self, S. G., 1989. Modeling paired survival data with covariates. *Biometrics* 45, 145-156.
- Lee, T. L., 1980. *Statistical Methods for Survival Data Analysis*. Wadsworth, Inc.
- McGilchrist, C. A., Aisbett, C.W., 1991. Regression with frailty in Survival Analysis. *Biometrics* 47, 461-466.
- Vaupel, J. W., Manton, K.G. and Stallard, E., 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 16, 439-454.
- Wienke, A., 2003. Frailty Models. Max Plank Institute of Demographic Research, Working Paper 2003-032 (www.demogr.mpg.de).
- World Health Organization Reports (2003).