World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:7, No:4, 2013

Dynamic Models versus Frailty Models for Recurrent Event Data

Entisar A. Elgmati

***Abstract*—**Recurrent event data is a special type of multivariate survival data. Dynamic and frailty models are one of the approaches that dealt with this kind of data. A comparison between these two models is studied using the empirical standard deviation of the standardized martingale residual processes as a way of assessing the fit of the two models based on the Aalen additive regression model. Here we found both approaches took heterogeneity into account and produce residual standard deviations close to each other both in the simulation study and in the real data set.

***Keywords*—**Dynamic, frailty, misspecification, recurrent events.

I. Introduction

THERE are several approaches that have been proposed in the literature to deal with heterogeneity in recurrent event data. One of the approaches is to use information on previous events for an individual as time-dependent covariates explaining the occurrence of future events. This is called a dynamic approach. Dynamic models are models that include time dependent covariates representing individual-specific histories not known at the outset of the study [1], [2], [3]. One way which uses the Cox model with dynamic covariates, is given in [4]. An alternative with which we will be concerned is based on the additive regression model [5], [6]. The other approach is to assume random effects or frailty is present in the data. A frailty is an unobserved random factor that provides a convenient way to describe unexplained heterogeneity between individuals or the influence of unobserved risk factors in the model. Usually we assume frailty acts multiplicatively on the intensity function. People with a large value of the frailty will experience more events and people with small frailty values will experience fewer events, in comparison with other people with the same observed covariates. The reason why we consider this approach is its connection with the dynamic modelling approach.

The term frailty was first introduced in this context by Vaupel [7] for univariate survival models to account for unobserved heterogeneity or missing covariates in the study population. Vaupel [8] and Aalen [9] discussed such heterogeneity and selection effects in more details. For full details see the book by Hougaard [10].

An important issue in the frailty model area is the choice of the frailty distribution. The most common one is the gamma distribution [11], [7], [10] because of the nice mathematical properties of the gamma family. Other distributions that can be chosen for the frailty are the positive stable distribution [12], the three parameter power variance distribution [13], the lognormal distribution [14] and the compound Poisson distribution [15], [16]. We will use gamma only.

There are two classes of frailty models. The first class is the univariate frailty model for univariate survival times. The second one is multivariate frailty models that describe multivariate survival times (e.g. multiple events or repeated events for the same individual). Recurrent event time data of interest in this work provide a special case of multivariate survival data. For this kind of data there are also two frailty models. The first considers the random effect across individuals and constant over time and it has been studied extensively in [10]. The aim of the second one is to model correlated event times [17], [18].

In this paper, a brief background to frailty modeling will be given and its connection to the dynamic model will be presented. This relationship will be examined through a simulation study and an application (Blue Bay data).

II. Frailty Model

The additive hazard frailty model assumes that, for a given vector of observed covariates $x_i(t)$ and unobserved frailty variable Z_i, which is considered as a random variable over the population of individuals, the counting process $N_i(t)$ for an individual i at time t has intensity

$$\lambda_i(t \mid F_{t^-}, x_i(t), Z_i) = Z_i Y_i(t)\alpha(t \mid x_i(t)), \qquad (1)$$

where F_{t^-} is the history or the filtration, $Y_i(t)$ is an at risk indicator with value one when individual at risk and zero otherwise and

$$\alpha(t \mid x_i(t)) = \beta_0(t) + \beta_1(t)x_{i1}(t) + \ldots + \beta_p(t)x_{ip}(t)$$

with $\beta_0(t)$ being the hazard baseline and $\beta_j(t),\ j = 1,2,\ldots,p$ the covariate coefficients.

The model is completed by assuming a parametric distribution for Z_i and working with the marginal intensity

$$\lambda_i(t \mid F_{t^-}, x_i(t)) = E[Z_i \mid F_{t^-}]Y_i(t)\alpha(t \mid x_i(t)), \qquad (2)$$

Entisar A. Elgmati is with the Depatment of Statistics, Faculty of Science, Tripoli University, Tripoli, Libya (e-mail: eelgmati@ hotmail.com).

When $Z_i's$ are gamma distributed with unit mean and variance ξ, then the conditional expectation of Z_i for individual i given the observed history for this individual is given by ([19] and [20]),

$$E[Z_i \mid F_{t^-}] = \frac{1+\xi N_i(t^-)}{1+\xi \Lambda_i(t)} \tag{3}$$

where $\Lambda_i(t) = \int_0^t Y_i(s)\alpha(s \mid x_i(s))ds$.

Estimation in this model can be based on an iterative approach, using the EM-algorithm. In the E-step Z_i is replaced by its conditional expectation ($\hat{Z}_i$) given the history. The M-step is then to calculate the estimates of the regression coefficients as if Z has been observed ($Z_i = \hat{Z}_i = E[Z_i \mid F_{t^-}]$) and by maximising the marginal likelihood of total event counts one can get an estimate of ξ [21]. This model is often called the shared frailty model, meaning that the same frailty is shared by all the event times pertaining to one individual.

III. Connection between Frailty and Dynamic Model

In this section we will look at the relationship between the dynamic model and the frailty model. Assume that the intensity function for an individual i with a frailty Z_i and the observable intensity process for this individual are as defined in (1) and (2). For ease of description we will assume a simple single time constant covariate ($\alpha(t) = \beta_0(t) + \beta_1(t)x_1$). And since $E[Z_i \mid F_{t^-}]$ is given by equ. (3) we have

$$\lambda_i(t \mid F_{t^-}) = \alpha(t)\frac{1+\xi N_i(t^-)}{1+\xi \Lambda_i(t)} \tag{4}$$

which can be simplified into

$$\lambda_i(t) = \beta_0^*(t) + \beta_1^*(t)x_1 + \beta_2(t)N_i(t^-) + \beta_3(t)x_1 N_i(t^-)$$

Thus the previous number of events comes into the model as an additive term, as when a dynamic covariate has been included. Therefore the dynamic model may alternatively be used instead of frailty model to explain the heterogeneity in the data. In the following we describe a simulation study to get a better insight into these situations.

A. Simulation Study

This simulation study has been conducted to investigate the behavior of the standard deviation of the standardized residual processes as a diagnostic tool for assessing the model fit, when frailty is present in the true data generating model.

As mentioned in [6] and [3], a model fit within the Aalen class of models can be summarized by the empirical standard deviation of the standardized residual processes. If the model is correctly specified then these should be close to one at all time points. We simulated from a frailty model with a sample of size 250 and $\tau = 100$ time points. Two time constant binary covariates (i.e. x_1 and x_2), and a frailty variable with gamma distribution with mean one and variance ξ were included. We used a variety of different values of ξ but the simulation study that is presented here is with $\xi = 1$ so $Z_i \sim G(1,1)$. Assume,

$$Model\ A: \lambda_i(t) = Z_i(0.1 + 0.05x_1 + 0.05x_2)$$

The frailty Z_i is fixed over time and each individual will have the same frailty over the study period. Once the data set was generated, the two following models (fixed and dynamic models) were fitted to these data assuming no knowledge of the right model,

$$Model\ A1: \lambda(t) = \beta_0(t) + \beta_1(t)x_1(t) + \beta_2(t)x_2(t)$$
$$Model\ A2: \lambda(t) = \beta_0(t) + \beta_1(t)x_1(t) + \beta_2(t)x_2(t) + \beta_3(t)D(t^-)$$

where $D(t^-)$ represents the average number of previous events in the process, i.e. $N(t^-)/t$. In addition to above simulation, we also simulated from

$$Model\ B1: \lambda(t) = 0.1 + 0.05x_1 + 0.05x_2$$
$$Model\ B2: \lambda(t) = 0.1 + 0.05x_1 + 0.2x_2$$

and fitted each of the following models to the two generated data sets

$$Model\ C1: \lambda(t) = \beta_0(t) + \beta_1(t)x_1(t)$$
$$Model\ C2: \lambda(t) = \beta_0(t) + \beta_1(t)x_1(t) + \beta_2(t)R(t)$$

where $R(t)$ is the residual from regressing $D(t^-)$ on the fixed covariate to keep the estimates of the fixed covariates the same.

Fig. 1 shows the mean of the standard deviation of the standardized residual processes for 100 simulations. The upper panel shows the standard deviation for fitting Model A1 and Model A2 to the frailty data (Model A). In the middle and lower panel the standard deviation of the standardized residual processes for fitting Model C1 and C2 to Model B1 (middle panel) and to Model B2 (lower panel) are shown.

Notice that, when a fixed model (Model A1) is fitted to frailty data (Model A) or when a model with less covariates (Model C1) was fitted to a data generated with more covariates (Model B1 and B2), the standard deviation of the standardized

residuals is above one indicating that the fitted model is not adequate for these data (Fig. 1, left column). However, the standard deviation for fitting the dynamic model C2 to Model B1 and B2 is almost one in all time points (right column middle and lower plot in Fig. 1). In addition fitting dynamic model A2 to frailty data, Model A, results in a standard deviation that is close to one. This supports the earlier discussion that a dynamic model may alternatively be used instead of frailty model. Furthermore, dynamic models can also capture heterogeneity induced by a missing covariate as we see in Fig. 1, right column. Thus dynamic models may well be richer than frailty models.

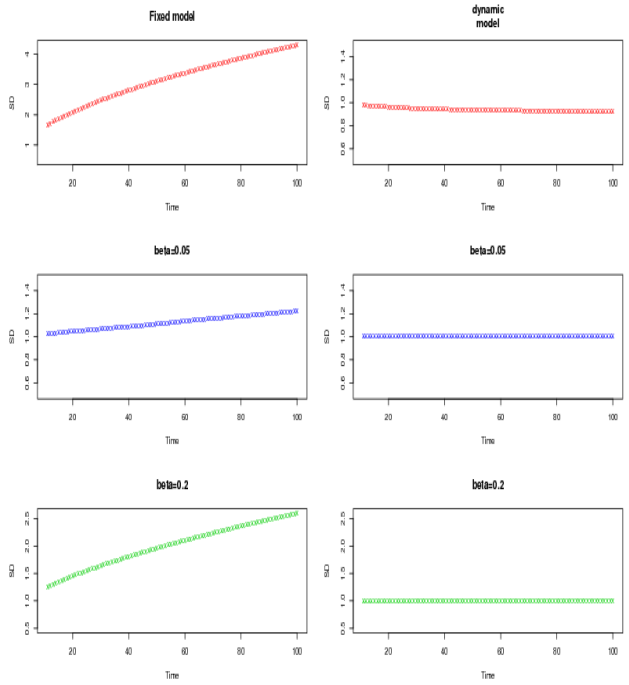

Fig. 1 The empirical standard deviations of standardized residual processes for; upper panel: simulate from Model A and fit Model A1 (left plot), fitting Model A2 (right plot). Middle panel: simulate from model B1 and fit C1 (left), fit model C2 (right) and lower panel: simulate from B2 and fit C1 (left), fit model C2 (right)

To see what is happening when the misspecified models were fitted, we did some calculations to see the expected standard deviation values for these models. Fig. 2 shows the average of 100 simulations of the empirical (black lines) and the expected standard deviation (red lines) for fitting a fixed model to frailty data (right plot), and fitting a model with one covariate to data generated with two covariates (left plot). We used the previously defined models (simulate from Model A and fit Model A1 and simulate from Model B2 and fit Model C1) but with sample size of $n = 500$. Looking at the figure one can see that the lines (empirical and expected) are very close to each other and linearly increasing in t for both fitted models indicating that these models did not fit the data well.

A third form of misspecification is when we fit a dynamic model to data generated with frailty (fit Model A2 to data from Model A). Unfortunately the algebra required to derive the theoretical properties is intractable because the exact times of previous events affect regression coefficients. We did work through the algebra for event times two and three, where $N(t^-)$ has only a small number of possibilities, and found that the empirical standard deviation of the standardized residual closely agreed with the theoretical one. For later event times we turned to simulation with the very large sample size of $n = 100{,}000$ to approximate the expected patterns. Fig. 3 and 4 show the patterns at $\xi = 1$ and $\xi = 0.1$ respectively. From the figures one can notice that the standard deviations fall initially especially with $\xi = 1$ and then climb back up. Although the values are close to one, and the deviations may not be noticeable in small samples, nonetheless, in theory the empirical standard deviations do not have expected value one. Being close to one indicate that the model fit the data and supports the previous discussion, Section III.

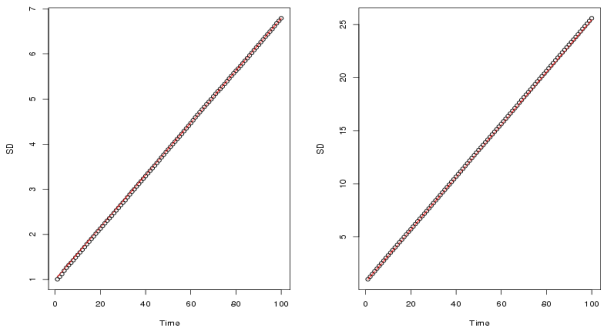

Fig. 2 The empirical standard deviations of standardized residual processes (black lines) and the large sample expected value (red lines) for; right plot: fitting A1 to A. Left plot: simulate from B2 and fit C1

Furthermore note that when the missing covariate has little effect (i.e. left column middle plot of Fig. 1) then the standard deviation will not be far from one, meaning, the fitted model still can explain the patterns in the data. Also including the interaction term as defined in (5), whether it is the interaction between the dynamic covariate and the fixed covariates or between the residuals from regressing the dynamic covariate on the fixed covariates, had no significant effect on the standard deviation of the standardized residual plots. Both give good fits to the data (plots not shown here).

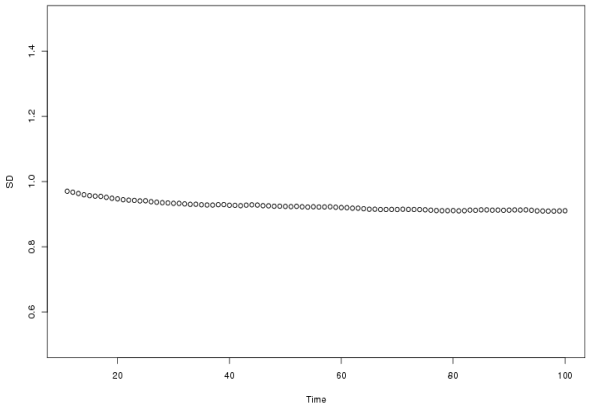

Fig. 3 The empirical standard deviations of standardized residual processes for fitting dynamic model to frailty data for $n = 100000$ and $\xi = 1$

We repeated the above simulation study for different sample sizes and different parameter values and the same results were obtained.

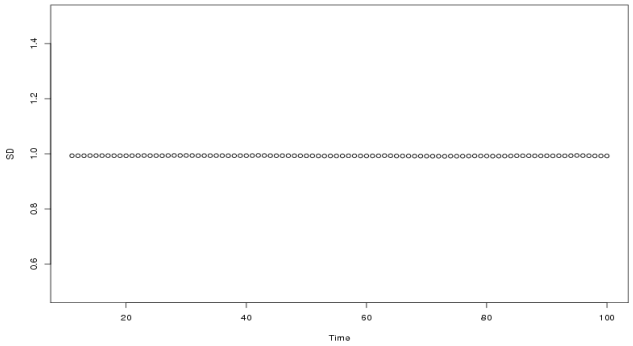

Fig. 4 The empirical standard deviations of standardized residual processes for fitting dynamic model to frailty data for $n = 100000$ and $\xi = 0.1$

IV. Diarrhoea Data: Blue Bay Data

As an illustration, we will study the recurrent incidence of infant diarrhoea data [3]. The empirical standard deviations of the standardized martingale residual processes were calculated for three models (Fig. 5): the Aalen model without heterogeneity, fixed covariates only (solid line), the Aalen model with dynamic covariates (dashed line), and the frailty model fitted here (the individual frailty has been estimated using routines provided by colleague Mahdi Mohammadi) without the interaction term (the dotted line). The fixed and dynamic covariates for this data set are presented in [3]. We also checked whether the inclusion of the interaction term between the dynamic covariates and fixed covariates or the interaction between the dynamic residuals and fixed covariates have significant effect in the standard deviation of the standardized processes. We found it did not make any difference.

From Fig. 5 one can see clearly that the heterogeneity needs to be taken into account and there is little difference between the fits under the dynamic and frailty approaches. The standard deviations of both models are close to one suggesting that each specified model is reasonable, although the frailty approach may be slightly preferred for these data.

V. Conclusion

In this paper we studied an alternative modelling approach, that accounts for heterogeneity in recurrent event data, to that discussed in our previous papers.

During the earlier papers we concentrated on the dynamic modeling approach. This paper introduced frailty modelling, which can be considered as another approach to deal with this kind of data. Frailty modeling itself is not a focus of this paper. Instead we were interested in methods to detect omitted frailty and in comparing the frailty and dynamic fits to data.

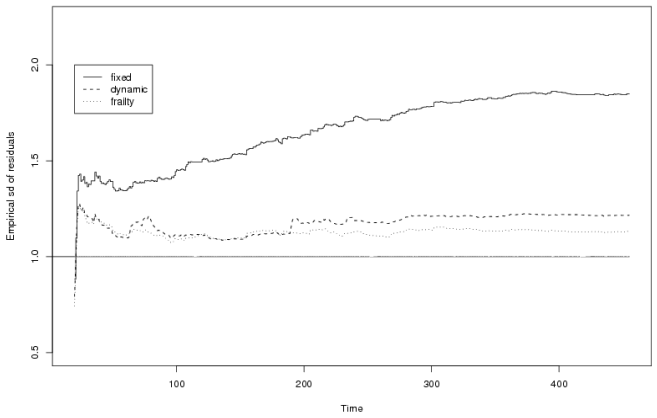

Fig. 5 Blue Bay data: the empirical standard deviations of standardized martingale residual processes for; solid line: fixed covariates, dashed line: dynamic covariate and dotted line: frailty model

A simulation study suggested that the dynamic model may be used instead of frailty model. Also we looked at the expected value of the standard deviation of the standardized residual processes for fitting different misspecified models and compared with the empirical value. We found that the empirical and the expected standard deviation of the standardized residual processes are very close to each other. For the case of fitting a dynamic model to data generated with frailty we needed to find $E[\hat{\Lambda}(t)]$ which depends not just on $N(t^-)$ but on the exact times of events over $(0, t^-)$. Thus the number of possible combinations quickly becomes unmanageable as t increases (Section III, A). Hence we did a very large simulation study to approximate the expected patterns.

In the final section we used the Blue Bay data as an illustration to compare the two procedures where we found

that both approaches took the heterogeneity into account and produce residual standard deviations close to each other.

ACKNOWLEDGMENT

My thanks go to Rosemerire Fiaccone and Mauricio Barreto for providing the diarrhoea data.

REFERENCES

[1] O. Aalen, J. Fosen, H. Wedon-Fekjaer, O. Borgan, and E. Husebye, "Dynamic analysis of multivariate failure time data," *Biometrics*, vol. 60, pp. 764-773, 2004.
[2] J. Fosen, O. Borgan, H. Weedon-Fekjaer, and O. Aalen, "Dynamic analysis of recurrent event data using the additive hazard model," *Biometrical Journal*, vol. 48, pp. 381-398, 2006.
[3] O. Borgan, R. L. Fiaccone, R. Henderson, M. L. Barreto, "Dynamic analysis of recurrent event data with missing observations, with application to infant diarrhoea in Brazil ," *Scandinavian Journal of Statistics* , vol. 34, pp. 53–69, 2007.
[4] P. Andersen, R. Gill, "Cox's regression model for counting processes: A large sample study," *The Annals of Statistics*, vol. 10, pp. 1100-1120, 1982.
[5] O. Aalen, "A linear regression model for the analysis of life times," *Statistics in Medicine*, vol. 8, pp. 907-925, 1989.
[6] O. Aalen, "Further results on the non-parametric linear regression model in survival analysis," *Statistics in Medicine*, vol. 12, pp. 1569-1588, 1993.
[7] J. Vaupel, K. Manton, E. Stallard, "The impact of heterogeneity in individual frailty on the dynamics of mortality," *Demography*, vol. 16, pp. 439-454, 1979.
[8] J. Vaupel, A. Yashin, "Heterogeneity's ruses: some surprising effects of selection on population dynamics," The American Statistician, vol. 39, pp. 176-185, 1985.
[9] O. Aalen, "Effects of frailty in survival analysis," *Statistical Methods in Medical Research*, vol. 3, pp. 227-243, 1994.
[10] P. Hougaard, "Analysis of multivariate survival data," Spring-Verlag:New York, 2000.
[11] D. Clayton, "A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence," *Biometrika,* vol. 65, pp.141-151, 1978.
[12] P. Hougaard, "Survival models for heterogeneous populations derived from stable distributions," *Biometrika,* vol. 73, pp. 671-678, 1986a.
[13] P. Hougaard, "A class of multivariate failure time distributions," *Biometrika*, vol. 73, pp. 671-678, 1986b.
[14] C. McGilchrist, and C. Aisbett, "Regression with frailty in survival analysis," *Biometrics*, vol. 47, pp. 461-466, 1991.
[15] O. Aalen, "Heterogeneity in survival analysis," *Statistics in Medicine*, vol. 7, pp. 1121-1137, 1988.
[16] O. Aalen, "Modelling heterogeneity in survival analysis by the compound poisson distribution," *Annals of Applied Probability*, vol. 4, pp. 951-972, 1992.
[17] A. Yashin, J. Vaupel, and I. Iachine, "Correlated individual frailty: An advantageous approach to survival analysis of bivariate data," *Mathematical Population Studies*, vol. 5, pp. 145-159, 1995.
[18] P. Andersen, O. Borgan, R. Gill, and N. Keiding, "Statistical Models Based on Counting Processes," Spring-Verlag:New York, 1993.
[19] P. Andersen, O. Borgan, R. Gill, and N. Keiding, "Statistical Models Based on Counting Processes," Spring-Verlag:New York, 1993.
[20] O. Aalen, O. Borgan, and H. Gjessing, "Survival and event history analysis. A process point of view," Spring-Verlag:New York, 2008.
[21] E. Elgmati, R. Fiaccone, R. Hendersen, and M. Mohammadi, "Frailty modeling for clustered recurrent incidence of diarrhoea," *Statistics in Medicine*, vol. 27, pp. 6489-6504, 2008.