Performance of recurrent event models on defect proneness data

M. K. Lintu¹ · Asha Kamath¹

Accepted: 19 November 2020 / Published online: 7 December 2020
© The Author(s) 2020

Abstract

The repeated occurrence of the same event in a process is commonly observed in many domains. Such events are referred to as recurrent events. The time to occurrence of these repeated events varies from unit to unit with a possibility of events not occurring among some of the units. Invariably such data are dealt with using some of the techniques in survival analysis called recurrent event models, which are commonly encountered in epidemiological studies and clinical trials. However, it applies to other domains in science and technology. We illustrate the usefulness of recurrent event models in the context of defect proneness analysis in quality assessment of software. Some of the models in practice are introduced on data collected to study the impact of module size on defect proneness in the Mozilla product. Module size plays a significant role in defect proneness and each defect fix makes the class more susceptible to further defects. The risk estimates obtained from the different models vary owing to the differences in the properties of the models as well as the assumptions underlying it.

Keywords Recurrent events · Survival model · Extended Cox models · Defect proneness

Mathematics Subject Classification 62N01 · 62N05

1 Introduction

Survival analysis builds a relationship between time to occurrence of an event to the covariates that influence it. This finds a lot of applications in medicine, economics, finance, and engineering. However, the likelihood of observing more than one event during the study period is very common regardless of the field of application and they are called recurrent or multiple events. The challenge underlying this problem is that the probability of one event is likely to get influenced by the earlier event even if they are of a different type.

✉ Asha Kamath
asha.kamath@manipal.edu

M. K. Lintu
lintu.mk@learner.manipal.edu

¹ Department of Data Science, Manipal Academy of Higher Education, Manipal, Karnataka 576104, India

Cox proportional hazards model (Cox 1972), a well-known approach for the analysis of survival data, has been extended to multiple events as well as events with non-proportional hazards. But the problem in extending the proportional hazards model to multiple events is the intra-subject correlation (Therneau and Grambsch 2013; Hougaard 2012). Some extensions of the Cox proportional model have been proposed for analysis, which belongs to the class of intensity processes such as Andersen and Gill model (AG) (Andersen and Gill 1982), Prentice, Williams, and Peterson models (PWP) (Prentice et al. 1981) and Wei, Lin and Weissfeld model (WLW) (Wei et al. 1989) to address the intrasubject correlation. Different types of frailty models (Vaupel et al. 1979), as well as multi-state models (MSM) (Hougaard 1999), are also popular in use for multiple event problems. The choice among the different models depends on the research question under investigation. These models are illustrated using data collected to study the effect of class size on defect proneness in the Mozilla product (Gunes et al. 2007, 2008).

Gunes et al. found out that the quality assurance activities in biomedical open-source software (OSS) were not sufficient to be used for mission-critical purposes, thus, justifying the significance of studying the characteristics of defect prone open-source software modules to take preventive actions (Gunes et al. 2007). The event of interest in this problem is the defect fixes made to C++ classes in the open-source Mozilla project. Since there was more than one defect per class, the data falls in recurring event setup. The purpose of the analysis is to obtain the instantaneous hazard rate of fixing a defect at a time t for a C++ class. Such data have been dealt with using various methodologies in epidemiology and clinical trials, some of which are introduced using the defect proneness data. The data layout for the application of each of the methods and comparison of the models have been discussed.

2 Data

Defect prediction modeling is an integral area in software maintenance. Most often, the size of the software is one of the reasons of defect proneness (El-Emam et al. 2000). Gunes et al. collected data from a Mozilla project to study this association (Gunes et al. 2007). The data consists of the complete history of all the C++ classes introduced to the Mozilla project from May 29, 2002, to February 22, 2006, which was available publically in the PROMISE repository (Sayyad Shirabad and Menzies 2005). The complete data was accessed from Open Machine Learning (OpenML) (Vanschoren et al. 2013). The event of interest of the study was defect fixes made to C++ classes added to the public code repository of the Mozilla product by developers. The revision history of the followed-up classes was extracted from the Concurrent Version System (CVS) repository. From the list of revisions, defect fixes were identified by the keywords “bug”, “fix”, and “defect” in the commit log message.

Time taken for repeated defect fixes of each C++ class was recorded in minutes. Each class introduced in the system was either followed up until the end of the observation period or until the class was deleted. If the class is removed to fix a defect, it was accounted as an event, otherwise censored. Even though classes were introduced at different calendar times to the system, starting time is defined as zero. The only covariate observed in the study (Gunes et al. 2007) was the size of the class as the study was conducted to establish the impact of module size on defect proneness. The size of the class was measured in terms of the number of lines of code without considering comment lines and blank lines. Class size is a time-varying covariate since the length of the code changes with time. A total of 8828 defects were fixed

during the study period which belonged to 4089 classes. The observations for each revision are recorded with the fields:

- ID: A unique identifier for each class.
- Start: Start = 0, when a class is introduced. Otherwise, the time when the revised class is reintroduced to the system.
- End: Next revision time of the class or the deletion time, or the time at which the study period ended, whichever occurs first.
- Event: An indicator that tells whether the revision was for defect fixing or not.
- Size: Number of lines of the code at the time of revision.

The median(Interquartile range) class size was 126 (36–418) lines. The data were trimmed for at most 4 defect fixes due to the anticipated instability of the estimates. Log transformed size was incorporated as a covariate in the models as done in previous studies (Gunes et al. 2007, 2008).

The analysis of repeated events can be approached in two different ways: (i) by considering the total time (in minutes) or (ii) by considering the gap time (in minutes) between the events. The data should be arranged appropriately to fit different models successfully. The formats for AG, PWP, WLW, MSM, and frailty models are described in Table 1 using the entries of first, second, and third classes.

3 Models

This section introduces some of the models using the data described in Sect. 2 with the demonstration of the relationship between the defect proneness and size of an OSS.

3.1 Andersen and Gill (AG) model

The intensity of recurring defect fixes is modeled using the realization of the counting process(counting observed defect fixes up to a particular time t) (Andersen and Gill 1982). Assuming the defect fixes to be independent and identical, defects were treated as the same irrespective of their order. AG model can give the best fit if the risk of further defects remains the same regardless of the previous defects (Hosmer Jr et al. 2008). The likelihood of a defect fix at time point t is unaffected by the previously observed defects and hence the baseline hazard function of all the defects is common (Amorim and Cai 2015).

The advantages of the model are its compatibility with complicated censoring schemes as well as time-dependent covariates, which can be used to bring in the dependency between the defects of the same class (Ullah et al. 2014). Moreover, the AG model is well incorporated in most of the statistical software.

3.2 Prentice, Williams and Peterson (PWP) model

This model is commonly used to analyze survival data with ordered recurrent events (Prentice et al. 1981). Modeling of hazard function is performed by stratification of the classes such that only those classes with a defect in the previous stratum are considered to be at risk in successive strata while all other classes are included in the first stratum. The risk set for the k th defect fix at time point t includes only those classes which had $(k - 1)$ th defect in advance and not experienced k th defect at time t (Amorim and Cai 2015; Castañeda and

Table 1 Data layout for different models

Model	Layout					
AG and frailty	id ^a	Start ^b	End ^c	Event	Size	
	1	0	913,439	0	14	
	1	913,439	1,292,405	1	14	
	1	1,292,405	1,540,749	0	14	
	2	0	50,170	0	14	
	3	0	913,439	0	18	
	3	913,439	1,292,405	1	18	
	3	1,292,405	1,540,749	0	18	
PWP	id	Start	End	Event	Size	Strata
	1	0	913,439	0	14	1
	1	913,439	1,292,405	1	14	1
	1	1,292,405	1,540,749	0	14	2
	2	0	50,170	0	14	1
	3	0	913,439	0	18	1
	3	913,439	1,292,405	1	18	1
	3	1,292,405	1,540,749	0	18	2
PWP GT	id	Start	End	Event	Size	Strata
	1	0	913,439	0	14	1
	1	913,439	1,292,405	1	14	1
	1	0	248,344	0	14	2
	2	0	50,170	0	14	1
	3	0	913,439	0	18	1
	3	913,439	129,2405	1	18	1
	3	0	248,344	0	18	2
WLW	id	Start	End	Event	Size	Event no.
	1	0	913,439	0	14	1
	1	0	1,292,405	1	14	2
	1	0	1,540,749	0	14	3
	1	0	1,540,749	0	14	4
	2	0	50,170	0	18	1
	2	0	50,170	0	18	2
	2	0	50,170	0	18	3
MSM	id	Start	End	Event	Size	State
	1	0	913,439	0	14	1
	1	913,439	1,292,405	1	14	2
	1	1,292,405	1,540,749	0	14	2
	2	0	50,170	0	14	1
	3	0	913,439	0	18	1

Table 1 continued

Model	Layout					
	3	913,439	1,292,405	1	18	2
	3	1,292,405	1,540,749	0	18	2

^aClass identification^bObservation starting time (in minutes)^cObservation ending time (in minutes)

Gerritse 2010). A separate column is included in data to take care of this stratification. The baseline hazard that varies across defects is a function of time since the study entry. The model is capable of incorporating both overall and defect-specific effects of the covariate and it is best suited for a small number of defects observed on a large number of classes. The hazard function for the k th defect of the i th class is given in Table 2.

3.3 Prentice, Williams and Peterson gap-time (PWP-GT) model

The PWP-GT model is a modification of the PWP model wherein time to a defect is observed since the immediately preceding defect fix (Prentice et al. 1981). The very next row of every defect fix will have a starting time zero in the data layout. Unlike the PWP model, the baseline hazard is a function of time since the previous defect fix. This model is highly preferred since it enables incorporating the effect of varying class sizes between any two successive defect fixes.

PWP models are conditional models since a class is at risk for a new defect only if it has sustained a previous defect and hence the hazard of a class at any time is conditional on the previous defect fixes made to the same class. More specifically, PWP as well as the PWP-GT model can be seen as a stratified version of the Cox model (Cox 1972). The dependency between the defects is incorporated through the strata formed. Hazard function for the k th defect of the i th class can be found in Table 2.

3.4 Wei, Lin, and Weissfeld (WLW) model

Unlike the other models, the mean number of defect fixes is modeled in the WLW model (Wei et al. 1989). The model can be used if the investigator is interested in the rate or the expected number of defect fixes, conditional on the covariate. All recurring defects of a class are treated as a single counting process. Moreover, if no time-varying (aka time-dependent) covariate is considered in the AG model to take care of the event history, the point estimates are the same as that of the WLW model (Amorim and Cai 2015). In such cases, the WLW model is preferred over the AG model due to its comparative flexibility. The model avoids specifying the dependence structure among different defect fixes received by the same class (Wei et al. 1989; Metcalfe and Thompson 2007).

One of the drawbacks of the model is an overestimation of the effect of size as all the classes are at risk for the maximum number of defects reported, even though they have developed a very low number of defects. In the data layout, every class has four rows since the maximum number of defects considered was four. The advantage of the model is the compatibility with unordered events.

Table 2 Model specifications for recurrent event survival models

Model	Risk set	Baseline hazard	Dependency	Hazard function
AG	Any class is at risk for any defect	Common across all defect fixes	Defects are assumed to be independent	$\lambda_i(t) = \lambda_0(t) \exp(X_i(t)\beta_k)$
PWP	Classes with the previous defect are at risk	Function of time since the study entry which differs across defects	Dependency is incorporated by preserving the order of defects	$\lambda_{ik}(t) = \lambda_{0k}(t) \exp(X_i(t)\beta_k)$
PWP-GT	Classes with the previous defect are at risk	Function of time since the previous defect fix which differs across defects	Dependency is incorporated by preserving the order of defects	$\lambda_{ik}(t) = \lambda_{0k}(t - t_{k-1}) \exp(X_i(t)\beta_k)$
WLW	All classes are at risk for all defects	Defect-specific baseline hazard	Does not specify the dependence structure	$\lambda_{ik}(t) = \lambda_{0k}(t) \exp(X_i(t)\beta_k)$
Frailty	Some classes are at more risk for some defects than others	Common baseline hazard	Random covariate introduced in the model takes care of the dependency	$\lambda_i(t) = \lambda_0(t) Z_i \exp(X_i(t)\beta_k)$

 $X_i(t)$: size of the class i at time t β_k : regression coefficient λ_0 : baseline hazard

Fig. 1 Graphical representation of the 5-state MSM fitted to the data

3.5 Frailty model

The frailty model takes care of the unobserved heterogeneity of different classes in the hazard function (Vaupel et al. 1979). The model is useful when the classes are of different risks. A random covariate is introduced in the AG model to account for the excess risk associated with different classes. Conditional frailty models can be fitted with the same methodology extended to the PWP models.

Just like the concept of mixed models (McLean et al. 1991) in the analysis of longitudinal data, class-specific interpretation of the estimates is introduced in this model. If the random effects are small, a larger number of defects fixes are required. Otherwise, a smaller number of defect fixes would be adequate to proceed with the frailty model.

3.6 Multi-state model (MSM)

The instantaneous hazard of progression to one particular state conditional on occupying the other state is modeled in MSM. The model provides a covariate effect on each transition. A five-state model is introduced for the defect fix data. Defect-free classes belong to state 1, classes with one, two, and three defects belong to state 2, state 3, and state 4 respectively, and state 5 indicates more vulnerable classes with more than three defects.

States are called absorbing if there is no path out from it and transient otherwise. In the data, states 1, 2, 3, and 4 have further transitions and hence they are transient. State 5 is absorbing since we defined it as the state where the more vulnerable classes belong to. Since a defective class has no going back to a defect-free state, the transitions are irreversible. The possible transitions are from state 1 to state 2, state 2 to state 3, state 3 to state 4, and state 4 to state 5. Figure 1 represents the 5-state model fitted to the data. A column is incorporated to indicate the state in which a class belongs to. The advantage of MSM is the estimates obtained for each transition in the data, which gives the influence of the covariates in each transition (Table 4).

Based on the different assumptions regarding the transition rates, different multi-state models such as Markov, semi-Markov, or non-Markov, etc. can be fit for multiple event survival data (Meira-Machado et al. 2009).

Brief description including the pieces of information like the risk set considered, baseline hazard function, dependence structure, etc. of the different models discussed from Sect. 3.1–3.6 can be found in Table 2. In the functional form of the hazard function, λ_0 denotes the baseline hazard, i.e., the underlying hazard without any covariate effect. X and β stand for covariate and unknown regression coefficients respectively.

4 Results

A total of 8828 defects were fixed during the observation period for 4089 classes. Above-mentioned models were fit using *survival* package (Therneau 2015) in RStudio (RStudio Team 2016). Parameter estimates obtained from the models are given in Tables 3 and 4. The

Table 3 Parameter estimates obtained from different models

Model	Parameter estimates	Hazard ratio (CI ^a)	LL ^b	AIC ^c	BIC ^d
AG	0.47*	1.60 (1.58,1.62)	− 66,618	133,239	133,246
PWP	0.26*	1.30 (1.27,1.34)	− 54,487	104,923	104,931
PWP-GT	0.24*	1.27 (1.25,1.30)	− 55,577	108,089	108,096
Frailty	0.15*	1.16 (1.14,1.19)	− 75,235	153,472	164,109
WLW	0.14*	1.15 (1.12,1.17)	− 77,611	155,224	155,231

^a95% confidence interval^bLog likelihood^cAkaike information criterion^dBayesian information criterion**p* value < 0.001**Table 4** Parameter estimates obtained from multi-state model

Transition	Parameter estimates	Hazard ratio (CI ^a)
Defect-free → Defective	0.19*	1.22 (1.16, 1.27)
First defect → Second defect	0.19*	1.21 (1.15, 1.26)
Second defect → Third defect	0.12*	1.13 (1.07, 1.20)
Third defect → Fourth defect	0.17*	1.19 (1.11, 1.27)

^a95% confidence interval**p* value < 0.001

hazard ratio is the exponential of the parameter estimate. A considerable difference exists in the estimates from the different models due to their assumptions and properties.

From the AG model, the hazard ratio observed is 1.60 with a 95% confidence interval (1.58, 1.62). i.e., one unit increase in the $\ln(\text{size})$ of the class is associated with 60% more risk of recurrence.

The hazard ratio from the PWP model is 1.32 (1.28, 1.36). The risk set for the PWP model was formed using the classes having the same number of previous defects and hence for a larger number of defects, the strata become too small. The number of defects was truncated to 4 to overcome this limitation.

According to the frailty model, the instantaneous risk of recurrence is increased by 16% or one unit increase in $\ln(\text{size})$.

According to the WLW model, the hazard for a defect fix increases by 15% with a unit increase in $\ln(\text{size})$.

The multi-state model gives the hazard ratio associated with each transition for a unit increase in the size of the class. The results can be found in Table 4.

The goodness of fit was assessed using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and log-likelihood. PWP models had the best performance in terms of AIC, BIC, and log-likelihood values. From the properties of the PWP model, it is clear that the model is most suited for the ordered events. Hence the result is obvious. AG model treats the defects as the same irrespective of the order. From Fig. 1, it is clear that the survival probability decreases as the number of prior defects increases. AG model gives the best fit if the risk of further defects remains the same regardless of the previous defects, which is not so in this data.

Fig. 2 Kaplan Meier estimates of the survival function stratified by number of previous defects

Figure 2 represents the Kaplan Meier estimates of the survival function, stratified by the number of previous defects. As the number of previous defects increases, survival considerably decreases. More specifically, classes with one previous defect are 2.3 times more susceptible to further defects. Classes with two, three, and four prior defects are 3.9, 4.3, and 9.6 times more prone to further defects according to the stratum-wise PWP model.

5 Conclusions

Characteristics of the different recurrent event models were introduced with the help of the defect fix data. As expected, each defect was making the class more susceptible to further defects. Parameter estimates obtained from various models were considerably different due to the disparity in their properties and assumptions. From the AIC, BIC, and log-likelihood values, we can infer that the PWP model gives the best fit for the defect fix data. Still, the selection of the appropriate model again depends on different factors such as research questions, type of defects, etc.

Most of the time, the above-mentioned models were introduced in the context of epidemiology and clinical trials. The same scenario occurs very often in science and technology also. In software quality assessment, recurrent event models can give more meaningful insights and the paper gives illustrations about the different approaches to recurrent event data which can be further applied in quality modeling.

Acknowledgements We would like to thank Gunes et al. for contributing the dataset to the PROMISE repository.

Funding Open access funding provided by Manipal Academy of Higher Education, Manipal.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give

appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Amorim, L. D., & Cai, J. (2015). Modelling recurrent events: A tutorial for analysis in epidemiology. *International Journal of Epidemiology*, 44(1), 324–333.
- Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10, 1100–1120.
- Castañeda, J., & Gerritse, B. (2010). Appraisal of several methods to model time to multiple events per subject: Modelling time to hospitalizations and death. *Revista Colombiana de Estadística*, 33(1), 43–61.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187–202.
- El-Emam, K., Benlarbi, S., Goel, N., Melo, W., Lounis, H., & Rai, S. (2000). The optimal class size for object-oriented software: A replicated study.
- Gunes, K. A., Dongsong, Z., & Hongfang, L. (2007). Modeling the effect of size on defect proneness for open-source software.
- Gunes, K. A., Khaled, E. E., Dongsong, Z., Hongfang, L., & Divya, M. (2008). Theory of relative defect proneness. *Empirical Software Engineering*, 13(5), 473.
- Hosmer, D. W., Jr., Lemeshow, S., & May, S. (2008). *Applied survival analysis: Regression modeling of time-to-event data* (Vol. 618). New York: Wiley.
- Hougaard, P. (1999). Multi-state models: A review. *Lifetime data analysis*, 5(3), 239–264.
- Hougaard, P. (2012). *Analysis of multivariate survival data*. Berlin: Springer.
- McLean, R. A., Sanders, W. L., & Stroup, W. W. (1991). A unified approach to mixed linear models. *The American Statistician*, 45(1), 54–64.
- Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suarez, C., & Andersen, P. K. (2009). Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research*, 18(2), 195–222.
- Metcalfe, C., & Thompson, S. G. (2007). Wei, lin and weissfeld's marginal analysis of multivariate failure time data: Should it be applied to a recurrent events outcome? *Statistical Methods in Medical Research*, 16(2), 103–122.
- Prentice, R. L., Williams, B. J., & Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, 68(2), 373–379.
- RStudio Team. (2016). *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio Inc.
- Sayyad Shirabad, J., & Menzies, T. (2005). *The PROMISE Repository of Software Engineering Databases*. Ottawa: School of Information Technology and Engineering, University of Ottawa.
- Therneau, T. M. (2015). *A package for survival analysis in S*. version 2.38.
- Therneau, T. M., & Grambsch, P. M. (2013). *Modeling survival data: Extending the Cox model*. Berlin: Springer.
- Ullah, S., Gabbett, T. J., & Finch, C. F. (2014). Statistical modelling for recurrent events: An application to sports injuries. *British Journal of Sports Medicine*, 48(17), 1287–1293.
- Vanschoren, J., van Rijn, J. N., Bischl, B., & Torgo, L. (2013). Openml: Networked science in machine learning. *SIGKDD Explorations*, 15(2), 49–60.
- Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3), 439–454.
- Wei, L.-J., Lin, D. Y., & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American statistical association*, 84(408), 1065–1073.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.