Semiparametric approaches for modelling mortality with application to university employee data

Danardono, R S Utami, R Fitriani, Zulaela and S H Kartiko

Department of Mathematics, Universitas Gadjah Mada, Indonesia

E-mail: danardono@ugm.ac.id

Abstract. Semiparametric approaches using Kernel, Splines and locally-weighted regression smoothing were used to model mortality pattern among the government university employee as alternative to the parametric approaches. The results from those three type of smoothing methods were compared. The final results of the estimates were also compared to the standard model of mortality pattern among the government employee in Indonesia, the TASPEN mortality table, and the population life table from the 2010 Indonesian population census.

Keywords: Kernel smoothing, Spline Smoothing, Locally-weighted Regression Smoothing, Life Table, Graduation

1. Introduction

Human mortality models, in the form of statistical models, are commonly constructed from raw estimates of death rate and then graduated or smoothed either parametrically or non-parametrically. The parametric human mortality models involve certain functional form with finite number of parameters. The Gompertz or Gompertz-Makeham models [1, 2] that have been known for centuries, are examples of the parametric models. More recent parametric models are the Heligman-Pollard models [3], Lee-Carter models [4] and Cairns-Blake-Dowd models [5]. Parametric approaches provide convenient ways to analyze mortality data, however, they require careful check on the assumptions and sometimes do not fit to the observed data.

Alternatively, semiparametric or non-parametric approaches, which do not specify any functional form of the models and are completely data-driven, can be used to model the mortality data. Moreover, these alternative methods can be used to parametric model checking.

The semiparametric or non-parametric approaches for mortality modeling are also known as graduation methods or smoothing methods in the literature. The graduation methods use the raw estimates of death rate and then construct a final mortality model in the form of life table or mortality table. The classical method of graduation commonly used in the life table estimates is the Whittaker or Whittaker-Henderson method [6]. The method is basically a Bayesian formulation using inverse probability argument to estimate the graduated values that minimizing certain objective function. The method is very popular in graduating a life table, however the choice of the balance between discrepancy and roughness is difficult [7]. The other approaches are based on smoothing principles such as Weighted Moving Average Methods [7], Kernel Smoothing [7, 8], Splines Smoothing [9] and Locally-weighted Regression (LOESS) [10]. The Weighted Moving Average Methods as the pioneer in the smoothing principle had been clearly improved by the other last three methods.
The aim of the paper is to compare three methods based on smoothing principles, i.e., the Kernel smoothing, the Spline smoothing and LOESS and implement the methods to analyze the mortality pattern among the university employee. All computation were performed using R [11]. The final result is compared to the standard model of mortality pattern among the government employee in Indonesia, i.e., the TASPEN mortality table [12] and the population life table from the 2010 Indonesian population census [13].

2. Semiparametric Approach
The methods to model mortality pattern in this article are considered as Semiparametric in the sense that these methods combine parametric approach to estimate the raw estimates of death rate, and then to smooth or graduate the raw estimate using non-parametric models. In this section, the method to estimate the death rates and three methods of semiparametric models are discussed.

2.1. Raw Estimates of Death Rate
The raw estimates of death rate can be obtained from data that provides the number of deaths for certain age \( x \) and the number of person or exposure in the age \( x \) for certain period of time. Such data may be obtained from a specific survey or census; or from administrative regular records. In the university human resource department for example, they have records on the birth of their employees and their current status (either active, pensioner, non-active, withdraw or deceased). In a particular period of time, the following quantities can be obtained: the number of deaths for a specific age \( (x) \) denoted by \( d_x \), and the exposure of the person in the age \( x \). The raw estimate of death rate typically use one year of age interval, such that \( d_x \) means the number of deaths during interval \([x, x + 1)\) of age.

Exposure can be determined simply by counting the number of persons aged \( x \) in the period. Another method is by calculating the exact observation for each individuals, in the sense that when the death occurs, the person’s exposure ends at the exact age of death \([14, 15]\). In this paper, the exact exposure is used, which has a similar concept as person-years in epidemiology.

Under assumption of constant hazard rate in the interval \([x, x + 1)\), the estimate of hazard rate \( \hat{\lambda} \) in that interval is

\[
\hat{\lambda}_x = \frac{d_x}{E_x}
\]

where \( d_x \) is the number of deaths in \([x, x + 1)\), and \( E_x \) is the exact exposure in the interval \([x, x + 1)\).

The quantities used to construct the semiparametric models are the estimate of \( q_x \), i.e., the conditional probability of death within one year interval for individual aged \( x \). The estimate is

\[
\hat{q}_x = 1 - \exp(\hat{\lambda}_x).
\]

This estimation is basically a maximum likelihood of a constant hazard model in the interval \([x, x + 1)\), therefore, it is the parametric part of the semiparametric models.

2.2. Kernel Smoothing
This technique is initially developed to estimate density functions. Kernel smoothing uses a weighted average with weights based on a function \( K \) called a kernel. The weights are

\[
\omega_{x_i} = \frac{K \left( \frac{x - x_i}{b} \right)}{\sum_{i=1}^{r} K \left( \frac{x - x_i}{b} \right)}
\]

2
where \( b \) is the bandwidth parameter and \( K \) is the kernel function. There are several types of kernel estimations: Normal, Gaussian, Triangular, Parzen and Epanechnikov. The choice of \( K \) does not have as much influence as the value of the smoothing parameter \( b \) [16].

To estimate \( q_x \), kernel estimators based on the Copas and Haberman or based on the Nadaraya-Watson estimator may be used [16]. The Nadaraya-Watson, however, is better than the other estimator for mortality data [17], and were used in this paper. The Nadaraya-Watson estimator for \( q_x \) is as follows

\[
\hat{q}_x^{NW} = \frac{\sum_i \tilde{q}_x K \left( \frac{x - x_i}{b} \right)}{\sum_i K \left( \frac{x - x_i}{b} \right)}
\]

where \( \tilde{q}_x \) is the raw estimate of \( q_x \), and NW stands for Nadaraya-Watson.

### 2.3. Splines Smoothing

Consider the \( q_x \) as continuous function \( f(x) \) and define the penalized sum of squares

\[
\sum_{j=1}^{n} \left( \tilde{q}_j - f(x_j) \right)^2 + \lambda \int_{x_1}^{x_n} \left( f''(t) \right)^2 dt
\]

(5)

where the first term measures the deviation between the raw estimates and the adjusted (graduated) values, and the second term increases with the fluctuations of the \( f(x) \). Parameter \( \lambda \) determines the relative importance of the terms as bandwidth in the Kernel Smoothing.

### 2.4. Locally-weighted Regression (LOESS)

This method locally adjusts polynomials of low degree. Locally means setting neighborhoods \( N(x) \) and assigning weight to each point of \( N(x) \). The weighting function is

\[
\omega_{x_i} = T \left( \frac{|x - x_i|}{\Delta(x)} \right)
\]

(6)

where \( \Delta(x) = \max_{x_i \in N(x)} |x - x_i| \) and

\[
T(u) = \begin{cases} 
(1 - u^3 P)^3, & \text{for } 0 \leq u < 1 \\
0, & \text{otherwise}
\end{cases}
\]

(7)

The degree of smoothness is determined by the span parameter, the proportion of the number of points in the neighborhood in comparison with the total number of points [16].

### 3. Application to University Employee Data

The data for this application were provided by human resource database in a government university in Yogyakarta, Indonesia. The period of observation for the study is from January 2013 until mid of August 2018. From this period, 9,127 records (employees) were obtained, and 113 individuals were deceased. The raw estimate of \( q_x \)’s are presented in figure 1 in natural logarithm scale. Unfortunately, the data for the lower ages (below 20) and the older ages (above 70) are quite limited both in the number of deaths and the amount of exposures, giving unreliable estimates in those age categories.
The first smoothing method used is LOESS smoothing method. The degree of smoothing was controlled by a span parameter, which was selected automatically using generalized cross validation method. The degree of local polynomial used in this study was 1. Figure 2 shows the raw estimates of $q_x$ in terms of its logarithmic values which were calculated for each age, and the smooth estimates of $\log(q_x)$ which were obtained from LOESS smoothing method.

Splines smoothing was the second technique of smoothing that was performed. Figure 3 shows the plot of estimates of $\log(q_x)$ that was obtained through Splines smoothing method. Cross validation method was used to select the optimum smoothing parameter for the Spline method. In comparison to the estimates of $\log(q_x)$ from the LOESS method, the estimates of $\log(q_x)$ produced by this method seemed to fall in a slightly smoother line.
The third nonparametric method of smoothing is Kernel smoothing. Nadaraya-Watson estimator was used to obtain the smooth estimates of \( \log(q_x) \). The estimation was based on a Normal kernel. Several choices of bandwidth were used and resulted in different degree of smoothness. The result presented in figure 4 is smooth estimates of \( \log(q_x) \) which was obtained from Kernel method with bandwidth of 10. Compared to the Loess method and Spline method, Kernel method provided the smoothest estimates, as can be seen in figure 4.

It can be seen from figure 2, figure 3 and figure 4 that the results are all quite similar. A certain semiparametric smoothing do not seem superior to the others. However, each models gave MSE of 7.543; 4.397; and 5.132 for Kernel, Spline and LOESS, respectively. Therefore, Spline has the best fit for this mortality data.
In figure 5, the mortality models for the university employee are compared to the standard mortality table for the government employee (TASPEN-2012 life table) and to the Indonesian Life Table constructed from the 2010 Census (ILT-2010). The mortality rates of the university employees are lower than that of the general population (ILT-2010). The TASPEN life table was constructed from all government employee records under the national pension program. The mortality rates of the university employees are also slightly lower than the TASPEN life table, except for age below 30 when the rates are higher; and for age 60 when the rates are close. However, the pattern of mortality of the government employee seems to increase after 40, in addition to the accidental hump in the younger age (15-20). This pattern is different with the pattern in the general population or in the general government employee.

4. Conclusions
Semiparametric smoothing using Kernel, Splines or LOESS can be used to graduate or adjust mortality rates given raw mortality estimates. In the application, Splines gives better fit than other methods. However, the difference in the adjusted estimates among the three methods were minor. The application signifies that the mortality rates among the state university employee are lower than that of general population, however, the pattern is different. The mortality rates of the state university employee tend to increase after age 40 before reaching the same rates as the rates in the general government employee.

References
[1] Gompertz B 1825 On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies Philos. Trans. R. Soc. Lond. A 115 513-85
[2] Makeham W M 1860 On the law of mortality J. Inst. Actuaries 13 325-58
[3] Heligman L and Pollard J H 1980 The age pattern of mortality J. Inst. Actuaries 107 49-80
[4] Lee R D and Carter L R 1992 Modeling and forecasting U.S. mortality JASA 87 659-75
[5] Cairns A J G, Blake D and Dowd K 2006 A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration J. of Risk and Insurance 73 687-718
[6] Henderson R 1924 A new method of graduation Transactions of American Society of Actuaries XXV
[7] Miller R B 2004 Graduation Encyclopedia of Actuarial Science Wiley
[8] Gavin J, Haberman S and Verrall R 1993 Moving weighted average graduation using kernel estimation Insurance: Mathematics and Economics 12(20) 113-126
[9] Hastie T J 1992 Generalized additive models Chapter 7 of Statistical Models in S ed J M Chambers and T J Hastie, Wadsworth & Brooks/Cole
[10] Cleveland W 1979 Robust locally-weighted regression and smoothing scatterplots. JASA 74 829-36
[11] R Core Team 2018 R: A language and environment for statistical computing R Foundation for Statistical Computing Vienna Austria URL https://www.R-project.org/
[12] TASPEN dan BPA 2012 TMT 2012 - Tabel Mortalita TASPEN 2012 PT. TASPEN (PERSERO) dan BPA (Biro Pusat Aktuaria) PT Sentra Jasa Aktuaria
[13] BPS 2015 Kajian lifetable Indonesia berdasarkan hasil SP2010 BPS
[14] London D 1997 Survival models and their estimation third eds ACTEX Publication
[15] Klugman S A, Panjer H H and Wilmott G E 2012 Loss Models: From Data to Decisions 4th Edition, John Wiley and Sons
[16] Debón A, Montes F and Sala R 2006 A Comparison of Nonparametric Methods in the Graduation of Mortality: Application to Data from the Valencia Region (Spain) International Statistical Review / Revue Internationale De Statistique 74(2) 215-33
[17] Gavin J, Haberman S and Verrall R 1994 On the choice of bandwidth for kernel graduation J. of the Institute of Actuaries 121 119-34