Estimation and Forecasting Survival of Diabetic CABG Patients (Kalman Filter Smoothing Approach)

M. Saleem¹, K. H. Khan², Nusrat Yasmin¹

¹Centre for Advanced Studies in Pure and Applied Mathematics, Bahauddin Zakariya University, Multan, Pakistan
²Department of Mathematics, College of Science and Humanities, Prince Sattam Bin Abdulaziz University, Al Khaj, Saudi Arabia

Received 9 September 2015; accepted 23 November 2015; published 26 November 2015

Abstract

In this paper, we present a new approach (Kalman Filter Smoothing) to estimate and forecast survival of Diabetic and Non Diabetic Coronary Artery Bypass Graft Surgery (CABG) patients. Survival proportions of the patients are obtained from a lifetime representing parametric model (Weibull distribution with Kalman Filter approach). Moreover, an approach of complete population (CP) from its incomplete population (IP) of the patients with 12 years observations/follow-up is used for their survival analysis [1]. The survival proportions of the CP obtained from Kaplan Meier method are used as observed values $y_t$ at time $t$ (input) for Kalman Filter Smoothing process to update time varying parameters. In case of CP, the term representing censored observations may be dropped from likelihood function of the distribution. Maximum likelihood method, in-conjunction with Davison-Fletcher-Powell (DFP) optimization method [2] and Cubic Interpolation method is used in estimation of the survivor’s proportions. The estimated and forecasted survival proportions of CP of the Diabetic and Non Diabetic CABG patients from the Kalman Filter Smoothing approach are presented in terms of statistics, survival curves, discussion and conclusion.

Keywords

CABG Patients, Complete and Incomplete Populations, Weibull & Distribution, Kalman Filter, Maximum Likelihood Method, DFP Method, Estimation and Forecasting of Survivor’s Proportions

1. Introduction

The Coronary Artery Disease (CAD) is a chronic disease, which progresses with age at different rates. CAD is a result of built-up of fats on the inner walls of the coronary arteries. Thus, the sizes of coronary arteries become narrow and as a result the blood flow to the heart muscles is reduced/blocked. Therefore, the heart muscles do not receive required oxygenated blood, which leads to the heart attack. CAD is a leading cause of death worldwide (see Hansson [3], John [4], and Sun and Hong [5], William, Stephen, Thomas and Robert [6]). The medical scientists Goldstein [7] and Jennifer [8], William [6] are of the opinion that CABG is an effective treatment option for CAD patients. The medical research organizations like Heart and Stroke Foundation Canada [9], American Heart Association [10] and Virtual Health Care Team Columbia have classified risk factors of CABG patients as modifiable (Hypertension, Diabetes, Smoking, High Cholesterol, Sedentary Lifestyle and Obesity) and non-modifiable (Age, Gender and Family History-Genetic Predisposition).

William, Ellis, Josef, Ralph and Robert [6] carried out the survival study on incomplete population (progressive censoring of type 1) of CABG patients comprising 2011 patients using Kaplan Meier method [11]. The patients were grouped with respect to Male, Female, Age, Hypertension, Diabetes, and Ejection Fraction, Vessels, Congestive Heart Failure, Elective and Emergency Surgery. The patients were undergone through a first re-operation at Emory University hospitals from 1975 to 1993 (see William [12]. The patients were observed/followed up for 12 years. In the article [13] [14] we proposed a procedure, to make an IP, and a CP.

The Weibull distribution model has been used for survival analysis by Abrenthy [15], Bunday [16], Cohen [17], Crow [18], Gross and Clark [19], Klein & Moeschberger [20], Lang [21], Lawless [22], and Paul [23]. In particular, the survival study of chronic diseases, such as AIDS and Cancer, has been carried out using Weibull distributions by Bain and Englehardt [24], Khan & Mahmud [23] [25], Klein & Moeschberger [20], Lawless [22] and Swaminathan and Brenner [26]. Lanju & William [27] used Weibull distribution to human survival data of patients with plasma cell and in response-adaptive randomization for survival trials respectively. We [14] have carried out survival analysis of CABG patients by parametric estimations-classical approach, in modifiable risk factors (Hypertension and Diabetes).

The dynamic linear model (DLM) and Kalman Filter (KF) equations have been described by Harrison and Steven [28]. According to the researchers, Sorenson [2] and Greg [29] Kalman Filter is a mathematical technique, used to estimate the state of a process by minimizing error of estimation. Kalman Filter extracts signals from a series of incomplete and noisy measurements. It removes noises from the process parameters and retains useful information. Kalman filter estimates the state of a dynamic linear model through its recurrence equations which minimizes the variance of estimation error. To implement Kalman filter, observed values as dependent variables are required for updating the process parameters. Though, since time of introduction, the Kalman Filter has been subject of research for engineering processes see Frank [30], however the KF methodology has been applied extensively in medical research/life-testing studies/survival analysis; for example, Meinhold and Singpurwalla [31] proposed a new method for inference and extrapolations in certain dose-response, damage-assessment, and accelerated-life-testing studies, using Kalman-filter smoothing. Anatoli, Kenneth and James [32] indicated that various multivariate stochastic process models have been developed to represent human physiological aging and mortality. These researchers considered the effects of observed and unobserved state variables on the age trajectory of physiological parameters. The parameters of the distribution used were estimated based on an extension of the theory of Kalman filters to include systematic mortality selection. Ludwig [33] considered models for discrete time panel and survival data; and used a generalized linear Kalman filter approach.

In our study, Kalman filter technique is applied to estimate parameters of Weibull probability distribution using Diabetic and Non Diabetic CABG patient’s data sets. For construction of KF equations, survivor function of the probability distribution is linearized by transformation of double-log. The procedure to construct linear form of the survivor function, as advocated by researchers (see Gross and Clark [19], Kalbfleisch and Prentice [34] and Lawless [22], Meinhold and Singpurwalla [35]) is followed. Survival proportions for complete population of Diabetic CABG patients obtained from Kaplan Meier method are used as observed values \( y_t \) at time \( t \), for updating the time varying parameters of the distribution. After defining the updating system of parameters of a probability distribution with KF approach (discussed in the methodology), the parameters are estimated at each time \( t \) by maximizing likelihood function of double-lognormal distribution, through Davidon-Fletcher-Powel method of optimization [36]. Since, in KF approach the observed values are from complete population, therefore, censored part is excluded (dropped) from log-likelihood function. The survival proportions obtained by the pro-
2. Methodology

\[ S(t) = \prod_{j|t \in j} \left(1 - \frac{d_j}{n_j}\right) \]
where \( d_j \) and \( n_j \) are the number of items failed (died individuals/patients) and number of individuals at risk at time \( t_j \) respectively, that is, the number of individuals survived and uncensored at time \( t_j \). This method may be applied to both censored and uncensored data, see Lawless [22]. In case of censored individuals (items) the analysis is performed on its IP. Khan, Saleem & Mahmud [1] proposed that the censored individuals \( c_j \) may be taken into account. The inclusion of splitted-censored individuals’ \( d_j \) respectively make populations complete. Thus the survival analysis may be performed on the \( \Phi ') s at time \( t \), \( \Phi \) may be taken into account. The inclusion of splitted-censored individuals’ \( d_j \) respectively make populations complete. Thus the survival analysis may be performed on the CP from its IP. We apply Kaplan Meier method on CP of \( D_p \) and \( D_a \) groups of CABG patients to obtain survival proportions \( \Phi ') s and use as input in the DLM and KF equations/process. In this study the observed values (survival proportions) are denoted by \( Y_t \in (0,1) \), where \( Y_t \) may take value \( y_1, y_2, \ldots, y_t \) at time \( t, t_1, t_2, \ldots, t \). Harrison and Stevens [28] described the DLM which may be reproduced as system of following two equations:

Observation Equation:
\[ Y_t = F_t \theta_t + \epsilon_t, \quad \left\{ \epsilon_t \sim N(0, \sigma^2_{\epsilon}) \right\} \]  

System Equation:
\[ \theta_t = G \theta_{t-1} + W_t, \quad \left\{ W_t \sim N(0, W_t) \right\} \]  

where \( Y_t \) and \( \theta_t \) are of arbitrary dimensions. \( Y_t \) is a scalar, \( \theta_t \) is vector of process parameters at time \( t \), \( F_t \) is matrix of independent variables, known at time \( t \), \( G \) is known system matrix (identity matrix), \( \sigma^2_{\epsilon} \) is error term, a difference between observed and expected value \( y_t \) and \( \hat{y}_t \) respectively at time \( t \). \( \sigma^2_{\epsilon} \) is the variance of \( \epsilon_t \). It is assumed that \( \epsilon_t \) has Gaussian distribution with mean 0 and variance \( \sigma^2_{\epsilon} \). The system equation describes the change which occurs when process parameter changes from preceding value \( \theta_{t-1} \) to current value \( \theta_t \) and \( W_t \) is the variance of disturbance term \( \epsilon_t \). According to Harrison and Stevens (1976), it is assumed that distribution of the parameter vector \( \theta_t \) at time \( t = 0 \) i.e. \( \theta_0 \) prior to the first observation \( y_1 \) in the form of normal probability distribution with mean say \( m_0 \) and variance \( C_0 \) i.e. \( \theta_0 \sim N(m_0, C_0) \). If the observed values; \( y_1, t = 1, 2, 3, \ldots \) are described through DLM, then the posterior distribution of parameter vector \( \theta_t \) is also normally distributed with mean say \( m_t \) and variance \( C_t \) i.e. \( \theta_t \sim N(m_t, C_t) \). Whereas, the values of \( m_t \) and \( C_t \) are recursively obtained as:
\[ \hat{y}_t = F_t G m_{t-1}; \quad \epsilon_t = y_t - \hat{y}_t; \quad \nabla_t = G C_t G^t + W_t; \quad \sigma^2_{\epsilon} = F_t \nabla_t F_t^t + \sigma^2_{\epsilon}; \quad A_t = R_t F_t^t (\sigma^2_{\epsilon})^{-1}. \]  

The Kalman filter equations are: \( m_t = G m_{t-1} + A_t \epsilon_t \) and \( C_t = R_t - A_t \sigma^2_{\epsilon} A_t^t \) (for detail see Harrison and Stevens [28]). \( \sigma^2_{\epsilon} \) is variance of \( \epsilon_t \) and \( A_t \) is a matrix which update \( m_t \) and \( C_t \) at each time \( t \) recursively.

The KF equations of Weibull probability distribution models are constructed by linearizing survival function of the distribution with transformation; double-log. The parameters of the probability distributions are estimated at each time \( t \), by maximizing log-likelihood function of lognormal distribution (which is transformed form of Weibull distribution), through the Davidon-Fletcher-Powel method of optimization. For the entire system, the parametric space at each time point \( t \) is \( \Phi = (m_t, C_t, W_t) \). Specification of starting values of the parameters is a common difficulty in implementing Kalman Filter. Practitioners have to check the sensitivity of the final results with different sets of assumed values (see Meinhold and Singpurwalla, [31]). After obtaining the prior values of the parameters of the probability distributions at time \( t = 0 \), the values \( \Phi = (m_t, C_t, W_t) \) are obtained recursively by using the Kalman filter updating equations. The survival proportions for complete population of \( D_p \) and \( D_a \) CABG patients are used as observed values \( y_t \)’s at time \( t \), for updating the time varying
parameters of the distributions. Since, in the Kalman filter approach the observed values are from complete population, therefore, censored part is dropped from the log-likelihood function. To find maximum likelihood estimates we take negative log-likelihood function of each distribution along with KF process is developed in FORTRAN program. The subroutine in-conjunction with the DFP optimization method is used to find the optimal initial estimates of the mean and variance parameters included in the model, $m_0, C_0$ and $W_0$, from final iteration of the program. For outside sample period (forecasting), due to non-availability of dependent values $y_i$, we stop the process of updating the mean parameters. Therefore, values of these optimal mean parameters remain constant and are utilized for updating the variance parameters for outside sample period, using the KF equations. The survival proportions $Y_i$’s of these probability distributions are estimated.

3. Application (Construction of KF Equations of Weibull Distribution)

Since the values of survival proportions $Y_i$ (observed values) lies in the interval $(0, 1)$, expected value $E(Y_i)$ of a probability distribution should also lie in the interval $(0, 1)$. Keeping in view the natural process of deaths with the passage of time, it is assumed that $E(Y_i)$ as a function of $t$ is monotonically decreasing. These researchers, Meinhold and Singpurwala [33] considered a quantity $E(Y_i) = e^{-\alpha t^\beta}$ ($\alpha_i$ and $\beta_i > 0$) which is a nonlinear, monotonically decreasing function of $t$ and is survival function $S(t)$ of the Weibull distribution. Moreover, the form $e^{-\alpha t^\beta}$ (where $\alpha_i$ and $\beta_i$ are scale and shape time varying parameters respectively in KF approach) has property with respect to linearity; may be linearized by taking its double logarithm. The linear form is a requirement for filtering techniques. Thus to implement KF a random quantity $Y_i^* = \ln (E(Y_i))$ is defined, which require that $Y_i^*$ has a Gaussian density with expectation $\mu_i$ and variance $\sigma_i^2$. This implies that the random quantity $Y_i^*$ must have double-lognormal distribution with $pdf$ at $y_i$ of the form:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}y_i(-\ln y_i)} \exp\left(-\frac{1}{2\sigma_i^2}(-\ln y_i - \mu_i)^2\right).$$

Now, $E(Y_i^*) = \ln (-\ln (S(t))) = \ln (-\ln (e^{-\alpha t^\beta})) = \beta_i \ln t + \ln \alpha_i = [1 \ln t] \ln \alpha_i \beta_i = [1 \ln t] \ln \alpha_i \beta_i$,

setting $\gamma_i = \ln \alpha_i$

$$Y_i^* = [1 \ln t] \ln \alpha_i \beta_i + e_i; \ e_i \sim N(0,\sigma_i^2)$$

(3)

The corresponding system equation is:

$$[\gamma_i \beta_i] = [\gamma_{i-1} \beta_{i-1}] + w_i; \ w_i \sim N(0,W_i)$$

(4)

Comparing Equations (3) and (4) with (1) and (2), we find that:

$$F_i = [1 \ln t],$$

$$\theta_i = [\gamma_i \beta_i] ; \theta_{i-1} = [\gamma_{i-1} \beta_{i-1}] \text{ and } G = I_2 \text{ (identity matrix)}.$$ To find maximum likelihood estimates we consider negative log-likelihood function say $l(\Phi)$ of the double-lognormal distribution, given as:

$$l(\Phi) = \sum l_i \ , \ l_i = f_{\delta_i} \ln f(y_i)$$

where, $y_i$ and $f_{\delta_i}$ are observed values from CP and $f_{\delta_i}$ number of failures at time $t_i$ respectively and $l_i$ may be obtained as: $l_i = f_{\delta_i} \left[ \ln \frac{1}{2\pi y_i(-\ln y_i - a_i + (\delta_i)^2/2)} \right]$.

For derivation of $l_i$ and its partial derivatives, see Appendix A.

A subroutine for maximizing log-likelihood function of the double-lognormal distribution along with KF process (subroutine) is developed in FORTRAN program.
The subroutine in-conjunction with DFP optimization method is used to find the optimal initial estimates of the parameters included in the model \( m_0, C_0 \) and \( W_0 \), from final iteration of the program.

The optimal initial estimates of parameters obtained by maximizing the log-likelihood function are presented in Table 1.

The results (survival proportions obtained by using Weibull distribution and KF approach \( (\hat{y}_{t_{KF}}) \) at each time point \( t \) as explained earlier) of \( D_a \) (Diabetic Absent) and \( D_p \) (Diabetic Present) groups of CABG patients are presented in Table 2 and Table 3 respectively.

Table 1. The estimates of parameters of Weibull distribution and KF using data of \( D_a \) and \( D_p \) groups of CABG patients.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>( D_a ) Group</th>
<th>( D_p ) Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_0 )</td>
<td>0.2093 \times 10^{-4}</td>
<td>0.8383 \times 10^{-5}</td>
</tr>
<tr>
<td>( \beta_0 )</td>
<td>1.2999</td>
<td>1.0546</td>
</tr>
<tr>
<td>( \sigma_a^2 )</td>
<td>1.2197</td>
<td>1.9704</td>
</tr>
<tr>
<td>( C_a )</td>
<td>\begin{pmatrix} 1.348 \times 10^{-4} &amp; 0 \ 0 &amp; 8.76 \times 10^{-4} \end{pmatrix}</td>
<td>\begin{pmatrix} 1.543 \times 10^{-4} &amp; 0 \ 0 &amp; 2.95 \times 10^{-4} \end{pmatrix}</td>
</tr>
<tr>
<td>( W_0 )</td>
<td>\begin{pmatrix} 1.8183 \times 10^{-8} &amp; 0 \ 0 &amp; 7.6737 \times 10^{-8} \end{pmatrix}</td>
<td>\begin{pmatrix} 2.38 \times 10^{-4} &amp; 0 \ 0 &amp; 8.75 \times 10^{-4} \end{pmatrix}</td>
</tr>
<tr>
<td>Value of Log-Likelihood</td>
<td>161.4857</td>
<td>84.236929169</td>
</tr>
</tbody>
</table>

Table 2. Survival proportions \( y_t \) of 12 years estimated and 3 years forecasted of CP (complete population) of \( D_a \) (diabetic absent group) of CABG patients obtained by Kalman Filter approach.

<table>
<thead>
<tr>
<th>Years (t)</th>
<th>( y_t )</th>
<th>( (\hat{y}<em>{t</em>{KF}}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0.90</td>
<td>0.979</td>
</tr>
<tr>
<td>2</td>
<td>0.87</td>
<td>0.949</td>
</tr>
<tr>
<td>3</td>
<td>0.85</td>
<td>0.916</td>
</tr>
<tr>
<td>4</td>
<td>0.83</td>
<td>0.880</td>
</tr>
<tr>
<td>5</td>
<td>0.79</td>
<td>0.843</td>
</tr>
<tr>
<td>6</td>
<td>0.77</td>
<td>0.805</td>
</tr>
<tr>
<td>7</td>
<td>0.73</td>
<td>0.767</td>
</tr>
<tr>
<td>8</td>
<td>0.69</td>
<td>0.729</td>
</tr>
<tr>
<td>9</td>
<td>0.65</td>
<td>0.692</td>
</tr>
<tr>
<td>10</td>
<td>0.61</td>
<td>0.656</td>
</tr>
<tr>
<td>11</td>
<td>0.56</td>
<td>0.620</td>
</tr>
<tr>
<td>12</td>
<td>0.53</td>
<td>0.586</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>0.552</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>0.520</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>0.489</td>
</tr>
</tbody>
</table>
4. Conclusion

The graphs of observed survival proportions from the complete population $y_i$ and expected survival proportions $(\hat{y}_i)_{KF}$ of $D_a$ and $D_p$ groups of CABG patients (Figure 1 & Figure 2) indicate that the behavior of $(\hat{y}_i)_{KF}$ from $D_a$ group is like linear throughout the sample period, whereas $(\hat{y}_i)_{KF}$ of the $D_p$ group is almost linear for the first 7 values and curved for the rest of values; due to more noises, however it remains around $y_i$ of the $D_p$. This reflects that the complete population (forecasting) data has been modeled adequately. Kalman Filter smoothing approach is appropriate and forecast of $D_a$ and $D_p$ groups of CABG patients is reliable outside the sample observations.

Table 3. Survival proportions $y_i$ of 12 years estimated and 3 years forecasted of CP (complete population) of $D_p$ (diabetic present group) of CABG patients obtained by Kalman filter approach $(\hat{y}_i)_{KF}$.

<table>
<thead>
<tr>
<th>Years ($t$)</th>
<th>$y_i$</th>
<th>$(\hat{y}<em>i)</em>{KF}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0.85</td>
<td>0.9200</td>
</tr>
<tr>
<td>2</td>
<td>0.82</td>
<td>0.8410</td>
</tr>
<tr>
<td>3</td>
<td>0.79</td>
<td>0.7667</td>
</tr>
<tr>
<td>4</td>
<td>0.73</td>
<td>0.6978</td>
</tr>
<tr>
<td>5</td>
<td>0.66</td>
<td>0.6345</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>0.5760</td>
</tr>
<tr>
<td>7</td>
<td>0.53</td>
<td>0.5225</td>
</tr>
<tr>
<td>8</td>
<td>0.51</td>
<td>0.4737</td>
</tr>
<tr>
<td>9</td>
<td>0.45</td>
<td>0.4291</td>
</tr>
<tr>
<td>10</td>
<td>0.36</td>
<td>0.3885</td>
</tr>
<tr>
<td>11</td>
<td>0.34</td>
<td>0.3515</td>
</tr>
<tr>
<td>12</td>
<td>0.23</td>
<td>0.3179</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>0.2873</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>0.2596</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>0.2345</td>
</tr>
</tbody>
</table>

Figure 1. Diabetes absent group.
Acknowledgements
We are thankful to our reviewers, whose constructive criticism has resulted in a clearer presentation of our work and inclusion of additional useful reference material.

References


Appendix-A
The Double-Lognormal Distribution

Consider p.d.f of log-normal distribution:

\[ f(y_i) = \frac{1}{\sqrt{2\pi \sigma_{\eta_i}}} \exp \left\{ -\frac{1}{2\sigma_{\eta_i}^2} \left( \ln y_i - \mu_{\eta_i} \right)^2 \right\}, \]

where \( \mu_{\eta_i} \) and \( \sigma_{\eta_i} \) are parameters of the distribution.

Let \( \frac{1}{\sigma_{\eta_i}} = \eta_i \) and \( a_i = \ln(\eta_i) \) or \( e^{a_i} = \eta_i \),

then,

\[ f(y_i) = \frac{e^{a_i}}{\sqrt{2\pi y_i}} \exp \left\{ \frac{\left[ \ln y_i - \mu_{\eta_i} e^{a_i} \right]^2}{2\sigma_{\eta_i}^2} \right\}. \]

Let \( \delta_i = \left[ \left( \ln y_i - \mu_{\eta_i} e^{a_i} \right) \right] \), we may write as:

\[ f(y_i) = \frac{e^{a_i}}{\sqrt{2\pi y_i}} \exp \left\{ \frac{\delta_i^2}{2} \right\}. \]

To find maximum likelihood estimates, we consider negative log-likelihood function say \( l(\Phi) \) of the double-lognormal distribution, given as:

\[ l(\Phi) = \sum l_i, \]

where \( l_i = f_i \ln(f(y_i)) \), by excluding the censored part since observed values \( y_i \) are from complete population, \( f_i \) are the number of failures (died) at time \( t_i \) and \( l_i \) may be obtained by replacing value of \( f(y_i) \). We get \( l_i \) as:

\[ l_i = f_i : \left[ \ln \sqrt{2\pi y_i} - a_i + \frac{\delta_i^2}{2} \right]. \]

For partial derivatives, differentiating above equation with respect to \( \delta_i \) and \( a_i \), we get

\[ \frac{\partial l_i}{\partial \delta_i} = f_i \delta_i \left( -e^{a_i} \right) \]

and

\[ \frac{\partial l_i}{\partial a_i} = -f_i \left( -1 + \delta_i^2 \right). \]