Parametric Models in Survival Analysis: Application in a Cohort of Older People in a Medium Size City of São Paulo – Brazil

Keywords: Older people; Survival time; Cohort

Abstract

Population ageing, Brazilian demographic facts in recent decades, became of the utmost importance for the social planning that emerges in this century with this new profile. Differences in survival in observed cohorts of older people according to various factors are referred in the literature. Then, the aim of this work was to develop more general parametric models to estimate survival time for grouped and censored data including covariates and apply them to data from the cohort of older people, considering as outcomes the death from all causes and some specific. A population survey was carried out in Botucatu, São Paulo, Brazil, in the years of 2003, 2006 and 2010, in order to evaluate the survival time of these people. The sample size was calculated in 365 older people considering the quality of life as unknown (50%). 5% of margin error and 95% of reliability, it was observed the risks were not proportional. Then these new models (Generalized Log-Normal and Generalized Weibull) were used to estimate the survival time rather than Cox model that could not be appropriate for these data. Including covariates, it was found the age more than 60 years, diabetes mellitus, another income and cardiovascular disease risk factor for survival. Already, social relationship was protective factor for survival, when including some aspects of quality of life. As a conclusion, the use of these models has become an interesting alternative to the Cox proportional hazard model and they seem more realistic with the factors involving survival time for older people.

Introduction

The global demographic transition, which began in the late 19th century and characterized by increasing the proportion of older people in the population [1,2], brought with it the concern with the factors involving survival time for older people. Another concern is the one with the factors involving survival time for older people. The association between lifestyle and cardiovascular mortality and cancer [6].

Models for survival data analysis has often been used as the known Kaplan-Meier method [7] and the Cox proportional hazard model [8]. Kaplan-Meier method estimates the survival function but does not allow the inclusion of covariates and provides only comparison between curves according to the outcomes of interest. Already, Cox proportional hazards model allow including covariates and provides the risks of them in the model, but these risks must be proportional hypothesis, which is not always verified.

Models for grouped and censored data

When we have data for survival time, it has been for grouped and censored data techniques.

Let the lifetime be \( T_i, i = 1, \ldots, n \) where \( n \) is the sample size, grouped in \( k \) intervals

\[ I_i = [a_{i-1}, a_i), l = 1, \ldots, k \text{ with } 0 = a_0 < a_1 < \ldots < a_k = \infty \]

and suppose all the censors occur at the end of the intervals. Let \( D \) be a set of individuals that failure in the interval \( I_l \), which is the set of individuals under risk at the beginning of \( I_l \), and \( \Lambda_l \), the indicator function of failure of \( l \)th individual in \( I_l \) that is zero when the \( l \)th observation is censored in the \( l \)th interval and one otherwise, \( l = 1, \ldots, k, i = 1, \ldots, n \). Then, the likelihood function can be constructed for the life table with the regressor variable \( (x) \), given by:

\[
L = \prod_{i=1}^{n} \prod_{j=1}^{\Lambda_i} p_i(x_i) \prod_{j=k+1}^{k} (1 - p_i(x_i)) \prod_{j=k+1}^{\Lambda_i} \left[ 1 - p_i(x_i) \right]^{1-\Lambda_i}, \tag{1}
\]

Where \( p_i(x_i) \) is the probability of the \( i \)th individual failure until \( a_i \).

given it survives until \( a_i \cdot 1 \) in the presence of the regressor variables \( x_i \).

The function (1) matches with the likelihood function for variable with Bernoulli distribution once \( \Delta_i \) is binary (that is, failure occurred or not in the interval \( I \)).

The probability \( p(x) \) can be modelled considering a hazard proportional Cox model, logistic and generalized log-normal and Weibull models.

**Cox proportional hazard model:** Adopting the Cox proportional hazard model for the lifetime, the hazard function for \( T \) given \( x_i \) is:

\[
h(t | x) = h_0(t) \exp(\beta x)
\]

and the survival function is:

\[
S(t | x) = \exp\left(\int_0^t h(u | x) \, du\right) = [S_0(t)]^{\exp(\beta x)}
\]

being \( S_0(t) \) the baseline survival function. Modelling \( p(x_i) \) by (2), we have:

\[
1 - p(x_i) = \frac{S(a_i | x_i)}{S(a_i | x_i)}
\]

or,

\[
p(x_i) = 1 - \left[ \frac{S(a_i)}{S(a_i)} \right]^{\exp(\beta x_i)}
\]

Replacing \( p(x_i) \), obtained from (3) in (1), the likelihood function for life table data with regressor variables \( (x_i) \) will be given by:

\[
\Lambda = \prod_{i=1, k=0}^{\Delta_i} \left[ 1 - \left( \frac{\sum_{a_i} S(a_i)}{\sum_{a_i} S(a_i)} \right) \exp(\beta x_i) \right]^{\Delta_i} \left[ \left( \frac{\sum_{a_i} S(a_i)}{\sum_{a_i} S(a_i)} \right) \exp(\beta x_i) \right]^{1-\Delta_i}
\]

Parameters are estimated using a maximum likelihood method, maximizing the function:

\[
\log L(\beta, \gamma) = \sum_{i=1}^k \left[ 1 - \Delta_i \exp(\beta x_i) \right] + \Delta_i \ln(1 - \exp(1 - \exp(\beta x_i)))
\]

Where \( \gamma = \log(-\log(Q_i)) \) is a reparametrized form and \( Q_i = \frac{S(a_i)}{S(a_i)}, i = 1, \ldots, k \).

**Logistic model:** Adopting the Logistic model, we have:

\[
p_i(x_i) = 1 - (1 + \exp(-\exp(\beta x_i)))^{-1}
\]

where \( \gamma = \frac{p_i(0)}{1 - p_i(0)} \), \( i = 1, \ldots, k \)

Then, the log-likelihood can be written:

\[
\log L(\beta, \gamma) = \sum_{i=1}^k \left[ \Delta_i (a_i + \beta x_i) - \ln(1 + \exp(\exp(\beta x_i))) \right]
\]

where \( a_i \), \( \gamma \) is a reparametrization to improve the normal approximation [10]. Parameters are estimated maximizing (5) using an iterative process.

**Generalized log-normal model:** The Generalized Log-Normal Family (GLN) is a family of distribution with three parameters [11]. It includes the log-normal distribution as a special case. Essentially, the GLN distribution is associated to the Box-Cox transformation. The general definition is the result of how to include the tails of the probability distribution that are generally neglected in the major studies about Box-Cox transformation Box-Cox [12].

Let \( \Phi \) be a standard normal distribution function. For \( t > 0 \) and for any \( \lambda \), we define:

\[
l_t(\lambda) = \begin{cases} 
\frac{t^{\lambda-1}}{\lambda}, & \text{if } \lambda \neq 0 \\
\log(t), & \text{if } \lambda = 0
\end{cases}
\]

Let be \( \Theta = \left( \lambda, \mu, \sigma^2 \right) \), \( \lambda \in R, \sigma > 0 \). Then, for some \( \theta \in \Theta \), and for \( t > 0 \) the Generalized Lognormal probability density function is defined as:

\[
f(t; \lambda, \mu, \sigma^2) = \begin{cases} 
\frac{\Phi\left(l_t(\lambda) \right)}{\sigma} \frac{\sigma}{\Phi(\delta)}, & \text{if } \lambda > 0 \\
\frac{\Phi\left(l_t(\lambda) \right)}{\sigma} \frac{\sigma}{\Phi(\delta)}, & \text{if } \lambda = 0 \\
\frac{\Phi\left(l_t(\lambda) \right)}{\sigma}, & \text{if } \lambda < 0
\end{cases}
\]

where \( \delta = \frac{(\mu + 1/\lambda)}{\sigma} \).

When \( \lambda = 0 \), this family of distributions is a particular case of the Log-Normal distribution. Also, if \( T \sim \mathcal{F}(t; \lambda, \mu, \sigma^2) \), the cumulative distribution for \( T \) is given by:

\[
F(t; \lambda, \mu, \sigma^2) = \frac{\Phi\left(l_t(\lambda) \right) + \Phi\left(\Phi^{-1}(\sigma) - \Phi^{-1}(\delta)\right)}{\Phi(\delta)}, \quad \text{if } \lambda > 0
\]

\[
F(t; \lambda, \mu, \sigma^2) = \frac{\Phi\left(l_t(\lambda) \right)}{\Phi(\delta)}, \quad \text{if } \lambda = 0
\]

\[
F(t; \lambda, \mu, \sigma^2) = \frac{\Phi\left(l_t(\lambda) \right)}{\Phi(\delta)}, \quad \text{if } \lambda < 0
\]

Adopting the model based on the generalized log-normal distribution for \( \lambda > 0 \), we have that \( p_i(x_i) \) is given by:

\[
p_i(x_i) = \frac{\Phi(\delta(a_i)) - \Phi(\delta(a_i))}{1 - \Phi(\delta(a_i))}
\]

where \( \Phi \) is the cumulative function of the standard normal distribution, \( \lambda \) the Box-Cox transformation parameter, \( \delta(a_i) = a_i(1 - \beta x_i) \), \( a_i(\lambda) = a_i^\lambda - 1 \) and \( \beta' = (\beta_0, \beta_1) \).

The parameter estimation is made using a maximum likelihood method and the likelihood function, the first and second order derivatives and the routine developed in SAS language [12].

**Generalized Weibull model (GW):** Let \( T \) be a random variable with Weibull distribution. The probability density function is given by:

\[
f(t; \alpha, \gamma) = \frac{\gamma}{\alpha} \frac{t^{\gamma-1}}{\exp\left(-\left(\frac{t}{\alpha}\right)\right)^\gamma}, \quad t \geq 0
\]
where $\gamma$ is the shape and $\alpha$ is the scale parameters, both positives.

The failure rate function $h(t)$ of the Weibull distribution is strictly increasing for $\gamma > 1$, strictly decreasing for $\gamma < 1$ and constant for $\gamma = 1$. When $\gamma = 1$, we have the failure rate of the exponential distribution which is a particular case of the Weibull distribution.

The Generalized Weibull family distribution is given, using quantile function, by:

$$
F(t;\alpha, \lambda, \sigma) = 1 - \left[ \frac{1 - \left( \frac{t}{\sigma} \right)^{1/\lambda} }{\lambda} \right]^\alpha,
$$

where $\alpha, \sigma > 0$ and $\lambda$ are real number.

Inverting the quantile function given by (11), the cumulative distribution function of the Generalized Weibull is:

$$
F(t;\alpha, \lambda, \sigma) = \left[ 1 - \frac{1 - \left( \frac{t}{\sigma} \right)^{1/\lambda} }{\lambda} \right]^\alpha
$$

which is a particular case of the Weibull distribution. The survival function and failure rate are, respectively:

$$
S(t;\alpha, \lambda, \sigma) = 1 - \left[ 1 - \lambda \left( \frac{t}{\sigma} \right)^{1/\lambda} \right]^\alpha
$$

$$
h(t;\alpha, \lambda, \sigma) = \alpha \sigma \left[ 1 - \lambda \left( \frac{t}{\sigma} \right)^{1/\lambda} \right]^{\alpha - 1}
$$

AIC values for the fitted models.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Cox</th>
<th>Generalized Weibull</th>
<th>Generalized Log-normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>929.39</td>
<td>397.16</td>
<td>441.25</td>
</tr>
<tr>
<td>Male</td>
<td>922.39</td>
<td>418.26</td>
<td>446.68</td>
</tr>
</tbody>
</table>

Table 2: Generalized Weibull fit for survival data with covariates for female.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age &gt; 69</td>
<td>0.5438</td>
<td>0.1363</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Presence of diabetes</td>
<td>0.4870</td>
<td>0.1633</td>
<td>0.0029</td>
</tr>
<tr>
<td>Have another income</td>
<td>0.3735</td>
<td>0.1930</td>
<td>0.0530</td>
</tr>
<tr>
<td>Presence of circulatory system disease</td>
<td>1.2955</td>
<td>0.1499</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Table 3: Generalized Weibull fit for survival data with covariates for male.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age &gt; 69</td>
<td>0.5886</td>
<td>0.1588</td>
<td>0.0002</td>
</tr>
<tr>
<td>Presence of diabetes</td>
<td>1.0055</td>
<td>0.3008</td>
<td>0.0008</td>
</tr>
<tr>
<td>Have another income</td>
<td>0.4798</td>
<td>0.2154</td>
<td>0.0259</td>
</tr>
<tr>
<td>Presence of circulatory system disease</td>
<td>1.5823</td>
<td>0.1904</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

An interesting characteristic of the Generalized Weibull family is the hazard function assumes four different forms in the four regions of the plane [13], divided by the lines $\alpha = 1$ and $\lambda = 0$.

Adopting a model based on the Generalized Weibull [13] and the likelihood function given by (1), we have:

$$
p_i(x) = \left[ 1 - \lambda \left( \frac{a_i - \beta' x_i}{\sigma} \right)^{1/\lambda} \right]^{\gamma - 1}
$$

where $\gamma, \alpha > 0, \lambda \in R$ and $\beta' = (\beta_0, \beta_1)$.

The implementation of the presented functions was made in SAS language [12].

Therefore, the aim of this paper was to apply these more general distribution families to a survival data for a cohort of older people from Botucatu city, Sao Paulo, Brazil, considering the mortality for all causes as the outcome. These models can be an interesting alternative to the Cox proportional hazard model and may result in better fittings.

**Methodology**

A survey was carried out in the Botucatu city, Sao Paulo, Brazil, firstly at the year of 2003, considered the baseline of the cohort and the main objective was to evaluate the quality of life of the older people. At that time, the population of the city was estimated in 110000 inhabitants where 12141 were older people with 60 years and over [14], representing 11% of the population.

For the baseline, the sample size was estimated considering 5% of error margin, 95% of reliability and the prevalence of quality of life in 50% (unknown), totaling 384 people.

It was a systematic sample in double stage after an update of the city map registered at the Census 2000. Nine thousands families were randomly selected using a 4:1 scale (26% of the households considering 1% of adjust) [14]. Next, 384 older people were randomly selected from these families. The inclusion criteria was to be capable to answer the questionnaire and excluded the ones that refused to participate, the ones with some cognitive impairment, dementia or any factor that could interfere in the interview.

The questionnaire was a fusion of several validated ones but the base was the Flanagan Quality of Life Scale (FQOLS) [15]. Adaptations were made in a pilot study in 2003, once some of elderly were illiterate, some presented difficulties to interpret the questions and problems with decreased visual acuity. Then, according to the pilot study, it was decided to use the “Motivation Scale” [16]. This scale consists of seven faces designed with the following expressions: a neutral face three motivated and three unmotivated. The size of the figure was adequate to the visual acuity of respondents to include all
randomly selected.

Seven years later (2010), the survival of the older people sample was studied considering the death of these people and their causes using a Mortality Information System (MIS) from Ministry of Health by means the municipal health secretary. One by one were searched in the system and, next, all households were revisited asking about the elder that lived there and had took part in the survey. All the death were confirmed by the death certificate from the system and from the family, in case of a not notified case. Elders not located were tracked using information from the neighborhood or from the closed trade next the old residence.

Data obtained in 2010 were analyzed using the survival techniques presented here considering time of death for any cause as an outcome.

The explanatory variables, obtained in 2003, were: sex, age and the score of quality of life measured by the FQOLS (transformed in binary variable using a median score as a cut off), cardiovascular disease, hypertension, lung disease and diabetes mellitus.

With the survival data, firstly the survival curve was estimated using a Kaplan-Meier method and tested for each explanatory variable using a log-rank test. Next, Cox model was used to evaluate the risk of death. The model was tested including hierarchical blocks of variables, beginning with the demographic, comorbidities and quality of life (binary) and another model including the five domains of the FQOLS: welfare: material comfort, familial relationship, social relationship, develop new abilities. The generalized log-normal and generalized Weibull models were also fitted with these variables. It were used Statistical Package of the Social Sciences (SPSS) for windows, v. 10.0, and Statistical Analysis System (SAS) for windows, v.9.2 softwares to analyze the data.

The use of this data set was authorized for the Ethical Committee for Research from the Botucatu Medical School (Process no. CAAE 14348113.0.0000.5411).

**Results**

From the 365 older people interviewed in 2003, 28 (8%) were not found after seven years and 71% were alive. Related to the sociodemographic data, in the year of 2003 there were a prevalence of female in the population (60%), most of them with more than 70 years old (61.8%). Quality of life self-reported measured by the FQOLS, the median was 75 points varying from 31 to 97 (maximum = 105 points), 25% referred cardiovascular disease, 45% hypertension, 15% diabetes mellitus and 7% lung disease.

Cox, Generalized Weibull and Generalized Log-normal models were applied to the survival data set and the comparisons were made using a Akaike Information Criteria (AIC). The analysis were made separately for gender, where the behaviour are different related to the survival (Table 1).

Comparing the AIC values, the generalized Weibull model was better considering each gender. Cox model has an assumption of proportional hazards and, it was observed in this data set that the major part of the covariates did not present proportional hazard, being this the motive of the worst fit.

The significant covariates in the generalized Weibull model were: age greater than 69 years old \((p < 0.0001)\), presence of diabetes \((p = 0.0029)\) and have circulatory system disease \((p < 0.0001)\) for female. For male, we found significance in age greater than 69 years old \((p < 0.0001)\), presence of diabetes \((p = 0.0029)\) and have circulatory system disease \((p < 0.0001)\) and have another source of income \((p = 0.0259)\) (Tables 2 and 3).

Next, variables referred about quality of life were included in the model using the domains of the FQOLS. In this case generalized log-normal model fitted better than others \((AIC = 550.56)\). Results are showed in Table 4.

Table 4 shows the variable age > 69 years old and the presence of cardiovascular disease are risk factor for survival. Social relations domain was a protection factor for survival, meaning that people who have good social relations have more chance to survival for more time.

---

**Table 4:** Generalized log-normal for survival data including the domains of quality of life.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age &gt; 69</td>
<td>0.4863</td>
<td>0.2068</td>
<td>0.0187</td>
</tr>
<tr>
<td>Presence of cardiovascular disease</td>
<td>0.5206</td>
<td>0.2265</td>
<td>0.0216</td>
</tr>
<tr>
<td>Social relationship domain</td>
<td>-0.398</td>
<td>0.20374</td>
<td>0.0259</td>
</tr>
</tbody>
</table>
Discussion

This study proposed to evaluate grouped and censored survival time for older people taking account some aspects of quality of life measured by the FQOLS and some morbidities more prevalent in this age using different approaches and fitting more general distributions comparing with the Cox model. As we can observe, the importance of develop new approaches in order to analyze grouped and censored survival data relay in the assumptions, once the considered data do not present proportional hazard. In this way, Cox model could be not appropriate for this sort of data.

The most significant for maintaining quality of life self reported for the older people from Botucatu, Sao Paulo, Brazil, as a protector factor of mortality was the domain of social relations. However, differently from cross-sectional studies, it is difficult to define the direction of causality. The prospective follow-up of the elderly could examine the relationship between baseline characteristics and its subsequent amendments, determining the precision and the magnitude of the influence of the variables in the outcome, which shows a very high score and therefore very positive for the quality of life of the elderly in this study.

The results of this study indicate that the lower survival of the population was related to older age and have age greater than 70 years old, presence of diabetes mellitus, have another income and cardiovascular disease both for men as for women. A study on the main causes of death in Sao Paulo showed that the largest proportion of preventable deaths was between heart disease (29.7%), which also had the highest mortality rate, 81.0 and 47.4 deaths per 10,000 elders male and female, respectively [17]. Therefore, it is important to reflect on prevention to minimize comorbidity of heart disease in this population.

Unlike of these findings, it was found that cognitive impairment (HR = 4.30), stroke (HR = 3.49), and dependence for basic activities of daily living (HR = 3.17) were the main risk factors for survival in older adults with functional limitations [18]. Also, diagnosis of dementia (p = 0.018) and depression (p = 0.007) as a statistically significant reduced in survival when the elderly suffer a hip fracture [19].

It was found with the fitted models that women had higher survival rate from men. Although women live longer, they do not enjoy better health, they tend to mention greater functional difficulties than men (in 2003, in Brazil, the prevalence of difficulty walking 100 m was 17.6% for older men and 26.6% for women) [20].

Related to the quality of life, a survey data from Avaré, São Paulo, Brazil, claim that the quality of life of the older adults was considered good, although they presented limitations [21]. The main factors associated with life satisfaction in general were the financial situation, comfort with the place where they live and even the lack of opportunity for leisure activities. Another study carried out in Botucatu city showed that the sequence of the domains for quality of life was the same the one proposed by the FQOLS, that is, physical and material well-being, relation with others, social activities, personal development and recreation [22]. As we can observe, social relationship is a protector factor for survival according to our findings.

Also in Botucatu, an analysis of quality of life of older people according to their own opinion, revealed, among other findings, a group who valued pleasure and material comfort [23]. In this same group, it was found factors associated with quality of life for the elderly were: age, marital status and the reference heart problems and that the high degree of satisfaction indicated that this step life where they have the opportunity to live according to their own values [24].

Quality of life self reported was assessed positively by the older people in this study and may be considered a predictor of survival time. The well-being and material comfort is the domain of Flanagan instrument that better explain this last result.

This study may help to propose actions for the older people within the municipal health policy. It is suggested initiatives to develop the welfare and comfort by providing incentives in the municipal infrastructure as well as the practice of physical activities.

This would benefit the development of the psychological, social and cognitive aspects in addition to improve performance in activities of daily living and enthusiasm for life, important indicators of quality of life and health promotion that reflects in an independent and successful ageing.

Conclusions

We can conclude that the proposed models could reveal important aspects for the survival time for older people considering covariates as demographic information, quality of life and more prevalent morbidities. It was found the risk factors for survival were age, diabetes mellitus, another income and cardiovascular disease. Already, social relationships were a protector factor for survival.

Also, the proposed models can be an alternative suggestion for Cox model in case on non proportional risks. Another study can be proposed in order to make some simulation study and more comparisons among them.

References


Acknowledgements

The authors would like to thank to FAPESP (Process no. 2012/03599-8) for the financial support to develop this research.