American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

http://asrjetsjournal.org/

Prediction of Life Expectancy

Nilushi Dias[a]*, Charith Sucharitharathna[b]

[a,b]*Sri Lanka Institute of Information Technology, Colombo, Sri Lanka*
[a]*Email: nilushi.d@sliit.lk*
[b]*Email: charith.s@sliit.lk*

Abstract

Life expectancy refers to the number of years a person is expected to live. The life expectancy for a particular person or population group depends on several variables such as their lifestyle, access to healthcare, diet, economical status and the relevant mortality and morbidity data. To make predictions on lifetime of a person, four independent variables were collected from death certificates and they are sex, cause of death, profession and race. During the study, impact of the four factors on life expectancy is measured. Independent samples t test and Kruskal wallis test were used to examine the independent variables. Except for the race, Kruskal wallis test indicated that age is not same across different categories of sex, cause of death and profession. But independent Sample's t test indicated that there was no significant difference between males and females and also it resulted in the fact that, there was no evidence to say that there is a difference between the lifetimes of cancer patients and languishing people. However, Kruskal wallis test and independent sample's t test gave contradictory results. After analyzing the variables, to make predictions, General Linear Models (GLM) and Kaplan Meier estimates were used. According to the results obtained from General Linear Models, sex and cause of death were statistically significant in the model. Being a non parametric test Kruskal wallis test always aligned with the model fitting results.

Keywords: General Linear Models; Kaplan Meier Estimates; Kruskal Wallis Test; Life Expectancy.

1. Introduction

Life expectancy is the most commonly used measure to describe population health and reflects the overall mortality level of a population. Life expectancy measures how long, on average, a person is expected to live [1]. In summarizing mortality patterns, life expectancy is often expressed as the number of years of life a person born today is expected to live.

---

* Corresponding author.

The life expectancy always depends on several variables such as their lifestyle, access to healthcare, diet, economical status and the relevant mortality and morbidity data. However, as life expectancy is calculated based on averages, a person may live for many years more or less than expected. The concept of life expectancy is also applied in ecological studies. Whether life expectancy is being calculated for plants, animals or humans, tables referred to as actuarial tables or mortality tables are used. Taking humans as an example, these tables can predict how likely it is that a person of a given age will die before their next birthday [2].

From here, several points can be calculated, including:

- The person's probability of surviving to any given age
- The life expectancy remaining for people of various ages

2. Research Elaborations

Doing a research in life expectancy domain is a challenging task, due to the difficulty of collecting data. If the research is done with the real data, then life time of a person should be collected with the independent variables. These independent variables include the factors which affect the life time of a person. The main problem in collecting data is, difficulty of collecting data When a person is alive it is not possible to predict the lifetime of that particular person. The other problem is, difficulty of collecting behavior of the person and the other important information once the person is died. Therefore most of the researches on life expectancy has been done using different other data, such as data collected from World Health Organization (WHO). Some of the researches have used economic factors such as GDP and values such as life expectancy at birth. Some of those previous researches are discussed below.

In [3] it has used statistical techniques to estimate the life time in cases of cerebral palsy. Comparison of life times of those who are strictly disabled with those who are not strictly disabled in different disability groupings is done. According to the analysis it has shown that, when the age increases, probability of surviving at that particular age has decreased.

In [4] it has estimated the effect of smoking and physical activity on active and disabled life expectancy using data from the Established Populations for Epidemiologic Studies of the Elderly (EPESE). Population-based samples of persons aged greater than or equal to 65 years from different areas were assessed at baseline between 1981 and 1983 and followed for mortality and disability over six yearly follow-ups. Active and disabled life expectancies were estimated using a Markov chain model. According to the results, nonsmokers survived 1.6-3.9 and 1.6-3.6 years longer than smokers, depending on level of physical motions. When smokers were disabled and close to death, most nonsmokers were still nondisabled. Higher physical activity was associated with fewer years of disability prior to death. These findings provide strong and explicit evidence that refraining from smoking and doing regular physical activity predict a long and healthy life.

In [5], they have studied educational differences in death and life expectancy among non-Hispanic blacks and whites in the 1980s and 1990s. Even though increased attention and dollars directed to groups with low socioeconomic status, within race and gender groups, the educational gap in life expectancy has risen, mainly

because of rising differentials among the elderly. With the exception of black males, all recent gains in life expectancy at age twenty-five have occurred among better-educated groups, raising educational differentials in life expectancy by 30 percent. Differential trends in smoking-related diseases explain at least 20 percent of this trend.

In [6], it examines healthy life expectancy by gender and education for whites and African Americans in the United States at three years: 1970, 1980 and 1990. They have found huge ethnic and educational differences in life expectancy at each year. Large racial differences exist in healthy life expectancy at lower levels of education. Educational differences in healthy life expectancy have been increasing over time because of widening differentials in both mortality and morbidity. In the last decade, a compression of morbidity has begun among those of higher educational status; those of lower status are still experiencing expansion of morbidity.

3. Results and Analysis

During the analysis, independent variables were analyzed individually and they were used to predict the life time of a person. In here, age specifies the life time of that person. Data was collected from death certificates and age, cause of death, sex, profession and race were collected. Death certificates were gathered from Colombo District, “Malabe” area.

Independent samples’ t test was carried out for Gender and the F value was not significant. The p-value is greater than 0.05 and therefore null hypothesis is rejected at 5% significance level. That indicates, mean life time of male persons is not significantly different from that of female persons. Then kruskal wallis test was also carried out in order to test the distribution of age across the categories of sex and under 5% significance level, null hypothesis is rejected and that indicated the distribution of age is not same across males and females. Therefore it indicated that sex has an effect on the life time of a person.

The analysis was carried out for individuals in the same area. Therefore most of the people were “Sinhala” and very few people were there in other races. According to the analysis, there are 90 respondents of Sinhala and very few from the other races. According to Kruskal wallis test, for race, under 5% significance level, null hypothesis is not rejected and that indicated the distribution of age is the same across different categories of race. Therefore it indicated that race has no effect on the life time of a person.

Cause of death of most of the people was languishing. Independent samples t test was carried out in order to compare the mean lifetimes of cancer affected individuals and mean life time of languishing people. Mean life time of cancer affected people have considerably lower than that of languishers. p value is 0.079 which is greater than 0.05. That indicates null hypothesis is not rejected at 5% significance level. According to Kruskal wallis test, for cause of death, under 5% significance level, null hypothesis is rejected and that indicated the distribution of age is not same across categories of cause of death. Therefore cause of death has no effect on the life time of a person.

According to Kruskal wallis test, for profession, under 5% significance level, null hypothesis is not rejected and that indicated the distribution of age is the same across categories of profession. Therefore job has no effect on

the life time of a person.

3.1. Kaplan Meier Estimate

The Kaplan Meier procedure is a method of estimating time-to-event models in the presence of censored cases. Censored cases are those for which the event of interest has not yet happened. This is a descriptive procedure for examining the distribution of time to event variables. Separate analysis can also be done to compare the distribution by levels of a factor variable.

Table 1: Mean and Median values of males and females

Sex	Mean[a]				Median			
	Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
Male	76.657	1.906	72.921	80.393	78.000	1.774	74.523	81.477
Female	81.613	1.414	78.842	84.384	82.000	1.432	79.194	84.806
Overall	79.825	1.155	77.560	82.089	82.000	.695	80.638	83.362

a. Estimation is limited to the largest survival time if it is censored.

Since there is an overlap between the confidence intervals, it is unlikely that there is much difference in the average survival times of males and females. If confidence intervals do not overlap between levels, differences in effect on time to event can be inferred.

Table 2: Test of equality of survival distributions for the different levels of Sex

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	6.500	1	.011
Breslow (Generalized Wilcoxon)	5.444	1	.020

Table 2 provides overall tests of the equality of survival times across groups. Since the significance values of the tests are all less than 0.05, it indicates that there is a statistically significant difference between males and females.

Figure 1 gives the survival curve of the life tables. The horizontal axis shows the time to event. Vertical axis shows the probability of survival. Therefore according to the plot, when the age gets longer probability of surviving gets decreased.

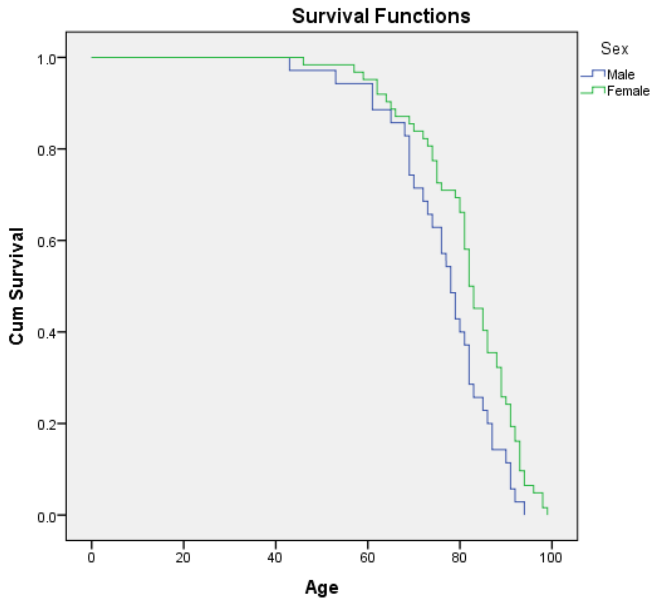

Figure 1: Survival Function for Age with Sex

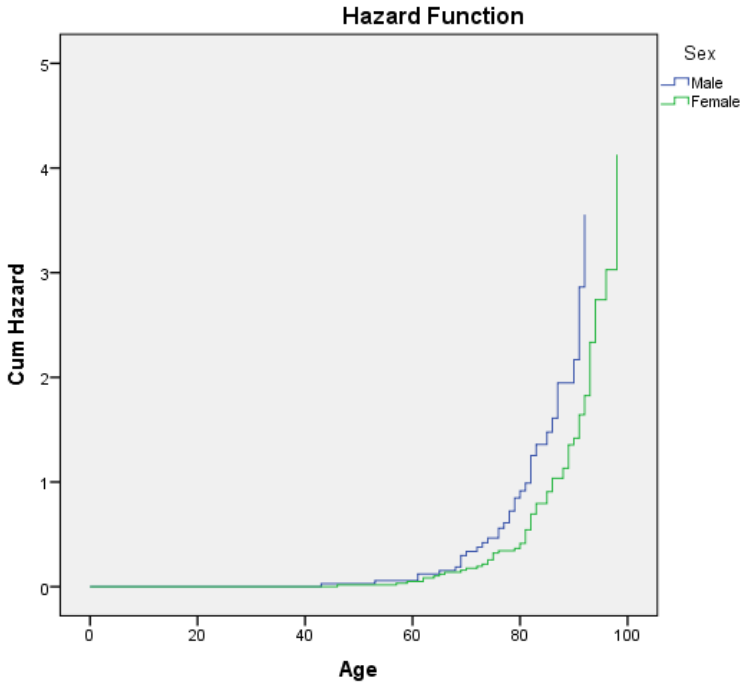

Figure 2: Hazard Function for Age with Sex

Figure 2 illustrates the hazard function and horizontal axis shows the time to event. Vertical axis shows the cumulative hazard, equal to the negative log of the survival probability. Cumulative hazard has increased with the age as can be seen from the above figure.

Kaplan Meier estimates were derived for cause of death and the resulting tables are given below.

Table 1: Means and Medians of Age with Cause of Death

CauseOfdeath	Mean[a]				Median			
	Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
KidneyProblem	83.000	5.495	72.229	93.771	87.000	11.635	64.195	109.805
Diabetes	84.667	1.509	81.709	87.625	85.000	2.981	79.156	90.844
Cancer	66.500	4.549	57.585	75.415	65.000	5.196	54.816	75.184
HeartProblem	76.500	5.195	66.319	86.681	82.000	11.635	59.195	104.805
Paralize	81.500	2.770	76.070	86.930	81.000	.866	79.303	82.697
Filaria	85.000	.000	85.000	85.000	85.000	.	.	.
Languishing	81.627	1.270	79.138	84.117	81.000	1.188	78.671	83.329
Overall	79.825	1.155	77.560	82.089	82.000	.695	80.638	83.362

a. Estimation is limited to the largest survival time if it is censored.

According to Table 3, confidence intervals for categories of cancer and heart problem lie in a very low value range. For example, for cancer, lower bound is 57.585 and for heart problem it is 66.319, which are comparatively lower than other categories. For the other categories, except the above 2, shows an overlap between the confidence intervals. Therefore it is unlikely that there is much difference in the average survival times of the other categories.

Table 2: Test of equality of survival distributions for the different levels of causeOfdeath

	Chi-Square	Df	Sig.
Log Rank (Mantel-Cox)	9.217	6	.162
Breslow (Generalized Wilcoxon)	22.653	6	.001

Table 4 provides overall tests of the equality of survival times across different categories of cause of death. Since the significance value of the Log rank test is less than 0.05, it indicates that there is a statistically significant difference between different categories of cause of death. But Breslow test has significance value less than 0.05. This test is used to test, the equality of survival functions by weighting all time points by the number of cases at risk at each time point. Therefore it gives evidence to say that, there is no statistically significant difference between different categories of cause of death.

3.2 Analysis with General Linear Model

The General Linear Model for univariate procedure allows to model the value of a dependent scale variable based on its relationship to categorical and scale predictors. When all the four categorical variables were used to fit a General Linear Model, all the terms were not statistically significant except sex and cause of death.

It indicates that, sex and cause of death are potentially important predictors of the dependent variable. Then all the terms were significant. Interaction term was not significant and the resulting table is as follows.

Table 5: Tests of between subjects effects

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	3173.476[a]	7	453.354	4.358	.000
Intercept	178132.851	1	178132.851	1712.345	.000
Sex	478.877	1	478.877	4.603	.035
CauseOfdeath	2624.051	6	437.342	4.204	.001
Error	9258.544	89	104.029		
Total	630515.000	97			
Corrected Total	12432.021	96			
a. R Squared = .255 (Adjusted R Squared = .197)					

Table 5 gives an analysis of variance table. Each term in the model, and the model as a whole, is tested for its ability to account for variation in the dependent variable.

Table 6: Parameter Estimates

Parameter	B	Std. Error	T	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	83.366	1.642	50.766	.000	80.104	86.629
[Sex=1]	-4.668	2.176	-2.146	.035	-8.991	-.345
[Sex=2]	0[a]	.	.	.	.	.
[CauseOfdeath=1]	1.189	4.403	.270	.788	-7.559	9.938
[CauseOfdeath=2]	3.375	3.691	.914	.363	-3.959	10.709
[CauseOfdeath=3]	-14.922	3.274	-4.558	.000	-21.427	-8.416
[CauseOfdeath=4]	-5.311	4.403	-1.206	.231	-14.059	3.438
[CauseOfdeath=5]	-.700	3.283	-.213	.832	-7.223	5.824
[CauseOfdeath=6]	1.634	10.331	.158	.875	-18.894	22.161
[CauseOfdeath=7]	0[a]	.	.	.	.	.
a. This parameter is set to zero because it is redundant.						

Table 6 gives the model coefficients in the first column "B". Therefore model can be derived as follows.

$$\begin{aligned}\text{Age} = 83.366 &- 4.668 * (\text{Sex} = 1) + 0 * (\text{Sex} = 0) + (\text{Cause of Death} = 1) * 1.189 \\ &+ (\text{Cause of Death} = 2) * 3.375 + (\text{Cause of Death} = 3) * -14.922 \\ &+ (\text{Cause of Death} = 4) * -5.311 + (\text{Cause of Death} = 5) * -0.7 \\ &+ (\text{Cause of Death} = 6) * 1.634 + (\text{Cause of Death} = 7) * 0\end{aligned} \quad \textit{(1)}$$

According to the above equation (1), when the cause of death is cancer, coefficient is -14.922. That indicates when a person is suffering from cancer, his/her lifetime will be reduced by 14.922 years. Further, model indicates that, when the heart problem is there, life time can again be reduced by 5 years, because coefficient is -5.333. Therefore this model can predict the life time of a person based on the gender and the disease they are having. The model is not developed for all the diseases, but it is developed for few diseases that people in this particular area are having. According to the model it indicates that cancer can cause reduction of lifetime by 14.9 years.

4. Conclusion

During the study, in order to predict the life expectancy, data was collected from death certificates. 4 variables were collected from death certificates and they were race, gender, profession and cause of death. They were analyzed individually with independent sample's t test, kruskals wallis test and using different graphical methods. According to the analysis of gender, Independent Sample's t test indicated that there was no significant difference between the lifetimes of males and females. But Kruskal wallis test indicated that age is not same across categories of sex. When the model is developed, Sex was significant and Kruskal wallis test also adhered to that result. The data set collected was, from an area where majority of the people are Sinhalese. Kruskal wallis test showed that, age is same across different categories of race. In the model also, race was not significant, which again adheres with the result of the Kruskals wallis test. Cause of death is considered as the third variable and in order to test whether cancer has a negative effect on life time of a person, Independent sample's t test was carried out. According to the test, average life time of cancer patients was lesser than that of languishing people. But under 5% significance level, there was no evidence to say that there is a difference between the lifetimes of cancer patients and languishing people. Kruskal wallis test also indicated that, age is not same across different categories of cause of death. In the model also, Cause of death was statistically significant. Profession of the person is analyzed with Kruskal Wallis test and indicated that, age is the same across different professions. In the model also, Profession was not significant and it again adheres with the Kruskal wallis test.

After analyzing all 4 variables, in order to do the prediction, General Linear Models (GLM) was used. The results indicated that, out of these four variables, sex and cause of death have significant influence on life time of a particular person. Univariate General Linear Model resulted in a model which can be used to predict the life time of a person.

5. Recommendations

Fitted model can be used to predict the lifetimes of people in this particular area. According to Gender and Cause of Death, predictions can be made about the life time.

References

[1] Australian Institute of Health and Welfare (2012). Australia's health 2012. Australia's health series no.13. [On-line]. Available:

http://www.aihw.gov.au/WorkArea/DownloadAsset.aspx?id=10737422169 [May 20, 2017]

[2] Dr. A. Mandal., "What is Life Expectancy". Internet: http://www.news-medical.net/health/What-is-Life-Expectancy.aspx, Oct 20, 2014 [May 21, 2017].

[3] Hutton J.L., Pharoah P.O., "Life expectancy in severe cerebral palsy". Arch Dis Child, 2006, 91: 254–258.

[4] L. Ferrucci, G. Izmirlian, S. Leveille, C. L. Phillips, M. C. Corti, D. B. Brock and J. M. Guralnik. "Smoking, Physical Activity and Active Life Expectancy". American Journal of Epidemiology, vol. 149, pp. 645-53, August 5, 1999.

[5] M. Ellen, S. Richards, and D. Cutler. "The Gap Gets Bigger: Changes in Mortality and Life Expectancy by Education, 1981−2000", Health affairs (Project Hope) 27.2 (2008): 350–360. PMC. Web. 19 June 2017.

[6] E. M. Crimmins and Y. Saito., "Trends in healthy life expectancy in US, 1970-1990: Gender, racial and ". Social Science and Medicine, 52(11), pp. 1629-1641, 2001.