Survival Analysis for Weaning Time of the Palestinian Children

Ali H. Abuzaid
Department of Mathematics
Al Azhar University – Gaza, Palestine

Raida F. Zaqout
Department of Applied Statistics
Al Azhar University – Gaza, Palestine

This study addresses the factors that have an effect on the weaning time of the Palestinian children based on the Palestinian family survey data in 2006. It was found that the Weibull parametric model is the most appropriate one to fit the data.

The study showed that factors such as child's weight at birth, child's age, mother's age at delivery, and mother's educational status have significant effects on the weaning time. The findings also revealed that factors such as mother's refugee status, locality type, total live births, and mother's smoking status do not have any significant effect at 0.05 level of significance on the duration of breastfeeding.

Keywords: *breastfeeding, censored, Cox proportional model, Wald statistic*

1. Introduction

The World Health Organization (WHO) recommends exclusive breastfeeding during the first six months of life for optimal growth, development and health (Kimani-Murage et al., 2011), when the child is more likely to get the health benefits of breastfeeding as his mother lengthens the weaning time (Bhimji and Hudda, 2005).

In the Palestinian Territories, the mean weaning time is 13 months and 62.8% of children started breastfeeding within the first hour of birth in 2015 (PCBS, 2015).

The percentage of children under six months who were exclusively breastfed was 38.6% in 2014, while 52.9% of children (12-15 months) were continuously breastfed for one year and 11.5% of children (20-23 months) were beastfed for two years. Furthermore, exclusive breastfeeding for the first six months has an inverse relationship with the level of mother's education (PCBS, 2015).

Based on a study conducted on the Palestinian family data survey in 2010, Zaid (2014) pointed out the importance of the awareness of breastfeeding for the mother and the child. It showed that the ratio of exclusive breastfeeding in Palestine between the years 2006 and 2010 remained the same.

After a continuous research, it has been shown that there is a scarcity of any published work on the factors affecting the weaning time in Palestine. Therefore, this paper aims to analyze the weaning time based on the data from the Survey of the Palestinian Family Health, 2006, to stand on the positive and negative factors that affects the lengthening or shortening of weaning time among Palestinian children.

The rest of the paper is organized as follows: Section 2 reviews some of the survival analysis statistical methods. Section 3 presents the data and its main characteristics. The survival analysis, which include the comparison between different groups, nonparametric and parametric modeling, is discussed in Section 4. Conclusions and recommendations are listed in Section 5.

2. Some Theories on Survival Analysis

This section discusses the theoretical side of the statistical methods which are used in the analysis of weaning time.

2.1 Kaplan-Meier survival curves

The Kaplan-Meier curves is the simplest way of comparing the survival over time, and it takes into account some types of censored data, particularly right-censoring, which occurs if a patient withdraws from a study.

The Kaplan–Meier estimator is the nonparametric maximum likelihood estimate of the survival function $S(t)$, where the maximum is taken over the set of all piece-wise constant survival curves with breakpoints at the event times t_i. It is a product of the form

$$\hat{S}(t) = \prod_{t_i \prec t} \frac{n_i - d_i}{n_i},$$

where n_i is the number of survivors just prior to time t_i when there is no censoring, and d_i is the number of survivors minus the number of losses for censored cases (Kaplan and Meier, 1958).

2.2 Log-rank test

It is the most commonly-used nonparametric statistical test for comparing the survival distributions of two or more groups (such as different treatment groups in a clinical trial). It is constructed by computing the observed and expected number of events in one of the groups at each observed event time, and then adding these to obtain an overall summary across all time points where there is an event (Shostak and Walker, 2010).

2.3 Common parametric models for survival data

Let T denote a continuous non-negative random variable representing survival time, with probability density function (pdf) and cumulative distribution function (cdf). Summarized below are some of the most important distributions which are popular in fitting the survival data:

2.3.1 Exponential distribution

The exponential is the simplest parametric duration model and it has constant hazard over time, where the survivor, hazard and probability density functions of exponential distribution are given by $S(t) = \{-\lambda t\}$, $h(t) = \lambda$ and $f(t) = \lambda \exp\{-\lambda t\}$, respectively.

2.3.2 Weibull distribution

Weibull distribution is one of the most popular distributions in analyzing the lifetime data. Much of its popularity is due to the wide variety of shapes it can assume by varying its parameters. The survivor, hazard and probability density functions are given by $S(t) = \exp(-\lambda t^{p})$, $h(t) = \lambda p t^{p-1}$ and $f(t) = \lambda p t^{p-1} \exp(-\lambda t^{p})$ for $0 \le t < \infty$, respectively.

2.3.3 Log-logistic distribution

For the lifetime, which has a log-logistic distribution with parameters a and b, the survivor, hazard and probability density functions are given by $S(t) = (1 + e^{a} t^{b})^{-1}$, $h(t) = b e^{a} t^{b-1} (1 + e^{a} t^{b})^{-1}$ and $f(t) = b e^{a} t^{b-1} (1 + e^{a} t^{b})^{-2}$, respectively.

2.3.4 Log-normal distribution

In the case when the logarithm of lifetime is normally distributed with mean μ and variance σ^2, the pdf is given by $f(t) = \frac{1}{\sigma\sqrt{2\pi}} t^{-1} \exp - \frac{1}{2\sigma^2}\left(\log t - \mu\right)^2$ for $0 \le t < \infty$. Furthermore, $S(t)$ and $h(t)$ can only be expressed in terms of integrals.

2.4 Survival models with covariates

Sometimes it is useful to add covariates or explanatory variables to describe the effects of factors which may affect the survival time of an individual. These variables can either be continuous such as age or categorical variables like gender. There are few approaches to modeling survival data with covariates, namely Cox proportional hazard model and parametric model.

2.4.1 Cox proportional hazard model

In survival models, the hazard function for a given individual describes the instant risk of experiencing an event of interest within an infinitesimal interval of time, given that the individual has not yet experienced that event. Cox (1972) proposed a semi-parametric model for the hazard function that allows the addition of explanatory variables, or covariates, but keeps the baseline hazard as a random, unspecified, nonnegative functional of time. The Cox hazard function for fixed-time covariates, x, is given by:

$$h_i\,(t) = h_o(t)\,\exp(\boldsymbol{\beta} x_i).$$

Due to the construction of Cox hazard function, the baseline hazard $h_o(t)$ is defined as the hazard function for the individual with zero on all covariates x, because the baseline hazard is not assumed to be of a parametric form (Cox, 1972).

The Cox proportional hazard regression model is a flexible tool for assessing the relationship of multiple predictors to a right-censored, time-to-event outcome and has much in common with linear and logistic models. (For further knowledge see Cox, 1972 and Collett, 2003).

2.4.2 Parametric survival model with covariates

Its idea is derived from picking one of the popular parametric distributions discussed in 2.3 and estimates its parameters depending on the covariates. (For further information see Kleinbaum and Klein, 2005).

3. Weaning Time Data

The considered data have been obtained from the Palestinian Central Bureau of Statistics (PCBS) - The Survey of the Palestinian Family Health 2006. The data were collected between November 2006 and January 2007.

Table 1 presents the sample distribution and response rate. The survey sample size totaled 13,238 households including 8,781 in the West Bank and 4,457 households in Gaza Strip. Number of children (below five years old) in the sample totaled 10,318 including 5,895 from the West Bank and 4,423 from Gaza Strip.

The total response rate was 99.2% distributed as 98.8% in the West Bank and 99.6% in the Gaza Strip.

Table 1. Sample Distribution and Response Rate

Sample and Response rate	Palestinian Territory	West Bank	Gaza Strip
Survey sample size (Household)	13,238	8,781	4,457
Children below 5 in the sample (Child)	10,318	5,895	4,423
Number of interviewed children	10,230	5,824	4,406
Response rate	99.2%	98.8%	99.6%

A set of explanatory analysis methods is used to clean the data. There were some errors with ages of children, e.g., age greater than 5 years, missing data, outliers, etc. Thus, the valid data for analysis consists of 7,041 cases.

From the variables contained in the original questionnaire, authors have selected a set of nine variables that are most relevant to the purpose of this study as described in the following:

<u>The dependent variable</u>

The time to wean has been considered as the dependent variable. It has been defined as the period between the date of the beginning until the end of breastfeeding.

<u>The independent variables</u>

There are nine independent variables that are expected to have impact on the length of weaning time which are listed and their values are explained in Table 2, associated with summary statistics i.e. minimum, maximum, mean, median and coefficient of variation (CoV) for continuous variables and relative frequency distribution for categorical variables.

Status was derived based on the number of months the child was breastfed and child's age in months. If the weaning time equals to the age of child, then the status equals 0 (censored), but if breastfed less than the age of the child, then status equals 1 (event).

Table 2. Description and Summary Statistics of the Study Variables, (*n* = 7,041)

Continuous variables					
Variables description	**min**	**max**	**Mean**	**Median**	**CoV**
Weaning time (month)	1	31	13.27	14.00	46.22%
Child's age (month)	1	60	36.20	37.00	37.23%
Child's weight at birth (grams)	500	5500	3249.28	3200.00	20.68%
Total live births	0	18	4.85	4.00	54.72%
Mother's age (years)	17	54	30.97	30.00	23.01%

Categorical variables		**n**	**(%)**
Status	0 = censored 1 = event	98 6943	1.4% 98.6%
Mother's refugee status	Non-refugee Registered refugee Non-registered refugee	3785 3154 102	53.8% 44.8% 1.4%
Locality type	Urban Rural Refugee camp	3726 2029 1286	52.9% 28.8% 18.3%
Mother's educational status	PhD Master degree Higher diploma Bachelor degree (B.Sc.) Intermediate level diploma Secondary Preparatory Elementary Acquainted (can read and write) Illiterate	1 10 14 518 409 1490 2566 1469 427 137	0.01% 0.1% 0.2% 7.4% 5.8% 21.2% 36.4% 20.9% 6.1% 1.9%
Mother's smoking status	Does not smoke and never smoked Yes, smoke Ex-smoker	6908 114 19	98.1% 1.6% 0.3%

4. Statistical Analysis and Results

This section presents the survival analysis of the weaning time, including the main characteristics, differences between the survival time of different groups, nonparametric and parametric modeling.

4.1 Survival characteristics

Table 3 presents descriptive statistics of survival time for infants. The results show that the measures for events and all data are much closed to each other, this may be referred to the small percentage of censoring (1.4%). The mean and median survival times are 13.265 and 14 months, respectively. Furthermore, additional investigation reveals that some of the censored data correspond to the extreme values of independent variables.

Table 3. Descriptive statistics for survival time

Statistic	Records	Event
N (%)	7041 (100%)	6943(98.6%)
Mean (S.E) 95% C.I.	13.344 (0.074) (13.200 – 13.489)	13.265 (0.073) (13.122 – 13.409)
Q_1 (S.E) 95% C.I.	18 (0.076) (17.85 - 18.15)	18 (0.075) (17.85 - 18.15)
Median (S.E) 95% C.I.	14 (0.074) (13.855 – 14.145)	14 (0.073) (13.856 – 14.144)
Q_3 (S.E) 95% C.I.	9 (0.136) (8.73 - 9.27)	9 (0.135) (8.74 - 9.26)

Figure 1 presents the survival plot of the weaning time, where the horizontal axis represents the period that the weaning time will continue. The vertical axis represents the rate of the weaning time that have started and are still ongoing. As anticipated, the function starts at one and drops monotonically to zero. It is important to note that the function does not take into consideration calendar dates. It is not important when the weaning time has started but rather how long it has lasted. This trend reflects the behavior of the end of breastfeeding in the Palestinian Territories, where the probability of stopping breastfeeding start increasing rapidly after 12 months**.**

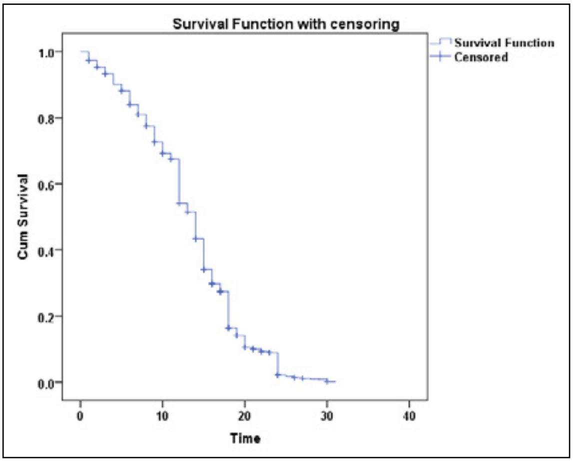

Figure 1. Survival Plots of Duration of Breastfeeding

4.2 Comparing the groups of survival curves

Table 4 presents the values of the log rank test associated with its degree of freedom and *p*-value for two variables, namely, mother's age at delivery and educational status, where there is a significant evidence of a difference between survival times for the groups of each mentioned variable at 0.05 level of significance. For the convenience of analysis, mother's age has been categorized into three groups (less than 25, 25-34 and 35 or more).

Table 4. Results of Significant Log Rank Test for Survival Time

Variable		Mean(S.E.)	Log Rank Test	df	Sig.
Mother's age at delivery (years)	Less than 25	14.49 (0.19)	233.17	2	0.00
	25-34	13.60 (0.11)			
	35 or more	12.39 (0.11)			
Mother's educational status	Illiterate	14.57 (0.53)	42.12	9	0.00
	Acquainted	13.31 (0.34)			
	Elementary	13.37 (0.17)			
	Preparatory	13.52 (0.12)			
	Secondary	13.35 (0.16)			
	Intermediate level diploma	13.37 (0.30)			
	Bachelor degree (BSc)	12.20 (0.25)			
	Higher diploma	9.57 (1.38)			
	Master degree	12.20 (1.87)			
	PhD	13.00 (0.00)			

For further explanations, the pairwise Scheffe's post hoc test is obtained and is found that there are significant differences between the mean weaning times of every pairwise categories of mother's age at delivery. Furthermore, there is a significant difference between the mean weaning times for children whom their mothers have higher diploma and other education levels. Further details on the summary of each group of all independent variables can be requested from the corresponding author.

4.3 Parametric survival models

It is worth to fit a parametric survival model of weaning time data. A set of four popular parametric survival models, namely, Exponential, Weibull, log-logistic and log-normal distributions are considered. The values of log likelihood, AIC, Chi-squared test associated with its p-values and the number of Newton-Raphson iterations are given in Table 5. Results show that the Weibull distribution has the smallest value of Chi-square test and has the smallest AIC values which indicates that it fits the survival data very well.

Table 5. Fitting parametric survival models

Statistic	Exponential distribution	Weibull distribution	Log-Logistic distribution	Log Normal distribution
Log likelihood	-24995.3	-22689.2	-22732.6	-24025.1
AIC	49992.6	45382.4	45469.2	48054.2
Chi-squared test	2577.63	1951.93	2998.86	2635.43
No. of Newton-Raphson iterations	4	7	4	4

For further investigation, the *log*(-*log*(*S*(*t*)) are plotted versus the *log*(*t*) as given in Figure 2, where *S*(*t*) is the estimated survival function. It is obvious that there is a linear relationship between the *log*(-*log*(*S*(*t*)) and *log*(*t*) which suggests an acceptable fitting of Weibull distribution with parameter estimates $\hat{\lambda} = 2.707$ and $\hat{p} = 0.447$. Therefore, the survival function is written by $S(t) = \exp(-2.707t^{0.447})$ and the hazard function is given by $h(t) = 1.21\ (2.707t^{0.553})$.

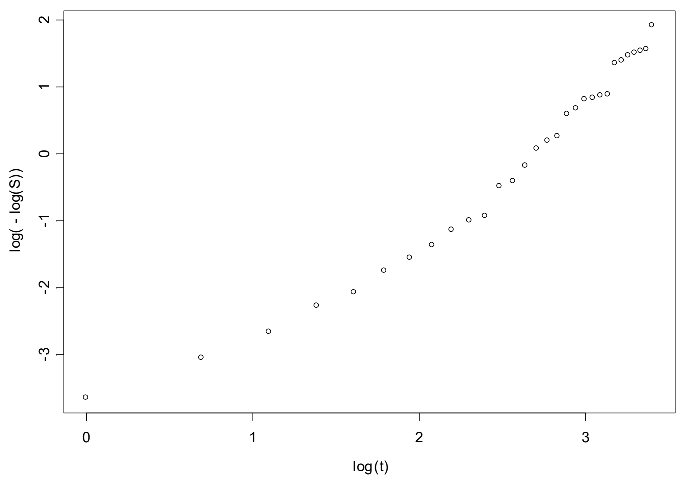

Figure 2: Scatter plot of *log(-log(S(t))* versus the log*(t)*

4.4 Estimation of Cox regression model parameters

Cox model assumes the risk function for the period of breastfeeding which associated with the impact of nine variables on a survival time which takes the following formula:

$$h(T) = h_0(T)\ exp(\boldsymbol{\beta}'\boldsymbol{x}),$$

where $h(T)$ represents the risk function for the model and $h_0(T)$ is the basic risk function for a period of breastfeeding when the values of covariates is $x = 0$.

Variable selection was done using backward method depending on Wald Statistic.

In the sixth step of the backward selection method as given in Table 6, results show that the following variables (child's weight at birth, child's age, mother's educational status and mother's age at delivery) have significant effect on the survival time of weaning and the rest of the variables are insignificant. Furthermore, the statistic (*-2 log likelihood*) equals 109868.879.

Table 6. Results of Backward Test for the Variables Influencing on the Survival Time

Variable	B	SE	Wald	d.f	p-value	Exp(B)
Child's weight at birth	0.00	0.00	13.65	1	0.00	1.00
Child's age	-0.02	0.00	235.82	1	0.00	0.99
Mother's educational status	0.03	0.01	9.76	1	0.00	1.03
Mother's age at delivery	-0.02	0.00	124.55	1	0.00	0.98

Therefore, the best fit model has the form:

$$h_i(t) = \exp\{(-0.000007)\,x_1 - (0.02)\,x_2 + (0.03)\,x_3 - (0.02)\,x_4\},$$

where x_1 is the child's weight at birth, x_2 is the child's age, x_3 is mother's educational status, and x_4 is the mother's age at delivery.

4.5 Parametric survival regression models with covariates

Four parametric models incorporate the covariates that were found to be significant in the previous fitted Cox model are obtained and summarized in Table 7. The results in Table 7 indicate that the Weibull survival model is the best fit model of weaning time where it has the smallest value of the (-2 *log likelihood*) statistic. Moreover, the difference between the (-2 *log likelihood*) of Cox model and Weibull model is 65271.28 which is statistically significant. Therefore, the Weibull model will be considered to fit the weaning time.

Assuming the survival time follows the Weibull distribution, then the Weibull regression model which is given by $h(t)=\exp(\boldsymbol{\beta}'\mathbf{x})\lambda p t^{p-1}$ is fitted. Since, the researchers used the S-Plus environment to obtain the estimates, then the parameterization of hazard function for Weibull distribution is given by $h(t) = \exp\left\{\frac{-\mu-\text{á}'\mathbf{x}}{\sigma}\right\}\sigma^{-1}t^{\sigma^{-1}-1}$ where μ and σ are called the intercept scale parameters, respectively.

Table 7. Parameter Estimates and (-2 log likelihood) Statistics for four Parametric Survival Models

Variable	Exponential		Weibull		Log-Logistic distribution		Log Normal distribution	
	parameter	p-value	parameter	p-value	parameter	p-value	parameter	p-value
Intercept	2.003	0.00	2.166	0.00	5.518	0.00	1.707	0.00
Child's weight at birth	0.000	0.02	0.014	0.00	0.001	0.00	0.000	0.00
Child's age	0.006	0.00	-0.094	0.00	0.094	0.00	0.008	0.00
Mother's educational status	-0.006	0.49	-0.011	0.00	-0.094	0.08	0.000	0.09
Mother's age at delivery	0.008	0.00	0.009	0.00	0.108	0.00	0.008	0.00
-2 log likelihood	49646		44597.6		44647.6		47503.8	

From the output of Weibull fitting we get the maximum likelihood estimates of the mean and scale parameters such that $\hat{\mu} = 2.166$ and $\hat{\sigma} = 0.431$. Hence, the maximum likelihood estimate of Weibull parameters are $\hat{\lambda} = \exp\left(-\frac{\mu}{\sigma}\right) = 0.0065$ and $\hat{p} = \hat{\sigma}^{-1} = 2.320$.

Table 8. The Estimates of Weibull Model Coefficients

Variable	β	exp (β)
Child's weight at birth	-0.032	0.968
Child's age	0.218	1.244
Mother's educational status	0.026	1.026
Mother's age at delivery	-0.021	0.979

Hence, the estimates of the Weibull model coefficients can be obtained by $\beta_i = -\frac{\alpha_i}{\sigma}$ as listed in Table 8. Therefore, the Weibull model is given by

$$h(t_i) = 0.015 t_i^{1.32} \exp(-0.032x_1 + 0.218x_2 + 0.026x_3 - 0.021x_4),$$

where x_1 is the child's weight at birth, x_2 is the child's age, x_3 is the mother's educational status, and x_4 is the mother's age at delivery.

The values of Weibull model coefficients in Table 8 show that, for every one gram increase in the child's weight at birth, the hazard of stopping breastfeeding decrease by 3.3%. Every month increase of the child's age results in an increase of the probability (24.4%) of stopping breastfeeding. For every education level increase of the child's mothers results in an increase of the probability (2.6%) of stopping breastfeeding. Finally, for every one year increase of the mother's age at delivery, the hazard of stopping breastfeeding is decreased by 2.1%.

5. Conclusions and Recommendations

Data for the study were drawn from the Survey of the Palestinian Family Health 2006. The analysis of survival time of weaning shows that the median among Palestinian infants is 14 months, and the probability of stopping breastfeeding rapidly increased after 12 months.

Log rank test show there are significant differences between the survival time of the groups of each of the following variables, mother's age at delivery and educational status.

The parametric Weibull distribution fits the weaning time. Furthermore, there are four variables contributed significantly in the determination of breast weaning time, namely, child's weight at birth, child's age, mother's age at delivery and mother's educational status.

It is recommended that the breastfeeding-promotion programs in Palestine should give special attention to: mothers whom delivered their babies at 35 years old or older.

REFERENCES

BHIMJI S, and HUDDA A., 2005, From Marriage to Parenthood the Heavenly Path. Islamic Humanitarian Service.

COLLETT D., 2003, *Modeling survival data for medical research*, 2nd ed, London, UK: Chapman-Hall.

COX D.R., 1972, Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187-220.

KAPLAN E.L., MEIER P., 1958, Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457-481.

KIMANI-MURAGE EW, MADISE NJ, FOTSO JCH, KYOBUTUNGI C, MUTUA MK, GITAU TM, YATICH N., 2011, Patterns and determinants of breastfeeding and complementary feeding practices in urban informal settlements, Nairobi Kenya. BMC Public Health, (DOI: 11396. 10.1186/1471-2458-11-396.)

KLEINBAUM D.G and KLEIN M., 2005, *Survival Analysis: A Self-Learning Text*, Wiley, New York, *2nd* ed.

PALESTINIAN CENTRAL BUREAU OF STATISTICS (PCBS), 2013, Annual Report. Children of Palestine - Issues and Statistics. Child statistics series (No. 16). Ramallah - Palestine.

PALESTINIAN CENTRAL BUREAU OF STATISTICS (PCBS), 2015, The Status of the Rights of Palestinian Children 2014 Ramallah - Palestine.

SHOSTAK J. and WALKER G.A., 2010, *Common Statistical Methods for Clinical Research with SAS Examples*, 3rd edn. Cary North Carolina, USA. SAS Institute Inc.

ZAID A., 2014, The Impact of Demographic, Economic and Social Factors on Infant Mortality in the Palestinian Territories Based on the Palestinian Family Data in 2010, Master Dissertation, An-Najah National University, Palestine.