

# **An Investigation into Social and Cultural Pressures of Vulnerable Women using Survival Analysis with Covariates**

## **Abstract**

In this study, we sought to model the time to the first birth interval from marriage for women in Nigeria, and identify the various factors affecting this timing. The study was also set to determine the average/median survival time for marriage to first birth interval among Nigerian women, to provide enlightenment in such areas and possibly reduce anxiety levels of women who have little or no knowledge of the median survival time to first birth who might be vulnerable to the exploitations of illicit religious and medical practitioners.

Data obtained from the Nigerian Demographic Health Survey (NDHS) 2018 was used for the purpose of this research. Information on the following variables was obtained: Time to first birth from marriage, Age Women's education, Wealth index, Place of residence, Employment, Contraceptive, Religion, and Region. The Kaplan-Meier estimator was used to estimate the median survival time, while the log-rank test was used to test the significance of the categories of the covariates used. The Density, Quantile, Survival and Probability plots were used to study candidate distributions that appropriately describe the data, and the Akaike Information Criterion (AIC) was used to select the best distribution for the Accelerated Failure Time Model.

The study found that the median Survival time of marriage to the first birth interval was 20 months. Level of education, religion, region, use of contraceptives and Wealth Index were found to significantly affect marriage to the first birth interval. A log-normal Accelerated Failure Time Model was fit to the data. Women with higher education were found to have a shorter time to first birth interval. Also, women from South-Western Nigeria had shorter marriage to first birth interval than the other regions.

**Keywords:** Survival Analysis, Marriage Birth Interval, NDHS 2018, Accelerated Failure Time Model, Kaplan-Meier Estimator

## Introduction

The first visible outcome of the fertility process is the birth of the first child. The first birth marks a woman's transition into motherhood. It plays a significant role in the future life of each individual woman and has a direct relationship with fertility (Tadesse, F., & Headey, D. 2010). The number of children a woman bears throughout her reproductive period in the absence of any active fertility control, and women who start giving the first birth very early in life tends to have a large number of children than those who start late is determined by the timing of the first birth according to Gyimah, S. O. (2003). However, one of determinant factors is the age at the start of marriage. Bongaarts, J. (2015), mentions that early childbearing can interrupt a young women's education and other activities which women need to accomplish. Clinical outcomes come in a variety of statistical forms, such as continuous systolic blood pressure that can easily be analyzed with linear regression. Others, such as mortality or myocardial infarction (MI), are distinct events and have forms that are slightly more complex to statistically analyze.

The growing issue of childlessness and delayed births has made women vulnerable, especially to fake religious leaders and quack medical practitioners in search of answers/solutions. Social and cultural pressures for children against couples often lead to desperate measures. Poor or even total lack of knowledge of the average marriage to birth interval (AMBI) increases the level of anxiety amongst couples in the event there is delay in child birth especially, the first one. Hence, this paper is to model marriage to first birth interval among women so as to provide some confidence to women and marriages experiencing delay since AMBI is a significant determinant of fertility.

Fertility is one of the factors that influence the fluctuation of the number of populations. One of the indicators of fertility rate is the total fertility rate (TFR), which can be defined as the average number of children that would be born to a woman over her reproductive age. According to (Islam, 2009), TFR can be reduced by increasing the age at marriage. However, this strategy is difficult to apply in the developed countries such as Indonesia due to the influence of social and cultural factors. Another alternative strategy is by controlling the FBI, which is defined as the time interval of a married woman to give birth to her first child since the time of first marriage. If the FBI is controlled, the next birth time would automatically be controlled (Islam, 2009).

## Data Structures and Methodology

The data used for this study was obtained from 2018 Nigeria Demographic Health Survey (NDHS). The response variable is time-to-first birth from marriage among women in Nigeria that is measured in months. For women who did not give birth (censored), the time was measured till the date of the interview. The Independent variables are shown in the table below. The table shows the various covariates used to model the survival time, and the categories for each covariate.

**Table 1: Variables and categories for covariates used in the model**

Variables	Description	Categories
Age	Age of women at marriage	Measured in years
Women education	Women's level of education	0 = No-education; 1 = Primary 2 = Secondary and 3 = higher
Wealth index	Household wealth index	0 = Poor; 1 = Middle; 2 = Rich
Place of residence	Place of residence	1 = Urban; 2 = Rural
Employment	Employment status	0 = unemployed, 1 = Employed
Contraceptive	Use of Contraceptive	0 = Non-User, 1 = User
Religion	Religion of respondents	0 = Christian, 1 = Muslim, 2 = Other

Region	Region of residence	1=North Central, 2=North East, 3=North West, 4= South East, 5= South-South, 6= South West
--------	---------------------	---

### Kaplan-Meier estimator

Let  $t_1, t_2, t_3, \dots$  denote the actual times of the occurrence of the event of interest of  $n$  individuals. Let  $d_1, d_2, d_3, \dots$  denote the number of event occurrences at each of these times, and let  $n_1, n_2, n_3, \dots$  be the corresponding number of subjects yet to experience the event of interest.

Note:  $n_2 = n_1 - d_1, n_3 = n_2 - d_2$

Then for any time  $t \in [0, t_1)$  we have  $S(t) = P(T > t) = 1$  ie propability of surviving beyond time  $t$

for any time  $t \in [t_1, t_2)$  we have  $\hat{S}(t) = 1 - \frac{d_1}{n_1}$

Similarly, for any time  $t \in [t_2, t_3)$  we have

$$\hat{S}(t) = \left(1 - \frac{d_1}{n_1}\right) \left(1 - \frac{d_2}{n_2}\right) \quad (1)$$

Hence in general for any time  $t \in [t_j, t_{j+1})$  we have

$$\hat{S}(t) = \left(1 - \frac{d_1}{n_1}\right) \left(1 - \frac{d_2}{n_2}\right) \dots \left(1 - \frac{d_j}{n_j}\right) = \prod_{i=1}^j \left(1 - \frac{d_i}{n_i}\right) \quad (2)$$

This is the Kaplan - Meier estimator of the survivor function  $S(t)$ .

The Kaplan-Meier estimator  $\hat{S}(t)$  can be regarded as a point estimate of the survival function  $S(t)$  at any time  $t$ .

### Cox Proportional Hazards Models

Proportional hazards models are a class of survival models in statistics. Survival models relate the time that passes, before some event occurs, to one or more covariates that may be associated with that quantity of time. In a proportional hazards model, the unique effect of a unit increase in a covariate is multiplicative

with respect to the hazard rate. For example, taking a drug may halve one's hazard rate for a stroke occurring, or, changing the material from which a manufactured component is constructed may double its hazard rate for failure. Other types of survival models such as accelerated failure time models do not exhibit proportional hazard

The purpose of the model is to evaluate simultaneously the effect of several factors on survival. In other words, it allows us to examine how specified factors influence the rate of a particular event happening (e.g., infection, death) at a particular point in time. This rate is commonly referred as the hazard rate. Predictor variables (or factors) are usually termed covariates in the survival-analysis literature.

The Cox model is expressed by the hazard function denoted by  $h(t)$ . Briefly, the hazard function can be interpreted as the risk of dying at time  $t$ . It can be estimated as follow:

$$h(t) = h_0(t) * \exp (\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p) \quad (3)$$

Where,

- $t$  represents the survival time
- $h(t)$  is the hazard function determined by a set of  $p$  covariates  $(x_1, x_2, \dots, x_p)$
- The coefficients  $(\beta_1, \beta_2, \dots, \beta_p)$  measure the impact (i.e., the effect size) of covariates.
- The term  $h_0$  is called the baseline hazard. It corresponds to the value of the hazard if all the  $x_i$  are equal to zero (the quantity  $\exp(0)$  equals 1). The  $t$  in  $h(t)$  reminds us that the hazard may vary over time.

The Cox model can be written as a multiple linear regression of the logarithm of the hazard on the variables  $x_i$ , with the baseline hazard being an 'intercept' term that varies with time. The quantities  $\exp(\beta_i)$  are called hazard ratios (HR). A value of  $\beta_i$  greater than zero, or equivalently a hazard ratio greater than one, indicates that as the value of the  $i^{th}$  covariate increases, the event hazard increases and thus the length of survival decreases. In other words, a hazard ratio above 1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival.

A key assumption of the Cox model is that the hazard curves for the groups of observations (or patients) should be proportional and cannot cross.

The hazard ratio for two subjects,  $k$  and  $k'$  with respective hazard functions

$$h_k(t) = h_0(t)e^{\sum_{i=1}^n \beta x_i} \quad (4)$$

$$h_{k'}(t) = h_0(t)e^{\sum_{i=1}^n \beta x'_i} \quad (5)$$

is given as:

$$\frac{h_k(t)}{h_{k'}(t)} = \frac{h_0(t)e^{\sum_{i=1}^n \beta x_i}}{h_0(t)e^{\sum_{i=1}^n \beta x'_i}} = \frac{e^{\sum_{i=1}^n \beta x_i}}{e^{\sum_{i=1}^n \beta x'_i}} \quad (6)$$

This hazard ratio is independent of time. The proportional hazard assumption however implies that the hazard of the event in any group is a constant multiple of the hazard in any other. In other words, if an individual has a risk of death at some initial time point that is twice as high as that of another individual, then at all later times the risk of death remains twice as high. It gives the effect size of covariates.

### Accelerated Failure Time (AFT) Model

Let  $S_1(t)$  and  $S_2(t)$  be the survival functions of two populations. The AFT models says that there is a constant  $c > 0$  such that

$$S_1(t) = S_2(ct) \quad \text{for all } t \geq 0 \quad (7)$$

This model implies that the survival time of population 1 is  $c$  times as much as that of population 2.

Let  $\mu_i$  be the mean survival time for population  $i$  and let  $\varphi_i$  be the population quantiles such that

$S_i(t)(\varphi_i) = \mu$  for some  $\mu \in (0,1)$ . Then

$$\begin{aligned} \mu_2 &= \int_0^{\infty} S_2(t)dt \\ &= c \int_0^{\infty} S_2(u)du \quad (t = cu) \end{aligned}$$

$$\begin{aligned}
&= c \int_0^{\infty} S_1(u) du \\
&= cu_1
\end{aligned} \tag{8}$$

And

$$S_2(\varphi_2) = \theta = S_2(c\varphi_1) \tag{9}$$

Assume that  $S_2(t)$  is a strictly decreasing function. Then we have

$$\varphi_2 = c\varphi_1 \tag{10}$$

This shows that under the accelerated failure time model, the expected survival time, median survival time of population 2 all are  $c$  times as much as those of population 1.  $c$  is sometimes called the acceleration factor.

Let  $T_i$  be the event time for individual  $i$ , and let  $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^T$  be a fixed covariate vector that allows a possibly non-null intercept. The AFT model can be represented by

$$\log T = \mathbf{x}_i' \boldsymbol{\beta} + \sigma \varepsilon_i, \quad i = 1, \dots, n \tag{11}$$

where  $\varepsilon_i$  are independent and identically distributed random errors with a distribution with support in the whole real line and that does not depend on  $\mathbf{x}_i$ .

The vector  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$  and  $\sigma$  are unknown parameters.

The above framework describes a general class of models: depending on the distribution we specify for  $\varepsilon$  that we will obtain a different model, but all will have the same general structure. Accelerated failure time models allow a wide range of parametric forms for the density function. For each distribution of  $\varepsilon$ , there is a corresponding distribution for  $T$ . The members of the AFT model class include the exponential AFT model, Weibull AFT model, log-logistic AFT model, log-normal AFT model, and gamma AFT model. The table below gives a brief summary of Parametric AFT models.

**Table 2: Summary of Popular Parametric AFT Models**

Distribution of $\varepsilon$	Distribution of T
Extreme value (1 parameters)	Exponential
Extreme value (2 parameters)	Weibull
Logistic	Log-logistic
Normal	Log-normal
Log-Gamma	Gamma

Given the values of the covariates  $\mathbf{x}$ , the density function has the following form

$$f(t) = (\sigma t)^{-1} f_0 \left( \frac{\log t - \log \varphi(\mathbf{x})}{\sigma} \right) \quad (12)$$

Where  $\sigma$  is the scale parameter, and  $\varphi(\mathbf{x})$  is some function of covariates; A popular choice for  $\varphi(\mathbf{x})$

$$\text{is } \varphi(\mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\beta}) \quad (13)$$

AFT models assume a survivor function of the following form,

$$\Pr(T \geq t) = S(t) = S_0^* \left[ \left( \frac{t}{\varphi(\mathbf{x})} \right)^{\frac{1}{\sigma}} \right] \quad (14)$$

where  $S_0^*$  is baseline survivor function.

The Weibull, lognormal, and log-logistic distributions for lifetime correspond to extreme value, normal, and logistic distributions for log of the lifetime, and the survivor function is given by

$$S(t) = S_0 \left( \frac{\log t - \log \varphi(\mathbf{x})}{\sigma} \right) \quad (15)$$

if  $\varphi(\mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\beta})$ , the survivor function can be rewritten as

$$S(t) = S_0 \left( \frac{\log t - \mathbf{x}'\boldsymbol{\beta}}{\sigma} \right) \quad (16)$$



**Table 3: The  $S_0(\varepsilon)$  functions for some common distributions**

Distribution	Survivor Function $S_0(\varepsilon)$
Normal	$1 - \Phi(\varepsilon)$
Extreme value	$\exp(-e^\varepsilon)$
Logistic	$(1 + e^\varepsilon)^{-1}$

### Parameter Estimation using Maximum Likelihood estimation

Survival times may be subject to right censoring. Here, the censoring times are represented by the independent random variables  $C_i$ , for  $i = 1, \dots, n$ , which are assumed to be independent of  $T_1, \dots, T_n$ . The censoring mechanism is assumed to be non-informative, that is, the distribution of the  $C_i$ s does not depend on unknown parameters. Let  $\delta_i = 1$ , if the observation for individual  $i$  is a failure time, and  $\delta_i = 0$ , if it is a censoring time. The observations can be represented by the pairs of random variables  $(Y_i, \delta_i)$ , where  $Y_i = \min(T_i, C_i)$ , and the covariate vectors  $x'_i$ , for  $i = 1, \dots, n$ .

The likelihood function for the unknown parameters is given by

$$L(\theta) = \prod_{i=1}^n \left[ \frac{1}{\sigma} f\left(\frac{y_i - x'_i \beta}{\sigma}\right) \right]^{\delta_i} \left( S\left(\frac{y_i - x'_i \beta}{\sigma}\right) \right)^{1-\delta_i} \quad (17)$$

where  $y_i$  is the observed value of  $Y_i$ ,  $f(\cdot)$  and  $S(\cdot)$  denote the density and survival functions of  $\varepsilon_i$ , respectively, and  $\theta = (\beta^T, \sigma)^T$  is the vector of unknown parameters

Using  $\varepsilon = \frac{y_i - x'_i \beta}{\sigma}$ , the log likelihood assumes the form

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^n \delta_i (\log f(\varepsilon_i) - \log \sigma) + (1 - \delta_i) \log S(\varepsilon_i) \quad (18)$$

The components of the score vector is given by

$$U_{\beta_j}(\theta) = \frac{\partial \ell(\theta)}{\partial \beta_j} = \frac{1}{\sigma} \sum_{i=1}^n a_i x_{ij}, \quad \text{for } j = 0, \dots, p \quad (19)$$

and

$$U_{\sigma}(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \sigma} = \frac{1}{\sigma} \sum_{i=1}^n \varepsilon_i a_i - \delta_i, \quad (20)$$

$$\text{where } a_i = - \left[ \delta_i \frac{d \log f(\varepsilon_i)}{d \varepsilon_i} + (1 - \delta_i) \frac{d \log S(\varepsilon_i)}{d \varepsilon_i} \right] \quad (21)$$

In matrix form, the score vector can be written as

$$U(\boldsymbol{\theta}) = \begin{bmatrix} U_{\beta}(\boldsymbol{\theta}) \\ U_{\sigma}(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \sigma^{-1} \mathbf{X}' \mathbf{a} \\ \sigma^{-1} (\boldsymbol{\varepsilon}^T \mathbf{a} - \mathbf{1}^T \boldsymbol{\delta}) \end{bmatrix}, \quad (22)$$

where  $\mathbf{X} = (x_1, \dots, x_n)^T$  is the  $n * (p + 1)$  matrix of covariates, and  $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)^T$ ,  $\boldsymbol{\varepsilon} = (\exp\{\varepsilon_1\}, \dots, \exp\{\varepsilon_n\})^T$  and  $\mathbf{a} = (a_1, \dots, a_n)^T$  are  $n$  dimensional column vectors.

Table 4 gives the expression for  $a_i$  in equation (21) for AFT models frequently used in survival data applications. The expression for  $a_i$  for the exponential distribution equals the corresponding  $a_i$  for the Weibull distribution with  $\sigma = 1$ . Maximum likelihood estimates (MLEs) for  $\beta$  and  $\sigma$  are obtained by solving the system of equations  $U(\boldsymbol{\theta}) = \mathbf{0}$ , which requires a numerical nonlinear optimization algorithm (such as Newton-Raphson and Fisher's scoring)

**Table 4: Expression for  $a_i$  in equation 21 for some common models**

Model	Error Distribution	$a_i$
Weibull	Standard extreme value	$\exp(\varepsilon_i) - \delta_i$
Log-Normal	Standard Normal	$\delta_i \varepsilon_i + \left( \frac{1 - \delta_i}{1 - \Phi(\varepsilon_i)} \right) \frac{d\Phi(\varepsilon_i)}{d\varepsilon_i}$
Logistic	Standard Logistic	$\frac{\exp(\varepsilon_i)}{1 - \exp(\varepsilon_i)} (1 + \delta_i) - \delta_i$
Note: $\Phi(\cdot)$ is the standard normal cumulative distribution function.		

The observed information matrix is

$$I(\boldsymbol{\theta}) = - \begin{bmatrix} \frac{d^2 l}{d\boldsymbol{\beta}^2} & \frac{d^2 l}{d\boldsymbol{\beta} d\sigma} \\ \frac{d^2 l}{d\sigma d\boldsymbol{\beta}'} & \frac{d^2 l}{d\sigma^2} \end{bmatrix}, \text{ where } \boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma) \quad (23)$$

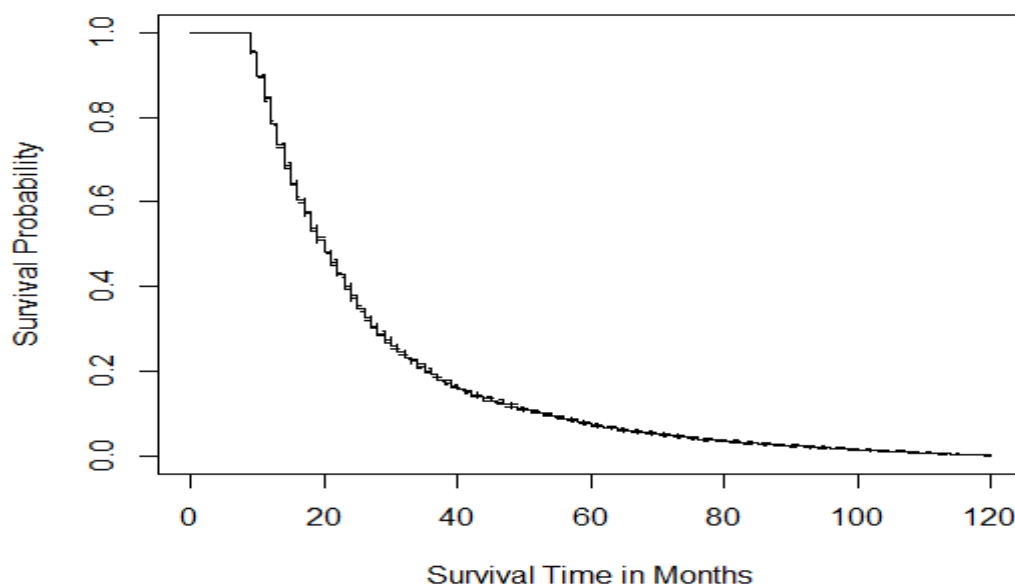
Other methods used in this study include: Newton Raphson Iteration while hypotheses were tested based on standard criteria such as the Likelihood Ratio and Wald tests. The Goodness of fit of the model were assessed using Akaike Information Criterion (AIC),  $R^2$ , Cox-Snell Residual. Cox-Snell residuals are calculated by using cumulative hazard  $H(t_i, \boldsymbol{\beta}, \sigma)$  function and standardized residual as:

$$r_{Si} = \frac{\log t_i - (\hat{\beta}_0 + \hat{\beta}_i x_i)}{\hat{\sigma}} \quad (24)$$

Where  $\hat{\beta}_0$ ,  $\beta$ , and  $\hat{\sigma}$  are maximum likelihood estimates of  $\beta_0$ ,  $\beta$ , and  $\sigma$ , respectively.

## Result and Discussion

The data obtained for the study were analyzed and interpreted in this section. The general median survival time was ascertained as well as those of the subcategories of various covariates. An appropriate survival model was built to model the first birth interval.



**Figure 1: Kaplan-Meier Plot of First Birth Interval**

The Kaplan Meier Plot above shows the probability of survival as the survival time increased. That is, the probability of not having a first child (i.e., not experiencing the event of interest) as the survival time increases.

**Table 5: General**

**Median Survival**

**Time**

n	events	median	0.95LCL	0.95UCL
22798	21514	20	20	20

Calculations from the Kaplan-Meier and its plot shows that the median survival time for the marriage to first birth interval for Nigerian women stood at 20 months using the NDHS 2018 datasets.

**Table 6: Median survival Time by Level of Education**

	n	events	median	0.95LCL	0.95UCL
edu=0	10780	10171	23	23	23
edu=1	3790	3651	18	18	19
edu=2	6291	5922	17	17	18
edu=3	1937	1770	16	16	17

The result shows that median survival times for Nigerian women with no education (edu=0) is 23 months, 18 months for those with primary education (edu=1), 17 months for those with secondary education (edu=2) and 16 months for respondents with higher educational qualification (edu=3). It is observed that the median survival time dropped with increasing educational qualification.

**Table 7: Median survival Time by Place of Residence**

	n	events	median	0.95LCL	0.95UCL
por=1	7958	7564	18	18	18
por=2	14840	13950	21	21	21

The result on type of place of residence shows that respondents who lived in urban areas (por=1) had a median survival time of 18 months while their rural counterparts had a median survival time of 21 months

**Table 8: Median survival Time by Working Status**

	n	events	median	0.95LCL	0.95UCL
work=0	7061	6457	21	21	22
work=1	15737	15057	19	19	20

The result on working status of the respondents shows that those who were not working had a median survival time of 21 months while those that worked had a 19 months median survival time.

**Table 9: Median survival Time by Religion**

	n	events	median	0.95LCL	0.95UCL
religion=0	8927	8462	17	17	17
religion=1	13711	12899	22	22	22

religion=2	160	153	21	19	25
------------	-----	-----	----	----	----

The results further reveal that the median survival time for Christians (religion=0) were 17 months, while the Muslims (religion=1) had a median survival time of 22 months and those of other religions (religion=2) had a median survival time of 21 months.

**Table 10: Median survival Time by Use of Contraceptive**

	n	events	median	0.95LCL	0.95UCL
contr=0	19600	18345	21	20	21
contr=1	3198	3169	16	16	17

Respondents who did not use contraceptives (contr=0) had a median survival time of 21 months while those who used contraceptives (contr=1) had a median survival time of 16 months.

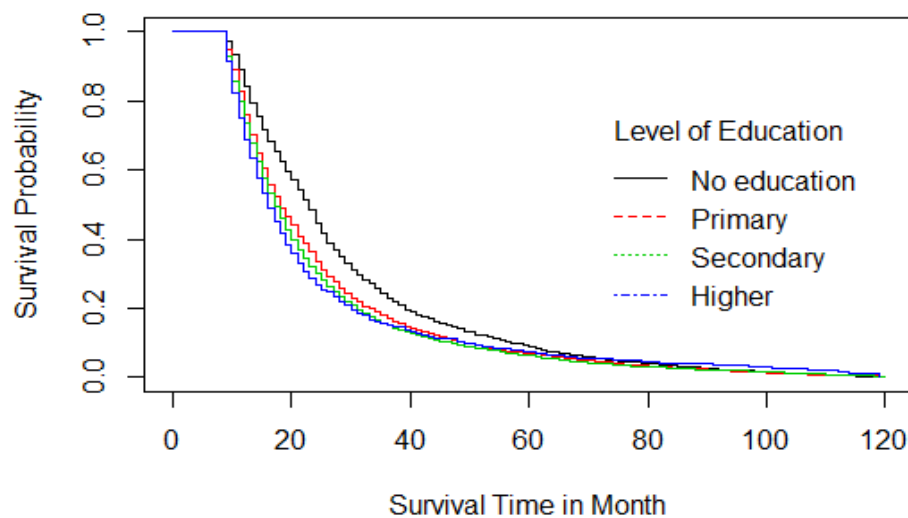
**Table 11: Median survival Time by Wealth Index**

	n	events	median	0.95LCL	0.95UCL
WI=1	10593	9968	22	22	22
WI=2	4540	4342	19	19	20
WI=3	7665	7204	17	17	18

The results on wealth index showed that respondents who were poor (WI=1) had a median survival time of 22 months, those of the middle class (WI=1) had a median survival time of 19 months whereas the rich had a median survival time of 17 months.

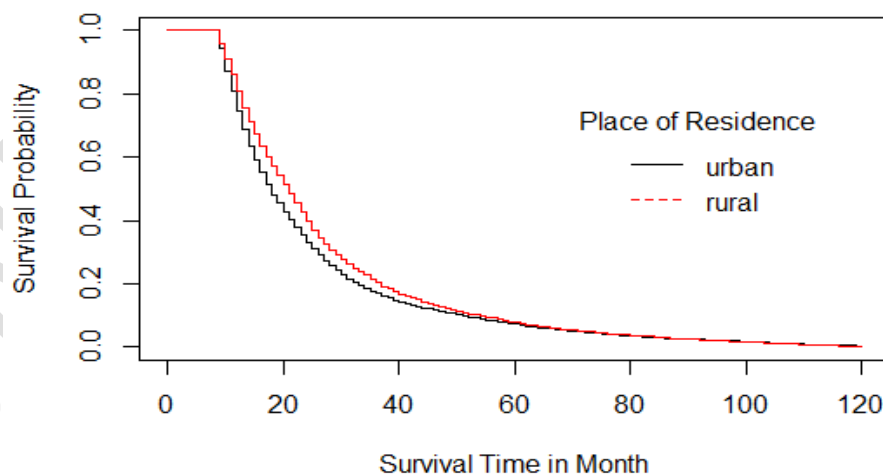
### KM Plots for Covariates

The following plots are the Kaplan Meier plots for all covariates



**Figure 2: KM plot for level of education**

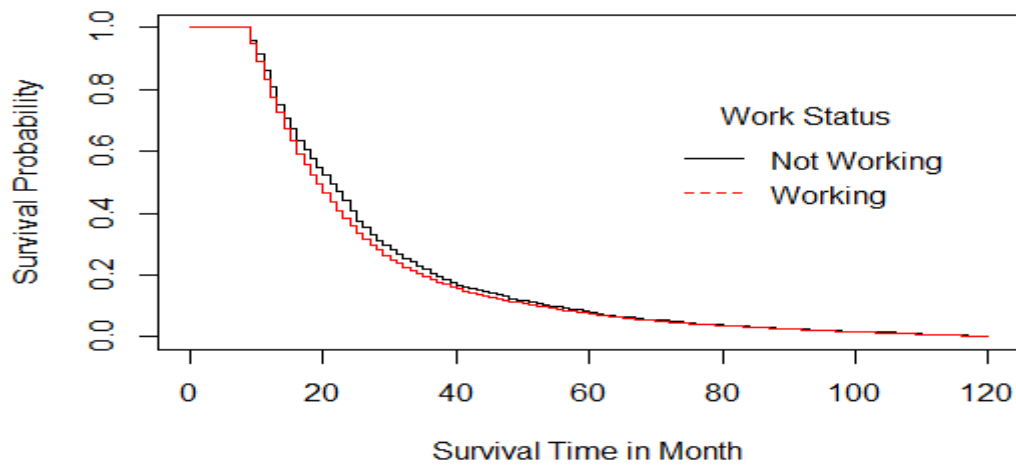
According to the plots in Figure 2 above, there is an observed difference between survival times for the different levels of education. The curve for no education was consistently above those of the other levels of education, which implies that Nigerian women with no education had a higher probability of not having their first baby relative to those with some form of education. On the other hand, Women with higher educational qualification had a lower probability of not having their first baby compared to the other women; in other words, women with higher education had a greater probability of having their first baby than the other women with lower educational qualifications.



**Figure 3: KM plot for Place of Residence**

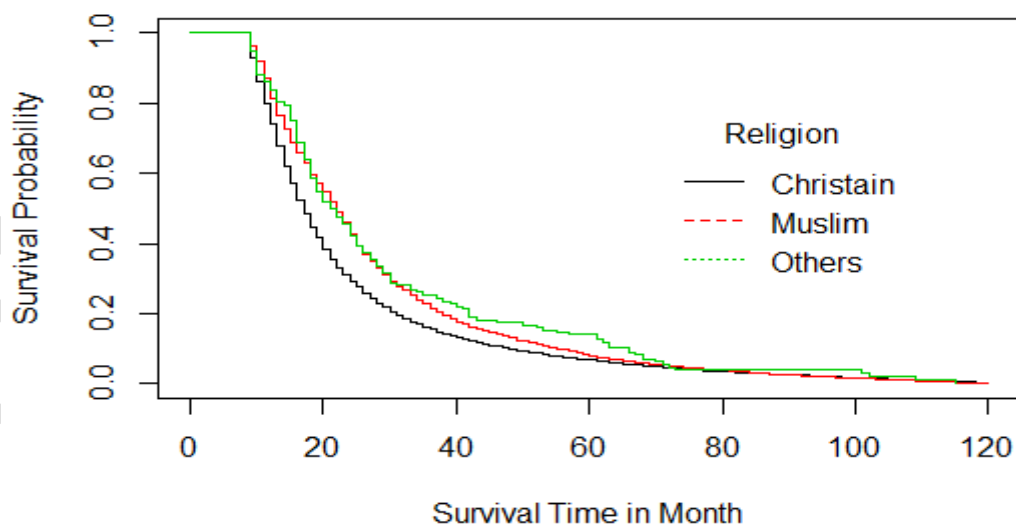
The Kaplan-Meier plot of the Place of residence above suggests there is difference in the survival of those who lived in urban residents and rural residents. The women who resided in rural areas as indicated by the

red curve have a higher probability of not having their first baby compared to their counterparts who lived in urban areas. This can alternately be put as, women who lived in urban areas have a higher probability of having their first babies than those who resided in rural areas.



**Figure 4: KM plot for Work Status**

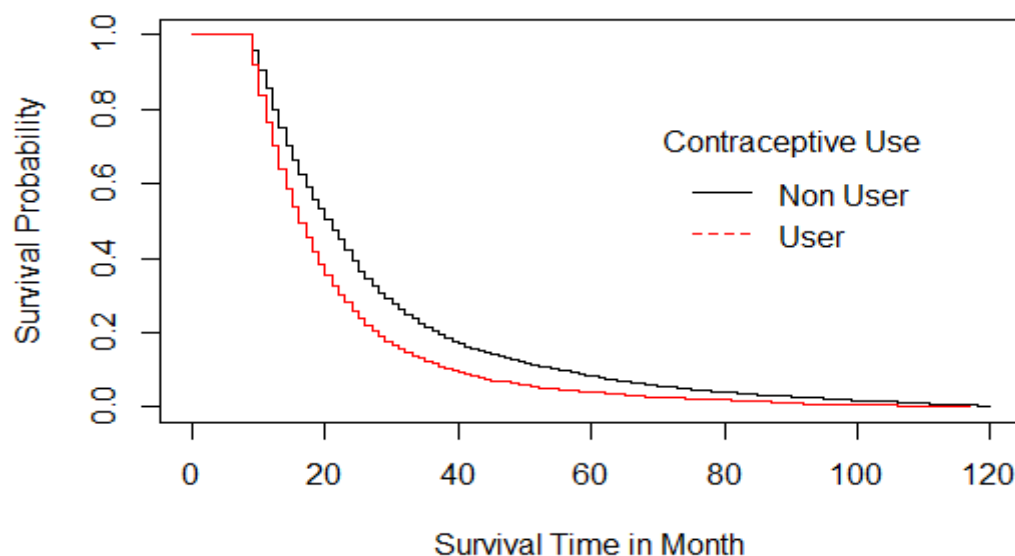
The Kaplan Meier plot of the work status suggests a difference in the survival time of respondents who worked and those who did not work, albeit very slight as the curves are quite close to each other. This implies that women who did not work have a slightly higher probability of not having their first babies than those who worked. Simply put women who worked have a higher probability of having their first baby.





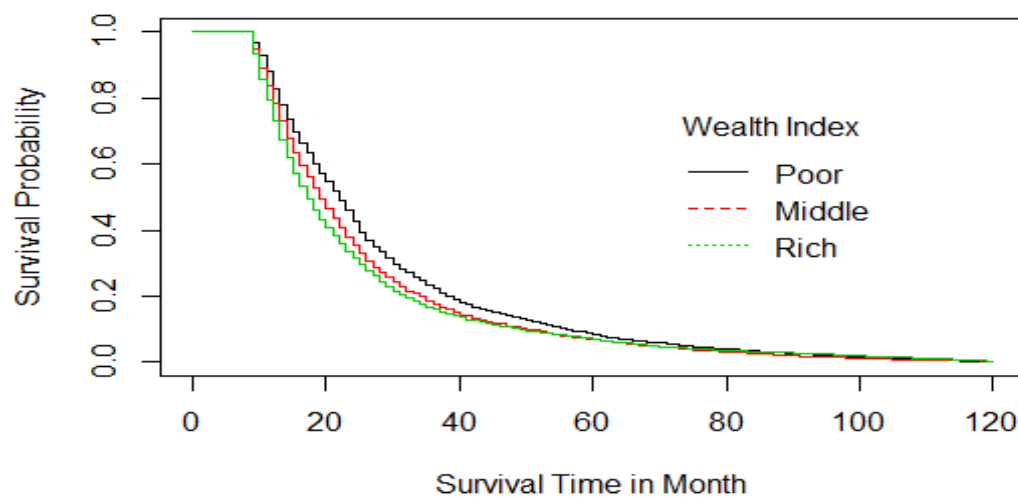
**Figure 5: KM plot for Religion**

The Kaplan-Meier plot for religion shows some differences in the survival time for the various religions. The survival times for Christians were consistently below those of Muslim and other religions. This means that Christian women have a lower probability of not having their first child when compared to Muslim women and women of other religions. The survival curve for Muslim didn't differ significantly from each other as they were very close to each other with some intersections.



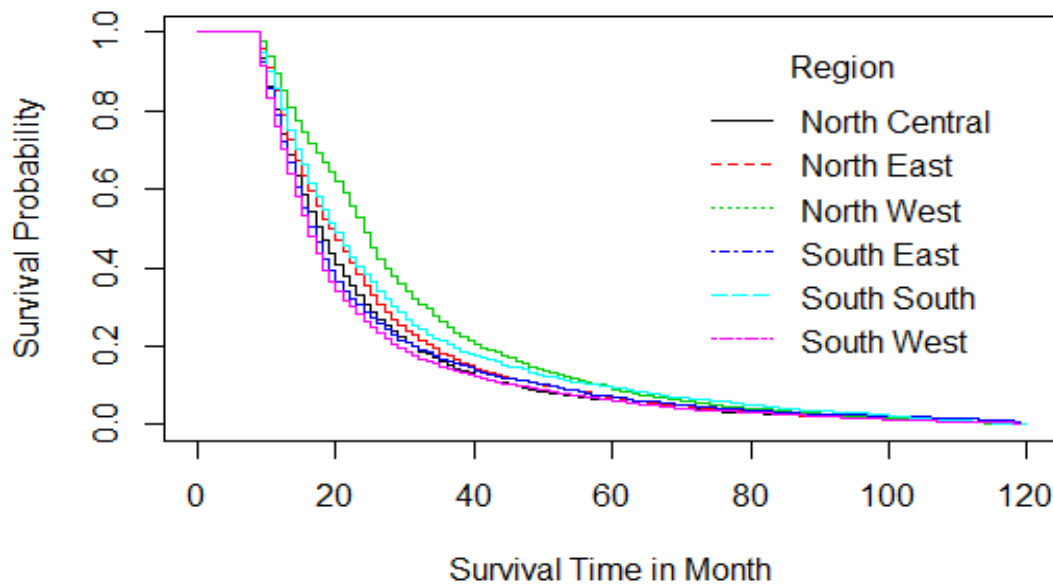
**Figure 6: KM Plot for Use of Contraceptive**

The plot on use of contraceptives suggests a difference in the survival times of Users and Non-Users of contraceptives, with the non-user curve above that of users. This implies that women who did not use contraceptives have a higher probability of not having their first babies than those who used contraceptives. That is to say, that women who did not use contraceptives have a lower probability of having their first babies.



**Figure 7: KM plot for Wealth Index**

The Kaplan-Meier plot in figure 7 above suggests a difference in the survival of time across different economic class. The curve for poor women was higher than those of the rich most of the time. This means that the poor women had a higher probability of not having their first baby than the richer women. In other, they poor women had a lower probability of having their first babies than the richer ones. On the other hand, the curve for the rich women was consistently lower, implying that the rich women have lower probability of not having their first babies that is they have a higher probability of having their first babies when compared to the poorer women.



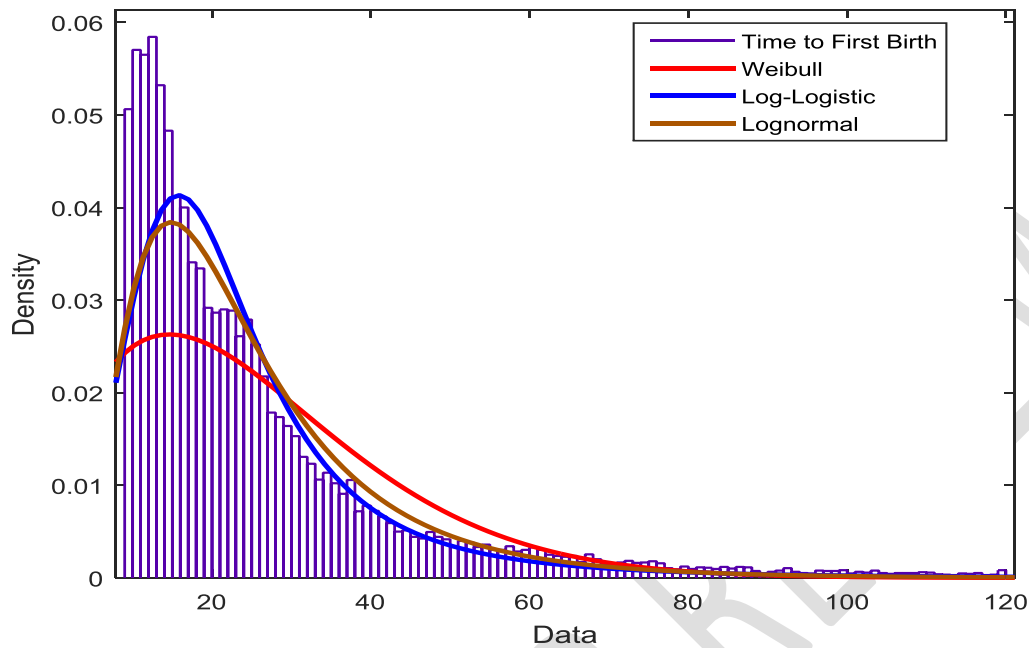
**Figure 8: KM plot for Region**

The plot above suggests a difference in the survival time for the different regions. The survival curve for North West is observed to be above all other curves, which means that women from the North West region of the country have a higher probability of not having their first babies than those from the other regions; their probability of having their first baby is thus lower than those from the other regions. While that of South West was consistently below those of the other regions, which means that women from the South West region (pink curve) of the country have a lower probability of not having their first baby than women from the other regions. It can be put simply as women from the South West region have a higher probability of having their first babies than the women from the other regions. In the same vein, albeit slightly, women from South (indicated by the blue curve) have a higher probability of experiencing their first babies than women from the North Central (black curve), and the women from the North Central than those from the North East (Red curve), and those of the North East than those of the South-South (blue curve), and lastly those of the South-South than those of the North West (green curve).

### **Accelerated Failure Time (AFT) Results**

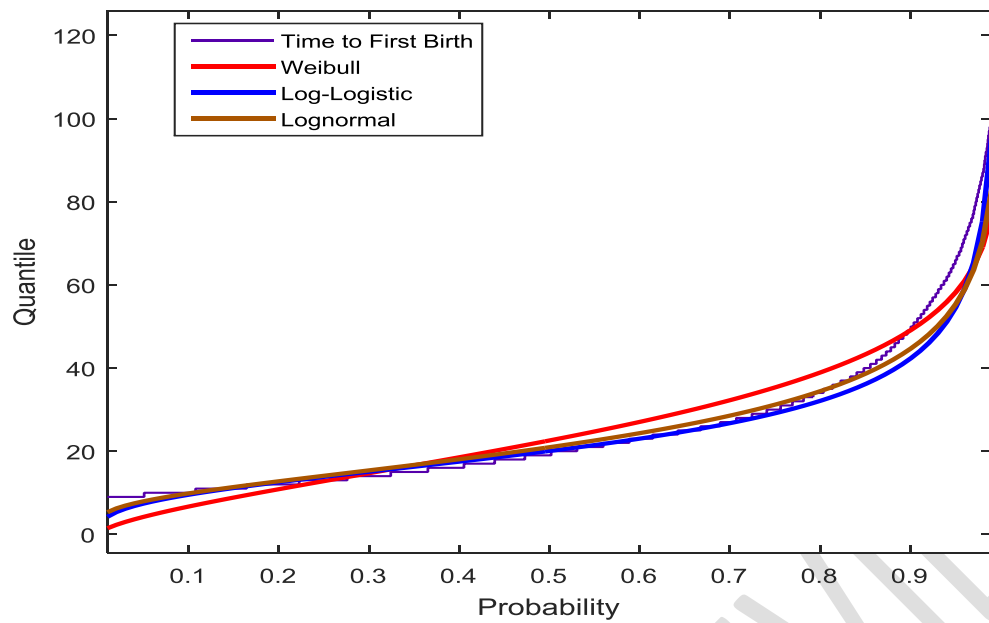
In this section we investigated possible/suitable distributions that describe the event of interest, which in this study is the time to first birth for married couples. Several distributions were already considered for

describing the First birth interval (too many to be included is single plot), but upon a prior visualization the ones shown below were the closest to describing the event of interest.



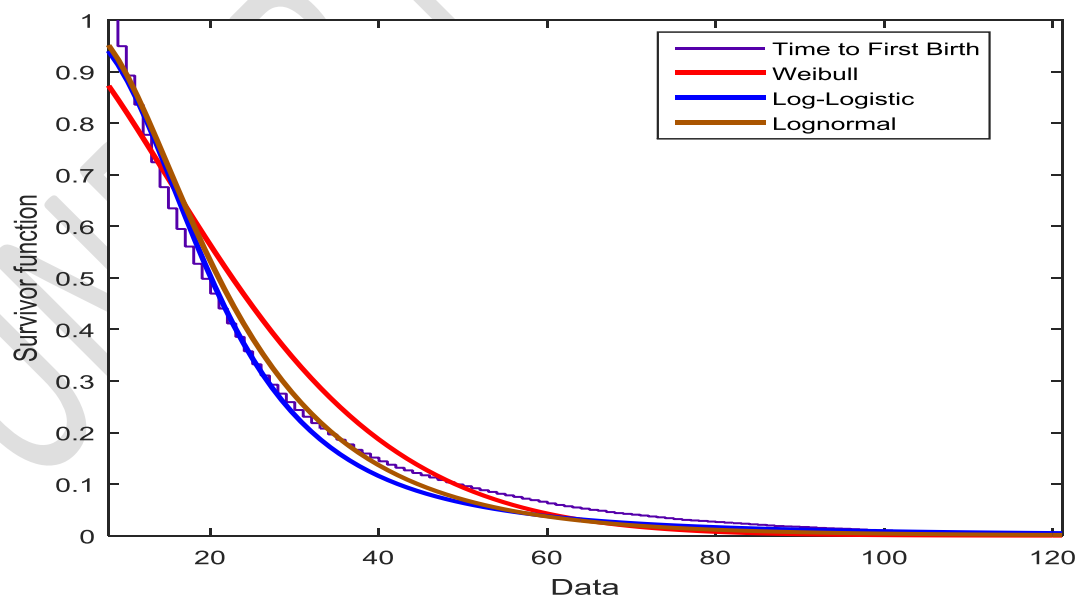
**Figure 9: Density Plot of First Birth Interval**

Probability density curves of all the distributions under investigation plotted with the histogram plot of the data on ‘First Birth Interval’ are displayed in figure 9. This is visualizing how well each of the distributions describes the data. With the distributions distinguishable with the different colors as shown in the plot above, we observe that the Weibull distribution does not appropriately describe the First Birth Interval, as it does poorly in covering the peak as well as the tail. This is an indication that the Weibull distribution may not very well describe our data. On the other hand, the other two distributions Log-logistic and Log-Normal distributions do better, with Log-Logistic describing the peak better while the Log-Normal describes the tail better.



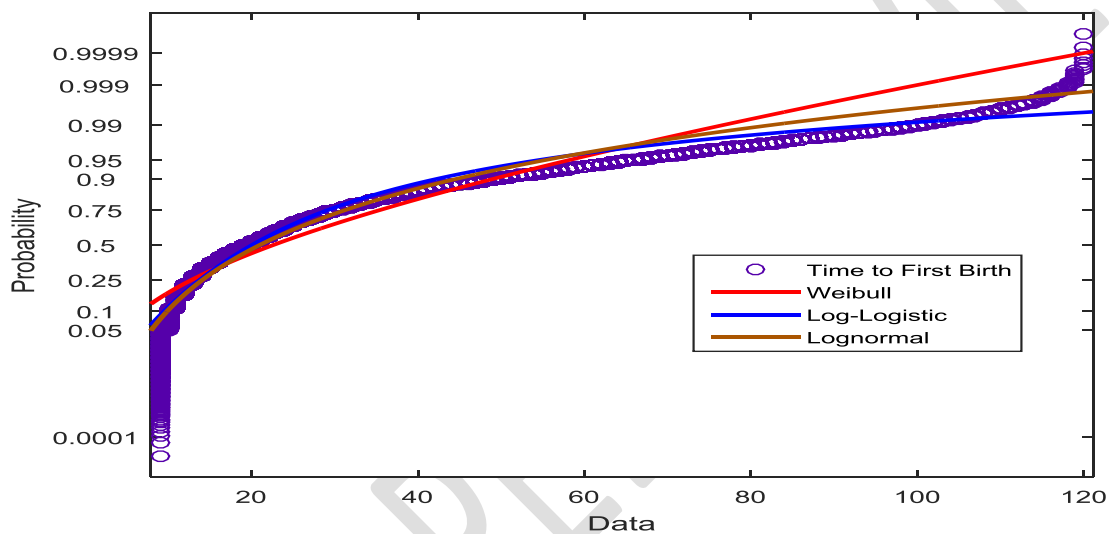
**Figure 10: Quantile Plot of First Birth Interval**

Quantile plot is another graphical method for determining whether sample data conform to a distribution. The plot shows a wider deviation of the Weibull distribution at the beginning with intersections at the middle and towards the end. The Log-Normal and Log-Logistic lay more consistently to the curve for the First Birth Interval with slight deviations toward the end.



**Figure 11 Survivor Plot of First Birth Interval**

Figure 11 displays the survival curves of all the distributions overlaid on the First Birth Interval curve. The plot measures how appropriately the distribution describe our data by how close the curve for each distribution is to the curve of the actual data. For most part, the curve for the Weibull distribution was further apart from the curve for the first birth interval, whereas, the log-normal and log-logistic curves were closer to the curve for the first birth interval. This is an indication that these two might describe the first birth interval better than the Weibull distribution.



**Figure 12: Probability Plot of First Birth Interval**

The probability plot once again shows how the Weibull distribution did not adequately mimic the path described by the data. The curve for the Log Normal and Log Logistics shows how closely the respective distributions follow the path shown by the data, with slight deviations towards the end. After the visualizations, further test was carried out using the AIC to determine the best distribution from amongst the three initially considered based on their strengths in describing the data via the plots.

**Table 12: Akaike Information Criterion (AIC) from the distributions**

Distribution	AIC
Log-Normal	170731.5
Log-logistic	170949.0

Weibull	177957.0
---------	----------

The above result shows the Log-Normal Distribution as a better fit to describing the event as it has the smallest AIC, and hence was used to obtain the following results. This means that for the most part and as the survival time increases, the Log-Normal distribution would still model the first birth interval better than the Log-Logistic and Weibull distribution. The table below contains the estimates of the Accelerated Failure time

model

**Table 13: Estimates of the Accelerated Failure Time Model (Log-Normal Distribution)**

	Value	Std. Error	z	p
(Intercept)	3.050391	0.024601	123.99	< 2e-16
age	-0.000300	0.000490	-0.61	0.54066
factor(edu)1	-0.094177	0.012550	-7.50	6.2e-14
factor(edu)2	-0.119160	0.013003	-9.16	< 2e-16
factor(edu)3	-0.130319	0.018285	-7.13	1.0e-12
factor(por)2	-0.000884	0.010145	-0.09	0.93056
factor(work)1	0.018905	0.009258	2.04	0.04115
factor(religion)1	0.071204	0.012598	5.65	1.6e-08
factor(religion)2	0.048772	0.047754	1.02	0.30710
factor(region)2	-0.006056	0.013678	-0.44	0.65792
factor(region)3	0.140730	0.013374	10.52	< 2e-16
factor(region)4	0.070983	0.017020	4.17	3.0e-05
factor(region)5	0.196680	0.017047	11.54	< 2e-16
factor(region)6	0.010055	0.016321	0.62	0.53783
factor(contr)1	-0.107340	0.011882	-9.03	< 2e-16

factor(WI)2	-0.036616	0.011492	-3.19	0.00144
factor(WI)3	-0.044212	0.012833	-3.45	0.00057
Log(scale)	-0.527278	0.004846	-108.81	< 2e-16

Scale= 0.59

Log Normal distribution

Loglik(model)= -85347.7    Loglik(intercept only)= -85916.1

Chisq= 1136.8 on 16 degrees of freedom, p= 5.4e-232

Number of Newton-Raphson Iterations: 3

n= 22798

The coefficients (Value)in the table above are logarithms of ratios of survival time, so a positive coefficient means longer survival. However, to get a more intuitive interpretation of the time ratio, the time ratios are transformed using the exponential function as shown in the table below.

**Table 14: Time Ratio from the AFT model**

Variables	Coefficient	Time Ratio (TR)
(Intercept)	3.05039122	21.12360687
age	-0.0002999	0.999700175
factor(edu)1	-0.0941773	0.910121395
factor(edu)2	-0.1191601	0.887665655
factor(edu)3	-0.1303191	0.877815273
factor(por)2	-0.0008841	0.999116327
factor(work)1	0.01890452	1.019084338
factor(religion)1	0.07120418	1.073800448



<b>factor(religion)2</b>	0.04877201	1.049980941
<b>factor(region)2</b>	-0.0060564	0.993961907
<b>factor(region)3</b>	0.14072973	1.15111349
<b>factor(region)4</b>	0.07098298	1.073562948
<b>factor(region)5</b>	0.19668029	1.217354775
<b>factor(region)6</b>	0.0100551	1.01010582
<b>factor(contr)1</b>	-0.1073402	0.898220072
<b>factor(WI)2</b>	-0.0366158	0.964046442
<b>factor(WI)3</b>	-0.0442118	0.956751336

The coefficients for the various factors from the table above are interpreted as follows:

**Education:** A time ratio of .91 shows that the survival time of respondents with a primary education is about 91% of the survival time of respondents with no education. In other words, the survival time of respondents with primary education is 9% shorter than the survival time of respondents with no education. Similarly, the survival time for respondents with a secondary education is about 88% of the survival time of respondents with no education. And finally, the survival time of respondents with higher educational qualification is about 87% of the survival time of respondents with no education.

**Religion:** The time ratio of 1.019 shows that the survival time of Muslim respondents is about a 102% of the survival time of Christian respondents. In other words, the survival time of Muslim respondents is about 2% longer than the survival time of Christian respondents. Also, the survival time of respondents who practiced other religions is about 5% longer than the survival time of the Christian respondents.

**Region:** The time ratio of 1.15 shows that the survival time of North-Western respondents is 115% of the survival time of North-Central respondents. In other words, the survival time of North-Western respondents

is about 15% longer than the survival time of North-Central respondents. The survival time of the South-Eastern respondents were about 7% longer than the survival time of the North-Central respondents. The survival time of South-Southern respondents is about 22% longer than the survival time of the respondents from the North-Central region. There were, however, no significant differences in the survival times of the North-East and South-Western respondents.

Contraceptive Use: the result shows that the survival time of respondents using a contraceptive is about 90% of the survival time of respondents not using contraceptives. In other words, the survival time of respondents who use contraceptives is about 10% shorter than the survival time of respondents who do not use contraceptives.

Wealth Index: the survival time of the middle class and rich respondents are both about 96% of the respondents who are poor, that is, to say that the survival times of the middle class and rich respondents are about 5% shorter than the survival time of respondents who are poor.

The place of Residence and working status of the respondents were however not very significant.

### Log Rank Test

A further analysis is carried out to provide backup to the observations from the Kaplan Meier plots above.

This is done using the log rank test as shown below:

**Table 15: Log rank test for Level of Education**

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
edu=0	10780	10171	11534	161	369.5
edu=1	3790	3651	3384	21	26.5
edu=2	6291	5922	5109	129	180.2
edu=3	1937	1770	1487	54	61.5

Chisq= 389 on 3 degrees of freedom,  $p = <2e-16$

The result of the log-rank test shows that there is a significant difference in the survival time of respondents with different levels of education with the p-value less than 0.05. In other words, the time to first birth differed significantly for at least two levels of education.

**Table 16: Log rank test for Place of Residence**

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
por=1	7958	7564	6975	49.8	78.1
por=2	14840	13950	14539	23.9	78.1

Chisq= 78.1 on 1 degrees of freedom,  $p = <2e-16$

With a p-value less than 0.05 the result shows that there is a significant difference in survival time between the urban and rural residents. That is to say that, the time it took women who lived in urban areas to have their first babies was indeed different from the time it took the women who lived in rural areas to have theirs.

**Table 17: Log rank test for Work Status**

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
work=0	7061	6457	6814	18.71	29
work=1	15737	15057	14700	8.67	29

Chisq= 29 on 1 degrees of freedom,  $p = 7e-08$

A p-value less than 0.05 shows that there is significant difference in the survival time of working and non-working respondents. This shows that the survival time or the time to first birth from marriage of working women differed significantly from those of non-working women.

**Table 18: Log rank test for Religion**

	N	Observed	Expected	(O-E) <sup>2</sup> /E	(O-E) <sup>2</sup> /V
religion=0	8927	8462	7354	167.03	269.93
religion=1	13711	12899	13983	84.10	255.53
religion=2	160	153	177	3.21	3.43

Chisq= 270 on 2 degrees of freedom, p= <2e-16

There is a significant difference in the survival time of respondents for the different religions practiced. It can be alternately put as, the survival time or the time to first birth from marriage differed significantly for at least two religions. This is however the case, as we observed a significant gap between the survival times of the Christian and Muslim Women, although, not much difference was observed for Muslim women and those of other religions.

**Table 19: Log rank test for Contraceptive Use**

	N	Observed	Expected	(O-E) <sup>2</sup> /E	(O-E) <sup>2</sup> /V
contr=0	19600	18345	19139	33	318
contr=1	3198	3169	2375	266	318

Chisq= 318 on 1 degrees of freedom, p= <2e-16

The result also reveals that the survival time of respondents using contraceptive is statistically different from those who do not contraceptives. This implies that the duration from marriage to first birth for women who used contraceptives was indeed different from those who did not use contraceptives.

**Table 20: Log rank test for Wealth Index**

	N	Observed	Expected	(O-E) <sup>2</sup> /E	(O-E) <sup>2</sup> /V
WI=1	10593	9968	10944	86.96	188.0
WI=2	4540	4342	4160	8.01	10.5

WI=3	7665	7204	6411	98.11	148.4
------	------	------	------	-------	-------

Chisq= 205 on 2 degrees of freedom,  $p = < 2e-16$

A p-value less than 0.05 shows that there is a significant difference in the survival time of respondents across the various wealth categories. This means that the duration from marriage to first birth of at least two categories of the wealth index classes were indeed different.

### Cox Proportional Hazard Results

In this section the Cox proportional hazard (a semi-parametric approach) was used to study the effect of the various covariates on the marriage to first birth interval.

**Table 21: Estimates of Cox Proportional Hazard model**

	coef	exp(coef)	se(coef)	z	Pr(> z )
age	0.0004445	1.0004446	0.0008570	0.519	0.6039
factor(educ)1	0.1347100	1.1442049	0.0216178	6.231	4.62e-10 ***
factor(educ)2	0.1776957	1.1944618	0.0226240	7.854	4.02e-15 ***
factor(educ)3	0.1739136	1.1899528	0.0319060	5.451	5.01e-08 ***
factor(por)2	0.0067438	1.0067666	0.0175573	0.384	0.7009
factor(work)1	-0.0153650	0.9847524	0.0160807	-0.955	0.3393
factor(religion)1	-0.1146443	0.8916833	0.0220516	-5.199	2.00e-07 ***
factor(religion)2	-0.0738940	0.9287701	0.0823788	-0.897	0.3697
factor(region)2	0.0156494	1.0157725	0.0238230	0.657	0.5112
factor(region)3	-0.1819309	0.8336590	0.0232398	-7.828	4.94e-15 ***
factor(region)4	-0.1327357	0.8756965	0.0294150	-4.513	6.41e-06 ***
factor(region)5	-0.3259824	0.7218179	0.0296735	-10.986	< 2e-16 ***
factor(region)6	-0.0242312	0.9760601	0.0282422	-0.858	0.3909
factor(contr)1	0.2250395	1.2523722	0.0202109	11.135	< 2e-16 ***

factor(wI)2	0.0476039	1.0487552	0.0199194	2.390	0.0169 *
factor(wI)3	0.0477812	1.0489411	0.0222617	2.146	0.0318 *

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Concordance= 0.583 (se = 0.002 )

Likelihood ratio test= 833.1 on 16 df, p=<2e-16

wald test = 855.1 on 16 df, p=<2e-16

Score (logrank) test = 864.7 on 16 df, p=<2e-16

The above result shows the significance level of the different tiers for each covariate on the survival time of respondents. The first category of each covariate is used as the reference group for the purpose of interpretation and comparison. The table further shows the hazard ratios for subcategories of each covariate. The result hence shows that the respondent's level of education has a significant impact on the marriage to first birth interval. Respondents with primary education have 14 percent higher risk of becoming a mother relative to those with no education, similarly respondents with secondary education about 19 percent greater chances of becoming mothers compared to those with no education while those of higher education are at 18 percent greater risk of becoming mothers. The result however shows that the place of residence and working status did not significantly affect the marriage to first birth interval.

The religion practiced had a significant effect in the marriage to first birth interval, respondents who practiced the Islamic religion were found to have 11% fewer chances of becoming mothers relative to their Christian counterpart, the sub-category "others" were however not significant.

The geographic regions of the respondents were also found to significantly impart the marriage to first birth interval. Respondents from the North-West region of the country have 18 percent lower chances of having their first birth after marriage compared to their colleagues from the North-Central, the respondents from the South-East have 13% fewer chances of becoming mothers relative to those from the North-Central region, similarly respondents from the South-South have 32% lower chances of having their first child after marriage relative to those from the North-Central part, whereas the sub category North-East and South-West were not significant. The use of contraceptive was found to have significant impact on the marriage to first

birth interval, as respondents who used contraceptives had 25% greater chances of having their first baby relative to the respondents who did not. The economic status of the respondents was equally significant in determining the marriage to first birth interval, as those of the middle class and the rich have about 5% chances higher of becoming mothers relative to the poor class.

**Table 22: Result of the Schoenfeld residual**

	chisq	df	p
age	49.2	1	2.4e-12
factor(edu)	404.5	3	< 2e-16
factor(por)	116.3	1	< 2e-16
factor(work)	23.7	1	1.1e-06
factor(religion)	261.3	2	< 2e-16
factor(region)	465.1	5	< 2e-16
factor(contr)	30.0	1	4.3e-08
factor(WI)	250.3	2	< 2e-16
GLOBAL	626.4	16	< 2e-16

The result from the Schoenfeld residual shows that on more general basis, the covariates listed above have a significant effect on the marriage to first birth interval. Specifically, with all the p-values less than 0.05, this implies that all the covariates (which includes education, place of residence, work status, religion, region, use of contraceptive and Wealth Index) all had a significant effect on the survival time.

## Conclusions

The study found that the median survival time of First Birth Interval for Nigerian women is 20 months. Furthermore, there was a significant difference in the survival time of the covariates, and the covariates generally had a significant effect on the survival time of First Birth Interval. The factors that significantly

impacted the survival time of First Birth Intervals includes Level of education, religion, region, use of contraceptive and Wealth Index. A Log-Normal Accelerated Failure Model was fit to the data. Women with higher education have a shorter time to first birth interval than women with lower educational attainment. The Christian women have the shortest time to first birth interval, followed by the Muslim women and then women who practiced other religions. Women from the South-West have shortest time to becoming mothers while North-West women have the longest time to becoming mothers. Finally, awareness should be promoted throughout the entire public regarding the median survival time to first birth interval in order to reduce anxiety among couples who may think they have waited too long for their first baby. Women should be exposed to better education, as those with a higher education showed a higher risk to first birth.

## REFERENCES

- Altman, D. G., De Stavola, B. L., Love, S. B., & Stepniowska, K. A. 1995. Review of survival analyses published in cancer journals. *British journal of cancer*, 72(2), 511-518.
- Andersen, P. K. 1991. Survival analysis 1982–1991: the second decade of the proportional hazards regression model. *Statistics in Medicine*, 10(12), 1931-1941.
- Bongaarts, J. 2015. World Fertility Report 2013: Fertility at the Extremes.
- Borsi, L., Lickes, M., & Soldo, L. 2011. The stratified Cox Procedure.
- Cox C, Chu H, Schneider M.F., Munoz A. 2007. Parametric Survival Analysis and Taxonomy of Hazard Functions for the Generalized Gamma Distribution. *Statist. Med*: 26: 4352-4374.
- Efron, B. 1977. The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72, 557-565. <http://dx.doi.org/10.1080/01621459.1977.10480613>



- Faruk, A. 2018. The comparison of proportional hazards and accelerated failure time models in analyzing the first birth interval survival data. *In Journal of Physics: Conference Series* (Vol. 974, No. 1, p. 012008). IOP Publishing.
- Francisco M. C., Antonio H. M., Dione M. and Silvia L. P. 2014. Testing Inference in Accelerated Failure Time Models; Canadian Center of Science and Education. *International Journal of Statistics and Probability*; Vol. 3, No. 2
- George, B., Seals, S., & Aban, I. 2014. Survival analysis and regression models. *Journal of nuclear cardiology*, 21(4), 686-694.
- Gijbels, I. 2010. Censored data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(2), 178-188.
- Gyimah, S. O. 2003. A cohort analysis of the timing of first birth and fertility in Ghana. *Population Research and Policy Review*, 22(3), 251-266.
- Hutton, J. L., & Monaghan, P. F. (2002). Choice of parametric accelerated life and proportional hazards models for survival data: asymptotic results. *Lifetime data analysis*, 8(4), 375-393.
- Kalbfleisch, J. D., & Prentice, R. L. 2002. *The Statistical Analysis of Failure Time Data* (Vol. 360). John Wiley & Sons.
- Kay, R., & Kinnersley, N. 2002. On the use of the accelerated failure time model as an alternative to the proportional hazards model in the treatment of time to event data: a case study in influenza. *Drug information journal*, 36(3), 571-579.
- Khanal, S. P., Sreenivas, V., & Acharya, S. K. 2014. Accelerated failure time models: an application in the survival of acute liver failure patients in India. *Int J Sci Res*, 3, 161-66.
- Klein, J. P. 2003; Moeschberger, ML. *Survival analysis: Techniques for censored and truncated data*. New York: Springer;
- Ko, J. (2017). Solving the Cox Proportional Hazards Model and Its Applications. <https://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-110.pdf>

- Kunnuji, M., Eshiet, I., & Nnorom, C. 2018. A survival analysis of the timing of onset of childbearing among young females in Nigeria: are predictors the same across regions?. *Reproductive health*, 15(1), 173.
- MacQuarrie, K. L. 2016. Marriage and Fertility Dynamics: The Influence of Marriage Age on the Timing of First Birth and Birth Spacing. dhs Analytical Studies, 56. *Recuperado de* <http://dhsprogram.com/pubs/pdf/AS56/AS56.pdf>.
- Nahar, M. Z., Zahangir, M. S., & Shafiqul Islam, S. M. 2013. Age at first marriage and its relation to fertility in Bangladesh. *Chinese Journal of Population Resources and Environment*, 11(3), 227-235.
- Nardi, A., & Schemper, M. 2003. Comparing Cox and parametric models in clinical studies. *Statistics in medicine*, 22(23), 3597-3610.
- Oakes, D. 1977. *The asymptotic information in censored survival data*. *Biometrika*, 64, 441-448.
- Obre J, Ferreira E, Nunez-Anton V. 2002. Comparing Proportional Hazards and Accelerated Failure Time Models for Survival Analysis. *Statistics in Medicine*: 21:3493-3510.
- OO, O., Akomolafe, A. A., & Musa, A. Z. 2018. Accelerated Failure Time Model with Application to Data on Tuberculosis/Hiv Co-Infected Patients in Nigeria.
- Patel, K., Kay, R., & Rowell, L. 2006. Comparing proportional hazards and accelerated failure time models: an application in influenza. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, 5(3), 213-224.
- Peng, Y., Yu, B., Kong, D. G., Zhao, Y. Y., Wang, P., Pang, B. B., & Gong, J. 2018. Reinfection hazard of hand-foot-mouth disease in Wuhan, China, using Cox-proportional hazard model. *Epidemiology & Infection*, 146(10), 1337-1342.
- Rahman, M., Mustafi, M., & Azad, M. 2013. Analysis of the determinant's of marriage to first birth interval in Bangladesh. *International Journal of Management and Sustainability*, 2(12), 208-219.
- Sood G.K. 2006. Acute liver failure. <http://emedicine.medscape.com/article/177354-overview>.

- Tadesse, F., & Headey, D. 2010. Urbanization and fertility rates in Ethiopia. *Ethiopian Journal of Economics*, 19(2), 35-72.
- Tolosie, K., & Sharma, M. K. 2014. Application of Cox proportional hazards model in case of tuberculosis patients in selected Addis Ababa health centres, Ethiopia. *Tuberculosis research and treatment*, 2014.
- Trey C, Davidson C.S. 1970. The management of fulminant hepatic failure. *Progress in Liver Diseases*, 3: 282 - 98.
- Wei, L. J. 1992. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, 11, 1871-1879.
- Whitehead, J. 1980. Fitting Cox's regression model to survival data using GLIM. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(3), 268-275.