Research Journal of Applied Sciences 8 (3): 215-220, 2013
ISSN: 1815-932X

Discrete Time Survival Analysis of Age at First Pregnancy Among Nigerian Women

O.M. Odeniya, A.A. Abiodun and B.A. Oyejola
Department of Statistics, University of Ilorin, Ilorin, Nigeria

Abstract: Discrete time survival analysis approach is often used when only the interval in which the event of failure occurs is known or the event itself occurs in discrete time scale. In this study, the approach was used to analyze data on age at first pregnancy among Nigerian women. Literature also reveals that, in some situations, in addition to observed covariates collected on each individual, there often exists unobserved heterogeneity in the data at individual or cluster level which if not accounted for during analysis may lead to biased estimates and unreliable conclusion. In this study, Discrete Time Logit Model was used to investigate the effect of some covariates (risk factors) on the hazard of age at first pregnancy. To account for unobserved heterogeneity (frailty) the random effects of ethnicity was added to the model. The data used for the study were extracted from the 2005 National HIV/AIDS and Reproductive Health Survey (NARHS). The results of the analysis showed that age at first pregnancy depended on geopolitical zone, location of residence, level of educational attainment, marital status, religion and the age of first sexual initiation. It was also observed, using Akaike Information Criterion (AIC) that the model that accounted for unobserved heterogeneity due to ethnicity was preferred to the one that did not.

Key words: Survival analysis, age at first pregnancy, Discrete Time Logit Model, random effects, akaike information criterion

INTRODUCTION

Pregnancy is sometimes viewed as an alternative path to economic independence and adult status (Brown and Barbosa, 2001). Pregnancy and the problems surrounding it are not specific to any age group but instead are treated as part of broader social issues. Analysis of survey data from 51 developing countries from the mid 1990s to the early 2000s showed that almost 10% of girls were mothers by age 16 with the highest rates in sub-Saharan Africa and South-Central and South-Eastern Asia (WHO, 2008). In 2005, the national median age at first pregnancy was 20 year in Nigeria.

In this study, researchers investigate the determinants of the timing of pregnancy among women in Nigeria. Many studies have shown that substantial numbers of teenagers have a positive or ambivalent attitude towards pregnancy (Jaccard *et al*., 2003; Condon *et al*., 2001; Stevens-Simon *et al*., 1996). A major area of research into planned and unplanned pregnancy concerns both teenage and adult pregnancy. Family structure, age at first sexual intercourse, future expectations, sexual abuse and knowledge, attitudes and beliefs are some of the factors that influence pregnancy. The interest of the study majorly dwells on examining the socioeconomic and demographic factors influencing the timing of first pregnancy among Nigerian women and the extent of variation in pregnancy rates within and across ethnicity. This requires fitting of discrete time survival model for age at first pregnancy and including random effects also known as frailty into the model in order to account for unobserved heterogeneity due to clustering of observations by ethnicity.

MATERIALS AND METHODS

Data description: The data used in this study were extracted from the 2005 National HIV/AIDS and Reproductive Health Survey (NARHS). NARHS is a Nationally Representative Household Survey of females (aged 15-49 years) and males (aged 15-64 years). The primary objective of NARHS is to provide information on the reproductive and sexual health situation as well as the factors that influence reproductive and sexual health and to provide data on the impact of ongoing HIV and reproductive health behaviours in Nigeria. For this study, a database from the main data of the survey for all female respondents aged 15-49 years was created and information were extracted on age at first pregnancy (age at which respondents got pregnant for the first time

Corresponding Author: A.A. Abiodun, Department of Statistics, University of Ilorin, Ilorin, Nigeria

expressed in years). This was considered as survival time and respondents who had never been pregnant as at the time of the survey were considered to be right censored. From a total of 10081 respondents, complete information was available for only 2183 females aged 15-49 years. It is believed that pregnancy can take place at any time after puberty with menarche (first menstrual period) normally taking place around age 12 or 13 (Onyiriuka *et al*., 2012). The analysis was therefore based on 2183 respondents.

Discrete time survival analysis: The two primary types of survival models often encountered in survival analysis are continuous time and discrete time. Continuous time models are typically employed when the exact timing of an event is known (e.g., numbers of days on admission before death). Discrete Time Models, in contrast, are generally used when only the interval in which an event occurs is known or the event itself occurs in discrete intervals (e.g., the year in which a respondent gets married the month of death of a child). One advantage of Discrete Time Model is that it allows for non-proportional hazards and time varying covariates. To record event occurrence in discrete intervals, continuous time is divided into an infinite sequence of contiguous time periods:

$$[0,a_1),[a_1,a_2),[a_2,a_3),[a_3,a_4),\ldots,[a_{t-1},a_t),[a_t,a_\infty)$$

Let T represent the discrete random variable that indicates the time period t when the event of failure occurs for a randomly selected individual from the population. Then, the discrete hazard function which is the conditional probability that an event will occur in time period t given that it has not already occurred in a earlier time period is given by:

$$h(t) = \Pr\left(T = t \middle| T \geq t\right) \tag{1}$$

and the discrete survival function is given by:

$$S(t) = \Pr(T > t) = \Pi\left(1 - h\left(t\right)\right) \tag{2}$$

which defines the probability of not experiencing the event up to (and including) time t. The probability of survival until the end of interval t is the product of probabilities of not experiencing event in each of the intervals up to and including the current one.

Data structure and model for discrete time analysis: The first step in a discrete time analysis is to restructure the data file into a person period format by expansion. One major drawback of the discrete time approach is that the expansion can lead to an extremely large dataset, particularly when time intervals are short and the observation period is long. One way to reduce the number of records is to use broader intervals and to weight by the exposure time within each interval.

Suppose that the data are collected on n individuals and survival information on each of them is recorded as t_i, δ_i i = 1, 2, 3, ..., n. From the observed data the event/censoring time t_i and the censoring indicator δ_i, a binary response y_{it} is created for each time interval t up to t_i which is coded as:

$$Y_{it} = \begin{cases} 1 & t = t_i,\ \delta_i = 1 \\ 0 & t = t_i,\ \delta_i = 0 \\ 0 & t < t_i \end{cases} \tag{3}$$

The restructured dataset in Eq. 3 is often called a person period formal. The failure process of individual i can then be considered as a sequence of binary response outcomes which follow a binomial distribution:

$$Y_{ti} = \begin{cases} 1 & \text{if } t = t_i \text{ and } \delta_i = 1 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where, δ_i is the censoring indicator which takes values 1 or 0 if the individual has failure event or is censored at time t_i respectively and the hazard function for an individual with covariate vector x_i can be given (Fahrmeir and Tut, 2001) as:

$$\Pr\left(y_{ti} = 1 \middle| x_i\right) = h\left(t \middle| x_i\right) \tag{5}$$

Discrete Time Survival Models: Discrete time survival models treat the indicators of event of interest at each time point as independent binary trials and use a series of Binary Regression Models at each time point (Allison, 1998; Arjas and Haar, 1987; Fahrmeir and Knorr-Held, 1997). The models often used are Logit and Probit Models. Logit Model is more commonly used and is often preferred for easy interpretation of parameter. Cox (1972) proposed an extension of the Proportional Hazards Model to discrete time by working with the conditional odds of getting the event at each time t_i given survival up to that point. Specifically, researcher proposed the model:

$$\frac{h\left(\dfrac{t}{x_i}\right)}{1 - h\left(\dfrac{t}{x_i}\right)} = \frac{h_0\left(t\right)}{1 - h_0\left(t\right)} \exp(x'_i\beta) \tag{6}$$

Where:

$h(t|x_i)$ = The hazard at time t for an individual with covariate values x_i, $h_0(t)$

$\exp(x'_i\beta)$ = The relative risk associated with covariate value x_i

By taking the log of Eq. 6, a model on the logit of the hazard or conditional probability of getting the event at t given survival up to that time is given as follows:

$$\text{logit}\, h\left(\frac{t}{x_i}\right) = \alpha_t + x'_i\beta \quad (7)$$

Where:

$\alpha_t = \text{logit}\, h_0(t)$ = The baseline logit-hazard

$x'_i\beta$ = The effect of the covariates on the logit of the hazard

The model in Eq. 7 essentially treats time as a discrete factor by introducing one parameter for each possible time t.

The discrete time logit model can be extended to account for unobserved heterogeneity (Lewis and Raftery, 1999; Biggeri *et al.*, 2001; Manda and Meyer, 2005) due to clustering of individuals. Suppose that the hazard, in addition to observed covariates x_i also depends on variable u ($u = u_1, u_2, ..., u_G$) which represents unobserved risk factors that are specific to groups (clusters) of individuals in the study then the standard approach to allow for such unobserved heterogeneity is to include it in the model as a random effect also known as frailty. In a Discrete Time Model, frailty is usually incorporated by including a normally distributed random effect in the linear predictor and is given by:

$$\text{logit}\, h\left(\frac{t}{x_i}\right) = \alpha_t + x'_i\beta + \mu_j \quad (8)$$

where, u_j is a random effect for individual due to cluster j. Ignoring this in survival data analysis may produce biased estimates especially when the variance of the unobserved clustering effect is large (Guo and Rodriguez, 1992).

Data analysis: The response variable for analyses in this study is age at first pregnancy of the respondent. The question on whether or not, the respondent has ever been pregnant was asked if yes her age at first pregnancy was asked and recorded in years and those who have never been pregnant were considered censored at their current age. The reported age at first pregnancy represents the event time.

In this study, apart from age at first pregnancy which is the response variable, information on covariates (categorical, continuous and frailty) which were thought to be associated with pregnancy were included in data analysis. These include: Geopolitical Zones South East, South West, South South, North West, North Central, North West (reference category); level of educational attainment koranic, primary, secondary and higher (reference category); current age of respondents (in years) which are grouped as: (15-19) (reference category), (20-24), (25-29), (30-39) and (40-49) years; marital status currently married, never married, widowed and divorced (reference category); location-rural (reference category), urban; religion Islam, Christianity others and no religion (reference category) whether the respondent is circumcised or not yes or no and age at first sexual initiation (in years) as the continuous covariate while ethnicity is frailty (random effect) covariate.

The dataset was restructured in a person period format as described in the study for Discrete Time Model. As a result of this expansion, the total number of observations used for this study increased from 2183-44936. Models (7) and (8) were then applied to the data. Model (7) included all the listed covariates as fixed effects while model (8) extended model (7) by including ethnicity as random effect (frailty). The essence of including random effects in model (8) is to enable us account for ethnic variations on age at first pregnancy among Nigerian women. All analyses were carried out using STATA (Version 10.0).

RESULTS

Results of the descriptive analysis: Distribution of respondents by geopolitical zone from North West (NW), North East (NE), North Central (NC), South West (SW), South East (SE), South South (SS) were 361 (16.5), 233 (10.7), 330 (15.1), 518 (23.7), 293 (13.4), 448 (20.5), respectively. With respect to locality, 938 (43.0%) were living in urban area and 1245 (57.0%) in rural area. By level of educational attainment, 262 (12.0%) have higher education while 298 (13.7), 728 (33.3), 895 (41.0) have koranic, primary, secondary education, respectively. There were 268 (12.3), 538 (24.6), 434 (19.9), 605 (27.7) and 338 (15.5%) in the age groups 15-19, 20-24, 25-29, 30-39 and 40-49 years, respectively. About 1314 (60.2%) were circumcised while the remaining 669 (39.8%) were not. With respect to their marital status, 1462 (67.0), 609 (27.9), 53 (2.4) and 59 (2.7%) were currently married, never married, divorced and widowed, respectively. About 788 (36.1%) practice Islam, 1378 (63.1%) were Christians while 17 (0.8%) practice other religions or have no religion.

Results of discrete time analysis: The results for models (7) and (8) are shown in Table 1 and 2, respectively. The two tables report the estimated Odds Ratio (OR), standard errors of the estimated regression parameters and the p-values (to justify the significance or otherwise of each predictor variable in the model) as well as the confidence intervals for the odds ratios. The contribution of each factor in the model is determined by its Odds Ratio (OR). Odd ratio measures the increase (+) or decrease (-) in the odds of age at first pregnancy due to the influence of the respective risk factor. The results obtained revealed significant change in the effects of some of risk factors on age at first pregnancy. As observed from Table 1, the estimated odds of early pregnancy for the first time among those that are in North West (NW), South West (SW) and South East (SE) are significantly different (p<0.001) from that of their counterparts in the Nort East (NE, reference category). Respondents from the NW have the highest odds (1.32) while women from SW and SE have the lowest odds (0.60 and 0.45, respectively). Respondents from NC, NE and SS have similar odds. Respondents living in urban areas have odds 0.92 of getting pregnant early (p = 0.004) and therefore get pregnant early compared to those living in rural areas.

Education is also found to be an important factor influencing age at first pregnancy. It is observed that the odds of having early first pregnancy is significantly (p<0.001) higher (1.49 and 1.25 times) for respondents with primary and secondary level of education, respectively compared to those with higher education while women with koranic, primary and secondary education are more likely (p<0.001) to get pregnant earlier relative to their counterparts with higher education. It is observed that the odds of early first pregnancy is 1.16 times and significantly (p<0.001) higher for those circumcised compared to those that are not circumcised.

Table 1: Results of Discrete-Time Logit Model (7) showing the odd ratios, standard errors and p-values

Covariates	Odds Ratio (OR)	SE (OR)	p-value	95% CI for OR
Zones				
North-West	1.3273	0.0739	<0.0010	(1.1902, 1.4803)
North-Central	0.9810	0.0547	0.7310	(0.8795, 1.0943)
South-West	0.6015	0.0332	<0.0010	(0.5398, 0.6702)
South-East	0.9230	0.0537	0.1690	(0.8235, 1.0345)
South-South	0.9230	0.0537	0.1690	(0.8235, 1.0345)
Urban	0.9246	0.0248	0.0040	(0.8772, 0.9746)
Level of education				
Koranic	1.3850	0.0247	<0.0010	(0.3395, 0.4365)
Primary	1.4912	0.6399	<0.0010	(1.3710, 1.6221)
Secondary	1.2514	0.0475	<0.0010	(1.1617, 1.3480)
Circumcised	1.1558	0.3163	<0.0010	(1.0954, 1.2194)
Age-group				
20-24	1.8399	0.0804	<0.0010	(1.6889, 2.0044)
25-29	2.8681	0.1365	<0.0010	(2.6126, 3.1485)
30-39	3.8790	0.1859	<0.0010	(3.5311, 4.2611)
40-49	3.6768	0.2011	<0.0010	(3.3030, 4.0929)
Marital status				
Currently married	2.8463	0.2005	<0.0010	(2.4793, 3.2676)
Never married	0.2712	0.0196	<0.0010	(0.2353, 0.3125)
Widowed	0.8373	0.0782	0.0570	(0.6973, 1.0056)
Religion				
Islam	0.6413	0.0418	<0.0010	(0.5678, 0.9599)
Christianity	0.5262	0.0232	<0.0010	(0.4956, 0.9578)
Others	0.9540	0.0768	0.0050	(0.7886, 0.9786)
Age at first sex	1.1046	0.0841	<0.0010	(1.0713, 1.5625)

Women within age-groups 20-24, 25-29, 30-39 and 40-49 years are significantly (p<0.001) more likely to get pregnant earlier than those aged 15-19 years. The odd ratios are 1.84, 2.88, 3.88 and 3.68, respectively. Those who practice Islam, Christianity and other religion are more likely (p = 0.005) to delay pregnancy than those who do not practice any religion. The odd ratios for those who practice Christianity, Islam and other religions are 0.53, 0.64 and 0.95, respectively. The respondent age at first sex is found to be a significant determinant of age at first pregnancy. As observed, a 1 year increase in age at first sexual initiation increases the age of getting pregnant by 1.105 years.

Table 2 shows the results of fitting random effects logit model. Here, ethnicity has been included as random effects. These results are similar to those in Table 1 where the random effects are not included. The odd ratios are

Table 2: Results of Discrete-Time Logit Model (8) with random effects of ethnicity showing the odd ratio, standard error, p-value and estimates of random effects parameters

Covariates	Odds Ratio (OR)	SE (OR)	p-value	95% CI for OR
Zones				
North-West	1.4060	0.0939	<0.001	(1.2335, 1.6025)
North-Central	0.8499	0.0525	0.008	(0.7529, 0.9593)
South-West	0.4874	0.0335	<0.001	(0.4260, 0.5577)
South-East	0.3485	0.0306	<0.001	(0.2934, 0.4140)
South-South	1.0024	0.0650	0.970	(0.8827, 1.1383)
Urban	0.9352	0.0265	0.018	(0.8847, 0.9887)
Level of education				
Koranic	1.3820	0.0257	<0.001	(0.3349, 0.4359)
Primary	1.4914	0.0647	<0.001	(1.3699, 1.6238)
Secondary	1.2411	0.0476	<0.001	(1.1512, 1.3379)
Circumcised	1.1140	0.0318	<0.001	(1.0533, 1.1782)
Age-group				
20-24	1.9129	0.0847	<0.001	(1.7540, 2.0862)
25-29	2.9514	0.1425	<0.001	(2.6850, 3.2442)
30-39	4.0421	0.1966	<0.001	(3.6746, 4.4464)
40-49	3.9130	0.2180	<0.001	(3.5082, 4.3645)
Marital status				
Currently married	3.0009	0.2143	<0.001	(2.6090, 3.4518)
Never married	0.2840	0.0209	<0.001	(0.2460, 0.3280)
Widowed	0.8527	0.0803	0.091	(0.7090, 1.0255)
Religion				
Islam	0.7634	0.0179	0 .997	(0.5274, 1.7094)
Christianity	0.6845	0.0186	0.979	(0.4217, 1.5024)
Others	0.9640	0.0124	0.987	(0.4217, 1.5024)
Age at first sex	1.1046	0.0841	<0.001	(1.0613, 1.4925)
Ethnicity	1.0342	0.0367	-	-
Sigma_u	1.0081	0.2706	-	-
Rho	0.2360	0.0968	-	-

Likelihood-ratio test of rho = 0: $\chi^2 = 284.59$, p<0.001

Table 3: Relative performance of the models with and without random effects

Models	No. of observations	Loglikelihood	AIC
Logit	44936	-20932.32	41906.64
Logit with random effects	44936	-20773.42	41590.84

only slightly different in most cases. For example, the respondent age at first sex is significant in the model; a 1 year increase in the age at sexual initiation decreases the odds of getting pregnant by a factor of 1.106 compared to 1.105 in model without random effects. The Sigma_u is the standard deviation of the heterogeneity while rho is the ratio of the heterogeneity variance. The estimate of (rho) of the frailty is 0.2360; a likelihood ratio test for the inclusion of random effect gives a value of 284.54 ($p<0.001$). Therefore, the frailty is important in the model. This implies that the addition of random effects gives account of unobserved risk factors due to ethnic variations among Nigerian females. It is therefore necessary that frailty be included in the analysis.

In order to know which of the two models (with or without random effects) fits the data better, Akaike Information Criterion (AIC) was used for model evaluation. The model with minimum AIC is often preferred since it is the model that minimizes the estimated information loss. The preferred model from Table 3 is therefore the model with random effect.

DISCUSSION

Discrete Time Survival Model was used with categorical and continuous covariates. Random effects were incorporated into the analysis in order to account for unobserved heterogeneity. Akaike Information Criteria (AIC) was used to select the preferred model for the data. Data on age at first pregnancy from National HIV/AIDS and Reproductive Health Survey were analysed. Ages at first pregnancy were compared based on geo political zones in Nigeria. Results showed that respondents from North West had highest risk of early pregnancy while respondents from South West and South East had lowest risk of early pregnancy. The other 3 zones had similar risk of early pregnancy. Based on the location of residence, the results revealed that those in urban areas had lower risk of getting pregnant earlier compared to those in rural areas. The results also showed that respondents with koranic education had lower risk of early pregnancy compared to those with higher education whereas those with primary and secondary education had higher risk. It was also observed that more respondents whose age fall between age 20-24, 25-29, 30-39 and 40-49 years at the time of the survey had earlier pregnancy than those whose age fall between 15-19 years. More respondents who were married had higher risk and those that never married got pregnant early compared to those that were separated. Those who practice Islam, Christianity and other religion had lower risk of early pregnancy and an increase in the age at first sexual initiation reduced the risk of getting pregnant early.

Since, the data is a national data, one main point is to investigate the geographical variation in age at first pregnancy. Geopolitical zone was therefore included as a factor in the analysis. In addition, ethnicity was included in the model to incorporate unobserved heterogeneity as random effects (frailty) in the study. Comparing models with and with no random effects, it was observed that model that accounted for frailty performed better than that which ignored it.

CONCLUSION

From the results of the analysis obtained, it can be concluded that the age at first pregnancy depends on geopolitical zone, location of residence, level of education attainment, marital status, religious belief and the age at first sexual initiation. It was also observed that the survival data contained some measure of unobserved heterogeneity (frailty) that should not be ignored during analysis. Accounting for unobserved heterogeneity in the categorical and continuous covariate provides more reliable results than when it is ignored.

REFERENCES

Allison, P., 1998. Discrete-Time Methods for the Analysis of Event Histories. In: Sociological Methodology, Leinhacdt, I.N.S. (Ed.). Jossey-Bass, San Francisco, pp: 61-68.

Arjas, E. and P. Haar, 1987. A logistic regression model for hazard: Asymptotic results. Scand. J. Statist., 14: 1-18.

Biggeri, L., M. Bini and L. Grilli, 2001. The transition from university to work: a multilevel approach to the analysis of the time to obtain the first job. J. R. Statist. Soc. Ser. A, 164: 293-305.

Brown, S.G. and G. Barbosa, 2001. Nothing is going to stop me now: Obstacles perceived by low-income women as they become self-sufficient. Public Health Nurs., 18: 364-372.

Condon, J.T., J. Donovan and C.J. Corkindale, 2001. Adolescents attitudes and beliefs about pregnancy and parenthood: Results from school-based intervention program. Int. J. Adelesc. Youth, 9: 245-256.

Cox, D.R., 1972. Regression models and life tables. J. Royal Stat. Soc. Ser. B (Methodological), 34: 187-220.

Fahrmeir, L. and G. Tut, 2001. Multivariate Statistical Modeling Based on Generalized Linear Models. 2nd Edn., Springer-Verlag, New York, USA., ISBN-13: 9780387951874, Pages: 548.

Fahrmeir, L. and L. Knorr-Held, 1997. Dynamic discrete-time duration models: Estimation via Markov Chain Monte Carlo. Soc. Methodol., 27: 417-452.

Guo, G. and G. Rodriguez, 1992. Estimating a multivariate proportional hazards model for clustered data using the EM algorithm, with an application to child survival in Guatemala. J. Am. Statis. Assoc., 87: 969-976.

Jaccard, J., T. Dodge and P. Dittus, 2003. Do adolescents want to avoid pregnancy? Attitudes towards pregnancy as predictors of pregnancy. J. Adolesc. Health, 33: 79-83.

Lewis, S.M. and A.E. Raftery, 1999. Comparing explanations of fertility decline using event history models and unobserved heterogeneity. Sociol. Methods Res., 28: 35-60.

Manda, S. and R. Meyer, 2005. Age at first marriage in Malawi: A Bayesian multilevel analysis using a discrete time-to-event model. J. Royal Statis. Soc. A, 168: 439-455.

Onyiriuka, A.N., P.O. Abiodun, R.C. Onyiriuka, F.A. Ehirim and E.P.A. Onyiriuka, 2012. Menarcheal age of Nigerian urban secondary school girls in Benin City. Reprod. Sys. Sex. Discord.

Stevens-Simon, C., L. Kelly, D. Singer and A. Cox, 1996. Why pregnant adolescents say they did not use contraceptives prior to conception. J. Adoles. Health, 19: 48-55.

WHO, 2008. Making pregnancy safer notes, volume 1. Department of Making Pregnancy Safer. World Health Organization, Geneva, Switzerland, October, 2008.