



PARAMETRIC DURATION MODELS FOR CENSORED AND UNCENSORED DATA

UPAHI, E.O., ABIODUN, A.A. AND ADENIYI, O.I.*

Department of Statistics, University of Ilorin, Ilorin, Nigeria.

ABSTRACT

Censoring is said to be present in data when information on time to outcome event is not available for all the study participants. The purpose of this study was to compare some parametric survival models when data are censored and when uncensored. The study used duration of labour data obtained from Record and files Department, Federal Medical Centre, Lokoja. Parametric proportional hazards model, namely, Exponential, Weibull and Gompertz models were fitted when the data were censored and when uncensored. Only variable full housewife was found to be significant in predicting duration of labour. Also, it was found that the three models had better fits when the data were censored than when they were uncensored. The study concludes that it is better to include censored observation when analysing duration data despite the incomplete information they provide.

Keywords: Censoring, proportional hazard, duration of labour, AIC, uncensored.

***Correspondence:** adeniyi.oi@unilorin.edu.ng

INTRODUCTION

Duration models are widely applied in many fields such as biology, medicine, public health, and epidemiology. A typical analysis of duration data involves the modelling of time-to-event such as the time until death. The time to the event of interest is called either survival time or failure time. An intrinsic characteristic of such data is inclusion of an individuals whose actual time until the occurrence of event is not observed. Such an individual is said to be censored. Censoring can arise as a result of withdrawal from the experiment or termination of the experiment. Because the response is usually a duration, some of the possible events may not yet have occurred when the period of data collection has terminated. For example, the clinical trials are conducted over time, and thus the length of follow-up varies by individuals; consequently, the time to the event may not be ascertained on all patients in the study. Also, some of the subjects may be lost to follow-up. For example, a subject may move or refuse to continue to participate in an experiment before termination of such experiment. In either case, only a lower bound on the failure time of the censored observation is known. These observations are said to be right censored. Thus, an additional variable is incorporated into the analysis to indicate which failure time is observed and which one is censored. More generally, the failure time might only be known to be smaller than a given value (left censored) or known to be within a given interval (interval censored). There are numerous possible censoring schemes that arise in survival analysis. Maddala [1] discusses several related types of censoring situations, the text by Klein and Moeschberger [2] discussed techniques for censored and truncated data and the text by Kalbfleisch and Prentice [3] also discusses several censoring schemes. Data with censored observations cannot be adequately analysed by ignoring such observation as this may result in estimates that do not represent the true survival function. The method of analysis must therefore take the censoring into account. A key assumption in

analysing censored data is that censoring distribution must be independent of event times Guo and Manatunga [4] and Abiodun *et al* [5]. Broadly classified, two types of censoring are often encountered, which are point and interval censoring Leung *et al* [6]. A point censoring is said to occur when despite continuous monitoring of outcome event, the patient is lost to follow-up or event does not occur within the study duration. It is also known as right censoring which can be either end-of-study censoring or loss- to-follow-up censoring. A subject is said to be censored if it has been on the risk set right before entering the study. However, left censoring is usually a problem in clinical trial, since starting point is defined by an event such as entry of patient in trial, randomization or occurrence of a procedure or treatment. Another form of censoring which commonly occurs in survival data is interval censoring. Interval censored data arises when failure time T can only be known to lie in an interval obtained from a sequence of examination times. Kongerud and Samuelsen [7] and Samuelsen and Kongerud [8] reported two studies on respiratory symptoms and asthmatic symptoms among Norwegians aluminum worker. In these studies, the time to development of respiratory symptoms or aesthetic symptoms is only known to be between two health examinations. Other examples of interval censored data in animal carcinogenicity and epidemiology studies have been reported Hoel and Walburg [9], Finkelstein and Wolfe [10], Finkelstein [11] and Self and Grossman [12]. Diamond *et al.* [13] and Diamond and McDonald [14] contain examples of interval-censored data from demography studies. The aim of our study is to apply Parametric Proportional Hazard Models (PPHM) to labour duration data in child delivery. Primarily with the interest of examining the importance of including censored observations (incomplete information) in analysing duration data rather than ignoring them.

MATERIALS AND METHODS

Proportional hazards model

Proportional hazards model is also known as multiplicative hazard model or log relative hazard model. It is generally written as

$$h(t, x) = h_0(t) \lambda \quad (3)$$

where $h_0(t)$ is the baseline hazard function, which depends on t but not on x . It summarizes the pattern of duration dependence, assumed to be common to all individuals. $\lambda = \exp(\beta'x)$ is a person-specific non-negative function of covariates x which does not depend on t , but by construction, scales the baseline hazard function common to all persons. For proportional hazards model (3), the explanatory variables act multiplicatively on the baseline hazards Ghilagaber [15].

A measure of how often a particular event happens in one group compared to the other is the hazard ratio. For example, a hazard ratio of one means that there is no difference in survival between the two groups. A hazard ratio of greater than or less than one means that survival was better in one of the groups. Hazard Ratio (HR) is mathematically defined as the hazard for one individual divided by the hazard for another individual and given by

$$HR = \frac{\hat{h}(t, X^*)}{\hat{h}(t, X)} \quad (4)$$

where X^* is the set of predictors for one individual and X is the set of predictors for the other individual.

Parametric survival hazard models

Exponential model

Exponential model is the simplest of parametric proportional hazard models. It assumes that the baseline hazard of an event is constant over time, which implies that the conditional probability of an event is the same, no matter when the observation is observed. It can accommodate both proportional hazard and accelerated failure time. Therefore, the Exponential distribution with rate λ is characterized by the density function given by

$$f(t, \lambda) = \lambda \exp(-\lambda t), \quad (5)$$

the survival function given by

$$s(t, \lambda) = \exp(-\lambda t) \quad (6)$$

and the hazard function given by

$$h(t, \lambda) = \lambda, \quad (7)$$

where $\lambda > 0$.

The model is stated in terms of hazard by parameterizing λ as

$$h(t) = \exp(\beta_0 x_1 + \beta_1 x_1 + \dots + \beta_p x_p) \quad (8)$$

Weibull model

Weibull model has a hazard that is monotonically increasing or decreasing. It can also accommodate both proportional hazards and accelerated failure time. The density function of Weibull distribution is

$$f(t, \lambda, \gamma) = \lambda \gamma t^{\gamma-1} \exp(-\lambda t^\gamma) \quad (9)$$

with survival function

$$s(t, \lambda, \gamma) = \exp(-\lambda t^\gamma) \quad (10)$$

and hazard function

$$h(t, \lambda, \gamma) = \lambda \gamma t^{\gamma-1} \quad (11)$$

$$\lambda > 0, \gamma > 0$$

γ is called the shape parameter with decreasing hazard over time when $\gamma < 1$, increasing hazard when $\gamma > 1$ and constant hazard (Exponential) when $\gamma = 1$. By parameterizing λ , the model is stated in terms of hazard as

$$h(t, x) = \gamma t^{\gamma-1} \exp(\beta_0 x_1 + \beta_1 x_1 + \dots + \beta_p x_p). \quad (12)$$

Gompertz model

Gompertz is also a proportional hazards model which assumes a hazard that is monotonically increasing or decreasing. It was used frequently by medical researchers and biologists in modelling the mortality ratio data. The respective density, survival and hazard functions are given as

$$h(t, \lambda, \gamma) = \lambda \exp(\gamma t) \quad (13)$$

$$S(t, \lambda, \gamma) = \exp\left[\frac{\lambda}{\gamma}(1 - \exp(\gamma t))\right] \quad (14)$$

$$f(t, \lambda, \gamma) = \lambda \exp(\gamma t) \exp\left[\frac{\lambda}{\gamma}(1 - \exp(\gamma t))\right] \quad (15)$$

where $\lambda > 0, t \geq 0, \gamma(-\infty, \infty)$

The model is stated in terms of hazard by

$$h(t) = \exp(\gamma t) \times \exp(\beta_0 x_1 + \beta_1 x_1 + \dots + \beta_p x_p) \quad (16)$$

Model comparison

Several methods are often used for comparing models in Generalized Linear Models (GLM). One popular method is the Akaike's Information Criterion (AIC), which is expressed as $AIC = -2\log L + 2p$, where L is the likelihood and p is the number of estimated parameters. Lower AIC values indicates a better model fit.

RESULTS AND DISCUSSION

This study used duration of labour data obtained from Record and Files Department, Federal Medical Centre, Lokoja. The response variable was duration of labour before child delivery, recorded in hours. Survival time was defined as time to normal delivery and those who delivered through caesarean section were right censored at the time the operation was carried out. Explanatory variables which were thought to be related to baby delivery were also collected. These include mother's age, Sex of the baby with male as reference category; parity which was categorized as first delivery or previous delivery; mother's occupation categorized as civil servant, self-employed, full house wife and other occupations (reference category); religion with Christianity as reference category.

The data are analyzed using parametric proportional hazards models namely, Exponential, Weibull and Gompertz models, first for subjects with only observed events (uncensored) and also with inclusion of censored observations. The results are presented in Tables 1-6.

Results of Exponential model for uncensored data are presented in Table 1. None of the variables is statistically significant. However, as observed, holding all other variables constant, with increase in age, the mothers are equally likely to have a normal delivery as having caesarean section. Mothers are 0.856 less likely to have normal delivery to a female child compared to a male child. The expected hazard of having a normal delivery is 0.897 times in mothers delivering a child for the first time compared to mothers that have had one or more previous deliveries. The hazard of having a normal delivery is 0.905 times for civil servants, 1.115 for self-employed and 1.302 for full housewives compared to mothers of other occupations. Mothers who are Muslims are less likely to have normal child delivery than their Christian counterparts.

Table 2 presents Exponential model for censored data. As observed, being a full housewife is statistically significant in predicting duration of labour, with a P-value 0.007 less than 0.05 level of significance. The pattern of results for the uncensored

data are the same in Weibull model (Table 3) and Gompertz model (Table 5) with Exponential model (Table 1). However, the effect of being a full housewife is observed to be highly significant in these models as against Exponential model where it is not significant. Comparing Weibull model for uncensored and censored data, the variable full housewife is observed to be statistically significant in predicting duration of labour in both cases. For Gompertz model also, the variable full housewife is the only variable that is statistically significant in predicting labour duration when data are censored and when uncensored.

Table 7 presents the values of log likelihood and AIC for the three models under investigation, namely; Exponential, Weibull and Gompertz models. Lower AIC indicates a better fit.

It is observed from the table that the three models have lower AIC when the data are censored compared to when they are uncensored. The AIC values presented in Table 7, it was revealed that each of these models gives a better fit when the data are censored than when they are uncensored. It is however observed that Weibull model has the least AIC under the uncensored data.

Comparing the models within the censored data and from the results, it is observed that Gompertz model has the lowest AIC with the value 703.309, indicating a best fit than other models.

The Kaplan-Meier survival curve shown in Fig 1-4 reveals that the median delivery time for a female child is 4.3 hours and 5.5 hours for male child, 6.0 hours for women who are civil servants by profession, 5.42 for the self-employed women, 3.6 hours for housewives and 5.0 hours for women in other categories of occupation; also, the median delivery time for first child delivery is 4.6 hours and 4.7 hours for other delivery while the median delivery time for Muslim women is 4.4 hours and 4.9 hours for Christian women. The log rank test for the equality of the survival experience for the categories of the covariates as shown in Table 8 reveals that the survival experience for the baby's gender, parity and religion of the mother are the same but are not the same for the mother's occupation.

Table 1: Exponential model for uncensored data

Variable	Haz. Ratio	Std. Err.	P > z	(95% Conf. interval)
Mothers age	1.007	0.015	0.628	0.9791.036
Female	0.856	0.109	0.223	0.6671.099
First delivery	0.897	0.141	0.491	0.6591.221
Civil servant	0.905	0.206	0.66	0.5791.413
Self-employed	1.115	0.262	0.642	0.7041.767
Full housewife	1.302	0.183	0.061	0.9881.714
Islam	0.945	0.122	0.665	0.7341.218

Table 2: Exponential model for censored data

Variable	Haz. ratio	Std. err.	P > z	(95% Conf. interval)
Mothers age	0.993	0.016	0.641	0.9621.024
Female	0.895	0.124	0.421	0.6821.173
First delivery	0.792	0.138	0.182	0.5631.115
Civil servant	0.994	0.243	0.982	0.6161.604
Self-employed	1.072	0.289	0.796	0.6321.819
Full housewife	1.497	0.225	0.007*	1.1152.009
Islam	0.929	0.131	0.604	0.7051.225

Asterisk (*) indicate the variables that are significant at 5% level of significance

Table 3: Weibull model for uncensored data

Variables	Haz. ratio	Std. err.	P > Z	(95% Conf. interval)
Mothers age	1.009	0.015	0.522	0.9811.039
Female	0.822	0.106	0.128	0.6381.058
First delivery	0.902	0.141	0.509	0.6641.225
Civil servant	0.882	0.201	0.582	0.5651.378
Self-employed	1.188	0.279	0.464	0.7291.882
Full housewife	1.411	0.201	0.015*	1.0681.865
Islam	0.931	0.121	0.583	0.7221.201

Asterisk (*) indicate the variables that are significant at 5% level of significance

Table 4: Weibull model for censored data

Variable	Haz. ratio	Std. err.	P > Z	(95% Conf. interval)
Mothers age	0.994	0.016	0.728	0.9641.026
Female	0.868	0.121	0.31	0.6591.141
First delivery	0.799	0.138	0.197	0.5691.122
Civil servant	0.969	0.237	0.897	0.6001.564
Self-employed	1.127	0.305	0.657	0.6641.915
Full housewife	1.603	0.244	0.002*	1.1902.159
Islam	0.916	0.129	0.533	0.6941.208

Asterisk (*) indicate the variables that are significant at 5% level of significance

Table 5: Gompertz model for uncensored data

Variables	Haz. ratio	Std. err.	P > Z	(95% Conf. interval)
Mothers age	1.008	0.015	0.61	0.9791.037
Female	0.821	0.106	0.129	0.6371.059
First delivery	0.876	0.137	0.398	0.6461.189
Civil servant	0.879	0.199	0.569	0.5631.371
Self-employed	1.238	0.291	0.366	0.7791.964
Full housewife	1.443	0.207	0.011*	1.0891.912
Islam	0.923	0.119	0.532	0.7151.189

Asterisk (*) indicate the variables that are significant at 5% level of significance

Table 6: Gompertz model for censored data

Variable	Haz. ratio	Std. err.	P > Z	(95% Conf. interval)
Mothers age	0.993	0.016	0.659	0.9621.025
Female	0.863	0.121	0.294	0.6551.137
First delivery	0.779	0.135	0.15	0.5561.094
Civil servant	0.953	0.233	0.843	0.5911.538
Self-employed	1.186	0.321	0.528	0.6982.017
Full housewife	1.664	0.256	0.001*	1.2312.250
Islam	0.901	0.128	0.463	0.6831.189

Asterisk (*) indicate the variables that are significant at

5% level of significance

Table 7: Log likelihood and AIC of model comparison for uncensored and censored data

Distribution	Criterion	Uncensored	Censored
Exponential	Log likelihood	-360.196	-351.937
	AIC	738.392	721.874
	Log likelihood	-346.982*	-343.978
	AIC	711.964*	705.956
Gompertz	Log likelihood	-349.227	-342.654*
	AIC	716.454	703.308*

Asterisk (*) indicate the variables that are significant at 5% level of significance

CONCLUSION

The study applied parametric proportional hazards models, namely: Exponential, Weibull and Gompertz models to data on duration of labour before childbirth. The models were fitted under censored and uncensored data structure. Only variable full housewife was significant in predicting labour duration. For Exponential model, the variable was significant only when the observations were censored. For Weibull and Gompertz models, it was significant under the two data structure.

Akaike Information Criterion (AIC) revealed that Weibull model had the best fit when the data were uncensored data while Gompertz model was best for the labour duration data for censored observations. The three models performed better with censored data than uncensored. The finding in this study establishes that it is better to include censored observation when analyzing duration data despite the incomplete information they provide.

REFERENCES

- MADDALA, G.S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, New York, USA, pp. 149-196.
- KLEIN, J.P. & MOESCHBERGER, M.L. (1997). *Survival Analysis. Techniques for Censored & Truncated Data*. Springer-Verlag, New York, USA, pp. 290-301.
- KALBFLEISCH, J.D. & PRENTICE, R.L. (1980). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, Inc., New York, USA, pp. 148-191.
- GUO, Y. & MANATUNGA, A.K. (2009). Measuring agreement of multivariate discrete survival times using a modified weighted Kappa coefficient. *Biometrics*, **65**: 125-134.
- ABIODUN, A.A., OYEJOLA, B.A., ADENIYI, O.I. & ODENIYA, O.M. (2014). Nested frailty model for infant mortality data. *ICASTOR Journal of Mathematical Science*, **1**: 39 – 53.
- LEUNG K.M., ELASHOFF, R.M. & AFIFI A.A. (1997). Censoring issues in survival analysis. *Annual Review of Public Health*; **18**: 83-104
- KONGERUD, J. & SAMUELSEN, S.O. (1991). A longitudinal study of Respiratory Symptoms in Aluminum Potroom Workers. *American Review of Respiratory Diseases*, **144**: 10-16.
- SAMUELSEN, S.O. & KONGERUD, J. (1993). Evaluation of Applying Interval Censoring on Longitudinal Data on Asthmatic Symptoms. Statistical Research Report 2, Institute of Mathematics, University of Oslo.
- HOEL, D.G. & WALBURG, H.E. (1972). Statistical analysis of survival experiments. *Journal of the National Cancer Institute*, **49**: 361-372.
- FINKELSTEIN, D.M. & WOLFE, R.A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* **41**: 933-945.
- FINKELSTEIN, D.M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, **42**: 845-854.
- SELF, S.G. & GROSSMAN, E.A. (1986). Linear rank tests for Interval-censored Data with Application to PCB levels in Adipose tissue of Transformer Repair Workers. *Biometrics*, **42**: 521-530.
- DIAMOND, I.D. & MCDONALD, J.W. and SHAH, I.H. (1986). Proportional hazard models for current status data: application to the study of differentials in age at weaning in Pakistan. *Demography*, **23**: 607-620.
- DIAMOND, I.D. & MCDONALD, J.W. (1991). *The Analysis of Current Status Data. Demographic Applications of Event History Analysis*, Eds. J. Trussell, R. Hankinson, Oxford University Press, Tilton, Oxford, U.K., pp. 231-252.
- GHILAGABER, G., ELISA, W. & GYIMAH, S.O. (2014). *A Family of Flexible Parametric Duration Functions and Their Applications to Modelling Child Spacing in Sub-Saharan Africa*. A Springer Dordrecht Heidelberg New York, London, pp. 185-209.