J. biosoc. Sci. 2005 **37**, 471–479
© 2004 Cambridge University Press
doi:10.1017/S0021932004006984

APPLYING A DATA DUPLICATION TECHNIQUE IN LINEAR REGRESSION ANALYSIS OF WAITING TIME TO PREGNANCY

HALIMAH AWANG

Faculty of Economics and Administration, University of Malaya, Kuala Lumpur, Malaysia

Summary. This analysis demonstrates the application of a data duplication technique in linear regression with censored observations of the waiting time to third pregnancy ending in two outcome types, using data from Malaysia. The linear model not only confirmed the results obtained by the Cox proportional hazards model, but also identified two additional significant factors. The method provides a useful alternative when Cox proportionality assumption of the hazards is violated.

Introduction

Studies on the determinants of birth intervals and their effect on fertility, infant and child mortality and health have been well documented (Trussell *et al*., 1985; DaVanzo & Starbird, 1991; Miller *et al*., 1992; Curtis *et al*., 1993; Kuate Defo, 1997; Awang, 2003). Previous research on pregnancy interval lengths has concentrated on the study of the determinants of the time to pregnancy ending in only a live birth: that is, the analysis is of inter-live birth intervals. However, it is possible that a pregnancy could end in a stillbirth, miscarriage or an induced abortion. With multiple outcomes the analysis should take these competing risks into consideration.

Separate analyses could be carried out for each specific outcome type by treating other outcome types as censored data. This approach does not treat the different outcome types jointly, complicating the comparison of the effects of different outcome types. Another approach is to treat the multiple outcome types simultaneously in the same model, thus facilitating comparisons of their effects on the waiting time. Previous research in competing risk analysis used models that are not only complex in nature, but which also require advanced computational programmes and software that are not readily available (Larson & Dinse, 1985; Pandey & Suchindran, 1988; Kuk, 1992; Narendrathan & Stewart, 1993). In this paper, the different types of outcomes are accommodated in the analysis using a data duplication technique, as

https://doi.org/10.1017/S0021932004006984 Published online by Cambridge University Press

Table 1. Third pregnancy outcome versus second pregnancy outcome

Third pregnancy	Second pregnancy Live birth	Non-live birth	Total
Live birth	840	75	915
Non-live birth	77	14	91
Censored	255	19	274
Total	1172	108	1280

described by Lunn & McNeil (1995). While Lunn and McNeil applied this technique in competing risk analysis using the Cox proportional hazards model, the present analysis incorporates the data duplication into censored linear regression analysis of the waiting time to third pregnancy using data from the Second Malaysia Family Life Survey (Awang, 2003).

Methods

Sample

There were 1280 women in the study who had completed their second pregnancy, of which 1006 had also completed their third conception. Completion here refers to a pregnancy termination leading to any outcome and the waiting time to pregnancy is the time from the date of second pregnancy completion to conception leading to the third pregnancy outcome. Even though there are four types of pregnancy outcome in the original data set, for the purpose of this analysis they are classified into just two: a live birth and a non-live birth. The latter category encompasses the three possible non-live outcomes: a stillbirth, a miscarriage and induced abortion. The reason for combining all the non-live births is that the sample size for each of these categories is small. Of the total number of women in this interval, 915 had completed their third pregnancy ending in a live birth, 91 had non-live births and 274 were censored. The censored cases consist of women who did not become pregnant and those who were currently pregnant at the end of the study period. Although the waiting time to pregnancy is known for women who were currently pregnant, they are considered censored in the present analysis because the outcomes of their pregnancies were not yet known. The cross-tabulation between the third pregnancy outcome and the second pregnancy outcome for this sample is shown in Table 1. The term 'birth' as used in this study refers to a pregnancy outcome consisting of both live and non-live outcomes.

Statistical methods

The method proposed by Lunn & McNeil (1995) requires that the waiting time for each woman is duplicated for each outcome type, where each duplicate is considered a censored case for the other outcome type. To further illustrate the method, consider

https://doi.org/10.1017/S0021932004006984 Published online by Cambridge University Press

a dichotomous pregnancy outcome type, denoted by d_i=1 if woman i has a live birth at time t_i, and d_i=0 if woman i has a non-live birth at time t_i. The record for each woman in the sample is entered twice to accommodate the two pregnancy outcomes as follows:

Woman	Waiting time	Status	Outome type	Covariate
i	t_i	pregnant	d_i	z_i, $d_i z_i$
i	t_i	censored	$1-d_i$	z_i, $(1-d_i)z_i$

When woman i conceives with a pregnancy outcome of type d_i at time t_i, the repeat observation is considered censored of outcome type $(1-d_i)$ having equal waiting time. If an event is censored (failing to conceive), then there will be two censored cases of equal waiting time: one for pregnancy outcome type d_i and the other for outcome type $(1-d_i)$. The data duplication method assumes that the competing risks are independent.

The waiting time to pregnancy ending in either outcome is the minimum of the two waiting times, and the risk of conception leading to either a live birth or a non-live birth is the combination of the risk processes associated with the two outcomes. The augmentation of covariates z_i allows for possible interactions with each type of pregnancy outcome ($d_i z_i$, $(1-d_i)z_i$). Data augmentation provides a solution to the incomplete data structure problem arising from censored observations, thus allowing regression to be run on the duplicated covariates without overparameterizing the model.

The censored linear regression model is of the form:

$$y_i = b_0 + b_{1j} z_{ij} + b_{2i} d_i + b_{3i} d_i z_{ij} + \varepsilon_i,$$

where $z_{ij}(j=1,\ldots,k)$ are explanatory variables, d_i is the failure type which takes a value 1 if a pregnancy ends in a live birth and 0 otherwise, and $d_i z_{ij}$ is the interaction term. The error terms ε_i are assumed to be independent and normally distributed with mean 0 and common variance σ^2. The regression parameters are estimated by the iterative maximum likelihood method using EM algorithm as discussed by Aitkin (1981). The model is fitted to the data set using the transformation $Y=\log_2(2+\text{Waiting Time})$, which gives the closest approximation to a symmetric distribution.

Variables

The explanatory variables selected for the analysis are: third birth outcome type, contraceptive method use, maternal education, the outcome type of second birth, age, the length of waiting time to the second pregnancy (LPI), ethnic group, the period of second pregnancy termination, place of residence, and work status of the woman. The outcome types of the third and second pregnancy are classified into two categories: a live birth and a non-live birth outcome using the latter as the referent. The length of previous interval (LPI) is classified into three categories: LPI<12 months; LPI 12–24 months; and LPI>24 months. Similarly, contraceptive method used is grouped into: use of contraceptive pills within 3 months of the second pregnancy termination; use of pills>3 months after the second pregnancy termination; use of any other

https://doi.org/10.1017/S0021932004006984 Published online by Cambridge University Press

Table 2. Waiting time to third pregnancy with two outcome types: censored linear regression analysis

Variable	Reduced model			Final model		
	Coeff.	SE	*p*	Coeff.	SE	*p*
Live third birth outcome	−2·444	0·584	0·000	−2·712	0·362	0·000
Pill						
Begin 0–3 mths	0·785	0·308	0·011	1·321	0·143	0·000
Begin 3+ mths	0·881	0·442	0·046	1·527	0·215	0·000
Other modern	0·207	0·446	0·642	0·431	0·406	0·288
Begin 0–3 mths						
Begin 3+ mths	0·685	0·630	0·277	1·717	0·360	0·000
Rhythm/withdrawal	0·813	0·369	0·027	1·353	0·175	0·000
Herbals	−0·018	0·441	0·967	0·540	0·201	0·007
Unknown	0·753	0·291	0·010	0·740	0·126	0·000
Live previous outcome	1·487	0·309	0·000	1·264	0·296	0·000
Age (years)						
20–24	0·633	0·274	0·021	0·248	0·131	0·059
25–29	1·102	0·320	0·001	0·713	0·253	0·005
30+	2·204	0·553	0·000	1·886	0·515	0·000
Education						
Primary	0·354	0·310	0·253	0·216	0·146	0·140
Secondary	0·680	0·353	0·054	0·424	0·163	0·009
LPI (months)						
12–24	0·271	0·257	0·291	0·239	0·117	0·042
24+/unknown	0·459	0·251	0·067	0·472	0·240	0·049
Previous birth						
1970–79	0·009	0·304	0·976	0·115	0·135	0·396
1980–88	−0·782	0·319	0·014	−0·742	0·228	0·001
Ethnicity						
Chinese	−0·357	0·263	0·175	0·075	0·122	0·542
Indian	−0·621	0·248	0·012	−0·445	0·221	0·044
Interaction between third birth outcome type with:						
Pill						
Begin 0–3 mths	0·676	0·348	0·052			
Begin 3+ mths	0·805	0·505	0·111			
Other modern						
Begin 0–3 mths	1·361	0·527	0·010	1·104	0·474	0·020
Begin 3+ mths	1·402	0·760	0·065			
Rhythm/withdrawal	0·684	0·419	0·103			
Herbal	0·692	0·495	0·162			
Unknown	−0·017	0·324	0·959			
Education						
Primary	−0·187	0·352	0·595			
Secondary	−0·339	0·398	0·394			
Live previous outcome	−0·754	0·361	0·037	−0·495	0·347	0·154

https://doi.org/10.1017/S0021932004006984 Published online by Cambridge University Press

Table 2. *Continued*

Variable	Reduced model			Final model		
	Coeff.	SE	*p*	Coeff.	SE	*p*
Interaction between third birth outcome type with *continued*:						
Age (years)						
20–24	− 0·498	0·313	0·111			
25–29	− 0·771	0·364	0·034	− 0·264	0·259	0·308
30+	− 1·546	0·606	0·011	− 1·110	0·546	0·042
Ethnicity						
Chinese	0·558	0·297	0·061			
Indian	0·728	0·282	0·010	0·516	0·248	0·037
Previous birth						
1970–79	0·161	0·340	0·636			
1980–88	0·885	0·359	0·014	0·802	0·227	0·000
LPI (months)						
12–24	− 0·067	0·289	0·817			
24+/unknown	0·646	0·288	0·025	0·637	0·274	0·020
− 2 log likelihood		4680·49			4699·42	

Reference groups: non-live failure type; no contraception; non-live previous birth; age <20 years; no education; LPI <12 months; previous birth year 1960–69; ethnicity Malay.

modern method within 3 months of the second pregnancy termination; use of any other modern method>3 months after; rhythm or withdrawal; herbal method; 'unknown' contraceptive practice; and no contraception.

About 45% of the total 1280 women in this interval practised some form of contraception, 31% did not use any, while the remaining 24% did not respond to the question on contraceptive method used (unknown). Among the users of contraception 48% used oral contraceptives, 15% used other modern methods, 24% practised either rhythm or withdrawal, and 13% used herbal or folk methods.

The husband's education and occupation, the sex of previous child and the duration of breast-feeding are excluded because they were found to be statistically not significant in the analyses of the waiting time to the third pregnancy ending in a live birth (Awang, 1997, 2003). For sex of previous child, not only was it found to be not statistically significant, but also the sex of the child could not be determined for many cases of pregnancy termination ending in a non-live birth.

Treating outcome type as a covariate in the model indicates if there is a significant difference in the length of waiting time to pregnancy ending in a live birth as compared with a non-live birth. The McNemar chi-squared statistic for testing the null hypothesis of no association between the second and third pregnancy outcomes, ignoring the censored cases, is 0·026. This does not give a statistically significant *p*-value, thus showing no evidence of an association between the two consecutive outcomes (a 95% confidence interval for the odds ratio is (0·708, 1·344)). While this finding is consistent with the result obtained by Majinge & Lema (1993), they found

https://doi.org/10.1017/S0021932004006984 Published online by Cambridge University Press

that having an adverse preceding pregnancy outcome shortens a woman's waiting time to her next pregnancy.

Duration of breast-feeding is not included in this analysis because of its high correlation with the previous pregnancy outcome. This is due to the fact that a non-live pregnancy outcome essentially eliminates the possibility of a woman breast-feeding, and the inclusion of both variables in the model is redundant. Janowitz & Smith (1984) also found that in the presence of contraception, the effect of breast-feeding on birth interval length is minimal.

Results

First, a full model is fitted in which all the selected explanatory variables are incorporated as covariates, including the interaction terms between third pregnancy outcome and each of the predictor variables. With the exception of place of residence and woman's work status, the other predictor variables show statistically significant coefficients either through their main effects or through their interactions with birth outcome type. Place of residence and woman's work status, together with their interaction terms, are omitted from the full model to obtain the reduced model, the results of which are presented in the left-hand panel of Table 2.

The presence of interaction terms between the third birth outcome type and the predictor variables complicates the interpretation of the regression coefficients. This is because the effect of any explanatory variable in the model with interactions cannot be interpreted without considering the effect of the variables that interact with it. For example, the coefficient associated with a live previous pregnancy outcome is 1·487 giving a lengthening effect of 2 to the power of 1·487 on the waiting time, or 2·803 times that of a non-live previous birth for a woman whose third pregnancy ends in a non-live outcome. However, if a woman's third pregnancy ends in a live birth, then the lengthening effect of having a live previous second birth on her waiting time to third pregnancy is reduced by 2 to the power of (1·487 − 0·754), or 1·662 times that of a non-live second birth.

Interpreting the effect of the coming outcome type on the waiting time is even more complicated because of its interaction with each of the explanatory variables incorporated in the model. To simplify the interpretation of this effect, the model is reduced further by including only those covariates and interaction terms that are statistically significant. The result of the final model is shown in the right-hand panel of Table 2.

The difference in deviance between the final model and the full model is 24·32 with 20 degrees of freedom suggesting that the simplified model is a reasonable one. No substantial change is observed in the coefficient associated with a live third birth outcome. Women who have a live third birth outcome have a shorter waiting time to their third pregnancy. Its coefficient of − 2·712 means that there is a change by a factor of 2 to the power − 2·712, or 0·15 in the waiting time to third pregnancy. This effect on shortening the waiting time is further increased by maternal age 30 years or older. However, the effect is reduced by use of another modern method within 3 months of the previous birth, Indian ethnic group, period of previous birth 1980–88, and length of previous interval longer than 24 months.

https://doi.org/10.1017/S0021932004006984 Published online by Cambridge University Press

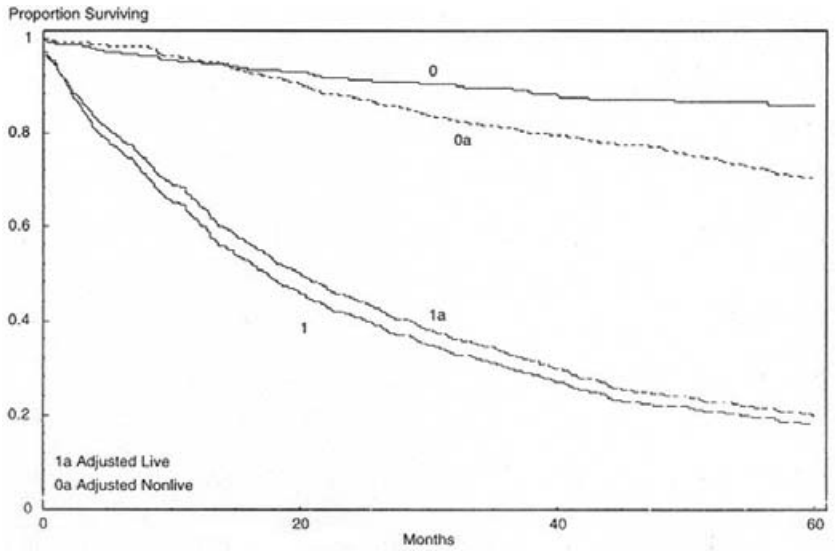


Fig. 1. Adjusted vs Kaplan–Meier survival curves live vs non-live third pregnancy outcome.

The effects of the other explanatory variables on the waiting time can be interpreted in a similar fashion. The results also shows that Indian ethnic group and period of previous birth 1980–88 significantly shorten the waiting time for women with a third non-live outcome. However, for women with a live-birth outcome of the third pregnancy, the two predictors are associated with no change or possibly a lengthening of the waiting time to their third pregnancy.

The adequacy of fit of the linear model is assessed by examining the diagnostics of the residuals, which suggests no substantial departure from the normality assumption. Figure 1 shows the predicted distribution from the linear model for the two types of third pregnancy outcome adjusting for all other covariates, with the Kaplan–Meier curves superimposed. The predicted model curves are adjusted Kaplan–Meier curves obtained by first adding the coefficients of the linear model to the empirical distribution of the residuals for each outcome type (live outcome=1, non-live outcome=0). The data are then re-expressed to counteract the effect of the transformation by taking anti-logarithms to base 2 and then subtracting 2. The linear model provides a reasonable fit judging by the closeness of the model curve to the unadjusted Kaplan–Meier curve.

Discussion

Most studies on pregnancy intervals concentrate on regression analysis based on the proportional hazards model and there is little research on the study of pregnancy intervals involving more than one outcome type. This analysis is the first attempt to demonstrate the simple linear model using a data duplication technique on the waiting time to third pregnancy ending in live and non-live births. The difference between

https://doi.org/10.1017/S0021932004006984 Published online by Cambridge University Press

these methods lies in the interpretation of the effects of the predictor variables included in the model. While the coefficients of the linear model assess the effects of the covariates on the waiting time to pregnancy, the coefficients of the proportional hazards model assess these effects on the probability of becoming pregnant.

The significance of contraceptive use, maternal age, length of previous birth interval, education, outcome of third birth and the outcome of previous birth on the waiting time to the third pregnancy confirms the results obtained by using the Cox proportional hazards model (Awang, 2003). In fact, the linear model is able to identify two additional significant factors, namely, the period of previous birth and ethnicity. While these two factors and the outcome of third birth significantly shorten the waiting time to the third pregnancy, the coefficients of the other covariates are positive indicating that their effect is to lengthen the waiting time.

This study shows that Indian women are more likely to have shorter waiting times compared with Malay women. This could be due to the fact that Indian women are less likely to breast-feed, and if they do they are less likely to breast-feed for long durations (Awang & Salleh, 2000). The significance of the period effect on shortening the waiting time to third pregnancy comes from having a second birth in the period 1980 to 1988. One possible explanation of the negative effect is the 'seventy million population policy' announced by the Malaysian Government in late 1982. As a result, medical expenses related to maternity and family were covered up to five children for all government employees and there was also an increase in the total fertility rate for Malaysia around the mid-1980s. The third birth outcome and the previous birth outcome have opposite effects on the waiting time to third pregnancy. In contrast to the third birth outcome, having a live previous birth significantly lengthens the waiting time to next pregnancy. This could be due to the fact that women who have a live second birth tend to take longer to conceive their third pregnancy because they are breast-feeding. It is also likely that these women practise some form of contraception to space out their pregnancies. On the other hand, women who experience pregnancy loss may desire another birth to replace the lost child and thus would not wish to delay conceiving their next pregnancy.

The censored linear model produces essentially the same significant factors of the waiting time to third pregnancy ending in either a live birth or a non-live birth. The linear model is a simple method of analysis since it is just an extension of the multiple regression where the adequacy of the model can be assessed by examining the diagnostics of the residuals. Thus the linear model for censored observations provides a useful alternative in cases where many of the significant factors do not obey the proportionality assumption of the proportional hazards model. Although the problem of non-proportionality of the hazards may be solved by fitting a time-stratified model, the procedure becomes cumbersome if the number of variables needed to be stratified is large because it may involve using a different set of time zones for different variables, depending on the behaviour of the hazards.

References

Aitkin, M. (1981) A note on the regression analysis of censored data. *Technometrics* **23**, 161–163.

https://doi.org/10.1017/S0021932004006984 Published online by Cambridge University Press

Awang, H. (1997) Survival analysis modelling of pregnancy intervals in Peninsular Malaysia. PhD Thesis, Macquarie University, Sydney, Australia.

Awang, H. (2003) Determinants of waiting time to third pregnancy using censored linear regression. *Journal of Biosocial Science* **35**, 59–70.

Awang, H. & Salleh, A. L. (2000) Determinants of breastfeeding duration in Peninsular Malaysia. *Asia-Pacific Journal of Public Health* **12**(2), 102–106.

Curtis, S., Diamond, I. & McDonald, J. W. (1993) Birth interval and family effects on postneonatal mortality in Brazil. *Demography* **30**(1), 33–43.

Davanzo, J. & Starbird, E. H. (1991) Correlates of short birth intervals in Peninsular Malaysia: their pathways of influence through breastfeeding and contraceptive use. *Studies in Family Planning* **22**(4), 241–254.

Janowitz, B. & Smith, J. (1984) Pregnancy intervals, breast-feeding and contraception. In Janowitz, B., Lewis, J. A., Burton, N. & Lamprey, P. (eds) *Reproductive Health in Africa: Issues and Options*. Research Triangle Park, North Carolina, Family Health International, pp. 40–47.

Kuate Defo, B. (1997) Effects of infant feeding practices and birth spacing on infant and child survival: a reassessment from retrospective and prospective data. *Journal of Biosocial Science* **29**, 303–326.

Kuk, A. Y. C. (1992) A semiparametric mixture model for the analysis of competing risks data. *Australian Journal of Statistics* **34**(2), 169–180.

Larson, M. G. & Dinse, G. E. (1985) A mixture model for the regression analysis of competing risk data. *Applied Statistics* **34**, 201–211.

Lunn, M. & McNeil, D. (1995) Applying Cox regression to competing risks. *Biometrics* **51**, 524–532.

Majinge, C. R. & Lema, V. M. (1993) Pregnancy intervals: their determinants and foetal outcome at the KCMC, Moshi, Tanzania. *East African Medical Journal* **70**(9), 544–550.

Miller, J. E., Trussell, J., Pebley, A. R. & Vaughn, B. (1992) Birth spacing and child mortality in Bangladesh and the Philippines. *Demography* **29**(2), 305–318.

Narendranathan, W. & Stewart, M. K. (1993) Modelling the probability of leaving unemployment: competing risks models with flexible base-line hazards. *Applied Statistics* **42**(1), 63–83.

Pandey, A. & Suchindran, C. M. (1988) *Modelling Fertility with Mortality as a Competing Risk: An Application*. Carolina Population Center Papers, No. 88–19, University of North Carolina.

Trussell, J., Martin, L.G., Feldman, R., Palmore, J.A., Conception, M. & Abu Bakar, N. L. (1985) Determinants of birth-interval length in the Philippines, Malaysia, and Indonesia: a hazard model analysis. *Demography* **22**(2), 145–168.

https://doi.org/10.1017/S0021932004006984 Published online by Cambridge University Press