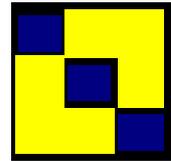


Modelling survival data in MLwiN 1.20



by

*Min Yang
Harvey Goldstein*

**Centre for Multilevel Modelling
Bedford Group for Lifecourse and Statistical Studies
Institute of Education, University of London**

Version 1.0: February 2003

Modelling survival data in MLwiN 1.20

Min Yang and Harvey Goldstein

© 2003 M. Yang, H. Goldstein

All rights reserved

No part of this document may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, for any purpose other than the owner's personal use without the prior written permission of one of the copyright holders.

ISBN: 0-9544036-1-4

Printed in the United Kingdom

Modelling survival data in MLwiN 1.20

by

Min Yang
Harvey Goldstein

**Centre for Multilevel Modelling
Bedford Group for Lifecourse and Statistical Studies
Institute of Education
University of London**

Version 1.0: February 2003

Web site: <http://multilevel.ioe.ac.uk>
email: m.yang@ioe.ac.uk

Acknowledgements

Juan Merlo of Lund University provided unidentified data from LOMAS (Longitudinal Database for Multilevel Analysis of Social Data in Skåne) with the consent from Statistics Sweden. A. Leyland, J. Merlo, I. Plewis, F. Steele and W. Browne commented on the draft. This work was funded by the Economic and Social Research Council (UK) through research grant R00023 8217.

Contents

1. Survival data.....	1
2. Questions for survival data analysis	3
3. Data exploration.....	4
3.1 Distribution of survival time.....	4
3.2 Kaplan-Meier estimate of survival and hazard functions	5
3.3 Comparison of survival times between groups	8
4. Accelerated lifetime (log-duration) models	10
4.1 Fitting a single level log-duration model	11
4.2 Fitting a three-level log-duration model.....	15
4.3 Calculating the survival function	17
4.4 Survival time for higher level units.....	19
5. Proportional Hazard models	20
5.1 Fitting a single-level model	22
5.2 Fitting a three-level model	22
5.3 Calculating the survival function	25
5.4 Residual analysis.....	27
5.5 Checking the assumption of proportional hazards	27
6. Discrete-time hazard models.....	28
References	29

Modelling survival data in MLwiN 1.20

1. Survival data

The term *survival data* refers to the length of time, t , that corresponds to the time period from a well-defined *start time* t_0 until the occurrence of some *particular event* or *end-point* t_c , i.e. $t = t_c - t_0$. It is a common outcome measure in medical studies for relating treatment effects to the survival time of the patients. In these cases, the typical start time is when the patient first received the treatment, and the end point is when the patient died or was lost to follow-up.

Generalizing the definition of survival time, we see other examples of similar data in social sciences, often called event history data. One example is in trials evaluating contraceptive methods: the particular event is the time (in days or months) from receiving contraception to, say, discontinuation of contraception. Another example would be in a life style study where the length of cohabitation time of partners until they get married or the transition time from education to employment during a certain age period in life are of interest.

In practice, survival data are often collected from a large clinical trial where many clinical centres are involved, or say, there is a contraception evaluation study implemented in different areas. Then the survival data have a two-level structure with patients or individuals nested within centres or areas. Individuals are level 1 units and centres are level 2 units. Other two-level data might come from repeated events within individuals, for example, birth intervals and employment episodes, or from population survey such as age-at-death or mortality by geographical areas.

In general, survival data have two distinctive features: non-symmetrical distributions and frequently censored observations. The frequency plot for most survival data shows a longer ‘tail’ to the right (known as *positive skew*) that would not meet the assumption of Normality. In the follow-up process, not every individual ends up having the event of interest observed. Some have left the study before the failure occurred, and some of them ended the observation because of other problems (i.e. competing risks), or were simply lost in the follow-up, or the study closed. Thus, their true failure time should be longer than the observed. These survival data are termed *right censored* survival times and we make the assumption that the censoring event is independent of the true survival time. There are also cases of left-censored and interval censored data, that will not be covered in this introduction.

Analysing survival data in MLwiN 1.20 is by means of menus and a set of macros labelled ‘*SURVIVAL-V2*’. We shall begin by exploring one data set.

Example one – Lifetime in relation to overall mortality

The example data, saved as a MLwiN worksheet file with the name ‘*LIFETIME.ws*’, shown below are the lifetimes in years of Malmö residents at the time of the 2000 Swedish Census. They are closed cohorts of people 65 ~ 69 years old at the 1970 Swedish Census and followed up over 30 years. Along with the mortality data, other variables such as individual gender, annual disposable family income and total

number of household members in 1970 were made available and matched with the individual identification. The age of an individual who was still alive at the 2000 Census or who was lost after the 1970 Census is treated as a censored time. The individuals are nested within 11,038 households, and households are nested within 21 parishes in the city.

The mortality and lifetime data are of interest to health authorities and epidemiologists in studying health and well being in populations and health inequalities between different socio-economic groups or geographical or administrative boundaries. One of the most important measures is life expectancy at birth or at some specific age groups (OECD, 2002). Using the example, we illustrate briefly how single level survival analysis and multilevel survival models can be applied to study life expectancy (LE) and the effects of individual background and social economic factors on LE using MLwiN.

Open the file 'LIFETIME.ws' to bring up the Names window as below.

Name	n	missing	min	max
1 parish	12587	0	128001	128021
2 household	12587	0	12492	481079
3 individual	12587	0	1	12587
4 age1970	12587	0	65	69
5 male	12587	0	0	1
6 death	12587	0	0	1
7 AGE2000	12587	0	66	98
8 familysize	12587	0	1	24
9 familyincome	12587	0	0	32605
10 log(t)	12587	0	4.189655	4.584968
11 cons	12587	0	1	1
12 censored	12587	0	0	1

Variable label

- parish: level 3 identification
- household: level 2 identification
- individual: level 1 identification
- age1970: age of individual in 1970
- male: gender coded as 1=male, 0=female
- death: life status 1=died, 0=otherwise
- age2000: age last seen
- familysize: total members in household
- familyincome: disposable family income, 100SEK
- log(t): nature logarithm of 'age2000'
- cons: constant vector = 1
- censored: censoring flag censored=1

In the data 86.2% of the households are single member families, and 5.4% of the records are censored. The censoring flag is in C12.

Use the **View or edit data** option in the **Data manipulation** menu to view the raw data in the **Data** window. Both 'familysize' and 'familyincome' are household level variables. The survival time is the age at death of individuals in 2000 in C7 with the status in C6. Note that C6 and C12 are complementary, i.e. the censored individuals are those who did not die.

	parish(household(individual(age1970(male(death(age2000(familysize(familyincome(
1	128001	12492	1	68	0	1	92	2	5257
2	128001	15643	2	65	0	1	88	1	1011
3	128001	263532	3	66	1	1	79	2	1377
4	128001	263532	4	67	0	1	85	2	1377
5	128001	263545	5	65	1	1	84	4	2651
6	128001	263545	6	65	0	1	85	4	2651
7	128001	263572	7	66	0	1	83	1	1306
8	128001	263583	8	67	1	1	80	5	1187
9	128001	263586	9	69	0	1	72	2	1295
10	128001	263588	10	66	0	1	95	2	4478
11	128001	263590	11	66	0	1	84	2	1065
12	128001	263595	12	68	0	1	88	2	2500
13	128001	263598	13	67	0	1	73	1	3432
14	128001	263600	14	68	0	1	77	1	2944
15	128001	263610	15	66	0	1	79	2	5278

2. Questions for survival data analysis

In substantive fields where a ‘treatment’ (e.g. a drug or surgery) may be introduced and evaluated in comparison to a control group, the main research questions can be summarised as follows.

- How long on average are the subjects going to survive after the treatment?
- Does a particular treatment result in a longer survival of subjects than other treatments?
- What are the risk factors that may affect the survival time?

In our example, no particular medical treatment is involved. The general term ‘survival time’ means the lifespan in years in this case. Life expectancy can be calculated readily based on the lifespan estimate. We shall be estimating and comparing lifespan of males and females in order to assess their health well being. We are also interested in investigating risk factors that may be associated with inequality in health, as measured by lifespan, between geographical boundaries and between social groups. The specific questions of interest are: What is the average lifespan of residents in this city? What is the gender difference in lifespan? Does lifespan vary between parishes, or between households? How does household income affect the lifespan of individuals? How does household size affect the lifespan of individuals?

To answer these research questions the following statistics are typically used:

- *Survivor function*: The probability that the random survival time variable T is greater than or equal to a specific t . Assuming $F(t)$ is the cumulative distribution function of t , the survivor function is the right tail probability, and so is defined

$$S(t) = P(T \geq t) = 1 - F(t) \quad (1)$$

- *Hazard function*: The probability that an individual dies at or just after time t , conditional on he or she having survived to that time. It represents the instantaneous death rate for an individual surviving to time t , and is defined as

$$h(t) = \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)} = \frac{F(t + \Delta t) - F(t)}{S(t)} \quad (2)$$

The term Δt represents a very small unit increment of time.

- *Cumulative hazard function*: The cumulative sum of the hazard probability function that can be expressed as,

$$H(t) = -\log S(t) \quad (3)$$

- *Median survival time*: The time when $S(t) = 0.5$. This statistic is termed the *life expectancy* in the population.

- Comparison of mean survival time or survivor function between treatment groups, or between background strata, by means of statistical tests such as Log-rank taking into account the stratification in the data.
- Regression analysis for multiple explanatory variables associated with the median survival time or survival function or hazard function, by means of parametric survival models and semi-parametric proportional hazard models.

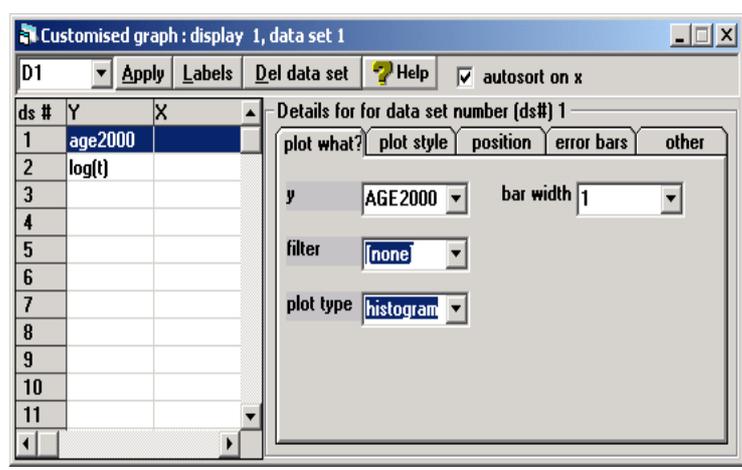
There are many well established statistical methods for carrying out these analyses. These are listed under the categories of non-parametric, parametric and semi-parametric approaches (Collett, 1999). In this chapter we only introduce how to use MLwiN for some basic data exploration (non-parametric) and for fitting single level and multilevel survival models (parametric and semi-parametric), illustrated by examples.

3. Data exploration

It is always advisable to carry out some simple descriptive or exploratory analysis of the data before fitting more complex models. We shall illustrate this assuming a single level data structure.

3.1 Distribution of survival time

To view the distribution of the variable 'age2000' we can use the **Graphs** menu to choose the **Customised graph** option to bring up the following window.



Steps for plotting:

1. Click on **ds # 1** to specify 'age2000' as y and choose **histogram** in **plot type**. Allow 1 as the **bar width**, although you could use any number. Click on the **Apply** button.
2. Click on **ds # 2** to specify 'log(t)' as y and choose **histogram** in **plot type**. Allow the appropriate bar width. Choose the **position** tab, tick in the second cell. Click on the **Apply** button.

The graph below shows the distribution of the raw survival time (on the left) and the distribution of the log survival times (on the right). The two distributions do not differ much.

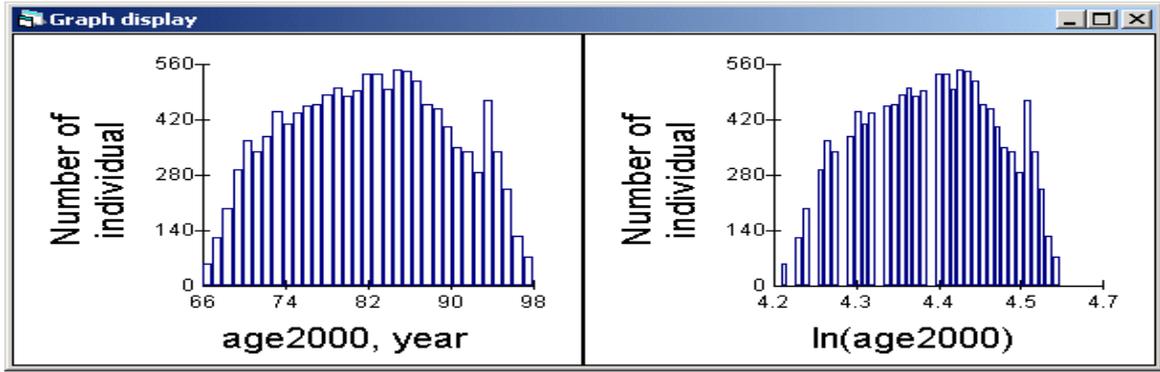


Figure 1 Histograms of the raw survival times and the log survival times

3.2 Kaplan-Meier estimate of survivor and hazard functions

Given n individuals with observed survival times, some of the observations may be censored and there may also be more than one individual who fails at the same observed time. We suppose that there are g ($g \leq n$) failure times amongst the individuals, and arrange these times in ascending order. The survival/failure times are recorded to the nearest year and this gives 34 distinct time intervals as listed in the first column of Table 1 for this example. We count the total number of individuals alive at the start of the interval (n_i , $i = 1, 2, \dots, g$) and the number of individuals who died (d_i) in the time interval. The Kaplan-Meier estimate of the survivor function is given by

$$\hat{S}(t_g) = \prod_{i=1}^g \left(\frac{n_i - d_i}{n_i} \right). \quad (4)$$

with the approximate standard error (Greenwood's formula)

$$s.e.\{\hat{S}(t_g)\} \approx [\hat{S}(t_g)] \left\{ \sum_{i=1}^g \frac{d_i}{n_i(n_i - d_i)} \right\}^{0.5} \quad (5)$$

Once $\hat{S}(t)$ is estimated, we can estimate the median survival time \hat{t}_M such that $\hat{S}(\hat{t}_M) = 0.5$. For different groups of individuals such as male and female, we can estimate a survivor function for each group and plot them for comparison.

The hazard rate is estimated as

$$\hat{h}(t) = \frac{d_i}{n_i(t_{i+1} - t_i)} \quad (6)$$

This is a non-parametric approach, and useful for showing the overall survival pattern, hazard pattern and differences between groups before further model fitting.

There is no Windows interface in the current version of MLwiN to do all the analysis. Instead we can call the macro “K-M-estimator” to carry out the estimation. It is essential to know how to use the **Command Interface** window and some basic commands of MLwiN in order to use the macros. The macro K-M-estimator also requires the fixed name “**time**” for the survival time column and the fixed name “**right**” for the censoring flag column.

This macro is one of the seventeen files in the SURVIVAL macro set that is stored in the directory c:\program files\mlwin1.20\survival by default. They can also be stored anywhere specified by users.

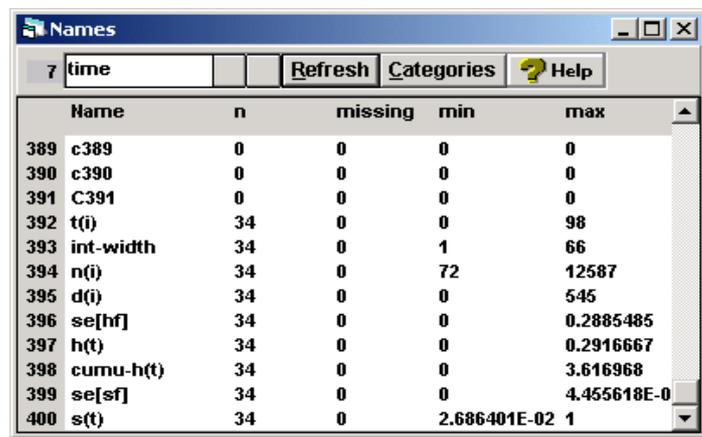
Working on the worksheet *LIFETIME.ws*, we first run the macro in the following steps:

Exercise one

► Estimate K-M survivor function for the whole data

- Step 1: In **Options** window change directory from the default to “User Defined Settings”, and specify the directory where the macros are stored;
- Step 2: In **Names** window change the column names ‘age2000’ to ‘time’, and ‘censored’ to ‘right’;
- Step 3: Open the **Command Interface** window and type the command line *OBEY K-M-estimator.txt* to execute the macro.

This macro returns results in columns C392 - C400 as below.

	<p>Description of the results</p> <p>C392: the start time of the interval; C393: the width of the interval; C394: Number of subjects at the beginning of the interval; C395: Number of failures acquired during the interval; C396: SE of the hazard rate; C397: Hazard rate; C398: Cumulative Hazard rate; C399: SE of survival probability; C400: Survival probability</p>
---	---

We can output some or all of the estimates by typing the command **Print** in the **Command Interface** window. For example the following command line gives the results in Table 1 where only part of the data are displayed.

Print c392 c394 c395 c400 c399

Table 1 K-M estimates of the survival function

	t(i)	n(i)	d(i)	S(t)	se[sf]
N =	34	34	34	34	34
1	0.00000	12587.	0.00000	1.0000	0.00000
2	66.000	12587.	55.000	0.99563	0.00058791
3	67.000	12532.	122.00	0.98594	0.0010495
4	68.000	12410.	196.00	0.97037	0.0015115
5	69.000	12214.	293.00	0.94709	0.0019953

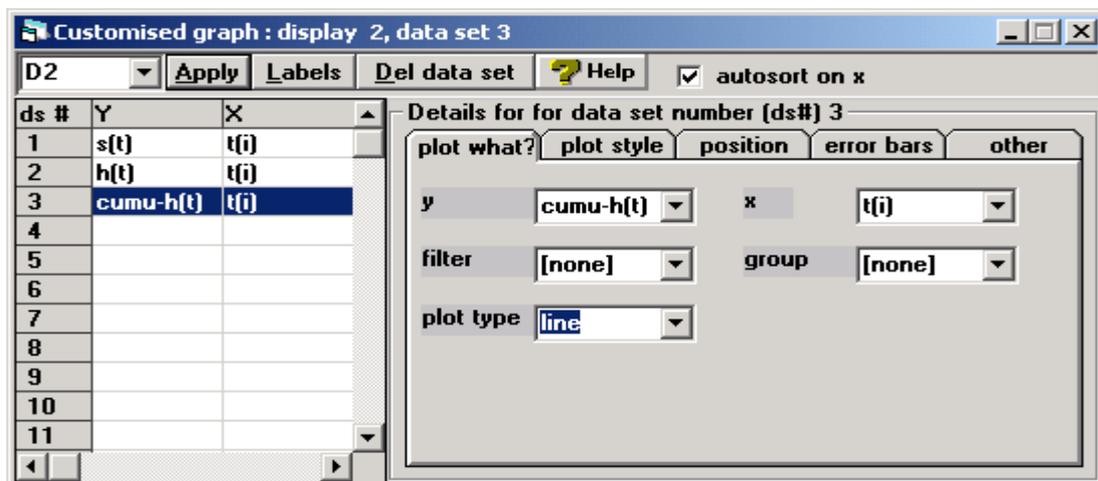
6	70.000	11921.	365.00	0.91809	0.0024443

17	81.000	7186.0	493.00	0.53174	0.0044477
18	82.000	6693.0	535.00	0.48923	0.0044556
19	83.000	6158.0	534.00	0.44681	0.0044314

30	94.000	1245.0	242.00	0.079813	0.0024159
31	95.000	777.00	159.00	0.063481	0.0022420
32	96.000	443.00	101.00	0.049008	0.0021441
33	97.000	199.00	45.000	0.037926	0.0022057
34	98.000	72.000	21.000	0.026864	0.0025627

From Table 1 we see that the survival probability is 0.5317 at $t=81$, and 0.4892 at $t=82$. This means that the median survival time (life expectancy) or $\hat{t}_{|\hat{S}(t)=0.5}$ should be between 81 and 82 years old. The common practice is to define the median time to be the smallest observed survival time for which the value of the estimated survival function is less than 0.5. In this case it is 82 years old for men and women jointly.

To plot the functions of survival, hazard and cumulative hazard, we use the **Customised Graphs** window in the **Graphs** menu. We plot $S(t)$, $h(t)$ and $H(t)$ against $t(i)$ using the *Line* style and place them in different *position* in the graph. All of these can be specified in the window as shown below.



Because of no death information before 1970 from the cohort of 65-69 year olds, the survival probability of them before 65 years was assumed one. This produces a flat survival probability of one and zero hazard risk in the following plots up to 65 years of age.

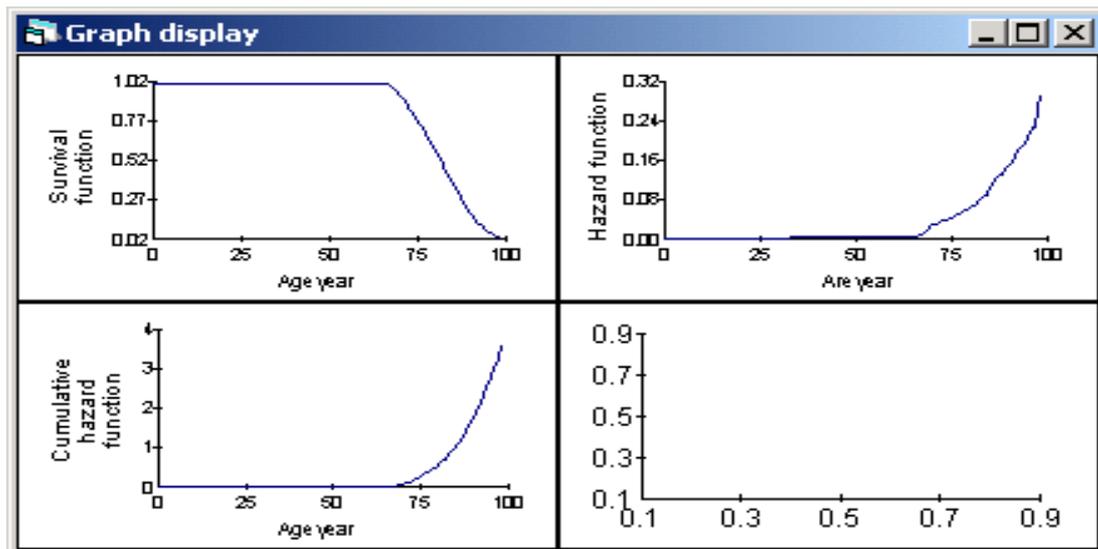


Figure 2 K-M estimates of the lifetime survival and hazard functions

These plots of the survivor and the hazard functions are not the exact K-M step functions. To get the step function you need to stack the lower and upper bound of each time interval in one column for the x-axis and repeat the survival probability of that interval twice for the y-axis and then plot the graph. This can be done using the VECT command, turn off the Autosort on x on the Graphs window. We shall leave this for the reader to explore.

3.3 Comparison of survival times between groups

To compare the median survival time for males and females, we can apply the same macro for males and females separately to obtain two survivor functions in different columns in the worksheet. A few commands will be used to separate gender, and to store their results in different columns. Then we can plot their survivor or hazard curves in the same graph.

Exercise two

► Estimate K-M survivor function by groups

Step 1: Open the worksheet "LIFETIME.ws".

Step 2: From **Options** window change from the default directory to "User Defined Settings" and specify the directory where the macros are stored;

Step 3: Open the **New Macro** window and type in the following command lines

```
Name C13 'grp' C14 'time' C15 'right'
Choo 0 'male' 'age2000' 'censored' C13 - C15
Obey K-M-estimator.txt
Eras C31-C36
Appen C31 C32 't(i)' 's(t)' C31 C32
Choo 1 'male' 'age2000' 'censored' C13-C15
Obey K-M-estimator.txt
Appen C33 C34 't(i)' 's(t)' C33 C34
Omit 0 60 C31 C32 C31 C32
Omit 0 60 C33 C34 C33 C34
```

Step 4: Click on the **Execute** button to run the commands

Step 5: Open the **Graphs** window to plot *Lines* for (x=C31, y= C32) and (x=C33, y=C34) in the same display position but different colors to obtain the following graph

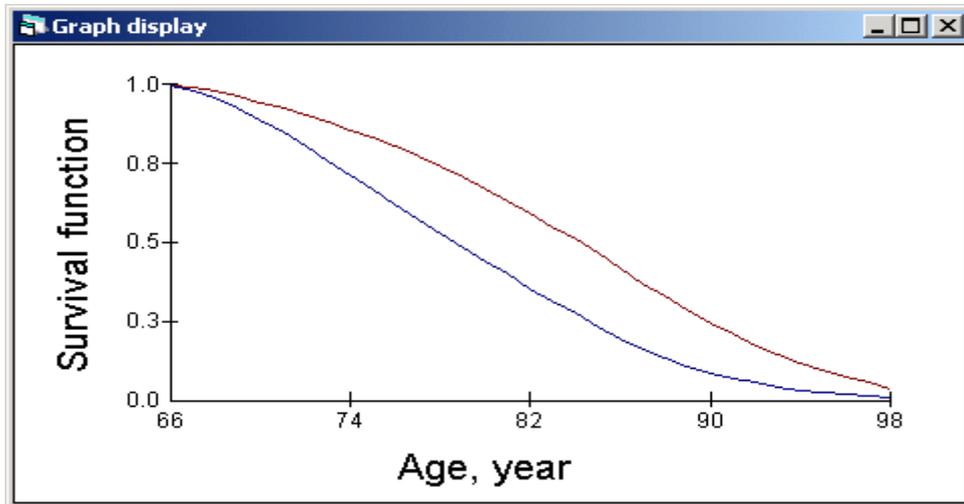


Figure 3 K-M estimates of survival function by gender (the bottom line for male and upper line for female)

The plot clearly shows a greater survival probability for females than for males. Viewing the results in C31-C34, we see the median life expectancy is estimated to be 79 years for males and 85 years for females.

Furthermore, we may want to test whether the gender difference in the K-M estimates is statistically significant. The commonly used non-parametric approaches are the Log-rank (Mantel-Haenzel) or Wilcoxon tests (*D. Collett, 1999*): the first is a global test of any differences between the distributions and the second compares their locations (medians). Both tests give approximate Chi-squared statistics. This can be done by using the macro “LOG-RANK.txt”. The macro requires the time variable named as “**Time**”, the censoring indicator as “**Right**” and the group indicator as “**GRP**”.

Exercise three

► Test for equality of several groups of survival data

Step 1: Open the worksheet “LIFETIME.ws”.

Step 2: From **Options** window change from the default directory to “User Defined Settings” and specify the directory where the macros are stored;

Step 3: Rename the columns ‘age2000’ to ‘Time’, ‘male’ to ‘Grp’ and ‘Censored’ to ‘Right’; Calculate ‘grp’=‘grp’+1;

Step 4: Open the **Command Interface** window and type the command line *OBEY Log-Rank.txt* to execute the macro.

Having done the exercise, you can see the following results displayed in the **Output** window

B1	B2	B3	B5	B6
311.62	377.69	1.0000	0.0000	0.0000

In boxes B1 and B2 are the χ^2 values of the Log-Rank and Wilcoxon tests with their tail probabilities in B5 and B6 respectively. The ‘degrees of freedom’ in this case is 1 in B3. Both tests show significant differences between the median life expectancies of males and females.

4. Accelerated lifetime (log-duration) models

The exploration of the example data has given us some idea of the overall survival probability and the differences in life expectancy between males and females. However, the question of how socio-economic status measured by family income, may affect the survival probability is still to be answered. One might think of grouping the income variable into categories of high, median and low incomes, and compare them using the Log-rank test. This would be adequate, but as we also wish to include covariates that are continuous, such an approach is impractical and we need to form an explicit (regression) prediction model. We first introduce the commonly used log-duration model or accelerated lifetime model for survival data, starting with the single level model that will then be extended to multilevel models.

We assume a general hazard function for the i^{th} individual at time t

$$h_i(t) = e^{X_i\beta} h_0(t e^{X_i\beta}) \quad (7)$$

where h_0 represents the baseline hazard function which is the hazard when the value of X_i is zero. The term $e^{X_i\beta}$ acts as the *acceleration factor* through the effects of explanatory variables on the hazard rate or the density function. Once the parameters β are estimated, the function is determined. Therefore, if t_0 is an event time sampled from the baseline distribution corresponding to values of zero for the covariates, then the accelerated life time model with the effects of covariates is $T = t_0 e^{X_i\beta}$. Under the log transformation the accelerated lifetime model has the following form,

$$y_i = \log(T_i) = X_i\beta + \log(t_0) \text{ or } y_i = X_i\beta + e_i \quad (8)$$

where the term e_i for the baseline survival time can be assumed to come from different distributions, such as the Normal, Extreme value, Logistic or Gamma distribution. The model is based on the assumption of proportional probability of the survival time and the baseline survival, $P(T > t | x) = P(t_0 > t e^{X_i\beta})$. One can get estimates of β by fitting this model using existing packages such as SAS.

The dependent variable is the logarithm of the survival time, hence log-duration model. The intercept of the model β_0 is the estimate of overall median survival time on the logarithmic scale.

In the two-level case, the log survival time for the i^{th} individual from the j^{th} cluster can be modelled as

$$y_{ij} = \log(T_{ij}) = X_{ij}\beta + u_j + \log(t_0) \text{ or } y_{ij} = X_{ij}\beta + u_j + e_{ij} \quad (9)$$

where we can assume the random effects $u_j \sim N(0, \sigma_u^2)$. Under the raw scale of the data, the exponential of the random effects $\exp(u_j)$ is also termed *frailty*; Frailty captures the difference of median survival times among clusters, parishes or households in our example. The model can be extended to allow some β coefficients to vary between clusters. When no censored observations present, this model is an ordinary two-level model.

The set of macros “SURVIVAL-V2” fits log-duration models for single level, two-level as well as more than two-level survival data (Model 9 with extension). The estimation procedure is Quasi-likelihood under IGLS (Goldstein, 2003). In the single level case, the estimates are ML under Normality when there are no censored data. In the multilevel case when there are many censored times, for example over 50% in the dataset, this estimation procedure tends to break down and is not recommended.

4.1 Fitting a single level log-duration model

To study the relationship between the lifetime of individuals and the covariates *gender* ($x_1=1$ for male, 0 for female), *family income* (x_2) and *family size* (x_3), we start with a single level model ignoring the structure of individuals nested within households nested within parishes. The model for the i^{th} log survival time can be written as

$$y_i = \beta_0 + \beta_1 x_{1i} + f(x_{2i}) + \beta_3 x_{3i} + e_i \quad (10)$$

The family income variable has a positive skewed distribution with large range of values 0 ~ 32,605 (100SEK); therefore the natural log transformation of income is applied. We also centre the transformed variable around the average income $\ln(1500)=7.31$, assuming a polynomial function f , typically a quadratic, between log-income and lifetime. Family size is centred at 2. The reference group in the model consists of women in two member families with average family income. The median lifespan for the reference group is estimated by $\exp(\beta_0)$. For males the same in socio-economic situation, the median lifespan is estimated as $\exp(\beta_0 + \beta_1)$.

Model (10) is specified following steps in Exercise four below.

Some preparation is required to be able to use the macros.

The default directory for these macros is c:\program files\mlwin\survival, however one can put the macros in any other directory. From the **Options** window one specifies the directory and the two important files PRE.SU and POST.SU in the pre and post file boxes accordingly.

Once in the correct directory, after typing the command *OBEY OPTIONS.SU* in the **Command Interface** window, we should see the following screen in the **Output** window.

```

          LOG-DURATION SURVIVAL MODEL OPTIONS (RELEASE 2.0)
          =====
ERROR DISTRIBUTION:  B10=*  - NORMAL(1), EXTREME VALUE(2), GAMMA(3), LOGISTIC(4)
MIXED RESPONSE :    B12=*  - YES(1), LINK SURVIVAL RELATED VARS. IN (G10)

*=UNSPECIFIED

```

This screen reminds us what number to set in which box for what distribution of e_i . For example, for a Normal distribution, we set B10 as 1, and for the Extreme value distribution we set B10 as 2. For mixed response models with a survival time and other Normal response, we set B12 as 1 and the explanatory variables associated with the survival time response should be linked to G10, in addition to the B10 setting. The mixed response model will not be covered in this chapter.

The column containing the event information (1=failed and 0=censored) should be named as “UNCENS”. This is required by the macro.

Exercise four

► Preparing to fit a single-level Log-duration model

Step 1: Open the worksheet “LIFETIME.ws”.

Step 2: From **Options** window change from the default directory to “User Defined Settings” and give the directory name where the macros are stored. Type PRE.SU and POST.SU as pre- and post- files.

Step 3: In **Names** window change the column names ‘death’ to ‘uncens’, ‘censored’ to ‘right’;

Step 4: Open the **New Macro** window and type in the following command lines to setup the model;

```

Iden 4 'parish' 3 'household' 2 'individual' 1 'cons'
Put 12587 0 C13
Calc C14='familysize' - 2
Calc C15=log('familyincome'+1) - 7.31
Calc C16=C15^2
Name C13 'zero' C14 'fmsize-2' C15 'L_income' C16 'L_income2'
Expl 1 'cons' 'male' C14 C15 C16 'zero'
Fpar 0 'zero'
Resp 'log(t)'
Setv 4 'zero'
Setv 3 'zero'
Set b10 2
Set b12 0

```

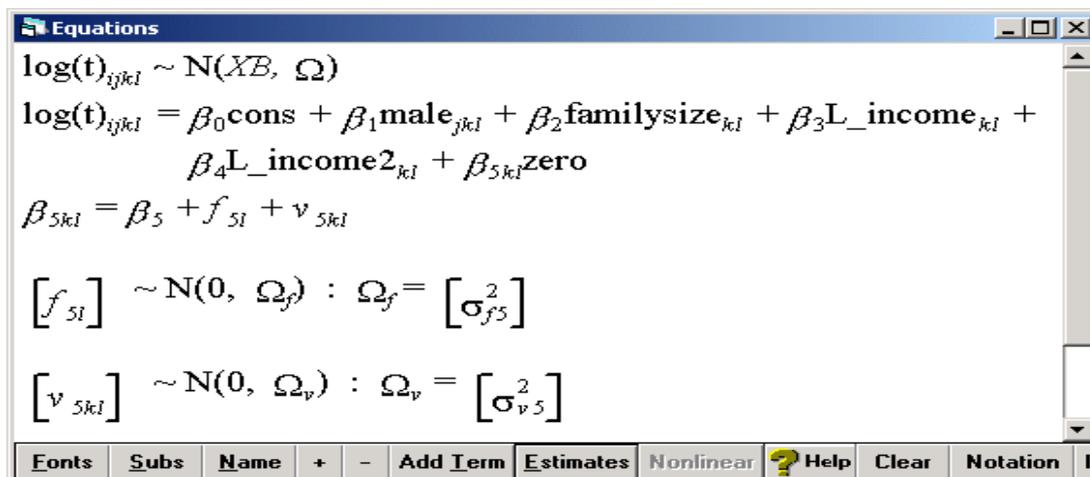
Step 5: Click on the **Execute** button to run the macro

In Exercise four, we have set up a three-level structure with individuals nested within household nested within parish. The true level 3 is shifted up to 4, and true level 1 is

moved up to 2. For a single level model, variances of the true levels 3 and 2 should be zero. So, in fitting such a model we put a zero column as the design vector and set it as the random term at parish and household levels. In this way we can fit a single level model by forcing a zero variance between parishes at level 4, and a zero variance between households at level 3. By setting B10=2, we assume an Extreme value distribution for errors. Her the two continuous variables ‘familysize’ and log ‘familyincome’ are centred around their own medians, and a quadratic polynomial function is fitted for the relationship between the log income and the age of death.

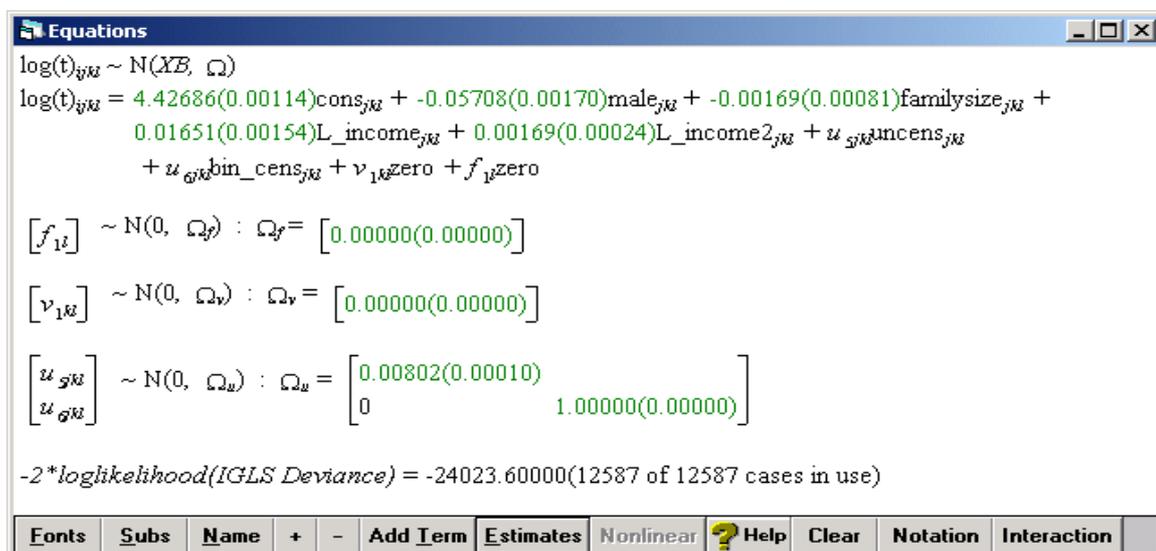
For users who have experience in specifying models through the **Equations** window, **Exercise four** can be done entirely in the **Equations Window**.

Having done Exercise 4, we open the **Equations** window to display the model specification as below.



Nothing has been set at levels 2 and 1, and the macros will set random terms at level 2 in the course of model fitting.

To fit this model, click on the **Start** button in the tool bar, the model should converge after a few iterations with the following estimates displayed in the **Equations** window.



Based on the fixed parameter estimates in the model, we obtain the estimated median lifespan as $\exp(4.4269)=83.67$ for females and $\exp(4.4269-0.05708)=79.03$ for males with a significant gender difference in log lifespan ($z = -0.05708/0.0017 = -33.58$). The quadratic form of the log scale of household income has a significantly positive effect too, and family size appears to have a marginal negative effect on lifespan.

Note in the window the $-2\log$ -likelihood value may not be reliable and the equation $\log(t)_{ijkl} \sim N(X\beta, \Omega)$ should be ignored.

We may want to display the predicted relationship between household income and lifespan graphically. This procedure is explained by **Exercise five** based on the model above.

Exercise five

► Predict and display the relationship between household income and lifespan

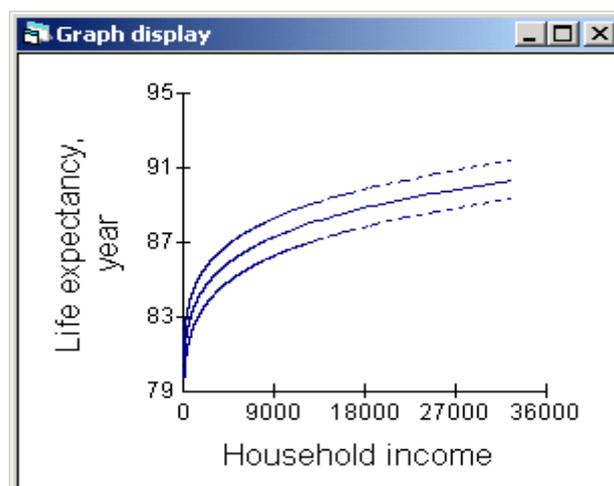
Step 1: Open **Prediction** window to highlight the terms 'Cons' 'L_income' and 'L_income2', set the predicted value in C17, and 1.96 SE of the quadratic function (Fixed effects) in C18. Then click on the **Calc** button.

Step 2: Open **Command Interface** window to type in and execute two command lines.

Calc C17=expo(C17)

Calc C18=expo(C18)

Step 3: Open **Graphs** window and use the **customised graph** to explore the line plot with the 95% confidence interval (using the *error bars* option) to get the following graph showing the trend of the higher household income, the longer the lifespan.



At the individual level in the model, the residual variance unexplained by the model is 0.008, and the term 'bin_cens' is a scalar with variance constrained to be 1 for the censored times. Nothing is set at the level below individual. This is a setting similar to the Normal multivariate model in MLwiN. To change the residual distribution from Extreme value to Normal, we type the command *SET B10 1* in the **Command Interface** window, then click on the **More** button to continue iterations till the model

converges. Similarly we can set B10=3 for a Gamma distribution or B10=4 for a Logistic distribution. The estimated median lifespan by gender adjusted for household income and family size of Model (10) under different error distributions are presented in Table 2, showing small difference between the four distributional forms.

Table 2 Median lifespan of 65-69 years old estimated by MLwiN macros under different residual distributions

	Raw K-M	B10=2 Extreme value	B10=1 Normal	B10=3 Gamma	B10=4 Logistic
Female	85	83.66	82.99	84.03	83.96
Male	79	79.03	78.97	79.16	79.09
Individual variance on raw scale		1.008	1.008	1.008	1.008

The differences between the K-M estimate and the log-duration model estimates on the lifespan of females is not surprising because the K-M estimate does not adjust for household income and family size while the single level log-duration model does.

To estimate the lifespan for each of the age groups in the population of 2000, or the age specific lifespan, we need to include dummy variables for the age groups in the fixed part. Based on the model we predict the overall survival probability $S(t)$. The life-table method would be used to derive the age specific life expectancy. We illustrate the procedure in Section 4.3 below.

4.2 Fitting a three-level log-duration model

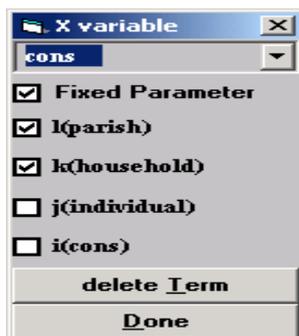
We can extend the single level Model (10) to a three-level model (11) by including random effects f_l for parishes and random effects v_{kl} for households,

$$y_{jkl} = (\beta_0 + f_l + v_{kl}) + \beta_1 x_{1jkl} + \beta_2 x_{2kl} + \beta_3 x_{2kl}^2 + \beta_4 x_{3kl} + e_{jkl} \quad (11)$$

$$f_l \sim N(0, \sigma_f^2), v_{kl} \sim N(0, \sigma_v^2)$$

The random effects f_l allow for mean differences of lifespan between parishes and the grand mean value with zero mean and variance σ_f^2 . The random effects v_{kl} estimate mean differences of lifespan between households and the grand mean value with zero mean and variance σ_v^2 .

To set up Model (11), we go back to the **Equations** window where Model (10) has been fitted. Click on the 'zero' term to delete it first, then click on the term 'cons' to bring up the box below, and to tick the parish and household boxes as shown.



Now click on the **More** button to run the model until convergence to show the following results.

```

Equations
log(t)ijk ~ N(XB, Ω)
log(t)ijk = β0iconsijk + -0.0580942(0.0016534)maleijk + -0.0011787(0.0009582)familysizeijk +
0.0163983(0.0016856)L_incomeijk + 0.0016802(0.0002567)L_income2ijk + u0iuncensijk
+ u1ibin_censijk
β0i = 4.4293510(0.0014227) + f0i + v0i

[f0i] ~ N(0, Ωf) : Ωf = [0.00000071(0.00000070)]
[v0i] ~ N(0, Ωv) : Ωv = [0.0035861(0.0001975)]

[u0i
u1i] ~ N(0, Ωu) : Ωu = [0.0051971(0.0001836)
0 1.0000010(0.0000000)]

```

We may test the joint significance of the parish and household level variances by a Wald test in the **Interval and Test** window as below.

	# 1	# 2
parish : cons/cons	0.00000(0.00000)	1.00000(0.00000)
household : cons/cons	1.00000(0.00000)	0.00000(0.00000)
individual : uncens/uncens	0.00000(0.00000)	0.00000(0.00000)
individual : bin_cens/bin_cens	0.00000(0.00000)	0.00000(0.00000)
constant(k)	0.00000(0.00000)	0.00000(0.00000)
function result(f)	0.00358(0.00000)	0.00000(0.00000)
f-k	0.00358(0.00000)	0.00000(0.00000)
chi sq. (f-k)=0. (1 df)	329.760(1.04467)	0.00001(0.00000)
+/- 95% sep.	0.00038; 0.00001	0.00001; 0.00000
+/- 95% joint	0.00048; 0.00001	0.00001; 0.00000

joint chi sq test(2df) = 331.2900000

The approximate χ^2 value is 331 with two degrees of freedom. This is highly significant, indicating an improvement of the three-level model over the single level model, although the parish level variance seems small enough to be ignored. Estimated as -0.00169 in the single-level model, 30% of the family size main effect has been explained by the large household level variance. Although all fixed effects estimated by the three-level model suggest conclusions similar to that obtained from the single-level model, the three-level model estimates have several noteworthy differences:

- The total variance of log survival time, estimated as 0.00802 in the single level model, has now been separated into three levels with 40.8% attributed to difference between households, and 59.1% to differences between individuals

within households. The proportion of variability among parishes is only 0.08% of the total variance.

- The standard error of the ‘familySize’ effect estimate is considerably larger in the three-level model than in the single-level model as 0.00097 v.s. 0.00081. It was under estimated in the single-level model. The same pattern is found for the effect of family income. Both variables were measured at the household level.
- The fixed effects of gender and household income are moderately different between the two models.

We can also estimate the household level residuals or parish level residuals, i.e. \hat{v}_{kl} or \hat{f}_l , using the **Residuals** window in the **Model** dropdown list to check for Normality and possible outliers. We leave this for the reader to explore.

4.3 Calculating the survival function

As the different survival functions for males and females are of interest, a column named ‘ $P(L>T)$ ’ in C161 of the worksheet stores the survival probability based on the survival time distribution assumed and is updated after each iteration. However, it is calculated based on the estimates of all covariates in the model. For the survival functions of the gender group conditional on other variables, for example, family size = 2 and family income at the average, we need to use commands. For different distributions assumed for the baseline survival rate, the formula for $S(t)$ is different as listed in Table 4. Remember that for the log-duration model we always work on the logarithmic scale of the observed time for any distribution, i.e. $y = \log(t)$.

In the single level case, we simply leave out the term Ω from the calculation. For the baseline survivor function, only the intercept parameter is involved in calculating $\hat{y}_{|x}$.

Table 4 Formula for calculating S(t) after fitting Log-duration model

Normal	$\phi\left(\frac{y - \hat{y}_{ x}}{\sqrt{\sigma_{uncens}^2 + \Omega}}\right)$	Note: Ω : total variance above the true level 1 for given explanatory variable values;
Extreme Value	$1 - \exp\left(-\exp\left(\frac{1.2826}{\sqrt{\sigma_{uncens}^2 + \Omega}}(\hat{y}_{ x} - y) - 0.5772\right)\right)$	$\hat{y}_{ x}$: predicted y given the value of the covariate x; To obtain S(t) from $\phi(z)$, use the command
Gamma	Gamma (b, α), $\alpha = \hat{y}_{ x}^2 / \sqrt{\sigma_{uncens}^2 + \Omega}, b = y / \alpha$	<i>NPRO</i> ‘z’ in MLwiN; To obtain S(t) from the Gamma (b, α), the commands are
Logistic	$\frac{1}{1 + \left(\exp(y - \hat{y}_{ x}) / (0.5513 \times \sqrt{\sigma_{uncens}^2 + \Omega})\right)}$	<i>GPRO</i> ‘b’ ‘a’ C50 <i>Calc</i> C50=1-C50

In Exercise 6, we calculate the survival functions for male and female conditional on family size = 2 and log of family income = 7.31, based on the three-level model (11) with Extreme Value distribution, and plot them in Figure 4 showing as before, a larger survival probability for females than for males.

Exercise six

► Calculating and plotting survival functions

Step 1: Open **Command Interface** window to run the command lines;

```
Calc C16=((4.42935-0.058094*'male')-'log(t)')*1.2826
Calc C16=(C16/sqrt(0.0051971+0.0000071+0.0035861))-0.5772
Calc C17=1-expo(-expo(C16))
Name C17 'S(t)'
```

Step 2: Choose **Customised graphs** from the **Graphs** window;

Step 3: Select C17 for the y-axis, 'age2000' for the x-axis, 'male' for the group indicator and Line for the plot type in the window;

Step 4: Click on the button **Plot style**, choose #16 rotated colour;

Step 5: Click on the button **Apply** to bring up the following plot without titles;

Step 6: Click on the plot to bring out the window for title specification.

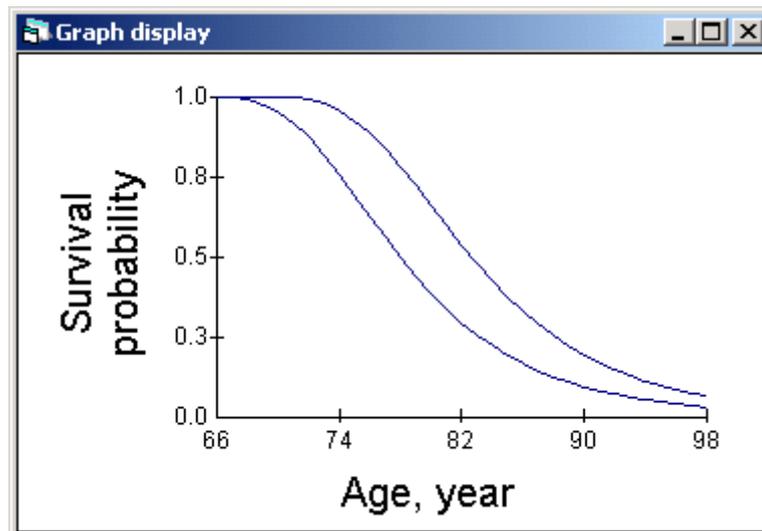


Figure 4 Survival functions estimated from three-level Log-duration model (11), upper line for females and lower line for males

The survival functions based on the other three distributions are very close in this example. We shall not present them all here.

To predict median age specific life expectancy for the age group 65-69 conditional on other variables in the model, we can use conventional life-table method based on the survival probability predicted by the model fitted. The following macro calculates this for females based on the three level model.

► Calculating life expectancy of females at age 65-69

Note calculate survival probability

Calc C16=(4.42935-'log(t)')*1.2826

Calc C16=(C16/sqrt(0.0051971+0.0000071+0.0035861))-0.5772

Calc C17=1-expo(-expo(C16))

Name C17 'S(t)'

Eras C18-C22

Sort 'age2000' 'S(t)' C18 C19

Take C18 C19 C18 C19

Note calculate mortality at time t

Calc C20=1-C19

Join C21 50000 C21

Count C18 B2

Calc B2=B2-1

Note calculate lives at time t and LE in B8

Loop B1 1 B2

Pick B1 C21 B3

Pick B1 C20 B4

Calc B5=B3*B4

Join C22 B5 C22

Calc B6=B3-B5

Join C21 B6 C21

Endloop

Sum C21 B7

Calc B8=B7/50000

In the Command Interface window, we print B8 to obtain the life expectancy estimate as 13.43, the years of life remaining for the female population aged 65-69 in year 2000. For the male population, we simply include the term $-0.058094 * 'male'$ in the first line of the macro above to get the estimate 10.19 years. In the model fitted, we did not fit the survival probability for each of the five age groups in the 1970 cohort. So the estimated age-specific life expectancy will be the same based on the same survival probability.

4.4 Survival time for higher level units

We may be interested in estimating the median survival time for each parish and comparing them as a measure of geographical equality in health. We show how it can be done as an example.

Based on model (11), the estimate of conditional median survival time for the l^{th} parish is given by $\exp(\hat{\beta}_0 + \hat{f}_{0l})$ for female and $\exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{f}_{0l})$ for males. The estimated median survival time for the k^{th} household in the l^{th} parish is given by $\exp(\hat{\beta}_0 + \hat{v}_{kl} + \hat{f}_{0l})$ for females and $\exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{v}_{kl} + \hat{f}_{0l})$ for males.

Exercise seven

► Obtaining the distribution of median survival time of level 3 units

Step 1: Choose **Residuals** tool from the **Model** window;

Step 2: Specify the level 4 id 'parish' in the **Level** box at the bottom of the screen;

Step 3: Click on the button **Calc** to get the residual estimates in C300, standardize residuals in C302 and the rank in C305;

Step 4: Open the Command Interface window to run the command lines to view the results;

```
Calc C299=expo(4.429+C300)
Print C299 C305
```

Step 5: Open **Customised Graphs** window to plot **Histogram** of C299.

The subject survival times estimated in C299 and their ranks in C305 show the median lifespan for females ranged from 83.4 to 83.9 among 21 parishes. The estimated survivor functions of parishes all overlap each other.

In the same way we can calculate the estimates for households. This shows a much wider distribution of median lifespan among 11,039 households of 73.4 - 91.9 years for females, and 69.3 – 86.7 years for males.

5. Proportional Hazard models

The general form of the proportional hazards model at time t for the i^{th} individual can be expressed as

$$h_i(t) = (e^{X_i\beta})h_0(t) \quad (12)$$

where t is treated as a continuous variable, and h_0 is the baseline hazard function, ie the hazard when the value of the explanatory variables is zero. At any time, the hazard function for the i^{th} individual dying in the period $t_{g+\Delta} - t_g$ can be specified as

$$h_i(t_g) = \frac{d_{gi}}{n_{gi} \times (t_{g+\Delta} - t_g)} \quad (13)$$

where $d_i=1$ if the individual dies in the time period and 0 otherwise. The term n_g is the total number of individuals at risk at time t_g , i.e. the number of ties, and is unity for a single subject at one time point. This is the K-M estimator of the hazard rate. Combining (12) and (13), we may express the hazard of death for the i^{th} individual at time interval t_g as

$$d_{gi} \approx n_{gi} \times (t_{g+\Delta} - t_g) \times h_0(t_g) \times e^{X_i\beta} \quad (14)$$

Taking the logarithm of (14), we get

$$\ln(d_{gi}) \approx \ln(n_{gi}) + \ln[(t_{g+\Delta} - t_g)h_0(t_g)] + X_i\beta \quad (15)$$

We may rewrite the expression as

$$y_{gi} = \ln(d_{gi}) \approx (offset) + \varphi(t_g) + X_i\beta \quad (16)$$

where the term $\varphi(t_g)$ is a function of time in relation to the baseline hazards. It can have different forms that we shall introduce later. When there are no tied survival

times, i.e. $n=l$ at any time point, the offset is zero. Model (16) is basically a Poisson model with log link.

If there are five survival times observed as (2, 5, 10*, 11, 30*) with * indicating censored time, we may rearrange the data as in Table 5.

Table 5 Restruction of survival data

Start Time, t_g	Individual i	Status	Number of failure, n_{gi}	Outcome d_{gi}	Time interval $t_{g+\Delta} - t_g$
2	1	Died	1	1	3
2	2	Alive	1	0	3
2	3	Alive	1	0	3
2	4	Alive	1	0	3
2	5	Alive	1	0	3
5	2	Died	1	1	5
5	3	Alive	1	0	5
5	4	Alive	1	0	5
5	5	Alive	1	0	5
10	3	Censored	0	0	1
10	4	Alive	0	0	1
10	5	Alive	0	0	1
11	4	Died	1	1	19
11	5	Alive	1	0	19
30	5	Censored	0	0	1

This expansion leads to several important features in fitting the hazard model (16) to survival data.

- (1) The hazard rate is assumed constant within the observed time intervals.
- (2) Fitting a Poisson model with log link to the outcome is straightforward.
- (3) If there are tied observations at any time interval, the number of failures in Table 5 would be greater than one. The logarithm of this number would be treated as an offset in the model.
- (4) Censored times do not provide information after the time censoring occurred, and their corresponding time blocks can be ignored, for example the last row in Table 5.
- (5) Time dependent covariates can be incorporated naturally in the expanded structure, and time dependent effects can be fitted by interacting covariates X with time t or $\log(t)$ in the model.

Several forms for $\varphi(t_g)$ can be considered. Here are just a few. The terms in α are parameters defining the time function.

— Polynomial function: $\alpha_1 \log(t_g) + \alpha_2 [\log(t_g)]^2 + \dots + \alpha_p [\log(t_g)]^p$

— Blocking factor or step function: $\alpha_1 Z_1 + \alpha_2 Z_2 + \dots + \alpha_g Z_g$

$$Z_g = \begin{cases} 1 & \text{for } t_g \\ 0 & \text{otherwise} \end{cases},$$

- Weibull distribution: $\log(\lambda) + \log(\gamma) + (\gamma - 1)\log(t_g)$
- Exponential distribution: $\alpha \log(t_g)$ and α to be constrained as unity.

The polynomial function is an effective form if there are large numbers of time points observed. The higher order the function is fitted, the better the approximation to the baseline hazard and other estimates in the model. For the data where the observed time could be grouped to a few intervals, the blocking factor approach is advisable.

In MLwiN v1.2, the program looks for a column named 'OFFS'. This is zero except where there are ties, as explained above.

5.1 Fitting a single-level model

Fitting Model (16) in MLwiN is the same as fitting any single level Poisson model using the **Equations** window on the expanded data. Extending the single level model to a two-level or more than two level model is straightforward. The following Exercise 8 takes you through the data expansion and fits a single level Poisson model on the LIFETIME example, and Exercise 9 extends it to a three-level Poisson model. The model estimates of the exercises are presented in Table 6.

The command **SURV** in Exercise 8 does the data expansion. Given the column for survival time, 'age2000' in the example, and the column for censored observation, 'censored' here, the **SURV** command returns five columns corresponding to each time interval: *response or death count*, *number of total failures*, *risk time indicator*, *survival time* and *number of subjects at the start point of the time*. The command also repeats other variables or level identifiers to the same length as the response column.

Exercise eight

► Expanding data and fitting a single level proportional hazard model

Step 1: Enlarge the worksheet size to 25,000 cells in the Options window, then open the worksheet "LIFETIME.ws".

Step 2: Open the **New macro** window to type in the following command lines and click on **Execute** button;

```

Eras C4 C6 C10 C11
Move
Sort 'age2000' C1-C8 'age2000' C1-C8
Surv 'age2000' 'censored' C1-C4 C6 C7 C10-C14 C1-C4 C6 C7
Eras 'age2000' 'censored'
Move
Name C7 'response' C8 'failure' C9 'rs-index' C10 'rs-time' C11 'rs-n'
Sort 2 'parish' 'household' C3-C11 'parish' 'household' C3-C11
Count C1 b1
Put b1 1 C12
Calc C13=C12
Calc C14=log('rs-time')
Aver C14 b1 b2
Calc C14=C14-b2
Calc 'familysize'='familysize'-2
Calc C15=C13-C12
Calc C16=log('failure')

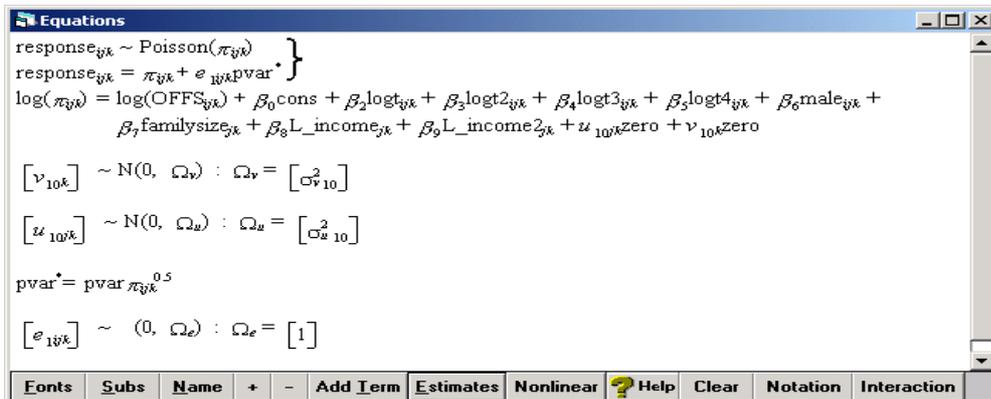
```

```

Calc C17=loge('familyincome'+1)-7.31
Calc C18=C17^2
Name C13 'cons' C14 'logt' C15 'zero' C16 'offs'
Name C17 'L_income' C18 'L_income2'
Name C12 'pvar' C19 'logt2' C20 'logt3' C21 'logt4'
Calc C19='logt'^2
Calc C20='logt'^3
Calc C21='logt'^4

```

Step 3: Open **Equations** window to specify a Poisson model shown below



Step 4: Set **Nonlinear** options to MQL1 Poisson error constrained and click on **Start** button to run the model till convergence

The baseline hazard is fitted adequately by a 4th order polynomial function. You will generally need to experiment with the order of the polynomial until adding further terms does not alter the model parameters. The estimated effects of gender, family size and household income in the column (5) of Table 6 are almost identical to the Cox model estimates that you will obtain from other packages such as SPSS. Due to the large sample size, removing the offset from the model does not make much difference to the fixed effects except that the baseline function is different.

5.2 Fitting a three-level model

Considering the structure of the data with i^{th} individual nested within j^{th} household within k^{th} parish, we now extend Model (10) to the three-level Model (17) illustrated in Exercise 9, working in the **Equations** window.

$$y_{ijk} = (\beta_0 + u_{0jk} + v_{0k}) + \sum_{h=1}^4 \alpha_h \log^h(t) + \beta_1 x_{1ijk} + \beta_2 x_{2jk} + \beta_3 x_{3jk} + \beta_4 x_{3jk}^2 \quad (17)$$

We may also allow random effects of other parameters in (17) at level 2, provided there are enough data points within individuals.

Exercise nine

► Fitting a three-level proportional hazard model

Step 1: Click to delete the term 'zero' in the equation

Step 2: Tick the term 'cons' in the parish and household level boxes

Step 3: Choose MQL1 procedure for the **nonlinear** specification

Step 4: Click **Start** button to run the model till converge

Estimates of the main effects of the three covariates in Table 6 are rather similar to the single level estimates, due to low variance between parishes and between households.

Table 6 Parameter estimates (SE) of the proportional hazards model (17)

Variable (1)	Parameter (2)	MLwiN 3- level with offset (3)	MLwiN single level without offset (4)	MLwiN single level with offset (5)	SPSS Cox model (6)
Intercept	β_0	-9.442 (0.021)	-3.308 (0.019)	-9.441 (0.019)	
$\text{Log}(t_k)$	α_1	4.535 (0.318)	5.468 (0.315)	4.532 (0.319)	
$\text{Log}(t_k)^2$	α_2	5.373 (2.297)	-10.74 (2.244)	5.373 (2.297)	
$\text{Log}(t_k)^3$	α_3	-39.36 (27.20)	303.76 (25.88)	-39.38 (27.21)	
$\text{Log}(t_k)^4$	α_4	682.5 (107.5)	-997.26 (102.2)	682.33 (107.5)	
Male	β_1	0.608 (0.020)	0.609 (0.020)	0.609 (0.020)	0.609 (0.020)
FamilySize	β_2	0.012 (0.010)	0.013 (0.010)	0.013 (0.010)	0.013 (0.010)
L_income	β_3	-0.184 (0.018)	-0.188 (0.017)	-0.188 (0.017)	-0.187 (0.017)
L_income^2	β_4	-0.021 (0.003)	-0.021 (0.003)	-0.021 (0.003)	-0.021 (0.003)
Parish var.	σ_v^2	.0007 (0.0007)	N/a	N/a	N/a
Household var.	σ_u^2	0.0000	N/a	N/a	N/a
Individual var.		Poisson constrained	Poisson constrained	Poisson constrained	

The sign of the main effects in the proportional hazard model is opposite to those in the accelerated-lifetime model because in the PH model the hazard is the outcome, whilst the outcome in the AL model is the survival time.

However, fitting the AL model in Section 4 we found significant random effects between households but not in the PH model. This could be due to the MQL1 estimation procedure that underestimates parameters, and PQL procedure did not converge on the data. Another possible reason could be that most households consisted of single member and the variation of those households would be pulled down at the individual level and constrained to be one. In fitting extra-Poisson variation model a sizable household variation was returned.

5.3 Calculating the survival function

According to equation (12) and based on the polynomial function fitted, the baseline hazard function at time g is approximated by

$$\hat{h}_0(t_g) = \exp(\hat{\beta}_0 + \sum_{p=1}^4 \hat{\alpha}_p \log^p(t_g)).$$

The baseline survival function at time g is

$$\hat{S}_0(t_g) = \exp[-\hat{h}_0(t_g)].$$

The estimated survivor function for the i^{th} individual given x value is

$$\hat{S}_i(t_g) = [\hat{S}_0(t_g)]^{\exp(\hat{\beta}_i)} \quad (18)$$

Cumulating $\hat{h}_0(t_g)$ to the end of time gives cumulative hazard function $\hat{H}_0(t)$, thus the overall survivor function of baseline is estimated as

$$\hat{S}_0(t) = \exp[-\hat{H}_0(t)]$$

In the next exercise, we calculate the survivor function using equation (18) for males and females based on the three level model presented in Table 6 above.

Exercise ten

► Calculating survivor function of the Cox model

Step 1: Open a **New macro** window and type in the following commands;

```
Sort 'rs-time' 'failure' 'logt' 'logt2' 'logt3' 'logt4' C22-C27
Take C22-C27 C22-C27
Calc C28=expo(-9.44+4.535*C24+5.373*C25-39.4*C26+682.5*C27)
Calc C28=C28*C23
Cumulative C28 C29
Calc C30=expo(-C29)
Calc C31=C30^expo(0.608)
Name C22 't' C30 'sf-female' C31 'sf-male'
```

Step 2: Click **Execute** button to run the macro

Step 3: View the columns C22 C30-C31 to find the median life expectancy in year for female and male.

The results in C22 C30 and C31 show that the median life expectancy is just over 84 years for females and just over 79 years for males. We can also plot the graph that looks very similar to Figure 4 of the log-duration model estimates.

In the three-level model (17) where random effects of the intercept β_0 are allowed among parishes and households, we can calculate a survival function for each parish or each household within a parish. For example, we include residual estimates of parish \hat{v}_{0k} in the $h_0(t)$ function to calculate survivor functions of parishes, and

include $\hat{v}_{0k} + \hat{u}_{0,jk}$ in the hazard function for the survivor function of households within parishes.

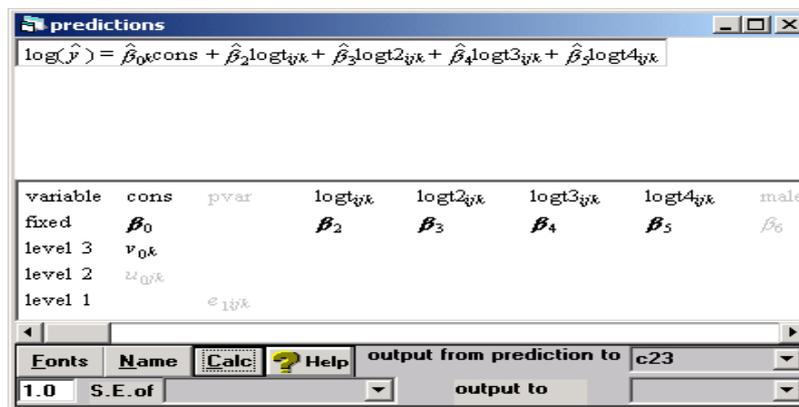
Exercise 11 leads to calculation of the parish survivor function graphed in Figure 5, based on the three-level model.

Exercise eleven

► Calculating parish survivor functions of a three-level Cox model

Step 1: Sort data on three columns in the order of 'parish', 'rs-time' and 'household', and carrying on the rest data, and put them all back to the same columns, using the **Sort** window.

Step 2: Predict $h_0(t)$ including level 3 residuals using the **Prediction** window as shown below and the results are stored in C23



Step 3: Open a **New macro** window and type in the following commands.

```
Calc C24=expo(C23+'offs')
Take 'rs-time' 'parish' C24 C25 C26 C27
Mlcv C26 C27 C28
Calc C29=expo(-C28)
```

Step 4: Plot *line graph* between C29 (*y*) and C25 (*age in year*) by C26 (*parish*) using Graphs window to get Figure 5.

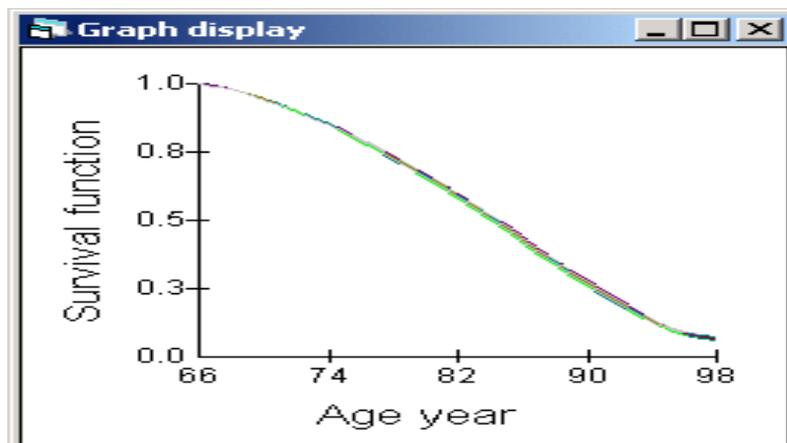


Figure 5 Estimated survivor function for parish from the three-level proportional hazard model

Showing no difference of lifespan between 21 parishes, this result is the same as what is estimated by the log-duration model.

The more general case allowing random effects of other covariates can be incorporated readily.

5.4 Residual analysis

In the single level survival analysis, it is of interest to know what distribution the survival function could be by graphically inspecting the relationship between the estimated survival probability $\hat{S}_j(t|x)$ and the cumulative hazard rate given the value of x . Typically the Cox-Snell residuals for individuals could be calculated as $\exp(X\hat{\beta})_i \hat{H}_0(t_i)$ in checking for the proportionality in the Cox model. In multilevel models aspect, the analysis of residuals at higher levels is similar to Poisson multilevel model analysis, using windows tools for residuals and graphs.

5.5 Checking the assumption of proportional hazards

Fitting Models (16) and (17), we assumed that the effects of covariates are independent of the time variable t . This means that the relative hazard for the i^{th} subject is proportional in relation to any change of the covariates, i.e. $h_i(t)/h_0(t) = e^{x_i\beta}$. In the case where the effect of a covariate x may depend on time t , the proportionality of hazards no longer holds. We can check for this in a number of ways.

1) Checking the relative hazards

We can simply introduce a time-dependent variable by creating an interaction term between the variable of interest and the log(t) term, and treat it the same as other time-independent variables. Consider a two-level model with one covariate,

$$y_{ij} = (\beta_0 + u_{0j}) + \alpha_1 \log(t) + \dots + \alpha_p \log^p(t) + \beta_1 x_{ij} + \beta_2 (x_{ij} \times \log(t)) \quad (19)$$

The parameter estimate $\hat{\beta}_2$ reflects the non-proportionality of the relative hazard in relation to x_{ij} . If x is a binary variable, we can easily plot the relative hazard functions of e^{β_0} and $e^{\beta_1 + \beta_2 \log(t)}$ against the log time for $x_{ij}=0$ and $x_{ij}=1$.

2) Checking the baseline hazards

As we have specified the form to approximate the entire baseline hazard function, we could directly interact x_{ij} with the baseline hazard function, i.e.,

$\sum_p \hat{\omega}_p (\log^p(t) \times x_{ij})$. The baseline functions e^{β_0} for $x_{ij}=0$ and $e^{\sum_p \hat{\omega}_p \log^p(t)}$ for $x_{ij}=1$ can be plotted against each other.

3) Allowing random effects of time-dependent variable in continuous scale.

The non-proportionality in relation to a time-dependent variable implies that the effect of the variable on the hazard rate is no longer a constant but varying over time. We may consider fitting random effects of this variable among time blocks in the expanded data structure. In doing this we may treat the survival time blocks as units a level above individuals, and sort the data by levels accordingly.

6. Discrete-time hazard models

In fitting proportional hazard models illustrated above, the data are restructured in time intervals corresponding to times when events occur. However, in many cases including the life expectancy study we can divide time span into predetermined intervals, for example, 20~29 years, 30~39 years, ..., 90~99 years, and restructure the data around each time interval.

Consider the same example of (2, 5, 10*, 11, 30*) used in Table 5, we may set 3 time intervals: 0~9, 10~19, 20~29 and 30_39 denoted by I_1 , I_2 , I_3 and I_4 . The data can be expanded as follows:

Time interval g	Individual i	Response d_{gi}	Indicator I_1	Indicator I_2	Indicator I_3	Indicator I_4
1	1	1	1	0	0	0
1	2	1	1	0	0	0
1	3	0	1	0	0	0
1	4	0	1	0	0	0
1	5	0	1	0	0	0
2	3	0	0	1	0	0
2	4	1	0	1	0	0
2	5	0	0	1	0	0
3	5	0	0	0	1	0
4	5	0	0	0	0	1

In each time interval the response has a code 1 if an individual died in the time period, and 0 otherwise. Thus the probability that an individual dies in the current period, given that they survived from the last period is

$$\pi_{gi} = P(d_{gi} = 1 | d_{(g-1)i} = 0).$$

A discrete-time model, assuming a piecewise constant baseline hazard can be written as a standard logistic model

$$\log\left(\frac{\pi_{gi}}{1 - \pi_{gi}}\right) = \sum_{p=1}^4 \alpha_p I_p + X_i \beta \quad (20)$$

where X_i are covariates and I_p are indicators for the time intervals shown in the table above. In some cases many more time intervals could be presented, and the first term in (20) can take the form of a continuous polynomial function (Goldstein, Pan and Bynner, 2002). Model (20) can be fitted using any package that fits standard logistic model, and readily extended to two or three level models by adding in random

effects in (20) allowing for variation between higher-level units. The **Equations** window in MLwiN can be used to specify and fit the model straightforwardly. The baseline hazard function of the time interval g is the exponential function of (20) with X_i is zero. The calculation of cumulative hazard and survivor functions is the same illustrated in Section 5.2 in the manual.

Applications of the model were presented by Goldstein et al. (2002), Goldstein (2003) and Steele et al. (1996). The extensions of the discrete event time model for competing risks and multistate competing risks can be found in Steele, et al (1996b) and Steele, et al. (2002).

References

D. Collett (1999). *Modelling Survival Data in Medical Research*, Chapman & Hall/CRC, Boca Raton, London.

Goldstein, H. (1995). *Multilevel Statistical Models*, Edward Arnold, London, New York.

Goldstein, H. Pan, H. and Bynner, J. (2002). A note on methodology for analysing longitudinal event histories using repeated partnership data from the national Child Development Study (NCDS). Working paper, Institute of Education, London University. (<http://kl.ioe.ac.uk/hgpersonal/>)

Steele, F., Diamond, I. And Amin, S. (1996b). Immunisation uptake in rural Bangladesh, a multilevel analysis. *JRSS A*, 159: 289-299.

Steele, F., Goldstein, H. and Browne, W. (2002). A general multilevel multistate competing risks model for event history data. Submitted for publication.

Goldstein, H. (2003). *Multilevel Statistical Models*, third edition. London, Edward Arnold.

OECD Health Data (2002). (<Http://www.oecd.org/m00029000/m00029907.pdf>)