

Survival Models for Grouped Data

Contents

1	MODELS FOR GROUPED- AND DISCRETE-TIME SURVIVAL DATA	2
1.1	INTRODUCTION	2
1.2	CHOOSING BETWEEN BINARY AND ORDINAL OUTCOME MODELS	3
1.2.1	<i>The data for a binary approach</i>	<i>3</i>
1.2.2	<i>The data for an ordinal approach</i>	<i>8</i>
1.3	THE MODELS	10
1.3.1	<i>Binary case: a 2-level model</i>	<i>10</i>
1.3.2	<i>Ordinal case: 2-level model</i>	<i>11</i>
1.4	EXAMPLE: A PROPORTIONAL HAZARDS MODEL- BINARY CASE	12
1.4.1	<i>Introduction</i>	<i>12</i>
1.4.2	<i>Setting up the analysis</i>	<i>12</i>
1.4.3	<i>Discussion of results</i>	<i>17</i>
1.4.4	<i>Interpreting the output</i>	<i>19</i>
1.5	EXAMPLE: SURVIVAL ANALYSIS MODEL FOR AN ORDINAL OUTCOME	21
1.5.1	<i>Introduction</i>	<i>21</i>
1.5.2	<i>Setting up the analysis</i>	<i>21</i>
1.5.3	<i>Discussion of results</i>	<i>25</i>
1.5.4	<i>Interpreting the output</i>	<i>28</i>
1.6	TWO-LEVEL SURVIVAL ANALYSIS MODELS	29
1.6.1	<i>The data</i>	<i>29</i>
1.6.2	<i>The model</i>	<i>30</i>
1.6.3	<i>Survival data as ordinal outcomes</i>	<i>31</i>
1.6.4	<i>Setting up the analysis</i>	<i>32</i>
1.6.5	<i>Discussion of results</i>	<i>36</i>

1 Models for grouped- and discrete-time survival data

1.1 Introduction

Models for grouped-time survival data are useful for analysis of failure time data when subjects are measured repeatedly at fixed intervals in terms of the occurrence of some event, or when determination of the exact time of the event is only known within grouped intervals of time. Additionally, it is often the case that subjects are observed nested within clusters (*i.e.*, schools, firms, clinics), or are repeatedly measured in terms of recurrent events. In this case, use of grouped-time models that assume independence of observations (Thompson, 1977; Allison, 1982; Prentice & Gloeckler, 1978) is problematic since observations from the same cluster or subject are usually correlated.

For data that are clustered and/or repeated, models including random effects provide a convenient way of accounting for association in correlated survival data. In terms of continuous-time survival data, several authors have developed survival analysis models including random effects that are usually assumed to be distributed as a gamma distribution. These models are often termed frailty models or survival models including heterogeneity, and review articles describe many of these models (Pickles & Crouchley, 1995; Hougaard, 1995).

Several authors have noted the relationship between ordinal regression models (using complementary log-log and logistic link functions) and survival analysis models for grouped and discrete time. Hedeker, Siddiqui, and Hu (2000) described a generalization of an ordinal random-effects regression model to handle correlated grouped-time survival data. This model accommodates multivariate normally-distributed random effects, and additionally, allows for a general form for model covariates.

Assuming a proportional or partial proportional, hazards or odds model, a maximum marginal likelihood solution is implemented using multi-dimensional quadrature to numerically integrate over the distribution of random-effects.

In this example, we explore various survival analysis models based on a study that was designed to test independent and combined effects of a school-based social-resistance curriculum and a television-based program in terms of tobacco use and cessation.

The structure of this study indicates a three-level hierarchical structure. However, for illustration purposes in this chapter we will consider a two-level structure in which students are nested within schools.

Two analysis approaches are considered for these data in the examples to follow. The first treats survival time as a set of dichotomous indicators of whether the event occurred for time periods up to the period of the event or censoring. This analysis, shown in Section 1.4, uses the data set mentioned above. The second approach treats survival time as an ordinal outcome, which is either right-censored or not. The same data, but in different format, is used for this second analysis (see Section 1.5).

1.2 Choosing between binary and ordinal outcome models

1.2.1 The data for a binary approach

An analysis of a data set where students are clustered within schools is used to illustrate features of random-effects analysis of clustered grouped-time survival data.

We focus on actual usage of tobacco products and on subsequent data collected from the respondents.

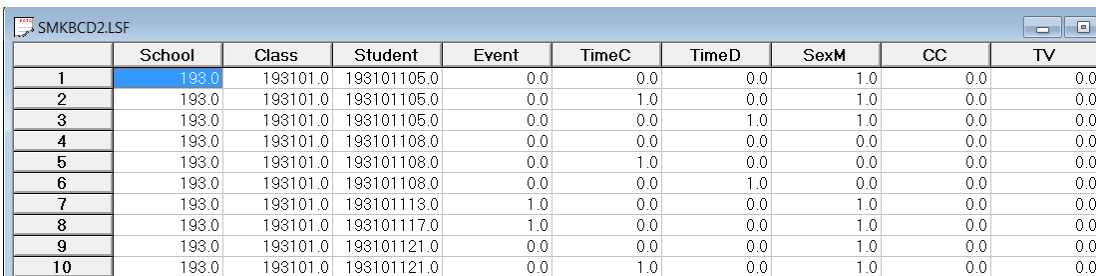
Schools were randomized to one of four study conditions: (a) a social-resistance classroom curriculum (CC); (b) a media (television) intervention (TV); (c) a combination of curriculum and TV conditions; and (d) a no-treatment control group. These conditions form a 2 x 2 factorial design of CC (yes or no) by TV (yes or no).

The outcome variable of interest in this chapter is the response the question "Have you ever tried a cigarette?". Students were assessed at 4 occasions:

- pre-intervention (January 1986, also referred to as Wave A)
- post-intervention (April 1986, *i.e.* Wave B)
- year follow-up (April 1987, *i.e.* Wave C)
- year follow-up (April 1988, *i.e.* Wave D)

As the intervention procedures were implemented following the pretest, we focus in the analyses to follow on the three post-intervention time points and include only those students who had not answered yes to this question at pretest. Of the original 1,600 respondents, 1,556 are included in the data considered here. Thus, our analysis examines the degree to which the intervention prevented or delayed students from initiating smoking experimentation. Because the intervention was also aimed at smoking cessation for individuals who had initiated smoking, here we are examining only a part of the intervention aims.

The first few lines of the LISREL spreadsheet **SMKBCD2.lsf** used in this section are shown below. Note that there is a maximum of 3 observations associated with each student – not all students have data at all 3 occasions.



	School	Class	Student	Event	TimeC	TimeD	SexM	CC	TV
1	193.0	193101.0	193101105.0	0.0	0.0	0.0	1.0	0.0	0.0
2	193.0	193101.0	193101105.0	0.0	1.0	0.0	1.0	0.0	0.0
3	193.0	193101.0	193101105.0	0.0	0.0	1.0	1.0	0.0	0.0
4	193.0	193101.0	193101108.0	0.0	0.0	0.0	0.0	0.0	0.0
5	193.0	193101.0	193101108.0	0.0	1.0	0.0	0.0	0.0	0.0
6	193.0	193101.0	193101108.0	0.0	0.0	1.0	0.0	0.0	0.0
7	193.0	193101.0	193101113.0	1.0	0.0	0.0	1.0	0.0	0.0
8	193.0	193101.0	193101117.0	1.0	0.0	0.0	1.0	0.0	0.0
9	193.0	193101.0	193101121.0	0.0	0.0	0.0	1.0	0.0	0.0
10	193.0	193101.0	193101121.0	0.0	1.0	0.0	1.0	0.0	0.0

The variables of interest are:

- School indicates the school a student is from.
- Class identifies the classroom to which a student belongs.
- Student represents the student identification number.
- Event indicates occurrence of the event (1 indicating "yes" and 0 "no.").
- TimeC is an indicator variable indicating the first follow-up occasion after the post-intervention measurement occasion. It assumes a value of 1 if a measurement was made at the first follow-up occasion, and 0 otherwise.
- TimeD is the indicator variable for the second follow-up occasion. It assumes a value of 1 if a measurement was made at the second follow-up occasion and 0 otherwise.
- SexM is an indicator variable for gender, with "1" indicating male respondents, and "0" female respondents.
- CC is a binary variable indicating whether a social-resistance classroom curriculum was introduced, with 0 indicating "no" and 1 "yes."
- TV is an indicator variable for the use of media (television) intervention, with a "1" indicating the use of media intervention, and "0" the absence thereof.

The post-intervention measurement, which is the first of the three measurement occasions in this data set, serves as the reference cell. In terms of the indicator variables TimeC and TimeD it would be a measurement for which $\text{TimeC} = \text{TimeD} = 0$.

	CCTV	SexTC	SexTD	CCTC	CCTD	TVTC	TVTD	CCTVTC	CCTVTD
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

In addition to these variables, **SMKBCD2.lsf** includes a number of interaction terms:

- CCTV was constructed by multiplying the variables TV and CC and represents the CC by TV interaction.
- SexTC denotes the SexM by TimeC interaction.
- SexTD denotes the SexM by TimeD interaction.
- CCTC denotes the interaction between classroom curriculum intervention CC and TimeC.
- CCTD denotes the interaction between CC and TimeD.

- TVTC denotes the interaction between media intervention TV and TimeC.
- TVTD denotes the interaction between TV and TimeD.
- CCTVTC represents the interaction between the CC by TV interaction at the TimeC.
- CCTVTD represents the interaction between the CC by TV interaction at the TimeD.

In all, there were 1556 students included in the analysis of smoking initiation. Of these students, approximately 40% ($n = 634$) answered yes to the smoking question at one of the three post-intervention time points, while the other 60% ($n = 922$) either answered no at the last time point or were censored prior to the last time point.

Consider a level-2 model, with schools as the level-2 units. In general, for $i = 1, \dots, N$ N level-2 units, containing $j = 1, \dots, n_i$ level-1 units (subjects or multiple failure times) the concept of a censoring or event indicator can be expressed as follows. First, we assume that the assessment time takes on discrete positive values $t = 1, 2, \dots, m$ representing time points or intervals and that each ij unit is observed until time t_{ij} . The censor/event indicator δ_{ij} is coded depending on what happens at time t_{ij} :

- an event occurs ($t_{ij} = t$ and $\delta_{ij} = 1$)
- the observation is censored ($t_{ij} = t$ and $\delta_{ij} = 0$)

The term censoring is used when a unit is observed at t_{ij} , but not at $t_{ij} + 1$ (and we know that the event has not occurred up to time t_{ij}).

As mentioned previously, the dichotomous variable EVENT indicated the occurrence of an event. Occurrence of an event was recorded at three time points (WaveB, WaveC, and WaveD), though some subjects dropped out of the study and were not measured at all three time points. To model the time until the event as the outcome variable in a binary analysis of the data, person-time indicators are created (Singer & Willett, 1993). For this, the number of records for each person depends on the timing of the event or censoring for that person. For example, if there were two follow-up points, the two person-time indicators T1 and T2 would be coded as follows:

- T1 = 1: event occurred at T1 (or in interval between T0 and T1)
- T1 = 0: event did not occur at T1 (or in interval between T0 and T1) and T1 was the subject's last measured time point
- T1 = 0 and T2 = 1: event did not occur at T1 but did occur at T2 (or in the interval between T1 and T2)
- T1 = 0 and T2 = 0: individual was censored at T2 (the subject did not experience the event at either T1 or T2)

Note that for the first two scenarios above, subjects would contribute a single record in the data set (for the T1 indicator), whereas they would contribute two records (one for each person-time indicator T1 and T2) for

the latter two scenarios. These indicators would represent the dependent variable in the analysis, akin to the variable named EVENT in our TVSFP data.

For this data, there were three follow-up occasions, and thus three person-time indicators are necessary to describe the occurrence of event/censoring. The three person-time indicators form the EVENT variable in the data set, and the timing of the event/censoring is represented by the two variables TimeC and TimeD in the data set. The coding of the person-time indicators (T1, T2, T3) that form the EVENT variable are given in Table 1.1.

Note that each person would contribute from one to three records in the data set depending on their outcome. For example, for the current data, the EVENT records and their corresponding time indicators are coded as shown in Table 1.2.

Table 1.1: Three time points with censoring

Outcome	Up to 3 records per person
Censor at T1	T1 = 0
Event at T1	T1 = 1
Censor at T2	T1 = 0; T2 = 0
Event at T2	T1 = 0; T2 = 1
Censor at T3	T1 = 0; T2 = 0; T3 = 0
Event at T3	T1 = 0; T2 = 0; T3 = 1

Table 1.2: Coding of time and event indicators for binary TVSFP analysis

EVENT records	Time indicators		Outcome description
	TimeC	TimeD	
T1 = 0	0	0	Censor at T1
T1 = 0	0	0	No event at T1
T2 = 0	1	0	Censor at T2
T1 = 0	0	0	No event at T1
T2 = 0	1	0	No event at T2
T3 = 0	0	1	Censor at T3
T1 = 1	0	0	Event at T1
T1 = 0	0	0	No event at T1
T2 = 1	1	0	Event at T2
T1 = 0	0	0	No event at T1
T2 = 0	1	0	No event at T2
T3 = 1	0	1	Event at T3

The breakdown of cigarette onset for gender and condition subgroups is presented in Table 1.3. Percentages given in the table are calculated relative to the totals for that subgroup at the time of response.

At Wave B (post-intervention time point; TimeC = 0 and TimeD = 0), 130 females (SexM = 0) and 156 males (SexM = 1) reported an event (Event = 1), while 105 females and 83 males were censored (Event = 0). These censored subjects did not experience the event at Wave B and were not measured at subsequent waves. The total numbers of females and males that provided data at Wave B were 814 and 742 respectively. The totals at Wave C (TimeC = 1) are only 579 and 503 females and males, respectively because the numbers of Wave B event and censored subjects are removed from the Wave C totals. For example, the total number of females at Wave C equals 814 (the number at Wave B) – 130 (females experiencing the event at Wave B) – 105 (censored females at Wave B) = 579. The male total of 503 is obtained in the same way. Of the 579 females, 117 experienced the event at Wave C and 154 were censored at Wave C. Similar calculations for Wave D (TimeD = 1) yield the total of 308 females (= 579 – 117 – 154), where 79 females experienced the event and 229 did not and were censored at this last time point. Regarding the differences between males and females, it can be seen that the proportion of males who experienced the event is relatively similar across the three waves. Alternatively, females were initially lower than males (16% versus 21% at Wave B) but increasingly experienced the event across the waves. At the end, the total proportion of males who experienced the event is 41.5% (156 + 89 + 63 of 742), and similarly it is 40.0% for females (130 + 117 + 79 of 814). Thus, the initial gender difference is largely gone by the end of the study.

In terms of the invention groups, the differences do not appear to be very large. If anything, there is some suggestion that control subjects have lower rates of the event, but this difference is not striking.

Table 1:3: Onset of cigarette experimentation across three time points

	TimeB			TimeC			TimeD		
	with event	censored	total	with event	censored	total	with event	censored	total
Males	156 (21.0)	83 (11.2)	742	89 (17.7)	134 (26.6)	503	63 (22.5)	217 (77.5)	280
Females	130 (16.0)	105 (12.9)	814	117 (20.2)	154 (26.6)	579	79 (25.6)	229 (74.4)	308
Control	66 (16.5)	60 (15.0)	401	53 (19.3)	69 (25.1)	275	34 (22.2)	119 (77.8)	153
CC only	75 (19.1)	27 (6.9)	392	53 (18.3)	61 (21.0)	290	49 (27.8)	127 (72.2)	176
TV only	71 (17.3)	54 (13.2)	410	60 (21.1)	79 (27.7)	285	38 (26.0)	108 (74.0)	146
CC & TV	74 (21.0)	47 (13.3)	353	40 (17.2)	79 (34.1)	232	21 (18.6)	92 (81.4)	113

In terms of clustering, these 1556 students were from 28 schools with between 13 and 151 students per school ($\bar{n} = 56$, S.D. = 38) Thus, the data are highly unbalanced with large variation in the number of clustered observations.

1.2.2 The data for an ordinal approach

The ordinal analysis illustrated in this chapter is again based on the TVSFP data. As shown in the previous section, one can also fit grouped-time survival models using dichotomous indicators of event/censoring across the study time points. To do so, requires additional data manipulation. The data set used for the ordinal approach differs from that previously discussed and is represented by the LISREL spreadsheet file **SMKCCLC.lsf**. The first 10 records of this data set are shown below.

	School	Class	Student	SmkOnset	Event	SexM	CC	TV	CCTV
1	193.0	193101.0	193101103.0	1.0	1.0	0.0	0.0	0.0	0.0
2	193.0	193101.0	193101105.0	4.0	0.0	1.0	0.0	0.0	0.0
3	193.0	193101.0	193101108.0	4.0	0.0	0.0	0.0	0.0	0.0
4	193.0	193101.0	193101111.0	1.0	1.0	1.0	0.0	0.0	0.0
5	193.0	193101.0	193101113.0	2.0	1.0	1.0	0.0	0.0	0.0
6	193.0	193101.0	193101117.0	2.0	1.0	1.0	0.0	0.0	0.0
7	193.0	193101.0	193101121.0	4.0	0.0	1.0	0.0	0.0	0.0
8	193.0	193101.0	193101132.0	4.0	0.0	0.0	0.0	0.0	0.0
9	193.0	193101.0	193101201.0	4.0	1.0	0.0	0.0	0.0	0.0
10	193.0	193101.0	193101204.0	4.0	0.0	0.0	0.0	0.0	0.0

The variables of interest are:

- School indicates the school a student is from.
- Class identifies the classroom to which a student belongs.
- Student represents the student identification number.
- SmkOnset indicates the time at which an event occurred. It assumes a value of 1 for a WaveA measurement (*i.e.*, the event occurred at Wave A), 2 for a WaveB measurement, 3 for a WaveC measurement, and 4 for a WaveD measurement.
- Event is an indicator variable indicating whether the subject experienced the event or was censored. A value of 1 indicates that the student did experience the event (*i.e.*, onset of cigarette experimentation) at one of the time points, while a value of 0 indicates that the subject was censored and never experienced the event (*i.e.*, no onset of cigarette experimentation) at any time point that they were assessed at.
- SexM is an indicator variable for gender, with "1" indicating male respondents, and "0" female respondents.
- CC is a binary variable indicating whether a social-resistance classroom curriculum was introduced, with 0 indicating "no" and 1 "yes."
- TV is an indicator variable for the use of media (television) intervention, with a "1" indicating the use of media intervention, and "0" the absence thereof.
- CC*TV was constructed by multiplying the variables TV and CC and represents the CC by TV interaction.

Survival data as ordinal outcomes

Assume 4 time points with no intermittent censoring and let y denote the ordinal outcome variable. Let us first consider subjects who initiated smoking at some point in the study. For these subjects, the variable Event will be coded as 1 and the coding of the SmkOnset variable will be as follows.

SmkOnset:

- $y_{ij} = 1$: Student first started to smoke at $t = 1$.
- $y_{ij} = 2$: Student did not smoke at $t = 1$, but first smoked at $t = 2$.
- $y_{ij} = 3$: Student did not smoke at $t = 1$ or 2, but first smoked at $t = 3$.
- $y_{ij} = 4$: Student did not smoke at $t = 1, 2, \text{ or } 3$, but first smoked at $t = 4$.

Similarly, subjects who were censored would have the variable Event coded as 0, and the following codes for the SmkOnset variable.

SmkOnset:

- $y_{ij} = 1$: Student did not smoke at $t = 1$ and no data beyond $t = 1$.
- $y_{ij} = 2$: Student did not smoke at $t = 1$ or 2, and no data beyond $t = 2$.
- $y_{ij} = 3$: Student did not smoke at $t = 1, 2, \text{ or } 3$, and no data beyond $t = 3$ (*i.e.*, no data at $t = 4$).
- $y_{ij} = 4$: Student did not smoke at $t = 1, 2, 3, \text{ or } 4$.

Here, the phrase "did not smoke" is more precisely "did not answer yes to the question have you ever smoked a cigarette." Table 1.4 shows how values are assigned to y_{ij} , and the relationship between the y_{ij} outcomes and the event indicator.

Table 1.4: Three time points with censoring

Outcome	Ordinal dep. Variable	Event indicator
Censor at T_1	1	0
Event at T_1	1	1
Censor at T_2	2	0
Event at T_2	2	1
Censor at T_3	3	0
Event at T_3	3	1
Censor at T_4	4	0
Event at T_4	4	1

1.3 The models

1.3.1 Binary case: a 2-level model

In the binary case, the survival time of individual i at occasion j is treated as a set of dichotomous observations indicating whether or not an individual failed in each time unit until a person either experiences the event or is censored. Thus, each survival time is represented as a $t_{ij} \times 1$ vector of zeros for censored individuals, while for individuals experiencing the event the last element of this $t_{ij} \times 1$ vector of zeros is changed to a one. These multiple person-time indicators are then treated as distinct observations in a dichotomous regression model. In the case of clustered data, a random-effects dichotomous regression model is used. This method has been called the pooling of repeated observations method by Cupples (1985). It is particularly useful for handling time-dependent covariates and fitting nonproportional hazards models because the covariate values can change across each individuals' t_{ij} time points.

For this approach, define p_{ijt} to be the probability of failure in time interval t , conditional on survival prior to t :

$$p_{ijt} = \Pr[t_{ij} = t \mid t_{ij} \geq t]$$

Similarly, $1 - p_{ijt}$ is the probability of survival beyond time interval t , conditional on survival prior to t . The proportional hazards model is then written as

$$\log[-\log(1 - p_{ijt})] = \alpha_{0t} + \mathbf{x}'_{ijt}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{v}_i$$

and the corresponding proportional odds model is

$$\log[p_{ijt}/(1 - p_{ijt})] = \alpha_{0t} + \mathbf{x}'_{ijt}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{v}_i$$

where now the covariates \mathbf{x} can vary across time and so are denoted as \mathbf{x}_{ijt} . Augmenting the model intercept, which we will denote α_{01} , the remaining intercept terms α_{0t} ($t = 2, \dots, m$) are obtained by including as regressors $m - 1$ time indicators representing deviations from the first time point. Because the covariate vector \mathbf{x} now varies with t , this approach automatically allows for time-dependent covariates, and relaxing the proportional hazards assumption only involves including interactions of covariates with the $m - 1$ time point dummy codes. It is further assumed that the random effects vector has a $N(\mathbf{0}, \Phi_{(2)})$ distribution.

In the examples to follow, two random intercept models are fitted to the data described in Section 1.2.2. The type of intervention (CC and/or TV), the gender of the student and the interactions between gender and time (SexTC and SexTD) are included as fixed effects, along with indicators of the time of assessment (TimeC and TimeD).

1.3.2 Ordinal case: 2-level model

Let y_{ij} denote an ordinal outcome variable that takes on discrete positive values $t = 1, 2, \dots, m$. In previous examples we assumed that y_{ij} has C categories or distinct values, however here to be consistent with the survival analysis notation we will use m to represent the number of ordinal categories. The subscript (i, j) denotes subject j , $j = 1, 2, \dots, n_i$ nested within level-2 unit i , $i = 1, 2, \dots, N$. In the present context the level-1 units j indicates students and the level-2 unit i indicates schools. Note, that as another example of this type of model, one could have multiple failure times nested within individuals.

Let δ_{ij} denote the censor/event indicator, then $\delta_{ij} = 1$ if the event occurs and $\delta_{ij} = 0$ if an observation is censored. In survival analysis each ij is observed until time t_{ij} and if an event occurs $t_{ij} = t$ and $\delta_{ij} = 1$. If the observation is censored at $t_{ij} = t$, then $\delta_{ij} = 0$.

In the case of censoring it is assumed that a unit is observed at t_{ij} but not at t_{ij+1} . As described in Hedeker, Siddiqui & Hu (2000), if events occur within continuous time intervals (*i.e.*, grouped-time), for example, a student initiates smoking experimentation in the past year, use of the complementary log-log link for an ordinal outcome is equivalent to a proportional hazards model in continuous time. Therefore, the grouped-time proportional hazards mixed model can be written as:

$$\log \left[-\log(1 - P_{ijt}) \right] = \gamma_t + \mathbf{x}_{ij}' \boldsymbol{\beta} + \mathbf{z}_{ij}' \mathbf{v}_i$$

where \mathbf{x}_{ij} is a vector of explanatory variables and \mathbf{z}_{ij} a vector of random effects. Typically, the elements of \mathbf{z}_{ij} are a subset of \mathbf{x}_{ij} . For example, the elements of \mathbf{z}_{ij} might correspond to the intercept and age, whereas \mathbf{x}_{ij} would include these two terms plus any additional model covariates. It is assumed that the random effects \mathbf{v}_i are from a normal distribution with mean zero and covariance matrix $\Phi_{(2)}$.

P_{ijt} denotes the probability that an event takes place up to and including the interval designated at time t_{ij} . Thus, P_{ijt} represents a cumulative probability of failure, whereas p_{ijt} is the interval-specific failure probability. Also, γ_t represent threshold values, and in the present context these reflect the baseline hazard (*i.e.*, the hazard when all covariates equal 0). These threshold parameters are akin to the intercept parameters α_{0t} in the dichotomous version of the model. The plus sign following γ_t means that a positive regression coefficient for a covariate indicates an increased hazard (*i.e.*, the event occurs sooner) as values of the covariate increase.

1.4 Example: A proportional hazards model- Binary case

1.4.1 Introduction

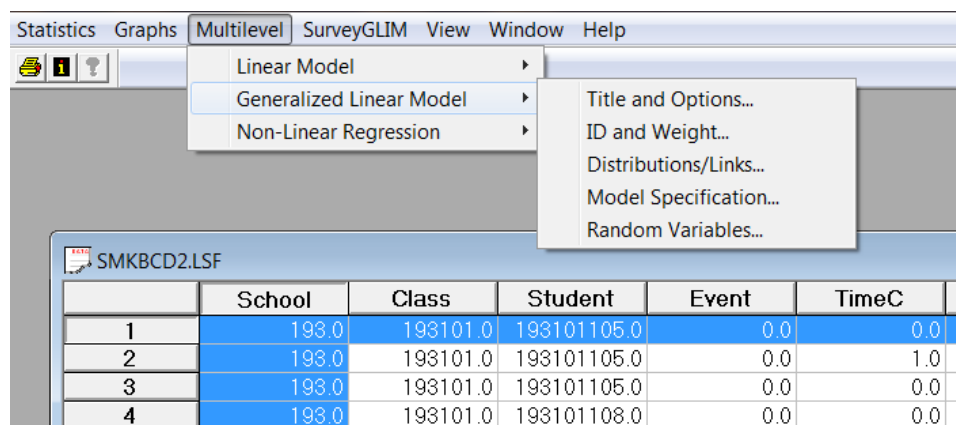
The first model fitted to the data will use the binary case and is of the form

$$\log[-\log(1 - p_{ijt})] = \alpha_{01} + (TimeC_{ij})\alpha_{02} + (TimeD_{ij})\alpha_{03} + (SexM_{ij})\beta_1 + (CC_j)\beta_2 + (TV_j)\beta_3 + v_{0i}.$$

In the current model specification, the baseline hazard is a function of the model intercept and the coefficients for the time indicators. Specifically, the baseline hazard estimate at the first time point equals the estimated model intercept, the baseline hazard estimate at the second time point is the sum of the model intercept and the estimated coefficient for the TimeC indicator, the baseline hazard at the third time point is the sum of the model intercept and the estimated coefficient for the TimeD indicator. Thus, two of these baseline hazard estimates involve sums of the estimated parameters.

1.4.2 Setting up the analysis

Start by opening the file **SMKBCD2.LSF** from the **Multilevel Generalized Linear Model Examples** folder selecting the **Multilevel, Generalized Linear** option from the main menu bar as shown below.



Enter (optional) a Title in the **Title and Options...** dialog.

Title and Options

Title:
TVSFP Ondet of Smoking (Waves B through D) Survival Analysis

Maximum Number of Iterations: 100

Convergence Criterion: 0.0001

Missing Data Value: -999999

Dependent Missing Value: -999999

Optimization Method

☐ MAP ☒ Quadrature

Number of Quadrature Points: 25

Additional Output

☐ Residual files ☐ No data summary

☐ Asymptotic covariance

Next >> Cancel OK

To build syntax, proceed to the Random Variables screen and click the Finish button

Change the number of quadrature points to 25, then click the **Next** button to obtain the **ID and Weight...** dialog.

ID and Weight Variables

Variables in data:

School
Class
Student
Event
TimeC
TimeD
SexM
CC
TV
CCTV
SexTC
SexTD
CCTC
CCTD
TVTC
TVTD

Add >> << Remove

Level 2 ID variable:
School

Add >> << Remove

Level 3 ID variable:

Add >> << Remove

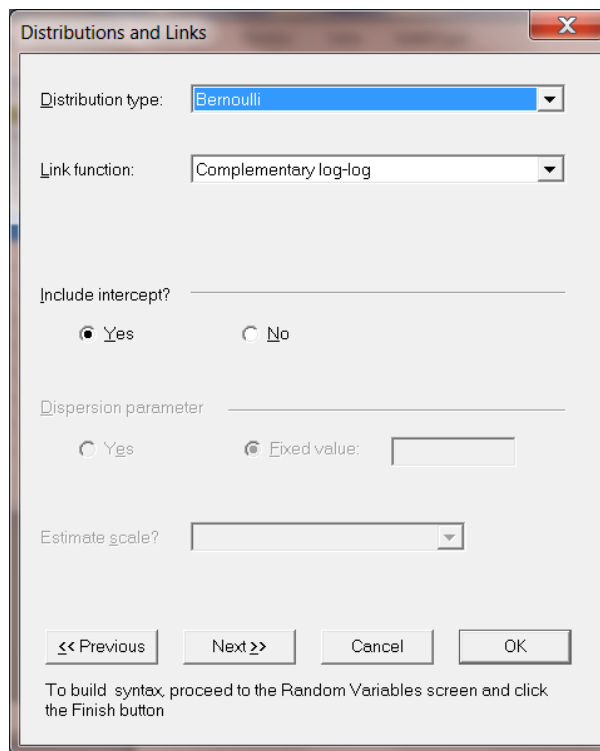
Weight variable:

<< Previous Next >> Cancel OK

To build syntax, proceed to the Random Variables screen and click the Finish button

The variable School, which defines the units within which students are nested, is selected as the Level-2 ID from the **Level-2 IDs** drop-down list box.

Next proceed to the **Distributions and Links** dialog. Select Bernoulli as the **Distribution type** and Complementary log-log as the **Link function** as shown below.



Distributions and Links

Distribution type: Bernoulli

Link function: Complementary log-log

Include intercept? ☒ Yes ☐ No

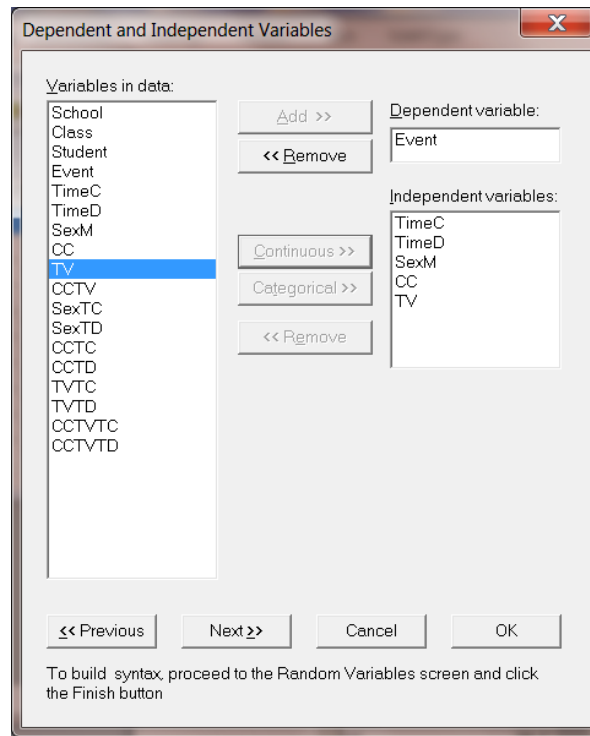
Dispersion parameter ☐ Yes ☒ Fixed value:

Estimate scale?

<< Previous Next >> Cancel OK

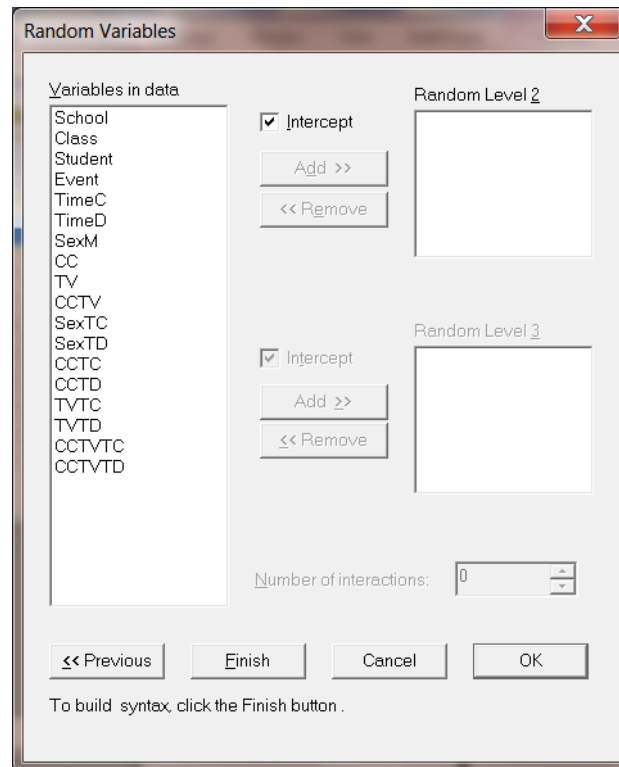
To build syntax, proceed to the Random Variables screen and click the Finish button

Click the **Include intercept?** Radio button and then click the **Next** button to obtain the **Dependent and Independent Variables** dialog.



Select the **binary** dependent (outcome) variable Event from the **Variables in data** list. Once done, TimeC, TimeD, SexM, CC, and TV are selected as the predictors (independent variables) of the fixed part of the model as shown below.

Finally, the **Random Variables** dialog is selected and Intercept is selected as the only random variable. To produce a syntax file, click the **Finish** button on the **Random Variables** dialog.



```

SMKBCD2.PRL
MGLimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999
                Method=Quad NQUADPTS=25 ;
Title=TVSFP Ondet of Smoking (Waves B through D) Survival Analysis;
SY='SMKBCD2.LSF';
ID2=School;
DEPENDENT_MISS=-999999;
Distribution=BER;
Link=CLL;
Intercept=Yes;
DepVar=Event;
CoVars=TimeC TimeD SexM CC TV;
RANDOM2=intcept;

```

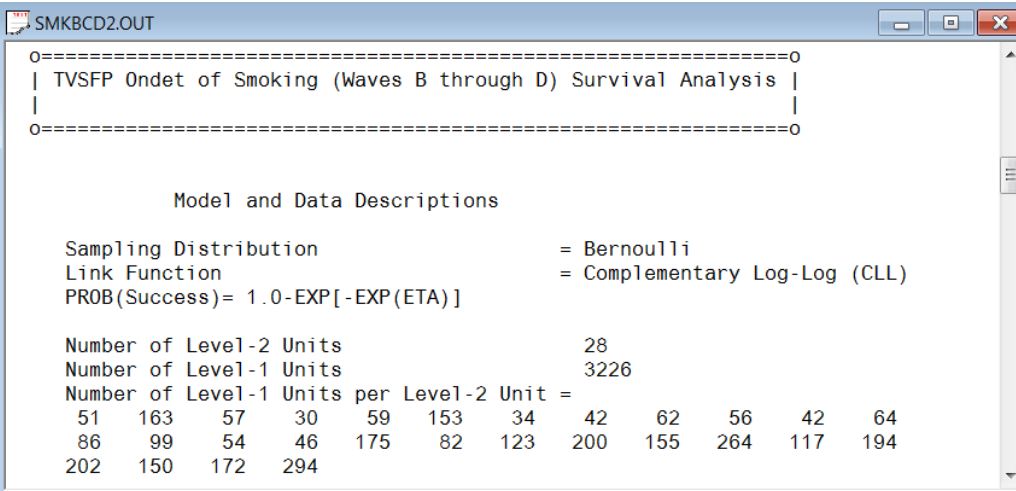
Next, click the **Run Prelis** icon on the main menu bar to run the analysis as shown below.



1.4.3 Discussion of results

Data summary

The portion of the output file shown below indicates that there are 28 schools. Nested within these level-2 units are 3226 measurements (note: this is not equal to the number of students because of the creation of person-time indicators in this binary version of the survival analysis model). A summary of the number of level-1 observations per level-2 unit is also given.



The screenshot shows a window titled "SMKBCD2.OUT" with a standard Windows interface. The main content area displays the following text:

```
=====0
| TVSFP Ondet of Smoking (Waves B through D) Survival Analysis |
|                                                                    |
0=====0
```

Model and Data Descriptions

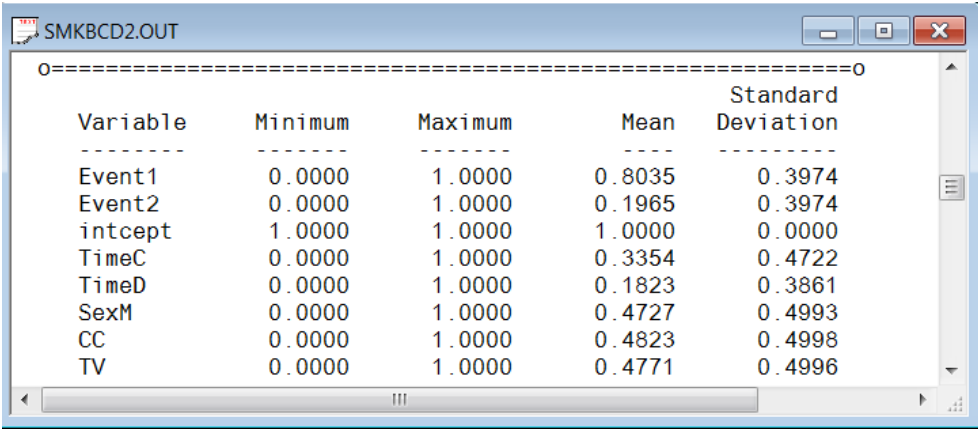
Sampling Distribution = Bernoulli
Link Function = Complementary Log-Log (CLL)
PROB(Success)= 1.0-EXP[-EXP(ETA)]

Number of Level-2 Units 28
Number of Level-1 Units 3226
Number of Level-1 Units per Level-2 Unit =

Level-2 Unit	Level-1 Units
51	163
86	99
202	150
57	57
30	30
59	59
153	153
34	34
42	42
62	62
56	56
42	42
64	64
82	82
123	123
200	200
155	155
264	264
117	117
194	194
172	172
294	294

Descriptive statistics

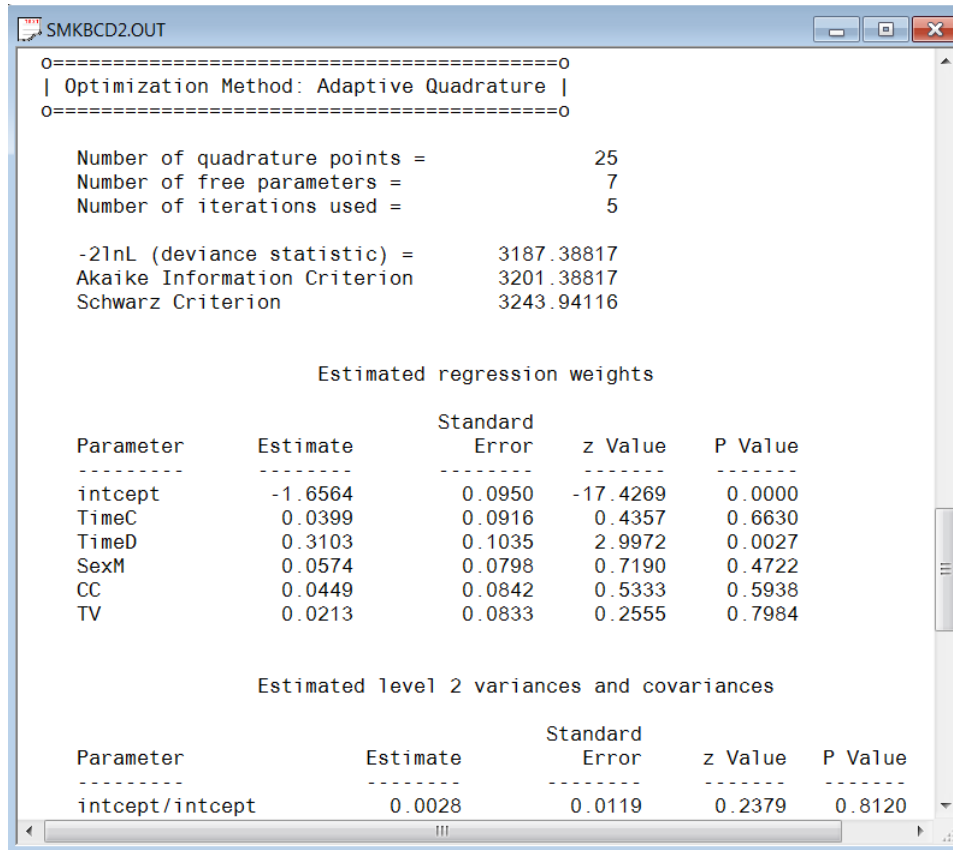
This is followed by descriptive statistics for all the variables. Except for the intercept term, the variables are all dichotomous. The proportions of subjects assigned a value of 0 or 1 are 0.80347 and 0.19653 respectively. In approximately 20% of the person-time indicators, an event occurred.



Variable	Minimum	Maximum	Mean	Standard Deviation
Event1	0.0000	1.0000	0.8035	0.3974
Event2	0.0000	1.0000	0.1965	0.3974
intcept	1.0000	1.0000	1.0000	0.0000
TimeC	0.0000	1.0000	0.3354	0.4722
TimeD	0.0000	1.0000	0.1823	0.3861
SexM	0.0000	1.0000	0.4727	0.4993
CC	0.0000	1.0000	0.4823	0.4998
TV	0.0000	1.0000	0.4771	0.4996

Fixed effects estimates

Parameter estimates are given in the next part of the output. The effect of SexM is positive and indicates that boys have a slightly, but non-significant, increased hazard (*i.e.*, a shorter time to the first occurrence), relative to girls. The coefficients associated with the TimeD indicator variable is significant at a 5% level. In contrast, the corresponding TimeC coefficient is not significant. These indicate that the baseline hazard does not significantly change between Waves B and C, however there is significant change between Waves B and D as relatively more students experiment with smoking at Wave D. Finally, the effects of the intervention variables CC and TV are not seen to be statistically significant, though the direction of their effects is positive (*i.e.*, increased hazard relative to the control group).



SMKBCD2.OUT

=====0
| Optimization Method: Adaptive Quadrature |
=====0

Number of quadrature points = 25
Number of free parameters = 7
Number of iterations used = 5

-2lnL (deviance statistic) = 3187.38817
Akaike Information Criterion 3201.38817
Schwarz Criterion 3243.94116

Estimated regression weights

Parameter	Estimate	Standard Error	z Value	P Value
intcept	-1.6564	0.0950	-17.4269	0.0000
TimeC	0.0399	0.0916	0.4357	0.6630
TimeD	0.3103	0.1035	2.9972	0.0027
SexM	0.0574	0.0798	0.7190	0.4722
CC	0.0449	0.0842	0.5333	0.5938
TV	0.0213	0.0833	0.2555	0.7984

Estimated level 2 variances and covariances

Parameter	Estimate	Standard Error	z Value	P Value
intcept/intcept	0.0028	0.0119	0.2379	0.8120

Intraclass correlation (ICC)

The last part of the output contains an estimate of the intra-cluster correlation. This estimate indicates a very modest school effect, and we also note that the random effect variance term is not significant. From this, we conclude that the time until the occurrence of an event does not vary significantly across schools. However, from a design point of view, because schools were randomized to the intervention conditions in this study, one can argue that the clustering attributable to schools is an important part of the model regardless of its significance.

Calculation of the intracluster correlation

```
-----
residual variance = pi*pi / 6 (assumed)
cluster variance = 0.0028

intracluster correlation = 0.0028 / ( 0.0028 + (pi*pi/6) ) = 0.002
```

Population Average Estimates

Parameter	Estimate	Standard Error	z Value	P Value
-----	-----	-----	-----	-----
intcept	-1.6553	0.0942	-17.5721	0.0000
TimeC	0.0399	0.0916	0.4357	0.6630
TimeD	0.3102	0.1034	2.9984	0.0027
SexM	0.0574	0.0798	0.7190	0.4721
CC	0.0449	0.0842	0.5334	0.5938
TV	0.0213	0.0833	0.2554	0.7984

1.4.4 Interpreting the output

Estimated unit-specific probabilities

We now use the estimated coefficients from the fitted model

$$\begin{aligned} \log \left[-\log(1 - \hat{p}_{ijt}) \right] &= \hat{\alpha}_{01} + (TimeC_{ij}) \hat{\alpha}_{02} + (TimeD_{ij}) \hat{\alpha}_{03} + (SexM_{ij}) \hat{\beta}_1 + (CC_j) \hat{\beta}_2 + (TV_j) \hat{\beta}_3 \\ &= -1.6564 + (TimeC_{ij})0.0399 + (TimeD_{ij})0.3103 + (SexM_{ij})0.0574 \\ &\quad + (CC_j)0.0449 + (TV_j)0.0213 \end{aligned}$$

and the inverse cumulative log-log link function

$$P(z) = 1 - \exp[-\exp(z)]$$

to calculate the probability of Event = 1 at various time points and for different covariate values.

At the first time point (Wave B), $TimeC_{ij} = TimeD_{ij} = 0$, and thus the relevant part of the fitted model (see above) is

$$\begin{aligned} \log \left[-\log(1 - \hat{p}_{ijt}) \right] &= \hat{\alpha}_{01} + (SexM_{ij}) \hat{\beta}_1 + (CC_j) \hat{\beta}_2 + (TV_j) \hat{\beta}_3 \\ &= -1.6564 + (SexM_{ij})0.0574 + (CC_j)0.0449 + (TV_j)0.0213 \end{aligned}$$

For female students (SexM = 0) from the control group (CC = TV = 0) the probability of smoking experimentation (Event = 1) at the point of post-intervention can be expressed as

$$P(Event = 1 \text{ at WaveB, female}) = 1 - \exp[-\exp(-1.6564)] \\ = 0.1737.$$

For male students in the control group adding the intercept with the SexM estimate together yields $z = -1.6564 + 0.0574 = -1.599$, and so

$$P(Event = 1 \text{ at WaveB, male}) = 1 - \exp[-\exp(-1.599)] \\ = .1830.$$

Results for all groups are summarized in Table 1.5. The probability of smoking experimentation at the time of post-intervention is larger for males than for females. The results also indicate an increased probability of failure with an increase of time. In the current model, it is assumed that the ratio of the estimated hazards over time will be constant for two individuals with the same values on the covariates. To check whether the effect of gender is dependent on time, and thus to check on the proportional hazards assumption, interactions with time indicators should be included in the model.

Table 1.5: Unit-specific probabilities for groups

Gender	CC	TV	WaveB (TimeC = 0, TimeD = 0)	WaveC (TimeC = 1, TimeD = 0)	WaveD (TimeC = 0, TimeD = 1)
Female	0	0	0.1737	0.1801	0.2291
	1	0	0.1809	0.1876	0.2383
	0	1	0.1771	0.1836	0.2335
	1	1	0.1844	0.1912	0.2428
Male	0	0	0.1830	0.1897	0.2409
	1	0	0.1905	0.1975	0.2505
	0	1	0.1865	0.1933	0.2454
	1	1	0.1942	0.2012	0.2551

Table 1.6 shows the differences between the estimated unit-specific probabilities and the observed proportions for each of the 24 subgroups formed by crossing all predictors currently in the model.

Looking at the direction of the differences, we note that for females all the estimated probabilities are larger in size than the observed ratios at WaveB, but consistently lower than the observed ratios at the next two time points, with the exception of the situation where TimeD = CC = TV = 1. It seems as if the model is overestimating the probabilities of failure at the first time point, but underestimating probabilities at the last time of measurement. However, the pattern for males is almost the opposite. At the first wave, only one estimated probability is larger than the observed proportion, at WaveC this is true for 2 of the four cells, and at WaveD for three of the four cells.

Table 1.6: Differences between unit-specific probabilities and observed proportions

Gender	CC	TV	Difference at WaveB (estimated – observed)	Difference at WaveC (estimated – observed)	Difference at WaveD (estimated – observed)
Female	0	0	0.0227	–0.0419	–0.0179
	1	0	0.0149	0.0016	–0.0117
	0	1	0.0091	–0.0174	–0.0875
	1	1	0.0204	–0.0058	0.0568
Male	0	0	–0.0150	–0.0073	–0.0361
	1	0	0.0165	–0.0025	0.0625
	0	1	–0.0275	0.0303	0.0064
	1	1	–0.0678	0.0613	0.0710

This trend could be the result of a gender effect (which we know to be non-significant in the current model) or from an interaction between gender and time. While only TimeD had a significant estimated coefficient, this apparent trend leads us to conclude that testing of the assumption of proportional hazards is appropriate. Specifically, the interaction between gender and the time of measurement will be explored.

1.5 Example: Survival analysis model for an ordinal outcome

1.5.1 Introduction

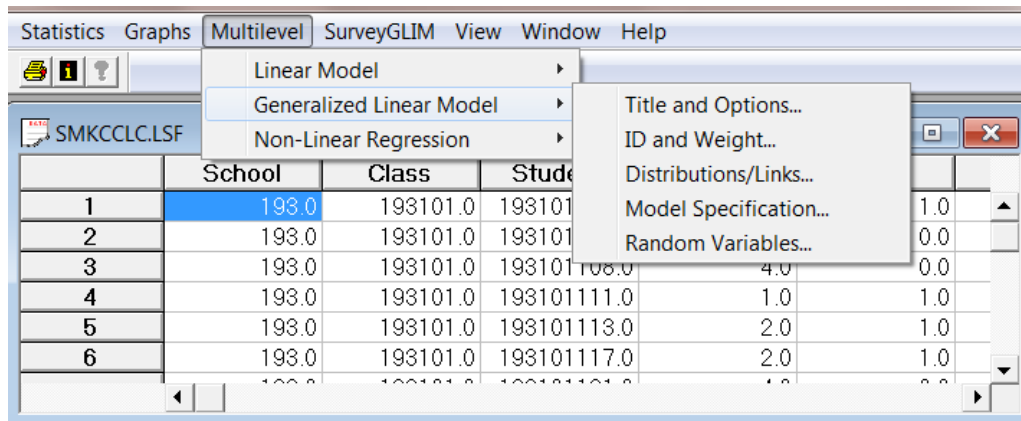
In this section, the re-formatted form of the data, as captured in **smkcclc.ss3** is used to fit a model to the data with the ordinal variable SmkOnset as outcome.

The model fitted to the data is of the form

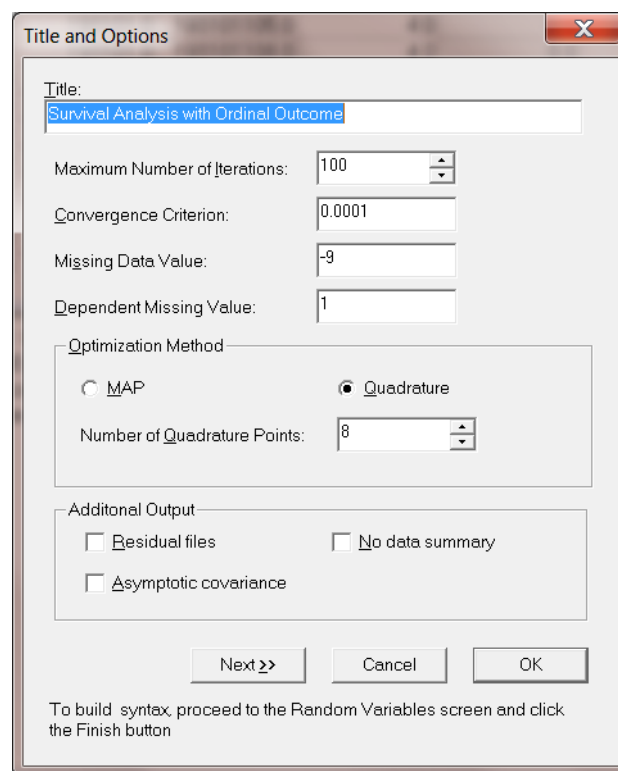
$$\log\left[-\log(1 - P_{ijt})\right] = \gamma_t + (SexM_{ij})\beta_1 + (CC_j)\beta_2 + (TV_j)\beta_3 + v_{0i}.$$

1.5.2 Setting up the analysis

Using the data in the LISREL spreadsheet **SMKCCLC.lsf**, we start by selecting the **Multilevel, Generalized Linear Model** from the main menu bar as shown below.



Start with the **Titles and Options** dialog and enter (optional) a title in the **Title** text box. Enter a value of -9 as the global missing value and a value of 1 as the dependent variable missing value.



Specify the number of quadrature points as 8. When done, click the **Next** button to proceed to the **ID and Weight Variables** dialog shown below and select School as the **level-2 ID variable**.

ID and Weight Variables

Variables in data:

- School
- Class
- Student
- SmkOnset
- Event
- SexM
- CC
- TV
- CCTV

Add >> << Remove

Level 2 ID variable: School

Add >> << Remove

Level 3 ID variable:

Add >> << Remove

Weight variable:

<< Previous Next >> Cancel OK

To build syntax, proceed to the Random Variables screen and click the Finish button

Next is the **Distributions and Links** dialog shown below. Select **Multinomial** as the distribution type and **Ordinal complementary log-log** as the link function.

Distributions and Links

Distribution type: Multinomial

Link function: Ordinal complementary log-log

Model terms: Add

Include intercept? ☒ Yes ☐ No

Dispersion parameter ☐ Yes ☒ Fixed value:

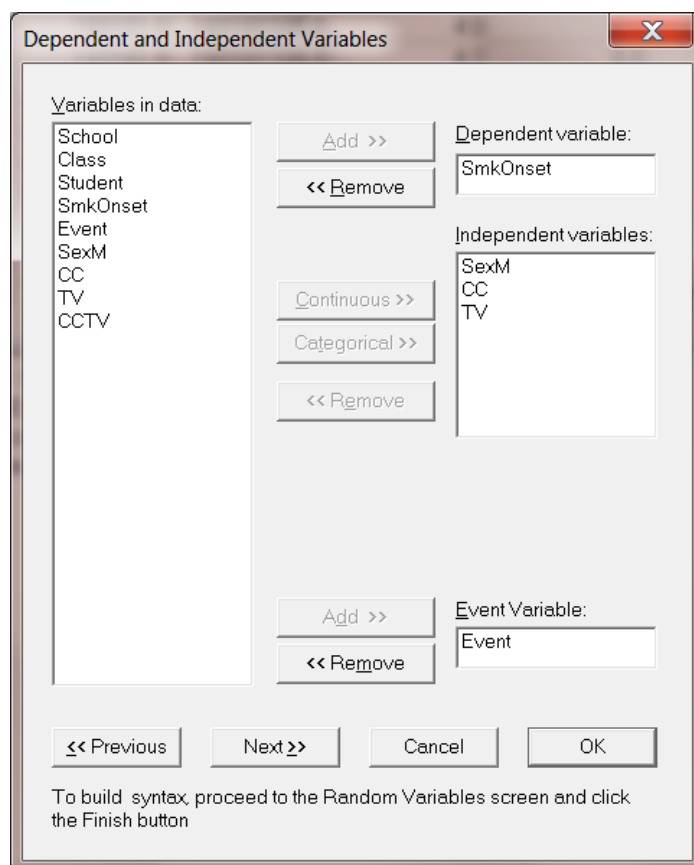
Estimate scale?

<< Previous Next >> Cancel OK

To build syntax, proceed to the Random Variables screen and click the Finish button

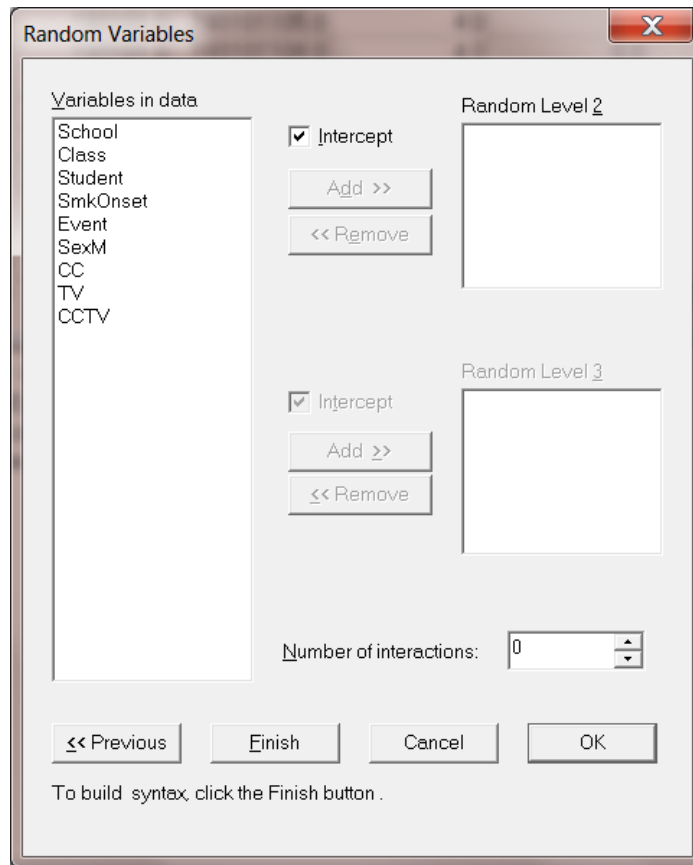
Choose **Add** for the model terms and then click the **Next** button to invoke the **Dependent and Independent Variables** dialog shown next.

Select SmkOnset as the dependent (outcome variable) and SexM, CC, and TV are specified as the predictors (independent variables) of the fixed part of the model. Before proceeding to the **Random Variables** dialog, select Event as the **Event Variable**.



The dialog box titled "Dependent and Independent Variables" contains the following elements:

- Variables in data:** A list box containing School, Class, Student, SmkOnset, Event, SexM, CC, TV, and CCTV.
- Buttons:** "Add >>", "<< Remove", "Continuous >>", "Categorical >>", and "<< Remove".
- Dependent variable:** A text box containing "SmkOnset".
- Independent variables:** A list box containing SexM, CC, and TV.
- Event Variable:** A text box containing "Event".
- Navigation buttons:** "<< Previous", "Next >>", "Cancel", and "OK".
- Footer text:** "To build syntax, proceed to the Random Variables screen and click the Finish button".



Click the **Finish** button (see above) to produce a syntax file and then click on the **Run Prelis** icon to run the analysis.

1.5.3 Discussion of results

Selected portions of the output file **smkccd1.out** are shown below.

Data summary and descriptive statistics

The portion of the output file shown below indicates that there are 28 schools, with 1556 students nested within these. This is followed by descriptive statistics for all the variables. Note that all three predictor variables are dichotomous in nature.

SMKCCCLC.OUT

Model and Data Descriptions

Sampling Distribution

Link Function

Number of Level-2 Units

Number of Level-1 Units

Number of Level-1 Units per Level-2 Unit =

217424132970181725271931

405127237939521067414251104

817494151

= Multinomial

= Cumulative CLL

28

1556

=====0

| Descriptive statistics for all the variables in the model |

=====0

Variable	Minimum	Maximum	Mean	Standard Deviation
Smk0nse1	0.0000	1.0000	0.3046	0.4604
Smk0nse2	0.0000	1.0000	0.3175	0.4656
Smk0nse3	0.0000	1.0000	0.3779	0.4850
SexM	0.0000	1.0000	0.4769	0.4996
CC	0.0000	1.0000	0.4788	0.4997
TV	0.0000	1.0000	0.4904	0.5001

Fixed effects estimates

This is followed by the results for the model specified, but without any random effects. In this format, none of the included predictors are significant. It will be interesting to compare these results with those obtained once the hierarchical structure of the data has been taken into account.

SMKCCCLC.OUT

```
=====0
| Optimization Method: Adaptive Quadrature |
=====0
```

```
Number of quadrature points =      8
Number of free parameters =      7
Number of iterations used =      5
```

```
-2lnL (deviance statistic) =      3187.38817
Akaike Information Criterion      3201.38817
Schwarz Criterion                3238.83729
```

Estimated regression weights

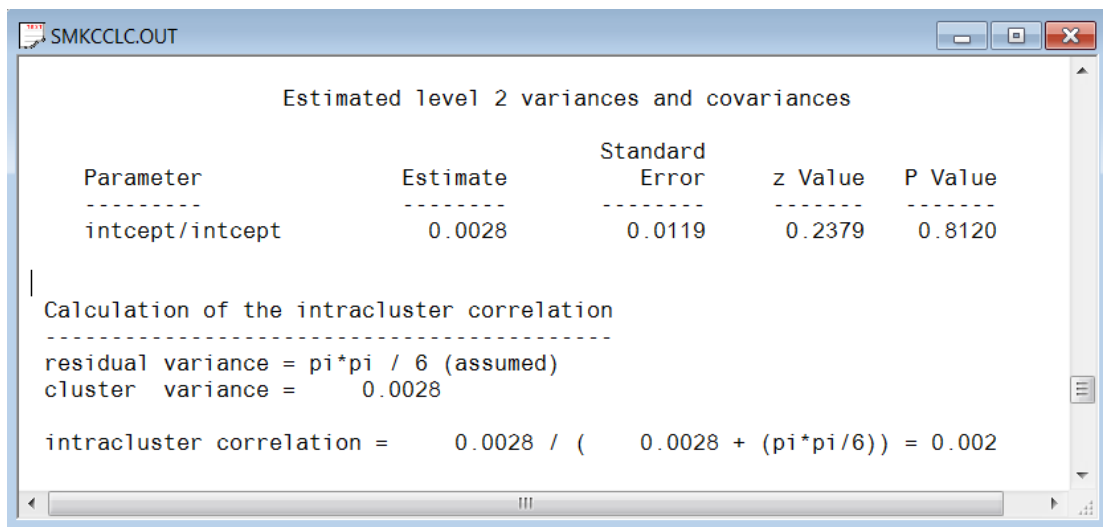
Parameter	Estimate	Standard Error	z Value	P Value
-----	-----	-----	-----	-----
Thresh1	-1.6564	0.0950	-17.4269	0.0000
Thresh2	-0.9431	0.0869	-10.8519	0.0000
Thresh3	-0.4313	0.0848	-5.0846	0.0000
SexM	0.0574	0.0798	0.7190	0.4722
CC	0.0449	0.0842	0.5333	0.5938
TV	0.0213	0.0833	0.2555	0.7984

111

Parameter estimates are given in the next part of the output. Taking the hierarchical structure into account and allowing for the intercept to vary randomly over the schools had little effect on the significance level of the 3 covariates: all are still non-significant. We note that the three thresholds, which represent the cumulative baseline hazard, are estimated as -1.6564 , -0.9431 , and -0.4313 respectively. An alternative parameterization is also given. Here, the first threshold has been set to zero and as a result, the intercept and second and third threshold estimates are calculated as -1.6564 , 0.7133 , and 1.2251 respectively.

Random effects estimates and intraclass correlation (ICC)

This part of the output shows the estimates of the random effects and an estimate of the intra-cluster correlation. There is no evidence of significant random variation in the intercept over the schools ($p = 0.8120$). The intra-cluster correlation coefficient shown is based on the use of the complementary log-log link function for these data, which results in a residual variance of $\pi^2/6$ (see Agresti, 2002).



Parameter	Estimate	Standard Error	z Value	P Value
intcept/intcept	0.0028	0.0119	0.2379	0.8120

Calculation of the intraclass correlation

residual variance = $\pi^2/6$ (assumed)

cluster variance = 0.0028

intraclass correlation = $0.0028 / (0.0028 + (\pi^2/6)) = 0.002$

Population Average Estimates

Parameter	Estimate	Standard Error	z Value	P Value
-----	-----	-----	-----	-----
Thresh1	-1.6553	0.0942	-17.5716	0.0000
Thresh2	-0.9422	0.0863	-10.9226	0.0000
Thresh3	-0.4308	0.0845	-5.0987	0.0000
SexM	0.0574	0.0798	0.7190	0.4722
CC	0.0449	0.0842	0.5334	0.5938
TV	0.0213	0.0833	0.2554	0.7984

1.5.4 Interpreting the output

Comparing binary and ordinal models

When the number of measurement occasions is not too large, the binary outcome model utilizing dummy variables to represent the measurement occasions can be useful in fitting survival analysis models. Additionally, the binary model easily allows relaxation of the proportional hazards assumption for model covariates through inclusion of interaction terms with the time point indicators. Finally, though not illustrated here, the binary model can also handle time-dependent covariates in the same manner as the covariate by time interactions. When the number of occasions is very large, however, the number of time point indicators that must be created for the binary model, and the resulting size of the data set, can get very large and unwieldy. In this case, the ordinal outcome model such as the model discussed in this section is perhaps the better analysis option (though covariates must follow the proportional hazards assumptions and time-dependent covariates are not allowed). If the complementary log-log link function is selected (*i.e.*, the model is specified as a proportional hazards model), the binary and ordinal outcome models yield identical estimates for parameters that do not depend on time (Laara & Matthews, 1985). This is shown in Table 1.12. The regression coefficients are exactly the same for Male, CC, and TV. This is also true of their standard errors and so the p -values for both sets are identical. However, the intercept and threshold parameters, which do represent time-related information, are not the same with the exception of the first intercept. The reason for this is that the intercepts in the binary model represent the *interval-specific* baseline hazard, whereas their corresponding threshold parameters in the ordinal model represent the *cumulative* baseline hazard across the time intervals. These are only equivalent only for the first time interval and thereafter diverge in value and meaning. Finally, it should be mentioned that if one uses the logit link, in place of the complementary log-log link, the estimates (of the parameters not involving time) from the binary and ordinal models are not equivalent, though similar.

Table 1.12: Comparison of results of binary and ordinal outcome models

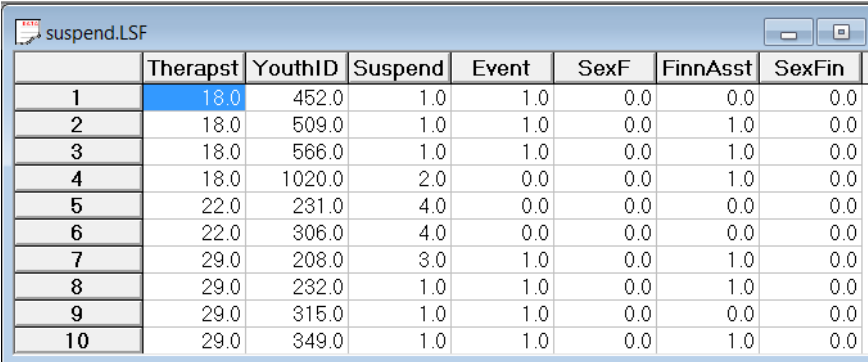
Term	Binary outcome (EVENT)	Ordinal outcome (SmkOnset)
Wave B baseline hazard binary α_{01} or ordinal γ_1	-1.6564	-1.6564
Wave C baseline hazard binary $\alpha_{01} + \alpha_{02}$ or ordinal γ_2	-1.65654+0.0399 = -1.6165	-0.9431
Wave D baseline hazard binary $\alpha_{01} + \alpha_{03}$ or ordinal γ_3	-1.6564 +0.3103 = -1.3461	-0.4313
Male β_1	0.0574	0.0574
CC β_2	0.0449	0.0449
TV β_3	0.0213	0.0213
$-2\ln L$	3187.38817	3187.38817
AIC	3201.38817	3201.38817
Schwarz	3243.94116	3238.83729
No. of parameters	7	7

Notice also that the likelihood values for the two representations are identical, as are the AIC values. The Schwarz values are not the same because the numbers of observations in the two representations are different. That is, because the binary-case data set consists of multiple person-time indicators for each outcome, the numbers of observations in the binary-case data set is inflated, relative to the ordinal case.

1.6 Two-level survival analysis models

1.6.1 The data

The data set for this example is taken from Schoenwald & Henggeler (2005). Children in the study were assigned to therapists and followed across time. At the child level, data were collected at baseline (pre-treatment, T_0), post-treatment (T_1), 6 months post-treatment (T_2), and 12 months post-treatment (T_3). The outcome of interest is whether a child was suspended in the current school year, assessed at T_0 , T_1 , T_2 , or T_3 . Specifically, here, we will focus on the time until the first school suspension as the "survival" outcome. As indicated in more detail later, this is indicated by a combination of the variables Event and Suspend: for example, if the student was suspended, the indicator Event is given the value 1 and Suspend will indicate the time period during which this occurred. However, there are also subjects who do not experience the event (*i.e.*, were not suspended), and who drop out of the study before its end. Such subjects are considered to be right-censored in the survival analysis literature, and for these subjects the Event variable is coded 0 and the Suspend variable indicated the last time period prior to their dropout from the study. For subjects who never experience the event and who never drop out, they receive Event codes of 0 and Suspend codes equal to the final time point. In addition to these data concerning school suspension, the gender of each student was also recorded, as well as whether or not the student's family was receiving financial assistance. The first 10 cases of the data set **suspend.lsf** stored in the **Multilevel Generalized Linear Model Example** folder are shown below.



	Therapst	YouthID	Suspend	Event	SexF	FinnAsst	SexFin
1	18.0	452.0	1.0	1.0	0.0	0.0	0.0
2	18.0	509.0	1.0	1.0	0.0	1.0	0.0
3	18.0	566.0	1.0	1.0	0.0	1.0	0.0
4	18.0	1020.0	2.0	0.0	0.0	1.0	0.0
5	22.0	231.0	4.0	0.0	0.0	0.0	0.0
6	22.0	306.0	4.0	0.0	0.0	0.0	0.0
7	29.0	208.0	3.0	1.0	0.0	1.0	0.0
8	29.0	232.0	1.0	1.0	0.0	1.0	0.0
9	29.0	315.0	1.0	1.0	0.0	0.0	0.0
10	29.0	349.0	1.0	1.0	0.0	1.0	0.0

The variables of interest are:

- Therapst is the patient therapist ID (443 level-2 units).
- YouthID is the child's ID (1914 level-1 units).

- Suspend is an ordinal outcome variable that assumes values 1, 2, 3 or 4, corresponding to the time points T_0 , T_1 , T_2 , and T_3 .
- Event is the event indicator, where 1 indicates suspension took place and 0 that the observation was censored.
- SexF indicates the child's gender (1 = female; 0 = male).
- FinnAsst equals 1 if financial assistance is given to the student's family and 0 otherwise
- SexFin equals $\text{SexF} \times \text{FinnAsst}$ and therefore assumes values of 0 and 1.

1.6.2 The model

Let y_{ij} denote an ordinal outcome variable that takes on discrete positive values $t = 1, 2, \dots, m$. In previous examples we assumed that y_{ij} has C categories. For example, 1 = not depressed, 2 = mildly depressed, 3 = depressed and 4 = extremely depressed. The subscript (i, j) denotes subject j , $j = 1, 2, \dots, n_i$ nested within level-2 unit i , $i = 1, 2, \dots, N$. In the present context the level-1 units j indicates children and the level-2 unit i indicates therapists. Note, that as another example of this type of model, one could have multiple failure times nested within individuals.

Let δ_{ij} denote the censor/event indicator, then $\delta_{ij} = 1$ if the event occurs and $\delta_{ij} = 0$ if an observation is censored. In survival analysis each ij is observed until time t_{ij} and if an event occurs $t_{ij} = t$ and $\delta_{ij} = 1$. If the observation is censored at $t_{ij} = t$ then $\delta_{ij} = 0$.

In the case of censoring it is assumed that a unit is observed at t_{ij} but not at t_{ij+1} . Hedeker, Siddiqui & Hu (2000) showed that if events occur within continuous time intervals (*i.e.*, grouped-time), for example, a student is suspended in the past year, use of the complementary log-log link for an ordinal outcome is equivalent to a proportional hazards model in continuous time. Therefore, the grouped-time proportional hazards mixed model can be written as:

$$\log \left[-\log \left(1 - P(t_{ij}) \right) \right] = \gamma_t + \mathbf{w}_{ij}' \boldsymbol{\alpha} + \mathbf{x}_{ij}' \boldsymbol{\beta}_i$$

where \mathbf{w}_{ij} is a vector of explanatory variables and \mathbf{x}_{ij} a vector of fixed effects. Typically, the elements of \mathbf{x}_{ij} are a subset of \mathbf{w}_{ij} . For example, the elements of \mathbf{x}_{ij} might correspond to the intercept and age, whereas \mathbf{w}_{ij} would include these two terms plus any additional model covariates. It is assumed that the random effects $\boldsymbol{\beta}_i$ are from a normal distribution with mean zero and covariance matrix $\boldsymbol{\Phi}$.

$P(t_{ij})$ denotes the probability that an event takes place in the interval designated at time t_{ij} . γ_t represent threshold values, and in the present context these reflect the baseline hazard (*i.e.*, the hazard when all covariates equal 0). The plus sign following γ_t means that a positive α indicates an increased hazard (*i.e.*, the event occurs sooner) as values of the covariate increase.

1.6.3 Survival data as ordinal outcomes

Assume 4 time points with no intermittent censoring and let y denote the outcome variable. Let us first consider subjects who were suspended at some point in the study. For these subjects, the variable Event will be coded as 1 and the coding of the Suspend variable will be as follows.

Suspend:

- $y_{ij} = 1$: Student first suspended at T_0 .
- $y_{ij} = 2$: Student not suspended at T_0 , but first suspended at T_1 .
- $y_{ij} = 3$: Student not suspended at T_0 or T_1 , but first suspended at T_2 .
- $y_{ij} = 4$: Student not suspended at T_0 , T_1 or T_2 , but first suspended at T_3 .

Similarly, subjects who were never censored would have the variable Event coded as 0, and the following codes for the Suspend variable.

Suspend:

- $y_{ij} = 1$: Student not suspended at T_0 and no data beyond T_0 .
- $y_{ij} = 2$: Student not suspended at T_0 or T_1 , and no data beyond T_1 .
- $y_{ij} = 3$: Student not suspended at T_0, T_1 , or T_2 , and no data beyond T_2 (*i.e.*, no data at T_3).
- $y_{ij} = 4$: Student not suspended at T_0, T_1, T_2 , or T_3 .

Table 2.1 shows how values are assigned to y_{ij} , and the relationship between the y_{ij} outcomes and the event indicator. It should be noted that one could also fit grouped-time survival models using dichotomous indicators of event/censoring across the study time points. This approach, which is described in Singer and Willett (1993), can also be done in SuperMix, though additional data setup and manipulation is required. The advantage of representing the survival data as ordinal outcomes is that there is no need to include time indicators since the thresholds take care of this. The ordinal presentation is also more efficient in terms of data set size, especially when the number of time points is large. More information on these two different approaches can be found in Hedeker, Siddiqui & Hu (2000).

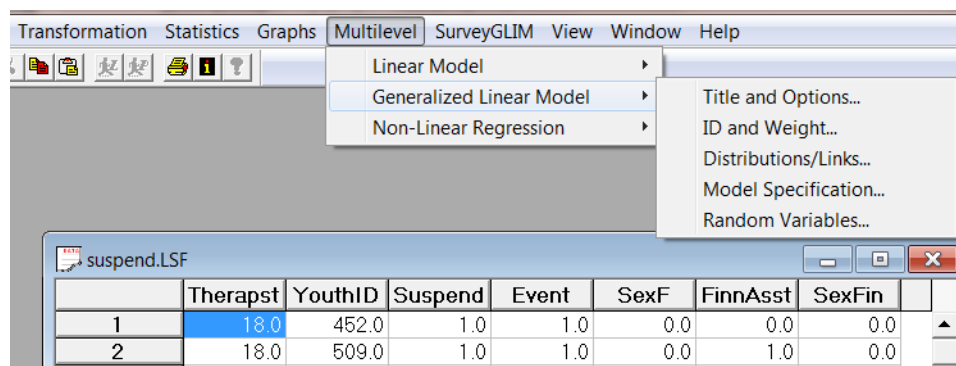
Table 2.1: Four time points with censoring

Outcome	Ordinal dep. Variable	Event indicator
Censor at T_1	1	0
Event at T_1	1	1
Censor at T_2	2	0
Event at T_2	2	1
Censor at T_3	3	0
Event at T_3	3	1
Censor at T_4	4	0
Event at T_4	4	1

1.6.4 Setting up the analysis

The model is fitted to the data in **suspend.lsf** as follows. The first step is to create the **lsf** file shown above from the Excel file **suspend.csv**. This is accomplished by using the **Import Data File** option on the **File** menu to load the **Open** dialog box. Next, browse for and open the file **suspend.csv**. The file is now displayed as a LISREL spreadsheet window **suspend.lsf**.

Using the data in the LISREL spreadsheet **suspend.lsf**, we start by selecting the **Multilevel, Generalized Linear Model** from the main menu bar as shown below.



Start with the **Titles and Options** dialog and enter (optional) a title in the **Title** text box.

Title and Options

Title: Survival Analysis Using Ordered Responses

Maximum Number of Iterations: 100

Convergence Criterion: 0.0001

Missing Data Value: -999999

Dependent Missing Value: -999999

Optimization Method

☐ MAP ☒ Quadrature

Number of Quadrature Points: 25

Additional Output

☐ Residual files ☐ No data summary

☐ Asymptotic covariance

Next >> Cancel OK

To build syntax, proceed to the Random Variables screen and click the Finish button

Specify the number of quadrature points as 25. When done, click the **Next** button to proceed to the **ID and Weight Variables** dialog shown below and select Therapist as the **level-2 ID variable**.

ID and Weight Variables

Variables in data:

Therapist
YouthID
Suspend
Event
SexF
FinnAsst
SexFin

Add >> << Remove

Level 2 ID variable: Therapist

Add >> << Remove

Level 3 ID variable:

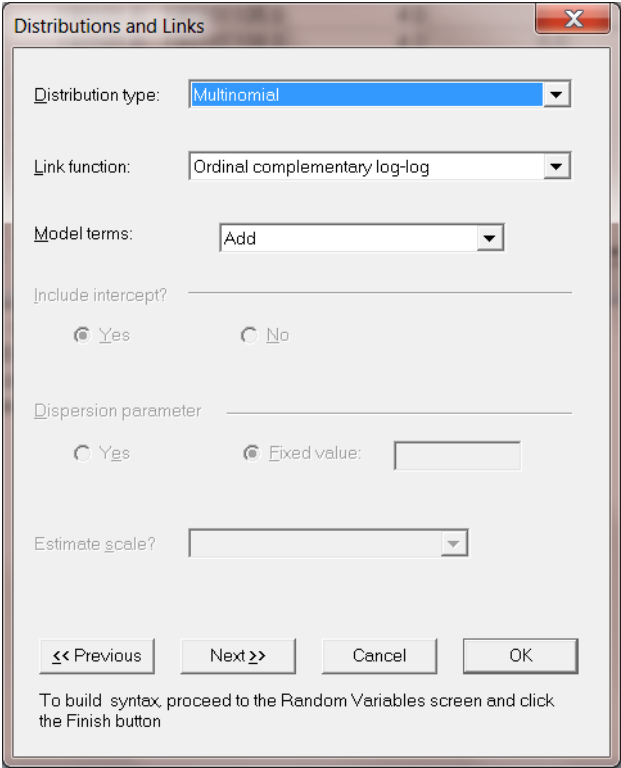
Add >> << Remove

Weight variable:

<< Previous Next >> Cancel OK

To build syntax, proceed to the Random Variables screen and click the Finish button

Next is the **Distributions and Links** dialog shown below. Select **Multinomial** as the distribution type and **Ordinal complementary log-log** as the link function.



Distributions and Links

Distribution type: Multinomial

Link function: Ordinal complementary log-log

Model terms: Add

Include intercept?

☒ Yes ☐ No

Dispersion parameter

☐ Yes ☒ Fixed value:

Estimate scale?

<< Previous Next >> Cancel OK

To build syntax, proceed to the Random Variables screen and click the Finish button

Choose **Add** for the model terms and then click the **Next** button to invoke the **Dependent and Independent Variables** dialog shown next.

Select **Suspend** as the dependent (outcome variable) and **SexF**, **FinnAsst** and **SexFin** are specified as the predictors (independent variables) of the fixed part of the model. Before proceeding to the last dialog, select **Event** as the **Event Variable**.

Dependent and Independent Variables

Variables in data:

- Therapst
- YouthID
- Suspend
- Event**
- SexF
- FinnAsst
- SexFin

Dependent variable:

Suspend

Independent variables:

- SexF
- FinnAsst
- SexFin

Event Variable:

Event

Buttons: << Previous, Next >>, Cancel, OK

To build syntax, proceed to the Random Variables screen and click the Finish button

Random Variables

Variables in data:

- Therapst
- YouthID
- Suspend
- Event
- SexF
- FinnAsst
- SexFin

Random Level 2

☒ Intercept

Random Level 3

☒ Intercept

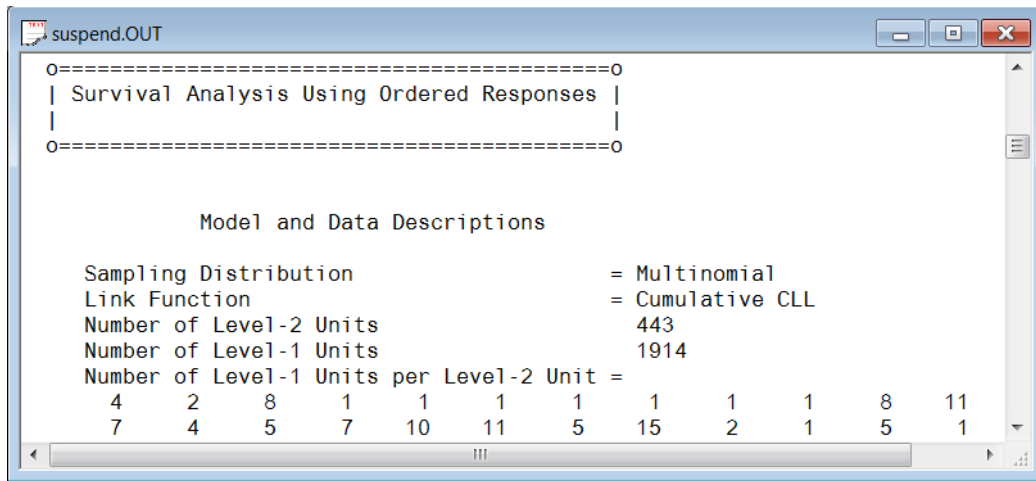
Number of interactions: 0

Buttons: << Previous, Finish, Cancel, OK

To build syntax, click the Finish button.

Click the **Finish** button to produce a syntax file and then click on the **Run Prelis** icon to run the analysis.

1.6.5 Discussion of results



```

suspend.OUT
=====0
| Survival Analysis Using Ordered Responses |
|                                         |
=====0

Model and Data Descriptions

Sampling Distribution                    = Multinomial
Link Function                          = Cumulative CLL
Number of Level-2 Units                 443
Number of Level-1 Units                1914
Number of Level-1 Units per Level-2 Unit =
  4   2   8   1   1   1   1   1   1   1   8   11
  7   4   5   7  10  11   5  15   2   1   5   1
  
```

The portion of the output file shown below indicates that there are 443 therapists. Nested within these level-2 units are 1914 subjects. A summary of the number of level-1 observations per level-2 unit (only first two lines shown) is also given.

This part of the output is followed by descriptive statistics for all the variables. The variable Suspend has four categories with values 1, 2, 3 and 4. Except for the intercept term, the remaining variables are all dichotomous.

=====0				
Descriptive statistics for all the variables in the model				
=====0				
Variable	Minimum	Maximum	Mean	Standard Deviation

Suspend1	0.0000	1.0000	0.4316	0.4954
Suspend2	0.0000	1.0000	0.1855	0.3888
Suspend3	0.0000	1.0000	0.1102	0.3133
Suspend4	0.0000	1.0000	0.2727	0.4455
SexF	0.0000	1.0000	0.3459	0.4758
FinnAsst	0.0000	1.0000	0.3626	0.4809
SexFin	0.0000	1.0000	0.1155	0.3197

The proportions of subjects assigned a value of 1, 2, 3 or 4 are 0.432, 0.185, 0.110 and 0.273 respectively. A crosstabulation of Suspend by Event is given in Table 2.2. It follows that, for example, 773 students out of the 1914 in the study were suspended prior to treatment (T_0). For 53 children, we only know that they were not suspended at T_0 , thereafter they are missing and treated as right-censored.

Table 2.2: Crosstabulation of Suspend by Event

T_0	T_1	T_2	T_3
773	255	106	72
53	100	105	450

Parameter estimates are given in the next part of the output. We conclude that there is no gender-financial assistance interaction and that all the remaining parameter estimates are significant. The effect of SexF is negative indicating that girls have a significantly decreased hazard (*i.e.*, a longer time to the first suspension), relative to boys. The FinnAsst estimate is positive indicating an increased hazard (shorter time to first suspension) for children from families receiving financial assistance, relative to children from families not receiving this assistance.

```

=====0
| Optimization Method: Adaptive Quadrature |
=====0

Number of quadrature points =      25
Number of free parameters =       8
Number of iterations used =       3

-2lnL (deviance statistic) =      4741.46567
Akaike Information Criterion      4757.46567
Schwarz Criterion                  4801.92128

```

Estimated regression weights

Parameter	Estimate	Standard Error	z Value	P Value
-----	-----	-----	-----	-----
Thresh1	-0.6565	0.0549	-11.9547	0.0000
Thresh2	-0.2238	0.0518	-4.3221	0.0000
Thresh3	-0.0326	0.0511	-0.6384	0.5232
Thresh4	0.1211	0.0512	2.3662	0.0180
SexF	-0.3208	0.0818	-3.9209	0.0001
FinnAsst	0.2005	0.0742	2.7012	0.0069
SexFin	-0.0236	0.1345	-0.1755	0.8607

The last part of the output contains an estimate of the intra-cluster correlation and the population average estimates.

Although the intra-cluster correlation estimate indicates a modest therapist effect, the random effect variance term is highly significant. From this we conclude that the time until suspension does vary significantly across therapists.

Calculation of the intraclass correlation

residual variance = $\pi^2 \pi / 6$ (assumed)

cluster variance = 0.0840

intraclass correlation = $0.0840 / (0.0840 + (\pi^2 \pi / 6)) = 0.049$

Population Average Estimates

Parameter	Estimate	Standard Error	z Value	P Value
Thresh1	-0.6374	0.0529	-12.0492	0.0000
Thresh2	-0.2161	0.0501	-4.3148	0.0000
Thresh3	-0.0317	0.0496	-0.6391	0.5227
Thresh4	0.1158	0.0497	2.3319	0.0197
SexF	-0.3147	0.0802	-3.9220	0.0001
FinnAsst	0.1956	0.0724	2.7032	0.0069
SexFin	-0.0219	0.1318	-0.1661	0.8681