EXERCISE 1: DISCRETE-TIME EVENT HISTORY MODELLING OF THE TIME TO A SINGLE EVENT, USING SPSS

In this first exercise, we will analyse a subsample of data from the National Child Development Study. This is a cohort study, following all individuals born in Britain in a particular week of March 1958. Partnership histories were collected when the respondents were aged 33. Here, we analyse the time from age 16 to an individual's first partnership (either a marriage or cohabitation). The SPSS data file is called `ex1.sav`.

The file is currently in the form of one record per individual and contains the following variables:

AGE1ST	Age at first partnership (equals 33 for censored cases)
EVENT	Indicator of event occurrence (1=partnered, 2=single, i.e. censored)
AGELEFT	Age at which respondent left full-time education
FEMALE	Respondent's gender (1=female, 0=male)
REGION	Region of residence 1=Scotland and the North 2=Wales and the Midlands 3=Southern and Eastern 4=South East, including London
FCLASS	Father's social class (defined by occupation) 1=class I or II (professional and managerial) 2=class III 3=class IV or V (manual)

1. Exploratory Analysis: Examining the Hazard and Survivor Functions

Before fitting any models, we will examine the hazard and survivor functions for each year from ages 16 to 33.

From the **Analyze** menu, select **Survival**, then **Life Tables**.

Specify the duration variable, by placing **AGE1ST** into the **Time:** box.

The next step is to specify the time intervals for which estimates of the hazard and survivor function will be calculated. We will calculate the hazard/survivor function for each year up to age 33. To do this, under **Display Time Intervals**, type **33** in the first box and **1** in the second to give '0 through 33 by 1'.

We now declare the censoring indicator. Put **EVENT** in the **Status:** box. Click on **Define Event**, and type 1 next to **Single value**. Event=1 indicates that an 'event', i.e. partnership formation, has been observed.

Now click on **Options**, and under **Plots**, click next to **Survival**, **Hazard**, and **One minus survival** (this is $F(t) = 1 - S(t)$).

Now click on **OK** to produce the life table and plots.

Notes:

i) The life table starts at t=0 even though our event time is measured from 16. For t=0 to t=16 and the hazard will be zero and the survivor function will equal 1. We can avoid this by creating a new duration variable which starts from zero. Simply create a new variable, called TIME say, where TIME=AGE1ST-16. (Use **Transform → Compute**.) Then construct a life table using TIME as the duration variable, for years 0 to 17.

ii) SPSS uses the alternative definition of the survivor function mentioned in the lecture notes, i.e. $S(t) = \Pr(T > t)$ rather than $S(t) = \Pr(T \geq t)$.

iii) The column headed 'Cumul Propn Surv at End' is the survivor function, $S(t)$. Here, the plot of the survivor function shows the proportion who have not partnered at the start of each year from ages 16 to 33. The plot of 'One minus survival' shows the proportion who have entered a partnership by a particular age.

We can also calculate life tables for subgroups of the sample, and make comparisons. Suppose we wish to compare the survivor functions for men and women.

Return to the Life Table window, and place **FEMALE** next to **Factor:**. Click on **Define Range**, and type **0** next to **Minimum** and **1** next to **Maximum**.

Click on **Options**, and deselect **Life table(s)** to suppress this part of the output, and select only the **Survival** plot.

2. Setting up a Person-Period File for a Discrete-time Analysis

Before we can fit a discrete-time model, we need to convert to person-period format so that, for each individual, we have a record for each year from age 16 until the age they partner or until age 33 for those who have not partnered by the time of interview.

So, for example, we have the following data for the first two individuals:

Individual	AGE1ST	EVENT	AGELEFT	FEMALE	REGION	FCLASS
1	21	1	18	0	2	2
2	31	1	21	0	4	2

These data will need to be restructured to give:

Individual	T	Y	FULLTIME	FEMALE	REGION	FCLASS
1	16	0	1	0	2	2
1	17	0	1	0	2	2
1	18	0	1	0	2	2
1	19	0	0	0	2	2
1	20	0	0	0	2	2
1	21	1	0	0	2	2
2	16	0	1	0	4	2
2	17	0	1	0	4	2
2	18	0	1	0	4	2
2	19	0	1	0	4	2
2	20	0	1	0	4	2
2	21	0	1	0	4	2
2	22	0	0	0	4	2
.	.	.	.	.	.	.
.	.	.	.	.	.	.
2	31	1	0	0	4	2

T is the age at a given year, Y is the binary response variable indicating whether the event (i.e. partnership) has occurred, and FULLTIME is the time-varying covariate indicating whether or not the individual is in full-time education.

A SPSS syntax file has been written to create this data structure. In SPSS, go to the **File** menu, then select **Open**, then **Syntax**. Select the file `ex1.sps`.

To run the syntax, select **Run** then **All**. The person-period file should appear in the Data Editor. It will be saved as `ex1pp.sav`.

3. Fitting a Discrete-Time Event History Model

In a discrete-time model, the dependent variable is the binary indicator Y. This can be analysed using logistic regression, just like any other binary variable.

To fit a logistic regression, go the **Analyze** menu, and select **Regression**, then **Binary Logistic**. Declare **Y** as the **Dependent** variable.

The effect of time (age here) on the hazard of a partnership is modelled by treating T as a covariate. We will begin by allowing the hazard to change with every year – the most flexible model. To do this, declare **T** as a **covariate** and then define as **categorical**. Take the **first** category (age 16) as the **reference category**; click on **Change** after you have selected the first category. A set of 17 dummy variables, corresponding to ages 17 to 33, will be created and added to the model.

You should obtain the following results:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	t			123.842	17	.000	
	t(1)	.840	.407	4.270	1	.039	2.317
	t(2)	1.380	.383	12.992	1	.000	3.977
	t(3)	1.992	.367	29.402	1	.000	7.330
	t(4)	2.068	.369	31.327	1	.000	7.909
	t(5)	2.339	.368	40.477	1	.000	10.373
	t(6)	2.527	.369	46.839	1	.000	12.517
	t(7)	2.374	.380	38.967	1	.000	10.742
	t(8)	2.421	.387	39.139	1	.000	11.252
	t(9)	2.533	.393	41.612	1	.000	12.590
	t(10)	2.462	.407	36.518	1	.000	11.727
	t(11)	2.648	.413	41.146	1	.000	14.120
	t(12)	2.538	.436	33.869	1	.000	12.651
	t(13)	2.337	.470	24.673	1	.000	10.347
	t(14)	2.758	.461	35.818	1	.000	15.760
	t(15)	2.127	.553	14.817	1	.000	8.393
	t(16)	1.514	.689	4.835	1	.028	4.546
	t(17)	.444	1.069	.173	1	.678	1.559
	Constant	-3.999	.336	141.353	1	.000	.018

a. Variable(s) entered on step 1: t.

If you look at the estimated coefficients of the time dummy variables, you can see that the logit-hazard (and therefore the hazard) increases rapidly in the teenage years, is fairly constant for most of the 20s, then declines in the early 30s. Instead of having dummy variables for each year, we could approximate the hazard by including a quadratic function of age. You will first need to create a new variable, called TSQ say, which is equal to T*T (**Transform → Compute** in the Data Editor). Then remove T from the Categorical covariates box. Add TSQ to the Covariates box. Fit this model. You should obtain these results:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	t	1.258	.139	82.311	1	.000	3.517
	tsq	-.025	.003	69.453	1	.000	.975
	Constant	-17.210	1.581	118.560	1	.000	.000

a. Variable(s) entered on step 1: t, tsq.

Now add the covariates FEMALE, FULLTIME, REGION and FCLASS to the model, remembering to treat REGION and FCLASS as categorical.

4. Testing the Proportional Hazards Assumption in a Discrete-Time Model

The model we have fitted assumes that the effects of the covariates are constant over time. This is the proportional hazards assumption. So, for example, we assume that the difference

between men and women in the hazard of partnering is the same for all ages from 16 to 33. We can test this assumption by fitting a model that includes an interaction between age and gender and testing whether the interaction effect is significant.

Add interactions between T and FEMALE, and between TSQ and FEMALE. To add an interaction, highlight both variables in the interaction (use Ctrl-Click to select more than one variable), then click on **>a*b>**. **female*t** and **female*tsq** should appear in the covariates box. The new model will have 2 extra terms, the coefficients of **female*t** and **female*tsq**.

To test whether the effect of gender depends on age, we can compare this model and the model without the interaction term using a **likelihood ratio test**. For each fitted model, the –2*log-likelihood value can be found in the 'Model Summary' table. You should obtain the following values:

Model 1: FEMALE, FULLTIME, REGION, FCLASS
-2 log-likelihood = 2284.706

Model 2: Model 1 + FEMALE*T + FEMALE*TSQ
-2 log-likelihood = 2272.776

The likelihood ratio statistic is the difference between these two values, i.e. 11.930. We compare this with a chi-squared distribution on 2 degrees of freedom, since 2 parameters have been added to Model 1 to get Model 2. If you look at tables of the chi-squared distribution, you will find that this difference is significant at the 5% level. We therefore conclude that the effect of gender does vary with age (or, equivalently, the effect of age varies with gender), i.e. the hazards for males and females are <u>not</u> proportional.