

Interpretation of Hazard or Event History Models

More on Paul Allison’s example (which is taken from his Sage paper, “Event History Analysis”):

Example: 200 Ph.D biochemists followed for 5 years from the time they first became assistant professors. End point: 1st change of job/leave university (source: Allison)

Covariates

Institutions: 1. Prestige of department
 2. Measure of funding (federal funds allocated to the institution for biomedical research)

Individual

Time Varying: 3. Cumulative number of published articles
 4. # citations made by other scientists
 5. Academic rank (1 = assoc. professor, 0 = assist. professor)

	Logit		OLS	
	<i>Coef</i>	<i>t</i>	<i>t</i>	
Constant	2.35			Dep. var.
1. Prestige	.056	.26	.25	$y_{ij} = 1$ if person i left in year j
2. Funding	-.078	-2.47*	-2.36*	= 0 if stayed
3. Publications	-.023	-.79	-.91	$j = \min(5 \text{ yrs.} - \text{ person } i \text{ left 1st job})$
4. Citations	.0069	2.33*	2.23*	
5. Rank	-1.6	-3.12*	-3.26*	
Year				
1	-.96	-2.11*	-2.07*	
2	-.025	-.06	.18	
3	-.74	-1.60	-1.54	L.S. 452.5
4	-.18	-.42	-.38	df 838
-2 log likelihood	-461.90			

Final Model:

$\logit \lambda_i(t) = 4.95 + .045 \text{ Prestige} - .077 \text{ Funding} - .021 \text{ Publ.} + .0072 \text{ Citations} - 1.4 \text{ Rank}$
 - the duration variables were not significant
 - the variables with bold coefficients are significant

-2 log likelihood -452.50

There were two points I wanted to make from this analysis.

1. When OLS was used instead of logit regression, the same variables were significant, despite the fact that we **know** OLS is not appropriate when the dependent variable is binary (has only two categories). This kind of result was found in many studies and was important in the past. OLS is much faster to compute than is logit regression. Therefore, when researchers had to pay for time on mainframe computers and computers were slow, they used OLS to screen for appropriate models and then estimated logit regression only for the final few models. Also, early pc’s were so slow and had so little memory that calculations of the type we’re doing could take hours -- or even a week!
2. The use of the logit regression gives us a way of testing whether the duration variables (years 1-4 vs. year 5) were significant. One (for year 1) was significant as a single category. We can carry out the likelihood ratio test for the difference between the two models above -- the one with the 4 duration dummy variables and the second without

these variables. The difference between the two values of $-2 \log$ likelihood is 9.4.

```
. display chiprob(4, 9.4)
.05184308
```

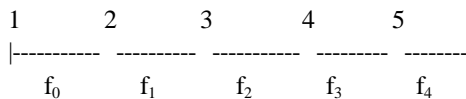
It just barely fails to reach significance, so we do not reject the null hypothesis that there is no difference in fit to the data between the two models -- the first with the duration variables, the second without. We therefore choose the model without the duration variables.

Allison does not give the model in which years 2-4 are dropped and only year 1 is left in as different. I would have liked to have tested that model. In any case, he assumes that there is *no difference in the hazard or odds of leaving in the first 5 years by year in job*.

For the second model, in which *funding*, *citations*, and *rank* are significant, we find that the hazard of leaving is lower the greater the *funding*, and if the person has been promoted -- in those first five years -- to associate professor, and higher the greater the number of *citations* to the persons work.

Use of logit regression for estimating event histories

- requires time to be in discrete units
- if an individual is observed only for a portion of the time interval and is last counted alive, all data for that final partial interval must be discarded
- similarly, if an individual died in time unit j but *could not* have been observed to the end of that time unit, then that person is counted as alive and last seen at time unit $j - 1$



Define a new set of variables for person i :

$$y_{ij} = \begin{cases} 0 & \text{if person } i \text{ survives time unit } j \\ 1 & \text{if person } i \text{ dies in time unit } j \end{cases}$$

Person i can only have 1 $y_{ij} = 1$

If person i is *censored*, no $y_{ij} = 1$

Then each person contributes an observation for each time unit for which the individual is observed and logistic regression can be used for estimation purposes.

- we look at survival times as 1 observation per person, or
- we treat each time unit as a separate observation -- each time unit is treated as a trial having 2 possible outcomes -- survival or end of lifetime
- if time units can be treated independently we can use logit regression. OLS can be used to develop models, but there's no reason to do so unless your computer is very slow or has very little memory

Back to Bangladesh

```
. use lifetabf
. expand year
```

(15524 observations created)

```
. save logitnew
. des
```

Save the new dataset under a new name!

Contains data from logitnew.dta

```
obs:      26,582
vars:      14                               13 Nov 2000 07:51
size:      824,042 (99.9% of memory free)
```

```
-----
1. idchild  str10  %10s      Child ID
2. Own0     byte   %8.0g     Own no goods, no remit
3. Mothed   byte   %8.0g     Mother has some education
4. FS1      byte   %8.0g     1 older female sibling
5. FS2      byte   %8.0g     2+ older female siblings
6. MS1      byte   %8.0g     1 older male sibling
7. MS2      byte   %8.0g     2+ older male siblings
8. Lghouse  byte   %8.0g     High per capita dwelling space
9. Female   byte   %8.0g     Index child female
10. Sibmale byte   %8.0g     Younger sibling male
11. Died    byte   %8.0g     Died during observation period
12. year    float  %9.0g     years of observation
13. area    byte   %8.0g     Born in Matlab
14. BCI     int    %8.0g     Birth-to-Conception Interval
-----
```

Sorted by:

Note: dataset has changed since last saved

```
. sort idchild
. quietly by idchild: gen seq = _n
. quietly by idchild: gen case = _N
. list idchild case seq in 1/20
      idchild      case      seq
1. 1A00000144         1         1
2. 1A00000242         3         1
3. 1A00000242         3         2
4. 1A00000242         3         3
5. 1A00000243         2         1
6. 1A00000243         2         2
7. 1A00001043         2         1
8. 1A00001043         2         2
9. 1A00001142         2         1
10. 1A00001142         2         2
11. 1A00001243         3         1
12. 1A00001243         3         2
13. 1A00001243         3         3
14. 1A00001346         1         1
15. 1A00001645         4         1
16. 1A00001645         4         2
17. 1A00001645         4         3
18. 1A00001645         4         4
19. 1A00001743         3         1
20. 1A00001743         3         2
```

```
. gen ageinyr = seq
. gen y2=age
```

This makes each obs represent a different year
Generate dummies for years 2-4

```
. recode y2 1=0 2=1 3/4=0
(26582 changes made)
```

```
. tab age y2 Check that the new variable is correct
```

ageinyr	y2		Total
	0	1	
1	11058	0	11058
2	0	7951	7951
3	5084	0	5084
4	2489	0	2489
Total	18631	7951	26582

```
. gen y3=age
. recode y3 1/2=0 3=1 4=0
(26582 changes made)
```

```
. gen y4=age
. recode y4 1/3=0 4=1
(26582 changes made)
```

```
. gen diedinyr = 0 A death can ONLY be recorded in the last year of observation
. replace diedinyr = Died if age==case
```

We can then ask whether the odds of dying are related to age, mother’s education, birth in the MCH-FP area, and sex

```
. logit died y2 y3 y4 Mothed area Female
```

```
Iteration 0: log likelihood = -2533.1949
Iteration 1: log likelihood = -2483.3418
Iteration 2: log likelihood = -2480.2116
Iteration 3: log likelihood = -2480.17
Iteration 4: log likelihood = -2480.17
```

```
Logit estimates Number of obs = 26582
LR chi2(6) = 106.05
Prob > chi2 = 0.0000
Log likelihood = -2480.17 Pseudo R2 = 0.0209
```

diedinyr	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
y2	-.1173032	.1003308	-1.169	0.242	-.3139481	.0793416
y3	-.6046619	.1389579	-4.351	0.000	-.8770144	-.3323094
y4	-1.131335	.2387924	-4.738	0.000	-1.599359	-.6633105
Mothed	-.5159085	.1081919	-4.768	0.000	-.7279608	-.3038562
area	-.3890968	.0923291	-4.214	0.000	-.5700585	-.2081352
Female	.3500518	.0899844	3.890	0.000	.1736857	.526418
_cons	-3.604997	.0903979	-39.879	0.000	-3.782173	-3.42782

All are significant. The interpretation of the coefficients here is that the risk of dying decreases with age, is lower for children whose mothers have some education, is lower in the area with the MCH-FP program, and is higher for girls than boys.

We can look at the ODDS:

```
. logit died y2 y3 y4 Mothed area Female, or
```

diedinyr	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
y2	.8893155	.0892258	-1.169	0.242	.730557 1.082574
y3	.5462591	.075907	-4.351	0.000	.4160231 .7172654
y4	.3226023	.077035	-4.738	0.000	.2020259 .5151431
Mothed	.596958	.064586	-4.768	0.000	.4828927 .737967
area	.6776686	.0625685	-4.214	0.000	.5654924 .8120972
Female	1.419141	.1277005	3.890	0.000	1.189682 1.692858

AGE: the odds of dying are highest for 1-year olds. Children aged 2 have odds of dying only 89% as high (but this difference is not significant. Children aged 3 have odds only 55% of those aged 1, and by age 4 the odds of dying are only 32% of the odds at age 1.

MOTHER'S EDUCATION: Children whose mothers had at least 1 year of education had odds of dying only 60% as high as those whose mothers had less than 1 year or no education.

AREA: The area that had access to the ICDDR,B Maternal and Child Health and Family Planning Program (MCH-FP) had only 68% as high mortality as the Comparison area.

FEMALE: Girls had 42% higher odds of dying compared to boys

WERE THESE EFFECTS TIME-DEPENDENT? For example, were girls more at risk at certain ages?

To investigate this question, we have to define age-female interaction terms.

```
. gen fy2 = Fem*y2
. gen fy3 = Fem*y3
. gen fy4 = Fem*y4

. logit died y2 y3 y4 Mothed area Female fy2 fy3 fy4 NONE were individually significant.

. test fy2 fy3 fy4
( 1)  fy2 = 0.0
( 2)  fy3 = 0.0
( 3)  fy4 = 0.0
      chi2( 3) =    5.03
      Prob > chi2 =    0.1697
```

Here the answer is no -- we do not need interactions with age -- girls seem to have the same higher risk of dying at all early childhood ages.

SO the effect of being a girl is NOT time dependent -- nor, obviously, is it time-varying.

I will ask you to test whether the effects of area and Mothed are time dependent.

MORE on INTERPRETATION

We can ask what the probability of dying is for a child, according to his or her characteristics:

```
. predict lambda
. gen surv = 1-lam
```

I then kept one kid of each Mothed, area, Female, age combination -- of which there are $2 \times 2 \times 2 \times 4 = 32$ patterns

```
. list Mothed area Fem ageinyr lambd surv
```

	Mothed	area	Female	ageinyr	lambda	surv
Boys whose mother had no ed and lived in the comparison area						
1.	0	0	0	1	.0264679	.9735321
2.	0	0	0	2	.0236075	.9763925
3.	0	0	0	3	.0146341	.9853659
4.	0	0	0	4	.0086945	.9913055
Girls whose mother has no ed and lived in comparison area						
5.	0	0	1	1	.0371496	.9628504
6.	0	0	1	2	.0331741	.9668258
7.	0	0	1	3	.0206412	.9793587
8.	0	0	1	4	.0122939	.9877061
Boys whose mother had no ed and lived in the MCH-FP area						
9.	0	1	0	1	.0180908	.9819092
10.	0	1	0	2	.0161207	.9838793
11.	0	1	0	3	.0099641	.990036
12.	0	1	0	4	.0059086	.9940915
Girls whose mothers had no ed and lived in the MCH-FP area						
13.	0	1	1	1	.0254802	.9745198
14.	0	1	1	2	.0227241	.977276
15.	0	1	1	3	.0140816	.9859184
16.	0	1	1	4	.0083644	.9916356
Boys whose mothers had some ed and lived in the comparison area						
17.	1	0	0	1	.0159706	.9840294
18.	1	0	0	2	.0142281	.985772
19.	1	0	0	3	.0087878	.9912122
20.	1	0	0	4	.0052085	.9947915
Girls whose mothers had some ed and lived in the comparison area						
21.	1	0	1	1	.0225139	.9774861
22.	1	0	1	2	.0200719	.9799281
23.	1	0	1	3	.0124253	.9875747
24.	1	0	1	4	.0073755	.9926245
Boys whose mothers had some ed and lived in the MCH-FP area						
25.	1	1	0	1	.0108788	.9891212
26.	1	1	0	2	.0096863	.9903136
27.	1	1	0	3	.0059721	.9940279
28.	1	1	0	4	.0035356	.9964644
Girls whose mothers had some ed and lived in the MCH-FP area						
29.	1	1	1	1	.0153685	.9846315
30.	1	1	1	2	.0136907	.9863093
31.	1	1	1	3	.0084541	.9915459
32.	1	1	1	4	.0050101	.9949899

We can then multiply the survival probabilities together ($age1 * age2 * age3 * age4$) for each group and see what the

differences are in survival to age 5 for kids with different characteristics.

```
. list Mothed Female area surviv
```

	Mothed	Female	area	surviv
1.	0	0	0	.9284954
2.	0	0	1	.9508027
3.	0	1	0	.9004852
4.	0	1	1	.93111
5.	1	0	0	.9564962
6.	1	0	1	.9702477
7.	1	1	0	.9389874
8.	1	1	1	.9581167

This gives us the estimated proportions surviving to age 5 of those children who lived to their first birthday.

GENERATING TIME-VARYING COVARIATES

I'd like to ask whether the fact that the mother had already conceived the next child affects the index child's survival. That is, at the second birthday, had mom conceived already. To test this, we need the child's age, in completed years, when the next sibling was conceived. We can get this from the BCI variable -- birth to next conception.

```
. gen gestyr=0
. replace BCI = 120 if BCI==0          no conception is coded as 0 - we need to change that.
(16261 real changes made)

. replace gestyr = int(BCI/12)
(26184 real changes made)

. list BCI gestyr in 1/10
```

	BCI	gestyr
1.	120	10
2.	27	2
3.	27	2
4.	27	2
5.	31	2
6.	31	2
7.	120	10
8.	120	10
9.	120	10
10.	120	10

```
. gen ygest =0
. replace ygest=1 if age==gestyr+1
(2888 real changes made)

. list age gestyr ygest in 1/10
```

	ageinyr	gestyr	ygest
1.	1	10	0
2.	1	2	0
3.	2	2	0
4.	3	2	1

5.	1	2	0
6.	2	2	0
7.	1	10	0
8.	2	10	0
9.	1	10	0
10.	2	10	0

. regress died Moth Female area y2 y3 y4 ygest

Source	SS	df	MS	Number of obs =	26582
Model	2.26603614	7	.323719448	F(7, 26574) =	17.18
Residual	500.833693	26574	.018846756	Prob > F =	0.0000
				R-squared =	0.0045
				Adj R-squared =	0.0042
				Root MSE =	.13728
Total	503.099729	26581	.018927043		

diedinyr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
Mothed	-.008701	.0018087	-4.811	0.000	-.0122462	-.0051559
Female	.0064935	.0016864	3.850	0.000	.003188	.009799
area	-.0068416	.0016897	-4.049	0.000	-.0101535	-.0035296
y2	-.0047985	.0020749	-2.313	0.021	-.0088655	-.0007315
y3	-.0129798	.0023847	-5.443	0.000	-.017654	-.0083057
y4	-.0165516	.0030516	-5.424	0.000	-.0225329	-.0105702
ygest	.013128	.0028141	4.665	0.000	.0076123	.0186437
_cons	.0262215	.0018091	14.494	0.000	.0226755	.0297674

Clearly, children whose mother recently conceived are at higher risk. But does this higher risk differ by gender?

. gen fgest = Fem*ygest

. regress died Moth Female area y2 y3 y4 ygest fgest

Source	SS	df	MS	Number of obs =	26582
Model	2.43527216	8	.304409021	F(8, 26573) =	16.16
Residual	500.664457	26573	.018841096	Prob > F =	0.0000
				R-squared =	0.0048
				Adj R-squared =	0.0045
				Root MSE =	.13726
Total	503.099729	26581	.018927043		

diedinyr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
Mothed	-.0086503	.0018085	-4.783	0.000	-.0121951	-.0051056
Female	.0047261	.0017863	2.646	0.008	.0012248	.0082275
area	-.0068536	.0016895	-4.057	0.000	-.010165	-.0035421
y2	-.0048087	.0020746	-2.318	0.020	-.0088751	-.0007424
y3	-.0130658	.0023845	-5.479	0.000	-.0177396	-.008392
y4	-.0166361	.0030513	-5.452	0.000	-.0226168	-.0106553
ygest	.0051657	.0038697	1.335	0.182	-.0024191	.0127506
fgest	.0162224	.0054128	2.997	0.003	.005613	.0268318
_cons	.027072	.001831	14.786	0.000	.0234832	.0306607

```
. test ygest
```

```
( 1) ygest = 0.0
```

```
      F( 1, 26573) =    1.78  
      Prob > F =    0.1819
```

It appears that only girls are at higher risk when the next sibling is coming along.