

Winter, 2007

Applying Discrete-Time Model:

What Factors Determine the Hazard Rate of
Korean Sociology Doctorates in the Academic Job Market?

Jae-Woo Kim

1. Introduction

I investigated so far what factors have determined how long it took for doctorates to get their jobs after they finished their PhDs in Korean sociology academic market and the chance of their getting their jobs through multiple regression, logistic regression, cumulative logit regression. I also examined whether the survival rates are different for different groups, given that the dependent variable is the year it took for them to experience that event (i.e., getting a job) by taking advantages of the Kaplan Meier method and the Life Table approach. At the very previous assignment, I employed Cox regression models while testing the PH assumption by using three different approaches. In this assignment, I applied discrete-time survival analysis instead because the unit of time is basically discrete in my data (i.e., year), but for the same research question: what factors influenced the hazard rate of the doctorates in the Korean academic job market. In the concluding part, I discussed under which conditions discrete-time models should be preferred to Cox regression.

2. Methods

1) Data

As I discussed in the earlier assignments, I gathered my data from the Korean Research Foundation (www.krf.or.kr). The data include 1) individual attributes and 2) individual ‘network’ attributes (that is, some variables about social capital. Exactly saying, ‘status centrality’ and ‘structural hole’), and the latter was calculated from changing affiliation networks – sociologists’ affiliation with their professional associations – by using UCINET 6, one of the most common program for social network analysis. Those who finished their doctoral dissertations from 1990 to 2004 were counted, and it turned out that the total case is 246. However, I randomly chose the half of the total case for this analysis to make it 123 because it took longer than I expected to transform the original dataset into another appropriate one case by case for discrete-time survival analysis.

Case\ Year	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04
1	1	1	2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2	NA	0	0	1	1	3	3	NA	NA	NA	NA	NA	NA	NA	NA
3	NA	NA	NA	0	0	0	0	1	1	1	2	2	3	4	4
4	NA	NA	NA	NA	3	4	4	4	7	7	NA	NA	NA	NA	NA
5	NA	NA	NA	1	1	5	6	6	NA	NA	NA	NA	NA	NA	NA
6	NA	NA	NA	0	0	0	2	4	6	6	6	9	NA	NA	NA
7	NA	NA	NA	2	5	6	6	NA	NA	NA	NA	NA	NA	NA	NA

In order to discuss how I handled time-varying covariates¹ in my data, I made up one data structure where the variable is the number of articles a doctorate had published since he or

¹ It is increasingly common to find longitudinal data sets with measurements of many variables at regular intervals. For most of kinds of events, such data are essential to get accurate estimates of the effects variables that change over time. (Allison, 1984: 11) For example, if an event occurred at time 10, and 15 individuals were at risk at that time, the values of the explanatory variables at time 10 must be known for all 15 individuals. Typically, that would require that the explanatory variables be measured continuously over time. Especially when time of event occurrence is measured more precisely than the interval at which the explanatory variables are measured, some ad hoc procedure might be useful. The simplest approach is to use the value closest in time to the event time as the estimated value. A better method is to use linear interpolation, which is equivalent to weighted averaging. (Allison, 1984: 37-8).

she received his or her Ph.D. although my data do not have the layout for Cox regression models that are able to deal with those covariates in the sense that I assigned only one value for each person.

In the table above, one case (3) is right-censored, while the rest of them are not (e.g. Case 1 entered the job market at 1990 and got a job at 1993, Case 2 entered the job market at 1991 and got a job at 1997...). The data used in the first and second assignment were built while assuming that the number of articles published by T-1 was assessed if a doctorate succeeded in getting a job at T (e.g. 2 for Case 1; 3 for Case 2... 6 for Case 7. That is, “bold” numbers at each row). I still think that this approach is quite reasonable, regardless of how many times he or she applied before getting his or her job.

However, this assumption would be really problematic for right-censored cases because I do not know when exactly and how many times they applied for their jobs. Let me discuss Case 3 for an example. If there is some effect of the number of published articles on the chance that he or she failed to get a job, is this because of 0(93-96), 1(97-99), 2(00-01), 3(02), or 4(03-04)? In a word, which value should be assigned to that person? I already discussed this problem in the first assignment shortly, but I would like to spend a little more time here.

To tackle this problem, I calculated its value as follows. First, I picked up cohort members as competitors in the data. Keeping in mind the concept of risk set, in this example, Person 3, 5, 6, and 7 can be competitors among 7 persons because they entered the job market at the same year of 1993. And, I calculated the average waiting time of those who succeeded in getting their jobs (i.e., Case 5, 6, and 7). It is 6 because of 5 years for Case 5, 9 years for Case 6, and 4 years for Case 7. I regarded this value as the “expected” waiting time of Case 3, which means that Person 3 should have succeeded within 6 years in order to be on a part with other cohort applicants, which led me to the approximation that the number of published articles for Case 3 is 1 (See First grey above). I applied this approximation to other time-varying covariates such as EIGEN, EFFISIZE, BOOK, and PROJECT in the new data, while 4 at the year of 2004 was assigned as the number of published papers for Case 3 in the old version of my data used in the first and second

assignment (See Second grey). Although it is not shown, it turned out that there are no significant differences between the results from the old data and the new data.

I think this approximation is better than, for example, using the average of scores – one measured at the starting point and the other from at the termination point. In the example above, if I applied the logic of averaging, the value for Case 3 would be 2 – 0 from the starting point (1994) and 4 from the censored point (2004).

Case\ Variable	Old → New (# of published papers)
1	2 → 2
2	3 → 3
3	4 → 1
4	7 → 7
5	6 → 6
6	9 → 9
7	6 → 6

In terms of the data layout, as I mentioned shortly before, discrete-time survival analysis requires me to make another new dataset based on the ‘person-year’ approach. As Allison(1984) described, for each unit of time that each individual is known to be at risk, a separate observational record should be created. There are a total of 1037 person-years in my reduced data to 50% as it shown below. The next step is that the indicator variable (STATUS) for each person-year is coded 1 if a person got a job in that year, otherwise it is coded 0. Also, covariates are assigned the values they took on in each person-year. The final step is to pool the 1037 person years into a single sample. The layout of the new dataset is captured in Figure 1 where ‘a1’ to ‘a11’ indicate the set of dummy variables for discrete-time period to which the record refers. See Table 1 where there are 11 intervals in the original data (N=246) and this is the case with the reduced data (N=123).

<Figure 1> here

2) Variables and Measurement

The dependent variables are the survival time - by the unit of year - it took for doctorates to get their jobs after finishing their PhDs (*PERIOD*). *JOB* indicates the event defined by

getting their jobs (1 if success; 0 if censored). I summarized the basic univariate statistics of them in Table 2 and 3 for the original data (N=246).

<Table 2> and <Table 3> here

I used *SEX* as a stratum for the test of PH assumption (1 for male; 2 for female) because there has been an alleged discrimination against female doctorates in the previous assignment. I also included in the full model the rest of covariates – the same with variables used in multiple regression and logistic regression: 1) status centrality (*EIGEN*, continuous) is defined by the eigenvector of the largest positive eigenvalue calculated as a measure of centrality; 2) effective size of ego's network (*EFFISIZE*, continuous) is measured by the number of alters minus the average degree of alters within the ego network, not counting ties to ego; 3) the birth year can be a good indicator to reflect the principle of seniority pervasive in Korean society (*BIRTH*, interval); 4) *INDEX* is created newly by $(\text{PAPER} + \text{BOOK} + \text{PROJECT})/3$ where *PAPER* is the number of papers an applicant has published in registered journals since he/she got PhD, *BOOK* is the number of books written or translated since getting PhD, and *PROJECT* is the number of projects in which an applicant has participated since getting PhD; 5) the number of interest areas an applicant has the ability to teach or research implies that he/she has a wide understanding of sociology (*COVER*, interval); 6) *STR_DUM*. 'Pulling strings' can be measured by whether an applicant tried to get a job in the same university from which they graduated (1 if the same university, 0 otherwise); 7) *SEX_DUM* is defined as follows. 1 if male; 0 if female; 8) *CHAN_DUM*. Whether an applicant changes his/her major from fields outside social sciences to sociology indicates his/her deep understanding of sociology (1 if not change; 0 otherwise); 9) *REG_DUM*. The area an applicant got his/her undergraduate degree is an indicator of his/her professional status (1 if Seoul, 0 if other areas); 10) The country where a doctorate got his/her PhD. 1 if US given Korea is the reference (*ABR_DUM1*); 1 if other foreign countries given Korea is the reference (*ABR_DUM2*). I provided the basic univariate information about those covariates in Table 4 and 5 for the original data (N=246).

<Table 4> and <Table 5> here

However, another data layout is required for discrete-time survival analysis mentioned before. I presented the summarized statistics of PERIOD and JOB from the pooled data in Table 6 and 7 (1037 Person-years).

<Table 6> and <Table 7> here

3) Statistical Technique

My introduction of discrete-time survival analysis begins with Allison (1984: 14-22) on unrepeatable events of a single kind. Cox regression models assume that continuity of dependent variables. However, exactly saying, the survival time in my study is measured as discrete rather than continuous. When discrete units are very small, it is usually acceptable to treat time as if it were measured on a continuous scale. But, when the time units are large such as months, years or something like that, it is more appropriate to use discrete-time methods. For this purpose, as a first approximation, one could express $P(t)$ as a linear function of explanatory variables: $\ln[P(t)/(1-P(t))] = a + b_1x_1 + b_2x_2(t)$. The coefficients Bs give the change in the logit for each one-unit increase in Xs . This model is still somewhat restrictive because it implies that the only changes that occur in the hazard over time are those which result directly from changes in X_2 , the time-varying explanatory variable. In most cases, there are reasons to suspect that the hazard changes autonomously with time. This is why Allison allows for any variation in the hazard by letting the intercept a be different at each point in discrete time: $\log [P(t)/(1-P(t))] = a(t) + b_1x_1 + b_2x_2(t)$ where $a(t)$ refers to “11” different constants in my study, one for each of fifteen observation years. These constants are estimated by including a set of dummy variables in the specified model. I created a set of “10” dummy variables based on survival time, each of the first eleven years of observation. There is no time-varying covariate such as $X_2(t)$ in my study officially, but I did take into account the change in the value for each variable to some extent as I discussed it before.

3. Results

I presented the result from discrete-time survival analysis in Table 8. When the hypothesis is that all parameters in the model is zero, the global fitness is statistically significant since $\chi^2 = -2(\ln L_{\text{null}} - \ln L_{\text{model}}) = 86.14$. This test leads to reject the null hypothesis in favor of the alternative hypothesis, which means that at least one coefficient significantly affects the outcome variable. According to the output, STR_DUM, SEX_DUM, and ABR_DUM1 are statistically significant at the alpha level of 0.01, but only CHAN_DUM at the level of 0.10. This result does not support any network effect. I could interpret the meaning of coefficients in a different way, that is, focusing the percentage change in the odds ratio by using the formula, $(e^b - 1) * 100$, but suffice to say here about the standard approach, i.e., focusing on the odds ratio, because I showed the three approaches in the previous assignment. First, it turned out that the probability that one who pulls strings (i.e., applying for the same university he or she graduates from) gets a job over the probability that he or she does not is a little bit more than 1/3 (0.371). This might sound like a non-sense, but I discussed the reason in the previous assignment. That ratio for men is about 2.8 times the ratio for women. In the same way, it could be said that the odd ratio for one who did not change his or her major (sociology) is higher than the odd ratio for one who did by about 2.5. Lastly, the ratio for one who received PhD in the US is about 3.3 times bigger than the ratio for one who finished PhD in Korea.

<Table 8> here

4. Conclusion

There are several reasons one could prefer discrete-time survival analysis (Willett and Singer, 2004: 200-1). First, it is intuitively more comprehensible than its continuous-time cousins, facilitating initial mastery and later transition to continuous-time methods if required. Second, it is very appropriate for much of the event history data collected by social scientists because data are often recorded only in terms of intervals for logistical and

financial reasons. Third, it facilitates inclusion of both time-invariant and time-varying predictors, whereas inclusion of the latter is more difficult under the continuous-time approach. Fourth, it fosters inspection of how the pattern of risk shapes up over time. The most popular continuous-time survival analysis strategy (i.e., Cox regression) ignores the shape of the temporal risk profile entirely in favor of estimating the influence of predictors on risk, under a restrictive assumption of proportionality. Fifth, under the discrete-time approach, the proportionality assumption is easily checked and nonproportional models fitted. Finally, in discrete-time survival analysis, all estimation can be conducted using standard statistical software packages that fit logistic regression models.

However, the inclusion of a set of dummy variables has two drawbacks. First, and most obviously, if the number of time points in the data set is large, then the number of temporal dummies included in the model will also be substantial. The temporal dummies can quickly consume many degrees of freedom. Second, if one is interested in ascribing substantive interpretation to temporal dependence, then interpreting the coefficients of many dummy variables can become unwieldy, particularly if the pattern of the coefficients is very noisy (Box-Steffensmeier et. al., 2004: 75).

5. References

- Allison, Paul. 1984. *Event History Analysis: Regression for Longitudinal Event Data*. Sage Publications.
- Box-Steffensmeier, Janet M. and Bradford S. Jones. 2004. *Event History Modeling: A Guide for Social Scientists*. Cambridge University Press.
- Willett, John and Judith Singer. 2004. "Discrete-Time Survival Analysis" in *The Sage Handbook of Quantitative Methodology for the Social Sciences* edited by David Kaplan. Sage Publications.

6. Appendix

<Figure 1> Data Layout for Discrete-time Survival Analysis (Esp. Time Dummy Below)

	eigen	effsize	birth	job	assoc	period	cover	sex	string	cha
1	.138	19.279	54	1	3	0	3	1	0	
2	0	0	55	1	0	1	6	1	0	
3	0	0	55	1	0	1	6	1	0	
4	.142	12.297	61	1	2	1	5	1	1	
5	.142	12.297	61	1	2	1	5	1	1	
6	.134	12.306	58	1	0	1	3	0	0	
7	.134	12.306	58	1	0	1	3	0	0	
8	.063	124.461	68	1	3	1	3	1	0	
9	.063	124.461	68	1	3	1	3	1	0	
10	.136	30.667	59	1	4	1	4	1	0	
11	.136	30.667	59	1	4	1	4	1	0	
12	.102	70.214	64	1	4	1	6	1	1	
13	.102	70.214	64	1	4	1	6	1	1	
14	.075	83.481	66	1	2	1	5	1	0	
15	.075	83.481	66	1	2	1	5	1	0	
16	.074	66.788	62	1	1	2	4	1	1	
17	.074	66.788	62	1	1	2	4	1	1	
18	.074	66.788	62	1	1	2	4	1	1	
19	.064	114.199	67	1	4	2	4	1	0	
20	.064	114.199	67	1	4	2	4	1	0	
21	.064	114.199	67	1	4	2	4	1	0	
22	.099	35.612	57	1	2	2	5	1	1	
23	.099	35.612	57	1	2	2	5	1	1	
24	.099	35.612	57	1	2	2	5	1	1	
25	.074	111.627	63	1	3	2	3	1	0	
26	.074	111.627	63	1	3	2	3	1	0	

	exptime	status	a1	a2	a3	a4	a5	a6	a7	a
1	0	0	1	0	0	0	0	0	0	
2	1	0	1	0	0	0	0	0	0	
3	1	1	0	1	0	0	0	0	0	
4	1	0	1	0	0	0	0	0	0	
5	1	1	0	1	0	0	0	0	0	
6	1	0	1	0	0	0	0	0	0	
7	1	1	0	1	0	0	0	0	0	
8	1	0	1	0	0	0	0	0	0	
9	1	1	0	1	0	0	0	0	0	
10	1	0	1	0	0	0	0	0	0	
11	1	1	0	1	0	0	0	0	0	
12	1	0	1	0	0	0	0	0	0	
13	1	1	0	1	0	0	0	0	0	
14	1	0	1	0	0	0	0	0	0	
15	1	1	0	1	0	0	0	0	0	
16	2	0	1	0	0	0	0	0	0	
17	2	0	0	1	0	0	0	0	0	
18	2	1	0	0	1	0	0	0	0	
19	2	0	1	0	0	0	0	0	0	
20	2	0	0	1	0	0	0	0	0	
21	2	1	0	0	1	0	0	0	0	
22	2	0	1	0	0	0	0	0	0	
23	2	0	0	1	0	0	0	0	0	
24	2	1	0	0	1	0	0	0	0	
25	2	0	1	0	0	0	0	0	0	
26	2	0	0	1	0	0	0	0	0	

<Table 1> Frequency Table of Waiting Time for Those Who Got Their Jobs (N=246)

Waiting	Frequency
0	1
1	16
2	18
3	24
4	15
5	22
6	5
7	6
8	8
9	3
10	5
Total	123
Missing	123

* Missing data indicate those who did not get their jobs.

<Table 2> PERIOD in the old data (N=123)

N	Mean	Standard Deviation	Min	Max
59	4.254	2.339	0	10

<Table 3> JOB in the old data (N=123)

JOB	Frequency	Percent
Censored	64	52
Event (Getting a job)	59	48
Total	123	100

<Table 4> Descriptive Statistics for Interval/Ratio Covariates in the old data (N=123)

Variable	Mean	Standard Deviation	Min	Max
EIGEN	.068935	.046604	0	.072
EFFISIZE	49.87712	42.63153	0	142.773
BIRTH	61.02439	4.599471	44	69
INDEX	4.75626	3.831687	0	25.33
COVER	4.056911	1.081141	1	6

<Table 5> Frequency for Categorical Variables (N=123)

Variable	Value	Frequency	Percent
STR_DUM (N=243)	Yes	54	43.9
	No	68	56.1
SEX_DUM (N=246)	Male	86	69.9
	Female	37	30.1
CHA_DUM (N=246)	Not changed	106	86.2
	Changed	17	13.8
REG_DUM (N=244)	Seoul	107	87.0
	Otherwise	15	13.0
ABR_DUM (N=246)	US	42	34.1
	Other	22	17.9
	Korea	59	48.0

<Table 6> PERIOD in the pooled data (1037 person-years)

N	Mean	Standard Deviation	Min	Max
333	5.099	2.453	0	10

<Table 7> JOB in the pooled data (1037 person-years)

JOB	Frequency	Percent
Censored	704	67.9
Event (Getting a job)	333	32.1
Total	123	100

<Table 8> Results from Discrete-Time Survival Analysis

Covariate	b	Std.Error	OR	95% CI	
EFFISIZE	.001346	.0037893	1.001347	.9939478,	1.008802
BIRTH	-.0467753	.0313277	.9543018	.8948343	1.017721
INDEX	.0046522	.0362285	1.004663	.9361076	1.078239
COVER	.2044197	.1816343	1.226813	.9178138	1.639842
STR_DUM Others				Reference	
Same school	-.9908828***	.1206887	.3712488	.1963132	.7020706
SEX_DUM Women				Reference	
Men	1.0225383***	1.059833	2.780243	1.317053	5.868973
CHA_DUM Change				Reference	
Not change	.8912314*	1.242494	2.43813	.8980008	6.619681
REG_DUM Others				Reference	
Seoul	.9348255	1.60981	2.546769	.7378125	8.790898
ABR_DUM Korea				Reference	
US	1.1913780***	1.145248	3.291614	1.664375	6.509785
Others	-.5680948	.5666039	1.286876	.5429475	3.050112
Year Dummy					
A2	-.0769453	.7774981	.9259405	.1785839	4.800915
A3	.3009442	1.108706	1.351134	.2705358	6.747958
A4	1.1602657	2.521233	3.190781	.6781254	15.01357
A5	.2711464	1.125421	1.311457	.243949	7.050322
A6	1.5498272**	3.751665	4.710656	.9889325	22.43862
A7	-.4547724	6595675	.6345924	.0827548	4.866272
A8	.1321574	1.085097	1.141288	.1770527	7.356781
A9	.2246416	1.19187	1.251874	.1937125	8.090285
A10	-.1120295	.9272073	.8940179	.1170967	6.825707
LRS	86.14				
Sample Size	1,037				
Df					