Survival Analysis

Hsueh-Sheng Wu

CFDR Workshop Series

Spring 2012

Outline

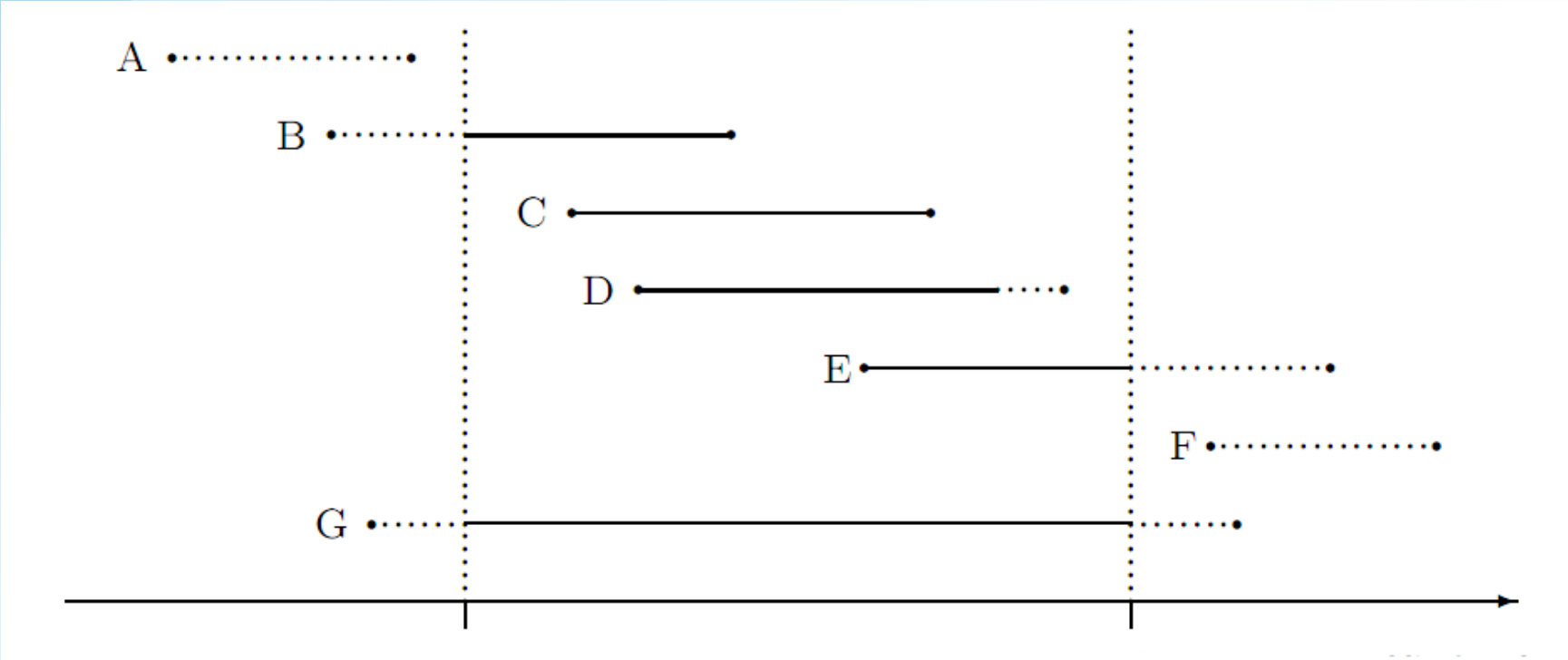
- What is survival analysis
- Survival analysis steps
- Create data for survival analysis
  - Data for different analyses
  - The dependent variable in Life Table analysis and Cox Regression
  - Reshape data for Discrete-time analysis
- Analyze data
    - Life Table
    - Cox Regression without time-varying variables
    - Discrete-time without time-varying variables
    - Discrete-time with time-varying variables
- conclusion

What is survival analysis

- Survival analysis is a “time to event” analysis, that is, we follow subjects over time and observe at which point in time they experience the event of interest
- Survival analysis establishes the causal relation between independent variables and the dependent variable
- Survival analysis can use incomplete information from respondents

What is survival analysis(continued)

Figure 1. Different types of censoring



Start of the study

End of the study

What is survival analysis(continued)

- A is fully censored on the left
- B is partially censored on the left
- C is complete
- D is censored on the right within the study period
- E is censored on the right
- F is completely censored on the right
- G represents a duration that is left and right censored

STEPS for Survival Analysis

- What is the research question
- Locate and select variables
- Establish analytic sample
- Recode variables
- Create timing data for survival analysis
  - Life Tables and Cox Regression
  - Discrete-time analysis
- Analyze data
  - Life Table
  - Cox regression
  - Discrete-time analysis

An example of conducting survival analysis

- Research Question:
  What factors are associated with the timing of first marriage ?

- Variables:
  - Dependent variable: Timing of first marriage

- Predictors:
  - Gender (male/female),
  - Race (black/non-black)
  - Age (continuous)
  - Expectation of marriage at Wave I (continuous)
  - High school graduation (yes/no)

- Weight variables:
  - Region: (West, Midwest, South, and Northeast)
  - Schools (Range 1 to 371)
  - Individual weights (Range 16.3183 to 6649.3618)

- An indicator of whether adolescents are included in the analytic sample
  - sub_pop (yes/no)

Analytic Sample

- The Sample Size:
  - 20, 745 adolescents participated in Wave 1 interview
  - 15, 170 adolescents provided information on marriages at Wave III interview
  - 14,253 adolescents has valid information on the timing of first marriage and weight variables at Wave I
  - 2,855 have married for the first time before Wave III interview

- Respondents who had first marriage before Wave III interview but were excluded from the analytic sample
  - 54 married before Wave I interview
  - 2 married before Age 14
  - 34 had first marriage, but did not have graduation time

- The analytic sample
  - Adolescents with valid responses to marital status, all the predictor variables, and weight variables. The final N = 13, 995.

Create data for survival analysis

- Data for different analysis

Table 1. Data for analyses not involving timing of first marriage

Name	Married	Female	High School Graduation
Tim	0	0	1
Sara	1	1	0
Tom	0	0	0
Sherry	1	1	1

Note:

Married: 1 = Married; 0 = Unmarried

Female: 1 = Female; 0 = Male

High School Graduation: 1 = Graduated from High School; 0 = Did not graduate from High School

Table 2. Data for Life Table and Cox Regression

Name	Married	Time (in months from W1) to getting married or being censored (reaching the W3 having never married)	Female	High School Graduation	Time (in months from W1 interview) to graduating from high school or being censored (i.e., reaching the W3 having not
Tim	0	3	0	1	3
Sara	1	3	1	0	3
Tom	0	5	0	0	5
Sherry	1	5	1	1	4

Note:

Married: 1 = Married; 0 = Unmarried

Female: 1 = Female; 0 = Male

High School Graduation: 1 = Graduated from High School; 0 = Did not graduate from High School

Table 3. Data for Discrete Time Analysis

Name	Month	Married	Female	High School Graduation
Tim	1	0	0	0
	2	0	0	0
	3	0	0	1
Sara	1	0	1	0
	2	0	1	0
	3	1	1	0
Tom	1	0	0	0
	2	0	0	0
	3	0	0	0
	4	0	0	0
	5	0	0	0
Sherry	1	0	1	0
	2	0	1	0
	3	0	1	0
	4	0	1	1
	5	1	1	1

Note:

Married: 1 = Married; 0 = Unmarried

Female: 1 = Female; 0 = Male

High School Graduation: 1 = Graduated from High School; 0 = Did not graduate from High School

Dependent Variable in Life Table and Cox Regression

- **Create the date indicator for:**

  - **Timing of first marriage**

```
gen marriage_t1 = ym(form_y1, form_m1)
label variable  marriage_t1 "century month
for getting married for the first time“
```

  - **Wave I interview**

```
gen interview_t1 = ym(iyear, imonth)
label variable interview_t1 "time for t1 interview"
```

  - **Wave III interview**

```
gen interview_t3 = ym(iyear3, imonth3)
label variable interview_t3 "time for t3 interview“
```

- **Calculate the number of months to first marriage since Wave I interview**

```
gen time1 = marriage_t1 - interview_t1  if (marriage_t1 ~=. & interview_t1~=.)
label variable time1 "time for those got married“
```

- **Calculate the number of months between Wave I and Wave III interview**

```
gen time2 = interview_t3-interview_t1
label variable time2 "time for those did not get married“
```

- **Calculate the number of months to first marriage or censoring**

```
gen time =.
label variable time "timing of the first marriage“
replace time = time1 if time1 ~=. & mar1 ==1
replace time = time2 if mar1 ==0
replace time =. if time1 <0
```

Reshape data for Discrete Time Analysis

- **Use the data created for Cox Regression**

```
use "t:\temp\cox.dta", clear
```

Table 4. Data for Cox regression

Name	mar1	time	female	gra	gra_tm
Tim	0	3	0	1	3
Sara	1	3	1	0	3
Tom	0	5	0	0	5
Sherry	1	5	1	1	4

Noted:

mar1: 1 = married for the first time, 0 = did not marry for the first time

time: the number of months to the first marriage since Wave I interview or having never married

Female: 0 = Male, 1 = Female

gra: 1 = Graduated from High School, 0 = Did not

gra_tm: the number of months to high school graduation or having never graduated.

- **Expand each observation into multiple observations, depending on the number of months that each original observation needs to get married for the first time or become censored.**

```
expand time
```

Table 5. Data after using Stata "expand" command

Name	mar1	time	female	gra	gra_tm
Tim	0	3	0	1	3
Tim	0	3	0	1	3
Tim	0	3	0	1	3
Sara	1	3	1	0	3
Sara	1	3	1	0	3
Sara	1	3	1	0	3
Tom	0	5	0	0	5
Tom	0	5	0	0	5
Tom	0	5	0	0	5
Tom	0	5	0	0	5
Tom	0	5	0	0	5
Sherry	1	5	1	1	4
Sherry	1	5	1	1	4
Sherry	1	5	1	1	4
Sherry	1	5	1	1	4
Sherry	1	5	1	1	4

Noted:
mar1: 1 = married for the first time, 0 = did not
time: the number of months to the first marriage since Wave I interview or having never married
Female: 0 = Male, 1 = Female
gra: 1 = Graduated from High School, 0 = Did not
gra_tm: the number of months to high school graduation or having never graduated.

- **Sort the data by the ID variable. Generate a variable "month" to indicate which month to which the observation now belongs.**

```
sort aid
by aid: gen month=_n
```

Table 6. Data after the "month" variable was generated

Name	mar1	time	female	gra	gra_tm	month
Tim	0	3	0	1	3	1
Tim	0	3	0	1	3	2
Tim	0	3	0	1	3	3
Sara	1	3	1	0	3	1
Sara	1	3	1	0	3	2
Sara	1	3	1	0	3	3
Tom	0	5	0	0	5	1
Tom	0	5	0	0	5	2
Tom	0	5	0	0	5	3
Tom	0	5	0	0	5	4
Tom	0	5	0	0	5	5
Sherry	1	5	1	1	4	1
Sherry	1	5	1	1	4	2
Sherry	1	5	1	1	4	3
Sherry	1	5	1	1	4	4
Sherry	1	5	1	1	4	5

Noted:

mar1: 1 = married for the first time, 0 = did not marry for the first time

time: the number of months to the first marriage since Wave I interview or having never married

Female: 0 = Male, 1 = Female

gra: 1 = Graduated from High School, 0 = Did not graduate from High School

gra_tm: the number of months to high school graduation or having never graduated.

- **Create a variable, married, to indicate the transition to first marriage.**

```
gen married=0
replace married=mar1 if month==time
```

Table 7. Data after the "married" variable was generated

Name	mar1	time	female	gra	gra_tm	month	married
Tim	0	3	0	1	3	1	0
Tim	0	3	0	1	3	2	0
Tim	0	3	0	1	3	3	0
Sara	1	3	1	0	3	1	0
Sara	1	3	1	0	3	2	0
Sara	1	3	1	0	3	3	1
Tom	0	5	0	0	5	1	0
Tom	0	5	0	0	5	2	0
Tom	0	5	0	0	5	3	0
Tom	0	5	0	0	5	4	0
Tom	0	5	0	0	5	5	0
Sherry	1	5	1	1	4	1	0
Sherry	1	5	1	1	4	2	0
Sherry	1	5	1	1	4	3	0
Sherry	1	5	1	1	4	4	0
Sherry	1	5	1	1	4	5	1

Noted:

mar1: 1 = married for the first time, 0 = did not marry for the first time
time: the number of months to the first marriage since Wave I interview or having never married

Female: 0 = Male, 1 = Female
gra: 1 = Graduated from High School, 0 = Did not graduate from High School

gra_tm: the number of months to high school graduation or having never graduated.

- **Check the accuracy of timing of first marriage**

```
tab1 married if sub_pop ==1
```

- **Create a variable, graduated, to indicate the timing of high school graduation.**

```
gen graduated=0
replace graduated = gra if month >= gra_tm
```

Table 8. Data after the "graduated" variable was generated

Name	mar1	time	female	gra	gra_tm	month	married	graduated
Tim	0	3	0	1	3	1	0	0
Tim	0	3	0	1	3	2	0	0
Tim	0	3	0	1	3	3	0	1
Sara	1	3	1	0	3	1	0	0
Sara	1	3	1	0	3	2	0	0
Sara	1	3	1	0	3	3	1	0
Tom	0	5	0	0	5	1	0	0
Tom	0	5	0	0	5	2	0	0
Tom	0	5	0	0	5	3	0	0
Tom	0	5	0	0	5	4	0	0
Tom	0	5	0	0	5	5	0	0
Sherry	1	5	1	1	4	1	0	0
Sherry	1	5	1	1	4	2	0	0
Sherry	1	5	1	1	4	3	0	0
Sherry	1	5	1	1	4	4	0	1
Sherry	1	5	1	1	4	5	1	1

Noted:
mar1: 1 = married for the first time, 0 = did not marry for the first time
time: the number of months to the first marriage since Wave I interview or having never married
Female: 0 = Male, 1 = Female
gra: 1 = Graduated from High School, 0 = Did not graduate from High School
gra_tm: the number of months to high school graduation or having never graduated.

- ***Check the accuracy of high school graduation***

```
sort aid
by aid: gen N=_N
tab1 gra if supop==1 & N==month
```

Table 9. Data for checking the accuracy of high school graduation

Name	mar1	time	female	gra	gra_tm	month	married	graduated	N
Tim	0	3	0	1	3	1	0	0	3
Tim	0	3	0	1	3	2	0	0	3
Tim	0	3	0	1	3	3	0	1	3
Sara	1	3	1	0	3	1	0	0	3
Sara	1	3	1	0	3	2	0	0	3
Sara	1	3	1	0	3	3	1	0	3
Tom	0	5	0	0	5	1	0	0	5
Tom	0	5	0	0	5	2	0	0	5
Tom	0	5	0	0	5	3	0	0	5
Tom	0	5	0	0	5	4	0	0	5
Tom	0	5	0	0	5	5	0	0	5
Sherry	1	5	1	1	4	1	0	0	5
Sherry	1	5	1	1	4	2	0	0	5
Sherry	1	5	1	1	4	3	0	0	5
Sherry	1	5	1	1	4	4	0	1	5
Sherry	1	5	1	1	4	5	1	1	5

Noted:
mar1: 1 = married for the first time, 0 = did not marry for the first time
time: the number of months to the first marriage since Wave I interview or having never married
Female: 0 = Male, 1 = Female
gra: 1 = Graduated from High School, 0 = Did not
gra_tm: the number of months to high school graduation or having never graduated.

Save the data set for future use

```
save "t:\temp\discrete.dta", replace
```

Analyze data

A. Life table

Stata commands:

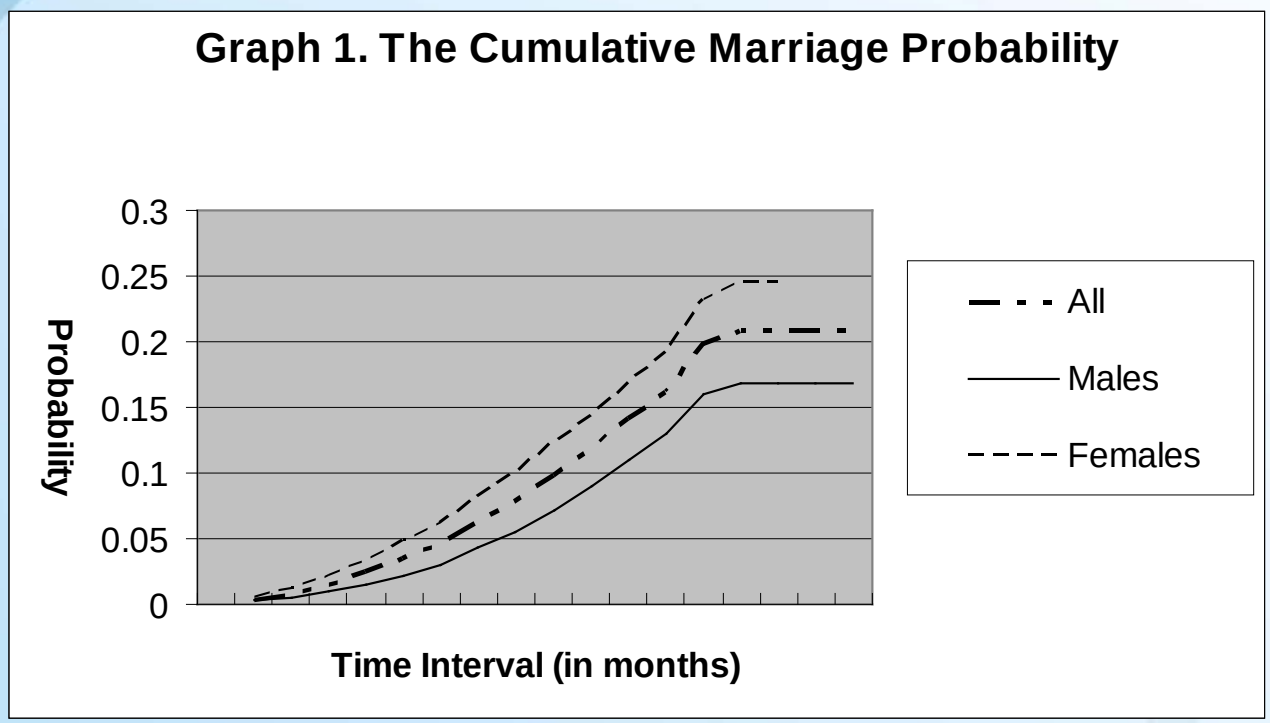
```
ltable time mar1 if sub_pop ==1, hazard
ltable time mar1 if sub_pop ==1
```

Results:

Table 1: Life Table for the Whole Sample

Interval (in months)			# of Single Adolescents	# of Adolescents Married	Lost to Follow-Up	Hazards	Cumulative Marriage Probability
0	→	6	13995	54	0	0.0039	0.0039
6	→	12	13941	68	0	0.0049	0.0087
12	→	18	13873	95	0	0.0069	0.0155
18	→	24	13778	128	0	0.0093	0.0247
24	→	30	13650	155	0	0.0114	0.0357
30	→	36	13495	153	0	0.0114	0.0467
36	→	42	13342	232	0	0.0175	0.0632
42	→	48	13110	220	0	0.0169	0.079
48	→	54	12890	274	0	0.0215	0.0985
54	→	60	12616	273	0	0.0219	0.118
60	→	66	12343	323	0	0.0265	0.1411
66	→	72	12020	290	400	0.0248	0.1622
72	→	78	11330	327	7288	0.0435	0.1978
78	→	84	3715	25	3682	0.0134	0.2085
84	→	90	8	0	6	0	0.2085
90	→	96	2	0	1	0	0.2085
96	→	102	1	0	1	0	0.2085

Life Table Graph



B. Cox regression without Time varying variables

•Stata commands

```
use "T:\temp\cox.dta", clear
svyset psuscid1 [pweight = gswgt1], strata(region1)
stset time, f(mar1)
svy, subpop(sub_pop): stcox female black age_t1 expect
```

• Results:

Survey: Cox regression

```
Number of strata   =         4                  Number of obs      =     14253
Number of PSUs     =       132                  Population size    =  16629862
                                                Subpop. no. of obs =     13995
                                                Subpop. size       =  16297823
                                                Design df          =       128
                                                F(   4,    125)    =    101.86
                                                Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   1.740813    .097873     9.86   0.000     1.557538    1.945654
       black |   .5463479   .0565109    -5.84   0.000     .4452316    .6704288
      age_t1 |   1.030068   .0019299    15.81   0.000     1.026256    1.033894
      expect |   1.266699   .0343744     8.71   0.000     1.200477    1.336573
------------------------------------------------------------------------------
```

C. Discrete-time Analysis without Time-varying Variables

- Stata commands:

```
use "T:\temp\discrete.dta", clear
svyset psuscid1 [pweight = gswgt1], strata(region1)
char month [omit] 77
xi: svy, subpop(sub_pop): logistic married i.month female black age_t1 expect
```

- Results:

```
Survey: Logistic regression
Number of strata   =         4                  Number of obs      =     1033582
Number of PSUs     =       132                  Population size    =  1209145097
                                                Subpop. no. of obs =     1010143
                                                Subpop. size       =  1178862615
                                                Design df          =         128
                                                F(  85,     44)    =       21.35
                                                Prob > F           =      0.0000
------------------------------------------------------------------------------
             |             Linearized
     married | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Imonth_1 |   .0855008   .0477686    -4.40   0.000     .0283055    .2582668
   _Imonth_2 |   .0622853   .0339932    -5.09   0.000     .0211541    .1833904
         .
         .
         .
  _Imonth_75 |    1.04427   .3475159     0.13   0.897     .5405591    2.017355
  _Imonth_76 |   1.187808   .3981339     0.51   0.609     .6119474     2.30557
  _Imonth_78 |   .3509662   .1625097    -2.26   0.025     .1404001    .8773308
  _Imonth_79 |   .1736188   .1291074    -2.35   0.020     .0398639    .7561599
  _Imonth_80 |   .6049959   .3388633    -0.90   0.371     .1997271    1.832601
  _Imonth_81 |   .3521969   .2508042    -1.47   0.145     .0860692    1.441196
  _Imonth_82 |   .1178069   .1170397    -2.15   0.033     .0164983    .8412027
      female |   1.745988   .0986846     9.86   0.000     1.561246     1.95259
       black |   .5448028   .0566048    -5.85   0.000     .4435634    .6691493
      age_t1 |   1.030225   .0019416    15.80   0.000     1.026391    1.034075
      expect |   1.268406    .03462     8.71   0.000     1.201722    1.338792
------------------------------------------------------------------------------
```

D. Discrete-time Analysis with a Time-varying Variable

- Stata commands:

```
use T:\temp\discrete, clear
svyset psuscid1 [pweight = gswgt1], strata(region1)
char month [omit] 77
 xi: svy, subpop(sub_pop): logistic married i.month female black age_t1 expect
graduated
```

- Results:

```
Survey: Logistic regression
Number of strata   =         4                  Number of obs      =     1033582
Number of PSUs     =       132                  Population size    =  1209145097
                                                Subpop. no. of obs =     1010143
                                                Subpop. size       =  1178862615
                                                Design df          =         128
                                                F(  86,     43)    =       21.55
                                                Prob > F           =      0.0000
------------------------------------------------------------------------------
             |             Linearized
     married | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Imonth_1 |   .0985339   .0562077    -4.06   0.000     .0318707    .3046348
   _Imonth_2 |   .0711091   .0398916    -4.71   0.000     .0234342    .2157742
         .
         .
         .
  _Imonth_75 |   1.043885   .3469749     0.13   0.897     .5407833    2.015034
  _Imonth_76 |   1.187321   .3974025     0.51   0.609     .6122765    2.302444
  _Imonth_78 |   .3518995   .1629764    -2.26   0.026      .140746    .8798348
  _Imonth_79 |   .1739343   .1292685    -2.35   0.020     .0399697    .7569009
  _Imonth_80 |   .6069465   .3397445    -0.89   0.374     .2005091    1.837244
  _Imonth_81 |   .3532947   .2515898    -1.46   0.146     .0863356    1.445719
  _Imonth_82 |   .1178734   .1171192    -2.15   0.033      .016504    .8418673
      female |   1.731455   .0973056     9.77   0.000     1.549238    1.935104
       black |   .5521323   .0567529    -5.78   0.000     .4505203    .6766624
      age_t1 |   1.028714   .0019135    15.22   0.000     1.024935    1.032508
      expect |   1.266885   .0345654     8.67   0.000     1.200305    1.337159
  graduated  |   1.232447   .1226013     2.10   0.038     1.012242    1.500556
------------------------------------------------------------------------------
```

Conclusion

- Survival analysis examines the timing of an event and allows researchers to test factors that may lead to the occurrence of the event.
- For life Table and Cox Regression, there is a need to construct the variables indicating when the event and its predicators occurred. For discrete-time analysis, the data need to be transformed into person-period format.
- Discrete-time analysis is more flexible than Cox Regression.
  - The dummy variables for time can delineate the magnitude of hazards at each time point.
  - Time-varying variables can be easily included in the models
  - It can be extended to examine Multinomial logistic regression.
- For more information on survival analysis
  - Dr. Alfred Demaris has written a book, “Regression With Social Data: Modeling Continuous and Limited Response Variables”. This book provides detailed information about assumptions and estimations of several survival models.
  - Dr. Judith Singer and Dr. John Willett have published a book, called “Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence”. Data sets, computer programs, outputs and PowerPoint slides for the examples used in this book can be found at http://gseacademic.harvard.edu/alda/
  - University of California at Los Angeles has helpful information on using SAS, Stata, and SPSS for conducting survival analysis at http://www.ats.ucla.edu/stat/seminars/.
  - Dr. David Garson has provided excellent documents on Life Table, Cox Regression, and Event History at http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm.