Stata Summer Series

**Stata 302 –Time Series in Stata**

Many of the projects at Urban involve repeated measurements of the same individuals (or sites or whatever) over time. With such data, we often model problems where we are interested in both the timing and the types of outcomes.

*What you will get out of this session:*

* + Examples of time series data and research projects at Urban that use time series data
  + Commands to set up the data to create suitable outcome variables for time series models
  + Commands to run the models for time series data (event history models, hazard models)
  + Learn about things that can go horribly wrong, and how to use STATA to notice and maybe fix problems

*Helpful resources for survival analysis*

Stata manual:

* + Time series: https://www.stata.com/manuals13/ts.pdf
  + Survival analysis: <https://www.stata.com/manuals13/st.pdf>
  + Hazard model: https://www.stata.com/manuals13/ststcox.pdf

UCLA IDRE:

* + <https://stats.idre.ucla.edu/stata/seminars/stata-survival/>

Other resources:

* + <http://hummedia.manchester.ac.uk/institutes/methods-manchester/docs/eha.pdf>

**STATA Training Session for 08/10/2018**

**A quick introduction to setting up and running an event history model**

**Here are some steps in turning an ordinary data set into an event history data set.**

**Fictitious data set: 10 cases in an RCT of an abstinence training program at age 14**

**Set up for a logit model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| case id | abst\_prog | age at 1st | age at int | logit outcome |
| 1 | 1 | 20 | 22 | 0 |
| 2 | 1 | 17 | 20 | 1 |
| 3 | 1 | 21 | 21 | 0 |
| 4 | 1 | . | 21 | 0 |
| 5 | 1 | . | 19 | . |
| 6 | 0 | . | 22 | 0 |
| 7 | 0 | 19 | 21 | 1 |
| 8 | 0 | 16 | 19 | . |
| 9 | 0 | 19 | 20 | 1 |
| 10 | 0 | . | 19 | . |

**Fictitious data set: 10 cases in an RCT of an abstinence training program at age 14**

**Set up for an event history model**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| case id | Abst  prog | age at 1st | age at int | Births by age 15-19 | duration | teen birth |
| 1 | 1 | 20 | 22 | 0 0 0 0 0 | 5 | 0 |
| 2 | 1 | 17 | 20 | 0 0 1 | 3 | 1 |
| 3 | 1 | 21 | 21 | 0 0 0 0 0 | 5 | 0 |
| 4 | 1 | . | 21 | 0 0 0 0 0 | 5 | 0 |
| 5 | 1 | . | 19 | 0 0 0 0 | 4 | 0 |
| 6 | 0 | . | 22 | 0 0 0 0 0 | 5 | 0 |
| 7 | 0 | 19 | 21 | 0 0 0 0 1 | 5 | 0 |
| 8 | 0 | 16 | 19 | 0 1 | 2 | 1 |
| 9 | 0 | 19 | 20 | 0 0 0 0 1 | 5 | 1 |
| 10 | 0 | . | 19 | 0 0 0 0 | 4 | 0 |

**Here are some important highlights from a log file for event history models**

------------------------------------------------------------------------------------- log: D:\Martin\_UI\STATA users group\stataclass08102018.log

log type: text

opened on: 9 Aug 2018, 19:40:58

. do "D:\Martin\_UI\STATA users group\stataclass08102018.do"

. infix hspnon 9-10 race 20-21 sex 31-32 numkids 41-43 ageint 52-54 /\*

> \*/ age1st 63-65 age2nd 74-76 using "D:\Martin\_UI\STATA users group\cps695dat.txt"

(2,952 observations read)

.

. \* there is no case id, so create one

. egen id = fill(1 2)

.

. \* create a dummy variable for whether a second birth occurred

. generate birth2 = 9

. replace birth2 = 1 if age2nd>100 & age2nd<600

(1,767 real changes made)

. replace birth2 = 0 if age2nd==.

(1,185 real changes made)

. tab birth2, missing

birth2 | Freq. Percent Cum.

------------+-----------------------------------

0 | 1,185 40.14 40.14

1 | 1,767 59.86 100.00

------------+-----------------------------------

Total | 2,952 100.00

.

. \* create a variable for interval from first birth to second or interview

.

. generate dur = 999

. replace dur = age2nd - age1st if birth2==1

(1,767 real changes made)

. replace dur = ageint - age1st if birth2==0

(1,185 real changes made)

.

. tab dur, missing

dur | Freq. Percent Cum.

------------+-----------------------------------

0 | 34 1.15 1.15

1 | 25 0.85 2.00

2 | 15 0.51 2.51

3 | 15 0.51 3.01

4 | 14 0.47 3.49

5 | 19 0.64 4.13

6 | 16 0.54 4.67

7 | 14 0.47 5.15

8 | 22 0.75 5.89

9 | 14 0.47 6.37

10 | 21 0.71 7.08

11 | 43 1.46 8.54

12 | 44 1.49 10.03

.

.

.

188 | 1 0.03 99.86

189 | 1 0.03 99.90

193 | 1 0.03 99.93

195 | 1 0.03 99.97

201 | 1 0.03 100.00

------------+-----------------------------------

Total | 2,952 100.00

.

. \* create categories for age at first birth le19,ge20

. generate age1teen = -9

. replace age1teen = 1 if age1st < 240 & age1st >= 0

(1,019 real changes made)

. replace age1teen = 0 if age1st >= 240 & age1st <= 888

(1,933 real changes made)

.

. \* first, tell STATA which is the duration, which is the event, and which is the id

. stset dur, fail(birth2) id(id)

id: id

failure event: birth2 != 0 & birth2 < .

obs. time interval: (dur[\_n-1], dur]

exit on or before: failure

------------------------------------------------------------------------------

2,952 total observations

34 observations end on or before enter()

------------------------------------------------------------------------------

2,918 observations remaining, representing

2,918 subjects

1,741 failures in single-failure-per-subject data

124,953 total analysis time at risk and under observation

at risk from t = 0

earliest observed entry t = 0

last observed exit t = 201

.

. \* life table of second birth intervals

. ltable dur birth2 if(dur>0), interval(0,9,21,33,69,129)

Beg. Std.

Interval Total Deaths Lost Survival Error [95% Conf. Int.]

-------------------------------------------------------------------------------

0 9 2918 13 127 0.9954 0.0013 0.9922 0.9974

9 21 2778 408 188 0.8441 0.0070 0.8299 0.8573

21 33 2182 561 164 0.6186 0.0096 0.5995 0.6372

33 69 1457 620 323 0.3226 0.0099 0.3032 0.3421

69 129 514 129 304 0.2076 0.0103 0.1877 0.2282

129 . 81 10 71 0.1620 0.0151 0.1337 0.1927

-------------------------------------------------------------------------------

.

. \* separate estimates by age at first birth

. ltable dur birth2 if(dur>0), by(age1teen) interval(0,9,21,33,69,129)

Beg. Std.

Interval Total Deaths Lost Survival Error [95% Conf. Int.]

-------------------------------------------------------------------------------

age1teen = 0

0 9 1901 9 127 0.9951 0.0016 0.9906 0.9974

9 21 1765 232 188 0.8569 0.0085 0.8393 0.8728

21 33 1345 312 164 0.6453 0.0122 0.6207 0.6686

33 69 869 340 322 0.3354 0.0137 0.3087 0.3623

69 129 207 44 163 0.2178 0.0168 0.1858 0.2516

age1teen = 1

0 9 1017 4 0 0.9961 0.0020 0.9896 0.9985

9 21 1013 176 0 0.8230 0.0120 0.7981 0.8451

21 33 837 249 0 0.5782 0.0155 0.5472 0.6079

33 69 588 280 1 0.3026 0.0144 0.2746 0.3310

69 129 307 85 141 0.1939 0.0132 0.1687 0.2204

129 . 81 10 71 0.1512 0.0157 0.1219 0.1835

-------------------------------------------------------------------------------

.

. \* add hazard intervals

. ltable dur birth2 if(dur>0), by(age1teen) hazard interval(0,9,21,33,69,129)

Beg. Cum. Std. Std.

Interval Total Failure Error Hazard Error [95% Conf. Int.]

-------------------------------------------------------------------------------

age1teen 0

0 9 1901 0.0049 0.0016 0.0005 0.0002 0.0002 0.0009

9 21 1765 0.1431 0.0085 0.0124 0.0008 0.0108 0.0140

21 33 1345 0.3547 0.0122 0.0235 0.0013 0.0209 0.0261

33 69 869 0.6646 0.0137 0.0176 0.0009 0.0158 0.0193

69 129 207 0.7822 0.0168 0.0071 0.0010 0.0050 0.0091

age1teen 1

0 9 1017 0.0039 0.0020 0.0004 0.0002 0.0000 0.0009

9 21 1013 0.1770 0.0120 0.0159 0.0012 0.0135 0.0182

21 33 837 0.4218 0.0155 0.0291 0.0018 0.0256 0.0327

33 69 588 0.6974 0.0144 0.0174 0.0010 0.0154 0.0193

69 129 307 0.8061 0.0132 0.0073 0.0008 0.0058 0.0088

129 . 81 0.8488 0.0157 . . . .

-------------------------------------------------------------------------------

.

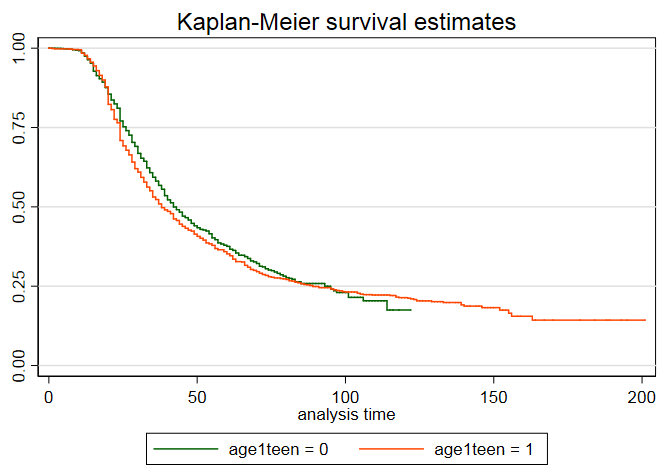
. \* graph results

. sts graph, by(age1teen)

failure \_d: birth2

analysis time \_t: dur

id: id



.

. \* calculate the event rate for the overall sample and by age groups at first birth

. strate

failure \_d: birth2

analysis time \_t: dur

id: id

Estimated rates and lower/upper bounds of 95% confidence intervals

(2918 records included in the analysis)

+-------------------------------------------------+

| D Y Rate Lower Upper |

|-------------------------------------------------|

| 1741 1.2e+05 0.013933 0.013294 0.014603 |

+-------------------------------------------------+

. strate age1teen

failure \_d: birth2

analysis time \_t: dur

id: id

Estimated rates and lower/upper bounds of 95% confidence intervals

(2918 records included in the analysis)

+-----------------------------------------------------------+

| age1teen D Y Rate Lower Upper |

|-----------------------------------------------------------|

| 0 937 6.8e+04 0.013814 0.012957 0.014727 |

| 1 804 5.7e+04 0.014075 0.013135 0.015082 |

+-----------------------------------------------------------+

.

. \* now, we are ready for a regression-style model

. \* use a cox model to automatically control for duration

. stcox age1teen hispanic nhblack nhother

failure \_d: birth2

analysis time \_t: dur

id: id

Iteration 0: log likelihood = -12569.256

Iteration 1: log likelihood = -12563.018

Iteration 2: log likelihood = -12562.965

Iteration 3: log likelihood = -12562.965

Refining estimates:

Iteration 0: log likelihood = -12562.965

Cox regression -- Breslow method for ties

No. of subjects = 2,918 Number of obs = 7,843

No. of failures = 1,741

Time at risk = 124953

LR chi2(4) = 12.58

Log likelihood = -12562.965 Prob > chi2 = 0.0135

------------------------------------------------------------------------------

\_t | Haz. Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

age1teen | 1.080141 .053695 1.55 0.121 .9798647 1.190678

hispanic | 1.157423 .0815311 2.08 0.038 1.008165 1.328778

nhblack | 1.038949 .0700246 0.57 0.571 .9103819 1.185672

nhother | 1.354747 .1486091 2.77 0.006 1.092661 1.679696

------------------------------------------------------------------------------

.

. \* if you are interested in a particular duration, you must make your own

. \* duration variables and interactions

.

. stsplit durcat, at(9 27 73)

(4,925 observations (episodes) created)

. egen durgroup = group(durcat)

(34 missing values generated)

. gen dur0008 = durgroup==1

. gen dur0926 = durgroup==2

. gen dur2772 = durgroup==3

. gen dur73p = durgroup==4

.

. generate teen0008 = age1teen\*dur0008

. generate teen0926 = age1teen\*dur0926

. generate teen2772 = age1teen\*dur2772

. generate teen73p = age1teen\*dur73p

.

. \* then STATA allows you to control your own duration variables and interactions

. \* then STATA allows you to control your own duration variables and interactions

. streg age1teen hispanic nhblack nhother, dist(exp)

failure \_d: birth2

analysis time \_t: dur

id: id

Iteration 0: log likelihood = -3226.7952

Iteration 1: log likelihood = -3222.8125

Iteration 2: log likelihood = -3222.765

Iteration 3: log likelihood = -3222.765

Exponential PH regression

No. of subjects = 2,918 Number of obs = 7,843

No. of failures = 1,741

Time at risk = 124953

LR chi2(4) = 8.06

Log likelihood = -3222.765 Prob > chi2 = 0.0894

------------------------------------------------------------------------------

\_t | Haz. Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

age1teen | 1.021616 .0496927 0.44 0.660 .9287184 1.123806

hispanic | 1.1239 .0791543 1.66 0.097 .9789922 1.290258

nhblack | 1.032775 .0696405 0.48 0.632 .9049172 1.178698

nhother | 1.320705 .1448353 2.54 0.011 1.065267 1.637393

\_cons | .0133384 .0004855 -118.61 0.000 .01242 .0143247

------------------------------------------------------------------------------

Note: \_cons estimates baseline hazard.

. streg age1teen hispanic nhblack nhother dur0008 dur0926 dur73p, dist(exp)

failure \_d: birth2

analysis time \_t: dur

id: id

Iteration 0: log likelihood = -3226.7952

Iteration 1: log likelihood = -2911.3742

Iteration 2: log likelihood = -2832.3141

Iteration 3: log likelihood = -2824.4367

Iteration 4: log likelihood = -2824.3615

Iteration 5: log likelihood = -2824.3614

Exponential PH regression

No. of subjects = 2,918 Number of obs = 7,843

No. of failures = 1,741

Time at risk = 124953

LR chi2(7) = 804.87

Log likelihood = -2824.3614 Prob > chi2 = 0.0000

------------------------------------------------------------------------------

\_t | Haz. Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

age1teen | 1.051414 .0521424 1.01 0.312 .9540269 1.158743

hispanic | 1.143459 .0805207 1.90 0.057 .9960476 1.312686

nhblack | 1.034758 .0697257 0.51 0.612 .9067382 1.180853

nhother | 1.351532 .148258 2.75 0.006 1.090066 1.675714

dur0008 | .0293801 .0076539 -13.54 0.000 .0176321 .0489556

dur0926 | .9165851 .0458006 -1.74 0.081 .8310733 1.010895

dur73p | .3591115 .0374795 -9.81 0.000 .2926791 .4406227

\_cons | .0187883 .0008353 -89.39 0.000 .0172203 .0204989

------------------------------------------------------------------------------

Note: \_cons estimates baseline hazard.

. streg age1teen hispanic nhblack nhother dur0008 dur0926 dur73p /\*

> \*/ teen0008 teen0926 teen73p, dist(exp)

failure \_d: birth2

analysis time \_t: dur

id: id

Iteration 0: log likelihood = -3226.7952

Iteration 1: log likelihood = -2905.7958

Iteration 2: log likelihood = -2824.3904

Iteration 3: log likelihood = -2815.0915

Iteration 4: log likelihood = -2814.9754

Iteration 5: log likelihood = -2814.9749

Iteration 6: log likelihood = -2814.9749

Exponential PH regression

No. of subjects = 2,918 Number of obs = 7,843

No. of failures = 1,741

Time at risk = 124953

LR chi2(10) = 823.64

Log likelihood = -2814.9749 Prob > chi2 = 0.0000

------------------------------------------------------------------------------

\_t | Haz. Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

age1teen | .9377954 .0646626 -0.93 0.352 .8192497 1.073495

hispanic | 1.137829 .0801446 1.83 0.067 .9911093 1.306269

nhblack | 1.025737 .0691419 0.38 0.706 .8987918 1.170612

nhother | 1.344263 .1474581 2.70 0.007 1.084207 1.666696

dur0008 | .0323194 .0098583 -11.25 0.000 .0177756 .0587627

dur0926 | .7983912 .053434 -3.36 0.001 .7002407 .9102993

dur73p | .5529736 .1041542 -3.15 0.002 .3822769 .799891

teen0008 | .7004167 .4117823 -0.61 0.545 .2212719 2.217108

teen0926 | 1.373591 .1375897 3.17 0.002 1.128742 1.671555

teen73p | .5996796 .1357383 -2.26 0.024 .3848121 .9345227

\_cons | .0198621 .0009725 -80.04 0.000 .0180446 .0218627

------------------------------------------------------------------------------

Note: \_cons estimates baseline hazard.end of do-file

. log close

closed on: 9 Aug 2018, 19:41:28

-------------------------------------------------------------------------------------