2013 Spanish Stata Users Group meeting

10th October 2013

Survival Analysis with Stata

Case studies of fertility of immigrant women

Rafael Grande

rgrande@usal.es

VNiVERSiDAD
Ð SALAMANCA

INTRODUCTION

- The *Survival Analysis* (*or Event History Analysis*) is being used progressively more in diverse research areas of social science, expanding its traditional use in demographics and health sciences.
  - Potential to analyze the development and occurrence of an event over time
  - More easy of doing analysis with statistical software package, for example STATA.

- **Aim of this presentation**:
  - ➢ To expose the tools available in STATA for Survival Analysis.
  - ➢ Example for the case of probability of having the first child in Spain after the arrival for latino-american immigrant women in Spain

SURVIVAL ANALYSIS

- Survival analysis examines the time to the occurrence of an event.
  - For example: death in biological organisms, failure in mechanical systems, effect of a drug, or fertility
- Basic Concepts:
  - **Event** → occur in an instant of time.
  - **Failure event** → the event to be analyzed.
  - **At risk** → the individual is at risk of the event of interest occurs.
  - **Origin** → At the moment in which the individual enters first risk.
  - **Time** → as measured in the data (days, dates, years, etc..)
  - **Analytical time** (time to event) → the time elapsed since the individual enters risk.
  - **Scale** → converter the time in analytic time

SURVIVAL ANALYSIS

The case of fertility of immigrants in Spain

- Period : “the wonder years” of immigration in Spain (1990-2007)
- Source: National Survey on Immigration 2007 (ENI)
  - This survey is statistically representative of the 4.5 million immigrants who lived in Spain in 2007
  - Sample → 15465 individuals (>16 years & who have resided for the least 1 year in Spain)
  - Sample used: Latin women arrivals between 1990 and 2007 → 3157
- Survival analysis allows adopt a longitudinal perspective on the fertility of immigrants in the residence time
- The dependent variable is the time until first birth after migration or until 10 years reside in Spain without having any son (censored cases).
- The duration of the transition was calculated based on the year of arrival and year of the first birth. We used continued-time models.

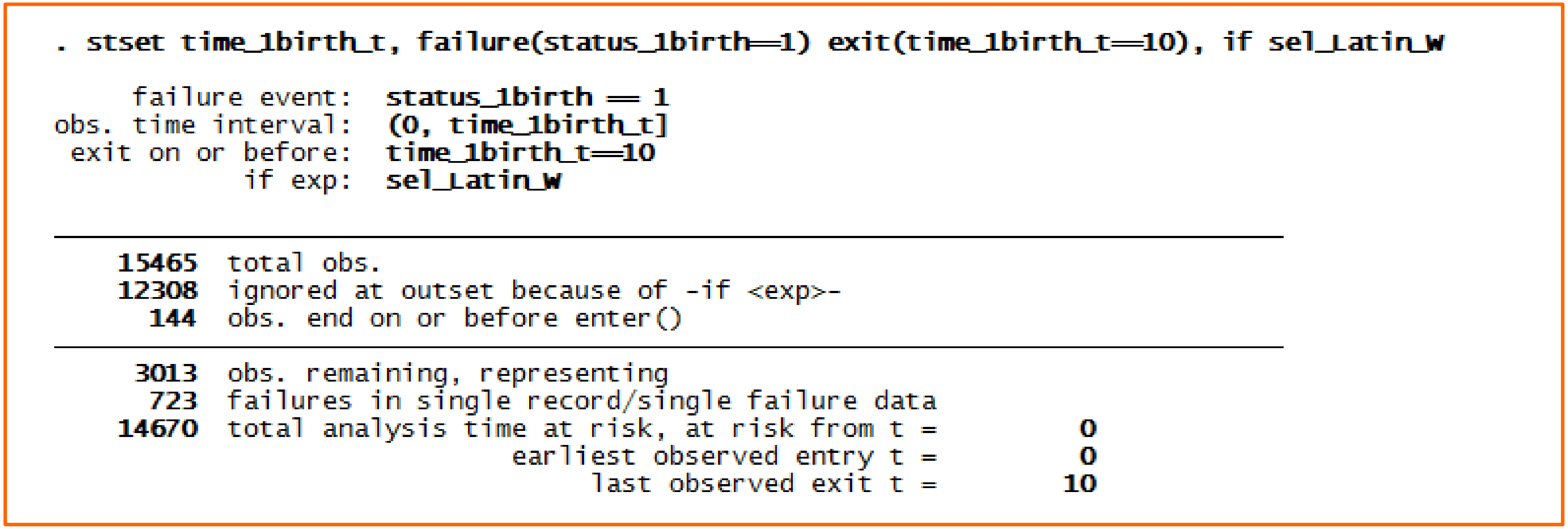
PREPROCESSING DATA (I)

- Survival analysis requires specialized data management and analysis procedures.

- Stata provides the **st** family of commands for organizing and summarizing survival data (Cleves et al. 2004)

  - **stset** → declare data to be survival-time data

    - **stset** create value analytical time



```
. stset time_1birth_t, failure(status_1birth==1) exit(time_1birth_t==10), if sel_Latin_W

     failure event:  status_1birth == 1
obs. time interval:  (0, time_1birth_t]
 exit on or before:  time_1birth_t==10
            if exp:  sel_Latin_W

------------------------------------------------------------------------------
    15465  total obs.
    12308  ignored at outset because of -if <exp>-
      144  obs. end on or before enter()
------------------------------------------------------------------------------
     3013  obs. remaining, representing
      723  failures in single record/single failure data
    14670  total analysis time at risk, at risk from t =         0
                                 earliest observed entry t =         0
                                      last observed exit t =        10
```

PREPROCESSING DATA (II)

- The **stset** command creates 4 variables. These variables contain all necessary information for the survival data. All the survival analysis (**st**) commands use these variables, as all information regarding

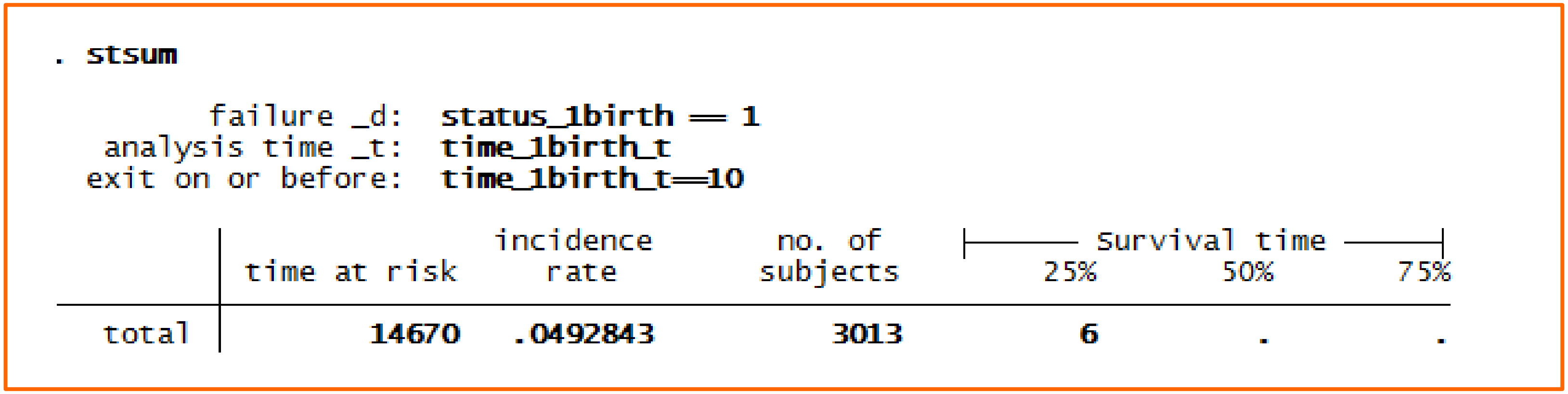
  _t0 → analysis time when record begins (time at which individual becomes at risk)

  _t → analysis time when record ends (time at which individual stops being at risk)

  _d → failure indicator: 1 if failure, 0 if censored

  _st → 1 if the record is included in st analyses, 0 if excluded

  - **stdes** → describe survival-time data
  - **stsum** → summarize survival-time data



```
. stsum

         failure _d:  status_1birth == 1
   analysis time _t:  time_1birth_t
  exit on or before:  time_1birth_t==10

         |               incidence       no. of     |------ Survival time -----|
         | time at risk     rate        subjects        25%       50%       75%
---------+---------------------------------------------------------------------
   total |        14670   .0492843          3013          6         .         .
```

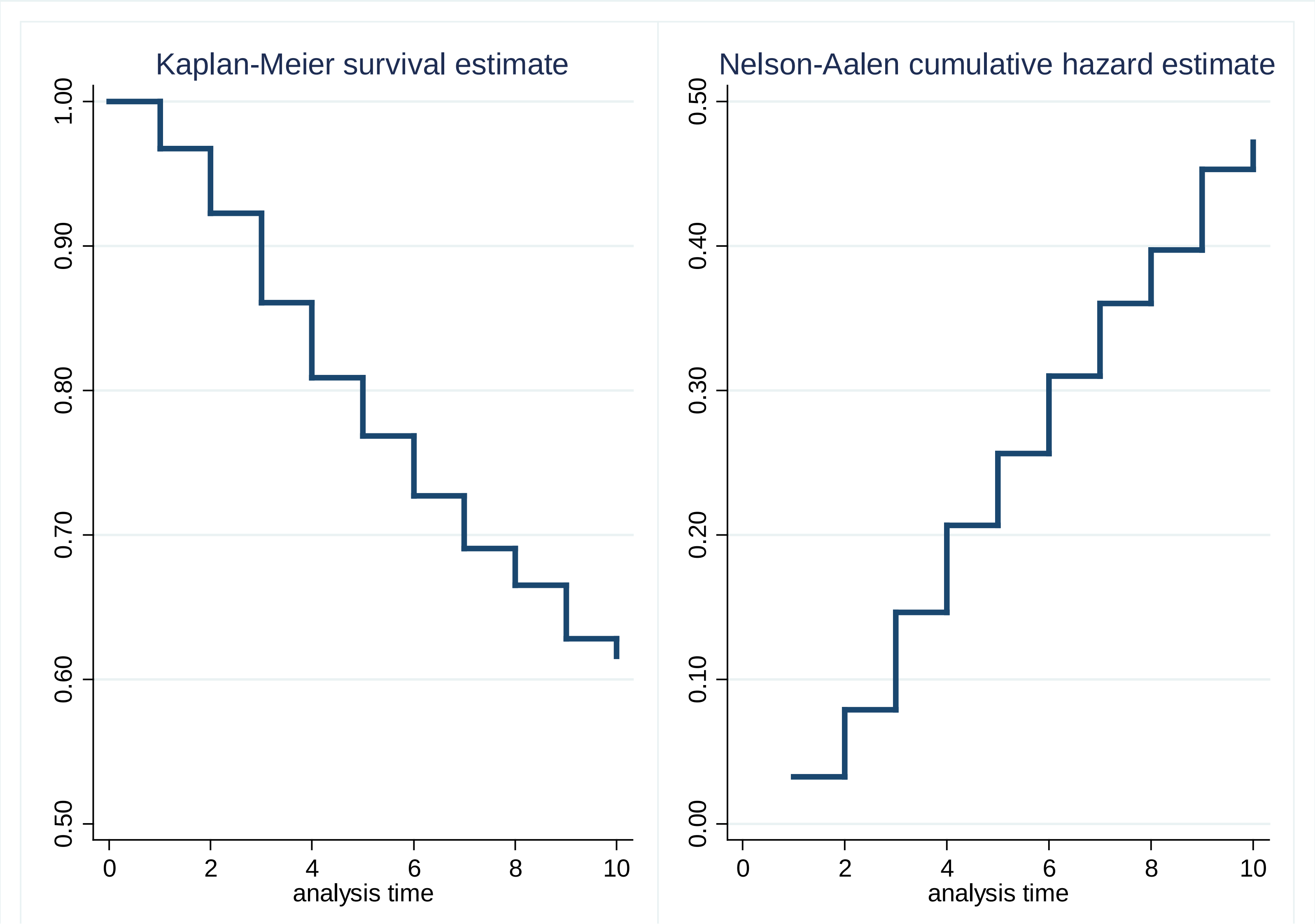
SURVIVAL FUNCTION (I)

- Estimation of the survival function (Kaplan-Meier estimate): It is a **nonparametric** method (assumes no probability function) and maximum likelihood.

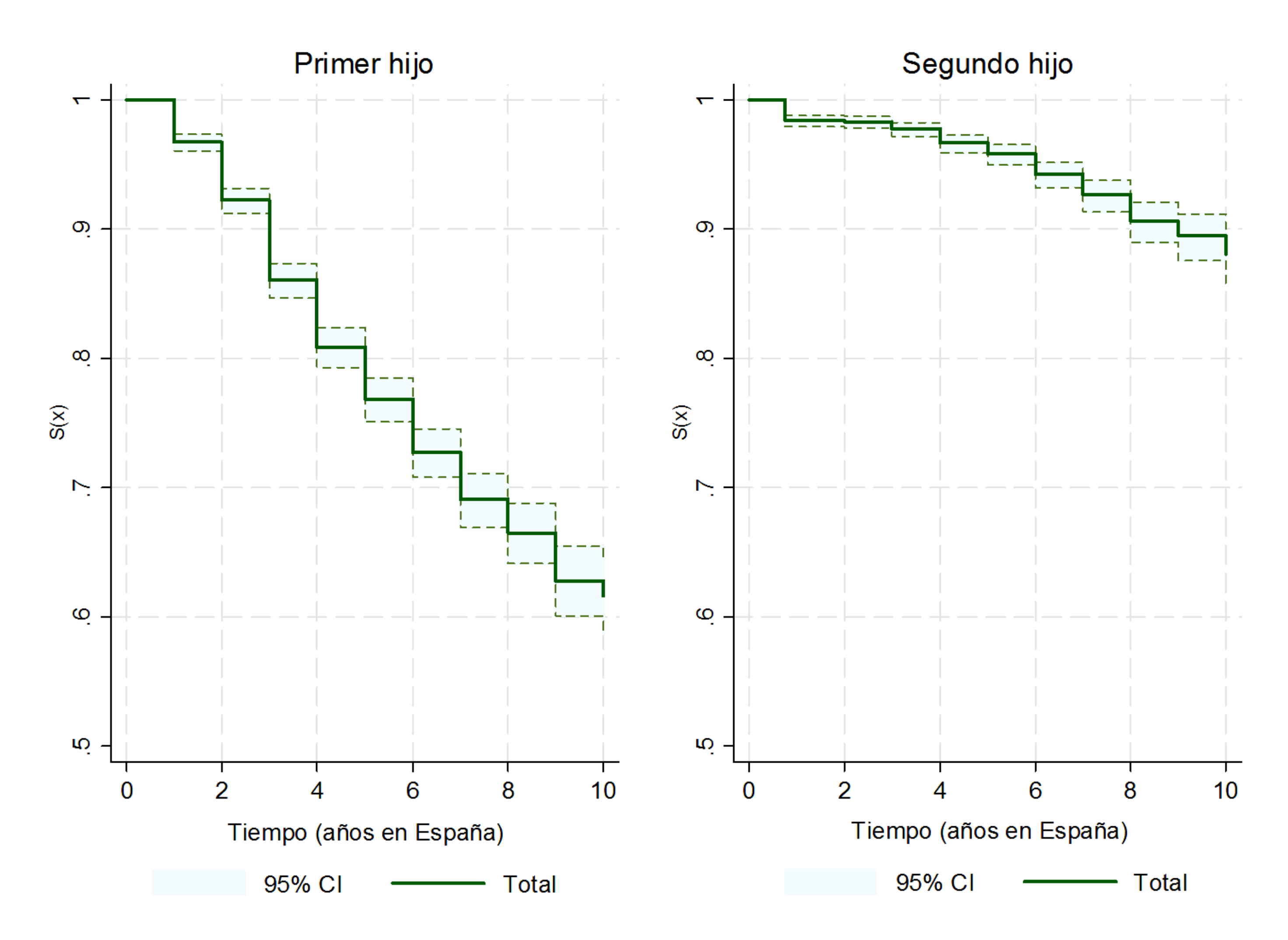
- `sts graph` → graph the survivor and cumulative hazard functions. By defect Kaplan-Meir survival estimate.
  - Example options:
    - `cumhaz` → Nelson-Aalen function of cumulative risk
    - `by(varlist)` → calculate separately on different groups of varlist
    - `risktable` → show table of number of individual at risk beneath graph
    - `gwood` → point-wise confidence bands be displayed

- `sts test` → test equality of survivor functions (perform log-rank, cox, wilcoxon, etc.

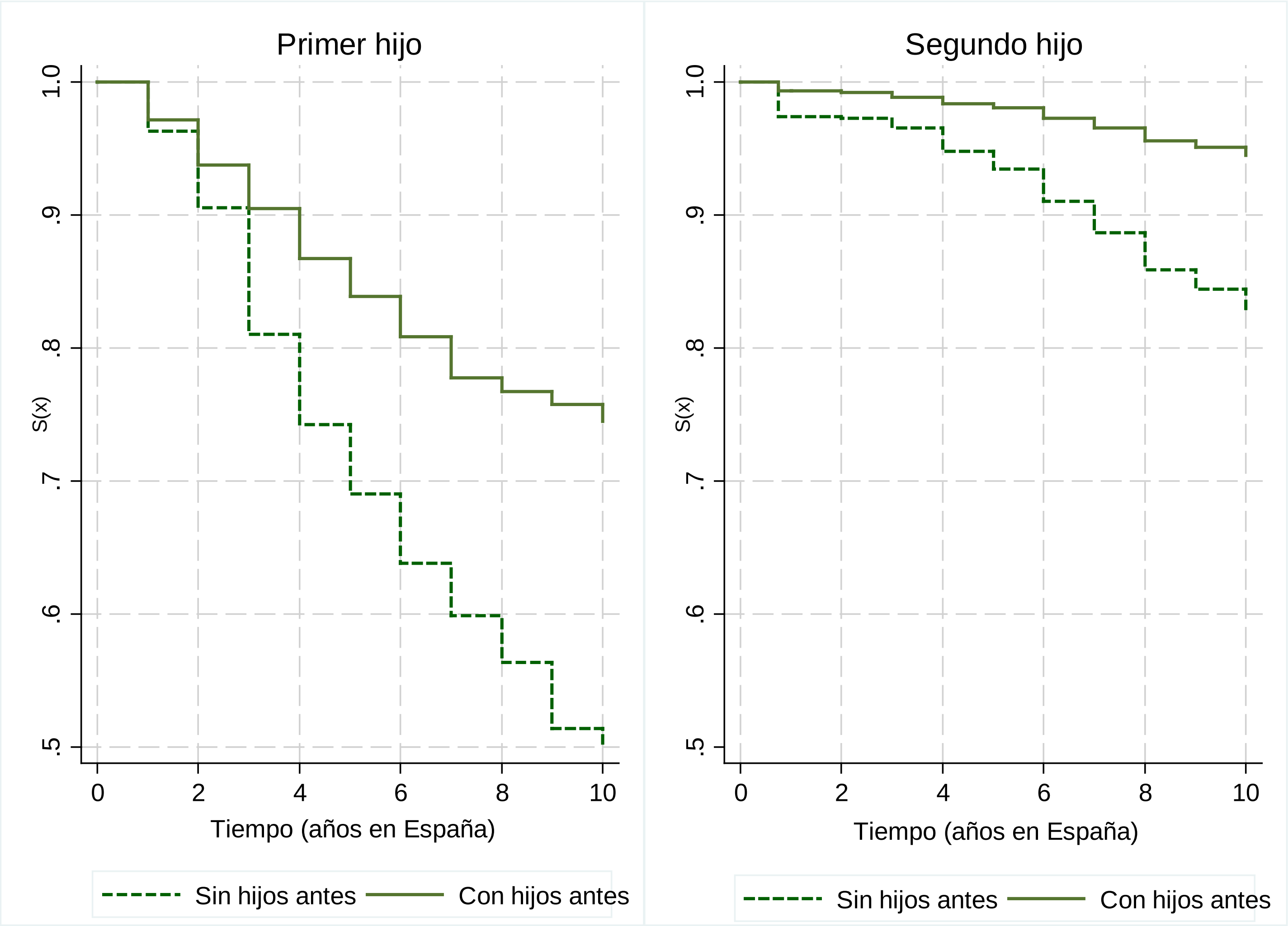
SURVIVAL CURVE (I)



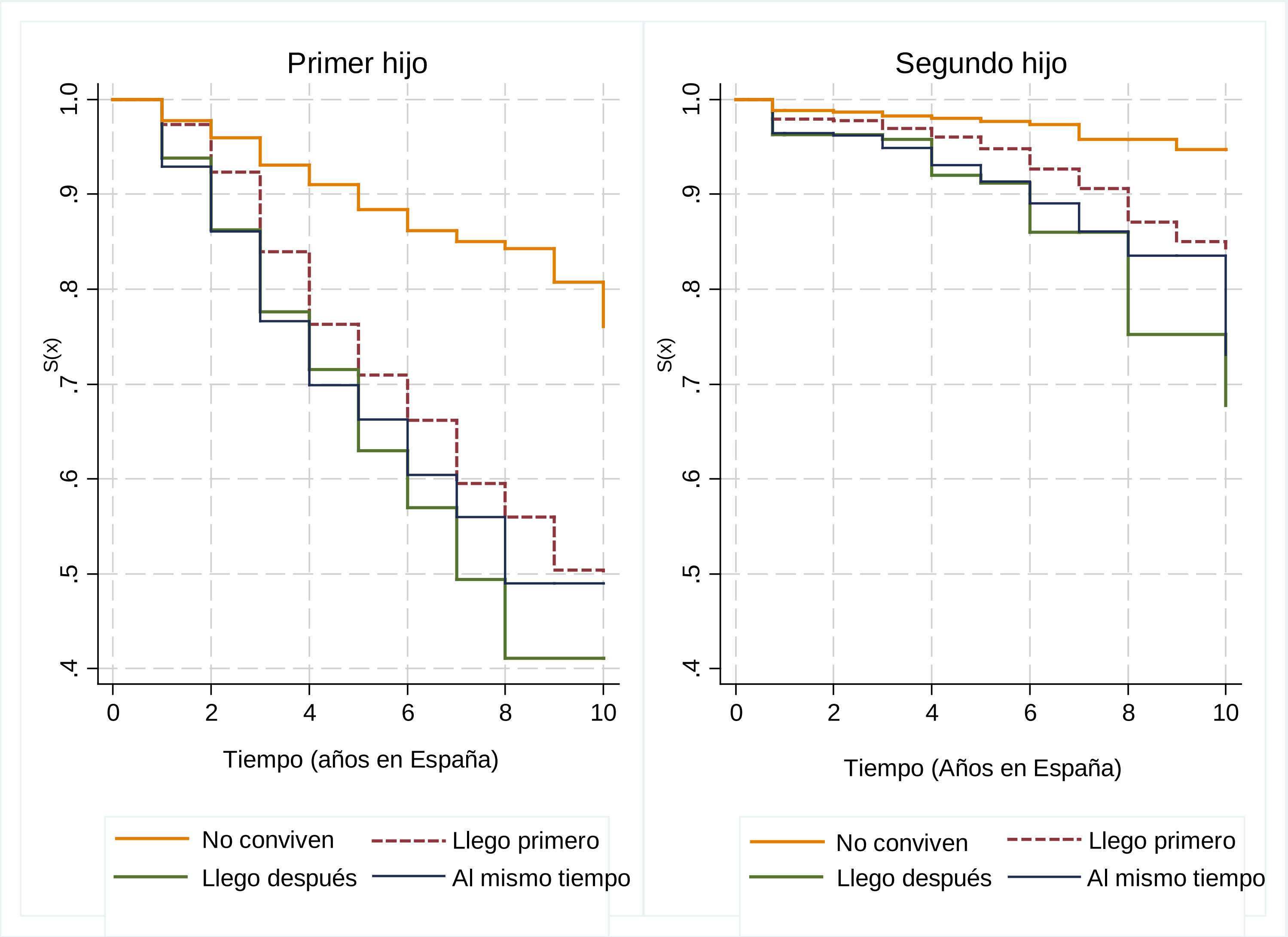
SURVIVAL CURVE (II)



SURVIVAL CURVE (III)



SURVIVAL CURVE (IV)



Regression Models

- Stata has commands for fitting both semiparametric and parametric regression models to survival data.

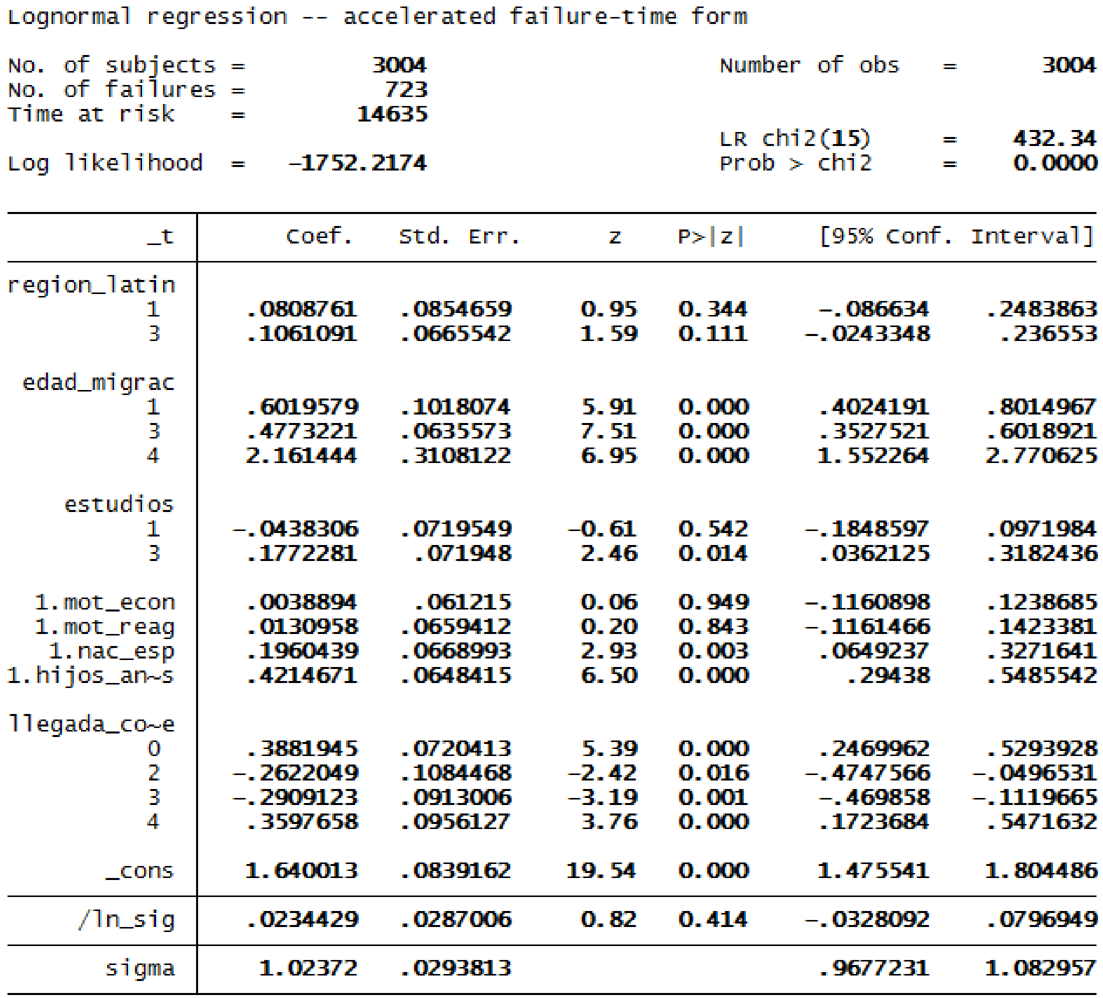
- `stcox [varlist] [if] [in] [,options]` - semiparametric model
  - stcox fits the Cox proportional hazards model and predict after stcox can be used to retrieve estimates of the baseline survivor function, the baseline cumulative hazard function, and the baseline hazard contributions.

- `streg [varlist] [if] [in], dist[distnamen]` - parametric model
  - Stata offers six parametric regression models for survival data: Exponential, Weibull, Lognormal, Loglogistic, Gompertz, and Gamma → `dist[distnamen]`
  - Stratified models may also be fit using streg.

- Noted Options:
  - `nohr` → report coefficients, not hazard ratios

Regression Models: streg

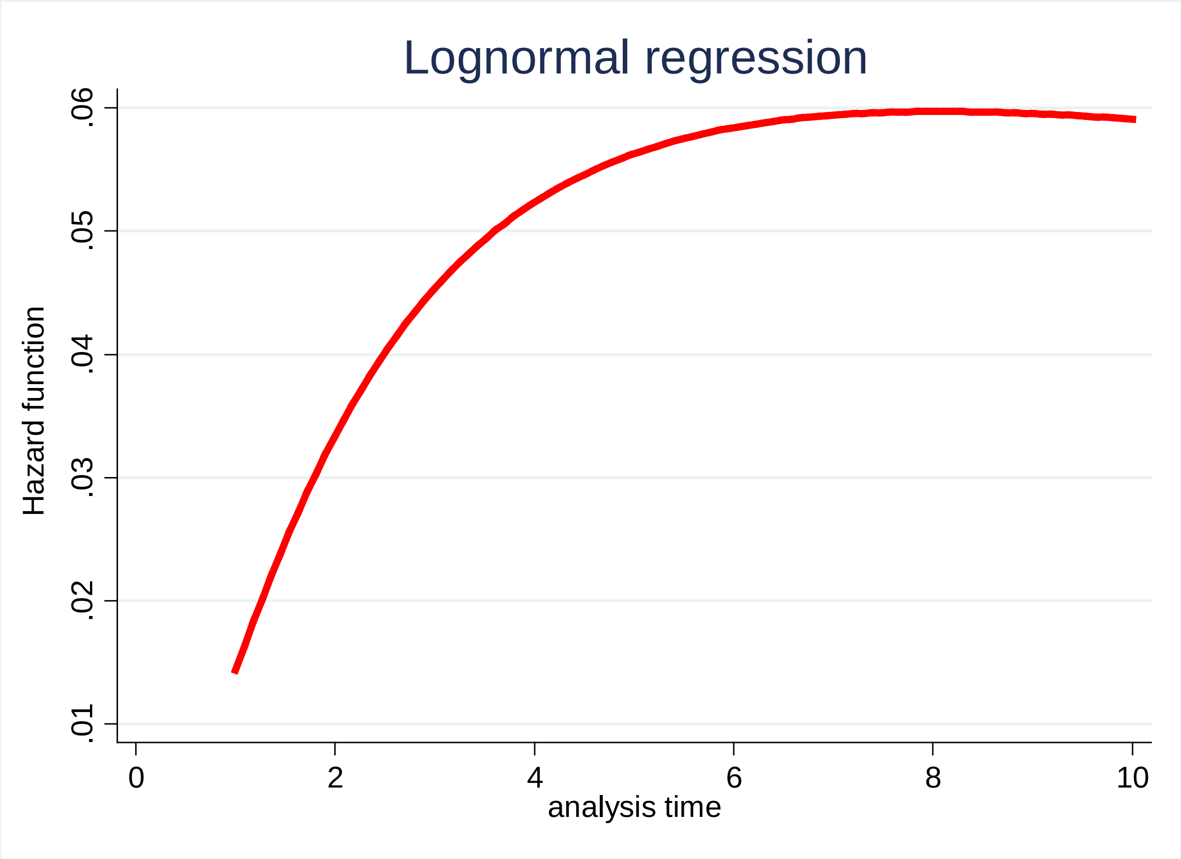
streg ib2.region_latin ib2.edad_migrac ib2.estudios i.mot_econ i.mot_reag i.nac_esp i.hijos_antes ib1.llegada_conyuge, dist(lnormal)



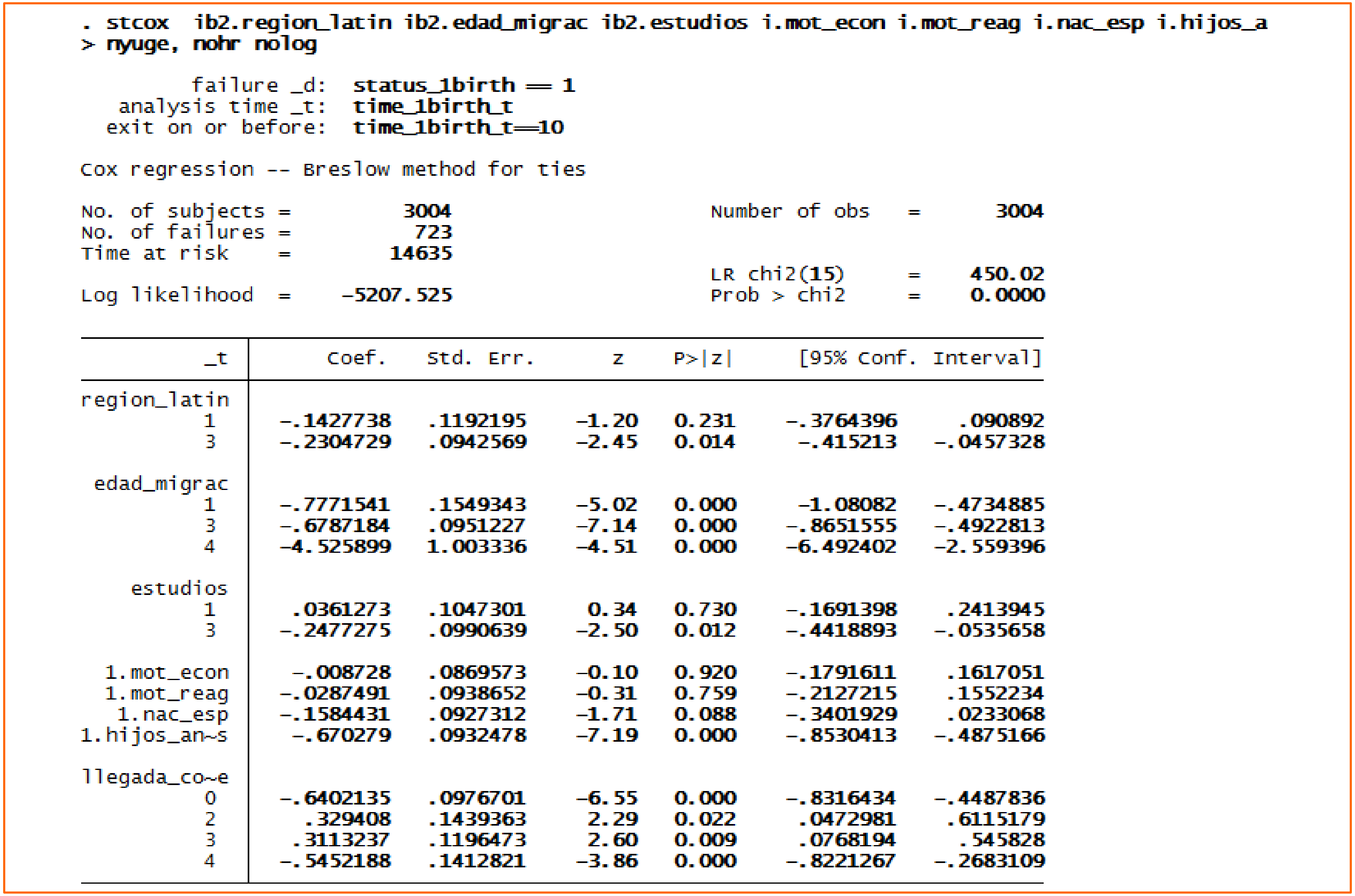
```
Lognormal regression -- accelerated failure-time form

No. of subjects =         3004                 Number of obs    =      3004
No. of failures =          723
Time at risk    =        14635
                                               LR chi2(15)      =    432.34
Log likelihood  =   -1752.2174                 Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
region_latin |
          1  |   .0808761   .0854659     0.95   0.344     -.086634    .2483863
          3  |   .1061091   .0665542     1.59   0.111    -.0243348     .236553
             |
 edad_migrac |
          1  |   .6019579   .1018074     5.91   0.000     .4024191    .8014967
          3  |   .4773221   .0635573     7.51   0.000     .3527521    .6018921
          4  |   2.161444   .3108122     6.95   0.000     1.552264    2.770625
             |
    estudios |
          1  |  -.0438306   .0719549    -0.61   0.542    -.1848597    .0971984
          3  |   .1772281    .071948     2.46   0.014     .0362125    .3182436
             |
  1.mot_econ |   .0038894    .061215     0.06   0.949    -.1160898    .1238685
  1.mot_reag |   .0130958   .0659412     0.20   0.843    -.1161466    .1423381
   1.nac_esp |   .1960439   .0668993     2.93   0.003     .0649237    .3271641
1.hijos_an~s |   .4214671   .0648415     6.50   0.000       .29438    .5485542
             |
llegada_co~e |
          0  |   .3881945   .0720413     5.39   0.000     .2469962    .5293928
          2  |  -.2622049   .1084468    -2.42   0.016    -.4747566   -.0496531
          3  |  -.2909123   .0913006    -3.19   0.001     -.469858   -.1119665
          4  |   .3597658   .0956127     3.76   0.000     .1723684    .5471632
             |
       _cons |   1.640013   .0839162    19.54   0.000     1.475541    1.804486
-------------+----------------------------------------------------------------
     /ln_sig |   .0234429   .0287006     0.82   0.414    -.0328092    .0796949
-------------+----------------------------------------------------------------
       sigma |    1.02372   .0293813                      .9677231    1.082957
------------------------------------------------------------------------------
```

Regression Models: streg

- `stcurve` → is for use after stcox and streg and will plot the estimated survivor, hazard, cumulative hazard, and cumulative incidence function for the fitted model.



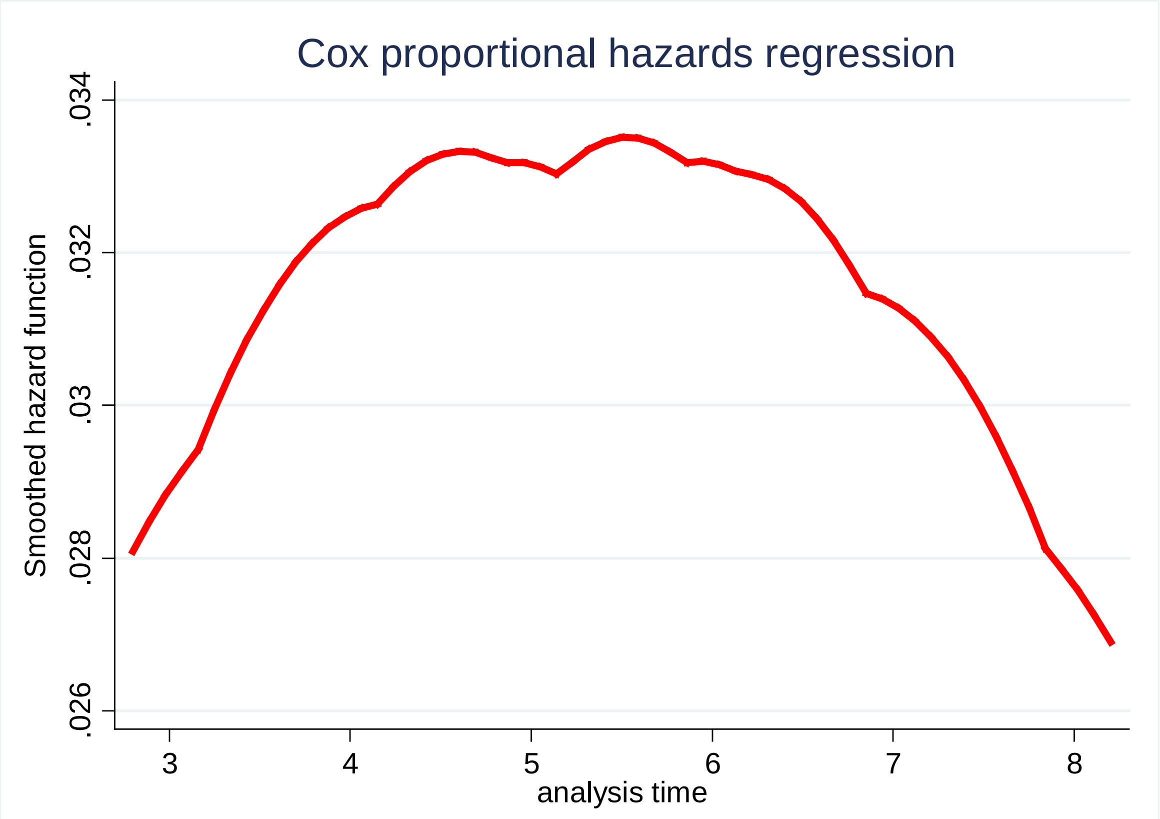
Regression Models: stcox



```
. stcox  ib2.region_latin ib2.edad_migrac ib2.estudios i.mot_econ i.mot_reag i.nac_esp i.hijos_a
> nyuge, nohr nolog

         failure _d:  status_1birth == 1
   analysis time _t:  time_1birth_t
  exit on or before:  time_1birth_t==10

Cox regression -- Breslow method for ties

No. of subjects =         3004                     Number of obs    =      3004
No. of failures =          723
Time at risk    =        14635
                                                   LR chi2(15)      =    450.02
Log likelihood  =    -5207.525                     Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
region_latin |
          1  |  -.1427738   .1192195    -1.20   0.231    -.3764396     .090892
          3  |  -.2304729   .0942569    -2.45   0.014     -.415213    -.0457328
             |
  edad_migrac |
          1  |  -.7771541   .1549343    -5.02   0.000     -1.08082    -.4734885
          3  |  -.6787184   .0951227    -7.14   0.000    -.8651555    -.4922813
          4  |  -4.525899   1.003336    -4.51   0.000    -6.492402    -2.559396
             |
    estudios |
          1  |   .0361273   .1047301     0.34   0.730    -.1691398     .2413945
          3  |  -.2477275   .0990639    -2.50   0.012    -.4418893    -.0535658
             |
  1.mot_econ |   -.008728   .0869573    -0.10   0.920    -.1791611     .1617051
  1.mot_reag |  -.0287491   .0938652    -0.31   0.759    -.2127215     .1552234
   1.nac_esp |  -.1584431   .0927312    -1.71   0.088    -.3401929     .0233068
1.hijos_an~s |   -.670279   .0932478    -7.19   0.000    -.8530413    -.4875166
             |
llegada_co~e |
          0  |  -.6402135   .0976701    -6.55   0.000    -.8316434    -.4487836
          2  |    .329408   .1439363     2.29   0.022     .0472981     .6115179
          3  |   .3113237   .1196473     2.60   0.009     .0768194      .545828
          4  |  -.5452188   .1412821    -3.86   0.000    -.8221267    -.2683109
------------------------------------------------------------------------------
```

Regression Models: stcox

- **stcurve** → is for use after stcox and streg and will plot the estimated survivor, hazard, cumulative hazard, and cumulative incidence function for the fitted model.



Thanks for your attention and comments!

REFERENCES

BERNARDI, Fabrizio. 2006. *El análisis de la Historia de Acontecimientos*. Madrid: CIS.

CLEVES, Mario; William W. GOULD; Roberto G. GUTIERREZ, and Yulia MARCHENKO. 2004. *An Introduction to Survival Analysis Using Stata*. College Station, Texas: Stata Press.

ESCOBAR MERCADO, Modesto; Enrique FERNÁNDEZ MACÍAS, BERNARDI, Fabrizio. 2010. *Análisis de datos con Stata*. Madrid: CIS