Gompertz regression parameterized as accelerated failure time model

Filip Andersson and Nicola Orsini

Biostatistics Team
Department of Public Health Sciences
Karolinska Institutet

2017 Nordic and Baltic Stata meeting

Karolinska Institutet

Content

- Introduction
- Proportional hazard model
- Accelerated failure time model
- The Gompertz distribution
- Structural equation models and mediation
- Mediation in survival models
- Estimating confidence intervals
- What I am working on

Content

- Example
  - → Data
  - → Pre-estimation
  - → Gompertz proportional hazard
  - → Cox regression
  - → Gompertz vs. Kaplan-Maier
  - → Gompertz ATF model
  - → Post-estimation
  - → Conclusion

Introduction

- Why use parametric surival models?
  - → Can handle right-, left- or interval-censored data
  - → Cox regression can't handle left- or interval-censored data
  - → Produce better estimation if you have a theoretical expectation of the baseline hazard
  - → Can estimate expected life, not only hazard ratios (AFT-models)
  - → Can include random effects – frailty models (not discussed here)

Karolinska Institutet

Introduction

- A model that is lacking an easy way to estimate in Stata
  → Gompertz regression parameterized as accelerated failure time model
  → Exist in R
    - eha-package, with command: aftreg

- Why use Stata?
  → Easy handling survival data
    - Data management
    - Setup
  → Good graphical possibility

Proportional hazard model

- Easy to compare with Cox regression
  - → Hazard ratios
  - → Plots
    - Cummulative hazard function
    - Survival function
  - → Commonly used

- Hazard function general form
  - → $h(t|x) = h_0(t)e^{xb}$

Accelerated failure time model

- Can be seen as a linear model (simplest form):
  → $\log(t) = a + bx + \varepsilon$
  → Useful in mediation

- Estimation on life scale
  → Estimation of expected baseline life
    - Area under the survival curve when all covariates are zero
  → Compare expected life between two groups
    - Logarithmic change in expected life compared to the baseline life expectancy
    - $\text{Expected life} = \text{Baseline life expectancy} * \exp(\text{effect})$

Accelerated failure time model

- Definiton of accelerated failure time model
  → For a group (X_1,X_2…X_p) , the model is written mathematically as $S(t|x) = S_0\left(\frac{t}{\eta(x)}\right)$, where $S_0(t)$ is the baseline survival function and $\eta(x)$ is an acceleration factor that is a ratio of survival times corresponding to any fixed value of S(t). The acceleration factor is given according to the formula $\eta(x) = e^{(a_1 x_1 + \cdots + a_p x_p)}$. (Qi, J (2009))

- Hazard function
  → $h(t|x) = \left[\frac{1}{\eta(x)}\right] h_0 \left[\frac{t}{\eta(x)}\right]$

- Log-linear from
  → $\log(t) = a + bx + \sigma\varepsilon$
  → Where t and ε following corresponing distributions

The Gompertz distribution

- When is it useful?
  → Adult and old age mortality for humans
    - Demographic models
    - Including models with treatment effects, such as cancer patiens
    - Can be problem with very old individuals

- Normal paramertization
  → $h(t) = \lambda e^{\gamma t}$
  → $\lambda > 0,\ \gamma \geq 0,\ t > 0$

The Gompertz distribution

- Suggested new parametrization by Broström, G & Edvinsson, S (2013)
  - → $\lambda \to \frac{\lambda}{\gamma},\ \gamma \to \frac{1}{\gamma}$
  - → $h(t) = \frac{\lambda}{\gamma} e^{t/\gamma}$
  - → $\lambda > 0,\ \gamma > 0,\ t > 0$

- Proof of new parametrization
  - → Hazard for AFT-model
  - → $h(t|x) = \left[\frac{1}{\eta(x)}\right] h_0 \left[\frac{t}{\eta(x)}\right]$
  - → Here, new gamma can be seen as an accelerated factor

Karolinska Institutet

The Gompertz distribution

- Linear model: $\log(t) = a + bx + \varepsilon$
  - Here, ε is following a log-Gompertz or inverse Weibull distribution
  - Compare to the Weibull model, where ε follows a Gumbel distribution

- Likelihood function

→ Survival function: $S(t) = exp\{-\lambda(e^{t/\gamma} - 1)\}$

→ Density function: $F(t) = h(t)S(t)$

→ Hazard function: $h(t) = \frac{\lambda}{\gamma} e^{t/\gamma}$

→ $L(\alpha, \mu, \sigma) = \prod_{i=1}^{n} \{h_i(t_i)S_i(t_i)\}^{\delta_i} \{S_i(t_i)\}^{1-\delta_i}$

Karolinska Institutet

Structural equation models and mediation

- Simple way to estimate linear models within a pathway framework
- Estimate all equations and combine for the direct and indirect effects
- Supported by most statistical programs
  → In Stata the gsem-command combined with simulation is preferable

Mediation in survival models

- What do we need to do?
  1. Estimate a parametric survival model
  2. Estimate the exposure on the mediator
     - First two steps directly from the gsem output
  3. Estimate the indirect, direct and total effect
  4. Estimate confidence intervals and significance
     - Step three and four can be done with either simulation or delta method
     - These models are simple for continous mediators, but can be tricky with binary or categorical mediators

Estimating confidence intervals

- Simulation
  - → Boostraping
    - Seems to be the more popular simulation method
    - Calculate point estimates for the indirect and direct effects
    - Simulate these point estimates
  - → Monte carlo simulation
    - More flexible to handle problematic correlations
    - Not as straight forward

- Delta method
    - Easiest method and probably most popular
    - Need a stronger assumption of normality

What I am working on

- A Stata command, **`staftgomp`**, to estimate the Gompertz regression parameterized as accelerated failure time model similar to what streg does

- A post-estimation command that would make it simple to estimate direct, indirect and total effect, with confidence intervals, for survival models

Example

- Scanian Economic Demographic Database (Bengtsson, T., Dribe, M. and Svensson, P. (2012))

- Longitudinal historical database
  - → Data from 17th century and onwards
  - → Here, data from individuals born between 1815-1860 are used
  - → Comes from five rural parishes in western Scania
  - → Consist of important life course events as birth and death, but also births of children, marriage or socioeconomic status are recorded

Data

- Variables used:
  → ”Treatment variable”:
    - Approximation of bad early life conditions
    - Infant mortality rate at the year of birth
    - High imr vs. low imr (binary)
    - Years of high diseaseload such as measles, smallpox and whooping cough (Quaranta, L. (2013))

  → Parental socioeconomic status
    - Socioceconomic status at birth (binary)
    - Confounder

  → Outcome
    - The individuals are followed until death or out-migration.

Pre-estimation

- Compare hazard estimations of Gompertz proportional hazard model and Cox regression
- Plot survival curve and compare with Kaplan-Maier
- If not acceptable test with different survival distribution until the parametric model is acceptable
  → Here, we choose Gompertz as it fits good and are supported theoretically for adult mortality

Gompertz proportional hazard

```
. streg imr_high ses, dist(gompertz)

Gompertz regression -- log relative-hazard form

No. of subjects =        3,756                  Number of obs    =      3,756
No. of failures =          880
Time at risk    =     19824107
                                                LR chi2(2)       =      26.53
Log likelihood  =   -1773.9194                  Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    imr_high |   1.259023   .0951873     3.05   0.002     1.085624    1.460119
         ses |   1.362878   .1010669     4.17   0.000     1.178513    1.576084
       _cons |   9.57e-06   8.25e-07  -134.05   0.000     8.08e-06    .0000113
-------------+----------------------------------------------------------------
      /gamma |   .0002332   8.35e-06    27.92   0.000     .0002168    .0002496
------------------------------------------------------------------------------
```

Cox regression

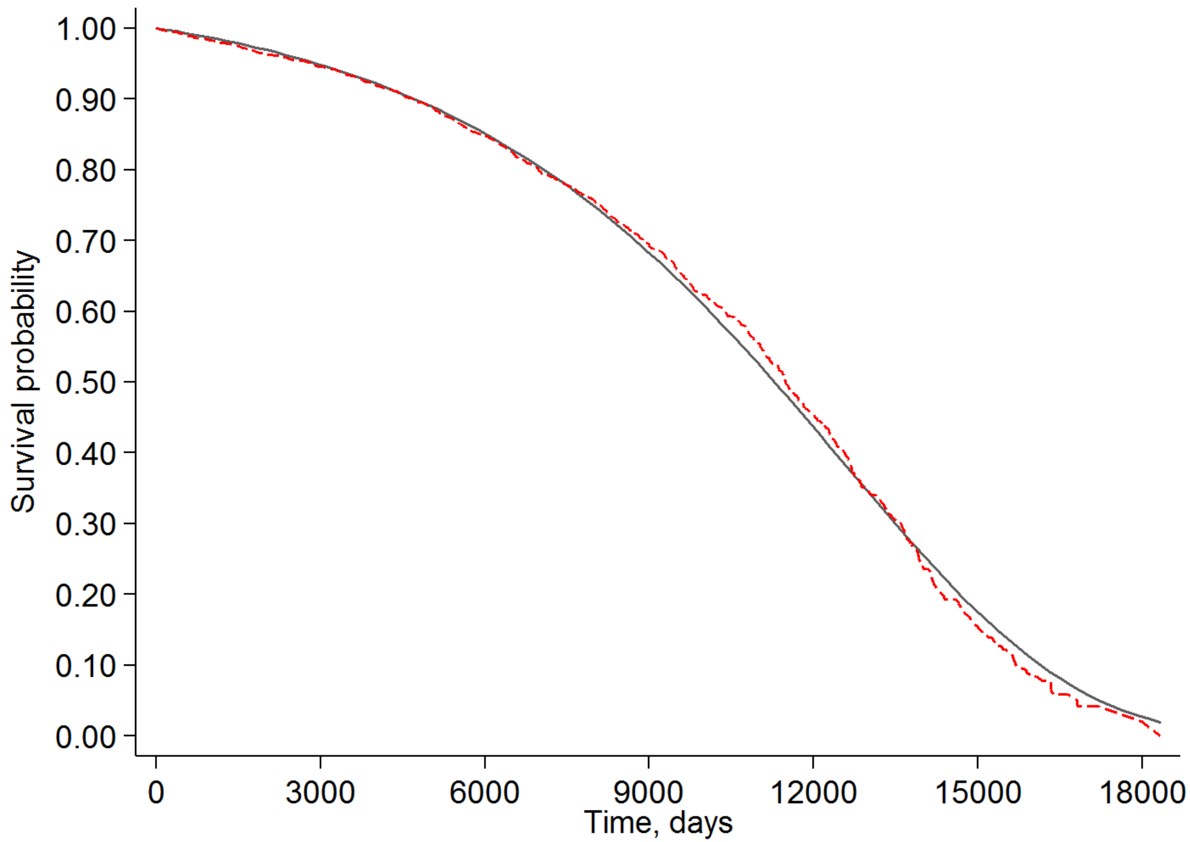
```
. stcox imr_high ses

Cox regression -- Breslow method for ties

No. of subjects =        3,756                  Number of obs    =      3,756
No. of failures =          880
Time at risk    =     19824107
                                                LR chi2(2)       =      28.17
Log likelihood  =   -5889.8259                  Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    imr_high |   1.261686   .0955679     3.07   0.002     1.087617    1.463614
         ses |   1.381581    .102833     4.34   0.000     1.194043    1.598573
------------------------------------------------------------------------------
```

Gompertz vs. Kaplan-Maier



Gompertz AFT model

```
. staftgomp imr_high ses

Gompertz AFT regression                               No. of obs =      3756
Log likelihood = -9325.8767                           LR chi2(2) =     14.34
Baseline life expectancy = 11669.94                  Prob > chi2 =    0.0008
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
    imr_high |  -.0496389   .0261027    -1.90   0.057    -.1007992    .0015214
         ses |  -.0873728   .0273233    -3.20   0.001    -.1409255   -.0338202
-------------+----------------------------------------------------------------
bp           |
      lambda |   8.434053   .0451675   186.73   0.000     8.345526    8.522579
       gamma |  -2.931995    .102271   -28.67   0.000    -3.132442   -2.731547
------------------------------------------------------------------------------
```

Post-estimation

```
. lincom imr_high, eform

 ( 1)  [xb]imr_high = 0

------------------------------------------------------------------------------
          _t |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    .951573   .0248386    -1.90   0.057     .9041146    1.001523
------------------------------------------------------------------------------


. nlcom exp([xb]imr_high)*11699

       _nl_1:  exp([xb]imr_high)*11699

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   11132.45   290.5866    38.31   0.000     10562.91    11701.99
------------------------------------------------------------------------------
```

Post-estimation

- Baseline life expectancy

  → $\frac{11699}{365}\text{days} = 32{,}1\ \text{years}$

- Estimating for individuals after 16000 days

  → $\frac{11699+16000}{365}\text{days} = 75{,}9\ \text{years of age}$

- Effect of high imr during birth

  → $\frac{11132+16000}{365}\text{days} = 74{,}3\ \text{years of age}$

Conclusion

- Conclusion
  → Even if you survive over the age of 40 you still have a mean shorter life expectancy of 1,6 years if you were born in a year with high imr
  → Latent effect
  → Support for the fetal origins hypothesis
  → Is the estimate reasonable?

- If needed
  → Mediation analysis and calculation of direct, indirect and total effect of treatment
  → Here, total effect = direct effect