Week 12

Hazard Models 1

Rich Frank
University of New Orleans
November 8, 2012

A few things

- APSA deadline is December 15th.
- It will be held in Chicago.

A few Stata things

- How to increase **memory** to Stata:

```
set mem 500m, permanently
```

- What is the **reference category**, and why is it important?

A few Stata things

- How to set graph schemes for black and white:

```
query scheme
set scheme s2mono
```

- Search documentation for “schemes intro”

A few Stata things

- Here's how you generate dummies automagically in Stata:

- To generate year dummies:

```
xi i.year
```

- To generate country dummies:

```
xi: i.ccode
```

- Then when you want to include them in the model you type `*year` or `*ccode`

Today

- We are going to be talking about a particular type of model that looks at time in a different way than we did last week.

- This group of models all look at the length of time until something happens.

- These are known as:
  - Duration models
  - Hazard models
  - Event history models
  - Survival models

- We are going to call them *hazard models* because they are commonly motivated by one of the two motivations for these models.

- These models come from the biostatistics literature where researchers were trying to model the **likelihood of death** (the event of interest) given some treatment effect.

- Political research began modeling hazards using OLS.

Why not OLS?

- However, duration data have some characteristics that make OLS unsuitable.

- Like event counts, **duration must be greater than or equal to zero**.

- Survival at time *t* means that you have survived since *t-1*. This means that observations at time *t* are **conditional** on observations at time *t-1*.

- Some observations will survive the length of the study. These observations are considered **censored**.

- **Time-varying covariates** are not taken into account.

Examples

- Criminal recidivism
- How long someone is unemployed
- How long a civil war lasts
- How long a peace between rivals lasts

- Why could we just not run a logit model where all observations are considered 0 until death, which is coded 1?

- Because of the **conditionality** of the observations.

- Therefore what we need is to figure out the conditional probability of t_i given the fact that a unit survived to t_i - 1.

Two Types of Hazard Models

- Continuous
  - Failure can happen and be captured at *any* time.
- Discrete
  - Observations are captured within certain regular measures of time (days, months, years).
  - The data we have are likely to be discrete time data.

Discrete

- The dependent variables are binary.
  - 0 if the event does not occur at time *t*.
  - 1 if the event does not occur at time *t*.

- These data are considered Binary Time Series Cross Section (BTSCS) data.

Example of Discrete Data

dyad	*year*	*dispute*	*jio*	*deml*	*peaceyears*
2020	1966	0	53	10	15
2020	1967	0	54.2	10	16
2020	1968	0	55.4	10	17
2020	1969	0	56.6	10	18
2020	1970	0	57.8	10	19
2020	1971	0	59	10	20
2020	1972	0	59.43	10	21
2020	1973	0	59.86	10	22
2020	1974	1	60.29	10	23
2020	1975	1	60.71	10	0
2020	1976	0	61.14	10	0
2020	1977	0	61.57	10	1
2020	1978	0	62	10	2

- As you can see you can have variables like `jio` that vary over time.
- Couldn't we just include a time count and use probit or logit?
- e.g. peace lasts 20 years, and then war broke out.
- Berry and Berry (1990) use this approach.

```
. probit dispute  border caprat ally jio deml depl

Iteration 0:   log likelihood = -3667.7447
Iteration 1:   log likelihood = -3352.8902
Iteration 2:   log likelihood =  -3313.131
Iteration 3:   log likelihood = -3309.6175
Iteration 4:   log likelihood = -3309.5866
Iteration 5:   log likelihood = -3309.5866

Probit regression                                 Number of obs   =      20142
                                                  LR chi2(6)      =     716.32
                                                  Prob > chi2     =     0.0000
Log likelihood = -3309.5866                       Pseudo R2       =     0.0977

------------------------------------------------------------------------------
     dispute |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      border |   .6709295   .0387312    17.32   0.000     .5950176    .7468413
      caprat |  -.0010414   .0001413    -7.37   0.000    -.0013184   -.0007644
        ally |  -.2973491   .0398825    -7.46   0.000    -.3755174   -.2191808
         jio |   -.006198   .0013492    -4.59   0.000    -.0088424   -.0035537
        deml |  -.0165374   .0034323    -4.82   0.000    -.0232646   -.0098102
        depl |  -23.45946    5.82678    -4.03   0.000    -34.87974   -12.03918
       _cons |  -1.651092   .0526098   -31.38   0.000    -1.754205   -1.547979
------------------------------------------------------------------------------
Note: 34 failures and 0 successes completely determined.
```

- So we can see that being allies and being joint members of international organizations decrease the risk of a dispute.

- How can we include time in this model?

- Looking at the data earlier we could see that the effect of time that we are interested in is the string of zeros…how long a dyad has not had a dispute.

- This suggests that the longer the peace has lasted the less likely a dispute is.

- In other words, the probability of conflict at time *t* is *conditional* on the probability of conflict at time t-1.

```
. probit dispute  border caprat ally jio deml depl peaceyears

Iteration 0:   log likelihood = -3667.7447
Iteration 1:   log likelihood = -2901.2495
Iteration 2:   log likelihood = -2737.4396
Iteration 3:   log likelihood = -2731.3405
Iteration 4:   log likelihood = -2731.3184
Iteration 5:   log likelihood = -2731.3184

Probit regression                                 Number of obs   =      20142
                                                  LR chi2(7)      =    1872.85
                                                  Prob > chi2     =     0.0000
Log likelihood = -2731.3184                       Pseudo R2       =     0.2553

------------------------------------------------------------------------------
     dispute |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      border |   .3515146   .0437673     8.03   0.000     .2657322     .437297
      caprat |  -.0009783    .000149    -6.56   0.000    -.0012704   -.0006862
        ally |  -.2935278   .0444728    -6.60   0.000    -.3806929   -.2063628
         jio |   .0108222   .0015529     6.97   0.000     .0077785    .0138659
        deml |  -.0275271   .0037987    -7.25   0.000    -.0349724   -.0200818
        depl |  -19.61321   6.234162    -3.15   0.002    -31.83194   -7.394476
  peaceyears |  -.1001257   .0036322   -27.57   0.000    -.1072446   -.0930068
       _cons |  -1.348062   .0575867   -23.41   0.000     -1.46093   -1.235194
------------------------------------------------------------------------------
Note: 77 failures and 0 successes completely determined.
```

- As you can see time matters—conflict becomes less likely as the number of peaceful years increases.

- Since the baseline (the constant) is negative and `peaceyears` is negative, as time goes by the *hazard* of conflict gets smaller.

Splines

- We could also include splines, which are a more general smooth function of all time point dummies.
- You can create splines in Stata:

```
. btscs dispute year dyad, g(peaceyears) nspline(3)
```

- This command creates both the peace counter as well as three points (known as knots) along a smooth function of time (see Beck, Katz, and Tucker 1998 for a more detailed explanation of splines).

```
. probit dispute  border caprat ally jio deml depl peaceyears _spline*

Iteration 0:   log likelihood = -3667.7447
Iteration 1:   log likelihood = -2451.3166
Iteration 2:   log likelihood = -2356.8304
Iteration 3:   log likelihood =  -2351.521
Iteration 4:   log likelihood = -2351.3876
Iteration 5:   log likelihood = -2351.3874

Probit regression                                 Number of obs   =      20142
                                                  LR chi2(10)     =    2632.71
                                                  Prob > chi2     =     0.0000
Log likelihood = -2351.3874                       Pseudo R2       =     0.3589

------------------------------------------------------------------------------
     dispute |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      border |   .3369159   .0468629     7.19   0.000     .2450664    .4287654
      caprat |  -.0007732   .0001482    -5.22   0.000    -.0010638   -.0004827
        ally |  -.2092891   .0471655    -4.44   0.000    -.3017319   -.1168464
         jio |   .0069146   .0016371     4.22   0.000     .0037059    .0101234
        deml |  -.0250491   .0040983    -6.11   0.000    -.0330816   -.0170165
        depl |  -15.20443   5.947518    -2.56   0.011    -26.86135   -3.547511
  peaceyears |  -.5315939   .0200018   -26.58   0.000    -.5707968    -.492391
    _spline1 |  -.0098482   .0006208   -15.86   0.000     -.011065   -.0086313
    _spline2 |   .0056757    .000479    11.85   0.000     .0047369    .0066146
    _spline3 |  -.0016084   .0002206    -7.29   0.000    -.0020408   -.0011761
       _cons |  -.5949071   .0663745    -8.96   0.000    -.7249988   -.4648154
```

- As you can see that time has a non-monotonic effect—the splines have significant effects but in different directions.

- Early periods of peace (`_spline 1`) shift the baseline hazard down while the middle period (`_spline 2`) shifts it upward.

- In essence looking at both the peace years and the splines it would appear that the chance that the dyad *fails* (has a dispute) decreases for every year that the dyad *survives* (stays at peace).

- We also *could* aggregate these data to just one row with the number of time units in which the event occurs.

- This would then necessitate *continuous* time models.

Continuous

- This only has one observation for each individual indicating the time the event happened.
- There are two variations dependent upon whether the independent variables vary over time (TVC) or not (NTVC).

Time Varying Covariates (TVC)

dyad	year	dispute	jio	deml	peaceyears
2020	1966	0	53	10	15
2020	1967	0	54.2	10	16
2020	1968	0	55.4	10	17
2020	1969	0	56.6	10	18
2020	1970	0	57.8	10	19
2020	1971	0	59	10	20
2020	1972	0	59.43	10	21
2020	1973	0	59.86	10	22
2020	1974	1	60.29	10	23
2020	1975	1	60.71	10	0
2020	1976	0	61.14	10	0
2020	1977	0	61.57	10	1
2020	1978	0	62	10	2

Non-Time Varying Covariates (NTVC)

dyad	*dispute*	*jio*	*deml*	*peaceyears*
2020	1	53	10	4
2091	1	21.14	-9	16
345710	0	.	-3	8
485623	1	13.88	-7	65
2020	0	69.17	4	23
2020	1	59	2	1

- As we can see from the dispute data, the time at peace is
  - strictly *positive*.
  - Is *reset* after a conflict, so there can be more than one dispute per dyad (multiple events).
  - And can *vary* with time-varying covariates.

- While the time counter and splines can capture the conditional effects of time, there is a whole class of models that more specifically model the hazard of events.

- As Box-Steffensmeier and Jones (2004) mention there are a number of models where you can explicitly assume a distribution for the hazard rate.

- I am going to go over the logic of event history before delving into the *parametric* models—models where we can estimate the hazard given a number of covariates and assumptions about the hazard distribution.

- Let’s begin by defining T as a positive random variable measuring survival time.
- We assume that T is continuous.
- What we observe is a value of T, called t.
- The possible values of T have a probability distribution characterized by a PDF: $f(t)$ and a CDF, $F(t)$.

- $f(t)$ can be considered as the estimate of the instantaneous probability that the event (a dispute) occurs.

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t}$$

- Thus as Δt gets infinitesimally small you get an instantaneous estimate of the probability of failure at time t.

- We are also interested in the survivor function, $S(t)$—the probability of surviving at time t.

- As you might guess the chances of surviving past t are related to the chance of dying at t.

$$S(t) = 1 - F(t) = P(T \geq t)$$

- Therefore the failure and the survival rates are related to each other.

- This relation is given by the hazard rate.

$$h(t) = \frac{f(t)}{S(t)}$$

- In words, the hazard rate is the conditional failure rate—the rate that units fail by *t* given that they have survived until *t*.

- The hazard rate, the survivor function, and the density functions are *mathematically linked* so that if you can specify one then you can determine the others.

- The hazard rate is what the literature (and we) are going to be focused on for the next week.

- The hazard rate can vary from 0 to infinity (and beyond!).
- When the risk of something is zero (me being declared president of the universe by Gary King), the hazard is zero.
- When the hazard rate nears infinity, it means the certainty of failure in that instant.
- Thus the hazard rate is not limited to a 0—1 range like probability.

- The (continuous) cumulative hazard function is given by:

$$H(t) = \int_0^t h(u)du$$

Which can also be (and more often is) seen as:

$$H(t) = -\ln\{S(t)\}$$

- As you might be able to guess, there are several issues we need to address when trying to model the hazard rate given the data that we have.

- For example, the democratic peace data ranged from 1951 to 1985.

- We lack information about what happened *before* 1951 and *after* 1985.

Censoring

- Censoring happens when the full event history of a unit is unobserved.
  - In political science we are likely to observe right censoring.
  - For example, two dyads (USA-Ecuador and China-South Korea) had disputes in 1971.
  - The peace year count would both start over in 1972.
  - But what if the US and Ecuador had a conflict in 1986 but China and S. Korea did not?
  - Both would have the same duration time, but the observations are clearly different.
- Also possible to have left censoring—US-Ecuador had a dispute in 1965 that was not coded.

Censoring

- Right-censored: _|______________|_X_ time

- Left-censored: _X_|______________|_ time

Truncation

- Now, the dataset also *begins* at a set time, so we lack information about what happened *prior* to the data.

- The time (history) before 1951 in the dispute data would therefore be *left-truncated*.

- Truncation means that the unit could have failed before we started measuring meaning that we never would have measured the unit.

  - A smoker dies before a study would be truncated from the study.

- In estimating the likelihood of the sampled duration times, you can account for right-censoring and left-truncation.

- The likelihood of observing the sample that we have is determined by t_i *, which represents the *i*th censored case equal to the last observed period even though *i* survives past *t**.

- If the case is uncensored then $t_i \leq t^*$.

- We can create a censor indicator δ_i that codes censored cases.

$$\delta_i = \begin{cases} 1 \; if \; t_i \le t^* \\ 0 \; if \, t_i > t^* \end{cases}$$

- Therefore when $\delta_i = 1$ the observation is uncensored and $\delta_i = 0$ it is censored.

- Given what we know about right-censoring we can now specify the likelihood of observing our duration data:

$$L = \prod_{i=1}^{n} \{f(t_i)\}^{\delta_i} \{S(t_i)\}^{1-\delta_i}$$

- But before we start with semi-parametric and parametric models (and down the rabbit hole), what can we learn through non-parametric means?

- And what on God's green earth are these terms?

Nonparametric, semi-parametric, & parametric

- What are the differences between them?

- **Nonparametric**—no x's, no distribution, lets the data "speak"

- **Semi-parametric**—probability of failure given x's but no assumptions about distribution of the error, ε. At a specific failure time (0,1)

- **Parametric**—assumes error distribution and probability given covariates. At all possible failure times.

- Let's use some actual data.
- I am going to use data on the duration of civil war from Collier, Hoeffler, and Soderbom (2004).

Stset

- First, you have to tell Stata about the structure of your data.

- It helps minimize error by having to specify these options with every command you run.

```
stset time_of_failure (or censoring)_var, ///
failure(one_if_failure_var)
```

Stset creates four variables

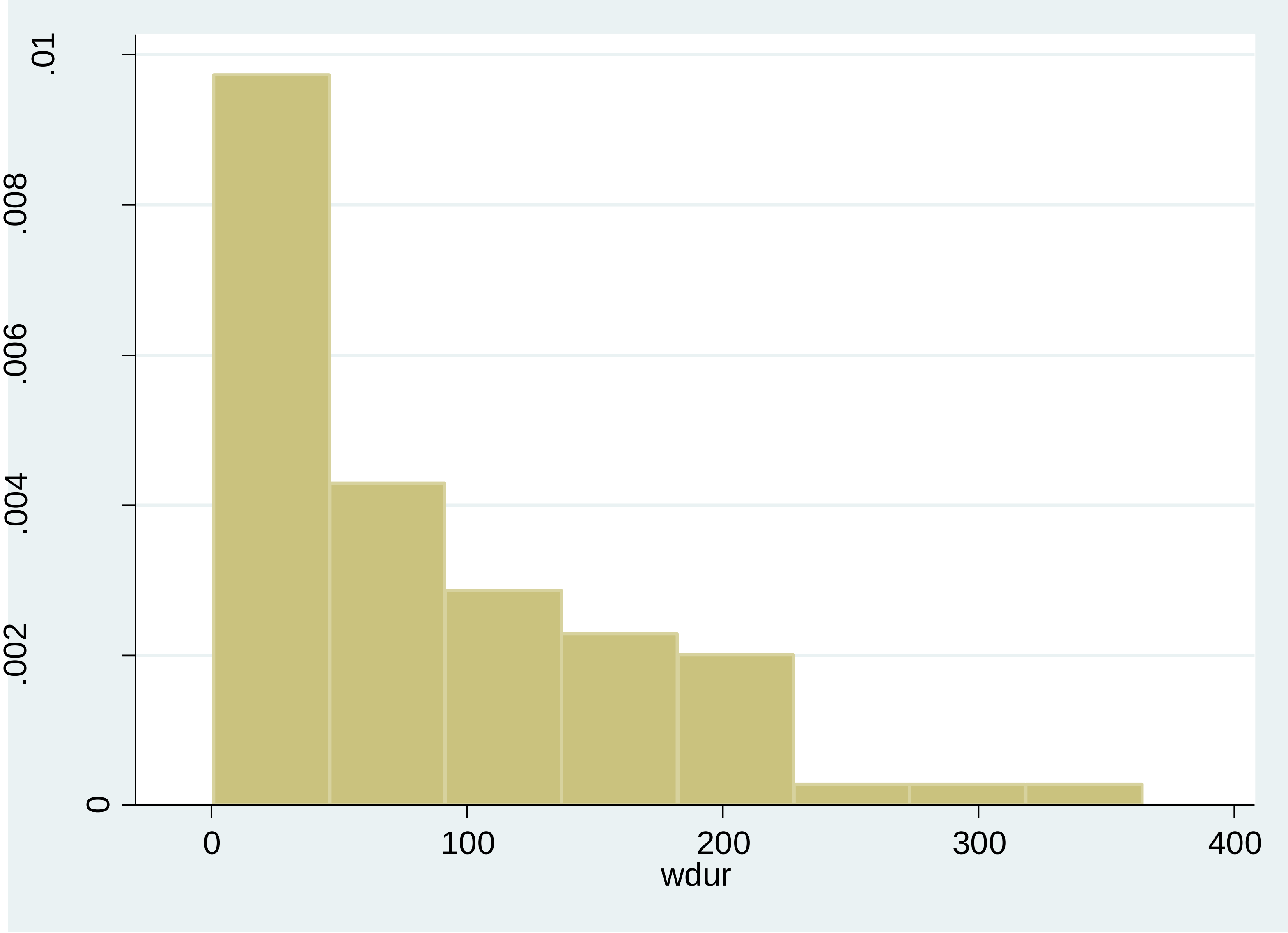
- `_t0` and `_t` = timespan starts at `_t0` and ends at `_t`
  - In these data _t0 is January 1960
- `_d` = outcome at end of time span
  - In these data it is 1 if the war ended before 1999, 0 otherwise.
- `_st` = 1 if the observation will be used in the analysis, 0 otherwise.
  - In these data since there are no civil wars that started before 1960, _st always ==1

```
. sts list

         failure _d:  cens
   analysis time _t:  mo
                 id:  indsp
```

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf.	Int.]
1	55	4	0	0.9273	0.0350	0.8177	0.9721
2	51	3	0	0.8727	0.0449	0.7515	0.9372
4	48	1	0	0.8545	0.0475	0.7301	0.9245
6	47	2	0	0.8182	0.0520	0.6884	0.8978
10	45	2	0	0.7818	0.0557	0.6479	0.8697
12	43	1	0	0.7636	0.0573	0.6280	0.8553
13	42	2	0	0.7273	0.0601	0.5890	0.8257
16	40	1	0	0.7091	0.0612	0.5698	0.8105
21	39	1	0	0.6909	0.0623	0.5508	0.7951
22	38	1	0	0.6727	0.0633	0.5320	0.7796
27	37	1	0	0.6545	0.0641	0.5134	0.7638
36	36	1	0	0.6364	0.0649	0.4950	0.7479
45	35	1	0	0.6182	0.0655	0.4768	0.7318
46	34	1	0	0.6000	0.0661	0.4587	0.7155
49	33	1	0	0.5818	0.0665	0.4408	0.6990
55	32	1	0	0.5636	0.0669	0.4231	0.6824
63	31	1	0	0.5455	0.0671	0.4056	0.6656
64	30	1	0	0.5273	0.0673	0.3882	0.6486
69	29	1	0	0.5091	0.0674	0.3710	0.6315
73	28	1	0	0.4909	0.0674	0.3539	0.6142
74	27	2	0	0.4545	0.0671	0.3204	0.5792
85	25	1	0	0.4364	0.0669	0.3038	0.5614
88	24	1	1	0.4182	0.0665	0.2875	0.5435
91	22	1	0	0.3992	0.0661	0.2704	0.5248
98	21	1	1	0.3802	0.0657	0.2535	0.5059
102	19	1	1	0.3602	0.0652	0.2356	0.4860
104	17	0	1	0.3602	0.0652	0.2356	0.4860
121	16	1	0	0.3376	0.0649	0.2152	0.4642
129	15	1	0	0.3151	0.0643	0.1953	0.4420
134	14	1	0	0.2926	0.0636	0.1759	0.4194
143	13	1	0	0.2701	0.0625	0.1570	0.3964
145	12	1	0	0.2476	0.0612	0.1387	0.3729
148	11	1	0	0.2251	0.0597	0.1209	0.3491
155	10	1	0	0.2026	0.0578	0.1037	0.3247
170	9	1	0	0.1801	0.0556	0.0872	0.2998
172	8	1	0	0.1576	0.0530	0.0714	0.2743
189	7	0	1	0.1576	0.0530	0.0714	0.2743
196	6	1	0	0.1313	0.0503	0.0530	0.2458
198	5	0	2	0.1313	0.0503	0.0530	0.2458
203	3	1	0	0.0875	0.0490	0.0219	0.2117
286	2	1	0	0.0438	0.0395	0.0041	0.1689
364	1	1	0	0.0000	.	.	.

- One of the simplest ways at looking at what the survivor function might look like is to histogram the length of the 55 conflicts in the estimate data.

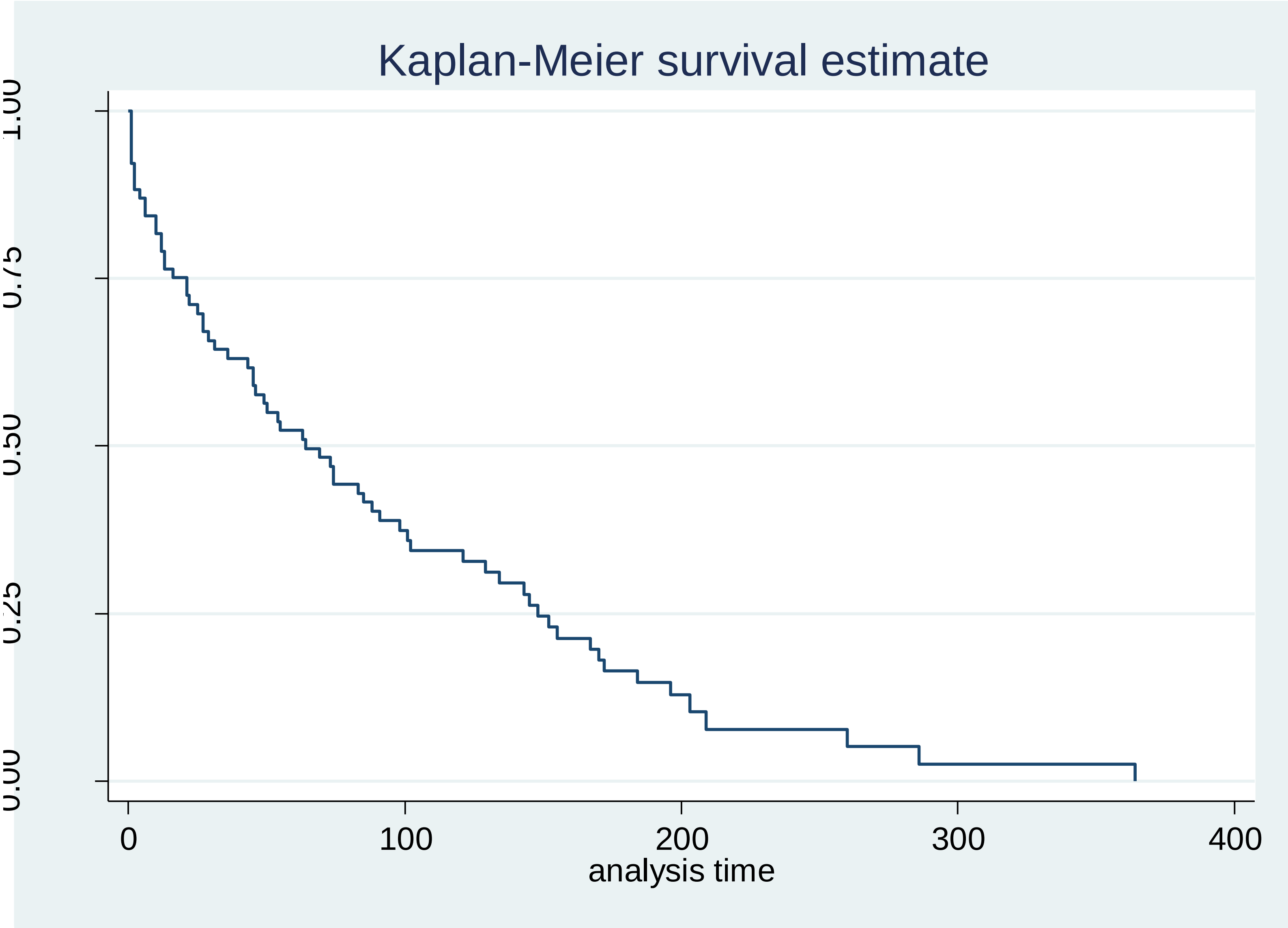
Histogram of survival



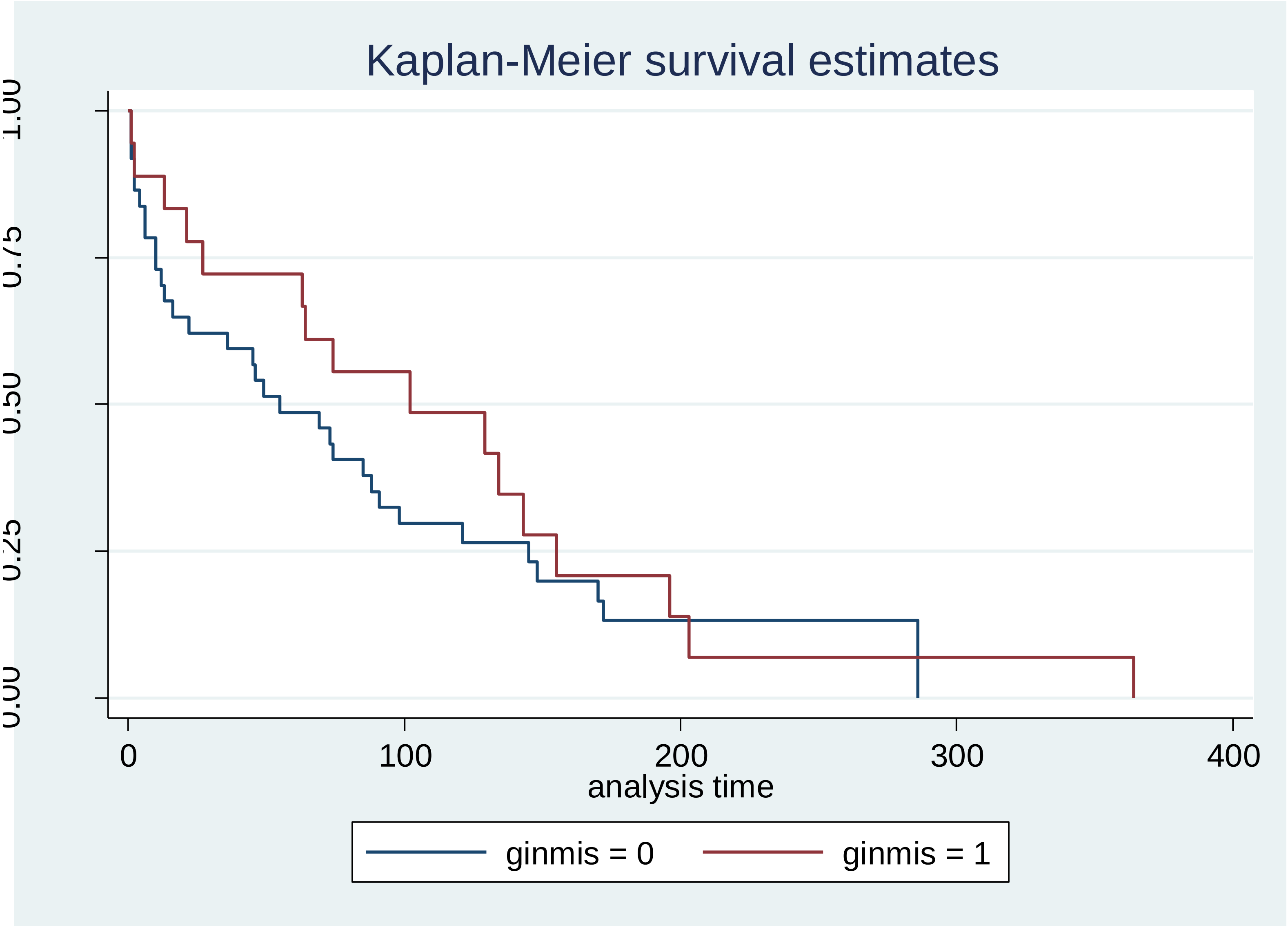
- The simplest means of survival analysis is the Kaplan-Meier estimator.
  - It is a non-parametric estimate of the survival function:

$$\hat{S}(t) = \prod_{j \,|\, t_j \leq t} \left(\frac{n_j - d_j}{n_j} \right)$$

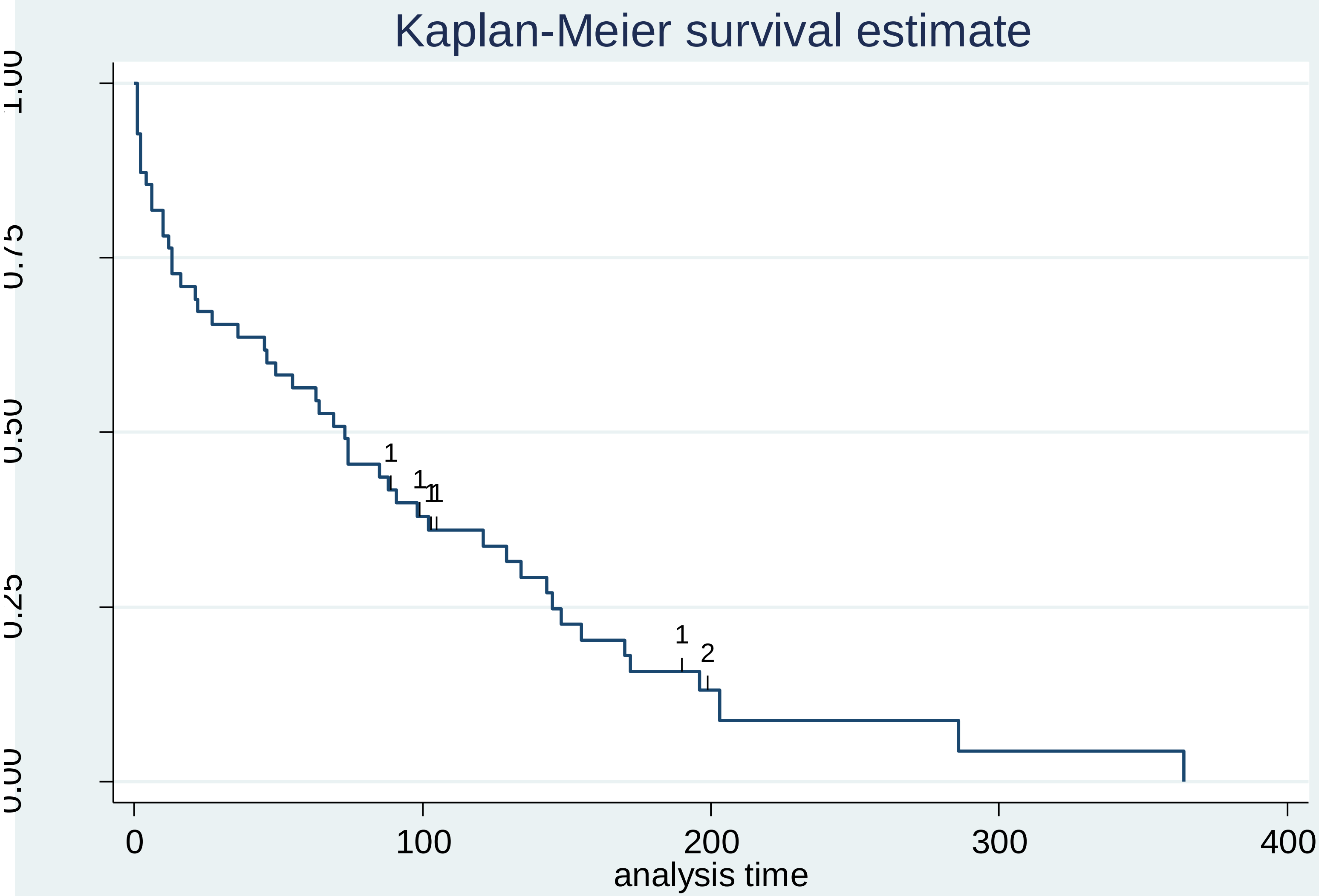
Where n_j is the number of individuals at risk at time t_j, and d_j is number of failures at time t_j


Kaplan-Meier survival estimate
1.00
0.75
0.50
0.25
0.00
0
100
200
300
400
analysis time

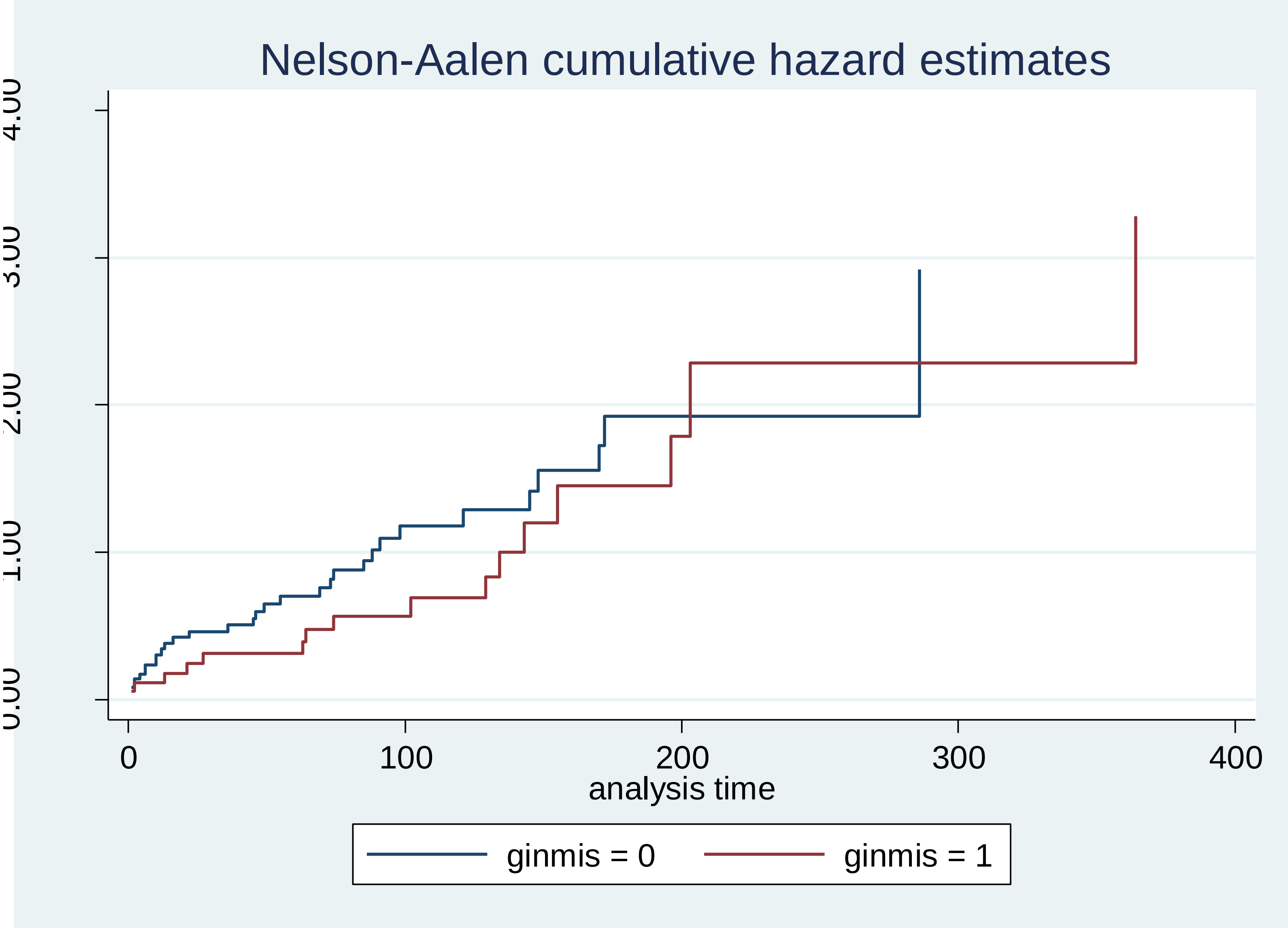
Can also graph by dichotomous variables

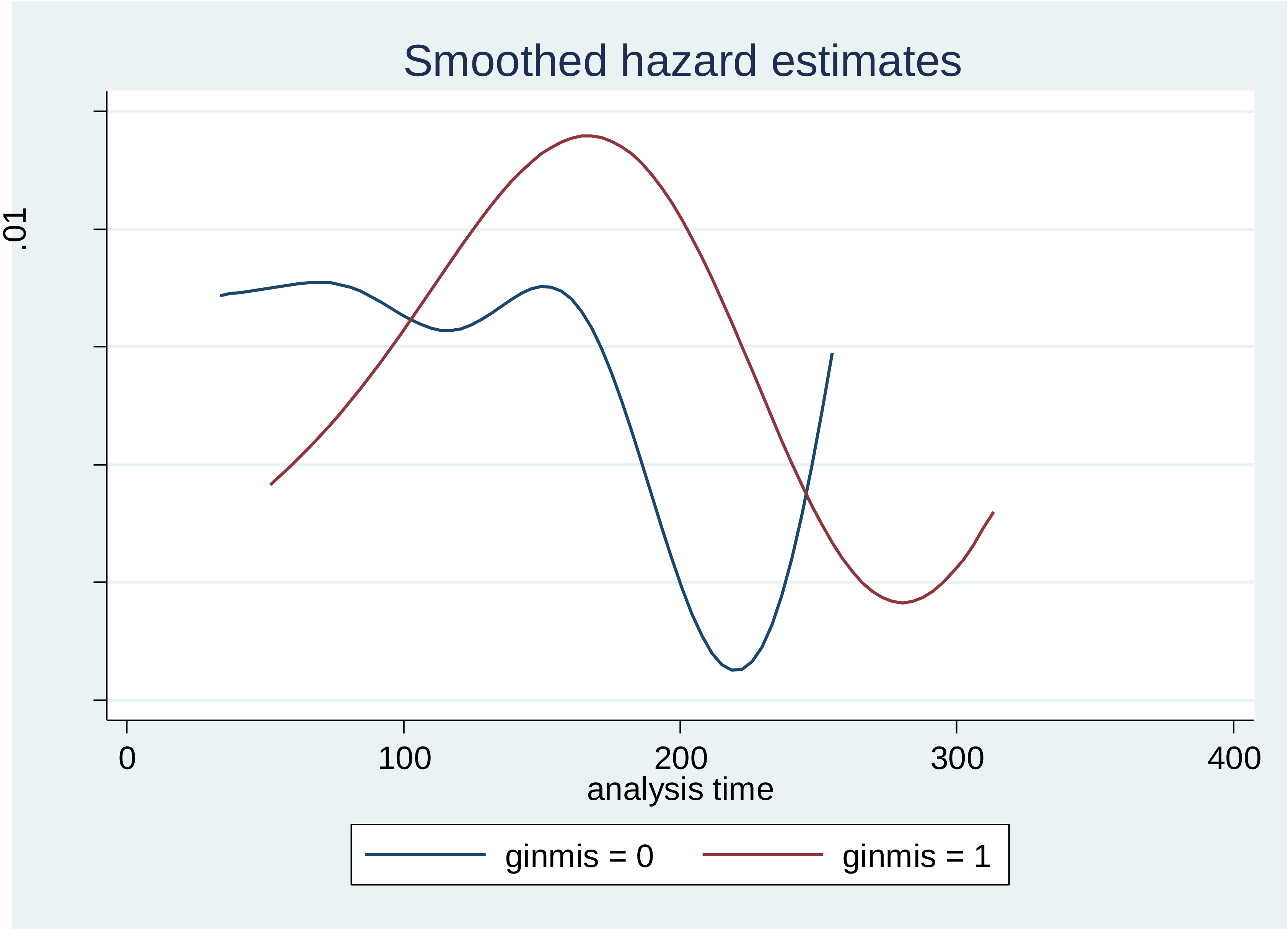


And whether units were censored



And the cumulative hazard





Methods of smoothing will be discussed next week.

Stata commands

```
** List Kaplan-Meier survival estimates **
sts list
sts graph

*By a dichotomous variable **
sts graph, by(ginmis)

** Showing where censored obs left data **
sts graph, censored(number)

** Cumulative Hazard **
sts graph, by(ginmis)   cumhaz

** Smoothed Hazard Function **
sts graph, hazard by(ginmis) kernel(gaussian)
```

- So now, how do we create a model that allows us to capture both the occurrence (or non-occurrence) of an event (death) as well as how long the unit lasted (lived) before the event?

- We move to looking at parametric models.

- We are starting from models that assume a distribution of the hazard rate.

- Next week, we will look at semi-parametric models where we do not specify the hazard distribution.

Exponential model

- The easiest model where we assume that the hazard rate is constant (flat) across time.

$$h(t) = \lambda \text{ where } \lambda > 0 \text{ and } t > 0$$

- Specifying the hazard rate allows us to determine the survival and density functions:

$$S(t) = e^{-\lambda(t)}$$

$$f(t) = \lambda(t)e^{-\lambda(t)}$$

- We can now parameterize a model to estimate what the expected duration time of observation *i*:

$$\mathrm{E}(t_i) = e^{\boldsymbol{\beta X}}$$

and parameterize the hazard rate:

$$h(t \mid \mathrm{x}) = e^{-(\boldsymbol{\beta X})}$$

- As we know from previous classes:

$$\boldsymbol{\beta X} = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_j x_{ij})$$

- This allows us to show an important characteristic of the exponential model:

$$h(t \mid \mathrm{x}) = e^{-(\beta_o)} \, e^{-(\boldsymbol{\beta X})}$$

- This shows that the *baseline hazard rate* is given by βo.

Proportional hazards property

- Changes to the baseline hazard rate in the exponential model are in multiples of the baseline hazard.

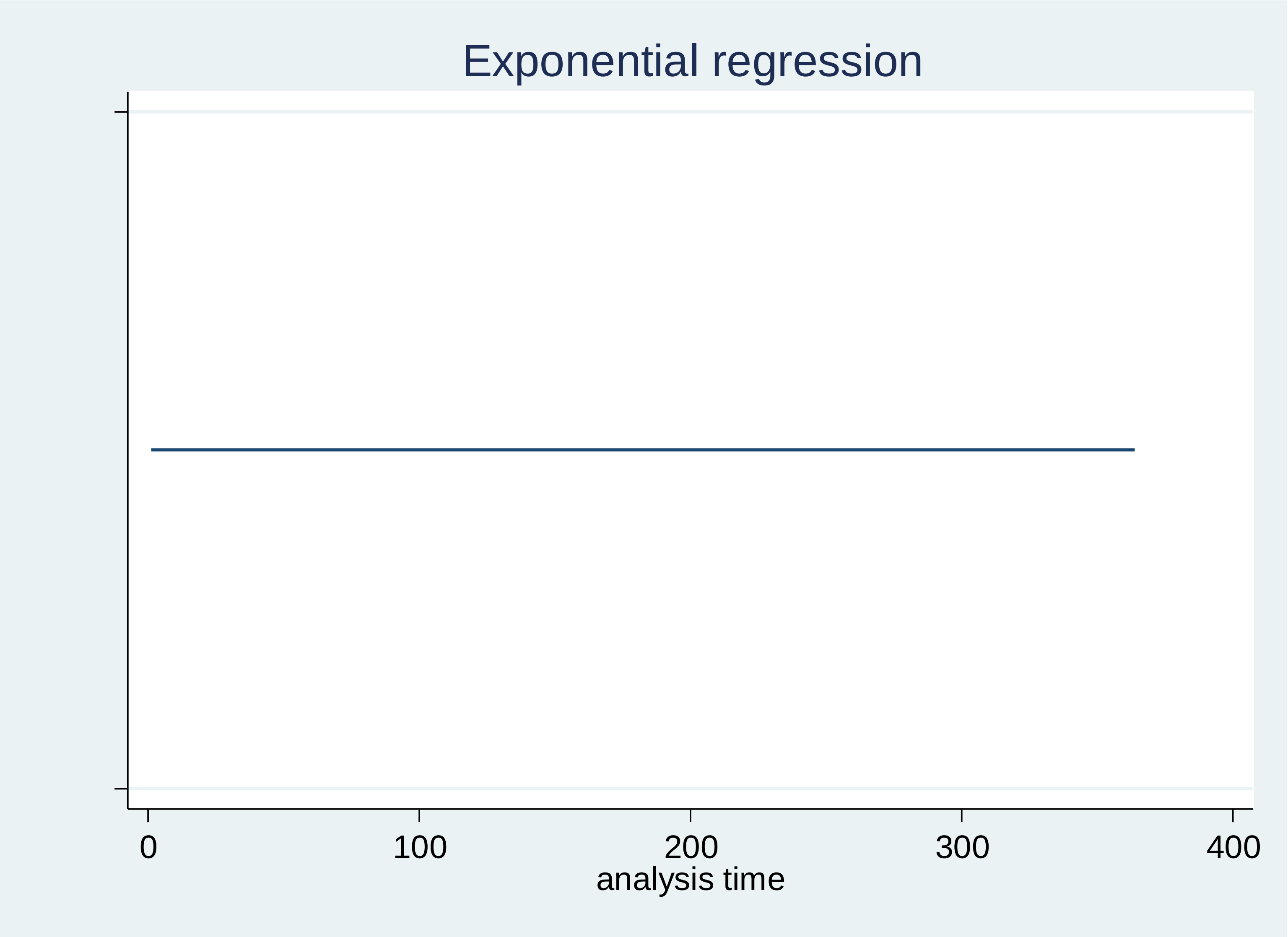
$$\frac{h_i(t\,|x_1{=}1)}{h_i(t\,|x_1{=}0)} = e^{-\beta_1}$$

- The exponential distribution is known as a "memoryless" because the distribution of the survival time is not affected by knowing how long the unit has survived.

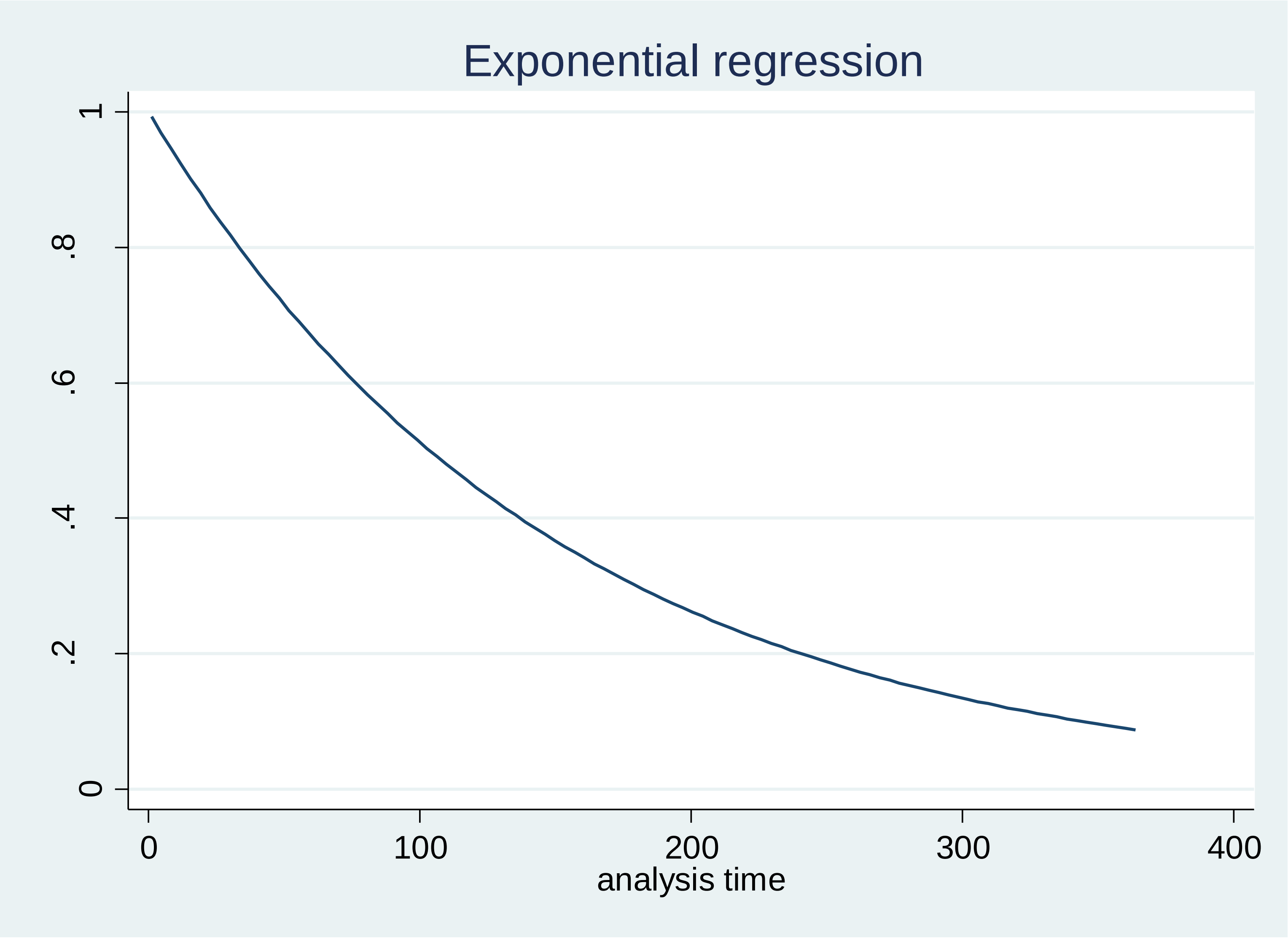
Collier et al. (2004), Exponential regression

```
. streg gini_m ginmis rgdpch elf elf2 logpop y70stv y80stv y90stv d2-d4,
dist(exponential) nohr
         failure _d:  cens
   analysis time _t:  mo
                 id:  indsp

Iteration 0:   log likelihood = -101.63735
Iteration 1:   log likelihood = -90.265345
Iteration 2:   log likelihood = -80.526611
Iteration 3:   log likelihood = -80.430147
Iteration 4:   log likelihood = -80.429995
Iteration 5:   log likelihood = -80.429995

Exponential regression -- log relative-hazard form

No. of subjects =           55                     Number of obs   =      4625
No. of failures =           48
Time at risk    =         4625
                                                   LR chi2(12)     =     42.41
Log likelihood  =   -80.429995                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gini_m |  -.1244463   .0284179    -4.38   0.000    -.1801444   -.0687482
      ginmis |  -5.867928   1.277403    -4.59   0.000    -8.371591   -3.364265
      rgdpch |   .3651031   .1322248     2.76   0.006     .1059472     .624259
         elf |  -.0628267   .0258742    -2.43   0.015    -.1135392   -.0121143
        elf2 |   .0581252   .0270411     2.15   0.032     .0051256    .1111247
      logpop |  -.3163905   .1230657    -2.57   0.010    -.5575948   -.0751863
      y70stv |   .0077905   .4625409     0.02   0.987    -.8987729    .9143539
      y80stv |  -1.420202   .5203341    -2.73   0.006    -2.440038   -.4003656
      y90stv |  -1.162059   .5416506    -2.15   0.032    -2.223675   -.1004433
          d2 |  -.8067415   .5742936    -1.40   0.160    -1.932336    .3188533
          d3 |  -.0010657   .5606172    -0.00   0.998    -1.099855    1.097724
          d4 |   .6098389   .4464024     1.37   0.172    -.2650937    1.484771
       _cons |   7.433105   2.707863     2.75   0.006     2.125791    12.74042
------------------------------------------------------------------------------
```

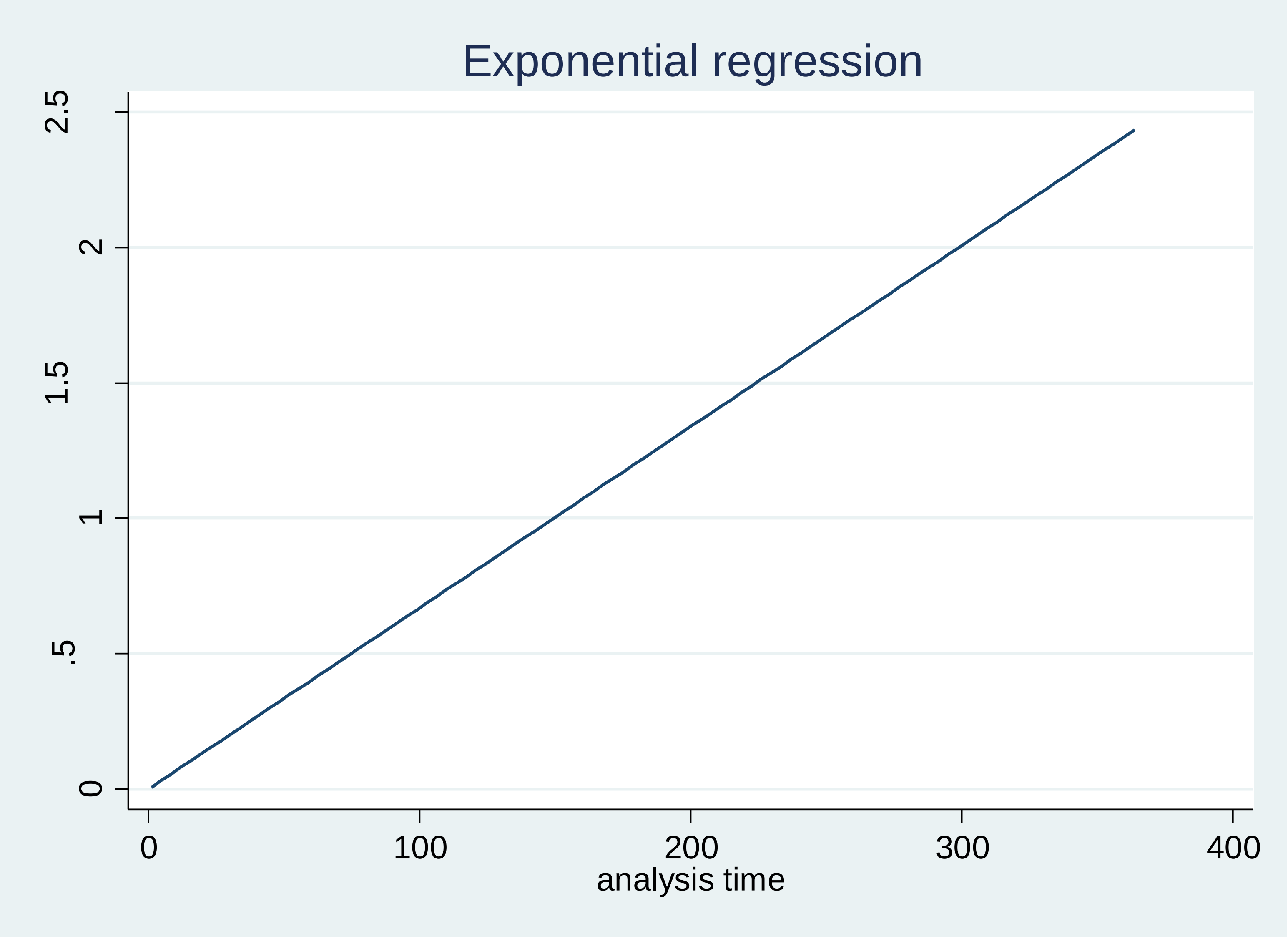
Exponential Hazard Function



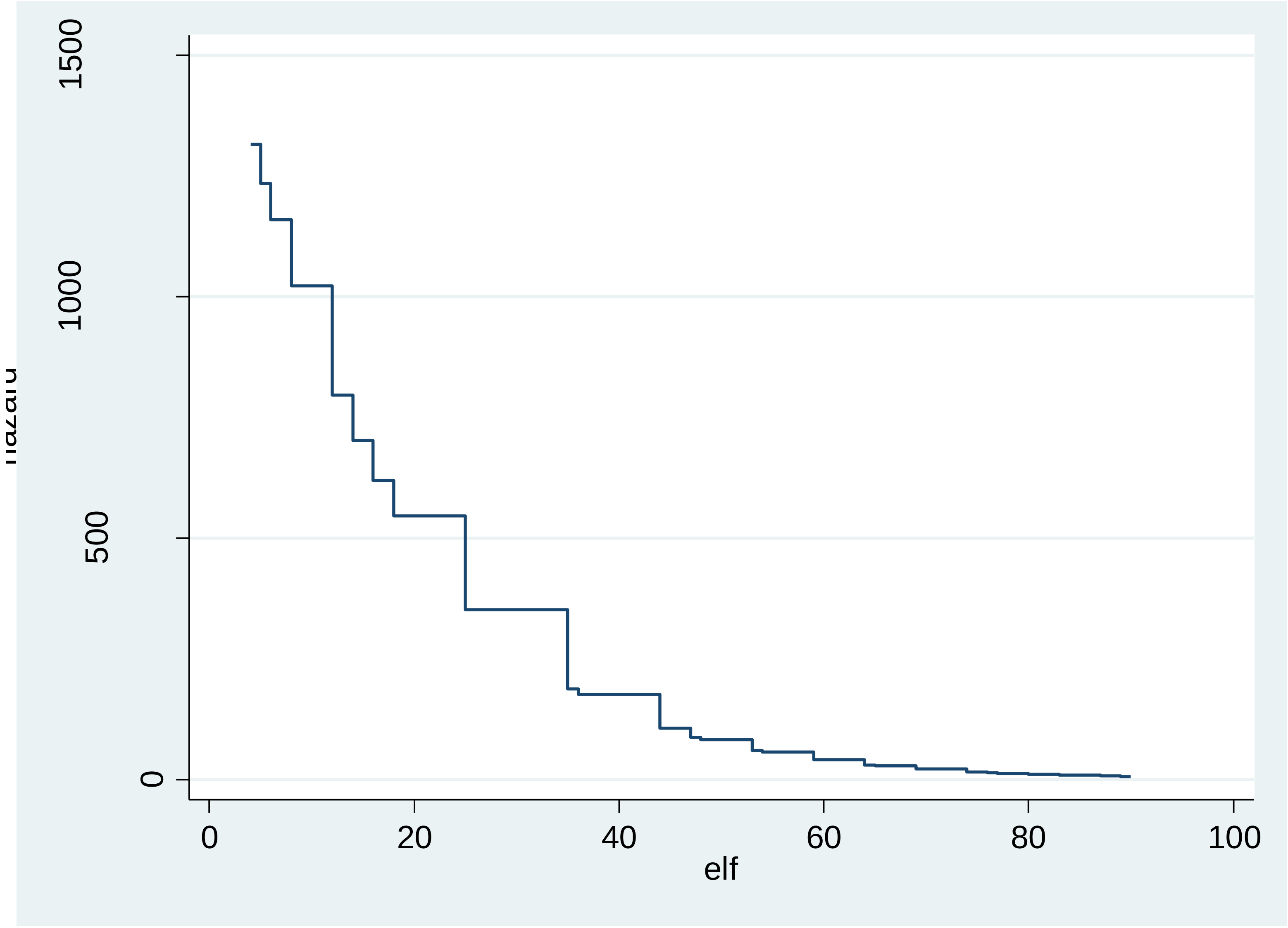
Survival function



Cumulative hazard



Over the range of ELF



Weibull

- Looking for a more flexible alternative, most turn to the Weibull, a distribution often seen in political science.

- It's defining characteristic is that the baseline hazard rate is monotonic—it can be always increasing, always decreasing, or flat.

Weibull hazard rate distribution

$$h(t) = \lambda p(\lambda t)^{p-1} \text{ where } t > 0, \lambda > 0, p > 0$$

- p is the **shape** parameter.
  - When p >1, the hazard is monotonically *increasing*.
  - When $p < 1$, the hazard is monotonically *decreasing*.
  - When $p = 1$, the hazard is flat at value λ (therefore the exponential is nested in the Weibull).
- λ is the **scale** parameter.

- You can then parameterize the hazard rate to model the effect of some x's.

$$h(t\,|x) = pt^{p-1}e^{(\beta_j x)}$$

- We can now maximize a log-likelihood equation based on the one we saw above:

$$L = \prod_{i=1}^{n}\{f(t_i)\}^{\delta_i}\{S(t_i)\}^{1-\delta_i}$$

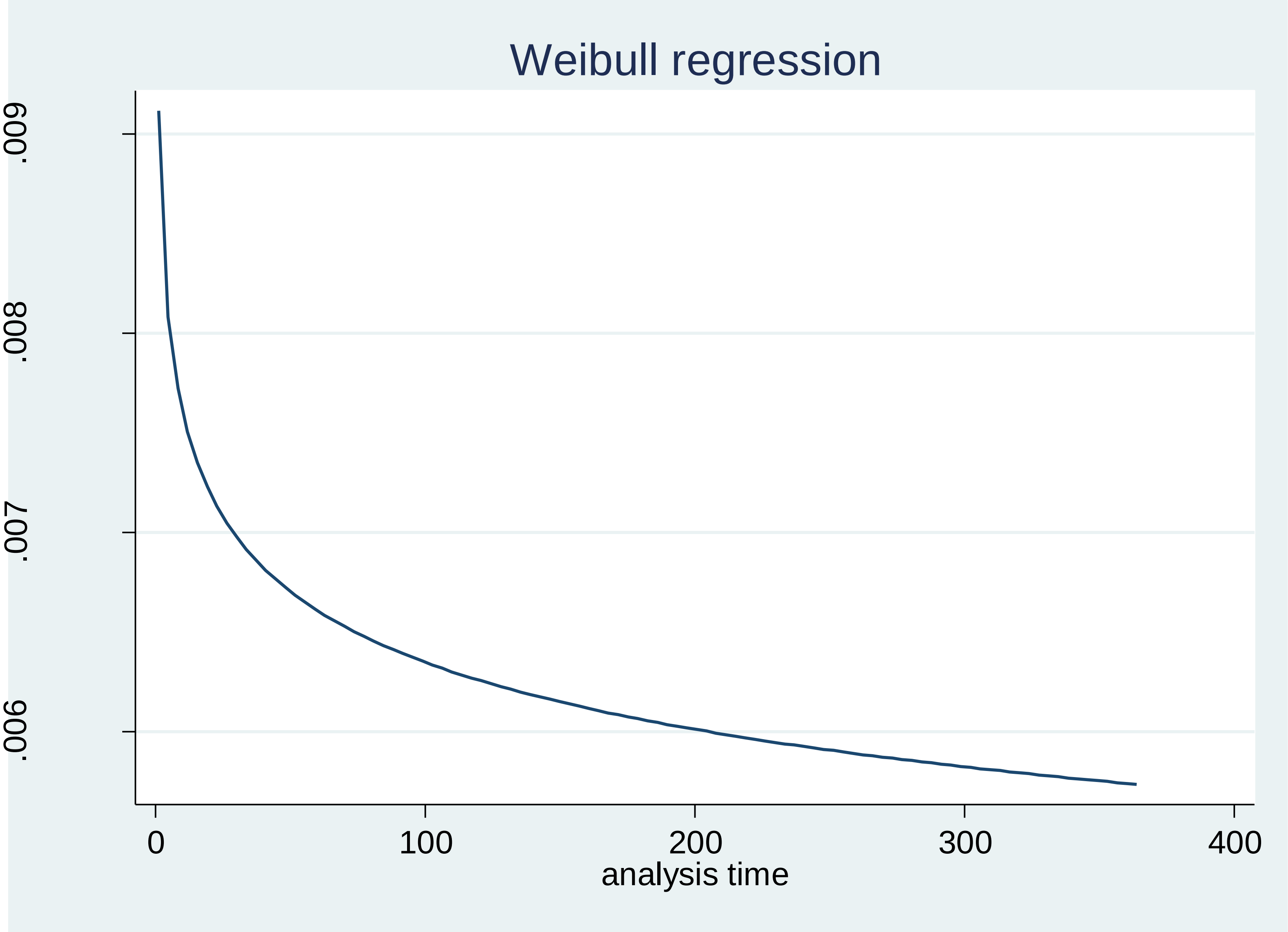
$$L(\,t\,|\,\lambda,p) = \prod_{i=1}^{n}\left\{\lambda p(\lambda t)^{p-1}e^{-(\lambda t)^p}\right\}^{\delta_i}\left\{e^{-(\lambda t)^p}\right\}^{1-\delta_i}$$

- Once you estimate the model, you can use the estimated shape parameter (p) to test whether the hazard is actually flat—e.g. the observations are duration independent.

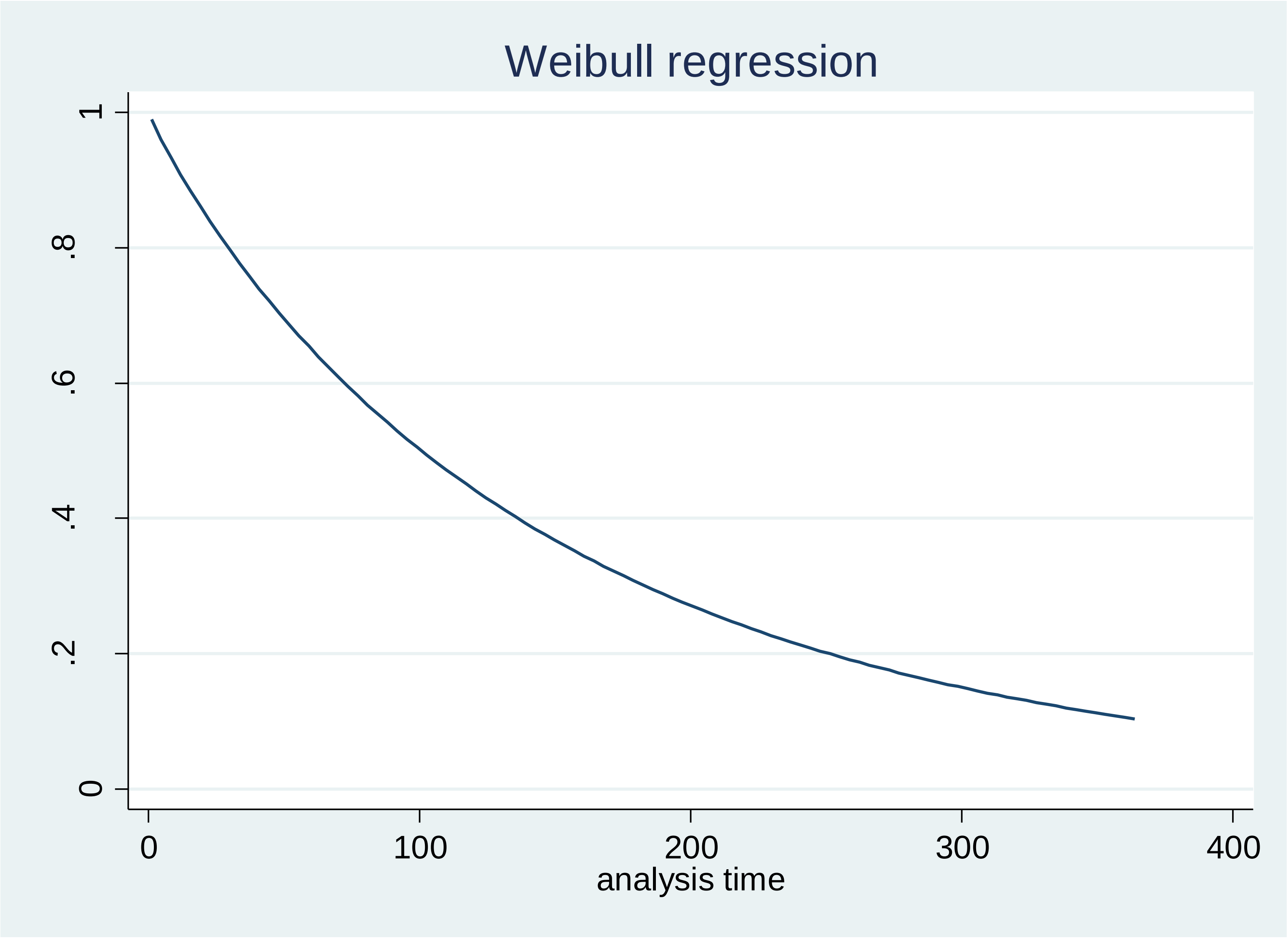
$$z = \frac{p-1}{se(p)}$$

```
. streg gini_m ginmis rgdpch elf elf2 logpop y70stv y80stv y90stv ///
>  d2-d4, dist(weibull) nohr nolog

         failure _d:  cens
   analysis time _t:  mo
                 id:  indsp

Weibull regression -- log relative-hazard form
No. of subjects =           55                   Number of obs   =       4625
No. of failures =           48
Time at risk    =         4625
                                                 LR chi2(12)     =      37.65
Log likelihood  =   -80.341859                   Prob > chi2     =     0.0002
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gini_m |  -.1221709   .0288244    -4.24   0.000    -.1786657    -.065676
      ginmis |  -5.746803   1.302304    -4.41   0.000    -8.299271   -3.194334
      rgdpch |   .3579029   .1331665     2.69   0.007     .0969013    .6189045
         elf |  -.0615579    .025954    -2.37   0.018    -.1124269   -.0106889
        elf2 |   .0572613   .0270164     2.12   0.034     .0043101    .1102125
      logpop |  -.3092118   .1241655    -2.49   0.013    -.5525717   -.0658519
      y70stv |   .0223796   .4641904     0.05   0.962     -.887417    .9321761
      y80stv |  -1.384908   .5263604    -2.63   0.009    -2.416556    -.353261
      y90stv |  -1.108178   .5550966    -2.00   0.046    -2.196148   -.0202086
          d2 |  -.6810668   .6493246    -1.05   0.294     -1.95372    .5915859
          d3 |   .1526106   .6702459     0.23   0.820    -1.161047    1.466268
          d4 |   .8091091   .6495045     1.25   0.213    -.4638963    2.082115
       _cons |   7.402131   2.691839     2.75   0.006     2.126223    12.67804
-------------+----------------------------------------------------------------
       /ln_p |  -.0818573   .1986233    -0.41   0.680    -.4711518    .3074373
-------------+----------------------------------------------------------------
           p |   .9214035   .1830122                      .6242828    1.359936
         1/p |   1.085301   .2155661                       .735329    1.601838
------------------------------------------------------------------------------
```

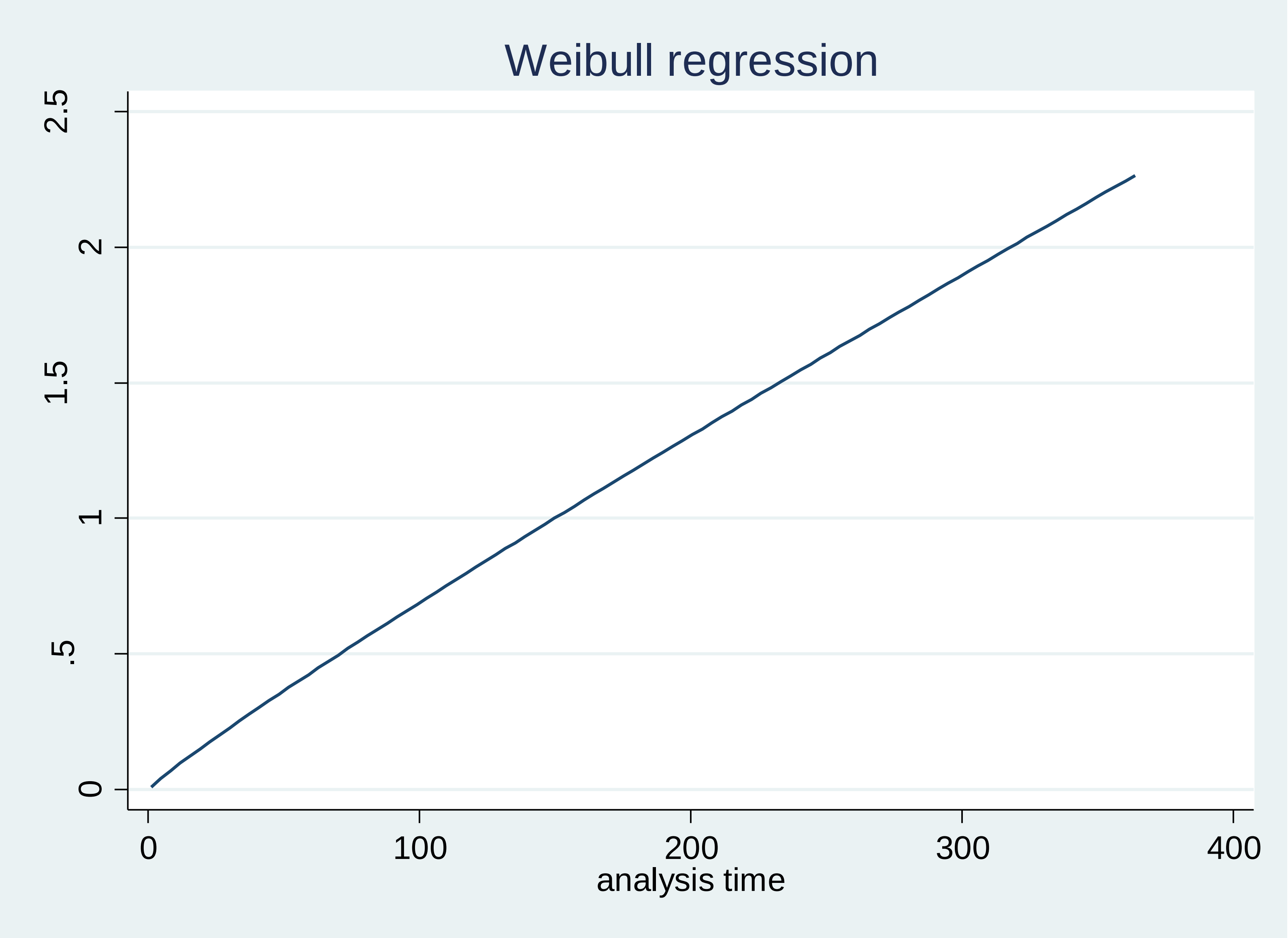
Hazard Function



Survival



Cumulative Hazard



Gompertz

- The Gompertz distribution is another popular distribution.

- It differs from the Weibull in that the hazard rate is considered an exponential function of duration times.

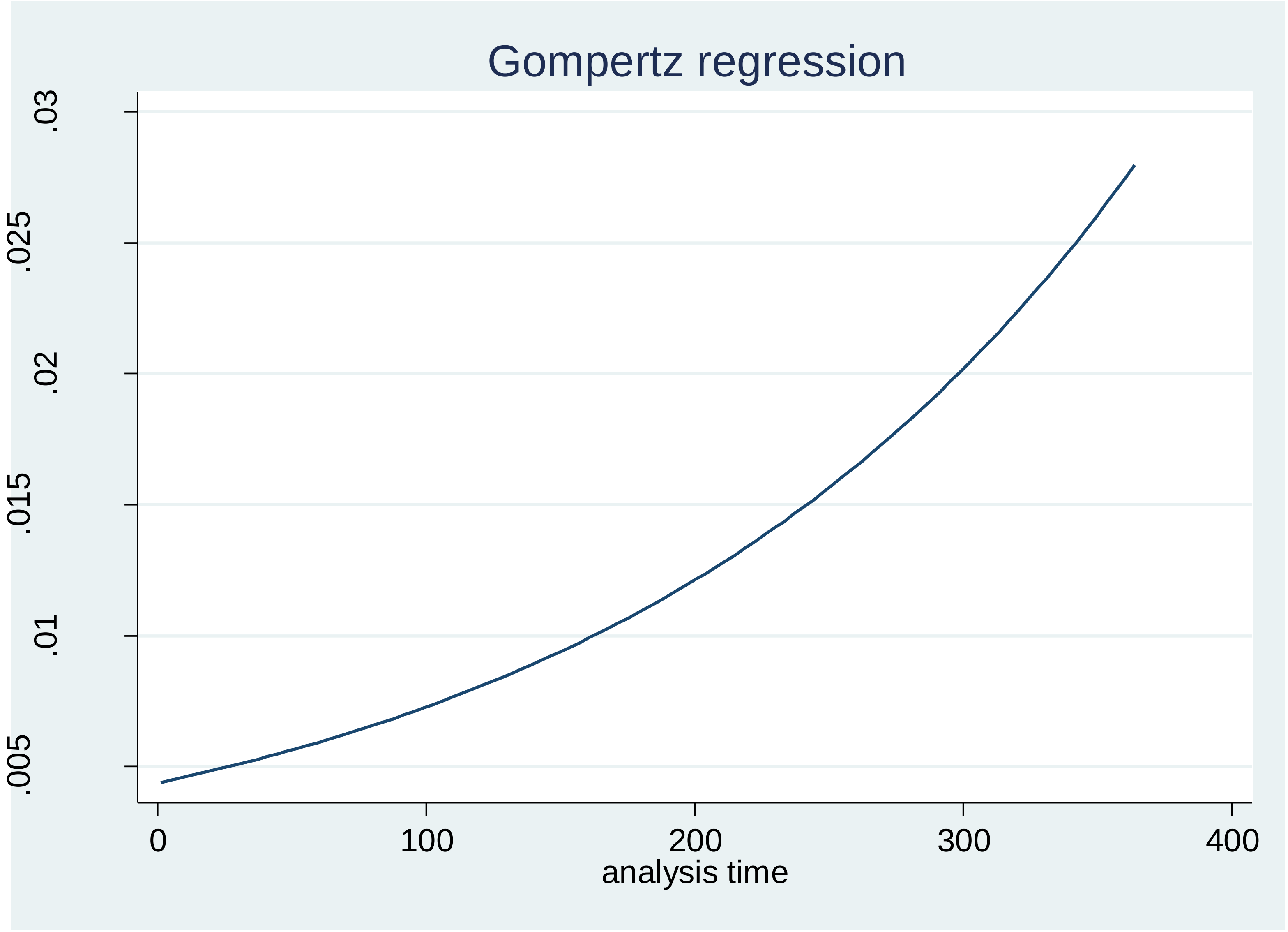
- It too is monotonic.

$$h(t) = e^{\gamma t} e^{\lambda}$$

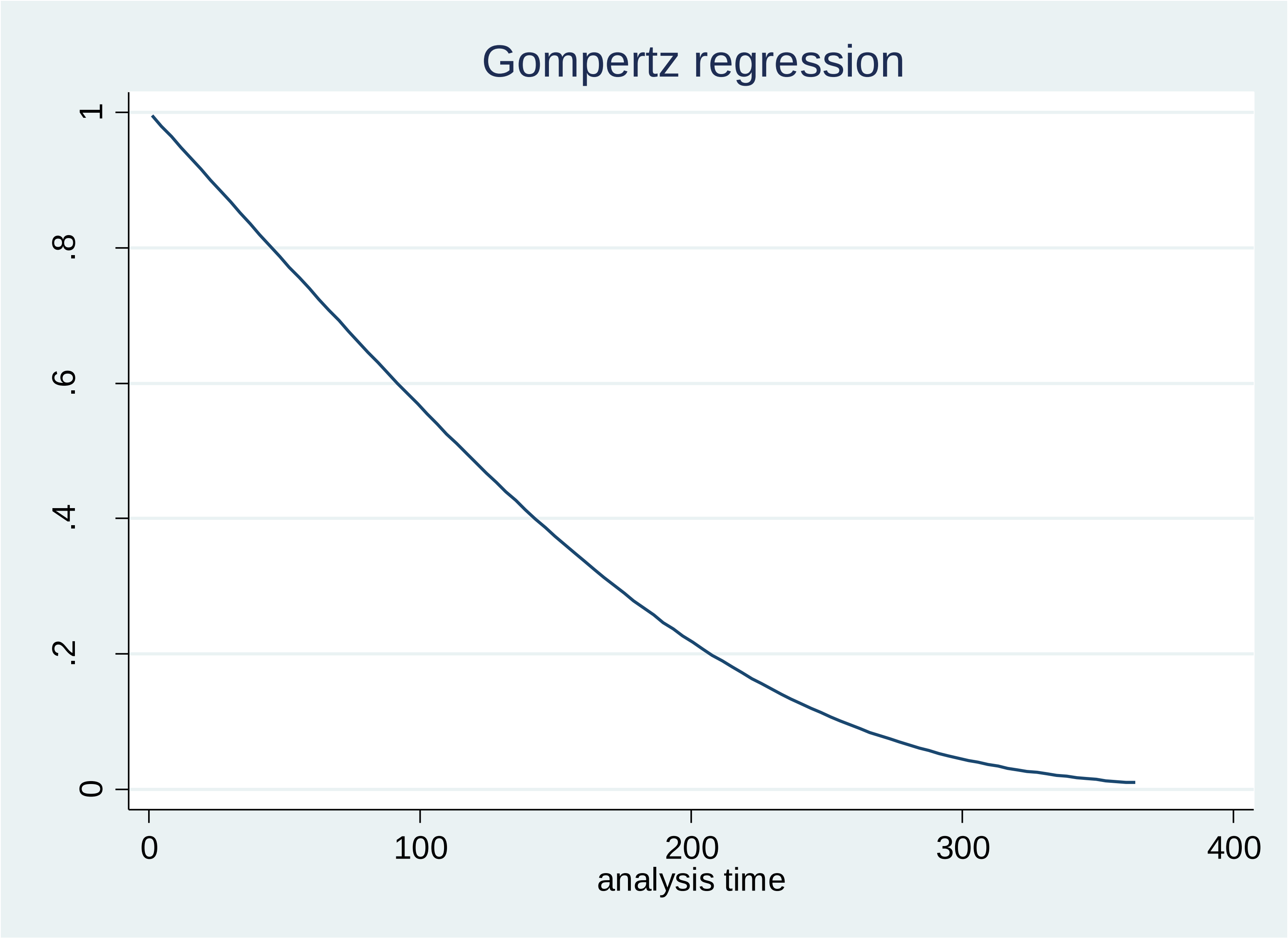
Where γ is the shape parameter, and $\lambda = e^{\beta X}$

```
.  streg gini_m ginmis rgdpch elf elf2 logpop y70stv y80stv y90stv ///
>  d2-d4, dist(gompertz) nohr nolog

         failure _d:  cens
   analysis time _t:  mo
                 id:  indsp

Gompertz regression -- log relative-hazard form

No. of subjects =           55                     Number of obs   =      4625
No. of failures =           48
Time at risk    =         4625
                                                   LR chi2(12)     =     44.15
Log likelihood  =   -79.524695                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gini_m |   -.127802   .0285724    -4.47   0.000    -.1838028   -.0718012
      ginmis |  -6.111802   1.299092    -4.70   0.000    -8.657975   -3.565629
      rgdpch |   .3803728   .1341579     2.84   0.005     .1174281    .6433174
         elf |  -.0652722   .0262721    -2.48   0.013    -.1167646   -.0137799
        elf2 |   .0588794   .0275237     2.14   0.032     .0049339    .1128248
      logpop |  -.3115859   .1224821    -2.54   0.011    -.5516464   -.0715254
      y70stv |  -.0404225   .4672147    -0.09   0.931    -.9561466    .8753015
      y80stv |  -1.530275   .5310187    -2.88   0.004    -2.571052   -.4894973
      y90stv |  -1.302327   .5539067    -2.35   0.019    -2.387964   -.2166894
          d2 |  -.9082049    .578435    -1.57   0.116    -2.041917    .2255069
          d3 |  -.1798546    .574289    -0.31   0.754     -1.30544    .9457313
          d4 |   .0726612   .6031641     0.12   0.904    -1.109519    1.254841
       _cons |   7.578502   2.697252     2.81   0.005     2.291986    12.86502
-------------+----------------------------------------------------------------
      /gamma |   .0050977   .0035916     1.42   0.156    -.0019418    .0121371
------------------------------------------------------------------------------
```

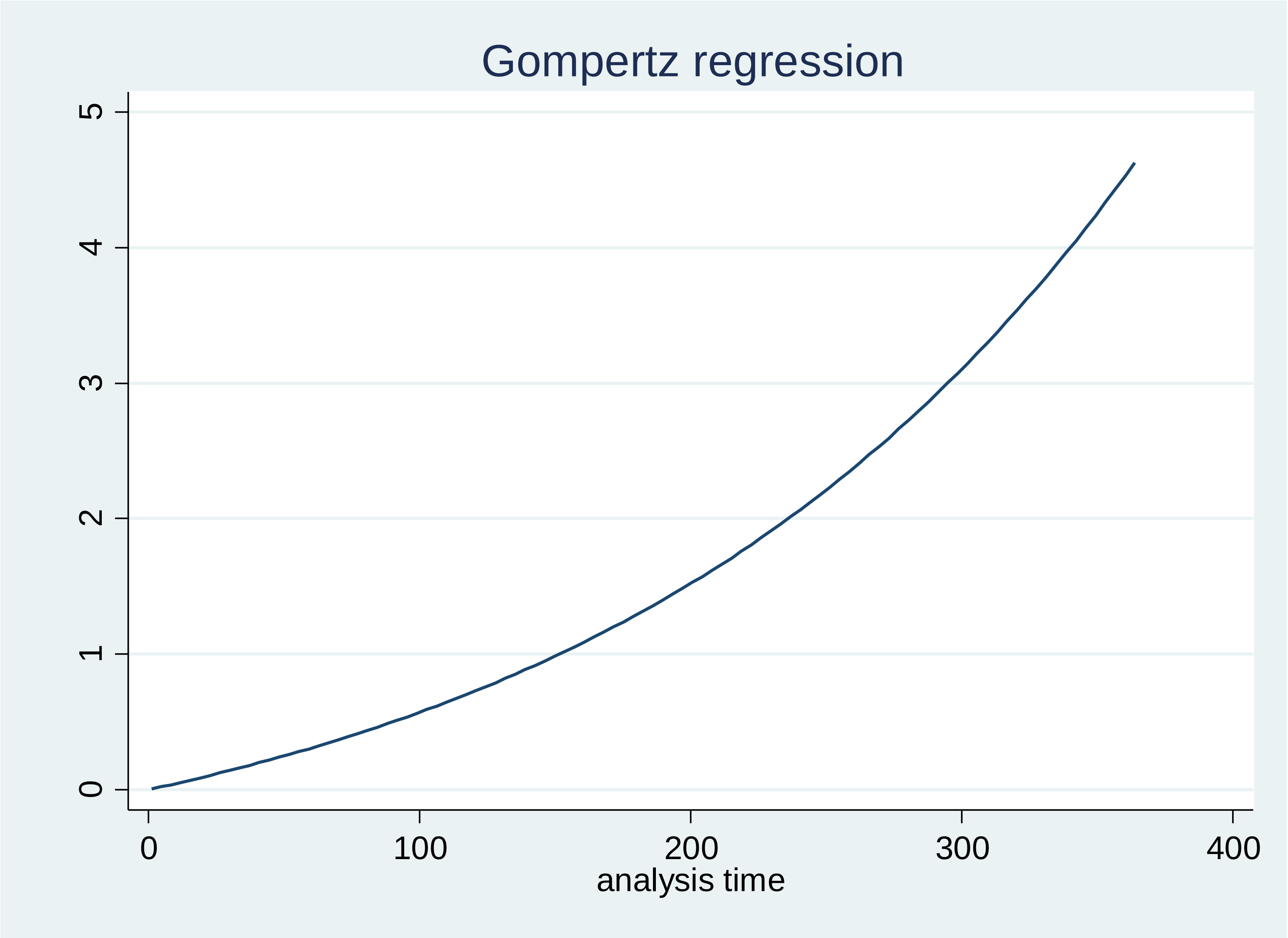
Hazard function



Survival



Cumulative hazard



```
. estout exponential weibull gompertz, cells(b(star) se(par)) ///
>     stats(N  ll  p gamma)

------------------------------------------------------------
              exponential         weibull        gompertz
                     b/se            b/se            b/se
------------------------------------------------------------
_t
gini_m         -.1244463***    -.1221709***     -.127802***
               (.0284179)      (.0288244)      (.0285724)
ginmis         -5.867928***    -5.746803***    -6.111802***
               (1.277403)      (1.302304)      (1.299092)
rgdpch          .3651031**      .3579029**      .3803728**
               (.1322248)      (.1331665)      (.1341579)
elf            -.0628267*      -.0615579*      -.0652722*
               (.0258742)       (.025954)      (.0262721)
elf2            .0581252*       .0572613*       .0588794*
               (.0270411)      (.0270164)      (.0275237)
logpop         -.3163905*      -.3092118*      -.3115859*
               (.1230657)      (.1241655)      (.1224821)
y70stv          .0077905        .0223796       -.0404225
               (.4625409)      (.4641904)      (.4672147)
y80stv         -1.420202**     -1.384908**     -1.530275**
               (.5203341)      (.5263604)      (.5310187)
y90stv         -1.162059*      -1.108178*      -1.302327*
               (.5416506)      (.5550966)      (.5539067)
d2             -.8067415       -.6810668       -.9082049
               (.5742936)      (.6493246)       (.578435)
d3             -.0010657        .1526106       -.1798546
               (.5606172)      (.6702459)       (.574289)
d4              .6098389        .8091091        .0726612
               (.4464024)      (.6495045)      (.6031641)
_cons           7.433105**      7.402131**      7.578502**
               (2.707863)      (2.691839)      (2.697252)
------------------------------------------------------------
ln_p
_cons                          -.0818573
                               (.1986233)
------------------------------------------------------------
gamma
_cons                                           .0050977
                                               (.0035916)
------------------------------------------------------------
N                    4625            4625            4625
ll                 -80.43       -80.34186        -79.5247
p                .0000284        .0001753        .0000144
gamma                                            .0050977
------------------------------------------------------------
```

- These parametric models—the Exponential, Weibull, and Gompertz—all made assumptions about the distribution of the errors.

- How to chose between them?
  - LR or Wald test for nested models
  - AIC for non-nested
  - Run a Generalized Gamma and test whether
    - $\kappa = 1$ for Weibull
    - $\kappa = p = 1$ for Exponential
    - $p = 1$ for Gamma (Stata calls it sigma rather than p)

- Other slightly less common models include the log-normal and log-logistic.

Distribution	$f(t)$	$S(t)$	$h(t)$
Exponential	$\lambda \exp(-\lambda t)$	$\exp(-\lambda t)$	λ
Weibull	$\lambda p t^{p-1} \exp(-\lambda t^p)$	$\exp(-\lambda t^p)$	$\lambda p t^{p-1}$
Log-logistic	$\frac{\lambda p t^{p-1}}{(1+\lambda t^p)^2}$	$\frac{1}{1+\lambda t^p}$	$\frac{\lambda p t^{p-1}}{1+\lambda t^p}$

- From Datwyler and Stucki (2011)

- Next week, we will examine the most common semi-parametric model, the Cox proportional hazards model, that does not make a distributional assumption.

- Questions?

- Questions on the readings?