

University of Oxford

Intermediate Social Statistics: Lecture Six

Raymond M. Duch

Nuffield College Oxford
www.raymond Duch.com
@RayDuch

February 21, 2012

Readings for the Lectures

- ▶ James Fearon. 2004 “Why Do Some Civil Wars Last so Much Longer than Others?” *Journal of Peace Research*, Vol 41, No. 3 (May). 275-301

Readings for the Lectures

- ▶ James Fearon. 2004 “Why Do Some Civil Wars Last so Much Longer than Others?” *Journal of Peace Research*, Vol 41, No. 3 (May). 275-301
- ▶ Gary King, James E. Alt, Nancy Elizabeth Burns, and Michael Laver 1990. “A Unified Model of Cabinet Dissolution in Parliamentary Democracies” *American Journal of Political Science*. 34(3) August: 846-71.

Readings for the Lectures

- ▶ James Fearon. 2004 “Why Do Some Civil Wars Last so Much Longer than Others?” *Journal of Peace Research*, Vol 41, No. 3 (May). 275-301
- ▶ Gary King, James E. Alt, Nancy Elizabeth Burns, and Michael Laver 1990. “A Unified Model of Cabinet Dissolution in Parliamentary Democracies” *American Journal of Political Science*. 34(3) August: 846-71.
- ▶ Cleeves, M., Gould, W., and Gutierrez, R. (2002) *An introduction to survival analysis using Stata*. Stata Press.

Readings for the Lectures

- ▶ James Fearon. 2004 “Why Do Some Civil Wars Last so Much Longer than Others?” *Journal of Peace Research*, Vol 41, No. 3 (May). 275-301
- ▶ Gary King, James E. Alt, Nancy Elizabeth Burns, and Michael Laver 1990. “A Unified Model of Cabinet Dissolution in Parliamentary Democracies” *American Journal of Political Science*. 34(3) August: 846-71.
- ▶ Cleeves, M., Gould, W., and Gutierrez, R. (2002) *An introduction to survival analysis using Stata*. Stata Press.

Readings for the Lectures

- ▶ James Fearon. 2004 “Why Do Some Civil Wars Last so Much Longer than Others?” *Journal of Peace Research*, Vol 41, No. 3 (May). 275-301
- ▶ Gary King, James E. Alt, Nancy Elizabeth Burns, and Michael Laver 1990. “A Unified Model of Cabinet Dissolution in Parliamentary Democracies” *American Journal of Political Science*. 34(3) August: 846-71.
- ▶ Cleeves, M., Gould, W., and Gutierrez, R. (2002) *An introduction to survival analysis using Stata*. Stata Press.

Today's Lecture: Overview

- ▶ Defining Duration Data

Today's Lecture: Overview

- ▶ Defining Duration Data
- ▶ Examples of Duration Data Related Research Questions

Today's Lecture: Overview

- ▶ Defining Duration Data
- ▶ Examples of Duration Data Related Research Questions
- ▶ Exponential Distribution

Today's Lecture: Overview

- ▶ Defining Duration Data
- ▶ Examples of Duration Data Related Research Questions
- ▶ Exponential Distribution
- ▶ Weibull Distribution

Today's Lecture: Overview

- ▶ Defining Duration Data
- ▶ Examples of Duration Data Related Research Questions
- ▶ Exponential Distribution
- ▶ Weibull Distribution

Definition of Duration Data

- ▶ The duration of time that units spend in a state before experiencing some event.
- ▶ We know when observations enter the process and when the process ends
- ▶ We are interested in relationship between length of the observed duration and covariates (independent variables)

Examples of Duration Data

- ▶ Why do coalition governments fail?
- ▶ How do politicians keep their legislative seats over time, in spite of unfavourable conditions?
- ▶ Why do military conflicts persist or fail to persist?

Duration Models are Comparative

- ▶ Event history data contains information on many observations (individuals, politicians, wars, patients, etc.).
- ▶ Explicit comparative inferences can be made regarding differences across the cases.

Examples of Duration Data

Table: Examples of Event History Data: Military Interventions

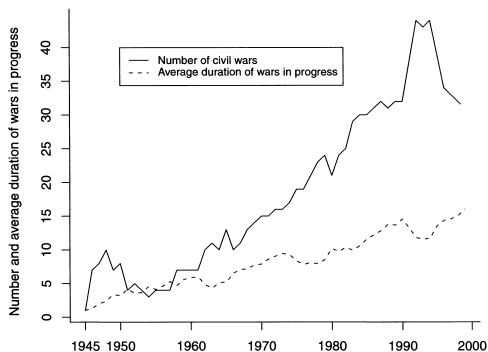
Intervenor	Target	Duration	Contiguous	Censored
UK	Albania	1	0	0
El Salvador	Honduras	657	1	0
U.S.	Panama	274	0	1
Bulgaria	Greece	12	1	0
Taiwan	China	7456	1	0
Botswana	S. Africa	1097	1	0
Uganda	Kenya	409	1	0
Israel	Egypt	357	1	0
Malawi	Mozambique	631	1	1
India	Pakistan	173	1	0

Why Some Civil Wars Last More than Others

276 *journal of PEACE RESEARCH*

volume 41 / number 3 / may 2004

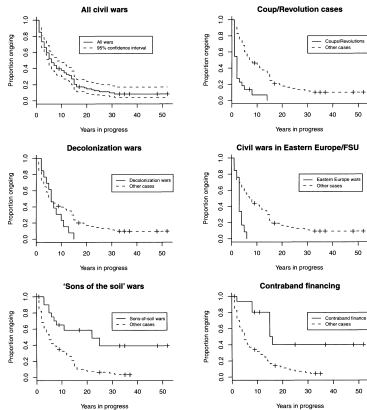
Figure 1. Number and Duration of Civil Wars in Progress



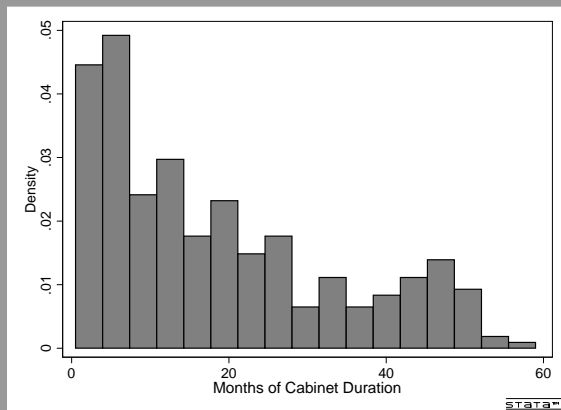
Why Some Civil Wars Last More than Others

James D. Fearon WHY SOME CIVIL WARS LAST LONGER 281

Figure 2. Proportion of Civil Wars Ongoing, by Year



Cabinet Duration Data from King et al



The Data: Coalition Durations

	DURAT	BELGIUM	SALIEN	OPPINF	ROPINF	VOLAT	RESPONSE
1.	3	yes	190.4412	high	4	93	split, majority winner
2.	7	yes	190.4412	high	4	93	split, majority winner
3.	20	yes	190.4412	high	4	93	winner only
4.	6	yes	190.4412	high	4	93	winner only
5.	7	yes	190.4412	high	4	93	winner only
6.	2	yes	261.2745	high	4	62	winner only
7.	17	yes	261.2745	high	4	62	winner only
8.	27	yes	261.2745	high	4	62	winner only
9.	49	yes	261.2745	high	4	62	even split
10.	4	yes	261.2745	high	4	62	winner only
11.	29	yes	261.2745	high	4	62	even split
12.	49	yes	365.4412	high	4	106	losers only
13.	6	yes	365.4412	high	4	106	losers only
14.	23	yes	365.4412	high	4	106	even split
15.	41	yes	365.4412	high	4	106	losers only
16.	10	yes	261.2745	high	4	71	even split
17.	12	yes	261.2745	high	4	71	split, majority loser
18.	2	yes	261.2745	high	4	71	even split
19.	33	yes	261.2745	high	4	71	split, majority loser
20.	1	yes	261.2745	high	4	71	even split

The Survival Function

The **Survival Function** is the probability that the duration of some episode (a cabinet for example) is at least t , and that the event by which the current episode comes to an end occurs later than t .

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

which gives the probability that the episode (the cabinet) is of length (duration) at least t .

The Hazard Rate

The **Hazard Rate** gives the rate at which episodes are completed at duration t , given that they lasted until t .

$$h(t) = \lim_{\Delta \rightarrow 0} \frac{\Pr(t \geq T < t + \Delta | T \geq t)}{\Delta} = \lim_{\Delta \rightarrow 0} \frac{F(t + \Delta) - F(t)}{S(t)} = \frac{f(t)}{S(t)} \quad (2)$$

The issue then is what distribution should we use for $F(\cdot)$

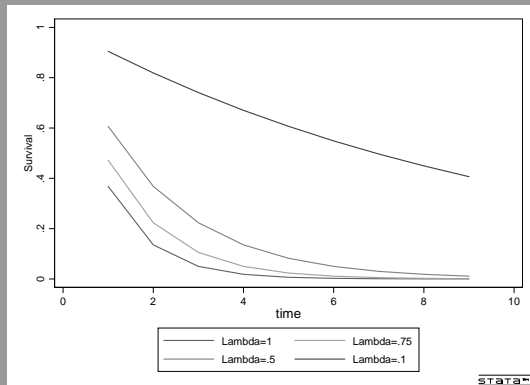
The Exponential Distribution

- ▶ This is the simplest distribution for modeling duration data.
- ▶ The Proportional Hazard version of the model is the following:
 - ▶ The hazard function $h(t) = \lambda$, where $\lambda = e^{xb}$
 - ▶ And the survival function $S(t) = e^{-\lambda t}$.
 - ▶ Positive coefficients on covariates indicate increasing hazards and decreasing survival time
 - ▶ Default specification in Stata

The Exponential Accelerated Failure Time Model

- ▶ The Accelerated Failure Time version of the model is the following:
 - ▶ The hazard function $h(t) = \lambda$, where $\lambda = e^{-xb}$
 - ▶ And the survival function $S(t) = e^{-\lambda t}$, where $\lambda = e^{-xb}$.
 - ▶ Positive coefficients on covariates indicate decreasing hazards (increasing failure time) and increasing survival time
 - ▶ time option in Stata

The Exponential Distribution



The Data: Coalition Durations

```
. streg, distribution(exponential) time
```

```
      failure _d: 1 (meaning all fail)
analysis time _t: DURAT
```

```
Iteration 0:  log likelihood = -460.73222
```

```
Iteration 1:  log likelihood = -460.73222
```

```
Exponential regression -- accelerated failure-time form
```

```
No. of subjects =          313                Number of obs   =          313
No. of failures =          313
Time at risk    =          5789
Log likelihood  = -460.73222                LR chi2(0)        =          0.00
                                                Prob > chi2       =          .
```

```
-----+-----
      _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      _cons |  2.917512   .0565233    51.62  0.000    2.806728    3.028295
-----+-----
```

```
.
end of do-file
```

The Exponential Distribution (AFT)

- ▶ The hazard rate is $\lambda = e^{-2.917} = .054$.
- ▶ The expected duration, expressed in months, of a typical cabinet government in this data set is then:

$$E(Y_i) = \frac{1}{\lambda} = \frac{1}{e^{-2.917}} = 18.5$$

Adding a Systematic Component to the Exponential Distribution

The Hazard Function incorporating explanatory variables is

$$h(t) = \lambda_i = \exp^{-(\beta_0 + \beta_{\mathbf{k}} \mathbf{x}_i)}$$

(3)

The Survivor Function then can be estimated as.

$$S(t) = \exp(-\lambda t) = \exp(-e^{-(\beta_0 + \beta_{\mathbf{k}} \mathbf{x}_i)} t) \quad (4)$$

The Data: Coalition Durations

```
. streg BELGIUM-SWEDEN, distribution(exponential) time
      failure _d: 1 (meaning all fail)
      analysis time _t: DURAT
```

```
Exponential regression -- accelerated failure-time form
No. of subjects =          313                Number of obs   =          313
No. of failures =          313
Time at risk    =          5789
Log likelihood  = -422.37542
LR chi2(14)    =          76.71
Prob > chi2    =          0.0000
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
BELGIUM	-.6171175	.3054608	-2.02	0.043	-1.21581 - .0184254
CANADA	-.0073727	.3483152	-0.02	0.983	-.690058 .6753126
DENMARK	-.3548504	.3170019	-1.12	0.263	-.9761626 .2664619
FINLAND	-.7440277	.3017973	-2.47	0.014	-1.335539 -.152516
FRANCE	-1.832701	.3054608	-6.00	0.000	-2.431394 -1.234009
ICELAND	-.0637419	.3429971	-0.19	0.853	-.7360039 .6085202
IRELAND	.0442898	.3542459	0.13	0.901	-.6500194 .738599
ISRAEL	-.6717962	.3170019	-2.12	0.034	-1.293108 -.0504839
ITALY	-1.077935	.288468	-3.74	0.000	-1.643322 -.5125485
NETHER	-.1903214	.3429972	-0.55	0.579	-.8625835 .4819407
NORWAY	-.19697	.3298841	-0.60	0.550	-.8435309 .449591
PORTUG	-1.015143	.3985267	-2.55	0.011	-1.796241 -.2340449
SPAIN	-.0180909	.6262242	-0.03	0.977	-1.245468 1.209286
SWEDEN	-.19697	.3298841	-0.60	0.550	-.8435309 .449591
_cons	3.385387	.2425356	13.96	0.000	2.910026 3.860748

Adding a Systematic Component to the Exponential Distribution

- ▶ Having added a systematic component to the exponential hazard function we can now calculate the expected cabinet duration for each of the countries in our data set.
- ▶ The UK case for example (which is captured by the intercept) is

$$E(Y_i) = \frac{1}{\lambda} = \frac{1}{e^{-3.385}} = 29.5$$

And the Italian case is.

$$E(Y_i) = \frac{1}{\lambda} = \frac{1}{e^{-(3.385-1.08)}} = 10.05$$

Average Coalition Duration

	country	meantime	haz
1.	Belgium	15.93103	.0627706
2.	Canada	29.31249	.0341151
3.	Denmark	20.70833	.0482897
4.	Finland	14.03226	.0712644
5.	France	4.724138	.2116788
6.	Iceland	27.70588	.0360934
7.	Ireland	30.86665	.0323974
8.	Israel	15.08333	.0662983
9.	Italy	10.04878	.0995146
10.	Netherlands	24.41176	.0409639
11.	Norway	24.25	.0412371
12.	Portugal	10.7	.0934579
13.	Spain	29	.0344828
14.	Sweden	24.25	.0412371
15.	UK	29.52941	.0338646

Adding a Systematic Component to the Exponential Distribution

Rather than simply distinguishing the hazard rates by country we can incorporate features of these countries as independent variables in the hazard function.

$$h(t) = \exp(-[\beta_0 + \beta_1 \text{Country Attributes} + \beta_2 \text{Party Structure Attributes} + \beta_3 \text{Coalition Attributes}]) \quad (5)$$

Censuring

- ▶ Observations of events are very often censored.
- ▶ Censoring occurs when the information about the duration in the origin state is incompletely recorded.
- ▶ *Left Censoring* occurs when the starting times of an episode are located before the beginning of the observational window.
- ▶ *Right Censoring* occurs because the observation is terminated at the right-side of the observation window.

Censuring: Coalition Example

- ▶ In the cabinet duration example, right censoring is a consideration that needs to be incorporated into the model.

Censuring: Coalition Example

- ▶ In the cabinet duration example, right censoring is a consideration that needs to be incorporated into the model.
- ▶ There are two distinct types of observed cabinet durations.

Censuring: Coalition Example

- ▶ In the cabinet duration example, right censoring is a consideration that needs to be incorporated into the model.
- ▶ There are two distinct types of observed cabinet durations.
- ▶ The largest group of observed coalitions are those that break apart due to some critical events. These are *uncensored observations* for which we have exact time of failure. We will define $d_i = 1$ for these uncensored observations.

Censuring: Coalition Example

- ▶ In the cabinet duration example, right censoring is a consideration that needs to be incorporated into the model.
- ▶ There are two distinct types of observed cabinet durations.
- ▶ The largest group of observed coalitions are those that break apart due to some critical events. These are *uncensored observations* for which we have exact time of failure. We will define $d_i = 1$ for these uncensored observations.
- ▶ The second group consists of those that come close to their constitutional inter-election period (CIEP). These are the *censored observations* and we only have information on the fact that they have survived at least to time T_i . We will define $d_i = 0$ for these censored observations.

Censuring: Coalition Example

- ▶ In the cabinet duration example, right censoring is a consideration that needs to be incorporated into the model.
- ▶ There are two distinct types of observed cabinet durations.
- ▶ The largest group of observed coalitions are those that break apart due to some critical events. These are *uncensored observations* for which we have exact time of failure. We will define $d_i = 1$ for these uncensored observations.
- ▶ The second group consists of those that come close to their constitutional inter-election period (CIEP). These are the *censored observations* and we only have information on the fact that they have survived at least to time T_i . We will define $d_i = 0$ for these censored observations.

Censuring: Defined

The complete censored likelihood is then a mixture of continuous (observed) and discrete (censored) observations.

$$\begin{aligned}\ln L &= \sum_{i=1}^N \left\{ d_i \ln \left[e^{-\mathbf{X}_i \beta} e^{-(e^{-\mathbf{X}_i \beta} t)} \right] + (1 - d_i) \ln \left[e^{-(e^{-\mathbf{X}_i \beta} t)} \right] \right\} \\ &= \sum_{i=1}^N \left\{ d_i \left[(-\mathbf{X}_i \beta) (-e^{-\mathbf{X}_i \beta} t) \right] + (1 - d_i) (-e^{-\mathbf{X}_i \beta} t) \right\} \quad (6)\end{aligned}$$

where d_i takes on the value of one for uncensored observations, and zero otherwise. Hence notice the only unknown in the likelihood function is β .

Fully Specified Model with Censoring

```
. streg IDENT POPINFL INVEST VOLAT ///
> RESPONSE FRACT POLAR NUMST2 CRISIS OPPCONC FORMAT, nohr distribution(exponential) time
      failure _d: CIEP12
      analysis time _t: DURAT
```

```
Exponential regression -- accelerated failure-time form
No. of subjects =          313          Number of obs   =          313
No. of failures =          270
Time at risk    =          5789

LR chi2(11)      =          119.85
Log likelihood   = -436.46511        Prob > chi2      =          0.0000
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
IDENT	.340261	.1432652	2.38	0.018	.0594663	.6210557
POPINFL	.071121	.2195679	0.32	0.746	-.3592242	.5014663
INVEST	-.3700965	.170434	-2.17	0.030	-.704141	-.0360519
VOLAT	.0002848	.0011019	0.26	0.796	-.0018749	.0024444
RESPONSE	-.000048	.0429063	-0.01	0.991	-.0845748	.0836148
FRACT	-.0005707	.0011007	-0.52	0.604	-.0027279	.0015866
POLAR	-.0232734	.0100109	-2.32	0.020	-.0428945	-.0036523
NUMST2	.4962359	.1531668	3.24	0.001	.1960345	.7964373
CRISIS	.0083382	.0023954	3.48	0.000	.0036434	.0130331
OPPCONC	.1294083	.087883	1.47	0.141	-.0428392	.3016557
FORMAT	-.0782494	.0485587	-1.61	0.107	-.1734227	.0169238
_cons	2.8698	1.02284	2.81	0.005	.8650692	4.87453

Average Coalition Duration in Fully Specified Model

	country	meantime	haz
1.	Belgium	17.03443	.0681892
2.	Canada	47.20558	.0240864
3.	Denmark	22.45923	.0580686
4.	Finland	16.95561	.0775819
5.	France	6.958303	.1641795
6.	Iceland	32.85577	.0425343
7.	Ireland	28.84973	.0377387
8.	Israel	22.38428	.0558742
9.	Italy	8.88159	.1245141
10.	Netherlands	34.23247	.0415503
11.	Norway	34.46441	.0305163
12.	Portugal	28.40756	.0375801
13.	Spain	21.55596	.057197
14.	Sweden	37.98434	.0306388
15.	UK	51.94485	.0201483

Hazard Ratios

For the Proportional Hazard Version of the Exponential Model

$$\lambda = \exp^{(X\beta)} \quad (7)$$

The Hazard Ratio is defined as follows:

$$\frac{h(t)|X_k + 1}{h(t)|X_k} = \exp^{(\beta_k)} \quad (8)$$

Weibull Model

- ▶ The exponential model is nice enough, but the restriction that the hazard be constant over time is often questioned in practice.

Weibull Model

- ▶ The exponential model is nice enough, but the restriction that the hazard be constant over time is often questioned in practice.
- ▶ we can imagine a number of processes where we might expect hazard rates to be changing over time.

Weibull Model

- ▶ The exponential model is nice enough, but the restriction that the hazard be constant over time is often questioned in practice.
- ▶ we can imagine a number of processes where we might expect hazard rates to be changing over time.
- ▶ If the (conditional) hazard is increasing or decreasing steadily over time, the exponential model will miss this fact.

Weibull Model

The Weibull can be thought of as a hazard rate model in which the hazard is:

Weibull Model

The Weibull can be thought of as a hazard rate model in which the hazard is:

$$h(t) = \lambda p(\lambda t)^{p-1} \quad (9)$$

Weibull Model

The Weibull can be thought of as a hazard rate model in which the hazard is:

$$h(t) = \lambda p(\lambda t)^{p-1} \quad (9)$$

The p parameter is called a “shape parameter” because it defines the shape of the Weibull distribution:

Weibull Model

The Weibull can be thought of as a hazard rate model in which the hazard is:

$$h(t) = \lambda p (\lambda t)^{p-1} \quad (9)$$

The p parameter is called a “shape parameter” because it defines the shape of the Weibull distribution:

- ▶ $p = 1$ corresponds to an exponential model (thus the Weibull “nests” the exponential model)

Weibull Model

The Weibull can be thought of as a hazard rate model in which the hazard is:

$$h(t) = \lambda p (\lambda t)^{p-1} \quad (9)$$

The p parameter is called a “shape parameter” because it defines the shape of the Weibull distribution:

- ▶ $p = 1$ corresponds to an exponential model (thus the Weibull “nests” the exponential model)
- ▶ $p > 1$ means that the hazards are rising monotonically over time, and

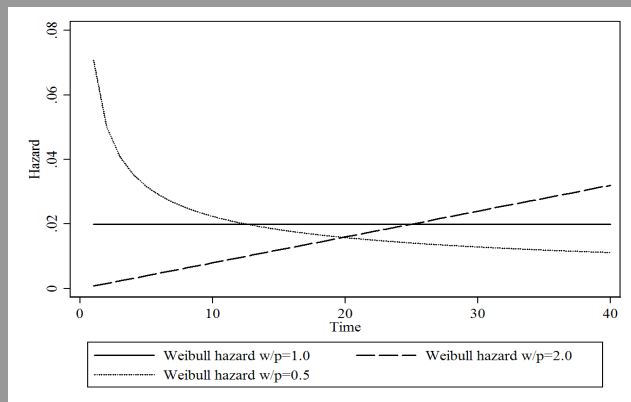
Weibull Model

The Weibull can be thought of as a hazard rate model in which the hazard is:

$$h(t) = \lambda p (\lambda t)^{p-1} \quad (9)$$

The p parameter is called a “shape parameter” because it defines the shape of the Weibull distribution:

- ▶ $p = 1$ corresponds to an exponential model (thus the Weibull “nests” the exponential model)
- ▶ $p > 1$ means that the hazards are rising monotonically over time, and
- ▶ $0 < p < 1$ means hazards are decreasing monotonically over time.

Weibull Hazard with $\lambda = 0.02$ 

Weibull Model Survival Function

Weibull survival function can be expressed as:

$$f(t) = \lambda p (\lambda t)^{p-1} x \exp^{(-\lambda t)^p} \quad (10)$$

- ▶ the Weibull is a two-parameter distribution

Weibull Model Survival Function

Weibull survival function can be expressed as:

$$f(t) = \lambda p (\lambda t)^{p-1} x \exp^{(-\lambda t)^p} \quad (10)$$

- ▶ the Weibull is a two-parameter distribution
- ▶ λ denotes the overall level of the hazard

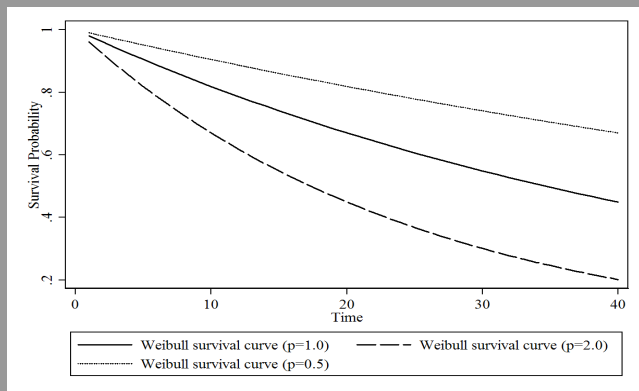
Weibull Model Survival Function

Weibull survival function can be expressed as:

$$f(t) = \lambda p (\lambda t)^{p-1} x \exp^{(-\lambda t)^p} \quad (10)$$

- ▶ the Weibull is a two-parameter distribution
- ▶ λ denotes the overall level of the hazard
- ▶ p determines its shape (increasing, decreasing or constant) – some refer to $\alpha = 1/p$.

Weibull Survival with $\lambda = 0.02$



Systematic Component of Weibull Model

As in the exponential case, we typically introduce covariates through an exponential function, to ensure that the hazard remains positive:

$$\lambda_i = \exp^{(X\beta)} \quad (11)$$

Determinants of Coalition Failure: Weibull

```
streg IDENT POPINFL INVEST VOLAT RESPONSE FRACT POLAR NUMST2 CRISIS OPPCONC FORMAT, nohr d(w
```

```
failure _d: 1 (meaning all fail)
```

```
Weibull regression -- accelerated failure-time form
```

```
No. of subjects = 313 Number of obs = 313
```

```
No. of failures = 313
```

```
Time at risk = 5789
```

```
LR chi2(11) = 117.36
```

```
Log likelihood = -397.81517
```

```
Prob > chi2 = 0.0000
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
IDENT	.2694487	.0948906	2.84	0.005	.0834665 .455431
POPINFL	.0156217	.1423621	0.11	0.913	-.2634029 .2946464
INVEST	-.2348876	.1109902	-2.12	0.034	-.4524244 -.0173509
VOLAT	-.0002645	.0007231	-0.37	0.715	-.0016818 .0011528
RESPONSE	-.0024215	.0290807	-0.08	0.934	-.0594185 .0545756
FRACT	-.0000695	.0007099	-0.10	0.922	-.0014608 .0013218
POLAR	-.0184121	.0065785	-2.80	0.005	-.0313057 -.0055186
NUMST2	.3259987	.1060599	3.07	0.002	.118125 .5338723
CRISIS	.0056751	.0015397	3.69	0.000	.0026573 .0086929
OPPCONC	.1074939	.0646413	1.66	0.096	-.0192007 .2341886
FORMAT	-.067827	.0331885	-2.04	0.041	-.1328752 -.0027787
_cons	2.717418	.6626347	4.10	0.000	1.418678 4.016158
/ln_p	.322616	.0455071	7.09	0.000	.2334237 .4118082
p	1.380735	.0628332			1.262916 1.509545
1/p	.7242519	.0329586			.6624513 .791818