

Some Quantities of Interest

Cumulative Distribution Function

$$F(t) = \int_0^t f(u)d(u) = \Pr(T \leq t),$$

Probability Density Function

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{F(t + \Delta t) - F(t)}{\Delta t},$$

(which is sometimes written as)

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T \leq t + \Delta t)}{\Delta t},$$

Survivor Function

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

Hazard Function

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

(which is sometimes written as)

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T \leq t + \Delta t \mid T \geq t)}{\Delta t}$$

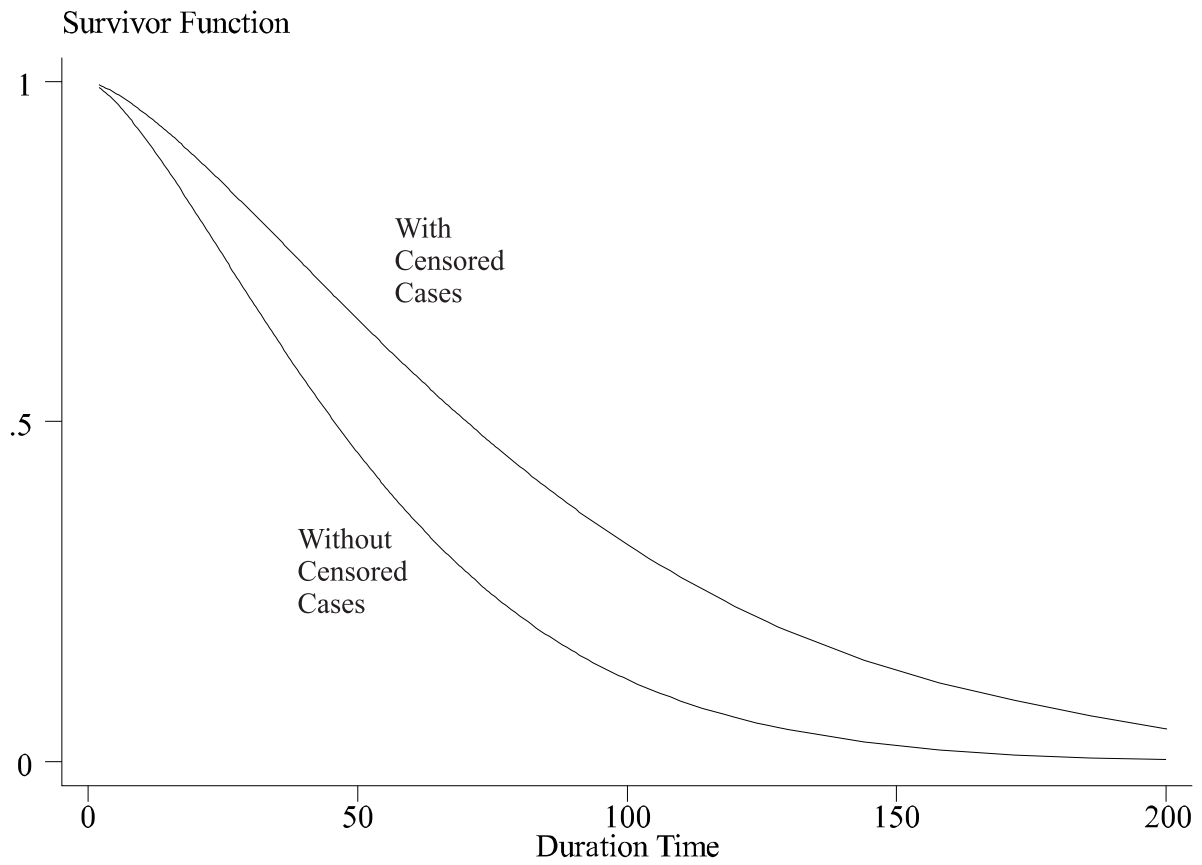


Figure 1: *This figure graphs the survivor function for a hypothetical data set. The top line denotes the survivor function for a data set with censored observations; the bottom line denotes the survivor function for uncensored data.*

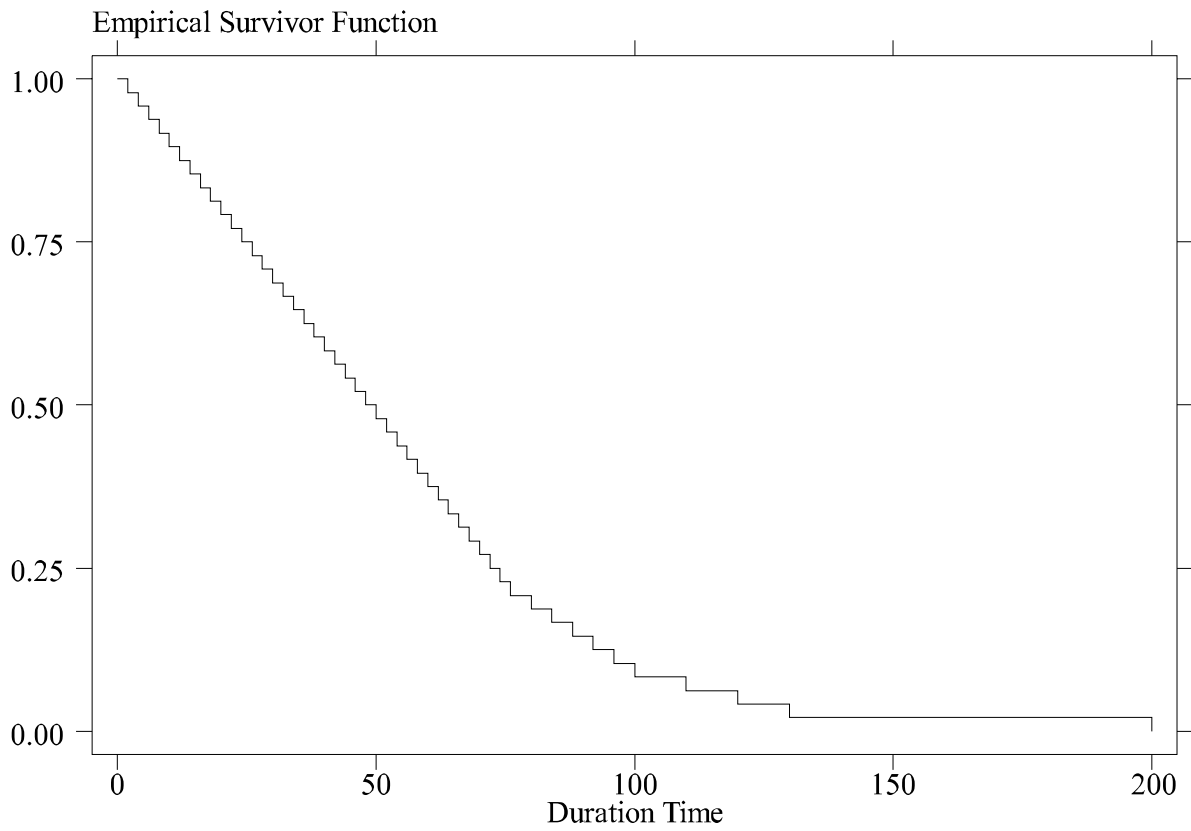


Figure 2: *This figure graphs the “empirical” survivor function for a hypothetical set of data. Note the stair-step nature of the function. This occurs because observations are observed as failing at discrete times, hence, the empirical survivor function is “flat” in between failures.*

## Censoring and Likelihood

$$h(t) = \frac{\int_t^{t+\Delta t} f(u)du}{\int_t^\infty f(u)du}. \quad (1)$$

$$S(t) = \int_t^{t=C_i} f(u)du. \quad (2)$$

$$S(t) = \int_{t_L}^\infty f(u)du, \quad (3)$$

$$\mathcal{L} = \prod_i^n f(t_i).$$

$$\mathcal{L} = \prod_{t_i \leq t^*} f(t_i) \prod_{t_i > t^*} S(t_i^*).$$

$$\delta_i = \begin{cases} 1 & \text{if } t_i \leq t^* \\ 0 & \text{if } t_i > t^* \end{cases}$$

$$\mathcal{L} = \prod_{i=1}^n \{f(t_i)\}^{\delta_i} \{S(t_i)\}^{1-\delta_i}, \quad (4)$$

## Usual Approaches to Modeling Duration Data

- Parametric Duration Models without TVCs are common.
- Concern (perhaps too much?) with “baseline hazard rates.”
- Extensive Use of Binary Link Models for Duration Data.
- With binary link models, time dependency is often ignored, thus leading to an exponential equivalent.
- Single-spell models dominate.
- Multi-spell data are treated *as if* they are single-spell.
- Events are broadly defined (which makes estimation of a standard single-spell model easy).
- Unobserved heterogeneity is acknowledged, but often not dealt with.

The logic of a parametric model is to define the time-dependency. It may be increasing:

$$\frac{dh(t)}{dt} > 0,$$

It may be decreasing:

$$\frac{dh(t)}{dt} < 0,$$

It may be flat:

$$\frac{dh(t)}{dt} = 0,$$

Or it could be increasing and then decreasing (nonmonotonic).

Popular choices include the:

- Weibull (under which the exponential is nested)
- Log-Normal
- Log-Logistic
- Gamma
- Gompertz
- Rayleigh
- ... and many other potential candidates.

→ Each of these are easy to implement in standard software packages like Stata, Limdep, SAS, or S-Plus.

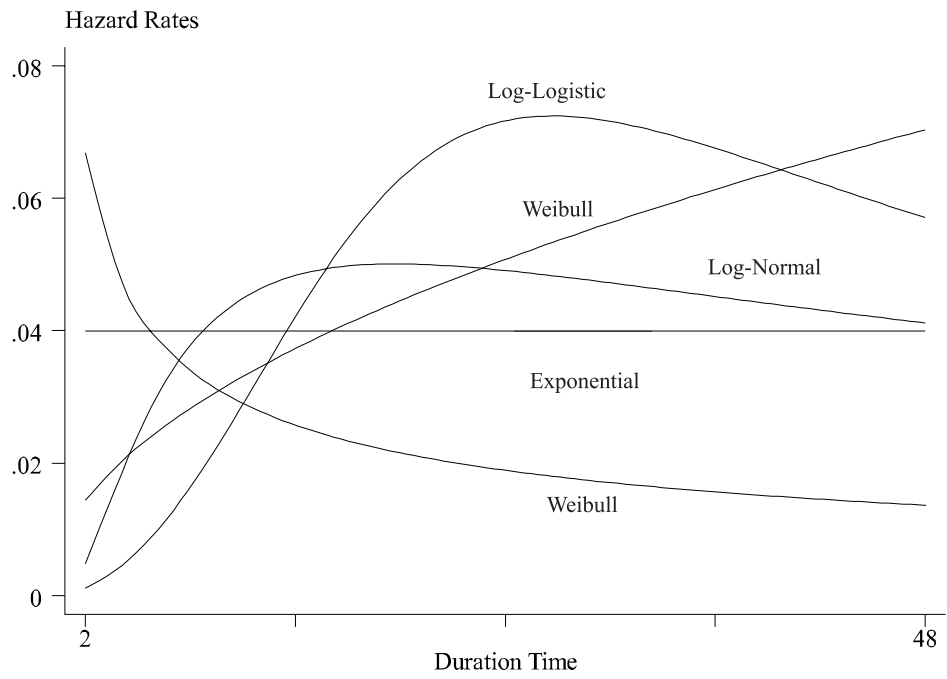


Figure 3: *This figure graphs typical functional forms for several common parametric distribution functions.*

## Densities for Two Encompassing Distribution Functions

The Generalized Gamma:

$$f(t) = \frac{\lambda p (\lambda t)^{p\kappa-1} \exp[-(\lambda t)^p]}{\Gamma(\kappa)},$$

If  $\kappa = 1$ , the Weibull distribution is implied.

If  $\kappa = p = 1$ , the exponential distribution is implied.

If  $\kappa = 0$ , the log-normal distribution is implied.

If  $p = 1$ , the gamma distribution is implied.

The Generalized  $F$ :

$$f(t) = \frac{\lambda p (\lambda t)^{-1}}{\beta(M_1, M_2) \mathbf{K}^{M_1} \times (1 + \mathbf{K})^{-(M_1 - M_2)}},$$

If  $(M_1, M_2) = (1, 1)$  the log-logistic is implied.

If  $(M_1, M_2) = (1, \infty)$ , the Weibull density is implied.

If  $(M_1, M_2) = (\infty, \infty)$ , the log-normal is implied.

If  $(M_1, M_2, \sigma) = (1, \infty, 1)$ , the exponential is implied.

If  $(M_2) = (\infty)$ , the gamma is implied.

Likelihoods defined in terms of the generalized  $F$  are *very* ill-behaved!

Why Not Just Estimate A Cox Model?

The Cox Model

$$h_i(t) = h_0(t) \exp(\beta' \mathbf{x}),$$

Illustration of the Proportional Hazard Property

$$\frac{h_i(t)}{h_0(t)} = \exp(\beta'(\mathbf{x}_i - \mathbf{x}_j)),$$

(The ratio of two hazards is a fixed proportion across time.)

Scalar Form of Cox Regression Model

$$h_i(t) = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}) h_0(t),$$

Re-expressed in terms of the log of the hazard ratios

$$\log \left\{ \frac{h_i(t)}{h_0(t)} \right\} = \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}.$$

Note the absence of a constant term (i.e. no  $\beta_0$ ).

Table 1: Example of Discrete-Time Event History Data

Case I.D.	Event Occurrence	Year	Time Elapsed
1	0	1974	1
1	0	1975	2
⋮	⋮	⋮	⋮
1	0	1986	13
1	1	1987	14
5	1	1974	1
45	0	1974	1
45	0	1975	2
⋮	⋮	⋮	⋮
45	0	1992	19
45	0	1993	20

These data are a portion of a data set originally analyzed in Brace, Hall, and Langer (1999). I thank Laura Langer for letting us use them.

## Discrete Models and Math

$$f(t) = \Pr(T = t_i) \quad (5)$$

$$S(t) = \Pr(T \geq t_i) = \sum_{j \geq i} f(t_j) \quad (6)$$

$$h(t) = \frac{f(t)}{S(t)} \quad (7)$$

## Likelihood

$$\mathcal{L} = \prod_i^n \left[ h(t_i) \prod_{i=1}^{t-1} (1 - h(t_i)) \right]^{y_{it}} \left[ \prod_{i=1}^t (1 - h(t_i)) \right]^{1-y_{it}} \quad (8)$$

$$\mathcal{L} = \prod_{i=1}^n \{f(t)\}^{y_{it}} \{S(t)\}^{1-y_{it}}, \quad (9)$$

## Models

$$h(t) = \Pr(T = t_i \mid T \geq t_i, \mathbf{x}) \quad (10)$$

$$\lambda_{it} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}. \quad (11)$$

Logit

$$\log\left(\frac{\lambda_i}{1-\lambda_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}. \quad (12)$$

$$\hat{\lambda}_i = \frac{e^{\beta' \mathbf{x}}}{1 + e^{\beta' \mathbf{x}}}, \quad (13)$$

Probit

$$\Phi^{-1}[\lambda_i] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}, \quad (14)$$

$$\hat{\lambda}_i = \Phi(\beta' \mathbf{x}). \quad (15)$$

Complementary Log-Log

$$\log[-\log(1 - \lambda_i)] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}. \quad (16)$$

$$\hat{\lambda}_i = 1 - \exp[-\exp(\beta' \mathbf{x})]. \quad (17)$$

## The Exponential Equivalent of a Logit EH Model

$$\log\left(\frac{\lambda_i}{1-\lambda_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i},$$

where  $x_{ki}$  are two covariates of interest that have a mean of 0, and  $\beta_0$  is the constant term. The “baseline” hazard under this model would be equivalent to

$$\hat{\lambda}_i = h_0(t) = \exp(\beta_0),$$

which is a constant. The baseline hazard is thus an exponential equivalent.

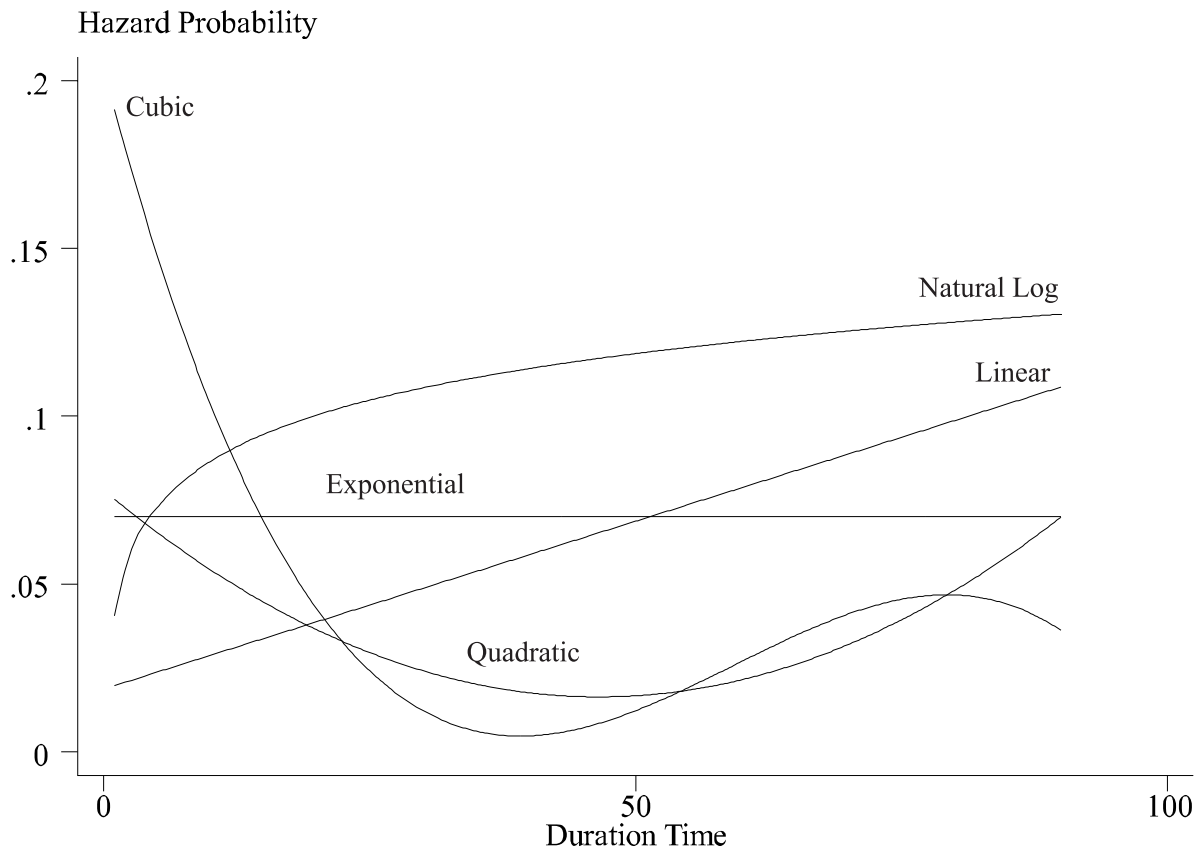


Figure 4: *This figure illustrates different ways that the duration time may be transformed to account for duration dependency in the discrete-time model. The lines in the graph represent a cubic transformation, natural log transformation, time entered linearly, no time dependency (exponential), and a quadratic transformation.*

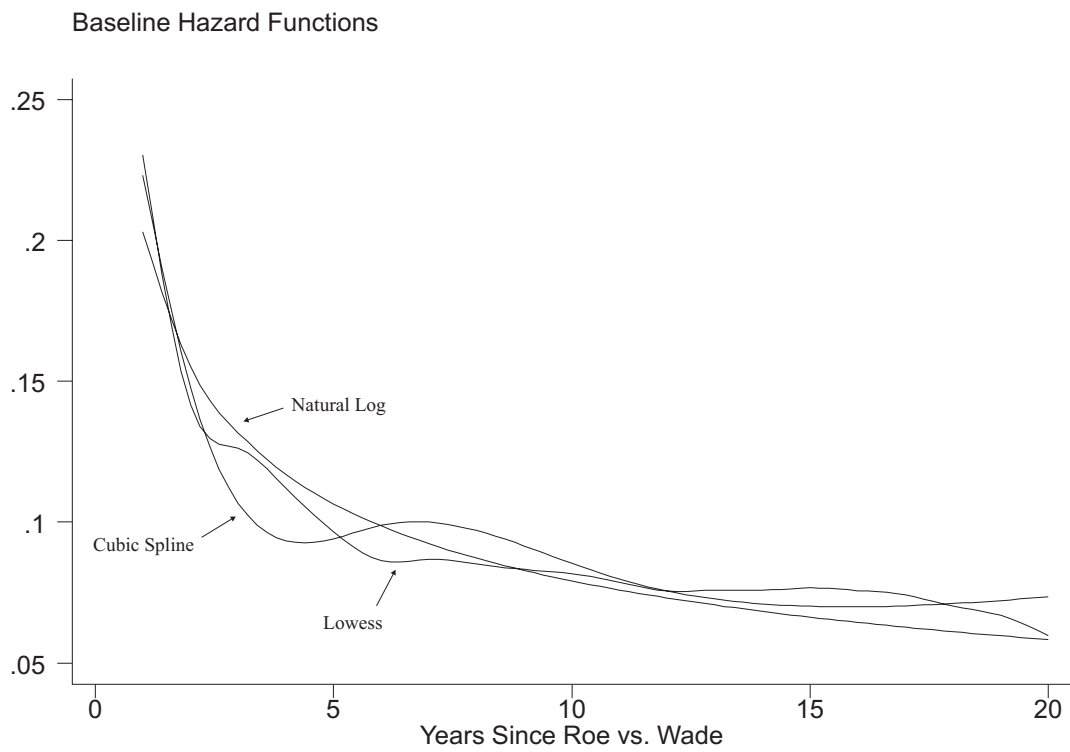


Figure 5: *This figure illustrates a cubic spline function, lowess, and natural log transformation of the baseline hazard function from a model of restrictive abortion policy adoption.*