Extended Lecture Notes for Short Course on Event History Analysis, University of Auckland, 31 May 2005

Professor Brad Jones
University of Arizona

1 Preliminaries

Now we have some motivation for applying duration models to social science problems. Let me establish some definitions in order to make subsequent topics a little clearer.

The mathematical quantities undergirding duration models are easily defined and let me quickly review them.

First, let's define a cumulative distribution function as

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

which in words, specifies the probability that a survival time T is less than or equal to some value t. For all points that $F(t)$ is differentiable, then a probability density function is defined, and can be expressed as

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

which in words gives us the unconditional failure rate (that is, the rate at which observations are failing) in an infinitesimally small differentiable area. Another common way to express the density function is as

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

which in words gives us the *instantaneous* probability that an event will occur in an infinitesimally small differentiable area.

If the distribution function tells us something about the failure rate, the *survivor function* tells us something about survival. This important quantity is simply

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

which in words gives us the probability that a survival time T is equal to or greater than some time t. The survivor function will tell us something about the proportion of observations surviving across time. As such, $S(t)$ must be *strictly decreasing*. At time 0, the proportion of survivors is 1; as time passes, this proportion must decline (we disallow reincarnation!).

We can see a stylized version of a survivor function in Figure 1. Note the implications

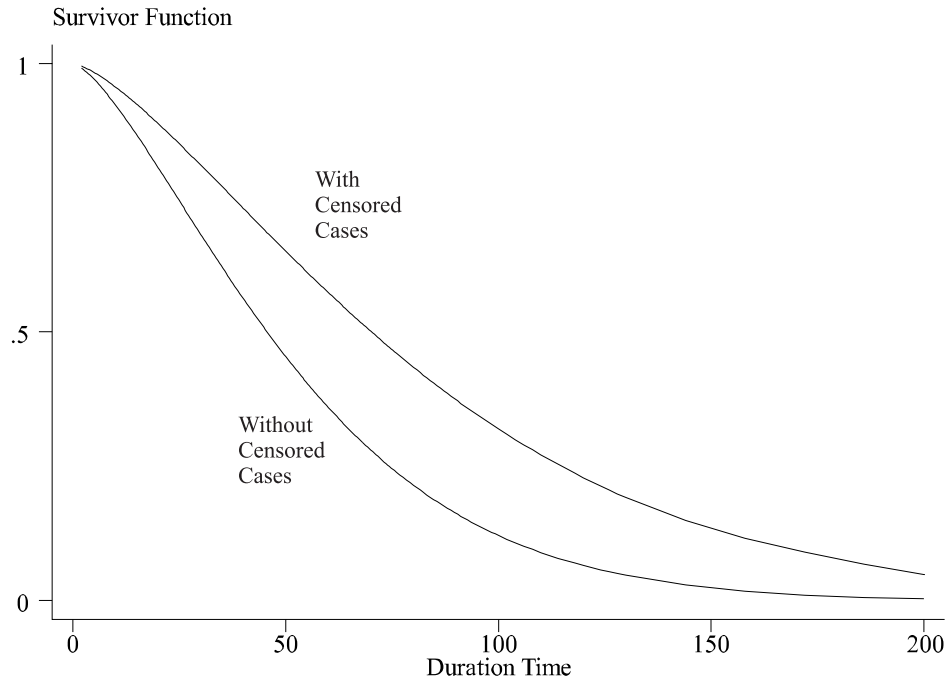

Figure 1: *This figure graphs the survivor function for a hypothetical data set. The top line denotes the survivor function for a data set with censored observations; the bottom line denotes the survivor function for uncensored data.*

of censored observations [$S(t)$ never hits 0.]

Of course empirically, $S(t)$ is a step-function. This is the case because we observe cases at discrete time points (i.e. we *don't* have continuous observation and measurement). An empirical survivor function may look like that shown in Figure 2.

I have said something about failures through $f(t)$, and survivors through $S(t)$: we now have the moving parts to tell us something about *risk*. Logically, risk is simply the relationship between failure and survival. As an observation survives or persists through time, the observation incurs a risk that at some point in time, it will fail (i.e. an event will occur). The quantity in EHA that gives us this kind of interpretation is the *hazard rate*.

Mathematically, we can define the hazard rate as

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

which illustrates clearly our intuition that risk (given by $h(t)$) is simply the relationship between failure and survival. Recall the discussion earlier, where I cast this notion in more conceptual terms.

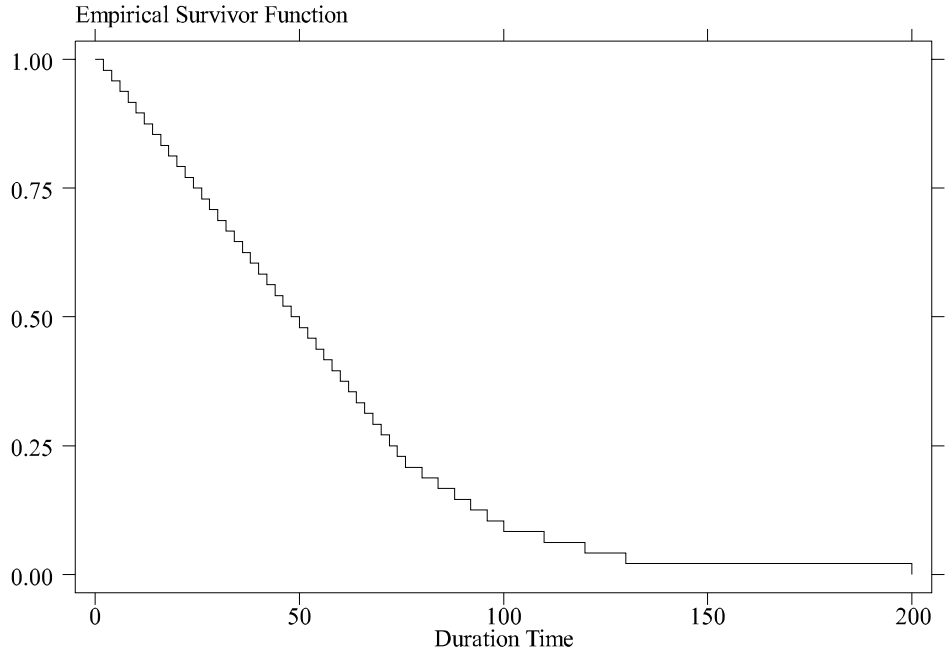

Figure 2: *This figure graphs the "empirical" survivor function for a hypothetical set of data. Note the stair-step nature of the function. This occurs because observations are observed as failing at discrete times, hence, the empirical survivor function is "flat" in between failures.*

Specifically, the hazard rate gives the rate at which units fail (or durations end) by t given that the unit had survived until t. Thus, the hazard rate is a *conditional* failure rate. To see this more clearly, we can reexpress the hazard as

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

which in words denotes the rate of failure per time unit in the interval $[t, t + \Delta t]$, conditional on survival at or beyond time t. Again, this is a conditional failure rate. Remember from before: this gives the rate of failure conditional on survival. "Given an M.P. has survived four elections, what are the chances she will lose in the next elections?"

As I said earlier, researchers often have a natural interest in risk, thus leading to interest in the hazard rate. Moreover, analysts are commonly interested in asking how the hazard rate—or the risk—of something happening increases or decreases with respect to covariates. In the P.M. example, one might be interested in how survival of a Premier is affected by the distribution of seats among parties in parliament. This may lead one to a "regression-like" strategy where the hazard rate (or some other function) is treated as a function of covariates. It is here that I want to begin the discussion of conventional applied practices in social sciences.

2 Why Event History Analysis?

Duration data present special challenges for statistical models thus rendering traditional linear regression models, for example ordinary least squares (OLS) regression, problematic. Consider a traditional linear regression model

$$y_i = \beta x_i + \epsilon_i, \tag{1}$$

where the response variable y_i denotes the duration time to some event, x_i is a covariate with associated parameter β, and ϵ_i is a stochastic disturbance term. The model in (1) may be estimated by OLS; however, some complications arise. As duration data must be positive, it is often the case that the response variable will exhibit considerable asymmetry, particularly if some observations have exceptionally long duration times. Since in OLS, we estimate the mean response on y_i as a function of covariates, inferences regarding the mean response may be misleading if the response variable is heavily skewed. Predicting negative durations, which is an impossibility, may occur. One common "fix" to this problem is to transform the response variable, for example, by taking the natural log, and then applying OLS. The resultant model,

$$\ln y_i = \beta x_i + \epsilon_i, \tag{2}$$

mitigates the skewness problem, but does not avoid other, more serious problems. We discuss some of these problems below.

2.1 Censoring and Truncation

In general, censoring occurs whenever an observation's full event history is unobserved. Thus, we may fail to observe the termination of a spell, or fail to observe the onset of a spell. In this sense, censored observations are akin to missing data, insofar as the portion of the history that is censored, is in fact, missing.

I'm going to make the assumption that the mechanisms producing the censoring are *independent* of the observed entry times and durations (or independent, conditional on the covariates). In this sense, I assume that the censored history is missing at random and the censoring mechanism is ignorable.

Another way to think of this is that the censoring is non-informative. As Cox and Oakes (1984, p. 5) note, non-informative censoring occurs when "an individual who is censored at c should be representative of all those subjects with the same values of the explanatory variable who survive to c."

2.1.1 Right-Censoring

Right-censoring is commonly observed in event history data sets. Typically, we encounter right-censoring because the time-frame of a study or observation plan concludes prior to the completion or termination of survival times.

- Careers
- Policy Adoption
- Marriage Failure
- . . . and so forth

The ubiquity of right-censoring provides a strong motivation for event history models. To illustrate this more clearly, consider Figure 3 with four hypothetical observations observed over t periods. The last period upon which we observe our sample units is time period t_n. Consider for now observations 1, 2, and 3. Observation 1 is observed until time j, at which time the duration culminates with an event occurrence. Observation 2 is observed up to the last observation period, at which time, an event is observed and the duration is terminated. Finally, observation 3 is observed up to the last observation period; however, at this point, case 3 is still "surviving." When this observation's history will end is unknown to us. Hence, observation 3's history is right-censored.

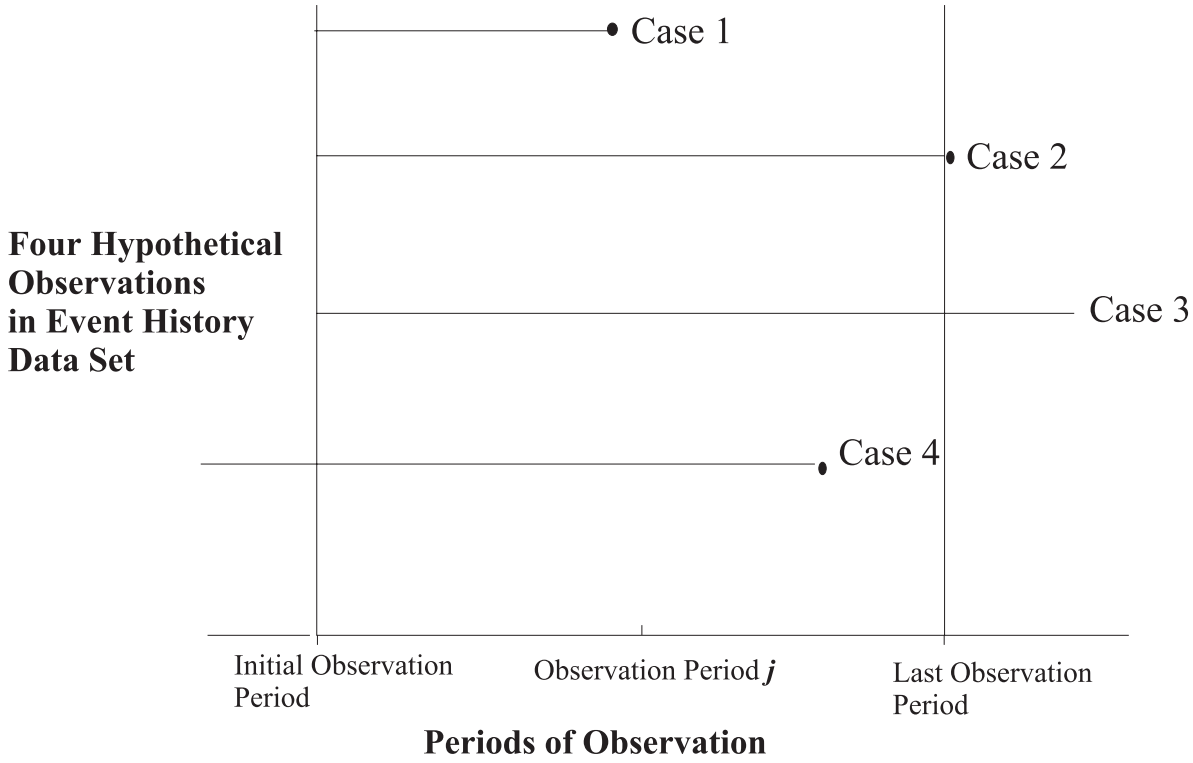

Figure 3: *This figure illustrates examples of right-censoring (Case 3) and left-truncation (Case 4), as well as observations that present no special problems (Cases 1 and 2).*

In the context of a traditional linear model, cases 1 and 2 presents no special problems with regard to censoring. We observe the full history to completion, and as such, all information for both cases is provided in our data set. The problem with right-censoring is

that case 2 and case 3 are treated as having equivalent event histories, when in fact they do not. Observation 2 has experienced an event, observation 3 has not. As noted in the example above, equating such observations is troublesome because the two observations are demonstrably not equivalent.

After all, we are interested in studying the duration times and case 2 has ended while case 3 has not. Case 3 may fail the next day, week, year, or never fail. Further, although systematic differences across observations may be accounted for by covariates, if censored and uncensored cases are treated equivalently, then parameter estimates of these covariates may be misleading (that is, the relationship between the covariates and the duration times may be under- or over-stated).

2.1.2 Left-Truncation

In contrast to right-censoring, some observations in an event history data set may be truncated. In Figure 3, observation 4 is an example of this. Although entering the observation period at time t_1, observation 4 has already amassed "history." Thus, this observation has a portion of its duration time that is unobserved prior to the onset of the observation period. Left-truncation emerges in event history data sets when it is infeasible, intractable, or impossible to determine precisely (in terms of "clock time") when a unit enters the process under study.[1] In the context of the regression model in (2), left-truncated observations are treated has having equivalent entry times as all non-left-truncated observations, even though they do not.

In contrast, event history methods can accommodate the presence of left-truncated observations. In some instances, however, accounting for left-truncation in the duration model framework is nontrivial. Nevertheless, as a general matter, right-censoring and left-truncation problems provides one impetus for the use of event history methods in political analysis.

2.1.3 Accounting for Censoring

The question, then naturally arises as to how event history methods overcome (or certainly help to mitigate) the problems of censoring? To understand this, it is instructive to return to the hazard rate. Formally, the hazard rate, $h(t)$, of a random variable is the ratio of the probability density to the survival function at value t. This is easily seen in (**??**). To illustrate further, it is also noted that $h(t)$ may be equivalently expressed in terms of the integrals of $f(t)$ and $S(t)$:

$$h(t) = \frac{\int_t^{t+\Delta t} f(u)du}{\int_t^{\infty} f(u)du}. \tag{3}$$

The area bounded by the definite integral in the numerator gives the probability of unconditional failure, while the denominator gives the area (or probability) of survival. The hazard thus gives the ratio of unconditional failure to survival. In practical applications, the upper

[1] I refer to left-censoring as observations that experience an event prior to the beginning of the study and thus are not part of the study.

limit of the integral of $S(t)$ is generally not ∞, because this limit is usually known to be some value $t = C_i$, where C_i denotes a *known right-censoring point* for the ith unit. That is, by the time the observation plan ends, some units may be right-censored and so the integral of $S(t)$ is thus given by

$$S(t) = \int_{t}^{t=C_i} f(u)du. \tag{4}$$

For left-truncated observations, the true entry time into the process may occur before the initial observation point, t. This suggests that the lower limit of the integral of $S(t)$ in (3) will, in reality, be a period occurring prior to t, but unobserved in the observation plan. We can think of the survivor function for left-censored cases as being

$$S(t) = \int_{t_L}^{\infty} f(u)du, \tag{5}$$

where t_L denotes the time period, i.e. the clock time, that the left-truncated observation actually enters the observation period.

The presentation of the survivor functions in (4) and (5) is nonstandard; however, it serves a useful pedagogical purpose. It is clear from (4) that right-censored cases *only* contribute information on survival up to the known censoring point, C_i; no information is contributed to $f(t)$, the numerator in (3). Likewise, by (5), it is clear that left-truncated observations *only* contribute information to the survivor function from point t_L forward.

If the censoring points are known in the data, it is possible to construct a likelihood function to accommodate censoring. To illustrate, suppose we have n observations on which the full duration time, t, is observed. In this case, there are no censored observations. To derive a likelihood for the sample, we need only specify a probability density function, $f(t)$. The likelihood of the sample under these conditions is given by

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

However, suppose that in the sample, some observations are right-censored. For censored cases, the duration, t_i, is observed only up to the last observation period t^*, after which, the duration continues, but is unobserved. The observed duration for right-censored cases is t_i^*, denoting that the duration time for the ith censored case is equal to the time of the last observation period (even though it is continuing beyond t^*). For uncensored observations, the full duration time is observed within the observation period, $t_i \leq t^*$. Under these conditions, it is clear that uncensored cases contribute information regarding failure times (as the event of interest is experienced), while censored observations only contribute information on survival. This suggests that the likelihood of the sampled observations will consist of two parts: the density of failure times, $f(t)$, and the survivor function, $S(t)$:

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

This likelihood function can be rewritten to explicitly show how censored and uncensored cases are treated. To see this, let us first define a censoring indicator, δ_i, in the following way:

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

When $\delta_i = 1$, the observation is uncensored; when $\delta_i = 0$, the observation is right-censored. Incorporating our knowledge of δ_i into the likelihood function, the likelihood of the sampled duration times may be expressed as

$$\mathcal{L} = \prod_{i=1}^{n} \Big\{ f(t_i) \Big\}^{\delta_i} \Big\{ S(t_i) \Big\}^{1-\delta_i}, \tag{6}$$

thus illustrating the point that censored duration times contribute to the overall likelihood only through $S(t)$, while uncensored duration times contribute to the overall likelihood through $f(t)$. Consequently, it becomes clear how event history methods can account for both censored and uncensored observations.

This approach is clearly preferable to the alternative "solution" of omitting censored cases from the data set. By deleting censored cases from the analysis, we may be induce a form of case selection bias into the results. If censored cases are systematically different from uncensored cases, then simply deleting the latter cases will produce a nonrepresentative "sample" and render coefficient estimates biased due to the case selection process. The implications of selection bias are well known, and should be avoided (Achen 1986, Geddes 1990, King, Keohane, and Verba 1994).

A "fix" to this problem is to convert the response variable to a binary indicator, and then model the likelihood (using logit or probit) that a spell will terminate. This approach, as Petersen (1995) notes, is troublesome, because it belies the logic of duration modeling; usually, we are concerned *both* with the occurrence or nonoccurrence of some event *as well as* the length of time the unit survived until the event occurred. This strategy precludes this kind of information.

2.2 Time-Varying Covariates

Another complication with the traditional regression-based approach shown in (2) is the inability of the model to account for covariates having values that change over time, that is, time-varying covariates. Analysts are frequently interested in TVCs; however, the regression model discussed earlier implicitly treats all covariates as if they are time-invariant. This is a substantial limitation of the OLS model, and thus provides another motivation for event history methods. In the context of event history modeling, TVCs may be readily incorporated into the analysis in a variety of ways. In the P.M. example, we saw how TVCs induce the "jump process" interpretation.

3 General Issues Regarding Modeling Strategies

The shortcomings of the regression model provide a statistical motivation for event history analysis. In recognizing that the duration modeling approach can overcome the obvious problems with the traditional linear model, our path is set toward consideration of a wide variety of statistical modeling choices. Some general issues remain to be discussed, however, before we consider specific kinds of event history models. In particular, I next discuss the issue of "continuous-time" versus "discrete-time" processes. Following this, we distinguish between parametric and nonparametric duration models.

3.1 Continuous-Time and Discrete-Time Processes

As alluded to earlier, models for event history data are frequently motivated by questions pertaining to risk. Event history data contain information allowing the researcher to assess the risk by considering event occurrences as well as the length of survival times prior to the event (if an event occurs). In terms of probability, analysts are therefore interested in determining the following quantity:

$$\Pr(t \leq T \leq t + \Delta t \mid T \geq t), \tag{7}$$

that is, the probability a duration T will end in some time interval given that the duration has persisted up to or beyond some time t.

If the random variable T in (7) is assumed to be continuous, this implies that change (or the event's occurrence) may occur anywhere in time. As such, by the definition of a continuous random variable, the probability in (7) is 0. Therefore, for a continuous-time event history process, T is assumed to be absolutely continuous. By dividing the probability in (7) by Δt, a ratio is established of the probability per time unit *to* the time unit. If the limit of this ratio is taken as Δt approaches 0, then we obtain

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

which of course is the familiar hazard rate for a continuous random variable shown in (**??**).

In contrast, if the random variable T is assumed to be discrete, this implies that change occurs at some observable time point. For example, legislative turnover in the United States Congress may be thought of as a discrete-time process: electoral turnover can only occur on (a Constitutionally mandated) election day. Event history models presuming a discrete-time process specify the probability in (7) directly. Therefore, the hazard "rate" for a discrete-time process is

$$h(t) = \Pr(T = t \mid T \geq t),$$

which is a probability.

Distinguishing continuous-time from discrete-time processes is only important insofar as

the quantities of interest (the hazard rate, survivor function, density or probability mass function) in event history analysis are defined differently, mathematically, thus leading to different kinds of modeling strategies. Nevertheless, while the distinction between the two kinds of processes may seem clear-cut, in fact, they are not.

While many processes may be absolutely continuous, techniques for observation and/or measurement fail to approximate the continuous nature of change. Data for continuous-time processes often are collected at discrete intervals, for example, by fiscal periods, months, quarters, or even years. Change may occur anywhere in the interval, but the data are only "observed" at predefined periods. So while the continuous-time process presumes knowledge of when change occurred in time, we sometimes only have an approximation as to when the change or transition actually occurred. Obviously, this kind of measurement problem is not solely confined to event history data, but special care needs to be paid by the analyst when defining the measurement units of T, if T is treated continuously.

Moreover, some longitudinal processes may conceivably be continuous-time processes, but knowledge of precisely when in time change occurred is largely unimportant. To illustrate, consider an example of state adoption of public policy. Presumably, a legislature could adopt a policy anytime within a legislative session. Because state legislatures routinely record votes, we could easily discern precisely when change occurred. In most analyses of state policy adoption, however, the crucial issue is not knowing exactly when adoption (the "event") occurred within a legislative session, but rather when adoption occurred relative to other states (c.f. Berry and Berry 1994).

In such analyses, the year in which a policy was adopted may be sufficient to mark the occurrence of an event. Therefore, while policy adoption may be a continuous-time process in principle, a discrete-time model may be suitable to the research question.

3.2 Parametric and Nonparametric Modeling Strategies

Apart from characterizing event history processes as being continuous or discrete, another way to think about the range of event history models is in terms of whether or not the distribution of failure times (or survival times) is specified. If this distribution is unspecified, the model (or statistic) is *nonparametric* or "distribution-free." In these kinds of models, the shape of the hazard rate, or analogously, the form of time dependency, is not directly specified (though nonparametric estimates *can* be obtained). In contrast, if the distribution of failure times is specified by a distribution function, then the model can be said to be parametric. Parametric approaches may be desirable if one has a theory about the distribution of failure times in the data.

4 Usual Approaches to Modeling Duration Data

To summarize the usual strategies applied to duration data, it would look something like this:

• 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.

I now want to consider each of these approaches in turn. Because of obvious time constraints, my presentation will be "survey-esque" in parts. I'm mostly concerned with getting across the major points. Let me pick up with parametric duration models.

5 Parametric Models

The logic of a parametric model is to define the hazard rate as increasing:

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

(which says the risk of an event occurrence (or failure) is increasing as the spell increases in length (sometimes referred to as "positive duration dependence"); as decreasing:

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

(which says the risk of an event occurrence is decreasing as the spell increases in length (sometimes referred to as "negative duration dependence"); as flat:

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

(which says that the risk of an event occurrence is invariant to the length of the spell; i.e, the risk is "flat" over time); or as increasing and then decreasing (nonmonotonic).

Because there are many distribution functions that are suitable for duration data, there are many kinds of parametric duration models. They all share the common feature that the *baseline hazard rate has a certain kind of functional form.*

Popular choices include the:

• Weibull (under which the exponential is nested)
• Log-Normal
• Log-Logistic
• Gamma

- Gompertz
- Rayleigh
- . . . and many other potential candidates.

Further, most of these are easy to implement in software packages like Stata, Limdep, SAS, R, or S-Plus.

Under a parametric model, ancillary parameters are estimated that give the "shape" of the time dependency in the data. From these parameters, nice smooth graphs can be constructed of the hazard function. To illustrate, consider Figure 4. Here I plot the functional form for several common parametric distribution functions.

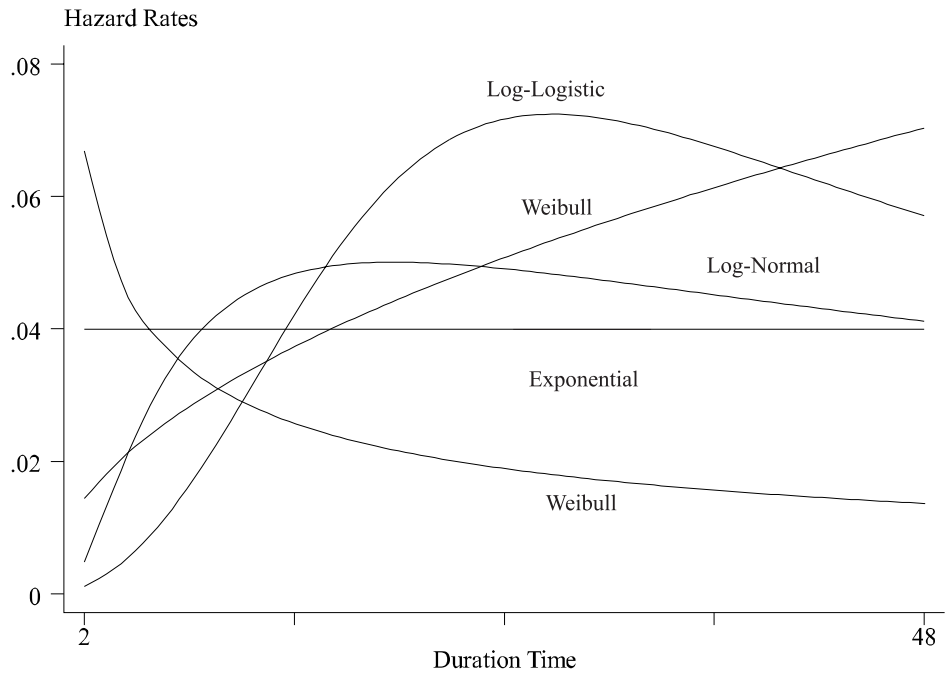

Figure 4: *This figure graphs typical functional forms for several common parametric distribution functions.*

Note that the Weibull produces *only* monotonic hazards (nonmonotonic hazards cannot be obtained through a Weibull); the exponential gives a "flat hazard;" the log-normal and log-logistic produce nonmonotonic hazards, which seems useful. Sometimes analysts justify the use of a log-normal or log-logistic on the grounds that it is more flexible than the Weibull; it is not. All three are two-parameter distributions for log-linear T; once the mean and variance are estimated, its shape is fixed.

In principle, there is absolutely nothing wrong with application of these kinds of models; indeed, they have been widely applied in social science settings. If the characterization of

the time-dependency is accurate, then parameter estimates will in general be more precise than estimates from models where the time dependency is unspecified (Collett 1994).

But what if the time dependency is misspecified? For example, what if the time dependency is nonmonotonic, but a Weibull is estimated? Misleading conclusions can be generated about the underlying risk simply because the underlying risk is incorrectly specified. As Larsen and Vaupel (1993) write, "In the analysis of duration data ... if the functional form of the hazard has the wrong shape, even the best-fitting model may not fit the data well enough to be useful."

The standard practice, at least in political science settings, has been an overreliance on the Weibull. This stems from computational ease in estimating the model. In the last few years, we've see more applications using the log-normal or log-logistic, but I continually see analysts inappropriately justifying these distribution functions on the grounds that they are "more flexible." As noted above, they are not.

Further, a bigger problem, in my view, has emerged regarding the use of parametric models in the social sciences and the problem roughly revolves around an "over-interest" in the shape of the time-dependency. It is often argued that "theory" should dictate the model that is selected. This, of course, is true.

This dictum has been extended, in the duration modeling framework, to covering time-dependency—that is, one's theory about the underlying time-dependency should help lead one to a certain kind of parametric model.

But let's understand what time-dependency *really* means. Time-dependency, in some sense, is the "left-over" effects of time on the hazard rate after one has conditioned the hazard on covariates. In principle, in a perfectly specified model, *there would be no time dependency because $h(t)$ would be fully characterized by covariates.* That is, the hazard will tend to the exponential, as a model becomes more fully specified.

Of course models are seldom perfectly specified and so there is almost always left-over time dependency. This occurs in just about all models. But in my view, the role of theory should come into play in positing a sensible model through the inclusion of well defined and hopefully well measured covariates.

As duration dependency, in a very real sense, is a *nuisance*, selecting parametric models on the grounds that there exists a theory about this nuisance seems an awkward way to proceed. The basic point is that ascribing substantive interpretations to the ancillary parameters in fully parametric models is, in my view, tenuous.

6 Estimation

If the hazard rate is treated as having a dependency on time *as well as covariates*, then we can reexpress the hazard rate as

$$h(t \mid \mathbf{x}) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T \leq t + \Delta t \mid T \geq t, \mathbf{x})}{\Delta t} \tag{8}$$

to accommodate this dependency. In (8), $\mathbf{x}$ are the covariates and the other terms are defined as before. Various statistical models may be constructed to describe (8).

Estimation of parametric models is, in principal, fairly straightforward and is achieved through maximum likelihood estimation. Any "survival" distribution has a characteristic signature called a distribution function (i.e. $F(t)$). Simply define the distribution function, write out the likelihood function, and maximize the likelihood (or the log of it) and you have your estimates. Sounds easy! (Perhaps too easy!)

There are a variety of distribution functions, each of which that can be used to describe the hazard function in different ways. Consider the "popular" choices:

Exponential

The risk of an event occurrence under the exponential is flat with respect to time. This implies that the *hazard rate* is constant (i.e. takes the same value at all time points). If such an expectation were warranted (and note that it usually will not be), then the hazard rate can easily be characterized as,

$$h(t) = \lambda \qquad t > 0, \lambda > 0, \tag{9}$$

where λ is a positive constant. The hazard, as expressed in (9), is "flat" in that the risk or the rate of event occurrences is equal to λ, and is the same at all observation points. Furthermore, if $h(t)$ is expressed as above, the survivor function, $S(t)$ and density function, $f(t)$ are defined from some results we saw earlier as

$$S(t) = \exp^{-\lambda(t)}, \tag{10}$$

and

$$f(t) = \lambda(t) \exp^{-\lambda(t)}. \tag{11}$$

If the density function is specified as in (11), then the duration time T has an exponential distribution with mean λ^{-1}. Graphically, the hazard rate plotted against time would look like that shown in Figure 5. The location of $h(t)$ on the y-axis in Figure 5 is fully determined by the value of λ.

Weibull

The exponential hazard is flat with respect to time. A more flexible alternative is one that allows $h(t)$ to monotonically increase or decrease. This gives rise to the Weibull. The hazard

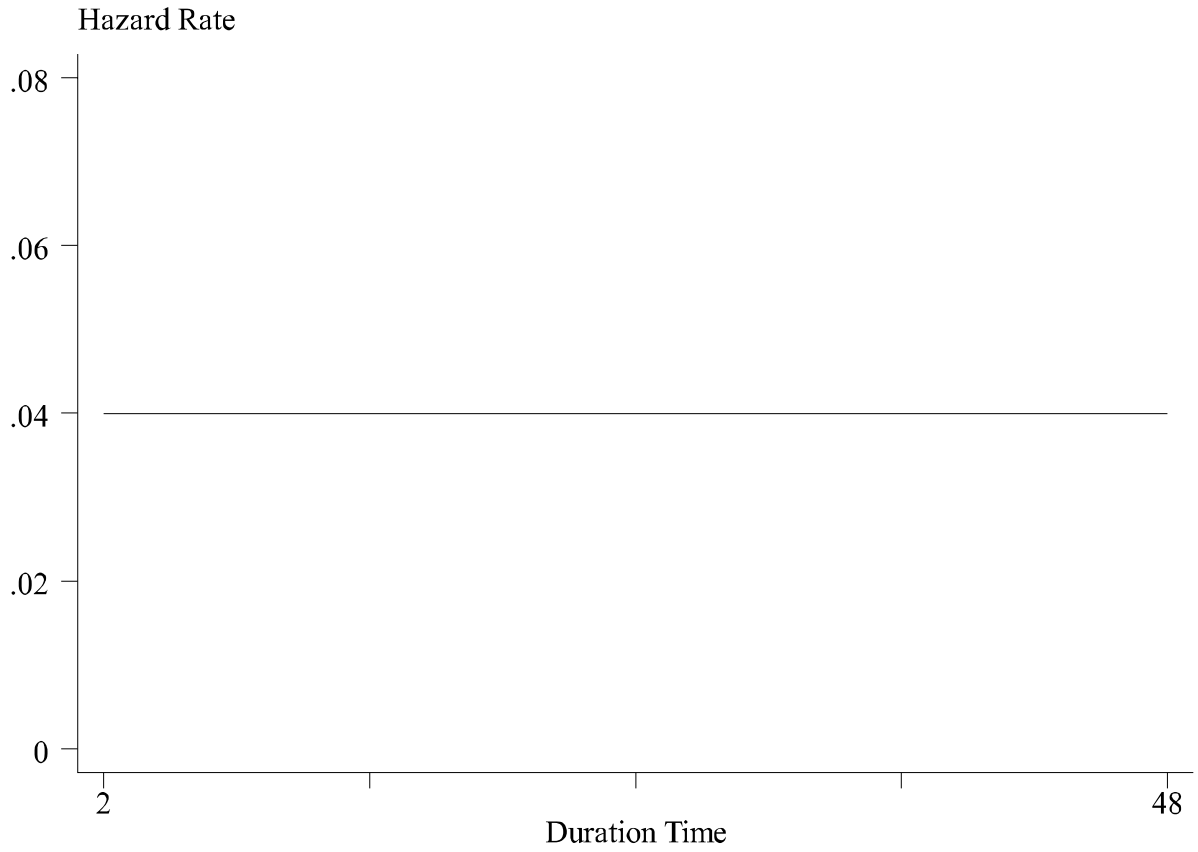

Figure 5: *This figure graphs a typical example of the exponential hazard rate.*

rate for the Weibull distribution is

$$h(t) = \lambda p(\lambda t)^{p-1} \qquad t > 0, \lambda > 0, p > 0, \tag{12}$$

where λ is a positive scale parameter and p is the shape parameter. The p term gets its name because the shape of the hazard rate depends on the value of this term.

- When $p > 1$, the hazard rate is *monotonically* increasing with time.
- When $p < 1$, the hazard rate is *monotonically* decreasing with time.
- When $p = 1$, the hazard is flat, taking a constant value λ.

For the case of $p = 1$, the hazard rate has an exponential distribution thus demonstrating the point that the exponential is merely a special case of the Weibull (i.e. it is encompassed by the Weibull [or nested within the Weibull]). The Weibull distribution is more flexible than the exponential since it is a function of two parameters, λ and p, and not a single parameter. Figure 6 shows the monotonicity of the Weibull hazard rate for p=1.5 and p=0.5 and the flat hazard rate given by p=1.0 (thus producing the exponential).

The survivor and density functions for the Weibull are, respectively,

$$S(t) = \exp^{-(\lambda t)^p}, \tag{13}$$

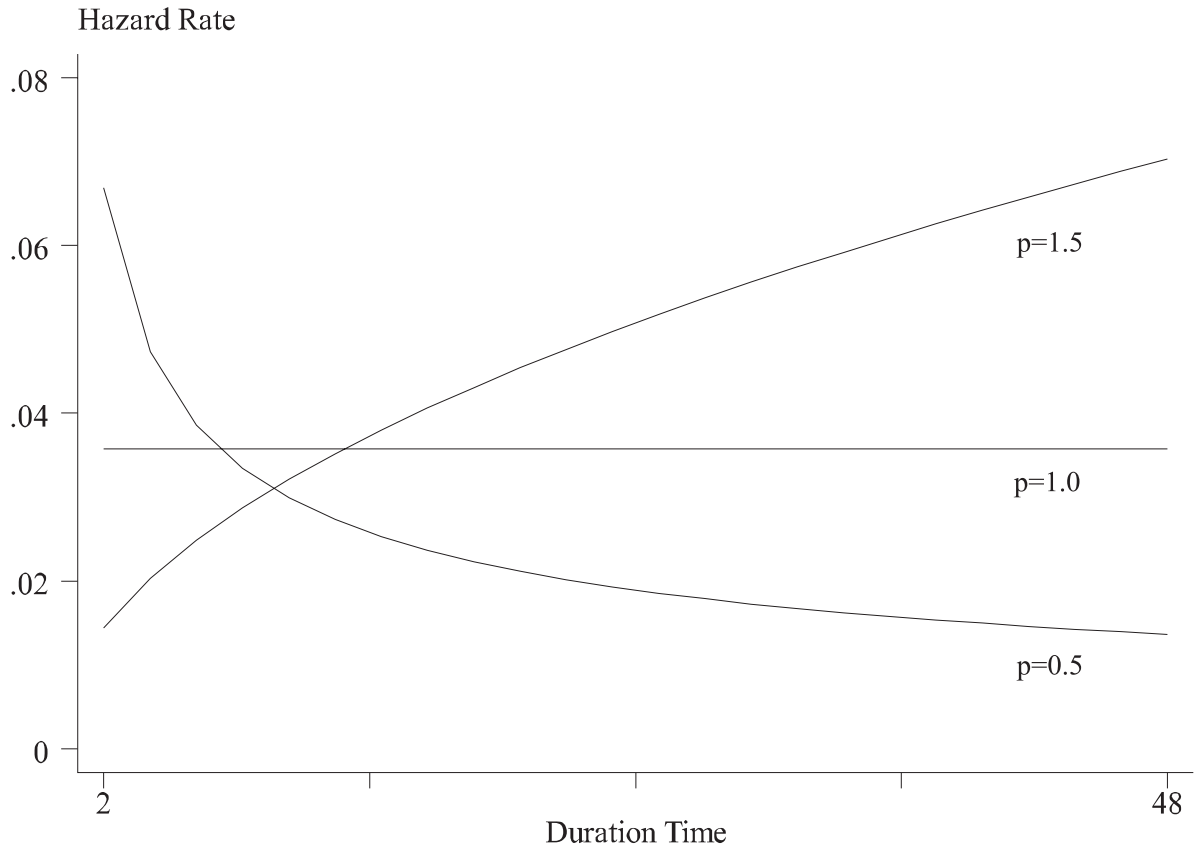

Figure 6: *This figure graphs three typically shaped Weibull hazard rates. Note the monotonicity of the Weibull hazard; note also that when the shape parameter is 1, the exponential hazard is obtained.*

and

$$f(t) = \lambda p(\lambda t)^{p-1} \exp^{-(\lambda t)^p}. \tag{14}$$

A Sidenote: Accelerated Failure Time Parameterization

Most software packages give users the choice in how survival model parameters are presented: either as hazard ratios or in terms of "accelerated failure rates." Let's briefly dispense with this issues. A common way to parameterize the Weibull model in terms of covariates is by constructing a linear model for $\log(T)$. This involves the specification of a log-linear model and treating the log of the survival times as the response variable. This parameterization is convenient for it permits easier comparison with the traditional linear model. The Weibull expressed as a log-linear model has the following form:

$$\log(T) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_j x_{ij} + \sigma\epsilon, \tag{15}$$

and in vector notation,

$$\log(T) = \beta_j' \mathbf{x} + \sigma\epsilon \tag{16}$$

where β_j are the regression coefficients, x_{ij} are time independent covariates, ϵ is a stochastic disturbance term with a *type-1 extreme-value* distribution scaled by σ, which is equivalent to $1/p$. Note that $F(\epsilon)$ in this parameterization is a type-1 extreme value distribution. There is a close connection between the Weibull distribution and the extreme value distribution. Specifically, the distribution of the log of a Weibull distributed random variable gives rise to the type-1 extreme value distribution. Hence, the log of a Weibull random variable is a type-1 extreme value random variable (indeed, sometimes this parameterization is referred to as a log-Weibull distribution). Because the regression parameters in (16) are expressed in terms of the log of the duration times, the coefficients convey information regarding expected failure times. For this reason, this model is sometimes referred to as a "log expected failure time model" or more conventionally as an *accelerated failure time* (AFT) model.

In contrast, the Weibull may be written in terms of the hazard rate:

$$h(t \mid \mathbf{x}) = h_{0t} \exp(\alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_j x_{ij}), \tag{17}$$

where h_{0t} is the baseline hazard rate and is equivalent to $\exp(\alpha_0) p t^{p-1}$ (where α_0 is the parameter for the regression constant term). We use α to denote the parameters in order to distinguish this model from the model given by (16). Sometimes this parameterization is called a "proportional hazards" model. Nevertheless, please note that these are *identical* models. The parameters in one are a simple transformation of the parameters in the other.

The Weibull (and by extension, the exponential) is unique among parametric models for duration data, for it is the only distribution that is *both* a proportional hazards *and* an AFT model. To see this, note that there is a direct correspondence between the parameters in (16) and (**??**). This correspondence is summarized in the table below. The top portion of the table, in the first two columns, presents the parameters of the proportional hazards and AFT models respectively. The third column shows the mathematical relationship between the parameters of the two models. The metric of the covariate parameters for the two models are rescaled by the shape parameter of the distribution and the interpretation of them differs, depending on the model that is estimated (see the last two columns of the table **??**). Under the proportional hazards model, a positively (negatively) signed coefficient implies that for changes in the value of covariate x_{ij}, the hazard rate, or "risk," of an event occurrence (or duration terminating) increases (decreases). In this sense, a positive (negative) coefficient implies the covariate is associated with shorter (longer) duration times, relative to the baseline. For the AFT model, a positively (negatively) signed coefficient implies that for changes in the value of covariate x_{ij}, the expected log duration time increases (decreases). This suggests that a positive (negative) coefficient implies longer (shorter) expected duration times. The two parameterizations produce different kinds of information about the data. Consequently, it is important to be aware of the parameterization used by the software package.

P.H. Parm.	A.F.T. Parm.	Relationship Between Parameters	Interp. of P.H. Parm.	Interp. of A.F.T. Parm.
α	β	$\beta = \frac{-\alpha}{p}$	$+\alpha \equiv \uparrow h(t \mid x_{ij})$	$+\beta \equiv \uparrow \log(T)$
		$\alpha = -\beta p$	$-\alpha \equiv \downarrow h(t \mid x_{ij})$	$-\beta \equiv \downarrow \log(T)$
p	σ	$\sigma = \frac{1}{p}$	$p > 1 \equiv \uparrow h(t \mid x_{ij})$	$\sigma > 1 \equiv \downarrow h(t \mid x_{ij})$
		$p = \frac{1}{\sigma}$	$p < 1 \equiv \downarrow h(t \mid x_{ij})$	$\sigma < 1 \equiv \uparrow h(t \mid x_{ij})$

The shape parameters of the two models are also sometimes presented differently depending on the software used. The relationship between the parameters p and σ are shown in the bottom part of the table in the third column. In the last two columns, the interpretation of these parameters, in terms of the hazard rate, is presented. If the shape parameter is reported in terms of p, then the hazard is increasing if $p > 1$ (recall Figure 6). If the parameter is reported in terms of the extreme-value scale parameter, σ, then the hazard is *decreasing* if $\sigma > 1$. Clearly, since the inverse of p produces σ, these are equivalent statements; nevertheless, it is important that the researcher understand how the parameters are reported from his or her software package. (If either $p = 1$ or $\sigma = 1$, then the hazard rate is flat.)

7 The Log-Logistic and Log-Normal Models

What about distributions that permit a non-monotonic hazard function? (The Weibull is strictly monotonic). Two parametric models most commonly used are the log-logistic and log-normal. A duration model using the log-logistic or log-normal models are only defined in terms of the log-linear (i.e. AFT) parameterization:

$$\log(T) = \beta_j' \mathbf{x} + \sigma\epsilon. \tag{18}$$

This parameterization is similar to the parameterization of the Weibull AFT model that we just talked about. Note that if an extreme value distribution is specified for ϵ, then the implied model is the Weibull. In contrast, if a logistic distribution is specified for ϵ, then the log-logistic model is implied. Finally, if a standard normal distribution is specified for ϵ, then the log-normal model is implied. For each model, the distribution is scaled by the parameter σ, which is equivalent to the inverse of the shape parameter, p, that is p^{-1}. Using the formulation of (18), it is clear that the Weibull, log-logistic, and log-normal models are similar in one respect: they each are two-parameter distributions, i.e., σ and ϵ, for $\log(T)$. This point is important because although the log-logistic and log-normal can produce non-monotonic hazard rates, *neither* of these models are any more flexible than the Weibull model. For the log-logistic, log-normal, *and* Weibull models, once the mean and variance of the distribution are estimated, its shape is fixed.

The hazard rate for the log-logistic can be nonmonotonic and unimodal. To see this, note that the hazard rate for the log-logistic,

$$h(t) = \frac{\lambda p(\lambda t)^{p-1}}{1 + (\lambda t)^p} \tag{19}$$

first increases and then decreases if $p > 1$, but is monotonically decreasing when $p \leq 1$. To demonstrate see Figure 7 where I illustrate the hazard rate for various values of p.

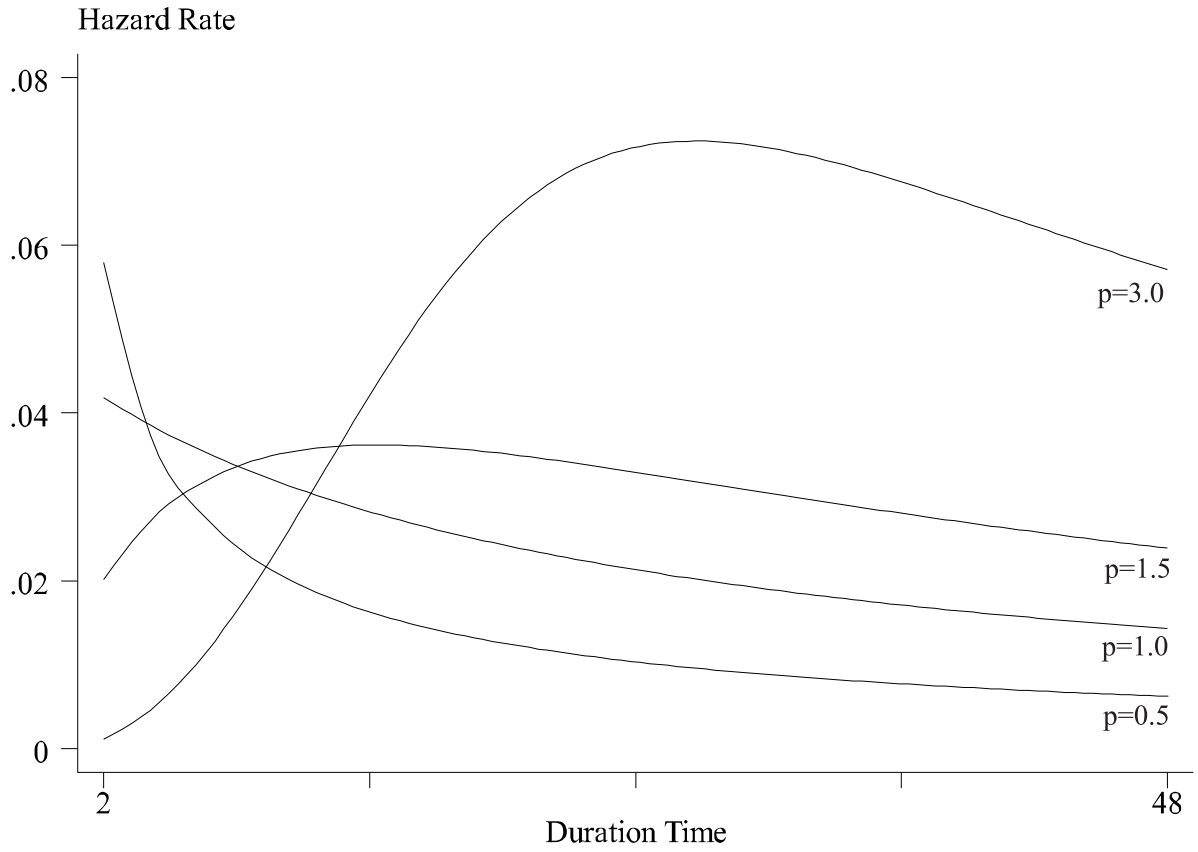

Figure 7: *This figure graphs some typically shaped hazard rates for the log-logistic model.*

The survivor function for the log-logistic model is given by

$$S(t) = \frac{1}{1 + (\lambda t)^p}, \tag{20}$$

while the probability density function is given by

$$f(t) = \frac{\lambda p (\lambda t)^{p-1}}{(1 + (\lambda t)^p)^2}, \tag{21}$$

which is a symmetric density. The logistic density is very similar to the normal density. The similarity between the log-logistic and log-normal model is analogous to the similarity between the logit and probit models that are often applied to models with discrete dependent variables. A logit model is specified in terms of the logistic distribution, while the probit is specified in terms of the standard normal distribution. The similarity across distributions produces the similarity in results given by logit and probit models. The derivation of the log-normal hazard rate is not as straightforward as it is for some of the other models we have

considered. $h(t)$ must be expressed in terms of integrals of the standard normal distribution. The survivor function for the log-normal model can be written as

$$S(t) = 1 - \mathbf{\Phi}\left(\frac{\log(t) - \beta'\mathbf{x}}{\sigma}\right), \tag{22}$$

where $\mathbf{\Phi}$ is the cumulative distribution function for the standard normal distribution and $\beta'\mathbf{x}$ are the covariates and parameter vector from (18). The probability density function for the log-normal model is given by

$$f(t) = \frac{1}{\sigma\sqrt{(2\pi)}}t^{-1}\exp\left[-\frac{1}{2}\left(\frac{\log(t) - \beta'\mathbf{x}}{\sigma}\right)^2\right], \tag{23}$$

where σ is the scale parameter of ϵ from (18). The hazard rate for the log-normal is

$$h(t) = \frac{f(t)}{S(t)}. \tag{24}$$

Noting that $\sigma = p^{-1}$, in Figure 8 we illustrate the nonmonotonicity of the log-normal hazard rate for various values of p. When p is small, the hazard rate rises to its peak very quickly and then falls.

8 Gompertz

The Gompertz distribution is a popular choice for use in demographic research because the hazard rate is treated as an exponential function of the duration times. Hazard rates produced by the Gompertz often do a good job of describing mortality data as well (Kalbfleisch and Prentice 1980). The hazard rate for the Gompertz, which is characterized as either being monotonically increasing, decreasing, or flat, is given by

$$h(t) = \exp^{\gamma t}\exp^{\lambda}, \tag{25}$$

where γ is the shape parameter for the Gompertz distribution and $\lambda = exp(x\beta)$. When $\gamma > 0$, the hazard rate is rising, when $\gamma < 0$, the hazard rate is decreasing, and when $\gamma = 0$, the hazard rate is flat, with respect to time. In Figure 9, we illustrate the form of the Gompertz hazard for these three cases.

The survivor function for the Gompertz distribution is

$$S(t) = e^{-\frac{e^{\lambda}}{\gamma}(e^{\gamma t}-1)}, \tag{26}$$

while the density function is given by,

$$f(t) = e^{(e^{\lambda}e^{\gamma t})-\frac{e^{\lambda}}{\gamma}(e^{\gamma t}-1)}. \tag{27}$$

The Gompertz model is a proportional hazards model; however, it is not an AFT model. If covariates are included in the estimation of the Gompertz model, $\lambda = e^{\beta'\mathbf{x}}$ in (25), which is in contrast to the AFT parameterization, where $\lambda = e^{-\beta'\mathbf{x}}$.

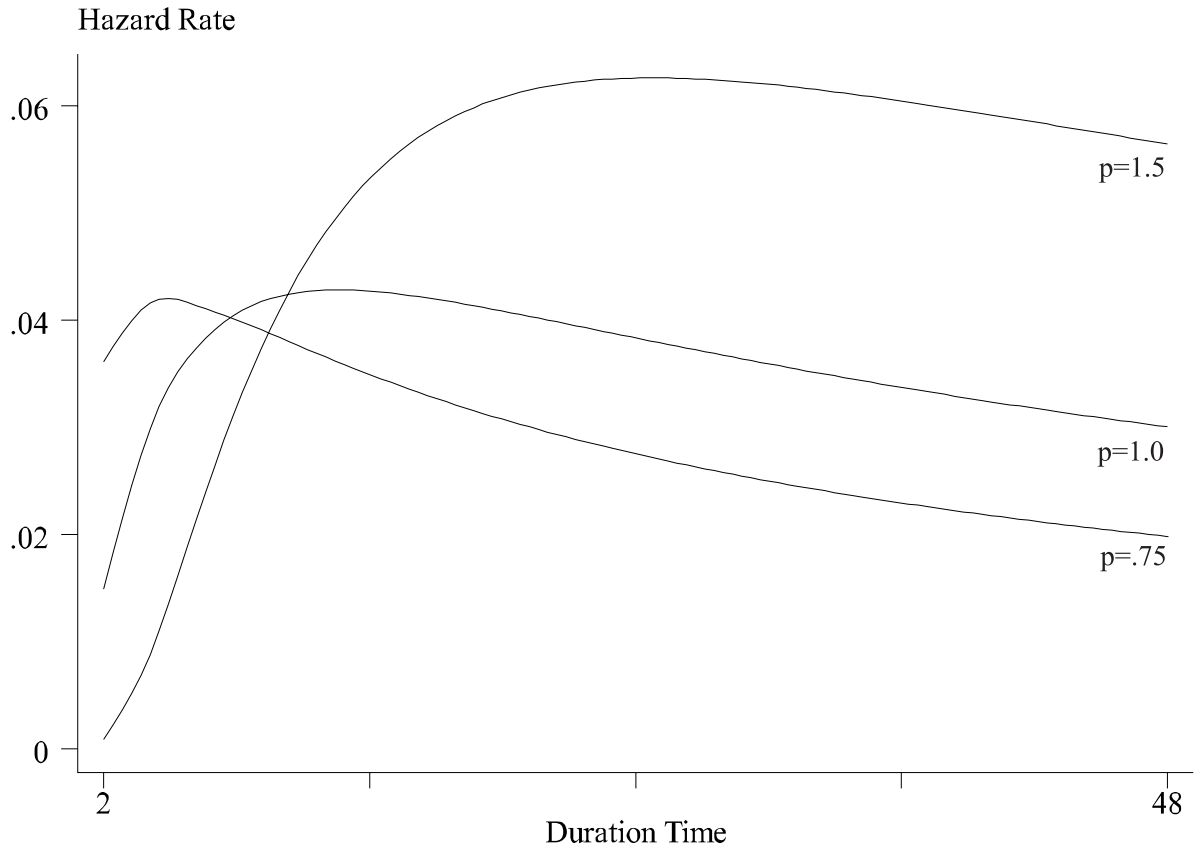

Figure 8: *This figure graphs some typically shaped hazard rates for the log-normal model.*

Others?

The exponential, Weibull, log-logistic, log-normal, and Gompertz distributions are probably the most commonly used in applied research. There are other distribution functions that are permissable for survival data including the Rayleigh and t distributions. I won't discuss these. It would be somewhat redundant.

Likelihood

The models discussed today can all be estimated by maximum likelihood (MLE). Suppose that we have n observations upon which we observe $t_1, t_2, \ldots t_n$ duration times. Furthermore, assume that these observed durations are independent (conditional on any covariates) realizations from the random variable T. Some duration times may be censored or uncensored; however, we assume that the censoring is non-informative.

Deriving an MLE entails specifying a probability density function for the random variable T. In so-doing, we assume that the probability of the t observed durations times can be derived from this probability density function. Hence, not only do we assume the observations are independent, but they are identically distributed, with the distribution being defined by the

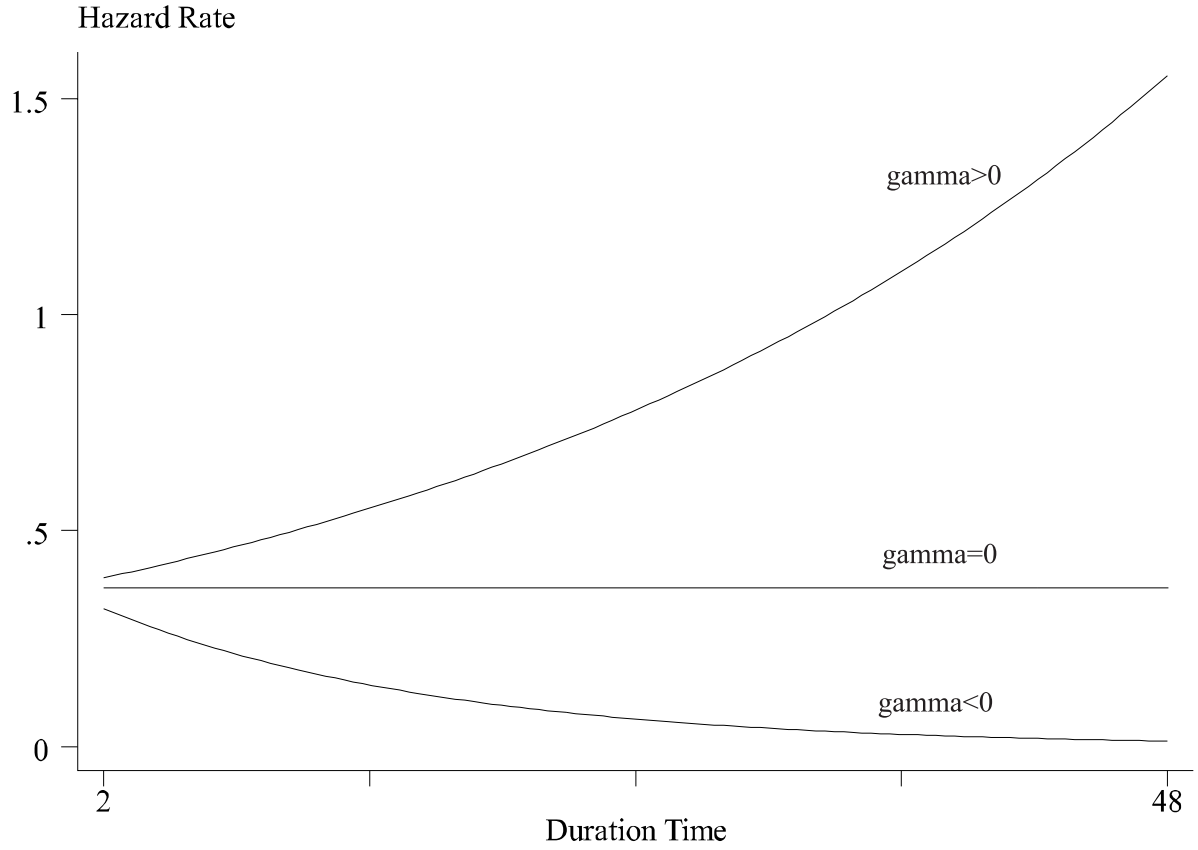

Figure 9: *This figures graphs some typically shaped hazard rates for the Gompertz model.*

density function.

Specifying a probability density function entails defining $f(t)$ for the data. From this, the survivor function, $S(t)$ can be retrieved. The likelihood function for event history data is a function of two components: the density of failure times, $f(t)$, and the survivor function, $S(t)$. The likelihood function is complete when the probability density function is defined for each observation. Hence, the general likelihood function for event history data is

$$\mathcal{L} = \prod_{i=1}^{n} \Big\{ f(t_i) \Big\}^{\delta_i} \Big\{ S(t_i) \Big\}^{1-\delta_i}, \tag{28}$$

where δ is a censoring indicator denoted 0 if the observed duration is censored, and 1 if it is not. In practice, deriving the likelihood implies defining $f(t)$ in terms of a distribution. So, for example, in the case of the Weibull distribution,

$$f(t) = \lambda p (\lambda t)^{p-1} \exp^{-(\lambda t)^p},$$

and the survivor function is

$$S(t) = \exp^{-(\lambda t)^p}.$$

The likelihood of the t duration times is obtained by substituting these functions into (28):

$$\mathcal{L} = \prod_{i=1}^{n} \left\{ \lambda p(\lambda t)^{p-1} \exp^{-(\lambda t)^p} \right\}^{\delta_i} \left\{ \exp^{-(\lambda t)^p} \right\}^{1-\delta_i}. \tag{29}$$

Through (29), it is easy to see that the likelihood of the duration times is solely a function of two parameters: λ and p. For a model like the exponential, the likelihood is a function of a single parameter, λ:

$$\mathcal{L} = \prod_{i=1}^{n} \left\{ \lambda(t) \exp^{-\lambda(t)} \right\}^{\delta_i} \left\{ \exp^{-\lambda(t)} \right\}^{1-\delta_i}, \tag{30}$$

which simplifies to,

$$\mathcal{L} = \prod_{i=1}^{n} \lambda^{\delta_i} \exp^{-\lambda(t)}. \tag{31}$$

Noting that $\lambda = e^{-\beta \mathbf{x}}$, it is straightforward to see how covariates are included in the likelihood function. The maximum value of this function gives the parameter estimates that maximize the likelihood of the observed data, hence the name maximum likelihood.

However, it is generally easier to solve the likelihood function by maximizing the log of the likelihood function. Continuing with the exponential distribution, the log-likelihood function is obtained by taking the logarithm of (30):

$$L = \sum_{i=1}^{n} \delta_i \log \lambda - \lambda \sum_{i=1}^{n} t_i. \tag{32}$$

The maximum likelihood estimate for the model is the value of λ for which the log-likelihood function is maximized. To find this, we differentiate the log-likelihood with respect to λ, which gives

$$\frac{\partial \log L}{\partial \lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\lambda} - \sum_{i=1}^{n} t_i. \tag{33}$$

Setting the derivative in (33) to 0 and evaluating λ at this point gives

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i}, \tag{34}$$

which is the maximum likelihood estimator of λ. When λ consists of several covariate parameters, there are as many likelihood equations to solve as there are parameters to estimate. For most problems, there is no analytical solution for these equations and so iterative procedures must be used. Under certain regularity conditions, maximum likelihood estimates are asymptotically normal, and so the usual kinds of hypothesis testing is generally possible with the estimates. Moreover, maximum likelihood estimates are consistent and asymptotically efficient, again, under certain regularity conditions.

9 Some Illustrations

To start with, I'll be using the UN Peacekeeping data that is available on the website. Let's consider first the exponential model. The syntax "dist(.)" is used to specify the distribution function, in this case, the exponential (i.e. `"exp"`). The option `"nohr"` tells Stata to provide the "proportional hazards" parameterization (see book and/or lecture notes). The `"nohr"` option gives the coefficient estimates in terms of proportional hazards. Because the coefficients are expressed in terms of hazards, a positive coefficient tells us that the risk of an event is increasing with the changes to the value of the covariate; a negative coefficient tells us the risk is decreasing with changes to the value of the covariate.

In Stata, I obtain the exponential in the following way:

```
. streg civil interst, dist(exp) nohr

        failure _d:  failed
  analysis time _t:  duration

 Iteration 0:   log likelihood = -103.03289
 Iteration 1:   log likelihood = -90.211473
 Iteration 2:   log likelihood =  -86.44131
 Iteration 3:   log likelihood = -86.354656
 Iteration 4:   log likelihood = -86.354481
 Iteration 5:   log likelihood = -86.354481

Exponential regression -- log relative-hazard form

 No. of subjects =           54         Number of obs = 54
 No. of failures =           39         Time at risk    = 3994
                                        LR chi2(2)      =     33.36
Log likelihood  =   -86.354481          Prob > chi2 = 0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       civil |   1.169344   .3588703     3.26   0.001     .4659714    1.872717
     interst |    -1.6401   .4954337    -3.31   0.001    -2.611132   -.6690679
       _cons |  -4.350864   .2132007   -20.41   0.000     -4.76873   -3.932999
------------------------------------------------------------------------------
```

Under this parameterization, exponentiating the coefficients yields the hazard ratios. Thus for the civil war coefficient, I obtain:

```
. display exp(_b[civil]) 3.2198805
```

We'll discuss this result in a moment.

I could also reparameterize these estimates in terms of $\log(T)$, or equivalently, in terms of accelerated failure times. This is done by including the `time` option in my `Stata` command:

```
. streg civil interst, dist(exp) time

        failure _d:  failed
  analysis time _t:  duration

 Iteration 0:   log likelihood = -103.03289
 Iteration 1:   log likelihood = -90.211473
 Iteration 2:   log likelihood =  -86.44131
 Iteration 3:   log likelihood = -86.354656
 Iteration 4:   log likelihood = -86.354481
 Iteration 5:   log likelihood = -86.354481

Exponential regression -- accelerated failure-time form

 No. of subjects =           54         Number of obs =        54
 No. of failures =           39
 Time at risk    =         3994
                                                 LR chi2(2)      =     33.36
 Log likelihood  =   -86.354481                  Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       civil |  -1.169344   .3588703    -3.26   0.001    -1.872717   -.4659714
     interst |     1.6401   .4954337     3.31   0.001     .6690679    2.611132
       _cons |   4.350864   .2132007    20.41   0.000     3.932999     4.76873
------------------------------------------------------------------------------
```

Note that for the exponential, the parameters of the AFT are simply -PH (where "PH" denotes the proportional hazards parameters from the previous model).

The sign shift makes sense: since AFT is expressed in terms of $\log(T)$, a positive coefficient implies the survival time is increasing with changes to the covariate; a negative coefficient implies the survival time is decreasing. Thus, interpretation of the coefficients in the AFT model are exactly *opposite* to that of the PH model (though either model must yield identical conclusions). This result will obviously hold for the Weibull (why?).

Thus, the hazard rate for civil wars using the AFT estimate is now

```
. display exp(-_b[civil])
3.2198805
```

which is identical to the PH estimate.

What does this quantity reveal? To understand, let's discuss an important property of the exponential (and Weibull and Cox models): proportional hazards.

To demonstrate this important property of the exponential model consider the following relationship (here, as in the book, I'm using the AFT parameterization):

$$h(t \mid \mathbf{x}) = \exp^{(-\beta_0)} \exp^{-(\beta' \mathbf{x})}, \tag{35}$$

where β_0 is the constant term. The *baseline hazard rate* is equal to β_0, a constant. Any increase or decrease to the baseline hazard rate is solely a function of the covariates. Changes to the baseline hazard in (35) are a multiple of the baseline hazard rate. If there is a single binary covariate, x_1, with parameter β_1. The baseline hazard rate in this case would be $\exp(-\beta_0)$ (because $x_1 = 0$), and the hazard rate for the case of $x_1 = 1$ would be $\exp(-\beta_0) \times \exp(-\beta_1 x_1)$ or equivalently, $\exp(-\beta_0 - \beta_1 x_1)$. Since the increase (or decrease) in the hazard rate when $x_1 = 1$ is a multiple of the baseline hazard rate, the change in the hazard rate is proportional to the baseline hazard. This is seen more clearly by noting that the ratio of the two hazards,

$$\frac{h_i(t \mid x_1 = 1)}{h_i(t \mid x_1 = 0)} = \exp^{-\beta_1} \tag{36}$$

is equal to the multiple of the baseline, i.e., $\exp^{-\beta_1}$. This result is known as the *proportional hazards* property.

Huh?

Let's compute the hazard rate for civil wars:

```
. display exp(-(_b[_cons]+_b[civil]*1)) .04152249
```

(note we're using the results from the previous paragraph to do this.) Now, let's display the hazards for the other two conflict types. Interstate conflicts gives us:

```
 . display exp(-(_b[_cons]+_b[interst]*1))
 .00250125
```

and internationalized civil wars (i.e. the constant) gives us:

```
 . display exp(-(_b[_cons]))
.01289566
```

Note an important result here pertaining to the exponential. Because the hazard rate is flat, the rate itself must be a constant (i.e. it's not changing with t). The above rates are constants.

Now we're in a position to see the PH property in action. The hazard ratios are as follows. For civil wars, we obtain

```
. display .04152249/.01289566
 3.219881,
```

for interstate conflicts, we obtain

```
 . display .00250125/.01289566
.1939606,
```

and for icw, we obtain:

```
 .   display .01289566/.01289566
 1
```

(Why 1?). More specifically, the ratio of two hazards is equal to the multiple of the baseline. For civil wars, the ratio of the two hazards (the civil war and the baseline hazard) is 3.22 which is equivalent to $\exp(-[-1.169344]) = 3.219881$. This is precisely the relationship shown in equation (36). For interstate conflicts, the ratio of the two hazards is equivalent to $\exp(-[1.6401]) = .1939606$. Finally, since the constant represents the baseline hazard, the hazard ratio *must be 1* as the change to the baseline is 0; hence $\exp(0) = 1$.

Software programs like `Stata` will usually compute these quantities directly. In `Stata`, some `predict` options can be used to derive the hazard rates and hazard ratios. I stress, however, that the user be aware as to how these quantities are derived before blindly using software options.

...but to illustrate. Suppose after the model, I asked `Stata` to compute the hazard rates? To do this I would invoke the `predict` option:

```
. predict hazard_rate, hazard (4 missing values generated)


----------------------
predicted | hazard    |      Freq.
----------+-----------
 .0025013 |         10
 .0128957 |         30
 .0415225 |         14
----------------------
```

Obviously, these hazard rates are the same as those computed directly from the AFT parameters. Similarly, I could use `Stata` to derive the hazard ratios.

```
. predict hazard_ratios, hr (4 missing values generated)
```

```
. table hazard_ratios

----------------------
hazard    | ratio     |      Freq.
----------+-----------
 .1939606 |         10
        1 |         30
 3.219881 |         14
----------------------
```

These are the same as those computed earlier. What do we make of all these numbers? The hazard ratio tells us that civil wars are associated with the highest risk of failure. Compared to the baseline case of internationalized civil wars, the failure rate is about 3.2 times higher. In contrast, interstate conflicts are associated with the lowest hazard (and hence longest survival times). The ratio of .19 implies that the hazard rate is about 80 percent lower than the baseline case of internationalized civil wars (i.e. $([\exp(-1.6401) - \exp(0)]/\exp(0)) * 100 = -80.6$).

In R, the exponential model (with the AFT parameterization) is estimated as:

```
> expo<-survreg(Surv(duration, failed) ~ civil+ interst , unR, dist='exponential' )
> summary(expo)

Call: survreg(formula = Surv(duration, failed) ~ civil + interst,
data = unR,
    dist = "exponential")
             Value Std. Error     z        p
 (Intercept)  4.35      0.213 20.41 1.44e-92
 civil       -1.17      0.359 -3.26 1.12e-03
 interst      1.64      0.495  3.31 9.32e-04

Scale fixed at 1

Exponential distribution Loglik(model)= -202.9   Loglik(intercept
only)= -219.5
        Chisq= 33.36 on 2 degrees of freedom, p= 5.7e-08
Number of Newton-Raphson Iterations: 5 n=54 (4 observations
deleted due to missing)
```

Usually, the exponential will be inadequate for most purposes. The reason is that its "memoryless" property is too restrictive. This prompts users to often consider the Weibull.

Using the same data as in the previous illustration, I obtain the Weibull (in Stata) as follows:

```
  . streg civil interst, dist(weib) time

        failure _d:  failed
  analysis time _t:  duration

Fitting constant-only model:

 Iteration 0:   log likelihood = -103.03289
 Iteration 1:   log likelihood = -93.501426
 Iteration 2:   log likelihood = -93.488663
 Iteration 3:   log likelihood = -93.488663

Fitting full model:

 Iteration 0:   log likelihood = -93.488663
 Iteration 1:   log likelihood = -86.548564
 Iteration 2:   log likelihood = -84.667898
 Iteration 3:   log likelihood = -84.655162
 Iteration 4:   log likelihood = -84.655157

Weibull regression -- accelerated failure-time form

 No. of subjects =           54                     Number of obs =        54
 No. of failures =           39
 Time at risk    =         3994
                                                    LR chi2(2)      =     17.67
 Log likelihood  =   -84.655157                     Prob > chi2     =    0.0001

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       civil |  -1.100421   .4457861    -2.47   0.014    -1.974146   -.2266966
     interst |   1.736832   .6165459     2.82   0.005     .5284242     2.94524
       _cons |    4.28793   .2652436    16.17   0.000     3.768062    4.807798
-------------+----------------------------------------------------------------
       /ln_p |  -.2145617   .1237889    -1.73   0.083    -.4571834      .02806
-------------+----------------------------------------------------------------
           p |    .806895   .0998846                      .6330642    1.028457
         1/p |   1.239319   .1534138                        .97233    1.579619
------------------------------------------------------------------------------
```

What's the difference? The difference is the inclusion of the shape parameter p (recall our discussion earlier). Under the exponential, it is assumed p is exactly 1, which gives rise to the flat hazard rate. Here we see $p < 1$ implying the hazard is *decreasing* with respect

to time. Because time dependency "matters" in the Weibull, we must account for p in our computation of quantities of interest (i.e. hazard rates, ratios, etc.). That is to say, it takes a little more effort to interpret the Weibull!

Recall that the hazard rate under the Weibull can be expressed as

$$h(t) = \lambda p(\lambda t)^{p-1} \qquad t > 0, \lambda > 0, p > 0. \tag{37}$$

Here, λ is our scale parameter and p is the shape parameter. Note that λ is a function of the covariates. Under the AFT parameterization, $\lambda = \exp(-\beta_k x)$. Let's compute the hazard rates "by hand." The following is cumbersome but useful to consider as an exercise!

```
gen lambda_civil=exp(-(_b[_cons]+_b[civil]))
gen lambda_interstate=exp(-(_b[_cons]+_b[interst]))
gen lambda_icw=exp(-(_b[_cons]))
```

These three statements generate three variables containing the scale parameter λ for each conflict type. Inserting the scale parameter into the function for the hazard rate (shown above), we can generate each mission-type-specific hazard rate:

```
gen haz_civil=lambda_civil*e(aux_p)*(lambda_civil*duration)^(e(aux_p)-1)
gen haz_interstate=lambda_interstate*e(aux_p)*
        (lambda_interstate*duration)^(e(aux_p)-1)
gen haz_icw=lambda_icw*e(aux_p)*(lambda_icw*duration)^(e(aux_p)-1)
```

In the statements above, the syntax `e(aux_p)` is `Stata` code to reference the shape parameter. One important thing to note from these statements is this: the hazard rate is *not* a constant. It changes as t changes; hence, the duration time t is include in the calculation of the hazard function. After I compute the hazards, I could graph them:

```
. scatter  haz_civil haz_interstate haz_icw duration, msymbol(O D S)
```

This would produce Figure 10. Again, note the non-constancy of the hazard rates (This graph is similar to Figure 3.1 in our book). As with the exponential, we see civil wars have the highest hazard and interstate conflicts have the lowest hazard.

We could also compute hazard ratios. Since the Weibull is a PH model, the hazard ratios are computed similarly to the exponential hazard ratios. Thus, taking the ratios of the hazard rates that we just computed (and graphed), we would obtain the hazard ratios. From this model, we find that the hazard ratios are:

```
. display haz_interst/haz_icw
 .24624184

. display haz_icw/haz_icw
```

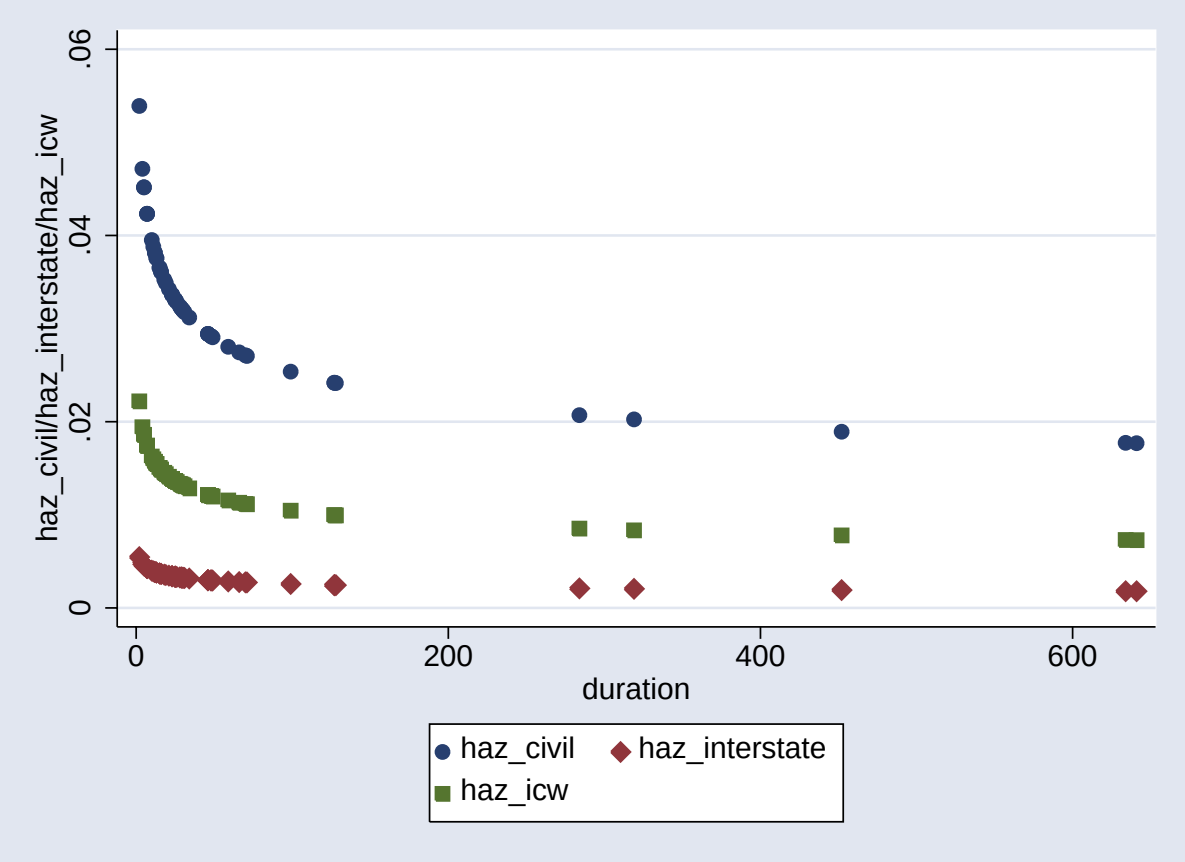

Figure 10: *This figures graphs the hazard rates from the Weibull.*

```
 1

. display haz_civil/haz_icw
 2.4300806
```

(We could have used the `predict, hazard` option in `Stata` if we wanted too ... but where would the fun be in doing it the easy way!) Importantly, I want you to see that although the hazard rates are monotonically decreasing across time, the hazard ratios are always proportional at each time t. This *is* the proportional hazards property (and it may not always hold!).

The AFT model in `R` is estimated as

```
 > weib<-survreg(Surv(duration, failed) ~ civil+ interst , unR, dist='weibull')
 > summary(weib)
Call: survreg(formula = Surv(duration, failed) ~ civil + interst,
data = unR,
    dist = "weibull")
```

```
             Value Std. Error     z        p
 (Intercept)  4.288      0.265 16.17 8.76e-59
 civil       -1.100      0.446 -2.47 1.36e-02
 interst      1.737      0.617  2.82 4.85e-03
 Log(scale)   0.215      0.124  1.73 8.30e-02

Scale= 1.24

Weibull distribution Loglik(model)= -201.2   Loglik(intercept
only)= -210
        Chisq= 17.67 on 2 degrees of freedom, p= 0.00015
Number of Newton-Raphson Iterations: 5 n=54 (4 observations
deleted due to missing)
```

We haven't included a covariate in any of our models that is continuous (or semi-continuous). Let's reestimate the mission-type model and include a covariate that measures the number of borders that are involved in a conflict (i.e. country borders).

In Stata (using the AFT parameterization) I obtain:

```
. streg civil interst borders, dist(weib) time nolog

        failure _d:  failed
  analysis time _t:  duration

Weibull regression -- accelerated failure-time form

 No. of subjects =           46                 Number of obs =        46
 No. of failures =           36
 Time at risk    =         3840
                                                LR chi2(3)      =     18.45
 Log likelihood  =   -76.493097                 Prob > chi2     =    0.0004

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       civil |  -1.380352   .4921063    -2.80   0.005    -2.344862   -.4158411
     interst |   1.806995   .6347777     2.85   0.004     .5628534    3.051136
     borders |  -.1368689   .0972727    -1.41   0.159    -.3275199     .053782
       _cons |   4.800974   .4777848    10.05   0.000     3.864533    5.737415
-------------+----------------------------------------------------------------
       /ln_p |  -.2278767   .1328443    -1.72   0.086    -.4882467    .0324932
-------------+----------------------------------------------------------------
```

```
           p |   .7962224   .1057736                      .6137014    1.033027
         1/p |   1.255931   .1668432                       .968029    1.629457
------------------------------------------------------------------------------
```

The negatively signed coefficient estimate for the `borders` covariate implies that the log survival time (remember: AFT is a linear mode for $\log(T)$) is decreasing as the number of country borders associated with a mission increases. The coefficient is not significantly different from 0 at standard levels ... but for our purposes, we won't worry about that just yet!

The hazard ratio for this covariate is 1.115. This ratio was computed as $\exp(-.137)^{.796}$ (where .796 is p); however, unlike the binary covariates, the range of this variable extends beyond 1. Hence the hazard ratio will increase over the range of the covariate. This is because the hazard ratio in this setting is computed as:

```
 . gen hazratio_borders=exp(-_b[borders]*borders)^e(aux_p)
(8 missing values generated)

. sort hazratio_

. scatter hazratio_borders borders, c(l)
```

This is illustrated in Figure 11. Note that although the estimated hazard ratio is increasing with the value of the covariates, the proportional hazards property still must hold. Below, I reproduce the hazard ratio estimates for each value of the `borders` covariate:

```
. table hazratio_borders borders

------------------------------------------------------------
          |                hazratio_borders
 borders  |   1     2     3    4     5     6     8     9    13
----------+-------------------------------------------------
 1.115138 |   10
 1.243533 |          7
  1.38671 |                6
 1.546373 |                     12
  1.72442 |                           8
 1.922966 |                                 3
 2.391271 |                                       2
 2.666597 |                                             1
 4.123554 |                                                   1
------------------------------------------------------------
```

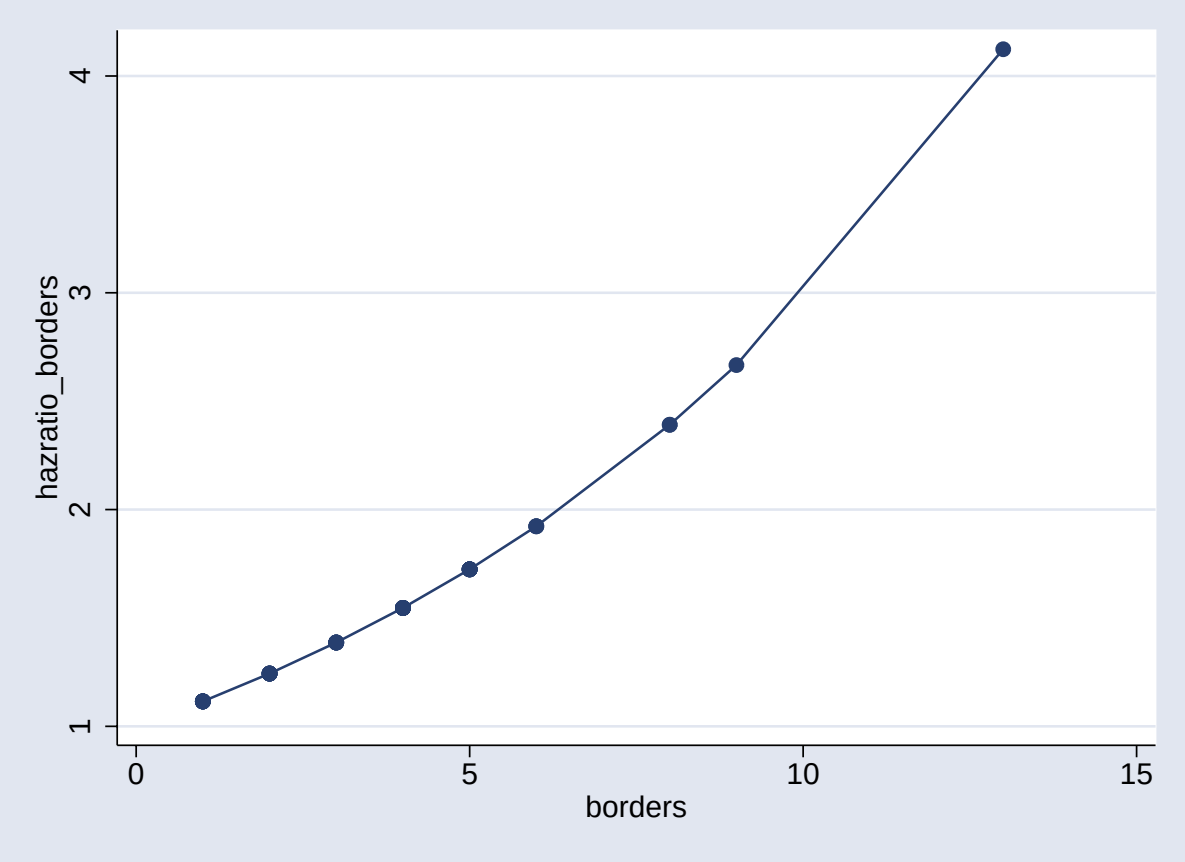

Figure 11: *This figures graphs the hazard ratios for the borders covariate.*

If you took the ratio of the ratios for each adjacent pair, it would be proportional to $\exp(-.137)^{.796}$, or more precisely, 1.115138. Thus, the ratio of the ratios when `borders=4` to `borders=3` is:

```
. display 1.546373/1.38671
 1.115138
```

Again, this illustrates the proportional hazards property. We've spent most of our time discussing the Weibull and the exponential. There are many other distribution functions that can be applied to parametric models. We discussed some of these earlier. The basic mode of analysis follows the presentation discussed above with some important exceptions. The log-normal and log-logistic survival models are not proportional hazards models. Consequently, these models are only interpretable as linear models for $\log(T)$. In contrast, the Gompertz model (also discussed above) is a proportional hazards model but not an AFT model. Indeed, what makes the Weibull (and the exponential) unique is it is the only model that gives rise to both the AFT and PH parameterizations. I will return to issues of interpretation but now I want to talk about model selection.

10 Inference and Model Selection

In regard to assessing statistical significance of the covariates, standard $z--$ or Wald tests can be used (so long as the asymptotic ML properties hold). In this sense, inference for the covariates is quite similar to inference in other settings that apply ML estimators. Additionally, standard likelihood ratio tests can be used to compare parametrically nested models. Again, this is all pretty standard stuff.

What is less standard is adjudicating among competing (and nonnested models). When applying parametric survival models, this can be an issue. While parametric models may be desirable because the baseline hazard is directly modeled, all this assumes the analyst has the "right shape" and therefore the "right" distribution function.

This may not always be the case, particularly if one simply settles on a particular CDF and then adjudicates among competing slates of covariates. If the CDF is "wrong" (or rather, less well fitting than other competitors), then the gains in efficiency fostered by parametric models may be lost.

Generalized Gamma

A distribution function that encompassed many parametric models might be desirable in this setting. One model, the generalized gamma is just such a distribution function. The density for the generalized gamma is given by

$$f(t) = \frac{\lambda p(\lambda t)^{p\kappa-1}\exp[-(\lambda t)^p]}{\Gamma(\kappa)}, \tag{38}$$

where $\lambda = e^{-\beta'\mathbf{x}}$, and p and κ are the two shape parameters of the distribution. The important difference between the generalized gamma and the other parametric models is the additional free parameter, κ, that is estimated.

The Weibull, log-logistic, and log-normal densities only have one free parameter; the generalized gamma has two free parameters. This flexibility permits researchers to assess the adequacy of other (nested) parameterizations. The generalized gamma is "general" because, depending on the value of the shape parameters, several parametric models are implied from this distribution.

Specifically, when $\kappa = 1$, the Weibull distribution is implied; when $\kappa = p = 1$, the exponential distribution is implied; when $\kappa = 0$, the log-normal distribution is implied; and when $p = 1$, the gamma distribution is implied.

11 Extended Illustration

We've now discussed a whole lot of stuff regarding parametric models. I turn to an extended illustration. I'll first illustrate the generalized gamma and then move to other issues. In this illustration I'm using the cabinet duration data discussed in the book. Also, for more details on some Stata code, see `do` file associated with the cabinet duration data on my website.

As there are a variety of parametric models, I could "plug and play" several distribution functions. To that end:

```
. streg invest polar  numst format postelec caretakr, dist(weib)
time nolog

         failure _d:  censor
   analysis time _t:  durat

Weibull regression -- accelerated failure-time form

 No. of subjects =          314                 Number of obs   =          314
 No. of failures =          271
 Time at risk    =       5789.5
                                                 LR chi2(6)     =    171.94
 Log likelihood  =   -414.07496                  Prob > chi2    =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      invest |  -.2958188   .1059024    -2.79   0.005    -.5033838   -.0882538
       polar |   -.017943   .0042784    -4.19   0.000    -.0263285   -.0095575
       numst |   .4648894   .1005815     4.62   0.000     .2677533    .6620255
      format |  -.1023747   .0335853    -3.05   0.002    -.1682006   -.0365487
    postelec |   .6796125    .104382     6.51   0.000     .4750276    .8841974
    caretakr |   -1.33401   .2017528    -6.61   0.000    -1.729438   -.9385818
       _cons |   2.985428   .1281146    23.30   0.000     2.734328    3.236528
-------------+----------------------------------------------------------------
      /ln_p |    .257624   .0500578     5.15   0.000     .1595126    .3557353
-------------+----------------------------------------------------------------
           p |   1.293852   .0647673                      1.172939     1.42723
         1/p |   .7728858   .0386889                       .700658    .8525593
------------------------------------------------------------------------------

. streg invest polar  numst format postelec caretakr, dist(exp)
time nolog

         failure _d:  censor
```

```
   analysis time _t:  durat

Exponential regression -- accelerated failure-time form

 No. of subjects =          314              Number of obs =       314
 No. of failures =          271
 Time at risk    =       5789.5
                                                LR chi2(6)     =    148.53
 Log likelihood  =   -425.90641                 Prob > chi2    =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      invest |  -.3322088   .1376729    -2.41   0.016    -.6020426   -.0623749
       polar |  -.0193017   .0055465    -3.48   0.001    -.0301725   -.0084308
       numst |    .515435   .1291486     3.99   0.000     .2623084    .7685616
      format |  -.1079432   .0435233    -2.48   0.013    -.1932474    -.022639
    postelec |   .7403427    .134558     5.50   0.000     .4766138    1.004072
    caretakr |  -1.319272   .2595422    -5.08   0.000    -1.827965   -.8105783
       _cons |   2.944518   .1663401    17.70   0.000     2.618498    3.270539
------------------------------------------------------------------------------



. streg invest polar  numst format postelec caretakr, dist(loglog)
time nolog

         failure _d:  censor
   analysis time _t:  durat

Log-logistic regression -- accelerated failure-time form

 No. of subjects =          314                Number of obs =       314
 No. of failures =          271
 Time at risk    =       5789.5
                                                LR chi2(6)     =    148.72
 Log likelihood  =   -424.10921                 Prob > chi2    =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      invest |  -.3367541   .1278083    -2.63   0.008    -.5872538   -.0862544
       polar |  -.0221958   .0052638    -4.22   0.000    -.0325127   -.0118789
       numst |   .4830709   .1212506     3.98   0.000     .2454241    .7207177
```

```
      format |  -.1093453   .0419715    -2.61   0.009    -.1916078   -.0270827
    postelec |   .6408808   .1240329     5.17   0.000     .3977807    .8839808
    caretakr |   -1.26921   .2310272    -5.49   0.000    -1.722015   -.8164046
       _cons |   2.728818   .1595866    17.10   0.000     2.416034    3.041602
-------------+----------------------------------------------------------------
     /ln_gam |  -.5657686   .0511353   -11.06   0.000     -.665992   -.4655451
-------------+----------------------------------------------------------------
       gamma |   .5679235    .029041                      .5137636    .6277928
------------------------------------------------------------------------------

. streg invest polar  numst format postelec caretakr,
dist(lognorm) time nolog

        failure _d:  censor
  analysis time _t:  durat

Log-normal regression -- accelerated failure-time form

 No. of subjects =          314                 Number of obs=       314
 No. of failures =          271
 Time at risk    =       5789.5
                                                LR chi2(6)      =    150.66
 Log likelihood  =   -425.30621                 Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      invest |  -.3738013   .1327055    -2.82   0.005    -.6338993   -.1137032
       polar |   -.021988   .0054825    -4.01   0.000    -.0327336   -.0112424
       numst |   .5717579   .1232281     4.64   0.000     .3302353    .8132805
      format |  -.1194982   .0432516    -2.76   0.006    -.2042698   -.0347266
    postelec |   .6668079   .1292366     5.16   0.000     .4135088     .920107
    caretakr |  -1.126047   .2576962    -4.37   0.000    -1.631122   -.6209713
       _cons |   2.632497    .164494    16.00   0.000     2.310095    2.954899
-------------+----------------------------------------------------------------
     /ln_sig |   .0078719   .0439881     0.18   0.858    -.0783432    .0940871
-------------+----------------------------------------------------------------
       sigma |   1.007903   .0443358                      .924647    1.098655
------------------------------------------------------------------------------

. streg invest polar  numst format postelec caretakr,
dist(gompertz) nohr nolog

        failure _d:  censor
```

```
   analysis time _t:  durat

Gompertz regression -- log relative-hazard form

 No. of subjects =          314                 Number of obs =       314
 No. of failures =          271
 Time at risk    =       5789.5
                                                 LR chi2(6)      =    159.11
 Log likelihood  =   -418.97771                  Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      invest |   .3661259   .1376137     2.66   0.008     .0964081    .6358437
       polar |   .0226296   .0056338     4.02   0.000     .0115875    .0336717
       numst |  -.6076508   .1322898    -4.59   0.000    -.8669341   -.3483675
      format |    .127332   .0440044     2.89   0.004      .041085    .2135791
    postelec |  -.8905733   .1422624    -6.26   0.000    -1.169402   -.6117441
    caretakr |   1.477549   .2642318     5.59   0.000     .9596641    1.995434
       _cons |  -3.238039   .1854011   -17.47   0.000    -3.601418   -2.874659
-------------+----------------------------------------------------------------
       gamma |   .0225632    .005949     3.79   0.000     .0109032    .0342231
------------------------------------------------------------------------------
```

Here, you see Weibull, exponential, log-logistic, log-normal, and Gompertz models, all using the same data (why are the signs on the Gompertz coefficients the opposite of the other models?) Which model should you report?

Perhaps an encompassing distribution like the generalized gamma could shed a little bit of light on this problem. Let's estimate:

```
. streg invest polar  numst format postelec caretakr, dist(gamma)
nolog

         failure _d:  censor
   analysis time _t:  durat

Gamma regression -- accelerated failure-time form

 No. of subjects =          314                 Number of obs =       314
 No. of failures =          271
 Time at risk    =       5789.5
                                                 LR chi2(6)      =    165.78
 Log likelihood  =   -414.00944                  Prob > chi2 =       0.0000
```

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
invest	-.3005269	.108745	-2.76	0.006	-.5136633	-.0873906
polar	-.0182998	.0044674	-4.10	0.000	-.0270559	-.0095438
numst	.4692142	.1030895	4.55	0.000	.2671626	.6712659
format	-.1031368	.0342637	-3.01	0.003	-.1702925	-.0359811
postelec	.6807161	.1061356	6.41	0.000	.4726942	.888738
caretakr	-1.328476	.2066422	-6.43	0.000	-1.733487	-.9234647
_cons	2.963114	.1447075	20.48	0.000	2.679492	3.246735
/ln_sig	-.234325	.0802121	-2.92	0.003	-.3915378	-.0771122
/kappa	.9241712	.2065399	4.47	0.000	.5193605	1.328982
sigma	.7911047	.0634561			.6760165	.9257859

Because some of the previously estimated models are nested within the generalized gamma, we can assess which *of these distributions* best fits the data using standard z tests.

Here we find that κ is significantly different from 0, thus providing evidence against the log-normal model. For the case of the gamma distribution, the appropriate test is for $\sigma = 1$. These results provide evidence against the null and hence against the gamma (this result also provides evidence against the exponential, which holds when $\sigma = \kappa = 1$). Finally, to test the suitability of the Weibull, we test for $\kappa = 1$. For this test, $z = -.38$. This tells us that among the nested distribution under the generalized gamma, the Weibull fits the best.

We have some semblance of a statistical test to follow and therefore may be able to avoid some ad hoc choices. Of course there are drawbacks. There are other survival distributions not nested under the generalized gamma. Additionally, because of the additional ancillary parameter that is estimated, the likelihood function can be ill-behaved . . . that is, it may fail to converge in some data sets. I've never had this become an issue but I know it is an issue.

To verify the similarity between the Weibull and the generalized gamma, I graph the predicted survival function from the two models in Figure 12 (code for this can be found on my website). The two functions are identical. This makes sense. The generalized gamma reduces to the Weibull.

What about adjudicating among non-nested models? The AIC (Akaike Information Criterion) may be helpful. The AIC is given by

$$AIC = -2(\log L) + 2(c + p + 1), \tag{39}$$

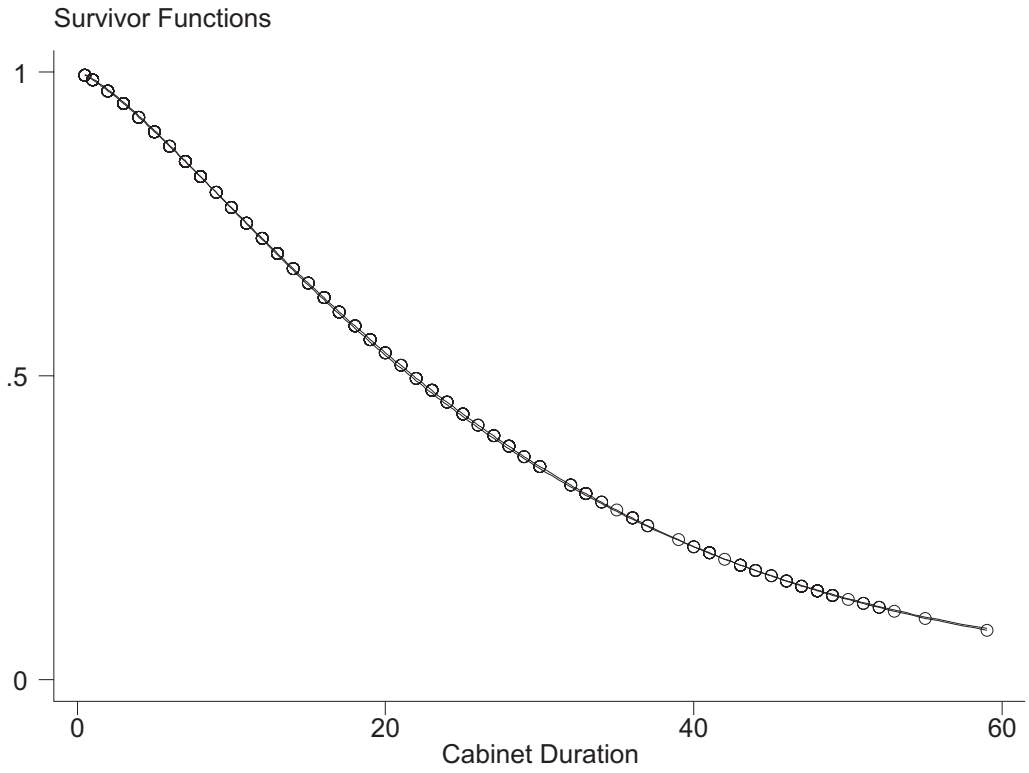

Figure 12: *The figure graphs the generalized gamma and Weibull survivor functions for the cabinet duration data. The Weibull estimates are denoted by the "O" symbol and the generalized gamma estimates are denoted by the line.*

where c denotes the number of covariates in the model and p denotes the number of structural parameters for the model. The idea behind the AIC is to "reward" parsimonious models by penalizing the log-likelihood for each parameter that is estimated. In Table 1, the log-likelihood and AIC for each model estimated above is reported. Based on minimizing the AIC, the preferred model for these data is the Weibull. This is consistent with the results of the generalized gamma. Notably, the generalized gamma model has the lowest log-likelihood; however, as we saw in the previous section, the generalized gamma model reduces to a Weibull model for these data. The difference in log-likelihoods for the Weibull and generalized gamma is trivial.

How should I interpret my model? This is a common question. There are lots of quantities and we've already seen a number of ways to interpret the parameters. I want to run through some possible ways to make your results more transparent.

My illustrations are keyed to Stata. Most of this can be done in `R`; I'm just not as proficient in `R`!

Each distribution can be characterized in terms of its survival function. As such, it's some-

Table 1: The AIC and Log-Likelihood for Cabinet Duration Models

Model	Log-Likelihood	AIC
Exponential	-425.91	865.82
Weibull	-414.07	844.14
Log-Logistic	-424.11	864.22
Log-Normal	-425.31	866.62
Gompertz	-418.98	853.96
Generalized Gamma	-414.01	846.02

Data for models are from King, et. al. 1990.

times nature to consider expected and median survival times. Computing these quantities allows you to say something about survival times for different kinds of covariate profiles.

Let's work with the cabinet duration data Weibull results (remember, we've already determined this is the best-fitting model among a few of the parametric models and has a superior AIC to several).

The survivor functions will (obviously) vary from one distribution to another. Our book and most other books discusses these functions. Under the Weibull, the expected survival time is given by

$$E(T) = \frac{\Gamma(1 + \frac{1}{p})}{\lambda}, \tag{40}$$

where Γ denotes the Gamma function. Since the Weibull density in (14) is defined for $0 \leq t \leq \infty$, the Weibull distribution will be right-skewed and the median may be a better summary of the duration times. The percentiles of the Weibull distribution are computed by

$$t(\text{p'tile}) = \lambda^{-1} \log \left(\frac{100}{100 - \text{p'tile}} \right)^{1/p}, \tag{41}$$

so for the median duration time, t(p'tile) in (41) is 50 percent, and is computed by $\lambda^{-1} \log(2)^{1/p}$.

Either of these quantities may be readily used to make results more transparent. To illustrate, I use `Stata` to derive the expected survival times for the cabinet data.

```
. gen expected_S=exp(lngamma(1+(1/e(aux_p))))*exp(_b[_cons]+_b[invest]*invest +
_b[polar]*polar + _b[numst]*numst + _b[format]*format +
_b[postelec]*postelec + _b[caretakr]*caretakr )
```

This is an ugly statement ... but it gets the job done. I computed the expected survival time for each observation. This may or may not be desirable (usually it will be undesirable). Instead, it may make sense to compute $E(T)$ for selected covariate profiles. (In the above syntax, instead of using all actual values, I could substitute "scenario values" for the covariates).

In Stata, the post-estimation command

```
. predict e_surv, time mean
```

will generate exactly what I did above. As with my version, Stata's computation will be for each observation. You may want to select a covariate profile. You could do this by

```
. predict e_surv, time mean, <if list>
```

where <if list> corresponds to some covariate profile (i.e. you specify the values of the covariates). Be careful. Sometimes the covariate profiles that interest you most don't actually exist in the data!

As with the expected survival times, I could computed the median survival times. This quantity refers to the estimated time at which about half of the observation fail. Using the result from above, I can compute the median directly:

```
. gen median_S= exp(_b[_cons] +_b[invest]*invest +
_b[polar]*polar + _b[numst]*numst + _b[format]*format
+ _b[postelec]*postelec + _b[caretakr]*caretakr)*log(2)^(1/e(aux_p))
```

which gives me the median for each observation. In Stata I could compute this using the post-estimation command:

```
. predict m_surv, time median
```

which does what I just did previously. Again, covariate profiles could be specified.

Similarly, computing the hazard rates (and ratios) for various covariate profiles may be a useful way to present and discuss EH results. In some of the illustrations earlier, I described this as pertaining to the Weibull. Again, Stata and most other software programs will directly compute these quantities for you. (Again, see above for examples).

A Stata specific option to be aware of is stcurve. This option allows you to graph survivor functions, cumulative hazard functions, and hazard functions for various covariate profiles. This option is useful because it visually describes the data and it avoids the problem of predict in that it derives the curves based on *average* covariate values (which may not make sense sometimes!!). This post-estimation command is easy to use and I ask you to use it in a problem set. Refer to the Cleves et al. book (pp. 245–250) for more details.

Lastly, analysis of residuals is extremely important. We will discuss this later this week!

12 Conclusion

This concludes our trip down parametric lane. We've covered a lot of ground. There are a lot of issues: model selection, duration dependency, and so forth. Some of these issues are avoided if we opt for an alternative modeling strategy: the Cox model (though other issues then become relevant!). This is where we want to turn next.

13 Introduction

We've spent a lot of time talking about parametric models. They have several advantages, many of which were discussed (and Bennett discusses these advantages in even more detail). There also are drawbacks.

If we believe that duration dependency is a nuisance *and* if our principle interest is in understanding the relationship between some set of covariates and the risk of an event happening, then alternative modeling strategies may be preferred. This leads to our consideration of the important Cox proportional hazards model.

The premise of the Cox model is to estimate the impact of the covariates on the hazard rate, without specifying the distribution of the duration dependency. That is, the baseline hazard rate is not directly estimated (though nonparametric estimates of the function can be retrieved).

The Cox model is named for the important statistician who derived the model: Sir David Cox. In the 30 plus years since Cox's pioneering work, the Cox model has become the workhorse of survival analysis in many, many disciplines.

Let's talk about the moving parts of this model.

As I note in our book, "The logic of the Cox model is simple and elegant." The hazard rate is given by

$$h_i(t) = h_0(t) \exp(\beta' \mathbf{x}), \tag{42}$$

where $h_0(t)$ is the baseline hazard function, and $\beta' \mathbf{x}$ are the time-independent covariates and regression parameters. The hazard rate for the Cox model is *proportional* which means that the ratio of two hazards can be written as,

$$\frac{h_i(t)}{h_0(t)} = \exp[\beta'(\mathbf{x}_i - \mathbf{x}_j)], \tag{43}$$

which demonstrates that this ratio is a fixed proportion across time (i.e. $\exp(\beta)$). The form of the baseline hazard rate, $h_0(t)$ is assumed to be unknown and is left unparameterized. This differs considerably from the parametric case.

For this reason, the Cox model is sometimes referred to as a "semi-parametric" model: the (ordered) duration times are parameterized in terms of a set of covariates, but the particular distributional form of the duration times is not parameterized.

Cox regression models do not have an intercept term. To see this, note that we can express the Cox model in scalar form as

$$h_i(t) = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}) h_0(t), \tag{44}$$

and if we re-express the model in terms of the log of the hazard ratios, we obtain,

$$\log\left\{\frac{h_i(t)}{h_0(t)}\right\} = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}. \tag{45}$$

Expressed as either (44) or (45), the model contains no constant term β_0. This term is "absorbed" into the baseline hazard function. We've seen this before: when we converted the Weibull (and exponential) into hazard ratios, the ratio for the "baseline case" was exactly 1 (why?). It's the same situation here ... which makes sense because the Cox and Weibull models are both proportional hazards models.

14 Partial Likelihood

To obtain estimates of the covariate parameters, Cox (1972, 1975) developed a nonparametric method he called *partial likelihood*. Estimation of the parameter values is then obtained by use of maximum partial likelihood estimation.

The partial likelihood method is based on the assumption that the intervals between successive duration times (or failure times) contributes no information regarding the relationship between the covariates and the hazard rate.

This is the case because the baseline hazard function for the Cox model is not directly specified. This rate is assumed to have an arbitrary form and could actually be zero in the intervals between successive failures. Therefore, it is the *ordered failure times*, rather than interval between failure times, that contributes information to the partial likelihood function.

This is in contrast to the parametric methods, where the actual survival times are used in the construction of the likelihood function. Because the Cox model only uses "part" of the available data ($h_0(t)$ is not estimated), the likelihood function for the Cox model is a "partial" likelihood function, hence the name.

Let's try and get a sense for how this thing works.

First, we'll think about the logic underlying the partial likelihood method. Consider the data in Figure 13. Here, I illustrate the survival times for nine cases. Of these nine cases, six of them experience an event, i.e., they "fail", and three of them are right-censored.

In the figure, the first failure time, t_1, occurs at 7 months, and represents the failure time for case 7; the last failure time, t_6, occurs at 51 months, and denotes the failure time for case 6.

It is clear that the failure times can be ordered such that $t_1 < t_2 < \ldots < t_6$. Note that the censored cases do not contribute a failure time. In Table 2, I have sorted the data from

Table 2: Data Sorted by Ordered Failure Time

Case Number	Duration Time	Censored Case
7	7	No
4	15	No
5	21	No
2	28	Yes
9	30	Yes
3	36	No
8	45	Yes
1	46	No
6	51	No

Data are sorted by the duration time. The duration time for censored cases denotes the time of last observation.

Figure 13 by the ordered failure time.

What are the main features of these data?

- Events can be ordered.
- At t_0 all cases are at risk of failing.
- After the first failure, the risk set decreases by 1.
- The risk set successively dwindles as events occur.

Now to motivate the partial likelihood estimator, let $\psi = \exp(\beta' \mathbf{x_i})$ (this notation is from Collett, 1994, p. 64).

The partial likelihood function for these data would be equivalent to,

$$\begin{aligned}
\mathcal{L}_p \;=\; & \frac{\psi(7)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)+\psi(6)+\psi(7)+\psi(8)+\psi(9)} \times \\
& \frac{\psi(4)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)+\psi(6)+\psi(8)+\psi(9)} \\
& \frac{\psi(5)}{\psi(1)+\psi(2)+\psi(3)+\psi 5+\psi(6)+\psi(8)+\psi(9)} \times \\
& \frac{\psi(3)}{\psi(1)+\psi(3)+\psi(6)+\psi(8)} \times \\
& \frac{\psi(1)}{\psi(1)+\psi(6)} \times
\end{aligned}$$

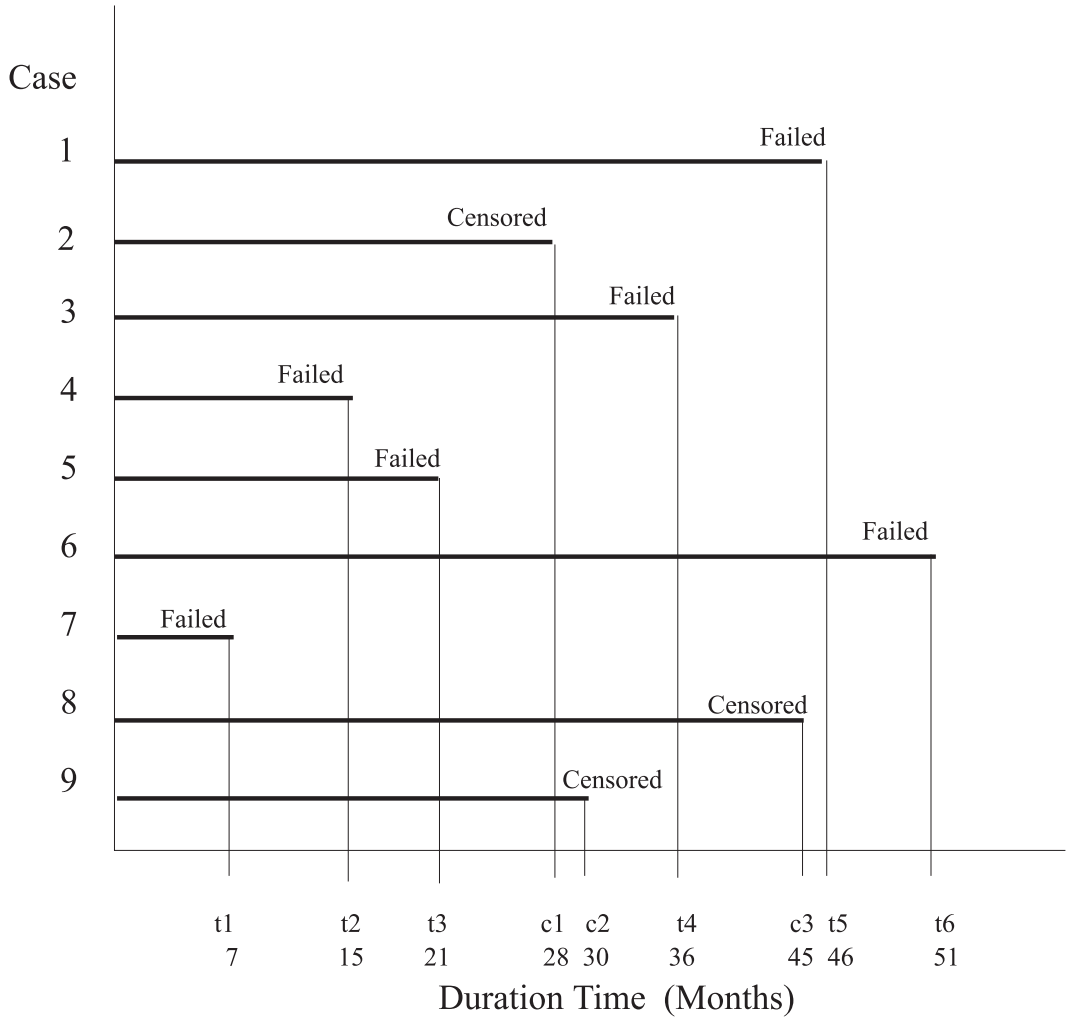

Figure 13: *Duration times for nine censored and uncensored (failed) cases.*

$$\frac{\psi(6)}{\psi(6)}.$$

In words, this tells us that each of the nine cases are at risk of experiencing an event up to the first failure time, t_1. After the first failure in the data set, the risk set decreases in size by 1; thus, the risk set up to the second failure time, t_2, includes all cases except case 7.

By the fourth failure time in the data, t_4, the risk set includes only cases 1, 6, and 8; cases 2 and 9 are right-censored before the fourth failure time is observed and do not contribute any information to this part of the likelihood function. By the last failure time, only case 6 remains in the risk set.

This exercise shows that the partial likelihood function is solely based on the ordered duration times, and not on the length of the interval between duration times.

Also, censored observations contribute information to the "risk set," that is, cases that are surviving to time t_i, but contribute no information regarding failure times.

To be a little more formal, suppose we have a data set with n observations and k distinct failure (event) times. Cox estimation first proceeds by sorting the ordered failure times, such that

$t_1 < t_2 < \ldots < t_k$,

where t_i denotes the failure time for the ith individual. For censored cases, we define δ_i to be 0 if the case is right-censored, and 1 if the case is uncensored. Finally, the ordered event times are modeled as a function of covariates, $\mathbf{x}$.

The partial likelihood function is derived by taking the product of the conditional probability of a failure at time t_i, given the number of cases that are at risk of failing at time t_i.

That is to say, given that some event has occurred, what is the probability the event occurred to the ith individual from a risk set of size n?

More formally, if we define $R(t_i)$ to denote the number of cases that are at risk of experiencing an event at time t_i, that is, the "risk set," then the probability that the jth case will fail at time T_i is given by

$$\Pr(t_j = T_i \mid R(t_i)) = \frac{e^{\beta' \mathbf{x_i}}}{\sum_{j \in R(t_i)} e^{\beta' \mathbf{x_j}}}, \tag{46}$$

where the summation operator in the denominator is summing over all individuals in the risk set. Taking the product of the conditional probabilities in (46) yields the partial likelihood function,

$$\mathcal{L}_p = \prod_{i=1}^{K} \left[\frac{e^{\beta' \mathbf{x_i}}}{\sum_{j \in R(t_i)} e^{\beta' \mathbf{x_j}}} \right]^{\delta_i}, \tag{47}$$

with corresponding log-likelihood function,

$$\log L_p = \sum_{i=1}^{K} \delta_i \left[\beta' \mathbf{x_i} - \log \sum_{j \in R(t_i)} e^{\beta' \mathbf{x_j}} \right]. \tag{48}$$

By maximizing the log-likelihood in (48), estimates of the β may be obtained.

What is the importance of this result?

- Specifying the baseline hazard, $h_0(t)$ is unnecessary.
- The interval between events does not inform the PL function.
- Censored cases contribute information only pertinent to the risk set (i.e. the denominator,

not the numerator)

The critical thing here is to note that no assumptions about the shape of the baseline hazard need to be made. Another way to see this is to think about the heuristic partial likelihood function above. All we need to know to compute a probability is ψ (or $\exp(\beta' \mathbf{x_i})$.

Cox (1972, 1975) demonstrated that maximum partial likelihood estimation produces parameter estimates that have the same properties as maximum likelihood estimates. This is convenient because under the same set of regularity conditions as maximum likelihood estimation (see Greene 1997), the parameter estimates from partial likelihood are asymptotically normal, asymptotically efficient, consistent, and invariant. So the usual kinds of hypothesis tests discussed in the context of parametric models are directly extended to the Cox model.

14.1 Ties

"Ties" occur when two or more cases fail at the same observed time. Ties cannot be accounted for in the partial likelihood function, as presented in (47).
The basic problem that tied events pose for the partial likelihood function is in the determination of the composition of the risk set at each failure time, and the sequencing of event occurrences.

For two or more observations that fail, or experience an event at the same time, it is impossible to tell which observation failed first. Consequently, it is not possible to discern precisely the composition of the risk set at the time of the failures.

In order to estimate the parameters of the Cox model with tied failure times, then, it becomes necessary to approximate the partial likelihood function in (47).

We address the issue of ties in some detail in our book. Other texts give even greater treatment to the issue. I'm not going to spend a lot of class time on ties except to say that the default approximation method for most software programs is the Breslow method. An alternative method, the Efron method, is somewhat more precise than the Breslow method. There are other approximation methods known as "exact" methods. This terminology is a little misleading. In any event, either of these methods are in general, more precise than either the Breslow or Efron method. What we call the exact discrete approximation in our book, it turns out, is equivalent to a discrete-choice model known as conditional logit (I'll talk more on this later).

Recommendations? Use Efron or exact methods. If the data are discrete, use the exact discrete method. As the number of ties decreases, the differences across the methods become negligible. The obverse is true as well!

15 Illustrations: Cox

Let's begin with the cabinet duration data. In `Stata`, I invoke the `stcox` command and in `R` I can use the `coxph` function. I'll use the Efron approximation for these data (in the book, I use the cabinet duration data to illustrate differences across approximation methods).

To obtain the Cox estimates in `Stata` I do the following:

```
. stcox invest polar  numst format postelec caretakr, efron nohr

        failure _d:  censor
  analysis time _t:  durat

 Iteration 0:   log likelihood =  -1369.664
 Iteration 1:   log likelihood = -1307.6939
 Iteration 2:   log likelihood = -1287.7995
 Iteration 3:   log likelihood = -1287.7389
 Iteration 4:   log likelihood = -1287.7389
 Refining estimates:
 Iteration 0:   log likelihood = -1287.7389

Cox regression -- Efron method for ties

 No. of subjects =          314                     Number of obs =       314
 No. of failures =          271
 Time at risk    =       5789.5
                                                   LR chi2(6)      =    163.85
 Log likelihood  =   -1287.7389                    Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      invest |   .3871388   .1371298     2.82   0.005     .1183693    .6559083
       polar |   .0233392   .0056193     4.15   0.000     .0123255    .0343528
       numst |  -.5826222   .1322266    -4.41   0.000    -.8417816   -.3234628
      format |    .130011   .0438699     2.96   0.003     .0440275    .2159945
    postelec |  -.8611202   .1406178    -6.12   0.000    -1.136726   -.5855144
    caretakr |   1.710397   .2828184     6.05   0.000     1.156084    2.264711
------------------------------------------------------------------------------
```

The coefficients are expressed as hazard rates. Exponentiating them would yield the hazard ratios. Thus exp(`invest`) = 1.472761. In `Stata` I can estimate this directly by omitting the `nohr` option:

```
. stcox invest polar  numst format postelec caretakr, efron
```

```
        failure _d:  censor
  analysis time _t:  durat

 Iteration 0:   log likelihood =  -1369.664
 Iteration 1:   log likelihood = -1307.6939
 Iteration 2:   log likelihood = -1287.7995
 Iteration 3:   log likelihood = -1287.7389
 Iteration 4:   log likelihood = -1287.7389
 Refining estimates:
 Iteration 0:   log likelihood = -1287.7389

Cox regression -- Efron method for ties

 No. of subjects =          314                     Number of obs =       314
 No. of failures =          271
 Time at risk    =       5789.5
                                                    LR chi2(6)      =    163.85
 Log likelihood  =   -1287.7389                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      invest |   1.472761   .2019595     2.82   0.005      1.12566    1.926892
       polar |   1.023614    .005752     4.15   0.000     1.012402     1.03495
       numst |   .5584321   .0738396    -4.41   0.000     .4309421    .7236389
      format |   1.138841   .0499609     2.96   0.003     1.045011    1.241096
    postelec |   .4226883   .0594375    -6.12   0.000     .3208678    .5568194
    caretakr |    5.53116   1.564314     6.05   0.000     3.177464    9.628346
------------------------------------------------------------------------------
```

The model is identical to the previous one; just the parameterization has changed. In R these models can be estimated through the command

```
> cabinetR <- read.dta("c:\\data\\statadata\\cabinet.dta")

> cabinet.coxph <- coxph(Surv(durat, censor)~ invest+polar+numst+ format+
 postelec+ caretakr,   cabinetR, method="efron")
> summary(cabinet.coxph)
Call: coxph(formula = Surv(durat, censor) ~ invest + polar + numst
 + format + postelec + caretakr, data = cabinetR, method =
"efron")

  n= 314
```

```
            coef exp(coef) se(coef)     z       p
invest    0.3871    1.473  0.13713  2.82 4.8e-03
polar     0.0233    1.024  0.00562  4.15 3.3e-05
numst    -0.5826    0.558  0.13223 -4.41 1.1e-05
format    0.1300     1.139  0.04387  2.96 3.0e-03
postelec -0.8611     0.423  0.14062 -6.12 9.1e-10
caretakr  1.7104     5.531  0.28282  6.05 1.5e-09

         exp(coef) exp(-coef) lower .95 upper .95
invest       1.473      0.679     1.126     1.927
polar        1.024      0.977     1.012     1.035
numst        0.558      1.791     0.431     0.724
format       1.139      0.878     1.045     1.241
postelec     0.423      2.366     0.321     0.557
caretakr     5.531      0.181     3.177     9.628

Rsquare= 0.407   (max possible= 1 )
Likelihood ratio test= 164  on 6 df,   p=0
Wald test = 176  on 6 df,   p=0
Score (logrank) test = 216  on 6 df,   p=0

>
```

Thus, you get both the hazard rates and the hazard ratios from this command. Let's turn attention to interpretation. I'll reference the coefficients produced by `Stata` for the next little while.

As with any duration model, there are a variety of ways to interpret the coefficients. Take for example the binary covariates from the model above: `invest`, `postelec`, and `caretakr`.

The hazard ratio for the `invest` covariate is

```
. display exp(_b[invest])
  1.4727609
```

(which of course we could have just taken directly from the `Stata nohr` results). This implies that compared to the case when `invest=0`, the risk of a cabinet terminating is about 1.47 times greater. With respect to the `postelec` covariate, when a the government is formed immediately after the election, the risk is about .42 that of the case when protracted negotiation is necessary to form the government. That is, the risk is *lower*. Another (perhaps clearer) way to say the identical thing is look at the ratio in "reverse": $1/.4227 \approx 2.366$. This tells us that the risk is about 2.37 times greater when negotiation is required (compared to the case of immediate formation, post-election). Note that this quantity is equivalent to $\exp(-\beta)$, where β is the hazard rate (i.e. coefficient based on the `nohr` option). In passing,

note that R computes this directly. Finally, for the caretakr variable, the risk is about 5.5 times greater when the government is in a "caretaker" role.

In general, a useful way to interpret coefficients is given by:

$$\%\Delta h(t) = \Big[\frac{e^{\beta(x_i=X_1)} - e^{\beta(x_i=X_2)}}{e^{\beta(x_i=X_2)}}\Big] * 100, \tag{49}$$

where x_i is the covariate, and X_1 and X_2 denotes two distinct values of the covariate. Through this, one can assess the impact a covariate has on increasing or decreasing the hazard rate.

Substituting in values for the invest variable (i.e. 1 vs. 0), we see that the risk increases by about 47 percent when invest=1 compared to when invest=0. Different values of a continuous covariate could be substituted into the above formula to compute differences in the hazard over various values of the covariate.

Below I do a mechanical illustration. First, note the range of the forma variable.

```
. table format

----------------------
Formation | attempts  |      Freq.
----------+-----------
        1 |        179
        2 |         63
        3 |         36
        4 |         14
        5 |         12
        6 |          5
        7 |          1
        8 |          4
----------------------
```

Now compute the hazard ratio for each value of the format covariate:

```
. gen format_ratio=exp(_b[format]*format)

. table format_ratio

----------------------
format_ra | tio       |      Freq.
----------+-----------
 1.138841 |        179
 1.296959 |         63
  1.47703 |         36
```

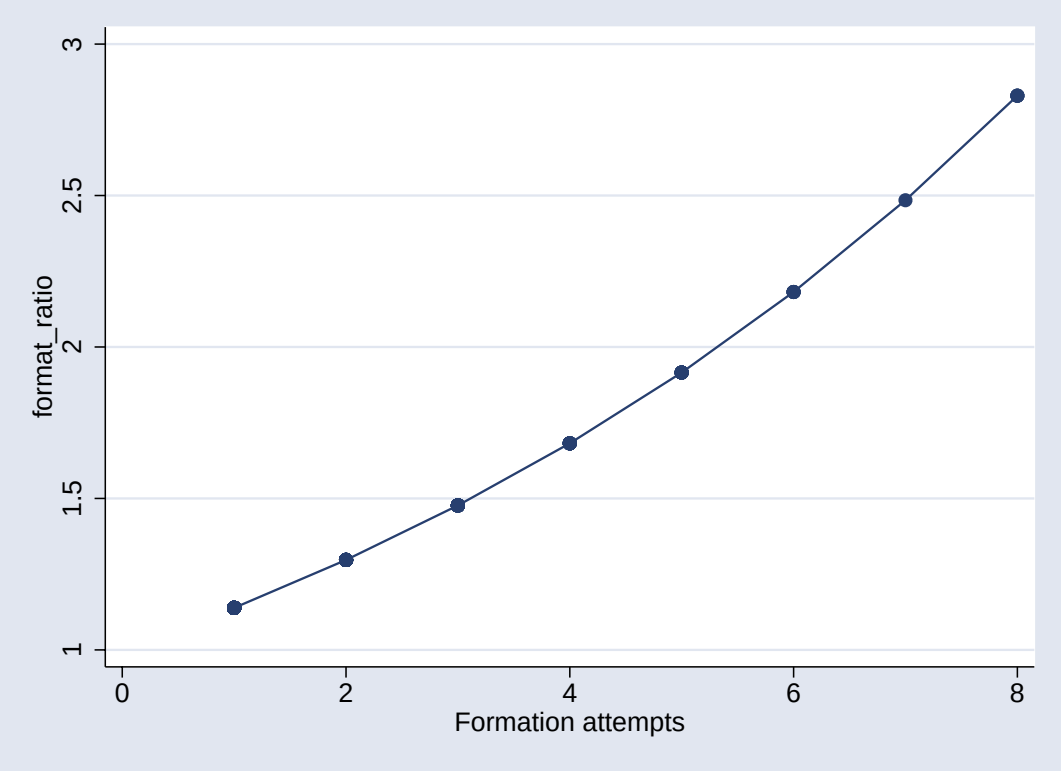

Figure 14: *Hazard ratios for various values of the formation attempts covariate.*

```
 1.682102 |          14
 1.915646 |          12
 2.181616 |           5
 2.484514 |           1
 2.829466 |           4
----------------------
```

To display the PH property, note that the ratio of each adjacent value of the `format` variable is:

```
. display 1.47703/1.296959 1.1388409
```

Why? I could graph these ratios if I wanted to:

```
. scatter format_ratio format
```

which produces Figure 14.

It's sometimes useful to visualize (and possibly report in your research) the "baseline" functions from the Cox model. It seems odd to speak of these functions as I've been saying they are not really directly estimated by the Cox model.

It turns out, however, that there is a convenient (and natural) link to the Kaplan-Meier estimator discussed the other day. Recall that the KM estimator (and the Nelson-Aalen estimator) can be used to construct nonparametric estimates of the survivor and hazard function. These two functions can be adapted readily to the Cox model to generate graphs of the survivor and hazard functions.

To illustrate this, I first mean-center the non-binary covariates (why might I want to do this?). Then I reestimate the model:

```
. egen meanpolar=mean(polar)

. egen meanform=mean(format)

. gen polarmean=polar-meanpolar

. gen formmean=format-meanform

 . stcox invest polarmean numst formmean postelec caretakr, nohr
exactm basech(inthaz) basehc(haz) basesurv(surv)

        failure _d:  censor
  analysis time _t:  durat

 Iteration 0:   log likelihood = -1000.4512
 Iteration 1:   log likelihood = -937.95835
 Iteration 2:   log likelihood = -918.35198
 Iteration 3:   log likelihood = -918.30728
 Iteration 4:   log likelihood = -918.30727
 Refining estimates:
 Iteration 0:   log likelihood = -918.30727

Cox regression -- exact marginal likelihood

 No. of subjects =          314                 Number of obs   =          314
 No. of failures =          271
 Time at risk    =       5789.5
                                                 LR chi2(6)      =    164.29
 Log likelihood  =   -918.30727                  Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      invest |   .3881819   .1372435     2.83   0.005     .1191896    .6571743
   polarmean |   .0234117   .0056238     4.16   0.000     .0123893    .0344341
       numst |  -.5842937   .1323193    -4.42   0.000    -.8436348   -.3249526
```

```
    formmean |   .1303069      .0439     2.97   0.003     .0442643    .2163494
    postelec |  -.8623465   .1407457    -6.13   0.000    -1.138203     -.58649
    caretakr |   1.735869   .2868057     6.05   0.000      1.17374    2.297997
------------------------------------------------------------------------------
```

Take a look at the `stcox` command. The options at the end of the statement will produce the Cox estimates of the baseline survivor, hazard, and integrated hazard functions. In Figure 15, I graph these functions (this figure is taken from the book, page 66; the code to produce this graph [based on Stata 7 syntax] can be found in the file `cabinet.do` on my website. Note that `Stata 8` will "do" `Stata 7`-type graphs; use the `gr7` option.)

In Figure 15, I overlay Weibull estimates of the same model on the Cox estimates. Sometimes analysts will use Cox functions to help inform them of the "ballpark" parametric model to use. This is fine . . . but it has some caveats!

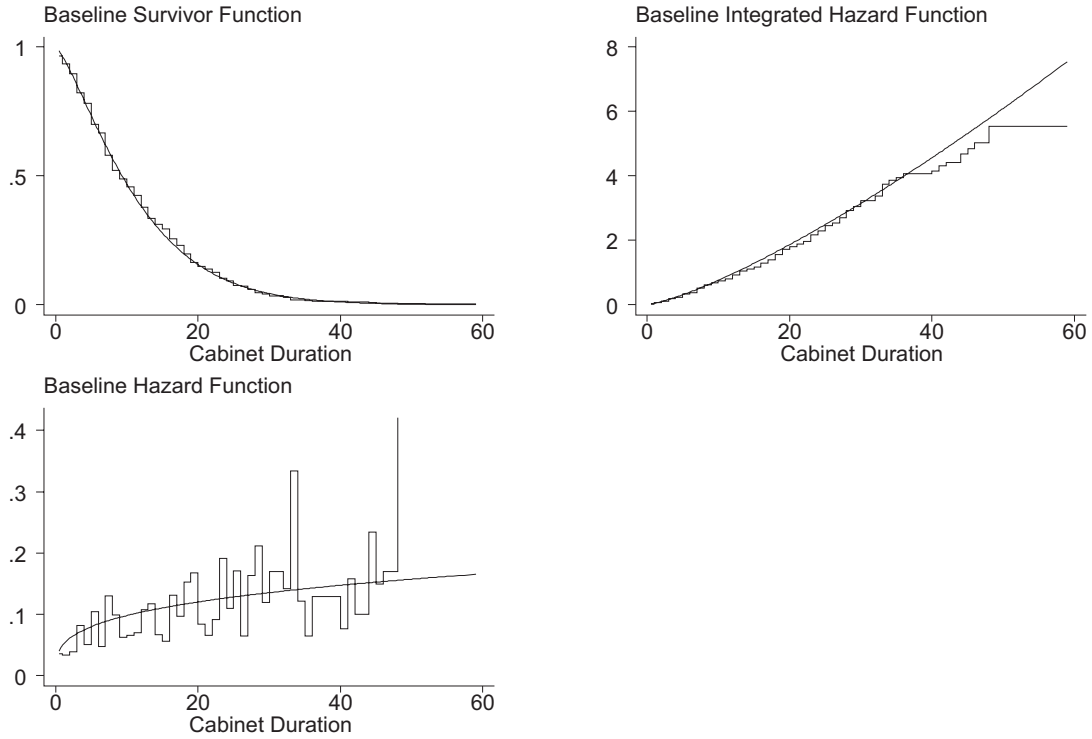

Figure 15: *This figure gives the estimated baseline survivor (top left panel), integrated hazard (top right panel) and hazard functions (bottom left panel) from the Cox Model and the Weibull Model for the cabinet duration data. The smooth function in the graphs is from the Weibull estimates; the non-smooth function in the graphs is from the Cox model.*

16 Discrete-Time Models

A very common practice in political science is the implementation of binary link models (like logit) on duration data. The reason for this is simple and we saw why vis-a-vis the counting process framework. Since duration data can be thought of in terms of a series of 0s (for censored observations) and 1s (for events), models suitable for binary dependent variables would seem suitable for duration models.

Event history data for discrete-time processes generally record the dependent variable as a series of binary outcomes denoting whether or not the event occurred at the observation point. To illustrate, consider Table 3. These data are from the study of state adoption of restrictive abortion policy (Brace, Hall, and Langer 1999) that we modeled earlier. The event of interest is whether or not a state adopted legislation that placed restrictions on abortion rights.

The starting point of the analysis is the first legislative session after the famous *Roe v. Wade* decision (1973). The question underlying these data is how long a state legislature goes before adopting restrictive legislation. Since legislation can only be adopted during a legislative session, the underlying process is assumed to be discrete.

The first column of data gives an identification number for each state. The second column of data is comprised of a sequence of zeroes and ones. A zero denotes that in that legislative session, no restrictive abortion legislation was adopted—i.e., no event occurrence is observed. A one denotes the adoption of restrictive abortion legislation—i.e., the event occurs. For discrete-time data, this is generally the form of the dependent variable used in an event history model. Finally, the third column gives the year of the legislative session in which the policy was adopted.

Table 3: 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.

Note what important thing: there is an equivalence between "duration time" variables

and binary sequences. To see this, consider case 1. We see that this state "enters" the process in 1974 (as do all states) and progresses through 14 legislative sessions until in the 1987 session, it adopts restrictive abortion legislation: the event occurs. Now, if we look at the column measuring the time that elapses from *Roe vs. Wade* until policy adoption, we see that at year 14, the state adopts restrictive legislation; again, the event occurs. *Both* measures of time lead to the same conclusion: at $t = 14$, state 1 adopted restrictive abortion legislation. The only difference between the two forms of the dependent variable is that in the case of the discrete-time formulation, the history, so to speak, is "disaggregated" into discrete intervals.

16.1 The Moving Parts of Discrete-Time Event History

We've seen the familiar functions (i.e. $f(t), S(t), h(t)$) in the context of the continuous models. Let's look at these quantities in the discrete-time model.

Let the random variable T denote a discrete random variable indicating the time of an event occurrence. Events are observable at specific points, t_i. Therefore, the probability mass function for a discrete random variable is

$$f(t) = \Pr(T = t_i), \tag{50}$$

and denotes the probability of an event occurring at time t_i. Through (50), it is clear that there can be multiple failures occurring at the same time. The survivor function for the discrete random variable T is given by

$$S(t) = \Pr(T \geq t_i) = \sum_{j \geq i} f(t_j), \tag{51}$$

where j denotes a failure time. Knowing the connection between the mass function and the survivor function, the hazard rate for the discrete-time case is given by

$$h(t) = \frac{f(t)}{S(t)}, \tag{52}$$

which demonstrates that the risk of an event occurrence is equivalent to the ratio of the probability of failure to the probability of survival.

16.2 Likelihood

Suppose that we have an event history data set consisting of n cases observed over t periods. For each observation, the dependent variable is a binary indicator coded 1 if an event occurs and 0 if an event does not occur at time t. If the event never occurs, the observation is right-censored and contributes to the data set a vector of zeroes. It can be shown (in our book, for example, pp. 71–72) the likelihood of such a data set is

$$\mathcal{L} = \prod_{i}^{n} \left[h(t_i) \prod_{i=1}^{t-1} (1 - h(t_i)) \right]^{y_{it}} \left[\prod_{i=1}^{t} (1 - h(t_i)) \right]^{1 - y_{it}}, \tag{53}$$

which is equivalent to

$$\mathcal{L} = \prod_{i=1}^{n} \Big\{ f(t) \Big\}^{y_{it}} \Big\{ S(t) \Big\}^{1-y_{it}}, \tag{54}$$

This looks very similar to the likelihood functions presented earlier this week.

For the likelihoods illustrated in the context of the parametric (and Cox) models, it was necessary to define a right-censoring indicator, which I called δ_i; for the discrete-time case with a dependent variable measured in terms of a series of binary outcomes, failure times are implicitly indexed when the dependent variable assumes the value of 1. For all other instances, the dependent variable is coded as 0.

Again we see that *only* cases experiencing an event contribute information regarding the probability of failure, i.e., $f(t)$, and cases not experiencing an event contribute information only regarding the probability of survival, i.e., $S(t)$.

16.3 Some Models

So we have binary data that *are* event history data. It therefore is natural to think about models for binary dependent variables. Guess what? We know these models already (last week you learned all about them!).

Given the structure of discrete-time data and the form the dependent variable, a wide variety of models are available to estimate covariate parameters of interest. We assume throughout this section that the analyst is not only interested in modeling the hazard probability, $h(t)$, but also interested in assessing the relationship between covariates, $\mathbf{x}$, and the probability that an event occurs. Inclusion of covariates into the model is straightforward, and is accomplished by treating the probability of failure as conditional on survival as well as covariates:

$$h(t) = \Pr(T = t_i \mid T \geq t_i, \mathbf{x}). \tag{55}$$

To simplify the presentation, let's denote the probability of an event's occurrence as $\Pr(y_{it} = 1) = \lambda_i$, and the probability of a nonoccurrence as $\Pr(y_{it} = 0) = 1 - \lambda_i$. It is assumed that this probability is a function of covariates, $\mathbf{x}$. To derive a discrete-time model, we first need to specify a distribution function for the following model:

$$\lambda_{it} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}. \tag{56}$$

Now we're cooking with gas. There are a lot of functions.

16.3.1 Logit

A commonly used function for this model is the logit function, which has the following form:

$$\log\left(\frac{\lambda_i}{1-\pi}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}. \tag{57}$$

This function specifies λ_i in terms of the log-odds ratio of the probability of an event occurrence to the probability of a nonoccurrence. The probability of an event occurrence, that

is $\hat{\lambda}_i$, can be retrieved from the logit model by reexpressing (57) directly in terms of the probability,

$$\hat{\lambda}_i = \frac{e^{\beta'\mathbf{x}}}{1 + e^{\beta'\mathbf{x}}}, \tag{58}$$

where $\exp(\beta'\mathbf{x})$ represents the exponentiated logit parameters for a given covariate profile.

16.3.2 Probit

An alternative function that can be specified for binary event history data is the probit function, which has the form

$$\Phi^{-1}[\lambda_i] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}, \tag{59}$$

where ϕ is the standard normal cumulative distribution function. The probability λ_i can be directly estimated by reparameterizing (59) as

$$\hat{\lambda}_i = \Phi(\beta'\mathbf{x}). \tag{60}$$

16.3.3 Complementary Log-Log

The complementary log-log link produces an event history model taking the form

$$\log[-\log(1 - \lambda_i)] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}. \tag{61}$$

Estimates from the complementary log-log model *can* depart substantially from those obtained by the logit or probit models. For logit and probit, the response curve is symmetric about $\lambda_i = .5$, whereas for the complementary log-log model, the response curve departs "slowly" from $\lambda_i = 0$ and approaches $\pi = 1$ very rapidly (Agresti 1990). The hazard probability, λ_i, can be computed for the complementary log-log model by reexpressing (61) as

$$\hat{\lambda}_i = 1 - \exp[-\exp(\beta'\mathbf{x})]. \tag{62}$$

These are three possible candidates; there are more but I won't discuss many more of them here. Implementing these models in conjunction with duration data is about as easy as implementing them with any kinds of data. There is an important hitch, however.

16.4 Incorporating Duration in the Discrete-Time Framework

In the parametric models, we gave structure to time dependency by specifying it in terms of some known distribution. In the Cox model, we did not directly model the baseline hazard. Instead, we assumed it takes some unknown and perhaps arbitrary form. In either case, the models gave us estimate of the covariate parameters that we could use to compute hazard rates/ratios and survivor functions.

What about the discrete-time models?

The answer is simple: ignoring time dependency in the baseline hazard produces a model that is very similar to an exponential model. Is this a problem?

Most likely. Remember we said the exponential is often an unrealistic model for most social processes. To see the "exponential equivalence" consider the logit model again. Suppose one estimates the following model

$$\log\left(\frac{\lambda_i}{1-\lambda_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \tag{63}$$

where x_{ki} are two covariates of interest that have a mean of 0, and β_0 is the constant term. The "baseline" hazard under this model would be equivalent to

$$\hat{\lambda_i} = h_0(t) = e^{\beta_0}, \tag{64}$$

which is a constant.

Importantly, the hazard probability is *flat with respect to time*. This is the same characterization that is forthcoming from an exponential. This *may* be "OK" but it usually won't be (i.e. there are other functional forms that will better fit the data).

The natural question to ask is how does one account for time dependency in the discrete-time model? There are several options:

- Ignore it
- Piecewise Functions
- Transformations on t
- Smoothing functions (like splines, lowess, etc.)

Of course "ignoring it" produces the exponential equivalent, which may not be what we want. The other choices are probably more reasonable. The advice here is this: if your primary interest are in covariate effects and you don't have a particularly compelling theory about the "shape" of temporal dependence, then try various approaches. Standard likelihood ratio tests can be used to adjudicate among different models. This is all pretty standard stuff but in my own experience, I can say three things: 1) lots of users still ignore duration dependency issues in discrete-time models; 2) accounting for t almost always produces better fitting models than not accounting for it; 3) covariate parameters are often sensitive to the function of t you choose.

16.5 Illustration

Let me briefly illustrate the points from above by replicating Table 5.2 from the book. This illustration uses data on congressional career paths from the House of Representatives. I

estimate a logit model with different functions of t.

Let's begin first with a model that does *not* model t (i.e. the exponential equivalent).

```
. logit _d rep

 Iteration 0:   log likelihood =  -1353.706
 Iteration 1:   log likelihood = -1352.1479
 Iteration 2:   log likelihood = -1352.1451

 Logit estimates                                   Number of obs  =        5054
                                                   LR chi2(1)     =        3.12
                                                   Prob > chi2    =      0.0773
 Log likelihood = -1352.1451                       Pseudo R2      =      0.0012

------------------------------------------------------------------------------
          _d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rep |   .1894179   .1069061     1.77   0.076    -.0201143    .3989501
       _cons |  -2.586274   .0720741   -35.88   0.000    -2.727536   -2.445011
------------------------------------------------------------------------------

. predict pexp, p
```

Now, I estimate the model with "linear" t:

```
. logit _d durat rep

 Iteration 0:   log likelihood =  -1353.706
 Iteration 1:   log likelihood = -1343.2309
 Iteration 2:   log likelihood = -1343.0146
 Iteration 3:   log likelihood = -1343.0144

 Logit estimates                                   Number of obs =        5054
                                                   LR chi2(2)    =       21.38
                                                   Prob > chi2   =      0.0000
 Log likelihood = -1343.0144                       Pseudo R2     =      0.0079

------------------------------------------------------------------------------
          _d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    duration |  -.0756131   .0184805    -4.09   0.000    -.1118342    -.039392
         rep |   .1585271   .1072663     1.48   0.139    -.0517111    .3687652
       _cons |  -2.257402   .1035782   -21.79   0.000    -2.460412   -2.054392
```

```
------------------------------------------------------------------------------

. predict plin, p
```

Now with lowess function (note I use the `ksm` command in `Stata`. This has been replaced in Stata 8 with `lowess` (although `ksm` still works):

```
. ksm _d duration, lowess gen(lowesst)

. logit _d lowesst rep

 Iteration 0:   log likelihood =  -1353.706
 Iteration 1:   log likelihood = -1331.9261
 Iteration 2:   log likelihood = -1330.8923
 Iteration 3:   log likelihood = -1330.8906

 Logit estimates                                   Number of obs =       5054
                                                   LR chi2(2)      =      45.63
                                                   Prob > chi2     =     0.0000
 Log likelihood = -1330.8906                       Pseudo R2       =     0.0169

------------------------------------------------------------------------------
          _d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lowesst |   13.17727   1.970735     6.69   0.000     9.314697    17.03984
         rep |   .1647638   .1074675     1.53   0.125    -.0458687    .3753964
       _cons |  -3.694489   .1877425   -19.68   0.000    -4.062457    -3.32652
------------------------------------------------------------------------------

. predict plowess, p (476 missing values generated)
```

Now with a cubic spline function. I'm using the `btscs` do file created by Richard Tucker (see Beck, Katz, and Tucker 1998).

```
. btscs _d durat memberid, g(t) nspline(3) dummy(k)

. logit _d t _spline1 _spline2 _spline3 rep

 Iteration 0:   log likelihood =  -1353.706
 Iteration 1:   log likelihood = -1330.9669
 Iteration 2:   log likelihood =  -1329.234
 Iteration 3:   log likelihood = -1329.2287
```

```
 Iteration 4:   log likelihood = -1329.2287

 Logit estimates                                    Number of obs =       5054
                                                    LR chi2(5)      =      48.95
                                                    Prob > chi2     =     0.0000
 Log likelihood = -1329.2287                        Pseudo R2 =         0.0181

------------------------------------------------------------------------------
          _d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |   -.700158   .1792494    -3.91   0.000     -1.05148   -.3488355
    _spline1 |  -.1286026   .0701528    -1.83   0.067    -.2660995    .0088942
    _spline2 |   .0465343   .0426414     1.09   0.275    -.0370412    .1301098
    _spline3 |  -.0021472   .0130642    -0.16   0.869    -.0277526    .0234582
         rep |   .1654837   .1077437     1.54   0.125    -.0456901    .3766574
       _cons |  -1.984141   .1073376   -18.49   0.000    -2.194519   -1.773763
------------------------------------------------------------------------------

. predict pspline, p (476 missing values generated)
```

Now $\log(t)$:

```
. gen logdur=log(duration)

. logit _d logdur rep

 Iteration 0:   log likelihood =  -1353.706
 Iteration 1:   log likelihood =  -1335.884
 Iteration 2:   log likelihood =  -1335.541
 Iteration 3:   log likelihood = -1335.5409

 Logit estimates                                    Number of obs  =       5054
                                                    LR chi2(2)      =      36.33
                                                    Prob > chi2     =     0.0000
 Log likelihood = -1335.5409                        Pseudo R2 = 0.0134

------------------------------------------------------------------------------
          _d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      logdur |    -.38962   .0679163    -5.74   0.000    -.5227335   -.2565064
         rep |     .15602   .1073906     1.45   0.146    -.0544618    .3665017
       _cons |  -2.136548   .1014298   -21.06   0.000    -2.335346   -1.937749
------------------------------------------------------------------------------
```

```
. predict plogt, p
```

Now quadratic:

```
. gen dur2=durat^2

. logit _d durat dur2 rep

 Iteration 0:   log likelihood =  -1353.706
 Iteration 1:   log likelihood = -1338.0372
 Iteration 2:   log likelihood = -1337.7571
 Iteration 3:   log likelihood =  -1337.757

 Logit estimates                                   Number of obs =       5054
                                                   LR chi2(3)      =      31.90
                                                   Prob > chi2     =     0.0000
 Log likelihood =  -1337.757                       Pseudo R2       =     0.0118

------------------------------------------------------------------------------
          _d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    duration |  -.2235403   .0477148    -4.68   0.000    -.3170597    -.130021
        dur2 |   .0122237   .0034901     3.50   0.000     .0053833    .0190641
         rep |   .1667169   .1075416     1.55   0.121    -.0440608    .3774947
       _cons |  -1.984996   .1296434   -15.31   0.000    -2.239093     -1.7309
------------------------------------------------------------------------------

. predict pquad, p
```

To evaluate the "best fitting" model, I could simply perform some likelihood ratio tests. Since each model is comparable to a null model of no duration dependency, the test statistic given by

$$LR = -2\log(L_0 - L_1), \tag{65}$$

could be used. For example, comparing the lowess model to the null model, I would obtain

```
. display -2*( -1352.1451 -  -1330.7979 ) 42.6944
```

On 1 degree of freedom, this is significant at any conventional level. That is, at least when compared to the null model, the lowess specification fits the data better. For the spline model, I find

```
. display -2*( -1352.1451 - -1329.2287  )
 45.8328
```

which again is statistically significant (on 4 degrees of freedom). Of the specifications tested, I might slightly prefer the cubic spline function over the lowess function.[2]

I plot the hazard probabilities using the `Stata 7` command below:

```
. gr pspline plowess durat, c(ss) s(ii) ylab xlab t1("Hazard
Functions") b2("Duration of House Career") saving(housecareer2,
replace)
```

This produces Figure 16. Here, the party identification variable is set to 0; hence the baseline category represents the hazard for Republicans. This corresponds to Figure 5.2 from our book.

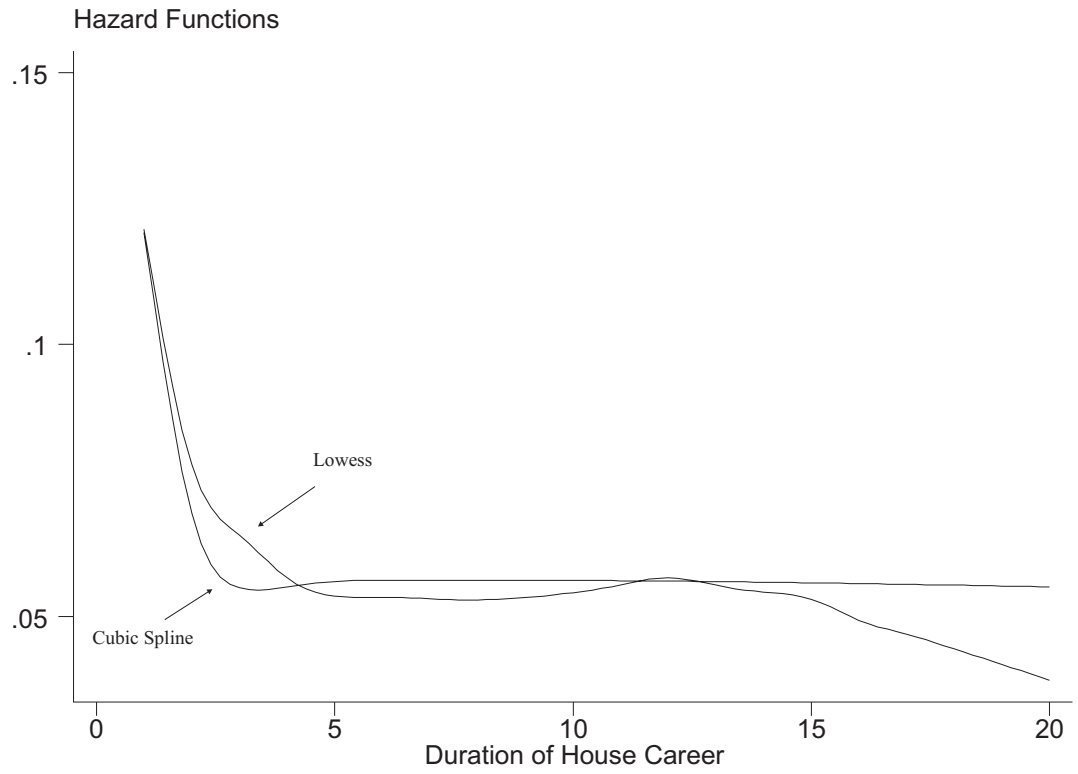

Figure 16: *The lines represent the estimated baseline hazard from the discrete-time logit model of House career paths. The cubic spline and lowess functions are presented.*

[2]Note that the estimates from the models above differ *very* slightly from the analysis in the book. I found a couple of coding errors after I did the initial analysis. Importantly, nothing changes regarding conclusions based on these models.

17 Conclusion

We've covered a lot of ground. Between the parametric, Cox, and discrete models, we have a lot of modeling strategies at our disposal. The natural question to ask is, which strategy is "best"?

This question doesn't have an easy answer. In our book, Chapter 6, we discuss these issues in great detail. In general, Janet and I argue that most applied work should naturally lead to the Cox model; however, there will be clear instances when other strategies are preferred. If one is interested in predicting hazards, especially out-of-sample predictions, parametric (or perhaps discrete) models may be preferable. There is another class of models I haven't discussed (but do so in the book). These so-called flexible parametric models developed by Royston and Parmar seem to provide a fruitful middle ground between Cox and parametric models.