An Introduction to Event History Analysis

Oxford Spring School
June 18-20, 2007

Day Two: Regression Models for Survival Data

Parametric Models

We'll spend the morning introducing regression-like models for survival data, starting with fully parametric (distribution-based) models. These tend to be very widely used in social sciences, although they receive almost no use outside of that (e.g., in biostatistics).

A General Framework (That Is, Some Math)

Parametric models are *continuous-time models*, in that they assume a continuous parametric distribution for the probability of failure over time. A general parametric duration model takes as its starting point the *hazard*:

$$h(t) = \frac{f(t)}{S(t)} \tag{1}$$

As we discussed yesterday, the density is the probability of the event (at T) occurring within some differentiable time-span:

$$f(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t)}{\Delta t}. \tag{2}$$

The survival function is equal to one minus the CDF of the density:

$$\begin{aligned} S(t) &= \Pr(T \geq t) \\ &= 1 - \int_0^t f(t)\, dt \\ &= 1 - F(t). \end{aligned} \tag{3}$$

This means that another way of thinking of the hazard in continuous time is as a conditional limit:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{4}$$

i.e., the conditional probability of the event as Δt gets arbitrarily small.

A General Parametric Likelihood

For a set of observations indexed by i, we can distinguish between those which are censored and those which aren't...

- *Uncensored observations* ($C_i = 1$) tell us both about the hazard of the event, and the survival of individuals prior to that event.
    - That is, they tell us the exact time of failure.
    - This means that, in terms of the likelihood, they contribute their *density*.
- *Censored observations* ($C_i = 0$) tell us only that that observation survived at least to time T_i.
    - This means that they contribute information through their *survival* function.

We can thus combine these in a simple fashion, into a general parametric likelihood for survival models:

$$L = \prod_{i=1}^{N} \left[f(T_i)\right]^{C_i} \left[S(T_i)\right]^{1-C_i} \tag{5}$$

with the corresponding log-likelihood:

$$\ln L = \sum_{i=1}^{N} \left\{C_i \ln\left[f(T_i)\right] + (1 - C_i) \ln\left[S(T_i)\right]\right\} \tag{6}$$

which we can maximize using standard (e.g. Newton / Fisher scoring) methods. To include covariates, we simply condition the terms of the likelihood on $\mathbf{X}$ and the associated parameter vector β, e.g.

$$\ln L = \sum_{i=1}^{N} \left\{C_i \ln\left[f(T_i|\mathbf{X}, \beta)\right] + (1 - C_i) \ln\left[S(T_i|\mathbf{X}, \beta)\right]\right\} \tag{7}$$

This general model can be thought of as encompassing all the various models we'll discuss this morning; the only differences among them are the distributions assumed for the probability of the event of interest.

The Exponential Model

The exponential is the simplest parametric duration model. There are (at least) two ways to motivate the exponential model. First, consider a model in which the hazard is constant over time:

$$h(t) = \lambda \tag{8}$$

This model assumes that the hazard of an event is constant over time (i.e., "flat"), which implies that the conditional probability of the event is the same, no matter when the observation is observed.

Put differently,

- Events occur according to a Poisson process (independent/"memoryless").
- This also means that the accumulated/integrated hazard $H(t)$ is simply the hazard times t, i.e.,

$$H(t) \equiv \int_0^t h(t)\, dt = \lambda t \tag{9}$$

Recalling that the survival function is equal to the exponentiated negative cumulative hazard:

$$S(t) = \exp[-H(t)]$$

we can write the survival function for the exponential model as:

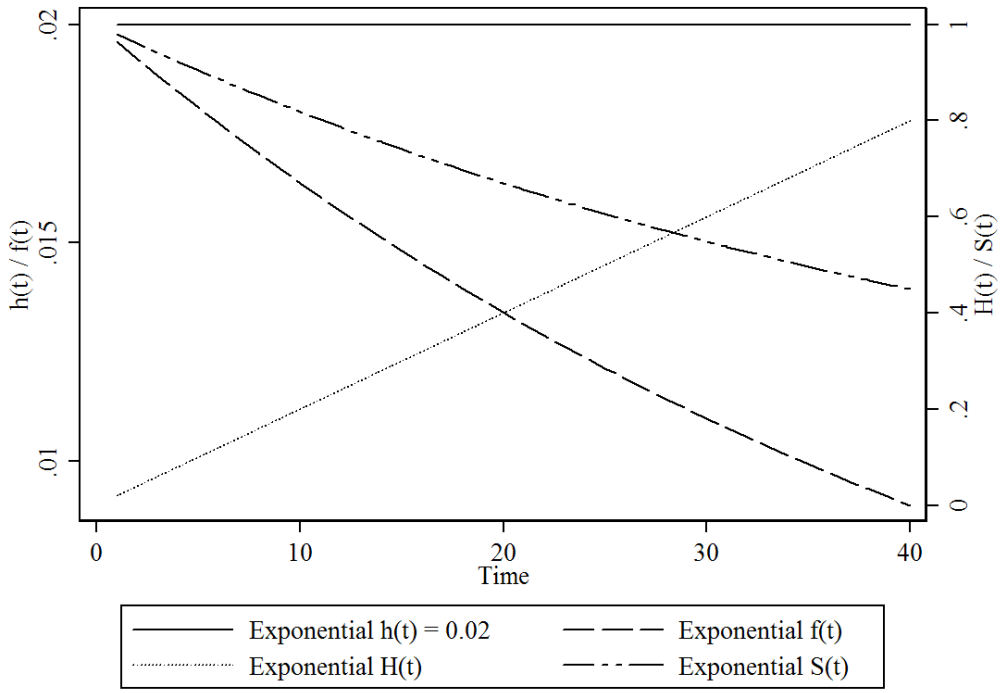
$$S(t) = \exp(-\lambda t) \tag{10}$$

This, in turn, means that the density $f(t)$ equals:

$$\begin{aligned} f(t) &= h(t)S(t) \\ &= \lambda \exp(-\lambda t) \end{aligned} \tag{11}$$

Given $f(t)$ and $S(t)$, we can plug these values into the log-likelihood equation (6) and estimate a value for $\hat{\lambda}$ via MLE.

Figure 1: Various Functions of an Exponential Model with $\lambda = 0.02$



Covariates

To introduce covariates, we have to consider that the hazard λ must always be positive. The natural solution is to allow the covariates to enter exponentially:

$$\lambda_i = \exp(\mathbf{X}_i\beta). \tag{12}$$

Note what this implies for the survival function: that

$$S_i(t) = \exp(-e^{\mathbf{X}_i\beta}t). \tag{13}$$

In other words, the survival function is "linked" to the covariates through a cumulative log-log function. Similarly, the full log-likelihood for the exponential model becomes

$$\begin{aligned} \ln L &= \sum_{i=1}^{N} \left\{ C_i \ln \left[\exp(\mathbf{X}_i\beta) \exp(-e^{\mathbf{X}_i\beta}t) \right] + (1 - C_i) \ln \left[\exp(-e^{\mathbf{X}_i\beta}t) \right] \right\} \\ &= \sum_{i=1}^{N} \left\{ C_i \left[(\mathbf{X}_i\beta)(-e^{\mathbf{X}_i\beta}t) \right] + (1 - C_i)(-e^{\mathbf{X}_i\beta}t) \right\} \end{aligned} \tag{14}$$

Estimation is accomplished via MLE of the likelihood in (14). We'll get to interpretation in a bit...

Another Motivation: The Accelerated Failure Time Approach

Another motivation for parametric models is via a regression-type framework, involving a model of the kind:

$$\ln T_i = \mathbf{X}_i\gamma + \epsilon_i \tag{15}$$

That is, as an explicit regression-type model of (the log of) survival time. In this instance, we consider the logged value mainly because survival time distributions tend to be right-skewed, and the exponential is a simple distribution with this characteristic.

This general format is known as the *accelerated failure-time* (AFT) form of duration models, and is most widely used in economics and engineering (though has also seen some use in biostatistics in recent years). It corresponds to a model of

$$T_i = \exp(\mathbf{X}_i\gamma) \times u_i \tag{16}$$

where $\epsilon_i = \ln(u_i)$. In this framework, we would normally posit that u_i follows an exponential distribution, so that ϵ_i takes on what is known as an *extreme value distribution*, here with mean zero and variance equal to 1.0.

As in a standard regression model, we can rewrite model (16) as

$$\epsilon_i = \ln T_i - \mathbf{X}_i\gamma \tag{17}$$

In standard fashion, then, the differences between the observed and predicted values (that is, the residuals) follow an extreme-value distribution. We can then substitute these into the likelihood given before, take derivatives w.r.t. the parameters of interest, set to zero and solve to get our MLEs. Standard errors are obtained in the usual way, as the negative of the "information" matrix (the matrix of second derivatives of the log-likelihood with respect to the parameters), and confidence intervals can be created by a "bounds" method.

A Note on the AFT Formulation

Parametric models are widely used in engineering, in studies of the reliability of components, etc. These studies often want to know, e.g., how long some piece of equipment will last before it "gives out." But it's often hard to do testing under typical circumstances – in particular, it can take too long.

In such situations, researchers expose equipment to more extreme conditions than will normally be experienced (i.e., accelerate the failure process), and then do out-of-sample predictions. Under such circumstances, a fully parametric model is better than a semiparametric one which relies on intercomparability of subjects to assess covariate effects.

Which is better for social scientists? I'd argue that (most of the time) the Cox model is, but we'll talk more about that later today...

Interpretation

The exponential model can be interpreted in either AFT or hazard form. In either case, the coefficient estimates are identical, but the sign is reversed: variables which increase the hazard decrease the expected (log-) duration, and vice-versa.

Here is an example, with data on U.S. Supreme Court justice retirements ($N = 107$) and no covariates (that is, we're just estimating a model of the "mean" hazard/failure time):

Hazard rate form:

```
. streg , nohr dist(exp)

Exponential regression -- log relative-hazard form

No. of subjects =          107                     Number of obs   =      1765
No. of failures =           51
Time at risk    =         1778
                                                    LR chi2(0)      =     -0.00
Log likelihood  =   -100.83092                      Prob > chi2     =         .

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -3.551419    .140028   -25.36   0.000    -3.825869   -3.276969
------------------------------------------------------------------------------
```

AFT form:

```
. streg , time dist(exp)

Exponential regression -- accelerated failure-time form

No. of subjects =          107                     Number of obs   =      1765
No. of failures =           51
Time at risk    =         1778
                                                    LR chi2(0)      =     -0.00
Log likelihood  =   -100.83092                      Prob > chi2     =         .

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   3.551419    .140028    25.36   0.000     3.276969    3.825869
------------------------------------------------------------------------------
```

The AFT form has a very convenient interpretation: Since the "dependent variable" is $\ln(T)$, we can directly translate the estimated coefficient $\hat{\gamma}_0$ by noting that $\hat{T} = \exp(\hat{\gamma}_0)$. That means that our estimated (mean) survival time is $\exp(3.5514) = 34.86$.

Similarly, if we include covariates:

```
. streg age pension pagree, nohr dist(exp)

Exponential regression -- log relative-hazard form

No. of subjects =          107                     Number of obs   =      1765
No. of failures =           51
Time at risk    =         1778
                                                    LR chi2(3)      =     44.96
Log likelihood  =   -78.351408                      Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .041146   .0210568     1.95   0.051    -.0001246    .0824165
     pension |   1.321105   .3870769     3.41   0.001     .5624487    2.079762
      pagree |   .1069193   .2854665     0.37   0.708    -.4525849    .6664234
       _cons |  -6.837742   1.341503    -5.10   0.000    -9.467039   -4.208445
------------------------------------------------------------------------------
```

```
. streg age pension pagree, time dist(exp)

Exponential regression -- accelerated failure-time form

No. of subjects =          107                     Number of obs   =      1765
No. of failures =           51
Time at risk    =         1778
                                                   LR chi2(3)      =     44.96
Log likelihood  =   -78.351408                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   -.041146   .0210568    -1.95   0.051    -.0824165    .0001246
     pension |  -1.321105   .3870769    -3.41   0.001    -2.079762   -.5624487
      pagree |  -.1069193   .2854665    -0.37   0.708    -.6664234    .4525849
       _cons |   6.837742   1.341503     5.10   0.000     4.208445    9.467039
------------------------------------------------------------------------------
```

Here,

- `age` is the justice's age in years,
- `pension` is coded 1 if the justice is eligible for a pension, 0 if not, and
- `pagree` is coded 1 if the party of the sitting president is the same as the party of the president that appointed the justice, and 0 otherwise.

Once again, the coefficients in the AFT model have a simple interpretation: Every one unit change in X_k corresponds to a change of $100 \times [1 - \exp(\hat{\gamma}_k)]$ percent in the expected survival time. That means that, in this case:

- Every year older a justice is decreases their time on the Court by $100 \times [1 - \exp(-0.041)] = 100 \times 0.0401 = 4.01$ – that is, reduces it by approximately four percent.
- The availability of a pension decreases the predicted survival time by $100 \times [1 - \exp(-1.32)] = 100 \times (1 - 0.267) = 73.3$ – a whopping seventy three percent decrease.

Hazard Ratios

In the hazard rate model, the coefficients have a similarly easy interpretation. In considering hazards, it is often useful to consider the *hazard ratio*: that is, the ratio of the hazard for two observations with different values of X_k:

$$\text{HR}_k = \frac{\hat{h}(t)|X_k = 1}{\hat{h}(t)|X_k = 0} \tag{18}$$

A hazard ratio of 1.0 corresponds to no difference in the hazard for the two observations; hazard ratios greater than one mean that the presence of the covariate increases the hazard of the event of interest, while a rate less than zero means that the covariate in question decreases that hazard.

In the context of the exponential model, note that, since we have defined $h(t) = \lambda$, a (conditional) constant, and $\lambda_i = \exp(\mathbf{X}_i\beta)$, then we can rewrite (12) as

$$h_i(t) = \exp(\beta_0)\exp(\mathbf{X}_i\beta)$$

This, in turn, means that the hazard ratio can be seen to be:

$$\begin{aligned}
\text{HR}_k &= \frac{\hat{h}(t)|X_k = 1}{\hat{h}(t)|X_k = 0} \\
&= \frac{\exp(\hat{\beta}_0 + X_1\hat{\beta}_1 + ... + \hat{\beta}_k(1) + ...)}{\exp(\hat{\beta}_0 + X_1\hat{\beta}_1 + ... + \hat{\beta}_k(0) + ...)} \\
&= \frac{\exp(\hat{\beta}_k)(1)}{\exp(\hat{\beta}_k)(0)} \\
&= \exp(\hat{\beta}_k)
\end{aligned} \tag{19}$$

In other words, the hazard ratio is just the exponentiated coefficient estimate from the hazard-rate form of the model. Thus, we would say that:

- A one-unit difference in the age of justices corresponds to a hazard ratio of $\exp(0.4115) = 1.042$. In words, the model suggests that the hazard of retirement for a given justice is roughly 1.04 times that that of (or four percent greater than) a justice one year her junior.
- Similarly, the hazard ratio of pension-eligible to non-pension-eligible justices is $\exp(1.32) = 3.748$; in words, the hazard for a retirement-eligible justice is nearly three times greater than one that is not so eligible.

More generally, the hazard ratio for two observations that differ by δ on the covariate X_k is just

$$\begin{aligned} \text{HR}_k &= \frac{\hat{h}(t)|X_k + \delta}{\hat{h}(t)|X_k} \\ &= \exp(\delta\,\hat{\beta}_k) \end{aligned}$$

Even more generally, the hazard ratio for two observations with covariate values $\mathbf{X}_i$ and $\mathbf{X}_j$ is:

$$\text{HR}_{\frac{i}{j}} = \frac{\exp(\mathbf{X}_i\hat{\beta})}{\exp(\mathbf{X}_j\hat{\beta})} \tag{20}$$

Finally, note that equation (19) illustrates a key property of the exponential model: that it is a model of *proportional hazards*. This means that the hazards for any two observations which vary on a particular covariate are proportional to one another – or, alternatively, that the ratio of two such hazards is a constant with respect to time. This is a topic we'll come back to, as it has some potentially important implications.

The Weibull Model

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

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

$$h(t) = \lambda p(\lambda t)^{p-1} \tag{21}$$

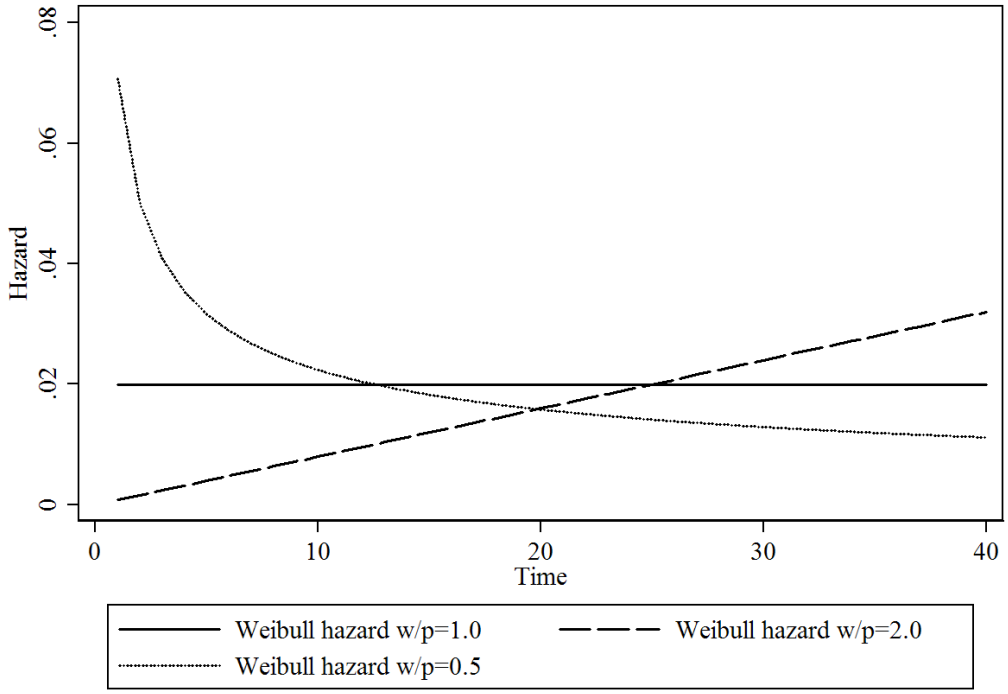
Here, the parameter p is sometimes called a "shape parameter," because it defines the shape of the Weibull distribution.

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

Recalling that the survival function can be expressed as the exponent of the negative integrated hazard, we can see that:

$$\begin{aligned} S(t) &= \exp\left[-\int_0^t \lambda p(\lambda t)^{p-1}\,dt\right] \\ &= \exp(-\lambda t)^p \end{aligned} \tag{22}$$

Figure 2: Various Weibull Hazards with $\lambda = 0.02$



and the corresponding density is the product of (21) and (22):

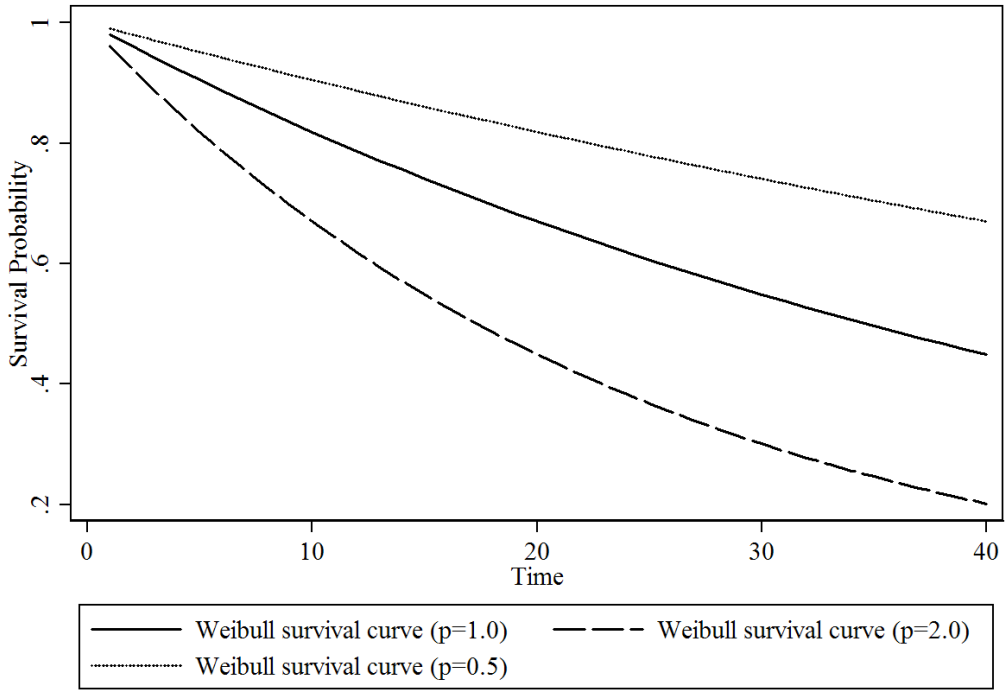
$$f(t) = \lambda p(\lambda t)^{p-1} \times \exp(-\lambda t)^p \tag{23}$$

The Weibull is thus a two-parameter distribution, with the initial parameter λ denoting the overall level of the hazard and the parameter p determining its shape (increasing, decreasing, or constant). As in the exponential case, we typically introduce covariates through an exponential function, to ensure that the hazard remains positive:

$$\lambda_i = \exp(\mathbf{X}_i\beta)$$

This can then be substituted into the density and survival functions in (22) and (23), which in turn can be "plugged in" to the general log–likelihood in (6) and maximized with respect to the parameters of interest β and p.

Figure 3: Various Weibull Survival Functions with $\lambda = 0.02$



The Weibull Model in an AFT formulation

As with the exponential model, the Weibull can also be considered a log-linear model of time, but one where the errors are generalized. To do so, we replace equation (16) with a more general version:

$$T_i = \exp(\mathbf{X}_i\gamma) \times \sigma u_i \tag{24}$$

In contrast to the AFT formulation of the exponential model, here the extreme-value distribution imposed upon the errors u_i is not constrained to have variance equal to 1.0. Instead, the variance of the "errors" can take on any positive value. The result is a model where we estimate both the direct effects of the covariates on the (log of) survival time ($\hat{\gamma}$) and the variance of the "errors" ($\hat{\sigma}$). Importantly, because of the nonlinearity of the model, these two quantities are related; in particular, it is important to note that $\partial T_i/\partial \mathbf{X}$ contains both γ and σ. This reflects why some parameterizations of the Weibull model (particularly in engineering and economics, e.g., that of Greene) refer to "σ" rather than p.

In fact, however, the two models are, as in the exponential case, fully (if somewhat more complicatedly) equivalent. Their equivalence can be seen through two factors:

1. $p = 1/\sigma$.
    - This means that the "shape" function is determined by the variance of the "residuals" in the AFT form.
    - More intuitively, data which have smaller error variances (that is, smaller σ) will also tend to exhibit positive duration dependence; this is because of their relative lack of heterogeneity, which in turn reduces adverse selection in the event process.
    - Conversely, data with high error variances will tend to exhibit negative duration dependence (and thus have $p > 1$).
2. $\beta = -\gamma/\sigma$
    - Or, equivalently, $\gamma = -\beta/p$.
    - In words, the parameters β and γ are equivalent up to a scale parameter p or σ.

Model Interpretation

As was the case for the exponential, the Weibull model is both a proportional hazards model and an AFT model. This means that the interpretations of parameter values used in the exponential case can, with some minor changes, also be used in the Weibull case as well. So, for example, in the hazards framework, the hazard ratio for two observations with different values i and j on a covariate vector $\mathbf{X}$ is the same as in the exponential case:

$$\text{HR}_{\frac{i}{j}} = \frac{\exp(\mathbf{X}_i\hat{\beta})}{\exp(\mathbf{X}_j\hat{\beta})}$$

This readily simplifies, so that (e.g.) the hazard ratio for two cases that differ on only one dichotomous variable is just $\exp(\hat{\beta})$. As with the exponential, estimates $\hat{\beta}$ reflect the impact of the covariate in question on the hazard rate, so positive values indicate that higher values of the variable correspond to higher hazards of the event (and thus to shorter expected durations) and vice versa. Conversely, a positive estimate for $\hat{\gamma}$ means that increases in that variable increase the expected duration until the event of interest (that is, lower the hazard). Also useful in the Weibull case are plots of the predicted hazards for various covariate values; we'll talk about how one can do this in a bit.

An Example

Consider again the models of Supreme Court retirements we discussed above:

```
. streg age pension pagree, nohr dist(weib)

Weibull regression -- log relative-hazard form

No. of subjects =          107                     Number of obs   =      1765
No. of failures =           51
Time at risk    =         1778
                                                   LR chi2(3)      =     26.79
Log likelihood  =   -78.350319                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0405689   .0244068     1.66   0.096    -.0072675    .0884052
     pension |   1.317683   .3935936     3.35   0.001     .5462542    2.089113
      pagree |   .1084476   .2872933     0.38   0.706    -.4546368    .6715321
       _cons |   -6.83356   1.344091    -5.08   0.000     -9.46793    -4.19919
-------------+----------------------------------------------------------------
       /ln_p |   .0103713   .2216177     0.05   0.963    -.4239915     .444734
-------------+----------------------------------------------------------------
           p |   1.010425   .2239281                      .6544294    1.560075
         1/p |   .9896823   .2193312                      .6409947    1.528049
------------------------------------------------------------------------------


. streg age pension pagree, time dist(weib)

Weibull regression -- accelerated failure-time form

No. of subjects =          107                     Number of obs   =      1765
No. of failures =           51
Time at risk    =         1778
                                                   LR chi2(3)      =     26.79
Log likelihood  =   -78.350319                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0401503   .0296687    -1.35   0.176    -.0982999    .0179993
     pension |  -1.304088   .5264274    -2.48   0.013    -2.335867   -.2723093
      pagree |  -.1073287   .2826184    -0.38   0.704    -.6612506    .4465932
       _cons |   6.763054     2.0678     3.27   0.001     2.710241    10.81587
-------------+----------------------------------------------------------------
       /ln_p |   .0103713   .2216177     0.05   0.963    -.4239915     .444734
-------------+----------------------------------------------------------------
           p |   1.010425   .2239281                      .6544294    1.560075
         1/p |   .9896823   .2193312                      .6409947    1.528049
------------------------------------------------------------------------------
```

Note a few things:

- In this (unfortunate, in this respect) example, the estimate for $\ln(p)$ is essentially equal to zero, which implies that $p = 1$ and therefore that the Weibull model is not a better "fit" than the exponential.
- Relatedly, the coefficient estimates $\hat{\beta}$ and $\hat{\gamma}$ are nearly identical, except for the signs; only if σ / p is significantly different from 1.0 will the exponential and Weibull models diverge.

Other Parametric Models

The Weibull and the exponential are (in that order) the most commonly used parametric survival models in the social sciences. There are, however, a pretty fair number of other parametric models that also have received some attention in the social sciences.

The Gompertz Model

In contrast to the log-normal and log-logistic, the Gompertz is *only* a proportional hazards model. It has been most widely used in demography, and to a lesser extent in sociology. It is a two-parameter distribution; its hazard is of the form:

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

with a corresponding survival function

$$S(t) = \exp\left[-\frac{e^{\lambda}}{\gamma}(e^{\gamma t} - 1)\right] \tag{26}$$

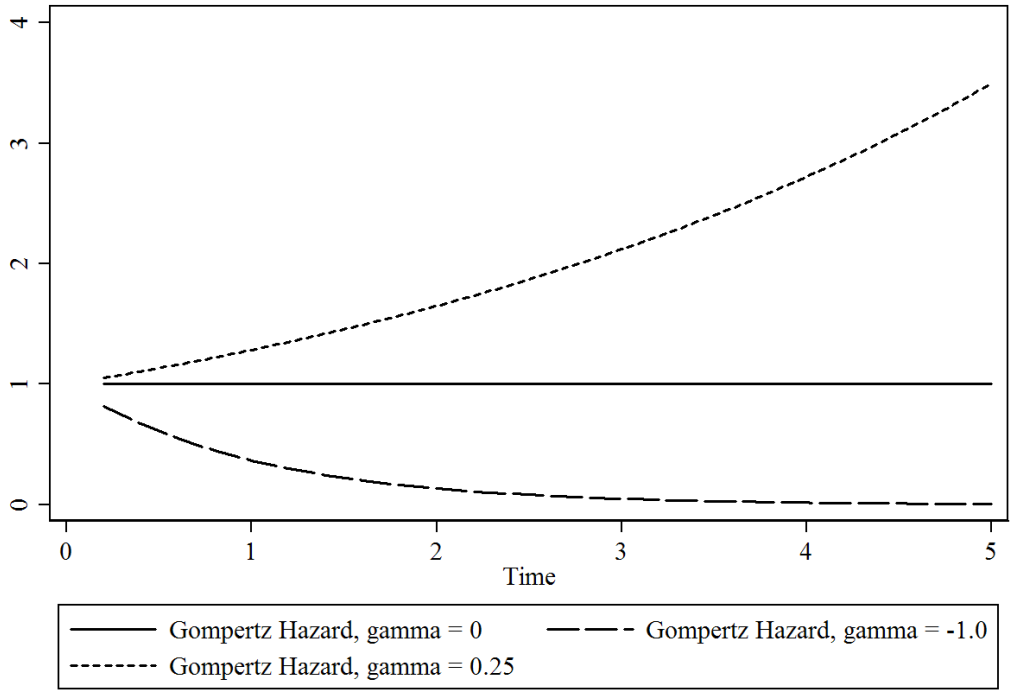
and a density equal to the product of (25) and (26). As above, the parameter λ models the first moment, such that we can allow covariates that we believe influence the level of the hazard to enter according to

$$\lambda_i = \exp(\mathbf{X}_i\beta).$$

The second parameter, γ, is a shape parameter along the lines of the p parameter in the Weibull model. Here,

- When $\gamma = 0$, the hazard is constant, and the model largely corresponds to an exponential model,
- when $\gamma > 0$, the hazards are monotonically increasing over time, and
- when $\gamma < 0$, the hazards are decreasing monotonically over time.

Figure 4: Various Gompertz Hazards with $\lambda = 0.001$



Accordingly, a test for $\hat{\gamma} = 0$ essentially tests for whether the conditional hazard is constant over time.

The Gompertz model is, as noted above, a proportional hazards model; as a result, the hazard ratio for a particular covariate X_k is simply $\exp(\hat{\beta}_k)$ – that is, the same as in the exponential and Weibull models.

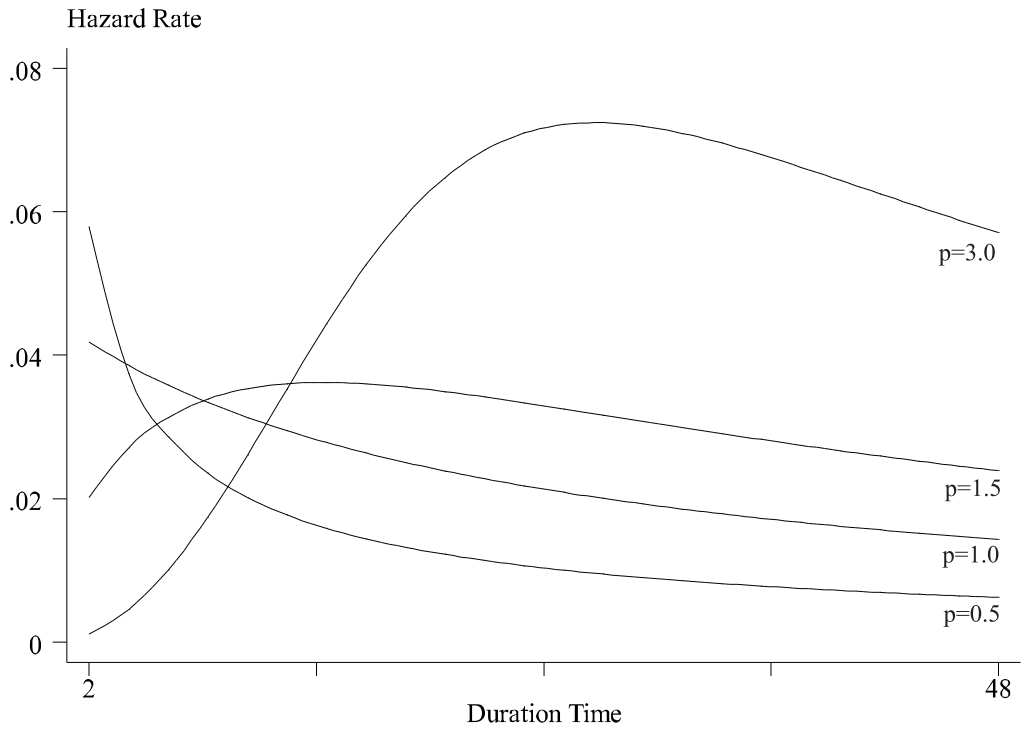
The Log-Normal and Log-Logistic Models

Probably the most commonly-used models in the social sciences beyond the exponential and Weibull are the log-normal and log-logistic models. Both of these are strictly AFT models, in that they begin with the log-linear model implicit in (24):

$$\ln(T_i) = \mathbf{X}_i\beta + \sigma\epsilon_i \tag{27}$$

If the errors ϵ in this model are assumed to follow a logistic distribution, then the resulting model is the log-logistic. Similarly, if the errors in (27) are assumed to be distributed according to a standard normal distribution, the log-normal model is implied. As with the Weibull distribution, both are two-parameter models, with a central tendency parameter λ and a "shape" parameter $p \equiv 1/\sigma$.

Figure 5: Various Log-Logistic Hazards (my thanks to Brad Jones for the figure)



The Log-Logistic Model

The log-logistic model has a survival function equal to

$$S(t) = \frac{1}{1 + (\lambda t)^p} \tag{28}$$

and a corresponding hazard function

$$h(t) = \frac{\lambda p (\lambda t)^{p-1}}{1 + (\lambda t)^p} \tag{29}$$

Because the terms in the denominator of (28) and (29) are identical, the density – which, again, is just the product of the two – takes on an especially simple form:

$$f(t) = \frac{\lambda p (\lambda t)^{p-1}}{[1 + (\lambda t)^p]^2} \tag{30}$$

Equation (30) defines a symmetric density for the instantaneous probability of the event. As in the Weibull case, covariates are typically entered into the model exponentially (as $\lambda_i = \exp(\mathbf{X}_i \beta)$).

Important: Note that many software packages – including Stata – parameterize p as $\gamma = \frac{1}{p}$. Bear this in mind as you consider your actual results.

The Log-Normal Model

The log–normal model is very similar to the log-logistic, in that the density of the "errors" ϵ are assumed to have a bell-shaped symmetrical (here, standard normal) distribution. If we think of the "errors" as being normally distributed (that is, as $\phi(\cdot)$), then the cumulative errors are cumulative normal. Because of that, we can write the survival function for the log-normal as:

$$S(t) = 1 - \Phi\left[\frac{\ln T - \ln(\lambda)}{\sigma}\right] \tag{31}$$

Log-Logistic / Log-Normal Commonalities

In general the log-logistic and log-normal models are very similar, and will yield similar results – think of them as somewhat like logit and probit in models for binary responses. For example, the results for the two models using our Supreme Court data are:

```
. streg age pension pagree, dist(loglog)

Log-logistic regression -- accelerated failure-time form

No. of subjects =          107                     Number of obs   =      1765
No. of failures =           51
Time at risk    =         1778
                                                   LR chi2(3)      =     19.36
Log likelihood  =   -84.359393                     Prob > chi2     =    0.0002

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0142307   .0220574     0.65   0.519     -.029001    .0574624
     pension |  -2.846252   3.764288    -0.76   0.450    -10.22412    4.531617
      pagree |   .0806037   .4969815     0.16   0.871    -.8934623     1.05467
       _cons |   2.770183   1.381703     2.00   0.045     .0620944    5.478271
-------------+----------------------------------------------------------------
     /ln_gam |    -.38227   .2162617    -1.77   0.077    -.8061351     .041595
-------------+----------------------------------------------------------------
       gamma |   .6823108   .1475577                      .4465807    1.042472
------------------------------------------------------------------------------
```

```
. streg age pension pagree, dist(lognorm)

Log-normal regression -- accelerated failure-time form

No. of subjects =          107                     Number of obs   =      1765
No. of failures =           51
Time at risk    =         1778
                                                    LR chi2(3)      =     26.77
Log likelihood  =    -81.42114                      Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0124893    .029725    -0.42   0.674    -.0707492    .0457705
     pension |  -3.598691   1.642479    -2.19   0.028    -6.817891    -.379491
      pagree |   .0354144   .4266008     0.08   0.934    -.8007078    .8715367
       _cons |   4.632599   1.873923     2.47   0.013     .9597768    8.305422
-------------+----------------------------------------------------------------
     /ln_sig |   .3819538   .2020201     1.89   0.059    -.0139983     .777906
-------------+----------------------------------------------------------------
       sigma |   1.465144   .2959887                      .9860992    2.176909
------------------------------------------------------------------------------
```

Also, a key trait of both the log-normal and log-logistic models is that they allow for the possibility of non-monotonic hazards. In particular, the log–logistic model with $p > 1$, and the log-normal for most values of p, both have hazards which first rise, then fall over time.

The Generalized Gamma Model

The generalized gamma is, as the model suggests, a general model that "nests" a number of others. The survival function is:

$$S(t) = 1 - \Gamma\left\{\kappa, \kappa \exp\left[\frac{\frac{\ln T_i - \lambda}{\sigma}}{\kappa^{1/2}}\right]\right\} \tag{32}$$

Here, $\Gamma(\cdot)$ is the indefinite gamma integral. The formulae for the hazard and the density are complicated, and largely beside the point. The more important thing is that the model nests the lognormal, Weibull and exponential models:

- The model is log-normal as $\kappa \to \infty$.
- the model is a Weibull when $\kappa = 1$,
- the model is exponential when $\kappa = \sigma = 1$.

This model is thus nice and general, but can be slooooow and difficult to get to converge. For example, using our Supreme Court retirements example data, the model wouldn't converge at all:

```
. streg age pension pagree, dist(gamma)

Fitting constant-only model:

Iteration 0:   log likelihood =  -179.4614  (not concave)
Iteration 1:   log likelihood = -101.39199
Iteration 2:   log likelihood = -92.677695
Iteration 3:   log likelihood = -90.945899
Iteration 4:   log likelihood = -90.415731
Iteration 5:   log likelihood = -90.114788
Iteration 6:   log likelihood = -89.913238
Iteration 7:   log likelihood =  -89.84974
Iteration 8:   log likelihood = -89.820951
Iteration 9:   log likelihood = -89.814303
Iteration 10:  log likelihood = -89.809908
Iteration 11:  log likelihood = -89.809583
Iteration 12:  log likelihood = -89.809157
discontinuous region encountered
cannot compute an improvement
r(430);
```

Estimation and Interpretation (With a Running Example)

For the rest of the morning, we'll focus on the mechanics of the estimation, selection, interpretation, and presentation of parametric models for survival data. To do this, we'll use a running example, based on the analysis and data in:

Brooks, Sarah M. 2005. "Interdependent and Domestic Foundations of Policy Change." *International Studies Quarterly* 49(2):273-294.

Brooks' major focus is the causes of pension privatization. In particular, Brooks is interested in the extent to which nations' decisions to privatize their pension plans are interdependent – that is, whether or not they reflect similar decisions by other nations. Her primary response variable is the onset of pension privatization in 59 countries in the OECD, Latin America, Eastern Europe, and Central Asia between 1980 and 1999.

Given this interest in diffusion, Brooks' key covariate of interest is *peer privatization* – operationalized as the proportion (really the percentage) of countries in each of the three "peer groups" (OECD, Latin America, and Eastern Europe and Central Asia) that have privatized. Brooks' main expectations are that the influence of peer privatization will vary across

the three peer groups, though she has no specific hypotheses about the relative magnitude of those variations. Because Brooks hypothesizes that the effects of this diffusion will be different for each of the three peer groups, she include dichotomous variables for two of the three groups, as well as interactions between those indicators and the *privatization* variable:

$$\begin{aligned} h_i(t) &= f[\beta_0 + \beta_1(\text{Peer Privatization}_{it}) + \beta_2(\text{Latin America}_i) \\ &\quad +\beta_3(\text{E. Europe/Cent. Asia}_i) + \beta_4(\text{Peer Privatization}_{it} \times \text{Latin America}_i) \\ &\quad +\beta_5(\text{Peer Privatization}_{it} \times \text{E. Europe/Cent. Asia}_i) + \mathbf{X}_{it}\gamma] \end{aligned} \tag{33}$$

Brooks also includes in the model a host of other control variables ($\mathbf{X}_{it}$) for such phenomena as international, demographic, and domestic political and economic pressures.[1] Here, we'll include these in the model as well, but focus out interpretation on the key covariates – that is, the effects of peer privatization across the three subgroups of countries.

The `stdes` of the survival variable, as well as summary statistics for the model's covariates, are on page 10 of today's handout. We'll get to more details of this example in a bit.

Estimation

Estimation in Stata

As we've already seen, the basic command for estimating parametric models is `streg`:

```
.streg ...  , dist() nohr / time ...
```

A few things here:

- As I mentioned earlier, one does not include the "dependent variable" among the variables listed after `streg`. Once one `stset`s the data, Stata then treats the survival variable you set as the response, and looks for covariates after the command.
- The available distributions are the ones we talked about yesterday: exponential, Weibull, gompertz, log-normal, log-logistic, and generalized gamma.
- For the three proportional hazards models (exponential, Weibull, and gompertz), the default is to report hazard ratios ($\exp(\hat{\beta})$s) rather than coefficient estimates ($\hat{\beta}$s). If, as I do, you find this annoying, be sure to include the `nohr` option at the end of the command.

[1] In her actual analysis, Brooks estimates a Cox proportional hazards model; here, we'll use parametric models, though we can return to the Cox formulation later. In addition, Brooks include a log-time interaction in her model; we'll worry about that later, when we discuss nonproportionality.

- Likewise, for those models that have both PH and AFT implementations, the default is the PH form. If you prefer the accelerated failure-time approach, the option to include is `time`.

Most of these models converge well and generally are well-behaved, with the notable exception of the generalized gamma model. The results of estimating each of these models for Brooks' data are on pages 2-4 of the handout.

Estimation in R

In general, R is not as good for parametric models as is Stata (this is in contrast to the Cox model). After loading the `survival` package, the basic command is `survreg`:

```
Results<-survreg(Surv(duration, censor)~X1+X2+X3, dist="exponential")
```

Note a couple things:

1. Available distributions are exponential, Weibull, normal, logistic, log-normal, and log-logistic.
2. To the best of my knowledge, R *will not estimate parametric models with time-varying data/covariates*. This is a big drawback, and really means that you're much better off using Stata for those models.

Model Selection

Theory

If possible, use it. It is generally better than all of the alternatives. Points to consider:

- Can you, from first principles, derive a data-generating process for the phenomenon under study that leads more-or-less directly to a **particular parametric distribution**? (If you can, you're very lucky, at least in the social scoences...).
- Are there theoretical expectations about how events ought to arise? Conditional on **X**, is there reason to expect that **events might not be independent** of one another (and so that the exponential might not be a good representation of the edata)?
- Is there some reason to expect the (conditional) **shape of the hazard** to take on a particular form?

Statistical Tests

If you have no theory, or if the theory is vague or otherwise indeterminate, then one can use formal statistical tests to choose among models There are two main sorts of these that we might be interested in here.

1. LR Tests

For nested models, we can use standard likelihood-ratio statistics to test for the superiority of a more general model versus a more parsimonious one. The general formula is:

$$LR = -2(\ln L_R - \ln L_U), \tag{34}$$

where $\ln L_R$ is the log-likelihood for the restricted (less general) model, and $\ln L_U$ is the log-likelihood for the unrestricted model. This test is asymptotically distributed as χ^2 with degrees of freedom equal to the difference in the number of restrictions between the two models.

In the context of parametric survival models, the main use of LR tests is to choose between the exponential and the Weibull models, and among the exponential, Weibull, and generalized gamma models, should we be so fortunate as to be able to estimate the latter. So, for example, the LR test for the Weibull versus the exponential here is equal to:

$$\begin{aligned} LR &= -2(-15.218 + 13.320) \\ &= 3.80 \end{aligned}$$

With one degree of freedom, this test statistics has a p-value of 0.11 or so – indicating that the Weibull is only a marginally better fit than the exponential.

2. AIC & BIC

Another set of tests are those based on estimates of the Kullback-Leibler information of a model. There are two of these in wide use. *Akaike's Information Criterion*, usually abbreviated AIC, is:

$$\text{AIC} = -2(\ln L) + 2(k) \tag{35}$$

where k in this context is the total number of parameters estimated in the model (including a constant term, if any). A similar statistic is the Bayesian Information Criterion (BIC), which is:

$$\text{BIC} = -2(\ln L) + 2(k)\ln(N) \tag{36}$$

Smaller AICs and BICs generally indicate better-fitting models. So, for example, a general rule of thumb for comparing AICs (Burnham and Anderson 2002, 70) is:

$AIC_{Model\ j}$ - $AIC_{Minimum\text{-}AIC\ Model}$	Level of Empirical Support for Model j
0-2	Substantial
4-7	Considerably Less
> 10	Essentially None

The AICs and BICs for each of the five models estimated (generated using Stata's `estat ic` command after estimation) are reported here:

Table 1: ln$L$, AIC, and BIC for Five Parametric Models

Model	ln$L$	AIC	BIC
Exponential	-15.218	64.436	144.14
Weibull	-13.320	62.640	147.03
Gompertz	-14.059	64.118	148.51
Log-Normal	-15.224	66.447	150.84
Log-Logistic	-14.310	64.620	149.01

By a pure AIC-minimization criterion, the Weibull model wins out, with the Gompertz coming in second, edging out the exponential. If we pay attention to the BIC – which penalizes additional parameters more than the AIC does – then the exponential is our model. Of course, none of these can tell you whether your model is any good in absolute terms; at best, they can be useful for choosing among relatively similar models.

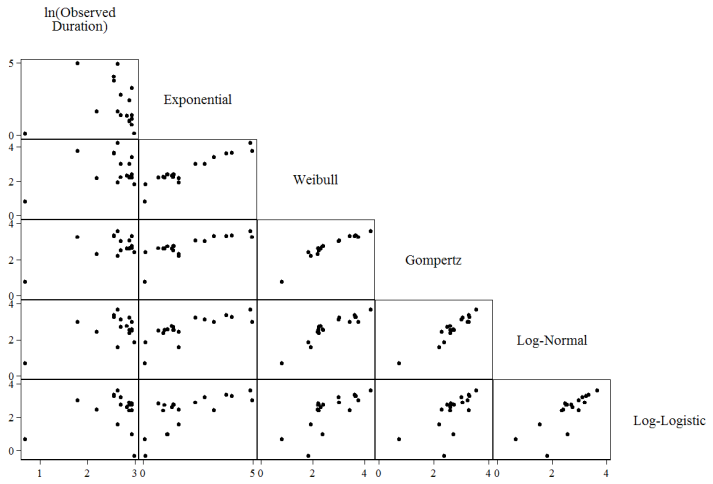
Fitted Values

A more impressionistic approach is to examine the fitted values of each model. These can be any number of things, though the most natural quantity to examine is the predicted duration or log-duration, since that is the easiest to compare to the actual values. Stata (and most other programs) makes generating such values relatively easy:

```
. predict lnThat_E, median lntime
```

When we plot the predicted log-median durations for our five models against their (also logged) actual values for those observations that privatized, we get:

Predicted (Median) vs. Actual $\ln(T_i)$



In addition, we can examine the (Pearson's) correlations between the various fitted values and the actual failure events:

```
. corr ln_duration lnThat_E lnThat_W lnThat_G lnThat_LN lnThat_LL if fail==1
(obs=18)

             | ln_dur~n lnThat_E lnThat_W lnThat_G lnThat~N lnThat~L
-------------+------------------------------------------------------
 ln_duration |   1.0000
    lnThat_E |  -0.0361   1.0000
    lnThat_W |   0.2647   0.9398   1.0000
    lnThat_G |   0.5873   0.7656   0.9294   1.0000
   lnThat_LN |   0.4870   0.7522   0.9067   0.9423   1.0000
   lnThat_LL |   0.1837   0.6758   0.7095   0.6644   0.7868   1.0000
```

In both of these instances:

- None of the fitted values are all that great, but
- Of the various ones, the Weibull, log-normal, and Gompertz are probably the best. Also,
- There are some high correlations among the fitted values, which suggest that none of the models are yielding predictions that are that much at odds with any of the others (a common phenomenon if the model fit isn't especially good).

Interpretation

Stata has a huge number of useful post-estimation commands. I won't go into all of them here – we'll skip a discussion of `test`, `testnl`, `mfx`, etc. – but discuss several other useful complements to `streg`.

Hazard Ratios

As we discussed earlier, for models that have a hazard-rate form, hazard ratios are a nice, intuitive way of talking about the marginal effects of covariates.

- Just $\exp(\hat{\beta}_k)$.
- Stata reports them automatically if you want it to (cf. p. 6 of the handout).
- Discussing them:
    - HRs reflect the relative/proportional change in $h(t)$ associated with a unit change in X_k. In other words, they are invariant to the values of the other covariates, or of the hazard itself.
    - Similarly, remember that 100 $\times$ (HR - 1) is the same thing as the percentage change in the hazard associated with a one-unit change in the covariate in question.
    - Also remember that the hazard ratios reported by Stata are for *one-unit* changes in X_k. That means that if a one-unit change in X_k is either very large or very small, the hazard ratios will be the opposite.
    - That, in turn, suggests that if you're going to use hazard ratios for reporting and discussing your results, it's a good idea to rescale your covariates so that a one-unit change is meaningful.

So, let's discuss some of the hazard ratios for the "full" Weibull model:

- `worldbank` ($\bar{X} = 0.07$, $\sigma = 0.48$, HR = 0.601)
    - A measure of "hard power," in the form of the level of World Bank loans or credits to the country.
    - The hazard ratio of 0.601 means that, among two otherwise-identical observations that differ by one unit on `worldbank`, the one with the higher level of the variable will have a hazard that is 60.1 percent the size of the former.
    - Likewise, each unit increase in `worldbank` decreases the hazard of privatization by 100 $\times$ (0.601 - 1) $\approx$ 40 percent.
    - The confidence intervals, however, are very wide, and encompass 1.0; this means that we cannot say this result is not due to chance.

- `age65` ($\bar{X} = 9.73$, $\sigma = 4.68$, HR = 1.35)
    - Measures the percentage of the population that is over the age of 65.
    - The hazard ratio of 1.35 means that, among two otherwise-identical observations that differ by one unit on `age65`, the one with the higher level of the variable will have a hazard that is 1.35 times as large as the former.
    - Likewise, each unit increase in `age65` increases the hazard of privatization by 100 $\times$ (1.35 - 1) $\approx$ 35 percent.
    - The confidence intervals suggest that this variable is "statistically significant" / "bounded away from zero" in its effect. The 95 percent credible range for the hazard ratio ranges from 1.04 to 1.76.
- `deficit` ($\bar{X} = -3.68$, $\sigma = 4.27$, HR = 1.24)
    - Brooks calls it *Budget Balance*; it's an indicator of the overall level of surplus or deficit in that country in that year, as a percentage of GDP.
    - Similar to `age65`, the hazard ratio of 1.24 means that, among two otherwise-identical observations that differ by one unit on `deficit`, the one with the higher level of the variable will have a hazard that is 1.24 times as large as the former.
    - If we want to know (say) the impact of a five-percent swing in the budget deficit, we can calculate that as $\exp(0.214 \times 5) \approx 2.92$. This means that a five-percent increase in the budget deficit (as a percent of GDP) nearly triples the hazard of pension privatization.
    - Once again, the 95-percent confidence interval (barely) excludes 1.0, meaning that the estimated effect is "statistically significant" at conventional ($p = .05$, two-tailed) levels.

We'll discuss the hazard ratios for the main (interactive) variables of interest in a bit...

Obtaining Linear (and Nonlinear) Combinations of Parameters

Stata offers some nice commands for obtaining linear and nonlinear combinations of estimated parameters, as well as the standard errors, p-values, and confidence intervals associated with those combinations. These are particularly valuable when – as we have here – there are interaction effects in our models.

Combinations of $\hat{\beta}$s: **`lincom`**

We can obtain estimates of linear combinations of parameters (as in, our estimated coefficients) using `lincom`. Suppose we want to know the "quasi-coefficient" of *Peer Privatization* for Latin American countries. In terms of our model in (33), this is equal to:

$$\hat{\beta}_{\textit{Peer Privatization}|\text{Latin America}} = \hat{\beta}_1 + \hat{\beta}_4$$

and the equivalent coefficient estimate for the Eastern European and Central Asian countries is:

$$\hat{\beta}_{\textit{Peer Privatization}|\text{E. Europe / Cent. Asia}} = \hat{\beta}_1 + \hat{\beta}_5.$$

After estimating our model, `lincom` has Stata calculate this quantity for us, along with its associated uncertainty quantities:

```
. lincom(peerprivat+LatAmxPeer)

 ( 1)  [_t]peerprivat + [_t]LatAmxPeer = 0

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .1055036   .0457227     2.31   0.021     .0158887    .1951185
------------------------------------------------------------------------------
```

This tells us that the effect of *Peer Privatization* on the hazard of privatization in Latin American countries is positive and "statistically significant" at conventional levels. We can do the same thing for the Eastern European and Central Asian countries:

```
. lincom(peerprivat+EECAxPeer)

 ( 1)  [_t]peerprivat + [_t]EECAxPeer = 0

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .0531117   .0448715     1.18   0.237    -.0348348    .1410581
------------------------------------------------------------------------------
```

...where we see that the effect of that variable is not significantly different from zero in those countries.

Combinations of Hazard Ratios: ***nlcom***

If combinations of coefficients are good, then combinations of things we actually might interpret (like hazard ratios) would be even better. To get those, however, requires that we combine nonlinear sorts of things.[2] This therefore requires `nlcom`:

```
. nlcom(exp(_b[peerprivat]+_b[LatAmxPeer]))

       _nl_1:  exp(_b[peerprivat]+_b[LatAmxPeer])

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |    1.11127   .0508103    21.87   0.000     1.011684    1.210856
------------------------------------------------------------------------------

. nlcom(exp(_b[peerprivat]+_b[EECAxPeer]))

       _nl_1:  exp(_b[peerprivat]+_b[EECAxPeer])

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   1.054547   .0473191    22.29   0.000     .9618037    1.147291
------------------------------------------------------------------------------
```

Note here that:

- These are regular, garden-variety hazard ratios, and can be interpreted as such (conditional, of course, on the interacted variables).
- The hypothesis tests reported are for the null that $\widehat{\text{HR}} = 0$, which is of course silly. You have to calculate your own z-scores based on the HRs and their standard errors if you want to test (e.g.) $\widehat{\text{HR}} = 1$, the more interesting case. (For these two, the z-scores on the hazard ratios are about 2.19 for the Latin American countries, and 1.16 for the Eastern European / Central Asian ones; these numbers are – unsurprisingly – very close to those for the coefficients, above).

[2] Because, in general, $\exp(A) + \exp(B) \neq \exp(A + B)$.

Predicted Hazard and Survival Functions – `stcurve`

Stata offers a useful additional command for generating predicted survival curves, `stcurve`. This command plots the estimated/predicted hazard, cumulative hazard, and/or survival functions for the model you've just estimated, and will do so for varying levels of covariates. The basic command is:

```
.stcurve, (haz/cumh/sur) at1() at2() ...
```

Important points:

- The initial option `haz / cumh / sur` is not optional, and specifies whether you want to plot the *predicted hazard*, the *predicted cumulative hazard*, or the *predicted survival curve*, respectively.
- The `at1()`, etc. commands are optional, but nonetheless extremely important, in that they specify the values of the covariates at which you want the curves to be plotted. Note that
  - If you don't specify an `at()` value, Stata holds the (unspecified) covariate constant at its mean.
  - You can plot up to 10 curves on a single graph (that is, you can go up to `at10()`).

See the handout for an example of this command in action...

Predicted Hazard and Survival Functions – `predict`

As with all models, Stata's `predict` command is very flexible and useful in survival analysis. Following `streg`, Stata's `predict` command will generate:

- **mean and median predicted survival times**,
- **mean and median predicted log-survival times** (as used above),
- **predicted hazards** ($\hat{h}_{it}$),
- **predicted hazard ratios** (that is, relative hazards),
- **predicted survival probabilities** ($\hat{S}_{it}$) (at the covariate values),
- **"index" values** / the linear predictor (that is, $\mathbf{X}_i\hat{\beta}$),
- the **estimated standard error of the linear predictor**, and
- gobs and gobs of **residuals** of various sorts.

As in all cases, you can use `predict` either in- or out-of-sample. That means you can generate "simulated" data with particular characteristics of interest, and then predict to them. This is, at the moment, the only way I'm aware of to get things like confidence intervals around predictions; moreover, it tends to be a bit clunky, since in order to predict from an `-st-` model, the data must be `-st-` as well. There are some ways to "fool" Stata into doing this, but they're all a bit artificial.

Finally: A Couple Things That Don't Work

1. Clarify (with any of the `-st-` commands).
2. `Zelig` will work, but only with non-time-varying data (since that's all that R will estimate as well).

If any of you nerdy folks care to write the code that will make either of these things more useful, the world would thank you (and I might even buy you a beer).

Cox's Semiparametric Model

The model introduced by Cox (1972) is arguably the single most important tool for survival analysis available today. It is by far the most widely-used model in biostatistics, epidemiology, and other "hard-science" fields, and is increasingly becoming so in the social sciences. The reasons for this popularity will become apparent in a minute.

This afternoon we'll introduce the Cox model, which is also sometimes (slightly inaccurately) known as "the proportional hazards model." We'll talk about estimation, and interpretation of that model, and we'll walk through an example in the Stata and R packages.

Cox's (1972) Semiparametric Hazard Model

Consider a model in which we start with a generic "baseline" hazard function; call that $h_0(t)$. For now, don't worry about its shape; turns out, it's not that important.

We might be interested in modeling the influence of covariates on this function; i.e., in allowing covariates $\mathbf{X}$ to increase or decrease the hazard. Since the hazard must remain positive, the covariates need to enter in a way that will keep them so as well. As we know from the parametric models we discussed, the natural way to do this is the exponential function; this suggests a model of the form:

$$h_i(t) = h_0(t)\exp(\mathbf{X}_i\beta) \tag{37}$$

Note several things about this model:

- The baseline hazard corresponds to the case where $\mathbf{X}_i = 0$,
- It is shifted up or down by an order or proportionality with changes in $\mathbf{X}$,
- That is, as in the exponential and weibull models, the hazards are *proportional* (hence the name).

This model was first suggested by Cox (1972). It is far and away the dominant survival analysis method in biostatistics, epidemiology, and so forth, and is also used in political science, sociology, etc. The model has a couple of useful and attractive characteristics:

- *The hazard ratio for a one-unit change in any given covariate is an easy constant.* Because it is a model of proportional hazards, then for a dichotomous covariate X the hazard ratio is

$$\begin{aligned} \text{HR} &= \frac{h_0(t)\exp(X_1\hat{\beta})}{h_0(t)\exp(X_0\hat{\beta})} \\ &= \exp[(1-0)\hat{\beta}] \\ &= \exp(\hat{\beta}) \end{aligned}$$

That is, $\exp(\hat{\beta})$ is the relative odds of an observation with $X = 1$ experiencing the event of interest, relative to those observations for which $X = 0$. As we saw with the exponential, Weibull, and Gompertz models, this is an easy, intuitive way of understanding the influence of covariates on the hazard.

- *The effect of covariates on the survival function is also straightforward.* Recall that

$$S(t) = \exp[-H(t)]$$

In the case of the Cox model, that means that

$$\begin{aligned} S(t) &= \exp\left[-\int_0^t h(t)\,dt\right] \\ &= \exp\left[-\exp(\mathbf{X}_i\beta)\int_0^t h_0(t)\,dt\right] \\ &= \left[\exp\left(-\int_0^t h_0(t)\,dt\right)\right]^{\exp(\mathbf{X}_i\beta)} \\ &= [S_0(t)]^{\exp(\mathbf{X}_i\beta)} \end{aligned}$$

In other words, the influence of variables is to shift the "baseline survivor function $S_0(t)$ by a factor of proportionality. (Remember that the baseline survivor function is always between 0 and 1; thus, the effect of a variable with a positive coefficient is to decrease the survival function relative to the baseline, while negative variable effects have the opposite effect on the survival curve.)

Partial Likelihood: Some Computational Details

Notice that throughout this, we didn't even have to specify a distribution. The Cox model completely avoids making potentially (probably) untenable distributional assumptions about the hazard. In fact, this is arguably its greatest strength.

How?

The basis of the Cox model is an argument based on conditional probability. Suppose that

- For each of the N uncensored observations in the data, we know both T_i (the time of the event) and C_i (the censoring indicator),
- observations are independent of one another, and
- there are no "tied" event times.

Then for a given set of data there are N_C distinct event times; call these t_j.

Now suppose we know that an event happened at a particular time t_j. One way we might go about estimation is to ask: Given that *some observation* experienced the event of interest at that time, what is the probability that it was observation k (with covariates $\mathbf{X}_k$)? That is, we want to know:

$$\Pr(\text{Individual } k \text{ experienced the event at } t_j \,|\, \text{One observation experienced the event at } t_j)$$

Since $\Pr(\text{A} \,|\, \text{B}) = \frac{\Pr(\text{A and B})}{\Pr(\text{B})}$, we can write this conditional probability as:

$$\frac{\Pr(\text{At-risk observation } k \text{ experiences the event of interest at } t_j)}{\Pr(\text{One at-risk observation experiences the event of interest at } t_j)}$$

Here,

- The numerator is just the hazard for individual k at time t_j, and
- The denominator is the sum of all the hazards at t_j for all the individuals at risk at t_j.

We can write this as:

$$\frac{h_k(t_j)}{\sum_{\ell \in R_j} h_\ell(t_j)}$$

where R_j denotes the set of all observations "at risk" for the event at time t_j. If we substitute the hazard function for the Cox model (37) into this equation, we get:

$$\begin{aligned} L_k &= \frac{h_0(t_j)\exp(\mathbf{X}_k\beta)}{\sum_{\ell \in R_j} h_0(t_j)\exp(\mathbf{X}_\ell\beta)} \\ &= \frac{h_0(t_j)\exp(\mathbf{X}_k\beta)}{h_0(t_j)\sum_{\ell \in R_j} \exp(\mathbf{X}_\ell\beta)} \\ &= \frac{\exp(\mathbf{X}_k\beta)}{\sum_{\ell \in R_j} \exp(\mathbf{X}_\ell\beta)} \end{aligned} \tag{38}$$

where the last equality holds because the because the baseline hazards "cancel out."

Each observed event thus contributes one term like this to the partial likelihood; the overall (joint) partial likelihood is thus:

$$L = \prod_{i=1}^{N} \left[\frac{\exp(\mathbf{X}_i \beta)}{\sum_{\ell \in R_j} \exp(\mathbf{X}_\ell \beta)} \right]^{C_i} \tag{39}$$

where the i denote the N distinct event times, and $\mathbf{X}_i$ denotes the covariate vector for the observation that actually experienced the event of interest at t_j. The log-partial-likelihood is then equal to:

$$\ln L = \sum_{i=1}^{N} C_i \left\{ \mathbf{X}_i \beta - \ln \left[\sum_{\ell \in R_j} \exp(\mathbf{X}_\ell \beta) \right] \right\} \tag{40}$$

Note a few things about the partial likelihood approach:

- *It takes account of the ordering of events, but not their actual duration.*
    - That is, it is tantamount to saying that the gaps between events tell us nothing about the hazards of events.
    - This is because, in theory, since $h_0(t)$ is unspecified, it could very well be zero during those times.
- This is also true of *censored observations*, which are simply either in our out of the risk set, with no attention payed to their precise time of exit.
- *The model doesn't allow for tied events.*
    - The Cox model assumes that time is continuous, so tied events can't really happen.
    - If we have tied events, it must be because our time-measuring device isn't sensitive enough (or so the logic goes).

Under the usual regularity conditions, Cox (1972, 1975) showed that estimates of $\hat{\beta}$ obtained by maximizing the partial likelihood in (40) are consistent, asymptotically normal, and asymptotically efficient, though not fully efficient compared to fully-parametric estimates. The big advantage, however, is that nothing about Cox's model requires making problematic parametric assumptions; in many cases, this fact more than makes up for the models (relatively slight) efficiency loss. We'll discuss the sometimes-thorny issue of choosing between parametric and Cox models a bit later.

Estimation

The Cox partial likelihood is relatively well-behaved under most conditions, and in general getting the partial-likelihood to converge is not a problem. Standard error estimates are obtained in the usual way, as the negative inverse of the Hessian evaluated at the maximized partial likelihood. One can also use "robust" / "sandwich" variance-covariance estimators in the Cox model; indeed, those are especially important when it comes to dealing with repeated/multiple event occurrences. There's an illustration of all this below.

Interpretation: An Example Using Interstate War, 1950-1985

To illustrate the Cox model, we'll look at some widely-used data on the occurrence of interstate war between 1950 and 1985. The data are dyad-years for "politically-relevant" dyads ($N = 827$); the total number of observations is 20,448. There are six "standard" covariates that have been used in a number of models of war onset (cf. Oneal and Russett 1997, 1999; Beck et al. 1998, etc.):

- Whether (=1) or not (=0) the two countries in the dyad are *allies*,
- Whether (=1) or not (=0) the two countries in the dyad are *contiguous*,
- The *capability ratio* of the two countries (that is, the capabilities of one over that of the other),
- The lower of the two countries economic (GDP) *growth* (rescaled),
- The lower of the two countries' *democracy* (POLITY IV) scores (rescaled to [-1,1]), and
- The amount of *trade* between the two countries, as a fraction of joint GDP.

Hazard Ratios

As suggested above, one key to interpretation is the *hazard ratio* (sometimes also known as the odds ratio, above), which is a very natural way of interpreting covariate effects. In the general case, the hazard ratio for two observations with covariate values X_j and X_k equals:

$$\exp[(X_j - X_k)\hat{\beta}] \tag{41}$$

The standard interpretation is as the ratio of the hazards of the two observations in question:

- A hazard ratio of one corresponds to $\hat{\beta} = 0$ – i.e., no effect for X.
- A hazard ratio greater than one implies a positive coefficient estimate, and
- A hazard ratio less than one implies a negative coefficient.

Moreover, the percentage difference in the hazards for the two observations is just:

$$100 \times \{\exp[(X_j - X_k)\hat{\beta}] - 1\}.$$

So:

- for a binary covariate with $\hat{\beta} = 1.1$, we would say that "the hazard of the event of interest is roughly $100 \times [\exp(1.1) - 1] = 200$ percent greater for observations with $X = 1$ than for those with $X = 0$." Similarly,
- For a continuous covariate Z with $\hat{\beta} = -0.03$, we could say that "a 20-unit increase in Z yields a decrease of $100 \times [\exp(-0.03 \times 20) - 1] = 45.1$ percent in the hazard of the event of interest."

Baseline Hazards and Survival Curves

As we said at the outset, the baseline hazard is undefined in the Cox model. But, that doesn't mean that it's not useful just the same. In fact, estimated hazards can also be useful, even though they're technically "undefined." Remember a few things:

- Baseline hazards for the Cox model are only defined at time points where events occurred. This means that if you plot them "'raw," they tend to be very jagged–looking (unlike the exponential, Weibull, and other parametric models).
- The default option is thus to "smooth" them, usually through some sort of running-mean or lowess smoother, to get a better sense of their overall shape.

The same is true for the baseline survival function; you can also generate and plot the estimated survivor functions. These can be especially useful if plotted for different values of discrete covariates. In Stata, you can use the `basehazard`, `basechazard`, and `basesurv` options to define variables equal to the baseline hazard and survival functions after you estimate the model; our friend `stcurve` is also available for the Cox model. More on this is in the handout.

Handling Ties

In general, the effect of "ties" in the Cox model is generally to bias the coefficient estimates toward zero; the extent to which this occurs is a more-or-less monotonic function of the number of "tied" observations in the data. While there are no clear guidelines, a rule of thumb is that if > 10 percent of your observations are "tied" (that is, have events occurring at the same time), ties might be an issue.

There are several solutions for using the Cox model if you have tied data. First, some notation:

- Call $d_j > 0$ the number of events occurring at t_j, and
- D_j the set of d_j observations that have the event at t_j.

Now, think about how those events happened.

- One possibility is that they all happened sequentially and independently (that is, discretely), but that we grouped them together.
- This means that we need to modify the numerator of (39) to include the covariates from all the observations that had the event at t_j, and
- To modify the denominator to account for the multiple possible orderings of those events.

A simple way to do this is like:

$$L_{\text{Breslow}}(\beta) = \prod_{i=1}^{N} \frac{\exp\left[\left(\sum_{q \in D_j} \mathbf{X}_q\right)\beta\right]}{\left[\sum_{\ell \in R_j} \exp(\mathbf{X}_\ell \beta)\right]^{d_j}} \tag{42}$$

This is known as Breslow's (1974) approximation, and is the standard/default way of dealing with tied event times in most statistical packages (e.g., Stata; SAS, etc., though not in the `survival` package in R). A modification of this was also suggested by Efron (1977);[3] it is generally accepted that Efron's is a better approximation than Breslow's, which is probably why it is the default way of dealing with ties in the `survival` package in R.

Finally, we can use "exact" methods, which take into account that, if there are d events at t, then there are $d!$ possible orderings of those events. This approach yields the *exact partial likelihood* and *exact marginal likelihood* methods; they are the most complex, but also the most "accurate," mathematically speaking.

Some Practical Advice About Ties in the Cox Model

In every case, the approximations reduce to the "standard" Cox partial likelihood when there are no ties (i.e., when $d_j = 1 \,\forall\, j$).

- This means that, if there are few ties, all of them will give very similar results.
- So, in that case, there is nothing to worry about.

If you have a lot of ties, things can get a bit dicier. A few rules:

- The Breslow approximation is the "worst."
- The Efron one is better.
- The exact methods are very computationally demanding, but give the "best" results. However, they can take forever to converge, and (especially in R) may lead to your machine freezing up. Be warned.

Generally, using Efron is OK; but use the exact method if you have lots of ties.

[3] I'm omitting the equation for the Efron approximation here; it's Eq. (3.35) in Hosmer and Lemeshow (1999) if you really care that much.

Software Matters

The Cox Model in Stata

Unsurprisingly, Stata will estimate Cox semiparametric hazard models. The basic command is `stcox`, which is similar to `streg` in a number of ways:

- Before using `stcox`, the data must be `stset`,
- As with `streg`, `stcox` does not take a "dependent variable" immediately after the command,
- A number of post-estimation commands, including `stcurve`, are allowed after `stcox`.

As we'll see a bit later, there are also a large variety of diagnostic commands available for the Cox model in Stata – we'll talk about those in a bit.

The Cox Model in R

The basic routine in the `survival` package is `coxph`, or, alternatively, `cph` in the `design` package. In contrast to the parametric models, R is fantastic for estimating, diagnosing, and interpreting Cox models – arguably better than Stata is.

Basic Commands

- `coxph` in `survival`
- `cph` in `design`
- Both work similarly...

`cph` and `coxph` output a *proportional hazards regression object*, which has a number of useful components.

Graphing Results

... can be done using `plot(survfit())` on the resulting PH object. See the handout for an example.

Discrete-Time Methods

Models for discrete duration data were certainly once – and, in many fields, remain – the most commonly-used approach for analyzing survival data. Discrete-time methods start with an intuitive presence: we are interested in an event, and we can collect data – both across units and over time – on if an when that event happened and relate it to covariates. The basic intuition, then, is a more-or-less standard GLM-type model, of the form:

$$\Pr(Y_{it} = 1) = f(\mathbf{X}_{it}\beta + u_{it}) \tag{43}$$

Something like this is the basis for nearly all *discrete-time models*, which is what we'll talk about today.

Concepts

A discrete duration process is one for which the hazard function $h(t)$ is defined only at some finite or countably infinite set of points in time. Some durations are just *discrete*. Box-Steffensmeier and Jones discuss Congressional careers in this respect. Alternatively, consider a study that focused on confirmation in the Catholic Church:

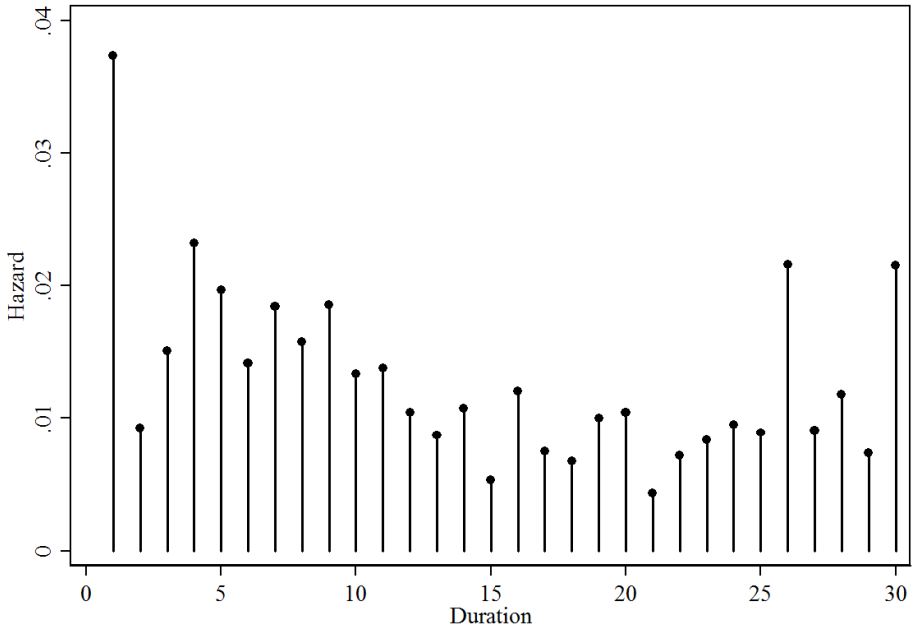
- Confirmation typically occurs between ages 13-18.
- But it takes a year to get through it. So,
- ...it is essentially discrete: occurs at either age 13 or age 14 or age 15 or ...
- We might want to model this as a function of church-related factors, etc.

One could think of these as being ranked, ordinal categories, in the sense that if you don't do it at 13, you might at 14, etc. But there are a discrete (and, here, relatively small) number of possible "times."

In similar fashion, there are many problems where, at some point, finer-grained time data just don't make sense; that is, the durations are *effectively discrete*. House members' position announcements on the Clinton impeachment (cf. Caldeira and Zorn 2004) is a good example: There was essentially one week in which members announced, but at what level ought we measure the data? Days? Hours? Minutes? Seconds? At some point, the additional information simply isn't telling us anything valuable.

Other times, we have effectively continuous data that are *grouped*, and in the process are "discretized." Of course, all duration data are grouped at some level; the question is whether the grouping is sufficiently imprecise to call into question whether we can still think of the response variable as continuous. In those instances, there may be some advantages to using discrete-time methods – for example, if the discretization is the result of imprecision in our measuring instrument, such that assuming a continuous-time process implies that we have much more information about the process than we really do.

Figure 6: A Discrete Hazard Function



A General Discrete-Time Model

Let's start with a model where we index the set of possible survival times:

$$t \in \{1, 2, ...t_{\max}\}$$

We'll use similar notation to the parametric case. So, as before, the density – that is, the unconditional probability of having the event of interest – is:

$$f(t) = \Pr(T = t) \tag{44}$$

With an integrated hazard equal to:

$$\begin{aligned} F(t) &= \Pr(T \leq t) \\ &= \sum_{j=1}^{t} f(t_j) \end{aligned} \tag{45}$$

Since the durations are discrete, we can simple sum the hazards across time points prior to t to get this quantity. Equation (45) implies that the survival function is:

$$\begin{aligned} S(t) &\equiv \Pr(T \geq t) \\ &= 1 - F(t) \\ &= \sum_{j=t}^{t_{\max}} f(t_j) \end{aligned} \tag{46}$$

This yields a hazard (i.e., the conditional probability of experiencing the event of interest) equal to:

$$\begin{aligned} h(t) &\equiv \Pr(T = t | T \geq t) \\ &= \frac{f(t)}{S(t)} \end{aligned} \tag{47}$$

Now, consider an observation that has survived to time t. Because, at time t, you either experience the event or you don't, the conditional probability of *survival* at t is:

$$\Pr(T > t | T \geq t) = 1 - h(t) \tag{48}$$

which means that we can rewrite the conditional probability of survival as the product of the previous probabilities of surviving up to that point:

$$\begin{aligned} S(t) &= \Pr(T > t | T \geq t) \times \Pr(T > t-1 | T \geq t-1) \times \Pr(T > t-2 | T \geq t-2) \times ... \\ &\quad \times \Pr(T > 1 | T \geq 2) \times \Pr(T > 1 | T \geq 1) \\ &= [1 - h(t)] \times [1 - h(t-1)] \times [1 - h(t-2)] \times ... \times [1 - h(2)] \times [1 - h(1)] \\ &= \prod_{j=0}^{t} [1 - h(t-j)] \end{aligned} \tag{49}$$

Combining this with (47) we can rewrite the density $f(t)$ in terms of these conditional prior probabilities:

$$\begin{aligned} f(t) &= h(t)S(t) \\ &= h(t) \times [1 - h(t-1)] \times [1 - h(t-2)] \times ... \times [1 - h(2)] \times [1 - h(1)] \\ &= h(t) \prod_{j=1}^{t-1} [1 - h(t-j)] \end{aligned} \tag{50}$$

An observational unit that experiences the event of interest in period t has $Y_{it} = 1$ in that period, and $t - 1$ periods of $Y_{it} = 0$ prior to that. Each of these "observations" thus tells us

something about the observation's (conditional) hazard and survival functions, in a manner analogous to that for the continuous-time case:

$$L = \prod_{i=1}^{N} \left\{ h(t) \prod_{j=1}^{t-1} [1 - h(t-j)] \right\}^{Y_{it}} \left\{ \prod_{j=0}^{t} [1 - h(t-j)] \right\}^{1-Y_{it}} \tag{51}$$

This is a general likelihood, in that (at least at this point) it isn't distribution-specific. The model we get, of course, depends on the distribution we chose for the density $f(t)$, as it did in the parametric case.

Approaches to Discrete-Time Data

There are a number of ways for analyzing discrete-time duration data in a GLM-like framework.

Ordered-Categorical Models

Suppose we have discrete time data on T_i – that is, the actual survival times in our data – and that the number of those event times K is small. Indexing event times as $k \in \{1, 2, ...K\}$, we can model the resulting duration as an ordered probit or logit, e.g.:

$$\Pr(T_i \leq k) = \frac{\exp(\tau_k - \mathbf{X}_i\beta)}{1 + \exp(\tau_k - \mathbf{X}_i\beta)} \tag{52}$$

which is sometimes written as:

$$\ln \left[\frac{\Pr(T_i \leq \kappa)}{\Pr(T_i > \kappa)} \right] = \tau_\kappa - \mathbf{X}_i\beta \tag{53}$$

Note:

- This approach was (first?) suggested by Han and Hausman (1990).
- It is an approach that has some nice properties:
  - It maintains the intrinsic ordering/sequence of the events.
  - It is relatively easy/straightforward to do and interpret.
  - It allows for the "spacing" of the events to differ, and be estimated, through the $\hat{\tau}$s.
  - Finally, it will probably give better estimates than the Cox if there are few event times (and therefore lots of ties).
- However, it also has all the potential pitfalls of ordered logit:
  - Covariates have the same effect across events (that is, the model maintains the *proportional odds* / "parallel regressions" assumption, though this can be relaxed).

- This model also requires a parametric assumption (though this effect is often pretty minimal).
- Finally, it is difficult (read: impossible) to handle time-varying covariates in such a setup.

The point of all this is that you can model a duration using an ordered logit/probit, though it isn't done much – I know of *one* published example in the social sciences...

Grouped-Data ("BTSCS") Approaches

The more common means of analyzing duration data in a discrete-time framework is by directly modeling the binary event indicator. Beck et al. (1998) (hereinafter BKT) refer to this as "binary time-series cross-sectional" ("BTSCS") data. The idea is to start with a model of the probability of an event occurring:

$$\Pr(Y_{it} = 1) = f\left(\mathbf{X}_{it}\beta\right)$$

and then apply standard GLM techniques to these data. The most commonly used "link" functions here are the ones we're all familiar with:

- The logit:

$$\Pr(Y_{it} = 1) = \frac{\exp(\mathbf{X}_{it}\beta)}{1 + \exp(\mathbf{X}_{it}\beta)} \tag{54}$$

- The probit:

$$\Pr(Y_{it} = 1) = \Phi(\mathbf{X}_{it}\beta) \tag{55}$$

- The complimentary log-log:

$$\Pr(Y_{it} = 1) = \exp[-\exp(\mathbf{X}_{it}\beta)] \tag{56}$$

There has been a lot of applied work done using this sort of approach, in political science (Berry and Berry, Mintrom, Squire, BKT, etc.) and in sociology (Allison, Grattet et al., etc.). In large part, this is because it is easy to do; as a practical matter, one simply:

- Organizes the data by unit/time point,
- Models the binary event (1) / no event (0) variable as a function of the covariates $\mathbf{X}$ using logit/probit/whatever, and
- Deals with issues like duration dependence on the right-hand side of the model.

Advantages

- Models like this are easily estimated, interpreted and understood by readers.
- There are natural interpretations for estimates:
  - The constant term is akin to the "baseline hazard," and
  - Covariates shift this up or down.
- The model can incorporate data in time-varying covariates, which is also a big plus.
- There are lots of software packages that will provide parameter estimates.

(Potential) Disadvantages

- A mode like this *requires* time-varying data in order to capture the duration element (one may not always have time-varying data in practice). One can, of course, always make one's data time-varying by "expanding" it, but if there is no additional information in the time-varying data (that is, if all the **X**s are constant over time), then one runs the risk of underestimating standard errors and confidence intervals.
- This approach also requires that the analyst takes explicit consideration of the influence of time, a topic to which we now turn.

Temporal Issues in Grouped-Data Models

It is important to remember that simply estimating a (say) logit on BTSCS data takes no account of the possibility of time dependence in the event of interest. Consider (for example) a monthly survey of cohabitating couples, to see when (if) they get married.

- A simple logit treats all the observations on a single couple as the same (that is, as *exchangeable*).
- More specifically, it doesn't take the *ordering* of them into account.
- Implicitly, this is like saying that (all else equal) the hazard of them getting married is the same in the first month of their living together as it is in the tenth, or the thirtieth.

Statistically, this is equivalent to something akin to an exponential model (independent event arrivals = memoryless process). We can see this if we consider the "baseline" hazard for such a model:

$$h_0(t) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

which is a constant term vis-'a-vis time. As we said when we discussed parametric models, this assumption is, in most circumstances, an insane one to make.

If we want to get away from this in a grouped-data context, we need to explicitly introduce variables into the covariates to capture changes over time.

- The simplest is a *time counter*; that is, a variable that counts the "duration" the observation has been in the data:

$$\Pr(Y_{it} = 1) = f(\mathbf{X}_{it}\beta + \gamma T_{it}) \tag{57}$$

  - Including this allows the "baseline hazard" (here, the constant term) to change monotonically over time:
    - $\hat{\gamma} > 0$ indicates that the hazard is rising over time,
    - $\hat{\gamma} < 0$ suggests the opposite, and
    - $\hat{\gamma} = 0$ corresponds to a "flat" (exponential) hazard.
  - In one sense, this approach is like a Weibull model.

- One can also include quadratics, cubics, etc. of the time counter if they make substantive sense:

$$\Pr(Y_{it} = 1) = f(\mathbf{X}_{it}\beta + \gamma_1 T_{it} + \gamma_2 T_{it}^2 + \gamma_3 T_{it}^3 + ...) \tag{58}$$

  This approach allows one to adopt a "test-down" strategy, estimating multiple degrees of polynomials and then using LR or Wald-type tests to see which offers the best fit (or AIC/BIC criteria to trade off fit for parsimony). See the handout (pp. 29-31) for an example of this approach.

- Another option is *time dummies*:

$$\Pr(Y_{it} = 1) = f[\mathbf{X}_{it}\beta + \alpha_1 I(T_{i1}) + \alpha_2 I(T_{i2}) + ... + \alpha_{t_{\max}} I(T_{it_{\max}})] \tag{59}$$

  - That is, include an indicator variable for each time point at which at least one event occurs.
  - This is more general than a time counter, in that it allows the "baseline" hazard to vary unrestrictedly at each duration point in the data.
  - In this sense, this approach is a lot like the Cox model (more on this in a bit...).
  - This also allows us to do testing on the joint significance of the dummy variables, to see whether the "hazard" remains constant over time.

This last approach by far the most flexible, but also means (potentially) lots of dummy variables eating up degrees of freedom. This latter concern is why BKT use *cubic splines* (smooth functions of the time-dummy variables), the adoption of which has subsequently become a shibboleth among people doing this kind of analysis.

In point of fact, there are a gazillion (that's a technical term) ways to get around the loss of degrees of freedom associated with the dummy-variables approach, cubic splines being only one of them. Others include:

- Using higher-order polynomials of survival time, or functions of fractional polynomials,
- Using something like a kernel smoother on the duration
- Including loess or lowess fits of the binary response variable to the survival duration,
- Using other kinds of splines to smooth the survival time variable (B-splines, P-splines, natural splines, etc.).

The handout shows how (e.g.) a lowess fit and the linear predictor from a cubic spline are effectively identical when it comes to capturing the shape of the underlying hazard in the militarized interstate dispute (MID) data. Moreover, the models' fits are all almost exactly the same.

Discrete-Time Model Selection

Choosing among the possible range of models offered here can be tough. Once again, consider (in order of importance):

1. Theory,
2. formal tests, and
3. fitted values.

The ln$L$s, AICs, and BICs for the various MID models are:

Table 2: AIC and BIC Values, Discrete-Time (Logit) Models

Model	ln$L$	d.f.	AIC	BIC
No Dependence / "Flat"	-1846.88	7	3707.8	3763.2
Linear	-1831.75	8	3679.5	3742.9
Polynomial	-1821.76	11	3665.5	3752.7
"Duration Dummies"	-1801.31	40	3682.6	3999.7
Lowess	-1822.75	8	3661.5	3724.9
Cubic Spline (linear predictor)	-1822.20	8	3660.4	3723.8
Cubic Splines	-1822.02	10	3664.0	3743.3

If we're sticking simply to tests for model selection, then both the AIC and the BIC seem to show a preference for the models incorporating the lowess and linear-predictor cubic spline functions of time. On pure log-likelihood terms, however, the dummy-variables model wins hands down.

Finally, note that in the discrete-time approach the fact that time dependence is an explicit covariate means that one can also interact other variables with survival time, to see if (for example) the effect of a covariate on the probability of the event increases or decreases over time. We'll talk more about this tomorrow.

Some Interesting/Useful Equivalencies

The Cox Model and Conditional Logit

In their book, Box-Steffensmeier and Jones discuss the "equivalency" between a (particular form of the) Cox model and a conditional logit model on the Y_{it} event indicator. The intuition behind this is that the conditional logit model can be thought of as a model of choice by individual i from a choice set $j \in \{1, 2, ...J\}$, of the form:

$$\Pr(Y_i = j) = \frac{\exp(\mathbf{X}_{ij}\beta + \mathbf{Z}_j\gamma)}{\sum_{\ell=1}^{J} \exp(\mathbf{X}_{i\ell}\beta + \mathbf{Z}_\ell\gamma)} \tag{60}$$

in which the $\mathbf{Z}_j\gamma$ – that is, the choice-specific covariates – are "conditioned" out of the model, in a manner akin to "fixed effects," yielding:

$$\Pr(Y_i = j) = \frac{\exp(\mathbf{X}_{ij}\beta)}{\sum_{\ell=1}^{J} \exp(\mathbf{X}_{i\ell}\beta)} \tag{61}$$

These are sometimes also referred to as logit for "matched case-control data," where the matching is on observations in a choice set.

In the context of duration data, rather than considering the j as being choices from among a set of options, we can instead think of them as the observations in the "risk set," and

thus having the potential of being "chosen" by the event process to experience the event of interest. Remember that, in the Cox model, we had:

$$L_k = \frac{\exp(\mathbf{X}_k\beta)}{\sum_{\ell \in R_j} \exp(\mathbf{X}_\ell\beta)}. \tag{62}$$

In this light, it becomes apparent that the Cox model is equivalent to a particular form of the conditional logit model, one in which the denominator reflects the various combinations of potential event orderings that might have happened at a given (discrete) survival time. In that sense, a slightly modified version of the conditional logit estimator is one form of an "exact" semiparametric partial-likelihood model for discrete survival data.

That is all well and good, but – that said – it is not entirely clear to me why we should care. In fact, the only application I've seen of this is in Jan & Brad's book. That's fine, but I'm not sure that this is something that is going to take the world by storm.

Cox-Poisson Equivalence

The discussion above about the use of time-dummies leads us to another interesting (and very important) fact:

Grouped-data duration models and the continuous-time Cox models are equivalent.

That is, it is possible to:

- Use a grouped-duration approach to estimate the parameters of a Cox model, and
- Generate the same baseline hazards, survival functions, and so forth from both models.

Why? How?

The answer lies at the intersection of the Cox model's form for the survival function and the "counting process" formulation for such models that we talked about yesterday. Recall that earlier, we noted (in passing) that, for the Cox model:

$$S_i(t) = \exp\left[-\exp(\mathbf{X}_i\beta)\int_0^t h_0(t)\,dt\right]$$

In words, this means that the survival function for the Cox model is a complimentary log-log function of the covariates $\mathbf{X}$ and the (cumulative) baseline hazard – the latter of which is undefined, according to the Cox model.

Now, recall that the probability distribution function for the Poisson is:

$$\Pr(Y = y) = \frac{\exp(-\lambda)\lambda^y}{y!}$$

where we normally introduce covariates as $\lambda_i = \exp(\mathbf{X}_i\beta)$.

Think about the counting process idea we talked about before. In that case, the variable of interest (here, the event indicator Y_{it}) is either equal to zero or one. For the Poisson, the probability of a zero is:

$$\begin{aligned} \Pr(Y_{it} = 0) &= \exp(-\lambda) \\ &= \exp[-\exp(\mathbf{X}_i\beta)] \end{aligned}$$

In this context, $Y_{it} = 0$ denotes the unconditional survival probability (that is, the absence of the event of interest), which in turn means that (at t) $1 - \exp[-\exp(\mathbf{X}_i\beta)]$ is the event probability. Thus, the key difference between the kernel of the Cox partial likelihood and the Poisson's is just the "baseline" hazard; so long as that baseline hazard does not have a specific form imposed on it, the two are equivalent.

In the Poisson context, not imposing a form on the "baseline" hazard means that we have to allow it to vary unrestrictedly at every event time in the data – in other words, we need to adopt a "dummy-variable" approach to dealing with time dependence. If and when we do that, the Poisson model is exactly equivalent to the Cox model (albeit the Breslow-for-ties version). The handout has an illustration of this equivalency on p. 35.

In practice, this means that *estimating a Poisson model on the binary event indicator, and including a series of dummy variables for each time point in the model, will yield precisely the same estimates as the Cox model.* At one time, we cared about this because it allowed us to use GLM software to estimate a Cox model. The more important thing about this fact now is that it means we can use some of the advances that have been developed for event-count models in a survival-analysis context, something we will come back to tomorrow.

Appendix: Using -stcurve-

Figure 6:

```
.  stcurve, surv
```

Figure 7:

```
. stcurve, hazard at1(LatinAm=0 EEurCentAs=0 LatAmxPeer=0 EECAxPeer=0)
at2(peerprivat=11.492 LatinAm=1 LatAmxPeer=11.492 EEurCentAs=0 EECAxPeer=0)
at3(peerprivat=6.351 EEurCentAs=1 EECAxPeer=6.351 LatinAm=0 LatAmxPeer=0)
ytitle(Predicted Survival Probability) xtitle(Time in Years) caption(, size(zero))
legend(order(1 "OECD" 2 "Latin America" 3 "Eastern Europe/Central Asia"))
lp(solid dash dot) lc(black black black) title(, color(white) size(zero))
```

Figure 8 (same as Figure 7, but with survival):

```
. stcurve, survival at1(LatinAm=0 EEurCentAs=0 LatAmxPeer=0 EECAxPeer=0)
at2(peerprivat=11.492 LatinAm=1 LatAmxPeer=11.492 EEurCentAs=0 EECAxPeer=0)
at3(peerprivat=6.351 EEurCentAs=1 EECAxPeer=6.351 LatinAm=0 LatAmxPeer=0)
ytitle(Predicted Survival Probability) xtitle(Time in Years) caption(, size(zero))
legend(order(1 "OECD" 2 "Latin America" 3 "Eastern Europe/Central Asia"))
lp(solid dash dot) lc(black black black) title(, color(white) size(zero))
```

Figure 9:

```
. stcurve, survival at1(peerprivat=0 LatinAm=1 EEurCentAs=0 LatAmxPeer=0
EECAxPeer=0) at2(peerprivat=10.297 LatinAm=1 LatAmxPeer=10.297
EEurCentAs=0 EECAxPeer=0) at3(peerprivat=31.3  LatinAm=1
LatAmxPeer=31.3 EEurCentAs=0 EECAxPeer=0) ytitle(Predicted Survival
Probability) xtitle(Time in Years) caption(, size(zero)) legend(order(1 "Zero
Peer Privatization" 2 "Mean Peer Privatization" 3 "+2 s.d. Peer Privatization"))
lp(solid dash dot) lc(black black black) title(, color(white) size(zero))
```

Figure 10:

```
. stcurve, hazard recast(line) lcolor(black) lpattern(solid) lwidth(medthick)
addplot(scatter basich duration, msymbol(smcircle) mcolor(black)
msize(small)) ytitle(Predicted Hazard) yscale(range(.004 .025))
ylabel(.004(.004).024) xtitle(Time) title(, size(zero) color(white))
```

Figure 11:

```
. stcurve, survival at1(allies=0) at2(allies=1) recast(line)
lcolor(black black) lwidth(medthick medthick) lp(dash solid)
ytitle(Estimated Baseline Survival Function) xtitle(Years)
title(, size(zero) color(white)) legend(order(1 "Non-Allies" 2 "Allies"))
```