An Introduction to Event History Analysis

Oxford Spring School
June 18-20, 2007

Day Three: Diagnostics, Extensions, and Other Miscellanea

(Non-)Proportional Hazards

As we've said from the outset, the exponential, Weibull, and Cox models are all proportional hazards (PH) models. That is,

- They assume that the effect of a covariate is to shift the hazard proportionally to the baseline.
- So, for two individuals A and B, their relative hazards will be:

$$h_A(t) = Ch_B(t)$$

  where C is the hazard ratio between A and B.

Note that the relationship in (1) is assumed to be true for all time points, and irrespective of the shape of the baseline hazard; this will become very important in just a bit.

What do proportional hazards look like?

- When hazards are "flat" (e.g., in an exponential model), proportional hazards correspond to parallel horizontal lines.
- Increasing hazards are necessarily diverging from one another, while
- Decreasing hazards are converging.

For lots of reasons, it may be the case that hazards aren't strictly proportional. Consider a few examples:

- In medical studies: *Resistance* to the therapy/drug yields hazards which are converging for treatment vs. control groups over time.
- We can see similar effects with *learning* phenomena: hazards between groups may converge.
- Conversely, the effects of a variable may grow more pronounced over time, leading to hazards that diverge as time passes.

- In some cases, hazards for two groups may actually cross. One example of this is common in oncology: The decision between surgery and radiation treatments.
    - Surgery has higher initial risk, but
    - Also a better long-term prognosis.
    - The result is that hazards for surgery are initially higher, but stay flat an even decrease over time, while those for radiation start out lower, but grow as time passes.

Thinking About Nonproportionality

PH are thus a set of assumptions about *parameter stability over time*; specifically, PH assumes that the influence of covariates $\mathbf{X}$ will be constant at any point in the duration. This suggests how we might go about addressing the issue of nonproportionality. Consider a general PH model:

$$h(t|\mathbf{X}_i) = h_0(t)\exp(\mathbf{X}_i\beta)$$

We can think of a more generalized PH model which relaxes the strict proportionality as something like:

$$h(t|\mathbf{X}_i) = h_0(t)\exp[\mathbf{X}_i\beta + \mathbf{X}_i g(t)\gamma] \tag{1}$$

This is a general way of thinking of non-proportionality in hazards: The effects of covariates are allowed to vary as a function of time. In fact, this intuition forms the basis for a number of tests for PH, as well as for a general way of addressing nonproportionality.

Tests

In general, there are three kinds of tests for nonproportionality:

1. Tests for changes in parameter values for coefficients estimated on a subsample of the data defined by t,
2. Tests based on plots of survival estimates and regression residuals against time, and
3. Explicit tests of coefficients on interactions of covariates and time.

Stratified/Piecewise Regressions

If the influences of our covariates on the hazard of the event of interest vary over time, the simplest way they could do so is as a *step function*. In this vein, consider the function $g(t)$ in (1) as:

$$\begin{aligned} g(t) &= 0 \; \forall \; t \leq \tau \\ &= 1 \; \forall \; t > \tau \end{aligned}$$

This amounts to estimating interactions of each of the covariates with a dummy variable which is "turned on" for $t > \tau$. A few things about this approach:

- It allows the effect of the variables be different earlier than later in the process.
- One can choose τ on the basis of events/structure in the data (e.g. Congressional redistrictings, etc.) or just use the median point (better than the mean).
- More generally, one can consider more than one "step," if data are plentiful.

In general, stepwise models of this sort are a good place to start, but are probably not adequate solutions to the more general issue of nonproportionality.

Graphs and Residual Plots (For the Cox Model Only)

There are a couple different graphical approaches to assessing the proportional hazards assumption, all of which have been developed exclusively for the Cox model.

Graphs of the log-log Survivor Function

Kalbfleisch and Prentice (1980) were the first to suggest that one could make use of the predicted survival plots for subgroups of the data to assess the PH assumption. The idea is based on the fact that, in the Cox model,

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

which means that

$$\ln\{-\ln[S(t)]\} = H_0(t) \times \mathbf{X}_i\beta \tag{2}$$

This, in turn, means that plots of the predicted log-negative-log survivor function for different values of $\mathbf{X}$ should be parallel to one another. The intuition is that, if the hazards are proportional, variables should shift the $\ln\{-\ln[S(t)]\}$ by a constant factor, based on the way they enter the survivor function. Accordingly,

- Lines which are diverging, converging or crossing suggest time-varying effects of the covariate in question.
- This, in turn, is a signal of violation of the proportional hazards assumption.

Log-log survivor plots (as they are called) have some advantages:

- They are easy to generate (e.g., using Stata's `stphplot` command), so
- They are widely used.

They also, however, are not especially reliable; they'll often signal the presence of nonproportionality when in fact there is none, while at other times they can miss nonproportionality completely. Accordingly, they're a good (graphical) place to start, but other tests are better.

Plots and Tests Based on Residuals

- *Martingale Residuals*

Recall that the counting process formulation of the Cox model gave us something that looked like a residual:

$$\hat{M}_i(t) = C_i(t) - \hat{H}_i(t) \tag{3}$$

where $C_i(t)$ is the censoring indicator at t and $\hat{H}_i(t)$ is the integrated hazard, which will depend on covariates and parameters. Note that, thanks to the proportionality assumption, we can further write (3) as:

$$\hat{M}_i(t) = C_i(t) - \exp(X_{it}\hat{\beta})\hat{H}_0(t) \tag{4}$$

This latter term – the combination of the integrated baseline hazard and the covariates and parameter estimates – is itself known as a Cox-Snell residual; some of you familiar with GLMs have undoubtedly encountered these before.

Martingale residuals are akin to the difference between observed and expected values at t; these "residuals" have the property that:

- $\mathrm{E}(M_i) = 0$ and
- $\mathrm{Cov}(M_i, M_j) = 0$ asymptotically.

Also, note that

- If you have one record per unit, this is just the residual, but

- More than one record per subject yields the "partial" martingale residual (Stata's term). The latter change over t (because $\hat{H}_i(t)$ is also changing over t); we can sum across all t for each observation i to get the "total" martingale residual.

One can make modifications to the martingale residuals to correct for their inherent skewness (Therneau et al. 1990). The usefulness of martingale residuals is largely in general model checking, particularly in detecting either influential observations or departures from linearity in the effects of covariates.

- *Schoenfeld Residuals*

More useful from a proportional-hazards-testing perspective are a class of variable-specific residuals, usually called *Schoenfeld* residuals. The formulas are hard, and are presented in Fleming and Harrington (1991); the Stata manuals also have a nice compact discussion, as does Box-Steffensmeier and Jones. The intuition is based on differentiating the Cox partial log-likelihood with respect to each of the β_ks:

$$\begin{aligned} \frac{\partial \ln L(\beta)}{\partial \beta_k} &= \sum_{i=1}^{N} C_i \left\{ X_{ik} - \frac{\sum_{j \in R(t)} X_{jk} \exp(X_j \beta)}{\sum_{j \in R(t)} \exp(X_j \beta)} \right\} \\ &= \sum_{i=1}^{N} C_i (X_{ik} - \bar{X}_{w_i k}). \end{aligned} \tag{5}$$

We can think of $\bar{X}_{w_i k}$ as a weighted mean of covariate X_k over the risk set at time t, with weights corresponding to $\exp(\mathbf{X}\beta)$. By substituting the estimated $\hat{\beta}$ into (5), we can get the estimated Schoenfeld residual for the ith observation on the kth covariate as $\hat{r}_{ik}$:

$$\hat{r}_{ik} = C_i \left[X_{ik} - \frac{\sum_{j \in R(t)} X_{jk} \exp(X_j \hat{\beta})}{\sum_{j \in R(t)} \exp(X_j \hat{\beta})} \right] \tag{6}$$

The (bad) intuition is that, at a particular time t, this quantity is the cumulative covariate-specific difference between the expected and the observed values of the hazard.

From (5) and (6), it is apparent that Schoenfeld residuals:

- Are defined only at event times (note the presence of C_i in equation (6)),
- are asymmetrical (Therneau introduces a scaling procedure for these as well), and
- sum to zero across subjects at a particular time.

Most important for our purposes: If the proportional hazards assumption holds, these residuals should be unrelated to survival time; that is, they should be a "random walk" vis-à-vis

survival time. Conversely, if covariates are nonproportional, then the residuals will vary systematically with survival time.

To get the intuition of why this ought to be the case, remember that Schoenfeld residuals are something like unit-specific marginal covariate effects. Now, visualize think of the case where rising hazards are converging:

- The PH model assumes that they are in fact proportional to one another.
- PH will thus impose diverging hazards on the "predictions."
- The result will be residuals which "underpredict" the marginal effect of X_k at earlier time points, and "overpredict" it at later ones.
- So plotting the residuals against time (or log-time) will yield a negatively-sloped "cloud."

The reverse is true for hazards which are diverging: a PH model will "overpredict" early, and "underpredict" later, with the result that a plot of the residuals against time will have a positive slope.

This suggests (at least) two things we can do with these residuals:

1. *Plot* them against (some function of) time, to see if there's a pattern.
    - It's generally a good idea to use a smoother to do this.
    - Example (see the handout) – Supreme Court retirements...
2. *Tests* based on the correlation between some function of time and the residuals.
    - E.g., a correlation-like coefficient developed by Grambsch and Therneau (1994).
        - that uses a rescaled Schoenfeld residual dsescribed in the paper.
        - They also have a global test, based on all the covariate-specific residuals.
        - In Stata, this is the `estat phtest` command, with the `detail` option.
    - The test is a chi-square test for the null of no systematic variation in the residuals over time.
    - It also yields results for each variable.
    - One can specify various functions of time for it, as well.
    - In R, the `cox.zph` command will do the same things.

From our Supreme Court example, above:

```
. estat phtest, detail

      Test of proportional hazards assumption

      Time:  Time
      ----------------------------------------------------------------
                  |       rho            chi2       df       Prob>chi2
      ------------+---------------------------------------------------
      age         |      0.34444         6.64        1         0.0100
      pension     |     -0.06250         0.20        1         0.6553
      pagree      |     -0.09512         0.51        1         0.4770
      ------------+---------------------------------------------------
      global test |                      7.02        3         0.0712
      ----------------------------------------------------------------
```

What To Do About Nonproportionality?

The most common thing to do when confronted with nonproportionality is to incorporate covariate interactions with time. That is, if we expect that the influence of some variable X on $h(t)$ changes over time, we can simply include a term of the form $X \times g(t)$ on the right–hand side of the model in question. Note:

- Note that this can be done with any duration model (Cox, Weibull, etc.).
- In practice, people usually use $\ln(T)$, rather than just T, for the interactions; this is because in nearly every case the covariates enter the model as $\exp(\mathbf{X}_i\beta)$.
- Also: Do *not* include time itself as a covariate on the right-hand side.
- Interpretation is standard for interaction terms: the marginal effect of the covariate in question is then dependent on the time at which the change in that covariate occurs.

Time-by-covariate interactions can be both a test for nonproportionality as well as the remedy for it; it can yield very different, interesting results. For example, consider the international conflict example B-S&Z (2001):

- Growth decreases the likelihood of conflict, but does so more at later points than in earlier ones, while
- We observe the opposite effect for alliances: Their pacifying influence wanes over time.

In our Supreme Court example, we would likely want to include a time interaction with the `age` variable, as described in the handout:

```
. gen lnT=ln(service)

. gen agexlnT=age*lnT

. stcox age pension pagree agexlnT, nohr efron

Cox regression -- Efron method for ties

No. of subjects =          109                     Number of obs   =      1783
No. of failures =           52
Time at risk    =         1796
                                                    LR chi2(4)      =     41.69
Log likelihood  =    -173.2178                      Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0023907   .0589481     0.04   0.968    -.1131454    .1179269
     pension |   2.039256   .5487846     3.72   0.000     .9636582    3.114854
      pagree |   .0849074   .2968391     0.29   0.775    -.4968865    .6667012
     agexlnT |   .0250956    .019777     1.27   0.204    -.0136665    .0638578
------------------------------------------------------------------------------
```

Here, the effect of a one–unit difference in age is effectively zero at $\ln(T) = 0$ (that is, at $T = 1$), but increases over time. This is what we would have expected, given our residual–based tests, above. Note that the Grambsch-Therneau test for nonproportionality now reflects that there is no systematic variation in the covariate effects over time:

```
. estat phtest, detail

      Test of proportional hazards assumption

      Time:  Time
      ------------------------------------------------------------------
                  |       rho            chi2       df       Prob>chi2
      ------------+-----------------------------------------------------
      age         |     -0.06453         0.21        1         0.6481
      pension     |     -0.03437         0.06        1         0.8055
      pagree      |     -0.10123         0.59        1         0.4434
      agexlnT     |      0.24114         2.13        1         0.1446
      ------------+-----------------------------------------------------
      global test |                      6.11        4         0.1913
      ------------------------------------------------------------------
```

Interestingly, the substance of the changing effect of age over time is what we also found in the Weibull model above as well. That is, one can think of nonproportionality in the Cox model as similar to systematic (variable-specific) duration dependence in a parametric context. In both cases, the result is to make the marginal effect of that covariate on the hazard change as time passes.

Duration Dependence

In all of our discussions so far, we've talked about duration dependence as if it were an intrinsic characteristic of the hazard. But, stop and think for a second: Why might we have duration dependence? In fact, there are two broad types of reasons: *state dependence* and *unobserved heterogeneity*.

State Dependence

State dependence is the situation in which the value of the hazard in some way depends directly on its own previous values, and/or the amount of time that has passed in a state. Consider a few substantive examples:

1. *Institutionalization*: Institutions often become "sticky" over time, causing the hazard of their termination to drop – think of government agencies.
2. *Degradation*: Think of "wear and tear;" the longer (say) a machine runs, the more likely it is to fail, as parts become worn out, friction builds up, etc.
3. Similarly, consider the old "coalition of minorities" argument from presidential politics: presidents/parliaments slowly anger subgroups of the group that got them elected; over time, those groups defect, yielding higher hazards of (say) no-confidence votes.

Each of these is an example of *state dependence*: the (conditional) hazard of "failure" depends on how long you've been in the state. The relationship between state dependence and duration dependence, however, is a negative one:

$$\text{Positive State Dependence} \longrightarrow \text{Negative Duration Dependence}$$

while:

$$\text{Negative State Dependence} \longrightarrow \text{Positive Duration Dependence}$$

That is, a process in which the longer you are in a state the less likely you are to leave it will show hazards which are falling over time. Conversely, a process in which the probability of "failure" grows the longer you are in the state will show positive duration dependence (i.e., rising hazards over time).

Substantively, we're often interested in these sorts of things. Social scientists often talk about duration dependence in substantive terms (e.g./ Scott Bennett's work on alliances), and it is thus the dominant way of thinking about duration dependence in empirical work.

Unobserved Heterogeneity

There's another source of duration dependence, though, one that receives relatively little attention among social scientists: unobserved heterogeneity. This amounts to nothing more than observations being conditionally different in their individual hazards – say, due to omitted variable bias. Remember:

- Models assume that observations which are the same on all the covariates are otherwise identical.
- If they're not, that's unobserved heterogeneity.

The presence of such heterogeneity has an effect in general (misspecification), but it is especially bad in the context of survival-data models. To see why, consider two groups (call them $X = 0$ and $X = 1$) with different, exponential hazards.

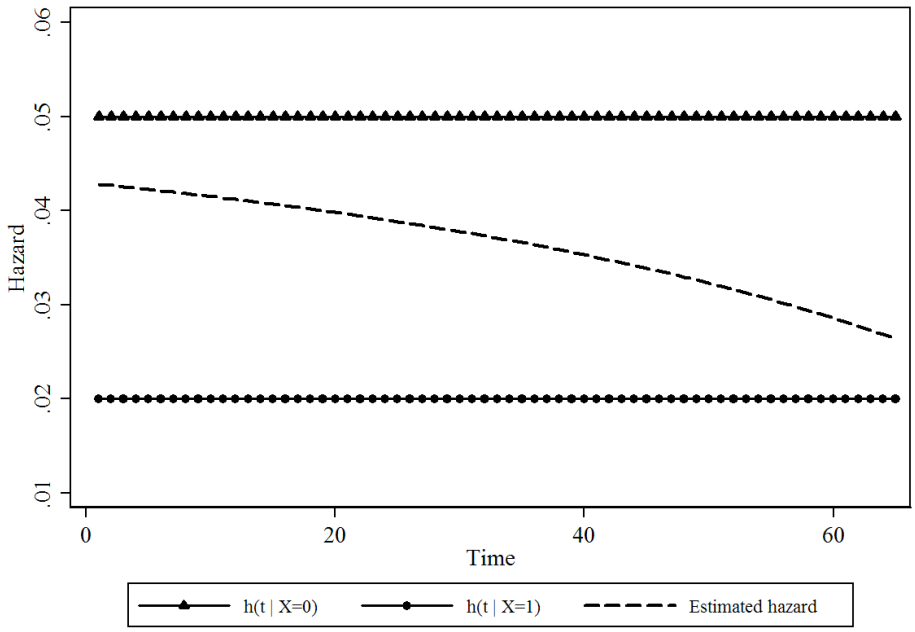
- Including X as a covariate tells us the magnitude of the differences; but
- If we don't include X in the model, then over time the observations with the higher hazards (i.e., those in group $X = 0$) will experience the event (and exit) at a higher rate.
- This means that, over time, the sample will increasingly be composed of observations with $X = 1$ (and lower hazards).
- Thus, the estimated hazard will appear to decline over time, even though there is no true state dependence at all (because the hazards are "flat").

The point of this is to note that unobserved heterogeneity yields hazards which will appear to be declining over time; and, more generally, hazards that are more downward–sloping than they truly are. This is referred to as "spurious duration dependence," and will occur even if the omitted covariate is independent of all the others, or of time. As a result of this,

- "Flat" hazards will be more negative, and
- "Rising" hazards will be flat, or even nonmonotonic (Omori and Johnson 1991).

This raises the issue that just talking about duration dependence in purely substantive / state–dependence terms confounds things. In fact, negative duration dependence (for example, a Weibull model in which $\hat{p} < 1.0$) may or may not indicate positive state dependence. The substantive interpretation of those $\hat{p}$s, then, is really potentially problematic.

Negative Duration Dependence Due To Unobserved Heterogeneity



What to Do About Duration Dependence?

Three things:

1. *Model specification.*
    - The better your model is, the less heterogeneity there will be, and the more confident you can be in substantively interpreting what you *do* observe/estimate.
    - For example, Lancaster (1979) noted that, in a model of strike duration, negative duration dependence decreased as variables were added (see also Bennett 1999).
2. *Models which address unit-level heterogeneity.*
    - These are called "frailty" models, and are akin to random effects models in a duration context.
    - We'll talk more about these a little later today; for now, note that if we can deal with the heterogeneity by (effectively) allowing each unit to have its own "intercept" (e.g., baseline hazard), then the problems associated with spurious duration dependence go away.
3. *Model the duration dependence itself.*
    - If we're substantively interested in duration dependence, we might also have hypotheses about it.

- For example, in a model of the duration of interstate wars, the "democratic peace" literature might lead us to think that democracies should have greater (positive) duration dependence than autocracies (since their electorates are more responsive to losing military efforts).
- To do this, one might allow the p parameter in the Weibull model to vary as a function of some covariates:
  - Formally, we'd likely specify $p = \exp(\mathbf{Z}_i\gamma)$, (or, equivalently, $\ln p = \mathbf{Z}_i\gamma$) since p needs to be strictly greater than zero.
  - We can then replace p with this expression in the usual Weibull likelihood, and estimate $\hat{\beta}$ and $\hat{\gamma}$ jointly.
  - Interpretation is straightforward: variables which increase (decrease) p cause the hazard to rise (fall) more quickly, or drop (rise) more slowly.
- We can also combine this with the model with unit-level frailties; e.g., the paper on alliance duration:
  - Theory: Bigger alliances will last longer.
  - Also: Larger alliances will be more institutionalized / "sticky," and so will also have lower (more negative) duration dependence.
  - This is in fact what happens: The effect of alliance size is to decrease both λ and p.
- There is (some) software to do this:
  - Stata allow you to introduce covariates into the "ancilliary" parameters of the various parametric models, though not while also estimating frailty terms.
  - A package called TDA (Blossfeld and Rohwer 1995) will also do this.

The handout (pp. 8-9) contains an example of this approach, using the Supreme Court retirement data.

Cure Models

"Cure" models are a means to relax yet another assumption we've so far been making about our event process. In particular, cure models are used for data in which not all units are, in fact, "at risk" for the event that we're studying. These have also been called "split-population" models in criminology and economics; I generally prefer the former term, as it more directly conveys what the models are about.

There are three general classes of cure models: parametric mixture, parametric non-mixture, and semi-parametric. We'll also discuss how one can implement cure models in a discrete-time approach as well. And we'll do (yet another) example using the familiar 1950-1985 MID data.

Mixture Models

Think for a moment about a standard controlled clinical trial. In certain circumstances, it may be the case that the drug in question cures the patient – that is, that for some fraction of the subjects receiving the drug, they will *never* experience the event that defines the end of the duration studied. Note several things about this:

- It is essentially impossible to know whether a particular subject is cured or not – that is, whether they will never have the event of interest, or whether they merely haven't had it yet. (In other words, "cured" is a *latent variable*).
- Similarly, the treatment may be neither necessary nor sufficient for a cure; some subjects may be "cured" on their own, even without receiving the drug.
- In either case, though, we have a problem: Standard survival models assume that $\int_0^\infty f(t)\,dt = 1 \,\forall\, i$ – that is, that all observations eventually have the event of interest. If there is a "cured" group in the data, this assumption is violated.

Now, think about this problem in the context of a standard parametric model for continuous-time duration data. Assume:

- $T_i > 0$ is the duration of interest, which
- has a density function $f(T_i|\mathbf{X}_i, \beta)$ with
- $\mathbf{X}_i$ a k–dimensional vector of covariates and θ a parameter vector to be estimated.
- The corresponding CDF is then $F(T_i|\mathbf{X}_i, \beta) = \Pr(T_i \leq t_i|\mathbf{X}_i, \beta), t_i > 0$, where t_i represents the duration defined by the end of the "follow–up" period for observation i. Also,
- call R_i to be the observable indicator of "failure," such that $R_i = 1$ when failure is observed and $R_i = 0$ otherwise.

The associated survival function is then equal to $1 - F(T_i|\mathbf{X}_i, \beta)$. We can then write the hazard in standard fashion as:

$$h(T_i|\mathbf{X}_i, \beta) = \frac{f(T_i|\mathbf{X}_i, \beta)}{S(T_i|\mathbf{X}_i, \beta)}$$

A Parametric Cure Model

Now, we need to consider a model for the duration T which splits the sample into two groups, one of which will eventually experience the event of interest (i.e., "fail") and the other which will not. To do so, define a latent (unobserved) variable Y such that $Y_i = 1$ for those observations that will eventually fail and $Y_i = 0$ for those that will not; define $\Pr(Y_i = 1) = \delta_i$. The corresponding conditional density and distribution functions are then defined as:

$$\begin{aligned} f(T_i|\mathbf{X}_i, \beta, Y_i = 1) &= g(T|\mathbf{X}_i, \beta) \\ F(T_i|\mathbf{X}_i, \beta, Y_i = 1) &= G(T|\mathbf{X}_i, \beta), \end{aligned}$$

while the corresponding density $f(T_i|Y_i = 0)$ and cdf $F(T_i|Y_i = 0)$ are undefined.[1]

For those observations that experience the event of interest during the observation period, we observe both $R_i = 1$ and their duration T_i. Since these observations also necessarily belong to the group in which $Y_i = 1$, we can write the unconditional density for these observations as:

$$\begin{aligned} L_i|R_i = 1 &= \Pr(Y_i = 1) \Pr(T_i = t|Y_i = 1, \mathbf{X}_i, \beta) \\ &= \delta_i g(T_i|\mathbf{X}_i, \beta) \end{aligned} \tag{7}$$

Intuitively, this illustrates that the observed duration is a function of two components: the probability that the observation would be among those that would eventually experience the event of interest (that is, $\Pr(Y_i = 1) \equiv \delta_i$) and, conditional on $Y_i = 1$, the conditional probability of failure at time T_i.

In contrast, for those observations in which we do not observe an event (that is, where $R_i = 0$), this fact may be due to two possible conditions:

1. It is possible that the observation in question is among those that will never experience the event defining the duration (that is, $Y_i = 0$).

2. It may also be the case, however, that the observation will experience the event, but simply did not do so during the observation period (that is, $Y_i = 1$ but $T_i > t_i$).

[1] Because $Y_i = 0$ implies that the observation will never experience the event of interest (and thus the duration will never be observed), the probabilities for $f(T_i|Y_i = 0)$ and $F(T_i|Y_i = 0)$ cannot be defined.

If, as is routinely the case, we assume that censoring is uninformative, the contribution to the likelihood for observations with $R_i = 0$ is therefore:

$$\begin{aligned} L_i|R_i = 0 &= \Pr(Y_i = 0) + \Pr(Y_i = 1)\Pr(T_i > t_i|Y_i = 1, \mathbf{X}_i, \beta) \\ &= (1 - \delta_i) + \delta_i[1 - G(T_i|\mathbf{X}_i, \beta)] \end{aligned} \tag{8}$$

Combining these values for each of the respective sets of observations, and assuming independence across observations, the resulting likelihood function is:

$$\mathbf{L} = \prod_{i=1}^{N} [\delta_i g(T_i|\mathbf{X}_i, \beta)]^{R_i} \{[(1 - \delta_i) + \delta_i[1 - G(T_i|\mathbf{X}_i, \beta)]\}^{(1-R_i)} \tag{9}$$

with the corresponding log–likelihood:

$$ln\mathbf{L} = \sum_{i=1}^{N} R_i\{\ln(\delta_i) + \ln[g(T_i|\mathbf{X}_i, \beta)]\} + (1 - R_i)\ln\{(1 - \delta_i) + \delta_i[1 - G(T_i|\mathbf{X}_i, \beta)]\} \tag{10}$$

Note a few things about this formulation:

- The hazard function for those who do experience the event may be any of the commonly–used parametric distributions (e.g., exponential, Weibull, log-logistic, etc.). Scholars are actively working on semi- and nonparametric approaches for estimating the latency; more on this below.

- The probability δ_i is typically modeled as a logit function of a set of covariates (call them $\mathbf{Z}_i$) and a vector of parameters γ:

$$\delta_i = \frac{\exp(\mathbf{Z}_i\gamma)}{1 + \exp(\mathbf{Z}_i\gamma)} \tag{11}$$

  although other specifications (e.g. probit, complimentary log-log, etc.) are also possible. Interpretation of these estimates is standard...

- Note that this model is identified even when the variables in δ_i are identical to those in the model of survival time. This means that one can test for the effects of the same set of variables on both the incidence of failure and the duration associated with it (Schmidt and Witte 1989).

Non-Mixture Models

An intuitive motivation for the non-mixture cure model arises in oncological studies of tumor recurrence following chemotherapy.[2] Prior to the treatment, we assume that the number of cancerous cell clusters is large; the effectiveness of the treatment, however, means that the probability of survival of any single cluster is vanishingly small (Yakovlev 1994, 1996).

- After treatment, an individual is left with N_i remaining clusters of precancerous cells, where (because of the process described above), $N_i \sim \text{Poisson}(\lambda)$.
- Define the probability of cure as $\pi_i = \Pr(N_i = 0)$.
- Now call the time for each of the remaining N_i clusters to develop into a detectable tumor Z_{ij}, $j = \{1, 2, ...N_i\}$, which has a distribution function $F(t)$.

Under these assumptions, the overall survival time until detection of the first post-therapy tumor is:

$$S(t) = \pi^{F(t)} \tag{12}$$

Note that, as $\pi \rightarrow 1.0$, the survival finction also converges on 1.0. The associated hazard function is then:

$$h(t) = -\ln(\pi) f(t) \tag{13}$$

which in turn reflects the fact that, in a model with a cured fraction, $\int_0^\infty h(t)dt < \infty$; specifically, $\int_0^\infty h(t)dt = -\ln(\pi)$. More usefully, $S(t)$ can be rewritten:

$$S(t) = \exp[\ln(\pi)F(t)]. \tag{14}$$

Note that this means that if $F(t)$ does not depend on covariates, the model is one in proportional hazards. In particular, if we assume a complimentary log-log link for the cured fraction:

$$\pi_i = \exp[-\exp(\mathbf{X}_i\beta)]$$

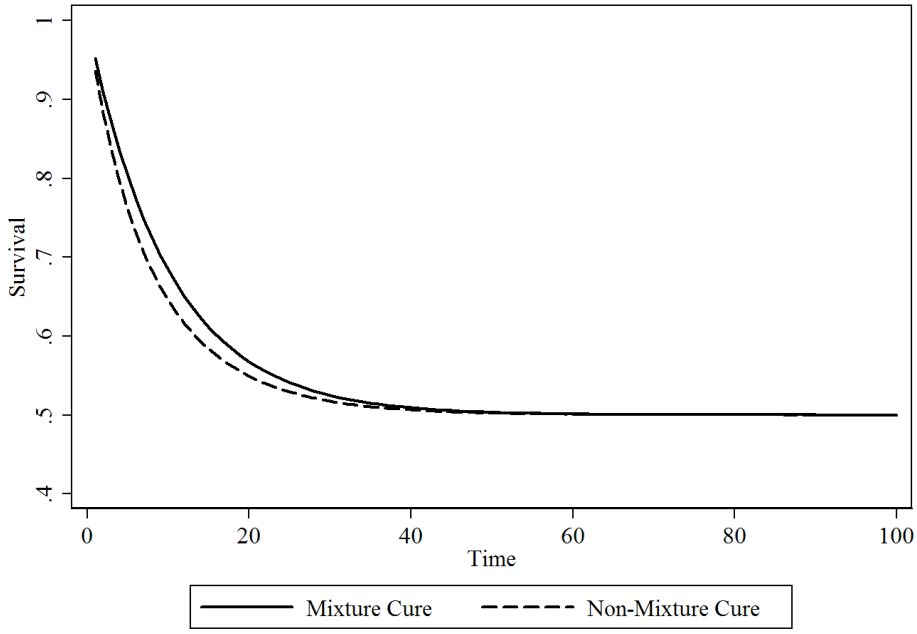
then the overall survival function conforms to that of the Cox model, where $F(t)$ is a transform of the "baseline" hazard (See Sposto 2002 for a derivation).

In comparison to the mixture model:

- In most instances, there is very little difference – if the chosen density function $f(t)$ and its associated quantities are the same, then the two models will generally yield very similar results (Figure 7).

[2]This section owes a great deal to Sposto (2002); other good references for non-mixture cure models include Chen 2002; Tsodikov 2003; Yin 2005; and Tournard and Ecochard 2006).

Figure 7: Mixture and Non-Mixture Cured-Fraction Survival Functions (Exponential Hazards with $\lambda = 0.1$ and $\pi = 0.5$)



- In terms of interpretation, the non-mixture model cannot and should not be interpreted as a model for a mixture of cured and non-cured individuals in the population. Instead, it is a multiplicative distribution of two otherwise independent probabilities.
- Some (e.g. Lambert et al.) have reported that non-mixture models have better model convergence properties than mixture models.

Semiparametric Cure Models

There have been a few papers making developments on semiparametric (Cox) cure models; all of them have appeared quite recently (cf. Tsodikov 1998; Sy and Taylor 2000; Peng and Carriere 2002; Peng 2003; Lu and Ying 2004). We won't discuss them at any length (but see below), but you are welcome to look into them.

Discrete-Time Cure Models

One of the benefits of the equivalence between Poisson/counting process models and survival analysis is the ability to use count-model advances in a survival analysis context. Specifically, a "zero-inflated" Poisson model (cf. Lambert 1992; Zorn 1998) (which is really nothing more than a Poisson model mixed with a point mass at zero) which includes period-specific dummies can exactly estimate a discrete-time cure model with time-varying covariates, in

the same fashion that a standard Poisson GLM with such dummies can replicate a Cox survival model.

Specifically, call:

- The latent probability of only observing zeros $= p_i^*$,
- The binary realization of p_i^* is $p_i \in \{0, 1\}$, and
- The underlying count variable $Y_i^* \sim \text{Poisson}(\lambda_i)$, which is only observed if $p_i = 1$.

Then, the probability of a zero outcome is:

$$\begin{aligned} \Pr(Y_{it} = 0) &= \Pr(p_{it} = 0) + [\Pr(p_{it} = 1) \times \Pr(Y_{it}^* = 0)] \\ &= (1 - p_{it}^*) + p_{it}^*[\exp(-\lambda_{it})] \end{aligned}$$

and the probability of a non-zero outcome is:

$$\begin{aligned} \Pr(Y_{it} = y) &= \Pr(p_{it} = 1) \times \Pr(Y_{it}^* = y) \\ &= p_{it}^* \times \frac{\exp(-\lambda_{it})\lambda_{it}^y}{y!} \end{aligned}$$

with

$$E(Y_{it}^*) \equiv \lambda_{it} = \exp(\mathbf{X}_{it}\beta)$$

and

$$\Pr(p_{it} = 1) = \frac{1}{1 + \exp(-\mathbf{Z}_{it}\gamma)} \text{ or } \Phi(\mathbf{Z}_{it}\gamma)$$

The likelihood and so forth are straightforward. Estimation can be a bit problematic, particularly if there is significant overlap between $\mathbf{X}$ and $\mathbf{Z}$, but in general is pretty well-behaved.

Practical Considerations

The use of cure models such as that outlined above should be considered whenever all observations cannot reasonably be assumed to "fail" at some point in the future (Chung, Schmidt, and Witte 1991). A particularly useful property of cure models is that they allow for separate estimation of the influence of covariates on the probability of experiencing the event from their effect on the time until the event of interest occurs for those observations that do experience the event. That is, covariates can have an independently positive or negative influence, or no effect at all, on both the incidence and the latency of an event. This fact makes cure models more flexible than other duration models; one may find that a particular

covariate affects incidence but not latency, or vice versa. Such interpretations are not available with other duration models.

Analysts should thus consider using cure models whenever there is a theoretical reason to suspect that not all observations will eventually "fail." Such an assessment is relatively straightforward and more common than the political science literature currently reflects. Political scientists have simply not been routinely asking whether all observations are expected to eventually fail, but they should.

In addition to theoretical considerations, one can empirically look at the data to get a general sense of the need for relaxing the assumption made by all other duration models, i.e., that eventually all observations experience the event. By plotting a Kaplan Meier (KM) figure of the survivor function versus time (Kaplan and Meier 1958), the analyst will gain a sense of whether observations in the data exist that will not experience the event of interest. Price (1999) and Sy and Taylor (2000) illustrate the use of a KM survival curve to empirically assess the need for a split population model. If it "shows a long and stable plateau with heavy censoring at the tail," there is strong reason to suspect that there is a subpopulation that will not experience the event (Sy and Taylor 2000, 227; see also Peng et al. 2001).

Several issues arise in the estimation and interpretation of cure models. Note, for example, that when $\delta_i = 1 \,\forall\, i$ (that is, when all observations will eventually fail), the likelihood reduces to that for a standard duration model with censoring. However, testing for $\delta_i = 1$ is a case of a boundary condition[3] and thus standard asymptotic theory does not apply (Price 1999). Maller and Zhou (1996) offer a corrected likelihood–ratio test for the proposition that all observations will eventually experience the event of interest. Issues of goodness-of-fit for split population models are an important, but currently ongoing, area of research (e.g., Sy and Taylor 2000). Finally, one can also test $\mathrm{H}_0 : \delta = 1$ (i.e., if the assumption that all observations will eventually fail is true) statistically. If so, the equation reduces to the standard general duration model with censoring; that is, a (e.g.) Weibull model is a special case of the Weibull cure model.

[3]Note that this does not correspond to the case of $\mathbf{Z}_i\gamma = 0$ (which yields $\delta_i = 0.5$).

Estimating Cure Models in Stata

In Stata, there are (at least) four options for estimating survival models with a cured fraction. Each has its pluses and minuses...

`lncure`

- Estimates a log-normal cure model *only*.
- Does not allow covariates to determine the cured fraction (that is, one can estimate only $\hat{\delta}$, not $\hat{\delta}_i$).
- It's `predict` is (currently) broken...
- Not, in general, a great option.

`spsurv`

- Will estimate discrete-time cure models (a la using `cloglog` in a discrete-time context).
- Duration dependence is, therefore, up to the analyst.
- As with `lncure`, does not allow covariates in the δ parameter.
- Also won't allow "robust" standard errors.
- Also not a great option.

`cureregr`

- Fits a very flexible generalized parametric cure model, as in (10), above.
- Allows both "mixture" and "non-mixture" cured fraction models.
- Allows a variety of distributions:
    - Exponential
    - Weibull
    - Log-Normal
    - Logistic
    - Gamma
- Allows for covariates in the "scale" parameter (that is, the duration part of the model), the "cured fraction"/mixture parameter, and even the "shape" parameter (cf. today's discussion of modeling duration dependence).

- Allows different "links" for the covariates to the cured fraction parameter:
  - Logistic: $\delta_i = \frac{1}{1+\exp(-\mathbf{Z}_i\gamma)}$.
  - Complimentary log-log: $\delta_i = \exp[-\exp(\mathbf{Z}_i\gamma)]$.
  - Linear: $\delta_i = \mathbf{Z}_i\gamma$.
- Does *not* allow "robust" standard errors.
- Is probably the best option for parametric models.

`zip / zinb`

- These are the commands for "zero-inflated" Poisson and negative binomial models, respectively.
- One can estimate cure models by:
  - Treating the event indicator as the response variable, and
  - Including the covariates believed to influence the cured fraction in the `inf()` part of the model.
- This is a discrete-time setup; temporal issues have to be dealt with explicitly on the right-hand side of the model.
- Allows "robust" variance-covariance estimates.
- Is very flexible – see the example in the handout (pp. 14).

Cure Models: A Quick Example

We return too data on disputes among "politically-relevant" international dyads between 1950 and 1985 for our example, using the same six standard variables as before. The handout presents:

- A standard Weibull model, for comparison,
- a model estimated using `spsurv` (i.e., a discrete-time model with a constant cured fraction),
- two parametric (here, Weibull) models – one mixture, one non-mixture – estimated using `cureregr`,
- a set of comparison plots for the values of interest in these two models, and
- a semiparametric model estimated using `zip`.

Heterogeneity and "Frailty" Models

Cure models represent a particular form of *heterogeneity* in survival data – some observations are essentially not "at risk" for the event of interest at all. We can extend this logic a bit to think of situations where each individual has some (individual–specific, latent) propensity toward the event of interest. This is the intuition behind *frailty* models in survival analysis.

A (Very) Little Math

Consider a model of the form:

$$h_i(t) = \lambda_i(t)\nu_i \tag{15}$$

Here, the ν_is can be thought of as unit–level factors which operates multiplicatively on the hazard.

- $\nu_i = 1$ can be thought of as a "baseline,"
- $\nu_i > 1$ means that individual i has a greater-than-average propensity toward having the event in question, while
- $\nu_i < 1$ means the opposite.

We might think of the ν_is as the combined influence of a bunch of unobserved variables. In this light, the cure models we discussed a bit ago correspond to a special case where some of the $\nu_i = 0$.

In a Cox model context, (15) might look like

$$h_i(t) = h_0(t)\nu_i \exp(\mathbf{X}_i\beta) \tag{16}$$

which is often reexpressed as:

$$h_i(t) = h_0(t)\exp(\mathbf{X}_i\beta + \alpha_i) \tag{17}$$

where $\alpha_i = \ln(\nu_i)$. In a regression-type survival model, the ν_is are analogous to unit-level effects in panel data, and the way(s) we deal with them are similar as well.

Implications

Q: What happens if such heterogeneity is present, but unaccounted for? That is, what happens of you ignore these ν_is?

A: Berry, berry bad stuff...

1. The parameter estimates from a model ignoring the effects are inconsistent (Lancaster 1985). In essence, because you've misspecified the model.
2. Ignoring the unit effects will also lead you to tend to underestimate the hazard (Omori and Johnson 1993). The intuition of this is straightforward:
    - There is more variability in the actual hazard than your model is picking up, so
    - Over time, this will cause observations to "select out" of the data (like we discussed earlier regarding duration dependence):
        - Low-frailty cases will stay in
        - High-frailty ones will drop out
    - The result is an underestimated hazard.
    - Think of the cure model as one example – in the end, you'll be left with only the "cured" folks, and so
    - You'll correspondingly overestimate the survival times.
3. You'll likely get the shape of the hazard wrong...
4. If the ν_is are correlated with the $\mathbf{X}$s, you'll get lousy estimates of the βs, too.

How To Deal

OK, its bad. So, what do we do?

Well, consider a linear model with unit effects, by comparison:

- We might use *fixed effects* to estimate the α_is, or
- Use *random effects*: Assume a distribution for the α_is, condition them out, and estimate the parameters of that distribution.

In the survival context, fixed effects are not a good option...

- There's the problem of incidental parameters (Lancaster 2000), which lead to inconsistency.
- This, in turn, tends to deflate our standard error estimates.
- As a result, fixedeffects models aren't generally used in a survival context.

A random-effects approach (called "frailty" in the survival literature) is much more common...

- E.g., Lancaster (1979) in economics, Vaupel et al. (1979, 1981) in demography, etc.

- As with random-effects models, frailty models involve making an assumption about the distribution of the (random) ν_is, and then conditioning them out of the resulting likelihood.

Not surprisingly, then, the most common frailty models are parametric ones; and the most often-used distribution for the frailties is the *gamma distribution*. So, for example, the Weibull distribution with a frailty term has a conditional survival function that looks like:

$$S(t|\nu) = \exp[-(\nu\lambda t)^p)], \tag{18}$$

which is to say, one that is a rescaled version of the standard Weibull survival function:

$$\exp(-\lambda t)^p.$$

Now, if we specify that $\nu_i \sim g(1, \theta)$, where the gamma density g is

$$g(\nu, \theta) = \frac{1}{\theta^{(1/\theta)}\Gamma(1/\theta)}\nu^{1/\theta-1}\exp\left(\frac{-\nu}{\theta}\right)$$

and $\Gamma(\cdot)$ is the indefinite Gamma integral. Making this assumption about the distribution of the νs means that the marginal survivor function is then equal to:

$$S(t) = [1 + \theta(\lambda t)^p]^{-1/\theta} \tag{19}$$

with a corresponding hazard function of

$$\begin{aligned} h(t) &= \frac{\lambda p(\lambda t)^{p-1}}{1 + \theta(\lambda t)^p} \\ &= \lambda p(\lambda t)^{p-1}[S(t)]^\theta. \end{aligned} \tag{20}$$

As usual, we introduce covariates as $\lambda_i = \exp(\mathbf{X}_i\beta)$. Note that:

- This looks more-or-less like a standard Weibull, but with an added "weighting" that is a function of $S(t)$ and θ, and
- If/when $\theta = 0$, the distribution reverts to the standard Weibull (that is, in the case where there is no unit-level variability).

It is also possible to derive a frailty model in the Cox context; in both parametric and semi-parametric cases, the idea is to integrate out the frailties to get a conditional hazard/survival function from which $\hat{\beta}$ can then be estimated.

Estimation

How do we estimate frailty models? There are several options...

The E-M algorithm (Klein 1992)

- This is outlined in Ch. 9 of Hosmer & Lemeshow.
- It essentially involves five steps:
  1. Fit a standard (e.g., Cox) model, and retain the estimate of the baseline hazard $\hat{H}_0(t_i)$ for each observation.
  2. Choose a set of possible values for θ (let's call them $\tilde{\theta}$, e.g. $\tilde{\theta} \in \{0, .1, .2, ...4, 4.5, 5\}$).
  3. For each value of $\tilde{\theta}$, generate an estimated "predicted frailty" $\hat{\tilde{\nu}}_i$ for each observation. This is the "E" step.
  4. Fit a second survival model, this time including the estimated $\hat{\tilde{\nu}}_i$s as an additional covariate, with a fixed coefficient of 1.0 (that is, as an offset):

     $$h(t) = h_0(t)\hat{\tilde{\nu}}_i \exp(\mathbf{X}_{it}\beta)$$

     and use those values to reestimate $\hat{\tilde{\nu}}_i$. That's the "M" step.
  5. Repeat steps 3 and 4, replacing with the value from the model including the generated frailties, until "convergence" (that is, things don't change any more).
- After doing steps 1 - 5 for each value of $\tilde{\theta}$,
- You can then evaluate the "profile log-partial-likelihood" for each of the values of $\tilde{\theta}$ to figure out what the MLE is.

Direct Estimation

It's also possible to directly estimate β and θ.

- This has been implemented for the Cox model with gamma, gaussian, and t frailties in R (via the `survival` package), and for the Cox model with gamma frailties in Stata's `-stcox-` command.
- Both packages adopt a "penalized likelihood" approach.
- See the R manuals, or the paper by Therneau, for more details...

Another Alternative

- Remember that, in event count models: Poisson event arrivals + gamma heterogeneity = negative binomial.
- In theory, we can port this idea over to the survival/counting process world, and
- use a negative binomial model to estimate the variance of the frailties.
- References are Lawless (1987), Thall (1988) Abu-Libdeh et al. (1990), Turnbull et al. (1997). In this case:
  - The "dispersion" parameter is equal to the estimate of the frailty variance. But...
  - I've personally never been able to get this to work...

A Few Other Relevant Points

- In the past, there have been a number of big fights over whether choosing a particular distribution for ν_i matters or not, particularly among economists. Earlier studies seemed to say that if you pick the wrong distribution for the frailties, the model was in deep trouble. More recent work (e.g. Pickles and Crouchley 1995) questions this assertion.
- Frailty models have the same strong requirements for consistency as do random–effects models in a panel context; in particular, that $\text{Cov}(\mathbf{X}_i, \nu_i) = 0$. In addition, they also require that the frailties be independent of the censoring mechanism (that is, that $\text{Cov}(C_i, \nu_i) = 0$ as well).
- Interpretation of the resulting estimates is standard, with the caveat that all interpretations are necessarily conditional on some level of frailty $\hat{\nu}_i$. In most instances, people set $\hat{\nu}_i = 1$, which is the natural (mean) frailty level.
- Frailty models are probably best used when the researcher suspects that one or more important unit-level covariates may have been omitted from the model. (See the example below for a bit more on this).

Frailty Models in Stata

Stata allows the analyst to estimate either parametric or semiparametric (Cox) models with "shared" (that is, unit-level) frailties. The relevant option in both `stcox` and `streg` is `shared()`, which takes as its argument the relevant identifier variable denoting the groups that have similar frailty values. Note that:

- While `streg` allows either gamma-distributed or inverse–gaussian (i.e., normally) distributed frailties, `stcox` permits only gamma-distributed frailties.

- These models can be a bit computationally challenging, particularly on large datasets.
- Note that Stata will also allow the analyst to generate $\hat{\nu}_i$s for each of the observations in the data. These can be useful, in that they can allow the analyst to see what observations are more or less "frail" vis-à-vis the event of interest.

Finally, it should be noted that frailty models have also been used to deal with a particular kind of heterogeneity: that due to repeated events. More on this below...

Frailty Models in R

R is also a strong package for estimating frailty models, particularly those rooted in the Cox model. R will estimate Cox models with frailties that are:

- Gamma-distributed,
- Gaussian, or
- t-distributed.[4]

The command is just the `frailty()` option on `coxph`:

```
> GFrail<-coxph(Surv(start, duration, dispute, type="counting")~contig+capratio
+allies+growth+democ+trade+frailty.gamma(dyadid, method=c("em")))
```

[4]That is, the α_is are either Normal or t on the scale of the linear predictor; the frailties ν_i are log-normal and log-t, respectively.

Options include the method of estimation (EM, penalized partial-likelihood) as well as controls over the parameters of those estimation subcommands. The result is a `coxph` object:

```
> summary(GFrail)
Call:
coxph(formula = Surv(start, duration, dispute, type = "counting") ~
    contig + capratio + allies + growth + democ + trade + frailty.gamma(dyadid,
        method = c("em")))

  n= 20448
                           coef   se(coef) se2     Chisq  DF  p
contig                      1.199  0.1673   0.1310  51.41   1 7.5e-13
capratio                   -0.199  0.0547   0.0495  13.29   1 2.7e-04
allies                     -0.370  0.1685   0.1252   4.82   1 2.8e-02
growth                     -3.685  1.3457   1.2991   7.50   1 6.2e-03
democ                      -0.365  0.1309   0.1108   7.78   1 5.3e-03
trade                      -3.039 12.0152  10.3084   0.06   1 8.0e-01
frailty.gamma(dyadid, met                           708.95 394 0.0e+00

         exp(coef) exp(-coef) lower .95 upper .95
contig      3.3182      0.301  2.39e+00  4.61e+00
capratio    0.8193      1.221  7.36e-01  9.12e-01
allies      0.6908      1.448  4.97e-01  9.61e-01
growth      0.0251     39.845  1.80e-03  3.51e-01
democ       0.6940      1.441  5.37e-01  8.97e-01
trade       0.0479     20.876  2.84e-12  8.09e+08

Iterations: 7 outer, 27 Newton-Raphson
     Variance of random effect= 2.42   I-likelihood = -2399.4
Degrees of freedom for terms=   0.6   0.8   0.6   0.9   0.7   0.7 394.2
Rsquare= 0.052   (max possible= 0.227 )
Likelihood ratio test= 1089  on 399 df,   p=0
Wald test            = 121  on 399 df,   p=1
```

R will also estimate parametric frailty models with the `survreg` command, in exactly the same fashion:

```
> W.GFrail<-survreg(Surv(duration, dispute)~contig+capratio+allies+growth+democ
+trade+frailty.gamma(dyadid, method=c("em")))

> print(W.GFrail)
Call:
survreg(formula = Surv(duration, dispute) ~ contig + capratio +
    allies + growth + democ + trade + frailty.gamma(dyadid, method = c("em")))

                          coef    se(coef) se2    Chisq   DF  p
(Intercept)                6.0133  0.1646  0.1438 1333.93   1 0.0e+00
contig                    -1.5687  0.1692  0.1409   85.99   1 0.0e+00
capratio                  -0.0164  0.0221  0.0198    0.55   1 4.6e-01
allies                     0.7220  0.1707  0.1386   17.90   1 2.3e-05
growth                    -0.5488  0.8454  0.8362    0.42   1 5.2e-01
democ                     -0.0431  0.0937  0.0860    0.21   1 6.5e-01
trade                     22.7762 10.4935  9.6378    4.71   1 3.0e-02
frailty.gamma(dyadid, met                          3103.90 323 0.0e+00

Scale= 0.541

Iterations: 8 outer, 41 Newton-Raphson
     Variance of random effect= 1.82   I-likelihood = -1746
Degrees of freedom for terms=   0.8  0.7  0.8  0.7  1.0  0.8  0.8 322.6  1.0
Likelihood ratio test=1525  on 327 df, p=0  n= 20448
```

In general, R is *much* faster than Stata at estimating shared-frailty models.

Extensions

Like other models with random effects (and random-parameter models in general), frailties turn out to be valuable in a host of different survival data contexts. Some of the readings give you a flavor for these:

- **Repeated Events**. As we'll see in a few minutes, one could use frailties to account for dependence due to observations having repeated events.
- **Multilevel Frailties**. Sastry (1997) applies a multilevel / hierarchical frailty approach to child mortality in northeast Brazil. The idea, of course, is that one can have more than one level of unit effect, in a fashion analogous to hierarchical linear and nonlinear models. This is a *big* growth area in survival modeling at the moment (2007).
- **Spatial Frailties**. An article by Banerjee et al. (2003) uses frailty terms in a Bayesian context, to fit a model with spatially-referenced frailty terms to data on infant mortality in Minnesota. For reasons that are readily apparent if you think about it, spatial survival models are another big growth area in these sorts of statistics.

I encourage you to explore any of these that you think might be useful in your own research.

An Example

The handout contains the results of estimating Cox and Weibull models with gamma frailties on the 1950-1985 international dispute data. Note several things:

- The results for the variables are generally the same as for the standard Cox model; they can be interpreted similarly, though it is important to note (as the output indicates) that the results – including the standard error estimates – are conditional on the random effects $\hat{\nu}_i$.
- R reports a Wald test for the null hypothesis that $\theta = 0$ – that is, that there are no unit-level random effects / frailties. Here, we can confidently reject that null.
- To get predictions for the unit-specific frailties, we'd use `predict` in R. Alternatively, if we were using Stata, we'd add the `effects(`*`newvar`*`)` option to the `stcox` command; this creates a new variable called *`newvar`* that contains the estimated log-frailties.

Competing Risks

The idea of *competing risks* addresses the potential for multiple failure types. For example,

- Members of Congress can retire, be defeated, run for higher office, or die in office.
- Supreme Court justices can either die or retire.
- Wars can end with victory by one side, or a negotiated settlement.

Motivation

Assume that observation i is at risk for R different kinds of events, and that each event type has a corresponding duration $T_{i1}, ..., T_{iR}$ associated with it, each of which has a corresponding density $f_r(t)$, hazard function $h_r(t)$ and a survivor function $S_r(t)$, $r \in \{1, 2, ...R\}$.

- Of these possible durations, we observe the shortest: $T_i = \min(T_{i1}, ...T_{iR})$.
- We also observe an indicator of which event the observation experiences: $D_i = r$ iff $T_i = T_{ri}$.
- Finally, we typically assume there can't be exact equality across the T_{ir}s (though this is not a big deal), and that, given long enough, the observation would have experienced each of the events in question eventually.

Estimation

If the risks for the various types of events are *conditionally independent* (that is, independent once the influence of the covariates $\mathbf{X}$ are taken into account – more on this in a bit), then estimation is easy. The contribution of each uncensored observation to the likelihood is:

$$L_i = f_r(T_i|\mathbf{X}_{ir}, \beta_r) \prod_{r \neq D_i} S_r(T_i|\mathbf{X}_{ir}, \beta_r) \tag{21}$$

That is, the contribution of a given observation with failure due to risk r to the likelihood function is identical to its contribution in a model where only failures due to risk r are observed and all other cases are treated as censored. The overall likelihood is then:

$$L = \prod_{i=1}^{N} \left\{ f_r(T_i|\mathbf{X}_{ir}, \beta_r) \prod_{r \neq D_i} S_r(T_i|\mathbf{X}_{ir}, \beta_r) \right\} \tag{22}$$

which – because we observe only one of the r events and its corresponding survival time – can be rewritten as:

$$L = \prod_{r=1}^{R} \prod_{i=1}^{N_r} \{ f_r(T_i|\mathbf{X}_{ir}, \beta_r)\, S_r(T_i|\mathbf{X}_{ir}, \beta_r) \} \tag{23}$$

where N_r denotes summation over the set of observations experiencing event r. If we modify our old familiar censoring indicator, such that $C_{ir} = 1$ indicates that observation i experienced event r, and $C_{ir} = 0$ otherwise, we can further rewrite (23) as:

$$L = \prod_{r=1}^{R} \prod_{i=1}^{N} \left[f_r(T_i|\mathbf{X}_{ir}, \beta_r) \right]^{C_{ir}} \left[S_r(T_i|\mathbf{X}_{ri}, \beta_r) \right]^{1-C_{ir}} . \tag{24}$$

and the log-likelihood is then just the sums (over r and i) of the logs of the terms on the right-hand side of (24).

- The proofs for this are in Cox and Oakes 1984; David and Moeschberger 1978; Diermeier and Stevenson 1999; and a number of other places.
- The intuition is easy: To the extent that the (marginal) risks are independent, their covariances are zero, and drop out of the likelihood functions, leaving us with easy products.
- So, if two risks j and k are conditionally independent of one another, we may analyze durations resulting from failure due to risk j by treating those failures due to k as censored, in the sense that they have not (yet) reached their theoretical time to failure from risk j; the same may then be done for risk k by treating failures due to risk j as censored.
    - That is, you can just run the model "both ways."
    - Interpretation is then standard for each of the two models.
    - Also, there is no identification problem with having similar (or even identical) sets of covariates in the models for the various failure events, if that is what theory suggests is the right thing to do.

Diermeier and Stevenson (1999) give a nice example of this in the context of the cabinet failures literature, where the competing risks are cabinet failures due to elections and replacements.

Independence of Risks

At first thought, it may seem that assuming conditionally independent risks is a pretty strong and/or unjustifiable thing to do. But, remember:

- The risks only need be independent *conditional on the effects of the covariates.*
- This means that, if a particular covariate affects the hazard of more than one event, and it is in the model, then its effect is "controlled for."
- In other words, only the *baseline* risks need be independent.

- If you've got a good model, then, this may not be such a strong assumption.

Unfortunately, there are no great "tests" for the presence of conditional dependence in competing risks. (And, for a long time, there was even a big debate over whether a dependent-risks model is even identified, or identifiable.) As a general rule, people tend to use independent-risks models. But, if you really, really think your risks are conditionally dependent, there are several options:

1. One can model dependent risks using frailties/random effects (e.g. Oakes 1989, Gordon 2001), as we discussed above. Sandy Gordon's 2001 *AJPS* paper is a nice introduction to this approach. The intuition is that one assumes that the correlation between the risks is due to individual-specific factors, which are then captured in the frailty term and integrated out of the likelihood.

2. One can also take advantage of the Poisson/duration equivalence we talked about last time...

    - If we think of standard model as a discrete-time logit, then
    - Competing risks are like a multinomial logit:
        - Independent risks correspond to the "independence of irrelevant alternatives" assumption in polychotomous choice models, which means that
        - Models which relax that assumption (e.g., multinomial probit, generalized extreme-value, nested logit, etc.) can be used to estimate dependent competing risks models.

3. Fukumoto (2005) also introduces a dependent discrete-time competing risks model.

4. Alternatively, something like a seemingly-unrelated Poisson model could, in principle, also be used this way, though I've never seen it done.

An Example: Supreme Court Vacancies

- Data: Supreme Court Vacancies, 1789-1992.
- Competing risks: Justices can leave the Court through retirement or through, ahem, mortality (cf. Zorn and Van Winkle 2000).
- Covariates:
    - *Chief justice* (naturally coded),
    - Justice from the *South* (naturally coded),
    - Justice's *age* (in years),
    - Justices' *pension eligibility* (naturally coded indicator),

- Justice's *party agreement* (coded one if the party of the sitting president is the same as that of the president that appointed the justice, and zero otherwise).

- Results for both types of risks are available in the handout:
  - Independent (Weibull) models for the two competing risks.
  - An (independent-risk) discrete-time approach, based on the application of multinomial logit to the data.
  - A dependent discrete-time model, using multinomial probit.
  - Note that the predictions (in this case, survival; probabilities) are essentially identical for the two models (p. 20).

Models for Repeated Events

A large number of the kinds of things that social scientists study are capable of repetition:

- International wars (that is, between dyads of countries),
- Marriages,
- Policy changes / shifts,
- Cabinet failures within countries, etc.

The possibility of repeated events leads to the potential for dependencies across events for the same unit. This, in turn, makes the usual PL- or ML-based inferences suspect:

- Treating such events as independent implies we have more information than we do.
- This leads to a tendency to (usually, but not always) underestimate s.e.s and/or overestimate the precision of our estimates.

Moreover, at times we may have theory about – or be otherwise interested in – the fact that second, third, etc. events are somehow different from "first events." For example, the "security dilemma" and the resulting spiral models of war suggest that, well, war will lead to additional war; on the other hand, informational models of war (a la Fearon, Gartzke, etc.) suggest that the occurrence of war decreases the hazard of future wars. Which is it? Absent better methods than we've introduced so far, we can't tell.

There are two general approaches to repeated events: Frailty models, and variance-correction models

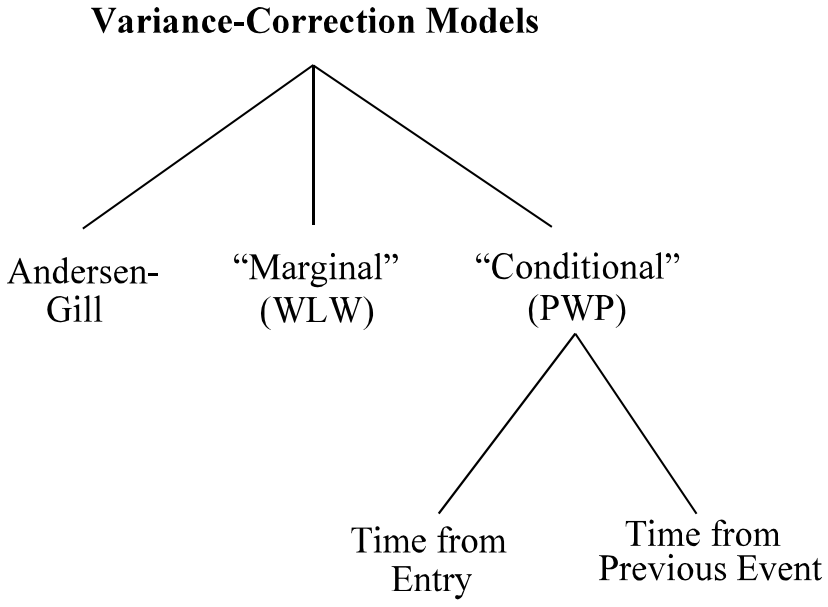
Frailty Models

We discussed these just a bit ago. In the context of repeated events, we can think of capturing the possible dependence across events within a subject with the frailty ν_i.

- This makes some substantive sense, if the frailty is fixed over time.
- Practically speaking, one just estimates a model where the shared frailty identifier denotes the unit that is experiencing the multiple events.
- Intuitively, this addresses the possibility of dependence *if we are willing to assume that such dependence is largely a function of unit-level influences.*
- So, for example, if some cabinets (say, Italy's) are more likely on average to collapse than others (say, Great Britain's), this would be a reasonable approach.

In a nutshell, the use of frailty terms can be – and often is – thought of as a means of dealing with data where events repeat.

Figure 11: Types of Variance-Correction Models



"Variance-Correction" Models

"Variance-Correction" (or "marginal") models do exactly what the name implies: they estimate a model as if the events were not repeated/dependent, and then "fix up" the variance-covariances after the fact. This is analogous to GEE-type models for panel data. As per the recent literature on the subject, there are four variance-correction models that have been widely used:

1. Andersen/Gill (AG)
2. Wei et al. (WLW)
3. Prentice et al. elapsed time (PWP-ET)
4. Prentice et al. gap (interevent) time (PWP-GT)

Figures 11 and 12 present a schematic of the relationship among these various types. There are three key things that define the different variance-corrected models:

1. The definition of the *risk set*,
2. The *time scale* / variable that is used,
3. Whether baseline hazards are constant across events, or allowed to be different (that is, the presence or absence of *stratification*).

Figure 12: A Comparison of Key Characteristics of Variance–Correction Models

Model Property	Andersen-Gill (AG)	Marginal (WLW)	Conditional (PWP), Elapsed Time	Conditional (PWP), Gap Time
Risk Set for Event k at Time t	Independent Events	All Subjects that Haven't Experienced Event k at Time t	All Subjects that Have Experienced Event k - 1, and Haven't Experienced Event k, at Time t	
Time Scale	Duration Since Starting Observation	Duration Since Starting Observation	Duration Since Starting Observation	Duration Since Previous Event
Robust standard errors?	Yes	Yes	Yes	
Stratification by Event?	No	Yes	Yes	

Risk Sets

- This tells us whether events develop sequentially or simultaneously – that is, can an observation be "at risk" for the second event before they experience the first event?
- Sometimes, the latter makes sense (e.g. development of tumors, or coups).
- Other (most) times, it doesn't (marriages, wars, etc.).
- How this is defined is critical to the model used.

Time Scale

- Does the clock "start over" after each event, or not?
- The "counting process" approach generally assumes the answer is no; this is sometimes known as "elapsed time."
- There may be "gaps" in this time.
- In contrast, the "interevent time" approach starts the clock over.
- This also is, in the end, a substantive question.

Stratification

- Some models assume a common baseline hazard $h_0(t)$ for the first, second, etc. events.
- Other approaches allow for stratified analysis by events, where each event has its own baseline hazard.
- The latter is generally more flexible/general.

Various combinations of these different characteristics can be combined into the models we see, as follows.

AG (Andersen/Gill 1982)

This approach adopts the counting process formulation to the Cox model. It assumes:

- Independent events, so...
- A single baseline hazard

As a practical matter, this amounts to nothing more than a Cox model with robust / clustered standard errors. It is, in every respect, the simplest and most restrictive alternative.

WLW (Wei et al. 1989)

The Wei, Lin, and Weissfeld (1989) model is best thought of as a generalization of the competing risks idea. It assumes:

- That observations are "at risk" for all possible k events from the beginning of the study, so
- All durations are measured from the beginning of the study.
- Repeated/dependent events are dealt with through
  - Allowing separate baseline hazards (strata) for different events, and
  - Allowing for the possibility of strata-by-covariate interactions.
- Note as well that the data set-up is different than the others, requiring in effect a complete set of data for each of the k event ranks.

PWP (Prentice et al.) – Elapsed Time

This approach is like the AG model, except that

- Each observation is not at risk for event k until it has experienced event $k-1$, and
- Different events have different baseline hazards. Also,
- The model allows for strata-by-covariate interactions.

PWP - Interevent Time

This model is like PWP-ET, except that the the "clock starts over" after every event. In my opinion, this model probably best corresponds to most of the data we analyze.

Event Contagion and Parameter Stability

In all the marginal models, if we are interested in addressing the question of whether and to what extent the effects of covariates change over time, we do so through a straightforward interactive approach. There are at least three possibilities:

- The effect on the hazard of the covariate in question is a smooth (read: linear, or at least monotonic) function of the "event count." In that case, a standard multiplicative interaction will capture what is going on nicely.
- The effect on the hazard of the covariate in question varies non-linearly/monotonically across events. In that case, a "strata-by-covariate" interaction is the recommended approach.
- If the latter is the case for a large number of covariates, you may simply be better off estimating separate models for each event count (as we did in the *JOP* paper).

(A Quick & Dirty Example: Capability Imbalances and War Onset)

In the handout...

Practical Advice

As a practical matter, estimating these models is simply a function of:

- Setting up the data correctly (so as to define the right risk sets),
- Stratifying when appropriate, and
- Calculating / using robust standard errors...

This is all outlined in our paper, or in Cleves (1999) and Kelly and Lim (2000).