Advances in Duration Modeling:
The Split Population Duration Model

By Brandon Bartels

In this Statistics Corner, I will review an exciting and important development in duration modeling—the split population duration model—which is applicable to many important questions in political science. This model accounts for a particular type of heterogeneity across observations, namely, it relaxes the assumption that all subjects will eventually experience the event of interest. Before discussing this model, I present a brief overview of duration models for political science.

Duration Models in Political Science

Duration models allow researchers to tackle important theoretical questions concerning the timing of events or the survival of particular states of being. The study of *when* events occur is often at the center of political inquiry, and the importance of timing in politics takes us back to Fenno's contention: "If we are to explain outcomes, who decides *when* may be as important as who decides *what*" (1986, 9). Indeed, in the past decade, scholars have used duration models to address interesting questions of timing and survival, such as the dissolution of cabinets in parliamentary governments (King et al. 1990; Warwick 1992; Diermeier and Stevenson 1999), the duration of wars (Bennett and Stam 1996), the survival of political regimes (Bueno de Mesquita and Siverson 1995), the timing of state policy adoption (Berry and Berry 1990; Volden 2003), challenger entry in congressional elections (Box-Steffensmeier 1996), the timing of position-taking in Congress (Box-Steffensmeier et al. 1997; Caldeira and Zorn 2003), delay in Senate confirmations of presidential nominees (McCarty and Razaghian 1999; Binder and Maltzman 2002; Martinek et al. 2002; Shipan and Shannon 2003), and the survival of precedent in the U.S. federal courts (Spriggs and Hansford 2001; Benesh and Reddick 2002).

While interest in questions of timing unifies scholars using duration models, some differences remain within the community, primarily those centering on how to treat the *hazard rate*, or a subject's risk of experiencing the event of interest, given that it has not yet done so.[1] Parametric models specify explicitly the distribution of the hazard; examples of parametric distributions include the Weibull, exponential, and Gompertz, as well as accelerated failure time (AFT) formulations such as the log-logistic and log-normal. Parametric models, then, explicitly account for *duration dependence*, or the extent to which the risk of experiencing the event increases or decreases as a function of time. On the other hand, the Cox model—the semi-parametric approach—leaves the baseline hazard unparameterized, which means that the hazard is not constrained to possess a specific distributional form. Thus, the Cox model estimates the effects of independent variables on the timing of the event of interest, and leaves duration dependence unspecified, essentially treating it as a nuisance factor. The choice of parametric versus semi-parametric models is typically left to the researcher's confidence in the theory

[1] Functionally, the dependent variable in duration models is the time until the event of interest occurs. Conceptually, though, the dependent variable is actually the unobserved hazard rate.

underlying the data-generating process. If one possesses strong theory as to the distributional form of the hazard, then using a parametric model may be appropriate. However, if one is uncertain about the shape of the underlying hazard rate for the process under study—indeed, many scholars remain skeptical that social science theory is precise enough to justify a parametric model—then the Cox model provides added flexibility over parametric models. Readers with further interest in the details surrounding parametric versus semi-parametric models should consult more general duration modeling sources (e.g., Box-Steffensmeier and Jones 1997, 2003; Blossfeld and Rohwer 1995; Hosmer and Lemeshow 1998).

The Split Population Duration Model

Split population duration (SPD) models account for a specific type of heterogeneity, i.e., the possibility that some cases will never experience the event of interest while some will. One of the assumptions of standard duration models is that every observation in the data *will eventually experience the event of interest*, which is sometimes an unreasonable assumption in violation of a particular theory or understanding of the process under examination. Developed in biostatistics about 50 years ago and popularized in economics and sociology by Schmidt and Witte (1984, 1989), SPD models relax this assumption by essentially "splitting" the observations under analysis into two subpopulations, one that will eventually experience the event of interest and one that will never experience the event. In their study of criminal recidivism, Schmidt and Witte (1984, 1989) were interested in the factors explaining the timing of criminals returning to prison after being released; but we know that not all criminals return to prison after release, a factor incapable of being accounted for by a standard duration model. To solve this problem, Schmidt and Witte specified a model that generates two sets of simultaneously estimated coefficients: one for the likelihood of the event ever occurring and the other for the timing of the event, conditional on the event ever occurring. The mathematical derivation of the model, drawing on Schmidt and Witte (1989) and Box-Steffensmeier et al. (2003), is located at the web archive (see below for website information).

It is worth emphasizing a few important points about the SPD model. First, two sets of coefficients are estimated in SPD models: (1) coefficients for the effects of covariates on the *incidence* of the event occurring, and (2) coefficients for the effects of covariates on the *timing* of the event, *conditional on the probability the event occurring*. It is also important to underscore that the censoring indicator (i.e., whether or not we observe the event occur within the analysis time) serves as the dependent variable in the incidence portion of the model. Second, a very powerful feature of these models is that different covariates can be included to explain *whether* and *when* the event occurred. For example, an independent variable may have a positive effect on *whether* the event occurred and a negative effect on *when* it occurred. This makes the split population model much more flexible than other duration models where the effect of timing and incidence are combined. Third, the SPD model estimates a "split parameter," δ, which is the estimated mean probability of cases experiencing the event of interest. This statistic allows the analyst to test whether relaxing the assumption that every observation will experience the event of interest is necessary. If it is not, i.e., if $\delta = 1$, the SPD model collapses into a typical duration model. So there is little cost to estimating the split population model, and as such, scholars should almost always use the SPD model when they have reason to believe that not all observations will experience the event of interest. The estimated split also serves as a sort of goodness-of-fit statistic in that it can be compared to the proportion of cases that actually experienced the event of interest. Fourth, SPD models are currently only estimable using parametric approaches. For those who advocate the semi-parametric approach to duration modeling, this is certainly the downside of the SPD model. However, work continues to be done to estimate a Cox-type SPD model, although problems of model identification have hampered these attempts so far (see, e.g., Sy and Taylor 2000; Kuk and Chen 1992). Fifth, regarding software capabilities, LIMDEP is currently the only software that has a canned routine for estimating SPD models. In addition, Forster and Jones (2001) have written programs for estimating these models in Stata (see the web archive for an example).

Finally, an exciting and relatively new development includes the capability of SPD models to generate estimates for time-varying covariates (TVCs). For many scholars, one of the disappointments of the Schmidt and Witte model was that it could only estimate the effects of time-invariant covariates. However, Forster and Jones (2001), in an effort to study the effect of increases and decreases in taxes on the timing of people starting smoking, have developed an SPD model to accommodate TVCs, providing even more flexibility to the model.

Applications

Surprisingly, very few applications of the SPD model exist in political science. In fact, I am aware of only one published paper, by Clark and Regan (2003), using an SPD model. In that paper, the authors analyze the timing of interstate conflict, and they account for the notion that not all dyads have the "opportunities" to engage in war. Two other unpublished papers use SPD models to study the incidence and timing of events. First, Box-Steffensmeier et al. (2003), recognizing that early money in campaigns influences the success of campaigns, examine the factors that affect the timing of PAC contributions to incumbent U.S. House members. Using datasets of incumbent-PAC dyads for both labor and corporate PACs, the authors argue that the use of a standard duration model would be unreasonable, since it would assume that both labor and corporate PACs would eventually give to *every* incumbent. From the literature on PAC contributions, we know that there are certain Representatives who would never receive contributions from labor or corporate PACs. For instance, we should not expect labor unions to contribute to Rep. Peter Hoestra, who has investigated the Teamsters, or Cass Ballenger, who has tried to reform OSHA. Thus, Box-Steffensmeier et al. use the SPD model to examine the factors that explain both the incidence and timing of PAC contributions.

Another application of the SPD model is by Hettinger and Zorn (2001), who analyze the timing of congressional overrides of Supreme Court decisions. A standard duration model would assume that all Supreme Court cases will eventually be overturned by Congress, certainly an unrealistic and unreasonable assumption that defies much of what we know about the survival of the Court's decisions. Thus, Hettinger and Zorn specify an SPD model estimating the effects of independent variables on both the incidence and timing of congressional overrides. In interpreting the timing coefficients, the effects are conditional on the probability of the case ever being overturned. Also, the estimated split parameter of .09 means that the model predicts that 9% of the cases will actually be overturned, certainly justifying the use of the SPD model. The actual percentage of cases overturned in the sample is 7%, indicating that the model does a reasonably good job in "splitting" the population. Certainly, a standard duration model, which would imply a splitting parameter of 1.0, would produce incorrect estimates of the effects of variables on the timing of overrides (see Hettinger and Zorn for a comparison of the SPD model with a standard model).

Conclusion

Given its capability to incorporate important information about the likelihood of different cases experiencing the event of interest, the SPD model certainly represents an exciting advance in duration modeling. With the recent development of SPD models to include TVCs, the constraints of using the SPD model have been reduced even more. However, those who advocate the use of semi-parametric duration models will undoubtedly be hesitant to use the SPD model, which currently can only accommodate the parametric approach. But in general, scholars who have reason to believe that not all observations will experience the event of interest should almost always estimate an SPD model instead of a standard duration model.

Web Archive

To view the mathematical derivation of the SPD model, an SPD program written for Stata, and the references cited in this Statistics Corner, go to http://psweb.sbs.ohio-state.edu/grads/bartels/statscorner.htm