Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. http://www.jstatsoft.org/

Splitting it up: the spduration split-population duration regression package

Daniel W. Hill University of Georgia

Nils W. Metternich University College London

Andreas Beger Duke University

Michael D. Ward Duke University

Abstract

We present an implementation of split-population duration regression in the **spduration** R package and an application to data on military coups. The statistical model accounts for units that are immune to a certain outcome and not part of the duration process the researcher is primarily interested in. We provide insights that if immune units exist, we can significantly increase the predictive performance compared to standard duration model. The package includes estimation and forecasting methods for split-population Weibull and Loglogistic models.

Keywords: keywords, not capitalized, R.

1. Introduction

This template demonstrates some of the basic latex you'll need to know to create a JSS article.

1.1. Code formatting

Don't use markdown, instead use the more precise latex commands:

- Java
- **plyr**
- `print("abc")`

2. Introduction

Duration models are an important class of statistical estimators that take into account the duration dependency of outcomes we are interested in. For example, the risk of dying is not independent of time. Newborns are at a great risk of dying, but as they grow older this risk declines and then gradually starts to increase again after the age of 9-10. This risk can be increased or decreased by behavioral (smoking, exercise, diet, etc) and structural factors (health care, regulations, urban vs rural, etc), but there is an underlying risk that is time dependent and impacts on all units.

However, sometimes we are interested in situations in which not all units are at risk of dying, failing, or acquiring some new characteristic. Imagine we are interested in the risk of acquiring the flu. First, we might think that everyone is at risk of getting the flu and it just depends on individual behavior (e.g. good hygiene) and structural factors (e.g. workplace) whether individuals will acquire the flu. If this would be the case we could use a standard duration model and estimate how different factors impact on the baseline risk. But there might also be individuals who are immune to the flu, because they received a flu shot, had the same flu virus in the past year, or some other factors that makes it impossible for them to get sick. Given that this is the case we have two underlying populations: An at risk population and an immune one. If the immune population is relatively large, estimates using a standard duration model will be biased and predictions resulting from such a model inaccurate.

3. Immune populations and inference of duration processes

Duration models, where the variable to be modeled describes a distribution of times until some event of interest, were originally developed in health and demographic research, where they grew naturally from life tables and survival records for medical patients. Basic formulations of such models, like the parametric exponential or Weibull regressions or semi-paramaetric Cox regression, implicitly assume that all subjects or units under observation, including right-censored observations, will eventually experience the event of interest. This assumption is outright wrong or at least disadvantageous in many substantive areas and empirical applications, where there instead is a sub-population of units or subjects that will never experience an event, and thus are effectively "cured".

This insight was realized as early as 1949 by Boag (1949) and Berkson and Gage (1952), who were researching survival rates in cancer patients following treatment, and where quite obviously some fraction of patients survived because their cancer was cured, while others relapsed after apparent remission due to levels of disease below detectable thresholds. As with conventional duration models, their origin in health and medicine has shaped the terminology conventionally used with such model, e.g. split-population duration models are sometimes also referred to in such contexts as cure rate models, and the basic concepts like survival and failure rates reference the survival of humans.

Yet the intuition underlying split-population duration models has led to applications in a broad range of subject areas outside demographics and medicine. In an early and foundational application that reached beyond criminal science, Schmidt and Witte (1989) examined criminal recidivism using data on close to 10,000 prisoners from the North Carolina prison system in the late 1970's and early 1980's to identify factors that influence whether a criminal lapses at all, and if so which factors are related to the amount of time between prison stints. This work already includes a full formulation of the model with independent covariates for both the duration equation and the risk or cure equation, although only with subject-specific covariates rather than time-varying covariates with multiple data points per subject.

In public health, Douglas and Hariharan (1994) use data from the US to model factors related to the age at which smokers started the habit, and Forster and Jones (2001) examines the impact of tobacco taxes on smoking and quitting decisions. DeYoung (2003) models the failure of new commercial banks in the US during the 1980's. Lastly, in our own domain, political science, Svolik (2008) used a split-population duration framework to study democratic consolidation, i.e. whether democratic regimes persist or slide back to authoritarianism, and when. Building on this effort, split-population duration models have also been used, along with other models, to produce regular predictions for five different forms of political conflict for the Integrated Crisis and Early Warning System (ICEWS) project (M. D. Ward et al. 2013) and to model irregular leadership changes for the Political Instability Task Force (PITF; Beger, Dorff, and Ward 2014).

4. Model development

Conventional duration models assume that all subjects will eventually fail, if we can only observe them long enough. The likelihood for a data point with survival time t is thus the failure rate at that time or the probability of survival beyond t, depending on whether the observation is right-censored $(1 - \delta_i)$ or not:

$$\mathcal{L} = \prod_{i=1}^{N} (f(t_i))^{\delta_i} \times (S(t_i))^{1-\delta_i} \tag{1}$$

The major modeling question in this setting, which we will return to below, is the choice of a function to describe the evolution of the hazard rate $h(t) = \frac{f(t)}{S(t)}$ over time, e.g. with exponential, Weibull, or log-logistic densities.

The cumulative failure rate $(F(t) = 1 - S(t))$ over time converges to 1, meaning all subjects fail eventually. This assumption is untenable in many situations. Some cancer patients are cured after treatment while others relapse, most young people do not start smoking even though most smokers started at a young age, and many states will not experience the kind of violence that persists in some parts of the world. A better assumption in such contexts is that there is a subpopulation of subjects who are at risk of experiencing an event at some point, and another subpopulation who will never experience the event.

The presence of a large sub-population which is not at risk for an event in practice will inflate estimates of the survival fraction, and reduce hazard estimates for all subjects by essentially, and falsely, distributing risk among subjects that genuinely will fail and those that are cured. A model will thus over predict hazard for subjects that are not at risk (cured), and under predict for those who are at risk of eventually experiencing the event of interest.

We can incorporate the presence of a sub-population, where we label the subpopulation at risk with π, by rewriting the likelihood as:[1]

[1]Usual presentation of the split-population duration framework in medical contexts focus on the "cured" subpopulation. In our applications events are typically rare and it thus is easier to emphasize the "risk" subpopulation. As risk = 1− cured, this difference is trivial.

$$\mathcal{L}\{\theta|(t_1,\ldots,t_n)\} = \prod_{i=1}^{N} (\pi_i f(t_i))^{\delta_i} \times ((1-\pi_i) + \pi_i S(t_i))^{1-\delta_i} \tag{2}$$

Crucially, this split-population framework is primarily useful in contexts where sub-populations are not clearly or easily identifiable. Even though there is a clear sub-populations in a model of the age at first pregnancy for humans—men—no researcher in their right mind would think to include male subjects in their data at all, for example. On the other hand, whether a cancer patient is cured or not cured given that they have no visible signs of cancer following treatment or have hit the 5-year disease free survival mark is much less clear and in fact the question that initially led to the idea for split-population duration modeling (Boag 1949, Berkson and Gage (1952)).

Early efforts focused only on the cure rate $(1-\pi)$ and treated it as a constant, but we can model membership in the subpopulation with its own covariates through a logistic link function:

$$\pi_i = \frac{1}{1+e^{-z_i\gamma}} \tag{3}$$

Where z_i is a vector of covariates for a subject at a given time. For interpretation, it is important to note that with time-varying covariates, the risk (or cured) estimate for a subject is particular to a given time point rather than constant for that subject over all time periods in the spell.[2] Depending on the covariates, the risk estimate for a subject can thus fluctuate rapidly over time. To ease interpretation, it might be convenient to restrict covariates in the logit risk model to slow-moving, stable covariates in order to produce stable risk estimates for subjects.

The last component to complete the likelihood is the choice of a distribution for the shape of the hazard rate. The **spduration** package implements two hazard rate shapes, Weibull[3] and log-logistic:

$$\begin{aligned}
&\text{Weibull} \\
f(t) &= \alpha\lambda(\lambda t)^{\alpha-1} e^{-(\lambda t)^\alpha} \\
S(t) &= e^{-(\lambda t)^\alpha} \\
h(t) &= \alpha\lambda(\lambda t)^{\alpha-1} \\
&\text{Log-logistic} \\
f(t) &= \frac{\alpha\lambda(\lambda t)^{\alpha-1}}{(1+(\lambda t)^\alpha)^2} \\
S(t) &= \frac{1}{1+(\lambda t)^\alpha} \\
h(t) &= \frac{\alpha\lambda(\lambda t)^{\alpha-1}}{1+(\lambda t)^\alpha}
\end{aligned}$$

[2]We use "spell" to designate all time periods observed for a subject up to the failure time. Subjects can theoretically have multiple spells, e.g. cancer patients who go into remission and relapse more than once, or states that experience multiple civil war onsets over their history.

[3]In the code, $f(t) = \alpha\lambda^\alpha t^{\alpha-1} e^{-(\lambda t)^\alpha}$.

Where $\lambda = e^{-x_i\beta}$ is a parameter of covariates.

With these, the main quantity of interest is the conditional hazard $h(t, \pi)$, where both the risk/cure probabilities and hazard are conditional on survival to time t:

$$h(t,\pi) = \frac{f(t,\pi)}{S(t,\pi)} = \frac{\pi(t) \times f(t)}{(1-\pi(t)) + \pi(t) \times S(t)} \tag{4}$$

$$\pi(t) = \frac{1-\pi}{S(t) + (1-\pi)(1-S(t))} \tag{5}$$

For a given unconditional risk rate π, the probability that a particular case with survival time t is in the risk set decreases over time because an increasing number of surviving cases consist of immune or cured $(1-\pi)$ cases that will never fail. In the hazard rate, the failure rate in the numerator is conditional on the probability that a case is in the risk set, give survival up to time t, and the numerator is an adjusted survivor function that accounts for the fraction of cured cases by time t, which is $1 - \pi(t)$.

5. Fit a split-population model on coups data

The **spduration** package for R implements a split-population duration regression for data with time-varying covariates and with two options for the hazard rate shape, Weibull and log-logistic. The Weibull density allows for hazard rates that are increasing, constant, or decreasing over survival time, while the log-logistic density also can fit rates that have a peak at a particular survival time.

As an example to illustrate the package functionality, we will replicate and extend the model of coup onsets in Belkin and Schofer (2003). Belkin and Schofer's paper lends itself to re-analysis with a split-populuation duration model because they explicitly distinguish long-term structural risk factors for coups from more short-term triggering causes that can explain the timing of a coup in a regime at risk. They argue that many countries never experience coups because coups are effectvely impossible due to structural factors, while others that never experience coups are nevertheless at risk due to a different configuration of those same factors. Using language which fits nicely with the class of models described above, they say "Triggers are not the source of the original risk, and in the absence of structural causes, the presence of triggering factors alone cannot lead to a coup. Hence, triggers should not be equated with coup risk. Rather, they are factors that may determine the exact timing of a coup in regimes that suffer from high coup risk" (p. 598 Belkin and Schofer 2003)

Belkin and Schofer develop a measure of "structural coup risk" which incorporates factors such as the strength of democratic institutions and civil society, and a recent history of successful coups. They employ a logistic regression model which does not adequately capture the process described in the quote above, since it implicitly assumes that all observations are at risk for a coup (i.e., the probability of a coup is non-zero for all observations). Their structural coup risk indicator is developed precisely to distinguish between cases where coups are possible and cases where they are not, so the split-population model allows one to examine whether the indicator effectively separates at-risk cases from cases where coups are impossible.

We begin by loading the package and the Belkin and Schofer replication data, and formatting the data to add several variables needed by the split-population duration model.

```
library("spduration")
#data(bscoup)
library("foreign")
bscoup <- read.dta("data/BelkinSchoferTable4.dta")
bscoup$coup <- ifelse(bscoup$coup=="yes", 1, 0)
bscoup      <- add_duration(bscoup, "coup", unitID="countryid", tID="year",
                            freq="year")
```

```
Warning in attempt_date(data[, tID], freq): Converting to 'Date' class with
yyyy-06-30
```

The `add_duration` function takes as input a data frame with a binary response variable (in this example, "coup"") that measures the occurrence of the event, or failure, recorded over discrete time periods, and produces a data frame with a duration variable (named"duration") for each spell as well as a binary variable ("atrisk") that is constant across the spell and indicates whether the spell ended in failure. These are then used as response variables for the formula in the `spdur` function. The `spdur` function is the primary function in the package and produces a regression model object of class `spdur` which can then be used with further methods.

Though the data used in this example are recorded annually, the function supports annual, monthly, or daily data. We begin by fitting first a Weibull and then a log-logistic split-population duration model using the coups data, including the measure of coup risk in the logit (risk) equation.

Using the `summary` function on either model object will produce standard output showing the model formula, estimates for the duration and risk equations, and test statistics with p-values. One may also use the `AIC` and `BIC` function to calculate the information criterion statistics for `spdur` objects.

```
AIC(weib_model)
```

```
[1] 1207.992
```

```
AIC(loglog_model)
```

```
[1] 1110.829
```

```
BIC(weib_model)
```

```
[1] 1290.452
```

```
BIC(loglog_model)
```

```
[1] 1193.289
```

Table 1: Coup model with log-logistic hazard

Parameter	Estimate	StdErr	p
(Dur. Intercept)	3.30	0.14	0.00
milreg	-1.53	0.15	0.00
instab	-0.22	0.03	0.00
regconf	3.04	2.22	0.17
log(alpha)	-0.52	0.07	0.00
(Risk Intercept)	-14.61	9.02	0.11
couprisk	9.11	4.37	0.04
wealth	3.51	1.75	0.05
milreg.1	21.76	0.24	0.00
rwar	-6.37	5.38	0.24
regconf.1	-20.22	54.04	0.71
samerica	-3.31	3.28	0.31
camerica	-6.46	2.57	0.01

The objects produced by `spdur` can also be used with functions in the `xtable` package to produce LaTeX code for tables of results. The `xtable` function used on an `spdur` object will produce code such as that used to create Table 1.

The table shows estimates from the duration equation, beginning with the intercept, and then estimates from the risk equation. Interestingly, the coup risk variable has a rather large and statistically significant coefficient in the risk equation.

The package also includes a function, `plot_hazard`, that can be used with an `spdur` object to plot the estimated hazard rate. `plot_hazard` plots the conditional hazard, which is the probability of survival conditional on the covariates in the risk and duration equations, and conditional on survival up to time t. The function calculates the estimated hazard rate as well as 90% confidence intervals, which are produced by simulating values from the estimated sampling distributions of the model parameters. By default the function uses the mean values of the covariates during the simulations, but users can choose specific covariate values by entering them as vectors in the arguments `xvals` and `zvals`, which correspond to the covariates in the duration and risk equations, respectively. The command below creates the graph A in Figure 1.

```
plot(loglog_model, type="hazard", main = "A")
```

```
plot(loglog_model, type="hazard",
     xvals = c(1, 1, 10, 0.05),
     zvals = c(1, 7, 8.64, 1, 1, 0.05, 0, 0),
     main = "B")
```

Importantly, the package also includes functions to evaluate model predictions. The generic function `predict` can be used on an object of class `spdur` to generate several kinds of predictions, including the probability that an observation is "at-risk" and the probability of failure for a given time period. In addition, the generic `plot` function can be used on `spdur` objects

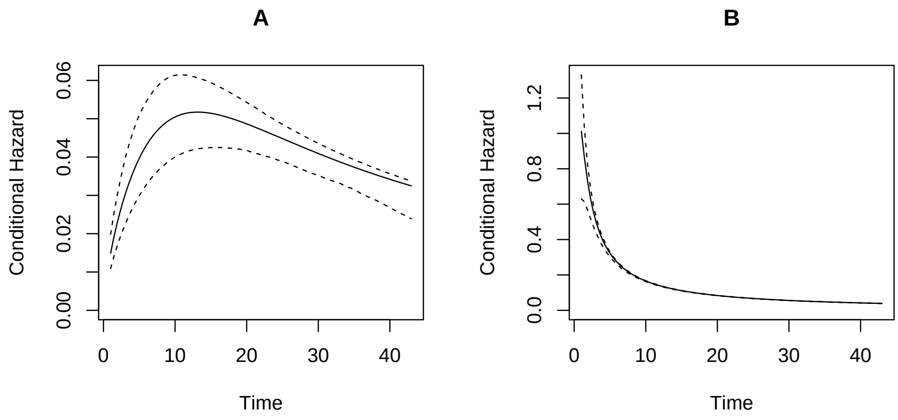


Figure 1: Plots of the hazard rate for the log-logistic model of coups. The left graph uses the default mean values for covariates, while graph B uses user-specified variable values for a high-risk military regime.

to produce a separation plot (Greenhill, Ward, and Sacks 2011), which is a graphical display for evaluating model predictions. The code below produces Figure 2.

```
plot(weib_model)
plot(loglog_model)
```

The option `endSpellOnly` is set to `FALSE` so that every observation, not only those at the end of a spell, is used in the plot. By default the `plot` function will calculate the conditional hazard for each observation. The separation plot sorts observations from left to right according to the predicted probability assigned by the model (higher values to the right), and shows each event/failure as red line, with non-events shown in beige. This makes it easy to see whether the model is assigning high probabilities of failure to actual cases of failure, and low probabilities to non-failures.

Finally, we demonstrate how to evaluate a model's out of sample predictions. We begin by splitting the data into training and test sets, using the last year of data (1999) as the test set and omitting incomplete observations from the test set.

```
data(bscoup)
bscoup$coup <- ifelse(bscoup$coup=="yes", 1, 0)
coup_train <- bscoup[bscoup$year < 1996, ]
coup_train <- add_duration(coup_train, "coup", unitID="countryid", tID="year",
                           freq="year")
```

```
Warning in attempt_date(data[, tID], freq): Converting to 'Date' class with
yyyy-06-30
```

```
coup_test  <- add_duration(bscoup, "coup", unitID="countryid", tID="year",
                            freq="year")
```

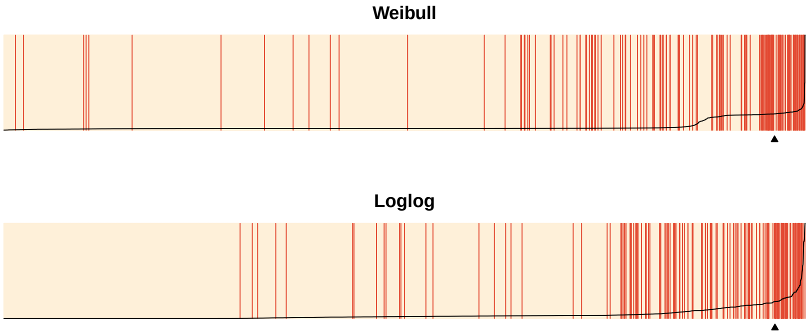


Figure 2: In-sample separation plots of Weibull and log-logistic model conditional hazard predictions

```
Warning in attempt_date(data[, tID], freq): Converting to 'Date' class with
yyyy-06-30
```

```
coup_test  <- coup_test[coup_test$year >= 1996, ]
```

The spdurCrisp function can then be used to fit a model on the training set and generate predictions for the test set.

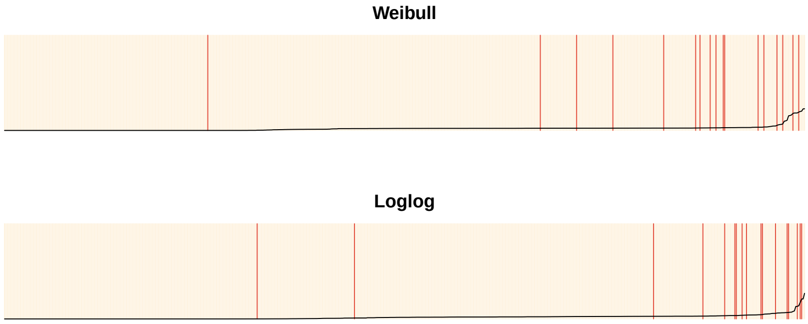
```
weib_model2   <- spdur(
  duration ~ milreg + instab + regconf,
  atrisk ~ couprisk + wealth + milreg + rwar + regconf + samerica + camerica,
  data = coup_train, silent = TRUE)

loglog_model2 <- spdur(
  duration ~ milreg + instab + regconf,
  atrisk ~ couprisk + wealth + milreg + rwar + regconf + samerica + camerica,
  data = coup_train, distr="loglog", silent = TRUE)
```

Foo

```
weib2_test_p   <- predict(weib_model2, newdata = coup_test)
loglog2_test_p <- predict(loglog_model2, newdata = coup_test)
```

The `train`,`test`, and `pred` arguments are used to specify data frames for the training set, test set, and forecasting set. We do not wish to generate forecasts here, so we simply specify the same data frame for the test and forecasting set. Here we use `spdurCrisp` to calculate the conditional risk, which is the default calculation for this function and is the probability that an observation is in the at-risk population. The test set predictions are stored in the model object as `test.p`, and we extract them to produce the plot shown below in Figure **??**, which was created using the `separationplot` package.

\begin{CodeChunk} \begin{CodeInput} library("separationplot")



obs_y <- coup_test$coup[complete.cases(coup_test)]

par(mfrow=c(2,1),mar=c(2,2,2,2)) separationplot(weib2_test_p, obs_y) separationplot(loglog2_test_p, obs_y) \end{CodeInput} \end{CodeChunk}

```
glm_model2    <- glm(
  coup ~ milreg + instab + regconf + couprisk + wealth + rwar + regconf +
    samerica + camerica,
  data = coup_train, family = "binomial"
)

c(Weibull = AIC(weib_model2),
  Loglog = AIC(loglog_model2),
  Logistic = AIC(glm_model2))

  Weibull    Loglog  Logistic
 990.5818  985.2633 1171.3556
```

6. Censoring considerations

Truncation and censoring are problematic for split-population duration models as they are for standard duration regression, but also pose some additional considerations. In left-truncation, we do not observe data for a spell prior to some date, and thus have incomplete and inaccurate values for the duration or time to failure for a spell. Since immune spells in the sample are over time going to distinguish themselves with exceptionally long survival times compared to spells at risk which fail periodically, left-censoring also makes it more difficult to distinguish the immune and at risk subpopulations.

Sometimes information about previous failures in the data is available beyond the time period over which covariates are observed, making it possible to ameliorate or eliminate left-censoring by using the information of previous failures when constructing the necessary duration variables with `add_duration()`.

Right-censoring, where spells end before outcomes are observed, also pose a unique problem in the split-population framework. Although right-censored spells themselves are accommodated in the modeling function, they impact the coding of at risk vs. immune spells. The add_duration() function retroactively codes all observations in a spell as at risk if the spell itself ended in failure. Right-censored spells are coded as immune over their entire duration. This can lead to some misclassification of observations as immune even though they experience failure at some point in the unobserved future.

Furthermore, in out-of-sample testing based on some kind of data partitioning scheme, this coding scheme can lead to unintentioned contamination of in-sample cases with knowledge of out-of-sample failures through the risk coding for failed spells. This leads to two recommendations. First, the duration data (add_duration) should not be built until after partitioning the data. Secondly, and as a result of the first point, data should be partitioned in a scheme that samples either by spell or block-wise, e.g. by witholding the last x years of data, in order to be able to proberly build the duration variables.

7. Conclusion

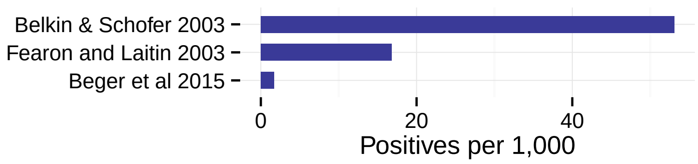


Figure 3: Rates of positive outcomes in select publications with binary outcomes

Many outcomes in the area that we work in, political science and political violence more specifically, are rare events. Coups, the onset of war and civil war, mass killings, and so on are exceedingly rare events when one considers that most countries most of the time are stable. Figure 3 shows a few examples for published research that models binary outcomes.

In the language of machine learning, we are dealing with highly imbalanced classes, although our data are also highly structured with clear dependence over space and time in countries and societies that may experience political violence. This is a well-recognized problem and has led to the development or use of several specialized mixture models like zero-inflated Poisson and negative binomial regression for count data, and a zero-inflated ordered probit for ordinal outcomes (Bagozzi et al. 2015). Split-population duration regression provides another principled solution to the challenges posed by data in this domain, but, unlike other solutions to the sparse outcome problem, also addresses underlying temporal dynamics.

Split-population duration models are not only appealing in this technical sense, but they also match the logic or intuituion many analysts of politics use when they distinguish long-term risk factors from more fleeting triggering causes. The example we have used, Belkin and Schofer (2003), is a particularly clear illustration of how well the language of theorists maps onto the model intuition.

References

Bagozzi, Benjamin E, Daniel W Hill, Will H Moore, and Bumba Mukherjee. 2015. "Modeling Two Types of Peace the Zero-Inflated Ordered Probit (ZiOP) Model in Conflict Research."

Journal of Conflict Resolution 59 (4). SAGE Publications: 728–52.

Beger, Andreas, Cassy L. Dorff, and Michael D. Ward. 2014. "Ensemble Forecasting of Irregular Leadership Change." *Research & Politics* 1 (3).

Belkin, Aaron, and Evan Schofer. 2003. "Toward a Structural Understanding of Coup Risk." *Journal of Conflict Resolution* 47 (5). Sage Publications: 594–620.

Berkson, Joseph, and Robert P Gage. 1952. "Survival Curve for Cancer Patients Following Treatment." *Journal of the American Statistical Association* 47 (259). Taylor & Francis Group: 501–15.

Boag, John W. 1949. "Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy." *Journal of the Royal Statistical Society. Series B (Methodological)* 11 (1). JSTOR: 15–53.

DeYoung, Robert. 2003. "The Failure of New Entrants in Commercial Banking Markets: A Split-Population Duration Analysis." *Review of Financial Economics* 12 (1). Elsevier: 7–33.

Douglas, Stratford, and Govind Hariharan. 1994. "The Hazard of Starting Smoking: Estimates from a Split Population Duration Model." *Journal of Health Economics* 13 (2). Elsevier: 213–30.

Forster, Martin, and Andrew M Jones. 2001. "The Role of Tobacco Taxes in Starting and Quitting Smoking: Duration Analysis of British Data." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 164 (3). Wiley Online Library: 517–47.

Greenhill, Brian, Michael D Ward, and Audrey Sacks. 2011. "The Separation Plot: A New Visual Method for Evaluating the Fit of Binary Models." *American Journal of Political Science* 55 (4). Wiley Online Library: 991–1002.

Schmidt, Peter, and Ann Dryden Witte. 1989. "Predicting Criminal Recidivism Using Split Population Survival Time Models." *Journal of Econometrics* 40 (1). Elsevier: 141–59.

Svolik, Milan. 2008. "Authoritarian Reversals and Democratic Consolidation." *American Political Science Review* 102 (02). Cambridge Univ Press: 153–68.

Ward, Michael D, Nils W Metternich, Cassy L Dorff, Max Gallop, Florian M Hollenbach, Anna Schultz, and Simon Weschle. 2013. "Learning from the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction." *International Studies Review* 15 (4). Wiley Online Library: 473–90.

Affiliation:

Daniel W. Hill
University of Georgia
First line Second line
E-mail: dwhill@uga.edu
URL: http://spia.uga.edu/directory/faculty_staff/hill-daniel

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:*
MMMMMM YYYY *Accepted:* yyyy-mm-dd