Appendices to Change Comes with Time: Substantive Interpretation of Nonproprotional Hazards in Event History Analysis.

Amanda A. Licht, PhD

Assistant Professor

Department of Political Science

University of South Carolina

3rd Floor Gambrell Hall

817 Henderson Street

Columbia, SC 29208

e: aalicht@gmail.com

www: http://people.cas.sc.edu/licht/

July 2010

Appendix I: Monte Carlo Experiment to Verify Appropriateness of Exponentiating Combined Coefficient Confidence Bounds

In order to verify that accurate confidence bounds can be located for relative hazards by exponentiating those of a combined coefficient, I demonstrate below with table A1 that the distribution of the combined coefficient and relative hazard will correspond to eachother at the same values of the modifying variable and estimated coefficients. For this table, I drew a random sample of 50 $\widehat{\boldsymbol{\beta}}$ assuming a joint normal distribution with mean vector $[.5 \quad -.2]$ and variance of $\begin{bmatrix} .04 & -.001 \\ -.001 & .02 \end{bmatrix}$. To further demonstrate the appropriateness of this method, I used the same distribution to draw 1,000 $\widehat{\boldsymbol{\beta}}$ for each of 10 values of the modifying variable x_2. I utilized this simulated sampling distribution to compare the confidence intervals derived from a sampling technique to those from the exponentiation of the combined coefficient. Figure A1 displays the result of this test. As the reader can see, there is a slight, though not statistically significant, difference between the median calculated relative hazard and the exponentiated mean of the combined coefficient. The confidence intervals from the two methods, however, match up exactly when a sufficiently large number of draws is taken from the distribution of the coefficient estimates.

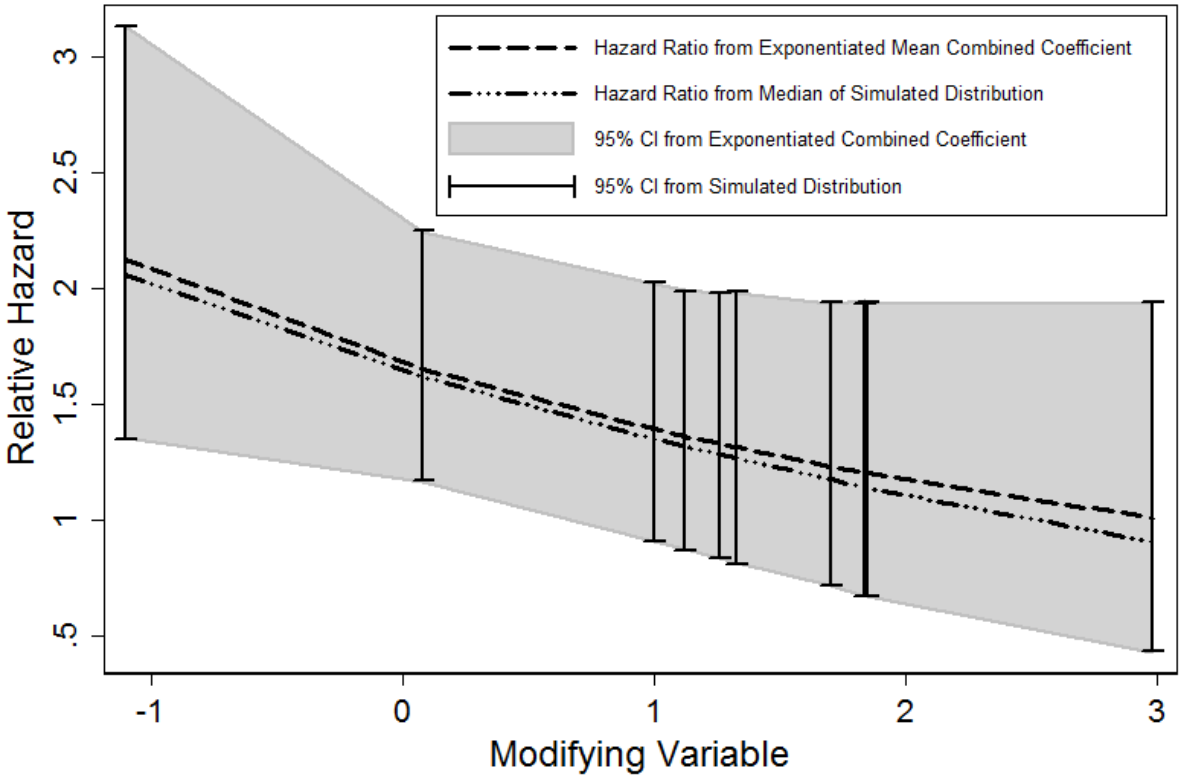
Table A1. Simulated Distribution of Combined Coefficient and Relative Hazard

Obs.	x_2	$\hat{\beta}_1$	$\hat{\beta}_2$	Combined Coefficient $(\hat{\beta}_1 + \hat{\beta}_2 x_2)$	Combined Coefficient Percentile	Relative Hazard $(e^{\widehat{\beta}_1 + \widehat{\beta}_2 x_2})$	Relative Hazard Percentile
1	1.75	0.48	-0.48	-0.36	2	0.69	2
2	1.75	0.21	-0.32	-0.34	4	0.71	4
3	1.75	0.36	-0.39	-0.32	6	0.73	6
4	1.75	0.45	-0.43	-0.30	8	0.74	8
5	1.75	0.07	-0.19	-0.26	10	0.77	10
6	1.75	0.14	-0.22	-0.24	12	0.78	12
7	1.75	0.41	-0.32	-0.16	14	0.85	14
8	1.75	0.48	-0.36	-0.15	16	0.86	16
9	1.75	0.47	-0.32	-0.09	18	0.91	18
10	1.75	0.58	-0.38	-0.09	20	0.92	20
11	1.75	0.42	-0.29	-0.08	22	0.92	22
12	1.75	0.31	-0.22	-0.08	24	0.92	24
13	1.75	0.60	-0.39	-0.08	26	0.93	26
14	1.75	0.76	-0.47	-0.07	28	0.93	28
15	1.75	0.16	-0.13	-0.06	30	0.94	30
16	1.75	0.58	-0.36	-0.05	32	0.95	32
17	1.75	0.35	-0.22	-0.04	34	0.96	34
18	1.75	0.24	-0.16	-0.04	36	0.96	36
19	1.75	0.72	-0.44	-0.04	38	0.96	38
20	1.75	0.67	-0.41	-0.04	40	0.96	40
21	1.75	0.40	-0.20	0.05	42	1.05	42
22	1.75	0.40	-0.20	0.06	44	1.06	44
23	1.75	0.29	-0.13	0.07	46	1.07	46
24	1.75	0.53	-0.26	0.07	48	1.07	48
25	1.75	0.82	-0.42	0.08	50	1.08	50
26	1.75	0.52	-0.23	0.11	52	1.12	52
27	1.75	0.43	-0.18	0.11	54	1.12	54
28	1.75	0.57	-0.24	0.15	56	1.16	56
29	1.75	0.35	-0.10	0.17	58	1.18	58
30	1.75	0.72	-0.30	0.19	60	1.21	60
31	1.75	0.65	-0.25	0.21	62	1.23	62
32	1.75	0.92	-0.40	0.22	64	1.25	64
33	1.75	0.41	-0.08	0.27	66	1.30	66
34	1.75	0.72	-0.25	0.28	68	1.32	68
35	1.75	0.56	-0.16	0.29	70	1.33	70
36	1.75	0.63	-0.19	0.29	72	1.34	72
37	1.75	0.30	0.00	0.31	74	1.36	74
38	1.75	0.56	-0.14	0.31	76	1.36	76

39	1.75	0.61	-0.16	0.33	78	1.39	78
40	1.75	0.50	-0.09	0.34	80	1.40	80
41	1.75	0.54	-0.11	0.34	82	1.41	82
42	1.75	0.74	-0.17	0.43	84	1.54	84
43	1.75	0.74	-0.17	0.44	86	1.56	86
44	1.75	0.50	-0.03	0.45	88	1.56	88
45	1.75	0.63	-0.02	0.60	90	1.82	90
46	1.75	0.51	0.07	0.63	92	1.88	92
47	1.75	0.80	-0.09	0.63	94	1.88	94
48	1.75	0.80	-0.05	0.72	96	2.05	96
49	1.75	0.74	-0.01	0.73	98	2.07	98
50	1.75	0.82	-0.03	0.76	100	2.14	100

NOTE: Draw of 50 $\widehat{\boldsymbol{\beta}}$ from hypothetical variance-covariance matrix of [.04, -.001\-.001, .02] with mean vector [.5, -.2].

Figure A1. Comparison of Methods for Obtaining Confidence Intervals around Relative Hazard



NOTE: Simulated data of 10,000 draws of beta1 and beta1 from mean vector of [.5, -.2] and covariance matrix of [.04, -.001\-.001, .02]

Appendix II: Sample STATA Code

The code below provides generic examples for producing the measures described in the text: the combined coefficient, the relative hazard, the hazard ratio and the first difference. For combined coefficients and relative hazards, simulation procedures are unnecessary as the simple formula for variance of summed random variables applies.

For the simulation procedures, when values of survival time are very numerous, it may be desirable to select key values rather than making the calculation at every point in time as the time required for the simulation procedure increases dramatically with the size of the dataset being used. An interesting alternative option would be to create a variable which tags the percentiles of time. The original data can then be collapsed to the mean values of time within each percentile. This would be accomplished easily with the following Stata commands:

```
xtile time = _t, n(100)
collapse (mean) _t, by(time)
```

Aside from saving time, this approach will also preserve the distribution of survival times, which will allow the analyst to array the resultant measures of substantive effect against the density of actual observations. Some nuance will be lost by calculating the effect at only key points, and the resulting plots of the statistics will not be smooth.

In the code, carrot brackets (< >) indicate items which should be filled in by the programmer. These include variable names and numerical values particular to the analysis at hand. For example, if the NPH-violating effect stems from the variable *age* and a new variable for the interaction with time is named *ageXlntime*, then the item `_b[<X>]` below should be filled in `_b[age]` and the item `_b[<X*ln(_t)>]` should be filled in `_b[ageXlntime]`. The items `V[<#,#>]` refer to elements in the variance-covariance matrix, where the first # should be filled in with the row number and the second # should be filled in with the column number.

```
/****************************/
/*** Combined Coefficient ***/
/****************************/

/*Following estimation of the NPH model, generate combined coefficient and standard
errors using parameters from the matrices stored in e(b) and e(V)*/

mat list e(b)
mat def V = e(V)
mat list V                /*identify position of variance of beta for
                    constitutive and interaction terms*/


/*generate combined coef*/

      gen combcoef = _b[<X>] + ln(_t)*_b[<X*ln(_t)>]


/*calculate standard error of combined coef*/

      gen se_combcoef = sqrt(V[<#,#>] + (ln(_t))^2*V[<#,#>] +2*ln(_t)*V[<#,#>])


/*calculate 95% confidence intervals*/

      gen combcoef_lo = combcoef - 1.96*se_combcoef

      gen combcoef_hi = combcoef + 1.96*se_combcoef


/*chart effect with confidence bounds*/

      twoway line combcoef combcoef_lo combcoef_hi _t, sort

/*compare to density of failure times*/

      twoway line combcoef combcoef_lo combcoef_hi _t, sort||kdensity _t



/****************************************/
/*** Relative Hazard for Binary Variable***/
/****************************************/

/*Following procedure to calculate combined coefficient, exponentiate to create the
relative hazard*/


/*calculate relative hazard*/

      gen relhaz = exp(combcoef)


/*calculate 95% confidence intervals*/

      gen relhaz_lo = exp(combcoef_lo)

      gen relhaz_hi = exp(combcoef_hi)


/*chart effect with confidence bounds*/

      twoway line relhaz relhaz_lo relhaz_hi _t, sort
```

```
/*********************************************/
/*** Relative Hazard for Continuous Variable ***/
/*********************************************/


/*Following estimation of NPH Cox model, preserve coefficient matrix and covariance
matrix stored in e(b) and e(V).  Generate dataset of values of time.*/

mat list e(b)
mat def b = e(b)

mat list e(V)
mat def V = e(V)                  /*Preserve matrices*/


/*Generate indicator for percentiles of time*/

       xtile time = _t, n(100)


/*Collapse to mean values within percentiles of time*/

       collapse (mean) _t, by(time)


/*Create 1000 observations at each value of time*/

       expand 1000


/*Generate natural log of time*/

       gen lnt = ln(_t)


/*Save dataset with values of time for use in simulation*/

       sum
       save times.dta, replace


/*** The program below performs the following steps
1. Opens dataset of values of time,
2. Draws a vector of coefficients from the sampling distribution N~(e(b), e(V))
3. Calculates the combined coefficient and relative hazard
4. Collapse to mean and standard deviation of combined coefficient
5. Append to dataset of results
It will be repeated multiple times, using different values of the PHA-violating
variable.  After the program, use statsby to get confidence intervals***/


program define intNPHvar

       syntax, num(real)

              use times.dta, clear

              drawnorm b1-b#, means(b) cov(V)

              gen combcoef = `num'*b# + `num'*lnt*b#

              gen relhaz = exp(combcoef)
              gen <var> = `num'
```

```
              append using results.dta

              save results.dta, replace

       end


clear
save results.dta, replace emptyok       /*Save empty dataset for results*/


/*Run loop of program at values of PHA-violating from <min> to <max> at intervals
<delta>*/

set more off
forvalues num = <min>(<delta)<max>      {

       simulate, reps(#): intNPHvar, num(`num')
                                        }

use results.dta, clear     /*Open dataset of results*/


statsby relhaz_lo=r(r1) relhaz_median=r(r2) relhaz_hi=r(r3), by(<var> lnt)
       saving(CIs_relhaz.dta, replace): _pctile relhaz, p(2.5, 50, 97.5);
#delimit cr



use CIs_relhaz.dta, clear


/*chart effect with confidence bounds*/

       twoway line relhaz relhaz_lo relhaz_hi _t, sort




/*****************************************************************/
/*** First Differences and Hazard Ratios for Continuous Variable ***/
/*****************************************************************/


/*Following estimate of NPH Cox model, preserve coefficient matrix and covariance
matrix.  Generate variables to hold values of interest for the PHA-violating
variable*/

mat def b=e(b)

mat list V
mat def V=e(V)

/*summarize PHA-violating variable*/

       sum <var>

/*generate variables from values of PHA-violating variables stored in memory*/

       gen max = r(max)

       gen mean = r(mean)
```

```
       gen min = r(min)

       gen sd = r(sd)

       gen 2sd= mean+2*sd


/*generate variable to tag percentiles of time*/

       xtile time = _t, n(100)

/*collapse to mean values within units of time*/

       collapse (mean) mean max min sd 2sd _t, by(time)

       sum

       save times.dta, replace          /*save dataset for use in simulations*/



/*** The program below performs the following steps
1. Opens dataset of values of time,
2. Draws a vector of coefficients from the sampling distribution N~(e(b), e(V))
3. Calculate the first difference and/or hazard ratio
5. Append to dataset of results
It will be repeated multiple times, using different values of the PHA-violating
variable. Then the <statsby> command can be used to get CIs based on the simulated
distribution. ***/

       program define firstdiff,

              syntax, hi(string) lo(string) term(string)


              use times.dta, clear

              drawnorm b1-b#, means(b) cov(V)

              gen hr = exp((`hi'-`lo')*(b# + ln(_t)*b#))

              gen D = exp((`hi' - `lo')*(b# + ln(_t)*b#)) - 1

              gen D_pct = D*100

              gen delta = "`hi'_`lo'"

              append using results.dta

              save results.dta, replace

              end


clear

save results.dta, emptyok replace       /*save empty dataset for results*/


/*run loop to perform program at multiple differences*/

set more off
```

```
foreach hi of varlist max 2sd  mean     {

       foreach lo of varlist mean min    {

simulate, reps(10): firstdiff, hi(`hi') lo(`lo')
                                        }
                                        }


/***The loop below opens up the dataset of results and performs _pctile across the
range of time for each value of delta, producing the 95% CIs***/


#delimit ;
foreach var of varlist  hr D D_pct      {;
use results.dta, clear;

statsby `var'_lo=r(r1) `var'_median=r(r2) `var'_hi=r(r3), by(delta lnt)
       saving(CIs_`var'.dta, replace): _pctile `var', p(2.5, 50, 97.5);
#delimit cr
                                        }



local vars  "D D_pct"            /*define local containing names of all but
                            first variable*/

use CIs_hr.dta.dta, clear        /*open dataset of CIs for first variable*/


/*Run loop to append all remaining datasets to make a master set */

foreach var of local vars        {

append using CIs_`var'.dta
                                 }


/*Chart differences or hazard ratios with CIs*/

twoway line hr_lo hr_hi hr_median lnt if delta==”<>”, sort


twoway line D_lo D_hi D_median lnt if delta==”<>”, sort
```