The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–9

Graphing the Effects of Variables that Violate the Nonproportional Hazards Assumption of the Cox Model

Michael Tiernay
New York University
michael.tiernay@nyu.edu

Abstract. This Program implements the suggestions of Licht (2011) for interpreting the effect of variables that violate the proportional hazards assumption of the Cox model. The standard correction for the problem of nonproportional hazards is the inclusion of time interactions with the violating variables. Including time interactions, however, makes interpreting the substantive effect of variables more difficult. With the program 'nonph_graph', I provide an easy to use solution to the problem of nonproportional hazards that graphs the effects over time.

Keywords: Not sure what keyword(s) to use

1 Introduction

Cox proportional hazards models are an increasingly common tool in the social sciences, yet many applications do not properly interpret variables that violate the model's proportional hazards assumption (Licht (2011), Keele (2010)). Following Box-Steffensmeier and Jones (2004), "if evidence of potential nonproportional hazards is found...one would estimate the event history model with the addition of an interaction effect between the offending covariate and some function of time. The most straightforward example of a model with the interaction term is to include the new covariate x_2, where $x_2 = x_1 * log(t)$, and x_1 is a covariate already in the model." (p. 136)

Licht (2011), however, notes that "correcting for NPH through the inclusion of time interactions for variables in violation of the proportional hazards assumption, however, is more than a quick statistical fix; it complicates the interpretation of statistical results and calls for more advanced postestimation techniques."(p. 227) The program introduced here, 'nonph_graph', implements four strategies for interpreting the effects of nonproportional hazards: (1) the combined coefficient; (2) the relative hazard; (3) the hazard ratio; and, (4) a first differences approach.[1]

2 Interpreting Estimates

The Cox proportional hazards model estimates the effect that covariates have on the hazard rate, $h_i(t)$, which is the instantaneous probability of failure at a specific time t,

[1] Note that these names are taken from Licht (2011) to provide consistency.

given that it has survived until then. The Cox model is written as:

$$h_i(t) = h_0(t)exp(\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik})$$

where $h_0(t)$ is the baseline hazard rate. The Cox model is semi-parametric because $h_0(t)$ is not estimated, meaning it can take any form. Assume, for example, that the variable x_1 violates the proportional hazards assumption. The model should then be re-estimated as:

$$h_i(t) = h_0(t)exp(\beta_1 x_{i1} + \beta_2 x_{i1} * ln(t) + \beta_3 x_{i2} + ... + \beta_k x_{ik})$$

where STATA's 'stcox' command estimates the coefficients. The four strategies for interpreting effects of x_1 are discussed below.

2.1 Combined Coefficient

The combined coefficient estimates:

$$\beta_1 + \beta_2 * ln(t)$$

This relates the contribution to the hazard rate over time when a variable changes from zero to one. Licht (2011) does not recommend using the combined coefficient because of the lack of substantive interpretation, but sometimes the graphs presented below are unsightly because the standard errors are so large, which makes the graph unreadable. The combined coefficient allows the analyst to asses a variable's statistical significance over time, but not the size of the effect.

2.2 Relative Hazard

The relative hazard estimates:

$$\frac{h_i(t)}{h_j(t)} = e^{X_i(\beta_1 + \beta_2 * ln(t))}$$

where X_i is a number set by the analyst. In the case of a binary variable, X_i is set to one, so the relative hazard is the change in the hazard rate when the variable goes from zero to one. For a continuous variable, X_i can still be set to one, and the relative hazard will still represent the change in the hazard rate when the variable goes from zero to one. If X_i is set such that $X_i \neq 0$, then the relative hazard represents the the change in the hazard rate when the variable goes from zero to X_i.

2.3 Hazard Ratio

The hazard ratio estimates:

$$\frac{h_i(t)}{h_j(t)} = e^{(X_i - X_j)(\beta_1 + \beta_2 * ln(t))}$$

where the difference between the relative hazard and the hazard ratio is the presence of X_j. In the relative hazard, X_j was set such that $X_j = 0$. Thus, the hazard ratio is a slightly more general specification for continuous variables that lets the analyst set the specific change in X.

2.4 First Differences

The first differences model estimates:

$$\%\Delta h_i(t) = 100 * (e^{(X_i - X_j)(\beta_1 + \beta_2 * ln(t))} - 1)$$

The first differences model displays the percent change in the hazard rate given the change in X, as opposed to a change in the ratio of the hazard rates.

The choice of model between the relative hazard, hazard ratio, or first differences depends on the substantive interest of the analyst. See Licht (2011) for more details on the differences between the models.

3 Using the 'nonph_graph' Command

3.1 STATA Syntax

```
nonph_binary graph ci low high
```

The nonph_binary command takes up to four arguments, described below (values the arguments can take are in parentheses).

- *graph* tells STATA what type of graph to create (i.e. combined_coefficient; hazard_ratio; first_difference)
- *ci* is the desired confidence interval for the graph ($0 < ci < 100$)
- *low* corresponds to X_j in the formulas above ($-\infty < low < high < \infty$)
- *high* corresponds to X_i in the formulas above ($-\infty < low < high < \infty$)

I will discuss below how to produce the desired graphs. However, two points are crucial. First, the 'nonph_binary' command must be run immediately following the 'stcox' command. This 'stcox' command must be written in a certain way for the 'nonph_binary' command to function properly. In particular, if the analyst wishes to produce a graph of a variable, the variable must be the first variable specified in the model. Additionally, the variable's interaction with the log of time must be the second variable specified in the model. For example, if x_1 is the variable of interest, the preceding Cox model must have the following syntax:

*stcox x1 x1*ln(t) x2 ...*

Second, note that the 'ci' argument provides the confidence interval for the graph. For example, writing 95 for the 'ci' argument will produce a graph with a 95% confidence interval, and writing 90 for the 'ci' argument will produce a graph with a 90% confidence interval. If no argument is specified for 'ci', a 95% confidence interval is used as the default. As discussed below, when 'low' and 'high' are not specified, default values of 0 and 1 are used.

Use the following data to produce the graphs below:

```
.webuse stan3, clear
.gen t_surgery = ln(_t)*surgery
.gen t_age = ln(_t)*age
```

3.2 Combined Coefficient

To produce the combined coefficient of the binary variable 'surgery':

```
.stcox surgery t_surgery age posttran year
.nonph_graph combined_coefficient
```

Or, with a 90% confidence interval (See Figure 1):

```
.stcox surgery t_surgery age posttran year
.nonph_graph combined_coefficient 90
```

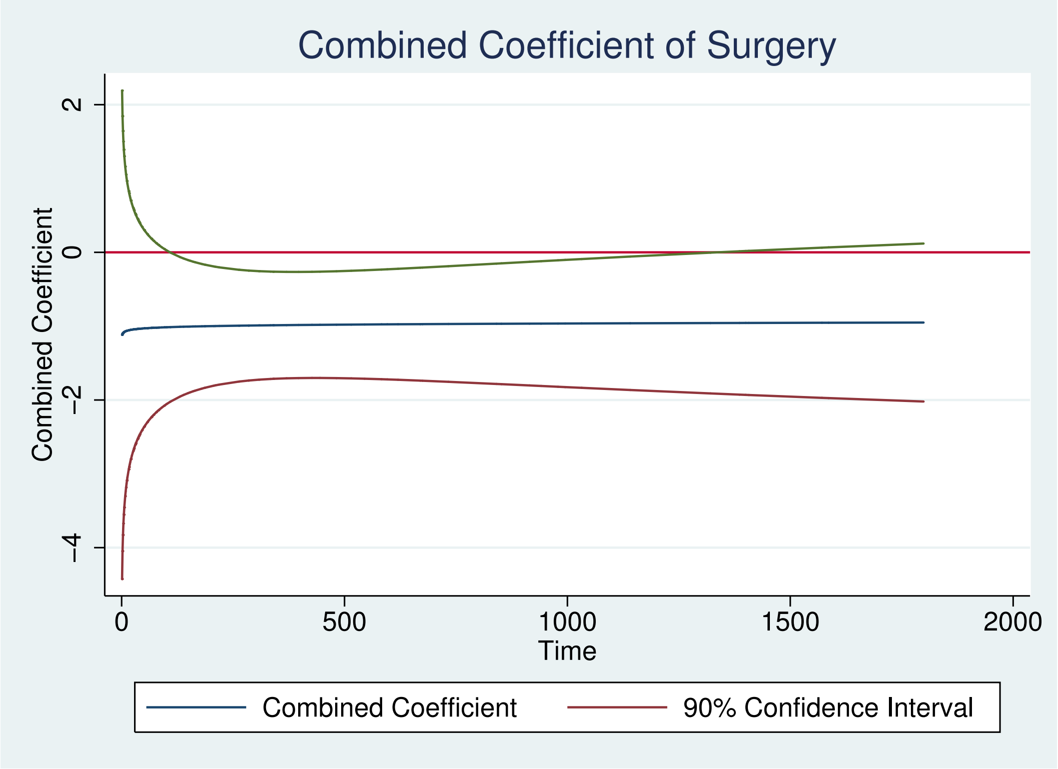

Figure 1: Combined Coefficient of Surgery with a 90% Confidence Interval

3.3 Relative Hazard

To produce the relative hazard of the variable 'surgery' with a 90% confidence interval (See Figure 2):

```
.stcox surgery t_surgery age posttran year
.nonph_graph hazard_ratio 90
```

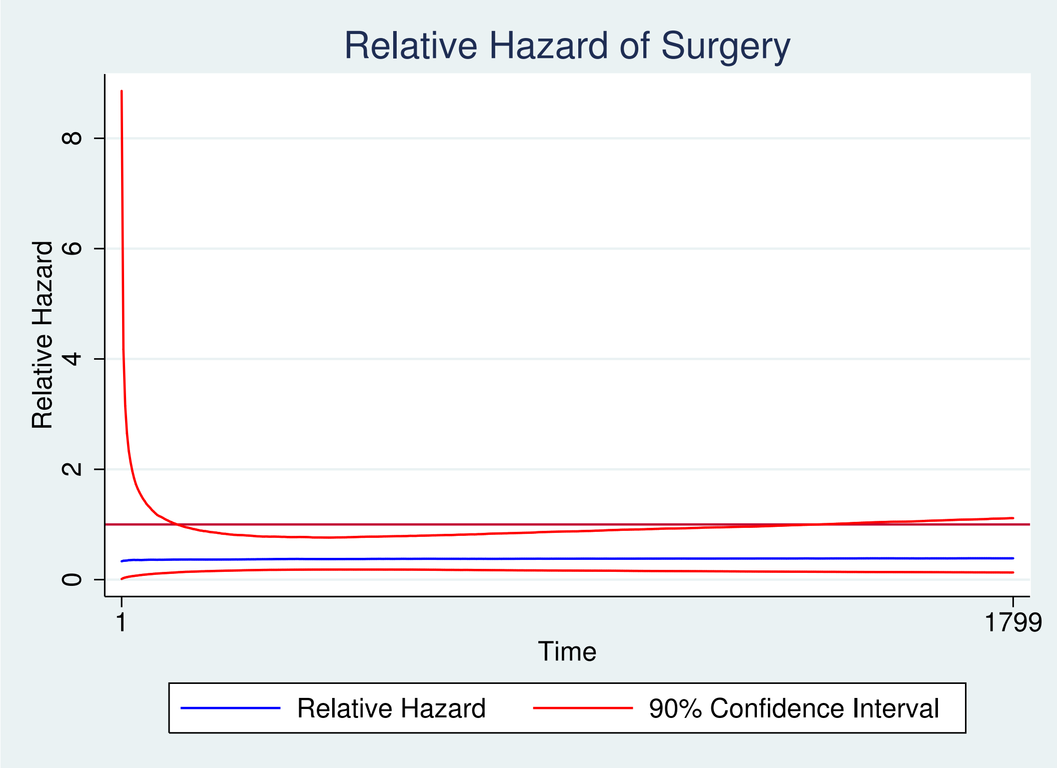

Figure 2: Relative Hazard of Surgery with a 90% Confidence Interval

Specifying the 'hazard_ratio' argument without the 'low' or 'high' arguments assumes that 'low' takes the default value of 0 and 'high' takes the default value of 1. This corresponds to the 'relative hazard' in Licht (2011).

To produce the relative hazard of the continuous variable ‘age’:

```
.stcox age t_age surgery posttran year
.nonph_graph hazard_ratio
```

This plots the effect of changing age from 0 to 1. Note that this change may not make intuitive sense, so the analyst may want to specify a specific range using the ‘hazard_ratio’ argument. To produce the hazard ratio as age changes from 40 to 43 with a 94% confidence interval (See Figure 3):

```
.stcox age t_age surgery posttran year
.nonph_graph hazard_ratio 94 40 43
```

Note that when specifying the ‘low’ and ‘high’ arguments, one must also specify the confidence interval.

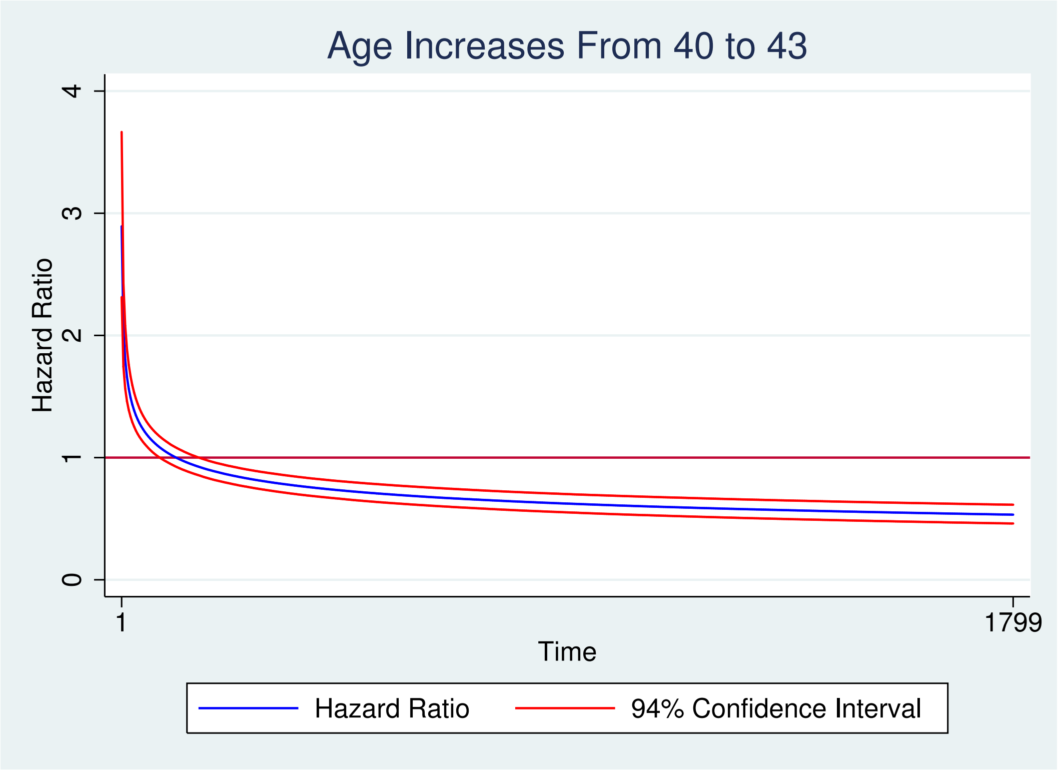

Figure 3: Hazard Ratio of Age with a 95% Confidence Interval

Finally, to graph the first differences as age changes from 40 to 44 with a 99% confidence interval (See Figure 4):

```
.stcox age t_age surgery posttran year
.nonph_graph first_difference 99 40 44
```

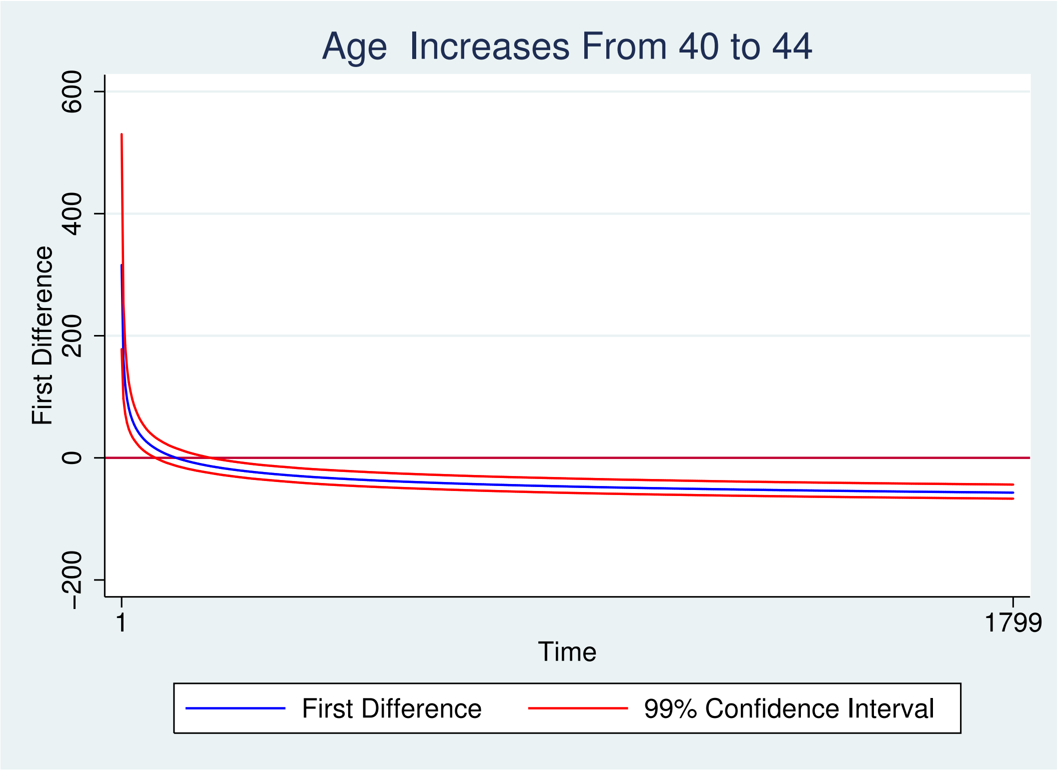

Figure 4: First Difference of Age with a 95% Confidence Interval

4 References

Box-Steffensmeier, J., and B. Jones. 2004. *Event history modeling: A guide for social scientists*. Cambridge University Press.

Keele, L. 2010. Proportionally difficult: testing for nonproportional hazards in Cox models. *Political Analysis* 18(2): 189–205.

Licht, A. 2011. Change comes with time: Substantive interpretation of nonproportional hazards in event history analysis. *Political Analysis* 19(2): 227–243.