Interpretation and parametrization of interaction effects in `Stata` (version 11 or later)

Caroline Weibull, Anna Johansson, Sandra Eloranta, Paul W. Dickman
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden

November 7, 2012

Introduction

With `Stata` version 11 came a new and improved way of handling factor variables. For details on all these updates, type `help factor variables` in the `Stata` command window. One of these improvements was the way `Stata` codes for interactions between factor variables. Previously `Stata` has coded interaction effects in the same way as other statistical software (e.g. SAS, R), by including the main effects (e.g. A and B) plus a product between the two ($A * B$) which is the interaction effect. This is still possible in version 11 and later, but now we can also use single (#) or double hashes (##) and get `Stata` to output specific contrasts of interest.
That is, to fit e.g. a linear regression model with outcome `Y` and exposures `A`, `B` and an interaction term between the two using the old syntax, we can write

```
. xi: regress Y i.A*i.B
```

which is equivalent to

```
. xi: regress Y i.A i.B i.A*i.B
```

Using the updated syntax we would instead write

```
. regress Y i.A##i.B
```

which is then equivalent to

```
. regress Y i.A i.B i.A#i.B
```

One important note is while a star always has the same interpretation, this is not always the case for double versus single hashes. E.g. `regress Y i.A##i.B` is not the same parametrization as `regress Y i.A i.A#i.B`. We will get back to this later.

This tutorial has two aims. Firstly, we will provide a hands-on example on how to interpret interaction effects models (irrespective of software used). Secondly, we will illustrate how to use the new syntax in `Stata` to estimate any contrast of interest (as opposed to those that are reported by default). In **Appendix A** you can find the motivation for the interpretations and calculations, explained in algebraic terms.

Example data

Assume we are running a survival analysis on a set of patients diagnosed with malignant melanoma, and that our outcome is all-cause mortality. Our exposure of interest is stage. We also have information on sex for our study participants. Thus we have:

$$\textbf{Stage} = \begin{cases} 1 & \text{if distant} \\ 0 & \text{if localized} \end{cases}$$

$$\textbf{Sex} = \begin{cases} 1 & \text{if male} \\ 2 & \text{if female} \end{cases}$$

We will model the mortality rate using Cox regression.

The main effects model

Let's start by looking at the main effects model and see what it estimates. We use the syntax

```
. stcox i.stage i.sex
```

and get the following output from `Stata`:

```
Cox regression -- Breslow method for ties

No. of subjects =         5794                     Number of obs    =      5794
No. of failures =         1421
Time at risk    =        39540
                                                   LR chi2(2)       =   1143.58
Log likelihood  =   -11275.135                     Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.stage |   11.55991   .7236842    39.10   0.000     10.22508    13.06899
       2.sex |    .755299   .0408707    -5.19   0.000     .6792955    .8398061
------------------------------------------------------------------------------
```

Here we estimate the effect of stage and sex on all-cause mortality, assuming that the effect size of each covariate is the same across all levels of the other covariate. So for example, the effect of stage, HR= 11.6, is assumed the same among both males and females. We can display the output in a table:

		Stage	
		0	1
Sex	1	1.00 *(ref.)*	11.6
	2	0.76	8.82

Table 1: Results from the main effects model.

In the right bottom cell we see that the effect of being female and having a distant stage as compared to being male and having localized stage is equal to 8.82. The estimated effect comes from multiplying the two main effects in our model: $11.6 \times 0.76 \approx 8.82$. Normally, this contrast is not of particular interest and is often not reported.

Assuming that all effects are the same across levels of all other covariates is, however, not always reasonable. In order to relax this assumption we can model interaction effects rather than just the main effects.

The interaction effects model using

If the following code is used to model the interaction effect

```
. stcox i.stage##i.sex
```

we get the following output from Stata:

```
Cox regression -- Breslow method for ties

No. of subjects =         5794                     Number of obs   =      5794
No. of failures =         1421
Time at risk    =        39540
                                                   LR chi2(3)      =   1160.40
Log likelihood  =   -11266.727                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.stage |   9.542748   .7568497    28.44   0.000     8.168896    11.14766
       2.sex |   .6608999   .0416637    -6.57   0.000     .5840839    .7478184
             |
   stage#sex |
       1 2   |   1.637429   .1949635     4.14   0.000     1.296619    2.067818
------------------------------------------------------------------------------
```

Note that we would get the exact same output if we used `stcox i.stage i.sex i.stage##i.sex`. We can start by answering the following questions:

1. What is the interpretation of the estimated hazard ratio that corresponds `1.stage`?
   It gives the effect of stage among males, i.e. males with distant metastases at diagnosis have an almost 10 times higher mortality rate than males without distant metastases at diagnosis.

2. What is the hazard ratio that corresponds to the joint effect of stage and sex (compared to the reference group)?
   The hazard ratio (HR) that compares females with distant metastases to the reference group is $9.54 \times 0.66 \times 1.64 \approx 10.3$.

3. Among females, what is the effect of having distant metastases at diagnosis?
   The hazard ratio (HR) for the effect of stage among females is approx. equal to $9.54 \times 1.64 \approx 15.6$. *I.e., among females, having distant metastases at diagnosis (compared to no distant metastases at diagnosis) implies a 15-fold increase of the all-cause mortality rate.*

4. What is the effect of sex within each level of stage?
   The hazard ratio (HR) for the effect of sex among patients with localized melanoma is approx. 0.66. *The hazard ratio (HR) for the effect of sex among patients with distant metastases at diagnosis is approx.* $0.66 \times 1.64 \approx 1.08$.

5. Is sex a significant effect modifier?
   Yes, the p-value for the interaction effect is below 0.05 and therefore significant on a 5% significance level. If any of our factor variables would have had more than two levels, we would not be able to determine this straight from the output. Instead we would perform, for example, a Wald test using

   ```
   testparm i.stage#i.sex.
   ```

   In this case, where both factor variables have only two levels, `testparm` *would give us the exact same test statistic and p-value that we got from the model output.*

6. The effect of stage among males is slightly lower here than in the main effects model. Why is that? *The effect of stage in the main effects model is averaged over the two sexes and, since the effect of stage is stronger among females than males, the male-specific effect size is expected to be smaller than the averaged.*

These results and calculations are more clearly displayed in tables.

		Stage	
		0	1
Sex	1	1.00 *(ref.)*	9.54
	2	0.66	10.3

Table 2: Results from the interaction effects model with one reference cell.

Table 2 displays the two main effects and the interaction effect, i.e. contrasts where males with localized melanoma are the reference group. To get the effect of stage within each level of sex we can perform some maths based on the estimates in Table 2:

		Stage	
		0	1
Sex	1	1.00 *(ref.)*	9.54
	2	$0.66 \div 0.66$	$10.3 \div 0.66$

$\Longrightarrow$

		Stage	
		0	1
Sex	1	1.00 *(ref.)*	9.54
	2	1.00 *(ref.)*	15.6

Table 3: Results from the interaction effects model with two reference cells.

Similarly, to compare within stage, we do the same as above but in the columns rather than the rows:

		Stage	
		0	1
Sex	1	1.00 *(ref.)*	$9.54 \div 9.54$
	2	0.66	$10.3 \div 9.54$

$\Longrightarrow$

		Stage	
		0	1
Sex	1	1.00 *(ref.)*	1.00 *(ref.)*
	2	0.66	1.08

Table 4: Results from the interaction effects model with two reference cells.

The command `lincom` in `Stata` can be used to calculate these contrasts for us and give us confidence intervals. The syntax and output equivalent to Table 3 is:

```
. lincom 1.stage, eform                    // Effect of stage among males

 ( 1)  1.stage = 0

------------------------------------------------------------------------------
          _t |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   9.542748   .7568497    28.44   0.000     8.168896    11.14766
------------------------------------------------------------------------------
```

```
. lincom 1.stage + 1.stage#2.sex, eform     // Effect of stage among females

 ( 1)  1.stage + 1.stage#2.sex = 0

------------------------------------------------------------------------------
          _t |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   15.62557   1.456428    29.49   0.000     13.01658    18.75749
------------------------------------------------------------------------------
```

and similarly the syntax and output equivalent to Table 4 is:

```
. lincom 2.sex, eform                    // Effect of sex among patients with localized stage

 ( 1)  2.sex = 0

------------------------------------------------------------------------------
          _t |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .6608999   .0416637    -6.57   0.000     .5840839    .7478184
------------------------------------------------------------------------------


. lincom 2.sex + 1.stage#2.sex, eform    // Effect of sex among patients with distant stage

 ( 1)  2.sex + 1.stage#2.sex = 0

------------------------------------------------------------------------------
          _t |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.082176   .1093058     0.78   0.434     .8878141    1.319089
------------------------------------------------------------------------------
```

The interaction effects model using

Another way of getting the parametrization in Table 3 directly from Stata is to use following syntax

```
. stcox i.sex i.stage#i.sex
```

which outputs:

```
Cox regression -- Breslow method for ties

No. of subjects =         5794                     Number of obs   =      5794
No. of failures =         1421
Time at risk    =        39540
                                                   LR chi2(3)      =   1160.40
Log likelihood  =   -11266.727                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       2.sex |   .6608999   .0416637    -6.57   0.000     .5840839    .7478184
             |
   stage#sex |
        1 1  |   9.542748   .7568497    28.44   0.000     8.168896    11.14766
        1 2  |   15.62557   1.456428    29.49   0.000     13.01658    18.75749
------------------------------------------------------------------------------
```

If we instead want to see the effect of sex within each level of stage (i.e., get the same results as in table 4) we simply use the following syntax

```
. stcox i.stage i.stage#i.sex
```

which in turn outputs:

```
Cox regression -- Breslow method for ties

No. of subjects =         5794                     Number of obs    =      5794
No. of failures =         1421
Time at risk    =        39540
                                                   LR chi2(3)       =   1160.40
Log likelihood  =   -11266.727                     Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.stage |   9.542748   .7568497    28.44   0.000     8.168896    11.14766
             |
   stage#sex |
        0 2  |   .6608999   .0416637    -6.57   0.000     .5840839    .7478184
        1 2  |   1.082176   .1093058     0.78   0.434     .8878141    1.319089
------------------------------------------------------------------------------
```

The one thing we do not get immediately from the `Stata` output is a significance test of the interaction effect. Testing this after running your model is, however, easy. Under the null hypothesis (i.e. no significant interaction effect) the two reported hazard ratios are the same. We can thus use the following syntax to formally test is the null hypothesis can be rejected (applies to the first model in this section).

```
. test 1.stage#1.sex==1.stage#2.sex

    ( 1)  1.stage#1b.sex - 1.stage#2.sex = 0

          chi2(  1) =   17.15
          Prob > chi2 =    0.0000
```

We see that the test statistic is 17.15 which in is approx. equal to 4.14^2 = the squared Wald test statistic from the model using the double hashes. The two test are equivalent.

As a final remark on this topic, note how the reported log likelihood for the three interaction models presented so far is the same (-11266.727). The reason is that we are conceptually fitting the same underlying statistical model to our data. The only difference between the models is the choice of contrast that we report.

A Some algebra

This section provides a more stringent motivation for why we can simply multiply and divide hazard ratios to get contrasts between different covariate patterns (table 2-4). To understand this it is important to be familiar with some algebraic rules.

$$e^{a+b} = e^a \cdot e^b$$
$$e^{a-b} = \frac{e^a}{e^b}$$
$$e^0 = 1$$

The Cox proportional hazards main effects model with n covariates can be written as

$$h(t) = h_0(t)e^{(\beta_1 \cdot x_1 + ... + \beta_n \cdot x_n)}$$

First, when comparing the hazard rate between two study participants, A and B, for whom everything is equal at some time point t except for the value on the i^{th} covariate, the ratio between their hazard functions (the hazard ratio, HR) equals:

$$\begin{aligned}
\text{HR} &= \frac{h_A(t|x_i = 1)}{h_B(t|x_i = 0)} \\
&= \frac{h_0(t)e^{(\beta_1 \cdot x_1 + ... + \beta_i \cdot 1 + ... + \beta_n \cdot x_n)}}{h_0(t)e^{(\beta_1 \cdot x_1 + ... + \beta_i \cdot 0 + ... + \beta_n \cdot x_n)}} \\
&= \{\text{everything cancels out except the } i^{th} \text{ coefficient}\} \\
&= \frac{e^{\beta_i}}{e^0} \\
&= \frac{e^{\beta_i}}{1} \\
&= e^{\beta_i}
\end{aligned}$$

That is, the only thing that is left, is e to the power of the coefficient for x_i. This is the *effect of covariate i on the outcome.*

When there are only two covariates, as in the example described in the tutorial, the Cox regression main effects model looks like

$$h(t) = h_0(t)e^{(\beta_1 \cdot x_1 + \beta_2 \cdot x_2)}$$

and a table with the rates for different covariate patterns:

		x_1	
		0	1
x_2	1	$h_0(t)$	$h_0(t)e^{\beta_1}$
	2	$h_0(t)e^{\beta_2}$	$h_0(t)e^{(\beta_1+\beta_2)}$

Table 5: Cox regression main effects model.

We see that this table correspond to Table 1 in the tutorial, but here with rates in the cells rather than contrasts.

If we add an interaction term to the model it becomes

$$h(t) = h_0(t)e^{(\beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_1 \cdot x_2)}$$

We now estimate an extra parameter (β_3). A table presenting rates for different covariate patterns now looks slightly different than for the main effects model:

		x_1	
		0	1
x_2	1	$h_0(t)$	$h_0(t)e^{\beta_1}$
	2	$h_0(t)e^{\beta_2}$	$h_0(t)e^{(\beta_1+\beta_2+\beta_3)}$

Table 6: Cox regression interaction effects model.

This table is similar to table 2. We can now use the covariate-specific expressions for the hazard rates and calculate hazard ratios for any contrast that we wish. For example, if we want to compare individuals A and B with covariate levels $x_1 = 1, x_2 = 1$ to individuals with $x_1 = 0, x_2 = 0$:

$$\begin{aligned}
\text{HR} &= \frac{h_A(t|x_1 = 1, x_2 = 1)}{h_B(t|x_1 = 0, x_2 = 0)} \\
&= \frac{h_0(t)e^{(\beta_1 \cdot 1 + \beta_2 \cdot 1 + \beta_3 \cdot 1)}}{h_0(t)e^{(\beta_1 \cdot 0 + \beta_2 \cdot 0 + \beta_3 \cdot 0)}} \\
&= \frac{e^{\beta_1+\beta_2+\beta_3}}{e^0} \\
&= \frac{e^{\beta_1+\beta_2+\beta_3}}{1} \\
&= e^{\beta_1+\beta_2+\beta_3} \\
&= e^{\beta_1} \cdot e^{\beta_2} \cdot e^{\beta_3} \\
&= \{\text{main effect for covariate 1} \cdot \text{main effect for covariate 2} \cdot \text{interaction effect}\}
\end{aligned}$$