

# Understanding interactions in the Cox model (Stata version)

Paul Dickman

October 2020

[www.pauldickman.com/video/interactions/](http://www.pauldickman.com/video/interactions/)

## Overview of this lecture

- ▶ Target audience is students and researchers in biomedical sciences without extensive training in statistics.
- ▶ This lecture will present an introduction to covariate by covariate interactions in the Cox model. The concepts are applicable to other models.
  - ▶ Interpreting parameter estimates (in both main effects and interaction models)
  - ▶ Reparameterising the interaction model.
- ▶ Slides available at <http://www.pauldickman.com/video/interactions/>
- ▶ R, Stata, and SAS code available on the same page as the slides.
- ▶ These slides use Stata; an R version also exists.

## Interactions between covariates

- ▶ In the 'Introduction to Cox' lecture we assumed estimated effects (hazard ratios) are constant across all levels of other covariates and constant over follow-up time.
- ▶ We'll now study and relax the assumption that effects are constant across all levels of other covariates.
- ▶ We'll analyse data for patients with localised melanoma (because there are interesting differences in survival between males and females).
- ▶ Outcome is death due to cancer.
- ▶ We will estimate the hazard ratio for sex (females/males), and study if it varies by calendar period and age group.
- ▶ Studying if the HR varies by time-since-diagnosis is conceptually similar, but technically difficult since we don't estimate the effect of time-since-diagnosis. This is the test of proportional hazards and will be covered in a separate lecture.

## Main effects model – localised melanoma

```
. stcox sex i.agegrp year8594, efron
```

```
No. of subjects =          5,318          Number of obs   =          5,318
No. of failures =           960
Time at risk    = 32376.66667
Log likelihood  = -7792.6678          LR chi2(5)       =          212.69
                                          Prob > chi2      =          0.0000
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
sex	.588247	.038501	-8.11	0.000	.5174261	.6687614
agegrp						
0-44	1	(base)				
45-59	1.326882	.124957	3.00	0.003	1.103244	1.595852
60-74	1.859043	.1689419	6.82	0.000	1.555736	2.221483
75+	3.380448	.3530315	11.66	0.000	2.754738	4.14828
year8594	.7164882	.0474192	-5.04	0.000	.6293238	.8157253

## Main effects model – localised melanoma (2)

- ▶ Technical: We specify the Efron methods for ties (default is Breslow) to obtain the same results as R.
- ▶ sex is coded as 1 for men and 2 for women.
- ▶ Females have an estimated 41% lower cancer-specific mortality than males.
- ▶ The 41% lower mortality is assumed to apply to both calendar periods, all age groups, and across the entire follow-up.
- ▶ We will shortly add an interaction between sex and period of diagnosis. This allows the effect of sex to potentially vary between the periods. We will then add an interaction between sex and age.
- ▶ The assumption that the effect of sex is constant across follow-up time is conceptually similar, but interactions with time are technically more difficult to model. This is covered in a separate lecture.
- ▶ The assumption of constant effects over time has a special name: proportional hazards.

## A look at the parameters in the main effects model

- ▶ First, let's review the interpretation of coefficients.
- ▶ Consider the Cox model:

$$\log(\lambda(t|X_1, X_2)) = \log(\lambda_0(t)) + \beta_1 X_1 + \beta_2 X_2$$

where

$$X_1 = \begin{cases} 0 & \text{if male} \\ 1 & \text{if female} \end{cases}$$

$$X_2 = \begin{cases} 0 & \text{if diagnosed 1975–84} \\ 1 & \text{if diagnosed 1985–94} \end{cases}$$

- ▶ We are, for simplicity, not including age in the model

## A look at the parameters in the main effects model

sex	year	$X_1$	$X_2$	$\log(\lambda(t X_1, X_2))$
male	1975–84	0	0	$\log(\lambda_0(t))$
male	1985–94	0	1	$\log(\lambda_0(t)) + \beta_2$
female	1975–84	1	0	$\log(\lambda_0(t)) + \beta_1$
female	1985–94	1	1	$\log(\lambda_0(t)) + \beta_1 + \beta_2$

- ▶ We see that  $\beta_1 = \log(\lambda(t|X_1 = 1, X_2 = 0)) - \log(\lambda(t|X_1 = 0, X_2 = 0))$   
and  
 $\beta_1 = \log(\lambda(t|X_1 = 1, X_2 = 1)) - \log(\lambda(t|X_1 = 0, X_2 = 1))$
- ▶  $\beta_1$  is the difference in log hazards between females and males, holding period of diagnosis constant. It is the same for both periods.
- ▶ That is,  $\beta_1 = \log(\lambda(t|X_1 = 1, X_2)) - \log(\lambda(t|X_1 = 0, X_2))$
- ▶ We see that the baseline hazards cancel out.
- ▶ If we had adjusted for age, then the additional parameters would cancel out when interpreting  $\beta_1$ .

## Adding an interaction

Extend the model to:

$$\log(\lambda(t|X_1, X_2)) = \log(\lambda_0(t)) + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

where

$$X_3 = \begin{cases} 1 & \text{if female \& diagnosed 1985-94} \\ 0 & \text{otherwise} \end{cases}$$



## A look at the parameters in the interaction model

sex	year	$X_1$	$X_2$	$X_3$	$\log(\lambda(t X_1, X_2, X_3))$
male	1975–84	0	0	0	$\log(\lambda_0(t))$
male	1985–94	0	1	0	$\log(\lambda_0(t)) + \beta_2$
female	1975–84	1	0	0	$\log(\lambda_0(t)) + \beta_1$
female	1985–94	1	1	1	$\log(\lambda_0(t)) + \beta_1 + \beta_2 + \beta_3$

- ▶ We see that  $\beta_1 = \log(\lambda(t|X_1 = 1, X_2 = 0)) - \log(\lambda(t|X_1 = 0, X_2 = 0))$
- ▶  $\beta_1$  is interpreted as the difference in log hazards between females and males, but only for 1975–84 ( $X_2 = 0$ ).
- ▶ The difference in log hazards between females and males for 1985–94 ( $X_2 = 1$ ) is  $\beta_1 + \beta_3$
- ▶ If  $\beta_3$  is zero then the effect of sex (difference in log hazards) is the same for both periods.
- ▶  $\beta_3$  represents the difference, between 1985–94 and 1975–84, between the difference in log hazards between females and males.

## Interaction between sex and period

```
. stcox i.sex i.year8594 i.sex#i.year8594 i.agegrp, efron
```

	_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
Male	1	(base)				
Female	.6038347	.0532172	-5.72	0.000	.5080423	.7176889
year8594						
Diagnosed 75-84	1	(base)				
Diagnosed 85-94	.7353338	.0650007	-3.48	0.001	.6183602	.8744349
sex#year8594						
Female#Diagnosed 85-94	.9437898	.1232728	-0.44	0.658	.7306273	1.219143
agegrp						
0-44	1	(base)				
45-59	1.325798	.1248814	2.99	0.003	1.1023	1.594612
60-74	1.8578	.1688542	6.81	0.000	1.554654	2.220056
75+	3.379666	.3529257	11.66	0.000	2.75414	4.147263

## Interaction between sex and period (2)

- ▶ The HR labelled Female is now the effect of sex for the reference level of period (1975–84).
- ▶ The HR for the interaction (0.9438) is the additional (multiplicative) HR for the second period compared to the first. If this is one then the effect of sex is the same in the two periods.
- ▶ The HR for sex is 0.6038 in the first period and  $0.6038 \times 0.9438 = 0.5699$  in the second period.

## Linear combinations of parameters

- ▶ The HR for sex, with CI, can be obtained using the `lincom` command (linear combinations of parameters).

```
. lincom 2.sex + 2.sex#1.year8594, eform
```

	_t	exp(b)	Std. Err.	z	P> z	[95% Conf. Interval]	
	(1)	.569893	.0553333	-5.79	0.000	.4711364	.6893503

- ▶ On the original scale we add the two parameters and exponentiate the result, which is equivalent to the product of the two HRs ( $0.6038 \times 0.9438 = 0.5699$ ) in the previous slide.
- ▶ I will shortly show how to reparameterise the model so this same contrast is represented by a single parameter.

## Same model but different Stata syntax

```
. stcox i.sex##i.year8594 i.agegrp, efron
```

```
-----+-----  
          _t | Haz. Ratio   Std. Err.      z    P>|z|    [95% Conf. Interval]  
-----+-----  
          Male |             1   (base)  
          Female |   .6038347   .0532172   -5.72   0.000   .5080423   .7176889  
          |  
          year8594 |  
          Diagnosed 75-84 |             1   (base)  
          Diagnosed 85-94 |   .7353338   .0650007   -3.48   0.001   .6183602   .8744349
```

[output truncated]

- ▶ The ## operator specifies the interaction plus the main effects.  
i.sex##i.year8594 is equivalent to  
i.sex i.year8594 i.sex#i.year8594
- ▶ Use help fvvarlist for more on factor variables in Stata.

## Reparameterising the model

```
. stcox i.year8594 i.sex#i.year8594 i.agegrp, efron
```

	_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
-----+-----						
year8594						
Diagnosed 75-84		1	(base)			
Diagnosed 85-94		.7353338	.0650007	-3.48	0.001	.6183602 .8744349
-----+-----						
sex#year8594						
Female#Diagnosed 75-84		.6038347	.0532172	-5.72	0.000	.5080423 .7176889
Female#Diagnosed 85-94		.569893	.0553333	-5.79	0.000	.4711364 .6893503

```
[output omitted: estimates for agegrp are unchanged]
```

- ▶ We are fitting the same model, but with a different parameterisation.
- ▶ The model is identical in that it has the same number of parameters, same predicted values (of the outcome), and same likelihood.
- ▶ One parameter has a different interpretation; the two effects of sex are now each represented by a single parameter.

## A look at the two parameterisations

Default parameterisation

Effect of sex for 1985–94 is represented by  $\beta_1 + \beta_3$

sex	year	$X_1$	$X_2$	$X_3$	$\log(\lambda(t X_1, X_2, X_3))$
male	1975–84	0	0	0	$\log(\lambda_0(t))$
male	1985–94	0	1	0	$\log(\lambda_0(t)) + \beta_2$
female	1975–84	1	0	0	$\log(\lambda_0(t)) + \beta_1$
female	1985–94	1	1	1	$\log(\lambda_0(t)) + \beta_1 + \beta_2 + \beta_3$

Alternative parameterisation

Effect of sex for 1985–94 is represented by  $\beta_3$

sex	year	$X_1$	$X_2$	$X_3$	$\log(\lambda(t X_1, X_2, X_3))$
male	1975–84	0	0	0	$\log(\lambda_0(t))$
male	1985–94	0	1	0	$\log(\lambda_0(t)) + \beta_2$
female	1975–84	1	0	0	$\log(\lambda_0(t)) + \beta_1$
female	1985–94	0	1	1	$\log(\lambda_0(t)) + \beta_2 + \beta_3$

## Yet another reparameterisation

```
. stcox i.year8594#i.sex i.agegrp, efron
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
year8594#sex						
Diagnosed 75-84#Male	1	(base)				
Diagnosed 75-84#Female	.6038347	.0532172	-5.72	0.000	.5080423	.7176889
Diagnosed 85-94#Male	.7353338	.0650007	-3.48	0.001	.6183602	.8744349
Diagnosed 85-94#Female	.4190616	.0408792	-8.92	0.000	.3461335	.5073551
agegrp						
0-44	1	(base)				
45-59	1.325798	.1248814	2.99	0.003	1.1023	1.594612
60-74	1.8578	.1688542	6.81	0.000	1.554654	2.220056
75+	3.379666	.3529257	11.66	0.000	2.75414	4.147263
-----						

- ▶ The effects of sex and period are now estimated compared to the joint reference group (males diagnosed in the early period).
- ▶ The coefficients now represent the difference in log hazards between each of three categories of sex and period compared to the joint reference.



## Interactions with a joint reference category

sex	year	$X_1$	$X_2$	$X_3$	$\log(\lambda(t X_1, X_2, X_3))$
male	1975–84	1	0	0	$\log(\lambda_0(t)) + \beta_1$
male	1985–94	0	1	0	$\log(\lambda_0(t)) + \beta_2$
female	1975–84	0	0	1	$\log(\lambda_0(t)) + \beta_3$
female	1985–94	0	0	0	$\log(\lambda_0(t))$

## Now study the interaction between sex and age group

```
. stcox i.year8594 i.sex##i.agegrp, efron
```

	_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
-----+-----						
year8594						
Diagnosed 75-84		1	(base)			
Diagnosed 85-94		.718027	.047512	-5.01	0.000	.6306909 .817457
sex						
Male		1	(base)			
Female		.4890204	.0705671	-4.96	0.000	.3685497 .6488703
agegrp						
0-44		1	(base)			
45-59		1.227955	.1502451	1.68	0.093	.966128 1.560739
60-74		1.710183	.2054475	4.47	0.000	1.351407 2.164208
75+		2.732236	.419661	6.54	0.000	2.02198 3.691981
sex#agegrp						
Female#45-59		1.188529	.2274608	0.90	0.367	.8167827 1.729469
Female#60-74		1.206162	.2212445	1.02	0.307	.8419199 1.727986
Female#75+		1.499611	.3158217	1.92	0.054	.9924611 2.265916
-----+-----						

## Now study the interaction between sex and age group (2)

- ▶ The row labelled Female gives the effect of sex at the reference level of age (0–45). The three interaction effects represent the additional effects for the other ages.
- ▶ The HR for sex is 0.489 for the youngest age group and  $0.489 \times 1.18853 = 0.581$  for age group 45–59.

## Why might the effect of sex depend on age?

- ▶ Can you think of a plausible biological reason as to why sex differences in survival might depend on age at diagnosis?
- ▶ Might an alternative categorisation of age be more appropriate?

## Reparameterise to get the four HRs for sex (one HR for each age group)

```
. stcox i.year8594 i.agegrp i.agegrp#i.sex, efron
```

	_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----							
year8594							
Diagnosed 75-84		1	(base)				
Diagnosed 85-94		.718027	.047512	-5.01	0.000	.6306909	.817457
agegrp							
0-44		1	(base)				
45-59		1.227955	.1502451	1.68	0.093	.966128	1.560739
60-74		1.710183	.2054475	4.47	0.000	1.351407	2.164208
75+		2.732236	.419661	6.54	0.000	2.02198	3.691981
agegrp#sex							
0-44#Female		.4890204	.0705671	-4.96	0.000	.3685497	.6488703
45-59#Female		.5812147	.0731102	-4.31	0.000	.4542189	.7437175
60-74#Female		.5898376	.0668412	-4.66	0.000	.47236	.7365323
75+#Female		.7333404	.1125519	-2.02	0.043	.5428308	.9907103
-----							

## Reparameterise to get the four HRs for sex (one HR for each age group) (2)

- ▶ Female superiority in survival decreases with increasing age.
- ▶ To test this, we need to test the null hypothesis that the three interaction effects are jointly zero. (equivalent to testing that the four coefficients above are equal).

## Test of interaction effects

```
. test 0.agegrp#2.sex = 1.agegrp#2.sex = 2.agegrp#2.sex = 3.agegrp#2.sex
```

```
( 1) 0b.agegrp#2.sex - 1.agegrp#2.sex = 0
```

```
( 2) 0b.agegrp#2.sex - 2.agegrp#2.sex = 0
```

```
( 3) 0b.agegrp#2.sex - 3.agegrp#2.sex = 0
```

```
      chi2( 3) =      3.71
```

```
      Prob > chi2 =      0.2940
```

- ▶ Fail to reject the null hypothesis that the effect of sex is the same for each age group.

## Now using a likelihood ratio test

```
. stcox i.year8594 i.agegrp##i.sex, efron  
[output omitted]
```

```
. estimates store interaction
```

```
. stcox i.year8594 i.agegrp i.sex, efron  
[output omitted]
```

```
. lrtest interaction
```

```
Likelihood-ratio test                                LR chi2(3) =      3.76  
(Assumption: . nested in interaction)              Prob > chi2 =     0.2886
```

- ▶ We are performing the same hypothesis test, but now using a likelihood ratio test rather than a Wald test.
- ▶ Test statistic and p-values are similar but not identical.
- ▶ Wald test is an approximation to the likelihood ratio test.