Understanding interactions in the Cox model (R version)

Paul Dickman

October 2020

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)
  - Reparamaterising 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 R; a Stata 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

```
> summary(
+ coxph(Surv(surv_mm, death_cancer)
+        ~ sex + year8594 + agegrp,
+          data = melanoma.12)
+         )
                            coef exp(coef) se(coef)      z
sexFemale                -0.53061   0.58825  0.06545 -8.107
year8594Diagnosed 85-94 -0.33339   0.71649  0.06618 -5.037
agegrp45-59               0.28283   1.32688  0.09417  3.003
agegrp60-74               0.62006   1.85904  0.09088  6.823
agegrp75+                 1.21801   3.38045  0.10443 11.663
```

- Sex is coded as 1 for men and 2 for women.
- The coefficient for `sexFemale` is the estimated difference in the log hazards between females and males, holding period and age constant.
- The exponentiated coefficient for `sexFemale` is the estimated ratio of hazards for females to males (i.e., the hazard ratio).

Main effects model – localised melanoma (2)

	coef	exp(coef)	se(coef)	z
sexFemale	-0.53061	0.58825	0.06545	-8.107
year8594Diagnosed 85-94	-0.33339	0.71649	0.06618	-5.037
agegrp45-59	0.28283	1.32688	0.09417	3.003
agegrp60-74	0.62006	1.85904	0.09088	6.823
agegrp75+	1.21801	3.38045	0.10443	11.663

- 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
  $\beta_1 = \log(\lambda(t|X_1 = 1, X_2 = 1)) - \log(\lambda(t|X_1 = 0, X_2 = 1))$
- β_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 β_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 $\beta_1 = \log(\lambda(t|X_1 = 1, X_2 = 0)) - \log(\lambda(t|X_1 = 0, X_2 = 0))$
- β_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 β_3 is zero then the effect of sex (difference in log hazards) is the same for both periods.
- β_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

```
> summary(  coxph(Surv(surv_mm, death_cancer)
+           ~ sex + year8594 + sex:year8594 + agegrp, data = melanoma.l2)
+ )
                                       coef exp(coef) se(coef)      z
sexFemale                           -0.50445   0.60383  0.08813 -5.724
year8594Diagnosed 85-94             -0.30743   0.73533  0.08840 -3.478
agegrp45-59                          0.28201   1.32580  0.09419  2.994
agegrp60-74                          0.61939   1.85780  0.09089  6.815
agegrp75+                            1.21778   3.37967  0.10443 11.662
sexFemale:year8594Diagnosed 85-94 -0.05785   0.94379  0.13061 -0.443
```

- The coefficient labelled `sexFemale` is now the effect of sex for the reference level of period (1975–84).
- The coefficient in the bottom row (-0.05785) is the additional log HR for the second period compared to the first. If this is zero then the effect of sex is the same in the two periods.
- Equivalently, exp(coef) is the multiplicative interaction effect. If it is equal to one, then the effect of sex is the same in both periods.
- The HR for sex is 0.6038 in the first period and $0.6038 \times 0.9438 = 0.5699$ in the second period.

Same model but different R syntax

```
> summary(
+     coxph(Surv(surv_mm, death_cancer)
+           ~ sex*year8594 + agegrp,
+           data = melanoma.l2)
+ )
                                    coef exp(coef) se(coef)      z
sexFemale                        -0.50445   0.60383  0.08813 -5.724
year8594Diagnosed 85-94          -0.30743   0.73533  0.08840 -3.478
agegrp45-59                       0.28201   1.32580  0.09419  2.994
agegrp60-74                       0.61939   1.85780  0.09089  6.815
agegrp75+                         1.21778   3.37967  0.10443 11.662
sexFemale:year8594Diagnosed 85-94 -0.05785   0.94379  0.13061 -0.443
```

- The * operator specifies the interaction plus the main effects. `sex*year8594` is equivalent to `sex + year8594 + sex:year8594`

Reparameterising the model

```
> summary(
+     coxph(Surv(surv_mm, death_cancer)
+           ~ year8594 + sex:year8594 + agegrp,
+           data = melanoma.l2)
+ )
                                      coef exp(coef) se(coef)      z
year8594Diagnosed 85-94           -0.30743   0.73533  0.08840 -3.478
agegrp45-59                        0.28201   1.32580  0.09419  2.994
agegrp60-74                        0.61939   1.85780  0.09089  6.815
agegrp75+                          1.21778   3.37967  0.10443 11.662
year8594Diagnosed 75-84:sexFemale -0.50445   0.60383  0.08813 -5.724
year8594Diagnosed 85-94:sexFemale -0.56231   0.56989  0.09709 -5.791
```

- 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 β_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

```
> summary(
+     coxph(Surv(surv_mm, death_cancer)
+           ~ sex:year8594 + agegrp,
+           data = melanoma.l2)
+ )
                                    coef exp(coef) se(coef)      z
agegrp45-59                      0.28201   1.32580  0.09419  2.994
agegrp60-74                      0.61939   1.85780  0.09089  6.815
agegrp75+                        1.21778   3.37967  0.10443 11.662
sexMale:year8594Diagnosed 75-84  0.86974   2.38628  0.09755  8.916
sexFemale:year8594Diagnosed 75-84 0.36528  1.44092  0.09790  3.731
sexMale:year8594Diagnosed 85-94  0.56231   1.75472  0.09709  5.791
sexFemale:year8594Diagnosed 85-94      NA        NA  0.00000     NA
```

- The effects of sex and period are now estimated compared to the joint reference group (females diagnosed in the latter 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

```
> summary(
+     coxph(Surv(surv_mm, death_cancer)
+           ~ year8594 + agegrp + sex + sex:agegrp,
+           data = melanoma.l2)
+ )
                            coef exp(coef) se(coef)      z
year8594Diagnosed 85-94 -0.33125   0.71803  0.06617 -5.006
agegrp45-59              0.20535   1.22796  0.12235  1.678
agegrp60-74              0.53660   1.71018  0.12013  4.467
agegrp75+                1.00512   2.73224  0.15360  6.544
sexFemale               -0.71535   0.48902  0.14430 -4.957
agegrp45-59:sexFemale    0.17272   1.18853  0.19138  0.902
agegrp60-74:sexFemale    0.18744   1.20616  0.18343  1.022
agegrp75+:sexFemale      0.40521   1.49961  0.21060  1.924
```

- The row labelled `sexFemale` gives the effect of sex at the reference level of age (less than 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)

```
> summary(  coxph(Surv(surv_mm, death_cancer)
+           ~ year8594 + agegrp + sex:agegrp, data = melanoma.l2)
+ )
                             coef exp(coef) se(coef)      z
year8594Diagnosed 85-94 -0.33125   0.71803  0.06617 -5.006
agegrp45-59              0.20535   1.22796  0.12235  1.678
agegrp60-74              0.53660   1.71018  0.12013  4.467
agegrp75+                1.00512   2.73224  0.15360  6.544
agegrp0-44:sexFemale    -0.71535   0.48902  0.14430 -4.957
agegrp45-59:sexFemale   -0.54263   0.58121  0.12579 -4.314
agegrp60-74:sexFemale   -0.52791   0.58984  0.11332 -4.659
agegrp75+:sexFemale     -0.31015   0.73334  0.15348 -2.021
```

- 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

```
> fit2 <- coxph(Surv(surv_mm, death_cancer)
+               ~ year8594 + agegrp + sex + sex:agegrp,
+               data = melanoma.l2)
> library(car)
> linearHypothesis(fit2, c("agegrp45-59:sexFemale",
         "agegrp60-74:sexFemale","agegrp75+:sexFemale"))
Linear hypothesis test

Hypothesis:
agegrp45 - 59:sexFemale = 0
agegrp60 - 74:sexFemale = 0
agegrp75  + :sexFemale = 0

Model 1: restricted model
Model 2: Surv(surv_mm, death_cancer) ~ year8594 + agegrp + sex + sex:agegrp

  Res.Df Df  Chisq Pr(>Chisq)
1   5313
2   5310  3 3.7145      0.294
```

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

Now using a likelihood ratio test

```
> fit1 <- coxph(Surv(surv_mm, death_cancer)
+       ~ year8594 + agegrp + sex,
+       data = melanoma.l2)
>
> fit2 <- coxph(Surv(surv_mm, death_cancer)
+               ~ year8594 + agegrp + sex + sex:agegrp,
+               data = melanoma.l2)
> anova(fit1,fit2,test="Chisq")
Analysis of Deviance Table
 Cox model: response is  Surv(surv_mm, death_cancer)
 Model 1: ~ year8594 + agegrp + sex
 Model 2: ~ year8594 + agegrp + sex + sex:agegrp
   loglik  Chisq Df P(>|Chi|)
1 -7792.7
2 -7790.8 3.7599  3    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.