

## Survival Analysis II

Su-Chun Cheng  
 April 7, 2009  
[scheng@biostat.ucsf.edu](mailto:scheng@biostat.ucsf.edu)

Reading VGSM 7.2.4 - 7.2.10

- > Project Description Due Today
- > Lab #1 Documents Revised

## Cox Model

- Can summarize effects based on coefficients,  $\beta$ , or in terms of hazard ratios,  $\exp(\beta)$
- Hazard ratios work better for interpretation
- Math works better based on coefficients
- Confidence intervals and tests are based on the fact that  $\hat{\beta}$  has an approximate normal distribution (given 15-25 **events**)

## Test and Confidence Interval in Cox Model

- Test and Confidence interval are based on estimators  $\hat{\beta}$  for coefficients  $\beta$
- 95% CI for HR is
- Upper limit:  $\exp(\hat{\beta} + 1.96 * SE(\hat{\beta}))$
- Lower limit:  $\exp(\hat{\beta} - 1.96 * SE(\hat{\beta}))$
- Wald test:  $Z = \hat{\beta} / SE(\hat{\beta})$

## Lung Cancer Data

- 40 subjects with BAC lung cancer
- Underwent PET scan
- Determined uptake of FDG (in standard units): variable `fdgavid` (tumor SUV > 2.5, Y/N)
- 12 subjects died during follow-up

## Wald test and CI hazard ratio scale

```
No. of subjects = 40          Number of obs = 40
No. of failures = 12
Time at risk = 1258.299998
Log likelihood = -31.394758
LR chi2(1) = 10.03
Prob > chi2 = 0.0015
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
fdgavid	11.7675	12.35468	2.35	0.019	1.503172 92.1212

$$11.77 - 1.96 * 12.4 = -12.5$$

$$11.77 + 1.96 * 12.4 = 36.1$$

very different

## Wald test and CI coefficient scale

```
. stcox fdgavid, nohr
No. of subjects = 40          Number of obs = 40
No. of failures = 12
Time at risk = 1258.299998
Log likelihood = -31.394758
LR chi2(1) = 10.03
Prob > chi2 = 0.0015
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
fdgavid	2.465341	1.049899	2.35	0.019	.4075778 4.523105

$$2.47 - 1.96 * 1.05 = 0.41$$

$$2.47 + 1.96 * 1.05 = 4.52$$

$$HR = \exp(2.47) = 11.8$$

$$95\% \text{ CI: } \exp(0.41) = 1.5$$

$$\exp(4.52) = 92.1 \text{ as in the previous page}$$

## Likelihood Ratio Tests

LR tests

- Tests for effect of predictor(s) by comparing log-likelihood between two models
- Fit models with and without predictor(s) to be tested
- -2 Times difference in log-likelihoods compared to a chi-square distribution
- Important to use when number of failures is small and the HR is far from 1 (strong effect)

## LR test for fdgavid

- `stcox fdgavid tumorsize multi`  
*fits model with all predictors (the reference model)*
- `est store A`  
*asks Stata to save log-likelihood for above model, call it "A"*
- `stcox tumorsize multi`  
*fits model leaving out fdgavid*
- `lrtest A`  
*compare log-likelihoods (default to the previous model)*

## Reference Model

few failures

```

No. of subjects = 40
No. of failures = 12
Time at risk = 1258.297998
Log likelihood = -29.48613
LR chi2(3) = 13.85
Prob > chi2 = 0.0031
    
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
fdgavid	7.4968	8.149509	1.85	0.064	.8903576 63.12297
tumorsize	1.249128	.1436471	1.93	0.055	.9970583 1.564924
multifocal	.296144	.3337985	-1.08	0.280	.0325141 2.497331

fairly large HR      non-significant Wald test

## Likelihood Ratio Test

Log likelihood = -32.03254 model w/o fdgavid  
 Log likelihood = -29.48613 model with fdgavid  
 -2 times diff = 5.09283

```

. lrtest A
Likelihood-ratio test
(Assumption: A nested in A)
LR chi2(1) = 5.09
Prb > chi2 = 0.0240
    
```

current model      significant likelihood ratio test

## Likelihood Ratio Test

A. `stcox fdgavid tumorsize multi`  
 B. `stcox tumorsize multi`

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
tumorsize	1.393288	.147124	3.14	0.002	1.132813 1.713655
multifocal	.2308219	.259914	-1.30	0.193	.0253976 2.097787

. lrtest A B  
 (Assumption: B nested in A)      Prob > chi2 = 0.0240

C. `stcox fdgavid multi`

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
fdgavid	11.51	12.08681	2.33	0.020	1.46966 90.14333
multifocal	.4677229	.506731	-0.70	0.483	.0559496 3.910033

. lrtest A C  
 (Assumption: C nested in A)      Prob > chi2 = 0.0725

fdgavid & tumorsize: high association

## Likelihood Ratio vs. Wald

- Two tests for the same null hypothesis
- Typically very close in results
- Will disagree when sample size small and HR are far from zero or collinearity
- When they disagree, the likelihood ratio test is more reliable.
- LR test always better -- just inconvenient to compute

## Binary Predictors

```

No. of subjects =      40          Number of obs =      40
No. of failures =      12
Time at risk    = 1258.299998
Log likelihood   = -35.78203      LR chi2(1)    =      1.26
                                      Prob > chi2    = 0.2623

```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
over3cm	1.950839	1.196869	1.09	0.276	.5861334 6.493017

“over3cm” coded 0/1

0 = tumor less than 3 cm

1 = tumor greater than 3 cm

relative hazard for  $\geq 3$ cm compared to  $< 3$  cm = 1.95

hazards about double!

## Binary Predictors

- Suggest 0/1 coding
- One-point change is easy to interpret
- Makes the baseline hazard an identifiable group  
e.g., those with tumors  $< 3$  cm
- Simplifies lincoms when we consider interactions  
and we will consider interactions
- Get the same answer if coded 10 vs. 11
- Get the significance but different HR if coded 0/2

## Reversed Coding

```

. recode over3cm 0=1 1=0, gen(less3cm)
. stcox less3cm

```

```

No. of subjects =      40          Number of obs =      40
No. of failures =      12
Time at risk    = 1258.299998
Log likelihood   = -35.78203      LR chi2(1)    =      1.26
                                      Prob > chi2    = 0.2623

```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
less3cm	.5125998	.3144878	-1.09	0.276	.1540116 1.706096

“less3cm” coded 0/1

0 = tumor greater than 3 cm

1 = tumor less than 3 cm

LR, Wald tests same. HR and it's CI are reciprocals

.5125998=1/1.950839

## Categorical Predictors

- Fit in Stata by `xi: stcox i.categoricalpredictor`
- Lots of different possible tests and comparisons
  - Overall versus trend tests (if ordinal)
  - Making pairwise comparisons

## PBC Data

- 312 patients: Primary Biliary Cirrhosis (PBC)
- Randomized trial: DPCA vs. Placebo
- 125 subjects died
- Dataset used to develop natural hx model
- 15 predictors: hepatomegaly, spiders, bilirubin, etc.
- Dickson, et al. *Hepatology* 10:1-7 (1989)

## Cox Model

```

. xi: stcox sex i._h1stol

```

```

No. of subjects =      312          Number of obs =      312
No. of failures =      125
Time at risk    = 1713.853528
Log likelihood   = -611.61794      LR chi2(4)    =      56.72
                                      Prob > chi2    = 0.0000

```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
sex	.6072455	.1433789	-2.11	0.035	.3822823 .9645939
_h1stol_2	5.488862	5.667663	1.65	0.099	.7253584 41.53478
_h1stol_3	9.459565	9.589963	2.22	0.027	1.296988 68.99321
_h1stol_4	23.05048	23.28112	3.11	0.002	3.183916 166.8778
3 vs 2	1.723411	.5056402	1.86	0.064	.9697295 3.06286
4 vs 3	2.436738	.4825026	4.50	0.000	1.652955 3.592168

```

. lincom _h1stol_4-_h1stol_3, hr

```

Is histology significant?

## Overall vs. Trend Tests

- Both have same null hypothesis:  
*no difference in event rates between the groups*
- But different alternative hypothesis:  
*overall: at least one group is different*  
*trend: there is a trend in the groups*
- Use trend tests only for **ordered** predictors  
*no trend test for ethnicity*
- When trends exist, a trend test is more powerful
- For ordinal predictors it is more interpretable

## Trend Test

```
xi: stcox sex i.histol
test -1* _Ihistol_2 + _Ihistol_3 + 3* _Ihistol_4=0
```

chi2( 1) = 10.69  
Prob > chi2 = 0.0011

p = 0.0011, there is survival trend with pathology grade

appropriate **linear combination** from p. 82 of VGSM

## Overall Test

Wald Test

```
. xi: stcox sex i.histol
(Output omitted)

. testparm _Ihistol*
( 1) _Ihistol_2 = 0
( 2) _Ihistol_3 = 0
( 3) _Ihistol_4 = 0
      chi2( 3) = 42.83
      Prob > chi2 = 0.0000
```

*at least one group different*

## Overall Test

Likelihood Ratio Test

```
xi: stcox sex i.histol
(Output omitted)
est store SexHist
Stcox sex
(Output omitted)
Lrttest SexHist
```

LR Chisq = 52.9  
Wald Chisq = 42.8  
*similar*

Likelihood-ratio test LR chi2(3) = 52.95  
(Assumption: . nested in Sex Hist) Prob > chi2 = 0.0000

Histology is a significant predictor of death after adjusting for age, p < 0.0001

## Survival by Tumor SUV

binary

	Alive	Dead
Tumor SUV=0	4	0
Tumor SUV> 0	24	12

No Deaths in Those with Tumor SUV=0

## Zero Hazard Ratio

LR test still OK

```
Cox regression -- no ties
No. of subjects = 40
No. of failures = 12
Time at risk = 1258.299998
Number of obs = 40
LR chi2(1) = 3.48
Prob > chi2 = 0.0621
Log likelihood = -34.670661
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
fdg0	6.53e-17	5.87e-09	-0.00	1.000	0

Hazard Ratio equals zero

Wald test and CI's have broken down

## Interpretation

“Zero of four subjects with a SUV of 0 died while 12/36 subjects with SUV > 0 died (hazard ratio = 0); the effect was borderline statistically significant (p=0.06)”

## Reverse the Reference

fdg\_gt0: I = SUV > 0, 0 if SUV=0

LR test is the same

```

Cox regression -- no ties
No. of subjects = 40          Number of obs = 40
No. of failures = 12        LR chi2(1) = 3.48
Time at risk = 1258.299998  Prob > chi2 = 0.0621
Log likelihood = -34.670661
-----+-----
      _t | Haz. Ratio  Std. Err.      z    P>|z|   [95% Conf. Interval]
-----+-----
fdg_gt0 | 2.07e+15    6.95e+22    0.00    1.000    0
    
```

Hazard Ratio equals  $\infty$

Wald test and CI's still don't work

## Zero/Infinite HR

- Two sides of the same coin  
*depends on reference*
- Category has either 0% or 100% events  
*often happens with lots of categories*
- Use likelihood ratio tests: they're fine  
*Wald test performs poorly*
- Confidence intervals: see statistician  
*who can calculate likelihood ratio based CI*
- Can consolidate categories to handle the issue

## PBC data: age (in days) as predictor

```

No. of subjects = 312          Number of obs = 312
No. of failures = 125        LR chi2(1) = 20.51
Time at risk = 1713.853528  Prob > chi2 = 0.0000
Log likelihood = -629.72592
-----+-----
      _t | Haz. Ratio  Std. Err.      z    P>|z|   [95% Conf. Interval]
-----+-----
agedays | 1.00011    .0000241    4.54    0.000    1.000062    1.000157
    
```

HR is nearly 1

Wald and LR tests highly significant

## PBC data: age (decades) as predictor

```

No. of subjects = 312          Number of obs = 312
No. of failures = 125        LR chi2(1) = 20.51
Time at risk = 1713.853528  Prob > chi2 = 0.0000
Log likelihood = -629.72592
-----+-----
      _t | Haz. Ratio  Std. Err.      z    P>|z|   [95% Conf. Interval]
-----+-----
age_decades | 1.491811    .1314533    4.54    0.000    1.250188    1.773041
    
```

HR is greater!

Wald and LR tests exactly the same

## Continuous Predictors

- HR greatly affected by the scale of measurement (e.g., age in decades or days)
- Statistical significance is unaffected because SE is proportional to coefficient
- Choose interpretable unit change in predictor
- Can rescale by
  - (1) defining new variable
  - (2) using `lincom`
  - (3) direct calculation

### (1) Define new variable

- Let `agedays` be age in days
  - `gen age_decade=agedays/(3650)` *About 3650 days per decade*
  - `stcox age_decade` *Gives HR for one-decade older*
- Works for every regression -- always
- Dividing by -3650: effect of one decade younger
- The most simple method
- Recommend

### (2) Lincom

- Let `agedays` be age in days
  - `stcox agedays`
  - `lincom 3650*agedays, hr` *A 1-unit change in decade is 3650 unit change in days*
- The above gives the effect of a decade (or effect of being 3650 days older)
- The HR option is important *otherwise get coefficient not the HR*
- less recommend, since *not intuitive and easy to make mistakes*

### (3) Direct Calculation

- Let  $HR_{agd}$  be the HR for age in days
  - HR for decade =  $(HR_{agd})^{3650}$
  - HR for conf limits: also raised to 3650
  - test, p-values exactly the same
- HR for  $k$ -days =  $(HR_{agd})^k$   
*k is any arbitrary number, even negative #*
- Little need to use this method  
*useful to know what calculations are going on*

### (3) Direct Calculation

	HR	Lower 95% CI	Upper 95% CI	Wald p-value	LR p-value
Day	1.0011	1.000062	1.00016	<0.0001	<0.0001
Decade	$1.011^{3650}$ = 1.49	$1.000062^{3650}$ = 1.26	$1.00016^{3650}$ = 1.77	same as above	same as above

### Confounding in the Cox Model

- Handled the same way as other regression models
- Confounders added into model
- Interpretation: HR of a 1-unit change holding all other predictors constant
- All predictors adjust for each other

### UNOS Kidney Example

- Interest: How recipients from cadaveric donors do compared to living kidney recipients
- crude HR = 1.97, 95% CI (1.63, 2.40)
- What might vary between living/cadaveric recipients : HLA match (0-2 loci v. 3+), year of transplant, previous transplant?
- Could lead to inflated crude HR



## Interaction

- Addressed the same way across regression
  - Create product terms
  - Test of product terms reveals interaction
  - Understand interaction through series of `lincom`
- Predictors of graft failure in UNOS
- Is there an interaction between previous transplant and year of transplant?

## Results

```

No. of subjects =      9678      Number of obs =      9678
No. of failures =      2501
Time at risk   = 38123.04385
Log likelihood = -20488.046      LR chi2(3) =      252.04
                                      Prob > chi2 =      0.0000
-----+-----
      _t | Haz. Ratio  Std. Err.      z  P>|z|  [95% Conf. Interval]
-----+-----
  prevtx |  8.39e+56   2.75e+58    4.00  0.000   1.14e+29   6.17e+84
    year |  1.047994   .0091017    5.40  0.000   1.030306   1.065986
     prod |  .9367223   .0153842   -3.98  0.000   .9070499   .9673652
-----+-----

```

## What gives?

- There is a huge HR for `prevtx`. Isn't this an example of colinearity?
  - There might be some colinearity. It is a minor issue
  - The big issue: the HR gives the effect of `prevtx` when all other predictors are equal to zero
- It's huge because it's a meaningless extrapolation!

## HR Interpretation

- `prevtx`: HR of previous transplant in year 0
- `year`: HR of year of tx with no prev tx effect of +1 year when `prevtx=0`
- `prod`: HR is not easily interpreted

## Advice

- Don't fixate on the model HRs
  - HRs may not correspond to something meaningful (sometimes yes, sometimes not)
  - Instead: look at test for product term
  - Followed by a series of `lincom` statements
- If you do this, there is no colinearity issue

## Comparisons

1. What is the effect of `prevtx` in 1990?
2. What is the effect of `prevtx` in 1995?
3. What is the effect of `prevtx` in 2000?

### Effect of Previous Transplant in 1990

	prevtx	year	prod
previous transp.	1	1990	1990
no prev tx	0	1990	0
diff	1	0	1990

lincom 1\*prevtx + 1990\*prod, hr

### Lincom

```

Effect of Previous transplant in 1990
. lincom prevtx + 1990*prod, hr
-----+-----
   _t | Haz. Ratio  Std. Err.   z  P>|z|  [95% Conf. Interval]
-----+-----
(1) |  2.686016   .1950779   13.60  0.000   2.329637   3.096914

Effect of Previous transplant in 1995
. lincom prevtx + 1995*prod, hr
-----+-----
   _t | Haz. Ratio  Std. Err.   z  P>|z|  [95% Conf. Interval]
-----+-----
(1) |  1.937148   .1054406   12.15  0.000   1.74113   2.155234

Effect of Previous transplant in 2000
. lincom prevtx + 2000*prod, hr
-----+-----
   _t | Haz. Ratio  Std. Err.   z  P>|z|  [95% Conf. Interval]
-----+-----
(1) |  1.397066   .1661109    2.81  0.005   1.106648   1.7637
    
```

### Interpretation

The effect of previous transplant on risk of graft failure varies by year of transplant ( $p < 0.001$ ).

The relative hazards (and 95% Conf. Int.) for the previous transplant are

2.7 (3.1-2.3), 1.9 (2.2-1.7) and 1.4 (1.8-1.1), in the years 1990, 1995 and 2000, respectively.

### Centered Regression

Year centered at 1996

```

Cox regression -- Breslow method for ties
No. of subjects = 9678           Number of obs = 9678
No. of failures = 2501
Time at risk = 38123.04385
Log likelihood = -20488.046      LR chi2(3) = 252.04
                                Prob > chi2 = 0.0000

   _t | Haz. Ratio  Std. Err.   z  P>|z|  [95% Conf. Interval]
-----+-----
prevtx |  1.81457   .1164537   9.28  0.000   1.600096   2.057791
cyear  |  1.043994   .0091017    5.40  0.000   1.030306   1.065986
cprod  |  .9367223   .0153842   -3.98  0.000   .9070499   .9673652
    
```

Only prevtx has changed: corresponds to effect of prev transplant in 1996

### Effect of Previous Transplant in 1990

	prevtx	cyear	cprod
previous transp.	1	-6	-6
no prev tx	0	-6	0
diff	1	0	-6

lincom 1\*prevtx + -6\*cprod, hr

### Lincom (centered)

```

Effect of Previous transplant in 1990
. lincom prevtx-6*cprod, hr
-----+-----
   _t | Haz. Ratio  Std. Err.   z  P>|z|  [95% Conf. Interval]
-----+-----
(1) |  2.686016   .1950779   13.60  0.000   2.329637   3.096914

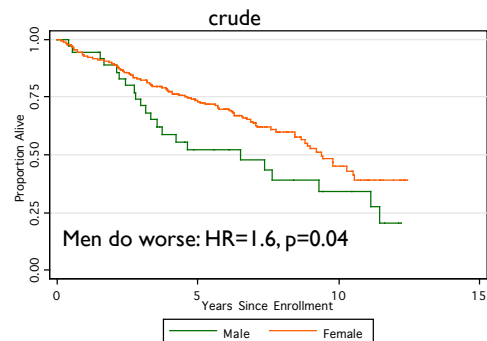
Effect of Previous transplant in 1995
. lincom prevtx-1*cprod, hr
-----+-----
   _t | Haz. Ratio  Std. Err.   z  P>|z|  [95% Conf. Interval]
-----+-----
(1) |  1.937148   .1054406   12.15  0.000   1.74113   2.155234

Effect of Previous transplant in 2000
. lincom prevtx+4*cprod, hr
-----+-----
   _t | Haz. Ratio  Std. Err.   z  P>|z|  [95% Conf. Interval]
-----+-----
(1) |  1.397066   .1661109    2.81  0.005   1.106648   1.7637
    
```

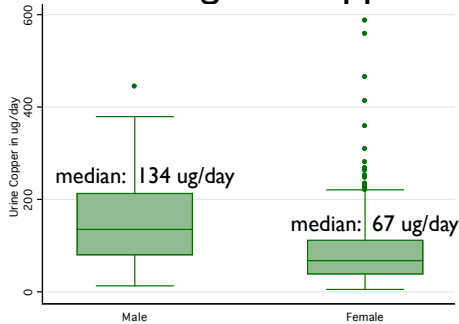
## Interaction

- Same advice as in previous models: *create product, test product, calculate lincom*
- Colinearity issue is specious
- Centering gives same test for interaction and doesn't change `lincom`
- Center makes some coefficients more interpretable. Can make `lincom` subject to mistake (if forget that centered variable)
- Don't forget "hr" in those `lincom`

## Effect of Sex: PBC data



## Men: Higher Copper



## Adjusted Survival Curves

- Would like to visualize the adjusted effects of variables
- Can make survival prediction based on a Cox model
- $S(t|x)$ : survivor function for someone with predictors  $x$
- $S(t|x)$ : proportion event free at time  $t$  with predictors  $x$

## Under the Cox Model

$$S(t|x) = \{ S_0(t) \}^{\exp(\beta_1 x_1 + \dots + \beta_p x_p)}$$

$\beta$  are the coefficients from the Cox model

$S_0(t)$ : baseline survivor function

$S_0(t)$ : survivor function when all predictors equal zero

In Cox model we see estimates of  $\exp(\beta_p)$

In background, Stata calculates estimates of  $S_0(t)$

## Adjusted Curve

- Look at effect of  $x_1$  (sex) adjusting for  $x_2$  (copper)
- Create two curves with same value for  $x_2$  otherwise we are not *adjusting* for copper adjustment: effect of sex w/ copper constant
- But differing by sex!
- But what value for  $x_2$ ? This value will affect the curves
- Let's use overall mean

## Adjusted Curves

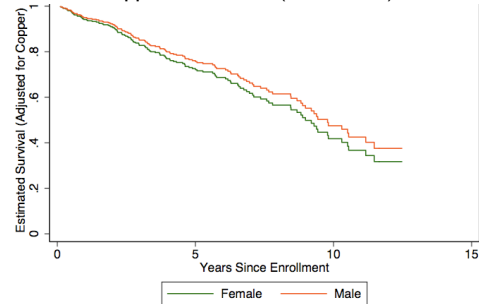
```
. stcox sex copper, basesurv(xx)
                                     basesurv(xx): makes Stata
                                     save the baseline survival in a
                                     variable "xx"

. stcurve, survival at1(sex=1) at2(sex=0)
                                     stcurve: gives predicted curves
                                     survival: graph survival (not hazard)
                                     at1: (value for curve 1)
                                     at2: (value for curve 2)
                                     copper default: fixed at overall mean

. stcurve, survival at1(sex=1 copper=97.6) at2(sex=0 copper=97.6)
```

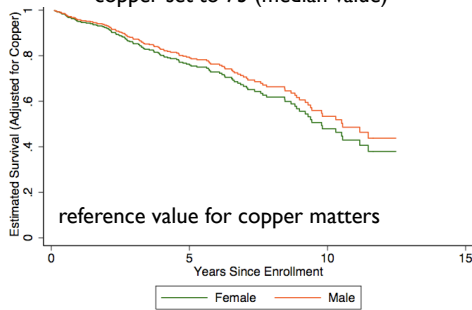
## Adjusted Curves

copper set to 97.6 (mean value)



## Adjusted Curves

copper set to 73 (median value)



## Adjusted/Predicted Curves

- Can be useful for visualizing effect of predictor
- Must choose reference values for confounders
  - often choose mean for continuous variable
  - most common category for categorical
- “stcurve” is a flexible tool for creating adjusted or predicted survival curves