Survival Analysis: Kaplan-Meier Method

Tuan V. Nguyen
Professor and NHMRC Senior Research Fellow
Garvan Institute of Medical Research
University of New South Wales
Sydney, Australia

What we will learn …

- Cox's proportional hazards model
- Computing the hazard ratio
- Adjusted survival curves using the Cox PH model
- The meaning of the PH assumption

Leukemia remission data

Leukemia Remission Data

Group 1 ($n = 21$)		Group 2 ($n = 21$)	
t(weeks)	log WBC	t(weeks)	log WBC
6	2.31	1	2.80
6	4.06	1	5.00
6	3.28	2	4.91
7	4.43	2	4.48
10	2.96	3	4.01
13	2.88	4	4.36
16	3.60	4	2.42
22	2.32	5	3.49
23	2.57	5	3.97
6+	3.20	8	3.52
9+	2.80	8	3.05
10+	2.70	8	2.32
11+	2.60	8	3.26
17+	2.16	11	3.49
19+	2.05	11	2.12
20+	2.01	12	1.50
25+	1.78	12	3.06
32+	2.20	15	2.30
32+	2.53	17	2.95
34+	1.47	22	2.73
35+	1.45	23	1.97

+ denotes censored observation

```
group = c(1,1,1,1,1, 1,1,1,1,1, 1,1,1,1,1,
1,1,1,1,1,1, 0,0,0,0,0, 0,0,0,0,0,
0,0,0,0,0, 0,0,0,0,0,0)

time =
c(6,6,6,7,10,13,16,22,23,6,9,10,11,17,19,2
0,25,32,32,34,35,1,1,2,2,3,4,4,5,5,8,8,8,8
,11,11,12,12,15,17,22,23)

status =
c(0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,
1, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0,
0,0,0,0,0,0)

wbc =
c(2.31,4.06,3.28,4.43,2.96,2.88,3.60,2.32,
2.57,3.20,2.80,2.70,2.60,2.16,2.05,2.01,1.
78,2.20,2.53,1.47,1.45,
2.80,5.00,4.91,4.48,4.01,4.36,2.42,3.49,3.
97,3.52,3.05,2.32,3.26,3.49,2.12,1.50,3.06
,2.30,2.95,2.73,1.97)

dat=data.frame(group,time,status,wbc)
```

Kaplan-Meier curves

```
km = survfit(Surv(time, status==0) ~ group); km

plot(km, lty=c(1,4), lwd=2, xlab="Weeks", ylab="S(t)")

legend("topright", c("Placebo", "Treatment"),
lty=c(1,4), lwd=2)
```

Leukemia remission data

- Outcome: time (+ status)
- Covariates / predictors: group, wbc

Log-rank test and Cox's model

- **Log-rank test:**
  - **No covariates**
- **Cox's proportional hazards model**
  - **Adjusted for covariates**

Cox's proportional hazards model

$$h(t, \mathbf{X}) = h_0(t)\, e^{\sum_{i=1}^{p} \beta_i X_i}$$

- **X = (X_1, X_2,…, X_p) , explanatory/predictor variables**
- **$h_0(t)$: baseline hazard, involves t but not X's**
- **exp($\beta_i X_i$) : exponential, involves X, but not t (Xs can be time-dependent)**

Cox's proportional hazards model

- Hazard(t) = (baseline risk) x (effects of covariates)
- Semi-parametric model

Proportional hazards model

- Let $X = (x_1, \ldots, x_p)$ - the set of predictors
- $h(t|x)$: hazard of someone with predictors x
- $h(t|x) = h_0(t) \exp(\beta_1 x_1 + \cdots + \beta_p x_p)$
- $h(t|x)/h_0(t) = \exp(\beta_1 x_1 + \cdots + \beta_p x_p)$
- $\log(\, h(t|x)\,) = \log(\, h_0(t)\,) + \beta_1 x_1 + \cdots + \beta_p x_p$ because $\log(a/b) = \log(a) - \log(b)$
- Much like logistic regression but change *odds* to *hazards*

Cox's model

- The "baseline" hazard $h_0(t)$ is unspecified

  plays the role of intercept

- Predictor effects in terms of hazard ratios

  relative rates of failure

- Don't need to know $h_0(t)$

  to understand these predictor effects

- Effect of one unit increase in predictor x_p is to multiply hazard by $\exp(\beta_p)$

  holding all other predictors constant

+1 unit change in x_p

$$h(t|x) = h_0(t)\exp(\beta_1 x_1 + \ldots + \beta_p\ V) \qquad x_p = V$$

versus

$$h(t|x) = h_0(t)\exp(\beta_1 x_1 + \ldots + \beta_p\ (V+1)) \qquad x_p = V+1$$

$$\text{Ratio} = \frac{h_0(t)\exp(\beta_1 x_1 + \ldots + \beta_p\ (V+1))}{h_0(t)\exp(\beta_1 x_1 + \ldots + \beta_p\ V)}$$

+1 unit change in x_p

$$\text{Ratio} = \frac{\cancel{h_0(t)}\exp(\beta_1 x_1 + \ldots + \beta_p\ (V+1))}{\cancel{h_0(t)}\exp(\beta_1 x_1 + \ldots + \beta_p V)}$$

$$= \frac{\exp(\beta_1 x_1 + \ldots + \beta_p\ (V+1))}{\exp(\beta_1 x_1 + \ldots + \beta_p V)}$$

Ratio does not depend on t !

+1 unit change in x_p

$$\text{Ratio} = \frac{\exp(\beta_1 x_1 + \ldots + \beta_p (V+1))}{\exp(\beta_1 x_1 + \ldots + \beta_p V)}$$

$$= \exp(\beta_1 x_1 + \ldots + \beta_p (V+1) - (\beta_1 x_1 + \ldots + \beta_p V))$$

because $\exp(a) / \exp(b) = \exp(a-b)$

$$= \exp(\beta_p (V+1) - \beta_p V)$$ b/c same other predictors

$$= \exp(\beta_p)$$ b/c $\beta_p V$ terms cancel

Hazard ratio

- β is the regression coefficient

  If no effect of a predictor variable then $\beta=0$

- HR for a unit increase in a predictor is $\exp(\beta)$

  If no effect of variable then $\exp(\beta)=1$

- A useful way to discuss predictor effects (increases or decreases Hazard by a factor)

Why Cox model?

- Can be fitted without an explicit model for the hazard
- Can model the effect of a continuous predictor
- Can model multiple predictors: *continuous, binary, categorical*
- Can adjust for confounders: *adjust by adding confounders to the model*
- Can incorporate interaction, mediation: *create and add product terms*
- Can detect and estimate predictors for patient-level prognosis

Comparison with other forms of regression

- ***Same issues as in linear and logistic regression:***

  predictor selection

- ***Differences:***

  interpretation, assumptions, model checking

Baseline data

```
group = c(1,1,1,1,1, 1,1,1,1,1, 1,1,1,1,1, 1,1,1,1,1,1,
0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0,0)

time =
c(6,6,6,7,10,13,16,22,23,6,9,10,11,17,19,20,25,32,32,34,35,1,1
,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23)

status = c(0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,
0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0,0)

wbc =
c(2.31,4.06,3.28,4.43,2.96,2.88,3.60,2.32,2.57,3.20,2.80,2.70,
2.60,2.16,2.05,2.01,1.78,2.20,2.53,1.47,1.45,
2.80,5.00,4.91,4.48,4.01,4.36,2.42,3.49,3.97,3.52,3.05,2.32,3.
26,3.49,2.12,1.50,3.06,2.30,2.95,2.73,1.97)

dat=data.frame(group,time,status,wbc)

baseline = Surv(time, status==0)

km = survfit(baseline ~ 1)

summary(km)
```

Baseline data

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	42	2	0.952	0.0329	0.8901	1.000
2	40	2	0.905	0.0453	0.8202	0.998
3	38	1	0.881	0.0500	0.7883	0.985
4	37	2	0.833	0.0575	0.7279	0.954
5	35	2	0.786	0.0633	0.6709	0.920
6	33	3	0.714	0.0697	0.5899	0.865
7	29	1	0.690	0.0715	0.5628	0.845
8	28	4	0.591	0.0764	0.4588	0.762
10	23	1	0.565	0.0773	0.4325	0.739
11	21	2	0.512	0.0788	0.3783	0.692
12	18	2	0.455	0.0796	0.3227	0.641
13	16	1	0.426	0.0795	0.2958	0.615
15	15	1	0.398	0.0791	0.2694	0.588
16	14	1	0.369	0.0784	0.2437	0.560
17	13	1	0.341	0.0774	0.2186	0.532
22	9	2	0.265	0.0765	0.1507	0.467
23	7	2	0.189	0.0710	0.0909	0.395

Cox's model using R

```
dat=data.frame(group,time,status,wbc)
baseline = Surv(time, status==0)
km = survfit(baseline)

cox = coxph(Surv(time, status==0) ~ group + wbc)
summary(cox)
```

Cox's model using R

```
  n= 42, number of events= 30

         coef exp(coef) se(coef)      z Pr(>|z|)
group -1.3861    0.2501   0.4248 -3.263   0.0011 **
wbc    1.6909    5.4243   0.3359  5.034  4.8e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      exp(coef) exp(-coef) lower .95 upper .95
group    0.2501     3.9991    0.1088    0.5749
wbc      5.4243     0.1844    2.8082   10.4776

Concordance= 0.852  (se = 0.062 )
Rsquare= 0.671   (max possible= 0.988 )
Likelihood ratio test= 46.71  on 2 df,   p=7.187e-11
Wald test            = 33.6  on 2 df,   p=5.061e-08
Score (logrank) test = 46.07  on 2 df,   p=9.921e-11
```

Interpretation of output

```
  n= 42, number of events= 30

          coef exp(coef) se(coef)      z Pr(>|z|)
group -1.3861    0.2501   0.4248 -3.263   0.0011 **
wbc    1.6909    5.4243   0.3359  5.034  4.8e-07 ***
```

- **Model**

$$h(t) = L_t \times e^{b1 \times group + b2 \times wbc}$$

- **$Risk(t) = L_t \times e^{-1.3861 \times group + 1.6909 \times wbc}$**

group: 0 = control, 1 = treatment

Interpretation of output

```
      exp(coef) exp(-coef) lower .95 upper .95
group    0.2501     3.9991    0.1088    0.5749
wbc      5.4243     0.1844    2.8082   10.4776
```

- Risk of remission in the treatment group is 75% lower than that in the controls (hazard ratio: 0.25; 95% CI 0.11 to 0.57)
- Risk of remission increased by 5.42 folds (95% CI 2.81 to 10.48) for each unit increase in wbc

Interpretation of output

```
Concordance= 0.852  (se = 0.062 )
Rsquare= 0.671   (max possible= 0.988 )
Likelihood ratio test= 46.71  on 2 df,   p=7.187e-11
Wald test            = 33.6  on 2 df,   p=5.061e-08
Score (logrank) test = 46.07  on 2 df,   p=9.921e-11
```

- Predicted risk and observed risk agree 85.2% of the times
- Group and wbc “explained” 67.1% variance in the risk of remission

What about “baseline risk” ?

- $\text{Risk}(t) = L_t \times e^{-1.3861 \times \text{group} + 1.6909 \times \text{wbc}}$

  group: 0 = control, 1 = treatment

- What is L_t ?
- L_t = baseline risk
- Can be estimated from

```
baseline = Surv(time, status==0)
km = survfit(baseline ~ 1)
summary(km)
```

Baseline probability of remission

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	42	2	0.952	0.0329	0.8901	1.000
2	40	2	0.905	0.0453	0.8202	0.998
3	38	1	0.881	0.0500	0.7883	0.985
4	37	2	0.833	0.0575	0.7279	0.954
5	35	2	0.786	0.0633	0.6709	0.920
6	33	3	0.714	0.0697	0.5899	0.865
7	29	1	0.690	0.0715	0.5628	0.845
8	28	4	0.591	0.0764	0.4588	0.762
10	23	1	0.565	0.0773	0.4325	0.739
11	21	2	0.512	0.0788	0.3783	0.692
12	18	2	0.455	0.0796	0.3227	0.641
13	16	1	0.426	0.0795	0.2958	0.615
15	15	1	0.398	0.0791	0.2694	0.588
16	14	1	0.369	0.0784	0.2437	0.560
17	13	1	0.341	0.0774	0.2186	0.532
22	9	2	0.265	0.0765	0.1507	0.467
23	7	2	0.189	0.0710	0.0909	0.395

Probability of remission with covariates

- Can estimate the probability of remission for
  - Any time point
  - A given group
  - AND a given wbc level

$$Risk(t) = L_t \times e^{-1.3861 \times group + 1.6909 \times wbc}$$

Estimation of risk probability

- Step 1: determine baseline risk during a period
- Step 2: calculate the average "linear term" (M)
- Step 3: calculate the individual "linear term" (L)
- Step 4: calculate D = L-M
- Step 5: calculate the risk of event

Estimation of risk probability: example

Example: An individual on treatment =1 with wbc = 3.0

What is the individual's probability of remission at 10 weeks?

Step 1 – determine baseline risk

- 10-week baseline "survival" is 0.565

$$S_0(10) = 0.565$$

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	42	2	0.952	0.0329	0.8901	1.000
2	40	2	0.905	0.0453	0.8202	0.998
3	38	1	0.881	0.0500	0.7883	0.985
4	37	2	0.833	0.0575	0.7279	0.954
5	35	2	0.786	0.0633	0.6709	0.920
6	33	3	0.714	0.0697	0.5899	0.865
7	29	1	0.690	0.0715	0.5628	0.845
8	28	4	0.591	0.0764	0.4588	0.762
10	23	1	0.565	0.0773	0.4325	0.739
….						

Step 2 – calculate the average "linear term" (M)

- Using the mean of risk factors to calculate M
- Mean group = 0.5
- Mean wbc = 2.93
- M = (-1.3861 x 0.5) + (1.6909 x 2.93)

  = 4.261

Step 3 - calculate the individual “linear term” (L)

- Using the individual’s group and wbc to calculate L
- group = 1 (on treatment)
- wbc = 3.0
- L = (-1.3861 x 1) + (1.6909 x 3)

  = 3.6866

Step 4 - calculate D = L-M

- Difference between the individual's linear term and average linear term

$$D = L - M$$

$$= 3.6866 - 4.261$$

$$= -0.5744$$

Step 5 - calculate the risk of event

We want to estimate the 10-week risk for the individual:

$$\text{Risk}(10) = 1 - [S_0]^{\exp(d)}$$

$$= 1 - 0.565^{\exp(-0.5744)}$$

$$= 0.682$$

The risk of remission at week 10 is 68.2%.

What is the risk of remission at week 10 for a control patient?

Exponential and Weibull models

Comparison

```
exponential = survreg(Surv(time, status==0) ~ group + wbc,
dist="exponential")

summary(exponential)


weibull = survreg(Surv(time, status==0) ~ group + wbc)

summary(weibull)
```

Summary

- Time-to-event data
- Kaplan-Meier analysis (actuarial analysis)
- Cox’s regression allows an assessment of risk factors
- Cox’s regression provides a very useful prognostic model in clinical medicine