Survival Analysis in R

于怡　yuyi1227@Gmail.com
Ph.D. Candidate
Of Mathematical Statistics,
Fudan Univ.

Outline

- What is Survival Analysis
- An application using R: PBC Data

  With Methods in Survival Analysis

  Kaplan-Meier Estimator

  Mantel-Haenzel Test (log-rank test)

  Cox regression model (PH Model)

What is Survival Analysis

- Model time to event (esp. failure)
- Widely used in medicine, biology, actuary, finance, engineering, sociology, etc.
- Able to account for censoring
- Able to compare between 2+ groups
- Able to access relationship between covariates and survival time

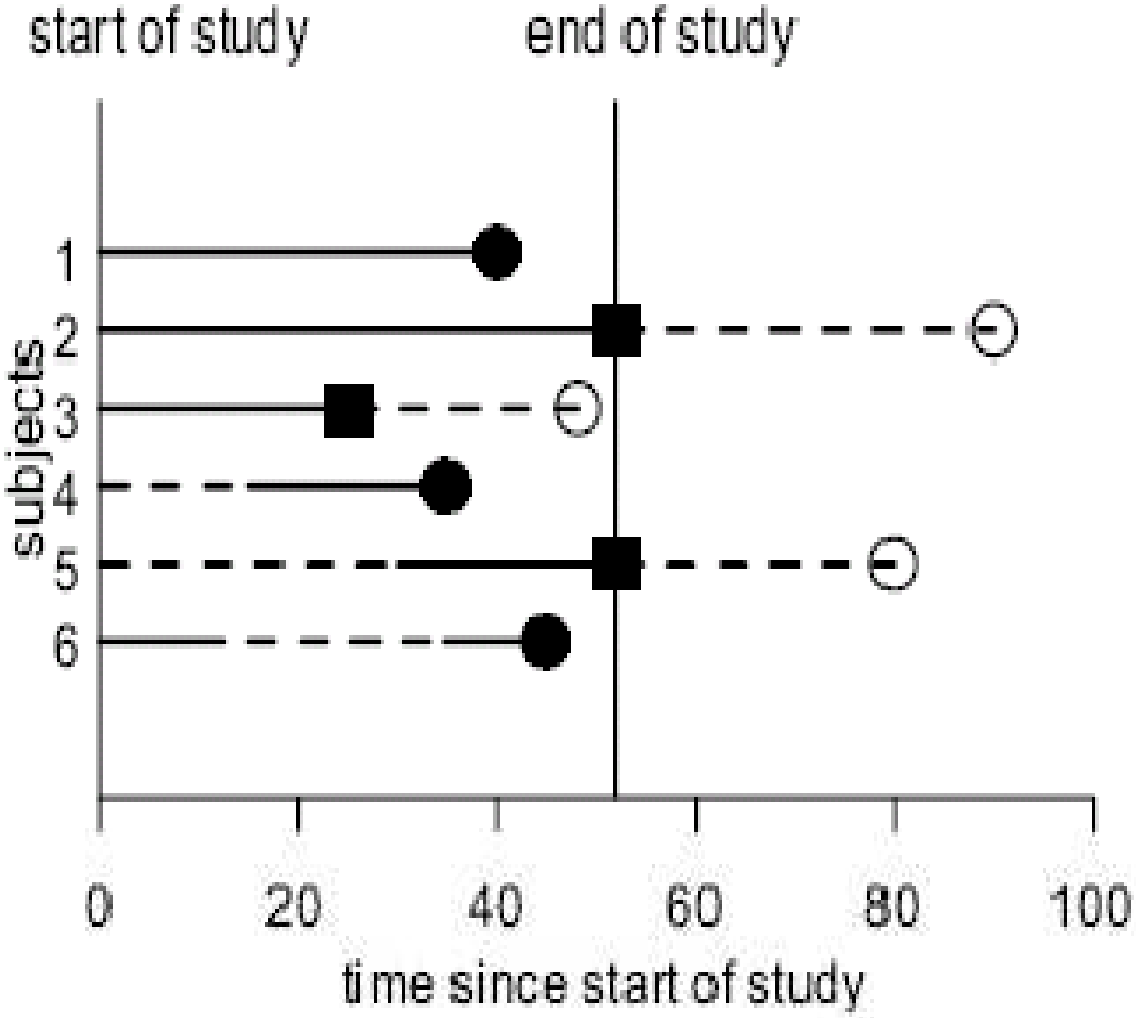


Figure 1. Data from an imagined study illustrating various kinds of subject histories: Subject 1, uncensored; 2, fixed-right censoring; 3, random-right censoring; 4 and 5, late entry; 6, multiple intervals of observation.

An application using R: PBC Data

Primary Biliary Cirrhosis

The data is from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases in the data set participated in the randomized trial and contain largely complete data. The additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis, so the data here are on an additional 106 cases as well as the 312 randomized participants. Missing data items are denoted by a period.

```
>data(pbc, package = "survival")
```

number	futime	status	drug	age	sex	edema	bilirubin		albumin	protime
1	400	2	1	21464	1	1	14.5	2.6	12.2	
2	4500	0	1	20617	1	0	1.1	4.14	10.6	
3	1012	2	1	25594	0	0.5	1.4	3.48	12	
4	1925	2	1	19994	1	0.5	1.8	2.54	10.3	
5	1504	1	0	13918	1	0	3.4	3.53	10.9	
6	2503	2	0	24201	1	0	0.8	3.98	11	
7	1832	0	0	20284	1	0	1	4.09	9.7	
8	2466	2	0	19379	1	0	0.3	4	11	
9	2400	2	1	15526	1	0	3.2	3.08	11	
10	51	2	0	25772	1	1	12.6	2.74	11.5	

Survival Analysis in R

- Package: survival

  ```
  >library (survival)
  ```

- Create a survival subject: `Surv`
- Kaplan-Meier Estimator: `survfit`
- Mantel-Haenzel Test: `survdiff`
- Cox Model: `coxph`

Creating the survival object

- Created by `Surv` function
- Usage

```
>Surv (time, time2, event, type=c
  ('right', 'left', 'interval',
  'counting', 'interval2'), origin=0)
```

- In our example

```
>Surv (pbc$time, pbc$status==2)
```

- Reference

```
>help (Surv)
```

Creating the survival object

- ```
  >surdays<-with (pbc, Surv
  (pbc$time, pbc$status==2))
  >surdays
  ```

```
 [1]  400  4500+ 1012  1925  1504+ 2503  1832+ 2466  2400    51
[26] 1444    77   549  4509+  321  3839  4523+ 3170  3933+ 2847
[51] 3853  2386  1000  1434  1360  1847  3282  4459+ 2224  4365+
[76]   71   326  1690  3707+  890  2540  3574  4050+ 4032+ 3358
[101] 3581+ 3099+  110  3086  3092+ 3222  3388+ 2583  2504+ 2105
[126]  824  3255+ 1037  3239+ 1413   850  2944+ 2796  3149+ 3150+
[151] 2870+ 1152  2863+  140  2666+  853  2835+ 2475+ 1536  2772+
[176] 1492  2609+ 2580+ 2573+ 2563+ 2556+ 2555+ 2241+  974  2527+
[201] 2294+ 2272+ 2221+ 2090  2081  2255+ 2171+  904  2216+ 2224+
[226] 1978+  999  1967+  348  1979+ 1165  1951+ 1932+ 1776+ 1882+
[251] 1457+ 1770+ 1765+  737+ 1735+ 1701+ 1614+ 1702+ 1615+ 1656+
[276] 1481+ 1434+ 1420+ 1433+ 1412+   41  1455+ 1030+ 1418+ 1401+
```
```

Kaplan-Meier Estimator

- Also known as product-limit estimator
- Just like the censoring version of empirical survival function
- Generate a stair-step curve
- Variance estimated by Greenwood's formula
- Does not account for effect of other covariates

Kaplan-Meier Estimator (Cont.)

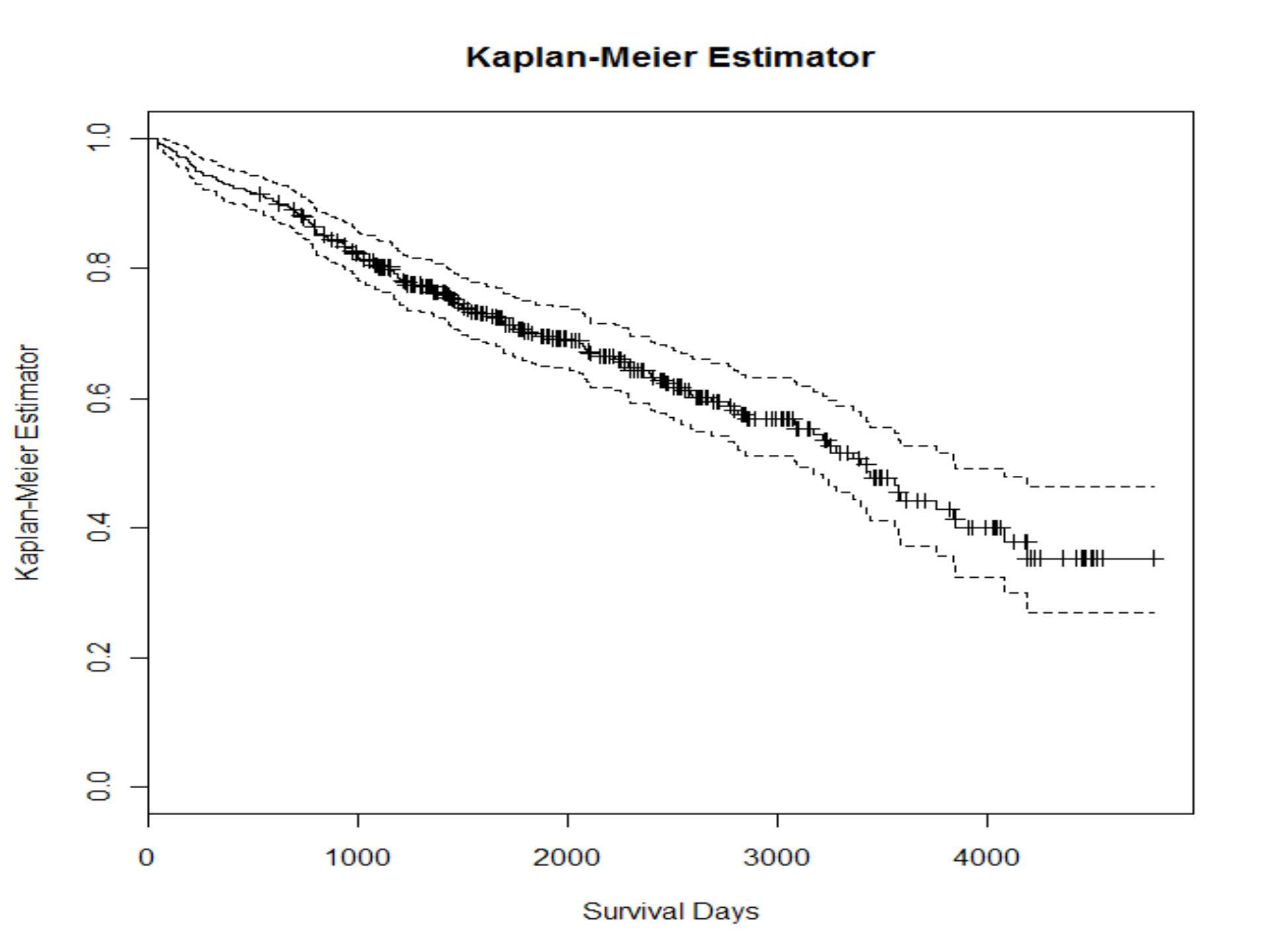
- Computed by the function: `survfit`
- Usage

  `>survfit (formula, …)`

- In our example

```
> survfit(Surv(pbc$time,pbc$status==2)~1)
Call: survfit(formula = Surv(pbc$time, pbc$status == 2) ~ 1)

records   n.max n.start  events  median 0.95LCL 0.95UCL
    418     418     418     161    3395    3090    3853
```


Kaplan-Meier Estimator
Kaplan-Meier Estimator
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
Survival Days

Mantel-Haenzel Test

- Also known as log-rank test
- Generated from a sequence of 2×2 tables
- Conditional independence
- Efficient in comparing groups differed by categorical variables, but not continuous ones

Mantel-Haenzel Test (Cont.)

- Computed by the function: `survdiff`
- Usage

  `>survdiff (formula, data, subset, na.action, rho=0)`

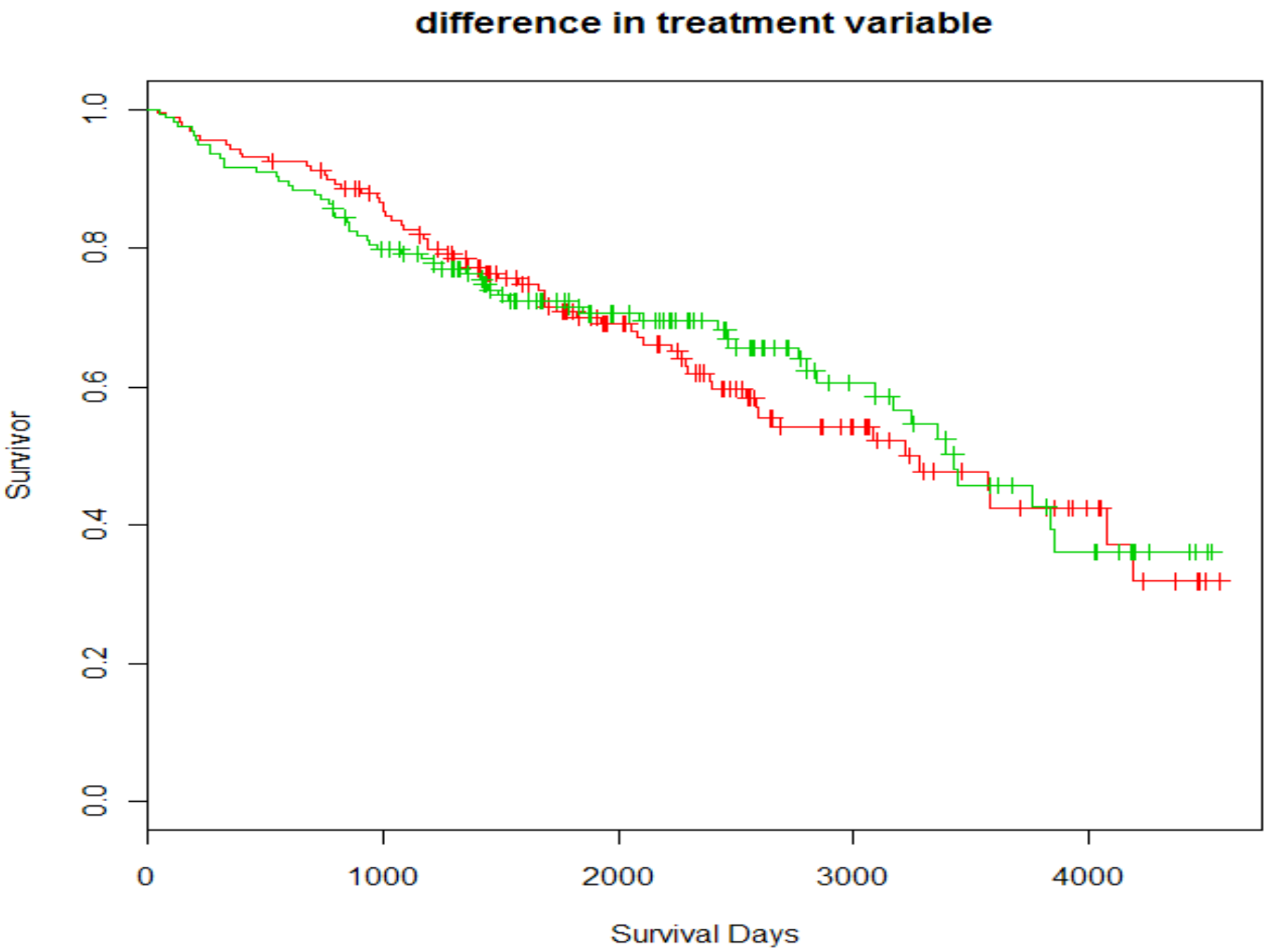
- In our example

```
> survdiff(Surv(pbc$time,pbc$status==2)~pbc$trt)
Call:
survdiff(formula = Surv(pbc$time, pbc$status == 2) ~ pbc$trt)

n=312, 106 observations deleted due to missingness.

             N Observed Expected (O-E)^2/E (O-E)^2/V
pbc$trt=1  158       65     63.2    0.0502     0.102
pbc$trt=2  154       60     61.8    0.0513     0.102

 Chisq= 0.1  on 1 degrees of freedom, p= 0.75
```


difference in treatment variable
Survivor
1.0
0.8
0.6
0.4
0.2
0.0
0
1000
2000
3000
4000
Survival Days

Cox Model

- Also known as proportional hazard model
- Conveniently access the effect of continuous and categorical variables
- Using partial likelihood to get inference even without knowledge of baseline hazard
- Assumption is quite strong…

Cox Model (Cont.)

- Computed by the function: `coxph`
- Usage:

```
>coxph (formula, data=, weights,
subset, na. action, init,
control, method=c
("efron","breslow","exact"),
singular. ok=TRUE, robust=FALSE,
model=FALSE, x=FALSE,
y=TRUE, ...)
```

```
> cfit <- coxph(Surv(time, status == 2) ~ age  + edema + log(bili)
+              + log(albumin) + log(protime), data = pbc)
> summary(cfit)
Call:
coxph(formula = Surv(time, status == 2) ~ age + edema + log(bili) +
    log(albumin) + log(protime), data = pbc)

  n=416 (2 observations deleted due to missingness)

                  coef exp(coef)  se(coef)      z Pr(>|z|)
age           0.039609  1.040404  0.007672  5.163 2.43e-07 ***
edema         0.896311  2.450547  0.271410  3.302 0.000959 ***
log(bili)     0.863551  2.371566  0.082941 10.412  < 2e-16 ***
log(albumin) -2.506923  0.081519  0.652916 -3.840 0.000123 ***
log(protime)  2.386839 10.879054  0.768509  3.106 0.001898 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

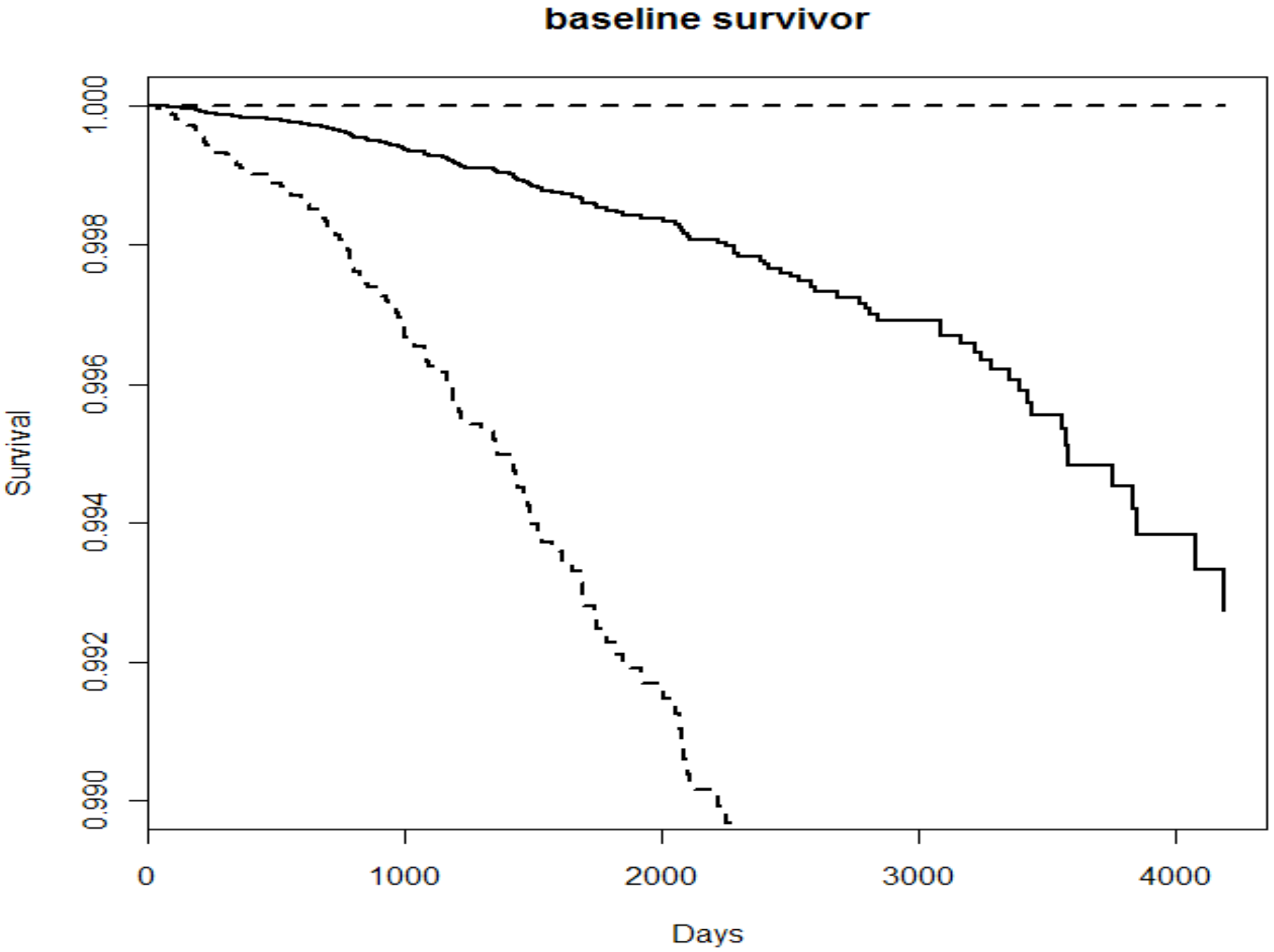
              exp(coef) exp(-coef) lower .95 upper .95
age             1.04040    0.96117   1.02488    1.0562
edema           2.45055    0.40807   1.43959    4.1715
log(bili)       2.37157    0.42166   2.01575    2.7902
log(albumin)    0.08152   12.26713   0.02267    0.2931
log(protime)   10.87905    0.09192   2.41232   49.0622

Rsquare= 0.426   (max possible= 0.985 )
Likelihood ratio test= 231  on 5 df,   p=0
Wald test            = 234.2  on 5 df,   p=0
Score (logrank) test = 301.8  on 5 df,   p=0
```

Cox Model (Cont.)

- For Baseline

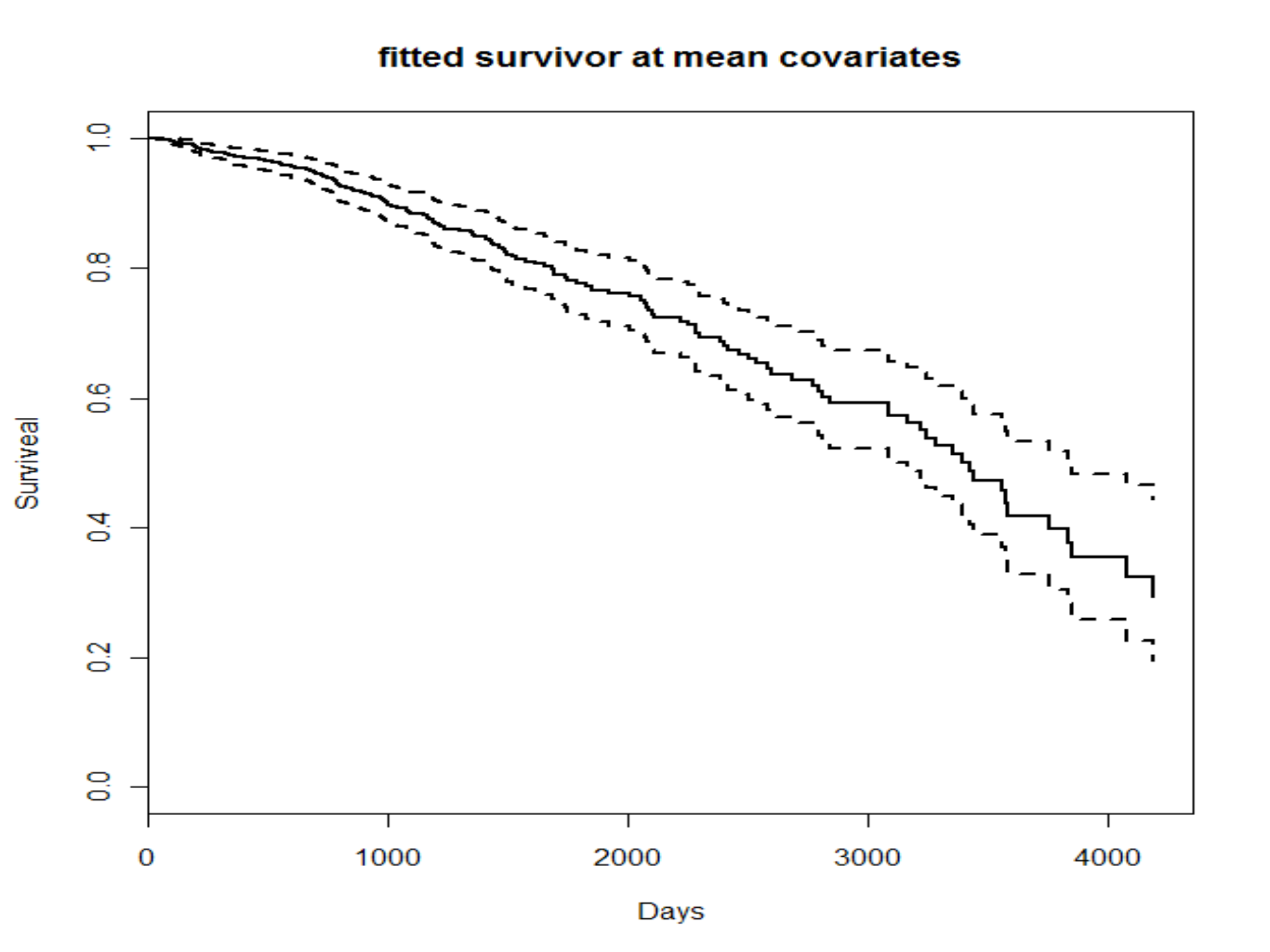
```
>pbc.null<-data.frame(age=rep(0,1),
edema=rep(0,1),bili=rep(1,1),albumin
=rep(1,1),protime=rep(1,1))
>plot(survfit(cfit,newdata=pbc.null),
lwd=2,ylim=c(.99,1),main='baseline
survivor‘,xlab ='Days',ylab=
'Survival',conf.int=T)
```


baseline survivor
Survival
1.000
0.998
0.996
0.994
0.992
0.990
0
1000
2000
3000
4000
Days

Cox Model (Cont.)

- For mean covariates

```
>plot(survfit(cfit),lwd=2,main=
 'fitted survival function at
mean covariates‘, xlab='Days',
ylab='Survival')
```


fitted survivor at mean covariates
Survival
1.0
0.8
0.6
0.4
0.2
0.0
0
1000
2000
3000
4000
Days

Diagnostic of Cox Model

- Cox model is amazing, but the assumption is really strong
- Schoenfeld residuals
- etc,.

Schoenfeld residuals

- Residuals are used to investigate the lack of fit of a model to a given subject.
- For Cox regression, there's no easy analog to the usual "observed minus predicted" residual of linear regression
- `>residuals(object, type=c("martingale", "deviance", "score", "schoenfeld", "dfbeta", "dfbetas", "scaledsch", "partial"), collapse=FALSE, weighted=FALSE, ...)`
- Schoenfeld (1982) proposed the first set of residuals for use with Cox regression packages
  - Schoenfeld D. Residuals for the proportional hazards regresssion model. *Biometrika*, 1982, 69(1):239-241.

Diagnostic of Cox Model (Cont.)

- Functions used here:

  `>residuals (object, ...)`

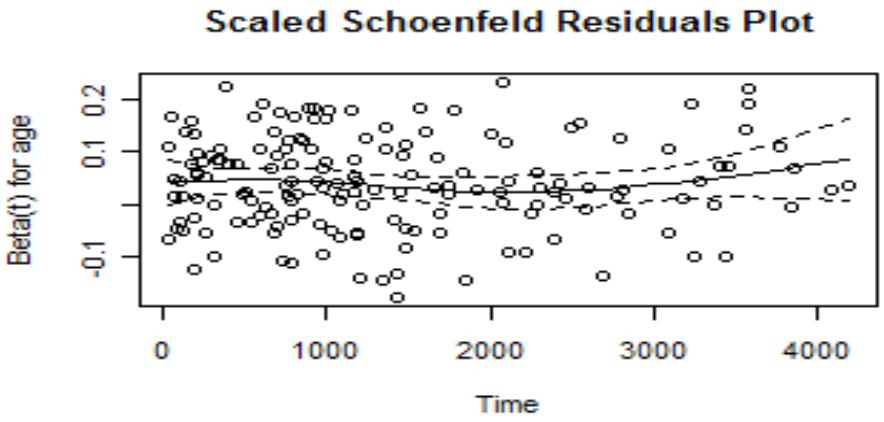
  to calculate different type residuals

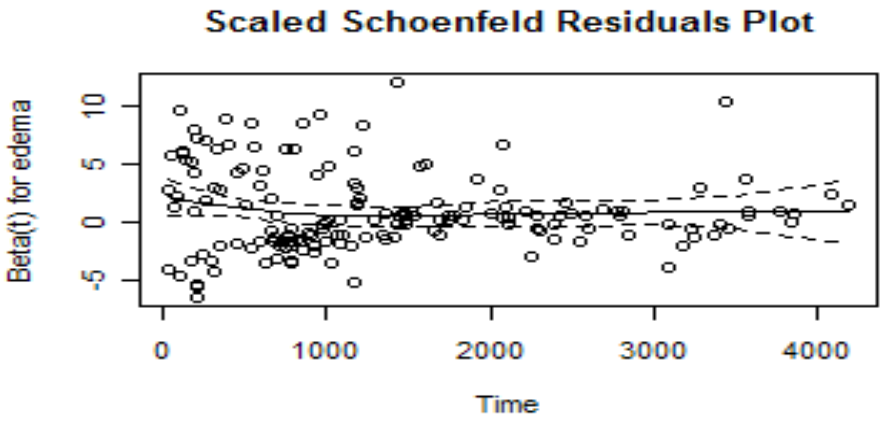
- `>cox.zph (fit, transform="km", global=TRUE)`

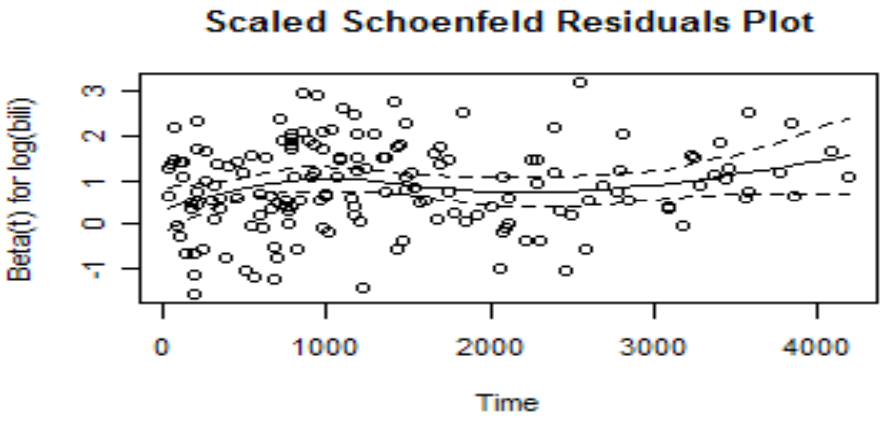
  to test the proportional hazards assumption for a Cox regression model fit.

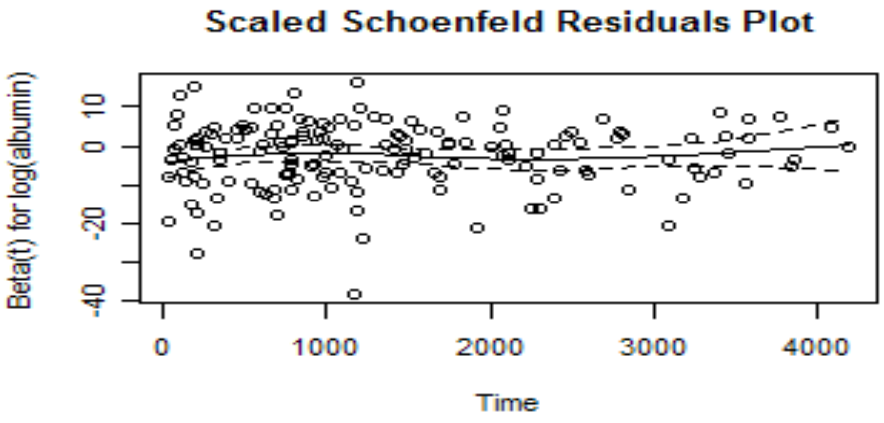
Diagnostic of Cox Model (Cont.)

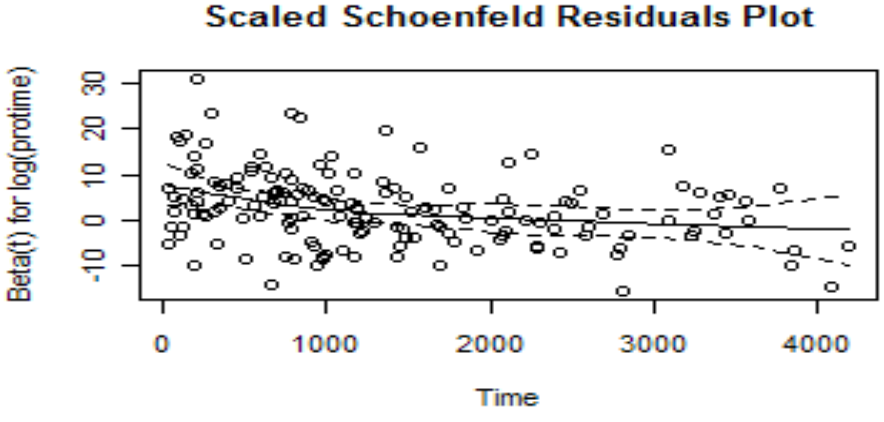
```
> cox.zph(cfit, transform = "identity")
                    rho    chisq       p
age             -0.00197 5.16e-04 0.98187
edema           -0.06684 7.26e-01 0.39432
log(bili)        0.11308 1.76e+00 0.18475
log(albumin)     0.02741 1.28e-01 0.72090
log(protime)    -0.30687 1.00e+01 0.00156
GLOBAL                NA 1.36e+01 0.01801
```


Scaled Schoenfeld Residuals Plot
Beta(t) for age
-0.1
0.1
0.2
0
1000
2000
3000
4000
Time


Scaled Schoenfeld Residuals Plot
Beta(t) for edema
-5
0
5
10
0
1000
2000
3000
4000
Time


Scaled Schoenfeld Residuals Plot
Beta(t) for log(bili)
-1
0
1
2
3
0
1000
2000
3000
4000
Time


Scaled Schoenfeld Residuals Plot
Beta(t) for log(albumin)
-40
-20
0
10
0
1000
2000
3000
4000
Time


Scaled Schoenfeld Residuals Plot
Beta(t) for log(protime)
-10
0
10
20
30
0
1000
2000
3000
4000
Time


THANK YOU