Using R for Basic Survival Analysis

Joseph Hogan

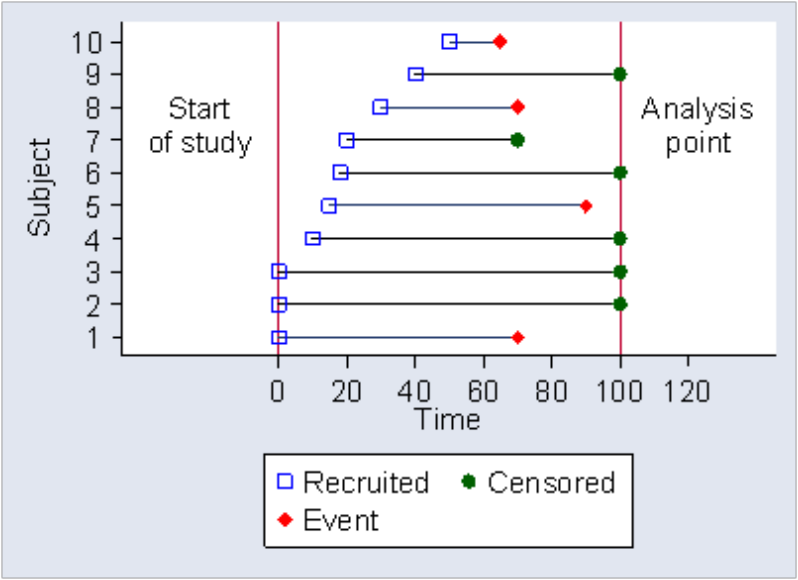
Department of Biostatistics
School of Public Health
Brown University

NAMBARI Workshop
June 23, 2017

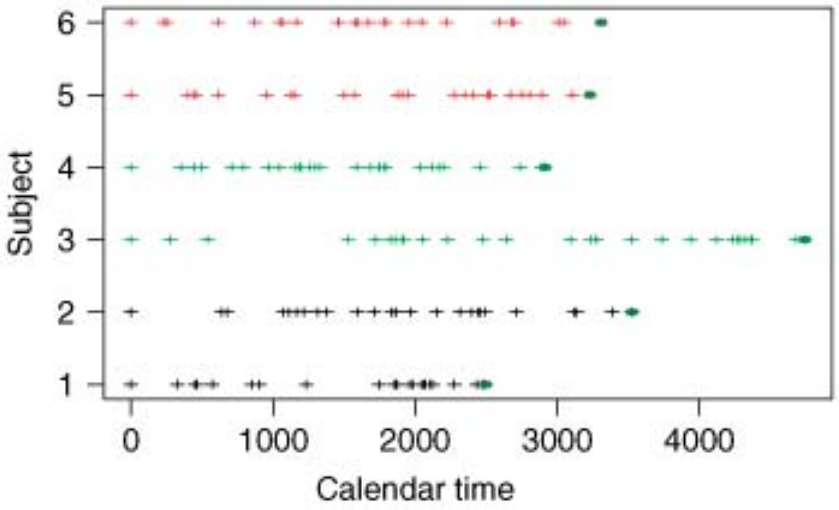
What makes 'survival data' different?

- Simplest version: time to event
  - Death, disease remission, etc
- More complicated versions: recurrent events
  - Clinic visits
  - Disease recurrence
- Always positively valued (greater than zero)
- Can be partially observed (censoring)

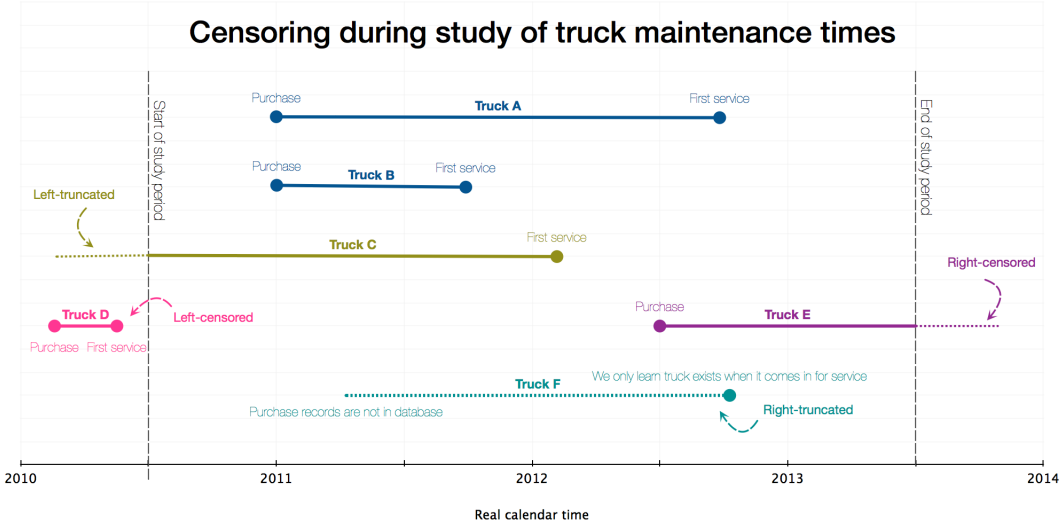
Simple time to event data (with censoring)



Recurrent event data



Different types of censoring



Important features of survival distribution

Suppose T is a time to event (survival time). The following two functions are important for summarizing the behavior of T

Survival function

$$\begin{aligned} S(u) &= \Pr(T > u) \\ &= \text{proportion surviving past time } u \end{aligned}$$

Hazard function

$$h(u) = \text{rate of failure at time } u \text{ among those still at risk}$$

If u is measured on a discrete time scale, the hazard is a conditional probability

$$h(u) = \Pr(T = u \mid T \geq u)$$

Statistical inference about survival distributions

- Summary plots
  - Survival curve, hazard curve
- Two-sample comparisons
  - log rank test
- Regression model
  - Cox proportional hazards regression

Data Example: Lung Cancer Survival Data

Description

- Cohort of 228 individuals with lung cancer
- Study was designed to evaluate prognostic value of patient- and physician-reported outcomes
  - Patient-reported performance scores
  - Caloric intake and weight loss

Data Example: Lung Cancer Survival Data

Objectives

- Summarize overall survival in the cohort
- Compare survival rates by gender and by cancer stage (ECOG Score)
- Fit regression having gender and stage as covariates
- Add patient-reported outcomes to see if they explain more variation
- Check proportional hazards assumption for regression

More information and link to dataset

Loprinzi et al. (1994). Journal of Clinical Oncology 12(3), 601–607.
`http://www.mayo.edu/research/documents/lunghtml/doc-10027247`

Load in dataset

1. Open `nambari.github.io`
   - Download `LungData.csv` from the website
2. Open RStudio
3. Open a new R script (follow directions)
4. Add these commands to the top of the R script
   - `library(survival)`
   - `library(survminer)`
   - `library(haven)`
5. Use the 'Import Dataset' tab to navigate to `LungData.csv` and import the data
6. Paste code generated by 'Import Dataset' into the top of the R script

Create survival object for graphs, models, etc.

```
# create survival object
Lung.km = survfit( Surv(time, status) ~ -1, data = LungData )
Lung.km

summary(Lung.km)
```

Structure of survival data

We will focus on data that may be *right censored*

Each individual has three data components:

$$(T, \Delta, \boldsymbol{X})$$

What these represent:

$$\begin{aligned} T &= \text{follow up time} \\ \\ \Delta &= \text{status indicator} \\ &= \begin{cases} 1 & \text{if } T \text{ is an event time} \\ 0 & \text{if } T \text{ is a censoring time} \end{cases} \\ \\ \boldsymbol{X} &= (X_1, \ldots, X_p) \\ &= \text{covariates} \end{aligned}$$

Create survival object for graphs, models, etc.

```
> Lung.km
Call: survfit(formula = Surv(time, status) ~ -1, data = LungData)

      n  events  median 0.95LCL 0.95UCL
    228     165     310     285     363

> summary(Lung.km)
Call: survfit(formula = Surv(time, status) ~ -1, data = LungData)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    5    228       1   0.9956 0.00438       0.9871        1.000
   11    227       3   0.9825 0.00869       0.9656        1.000
   12    224       1   0.9781 0.00970       0.9592        0.997
   13    223       2   0.9693 0.01142       0.9472        0.992
   15    221       1   0.9649 0.01219       0.9413        0.989
   26    220       1   0.9605 0.01290       0.9356        0.986
   30    219       1   0.9561 0.01356       0.9299        0.983
   31    218       1   0.9518 0.01419       0.9243        0.980
   53    217       2   0.9430 0.01536       0.9134        0.974
```

Summary survival curve: Kaplan-Meier Estimator

The Kaplan-Meier (KM) estimator is also known as the 'product limit' estimator

It calculates the survival probability at each observed event time $T_1, T_2, \ldots$

Some notation

$$\begin{aligned} d_j &= \text{number of deaths (events) at time } T_j \\ n_j &= \text{number still at risk at time } T_j \end{aligned}$$

Calculation of Kaplan Meier estimator

$$
\begin{aligned}
\widehat{S}(T_1) &= 1 - d_1/n_1 \\
&= \Pr(\text{survive beyond } T_1)
\end{aligned}
$$

$$
\begin{aligned}
\widehat{S}(T_2) &= (1 - d_1/n_1) \times (1 - d_2/n_2) \\
&= \widehat{S}(T_1) \times (1 - d_2/n_2) \\
&= \Pr(\text{survive beyond } T_1 \text{ and survive beyond } T_2)
\end{aligned}
$$

$$
\begin{aligned}
\widehat{S}(T_3) &= (1 - d_1/n_1) \times (1 - d_2/n_2) \times (1 - d_3/n_3) \\
&= \widehat{S}(T_2) \times (1 - d_3/n_3) \\
&\vdots
\end{aligned}
$$

Kaplan-Meier example

T_j	n_j	d_j	S(T_j)	std.err	lower 95% CI	upper 95% CI
5	228	1	0.9956	0.00438	0.9871	1.000
11	227	3	0.9825	0.00869	0.9656	1.000
12	224	1	0.9781	0.00970	0.9592	0.997
13	223	2	0.9693	0.01142	0.9472	0.992
15	221	1	0.9649	0.01219	0.9413	0.989
.....						
183	156	1	0.7035	0.03041	0.6464	0.766
186	154	1	0.6989	0.03056	0.6416	0.761
189	152	1	0.6943	0.03070	0.6367	0.757
194	149	1	0.6897	0.03085	0.6318	0.753

Kaplan-Meier example

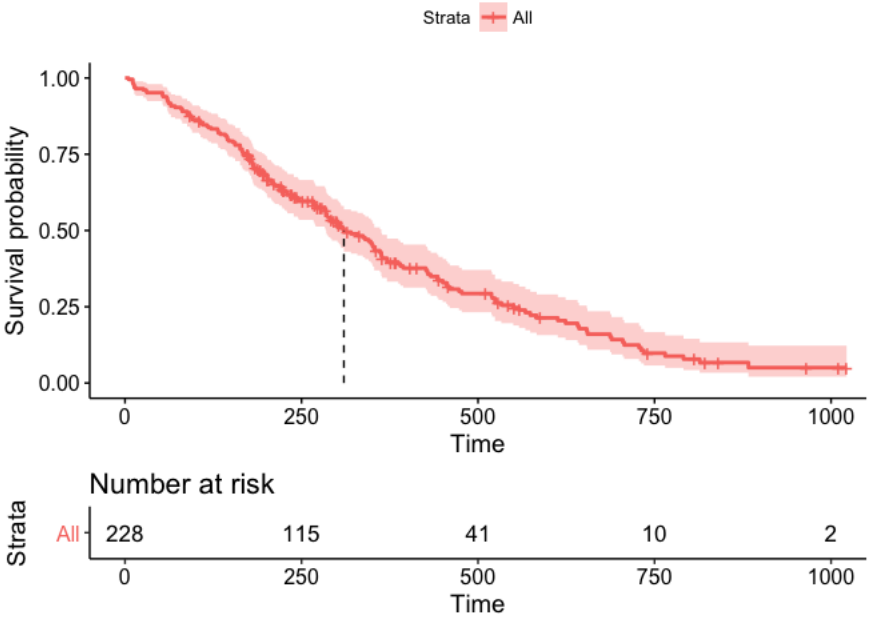
Some options for plotting K-M curve from a survival object

```
plot(Lung.km)

ggsurvplot(Lung.km, risk.table=T)

ggsurvplot(Lung.km, risk.table=T, surv.median.line="v")
```

Kaplan Meier plot (using ggsurvplot)



Calculation of the hazard estimator

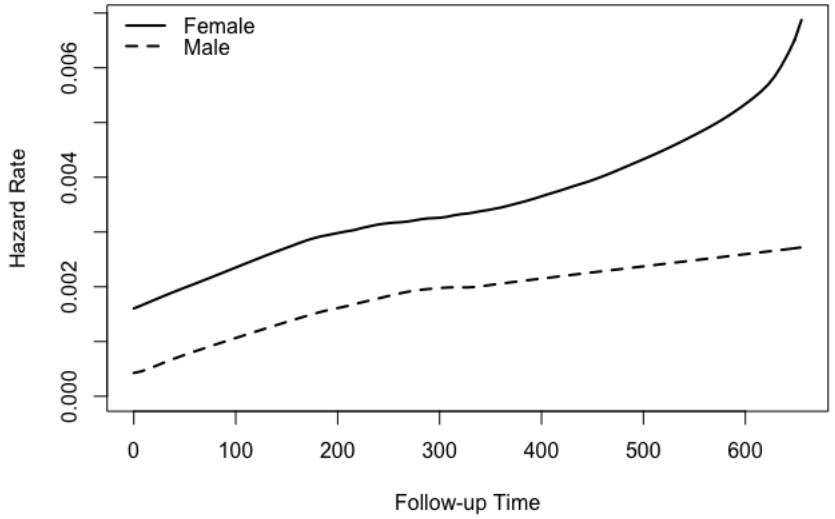
The most commonly-used estimate of the hazard is the *Nelson estimator*

The Nelson estimator is simple: at each observed event time, it calculates the probability of experiencing an event

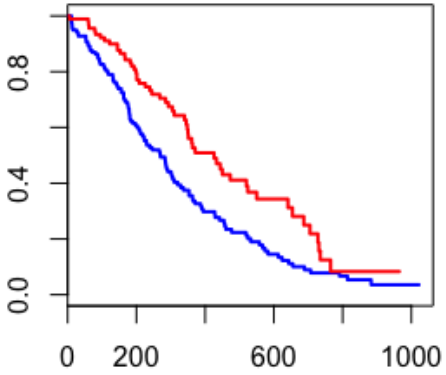
$$\widehat{h}(T_j) \quad = \quad d_j/n_j$$

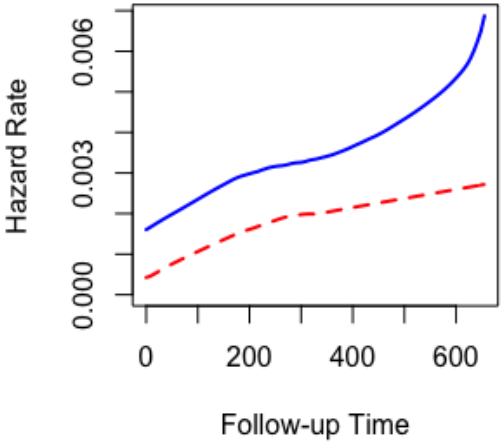
One problem with this estimator is that many time points have only one or two events, so the graph is 'spikey'. For those situations we use a smoothing function.

Hazard plot by gender



Hazard and survival plots side by side





Two sample comparisons

You are probably familiar with the t-test (for means) and the rank sum test (for medians).

Comparison of survival distributions uses a modification of the rank sum test known as the *log rank test*. It addresses the complications due to censoring.

The logrank test compares observed and expected number of events in the two samples. The null hypothesis is that the survival curves are equal, and the alternative hypothesis is that they are not equal.

Logrank test example

```
> survdiff( Surv(time, status) ~ sex, data = LungData )


Call:
survdiff(formula = Surv(time, status) ~ sex, data = LungData)

        N Observed Expected (O-E)^2/E (O-E)^2/V
sex=1 138      112     91.6      4.55      10.3
sex=2  90       53     73.4      5.68      10.3

 Chisq= 10.3  on 1 degrees of freedom, p= 0.00131
```

Regression modeling for survival data

Regression analysis is designed to assess the effect of covariates on an outcome.

Linear regression: Goal is to model the *mean* of an outcome as a function of covariates

- Covariates: $\boldsymbol{X} = X_1, X_2, X_3, \ldots, X_p$
- Outcome: Y
- Mean of Y given $\boldsymbol{X}$ denoted by $\mu(X)$
- Model is

$$\mu(\boldsymbol{X}) \quad = \quad \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

What is β_0 ?

What is β_1 ?

Regression modeling for survival data

In survival analysis, instead of the mean, we typically model the effect of covariates on the *hazard function*.

(Remember that 'hazard' is just another term for 'event rate'.)

It is easier to write the model in terms of the log hazard.

$$\log h(t \,|\, \boldsymbol{X}) \;\; = \;\; \log h_0(t) + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

In this case, the term $h_0(t)$ is called the *baseline hazard* or *reference hazard.* It can be viewed in the same way as an *intercept*, except it's a function (curve).

This model is known as the *proportional hazards model*

Proportional hazards model for survival data

In many cases, the proportional hazards model is written directly in terms of the hazard.

It takes the form

$$\begin{aligned} h(t \mid \boldsymbol{X}) &= h_0(t)\, e^{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p} \\ &= h_0(t)\, \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) \end{aligned}$$

Why is this called the proportional hazards model?

In the next slide, we will consider the case of a single binary covariate, where $X = 1$ or $X = 0$. This would be the setup for a two-sample comparison of the hazard functions.

Proportional hazards model for survival data

The PH model in this case is

$$h(t \mid X) \;=\; h_0(t)\ \exp(\beta_1 X)$$

When $X = 1$, the hazard is

$$h(t \mid X = 1) \;=\; h_0(t)\ \exp(\beta \times 1) \;=\; h_0(t) \exp(\beta)$$

When $X = 0$, the hazard is

$$h(t \mid X = 0) \;=\; h_0(t)\ \exp(\beta \times 0) \;=\; h_0(t)$$

The *hazard ratio* is

$$\frac{h(t \mid X = 1)}{h(t \mid X = 0)} = \frac{h_0(t) \exp(\beta)}{h_0(t)} \;=\; \exp(\beta)$$

so that the two hazard rates are *proportional* for every point in time.

Two-group comparison using proportional hazards model

In this example we examine the effect of gender. The model is

$$h(t \mid X_1) \;=\; h_0(t)\exp(X_1\beta_1)$$

```
> Model.1 = coxph( Surv(time, status) ~ sex, data = LungData  )
> summary(Model.1)

  n= 228, number of events= 165

       coef exp(coef) se(coef)      z Pr(>|z|)
sex -0.5310    0.5880   0.1672 -3.176  0.00149 **


    exp(coef) exp(-coef) lower .95 upper .95
sex     0.588      1.701    0.4237     0.816
```

Multiple regression using proportional hazards model

Add the effect of ECOG score, $X_2 = 0, 1, 2, 3$.

$$h(t \mid X_1, X_2) \;=\; h_0(t) \exp(X_1\beta_1 + X_2\beta_2)$$

```
> Model.2 = coxph( Surv(time, status) ~ ecog + sex, data = LungData, subset = (ecog<3
> summary(Model.2)

  n= 226, number of events= 163
   (1 observation deleted due to missingness)

        coef exp(coef) se(coef)      z Pr(>|z|)
ecog  0.4759    1.6094   0.1137  4.186 2.84e-05 ***
sex  -0.5484    0.5779   0.1678 -3.268  0.00108 **

     exp(coef) exp(-coef) lower .95 upper .95
ecog    1.6094     0.6213    1.2879    2.0111
sex     0.5779     1.7305    0.4159    0.8029
```

Adding additional covariates: ECOG score

In the next models, we will add patient-reported functioning score to see if it is associated with survival time, even after conditioning on the ECOG score.

The ECOG score is a widely-used staging variable that is ascertained by the treating physician. The patient-reported functioning score is the Karnofsky score, scaled from 0 to 100.

We will add this variables:

$$X_3 \quad = \quad \text{patient-reported Karnofsky score}$$

The model is

$$h(t \mid X_1, X_2, X_3) \quad = \quad h_0(t) \exp(X_1\beta_1 + X_2\beta_2 + X_3\beta_3)$$

Adding patient-reported score

```
  n= 223, number of events= 160
    (4 observations deleted due to missingness)

               coef exp(coef)  se(coef)      z Pr(>|z|)
sex       -0.533948  0.586285  0.169584 -3.149  0.00164 **
ecog       0.371346  1.449684  0.139006  2.671  0.00755 **
karno_pat -0.009201  0.990841  0.006877 -1.338  0.18093


          exp(coef) exp(-coef) lower .95 upper .95
sex          0.5863     1.7057    0.4205    0.8174
ecog         1.4497     0.6898    1.1040    1.9037
karno_pat    0.9908     1.0092    0.9776    1.0043
```

Questions to answer

1. What happens when you fit the model that has Karnofsky score but not ECOG score; that is, the model with X_1 and X_3 only?

2. Why is the effect of Karnofsky score different in the model without X_2? To shed some light on this, you can try looking at the boxplot of Karnofsky score, stratified by ECOG score. What does this tell you?

   ```
   boxplot(karno_pat ~ ecog, data=LungData)
   ```

3. You can check the proportional hazards assumption with the command `cox.zph`. This runs a hypothesis test to check the PH assumption for each covariate, and for the model as a whole. The null hypothesis is that the hazards are proportional across levels of the covariate. Low p-values indicate violations of the PH assumption. Higher p-values indicate that there is no evidence for violation of the assumption. But this does not prove the PH assumption is true.

   Type `help(cox.zph)` to learn how to test the PH assumption