Introduction to Survival Data/Analysis

LAs BeST 2023

Eric S. Kawaguchi

Division of Biostatistics and Epidemiology
Department of Population and Public Health Sciences
University of Southern California

July 10, 2023

Keck School of
Medicine of USC

General outline

- What is time-to-event (survival) data?
- Common quantities in survival analysis
- Basic inference (based on these common quantities)
- Regression modeling via Cox model

What is time-to-event (survival) data?

Working example: Primary Biliary Cirrhosis.

- Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 - 1984;
- Randomized control trial of the drug D-penicillamine;
- Recorded various demographic and clinical covariates;
- Outcome of interest: Death (Time-to-death)

What characterizes time-to-event data

Survival data are data where the *outcome of interest* is quantified as a "time-to-event".

NOTE: In what follows, we will focus on the *continuous* time setting.

Survival data arise in a number of applied fields:

Keck School of Medicine of USC

What characterizes time-to-event data

Survival data are data where the *outcome of interest* is quantified as a "time-to-event".

NOTE: In what follows, we will focus on the *continuous* time setting.

Survival data arise in a number of applied fields:

- Biomedical

Keck School of Medicine of USC

What characterizes time-to-event data

Survival data are data where the *outcome of interest* is quantified as a "time-to-event".

NOTE: In what follows, we will focus on the *continuous* time setting.

Survival data arise in a number of applied fields:

- Biomedical
- Engineering

Keck School of Medicine of USC

What characterizes time-to-event data

Survival data are data where the *outcome of interest* is quantified as a “time-to-event”.

NOTE: In what follows, we will focus on the *continuous* time setting.

Survival data arise in a number of applied fields:

- Biomedical
- Engineering
- Business/commerce

Keck School of Medicine of USC

What characterizes time-to-event data

Survival data are data where the *outcome of interest* is quantified as a "time-to-event".

NOTE: In what follows, we will focus on the *continuous* time setting.

Survival data arise in a number of applied fields:

- Biomedical
- Engineering
- Business/commerce
- Sociology

What characterizes time-to-event data

Survival data are data where the *outcome of interest* is quantified as a "time-to-event".

NOTE: In what follows, we will focus on the *continuous* time setting.

Survival data arise in a number of applied fields:

- Biomedical
- Engineering
- Business/commerce
- Sociology

In public health/preventive medicine we often refer to time-to-event data as survival data (e.g. time-to-death).

Treating the event as binary

Question: Can we treat the event of interest (eg. dead/alive) as a binary outcome?

- Yes, there is nothing wrong with treating the endpoint as a binary outcome.
- Analyses can be performed using χ^2 tests, logistic regression, etc.
- However, modeling the endpoint as a time-to-event outcome over a binary outcome can increase power.
  - Ref: van der Net et al. (2008)
  - Ref: Hughey et al. (2019)
- Key: More information and less assumptions when modeling the endpoint as a time-to-event outcome.

Treating the time-to-event as continuous

Question: Can we treat the time-to-event as (non-negative) continuous data?

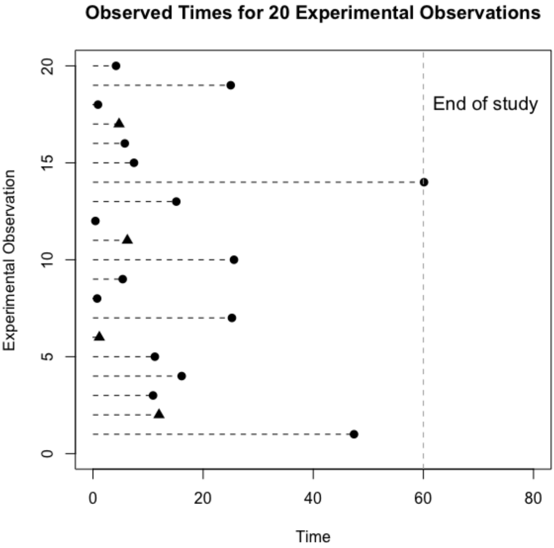
- Analyses can be performed using t-tests, ANOVA, linear regression etc.
- However, we may only know that events have occurred only within certain intervals.
  - Event may have occurred prior to the start of the study.
  - Event may have not yet occurred by the end of the study.
  - Event may have occurred but we do not know exactly when it occurred.
- These are all examples of *censoring*.
  - Not taking censoring into account (appropriately) will lead to biased inference.

Censoring

Survival data present a challenge not seen in typical data.

- **Censoring:** When event times of a subject are not *fully* known.
    - Right censoring
    - Left censoring
    - Interval censoring
- Censoring must be adequately accounted for when analyzing survival data.
- We will focus on right censoring, which is the most common censoring in biomedical applications.

Toy example



Keck School of Medicine of USC

Censoring

Survival data present a challenge not seen in typical data.

- **Censoring:** When event times of a subject are not *fully* known.
  - Right censoring
  - Left censoring
  - Interval censoring
- Censoring must be adequately accounted for when analyzing survival data.
- We will focus on right censoring, which is the most common censoring in biomedical applications.
- Censoring can be viewed as “partially” observed data

Censoring

Survival data present a challenge not seen in typical data.

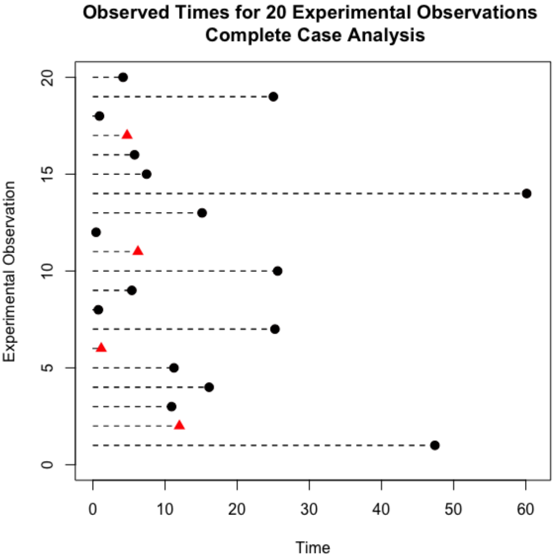
- **Censoring:** When event times of a subject are not *fully* known.
    - Right censoring
    - Left censoring
    - Interval censoring
- Censoring must be adequately accounted for when analyzing survival data.
- We will focus on right censoring, which is the most common censoring in biomedical applications.
- Censoring can be viewed as "partially" observed data

Question: So how do we deal with right censoring?

Three “out-of-the-box” solutions

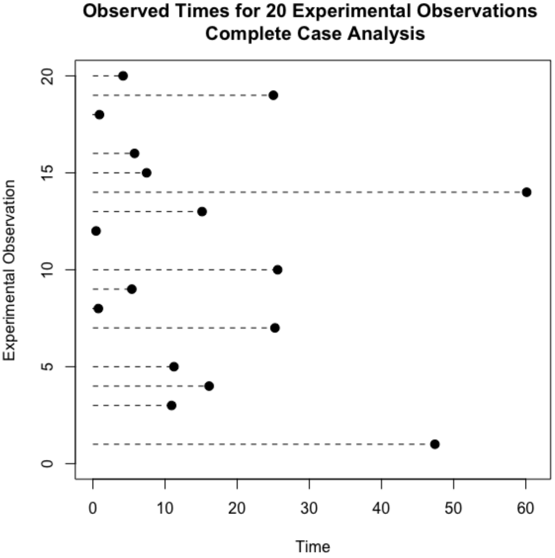
1. Complete case analysis: Remove subjects who are censored
2. Last observation as event: Assume that the observed event time is the true event time
    - Example: If the last available follow up for an individual was 1 year, we assume that the individual died at 1 year.
3. Last observation carried forward: Assume that the individual survived until the end of study.

Toy example



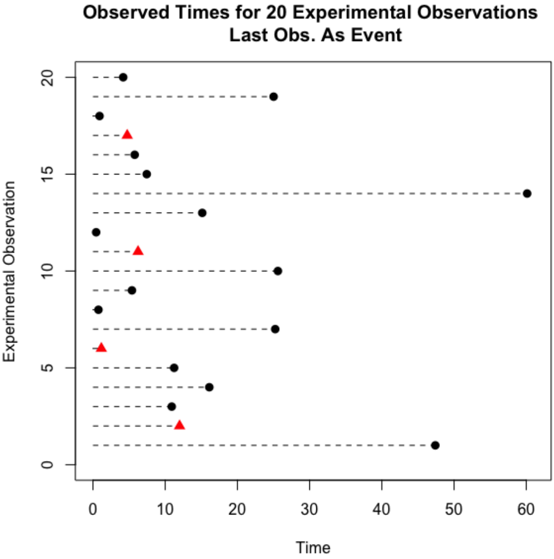
Keck School of Medicine of USC

Toy example



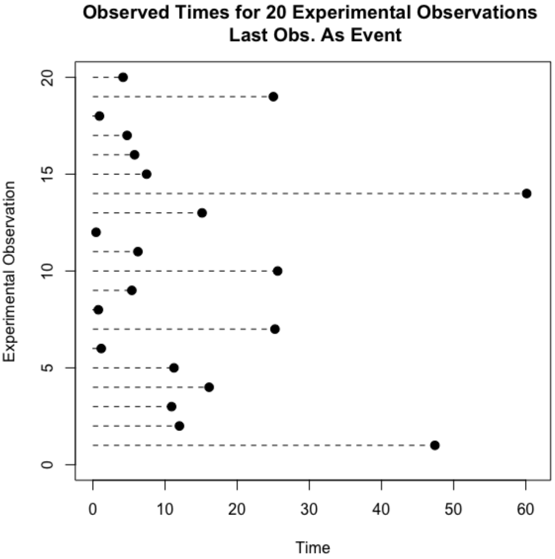
Keck School of Medicine of USC

Toy example

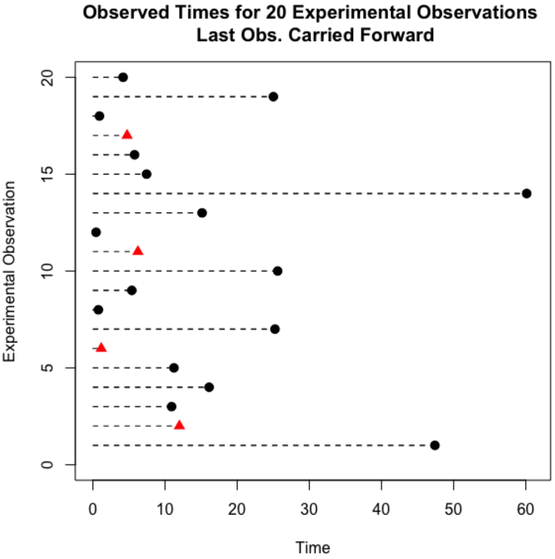


Keck School of Medicine of USC

Toy example

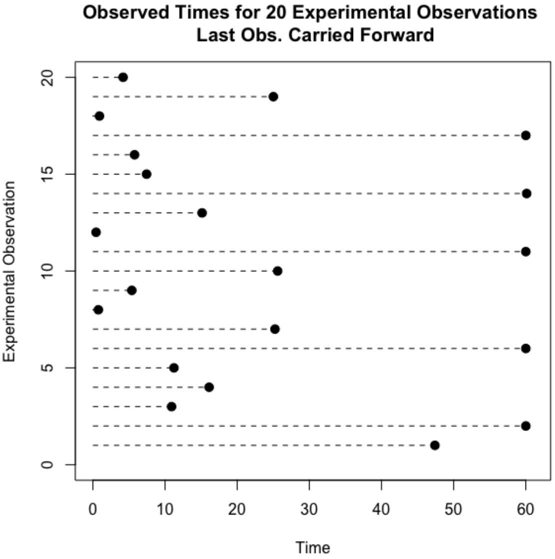

Observed Times for 20 Experimental Observations
Last Obs. As Event
Experimental Observation
0
5
10
15
20
Time
0
10
20
30
40
50
60

Toy example



Keck School of Medicine of USC

Toy example


Observed Times for 20 Experimental Observations
Last Obs. Carried Forward
Experimental Observation
0
5
10
15
20
Time
0
10
20
30
40
50
60

Three "out-of-the-box" solutions

1. Complete case analysis: Remove subjects who are censored
2. Last observation as event: Assume that the observed event time is the true event time
    - Example: If the last available follow up for an individual was 1 year, we assume that the individual died at 1 year.
3. Last observation carried forward: Assume that the individual survived until the end of study.

Three “out-of-the-box” solutions

1. Complete case analysis: Remove subjects who are censored
2. Last observation as event: Assume that the observed event time is the true event time
    - Example: If the last available follow up for an individual was 1 year, we assume that the individual died at 1 year.
3. Last observation carried forward: Assume that the individual survived until the end of study.

All of these approaches will lead to biased estimates.

Three “out-of-the-box” solutions

1. Complete case analysis: Remove subjects who are censored
2. Last observation as event: Assume that the observed event time is the true event time
    - Example: If the last available follow up for an individual was 1 year, we assume that the individual died at 1 year.
3. Last observation carried forward: Assume that the individual survived until the end of study.

All of these approaches will lead to biased estimates.

NOTE: Censoring must be appropriately handled to ensure valid statistical inference

Keck School of
Medicine of USC

Basic quantities in survival analysis

Basic quantities in survival analysis

List of some common quantities in survival analysis

- The Survival Function: $S(t)$
- The Hazard Function: $h(t)$

The Survival Function

- For a nonnegative random variable T, the survival function is defined as
$$S(t) = \Pr(T > t).$$
i.e. "The probability of an individual experience the event of interest after time t".
- For right censored data, $CIF(t) = \Pr(T \leq t) = 1 - \Pr(T > t) = 1 - S(t)$ where $CIF(t)$ is the cumulative incidence function at time t (cumulative probability of experiencing the event of interest by time t).

The Hazard Function

- Also known as the “intensity function” in stochastic processes or the “age-specific failure rate” in epidemiology.
- The hazard function (rate) is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}$$

- If T is continuous,
  - $h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log[S(t)]$
  - Cumulative hazard: $H(t) = \int_0^t h(s) ds$

The Hazard Function: Some notes

- By construction, $h(t) \geq 0$.
- $h(t)$ is NOT a probability.
- However, $h(t)\Delta t$ can be viewed as the "approximate" probability of an individual of age t experiencing the event in the next instant.

The Kaplan-Meier Estimator

- The objective of the Kaplan-Meier (KM) estimator is to estimate the population survival curve from a sample.
- The KM estimator is also often referred to as the "product-limit" estimator and is defined as

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \le t}[1 - d_i / Y_i] & \text{if } t_1 \le t \end{cases}$$

where

  - $t_1 < t_2 < \ldots < t_D$ are the D distinct event times;
  - d_i is the number of events at time t_i;
  - Y_i is a count of the number of individuals with a study time $\geq t_i$ (generally referred to as the "risk set").

Keck School of Medicine of USC

Some notes on the KM estimator

- The KM estimator is a non-parametric estimator of the survivor function.
- The KM estimator is a step function with jumps at the D observed event times.
- The size of the jumps depends on both the number of observed events at time t_i and the pattern of the censored observations prior to t_i.
- For $t > t_{max}$, the largest observation time, the KM estimator is not well defined.
- Side note:
  - The KM estimator can also be used to estimate the cumulative hazard $H(t)$ since $H(t) = -\log S(t)$;
  - An alternative, with better finite sample performance, is the Nelson-Aalen estimator (not covered here).

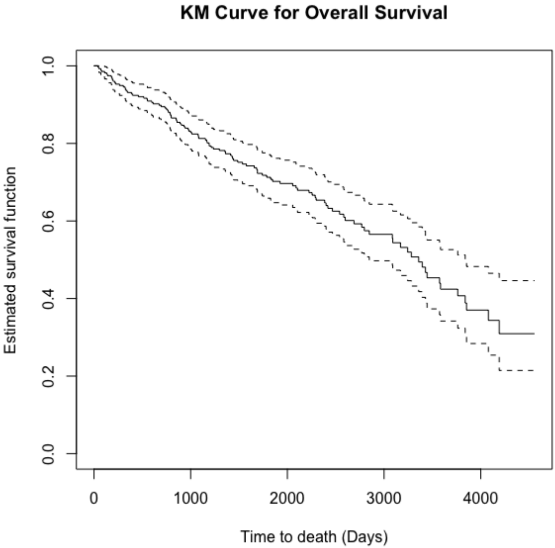
PBC Data Revisited

- Outcome of interest: Death
- Subjects who receive a liver transplant no longer participate in the study.
- Censoring: Alive at study time or received a transplant.

PBC Data Revisited

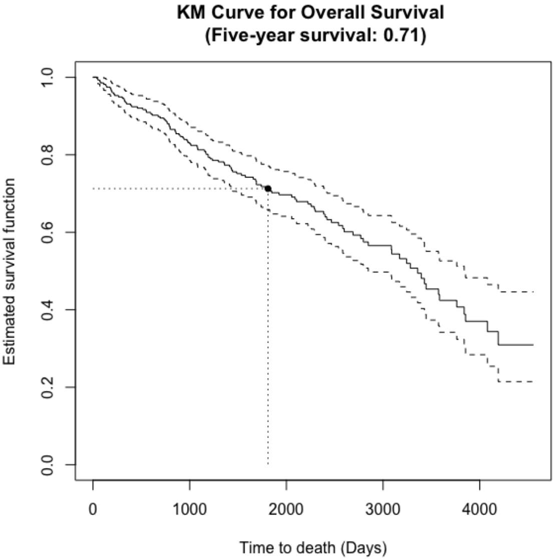
- Outcome of interest: Death
- Subjects who receive a liver transplant no longer participate in the study.
- Censoring: Alive at study time or received a transplant.
- Some questions of interest:
    - What is the estimated probability of death?
    - How variable are these estimates?
    - What is the five-year survival probability?
    - What is the median survival time?
    - How does survival differ by gender or by cancer stage?

PBC Data: KM Curve



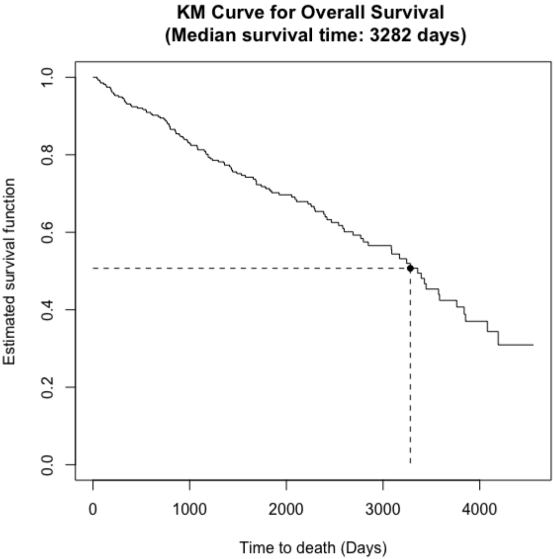
Keck School of Medicine of USC

PBC Data: KM Curve



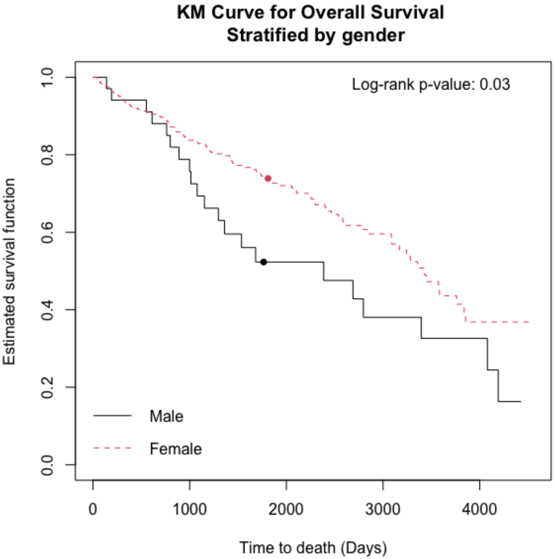
Keck School of Medicine of USC

PBC Data: KM Curve

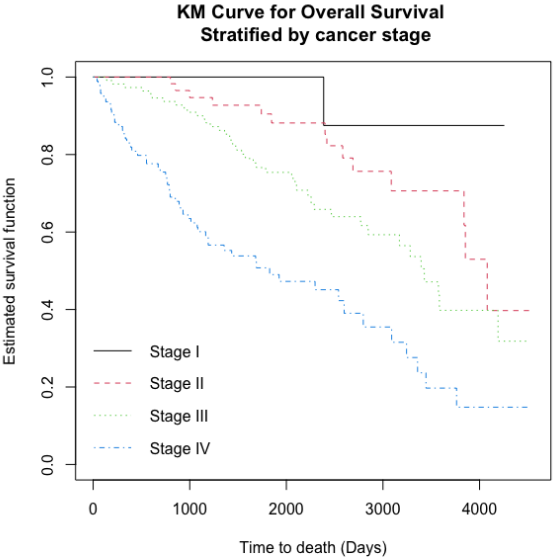


Keck School of
Medicine of USC

PBC Data: KM Curve



PBC Data: KM Curve



Keck School of Medicine of USC

Formal statistical inference

Statistical inference can be used to:

- Compare times and curves against some *a priori* values/distributions;
- Compare survival times (point wise) between two different groups;
- Compare if the survival curves between two (or more) groups are different;
- Compare survivor quantiles (not covered)
- Model covariate effects on survival.

	Continuous	Binary	Survival
Display	Histogram/Box plots	$R \times C$ table	KM Curve
K-sample test	t-test/ANOVA	Fisher's exact test/χ^2	Log-rank test
Regression	Linear	Logistic	Cox PH

Modeling time-to-event data via the Cox proportional hazards model

Regression modeling

- Oftentimes we are interested in quantifying the relationship between the time to event and a set of explanatory variables.
- One of the most widely used regression models for right-censored data is due to Cox (1972):

  $$h(t|Z) = h_0(t)\exp(Z\beta),$$

  where $h_0(t)$ is an unspecified baseline hazard, Z is an $n \times p$ design matrix, and $\beta = (\beta_1, \ldots, \beta_p)$ is a p-dimensional parameter vector.
- This model is often referred to as: The Cox proportional hazards (PH) model.

Keck School of Medicine of USC

Interpreting β

Assume that $z \in \{0, 1\}$. The Cox model assumes,

$$h(t|z) = h_0(t)\exp(z\beta)$$

- Therefore

$$\frac{h(t|z=1)}{h(t|z=0)} = \frac{h_0(t)\exp(\beta)}{h_0(t)\exp(0)} = \frac{\exp(\beta)}{1} = \exp(\beta).$$

- Individuals with $z = 1$ have a hazard that is $\exp(\beta)$ times the hazard for individuals with $z = 0$.
- β is often referred to as the "log hazard ratio".
- Multivariate setting: hazard ratio is conditional on the values of the other covariates in the model.

Keck School of Medicine of USC

Proportional hazards

IMPORTANT NOTE: The hazard ratio is constant!

$$\frac{h(t|z=1)}{h(t|z=0)} = \frac{h_0(t)\exp(\beta)}{h_0(t)\exp(0)} = \frac{\exp(\beta)}{1} = \exp(\beta).$$

For any $t > 0$, we assume that the hazard ratio is $\exp(\beta)$.
We assume that the effect of z is *proportional* across t.

Keck School of
Medicine of USC

Relation to survival

- Mathematically, $S(t) = \exp\{-H(t)\}$, where $H(t) = \int_0^t h(s)ds$.
- Under the Cox model, $h(t|z) = h_0(t)\exp(z^T\beta)$.
- Therefore, $\hat{S}(t|z) = \hat{S}_0(t)^{\exp(z^T\beta)}$, where $S_0(t) = S(t|z=0)$.
- Note that if $z \in \{0,1\}$ then $\hat{S}(t|z=0) = \hat{S}_0(t)$ and $\hat{S}(t|z=1) = \hat{S}_0(t)^{\exp(\beta)}$.
- If $\beta > 0$, then $\hat{S}_0(t) > \hat{S}_0(t)^{\exp(\beta)}$ (Lower survival)
- If $\beta < 0$, then $\hat{S}_0(t) < \hat{S}_0(t)^{\exp(\beta)}$ (Higher survival)

Working example: Primary Biliary Cirrhosis.

Covariates of interest: Treatment, age at study entry, sex, cancer stage, serum bilirubin (bili), serum albumin (albumin), serum cholesterol (chol), platelet count, triglycerides (trig)

Working example: Primary Biliary Cirrhosis.

Covariates of interest: Treatment, age at study entry, sex, cancer stage, serum bilirubin (bili), serum albumin (albumin), serum cholesterol (chol), platelet count, triglycerides (trig)

```
                      coef  exp(coef)   se(coef)      z  Pr(>|z|)
trt             -0.1290490  0.8789309  0.2076922 -0.621   0.5344
age              0.0228666  1.0231300  0.0111791  2.045   0.0408 *
sex             -0.5715253  0.5646635  0.2809864 -2.034   0.0420 *
factor(stage)2   1.0786363  2.9406667  1.0417370  1.035   0.3005
factor(stage)3   1.5921009  4.9140621  1.0211162  1.559   0.1190
factor(stage)4   2.0841263  8.0375659  1.0266754  2.030   0.0424 *
bili             0.1404800  1.1508260  0.0190757  7.364 1.78e-13 ***
albumin         -1.1185769  0.3267445  0.2704679 -4.136 3.54e-05 ***
chol             0.0002802  1.0002802  0.0004151  0.675   0.4996
platelet        -0.0006750  0.9993252  0.0010939 -0.617   0.5372
trig            -0.0011835  0.9988172  0.0012659 -0.935   0.3498
```

Conclusion

What we went over:

- Why is survival analysis a necessary sub field of statistics (esp. in biomedical settings)
- Basic quantities in survival analysis
- The Cox proportional hazards model

Thank You!

ekawaguc [at] usc.edu

Keck School of
Medicine of USC

Appendix: Intuition behind KM estimator

Recall: $\Pr(A \cap B) = \Pr(A) \times \Pr(B|A)$

- For any time $t \in [t_1, t_2)$,

$$S(t) = \Pr(T > t) = \Pr(\text{survive in } [0, t_1)) \times \Pr(\text{survive in } [t_1, t)|\text{survive in } [0, t_1))$$

$$\hat{S}(t) = 1 \times \frac{Y_1 - d_1}{Y_1} = 1 - \frac{d_1}{Y_1}$$

- For any time $t \in [t_2, t_3)$,

$$S(t) = \Pr(T > t) = \Pr(\text{survive in } [t_1, t_2)) \times \Pr(\text{survive in } [t_2, t)|\text{survive in } [t_1, t_2))$$

$$\hat{S}(t) = \left(1 - \frac{d_1}{Y_1}\right) \times \left(1 - \frac{d_2}{Y_2}\right)$$

$$\vdots$$

Keck School of Medicine of USC