# BIS630 HW1

*Joanna Chen*

*1/28/2020*

**Question 1**

Below are data from a study conducted on patients awaiting a heart transplant. The goal of the study was to describe the **survival of patients during their time on the heart transplant waiting list**, which we'll call "transplant-free survival." A patient remains on the heart transplant waiting list until they receive a heart transplant and is no longer on the transplant waiting list after they receive the transplant. The event of interest is **death before transplant**.

**Data description:**

- birth.dt: birth date
- wtlist.dt: date patient is put on the heart transplant waiting list
- tx.date: transplant date
- fu.date: end of follow-up (Note: Equals date of death for those who died. For some patients, follow-up continued past date of transplant but we are only interested in their survival experience before transplant.)
- transplant: transplant indicator (1=yes)
- death: death indicator (1=yes)

**Create three new variables in this data set for these 20 patients:**

(1). Age at which the patient is put on the transplant waiting list (in years)

(2). Appropriate survival time variable for the analysis of transplant-free survival (in days)

Since we are interested in the survival of patients during their time on the heart transplant waiting list, the time origin is *wtlist.dt*, while the ending event is patients is no longer on the waitlist and that is the time they receive a transplant *tx.date*. For patients who do not receive a transplant, the end date is *fu.date*.

(3). Event indicator for the analysis of transplant-free survival During patients' time on the heart transplant waiting list, if the event of interest is **death before transplant** happens, then event indicator equals 1. Otherwise, 0.

Append the table above with the three new variables that you created and submit the table as your answer to this question.

| ID | birth.dt | wtlist.dt | tx.date | fu.date | transplant | death | age_wtlist | survival_days | tf_survival |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1937-01-10 | 1967-11-15 | NA | 1968-01-03 | 0 | 1 | 30.8 | 49 | 1 |
| 2 | 1916-03-02 | 1968-01-02 | NA | 1968-01-07 | 0 | 0 | 51.8 | 5 | 0 |
| 3 | 1913-09-19 | 1968-01-06 | 1968-01-06 | 1968-01-21 | 1 | 0 | 54.3 | 0 | 0 |
| 4 | 1927-12-23 | 1968-03-28 | 1968-05-02 | 1968-05-05 | 1 | 0 | 40.3 | 35 | 0 |
| 5 | 1947-07-28 | 1968-05-10 | NA | 1968-05-27 | 0 | 0 | 20.8 | 17 | 0 |
| 6 | 1913-11-08 | 1968-06-13 | NA | 1968-06-15 | 0 | 1 | 54.6 | 2 | 1 |
| 7 | 1917-08-29 | 1968-07-12 | 1968-08-31 | 1970-05-17 | 1 | 1 | 50.9 | 50 | 0 |

| ID | birth.dt | wtlist.dt | tx.date | fu.date | transplant | death | age_wtlist | survival_days | tf_survival |
|---|---|---|---|---|---|---|---|---|---|
| 8 | 1923-03-27 | 1968-08-01 | NA | 1968-09-09 | 0 | 1 | 45.3 | 39 | 1 |
| 9 | 1921-06-11 | 1968-08-09 | NA | 1968-11-01 | 0 | 0 | 47.2 | 84 | 0 |
| 10 | 1926-02-09 | 1968-08-11 | 1968-08-22 | 1968-10-07 | 1 | 1 | 42.5 | 11 | 0 |
| 11 | 1920-08-22 | 1968-08-15 | 1968-09-09 | 1969-01-14 | 1 | 0 | 48.0 | 25 | 0 |
| 12 | 1915-07-09 | 1968-09-17 | NA | 1968-09-24 | 0 | 0 | 53.2 | 7 | 0 |
| 13 | 1914-02-22 | 1968-09-19 | 1968-10-05 | 1968-12-08 | 1 | 0 | 54.6 | 16 | 0 |
| 14 | 1914-09-16 | 1968-09-20 | 1968-10-26 | 1972-07-07 | 1 | 1 | 54.0 | 36 | 0 |
| 15 | 1914-12-04 | 1968-09-27 | NA | 1968-09-27 | 0 | 0 | 53.8 | 0 | 0 |
| 16 | 1919-05-16 | 1968-10-26 | 1968-11-22 | 1969-08-29 | 1 | 1 | 49.4 | 27 | 0 |
| 17 | 1948-06-29 | 1968-10-28 | NA | 1968-12-02 | 0 | 1 | 20.3 | 35 | 1 |
| 18 | 1911-12-27 | 1968-11-01 | 1968-11-20 | 1968-12-13 | 1 | 1 | 56.8 | 19 | 0 |
| 19 | 1909-10-04 | 1968-11-18 | NA | 1968-12-24 | 0 | 1 | 59.1 | 36 | 1 |
| 20 | 1913-10-19 | 1969-01-29 | 1969-02-15 | 1969-02-25 | 1 | 0 | 55.3 | 17 | 0 |

**Question 2**

A large number of individuals were enrolled in a study and were followed for 30 years after enrollment to assess age at which the disease symptom first appeared. The symptom could only be identified by a clinical exam and clinical testing and was assessed at yearly at scheduled follow-up visits. For ten selected individuals described below,

- determine if the individual is censored or not.
- If censored, what type of censoring is it (left, right, interval)?
- Report the value of the "time" variable for each participant (i.e., the censoring time or the event time, as appropriate). Note that time is the age of the individual. For interval censored data, you will have two values of time to define the interval. Assume the patients were followed for the full 30 years unless you are told that they died or moved.

*a. The first individual, enrolled in the study at age 45, entered the study with the symptom already present.*
The individual is censored. It's left censored because the event happened before the observation. The censoring time is 45.

*b. The next two healthy (without symptom present) individuals enrolled in the study at ages 30 and 42 and never showed the symptom.*
Censored and right censored because the event happened to the right of the end of the study are not known precisely. The first individual's censoring time is 30+30 = 60 and the second is 42+30 = 72.

*c. The next healthy individual, enrolled in the study at age 35, exhibited the symptom at the second exam after enrollment (2 years after enrollment). The symptom may have appeared between exams.*
Censored and it's interval censored because the event happened between two observations. (35+1, 35+2] = (36,37].

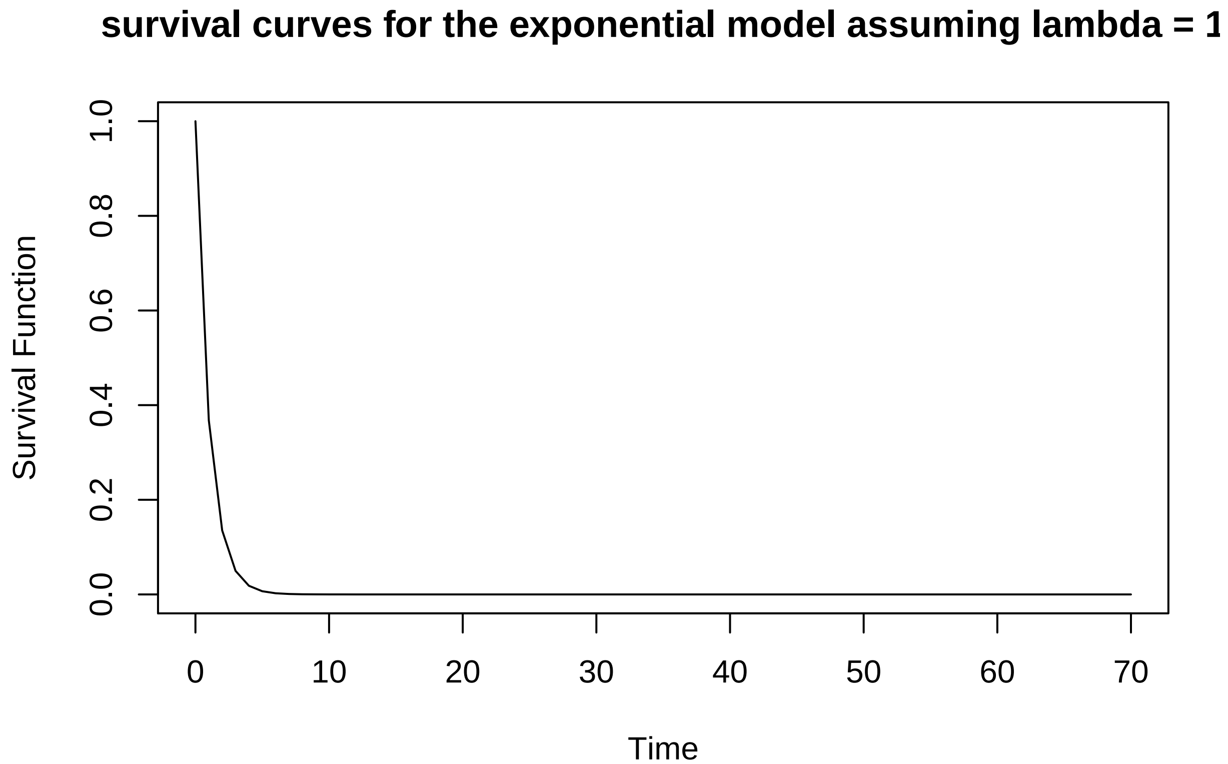
*d. The next healthy individual, enrolled in the study at age 40, exhibited the symptom at the fifth exam after enrollment (5 years after enrollment). The symptom may have appeared between exams.*
Censored and it's interval censored. The reason is the same as c. (40+4, 40+5) = (44, 45).

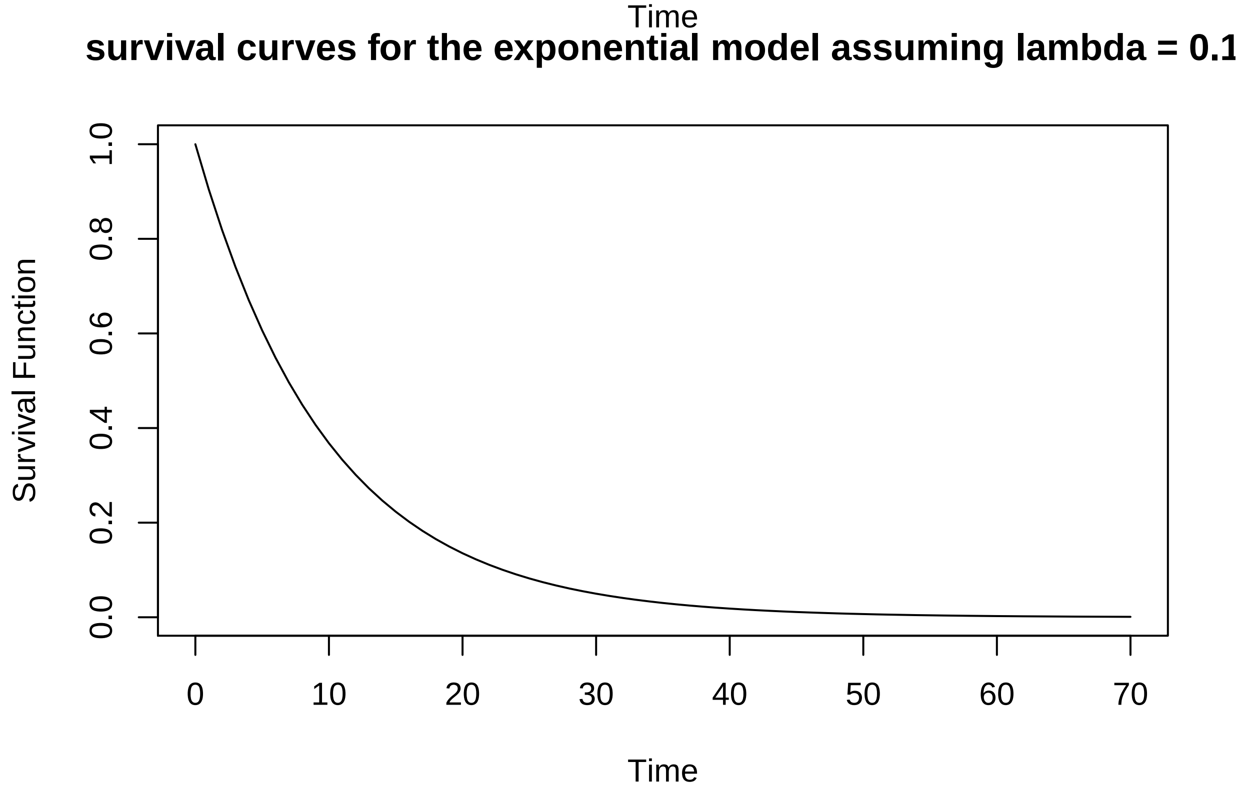
*e. The next two healthy individuals, enrolled in the study at ages 50 and 47, died from causes unrelated to the disease (the symptom was absent) at ages 61 and 65, respectively.*
Right censored. This is because the situation doesn't fit the other censoring definition. The censoring time is 61 and 65.

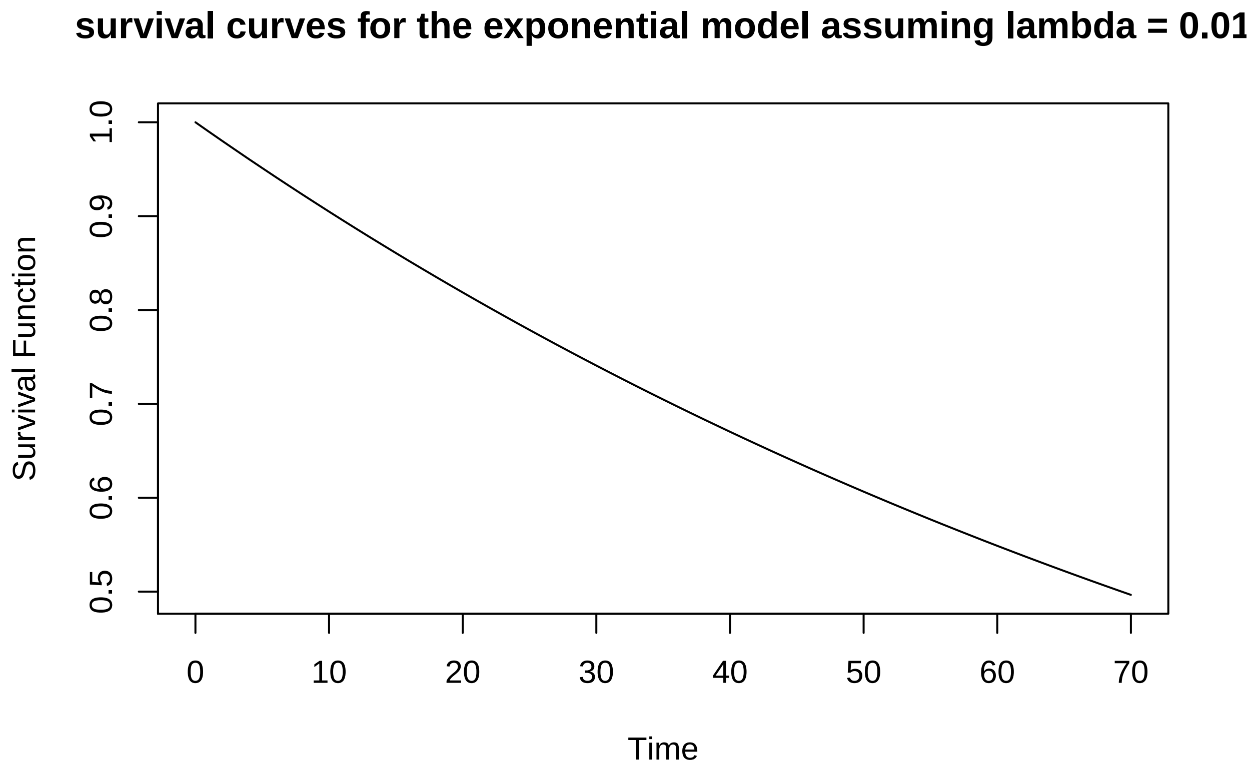
*f. The last three individuals, enrolled in the study at ages 36, 42, and 50, moved away from the area at ages 40, 55, and 60, respectively, never having showed the symptom.*
Right censored. Recall the similar example given in class. The censoring times are 40, 55, 60.

**Question 5**

Plot the survival curves for the exponential model assuming $\lambda = 1$, 0.1, 0.01 events/day over a range of time from 0 to 70 days.
The survival function under the exponential model is $S(t) = exp(-\lambda t)$. We will plot 3 curves. They are $S(t) = exp(-t)$, $S(t) = exp(-0.1t)$ and $S(t) = exp(-0.01t)$.






*Does the relative positioning of the survival curves make sense given the hazard rates? Explain.*
Yes. We know that the survival function focuses on not failing, while the hazard rate focuses on failing. So the larger the hazard rate (the instantaneous rate of failure) is, the quicker the survival curve decaying. The smaller the hazard rate is, the slower the survival cruve decaying (less failing and survive longer).

*Assuming the event of interest is death, which hazard corresponds to the better survival?*
The one with lambda 0.01 one is the better survival. It shows a steady decrease in survival probabilities throughout the time range.

*Report the median survival time for each model.*
for the model with lambda = 1, $exp(-t) = 0.5$ gives $t = -ln(0.5)$.
for the model with lambda = 0.1, $exp(-0.1t) = 0.5$ gives $t = \frac{-ln(0.5)}{0.1}$
for the model with lambda = 0.01, $exp(-0.01t) = 0.5$ gives $t = \frac{-ln(0.5)}{0.01}$

**Question 6**

Perform a brief literature review and find at least two studies in your research area of interest that use survival analysis. Describe the primary question the researchers are trying to address, the time origin, ending event of interest, and measurement scale for the passage of time. How is censoring defined? What are the key independent variables? Attach a print out of the first page of each article with your homework submission.

I am interested in survival analysis pediatric application, particularly, on the Pediatric intensive care unit (PICU) admission and survival rates. I found two studies in the research area. They are:

1. Straney et al. 2018. Trends in PICU dmission and Survival Rates in Children in Australia and New Zealand Following Cardiac Arrest, PLoS One

2. Hakan et al. 2015. Survival after PICU admission: The impact of multiple admissions and complex chronic conditions. Pediatric Critical Care Medicine

Primary question the researchers are trying to address:
For the first article, researchers are trying to address if multiple admissions compared to single PICU

admissions were associated with poor survival over time after the PICU admission.
For the second article, researchers are aiming to describe the trends in PICU admission rates and ICU survival for admitted children in PICU due to in-hospital and out-of-hospital cardiac arrest.

*The time origin:*
In the first article: the time that children admitted into PICU
In the second article: the time that children admitted into PICU

*Ending event of interest:*
In the first article: Death
In the first article: Death

*Measurement scale for the passage of time:*
In the first article: Years
In the first article: Hours

*How is censoring defined?*
Due to the fact that only some of children have experienced the event or were lost to follow-up, survival time was incomplete for a subset of the study group and therefore it's censored. If the event is beyond the end of the follow up period, then it's right censoring. Although most survival data include right censored observation, it wasn't the case in the first paper. In the first paper, individuals are known to experience death within the interval of i_th time admitted to the PICU and i+1_th time admitted to PICU, therefore it's interval censored. In the second paper, right censored is performed. This is because the event, death, will occur eventually and can be beyond the end of the PICU discharge hours (censoring time).

*What are the key independent variables?*
In the first paper, patient demographics, age groups are the key independent variables. In the second paper, age, gender are the key independent variables.

**Apendix**

```
library(readr)
library(data.table)
data <- read_csv("~/Desktop/data.csv")
setDT(data)

data$birth.dt = as.Date(data$birth.dt, format='%d/%m/%Y')
data$wtlist.dt = as.Date(data$wtlist.dt, format='%d/%m/%Y')
data$tx.date = as.Date(data$tx.date, format='%d/%m/%Y')
data$fu.date = as.Date(data$fu.date, format='%d/%m/%Y')

#(1)
data[,age_wtlist := as.numeric(round((wtlist.dt - birth.dt)/365.25,1))]
#(2)
data[,survival_days := ifelse(!is.na(tx.date),tx.date-wtlist.dt,fu.date-wtlist.dt)]
#(3)
data$tf_survival = c(1,0,0,0,0,1,0,1,0,0,0,0,0,0,0,0,1,0,1,0)
knitr::kable(data)

#Q5
t = 0:70
s1 = exp(-t)
s2 = exp(-0.1*t)
s3 = exp(-0.01*t)
plot(t, s1, type='l',main="survival curves for the exponential model assuming lambda = 1", xlab="Time",
plot(t, s2, type='l',main="survival curves for the exponential model assuming lambda = 0.1", xlab="Time
```

```
plot(t, s3, type='l',main="survival curves for the exponential model assuming lambda = 0.01", xlab="Tim
```