Untitled

Denis Ostroushko

2023-02-08

```
knitr::opts_chunk$set(echo = F, message = F, warning = F, fig.height=4, fig.width=7,
                      fig.align='center', fig.pos = "H")
options(scipen=999)
```

1 Problem 1

1.0.1 (i)

We define retirement as the event of interest in this cross-sectional study. In order to retire, a person needs to be working. Therefore, a person enters the 'risk pool' when they start working, or entering the work force. So, when the researchers ask "At which age did you first enter the job market", they will know when each individual started being at the risk of an event.

1.0.2 (ii)

First, we need to understand if there will be any possibility of censoring.

Case I - No censoring

If we can observe the actual retirement age for each person, and can accurately record when an individual entered the work force, we can define our time to retirement, X, as the difference between retirement date, or year, and the year when they enter the work force.

Case II - Only Right Censoring

In this case we will define X as the difference between censoring or retirement, whichever occurs first, and the age at which an individual starts working.

Case III - Left censoring

If we take a survey and some people in the sample of 10,000 already retired, we would only be able to know the age by which they retire. Supposedly, everyone should remember when they started working, so we would be able to calculate the time to event. If they do not recall that information, we would only be able to know that they retire by a certain age.

1.0.3 (iii)

One example of right censoring that can occur is termination of study. Suppose that we stop observing these individuals on Feb 1, 2025. Then, all people who are did not retire by that point would be censored observations in our study.

2 Problem 2

2.0.1 (i)

As we discussed in class, left censoring can occur if we want to study how senior adults develop memory disabilities in the population of retirement homes.

We will record that event happened after a diagnosis occurs. We will have to periodically check in with the members.

Time-to-event will be the difference between diagnosis, or right censoring date, and the study start date, or enrollment date.

Left censoring can occur if at the time when we sample adults for the study some of randomly chosen members already have developed conditions that we study.

2.0.2 (ii)

When we sample players from a professional sport league in the middle of the season and try to analyze the number of games it takes for players to reach Y points total in the season. Some extraordinary players may be left censored because by the time we conduct the study they already achieve Y points before the start of the study.

2.0.3 (iii)

Left censoring can occur if HR studies tenure of employees, or time-to-quitting. Only employees that remain with the company will be included at the beginning of the study. Employees who have already left the company by the time study begins are left censored. Left censoring occurred because we record their tenure before the study begins.

3 Problem 3

3.0.1 (i)

As we discussed in class, pretty much any study that involves periodic check in will result in the interval censoring because we will not be able to tell exactly at what point does the event occur in between the check ins.

For example, we can study a treatment that aids patients in quitting smoking. Patients check in with the doctor once a month. If a patient reports that they have smoked in between appointments, we can know that a patient went 6 periods without smoking, but that is as accurate as we can be.

Event will be a patient smoking.

Time-to-event: time difference between study start and relapsing.

Reason: we check in with patients once a month.

3.0.2 (ii)

In economics we can study time to recession declaration after a certain economic metric or marker reached a certain value. The recession is declared when other certain economic markers reach their levels. There variables are recorded and reported on the quarterly basis, but if we were able to track these data in real time, there surely is a way for us to know when exactly unemployment reaches 10%, or whatever.

So, an event is an economic variable reaching a pre-defend level.

Time-to-event: number of quarters that occurs between a leading indicator and a recession.

Reason for censoring: inability to track data in real time, data can only be collected and organized on the quarterly basis.

3.0.3 (iii)

A medical device company wants to study time-to-event of device failure. Suppose we can only check implants when we send a technician to take a look at the devices, so if failure occurs, we would not be able to know exactly when the failure occurs.

Event: failure of a medical device.

Time-to-event: time between failure and implantation. Let's suppose that devices are to be replaced every 5 years due to regulations. Then, if devices does not fail through all of its life, I suppose this means that the observation is right censored.

Reason for censoring: we send a specialist to check the device only periodically.

A time-to-event study of the failure of a mechanical device: In this study, some devices may have failed between two study visits. The exact time of failure is unknown and can only be estimated to have occurred within a certain time interval. This is an example of interval censoring. The reason for censoring is that the exact time of failure is unknown.

4 Problem 4

The function $h(x) = a * e^{-b*X}$ for x, a, b all greater than 0 is a valid hazard function. We need to verify that it is non-negative everywhere, and although this function is decreasing rapidly, is is positive everywhere. Constant a is positive, and exponential functions are always positive everywhere, therefore their product is positive.

This requirement is given in the textbook in section 2.3 on page 27.

5 Problem 5

We are working with an exponential survival model. In order to compute the percentile for time-to-event random variable X, we need to find a function of percentile.

Let us denote p^{th} percentile is x_p. Then, the p^{th} percentile is given by

$$S(x_p) = e^{\frac{-x}{12.5}}$$

Solving for x_p we obtain a function:

$$x_p = -12.5 \times ln(Quantile)$$

.

So, 75^{th} percentile, we need to plug in 0.25 since, which is given by $-12.5 * ln(0.25) = 17.32868$. Therefore, for such an exponential model, at time 17.33 only 25% of initial observations remain.

Similarly, the median is the 50^{th} percentile, which occurs at time $-12.5 * ln(0.5) = 8.66434$.

Since we are working with an exponential model we can compute the mean directly using model parameter, λ, which is $\frac{1}{12.5}$. So, the mean is 12.5.

6 Problem 6

Given a function $h(x)$, we need obtain $S(x)$ first.

The general form of a gompertz hazard function is $h(x) = \theta * e^{\alpha * x}$, so in our case $\theta = 0.001$ and $\alpha = 0.01$.

The general form of a gompertz survival function is $S(x) = e^{\frac{\theta}{\alpha}*(1-e^{\alpha*x})}$. Plugging in obtained values for parameters gives us $S(x) = e^{-0.1*(1-e^{0.01*x})}$.

This solution is more involved than the previous problem, so I will lay out my steps in case I made a computational error. Additionally, I will use general form of parameters so I can translate the solution into an R function later:

$$S = e^{-\frac{\theta}{\alpha}(e^{\alpha*x}-1)},$$
$$e^{\alpha*x} = 1 - \frac{\alpha}{\theta}ln(S(x)),$$
$$\alpha * x = ln[1 - \frac{\alpha}{\theta}ln(S(x))],$$
$$x = \frac{1}{\alpha} * ln[1 - \frac{\alpha}{\theta}ln(S(x))]$$

```
gompertz_q2 <-
  function(x, theta, alpha){

    1/alpha  * log(1 - alpha/theta * log(x))

  }
```

Using specified parameters, we obtain the median survival time 207.08

7 Problem 7

7.0.1 (a)

First, we get KM survival curves, displayed of Figure 1. It appears that the survival probabilities are quite different for the two groups, as indicated by mostly non-overlapping confidence bands around the fitted curves.

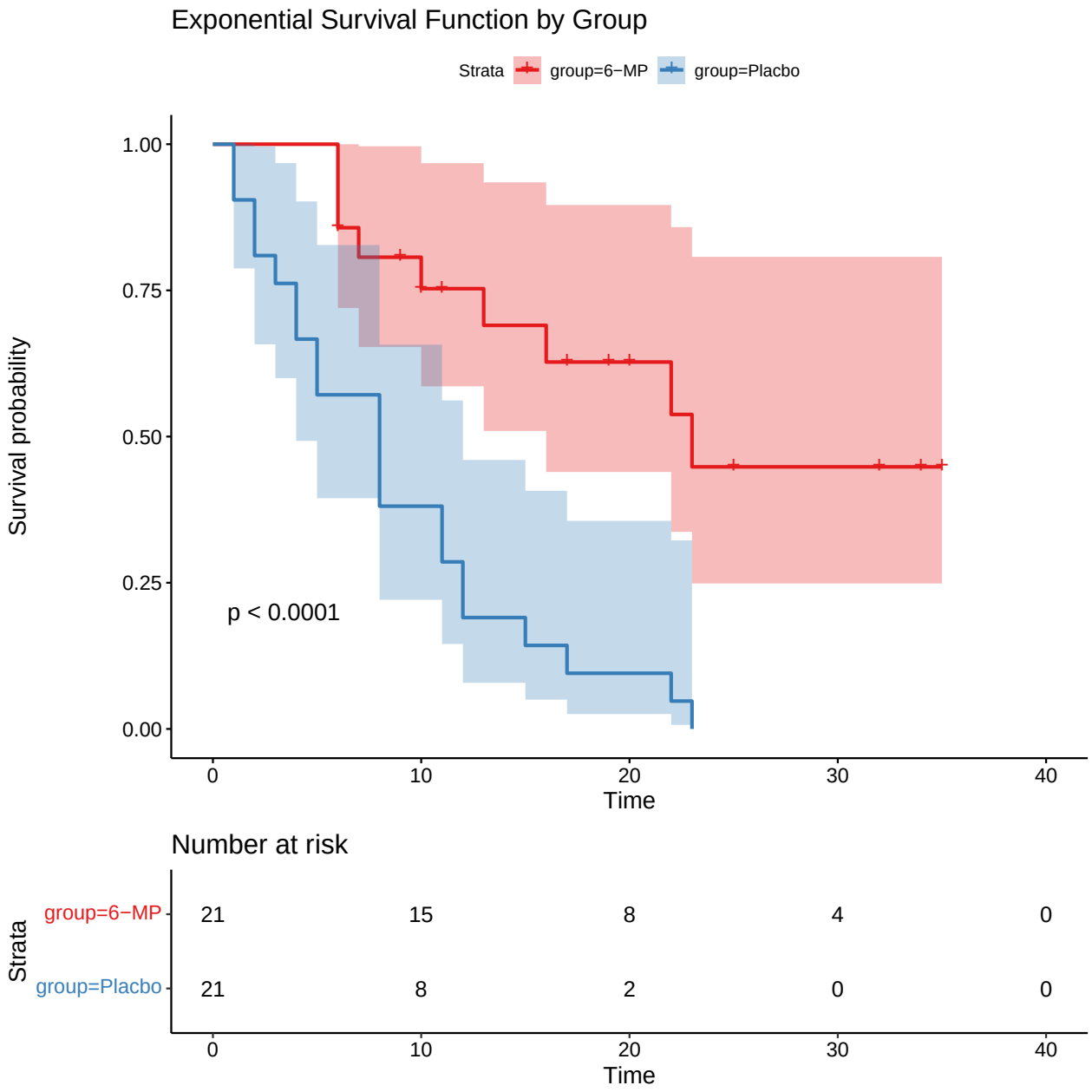

Figure 1: Kaplan Meier Survival curves

Now we fit an exponential survival model for each group. Results of model are given below. It appears that the two groups have different parameters, which are statistically significantly different.

```
surv_m <- survreg(Surv(weeks, relapse) ~group ,data = l_df,  dist = "exponential")
summary(surv_m)
```

```
##
## Call:
## survreg(formula = Surv(weeks, relapse) ~ group, data = l_df,
##     dist = "exponential")
##              Value Std. Error     z                    p
## (Intercept)  3.686      0.333 11.06 < 0.0000000000000002
```

```
## groupPlacebo -1.527      0.398 -3.83              0.00013
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -108.5   Loglik(intercept only)= -116.8
##  Chisq= 16.49 on 1 degrees of freedom, p= 0.000049
## Number of Newton-Raphson Iterations: 4
## n= 42
```

```
obj <- survfit(Surv(weeks, relapse) ~1 ,data = l_df[l_df$group == "6-MP", ])

lm(obj$cumhaz ~ l_df[l_df$group == "6-MP", ]$weeks %>% unique() %>% sort())
```

```
##
## Call:
## lm(formula = obj$cumhaz ~ l_df[l_df$group == "6-MP", ]$weeks %>%
##     unique() %>% sort())
##
## Coefficients:
##                                                  (Intercept)
##                                                      0.04648
## l_df[l_df$group == "6-MP", ]$weeks %>% unique() %>% sort()
##                                                      0.02277
```

7.0.2 (b)

We can use the fact that MLE estimates are asymptotically normally distributed, so we can compute a Wald confidence interval for each group using estimate and standard errors from the model output above.

In order to obtain a confidence interval for each estimate, we fit two separate models for two groups:

```
##
## Call:
## survreg(formula = Surv(weeks, relapse) ~ 1, data = l_df %>% filter(group ==
##     "6-MP"), dist = "exponential")
##             Value Std. Error    z                    p
## (Intercept) 3.686      0.333 11.1 <0.0000000000000002
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -42.2   Loglik(intercept only)= -42.2
## Number of Newton-Raphson Iterations: 4
## n= 21
```

```
##
## Call:
## survreg(formula = Surv(weeks, relapse) ~ 1, data = l_df %>% filter(group !=
##     "6-MP"), dist = "exponential")
##             Value Std. Error   z                    p
## (Intercept) 2.159      0.218 9.9 <0.0000000000000002
##
## Scale fixed at 1
##
## Exponential distribution
```

```
## Loglik(model)= -66.3   Loglik(intercept only)= -66.3
## Number of Newton-Raphson Iterations: 4
## n= 21
```

The upper and lower bounds for the 95% confidence interval of the 6-MP group are given by 3.686 $\pm$ 1.96 * 0.333. Thus, parameter λ for the 6-MP group is 3.686, bounded by (3.033, 4.339).

Parameter λ for placebo group is given by 3.686 - 1.527 = 2.159.

The upper and lower bounds for the 95% confidence interval of the placebo group are given by 2.159 $\pm$ 1.96 * 0.398 Thus, parameter λ for the 6-MP group is 2.159, bounded by (1.732, 2.586).

We can check our work by calling the `confint` function in R:

```
##                2.5 %   97.5 %
## (Intercept) 3.032776 4.339419
```

```
##                2.5 %   97.5 %
## (Intercept) 1.731785 2.587183
```

If we look at the confint of the original model we fit, the confidence interval for placebo is given for the difference in rates, and not for the λ of placebo group. However, it will be useful in the next section.

7.0.3 (c)

We set up a log-rank test and give results below:

```
# perform the log-rank test for two parameters being equal
result <- survdiff(Surv(weeks, relapse) ~ group,data = l_df, rho=0)
```

- H_0: the exponential rates of the 6-MP and Placebo groups are identical
- H_a: the exponential rates of the 6-MP and Placebo groups differ
- Chi-square test statistic: 16.8
- P-value: <0.0001. We have enough statistical evidence to conclude that the rates for two groups are different. Therefore those in the experimental arm 6-MP are more likely to survive for an extended period of time.

7.0.4 (d)

For a one sided test we will use a t test.

- H_0 : exponential rates for two groups are the same
- H_a : exponential rate of the 6-MP group is lower than that of Placebo group
- Estimates with the data: $\hat{\lambda}_{6-MP} = 0.02507$; $\hat{\lambda}_{Placebo} = 0.115385$
- Variance for the difference is the sum of varainces of two estimtes, since they are independent normal variables. Variance for the difference is 0.000704
- Cutoff z-value at the lower 5% of the standard normal distribution is 1.65, and the test statistic is -3.4
- P-value is 0.000332
- Conclusion: we