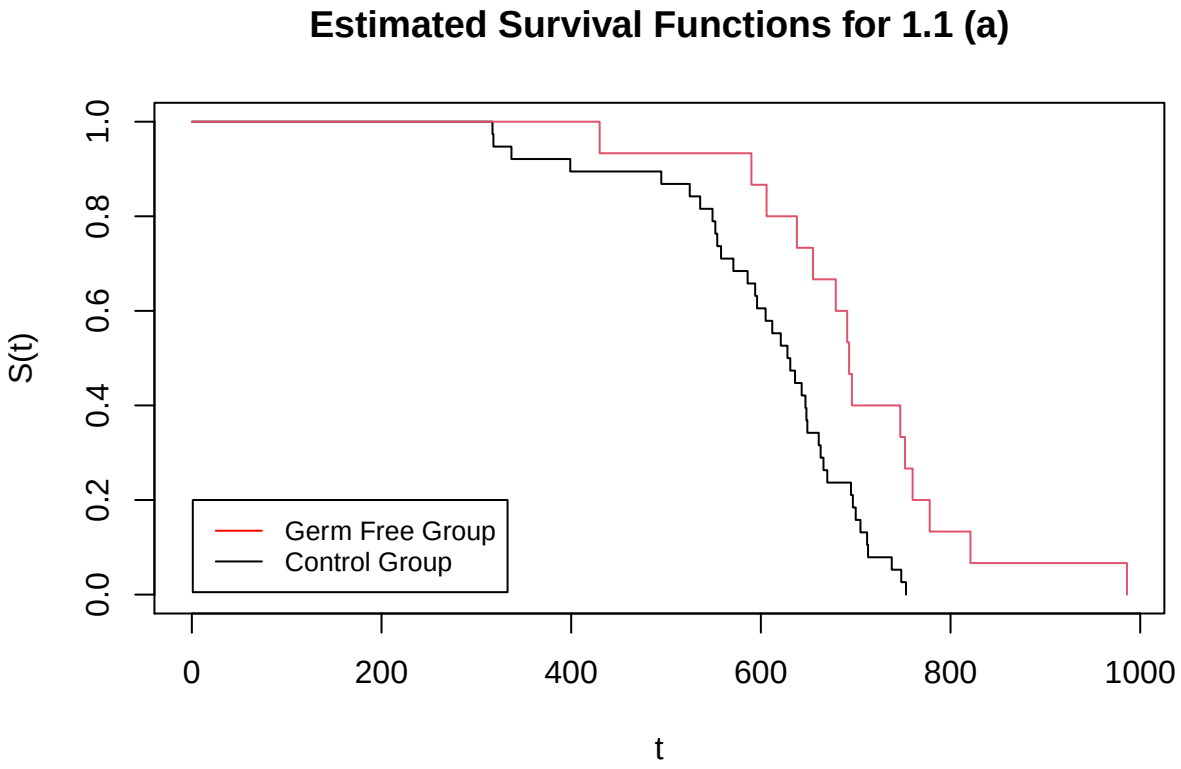
Problem 1.1

Youngjin Cho

2021 9 6

(a) : Ignore Censored Data



I plotted estimated survival functions for the two groups. The red line is for the $\hat{S}_G(t)$, which is estimated survival function for the Germ Free group. The black line is for the $\hat{S}_C(t)$, which is estimated survival function for the Control group. Considering the estimated survival plots, it seems like mice in Germ Free group are likely to survive longer than mice in Control group since their estimated survival function is higher than that of mice in the other group. So one can think that germ free environment reduce the risk of reticulum cell sarcoma.

```r
## Endpoint Survival Function Estimates for Control Group
fit.lm.a.C <- survfit(Surv(time)~1,data=dataset_v_a[dataset_v_a$group==1,])
summary(fit.lm.a.C,times=c(753))$surv
```

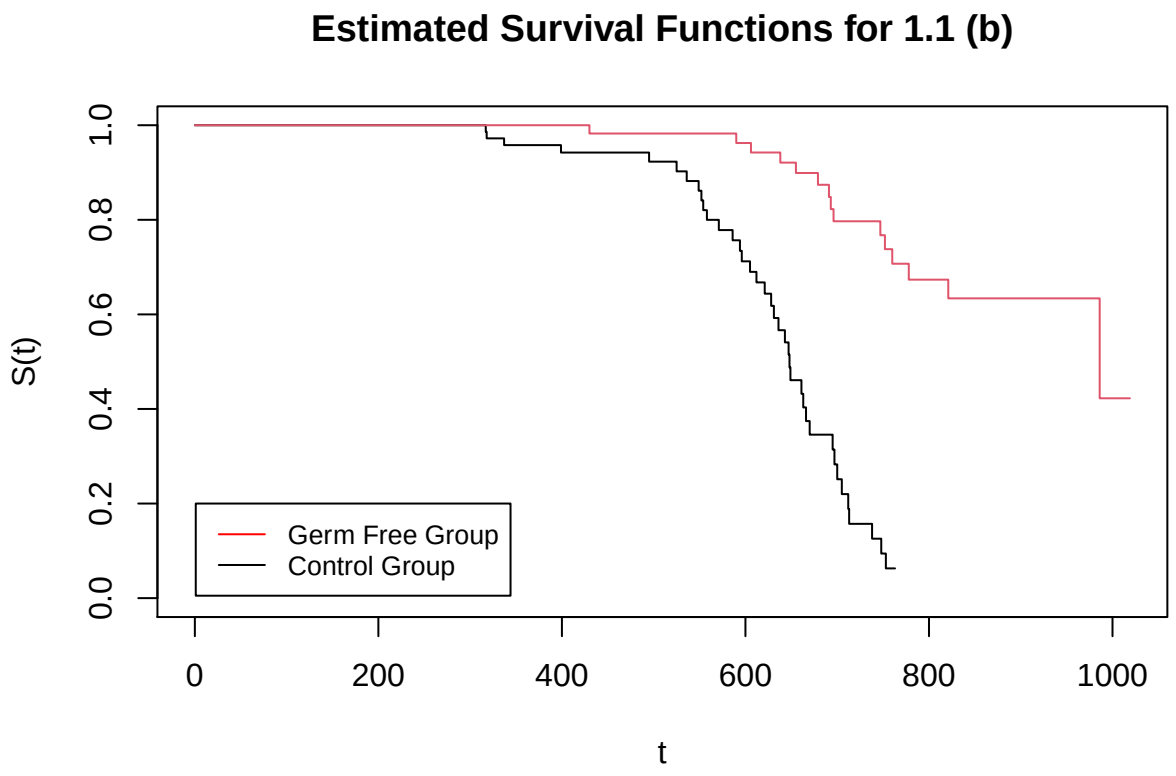
```
## [1] 0
```

```r
## Endpoint Survival Function Estimates for Germ Free Group
fit.lm.a.G <- survfit(Surv(time)~1,data=dataset_v_a[dataset_v_a$group==2,])
summary(fit.lm.a.G,times=c(986))$surv
```

```
## [1] 0
```

Now I calculated estimated survival function for the endpoint for each group. For Control group, the endpoint is 753. So I calculated $\hat{S}_C(t = 753) = 0$. For Germ Free group, the endpoint is 986. So I calculated $\hat{S}_G(t = 986) = 0$. It is natural since we do not used any censored data, which makes estimated survival function as $(\#\text{ obs } > t)/n$, making endpoint estimate as 0. The good point of ignoring censored data points is we can concentrate on exact information for the death of our interest : reticulum cell sarcoma.

(b) : Include Censored Data



I plotted estimated survival functions for the two groups. The red line is for the $\hat{S}_G(t)$, which is estimated survival function for the Germ Free group. The black line is for the $\hat{S}_C(t)$, which is estimated survival function for the Control group. Considering the estimated survival plots, it seems like mice in Germ Free

group are likely to survive longer than mice in Control group since their estimated survival function is higher than that of mice in the other group. Also, comparing to the result in (a), one can see that the shape of estimated survival functions changed much, which is the result of partial information in censored data. Due to the information from censored data, the difference between $\hat{S}_G(t)$ and $\hat{S}_C(t)$ is more severe than that of (a), which makes us think more likely that germ free environment reduce the risk of reticulum cell sarcoma.

```
## Endpoint Survival Function Estimates for Control Group
fit.lm.b.C <- survfit(Surv(time,delta)~1,data=dataset_v_b[dataset_v_b$group==1,])
summary(fit.lm.b.C,times=c(753))$surv
```

```
## [1] 0.06284176
```

```
## Endpoint Survival Function Estimates for Germ Free Group
fit.lm.b.G <- survfit(Surv(time,delta)~1,data=dataset_v_b[dataset_v_b$group==2,])
summary(fit.lm.b.G,times=c(986))$surv
```

```
## [1] 0.4225683
```

Now I calculated estimated survival function for the endpoint for each group. For Control group, the endpoint is 753. So I calculated $\hat{S}_C(t = 753) = 0.063$. For Germ Free group, the endpoint is 986. So I calculated $\hat{S}_G(t = 986) = 0.423$. The result is totally different from that in (a), which calculates all end point estimated survival function as 0. As (b) used partial information in censored data, it estimates survival function in the enpoint larger than 0 (there are censored data after endpoint). The good point of using censored data is that we can use given information as many as possible to estimate the survival function, not only using exact information, but also using partial information for the risk of reticulum cell sarcoma.

R Code

```
# Data for Control Group
TL_C <- c(159,189,191,198,200,207,220,235,245,
          250,256,261,265,266,280,343,356,383,
          403,414,428,432)
RS_C <- c(317,318,399,495,525,536,549,552,554,
          337,558,571,586,594,596,605,612,621,
          628,631,636,643,647,648,649,661,663,
          666,670,695,697,700,705,712,713,738,
          748,753)
OC_C <- c(040,042,051,062,163,179,206,222,228,
          252,249,282,324,333,341,366,385,407,
          420,431,441,461,462,482,517,517,524,
          564,567,586,619,620,621,622,647,651,
          686,761,763)

# Data for Germ Free Group
TL_G <- c(158,192,193,194,195,202,212,215,229,
          230,237,240,244,247,259,300,301,321,
          337,415,434,444,485,496,529,537,624,
          707,800)
RS_G <- c(430,590,606,638,655,679,691,693,696,
          747,752,760,778,821,986)
```

```r
OC_G <- c(136,246,255,376,421,565,616,617,652,
          655,658,660,662,675,681,734,736,737,
          757,769,777,800,807,825,855,857,864,
          868,870,870,873,882,895,910,934,942,
          1015,1019)

# Dataframe
group <- c(rep(1,length(TL_C)+length(RS_C)+length(OC_C)),
           rep(2,length(TL_G)+length(RS_G)+length(OC_G)))
type <- c(rep("TL",length(TL_C)),rep("RS",length(RS_C)),
          rep("OC",length(OC_C)),rep("TL",length(TL_G)),
          rep("RS",length(RS_G)),rep("OC",length(OC_G)))
time <- c(TL_C,RS_C,OC_C,TL_G,RS_G,OC_G)
dataset_v <- data.frame(group,type,time)

# 1.1 a
dataset_v_a <- dataset_v[dataset_v$type=="RS",]
## Comparing Survival Functions
fit.lm.a <- survfit(Surv(time)~group,data=dataset_v_a)
plot(fit.lm.a,col=1:2, xlab="t", ylab="S(t)", main="Estimated Survival Functions for 1.1 (a)")
legend(1, 0.2, legend=c("Germ Free Group", "Control Group"),
       col=c("red", "black"), lty=1:1, cex=0.8)
## Endpoint Survival Function Estimates for Control Group
fit.lm.a.C <- survfit(Surv(time)~1,data=dataset_v_a[dataset_v_a$group==1,])
summary(fit.lm.a.C,times=c(753))$surv
## Endpoint Survival Function Estimates for Germ Free Group
fit.lm.a.G <- survfit(Surv(time)~1,data=dataset_v_a[dataset_v_a$group==2,])
summary(fit.lm.a.G,times=c(986))$surv

# 1.1 b
dataset_v_b <- dataset_v
dataset_v_b$delta <- as.numeric(dataset_v$type=="RS")
## Comparing Survival Function Estimates
fit.lm.b <- survfit(Surv(time,delta)~group,data=dataset_v_b)
plot(fit.lm.b,col=1:2, xlab="t", ylab="S(t)", main="Estimated Survival Functions for 1.1 (b)")
legend(1, 0.2, legend=c("Germ Free Group", "Control Group"),
       col=c("red", "black"), lty=1:1, cex=0.8)
## Endpoint Survival Function Estimates for Control Group
fit.lm.b.C <- survfit(Surv(time,delta)~1,data=dataset_v_b[dataset_v_b$group==1,])
summary(fit.lm.b.C,times=c(756))$surv
## Endpoint Survival Function Estimates for Germ Free Group
fit.lm.b.G <- survfit(Surv(time,delta)~1,data=dataset_v_b[dataset_v_b$group==2,])
summary(fit.lm.b.G,times=c(986))$surv
```