ST227 - Exercise 4

Exercise 1

Cancer patients who are in remission are observed and the number of days until the symptoms reappear is recorded. Some records have been right-censored. The data set is provided in a spreadsheet named `smoker_data.xlsx` and the columns therein are:

- `time`: the time until reappearance of symptoms in number of days.
- `event`: an indicator variable taking value 0 if the record has been right-censored and 1 if fully observed.
- `fullyObserved`: logical variable indicating whether the record has been fully observed.
- `smoker`: categorical variable with value 0 for non-smoker (the reference group) and 1 for smoker.

1.a

In R, calculate the Kaplan-Meier estimate for survival probabilities and plot them.

```
library("readxl")

cancer <- as.data.frame(read_excel("smoker_data.xlsx"))
head(cancer)
```

```
##   time event fullyObserved smoker
## 1   13     1          TRUE      1
## 2   55     1          TRUE      0
## 3   20     1          TRUE      0
## 4   17     1          TRUE      0
## 5   42     1          TRUE      1
## 6    4     0         FALSE      1
```

We observe that the dataset is not sorted according to `time`. As a first step, let us sort it.

```
inds <- sort.int(cancer$time, index.return = TRUE)$ix
cancer <- cancer[inds, ]
head(cancer)
```

```
##    time event fullyObserved smoker
## 6     4     0         FALSE      1
## 12    5     1          TRUE      0
## 11    6     1          TRUE      1
## 19    8     1          TRUE      1
## 23   12     1          TRUE      0
## 1    13     1          TRUE      1
```

Now recall the Kaplan-Meier estimator for the survival function

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right),$$

with t_i a time when at least one event happened, d_i the number of events (e.g., deaths) that happened at time t_i, and n_i the individuals known to have survived (have not yet had an event or been censored) up to time t_i.

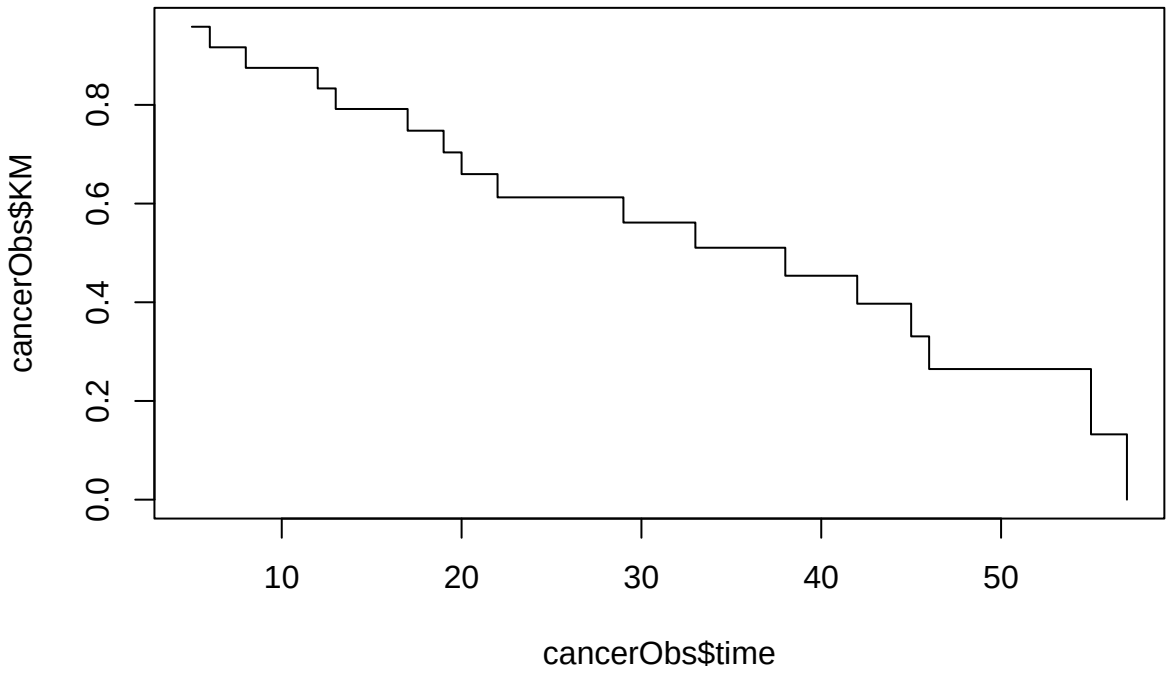
We implement this in R:

```r
# vector with entries n_i
cancer$atRisk <- nrow(cancer) : 1

# filter the fully observed rows only
cancerObs <- cancer[cancer$fullyObserved, ]

# vector with entries d_i
cancerObs$death <- 1

# vector with entries 1 - d_i / n_i
cancerObs$survProb <- (cancerObs$atRisk - cancerObs$death) / cancerObs$atRisk

# caluclate survival function
cancerObs$KM <- cumprod(cancerObs$survProb)

cancerObs$KM
```

```
##  [1] 0.9583333 0.9166667 0.8750000 0.8333333 0.7916667 0.7476852 0.7037037
##  [8] 0.6597222 0.6125992 0.5615493 0.5104993 0.4537772 0.3970550 0.3308792
## [15] 0.2647034 0.1323517 0.0000000
```

We can now plot the Kaplan-Meier curve

```r
plot(cancerObs$time, cancerObs$KM, type = "s")
```



1.b

Using the Greenwood's formula:

$$\mathrm{Var}(\hat{S}(t)) \approx (\hat{S}(t))^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}, \quad t \in [t_{(k)}, t_{(k+1)}), \tag{1}$$

calculate the variance of $\hat{S}(t)$ at the fully observed times $t_{(i)}$, $i = 1, 2, \ldots$ in R.

```
cancerObs$greenwoodVar <- cancerObs$KM^2 *
    cumsum(
        cancerObs$death / (cancerObs$atRisk * (cancerObs$atRisk - cancerObs$death))
    )
cancerObs$greenwoodVar
```

```
##  [1] 0.001663773 0.003182870 0.004557292 0.005787037 0.006872106 0.007956655
##  [7] 0.008868694 0.009608223 0.010346607 0.011082945 0.011528637 0.011968959
## [13] 0.011978961 0.011968091 0.011162972 0.011549227         NaN
```

1.c

By using the R `survival` package or otherwise, calculate the MLE for the Cox Proportional Hazard Model with `time` as the response variable and `smoker` as the categorical covariate.

```
library(survival)
```

```
##
## Attaching package: 'survival'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     cancer
```

```
survivalObject <- Surv(cancer$time, cancer$event)
coxmodel <- coxph(survivalObject ~ smoker, data = cancer)
summary(coxmodel)
```

```
## Call:
## coxph(formula = survivalObject ~ smoker, data = cancer)
##
##   n= 25, number of events= 17
##
##           coef exp(coef) se(coef)     z Pr(>|z|)
## smoker 0.08854   1.09258  0.52274 0.169    0.865
##
##        exp(coef) exp(-coef) lower .95 upper .95
## smoker     1.093     0.9153    0.3922     3.044
##
## Concordance= 0.505  (se = 0.072 )
## Likelihood ratio test= 0.03  on 1 df,   p=0.9
## Wald test            = 0.03  on 1 df,   p=0.9
## Score (logrank) test = 0.03  on 1 df,   p=0.9
```

1.d

Based on the output you have generated, perform the z-test, Score test, and Likelihood Ratio test on the following hypotheses:

$$H_0 : \beta = 0, \quad \text{vs}\, H_1 : \beta \neq 0\,..$$

The Cox Proportional Hazard model has only one parameter, β, which has a point estimate value of 0.08854. With the p-value being approximately 0.9 at all tests, we do not reject the null hypothesis at any reasonable significance level (be careful, one never accepts the null hypothesis).

Exercise 2

The file `ST227_exam.xlsx` contains data of students taking an old ST227 exam, and consists of the following variables:

- `time`: time in minutes taken to complete the exam
- `revised`: a categorical variable taking value 1 if the student revised for the exam and 0 else

2.a

If a student's recorded time is exactly the allotted time we may assume they did not finish the whole exam in the allotted time of 120 minutes, and given more time they would have continued writing. Add a categorical variable called `censored` to your data frame taking value 0 if you believe the student did not finish the whole exam and 1 else.

```
library("readxl")

ST227 <- as.data.frame(read_excel("ST227_exam.xlsx"))

ST227$censored <- as.numeric(ST227$time < 120)
```

2.b

By using the `survival` package or otherwise, calculate the MLE for the Cox Proportional Hazard Model. Using an appropriate statistical test, test at the 95% level $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$. Interpret the output of your test.

```
library(survival)
survivalObject <- Surv(ST227$time, ST227$censored)
coxmodel <- coxph(survivalObject ~ revised, data = ST227)
summary(coxmodel)
```

```
## Call:
## coxph(formula = survivalObject ~ revised, data = ST227)
##
##   n= 1000, number of events= 658
##
##             coef exp(coef) se(coef)     z Pr(>|z|)
## revised 0.86679   2.37926  0.07946 10.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## revised     2.379     0.4203     2.036      2.78
##
## Concordance= 0.62  (se = 0.009 )
## Likelihood ratio test= 120.6  on 1 df,   p=<2e-16
## Wald test            = 119  on 1 df,   p=<2e-16
## Score (logrank) test = 126.3  on 1 df,   p=<2e-16
```

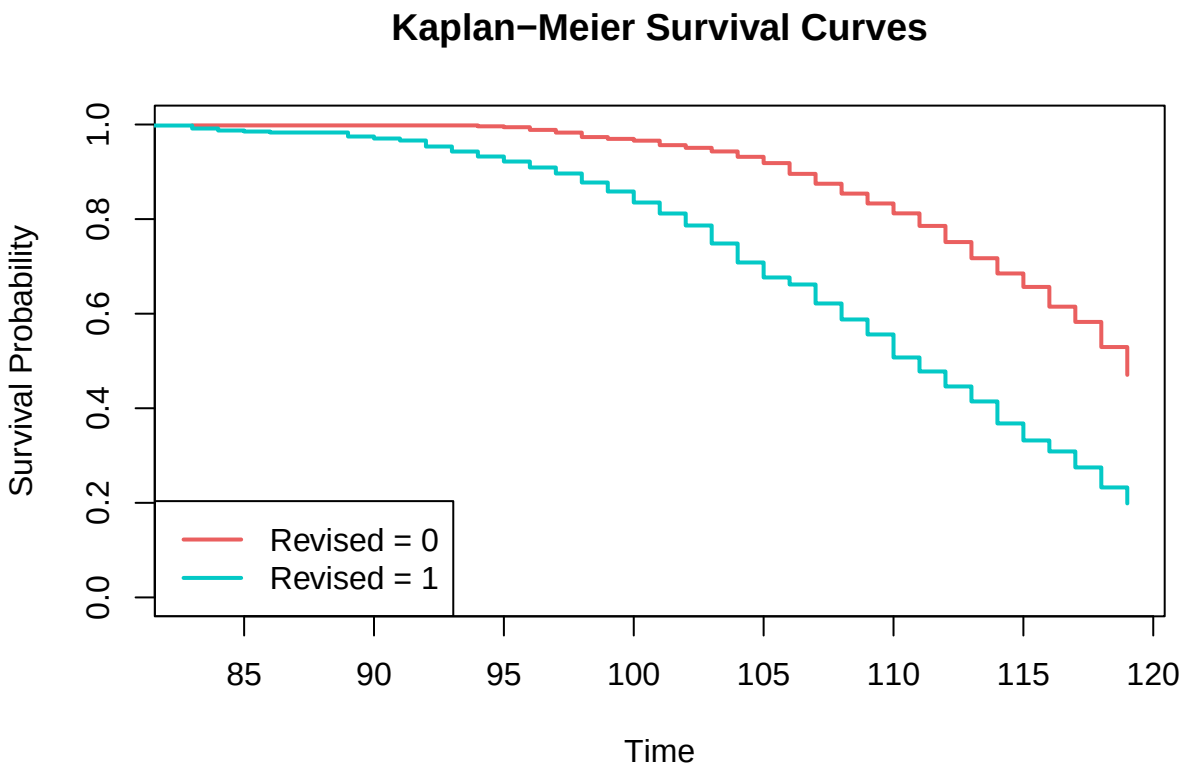
We reject the null at the given significance level.

2.c

Repeat the same analysis using the Kaplan-Meier estimator.

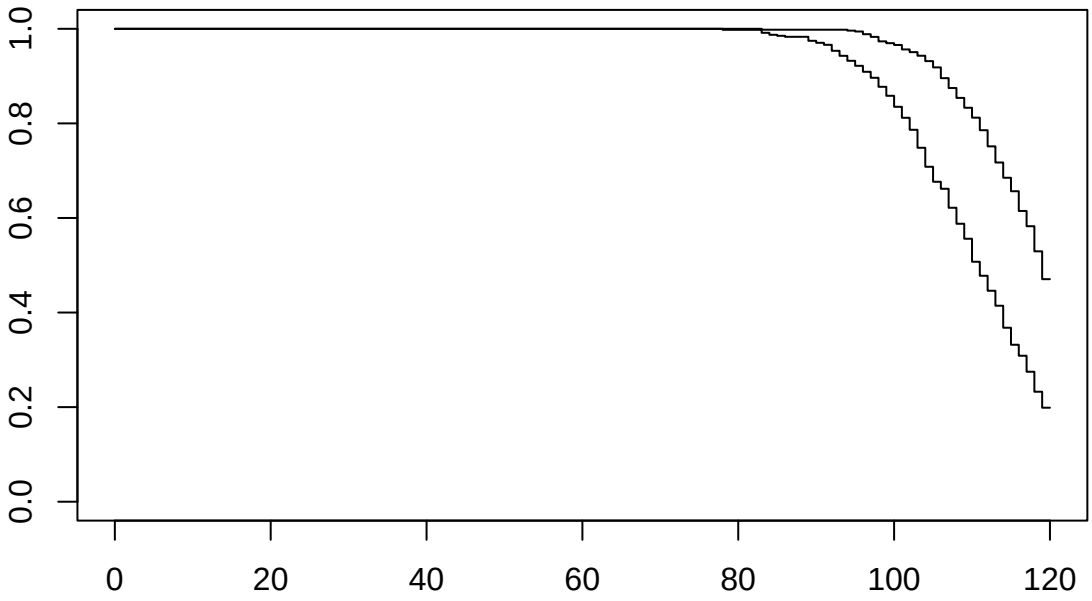
```
# Subset data based on 'revised' variable
ST227_0 <- ST227[ST227$revised == 0, ]
```

```r
ST227_1 <- ST227[ST227$revised == 1, ]

# Order subset data
inds <- sort.int(ST227_0$time, index.return = TRUE)$ix
ST227_0 <- ST227_0[inds, ]

inds <- sort.int(ST227_1$time, index.return = TRUE)$ix
ST227_1 <- ST227_1[inds, ]


# Compute Kaplan-Meier estimates for revised == 0
ST227_0$atRisk <- nrow(ST227_0) : 1
ST227_0_censored <- ST227_0[ST227_0$censored == 1, ]
ST227_0_censored$submit <- 1
ST227_0_censored$survProb <- (ST227_0_censored$atRisk - ST227_0_censored$submit) / ST227_0_censored$atRi
ST227_0_censored$KM <- cumprod(ST227_0_censored$survProb)

# Compute Kaplan-Meier estimates for revised == 1
ST227_1$atRisk <- nrow(ST227_1) : 1
ST227_1_censored <- ST227_1[ST227_1$censored == 1, ]
ST227_1_censored$time_out <- 1
ST227_1_censored$survProb <- (ST227_1_censored$atRisk - ST227_1_censored$time_out) / ST227_1_censored$a
ST227_1_censored$KM <- cumprod(ST227_1_censored$survProb)

# Plot the Kaplan-Meier curves
plot(ST227_0_censored$time, ST227_0_censored$KM, type = "s", col = "#ea5f5f", xlab = "Time", ylab = "Su
     main = "Kaplan-Meier Survival Curves", lwd = 2, ylim = c(0,1))
lines(ST227_1_censored$time, ST227_1_censored$KM, type = "s", col = "#04c9c6", lwd = 2)
legend("bottomleft", legend = c("Revised = 0", "Revised = 1"), col = c("#ea5f5f", "#04c9c6"), lwd = 2)
```



Or alternatively, you can use the `survival` package.

```
km_fit <- survfit(Surv(time, censored) ~ revised, data = ST227)
plot(km_fit)
```



Exercise 3

Cancer patients who are in remission are observed and the number of days until the symptoms reappear is recorded. Some records have been right-censored. The data set is provided in a spreadsheet named `cancer.xlsx` and the columns therein are:

- `time`: the time until reappearance of symptoms in number of days.
- `event`: an indicator variable taking value 0 if the record has been right-censored and 1 if fully observed.
- `fullyObserved`: logical variable indicating whether the record has been fully observed.
- `smoksexer`: categorical variable with value 0 for male (the reference group) and 1 for female.

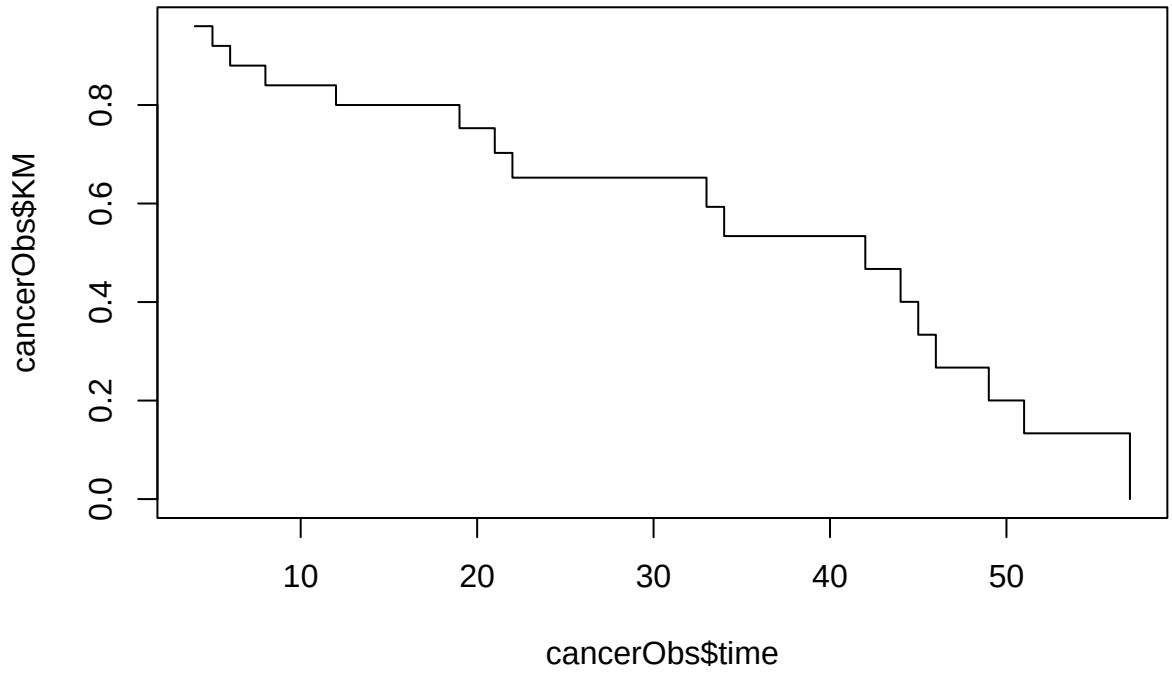
3.a

In R, calculate the Kaplan-Meier estimate for survival probabilities and plot them.

```
library("readxl")

cancer <- as.data.frame(read_excel("cancer.xlsx"))
inds <- sort.int(cancer$time, index.return = TRUE)$ix
cancer <- cancer[inds, ]
cancer$atRisk <- nrow(cancer) : 1

cancerObs <- cancer[cancer$fullyObserved, ]
cancerObs$death <- 1
cancerObs$survProb <- (cancerObs$atRisk - cancerObs$death) / cancerObs$atRisk
cancerObs$KM <- cumprod(cancerObs$survProb)

cancerObs$KM
```

```
##  [1] 0.9600000 0.9200000 0.8800000 0.8400000 0.8000000 0.7529412 0.7027451
##  [8] 0.6525490 0.5932264 0.5339037 0.4671658 0.4004278 0.3336898 0.2669519
## [15] 0.2002139 0.1334759 0.0000000
```

```r
plot(cancerObs$time, cancerObs$KM, type = "s")
```



3.b

Denote by T the time until reappearance of cancer. Using the following formula:

$$\mathrm{E}\,(T^n) = \int_0^\infty nt^{n-1}\,\Pr(T > t)dt\,,$$

propose and calculate a suitable estimation for $\mathrm{Var}(T)$. (Hint: you can use Midpoint Rule, Trapezoidal/Trapezium Rule or any geometric method of approximating the area under the curve.)

The variance is given by

$$\mathrm{E}(T^2) - \mathrm{E}(T)^2\,,$$

and

$$\mathrm{E}(T^2) = \int_0^\infty 2t\,\Pr(T > t)dt, \quad \mathrm{E}(T) = \int_0^\infty \Pr(T > t)dt\,.$$

The survival function is by definition $S(t) = \Pr(T > t)$ and we have a Kaplan-Meier estimate $\hat{S}(t)$ of $S(t)$, which we can substitute for $\Pr(T > t)$.

We see that our KM-estimate of $S(t)$ is 0 for $t \geq 57$. Hence we can use the follwing approximation alagorithm:

1. Place N equally spaced grid-points on the intervals $[0, 57]$
2. For each grid-point x_i, find $\hat{S}(x_i)$
3. In each interval, use the Trapezoidal rule to approximate the area under the curve

```r
N <- 10000
xs <- seq(from = 0, to = 57, length.out = N)
ts <- cancerObs$time

inds <- vapply(xs, function(x){which(ts - x >= 0)[1]}, 0)
S_of_xs <- (c(1, cancerObs$KM))[inds]
```

```r
E_aprrox <- (2 * sum(S_of_xs) - S_of_xs[1] - S_of_xs[N]) / (2 * N)
E_sq_integrand <- 2 * xs * S_of_xs
E_sq_approx <- (2 * sum(E_sq_integrand) - E_sq_integrand[1] - E_sq_integrand[N]) / (2 * N)

Var_approx <- E_sq_approx - E_aprrox^2

Var_approx
```

```
## [1] 25.92466
```

3.c

Using the R `survival` package or otherwise, calculate the MLE for the Cox Proportional Hazard Model with `time` as the response variable and `sex` as the categorical covariate.

```r
library(survival)

survivalObject <- Surv(cancer$time, cancer$event)
coxmodel <- coxph(survivalObject ~ sex, data = cancer)
summary(coxmodel)
```

```
## Call:
## coxph(formula = survivalObject ~ sex, data = cancer)
##
##   n= 25, number of events= 17
##
##       coef exp(coef) se(coef)     z Pr(>|z|)
## sex 0.1526    1.1649   0.5251 0.291    0.771
##
##     exp(coef) exp(-coef) lower .95 upper .95
## sex     1.165     0.8584    0.4162      3.26
##
## Concordance= 0.48  (se = 0.077 )
## Likelihood ratio test= 0.09  on 1 df,   p=0.8
## Wald test            = 0.08  on 1 df,   p=0.8
## Score (logrank) test = 0.08  on 1 df,   p=0.8
```

3.d.

Based on the output you have generated, perform the z-test, Score test, and Likeli- hood Ratio test on the following hypotheses:

$$H_0 : \beta = 0, \quad \text{vs}\, H_1 : \beta \neq 0\,.$$

The Cox Proportional Hazard model has only one parameter, β, which has a point estimate value of 0.1526. With the p-value being approximately 0.8 at all tests, we do not reject the null hypothesis at any reasonable significance level (be careful, one never accepts the null hypothesis).