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.

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.

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.

1.d

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

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

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.

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.

2.c

Repeat the same analysis using the Kaplan-Meier estimator.

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.

3.b

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

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

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.)

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.

3.d.

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

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