MATH-449 - Biostatistics
EPFL, Spring 2025
Problem Set 9

1. As before we let $X \in \{0,1\}$ be a treatment, T be a survival time, and T^x for $x \in \{0,1\}$ be the survival time if X is set to x. We assume that causal consistency holds, i.e. that $T = XT^{x=1} + (1-X)T^{x=0}$, and we are interested in estimating functions of T^x, e.g. $P(T^x > t)$. We let L be a set of pretreatment variables (we typically take L to be a discrete variable, but this is unimportant). We have earlier seen that, under conditional exchangeability given L, $T^x \perp\!\!\!\perp X|L$[1], $P(T^x > t)$ can be written as

$$E\left[I(T^x > t)\right] = P(T^x > t) = \sum_l P(T > t \mid X = x, L = l)P(L = l). \tag{1}$$

Suppose we know how to model $P(T > t|X = x, L = l)$. The formulation (1) then suggests an estimator for $P(T^x > t)$: By using estimates $\hat{P}(L = l)$ and $\hat{P}(T > t|X = x, L = l)$ from the model, multiplying them together, and summing over l, we can estimate $P(T^x > t)$ with

$$\sum_l \hat{P}(T > t \mid X = x, L = l)\hat{P}(L = l). \tag{2}$$

(a) Verify that (1) can be written as

$$E\left[\frac{I(X = x)}{P(X = x \mid L)} I(T > t)\right] \tag{3}$$

by the law of total probability, assuming that $P(X = x|L) > 0$ almost surely.

(b) Suppose you are given an i.i.d. sample $\{(X_i, T_i, L_i)\}_{i=1}^n$ (i.e. assume there is no censoring), and that you have access to estimates $\hat{P}(X = x|L = l)$ of $P(X = x|L = l)$ for $x \in \{0,1\}$ and all l. Argue that the formulation (3) suggests the *inverse probability weighted estimator*

$$\frac{1}{n}\sum_{i=1}^n \frac{I(X_i = x)}{\hat{P}(X_i = x \mid L = L_i)} I(T_i > t) \tag{4}$$

for $P(T^x > t)$.

(c) We will now inspect the performance of the estimators (2) and (4) using simulations. We will do this by

(i) Generate an i.i.d. sample $\{L_i\}_{i=1}^n$ from a known distribution.
(ii) Generate $\{X_i\}_{i=1}^n$ independently from the same distribution, but such that X_i depends on the realisation L_i.
(iii) Generate $\{T_i\}_{i=1}^n$ as in ii), such that T_i depends on the realisations L_i and X_i.

A simple example of this is shown below:

```
library(zoo)
library(survival)
set.seed(42)
n = 1e3
timegrid = seq(0,10,length.out=1e4)

L_prob_vec = c(0.1,0.25,0.15,0.1,0.4)
len = length(L_prob_vec)
# Sampling L with possible values 1,...,5 based
```

[1]Recall that, under exchangeability ($T^x \perp\!\!\!\perp X$ for $x \in \{0,1\}$) we have that the identity $P(T^x > t) = P(T > t|X = x)$ holds. $P(T^x > t)$ can thus be assessed as $P(T > t|X = x)$ can be estimated from observed data. Exchangeability will often fail to hold in real-life settings, so we will need to come up with formulas for estimating $P(T^x > t)$ under weaker assumptions.

```
# on the probabilities in L_prob_vec
L = sample(1:len,n,replace = T,prob = L_prob_vec)
# Defining probabilities for outcomes of X given realisations of L
X_prob = 1 - (0.90/L) # Sampling X
X = rbinom(n,1,prob=c(X_prob,1-X_prob))
# Sampling exponentially distributed survival times depending on
# the realisations of X and L (no censoring)
TT = rexp(n,1+0.5*X - L*0.9/len)
# Estimation:
# Estimating the distribution of L
L_dist = table(L)/length(L)
# We can implement the estimator (2) without modelling assumptions since L
# is discrete and the sample size is large
survMat_X0 = survMat_X1 = matrix(NA,ncol=length(timegrid),nrow=len)
survMat_X0[,1]=survMat_X1[,1] = 1
for(i in 1:len){ if(any(X==0 & L==i)){
  SFT_X0Li = survfit(Surv(TT[X==0 & L==i])~1)
  tms0 = sapply(SFT_X0Li$time, function(tm) max(which( timegrid <= tm )))
survMat_X0[i, tms0 ] = SFT_X0Li$surv
}
if(any(X==1 & L==i)){
SFT_X1Li = survfit(Surv(TT[X==1 & L==i])~1)
tms1 = sapply(SFT_X1Li$time, function(tm) max(which( timegrid <= tm )))
survMat_X1[i, tms1 ] = SFT_X1Li$surv
}
survMat_X0[i,] = na.locf(survMat_X0[i,])
survMat_X1[i,] = na.locf(survMat_X1[i,])
survMat_X0[i,] = survMat_X0[i,] * L_dist[i]
survMat_X1[i,] = survMat_X1[i,] * L_dist[i] }
# The estimator (2):
S_X0 = apply(survMat_X0,2,sum)
S_X1 = apply(survMat_X1,2,sum)
# Estimating the probability of X=0 given L=i and X=1 given L=i
PX0_L = PX1_L = rep(0,len)
for(i in 1:len){
PX0_L[i] = sum( X==0 & L==i )/sum(L==i)
PX1_L[i] = sum( X==1 & L==i )/sum(L==i) }
# The estimator (3):
S_X0_ipw = sapply(timegrid,function(tm) mean(1*(X==0)/PX0_L[L] * 1*(TT>= tm)))
S_X1_ipw = sapply(timegrid,function(tm) mean(1*(X==1)/PX1_L[L] * 1*(TT>= tm)))
# The estimated survival differences:
Surv_difference = S_X1 - S_X0
Surv_difference_ipw = S_X1_ipw - S_X0_ipw
# plot(timegrid,Surv_difference,type="s")
# lines(timegrid,Surv_difference_ipw,type="s",col=3)
```

Extending on the above code, write a routine for comparing the variances of the estimators (2) and (4) in this simulation setting (i.e. with a discrete variable L taking values in $\{1,2,3,4,5\}$ with probabilities given in `L_prob_vec`, $P(X=1|L=l) = 1 - \frac{0.9}{l}$, and T given X and L exponentially distributed with rate $1+0.5X-\frac{0.9}{len}L$, where len is the length of the vector `L_prob_vec`). You can do this by repeating the simulation m times and store the obtained estimates `Surv_difference` and `Surv_difference_ipw` after each repetition. Then, using the stored simulations, obtain approximate 95% confidence intervals for the estimators (2) and (4) by calculating the 2.5th and 97.5th percentiles for each time point in `timegrid`. You may find the following lines of code helpful:

```
Surv_difference_matrix = Surv_difference_matrix_ipw
=matrix(0,nrow = m, ncol=length(timegrid))
```

Then store the j-th realisation of `Surv_difference` in the j-th row of `Surv_difference_matrix`

etc. You can then use the commands `Qfun = function(cl)quantile(cl,prob=c(0.025,0.975))`
`quantile_S = apply(Surv_difference_matrix,2,Qfun)`
`quantile_S_ipw = apply(Surv_difference_matrix_ipw,2,Qfun)` to obtain the quantiles, i.e. the approximate confidence intervals.

(d) Set $m = 500$ and compare the estimates of $P(T^{x=1} > t) - P(T^{x=0} > t)$ obtained by (2) and (4) with their approximate 95% confidence intervals. Why are they so similar in this case?

(e) Set `X_prob = 1 -(0.95/L)` instead of `X_prob = 1 - (0.90/L)` and perform the calculation in point d) again. Why is the performance of the two estimators different this time? Note that we have not made any model assumptions in our estimation procedure.

(f) Perform the same calculation as in d), but this time set
`L_prob_vec = c(0.1,0.25,0.15,0.1,0.1,0.05,0.05,0.05,0.1,0.05)`. What do you see?

2. You have access to time-to-event data that follow a Cox model, so that the hazard takes the form $\alpha(t \mid X = x) = \alpha_0(t)e^{\beta x}$ for $x \in \{0, 1\}$ for $t \in [0, \tau]$.

(a) Starting from the definition of the hazard, verify that the hazard ratio under the Cox model is

$$e^{\beta} = \frac{\lim_{h\to 0^+} \frac{1}{h} P(t \leq T < t + h \mid t \leq T, X = 1)}{\lim_{h\to 0^+} \frac{1}{h} P(t \leq T < t + h \mid t \leq T, X = 0)} \tag{5}$$

for all times $t \in [0, \tau]$.

(b) Suppose that the data were from a randomized trial. Suppose also that causal consistency holds for the failure times, i.e. $T = XT^{x=1} + (1 - X)T^{x=0}$, where T denotes the failure time, X is the treatment, and T^x is the failure time if X is set to the value x. Because the failure times come from a randomised trial, we have that $T^x \perp\!\!\!\perp X$ for $x \in \{0, 1\}$. Starting from (5), show that

$$e^{\beta} = \frac{\lim_{h\to 0^+} \frac{1}{h} P(t \leq T^{x=1} < t + h \mid t \leq T^{x=1})}{\lim_{h\to 0^+} \frac{1}{h} P(t \leq T^{x=0} < t + h \mid t \leq T^{x=0})} \tag{6}$$

for all times $t \in [0, \tau]$. State the assumptions you use in each step from (5) to (6).

(c) A friend of you struggles to formalise an interpretation of the coefficient e^{β} from (a). He comes up with the following candidates:

i) e^{β} is the marginal hazard in the population when X is set to one divided by the marginal hazard in the population when X is set to zero.

ii) e^{β} is the hazard when X takes the value one divided by the hazard when X takes the value zero.

We assume that the model $\alpha(t \mid X = x) = \alpha_0(t)e^{\beta x}$ for $x \in \{0, 1\}$ is true. Under what assumptions is the interpretation i) correct. Under what assumptions is the interpretation ii) correct? Are the assumptions needed for the interpretation ii) weaker or stronger than the assumptions needed for the interpretation i)?[2]

3. Suppose the hazard of a failure time T given the covariates X and V takes the form

$$\alpha(t \mid X, V) = \alpha_0(t) + X\alpha_X(t) + V\alpha_V(t), \tag{7}$$

where $X \in \{0, 1\}$, and V is a discrete random variable. We assume that causal consistency holds, i.e. that $T = XT^{x=1} + (1 - X)T^{x=0}$, and that $T^x \perp\!\!\!\perp X \mid V$ for $x \in \{0, 1\}$.

[2]Hint: Look at the assumptions used in (a) and (b).

(a) Show that

$$\alpha(t \mid X = x, V) = \lim_{h \to 0^+} \frac{1}{h} P(t \leq T^x < t + h \mid t \leq T^x, V). \tag{8}$$

Let $h[f(T^{x=0}, T^{x=1})](t)$ be the hazard difference

$$\lim_{h \to 0^+} \frac{1}{h} \left(P(t \leq T^{x=1} < t + h \mid t \leq T^{x=1}) - P(t \leq T^{x=0} < t + h \mid t \leq T^{x=0}) \right). \tag{9}$$

We will in steps (b)–(d) show that $h[f(T^{x=0}, T^{x=1})](t)$ is collapsible on V if the model (7) holds.

(b) Use the identity (8) to show that

$$h[f(T^{x=0}, T^{x=1} \mid V = v)](t) = \alpha_X(t). \tag{10}$$

(c) Use laws of probability to show that (9) is equal to

$$\alpha_X(t) + \alpha_V(t) \left(E[V \mid t \leq T^{x=1}] - E[V \mid t \leq T^{x=0}] \right). \tag{11}$$

(d) Use laws of probability and the model assumption (7) to show that

$$E[V \mid t \leq T^x] = \frac{\sum_v v e^{-\int_0^t v \alpha_V(s) ds} P(V = v)}{\sum_{v'} e^{-\int_0^t v' \alpha_V(s) ds} P(V = v')}. \tag{12}$$

Conclude that $E[V \mid t \leq T^x]$ is constant as a function of x, and combine this fact with the result from c) to infer that (9) is equal to $\alpha_X(t)$. Conclude that $h[f(T^{x=0}, T^{x=1})](t)$ is collapsible on V for any choice of weights $\{w_v(t)\}_v$.

4. [3]

In R, write the following commands:

```
library(boot)
library(survival)
set.seed(42)
```

The data set `melanoma` is automatically loaded.

We will compute Kaplan-Meier estimates for the all-cause mortality (death from melanoma and death from other causes). If we consider all patients, this may be performed by the command:

```
fit.mel0 = survfit(Surv(time, status %in% c(1,3)) ~ 1, data = melanoma, conf.type = "plain")
```

We can plot the Kaplan-Meier estimate with standard confidence limits by:

```
plot(fit.mel0, mark.time = FALSE, xlab = "Days after operation")
```

The following command gives a summary of the results:

```
summary(fit.mel0)
```

a) Make Kaplan-Meier plots for the total population and interpret the plots. Determine the lower quartile with 95% confidence limits using the output from the summary-command. (Note that the lower quartile corresponds to 75% survival probability.)

[3] Inspired by "Practical exercise 2" from "STK4080 - Forløpsanalyse", autumn 2014.

b) Make Kaplan-Meier plots for patients with ulceration present and absent and interpret the results. Is it possible to estimate the lower quartile for both ulceration groups? Estimate the lower quartile with confidence limits if possible.

c) Can we interpret the difference between the Kaplan-Meier curves in b) as the causal effect of an ulceration being present versus absent?

d) Use the function `coxph` to fit a Cox model with ulceration as the only covariate. Print and interpret the hazard ratio. Let $X = 1$ denote presence of ulceration and $X = 0$ denote absence of ulceration. Plot $\log(-\log(\hat{S}(t \mid X = 1)))$ and $\log(-\log(\hat{S}(t \mid X = 0)))$ to visually inspect whether the proportional hazards assumption is reasonable.

e) Make Kaplan-Meier plots for the three thickness groups 0–1 mm, 2–5 mm, 5+ mm and interpret the plots. Estimate the lower quartile with confidence limits if possible.