P8108 Homework 6

Ryan Wei, rw2844

2022-10-17

Remaining Questions in Homework 5

Problem 1

- A study design has been discussed in a study team for treating B-cell lymphoma in the second line patient population. The study will randomize subjects in 1:1 ratio

- After thorough literature search, the study team would like to assume 20 months for the median survival time in the standard of care. The expected median survival time in the new treatment arm is 28 months

  - Hazard rate
    * $h_0 = \frac{\log 2}{20} = 0.03465736$
    * $h_1 = \frac{\log 2}{28} = 0.02475526$
  - Hazard ratio $\lambda = 20/28 = 0.7142857, \log \lambda = -0.3364722$

If we assume a 90% power ($\beta = 0.10$) at the 1-sided significant level of 0.025 ($\alpha = 0.025$), the total number of events needed is (assuming equal events in both groups):

$$d = 4 \left(\frac{z_{1-\alpha} - z_\beta}{\log \lambda} \right)^2 = 4 \left(\frac{1.96 + 1.28}{-0.336} \right)^2 = 370.8955 \approx 371$$

Since the study will randomize subjects in 1:1 ratio, let n_0, d_0, n_1, d_1 denotes the number of events and subjects we needed in each group respectively, we have $n_0 = n_1 = N/2$, where N is the total sample size we needed. We also have $n_0 = d_0/\Pr_0, n_1 = d_1/\Pr_1$, from the power formula:

$$z_\beta = z_{1-\alpha} - \log \lambda / \sqrt{(\frac{1}{d_0} + \frac{1}{d_1})}$$

we can get:

$$\left(\frac{z_{1-\alpha} - z_\beta}{\log \lambda} \right)^2 = \frac{N}{2} \frac{\Pr_0 \Pr_1}{(\Pr_0 + \Pr_1)}$$

then we can solve the above equation for N:

$$N = \frac{2(\Pr_0 + \Pr_1)}{\Pr_0 \Pr_1} \left(\frac{z_{1-\alpha} - z_\beta}{\log \lambda} \right)^2$$

Since the enrollment period (τ_a) is 18 months and the minimum follow-up time (τ) for each subject is 24 months, assuming there are no censoring, we have:

$$\mathrm{Pr}_0 = \left(1 - \frac{e^{-h_0\tau}\left(1 - e^{-h_0\tau_a}\right)}{h_0\tau_a}\right)$$

$$= \left(1 - \frac{e^{-(0.03465736)\times 24}\left(1 - e^{-(0.03465736)\times 18}\right)}{(0.03465736)\times 18}\right)$$

$$= 0.6761678$$

$$\mathrm{Pr}_1 = \left(1 - \frac{e^{-h_1\tau}\left(1 - e^{-h_1\tau_a}\right)}{h_1\tau_a}\right)$$

$$= \left(1 - \frac{e^{-(0.02475526)\times 24}\left(1 - e^{-(0.02475526)\times 18}\right)}{(0.02475526)\times 18}\right)$$

$$= 0.5545472$$

Then we can get:

$$N = \frac{2(\mathrm{Pr}_0 + \mathrm{Pr}_1)}{\mathrm{Pr}_0\,\mathrm{Pr}_1}\left(\frac{z_{1-\alpha} - z_\beta}{\log\lambda}\right)^2$$

$$= 609.2454$$

Therefore, the subjects we needed for each group is $n_0 = n_1 = N/2 = 609.2454/2 = 304.6227 \approx 305$. The total subjects we needed is 610.

If more investigate sites are available and the enrollment period is shortened to 12 month, that means $\tau_a' = 12$. Then the expected event rates for the two groups become:

$$\mathrm{Pr}_0' = \left(1 - \frac{e^{-h_0\tau}\left(1 - e^{-h_0\tau_a'}\right)}{h_0\tau_a'}\right)$$

$$= \left(1 - \frac{e^{-(0.03465736)\times 24}\left(1 - e^{-(0.03465736)\times 12}\right)}{(0.03465736)\times 12}\right)$$

$$= 0.6438931$$

$$\mathrm{Pr}_1' = \left(1 - \frac{e^{-h_1\tau}\left(1 - e^{-h_1\tau_a'}\right)}{h_1\tau_a'}\right)$$

$$= \left(1 - \frac{e^{-(0.02475526)\times 24}\left(1 - e^{-(0.02475526)\times 12}\right)}{(0.02475526)\times 12}\right)$$

$$= 0.5224008$$

Then we can get:

$$N' = \frac{2(\mathrm{Pr}_0' + \mathrm{Pr}_1')}{\mathrm{Pr}_0'\,\mathrm{Pr}_1'}\left(\frac{z_{1-\alpha} - z_\beta}{\log\lambda}\right)^2$$

$$= 643.6031$$

Therefore, the subjects we needed for each group is $n_0' = n_1' = N'/2 = 643.6031/2 = 321.8015 \approx 322$. The total subjects we needed is 644.

If the hazard ratio is 1 during the first 4 months of treatment, we will lose some power since the event rate are the same in the first 4 months. However, since the median survival time for the control group and treatment group are 20 and 28 months respectively, the event rate of the first 4 months would be quiet small

such that there would not be many events occur. The hazard ratio 1 for the first 4 months may not cause significant power loss. If the investigators want to recover the power, they could increase the sample size, they could also over power the test with a power 95% instead of 90% to overcome the loss of power.

Now we make further assumptions on the censoring.

The total number of events needed is still (assuming equal events in both groups):

$$d = 4 \left(\frac{z_{1-\alpha} - z_\beta}{\log \lambda} \right)^2 = 4 \left(\frac{1.96 + 1.28}{-0.336} \right)^2 = 370.8955 \approx 371$$

Assume that independent censoring follows $C_0 \sim \exp(h_{C_0}), C_1 \sim \exp(h_{C_1})$ and $h_{C_0} = h_{C_1} = 0.002$. Since the enrollment period (τ_a) is 18 months and the minimum follow-up time (τ) for each subject is 24 months, we have:

$$
\begin{aligned}
\Pr_{0c} &= \left(1 - \frac{e^{-(h_0 + h_{C_0})\tau} \left(1 - e^{-(h_0 + h_{C_0})\tau_a} \right)}{(h_0 + h_{C_0})\tau_a} \right) \frac{h_0}{h_0 + h_{C_0}} \\
&= \left(1 - \frac{e^{-(0.03465736 + 0.002) \times 24} \left(1 - e^{-(0.03465736 + 0.002) \times 18} \right)}{(0.03465736 + 0.002) \times 18} \right) \frac{0.03465736}{0.03465736 + 0.002} \\
&= 0.6582824 \\
\Pr_{1c} &= \left(1 - \frac{e^{-(h_1 + h_{C_1})\tau} \left(1 - e^{-(h_1 + h_{C_1})\tau_a} \right)}{(h_1 + h_{C_1})\tau_a} \right) \frac{h_1}{h_1 + h_{C_1}} \\
&= \left(1 - \frac{e^{-(0.02475526 + 0.002) \times 24} \left(1 - e^{-(0.02475526 + 0.002) \times 18} \right)}{(0.02475526 + 0.002) \times 18} \right) \frac{0.02475526}{0.02475526 + 0.002} \\
&= 0.5388828
\end{aligned}
$$

Then we can get:

$$
\begin{aligned}
N_c &= \frac{2(\Pr_{0c} + \Pr_{1c})}{\Pr_{0c} \Pr_{1c}} \left(\frac{z_{1-\alpha} - z_\beta}{\log \lambda} \right)^2 \\
&= 626.434
\end{aligned}
$$

Therefore, the subjects we needed for each group is $n_{0c} = n_{1c} = N_c/2 = 626.434/2 = 313.217 \approx 314$. The total subjects we needed is 628.

If more investigate sites are available and the enrollment period is shortened to 12 month, that means $\tau'_a = 12$. Then the expected event rates for the two groups become:

$$
\begin{aligned}
\Pr'_{0c} &= \left(1 - \frac{e^{-(h_0 + h_{C_0})\tau} \left(1 - e^{-(h_0 + h_{C_0})\tau'_a} \right)}{(h_0 + h_{C_0})\tau'_a} \right) \frac{h_0}{h_0 + h_{C_0}} \\
&= \left(1 - \frac{e^{-(0.03465736 + 0.002) \times 24} \left(1 - e^{-(0.03465736 + 0.002) \times 12} \right)}{(0.03465736 + 0.002) \times 12} \right) \frac{0.03465736}{0.03465736 + 0.002} \\
&= 0.6280987 \\
\Pr'_{1c} &= \left(1 - \frac{e^{-(h_1 + h_{C_1})\tau} \left(1 - e^{-(h_1 + h_{C_1})\tau'_a} \right)}{(h_1 + h_{C_1})\tau'_a} \right) \frac{h_1}{h_1 + h_{C_1}} \\
&= \left(1 - \frac{e^{-(0.02475526 + 0.002) \times 24} \left(1 - e^{-(0.02475526 + 0.002) \times 12} \right)}{(0.02475526 + 0.002) \times 12} \right) \frac{0.02475526}{0.02475526 + 0.002} \\
&= 0.5088277
\end{aligned}
$$

Then we can get:

$$N'_c = \frac{2(\mathrm{Pr}'_{0c} + \mathrm{Pr}'_{1c})}{\mathrm{Pr}'_{0c} \mathrm{Pr}'_{1c}} \left(\frac{z_{1-\alpha} - z_\beta}{\log \lambda} \right)^2$$
$$= 660.3306$$

Therefore, the subjects we needed for each group is $n'_{0c} = n'_{1c} = N'_c/2 = 660.3306/2 = 330.1653 \approx 331$. The total subjects we needed is 662.

The discussion of power loss is same as before.

Problem 2

The study team learned from clinicaltrial.gov that a competitor's trial for the same indication requires only 300 subjects and will take only 3 years to complete. Discuss what you think of the competitor's trial design.

From our calculation in Problem 1, even if without censoring, in order to complete the study with 90% power and at a 1-sided significant level 0.025, assuming our hazard ratio is 0.714 and we have a balanced design, the sample size we needed is 610 and 644 respectively for a total 3.5 years and 3 years study, which are far greater than the 300 in the competitor's trial.

One possible explanation for this significant difference in sample sizes is that the competitor assume their drug or treatment has a larger treatment effect, hence they expect a greater reduction in hazard rate. As a result, from the sample size calculation formula (assuming 1:1 randomization ratio):

$$N = \frac{2(\mathrm{Pr}_0 + \mathrm{Pr}_1)}{\mathrm{Pr}_0 \mathrm{Pr}_1} \left(\frac{z_{1-\alpha} - z_\beta}{\log \lambda} \right)^2$$

if the reduction in hazard is large, then the hazard ratio λ will be closer to zero. Therefore, the $(\log \lambda)^2$ will be larger and thus we get a smaller sample size. For example, if the log hazard ratio of the competitor's drug is $\sqrt{2}$ times than ours, then the sample size will roughly reduced by $\frac{1}{2}$, assuming everything else are the same.

In addition, the shorter accrual period may also have some effect in the reduction in sample size, but this effect will not be that significant compared with the effect of larger treatment effect.

Homework 6

Problem 1

A trial is designed to evaluate a biomarker reduction post-treatment from baseline. The biomarker is a continuous variable. The reduction in the geometric mean is 33% with 95% CI (15%,47%) reported in literature. The team would like to know the sample size with 30% and 50% risk reduction. The statistician in the team suggests to calculate sample size for 80% and 90% power at the 2-sided significance levels of 0.05 and 0.1. Please show steps for the sample size calculation.

Assuming equal randomization ratio, which means $n_0 = n_1 = \frac{n}{2}$.

Let Y_{pi} denotes the post-treatment biomarker measure and $Y_{bi}-$ denotes the pre-treatment biomarker measure, $i = 1, \ldots, n$.

$$\frac{1}{n}\sum_{i=1}^{n}\{\log Y_{pi} - \log Y_{bi}\} = \log \sqrt[n]{\prod_{i=1 \text{ to } n}\frac{Y_{pi}}{Y_{bi}}}$$

$$= \log \frac{\bar{Y}_p}{\bar{Y}_b}$$

The reduction $1 - \frac{\bar{Y}_p}{\bar{Y}_b} = 33\%$, $\frac{\bar{Y}_p}{\bar{Y}_b} = 67\%$ with 95% CI $= (53\%, 85\%)$. Take log transformation for the ratio and CI, $\log \bar{Y}_p - \log \bar{Y}_b = \log 0.67$ with 95% CI $= (\log 0.53, \log 0.85) = (-0.6348783, -0.1625189)$. Then we can get the standard error of $\log \bar{Y}_p - \log \bar{Y}_b$ is $se(\log \bar{Y}_p - \log \bar{Y}_b) \approx 0.12$. Assuming the sample size for this literature report is 18, then the standard deviation for this log-scaled reduction is $sd = 0.12/\sqrt{18} \approx 0.51$, then $sd^2 = 0.26$.

Since we want to know the sample size with 30% and 50% reduction, that means the effect size should be $\delta_1 = \log(1 - 30\%)$ and $\delta_2 = \log(1 - 50\%)$ respectively. From the sample size calculation formula for 2 arms randomized trial with 1:1 randomization ratio, let n denotes the total sample size needed:

$$n = \frac{2(z_{1-\alpha/2} - z_\beta)^2 sd^2}{\delta^2}$$

where α is the significance level and $1 - \beta$ is the statistical power we want to achieve. In our case, $\alpha_1 = 0.05, \alpha_2 = 0.1; 1 - \beta_1 = 0.8, 1 - \beta_2 = 0.9$.

```
samplesize_tab = expand.grid(c(0.05, 0.1),c(1-0.8, 1-0.9), c(log(1-0.3), log(1-0.5)))
colnames(samplesize_tab) = c("alpha","beta","delta")
samplesize_tab = samplesize_tab[order(samplesize_tab$alpha),]
samplesize_tab$z_1_alpha_2 = qnorm(1-(samplesize_tab$alpha/2))
samplesize_tab$z_beta = qnorm((samplesize_tab$beta))
samplesize_tab$risk_reduction =1- exp(samplesize_tab$delta)
samplesize_tab %>%
  mutate(n = ceiling((2*(z_1_alpha_2-z_beta)^2*0.26)/(delta^2))) %>%
  select("risk_reduction","delta",everything()) %>%
  kable(digits = 2, booktab = T, escape = F, col.names = c("Risk Reduction", "$\\delta$", "$\\alpha$", 
```

Table 1: Sample Size Calculation for Different Study Design Parameters

	Risk Reduction	δ	α	β	$z_{1-\alpha/2}$	z_β	n
1	0.3	-0.36	0.05	0.2	1.96	-0.84	33
3	0.3	-0.36	0.05	0.1	1.96	-1.28	43
5	0.5	-0.69	0.05	0.2	1.96	-0.84	9
7	0.5	-0.69	0.05	0.1	1.96	-1.28	12
2	0.3	-0.36	0.10	0.2	1.64	-0.84	26
4	0.3	-0.36	0.10	0.1	1.64	-1.28	36
6	0.5	-0.69	0.10	0.2	1.64	-0.84	7
8	0.5	-0.69	0.10	0.1	1.64	-1.28	10

Problem 2

A randomized, controlled, and three-arm phase II study for Alzheimer's disease is designed using 2 active doses and a placebo control. The randomization ratio is 1:1:1. The primary endpoint will be clinical dementia rating scale sum of boxes (CDR-SB). Two recent phase III studies for Aducanumab showed a mean change from baseline to be 1.74 with standard error of 0.11 in the placebo arm and 548 subjects in Study Emerge 302; and a mean change from baseline of 1.56 with standard error also 0.11 calculated from 545 subjects. Please choose the significance level and power and calculate the sample for each arm and the total sample size. Please explain your rationale of the sample size calculation, including the assumptions used and if multiplicity adjustment should be considered.

The mean change from baseline is 1.74 with standard error 0.11 and sample size 548 in one active dose, the mean change from baseline is 1.56 with standard error 0.11 and sample size 545 in another active dose. Therefore, the effect sizes for the two doses are $\delta_1 = 1.74$ and $\delta_2 = 1.56$ respectively. The standard error of two experiments are $sd_1 = 2.58$ and $sd_2 = 2.57$ respectively.

If we assume one-sided significance level $\alpha = 0.05$ and power $1 - \beta = 0.9$. Since there are multiple comparison issue, we need adjustment for multiplicity. Here I use Bonferroni adjustment, in particular, for each test, we assume a one-sided significance level $\alpha' = \alpha/2 = 0.025$.

For dose one:

$$\begin{aligned} n_1 &= \frac{2(z_{1-\alpha/2} - z_\beta)^2 sd_1^2}{\delta_1^2} \\ &= \frac{2 \times (1.96 + 1.28) \times (2.58)^2}{1.74^2} \\ &= 47 \end{aligned}$$

For dose two:

$$\begin{aligned} n_1 &= \frac{2(z_{1-\alpha/2} - z_\beta)^2 sd_1^2}{\delta_1^2} \\ &= \frac{2 \times (1.96 + 1.28) \times (2.57)^2}{1.56^2} \\ &= 57 \end{aligned}$$

The sample size for two comparisons are 47 and 57, respectively. To achieve the power in both tests, we choose the larger sample size, 57, to be the sample size needed per group. Therefore, the total sample size needed is 171.

Problem 3

Construct 95% CI for the hazard ratio from a PH model shown in the follow table for the risk reduction between two treatment groups

From the table, we know that the parameter estimate is -1.62822 with standard error 0.43313. Therefore, the 95% CI for the parameter estimate is calculated by:

$$\left(\hat{\beta} - 1.96 \times se(\hat{\beta}), \hat{\beta} + 1.96 \times se(\hat{\beta})\right)$$
$$\Rightarrow (-2.4771548, -0.7792852)$$

Therefore, the 95% CI for hazard ratio is calculated by:

$$\left(\exp(\hat{\beta} - 1.96 \times se(\hat{\beta})), \exp(\hat{\beta} + 1.96 \times se(\hat{\beta}))\right)$$
$$\Rightarrow (\exp(-2.4771548), \exp(-0.7792852))$$
$$\Rightarrow (0.08398183, 0.45873380)$$

Problem 4

There are some non-randomized subjects in the `pbc` dataset. Therefore, I need to remove the non-randomized data. In this model, I only consider censored and dead endpoints.

I transformed the `edema` into categorical variable. Therefore, the model I am going to fit is:

$$h(t, Z) = h_0(t) \exp(\beta_1 \text{sex} + \beta_2 \text{edema0.5} + \beta_2' \text{edema1} + \beta_3 \text{bili} + \beta_4 \text{albumin} + \beta_5 \text{copper} + \beta_6 \text{stage})$$
$$= h_0(t) \exp(\beta_1 z_1 + \beta_2 z_2 + \beta_2' z_2' + \beta_3 z_3 + \beta_4 z_4 + \beta_5 z_5 + \beta_6 z_6)$$

Table 2: Cox Proportional Hazard Model Summary

	$\hat{\beta}$	HR	$se(\hat{\beta})$	z	$\Pr(>\lvert z \rvert)$	95% CI of HR lower bound	95% CI of HR upper bound
sexf	-0.5385	0.5836	0.2538	-2.1216	0.0339	0.3549	0.9598
edema0.5	0.1922	1.2119	0.2969	0.6474	0.5174	0.6773	2.1686
edema1	1.0496	2.8566	0.3079	3.4088	0.0007	1.5622	5.2233
bili	0.1167	1.1238	0.0174	6.6929	0.0000	1.0860	1.1629
albumin	-0.9608	0.3826	0.2530	-3.7976	0.0001	0.2330	0.6282
copper	0.0028	1.0029	0.0009	3.0684	0.0022	1.0010	1.0047
stage	0.5013	1.6508	0.1329	3.7721	0.0002	1.2723	2.1420

a) identify the information matrix

The information matrix of the model is:

```
solve(vcov(pbc_cox))
```

```
##               sexf    edema0.5     edema1        bili   albumin
## sexf     17.8628832   2.282649   0.604611   44.470298 -1.582558
## edema0.5  2.2826489  14.337896  -2.910413   70.222110 -3.078008
## edema1    0.6046110  -2.910413  13.253027   41.359721 -4.691793
```

```
## bili          44.4702985  70.222110  41.359721  4229.203897  -48.070689
## albumin       -1.5825575  -3.078008  -4.691793   -48.070689   20.056573
## copper     -1276.3211502 -91.462439 657.851088 19184.094045 -563.824948
## stage          0.6373017   1.803368   3.662610     7.914929  -11.704603
##                   copper       stage
## sexf          -1276.32115   0.6373017
## edema0.5        -91.46244   1.8033685
## edema1          657.85109   3.6626103
## bili          19184.09404   7.9149294
## albumin        -563.82495 -11.7046025
## copper      1420084.38169 1365.7937460
## stage          1365.79375  64.7050276
```

b) Construct likelihood ratio test for hypothesis $H_0 : \beta_4 = \beta_5 = \beta_6 = 0$

```
# fit sub model and do likelihood ratio test
pbc_cox_sub <- coxph(Surv(time, status==2) ~ sex + edema + bili, data =  pbc_data)
# get loglik
pbc_full_loglik = pbc_cox$loglik[2]
pbc_sub_loglik = pbc_cox_sub$loglik[2]

lrt_stats = 2*(pbc_full_loglik-pbc_sub_loglik)
anova(pbc_cox_sub, pbc_cox)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time, status == 2)
##  Model 1: ~ sex + edema + bili
##  Model 2: ~ sex + edema + bili + albumin + copper + stage
##    loglik  Chisq Df Pr(>|Chi|)
## 1 -573.80
## 2 -545.94 55.716  3  4.829e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#qchisq(0.975,3)
```

Let the full covariates be $Z = (z_1, z_2, z_2', z_3, z_4, z_5, z_6)$, and the subset of covariates $Z' = (z_1, z_2, z_2', z_3)$.

From the fitted full model and sub-model, we can get the log likelihood of the full model $l_{\mathrm{full}}(\hat{\beta}) = -545.94$ and the log likelihood of the sub-model $l_{\mathrm{sub}}(\hat{\beta}) = -573.80$.

The likelihood ratio test statistic is:

$$-2 \log \Lambda = -2 \left\{ l_{\mathrm{sub}}(\hat{\beta}) - l_{\mathrm{full}}(\hat{\beta}) \right\} = 48.799 \sim \chi^2(7 - 4)$$

Assuming a one-sided significance level $\alpha = 0.025$, we get the p-value of the test is 1.44e-10. Therefore, we reject the null hypothesis and conclude that the model with all variabless offers a significant improvement in fit over the model with four variables.

c) Compare the p-values for sex between Model (1) and the log-rank test

```
survdiff(Surv(time, status==2) ~ sex, data = pbc_data, rho = 0)
```

```
## Call:
## survdiff(formula = Surv(time, status == 2) ~ sex, data = pbc_data,
##     rho = 0)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=m  36       22     14.6     3.799      4.35
## sex=f 274      102    109.4     0.506      4.35
##
##  Chisq= 4.4  on 1 degrees of freedom, p= 0.04
```

As shown in the regression summary, the p-value for sex in the full model is 0.0339. In the log rank test, the p-value is 0.0369. Since the Cox PH model adjust several potential confounders in the model, as a result, it gives us a more accurate estimate of the effect of the covariates on the survival. However, the log-rank test only works for mutually exclusive categories and does not handle a continuous exposure variable. Therefore, I would recommend to rely on the Wald test from the Cox PH model.

d) Suppose that the team plans to publish a paper about the study result. You need to describe the statistical methods in the method section of the paper. Please write all needed analyses in the method section.

- The total number of subjects under randomization is 312, however, there are 2 subject with missing data in `copper` so we drop those two subject's data.

- We considered the transplant patients as survived subjects at the end of the study.

- The main tool used was the Cox proportional-hazards regression model. In this model, each patient is given a risk score:

$$h(t, Z) = h_0(t) \exp(\beta_1 \text{sex} + \beta_2 \text{edema0.5} + \beta_2' \text{edema1} + \beta_3 \text{bili} + \beta_4 \text{albumin} + \beta_5 \text{copper} + \beta_6 \text{stage})$$
$$= h_0(t) \exp(\beta_1 z_1 + \beta_2 z_2 + \beta_2' z_2' + \beta_3 z_3 + \beta_4 z_4 + \beta_5 z_5 + \beta_6 z_6)$$

- Each coefficient β_i has the interpretation that every unit increase in the ith covariate, z_i, increases the risk of dying by the multiplicative factor $\exp(\beta_i)$. High-risk scores correspond to poor prognoses.

- Cumulative survival curves were constructed as time-to-event plots by Kaplan-Meier methods

- One-sample log-rank tests were used to compare individuals patients' predict survival with their observed survival.

e) You are also responsible for the result section of the potential publication. Please describe the results, including interpretation of the hazard ratios, p-values, 2-sided 95% CIs of each covariate.

For this part, I only include the result of the Cox model.

Here is the descriptive statistical analysis for this dataset,

	Censored	Transplant	Dead	Overall
	(N=167)	(N=19)	(N=124)	(N=310)
Sex				
Male	11 (6.6%)	3 (15.8%)	22 (17.7%)	36 (11.6%)
Female	156 (93.4%)	16 (84.2%)	102 (82.3%)	274 (88.4%)
Histologic stage of disease				
Stage 1	15 (9.0%)	0 (0%)	1 (0.8%)	16 (5.2%)
Stage 2	47 (28.1%)	3 (15.8%)	16 (12.9%)	66 (21.3%)
Stage 3	69 (41.3%)	8 (42.1%)	43 (34.7%)	120 (38.7%)
Stage 4	36 (21.6%)	8 (42.1%)	64 (51.6%)	108 (34.8%)
Edema status				
No edema	157 (94.0%)	17 (89.5%)	88 (71.0%)	262 (84.5%)
Untreated or successfully treated	9 (5.4%)	2 (10.5%)	17 (13.7%)	28 (9.0%)
Edema despite diuretic therapy	1 (0.6%)	0 (0%)	19 (15.3%)	20 (6.5%)
Serum albumin (g/dL)				
Mean (SD)	3.63 (0.334)	3.57 (0.377)	3.36 (0.479)	3.52 (0.421)
Median [Min, Max]	3.63 [2.75, 4.64]	3.57 [2.83, 4.20]	3.43 [1.96, 4.40]	3.56 [1.96, 4.64]
Serum bilirubin (mg/dL)				
Mean (SD)	1.51 (1.75)	3.13 (2.02)	5.67 (6.09)	3.27 (4.54)
Median [Min, Max]	0.900 [0.300, 13.0]	3.20 [0.500, 8.70]	3.25 [0.300, 28.0]	1.40 [0.300, 28.0]
Urine copper (μg/day)				
Mean (SD)	66.6 (57.1)	124 (100)	135 (98.5)	97.6 (85.6)
Median [Min, Max]	52.0 [4.00, 412]	102 [32.0, 444]	111 [13.0, 588]	73.0 [4.00, 588]

The Cox Proportional Hazard Model Results is as follow:

Table 3: Cox Proportional Hazard Model Summary

	$\hat{\beta}$	HR	$se(\hat{\beta})$	z	$Pr(>\|z\|)$	95% CI of HR lower bound	95% CI of HR upper bound
sexf	-0.5385	0.5836	0.2538	-2.1216	0.0339	0.3549	0.9598
edema0.5	0.1922	1.2119	0.2969	0.6474	0.5174	0.6773	2.1686
edema1	1.0496	2.8566	0.3079	3.4088	0.0007	1.5622	5.2233
bili	0.1167	1.1238	0.0174	6.6929	0.0000	1.0860	1.1629
albumin	-0.9608	0.3826	0.2530	-3.7976	0.0001	0.2330	0.6282
copper	0.0028	1.0029	0.0009	3.0684	0.0022	1.0010	1.0047
stage	0.5013	1.6508	0.1329	3.7721	0.0002	1.2723	2.1420

From the proportional hazard model, we see that:

- The risk of death is significantly lower in female than that in male, with HR = 0.5836, 95% confidence interval (0.3549, 0.9598), p-value 0.0339. The risk reduction of death is 0.4164 in female compared with male.

- The risk of death is significantly higher in subjects with edema despite diuretic therapy compare with subjects with no edema, with HR = 2.8566, 95% confidence interval (1.5622, 5.2233), p-value 0.0007. The risk of death is about 3 times in subjects with edema despite diuretic therapy compare with subjects with no edema.

- The risk of death is significantly higher in unit increase in serum bilirunbin, with HR = 1.1238, 95%

confidence interval $(1.0860, 1.1629)$, p-value <0.0001. The risk of death with unit increase in serum bilirunbin is about 1.12 times larger.

- The risk of death is significantly higher in unit increase in urine copper, with HR $= 1.0029$, 95% confidence interval $(1.0010, 1.0047)$, p-value 0.0022. The risk of death with unit increase in urine copper is about 0.29% increased.

- The risk of death is significantly higher in unit increase in stage, from lower stages to higher stages, with HR $= 1.6508$, 95% confidence interval $(1.2723, 2.1420)$, p-value 0.0002. The risk of death increase as the subjects stage increased from lower one to a level higher, on average, the risk increased about 65%.

Problem 5

Proof

The log-rank test statistic is:

$$\frac{L}{\sqrt{\mathrm{Var}(L)}} \sim N(0,1)$$

where L is the total deviation from the null, $L = \sum_{i=1}^{k}(d_{0i} - e_{0i})$. $\mathrm{Var(L)} = \mathrm{Var}\left(\sum_{i=1}^{k}(d_{0i} - e_{0i})\right) \approx \sum_{i=1}^{k}\mathrm{Var}(d_{0i})$.

From hypergeometric distribution, we have:

$$\mathrm{Var}(d_{0i}) = \frac{n_{0i}n_{1i}d_i(n_i - d_i)}{n_i^2(n_i - 1)}$$

$$\mathrm{Var}(L) = \sum_{i=1}^{k}\frac{n_{0i}n_{1i}d_i(n_i - d_i)}{n_i^2(n_i - 1)}$$

Consider a special case of a two-sample problem and no ties, $z_i = 0, 1$ is the group indicator for the ith subject. Under null $H_0 : \beta = 0$, the Score test statistic is:

$$\frac{U(0)}{\sqrt{I(0)}} \sim N(0,1)$$

and we have:

$$
\begin{aligned}
U(0) &= \sum_{i=1}^{n}\Delta_i\left\{z_i - \frac{\sum_{l \in R(t_i)}z_i}{\sum_{l \in R(t_i)}1}\right\} \\
&= \sum_{j=1}^{J}\left(z_jd_j - d_j\frac{n_{1j}}{n_j}\right) \\
&= \sum_{j=1}^{J}\left(d_{1j} - d_j\frac{n_{1j}}{n_j}\right) \\
&= \sum_{j=1}^{J}(d_{1j} - e_{1j})
\end{aligned}
$$

The information under null and when $d_j = 1$, assuming no ties, is:

$$I(\beta) = I(0) = \sum_{i=1}^{n} \Delta_i \left\{ \frac{\sum_{l \in R(t_i)} z_l^2}{\sum_{l \in R(t_i)} 1} - \left(\frac{\sum_{l \in R(t_i)} z_l}{\sum_{l \in R(t_i)} 1} \right)^2 \right\}$$

$$= \sum_{j=1}^{J} d_j \left\{ \frac{n_{1j}}{n_j} - \left(\frac{n_{1j}}{n_j} \right)^2 \right\}$$

$$= \sum_{j=1}^{J} \left\{ \frac{n_{1j}}{n_j} - \left(\frac{n_{1j}}{n_j} \right)^2 \right\}$$

$$= \text{Var}(L)$$

Therefore, the score statistic is:

$$\frac{U(0)}{\sqrt{I(0)}} = \frac{L}{\text{Var}(L)} \sim N(0,1)$$

$\square$

Problem 6

Thee observed survival data T Δ Z are(16,1,1),(20,0,1),(12, 1,0), (14, 0,0), (11, 1,0), (9, 1, 1). Please construct the partial likelihood.

First we ordered the data by event/censor time and group the data:

- Group 0: 11, 12, 14+

- Group 1: 9, 16, 20+

- There are 4 event time: 9, 11, 12, 16

Group 0	11,12,14+	11,12,14+	12,14+	
Group 1	9,16,20+	16,20+	16,20+	16,20+
Event Time	9	11	12	16

Therefore, assuming there is only one binary indicator z for the group allocation in the model, the partial likelihood is:

$$L_p(\beta) = \frac{e^\beta}{3e^\beta + 3} \frac{1}{2e^\beta + 3} \frac{1}{2e^\beta + 2} \frac{e^\beta}{2e^\beta}$$

Reference:

1. Grambsch, P. M., Dickson, E. R., Wiesner, R. H., & Langworthy, A. (1989). Application of the Mayo primary biliary cirrhosis survival model to Mayo liver transplant patients. Mayo Clinic proceedings, 64(6), 699–704. https://doi.org/10.1016/s0025-6196(12)65350-6

Appendix: Code for this report

```r
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE)
library(tidyverse)
library(knitr)
library(kableExtra)
library(survival)
h0 = log(2)/20
h1 = log(2)/28

hr = 20/28

hc = 0.002
expected.event.rate <- function(h,hc,t,ta){
  frac = (exp(- t*(h+hc))*(1-exp(-ta*(h+hc))))/(ta*(h+hc))
  return((1-frac)*(h/(h+hc)))
}
tau = 24
tau_a = 18
pr0 = expected.event.rate(h0,0,tau, tau_a)
pr1 = expected.event.rate(h1,0,tau, tau_a)
#N0 = 4*((qnorm(0.975)+qnorm(0.90))/log(20/28))^2/pr0
#N1 = 4*((qnorm(0.975)+qnorm(0.90))/log(20/28))^2/pr1
pr0_ = expected.event.rate(h0,0,tau,12)
pr1_ = expected.event.rate(h1,0,tau,12)
#N0_ = 4*((qnorm(0.975)+qnorm(0.90))/log(20/28))^2/pr0_
#N1_ = 4*((qnorm(0.975)+qnorm(0.90))/log(20/28))^2/pr1_
pr0_c = expected.event.rate(h0,hc,tau, tau_a)
pr1_c = expected.event.rate(h1,hc,tau, tau_a)
pr0_c_ = expected.event.rate(h0,hc,tau,12)
pr1_c_ = expected.event.rate(h1,hc,tau,12)
#N0 = 4*((qnorm(0.975)+qnorm(0.90))/log(20/28))^2/pr0
#N1 = 4*((qnorm(0.975)+qnorm(0.90))/log(20/28))^2/pr1
logyp = log(0.67)
logyp_ci = c(log(0.53), log(0.85))
samplesize_tab = expand.grid(c(0.05, 0.1),c(1-0.8, 1-0.9), c(log(1-0.3), log(1-0.5)))
colnames(samplesize_tab) = c("alpha","beta","delta")
samplesize_tab = samplesize_tab[order(samplesize_tab$alpha),]
samplesize_tab$z_1_alpha_2 = qnorm(1-(samplesize_tab$alpha/2))
samplesize_tab$z_beta = qnorm((samplesize_tab$beta))
samplesize_tab$risk_reduction =1- exp(samplesize_tab$delta)
samplesize_tab %>%
  mutate(n = ceiling((2*(z_1_alpha_2-z_beta)^2*0.26)/(delta^2))) %>%
  select("risk_reduction","delta",everything()) %>%
  kable(digits = 2, booktab = T, escape = F, col.names = c("Risk Reduction", "$\\delta$", "$\\alpha$", 
delta_1 = 1.74
n_1 = 548
sd_1 = 0.11*sqrt(n_1)
alpha_1 = 0.025
beta_1 = 0.1
N1 = ceiling(2*(qnorm(1-alpha_1)-qnorm(beta_1))^2*sd_1^2/delta_1^2)
delta_2 = 1.56
n_2 = 545
```

```r
sd_2 = 0.11*sqrt(n_2)
alpha_2 = 0.025
beta_2 = 0.1
N2 = ceiling(2*(qnorm(1-alpha_2)-qnorm(beta_2))^2*sd_2^2/delta_2^2)
hat_beta = -1.62822
hat_se = 0.43313
ci_beta = hat_beta+1.96*c(-1,1)*hat_se
ci_hazard = exp(ci_beta)
#status at endpoint, 0/1/2 for censored, transplant, dead
# remove NA from trt variable (non-randomized participants)
pbc_data <- subset(survival::pbc, !is.na(trt))


#  0 no edema, 0.5 untreated or successfully treated
pbc_data$edema <- as.factor(pbc_data$edema)

pbc_data <- pbc_data %>% dplyr::select(c(sex,edema , bili , albumin ,copper, stage,status,time)) %>% na

pbc_data$status <- as.numeric(pbc_data$status)
pbc_cox <- coxph(Surv(time, status==2) ~ sex + edema + bili + albumin + copper + stage, data =  pbc_data
pbc_cox_sum = summary(pbc_cox)
pbc_cox_detail = coxph.detail(pbc_cox)

cbind(pbc_cox_sum$coefficients,pbc_cox_sum$conf.int[,3:4]) %>% kable(col.names = c("$\\hat\\beta$", "HR
solve(vcov(pbc_cox))
# fit sub model and do likelihood ratio test
pbc_cox_sub <- coxph(Surv(time, status==2) ~ sex + edema + bili, data =  pbc_data)
# get loglik
pbc_full_loglik = pbc_cox$loglik[2]
pbc_sub_loglik = pbc_cox_sub$loglik[2]

lrt_stats = 2*(pbc_full_loglik-pbc_sub_loglik)
anova(pbc_cox_sub, pbc_cox)
#qchisq(0.975,3)
survdiff(Surv(time, status==2) ~ sex, data = pbc_data, rho = 0)
require(table1)
dat <- pbc_data
dat$sex <- factor(dat$sex, levels=c("m", "f"), labels=c("Male", "Female"))
dat$stage <- factor(dat$stage, levels=1:4, labels=paste("Stage", 1:4))
dat$edema <- factor(dat$edema, levels=c(0, 0.5, 1),
labels=c("No edema",
"Untreated or successfully treated",
"Edema despite diuretic therapy"))
dat$status <- factor(dat$status, levels = c(0,1,2), labels = c("Censored", "Transplant", "Dead"))

label(dat$sex) <- "Sex"
label(dat$stage) <- "Histologic stage of disease"
label(dat$edema) <- "Edema status"
label(dat$albumin) <- "Serum albumin (g/dL)"
label(dat$bili) <- "Serum bilirubin (mg/dL)"
label(dat$copper) <- "Urine copper ($\\mu$;g/day)"

table1(~ sex + stage + edema + albumin + bili + copper|status , data=dat)
```

```
cbind(pbc_cox_sum$coefficients,pbc_cox_sum$conf.int[,3:4]) %>% kable(col.names = c("$\\hat\\beta$", "HR
risk_tab = tibble(
  item = c("Group 0", "Group 1", "Event Time"),
  e9 = c("11,12,14+", "9,16,20+", "9"),
  e11 = c("11,12,14+", "16,20+", "11"),
  e12 = c("12,14+", "16,20+", "12"),
  e16 = c("", "16,20+", "16"),
)
risk_tab %>% kable(col.names = NULL, booktab = T) %>% kable_styling(latex_options = "HOLD_position")
```