BST 222 Homework 3

Bohao Zou (bhzou@ucdavis.edu)

November 30, 2020

Question 1.

(1)

We can use histogram to show the distribution of variable "*Time in days until relapse*" of "*patch only*" or "*combination*" group.

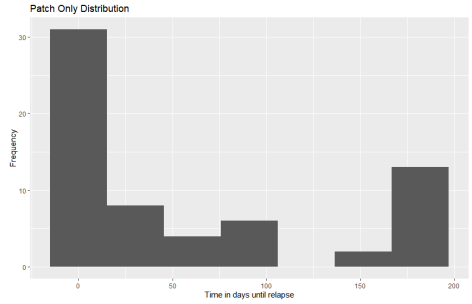

Figure 1: *The histogram of survival data of patch only group.*

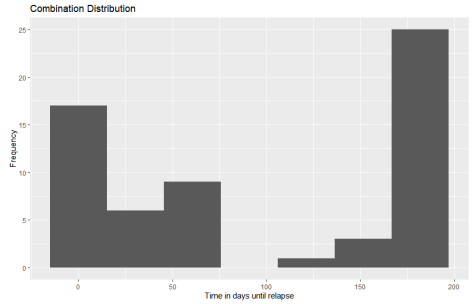

Figure 2: *The histogram of survival data of combination group.*

From those plots we can know that in the *patch only* group, most people have short time in days until relapse. However, in the *combination* group, the number of people who have short time in days until relapse are roughly equal with the number of people who have long time in days until relapse. This may indicate that combination can prolong time to relapse.

(2)

We can use test for trend to answer this question that "whether the combination therapy can prolong time to relapse?" We can compare the hazard rates of K, $K \geq 2$ populations. The hypotheses is :

$$H_0 : h_1(t) = h_2(t) \quad for \quad all \quad t \leq \tau$$
$$H_1 : h_1(t) < h_2(t) \quad for \quad all \quad t \leq \tau$$

where $h_1(t)$ and $h_2(t)$ are the hazard rates for *combination* group and *patch only* group respectively. τ is the largest of the observed relapse time.

(3)

The result of trend test in SAS is showed below:

Test	Test Statistic	Standard Error	z-Score	P-value
Log-Rank	12.9475	4.5697	2.8333	0.0046

From this table we can know, under the significant level of $\alpha = 0.05$, there is a significant difference between the combination therapy and patch only therapy and the hazard rate of combination is smaller than the hazard rate of patch only.

The survival probability vs. time plot also shows the trend that combination group has a longer relapse time.

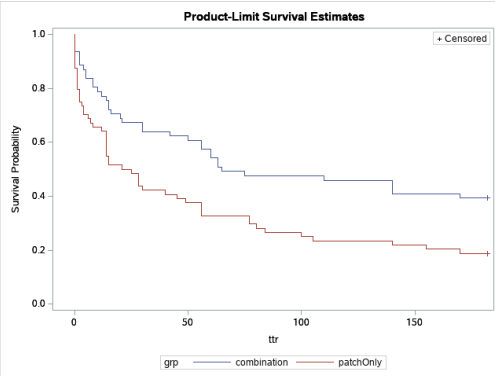

Figure 3: The survival probability curve of combination therapy and patch only therapy.

From this plot and the trend test we can give a conclusion that combination therapy can prolong time to relapse.

(4)

We can use stratified test to account for if age at randomization could affect the intervention effect. In this dataset, there are two variables which relate with age. So, we can do two separate stratified test. After bonferroni correction, if one of P-value smaller than 0.05, we can conclude that age can affect the intervention effect.

We can assume that our test is to be stratified on M levels of a set of age. The hypothesis is :

$$H_0 : h_{1s}(t) = h_{2s}(t) \quad for \quad s = 1...M, \quad t \leq \tau$$
$$H_1 : At \quad least \quad one \quad of \quad h_{js}(t) \quad is \quad different \quad for \quad some \quad t \leq \tau$$

where $h_{1s}(t)$ and $h_{2s}(t)$ are the hazard rates for *patch only* group and *combination* group for different level of age. τ is the largest of the observed relapse time.

(5)

The test result of *ageGroup2* variable shows below:

Test	Chi-Square	Df	P-value
Log-Rank	10.0707	1	0.0015

The test result of *ageGroup4* variable shows below:

Test	Chi-Square	Df	P-value
Log-Rank	11.5198	3	0.0092

After bonferroni correction, those P-value are all lower than 0.05, so we can conclude that age at randomization could affect the intervention effect.

(6)

Compare the Chi-Square of Q3 and Q5, we can find that the Chi-Squares in Q5 are all larger than Chi-Squares in Q3. This may indicate that the different of *time in days until relapse* between *patch only* and *combination* may conduct by the age at randomization.

Question 2.

(1)

The proportional hazard model is :

$$h(t|\overrightarrow{Z}) = h_0(t)exp(\overrightarrow{\beta}^T\overrightarrow{Z}) = h_0(t)exp(\sum_{k=1}^{p}\beta_k z_k)$$

where $\overrightarrow{\beta}$ is the coefficients of covarites. We need to estimate $\overrightarrow{\beta}$. $\overrightarrow{Z}$ is the covarites. $h_0(t)$ is the baseline of hazard ratio. $h(t|\overrightarrow{Z})$ is the Cox PH regression model.

$\theta = exp(\beta_i)$ is the hazard ratio for the $i-th$ covariate.

$$if \quad \theta = 1, \quad this \quad variate \quad has \quad no \quad effect$$
$$if \quad \theta > 1, \quad increasing \quad risk$$
$$if \quad \theta < 1, \quad decreasing \quad risk$$

(2)

The partial likelihood of β based on the above 4 observations is :

$$L(\overrightarrow{\beta}) = \frac{exp(4\beta)}{exp(3\beta) + exp(4\beta) + exp(5\beta) + exp(6\beta)} \times \frac{exp(3\beta)}{exp(3\beta) + exp(5\beta) + exp(6\beta)}$$

(3)

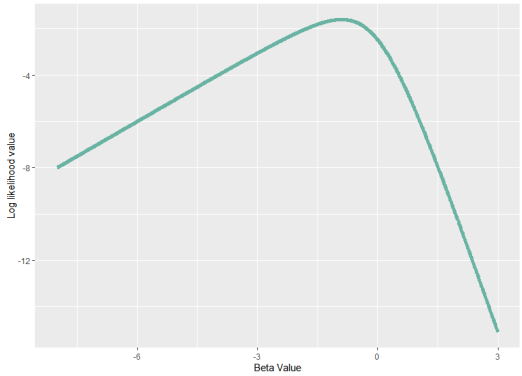

Figure 4: The Plot of the log partial likelihood for β in [-8, 3]

From this plot we can know that when $\beta = 0.9$, the log partial likelihood will get the maximum value. Obviously, it is a concave function.

(4)

From the conclusion of (3), I will assign -0.9 to β, the $\hat{\beta} = -0.9$. The value of second derivative of the log partial likelihood function at $\hat{\beta}$ is -5.010615. It is a very complex function and the main idea of it is to infer the formula of second derivative.

$$\frac{d^2LL(\beta)}{d\beta^2} = -\frac{(4e^{\beta} + 10e^{2\beta} + 18e^{3\beta}) \times (1 + e^{\beta} + e^{2\beta} + e^{3\beta}) - (e^{\beta} + 2e^{2\beta} + 3e^{3\beta}) \times (3 + 4e^{\beta} + 5e^{2\beta} + 6e3\beta)}{(1 + e^{\beta} + e^{2\beta} + e^{3\beta})^2} - \frac{(10e^{2\beta} + 18e^{3\beta}) \times (1 + e^{2\beta} + e^{3\beta}) - (2e^{2\beta} + 3e^{3\beta}) \times (3 + 5e^{2\beta} + 6e^{3\beta})}{(1 + e^{2\beta} + e^{3\beta})^2}$$

(5)

The SAS output shows below:

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
z	1	-0.90291	0.80927	1.2448	0.2646	0.405

In the SAS output, the estimated $\hat{\beta} = -0.90291$. It is very close to my result that $\hat{\beta} = -0.9$.

Question 3.

(1)

We can use histogram to show the distribution of variable "Time to death" of "Aneuploid" and "Diploid" group.

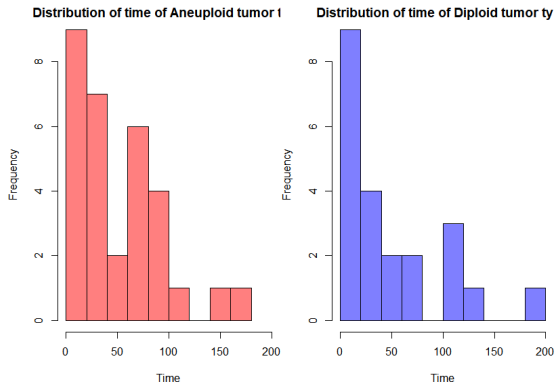

Figure 5: *The distributions of time of "Aneuploid" and "Diploid" group.*

From this plot we can know that in the group Aneuploid, most people dead before 100 and in the group Diploid, most of people dead before 150. The mean survival time of Aneuploid is 51.6129 and the mean survival time of Diploid is 47.

(2)

The SAS output is:

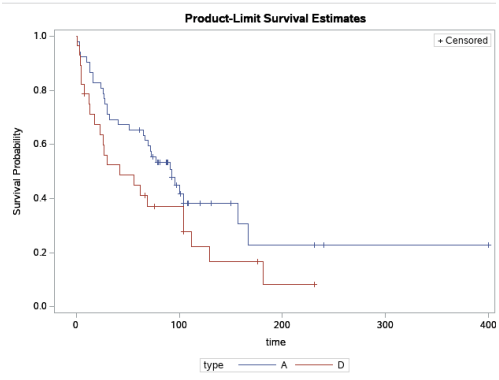

Figure 6: *The distributions of time of "Aneuploid" and "Diploid" group.*

The statistic table is :

Test	Chi-Square	Df	P-value
Log-Rank	2.7897	1	0.0949

From this table we can know that under the significant level of 0.05, we can conclude that the risk of death overtime is no significant different between the group of Aneuploid and Diploid.

(3)

The Cox PH model is

$$h(t|z) = h_0(t)exp(\beta z)$$

Z = 1, if Aneuploid group and Z = 0 if Diploid group. The SAS output is:

	DF	Parameter Estimate	Standard Error	Chi-Square	Pr> ChiSq	Hazard Ratio	Label
Aneuploid	1	-0.46104	0.28053	2.7009	0.1003	0.631	Aneuploid

The table of global testing is :

Test	Chi-Square	Df	P-value
Likelihood Ratio	2.6115	1	0.1061
Score	2.7464	1	0.0975
Wald	2.7009	1	0.1003

From those table we can know that under significant level of 0.05, the risk of death overtime is no significant different between the group of Aneuploid and Diploid.

(4)

The hazard ratio (Diploid vs. Aneuploid) is 0.631.
The median survival time derived from Kaplan-Meier survival curve of Aneuploid is 93. The median survival time derived from Kaplan-Meier survival curve of Diploid is 42. The ratio of median survival time of two groups by KM estimator is :

$$\frac{S_{Diploid}}{S_{Aneuploid}} = \frac{42}{93} = 0.4516$$

It seems that the ratio of median survival time is slightly lower than the hazard ratio calculated by Cox model. Under the exponential distribution, those two would be identical.

(5)

The score test of the Cox PH model is shown in above table, the P-value 0.0975 indicates there is no significant difference of survival time between Aneuploid and Diploid tumor type. The log-rank test also shows this conclusion. The Chi-Square statistics and P-values are very close with each other.

CODE

```
proc import datafile = "/folders/myshortcuts/MyFolder/Q2.csv" out=HWQ2;

PROC PHREG DATA= HWQ2;
model time*censor(1) = z;
run;

proc import datafile= "/folders/myshortcuts/MyFolder/Book1.csv" out = book;

proc lifetest data = book;
time time*censor(1);
strata type/test=logrank;
run;

proc phreg data=book;
class type;
model time*censor(1) = type;
run;

* KM;
proc lifetest data=book atrisk;
time time*censor(1);
strata type;
run;

proc import datafile = "/folders/myshortcuts/MyFolder/pharmacoSmoking.csv" out=PS;
proc lifetest data = PS;
time ttr*relapse(0);
strata grp/trend test=logrank;
run;
```