Seminar 3

Group A2

Yufeng Deng Yuanqing Wang

Task 1

1.1 Feature Selection & Model Development

First, we calculated the correlation coefficient between each variable. From figure 1.1 we can see that the correlation values between TC-LDL, HDL-APOA1, WBC-NEU are almost 1, which means these pairs are highly linearly related, leading to multicollinearity. We used Elastic Net to simplify the model and reduce the probability of multicollinearity.

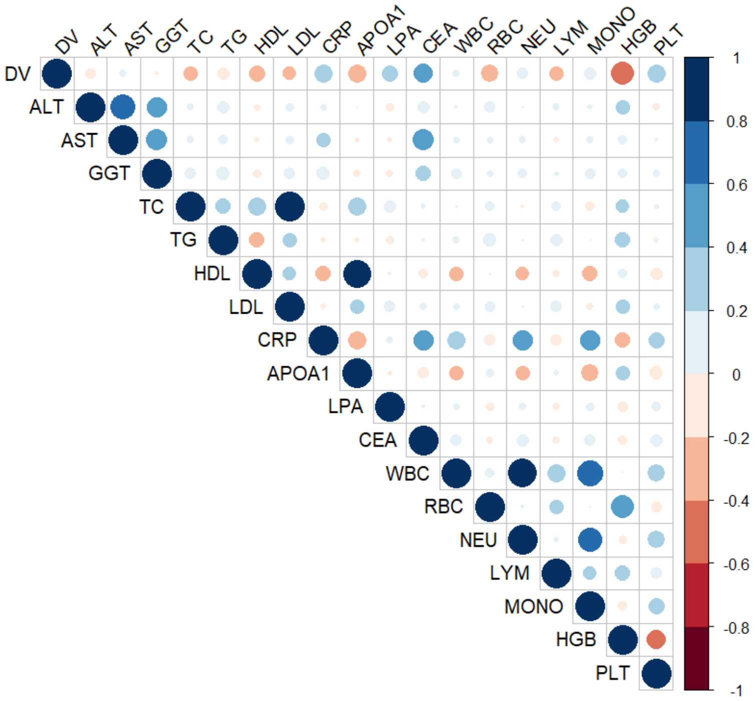

Figure 1.1 Correlation Matrix

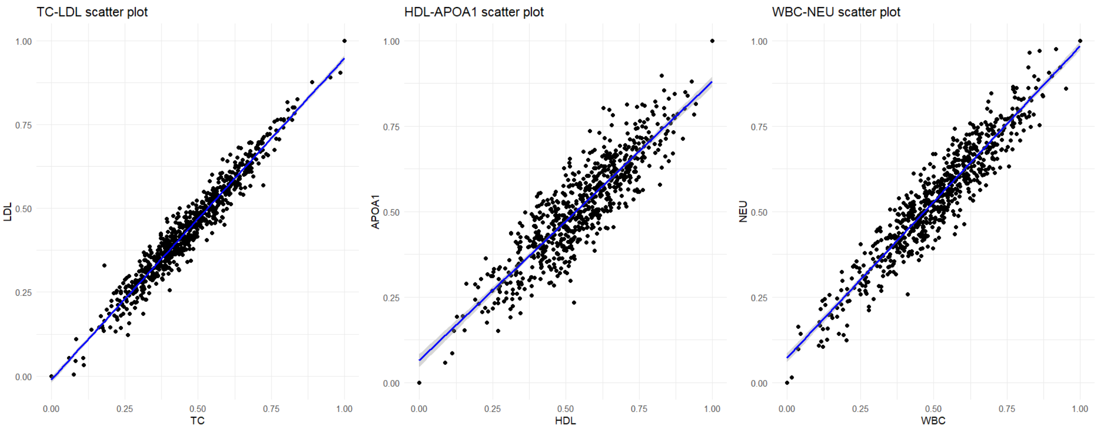

Figure 1.2 Linear Relationship

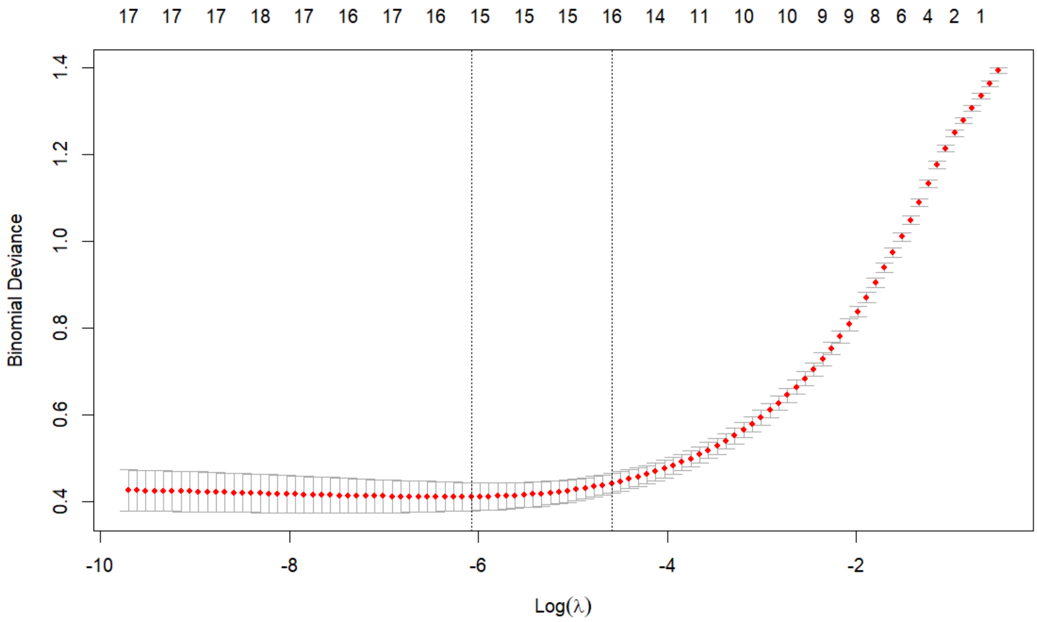

Figure 1.3 Cross Validation of Elastic Net

We used the package ‘glmnet’ to implement Elastic Net. Figure 1.3 shows how does it find the best lambda value, which is 0.002294098 in this case. After Regression, the coefficients are shown in Table 1.1. It obvious that the regression deleted the features GGT, TC and CRP, and gave other features weight coefficients. Using this result, we constructed a linear model to do the binary classification task.

Table 1.1 Coefficients of Elastic Net

Feature	Coefficient
Intercept	0.1096291
ALT	-0.2006522
AST	-0.3814531
GGT	0
TC	0
TG	-0.1771271
HDL	-1.2138212
LDL	-0.8198785
CRP	0
APOA1	-0.1409414
LPA	1.6898303
CEA	2.5336365
WBC	-0.1975546
RBC	0.1212922
NEU	-0.2468349
LYM	-0.8486259
MONO	0.2116491
HGB	-2.5728811
PLT	0.7236508

1.2 Diagnose Model Performance.

We calculated several Evaluation Metrics of the prediction of our model and compared them with the prediction of the linear model without feature selection. Table 1.2 shows the results, and we can see that with Elastic Net, all the Metrics of the linear model increased.

Table 1.2 Evaluation Metrics of the prediction (transform threshold is 0.5)

Metrics	Before Elastic Net	After Elastic Net
Accuracy	0.8832	0.9112
Precision	0.8558	0.8846
Recall	0.8990	0.9293
F1	0.8768	0.9064

1.3 Threshold Value

In diagnosis, detecting a CRC patient as a healthy person (False Negative) is more costly than detecting a healthy person as a CRC patient (False Positive). So, thresholds may need to be biased to increase the Recall rates to reduce the likelihood of False Negative. To achieve that, we can lower the threshold of the transform fromprobability to binary class. For instance, when we turn the threshold from 0.5 (default) to 0.3, the Recall rate of our model will increase from 0.9293 to 0.9596.

1.4 Clinical Implementation

In clinical implementation, doctors can check the coefficients of the model like the Table 1.1, and they can decide to manually delete a feature by considering other information, and they also can modify the value based on their knowledge, because this model is simple and easy to explain. On the other hand, if we know a certain patient's information, we can select the data from patients with similar physical conditions to increase the performance of the model. This may involve privacy risk, so the hospitals need to consider of it and develop a reasonable system to protect the data.

Task 2

1. Analyze trends in CRC stage

We checked the dataset, the number of people in each stage is equal, all are 50. The distributions of Age are different among all the stages. From Table 2.1 and Figure 2.1, we can briefly tell that the trend is people detected in later stage are older. That means when getting older, the probability of late-stage detection increases. For others, the values are shown in Table 2.2 ~ 2.4 and Figure 2.2 ~ 2.4. We applied Chi-square tests for these categorical variables. And we concluded that All the three variables have statistically significant association with the stage.

Table 2.1 Age distribution by Stage

Stage	Mean Age	SD Age	Min Age	Max Age
1	26.1	3.78	21	37
2	29.4	3.56	22	37
3	31.9	3.77	23	39
4	36.0	3.78	28	43

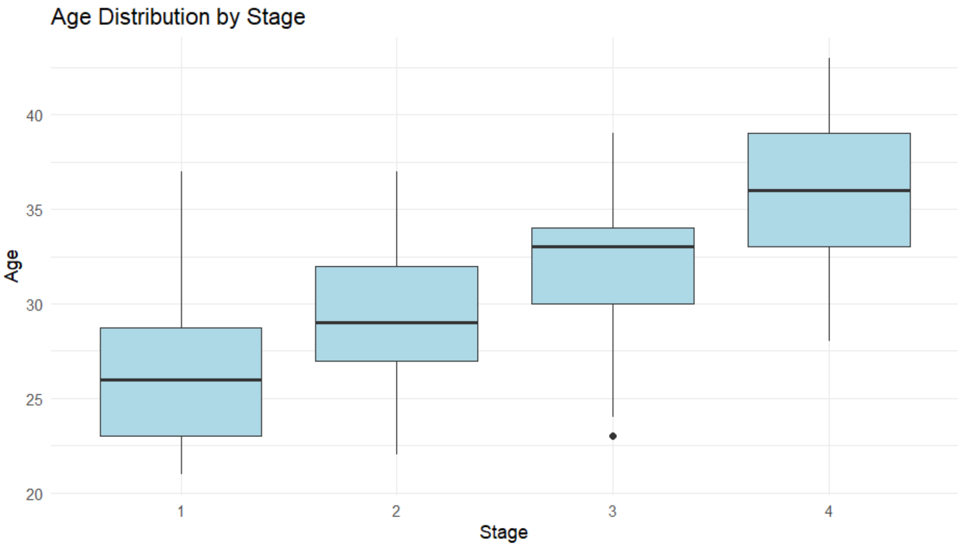

Figure 2.1 Age distribution by Stage

Regarding racial distribution, we found that the sample primarily consists of White individuals. We noted that ethnic minorities (Black and Asian populations) appear to have a higher proportion of late-stage diagnoses. Concerning the impact of lifestyle, we discovered that individuals living alone are more likely to receive late-stage diagnoses compared to those with partners. We believe this may be related to Mutual care between partners. In terms of tumor location characteristics, we observed that rectal cancer cases show a relatively high proportion of late-stage diagnoses. We found that right-sided colon cancers have a slightly better early detection rate compared to left-sided colon cancers.

Table 2.2 Stage distribution by Lifestyle

Stage	Alone	Partnered
1	20	30
2	21	29
3	27	23
4	29	21

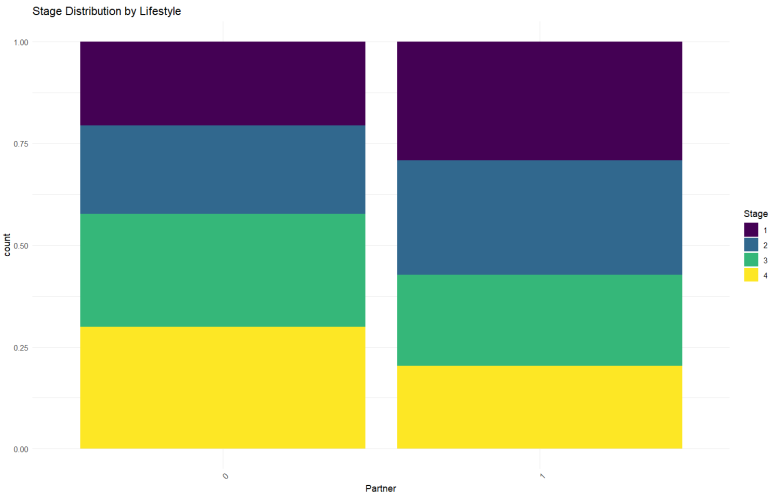

Figure 2.2 Stage distribution by Lifestyle

Table 2.3 Stage distribution by Ethnicity

Stage	Asian	Black	White
1	1	9	40
2	3	6	41
3	4	9	37
4	5	12	33

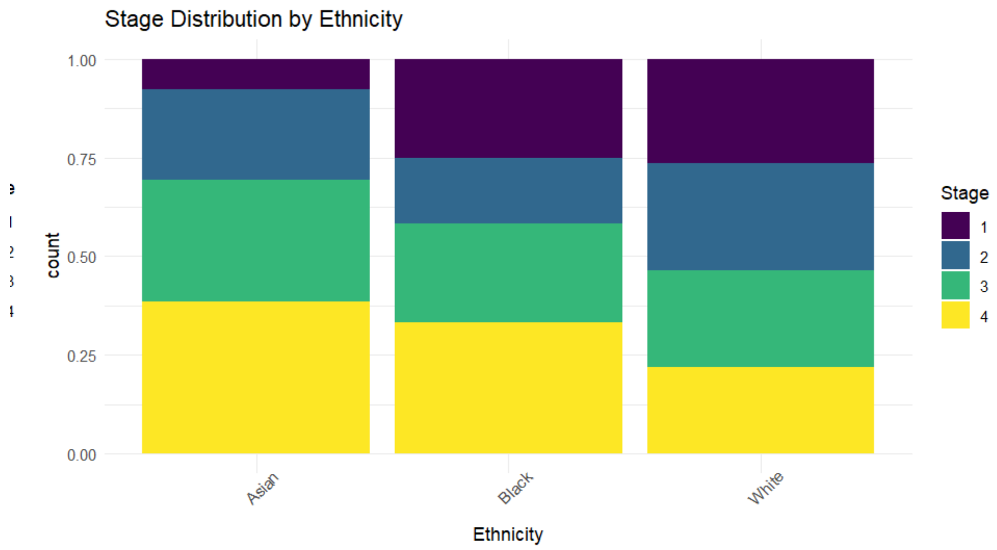

Figure 2.3 Stage distribution by Ethnicity

Table 2.4 Stage distribution by Site

Stage	Left	Rectum	Right
1	10	14	26
2	21	13	16
3	19	14	17
4	16	16	18

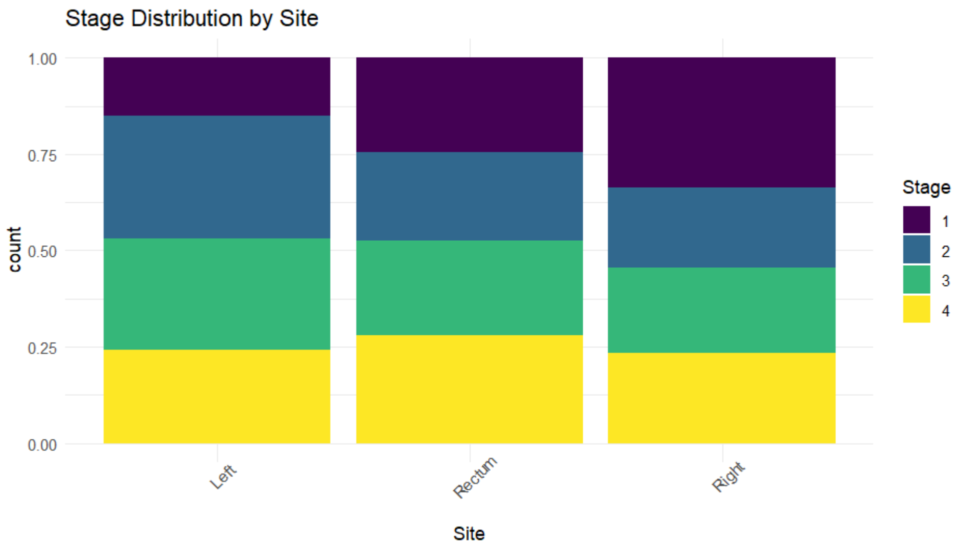

Figure 2.4 Stage distribution by Site

2. Recommendations based on the data analysis

According to the Age impact, we recommend strengthening early screening efforts for CRC and lowering the screening starting age to 20-25 years old. Besides, we strongly recommend to regularly visit young people for diagnosis, especially for those living alone. In terms of public health strategies, we advocate developing targeted health education for ethnic minorities and considering racial differences in healthcare resource allocation. For researchers, we suggest conducting in-depth research on disease mechanisms in young populations and different mechanisms among different cancer site.

Task 3

First, plot the change in log concentration over time by sex as a classification, observing its trend and linearity, as shown in the Figure3.1.

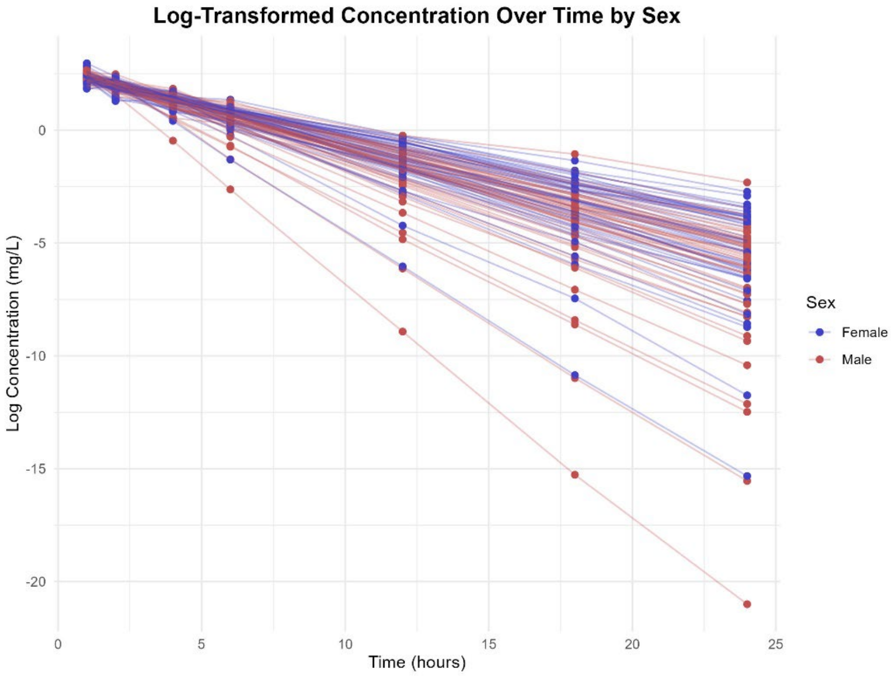

Figure3.1 Log Concentration Over Time

Through observation, it is evident that the linearity is significant. Furthermore, we use a linear mixed-effects model to analyze and evaluate the impact of various variables on the concentration-time change.

```
# Build the mixed-effects model with Time, BW, SEX, and AGE as fixed effects, and ID as
a random effect
model <- lmer(log_DV ~ Time + Bw + Sex + Age + (1 | ID), data = data)
# Display model summary to examine the significance of each fixed effect
summary(model)
```

This code constructs a linear mixed-effects model that incorporates both fixed and random effects, where Time, BW (body weight), SEX (gender), and AGE (age) are treated as fixed effects to assess their impact on log concentration (log_DV). Additionally, ID (each patient) is included as a random effect, allowing for individual variability in concentration changes and enhancing the model's fit by accounting for patient-specific deviations in log concentration trends.

```
Random effects:
 Groups   Name        Variance Std.Dev.
 ID       (Intercept) 0.7519   0.8671
 Residual             1.2377   1.1125
Number of obs: 700, groups:  ID, 100

Fixed effects:
             Estimate Std. Error t value
(Intercept)  7.986750   2.290779   3.486
Time        -0.364388   0.005189 -70.227
Bw          -0.043189   0.022434  -1.925
Sex         -0.111348   0.218560  -0.509
Age         -0.027994   0.024580  -1.139

Correlation of Fixed Effects:
     (Intr) Time   Bw     Sex
Time -0.022
Bw   -0.771  0.000
Sex   0.346  0.000 -0.460
Age  -0.618  0.000 -0.019 -0.041
```

The results indicate that time has a strong and highly significant negative effect on log concentration, suggesting that as time progresses, log concentration decreases significantly. Body weight also shows a negative relationship with log concentration, although the effect is weaker and only marginally significant. In contrast, sex and age do not exhibit significant effects on log concentration, with the estimates being small and the corresponding t-values not reaching statistical significance. The random effects analysis reveals that there is significant individual variability in baseline log concentration, as indicated by the substantial variance in the random intercept for patients. The residual variance reflects within-patient variability in log concentration changes over time. These findings highlight the importance of time and body weight in influencing log concentration, while suggesting that sex and age may not be as relevant in this context.

Task 4

For this task, we used the built-in R libraries **survival** and **survminer**.

```
library(survival)
library(survminer)
```

The survival package in R is a fundamental tool for survival analysis, providing functions for the estimation and modeling of survival data. It supports Kaplan-Meier estimation, Cox proportional hazards models, and parametric survival models (e.g., Weibull). The package is widely used for analyzing time-to-event data, particularly in clinical research and epidemiology. Key functions include Surv() for creating survival objects, survfit() for estimating survival curves, and coxph() for fitting Cox regression models.

The survminer package is an extension designed for visualizing survival analysis results. It integrates well with the survival package and offers various functions for creating high-quality survival plots, such as Kaplan-Meier curves and Cox model diagnostics. The ggsurvplot() function is a key tool for visualizing survival curves, and other functions help assess the proportional hazards assumption and generate diagnostic plots for survival models.

```
data <- read.csv("data_task4.csv")
surv_obj <- Surv(time = data$Time, event = data$Status)
surv_diff <- survdiff(surv_obj ~ Group, data = data)
print(surv_diff)
cox_model <- coxph(surv_obj ~ Group, data = data)
summary(cox_model)
fit <- survfit(surv_obj ~ Group, data = data)
ggsurvplot(fit, data = data, pval = TRUE,
           xlab = "Time (Months)", ylab = "Survival Probability",
           legend.labs = c("Control", "Treatment"),
           legend.title = "Group")
```

Then, we obtain the following results.

```
Call:
survdiff(formula = surv_obj ~ Group, data = data)

                  N Observed Expected (O-E)^2/E (O-E)^2/V
Group=Control   100       73     90.6      3.43      9.53
Group=Treatment 100       72     54.4      5.72      9.53

 Chisq= 9.5  on 1 degrees of freedom, p= 0.002
Call:
coxph(formula = surv_obj ~ Group, data = data)

  n= 200, number of events= 145

                 coef exp(coef) se(coef)     z Pr(>|z|)
GroupTreatment 0.5204    1.6828   0.1703 3.055  0.00225 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

               exp(coef) exp(-coef) lower .95 upper .95
GroupTreatment     1.683     0.5943     1.205      2.35

Concordance= 0.573  (se = 0.023 )
Likelihood ratio test= 9.26  on 1 df,   p=0.002
Wald test            = 9.34  on 1 df,   p=0.002
Score (logrank) test = 9.53  on 1 df,   p=0.002
```

The results of the survival analysis reveal a statistically significant difference in survival between the Control and Treatment groups. The log-rank test (Chi-square = 9.5, p = 0.002) indicates that the two groups have significantly different survival profiles.

Further analysis with the Cox proportional hazards model shows that the hazard ratio (HR) for the Treatment group is 1.683 (95% CI: 1.205 - 2.350), suggesting that the Treatment group has a

68.3% higher risk of experiencing the event (death) compared to the Control group. This result is statistically significant (p = 0.00225). The concordance index of 0.573 suggests that the model has moderate predictive ability.

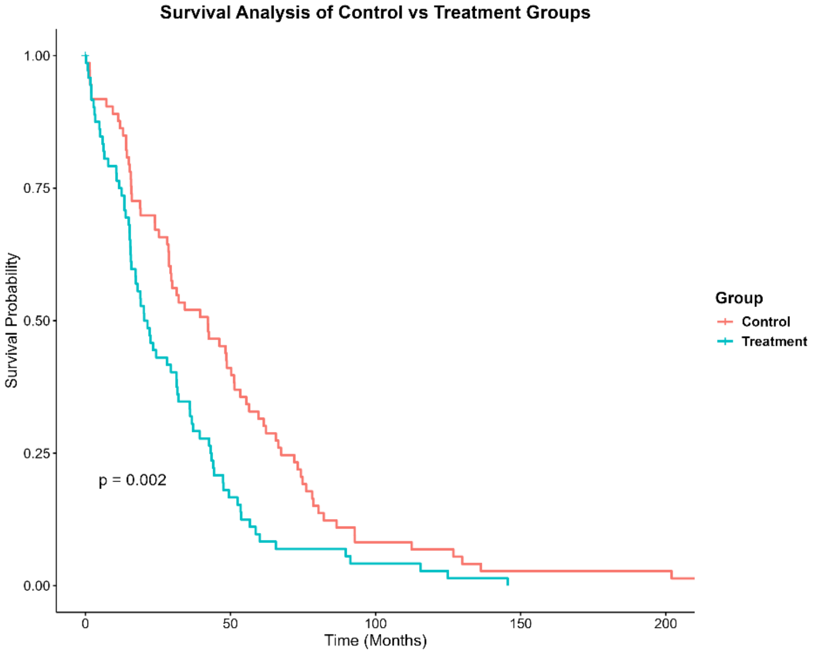

Figure 4.1 Survival Analysis of Control and Treatment Group

Despite the significant survival difference observed in the log-rank test, the Cox model indicates that the Treatment group has a higher risk of death. This suggests that the experimental treatment may not improve overall survival compared to the control. As shown in the Figure 4.1, the Control group has better survival rates at the same time points, with a significantly lower risk of death. In contrast, the Treatment group exhibits a higher risk of death. This result suggests that the experimental treatment may not have significantly improved patient survival, and it may even be worse than the Control group.