**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.

# correlation matrix, representing linear relationship

corrMat <- cor(scaledData)

corrplot(corrMat,

method = "circle",

type = "upper",

col = COL2("RdBu", 10),

tl.col = "black",

tl.srt = 45,

order = "original")

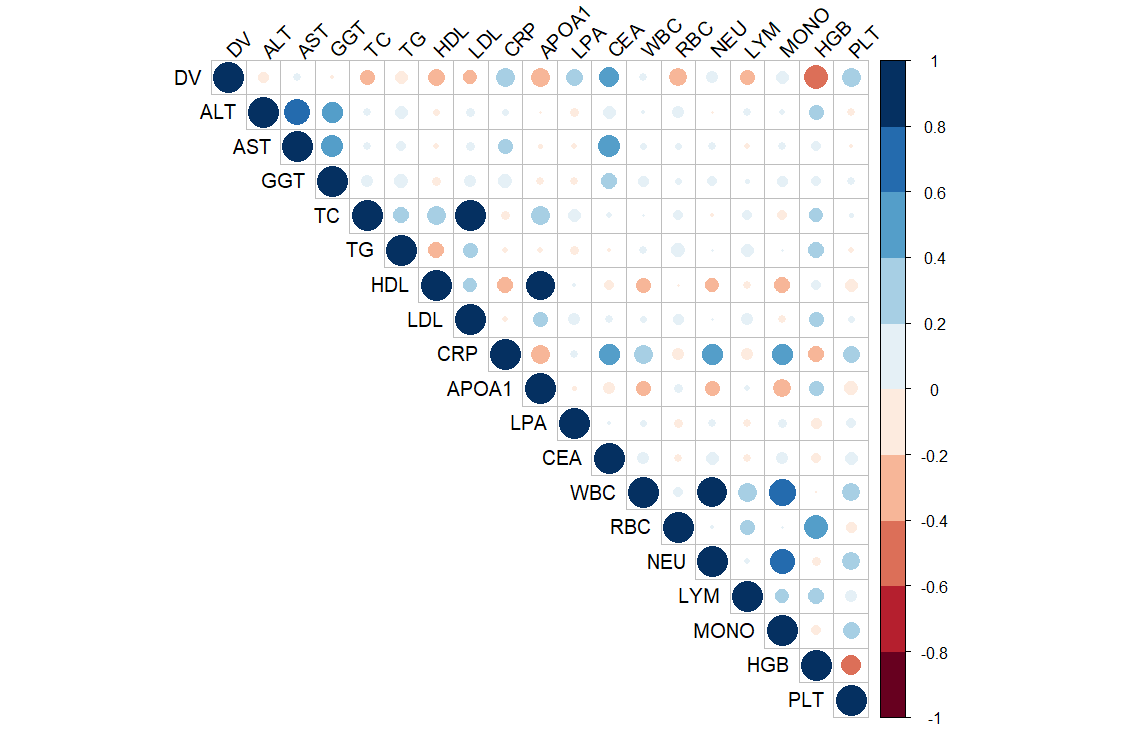


Figure 1.1 Correlation Matrix

# plot each pair of high-related features

p1 <- ggplot(data, aes(x = TC, y = LDL)) +

geom\_point() +

geom\_smooth(method = "lm")+

theme\_minimal()

p2 <- ggplot(data, aes(x = HDL, y = APOA1)) +

geom\_point() +

geom\_smooth(method = "lm")+

theme\_minimal()

p3 <- ggplot(data, aes(x = WBC, y = NEU)) +

geom\_point() +

geom\_smooth(method = "lm")+

theme\_minimal()

grid.arrange(p1, p2, p3, ncol = 3)

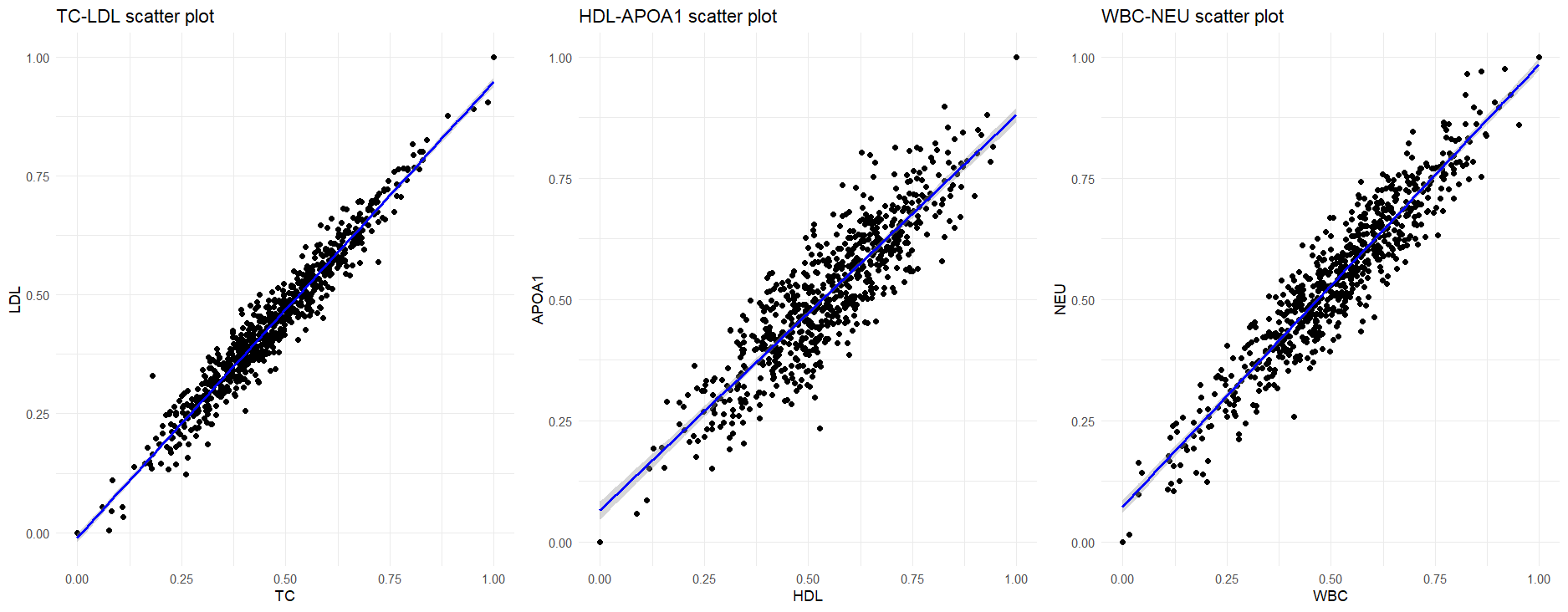


Figure 1.2 Linear Relationship of 3 high-related pairs

# Divide into training and testing sets

set.seed(2024)

# split the data into training and testing sets

trainIndex <- sample(1:nrow(data), 0.7\*nrow(data))

trainFea <- scaledFeature[trainIndex,]

trainLabel <- scaledData[trainIndex,1]

testFea <- scaledFeature[-trainIndex,]

testLabel <- scaledData[-trainIndex,1]

x = as.matrix(trainFea) # transfer the data frame to matrix

y = as.matrix(trainLabel)

newx <- as.matrix(testFea)

# Logistic Regression

logic\_reg = glm(y ~ ., data = trainFea, family = "binomial")

#Elastic Net

elastic\_net\_model <- cv.glmnet(x, y, alpha = 0.5, family = "binomial")

plot(elastic\_net\_model)

best\_lambda <- elastic\_net\_model$lambda.min

print(best\_lambda)

0.002294098

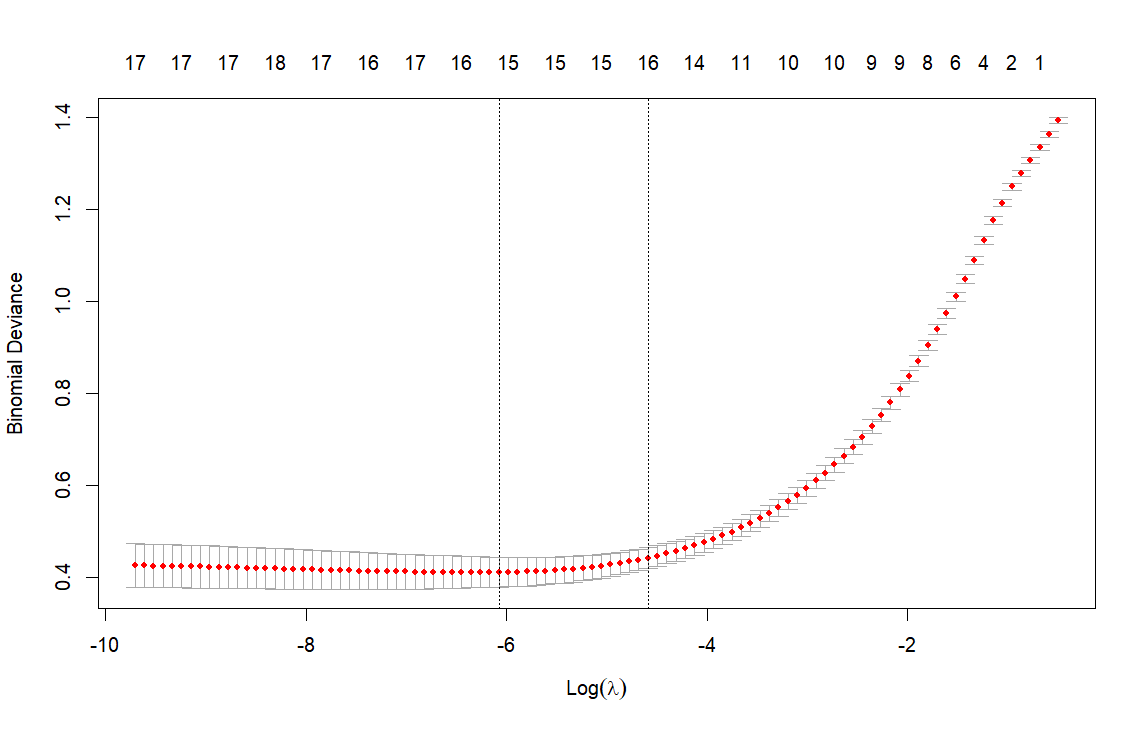


Figure 1.3 Cross Validation of Elastic Net

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.

best\_model\_coeff <- coef(elastic\_net\_model, s = "lambda.min")

Table 1.1 Best 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 logistic 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.

# Accuracy

accuracy\_logic <- sum(pre\_logic\_class == testLabel) / length(testLabel)

accuracy\_elas <- sum(pre\_elas\_class == testLabel) / length(testLabel)

# Precision

precision\_logic <- sum(pre\_logic\_class == 1 & testLabel == 1) / sum(pre\_logic\_class == 1)

precision\_elas <- sum(pre\_elas\_class == 1 & testLabel == 1) / sum(pre\_elas\_class == 1)

# Recall

recall\_logic <- sum(pre\_logic\_class == 1 & testLabel == 1) / sum(testLabel == 1)

recall\_elas <- sum(pre\_elas\_class == 1 & testLabel == 1) / sum(testLabel == 1)

# F1

f1\_logic <- 2 \* precision\_logic \* recall\_logic / (precision\_logic + recall\_logic)

f1\_elas <- 2 \* precision\_elas \* recall\_elas / (precision\_elas + recall\_elas)

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

|  |  |  |
| --- | --- | --- |
| Metrics | Before Elastic Net | After Elastic Net |
| Accuracy | 0.8972 | 0.9112 |
| Precision | 0.8812 | 0.8846 |
| Recall | 0.8990 | 0.9293 |
| F1 | 0.8900 | 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. Table 2.1 and Figure 2.1 show the distribution of Age by Stage. Table 2.2~2.4 and Figure 2.2~2.4 show distributions of Partner Status, Ethnicity and Site by 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 |

图表, 条形图

描述已自动生成

Figure 2.1 Age distribution by Stage

Table 2.2 Stage distribution by Partner Status

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

图表, 条形图

描述已自动生成

Figure 2.2 Stage distribution by Partner Status

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 |

图表, 条形图

描述已自动生成

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 |

图表, 条形图

描述已自动生成

Figure 2.4 Stage distribution by Site

We performed a multinomial logistic regression analysis, with CRC staging as the dependent variable and age, ethnicity, Partner Status, and cancer site as independent variables. The analysis was conducted using the multinom function in R. The multinomial logistic regression model successfully converged after 30 iterations, with an AIC of 440.60.

From the results we can see that Age had a significant impact on CRC staging. Increasing age was positively correlated with Stage 4, with a coefficient of 0.777 and a p-value less than 0.05.

Ethnicity did not show a significant effect on CRC staging. While both Black and White patients had negative coefficients compared to the Asian category, the p-values were above 0.05, indicating that ethnicity had no significant impact on staging in this data.

Partner status significantly influenced CRC staging. Patients without a partner were more likely to present with advanced stages (Stage 3 and Stage 4). Specifically, in Stage 4, the p-value was 0.0038, indicating a significant relationship between no partner and later-stage cancer.

Site had a partial significant effect on staging. Rectal cancer (Rectum) and right-sided colon cancer (Right) were more likely to present in earlier stages (Stage 2), compared to left-sided colon cancer (Left), with a p-value of 0.0096 in Stage 2.

# multinom Logistic Regression

model <- multinom(Stage ~ Age + Ethnicity + Partner + Site, data = data)

summary(model)

z\_values <- summary(model)$coefficients / summary(model)$standard.errors

p\_values <- (1 - pnorm(abs(z\_values), 0, 1)) \* 2

> summary(model)

Call:

multinom(formula = Stage ~ Age + Ethnicity + Partner + Site,

data = data)

Coefficients:

(Intercept) Age EthnicityBlack EthnicityWhite Partner1 SiteRectum SiteRight

2 -4.91196 0.2552062 -1.8758858 -1.050751 -0.2187815 -0.9529227 -1.492296

3 -10.39972 0.4552287 -1.4888519 -1.512040 -1.2036202 -0.7115117 -1.142369

4 -21.58138 0.7773201 -0.4197205 -1.281200 -1.8421940 -0.5866978 -0.817067

Std. Errors:

(Intercept) Age EthnicityBlack EthnicityWhite Partner1 SiteRectum SiteRight

2 2.229171 0.06694487 1.357086 1.242041 0.4836505 0.6107411 0.5763187

3 2.544499 0.07795927 1.400915 1.304087 0.5353692 0.6715391 0.6220647

4 3.474684 0.10015148 1.568279 1.456569 0.6356305 0.7821828 0.7270260

Residual Deviance: 398.5969

AIC: 440.5969

> p\_values

(Intercept) Age EthnicityBlack EthnicityWhite Partner1 SiteRectum SiteRight

2 2.756012e-02 1.377440e-04 0.1668829 0.3975599 0.651013594 0.1186954 0.009615504

3 4.367240e-05 5.241619e-09 0.2878859 0.2462678 0.024563063 0.2893613 0.066296396

4 5.263718e-10 8.437695e-15 0.7889832 0.3790754 0.003752928 0.4532079 0.261077311

> z\_values

(Intercept) Age EthnicityBlack EthnicityWhite Partner1 SiteRectum SiteRight

2 -2.203492 3.812184 -1.3822891 -0.8459873 -0.4523545 -1.5602729 -2.589358

3 -4.087140 5.839314 -1.0627707 -1.1594624 -2.2482059 -1.0595238 -1.836414

4 -6.211034 7.761444 -0.2676312 -0.8796011 -2.8982153 -0.7500776 -1.123848

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.

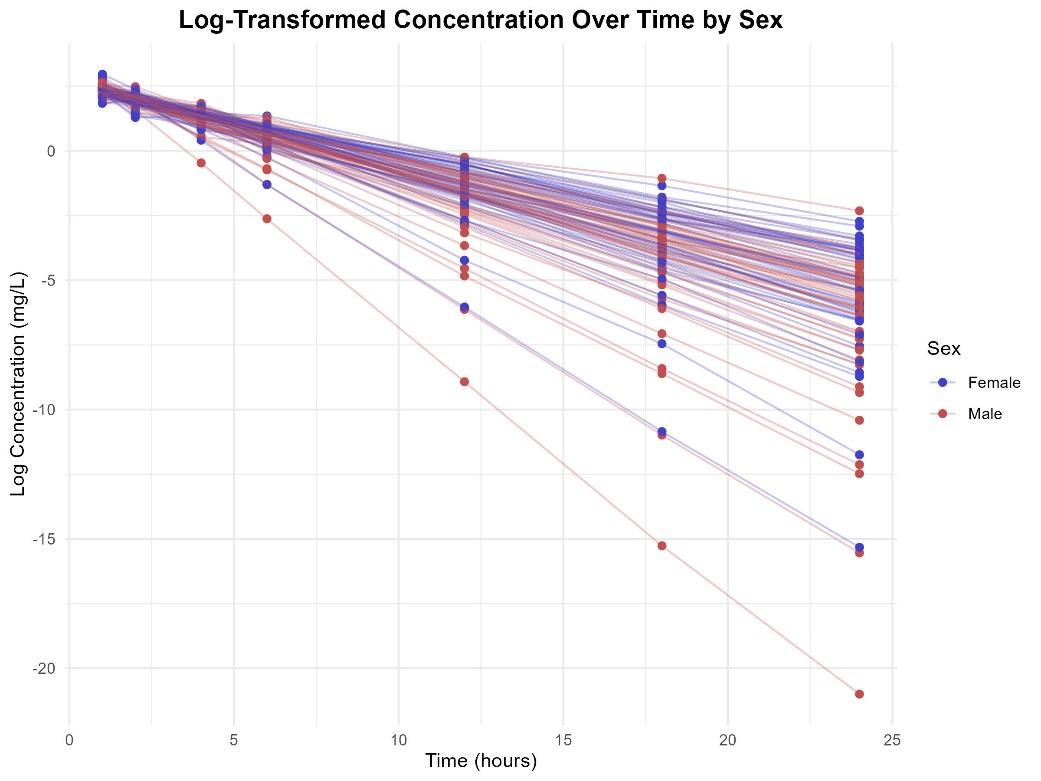


Figure3.1 Log Concentration Over Time

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.

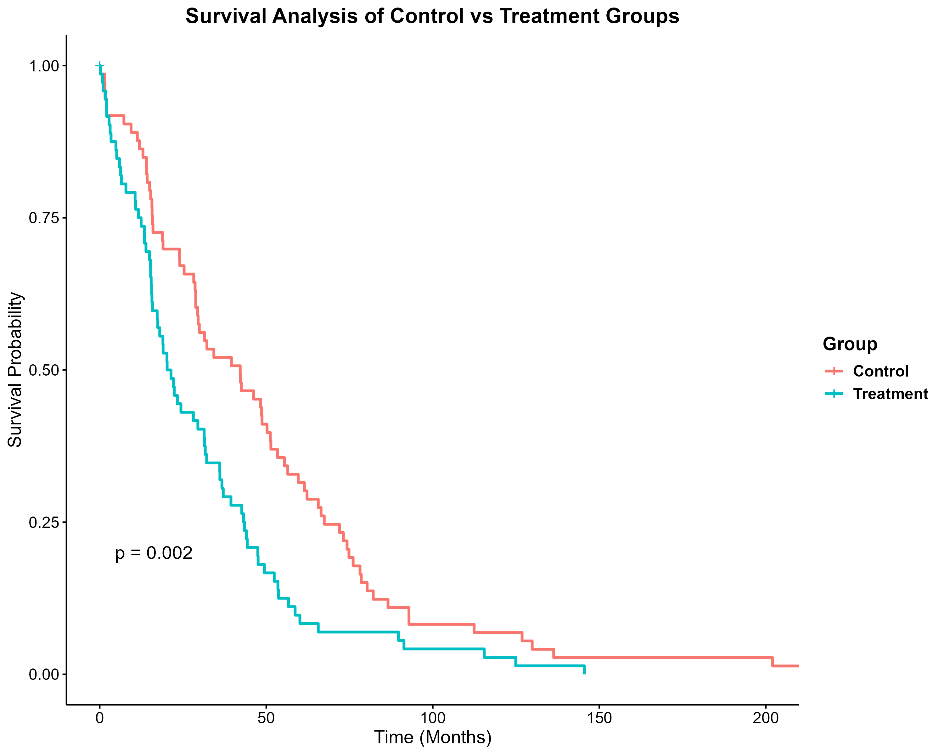


Figure 4.1 Survival Analysis of Control and Treatment Group

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.