Seminar 2

Group A2

Yufeng Deng Yuanqing Wang

Task 1

The analysis of the distribution patterns of the three scores for cognitively normal and cognitively impaired individuals is the most straightforward and evident task. The distributions are shown in the figures below. We used box plots and histograms to illustrate the differences, providing an overall understanding of the data. Where the LABEL=0 is cognitively normal and LABEL=1 is cognitively impaired.

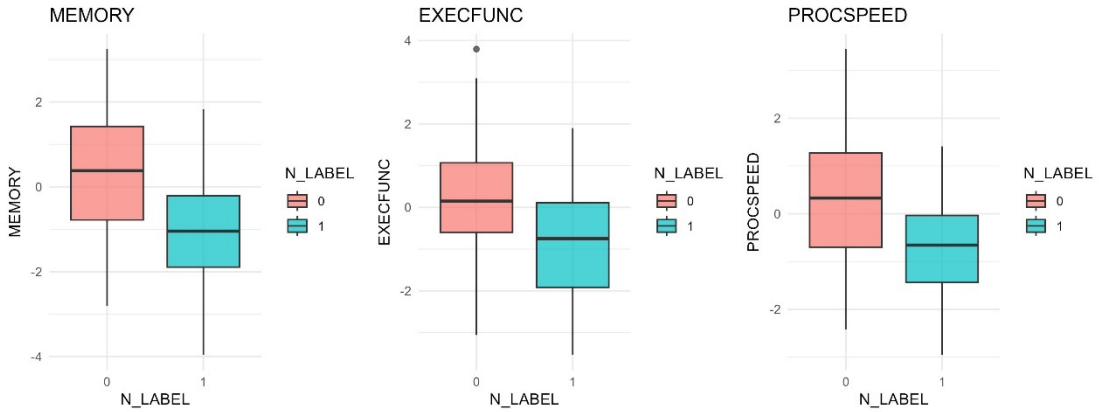

Figure 1.1 Box plots of scores

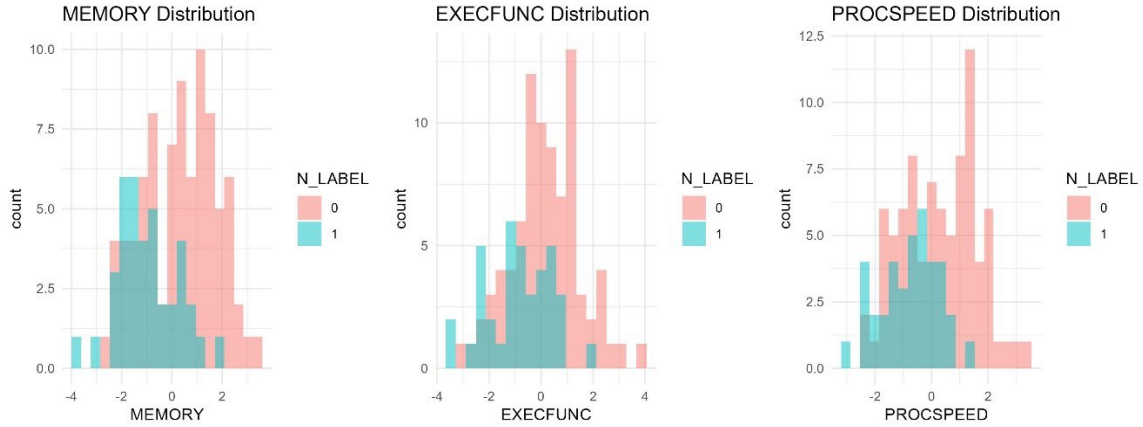

Figure 1.2 Histograms of scores

From the figures, it can be observed that individuals with cognitive impairment have lower mean scores in all three assessments. This is reasonable for a behavioral test. No significant differences were found between the standard deviations of the data for impaired individuals and those for cognitively normal individuals.

Then, the six variables—SEX, HYPERTENSION, DIABETES, ATRIALFIBR, INFARCTION, and AMYLOIDVIS—are subjected to chi-square tests against N_LABEL. During the tests, it was found that some frequencies were less than 5, which can reduce the confidence level of the chi-square test results. The Fisher test is more accurate for small frequency samples, so it should be used to replace the chi-square test for assessing the correlation between variables.

```r
data$N_LABEL <- as.factor(data$N_LABEL)
variables <- c("SEX", "HYPERTENSION", "DIABETES", "ATRIALFIBR", "INFARCTION",
"AMYLOIDVIS")
for (var in variables) {
  contingency_table <- table(data$N_LABEL, data[[var]])
  #cat("Contingency table for", var, ":\n")
  #print(contingency_table)
  #cat("\n")

  fisher_result <- fisher.test(contingency_table)

  cat("Fisher's exact test for", var, ":\n")
  print(fisher_result)
  cat("\n")
}
```

```
Fisher's exact test for SEX :
        Fisher's Exact Test for Count Data
data:  contingency_table
p-value = 0.4407
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.5932765 3.2210470
sample estimates:
odds ratio
  1.371404

Fisher's exact test for HYPERTENSION :
        Fisher's Exact Test for Count Data
data:  contingency_table
p-value = 0.1247
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.2018769 1.2682876
sample estimates:
odds ratio
 0.5041308

Fisher's exact test for DIABETES :
        Fisher's Exact Test for Count Data
data:  contingency_table
p-value = 1
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.01829386 21.89579366
```

```
sample estimates:
odds ratio
  1.107154

Fisher's exact test for ATRIALFIBR :
        Fisher's Exact Test for Count Data
data:  contingency_table
p-value = 0.002208
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  1.616074 16.709141
sample estimates:
odds ratio
  4.996928

Fisher's exact test for INFARCTION :
        Fisher's Exact Test for Count Data
data:  contingency_table
p-value = 0.1702
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.005009745 1.748740856
sample estimates:
odds ratio
 0.2272486

Fisher's exact test for AMYLOIDVIS :
        Fisher's Exact Test for Count Data
data:  contingency_table
p-value = 0.2973
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.6244173 3.7492342
sample estimates:
odds ratio
  1.541436
```

In the Fisher tests conducted on N_LABEL and six variables, it was found that SEX (p=0.4407), HYPERTENSION (p=0.1247), DIABETES (p=1.000), INFARCTION (p=0.1702) and AMYLOIDVIS (0.2973) showed no significant relationship with cognitive status, suggesting that these variables may have a minimal impact on cognitive status. Although the odds ratio for ATRIALFIBR is 4.996, indicating a strong association with cognitive impairment (p=0.0022), it is important to note that while the odds ratios for other variables suggest potential associations, these results lack statistical significance. However, whether these variables influence the final scores

and the patients' cognitive status requires further analysis.

Then, based on the grouping of N_LABEL, perform Wilcoxon rank-sum tests on the remaining continuous variables to observe the differences between groups and examine their correlations with cognitive status.

```
# Wilcoxon rank-sum test
variables <- c("AGE", "EDUCATION", "MMSE", "CERAD", "PASTCAQ", "HANDGRIP", "MUSCLE",
"SPPB", "MNA", "BMI", "CRP", "LEUK", "BPSYS", "BPDIA", "HBA1C", "CHOLESTEROL", "WMHVOL",
"HIPPOVOL", "AMYLOIDBIND")
results <- data.frame(Variable = character(), p.value = numeric(), stringsAsFactors =
FALSE)

for (var in variables) {
  test_result <- wilcox.test(data[[var]] ~ data$N_LABEL)
  results <- rbind(results, data.frame(Variable = var, p.value = test_result$p.value))
}
print(results)
```

Table1.1 Results of rank-sum test

Variable	p-value
AGE	1.937940e-05
EDUCATION	1.856402e-01
MMSE	1.162747e-12
CERAD	5.055813e-14
PASTCAQ	8.234381e-01
HANDGRIP	1.122971e-04
MUSCLE	5.138574e-01
SPPB	7.370349e-02
MNA	4.660691e-04
BMI	5.860135e-01
CRP	5.972583e-02
LEUK	3.994737e-02
BPSYS	5.394276e-01
BPDIA	5.213018e-01
HBA1C	1.624442e-01
CHOLESTEROL	2.614635e-01
WMHVOL	1.560449e-02
HIPPOVOL	5.087277e-06
AMYLOIDBIND	3.661085e-04

Meanwhile, in order to keep the accuracy of test, we performance multiple test correction.

```
# Multiple Testing Correction
results <- data.frame(
  Variable = c("AGE", "EDUCATION", "MMSE", "CERAD", "PASTCAQ", "HANDGRIP", "MUSCLE",
               "SPPB", "MNA", "BMI", "CRP", "LEUK", "BPSYS", "BPDIA", "HBA1C",
               "CHOLESTEROL", "WMHVOL", "HIPPOVOL", "AMYLOIDBIND"),
  p.value = c(1.937940e-05, 1.856402e-01, 1.162747e-12, 5.055813e-14, 8.234381e-01,
              1.122971e-04, 5.138574e-01, 7.370349e-02, 4.660691e-04, 5.860135e-01,
              5.972583e-02, 3.994737e-02, 5.394276e-01, 5.213018e-01, 1.624442e-01,
              2.614635e-01, 1.560449e-02, 5.087277e-06, 3.661085e-04)
```

```
)
# cal Bonferroni corrections
results$Bonferroni_p <- p.adjust(results$p.value, method = "bonferroni")

# cal BH corrections
results$BH_p <- p.adjust(results$p.value, method = "BH")

results$Significant_Bonferroni <- results$Bonferroni_p < 0.05
results$Significant_BH <- results$BH_p < 0.05

print(results)
```

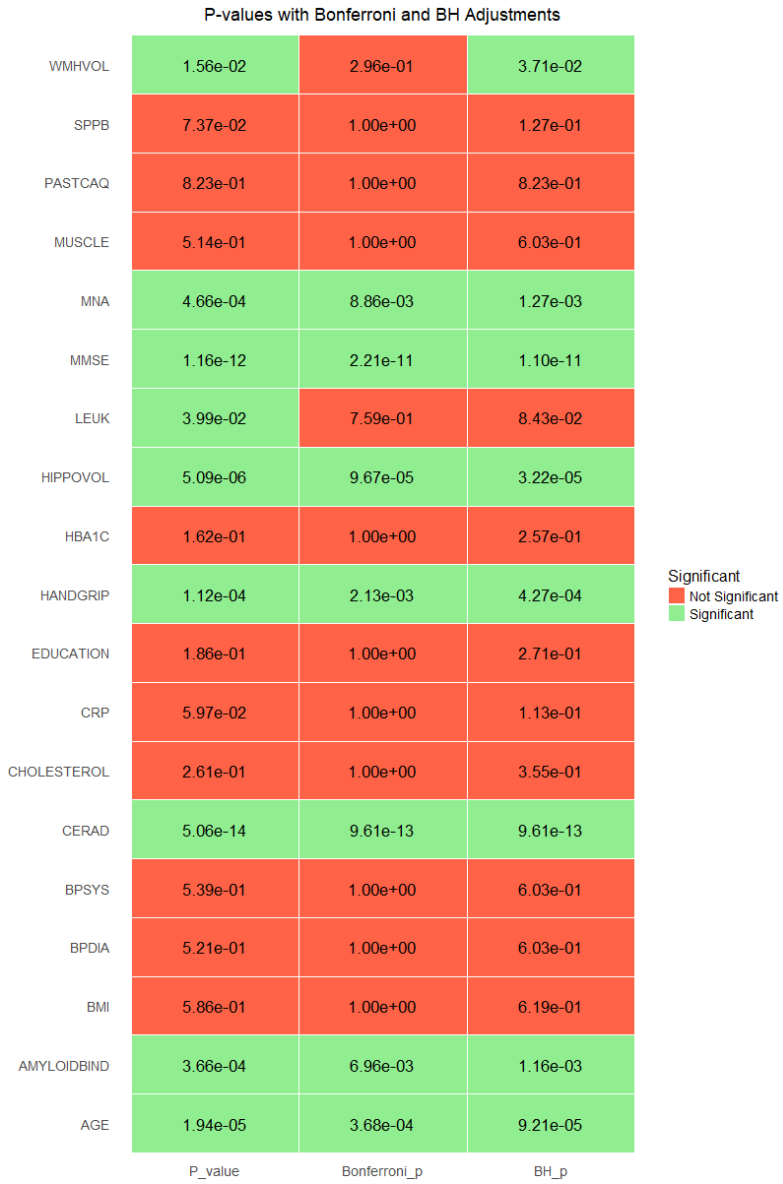

Figure1.3 Result of multiple test correction

Based on the above tests, we can conclude that the categorical variable ATRIALFIBR shows a significant association with cognitive status after performing the chi-square test. Seven continuous

variables—AGE, MMSE, CERAD, HANDGRIP, MNA, HIPPOVOL, and AMYLOIDBIND—remain significant after multiple corrections, indicating substantial differences between the two groups and suggesting a strong association with the group classification (N_LABEL). WMHVOL is significant under a more lenient correction (BH) but not under the stricter Bonferroni correction, implying a moderate difference between the groups. Although LEUK is not significant after correction, it still warrants some attention. These ten variables demonstrate a statistical association with cognitive status, but this does not necessarily mean they contribute more weight to the three scores. However, the results remain statistically significant, and we will focus on these variables in the subsequent regression models.

As shown in Figure 1.4, these are the boxplots of variables with significant differences between groups.

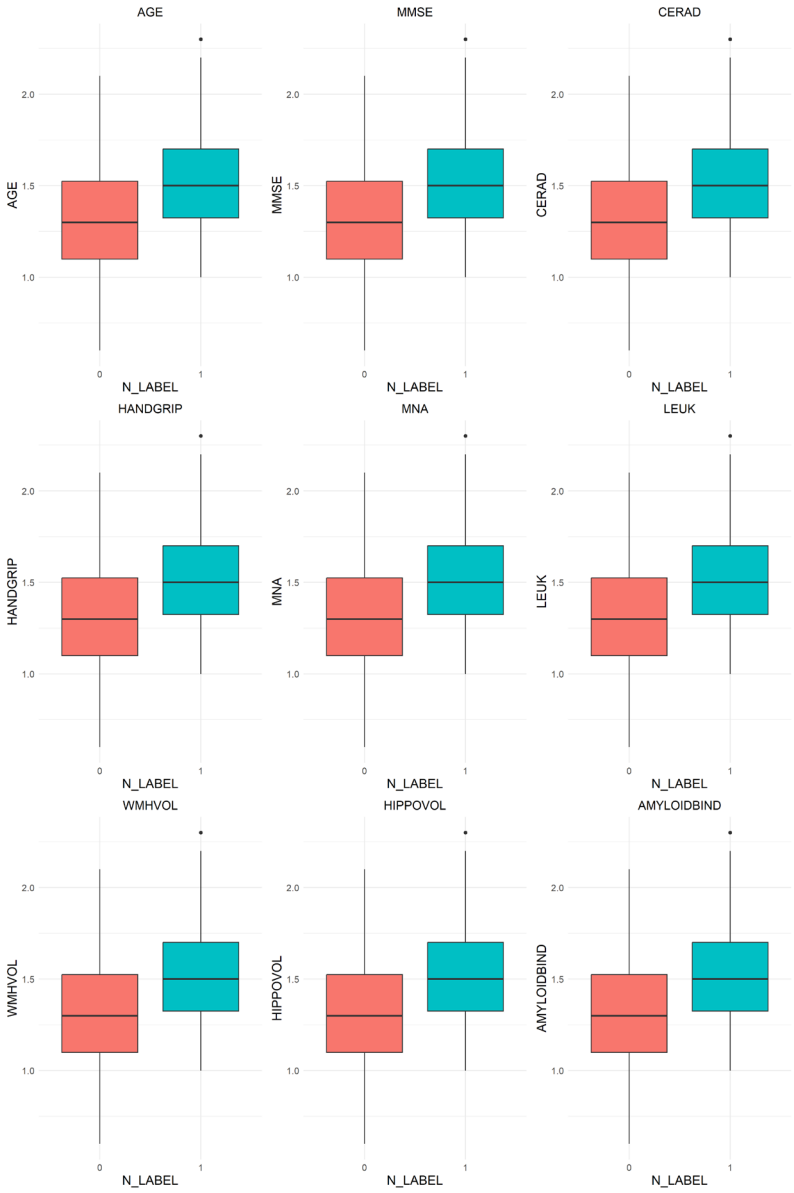

Figure1.4 Box plots of significant variables

As shown in Figure 1.5, these are the boxplots of variables with no significant differences between

groups. Figures 1.4 and 1.5 visually illustrate whether the differences in group variables are significant.

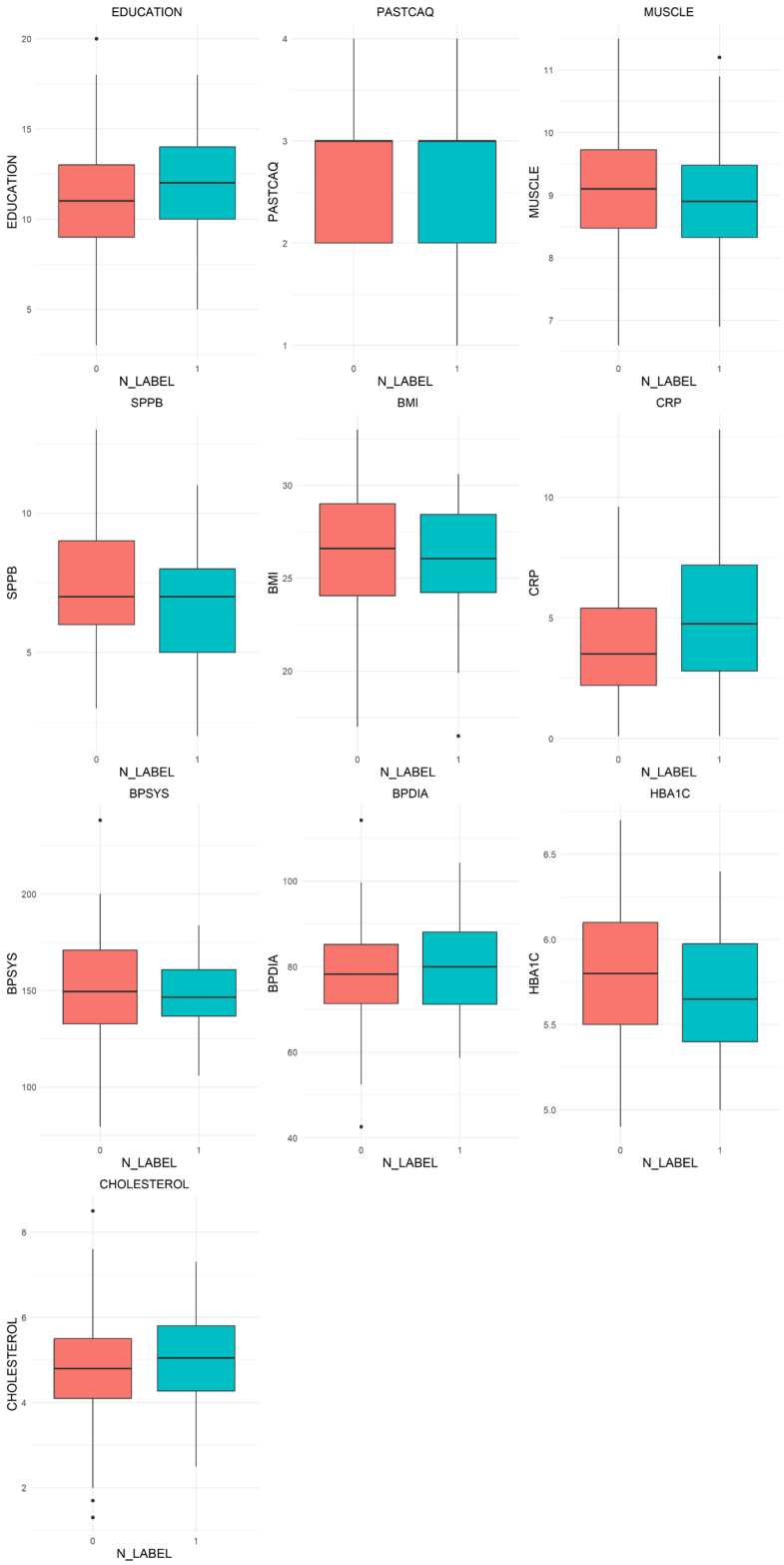

Figure1.4 Box plots of insignificant variables

The correlation matrix for the continuous variables was calculated. As shown in the figure, there is a relatively weak linear correlation among the variables. There is no significant relation between

any pairs of variables.

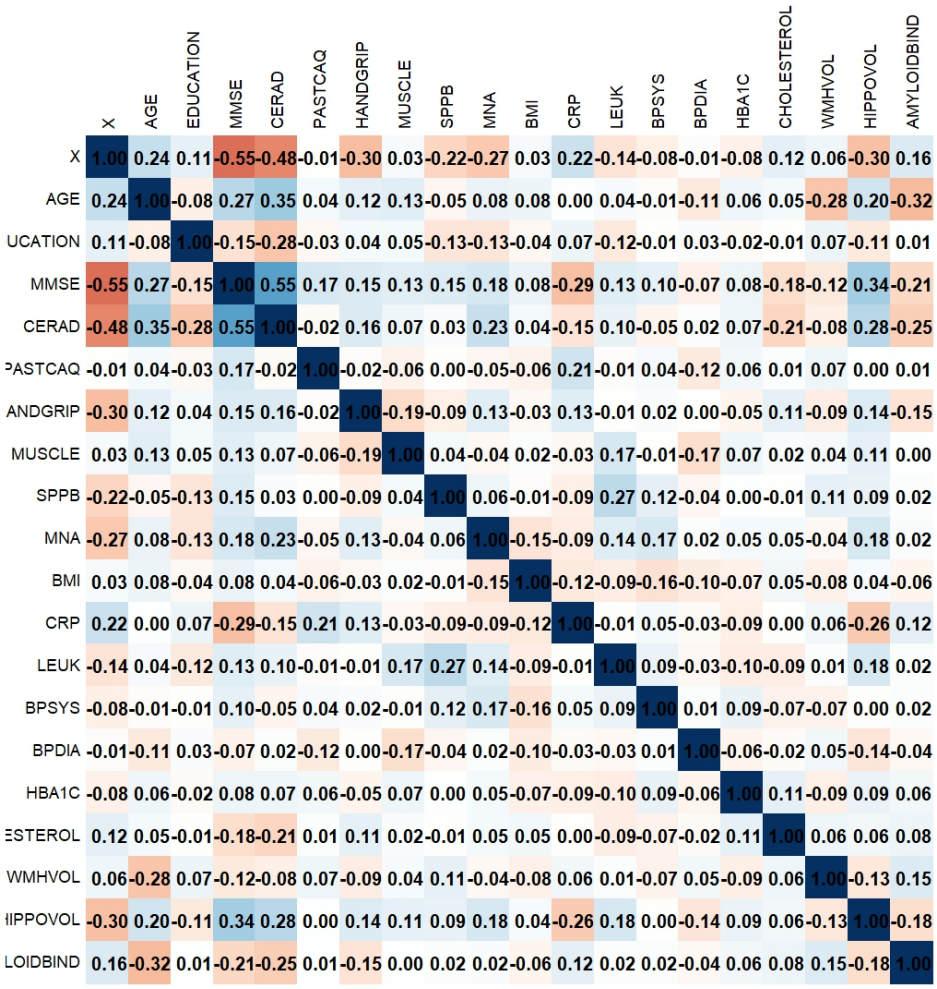

Figure 1.4 Correlation matrix of different variables

Due to the weak correlations among the variables and the large number of variables, we used the AIC (Akaike Information Criterion) method to simplify our model and obtain a more streamlined result

```
# AIC
full_model_memory <- lm(MEMORY ~ N_LABEL + AGE + EDUCATION + MMSE + CERAD + PASTCAQ +
HANDGRIP + MUSCLE + SPPB + MNA + BMI + CRP + LEUK + BPSYS + BPDIA + HBA1C + CHOLESTEROL
+ WMHVOL + HIPPOVOL + AMYLOIDBIND, data = data)

optimized_model_memory <- step(full_model_memory, direction = "both", trace = FALSE)

summary(optimized_model_memory)

full_model_execfunc <- lm(EXECFUNC ~ N_LABEL + AGE + EDUCATION + MMSE + CERAD + PASTCAQ
+ HANDGRIP + MUSCLE + SPPB + MNA + BMI + CRP + LEUK + BPSYS + BPDIA + HBA1C +CHOLESTEROL
+ WMHVOL + HIPPOVOL + AMYLOIDBIND, data = data)
optimized_model_execfunc <- step(full_model_execfunc, direction = "both", trace = FALSE)

summary(optimized_model_execfunc)
```

```r
full_model_procspeed <- lm(PROCSPEED ~ N_LABEL + AGE + EDUCATION + MMSE + CERAD +
PASTCAQ + HANDGRIP + MUSCLE + SPPB + MNA + BMI + CRP + LEUK + BPSYS + BPDIA + HBA1C +
CHOLESTEROL + WMHVOL + HIPPOVOL + AMYLOIDBIND, data = data)

optimized_model_procspeed <- step(full_model_procspeed, direction = "both", trace =
FALSE)

summary(optimized_model_procspeed)
```

```
Call:
lm(formula = MEMORY ~ MUSCLE + SPPB + MNA + BMI + BPSYS + BPDIA +
    HBA1C + WMHVOL + HIPPOVOL + AMYLOIDBIND, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.58020 -0.65142 -0.04628  0.53095  2.49511

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.344647   2.346479  -6.113 1.49e-08 ***
MUSCLE        0.512798   0.100976   5.078 1.55e-06 ***
SPPB          0.154524   0.039434   3.919 0.000154 ***
MNA           0.191054   0.074757   2.556 0.011951 *
BMI           0.048656   0.026912   1.808 0.073315 .
BPSYS         0.008202   0.004308   1.904 0.059509 .
BPDIA         0.012387   0.008439   1.468 0.144963
HBA1C         0.619109   0.242742   2.550 0.012120 *
WMHVOL       -0.360692   0.115034  -3.136 0.002196 **
HIPPOVOL      2.531966   0.527605   4.799 5.01e-06 ***
AMYLOIDBIND  -0.702964   0.309379  -2.272 0.025002 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.059 on 111 degrees of freedom
Multiple R-squared:  0.5488,    Adjusted R-squared:  0.5081
F-statistic:  13.5 on 10 and 111 DF,  p-value: 2.991e-15


Call:
lm(formula = EXECFUNC ~ EDUCATION + MMSE + PASTCAQ + MUSCLE +
    SPPB + MNA + BPSYS + HBA1C + WMHVOL + HIPPOVOL, data = data)

Residuals:
```

```
     Min       1Q   Median       3Q      Max 
-2.45251 -0.63954 -0.01434  0.79976  2.65366 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept) -6.669096   2.238373  -2.979  0.00355 **
EDUCATION   -0.054679   0.032697  -1.672  0.09728 .
MMSE         0.053551   0.038432   1.393  0.16628  
PASTCAQ     -0.255481   0.172290  -1.483  0.14095  
MUSCLE       0.306640   0.110455   2.776  0.00646 **
SPPB         0.090581   0.043731   2.071  0.04065 *
MNA          0.162850   0.082117   1.983  0.04982 *
BPSYS       -0.009325   0.004685  -1.990  0.04900 *
HBA1C        0.398464   0.264416   1.507  0.13466  
WMHVOL      -0.244557   0.125541  -1.948  0.05394 .
HIPPOVOL     1.620409   0.589421   2.749  0.00698 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.158 on 111 degrees of freedom
Multiple R-squared:  0.3528,    Adjusted R-squared:  0.2944 
F-statistic:  6.05 on 10 and 111 DF,  p-value: 2.726e-07


Call:
lm(formula = PROCSPEED ~ N_LABEL + EDUCATION + CERAD + MUSCLE + 
    SPPB + MNA + BMI + CRP + LEUK + CHOLESTEROL + WMHVOL + HIPPOVOL + 
    AMYLOIDBIND, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.23259 -0.49948 -0.03175  0.53399  2.42630 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.34148    1.47367  -4.982 2.41e-06 ***
N_LABEL      0.79159    0.27103   2.921 0.004253 ** 
EDUCATION    0.05064    0.02516   2.013 0.046633 *  
CERAD        0.08025    0.02269   3.536 0.000600 ***
MUSCLE       0.11857    0.08140   1.457 0.148114    
SPPB         0.14909    0.03386   4.403 2.52e-05 ***
MNA          0.13447    0.06263   2.147 0.034024 *  
BMI          0.06003    0.02153   2.789 0.006257 ** 
CRP         -0.08509    0.03216  -2.646 0.009358 ** 
```

```
LEUK          0.08864    0.03454   2.567 0.011636 *
CHOLESTEROL   0.10037    0.06229   1.611 0.110015
WMHVOL       -0.27255    0.09479  -2.875 0.004864 **
HIPPOVOL      2.95996    0.45467   6.510 2.43e-09 ***
AMYLOIDBIND  -0.93643    0.26107  -3.587 0.000504 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8487 on 108 degrees of freedom
Multiple R-squared:  0.6338,    Adjusted R-squared:  0.5898
F-statistic: 14.38 on 13 and 108 DF,  p-value: < 2.2e-16
```

We will take the intersection of the significant variables mentioned above and the variables identified for further attention in the previous discussion, and form a new linear model. We will then perform linear regression again to obtain the final model.

```
model_memory <- lm(MEMORY ~ MUSCLE + SPPB + MNA + HBA1C + WMHVOL + HIPPOVOL +
AMYLOIDBIND, data = data)
summary(model_memory)

model_execfunc <- lm(EXECFUNC ~ MUSCLE + SPPB + MNA + WMHVOL + HIPPOVOL, data = data)
summary(model_execfunc)

model_procspeed <- lm(PROCSPEED ~ N_LABEL + CERAD + SPPB + MNA + BMI + CRP + LEUK +
WMHVOL + HIPPOVOL + AMYLOIDBIND, data = data)
summary(model_procspeed)
```

```
Call:
lm(formula = MEMORY ~ MUSCLE + SPPB + MNA + HBA1C + WMHVOL +
    HIPPOVOL + AMYLOIDBIND, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5712 -0.6498 -0.1381  0.7388  2.7857

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.64934    1.90139  -5.601 1.50e-07 ***
MUSCLE        0.49134    0.10182   4.826 4.36e-06 ***
SPPB          0.16209    0.03995   4.058 9.11e-05 ***
MNA           0.19815    0.07462   2.655  0.00906 **
HBA1C         0.61407    0.24633   2.493  0.01411 *
WMHVOL       -0.38546    0.11647  -3.310  0.00125 **
HIPPOVOL      2.42830    0.53340   4.553 1.33e-05 ***
AMYLOIDBIND  -0.74435    0.31501  -2.363  0.01982 *
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.081 on 114 degrees of freedom
Multiple R-squared:  0.5169,    Adjusted R-squared:  0.4872
F-statistic: 17.43 on 7 and 114 DF,  p-value: 1.556e-15



Call:
lm(formula = EXECFUNC ~ MUSCLE + SPPB + MNA + WMHVOL + HIPPOVOL,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5899 -0.7936  0.0641  0.7928  2.8647

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.10821    1.48182  -4.122 7.08e-05 ***
MUSCLE       0.33536    0.11222   2.988 0.003423 **
SPPB         0.09777    0.04412   2.216 0.028632 *
MNA          0.17475    0.08224   2.125 0.035730 *
WMHVOL      -0.28425    0.12720  -2.235 0.027356 *
HIPPOVOL     2.02643    0.57867   3.502 0.000657 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.195 on 116 degrees of freedom
Multiple R-squared:  0.2807,    Adjusted R-squared:  0.2497
F-statistic: 9.053 on 5 and 116 DF,  p-value: 2.734e-07



Call:
lm(formula = PROCSPEED ~ N_LABEL + CERAD + SPPB + MNA + BMI +
    CRP + LEUK + WMHVOL + HIPPOVOL + AMYLOIDBIND, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.15058 -0.48419 -0.07701  0.61037  2.29942

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.82154    1.20823  -3.991 0.000119 ***
N_LABEL      0.67731    0.27367   2.475 0.014839 *
CERAD        0.06080    0.02156   2.820 0.005697 **
```

```
SPPB          0.13747    0.03439   3.998 0.000115 ***
MNA           0.12571    0.06364   1.975 0.050721 .
BMI           0.06175    0.02203   2.803 0.005981 **
CRP          -0.08311    0.03300  -2.518 0.013222 *
LEUK          0.08699    0.03468   2.508 0.013578 *
WMHVOL       -0.23414    0.09639  -2.429 0.016746 *
HIPPOVOL      3.05176    0.46241   6.600 1.46e-09 ***
AMYLOIDBIND -0.92336    0.26770  -3.449 0.000796 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8717 on 111 degrees of freedom
Multiple R-squared:  0.603,     Adjusted R-squared:  0.5673
F-statistic: 16.86 on 10 and 111 DF,  p-value: < 2.2e-16
```

In the memory model, muscle strength (MUSCLE), physical performance (SPPB), and hippocampal volume (HIPPOVOL) were found to have significant positive associations with memory scores, indicating that greater muscle strength, physical capability, and hippocampal volume may contribute to improved memory. In contrast, amyloid beta binding (AMYLOIDBIND) and white matter hyperintensity volume (WMHVOL) showed significant negative associations with memory, suggesting a potential link to memory decline.

In the executive function model, MUSCLE, SPPB, and HIPPOVOL also showed significant positive effects, suggesting that these physiological characteristics not only influence memory but are also strongly associated with improvements in executive function. Additionally, WMHVOL and MNA (Mini Nutritional Assessment) were significant in this model, highlighting the role of white matter changes and nutritional status in executive function.

In the processing speed model, beyond the previously mentioned factors, CERAD (Clinical Dementia Rating) was positively associated with processing speed, while WMHVOL and AMYLOIDBIND again showed significant negative associations, indicating their potential contribution to processing speed decline. Glycated hemoglobin (HBA1C) and body mass index (BMI) were also positively associated with processing speed, suggesting that metabolic factors may play a role in cognitive speed.

These analyses highlight the complex interplay of physiological, structural, and metabolic factors influencing cognitive function in older adults. The adjusted R^2 for the memory model was 0.4872, 0.2497 for the executive function model, and 0.5673 for the processing speed model, suggesting that these variables have the highest explanatory power in processing speed and the lowest in executive function.

Task 2

First, a scatter plot (Figure 2.1) was generated to provide an overview of the dataset. The plot illustrates a positive correlation between DV_Amyloid (Amy) and Age, indicating that as Age

increases, DV_Amyloid values tend to rise as well. However, due to the presence of multiple corresponding DV_Amyloid values for a single Age value, it was necessary to create a new dataset that establishes a one-to-one correspondence between Age and DV_Amyloid. To achieve this, we applied the method of averaging to obtain a mean value of DV_Amyloid for each unique Age. The resulting dataset was then visualized in a plot, as shown in Figure 2.2.

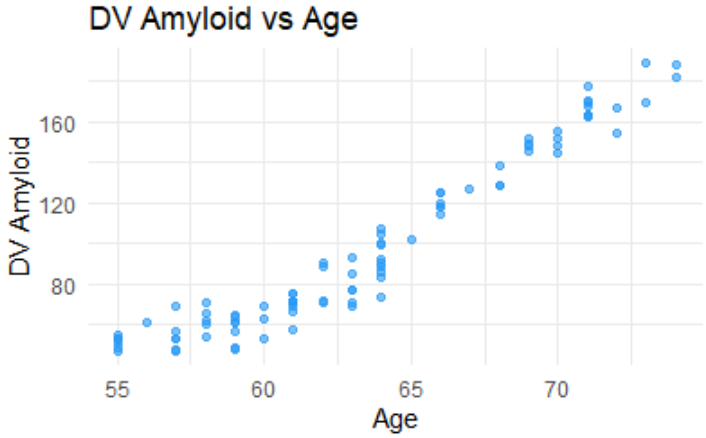

Figure 2.1 Scatter plot of the original dataset

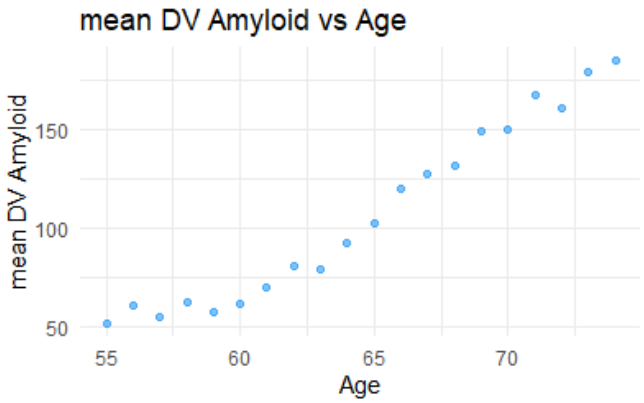

Figure 2.2 mean DV_Amyloid plot

The trend observed in Figure 2.2 appears to be linear; therefore, we initially employed a linear model for our analysis.

```
model_linear <- lm(average_amyloid$mean_amyloid ~ average_amyloid$age)
summary(model_linear)
```

```
Call:
lm(formula = average_amyloid$mean_amyloid ~ average_amyloid$age)
Residuals:
    Min      1Q  Median      3Q     Max
-16.7781  -7.9167   0.9219   5.9404  18.1809
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -381.2828    24.8251  -15.36 8.67e-12 ***
average_amyloid$age    7.5727     0.3834   19.75 1.19e-13 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.886 on 18 degrees of freedom
Multiple R-squared:  0.9559,    Adjusted R-squared:  0.9535
F-statistic: 390.2 on 1 and 18 DF,  p-value: 1.193e-13
```

The R-squared value obtained from the linear model is 0.9559, indicating a very strong fit to the data. Subsequently, we plotted the regression line alongside both the original dataset and the mean dataset, as illustrated in Figures 2.3 and 2.4.

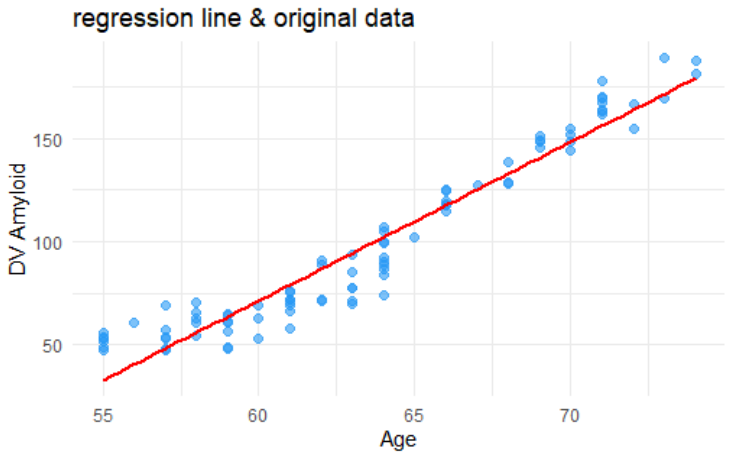

Figure 2.3 regression line & original data

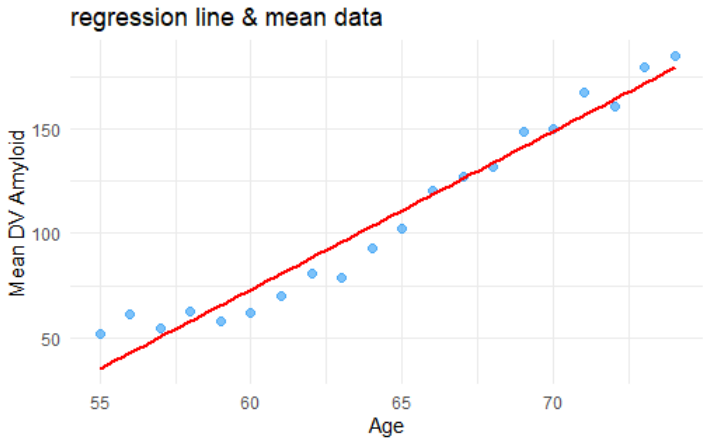

Figure 2.4 regression line & mean data

To evaluate the model, we analyzed the residuals versus fitted values plot and assessed the normality of the residuals.

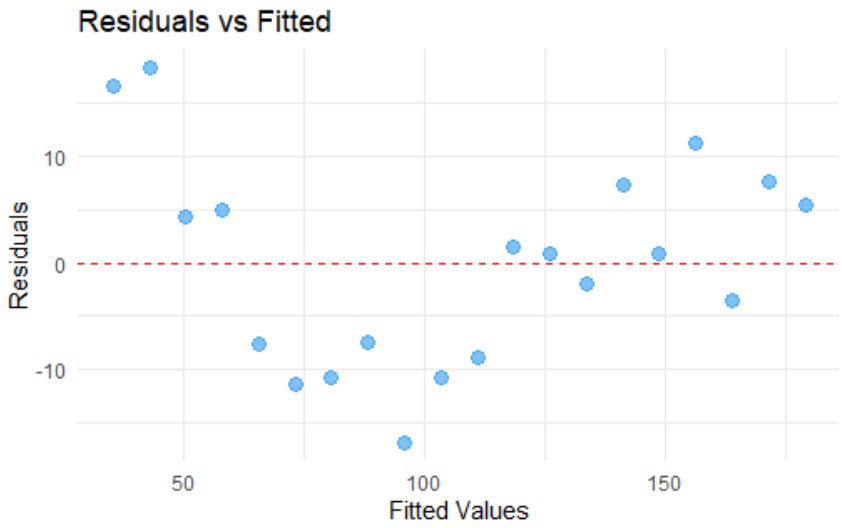

Figure 2.5 residuals value vs fitted value

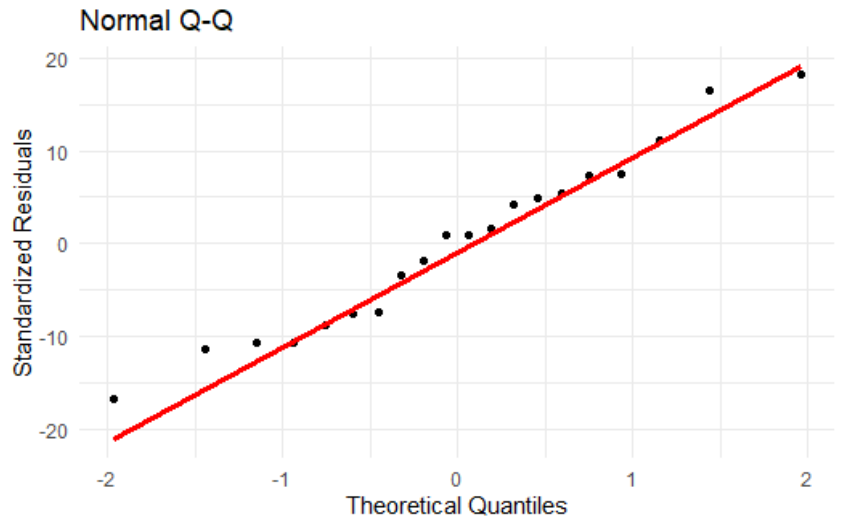

Figure 2.6 Q-Q plot

```
> shpiro_result <- shapiro.test(residuals(model_linear))
> print(shpiro_result)

        Shapiro-Wilk normality test

data:  residuals(model_linear)
W = 0.97047, p-value = 0.7647
```

As shown in Figure 2.5, the residuals exhibit considerable variability. The Q-Q plot and the results of the Shapiro-Wilk test indicate that the residuals do not follow a normal distribution. Consequently, the performance of the linear model is not sufficient.

Then, we tried Logarithmic Model, Polynomial Model (2), Mixed-effect Model. The workflow of the first two models is just as same as it in the Linear Model and the results are as follows.

Logarithmic Model

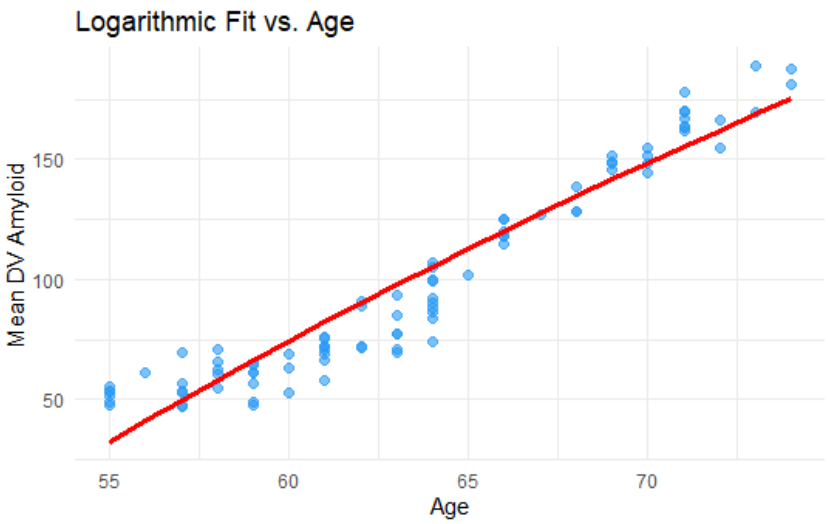

Figure 2.7 Logarithmic Model regression

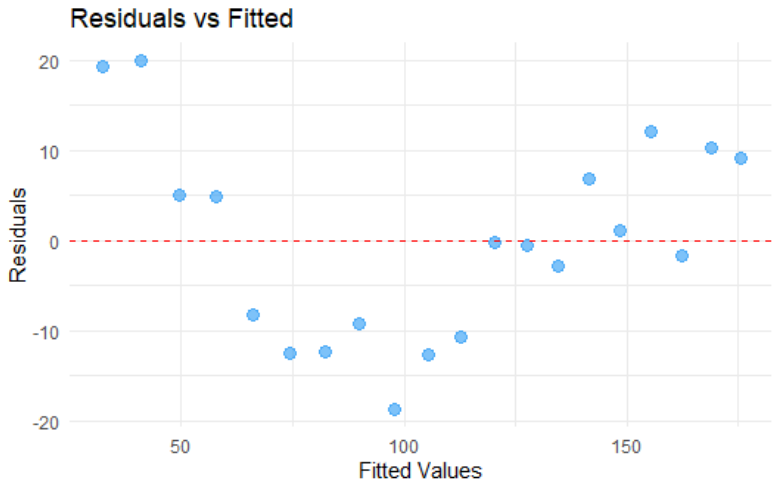

Figure 2.8 Logarithmic Model residuals

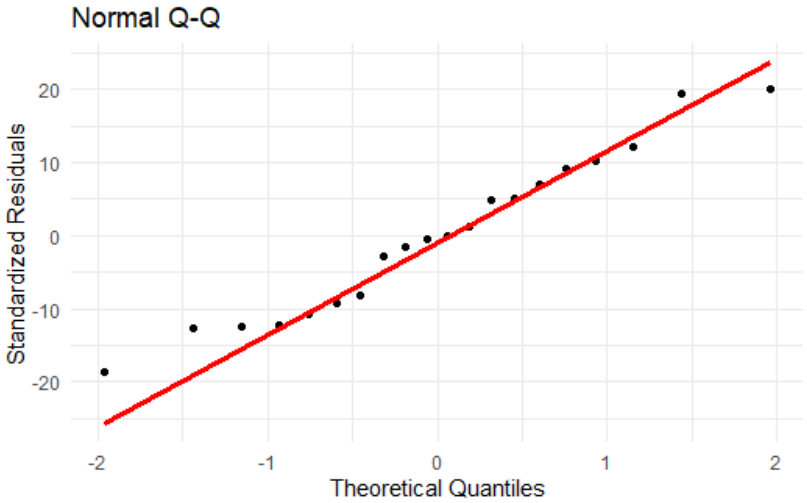

Figure 2.9 Logarithmic Model Q-Q plot

```
> # Shapiro-Wilk test
> shpiro_result <- shapiro.test(residuals(model_log))
> print(shpiro_result)

        Shapiro-Wilk normality test

data:  residuals(model_log)
W = 0.96426, p-value = 0.632
```

Polynomial(2) Model

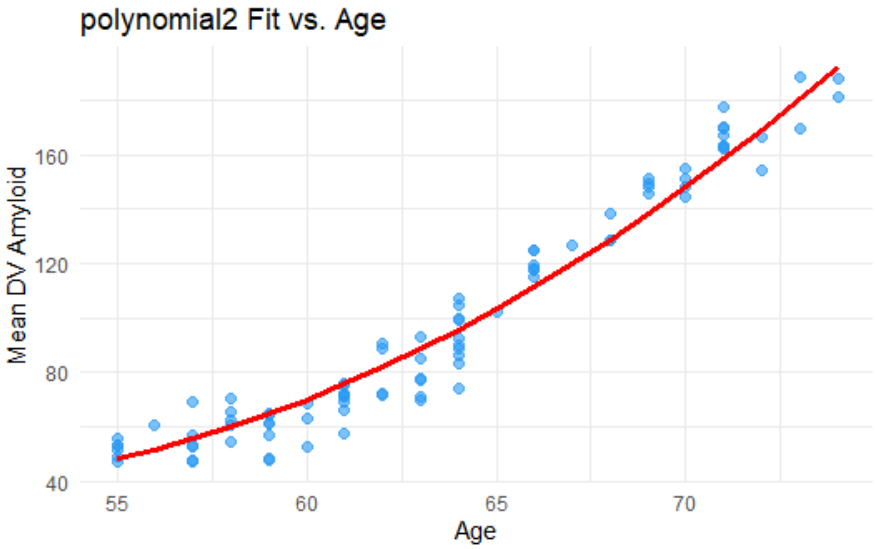

Figure 2.10 Polynomial 2 Model regression

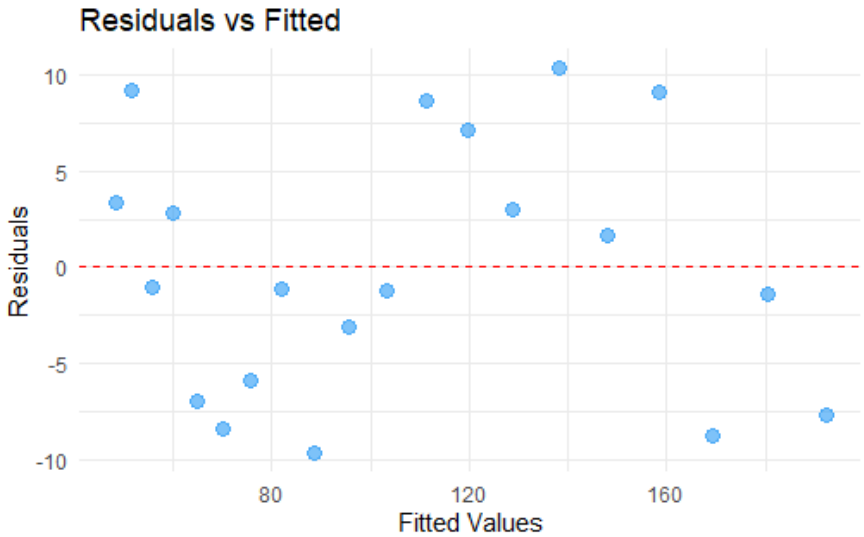

Figure 2.11 Polynomial 2 Model residuals

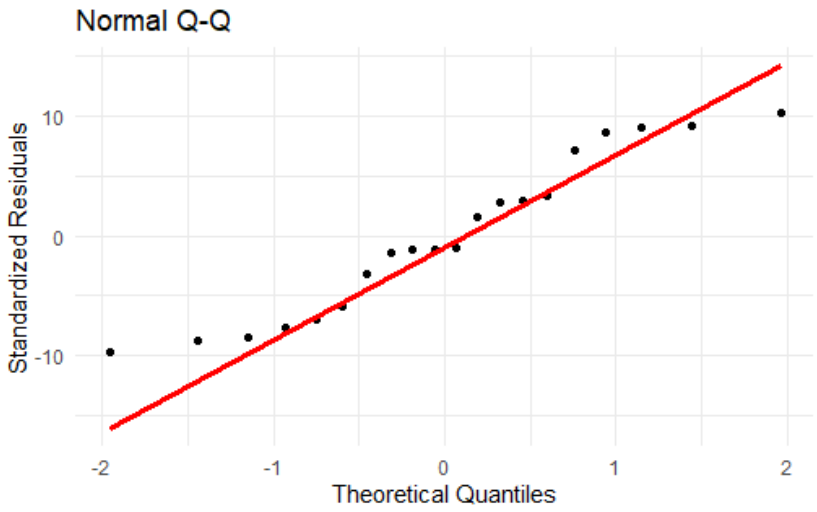

Figure 2.12 Polynomial 2 Model Q-Q plot

```
> shpiro_result <- shapiro.test(residuals(model_poly2))
> print(shpiro_result)

        Shapiro-Wilk normality test

data:  residuals(model_poly2)
W = 0.93027, p-value = 0.1563
```

Mixed-effect Model

In this case, the fixed-effect is the age, and the random effect is the difference among all the participants, which can be caused by weight, diet, education level, average sleeping time, etc. Because there are no independent variables except the age in this dataset, we manually constructed a variable, acting as the random effect.

```
# construct the random variable
data_2$Y <- floor(data_2$X / 17)          # Y = X/17, dividing the dataset into 6 groups.
# use lmer() to construct the model,  age as the fixed effect,  Y as the random effect
model_mixed <- lmer(DV_amyloid ~ age + (1|Y), data = data_2)
summary(model_mixed)
# plot prediction
set.seed(123)  # random seed
new_data$Y <- sample(data_2$Y,20,replace = TRUE)  # randomly select 20 values
new_data$predicted_mix_Amyloid <- predict(model_mixed, newdata = new_data)
```

```
> summary(model_mixed)
Linear mixed model fit by REML ['lmerMod']
Formula: DV_amyloid ~ age + (1 | Y)
   Data: data_2
```

```
REML criterion at convergence: 678.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.26820 -0.68992  0.02908  0.62157  1.89081

Random effects:
 Groups   Name        Variance Std.Dev.
 Y        (Intercept)  4.91    2.216
 Residual             131.90   11.485
Number of obs: 88, groups:  Y, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept) -391.2014    14.9362  -26.19
age            7.7080     0.2339   32.95

Correlation of Fixed Effects:
    (Intr)
age -0.995
```

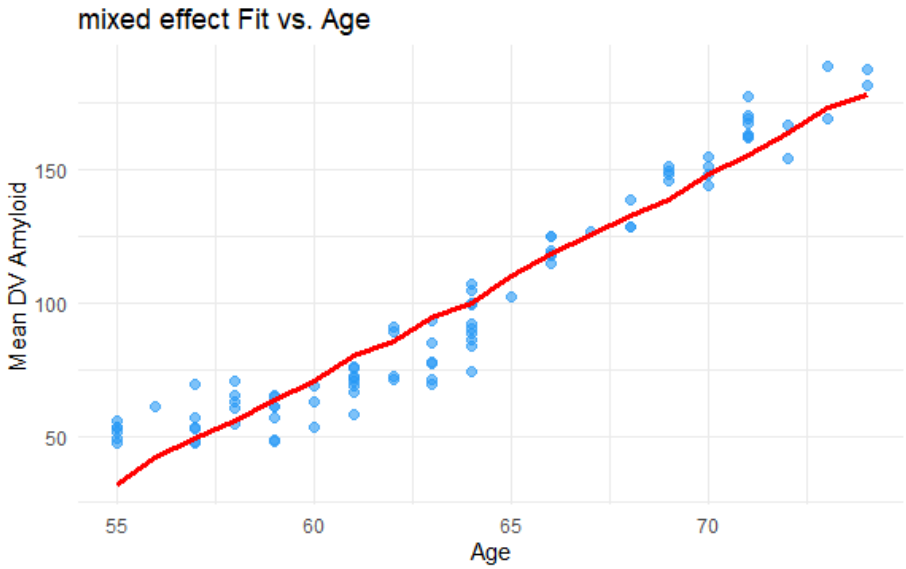

Figure 2.13 Mixed-effect Model regression

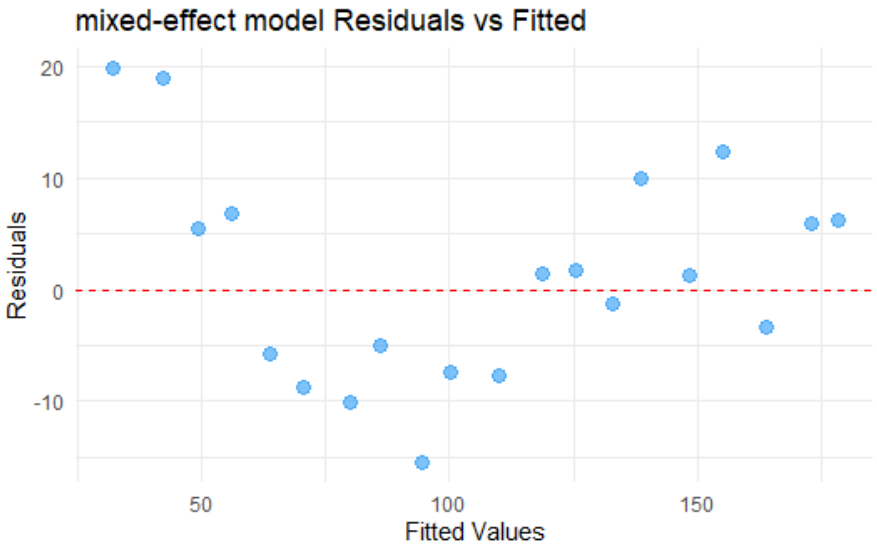

Figure 2.14 Mixed-effect Model residuals

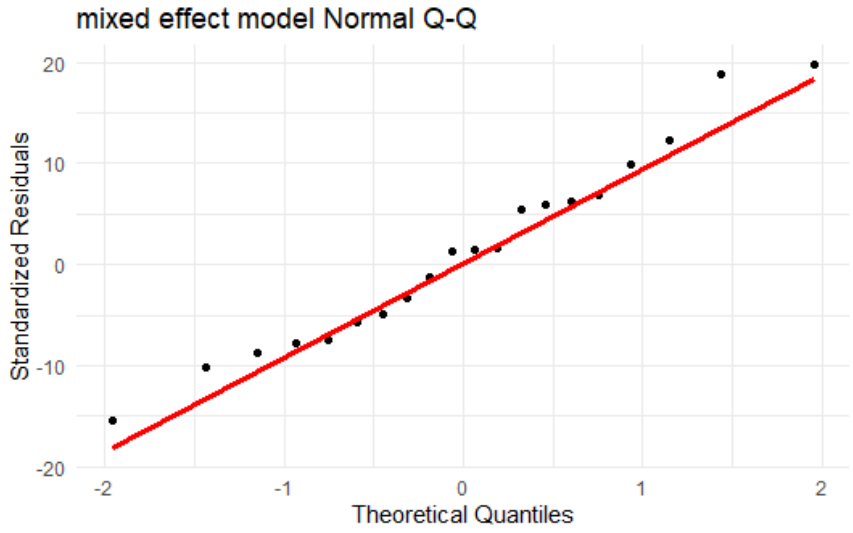

Figure 2.15 Mixed-effect Model Q-Q plot

```
> shpiro_result <- shapiro.test(residuals(model_mixed))
> print(shpiro_result)

        Shapiro-Wilk normality test

data:  residuals(model_mixed)
W = 0.98445, p-value = 0.3756
```

From all the results, we can see that the residuals of the Polynomial 2 Model and the Mixed-effect Model are normally distributed in this dataset.

Prediction

Next, we used the four types of models to predict the amyloid beta at the ages of 75 to 110 years old.

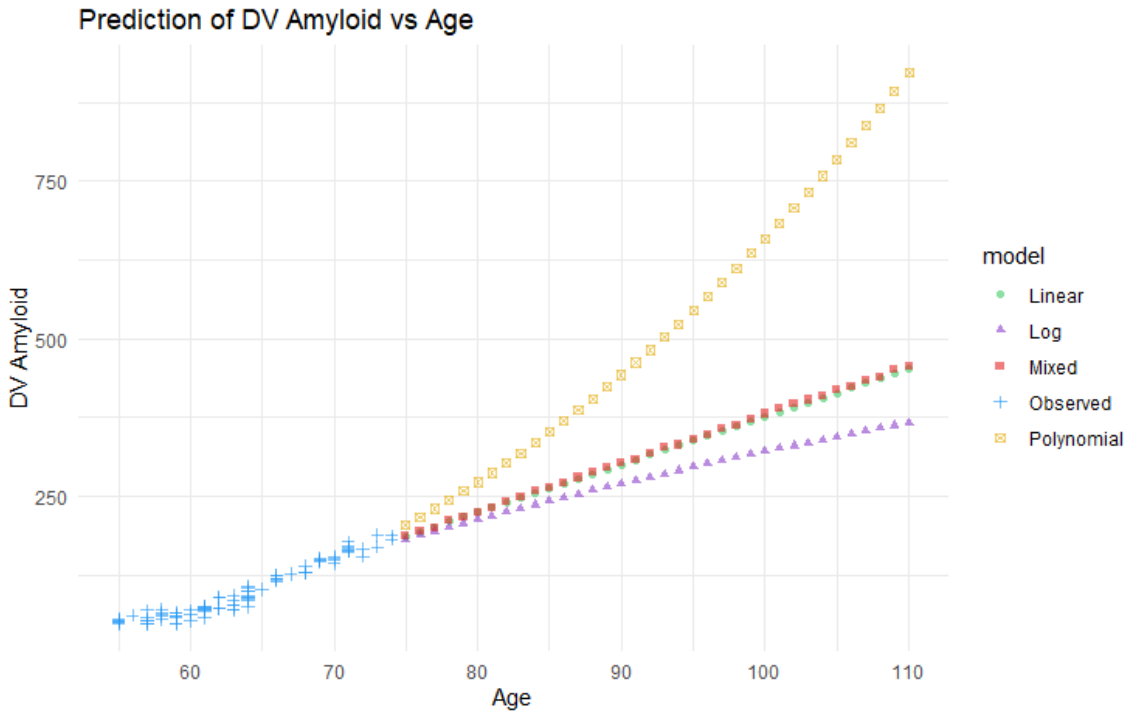

Figure 2.16 Prediction of the amyloid

Recommendations

The four types of model all gave a rising trend of the Amyloid, and the Linear Model and the Mixed-effect Model gave similar results while others' results are significant different. The result of the Polynomial Model seems ridiculous, especially in the highest age region. One of the reasons could be that the prediction age region is very high for the AD patients.

Task 3

Here we drew the comparison of our predictions and the groud truth. It is obvious that the Logarithmic Model has the best result, but still not good enough.

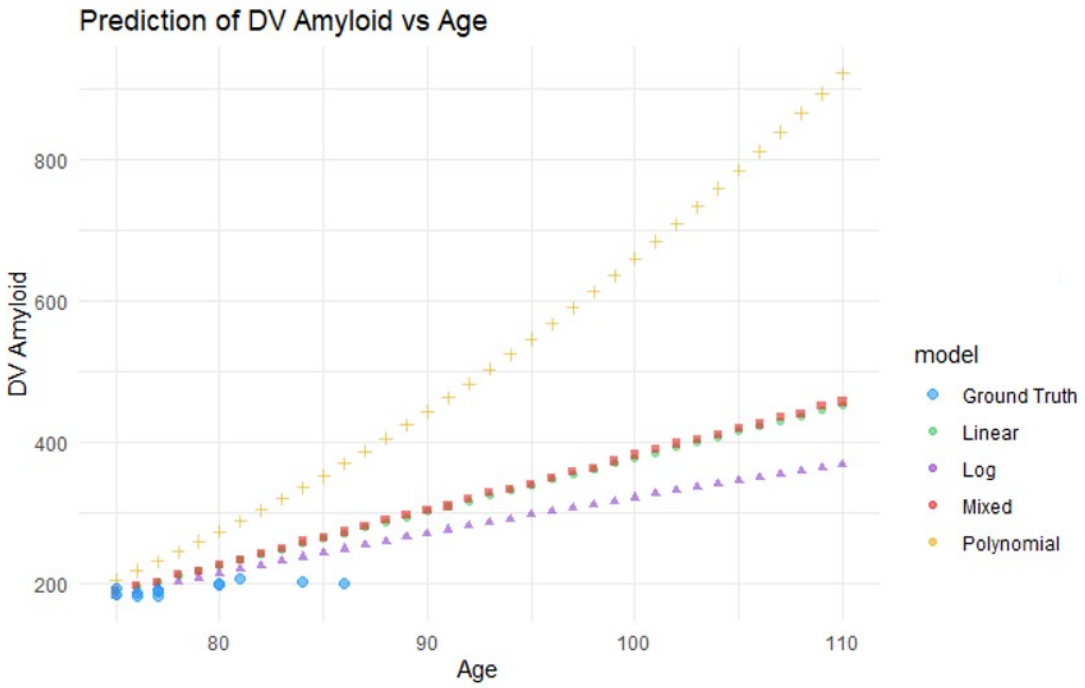

Figure 3.1 Comparison of the predictions and the ground truth

Improved model

To better fit the data of kind like a 'S' shape, we tried the Polynomial 3 Model, the results are as follows.

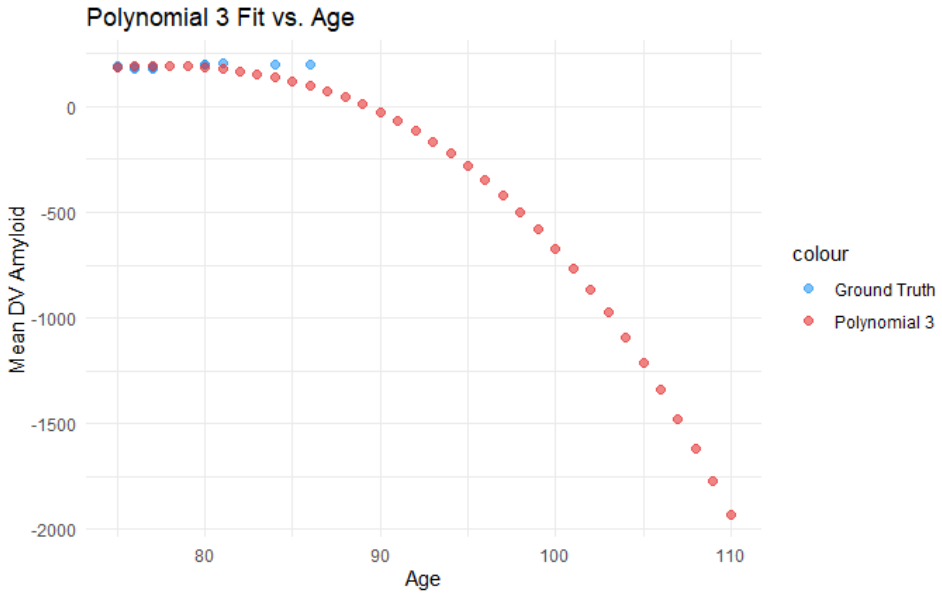

Figure 3.3 Polynomial 3 Model

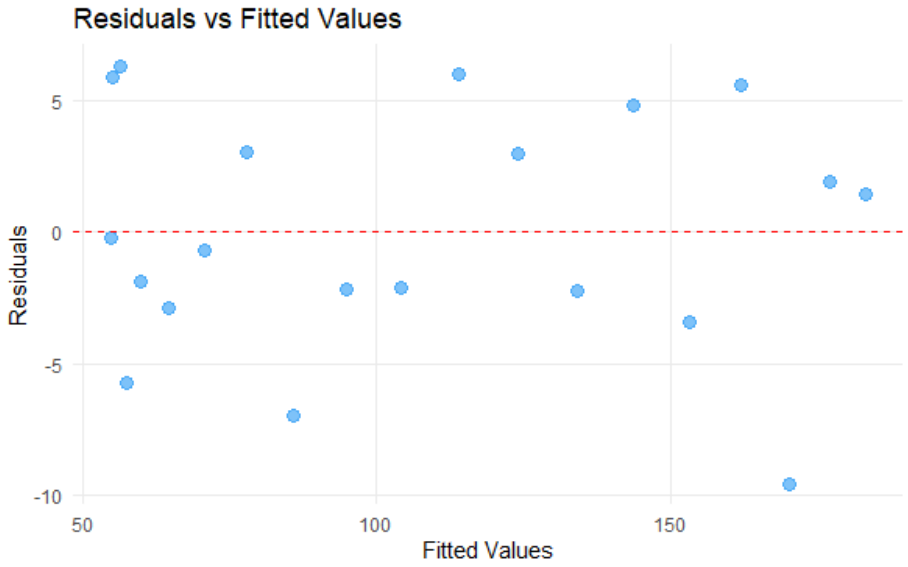

Figure 3.3 Polynomial 3 Model residuals

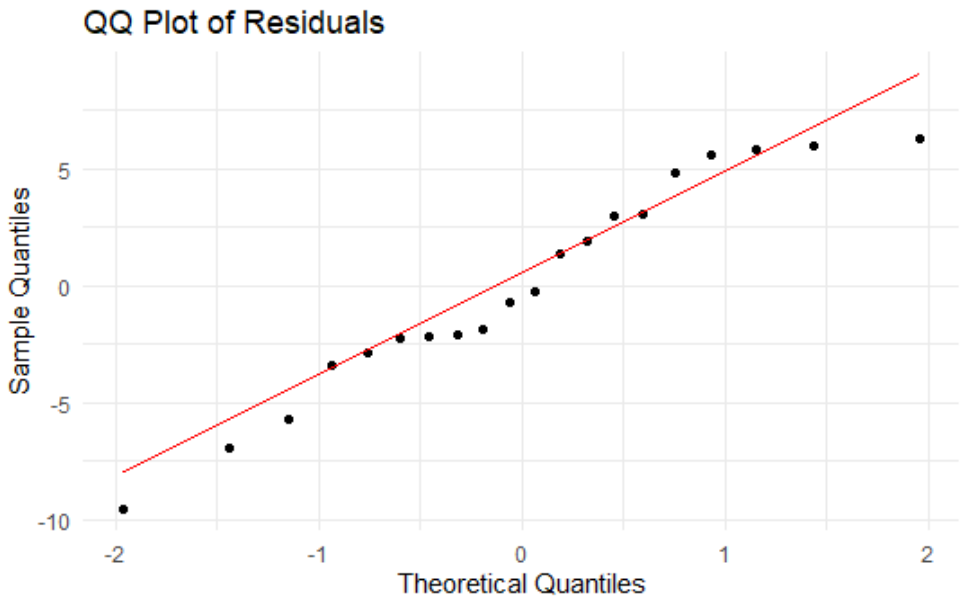

Figure 3.3 Polynomial 3 Model Q-Q plot

```
> shpiro_result <- shapiro.test(residuals(model_poly3))
> print(shpiro_result)

        Shapiro-Wilk normality test

data:  residuals(model_poly3)
W = 0.95035, p-value = 0.3725
```

General learning

When a model performs exceptionally well on the training data, capturing all the details and noise,

it can lead to poor performance on new, unseen data. This happens because the model learns not just the underlying patterns but also the random fluctuations in the training data.
As a result, while the model may have a low error on the training set, it may have a high error when making predictions on new data.

Task 4

From the Shapiro-Wilk test and the Q-Q plot, we can see that the control group is not normally distributed.

```
> print(shapiro_test_control)
        Shapiro-Wilk normality test
data:  data$AMYLOIDB[data$GROUP == "control"]
W = 0.94804, p-value = 0.02836
> print(shapiro_test_gene_exp1)
        Shapiro-Wilk normality test
data:  data$AMYLOIDB[data$GROUP == "gene_exp1"]
W = 0.98604, p-value = 0.8154
> print(shapiro_test_gene_exp2)
        Shapiro-Wilk normality test
data:  data$AMYLOIDB[data$GROUP == "gene_exp2"]
W = 0.98033, p-value = 0.5661
```

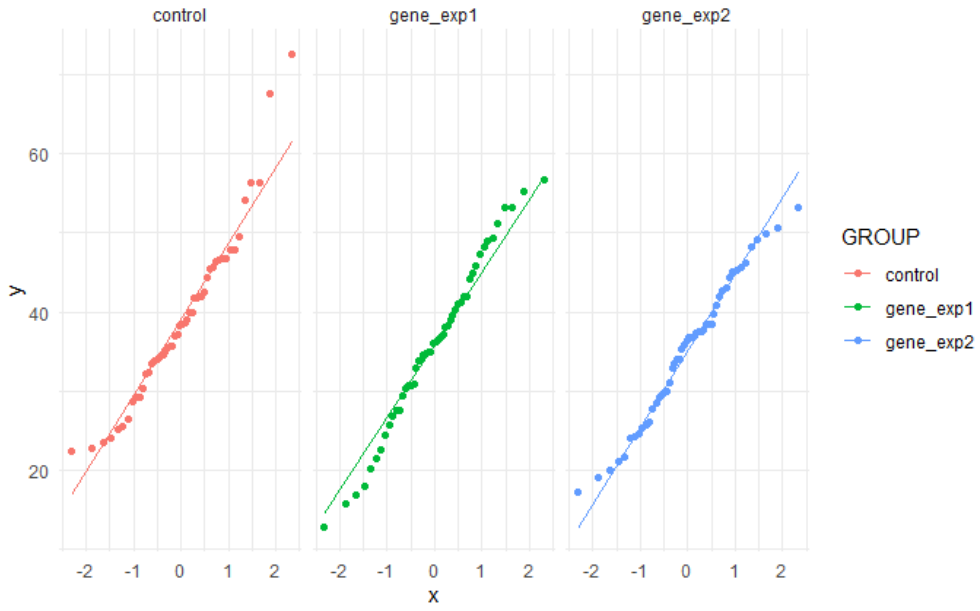

Figure 4.1 three groups Q-Q plot

If one group is not normally distributed, we cannot do the ANOVA test. So we did a root transformation to the data, the results are as follows.

```
> print(shapiro_test_log_control)
        Shapiro-Wilk normality test
data:  data$root_AMYLOIDB[data$GROUP == "control"]
W = 0.97362, p-value = 0.3228
```

```
> print(shapiro_test_log_gene_exp1)
        Shapiro-Wilk normality test
data:  data$root_AMYLOIDB[data$GROUP == "gene_exp1"]
W = 0.97341, p-value = 0.3169
> print(shapiro_test_log_gene_exp2)
        Shapiro-Wilk normality test
data:  data$root_AMYLOIDB[data$GROUP == "gene_exp2"]
W = 0.97474, p-value = 0.3567
```

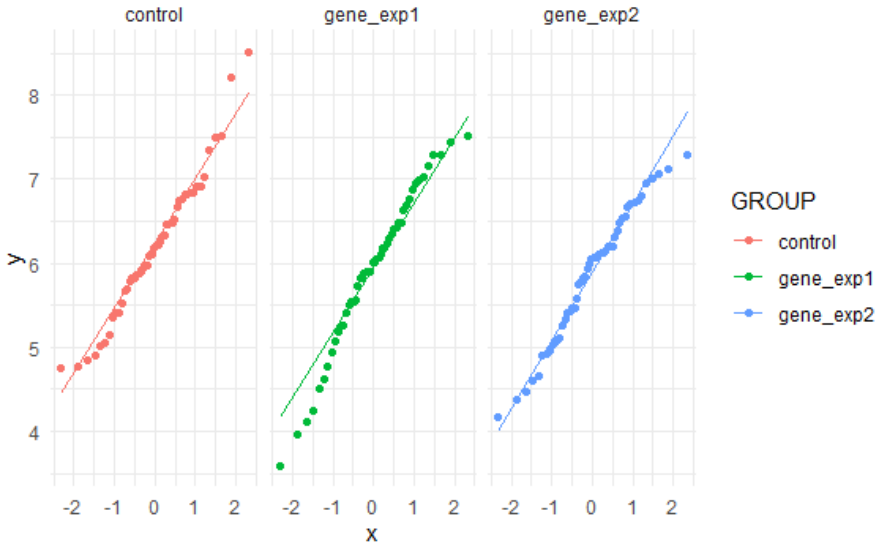

Figure 4.2 three groups Q-Q plot after root transformation

After that, we did the Levene's test to check the homogeneity of variance. Because the p value is bigger than 0.05, the root data is homogeneous in variance.

```
> print(levene_test)
Levene's Test for Homogeneity of Variance (center = median)
       Df   F value   Pr(>F)
group   2   0.3898   0.6779
      147
```

Finally, we did the ANOVA test.

```
> anova_result <- aov(root_AMYLOIDB ~ GROUP, data = data)
> summary(anova_result)
             Df Sum Sq Mean Sq F value Pr(>F)
GROUP         2   3.07  1.5364   2.123  0.123
Residuals   147 106.38  0.7237
```

Because the p value is bigger than 0.05, we concluded that the three groups do not have significant differences.