**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.

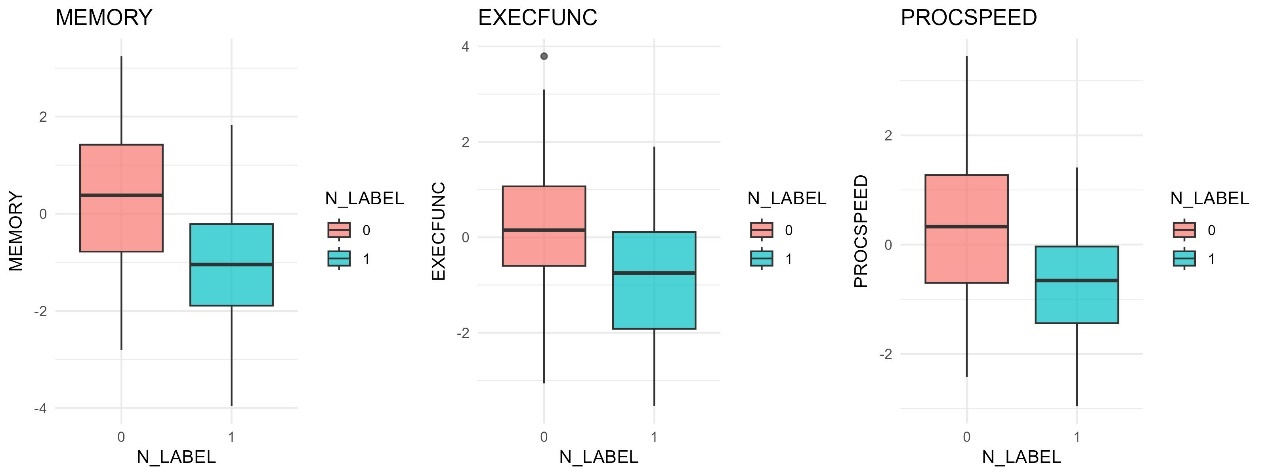


Figure 1.1 Box plots of scores

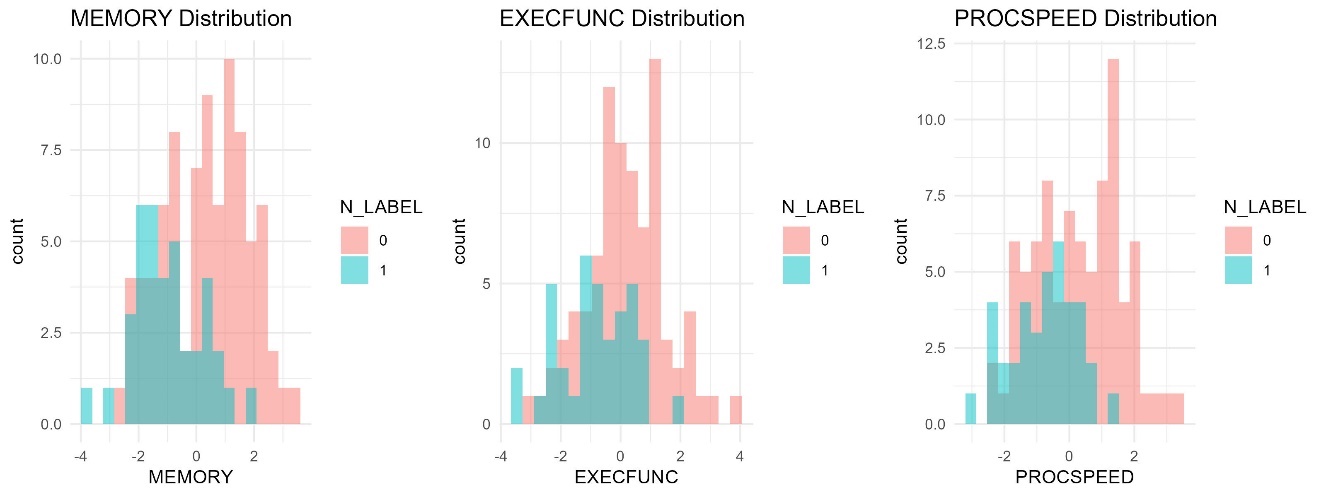


Figure 1.2 Histograms of scores

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.

Table 1.1 Fiser test results

|  |  |  |
| --- | --- | --- |
| variable | p-value | odds ratio |
| AMYLOIDVIS | 0.297 | 1.541 |
| INFARCTION | 0.170 | 0.227 |
| ATRIALFIBR | 0.002 | 4.997 |
| DIABETES | 1.000 | 1.107 |
| HYPERTENSION | 0.125 | 0.504 |
| SEX | 0.441 | 1.371 |

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.

Table1.2 Results of Wilcoxon rank-sum test

|  |  |
| --- | --- |
| Variable | p-value |
| AGE | 1.938e-05 |
| EDUCATION | 1.856e-01 |
| MMSE | 1.163e-12 |
| CERAD | 5.056e-14 |
| PASTCAQ | 8.234e-01 |
| HANDGRIP | 1.123e-04 |
| MUSCLE | 5.139e-01 |
| SPPB | 7.370e-02 |
| MNA | 4.661e-04 |
| BMI | 5.860e-01 |
| CRP | 5.973e-02 |
| LEUK | 3.995e-02 |
| BPSYS | 5.394e-01 |
| BPDIA | 5.213e-01 |
| HBA1C | 1.624e-01 |
| CHOLESTEROL | 2.615e-01 |
| WMHVOL | 1.560e-02 |
| HIPPOVOL | 5.087e-06 |
| AMYLOIDBIND | 3.661e-04 |

Meanwhile, in order to maintain the accuracy of the test, we performed multiple testing corrections using Bonferroni and BH method. Figure 1.3 shows the results, where the first column represents the p-value before adjustment, the second column shows the p-value after Bonferroni adjustment, and the third column shows the p-value after the BH adjustment.

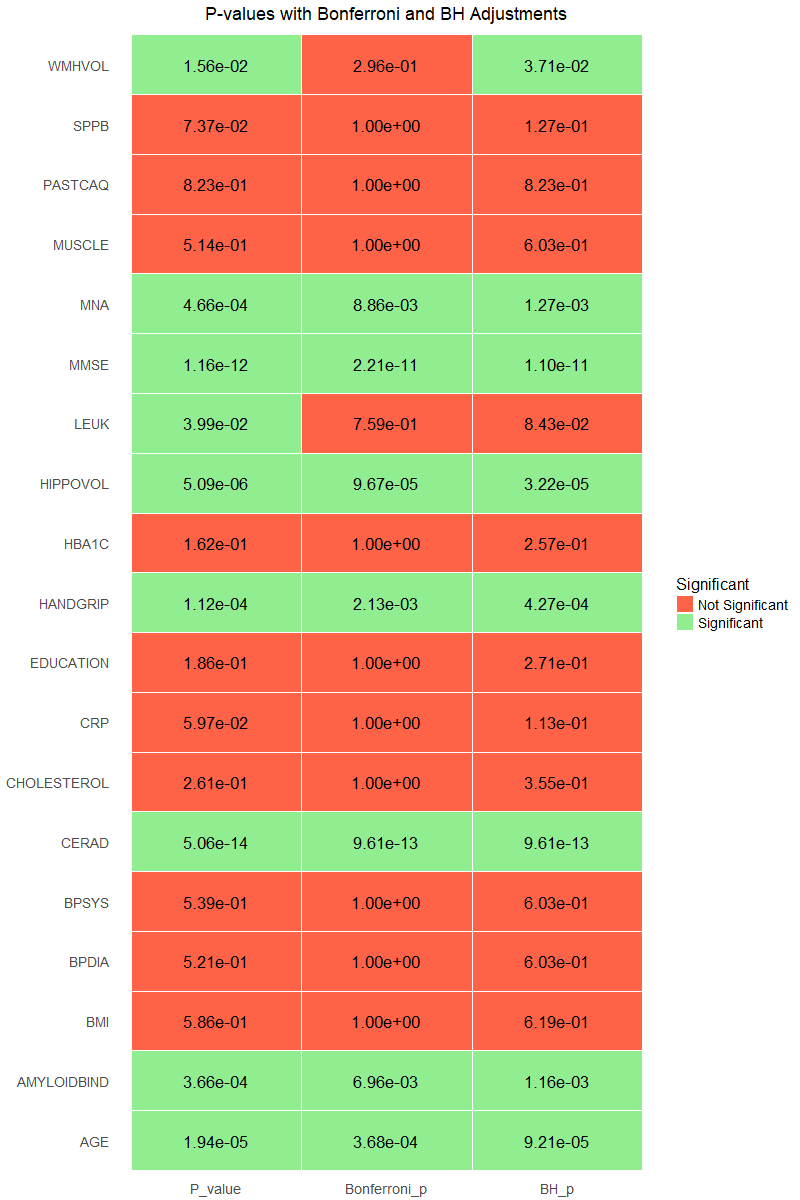


Figure1.3 Result of multiple test correction

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.

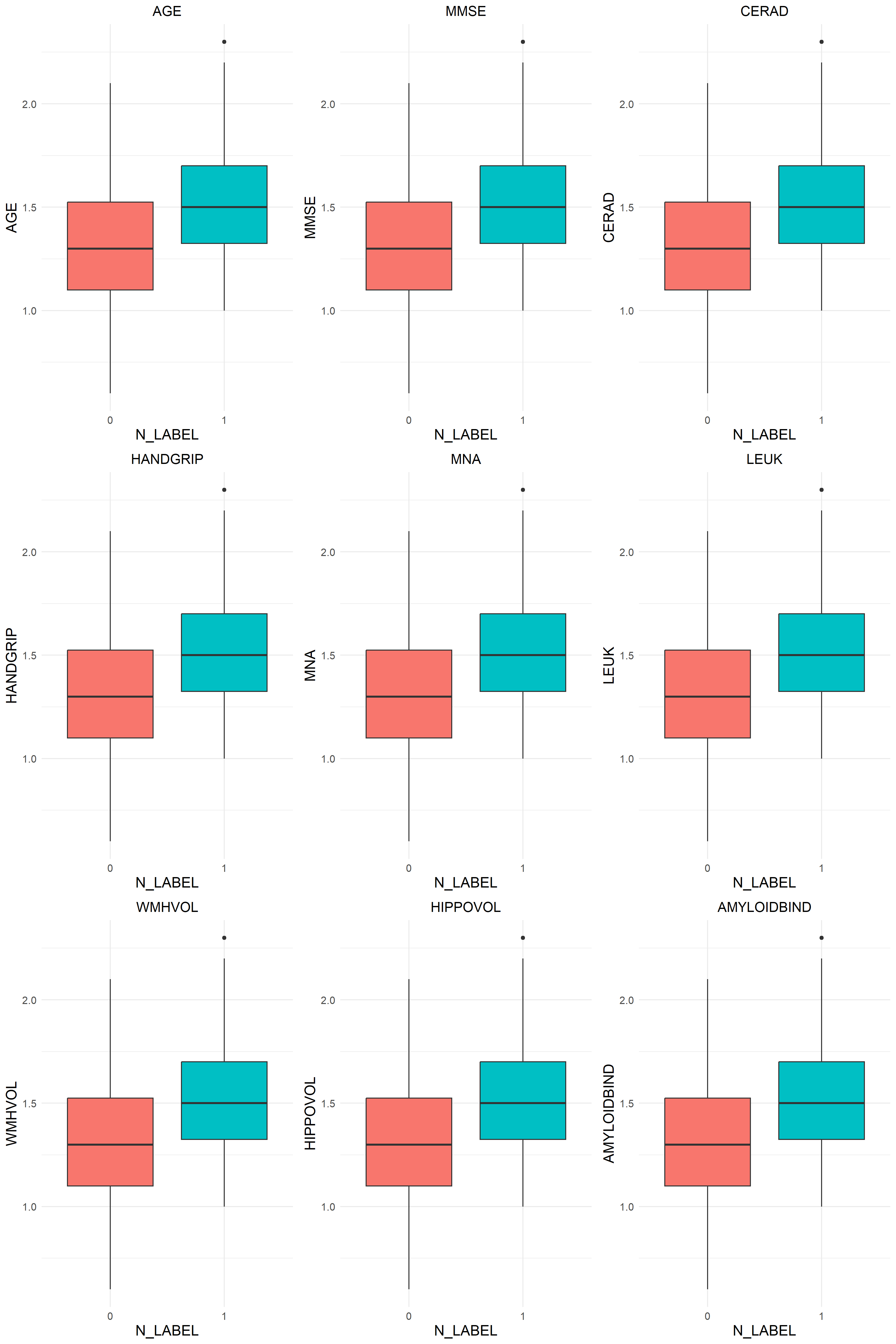


Figure1.4 Box plots of significant variables

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.

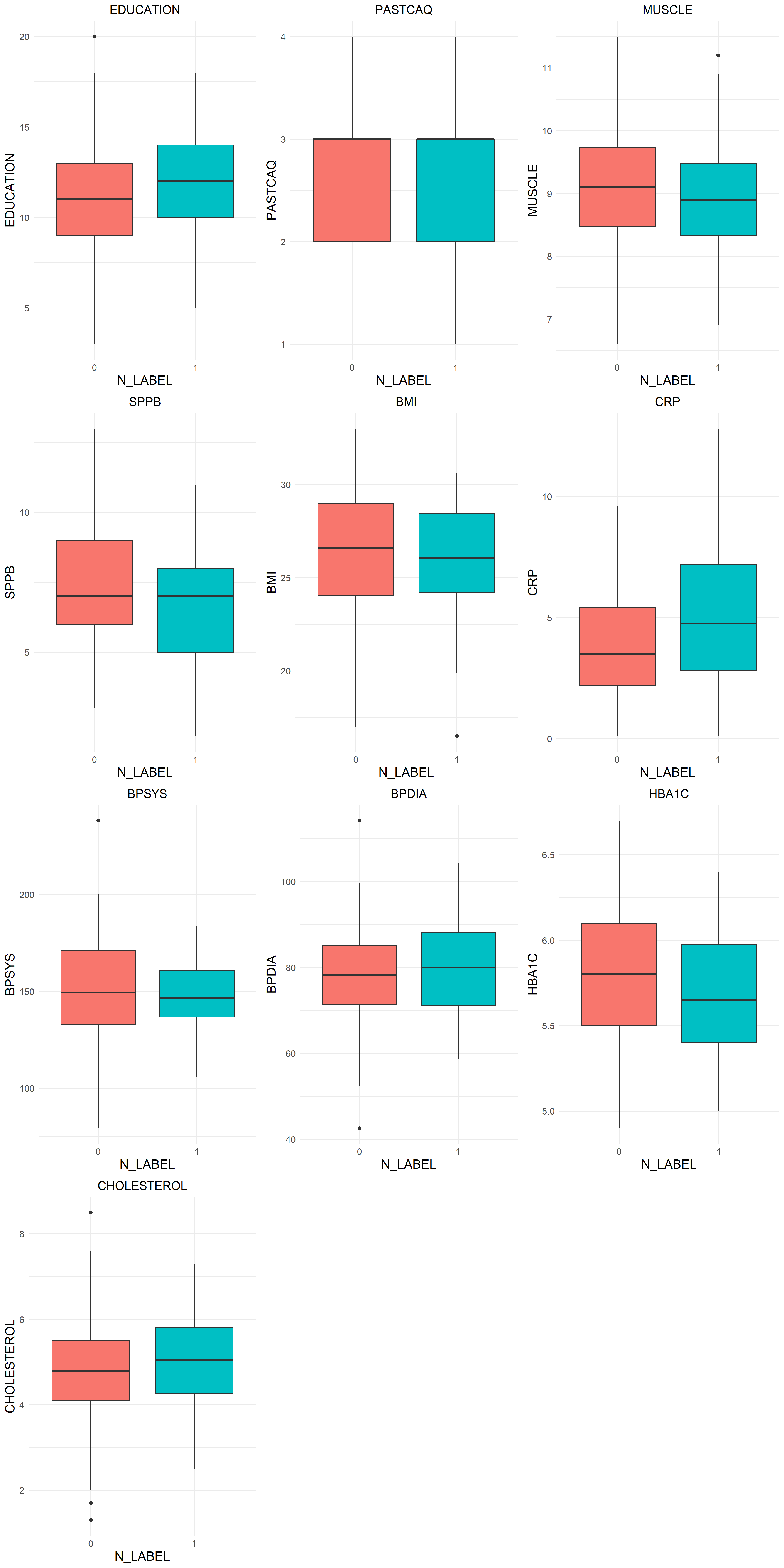


Figure1.4 Box plots of insignificant variables

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.

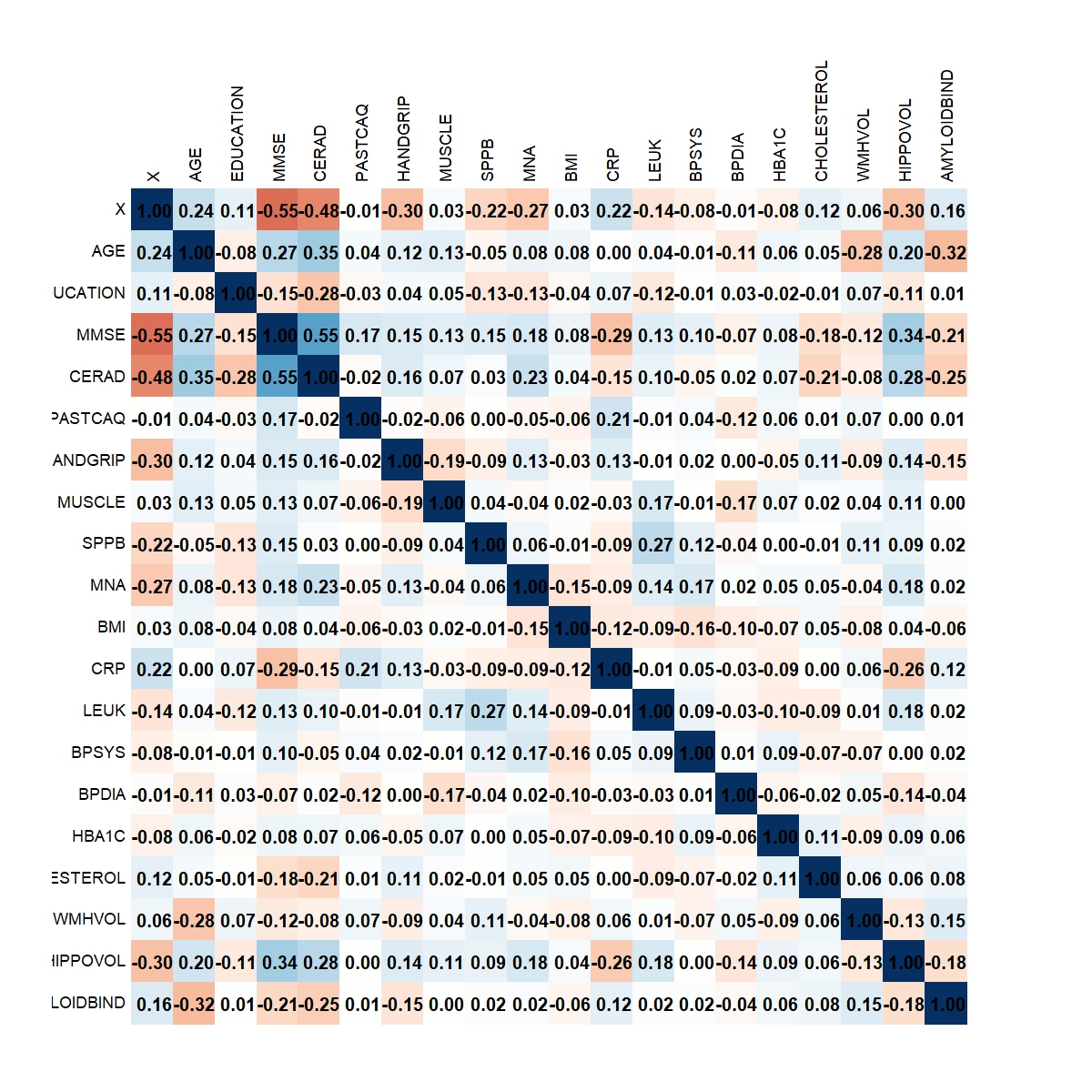


Figure 1.4 Correlation matrix of different variables

Given the weak correlations among the variables and the high dimensionality of the dataset, we applied the Akaike Information Criterion (AIC) to simplify the model and achieve a more parsimonious solution. As an example, we used AIC to refine the memory model. The step() function was employed for stepwise regression, allowing the automatic selection of variables to be included or removed from the model. We specified the direction as “both,” enabling the function to add or remove variables based on their contribution to model fit.

# 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)

We took the intersection of the significant variables found by AIC and the variables identified for further attention in the previous discussion, and then formed a new linear model. We then performed 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 R² 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.

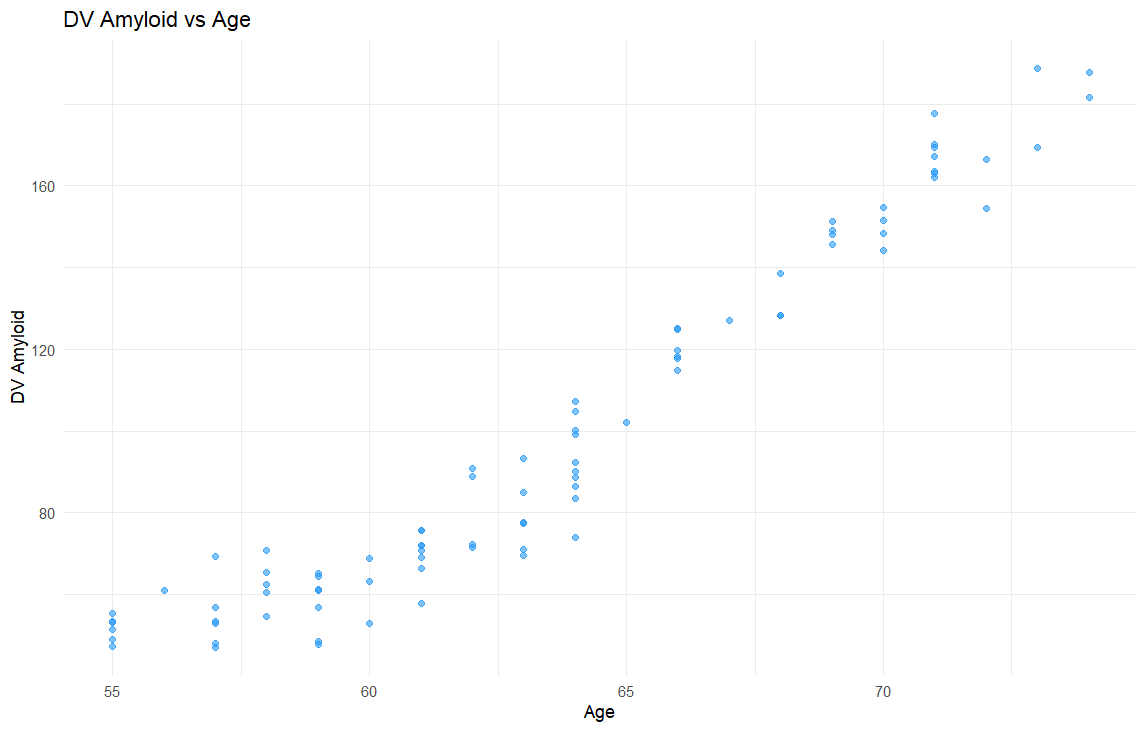


Figure 2.1 Scatter plot of the original dataset

The trend observed in Figure 2.1 appears to be linear; therefore, we initially employed a linear model for our analysis. The Adjusted R-squared of the linear model is 0.9253 and the regression result is shown in figure 2.2.

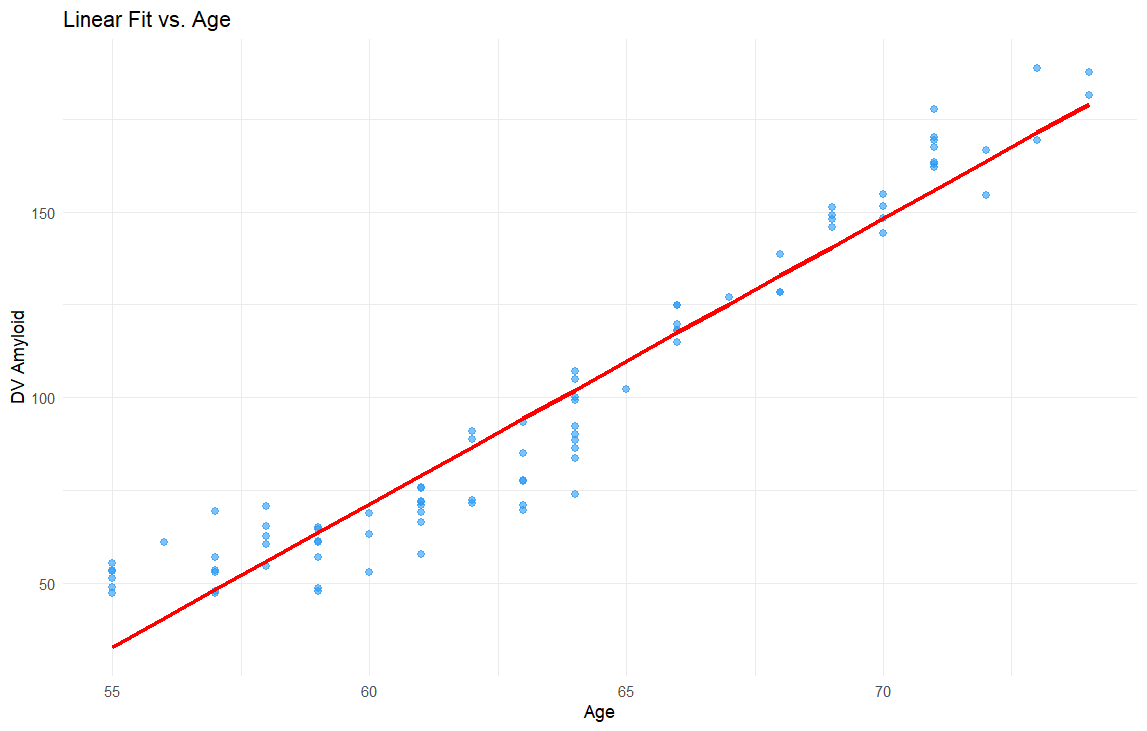


Figure 2.2 regression line & original data

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

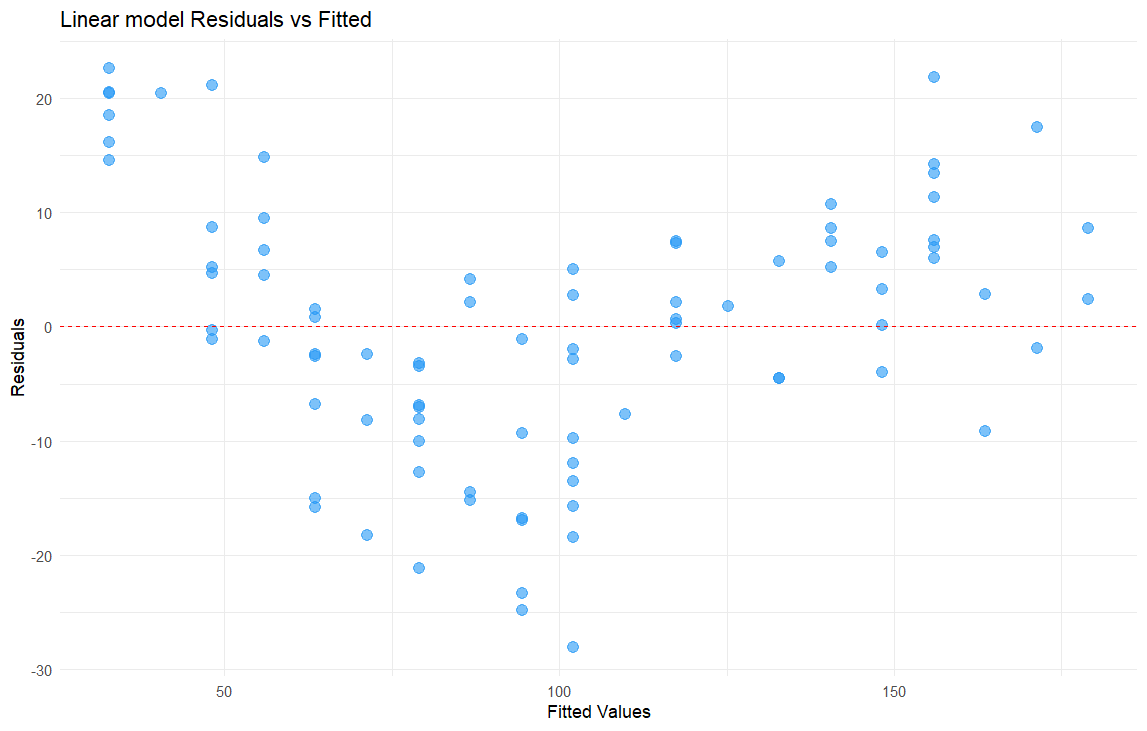


Figure 2.3 residuals value vs fitted value

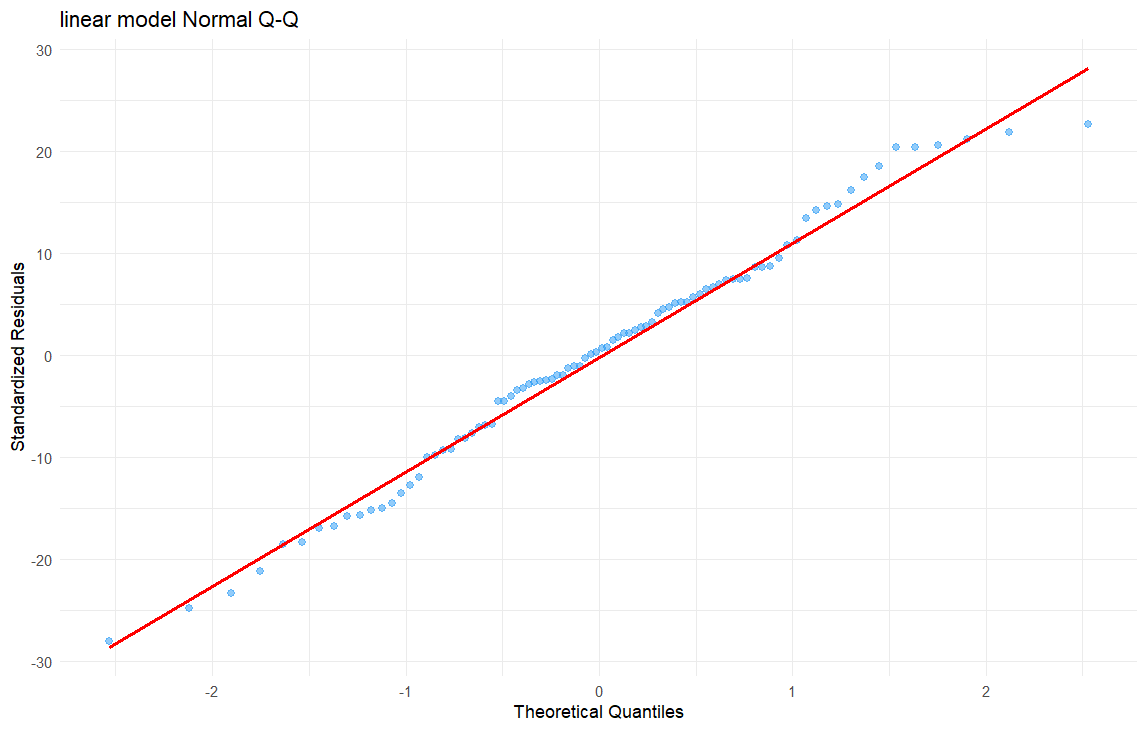


Figure 2.4 Q-Q plot

> shpiro\_result <- shapiro.test(merged\_data$residual\_linear)

> print(shpiro\_result)

Shapiro-Wilk normality test

data: merged\_data$residual\_linear

W = 0.98619, p-value = 0.4782

The Adjusted R-squared value of 0.9253 suggests that the linear model captured a substantial portion of the data's variability. Furthermore, the Q-Q plot and the results of the Shapiro-Wilk test indicate that the residuals of the linear model appear to be approximately normally distributed, which corroborates the validity of model assumptions.

Then, we tried Logarithmic Model, Polynomial Model (2). The workflow is just as same as it in the Linear Model and the results are as follows.

**Logarithmic Model**

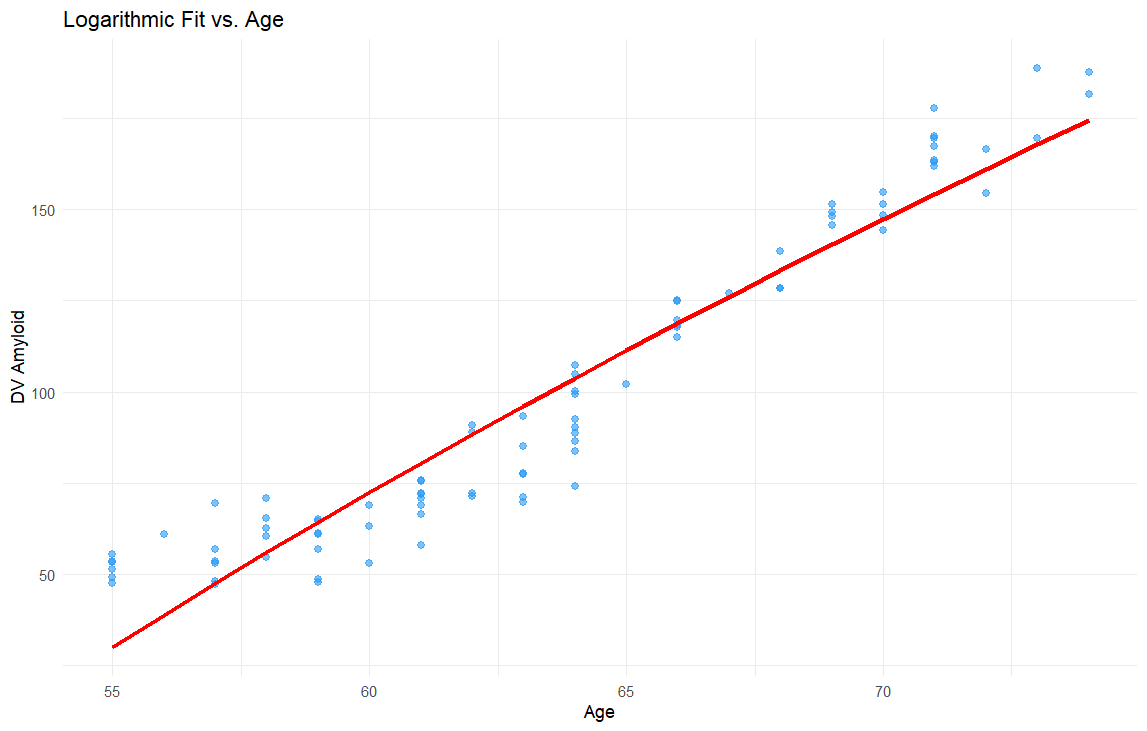


Figure 2.5 Logarithmic Model regression

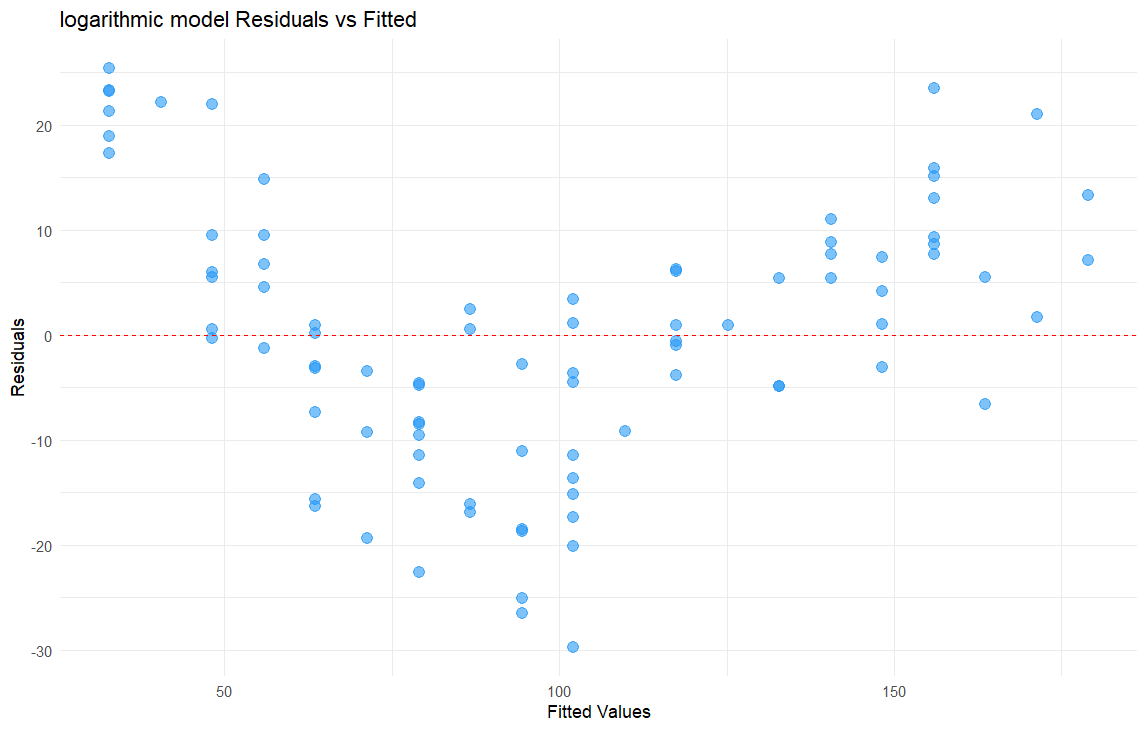


Figure 2.6 Logarithmic Model residuals

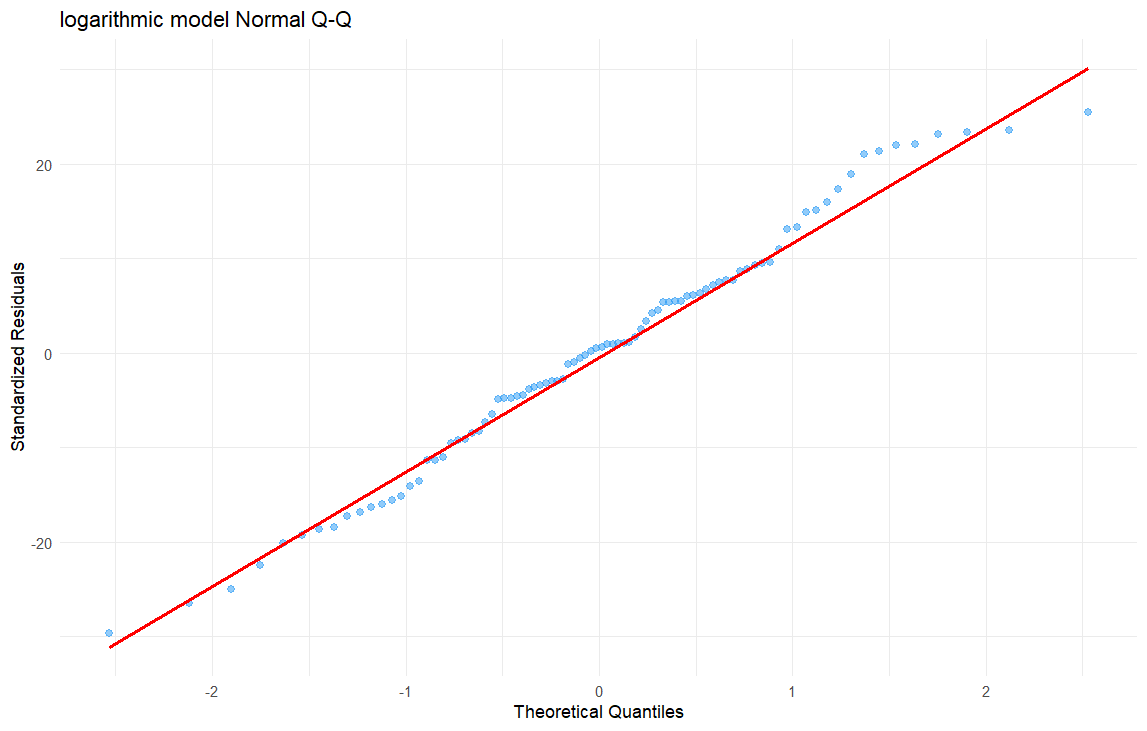


Figure 2.7 Logarithmic Model Q-Q plot

> # Shapiro-Wilk test

> shpiro\_result <- shapiro.test(merged\_data$residual\_log)

> print(shpiro\_result)

Shapiro-Wilk normality test

data: merged\_data$residual\_log

W = 0.98566, p-value = 0.4448

**Polynomial(2) Model**

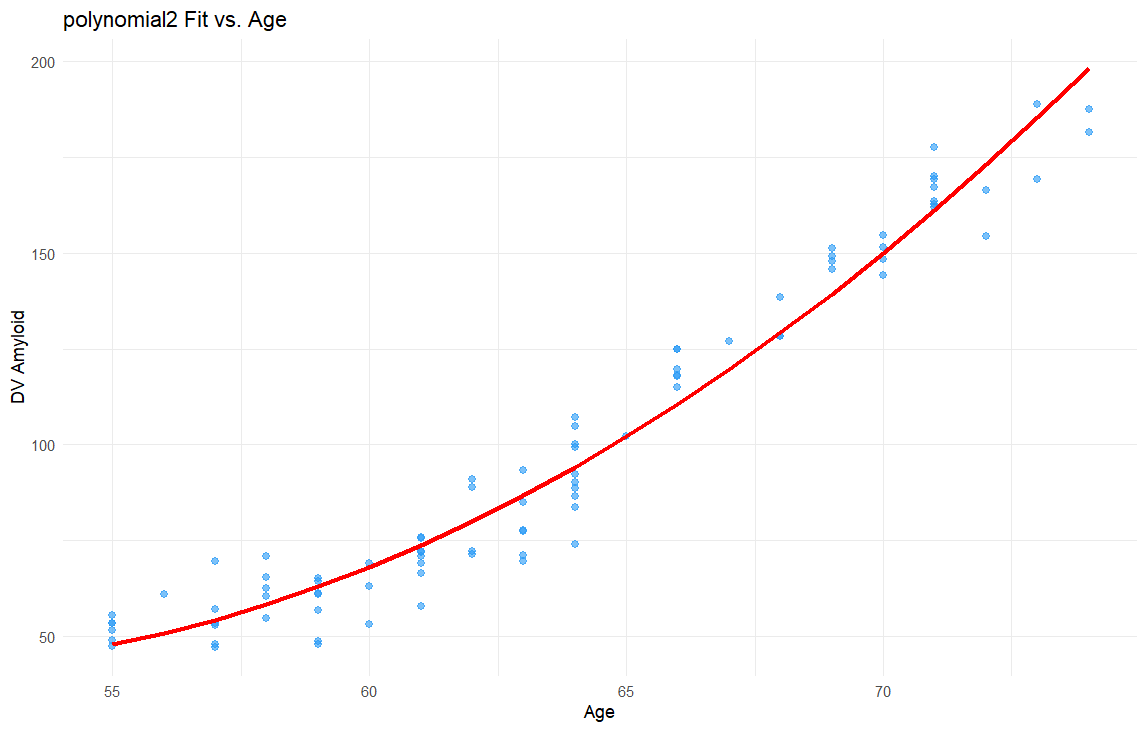


Figure 2.8 Polynomial 2 Model regression

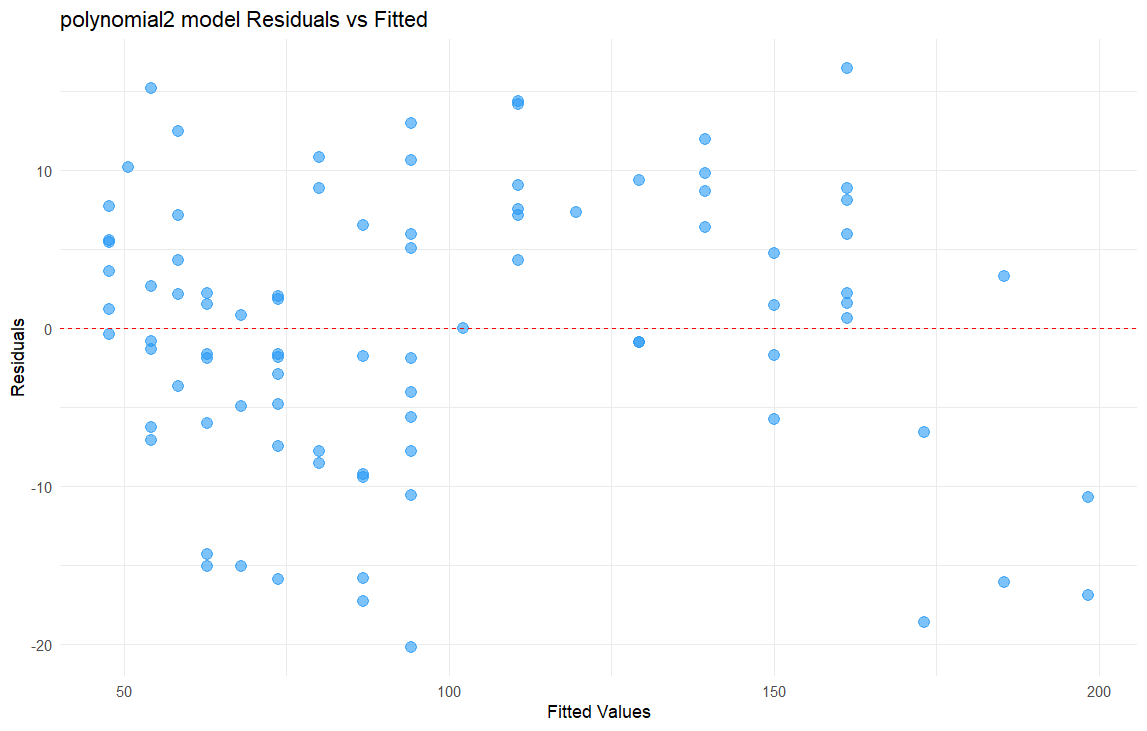


Figure 2.9 Polynomial 2 Model residuals

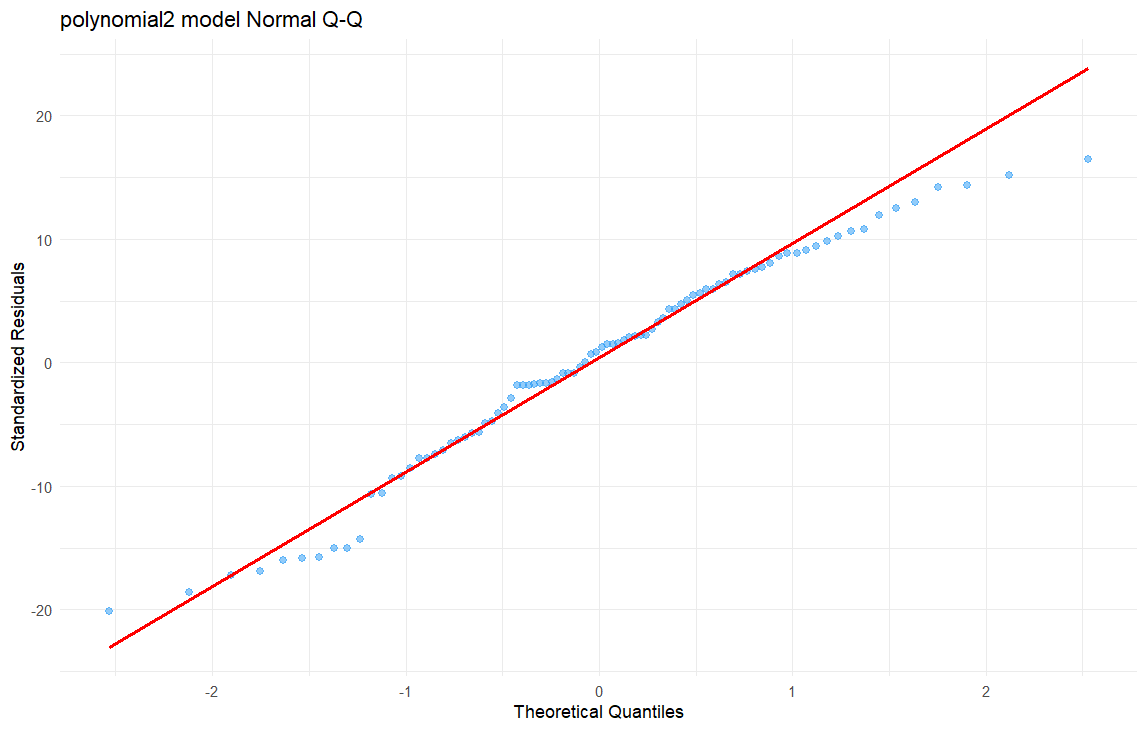


Figure 2.10 Polynomial 2 Model Q-Q plot

> # Shapiro-Wilk test

> shpiro\_result <- shapiro.test(merged\_data$residual\_poly2)

> print(shpiro\_result)

Shapiro-Wilk normality test

data: merged\_data$residual\_poly2

W = 0.97372, p-value = 0.07047

Table 2.1 Comparison of the Adjusted R-Squared value

|  |  |
| --- | --- |
| Model | Adjusted R-Squared |
| Linear | 0.9253 |
| Logarithmic | 0.9093 |
| Poly2 | 0.9570 |

From all the results, we can see that the residuals of all the 3 models are normally distributed in this dataset. And the Polynomial 2 model has the highest Adjusted R-Squared value.

**Prediction**

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

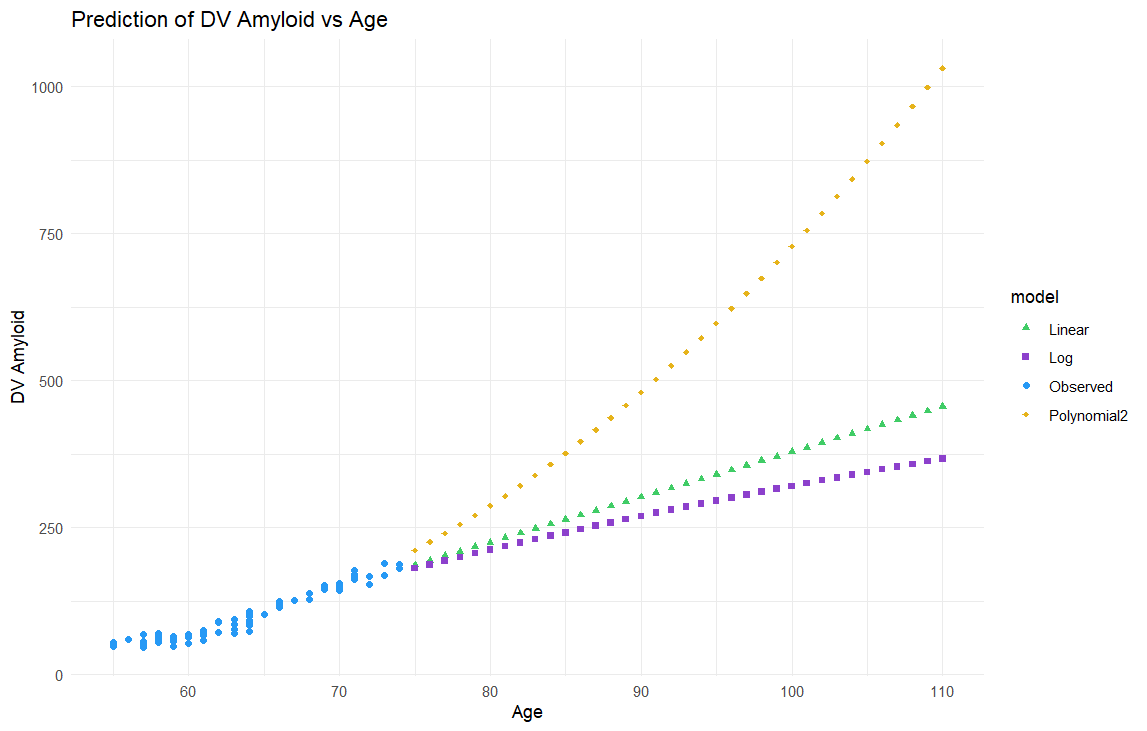


Figure 2.11 Prediction of the amyloid

**Recommendations**

The three types of model all gave a rising trend of the Amyloid. 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**

Figure 3.1 shows the comparison of our predictions and the groud truth and table 3.1 provides the exact value of the Mean Squared Error (MSE) of each models. It is obvious that the Logarithmic Model has the best result.

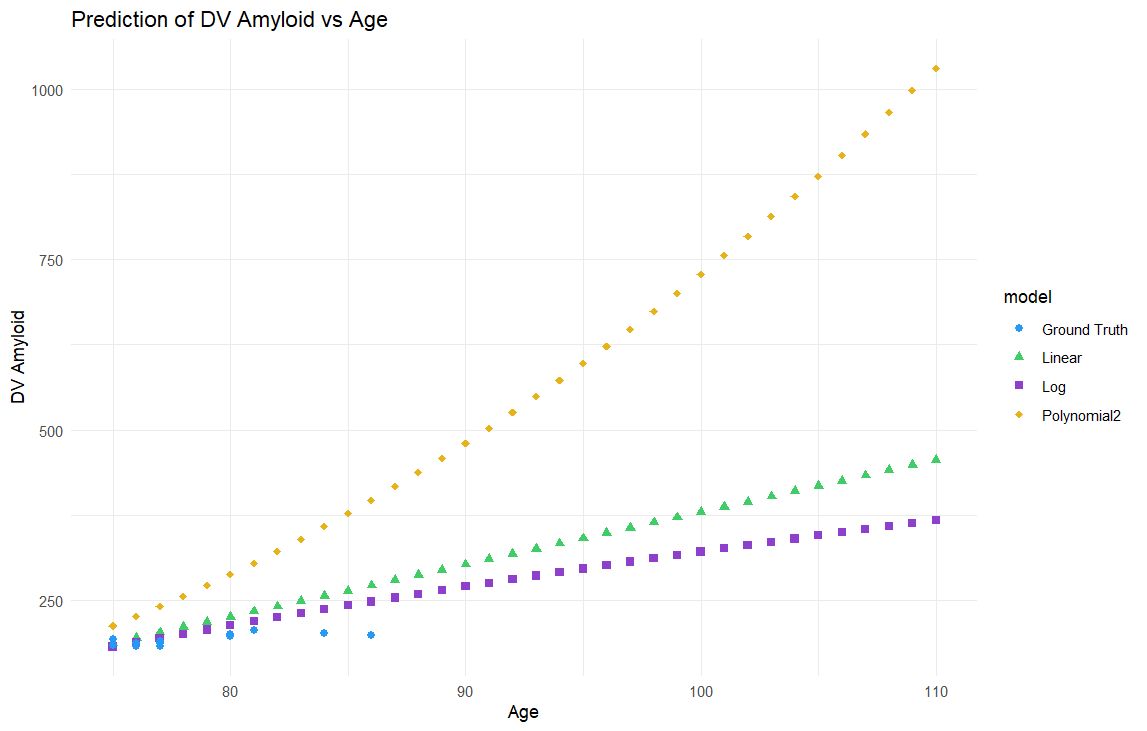


Figure 3.1 Comparison of the predictions and the ground truth

Table 3.1 comparison of the MSE value

|  |  |
| --- | --- |
| Model | MSE |
| Linear | 974.83 |
| Logarithmic | 381.30 |
| Poly2 | 8617.83 |

**Improved model**

To better fit the data of kind like a ‘S’ shape, we tried the Polynomial 3 model, the results are as follows. The MSE value of the Polynomial 3 model is 1020.13, which is worse than the Linear model and the Logarithmic model. The Adjusted R-Squared value of the Polynomial 3 model is 0.967, and figure 3.3 shows it fits the train data very well. So, the Polynomial 3 model is kind of overfitted. On the other hand, Figure 3.2 shows that the output of the Polynomial 3 model has negative values, but the concentration cannot be negative. We think this is because the age range is to high for AD patient, causing a ridiculous result.

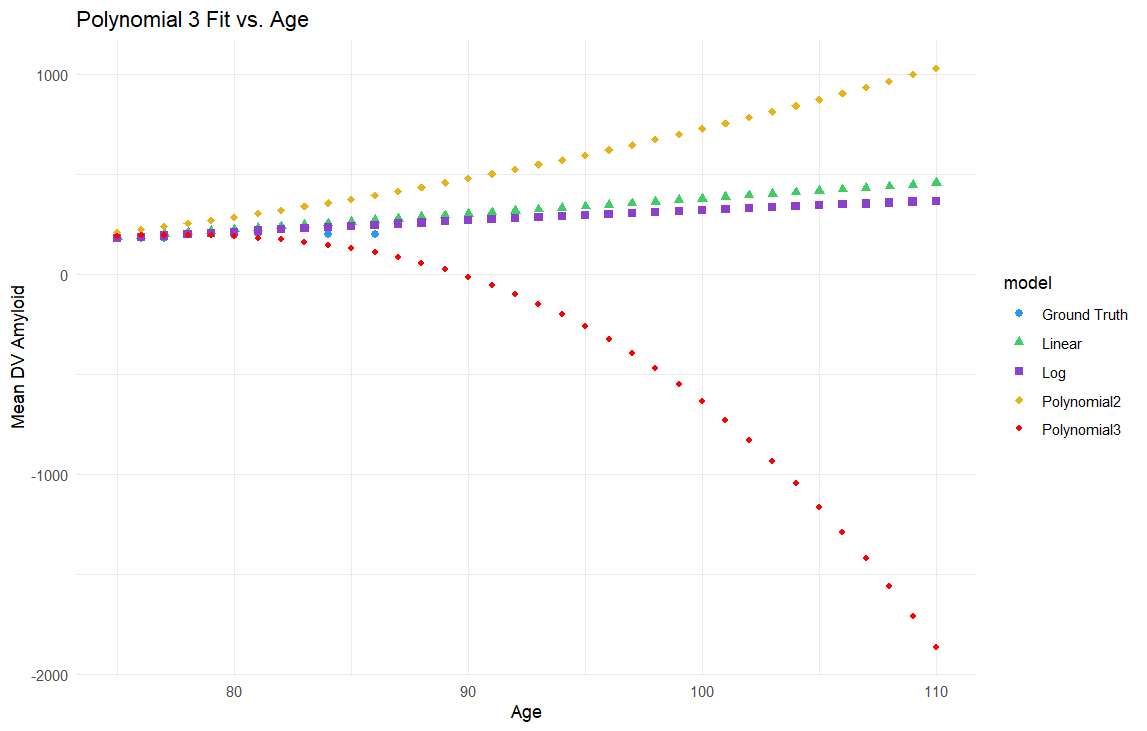


Figure 3.2 Prediction results

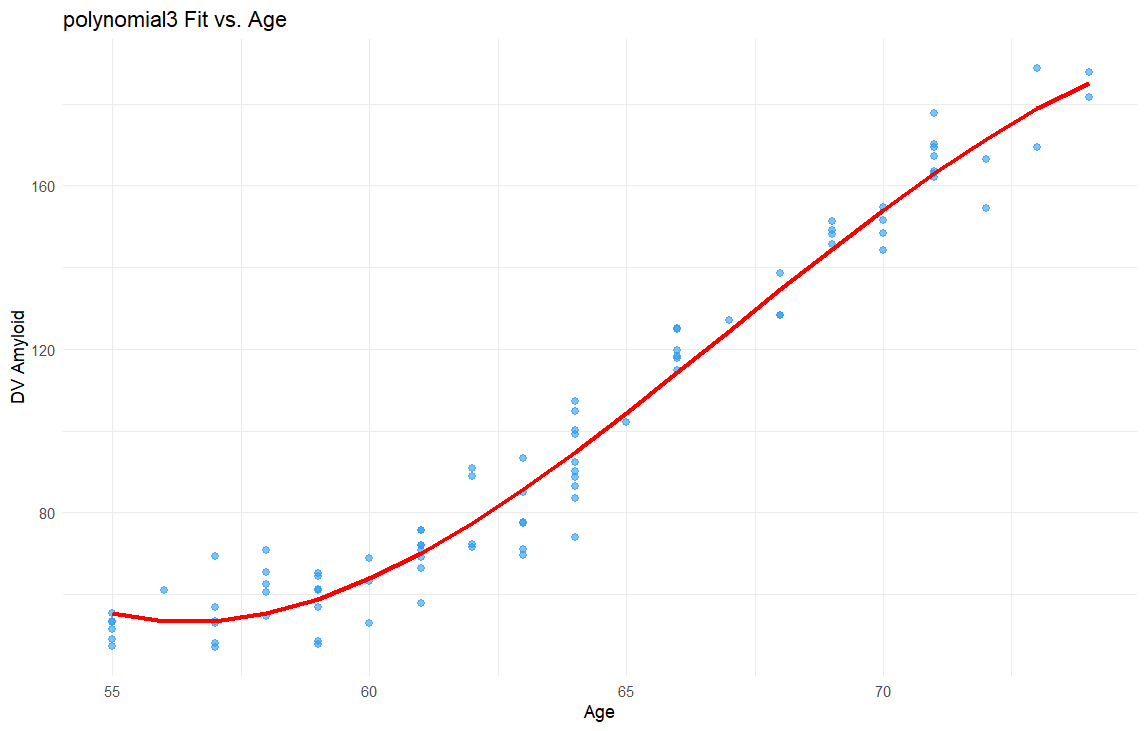


Figure 3.3 Polynomial 3 Model regression

**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

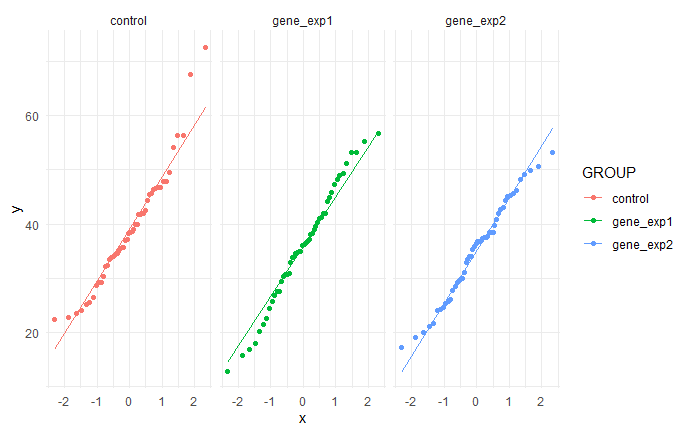


Figure 4.1 three groups Q-Q plot

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

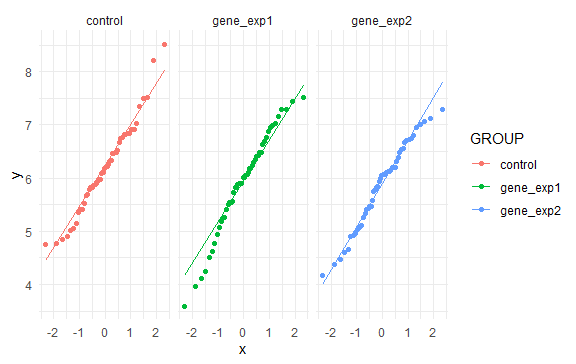


Figure 4.2 three groups Q-Q plot after root transformation

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.