**Seminar 2**

**Group A2**

**Yufeng Deng Yuanqing Wang**

**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.

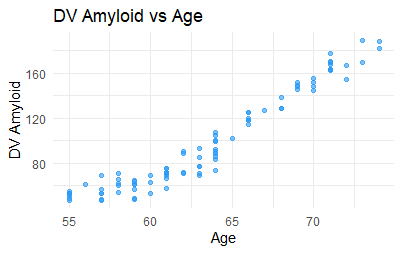


Figure 2.1 Scatter plot of the original dataset

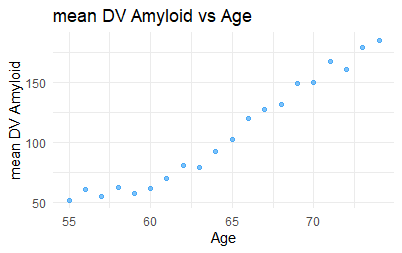


Figure 2.2 mean DV\_Amyloid plot

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.

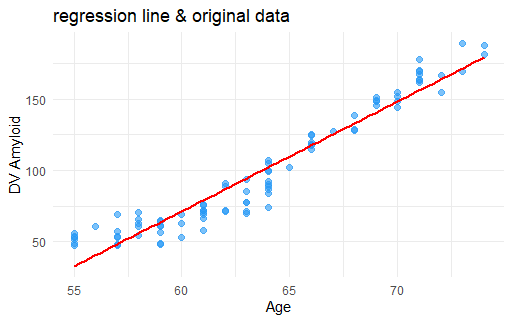


Figure 2.3 regression line & original data

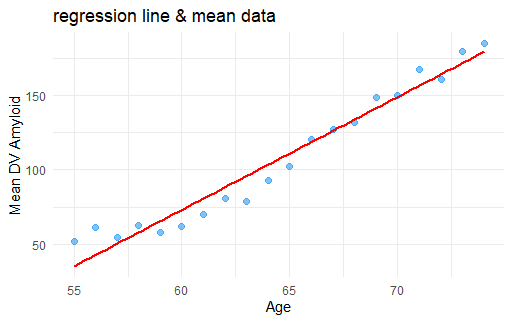


Figure 2.4 regression line & mean data

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

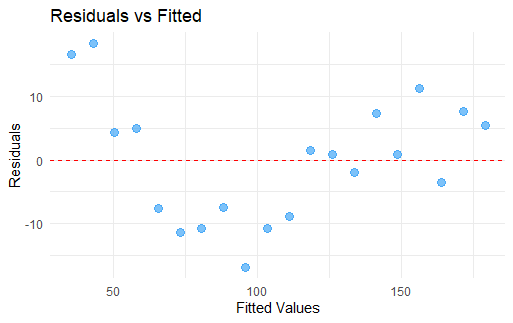


Figure 2.5 residuals value vs fitted value

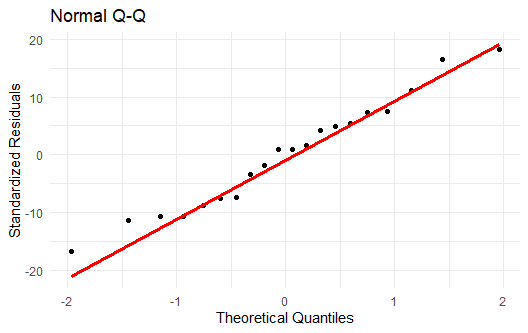


Figure 2.6 Q-Q plot

> 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**

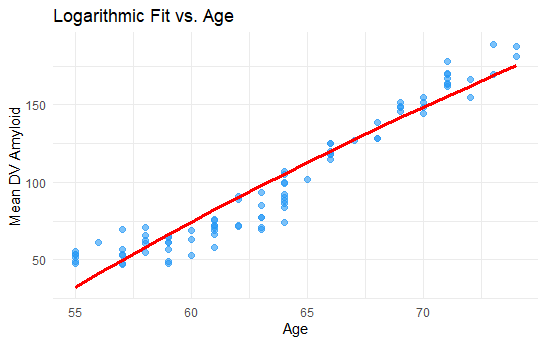


Figure 2.7 Logarithmic Model regression

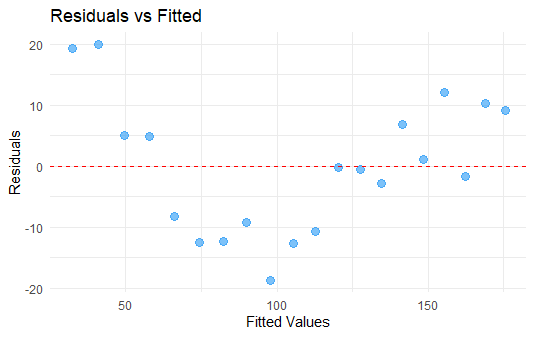


Figure 2.8 Logarithmic Model residuals

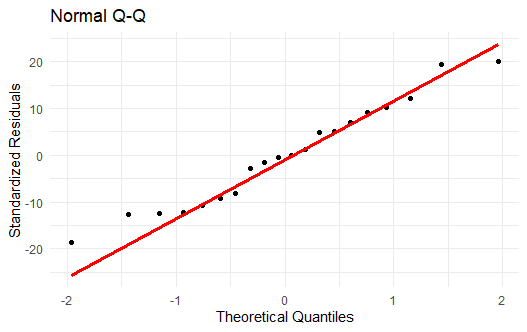


Figure 2.9 Logarithmic Model Q-Q plot

> # 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**

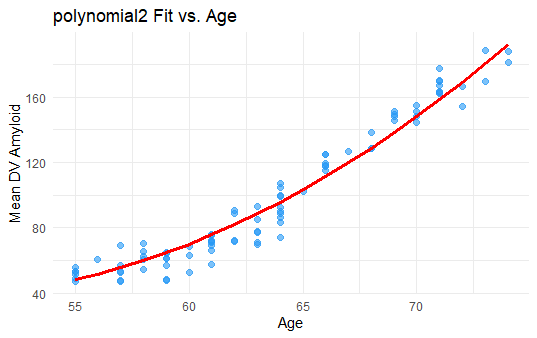


Figure 2.10 Polynomial 2 Model regression

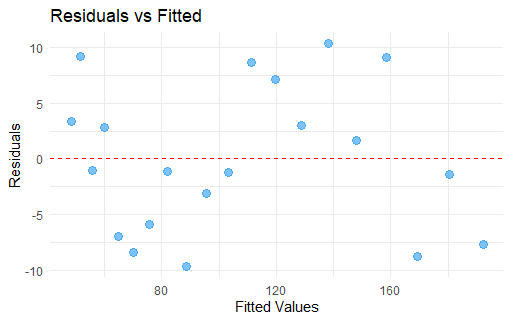


Figure 2.11 Polynomial 2 Model residuals

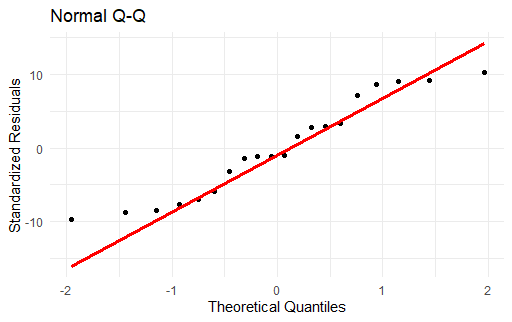


Figure 2.12 Polynomial 2 Model Q-Q plot

> 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

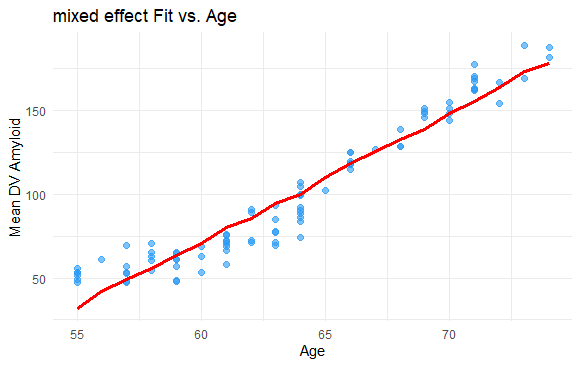


Figure 2.13 Mixed-effect Model regression

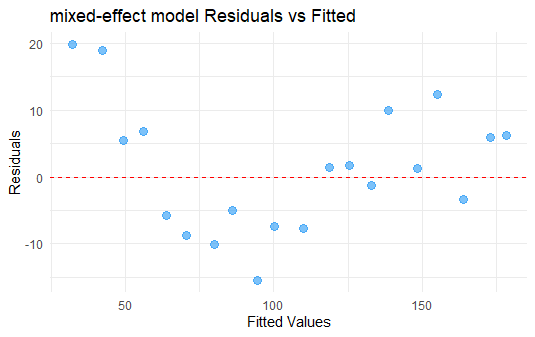


Figure 2.14 Mixed-effect Model residuals

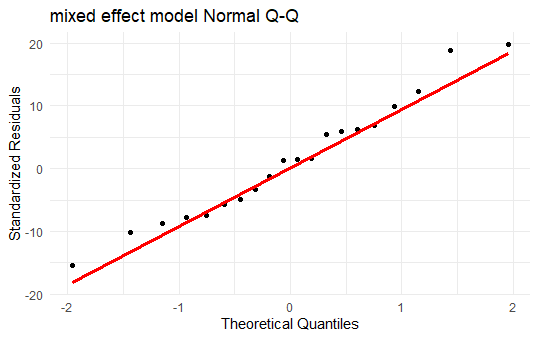


Figure 2.15 Mixed-effect Model Q-Q plot

> 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.

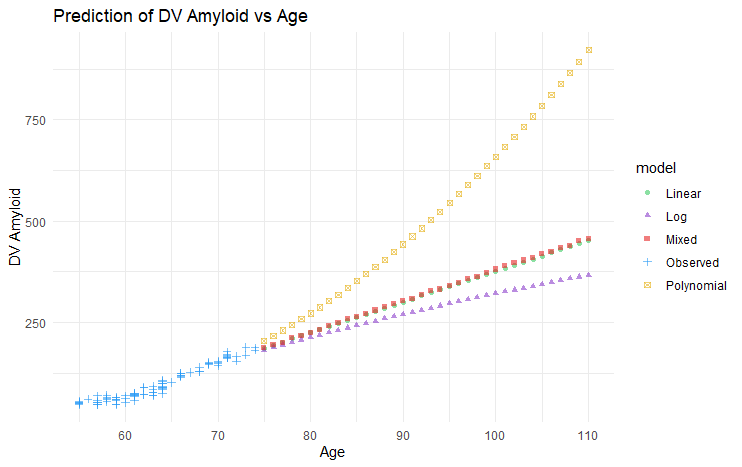


Figure 2.16 Prediction of the amyloid

**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.

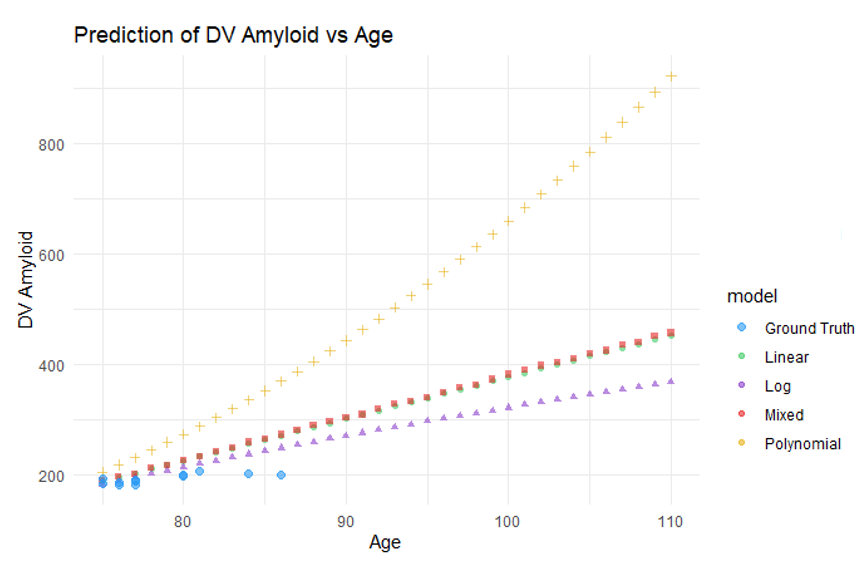


Figure 3.1 Comparison of the predictions and the ground truth

**Improved model**

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

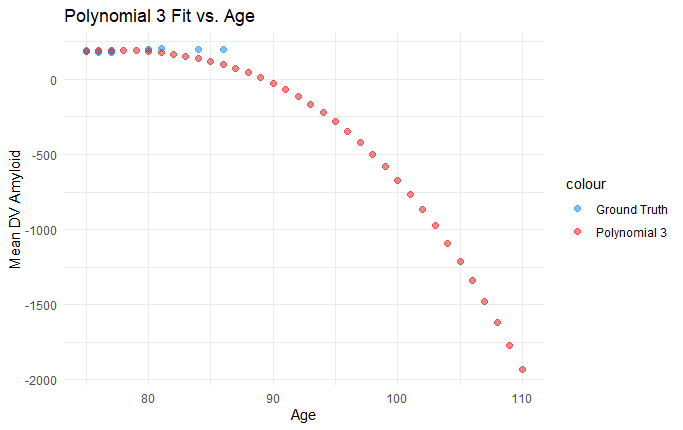


Figure 3.3 Polynomial 3 Model

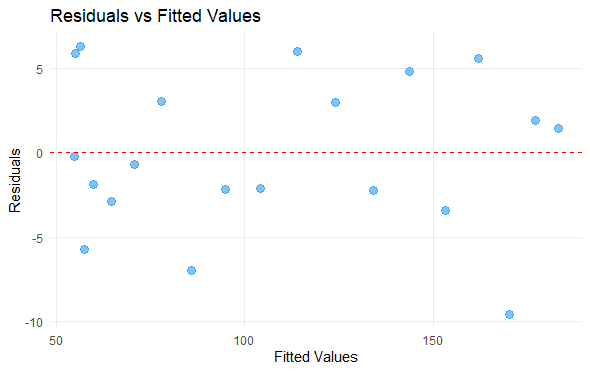


Figure 3.3 Polynomial 3 Model residuals

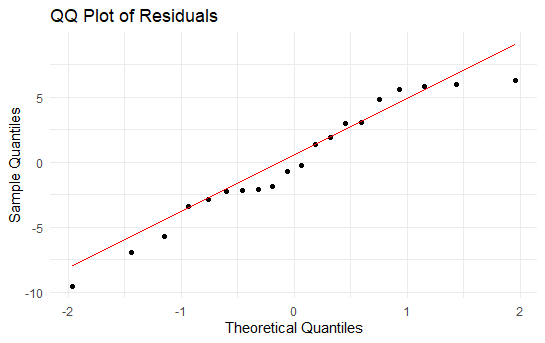


Figure 3.3 Polynomial 3 Model Q-Q plot

> 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

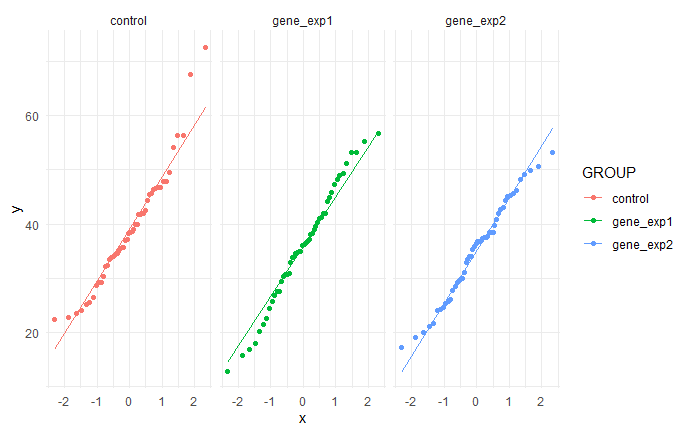


Figure 4.1 three groups Q-Q plot

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

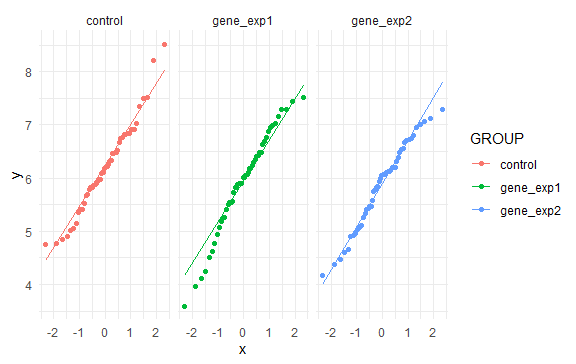


Figure 4.2 three groups Q-Q plot after root transformation

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.