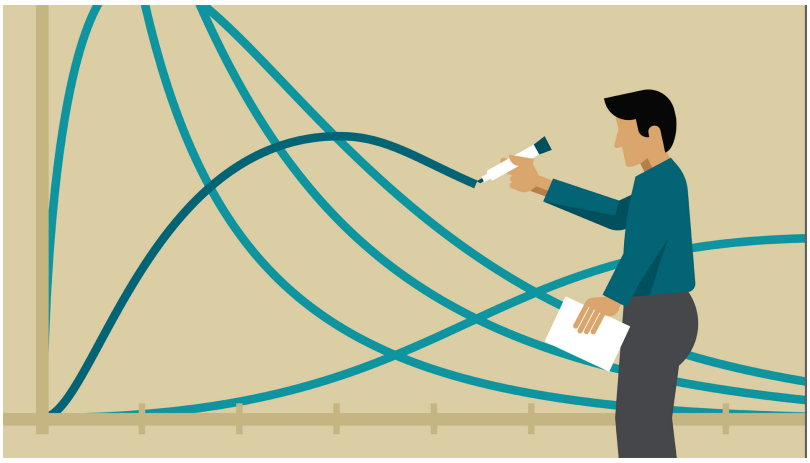
Group Project - Stats 112

Final Report



Nhat Dao

Lucy Lu

Shou Shimaya

Table of Contents

I. Dataset Introduction and Analysis

1. Dataset and Experiment Introduction

There are a total of 1309 subjects in this randomized study with 4 covariates available in this dataset, which are treatment, age, gender, and week.

There are total of 4 types of treatment: zidovudine alternating monthly with 400mg didanosine, zidovudine plus 2.25mg of zalcitabine, zidovudine plus 400mg of didanosine, and zidovudine plus 400mg of didanosine plus 400mg of nevirapine.

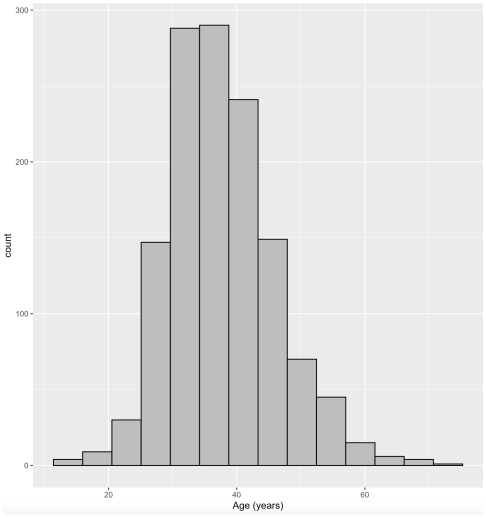
The response variable is the transformed CD4 counts which measure the number of CD4 cells within the AIDS patients.

In this report, we want to see how the different treatment types change the log CD4 counts over time.

2. Univariate summaries (numerical and graphical) of each covariate

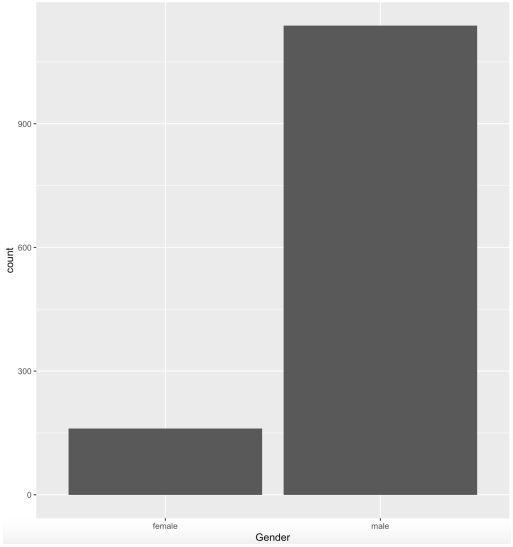
```
      id         treatment      age           gender              week      log_cd4
Min.   :   1.0   1:320     Min.   :14.90   Length:1299        Min.   :0   Min.   :0.000
1st Qu.: 325.5   2:322     1st Qu.:31.78   Class :character   1st Qu.:0   1st Qu.:2.398
Median : 652.0   3:327     Median :36.86   Mode  :character   Median :0   Median :3.045
Mean   : 655.9   4:330     Mean   :37.73                      Mean   :0   Mean   :2.913
3rd Qu.: 986.5             3rd Qu.:42.45                      3rd Qu.:0   3rd Qu.:3.584
Max.   :1313.0             Max.   :74.19                      Max.   :0   Max.   :5.198
      id         treatment      age           gender              week          log_cd4
Min.   :   1.0   1:1239    Min.   :14.90   Length:5036        Min.   : 0.00   Min.   :0.000
1st Qu.: 318.8   2:1251    1st Qu.:31.76   Class :character   1st Qu.: 0.00   1st Qu.:2.303
Median : 640.5   3:1254    Median :36.85   Mode  :character   Median :15.86   Median :2.944
Mean   : 654.5   4:1292    Mean   :37.73                      Mean   :15.46   Mean   :2.872
3rd Qu.: 988.0             3rd Qu.:42.54                      3rd Qu.:25.00   3rd Qu.:3.570
Max.   :1313.0             Max.   :74.19                      Max.   :40.00   Max.   :6.297
```

Age:



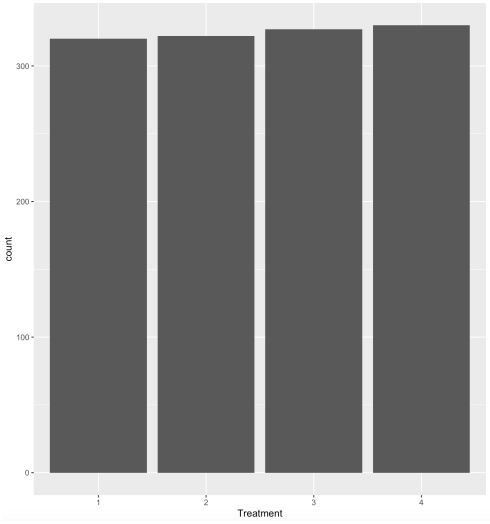
Age refers to the patient's age in years from a minimum of 14.9 to a maximum of 74.19 with a mean of 37.73 and median of 36.85. The histogram of age is approximately normally distributed with no obvious skewness.

Gender:



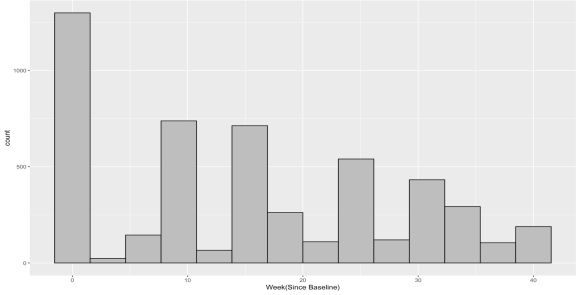
Gender refers to a patient being male or female. From the bar graph, the number of patients being male and female have a significant difference with a larger population of subjects being male (right bar). There are a total of 1147 male subjects and 162 female subjects.

Treatment:



Treatment refers to the 4 treatments in the study: 1, 2, 3, 4. From the bar graph we can see that the number of patients that are being assigned to each treatment are roughly the same with 320 subjects in treatment 1, 322 in treatment 2, 327 in treatment 3, and 330 in treatment 4.

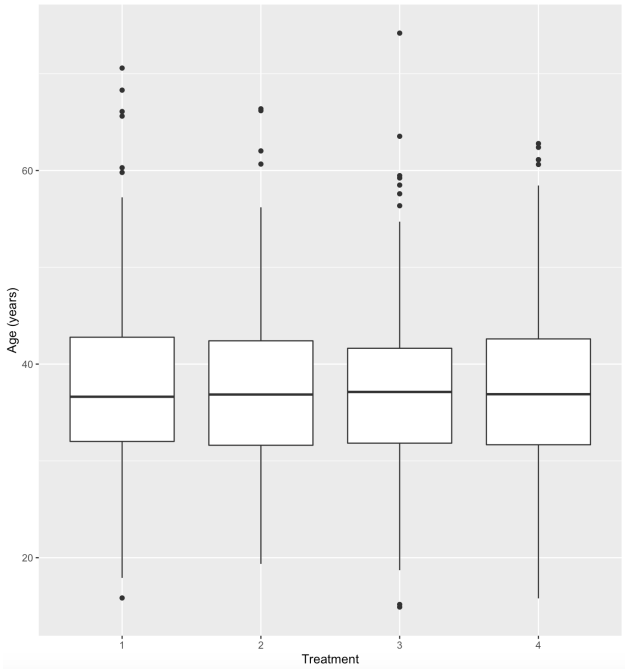
Week:



Week refers to the time since baseline (in weeks) from week 0 to week 40. There is a minimum of 0 week and maximum of 40 weeks with a mean of 15.46 and median of 15.86. For the histogram, we can see that there is some skewness.

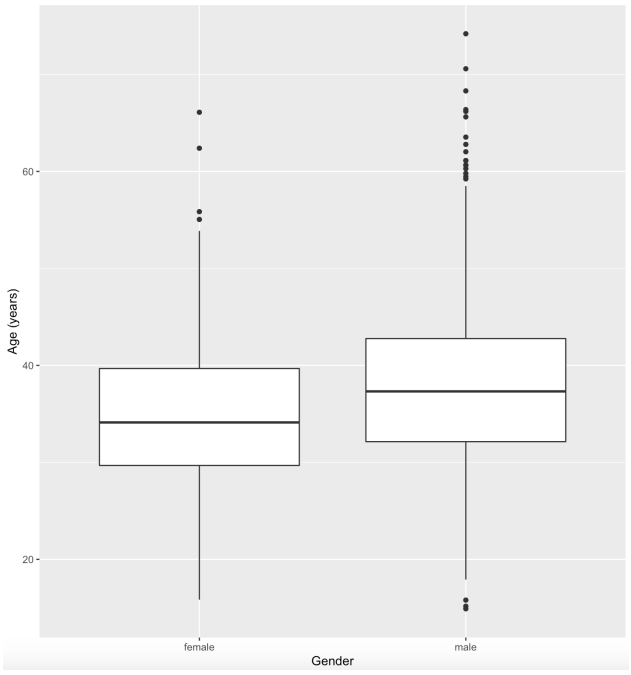
3. Bivariate summaries (numerical and graphical) of variables in the dataset

Treatment vs Age:



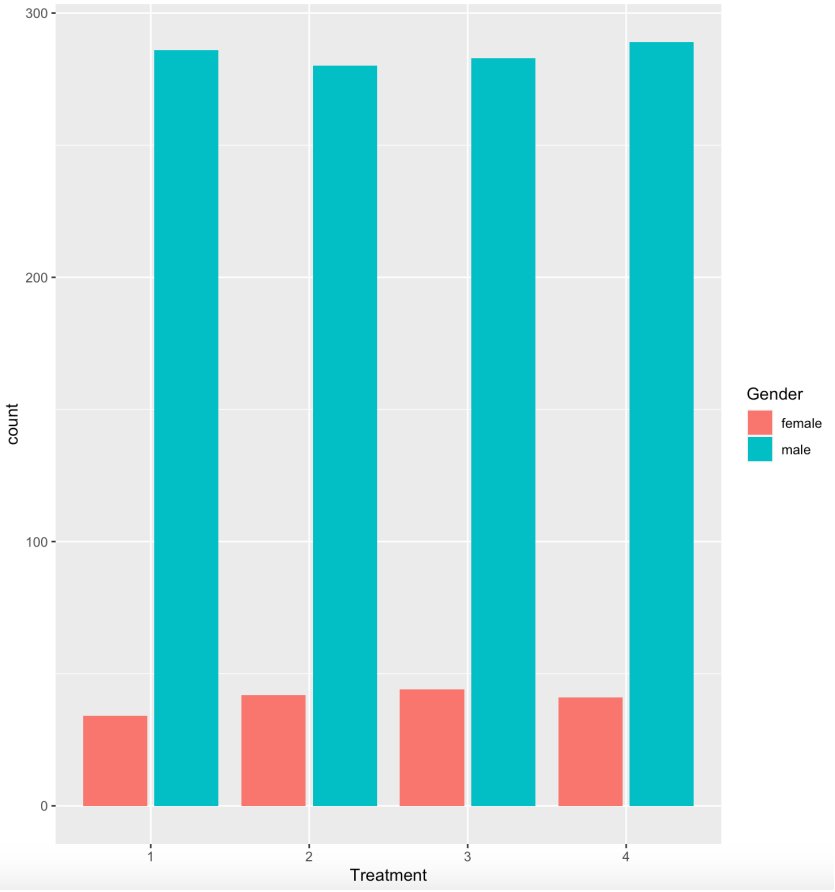
This boxplot shows the relationship between treatment and age. The median age of the four treatments are relatively the same, and the distributions of age are very similar except for the potential outliers.

Gender vs Age:



The median age of male subjects is higher than the female subjects. The minimum age for female and male subjects is approximately the same. However, the maximum age for male is higher than female subjects.

Male and female counts in each treatment:



This bar graph shows the amount of females and males within each treatment group. Because there are more male subjects in the study than female subjects, hence, the number of males in each treatment group is greater than females. And this is a randomized study, thus, the number of subjects assigned to each treatment group is roughly the same.

log_cd4 vs gender:

mean log_cd4 for male: 2.91

mean log_cd4 for female: 2.94

log_cd4 vs treatment:

mean log_cd4 for treatment 1: 2.98

mean log_cd4 for treatment 2: 2.93

mean log_cd4 for treatment 3: 2.91

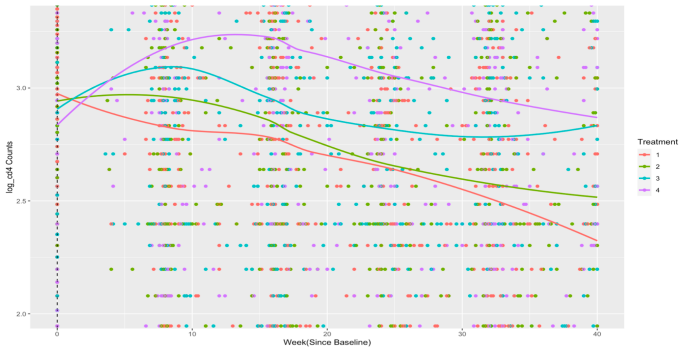
mean log_cd4 for treatment 4: 2.84

4. Overall trends of the response variable relative to other variables

week vs log_cd4 grouped by gender

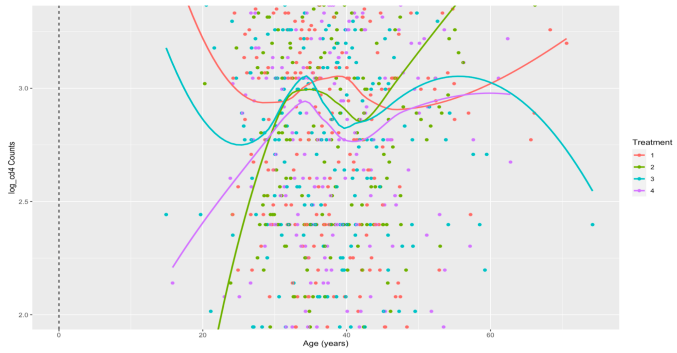
From the scatter plot, we can see that the log_cd4 counts have a decreasing trend as the number of weeks increases for both male and female. However, the log_cd4 counts for females (red line) are consistently higher than males.

week vs log_cd4 grouped by treatment



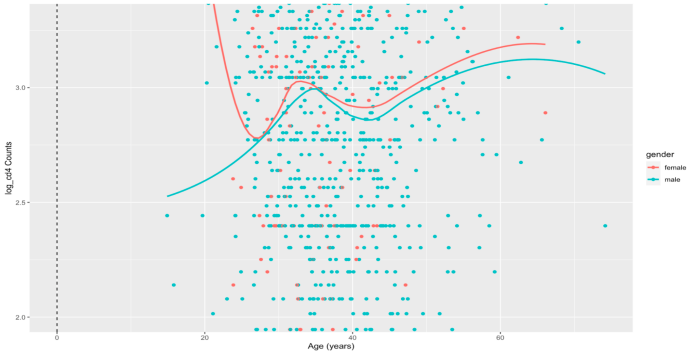
From the scatter plot, we can see that the log_cd4 counts have an increasing and decreasing trend as the number of weeks increases for all treatment groups. The log_cd4 counts for treatment group 4 (purple line) are consistently higher than other treatment groups, treatment group 1 has the lowest log_cd4 counts (red line). From this graph, we see that an interaction term of treatment and week is needed, possibly up to degree of 3 for week.

Age vs log_cd4 grouped by treatment



From the scatter plot, we can see that the log_cd4 counts have no obvious trend as age increases for all treatment groups. And there is an interaction effect between the variable age and treatment, since there is intersection on the graph.

Age vs log_cd4 grouped by gender



From the scatter plot, we can see that the log_cd4 counts have no obvious pattern as age increases for both male and female. And there is some interaction effect between the variable age and gender.

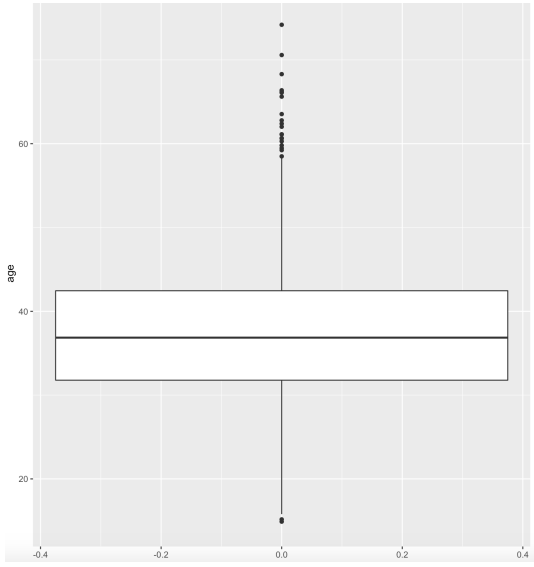
5. Imbalance in the dataset

The dataset has an imbalance in the numbers of male and female patients (1147 males and 162 females).

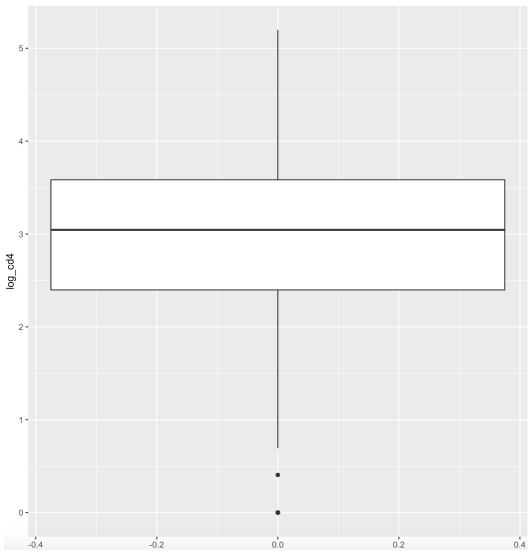
Additionally, there is an imbalance in the measurement occasions for some patients. For example, patient with id 1 has 6 measurement occasions while patient with id 3 has only 1 measurement occasion.

6. Outlier in the dataset

Age:



log_cd4:



There are some potential outliers above the age of 60 of the patients.

There is also one outlier with a log_cd4 count of 0.

II. LME Model Selection

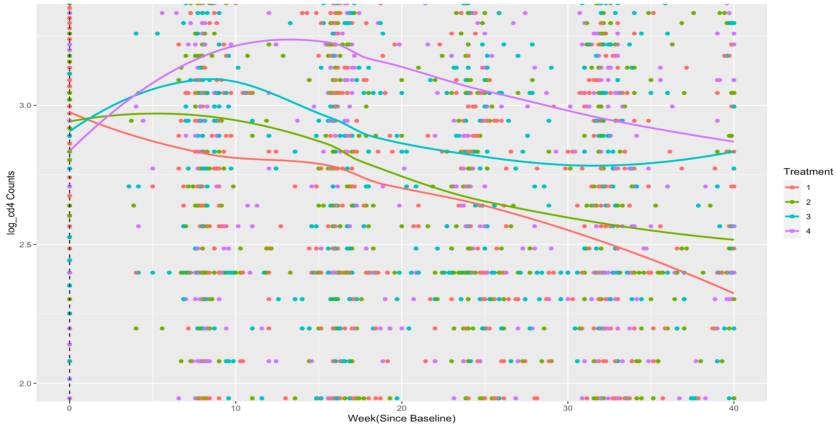
1. Visualizations to get ideas of what covariates should be included:

week vs log_cd4 grouped by Gender

From the graph above, we see that there should be some interaction between gender and week. Based on the male curve (red line), it looks like the model should have up to a cubic term for week. We see a slight quartic pattern in the female curve, but we will not include quartic terms in the model because it might overfit.

The curvature pattern is different in different genders, so we will also include up to a cubic term of week for interaction with gender.

week vs log_cd4 grouped by Treatment



We also see cubic patterns here as well. The curvature pattern is different in different treatment groups, we will include up to a cubic term for week*treatment interaction.

From the graphs *age vs log_cd4 grouped by treatment* and *age vs log_cd4 grouped by gender* graphs from section I part 4, we do not see a clear trend between age and log_cd4. However, we decided to include age, age^2, age:treatment, and age:gender in our model to start with.

2. Choosing mean model

We will first find the best mean model using intercept and week as random effects, and then find the best random effects for the best mean model we come up with. For choosing a mean model, our approach is to make a starting model, which includes all the covariates that we thought might be useful. Then, we look at p-values of coefficients and take out unnecessary covariates.

Our starting model:

$$\text{log\_cd4} \sim week + week^2 + week^3 + treatment:week + treatment:week^2 + treatment:week^3 + gender + gender:week + gender:week^2 + gender:week^3 + age + age^2 + age:treatment + age:gender$$

(random = ~ $week$)

AIC of starting model with ML method: 11946.44

Anova tests for reduced models:

H_0	H_a	AIC	p-value for anova test with previous model	Conclusion
Model 2 (removed gender) is better	Starting model is better	11944.45	0.9085	model 2 is better
Model 3 (removed age^2) is better	Model 2 is better	11942.67	0.6367	model 3 is better
Model 4 (removed gender:week) is better	Model 3 is better	11941.17	0.4803	model 4 is better
Model 5 (removed	Model 4 is	11939.27	0.7579	model 5 is better

gender:week^3) is better	better			
Model 6 (removed age:gender) is better	Model 5 is better	11938.16	0.3450	model 6 is better
Model 7 (removed age:treatment) is better	Model 6 is better	11935.44	0.3495	model 7 is better
Model 8 (removed gender:week^2) is better	Model 7 is better	11936.73	0.0699	model 8 is better

Coefficients and their p-values for model 8:

	Value	Std.Error	DF	t-value	p-value
(Intercept)	2.5521725	0.11590980	3715	22.018608	0.0000
week	-0.0088684	0.00922833	3715	-0.960999	0.3366
I(week^2)	-0.0003921	0.00064235	3715	-0.610422	0.5416
I(week^3)	0.0000051	0.00001185	3715	0.426884	0.6695
age	0.0096218	0.00299484	1307	3.212779	0.0013
week:treatment2	0.0281978	0.01295470	3715	2.176648	0.0296
week:treatment3	0.0477274	0.01294858	3715	3.685920	0.0002
week:treatment4	0.0753505	0.01276589	3715	5.902487	0.0000
I(week^2):treatment2	-0.0018137	0.00090241	3715	-2.009829	0.0445
I(week^2):treatment3	-0.0027267	0.00090750	3715	-3.004647	0.0027
I(week^2):treatment4	-0.0033425	0.00088851	3715	-3.761944	0.0002
I(week^3):treatment2	0.0000311	0.00001658	3715	1.876839	0.0606
I(week^3):treatment3	0.0000448	0.00001680	3715	2.669674	0.0076
I(week^3):treatment4	0.0000465	0.00001634	3715	2.842494	0.0045

Since most of the p-values are less than 0.05, we know to stop the model selection here. Although the p-values for week, week^2, and week^3 have p-values greater than 0.05, we still need to include them in the model to ensure we have an accurate interpretation of the interaction terms.

Model 8: (mean model is carefully chosen, but random effects is not optimized yet)

$$\text{log\_cd4} \sim week + week^2 + week^3 + age + treatment{:}\,week + treatment{:}\,week^2 + treatment{:}\,week^3 \quad (\text{random} = \sim week)$$

3. Choosing random effects

Using the mean model we found, we try out different combinations of week, week^2, and week^3 for random effects. Here are the random effects that we tested and their AIC for ML:

1. week, AIC = 12143.82
2. week^2, AIC = 12211.51
3. week + week^2, AIC = 12060.26
4. week + week^3, AIC = 12082.50
5. week^2 + week^3, AIC = 12132.87

*intercept is included as a default

We tried to put 3 terms for random effects, but R could not handle it and it gave an error. Thus, we choose "week + week^2" to be the best possible terms to include for the random effects.

4. Final model

- $\log\_cd4 \sim week + week^2 + week^3 + age + treatment{:}\,week + treatment{:}\,week^2 + treatment{:}\,week^3$ (random = $\sim week + week^2$)
- $\log\_cd4 = \beta_1 + \beta_2 week + \beta_3 week^2 + \beta_4 week^3 + \beta_5 age + \beta_6 Treatment2 * week + \beta_7 Treatment3 * week + \beta_8 Treatment4 * week + \beta_9 Treatment2 * week^2 + \beta_{10} Treatment3 * week^2 + \beta_{11} Treatment4 * week^2 + \beta_{12} Treatment2 * week^3 + \beta_{13} Treatment3 * week^3 + \beta_{14} Treatment4 * week^3 + b_1 + b_2 week + b_3 week^2$

5. Testing for interaction between week and treatment

Ho: $\beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = 0$

Ha: At least one of the beta is not 0

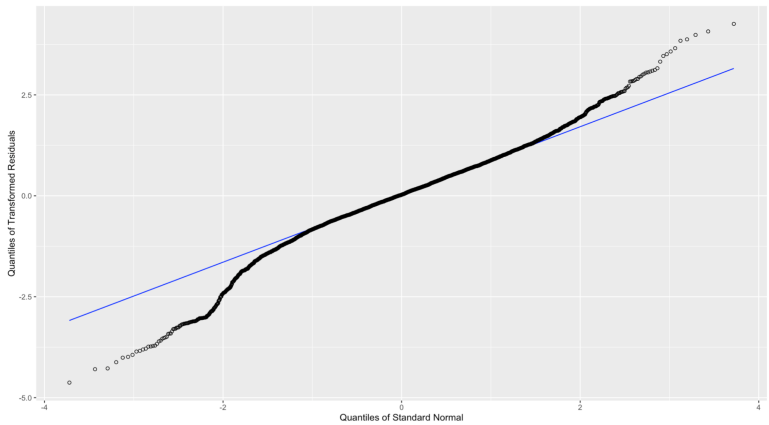
Test Statistic: 81.27594

P-value: 9.012486e-14

Since p-value is less than 0.05, we can reject the null hypothesis, and conclude that the expected change in log_cd4 counts over time is different between treatments.

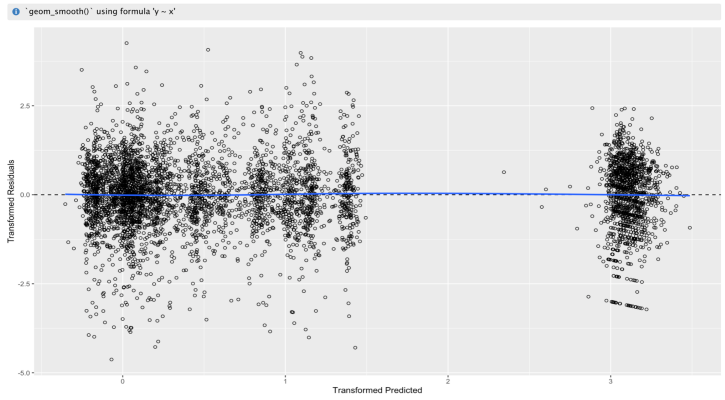
III. Residual Analysis for LME

QQ Plot:



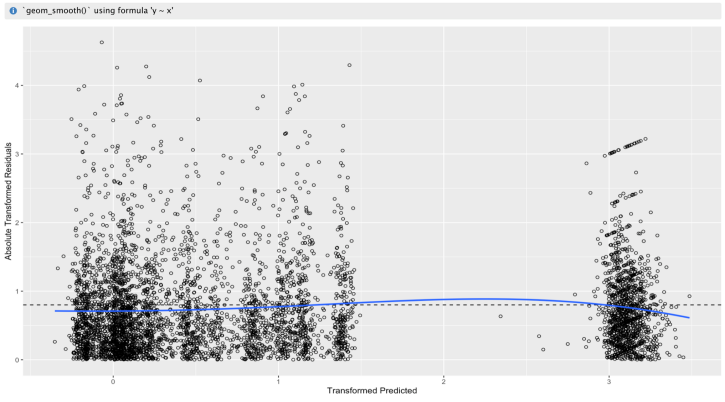
Although our transformed residuals do not perfectly follow standard normal distribution, this was the best we could do because we found the best pair of random effects, and R couldn't handle 3 random effects

Transformed Predicted Value vs Transformed Residuals:



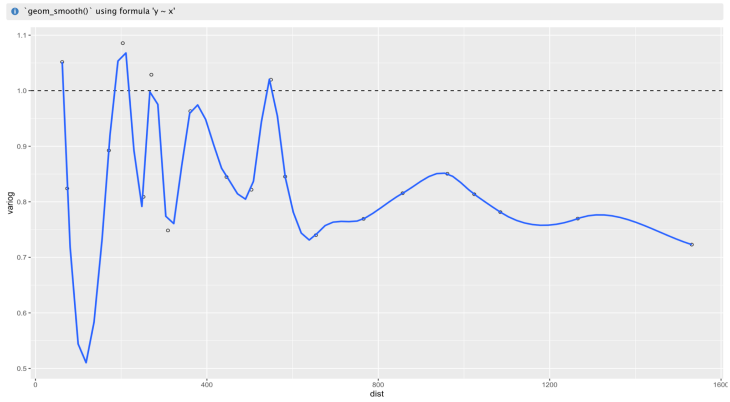
The points are randomly scattered around y=0, which means the model for the mean is correctly specified.

Transformed predicted value vs Absolute transformed residual:



There is no obvious pattern, which means that the model for variance is adequate

Semi-Variogram



The corresponding observations did not fluctuate randomly around the horizontal line centered at 1, therefore, the model for covariance is not correctly specified. But since we can only include two random effects in the model, the one we currently have is the best model.

Mahalanobis Distance Analysis:

There are 129 potential outlying individuals. We don't expect to have these many outliers because there are total of 1309 individuals, and 1309*0.05 ≈ 66, but we have 129 individuals who have a p-value less than 0.05. We don't expect this to happen by chance.

IV. GLME Modeling

When we tried to fit the GLME using the same equation as our LME model, it gave us warning and error, so we standardized the "week" in the dataset before modeling GLME. After rescaling, we were able to fit the GLME model with same equation as our LME model

1. Final Model

$$\text{CD4 Counts} = \beta_1 + \beta_2 week + \beta_3 week^2 + \beta_4 week^3 + \beta_5 Age + \beta_6 Treatment2 * week + \beta_7 Treatment3 * week + \beta_8 Treatment4 * week + \beta_9 Treatment2 * week^2 + \beta_{10} Treatment3 * week^2 + \beta_{11} Treatment4 * week^2 + \beta_{12} Treatment2 * week^3 + \beta_{13} Treatment3 * week^3 + \beta_{14} Treatment4 * week^3 + b_1 + b_2 week + b_3 week^2$$

2. Testing for interaction between week and treatment

Ho: $\beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = 0$

Ha: At least one of the betas is not 0

Test Statistic: 130.351

P-value: 1.004307e-23

Since p-value is less than 0.05, we can reject the null hypothesis and conclude that the expected change in CD4 counts over time is different between treatments.

V. Conclusion

After examining the effect of treatment types on the changes in CD4 counts over time, both the LME model and the GLME model agree that the expected change in CD4 counts over time is different between treatments. Since the two models have the same mean and variance covariates, both models are adequate to use for analyzing the primary interest of the experiment.