Stats 112 Final Project

Hursh Naik, Ulises Jimenez, Kevin Tobias

6/8/2021

Exploratory Data Analysis

Number of Subjects and Covariates

There are 1313 subjects in the dataset. There are 5 covariates in the dataset: treatment which is a categorical covariate, age which is a numerical covariate, gender which is a categorical covariate, week which is a numerical covariate and log_cd4 which is a numerical covariate.

Univariate Summaries

We will analyze the univariate summaries of covariates in our dataset in this section.
Let us first assign the treatment groups with codes.
1 is Zidovudine alternating monthly with 400mg didanosine.
2 is Zidovudine plus 2.25mg zalcitabine.
3 is Zidovudine plus 400mg didanosine.
4 is Zidovudine plus 400mg didanosine plus 400mg nevirapine.

The summary of the categorical covariate treatment is as follows:

```
##    1    2    3    4
## 1239 1251 1254 1292
```

The summary for age is:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.90   31.76   36.85   37.73   42.54   74.19
```

The summary for categorical covariate gender is:

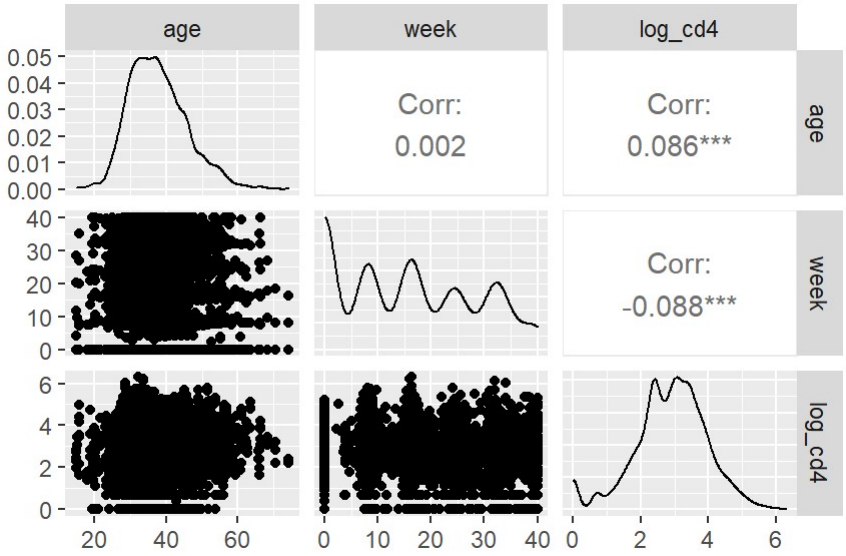
```
## female   male
##    561   4475
```

The summary of log cd4 is:

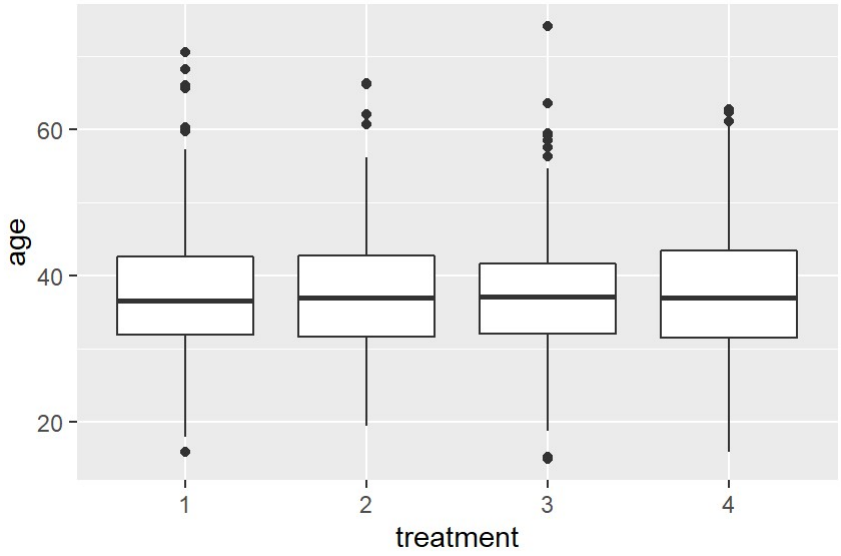
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.303   2.944   2.872   3.570   6.297
```

Bivariate Summaries

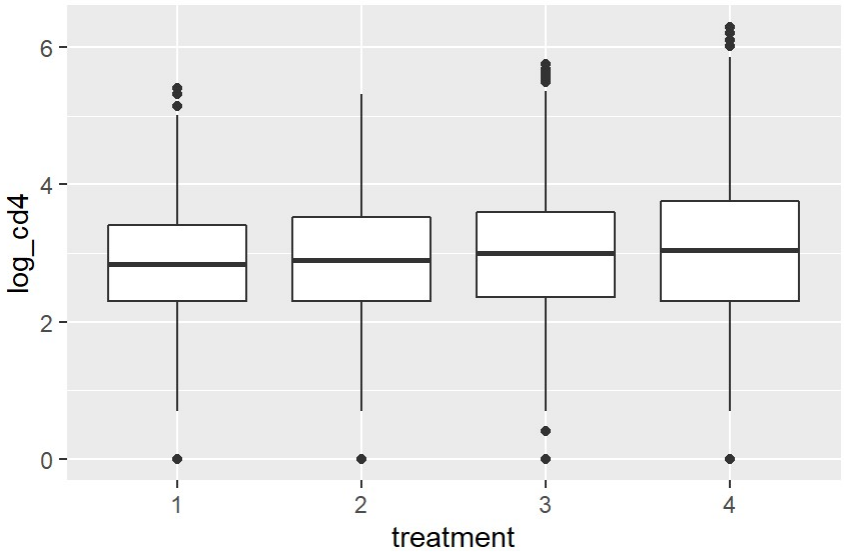
Here we will analyze the bivariate summaries between two covariates from our dataset.
Here we can see the variable distributions on the diagonal, scatterplots on the lower triangle and correlation coefficients on the upper triangle for age, week and log cd4.



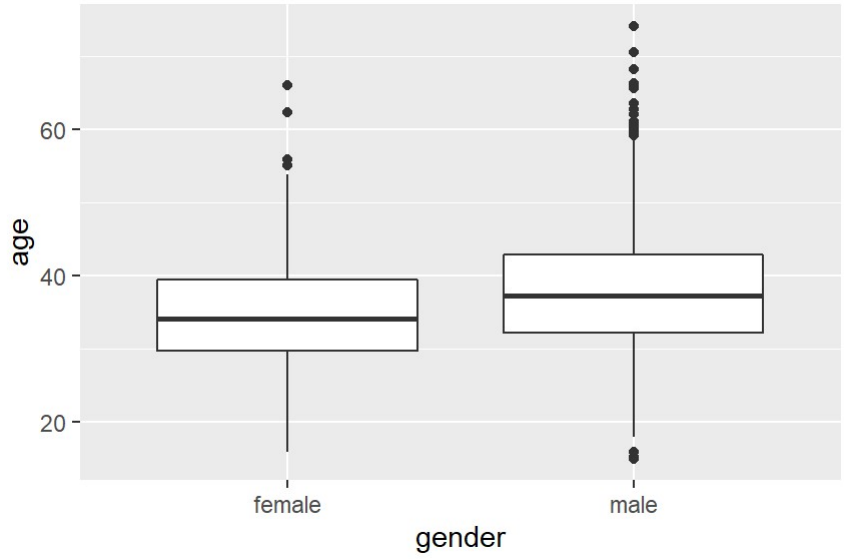
The boxplots for age among different treatment groups are:



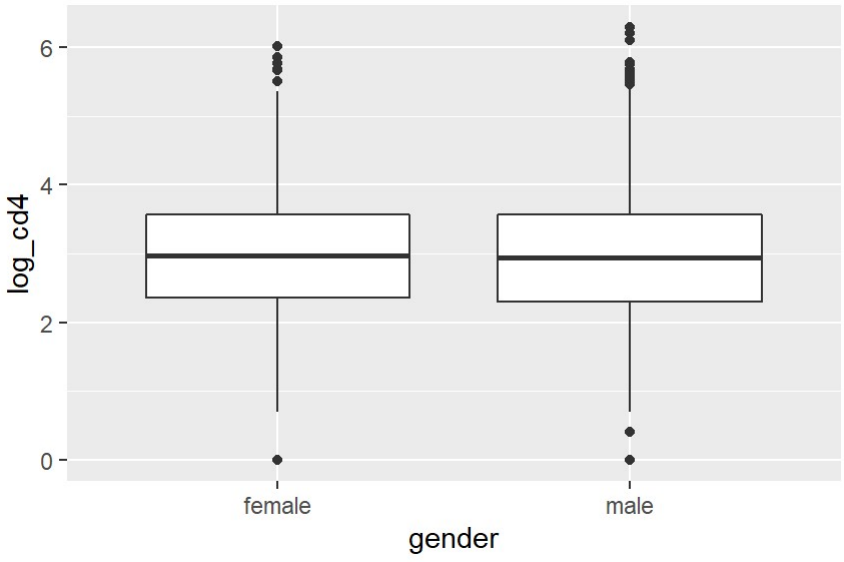
The boxplots for log cd4 among different treatment groups are:



The boxplots for age among different genders are:



The boxplots for log cd4 among different genders are:



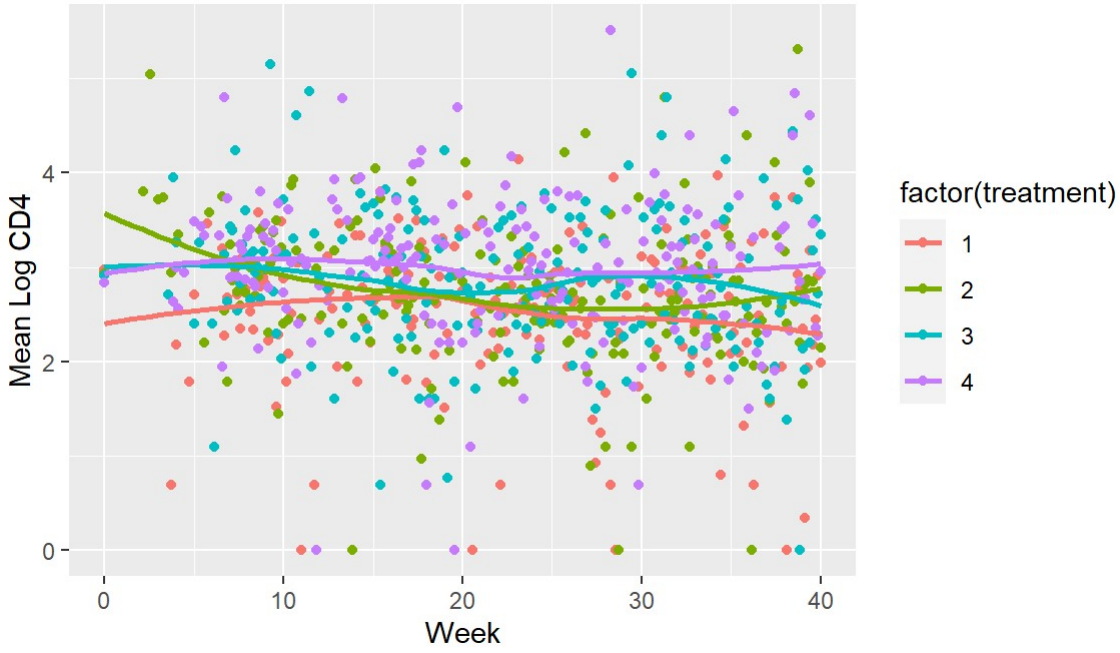
The covariance between week and log cd4, age and log cd4 are following respectively:

```
## [1] -1.188937
```

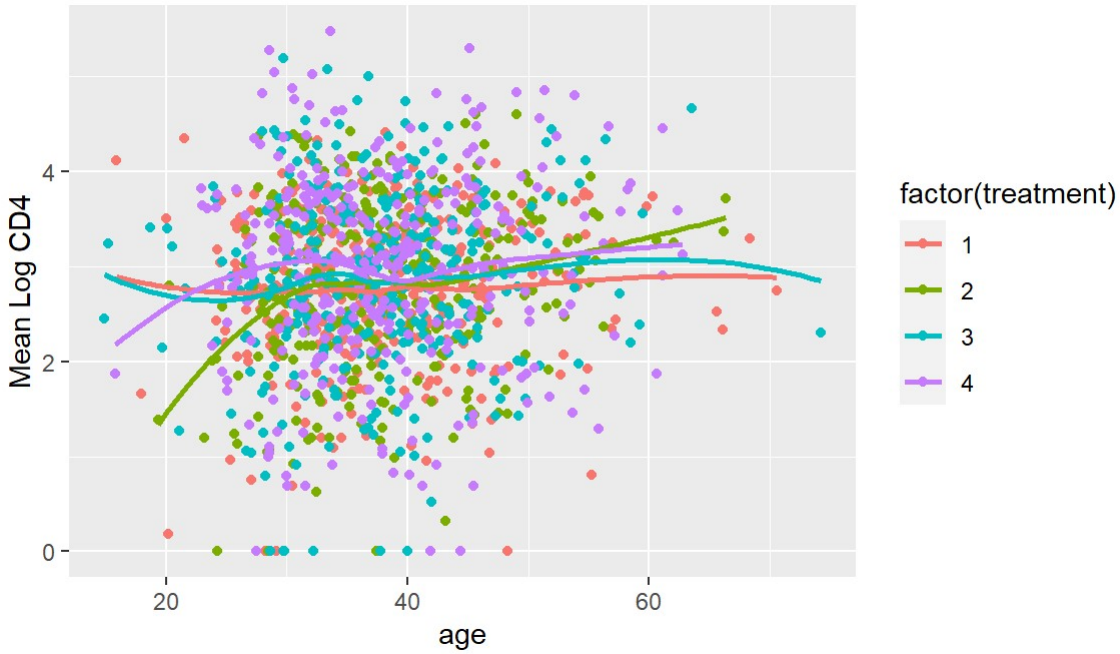
```
## [1] 0.7553842
```

Overall Trends

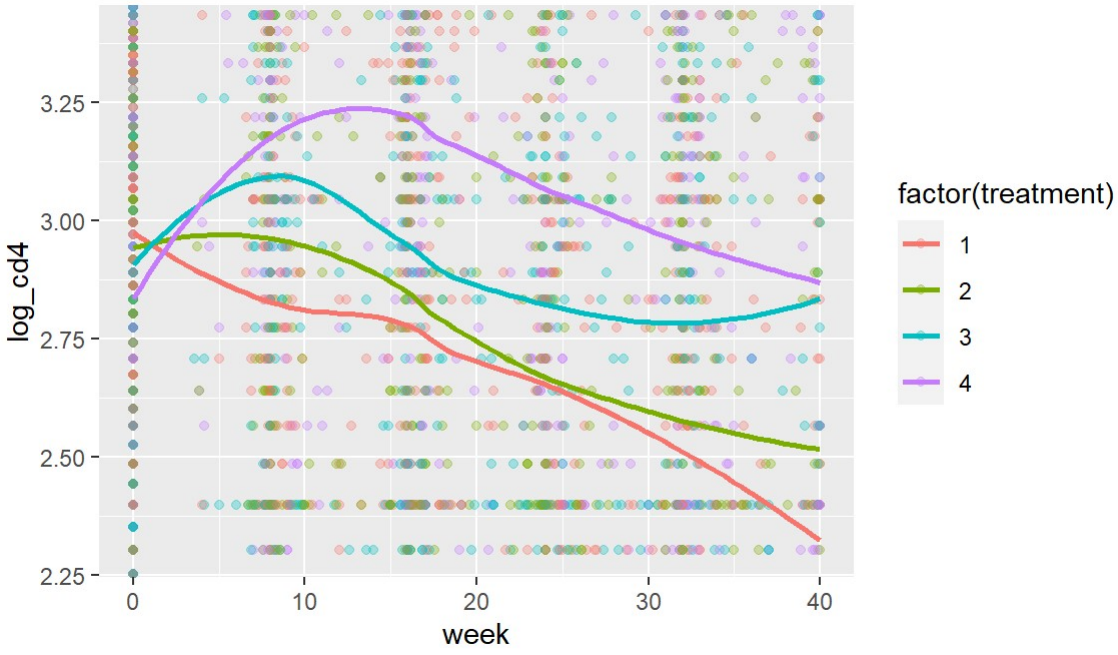
The mean log CD4 versus week grouped by the treatment group is as follows:



The mean log CD4 versus age grouped by the treatment group is as follows:



Below is a graph showing the trend of the log CD4 response across different treatment groups:



From this graph, we can see that there is varying degrees of curvature between the different treatment groups.

Imbalances

There were no fixed occasions for measurement of the response variable. The number of recorded measurements varies per id. There was a large imbalance in the gender ratio as males outnumbered females by a large margin. These were the imbalances observed in this dataset.

Outliers

All the subjects who have $cd4 = 0$ are possible outliers. And one subject with age 75 is also an outlier due to his old age potentially resulting in a different outcome of the treatment.

Modeling LME

Since the study is a randomized experiment, we already know the covariate of treatment will not be a good predictor for the model. So, we will be looking to incorporate the effect of treatment into the model by using an interaction term, most likely with week.

LME1: Our first model starts simple and looks at using week, age, and gender without any treatment tied into the model yet. We chose to use random effects on the intercept, week and age for this model, along with all subsequent models.

$$\begin{aligned} Y_{ij} =& \beta_1 + \beta_2 Week_{ij} + \beta_3 Age_i + \beta_4 Gender_i + \\ & b_{1i} + b_{2i} Week_{ij} + b_{3i} Age_i + \epsilon_{ij} \end{aligned}$$

LME2: From our exploratory data analysis, we thought that adding in a week squared term could benefit the model, as there appears to be some curvature present within the created LOESS line on the data. The revised model is as follows:

$$Y_{ij} = \beta_1 + \beta_2 Week_{ij} + \beta_3 Week_{ij}^2 + \beta_4 Age_i + \beta_5 Gender_i + b_{1i} + b_{2i} Week_{ij} + b_{3i} Age_i + \epsilon_{ij}$$

To check which model is better the null and alternative hypotheses are:

$H_0 : \beta_3 = 0$ (Referring to model LME2)

$H_A : \beta_3 \neq 0$

```
##      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## lme1     1 11 12105.28 12177.05 -6041.641
## lme2     2 12 12049.41 12127.70 -6012.706 1 vs 2 57.86908  <.0001
```

Since the p-value is less than $\alpha = 0.05$, we reject the null and conclude evidence for the alternative. Therefore, the model adding in week squared as a covariate is a better predictor of the log_cd4 of the patient than without. Thus, LME2 is a better model than LME1.

LME3 Next, we decided to add our first interaction term with treatment, in the form of an interaction between treatment and week. The model LME3 is as follows:

$$\begin{aligned} Y_{ij} = &\beta_1 + \beta_2 Week_{ij} + \beta_3 Week_{ij}^2 + \beta_4 Age_i + \beta_5 Gender_i + \\ &\beta_6 Treatment_2 Week_{ij} + \beta_7 Treatment_3 Week_{ij} + \beta_8 Treatment_4 Week_{ij} + \\ &b_{1i} + b_{2i} Week_{ij} + b_{3i} Age_i + \epsilon_{ij} \end{aligned}$$

To check which model is better the null and alternative hypotheses are:

$H_0 : \beta_6 = \beta_7 = \beta_8 = 0$ (Referring to model LME3)

H_A : the null is not true

```
##      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## lme2     1 12 12049.41 12127.70 -6012.706
## lme3     2 15 12001.19 12099.06 -5985.597 1 vs 2 54.21857  <.0001
```

Since the p-value is less than $\alpha = 0.05$, we reject the null and conclude evidence for the alternative. Therefore, the model adding in the interaction term between treatment and week is a better predictor of the log_cd4 of the patient than without. Thus, LME3 is a better model than LME2.

LME4: Next, we decided to see if we could remove the covariate of gender from our model. The model LME4 is as follows:

$$\begin{aligned} Y_{ij} = &\beta_1 + \beta_2 Week_{ij} + \beta_3 Week_{ij}^2 + \beta_4 Age_i + \\ &\beta_5 Treatment_2 Week_{ij} + \beta_6 Treatment_3 Week_{ij} + \beta_7 Treatment_4 Week_{ij} + \\ &b_{1i} + b_{2i} Week_{ij} + b_{3i} Age_i + \epsilon_{ij} \end{aligned}$$

To check which model is better the null and alternative hypotheses are:

$H_0 : \beta_5 = 0$ (Referring to model LME3)

$H_A : \beta_5 \neq 0$

```
##      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## lme3     1 15 12001.19 12099.06 -5985.597
## lme4     2 14 12000.26 12091.60 -5986.132 1 vs 2 1.069812   0.301
```

Since the p-value is greater than $\alpha = 0.05$, we fail to reject the null. Therefore, the model removing the covariate for gender is a better predictor of the log_cd4 of the patient. Thus, LME4 is a better model than LME3.

LME5: Next, we decided to try an interaction term between treatment and week squared. This is because we noticed that the LOESS curves on the different treatments varied, so we thought the model could benefit from an interaction between the two. The model LME5 is as follows:

$$\begin{aligned} Y_{ij} =& \beta_1 + \beta_2 Week_{ij} + \beta_3 Week_{ij}^2 + \beta_4 Age_i + \\ & \beta_5 Treatment_2 Week_{ij} + \beta_6 Treatment_3 Week_{ij} + \beta_7 Treatment_4 Week_{ij} + \\ & \beta_8 Treatment_2 Week_{ij}^2 + \beta_9 Treatment_3 Week_{ij}^2 + \beta_{10} Treatment_4 Week_{ij}^2 + \\ & b_{1i} + b_{2i} Week_{ij} + b_{3i} Age_i + \epsilon_{ij} \end{aligned}$$

To check which model is better the null and alternative hypotheses are:

$H_0 : \beta_8 = \beta_9 = \beta_{10} = 0$ (Referring to model LME5)

H_A : the null is not true

```
##      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## lme5     1 17 11977.88 12088.79 -5971.940
## lme2     2 12 12049.41 12127.70 -6012.706 1 vs 2 81.53342  <.0001
```

Since the p-value is less than $\alpha = 0.05$, we reject the null and conclude evidence for the alternative. Therefore, the model adding in the interaction term between treatment and week squared is a better predictor of the log_cd4 of the patient than without. Thus, LME5 is a better model than LME4.

LME6: Finally, we still thought that there might be some curvature not explained alone by the week squared covariate and treatment week squared interaction term, so we decided to test a 6th model utilizing both a week cubed covariate and a treatment week cubed interaction term. The model is as follows:

$$\begin{aligned} Y_{ij} =& \beta_1 + \beta_2 Week_{ij} + \beta_3 Week_{ij}^2 + \beta_4 Week_{ij}^3 + \beta_5 Age_i + \\ & \beta_6 Treatment_2 Week_{ij} + \beta_7 Treatment_3 Week_{ij} + \beta_8 Treatment_4 Week_{ij} + \\ & \beta_9 Treatment_2 Week_{ij}^2 + \beta_{10} Treatment_3 Week_{ij}^2 + \beta_{11} Treatment_4 Week_{ij}^2 + \\ & \beta_{12} Treatment_2 Week_{ij}^3 + \beta_{13} Treatment_3 Week_{ij}^3 + \beta_{14} Treatment_4 Week_{ij}^3 + \\ & b_{1i} + b_{2i} Week_{ij} + b_{3i} Age_i + \epsilon_{ij} \end{aligned}$$

```
##        (Intercept)                week            week_sq            week_cu
##       2.523510e+00       -7.641561e-03      -4.496985e-04       5.578070e-06
##                age    week:treatment2    week:treatment3    week:treatment4
##       1.035579e-02        2.762428e-02       4.681152e-02       7.366294e-02
## treatment2:week_sq treatment3:week_sq treatment4:week_sq treatment2:week_cu
##      -1.795622e-03       -2.644674e-03      -3.244397e-03       3.099304e-05
## treatment3:week_cu treatment4:week_cu
##       4.291019e-05        4.509318e-05
```

To check which model is better the null and alternative hypotheses are:

$H_0 : \beta_4 = \beta_{12} = \beta_{13} = \beta_{14} = 0$ (Referring to model LME6)

H_A : the null is not true

```
##      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## lme5     1 17 11977.88 12088.79 -5971.940
## lme6     2 21 11942.73 12079.74 -5950.365 1 vs 2 43.14889  <.0001
```

Since the p-value is less than $\alpha = 0.05$, we reject the null and conclude evidence for the alternative. Therefore, the model adding in week cubed covariate and a treatment week cubed interaction term is a better predictor of the log_cd4 of the patient than without. Thus, LME6 is a better model than LME5. We believe this occurs because after the base treatment of zidovudine and treatment type (1,2,3, or 4) is given there must be time for the medication to change the CD4 counts. Thus the interaction terms between treatment and weeks are significant despite treatment as a covariate not being significant.

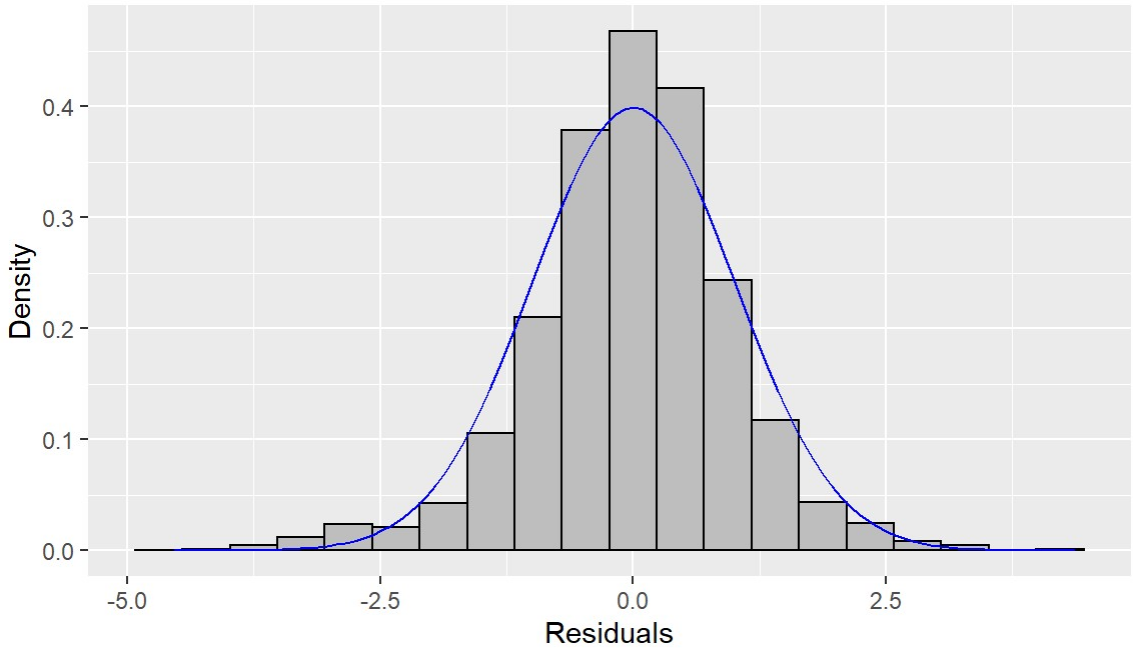
Therefore, after our initial model exploration we decided to settle on model LME6 as our best predictor of the relationship between treatment and log_cd4 counts.

Residual Analyses

Residuals of LME6:

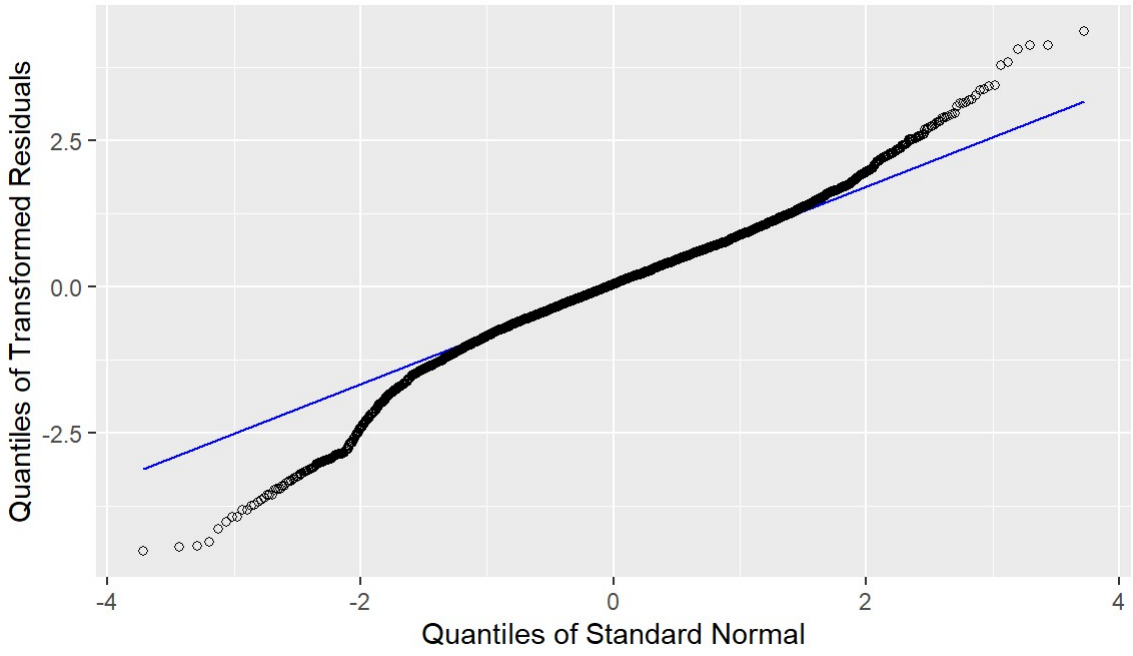
Since this is a longitudinal study, the components of the vector residuals are correlated. Thus, we will transform the residuals in order to prevent any systemic trend in the scatterplot of the residuals against any of the covariates of our model. When we transform the residuals we obtain constant variance and zero correlation so that we can use our typical residual diagnostics for standard linear regression.

Histogram of the transformed residuals of LME6:



Notice that the transformed residual histogram follows a normal distribution centered around mean 0 without any skewness.This indicates we have normality.

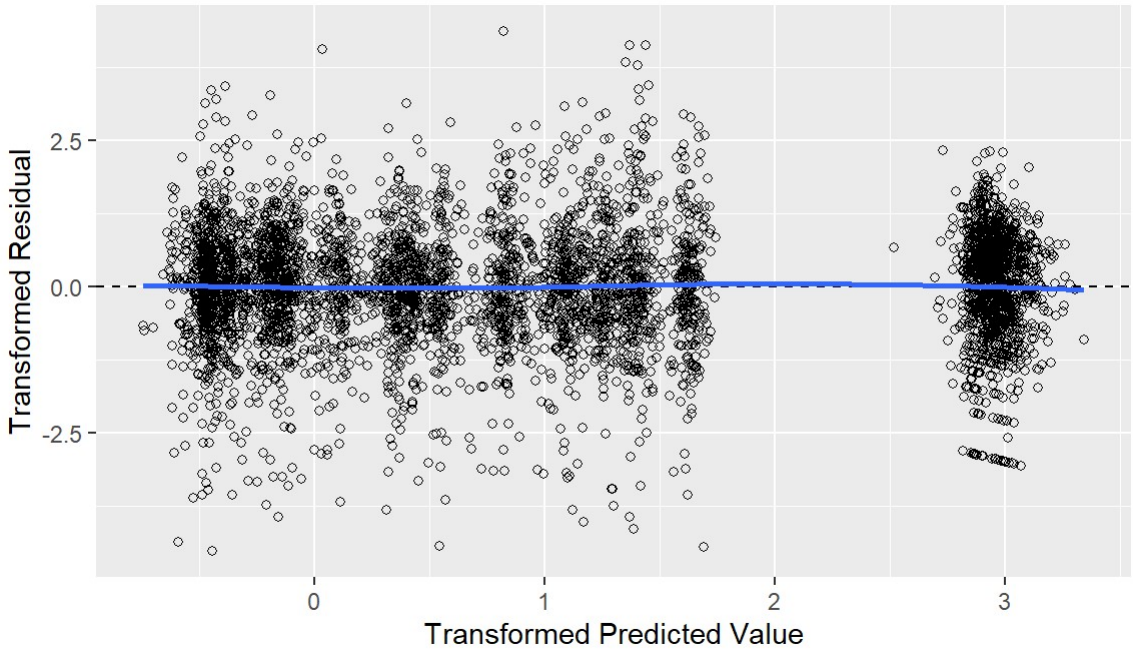
QQplot of the transformed residuals of LME6:



Notice that on the lower half of the QQplot many points diverge from the blue linear line indicating a couple of outliers (around 100 or so). This can also be seen at the upper right part of the QQplot which also shows many extreme values. Eventhough around 100 points seem to be outliers this is to be expected since we have a large sample size of 1313 with 5036 observations. However, the majority of the points are on the the blue linear line so we can assume normal distribution.

Scatterplot of the transformed residuals of LME6:

```
## `geom_smooth()` using formula 'y ~ x'
```



The plot shows no systemic trend or pattern but it does show some curvature. However, some curvature is to be expected so it is within a normal range. In addition, all the points seem to be equally scattered around 0 and majority of the cluster seems to be located to the left and center of the plot.

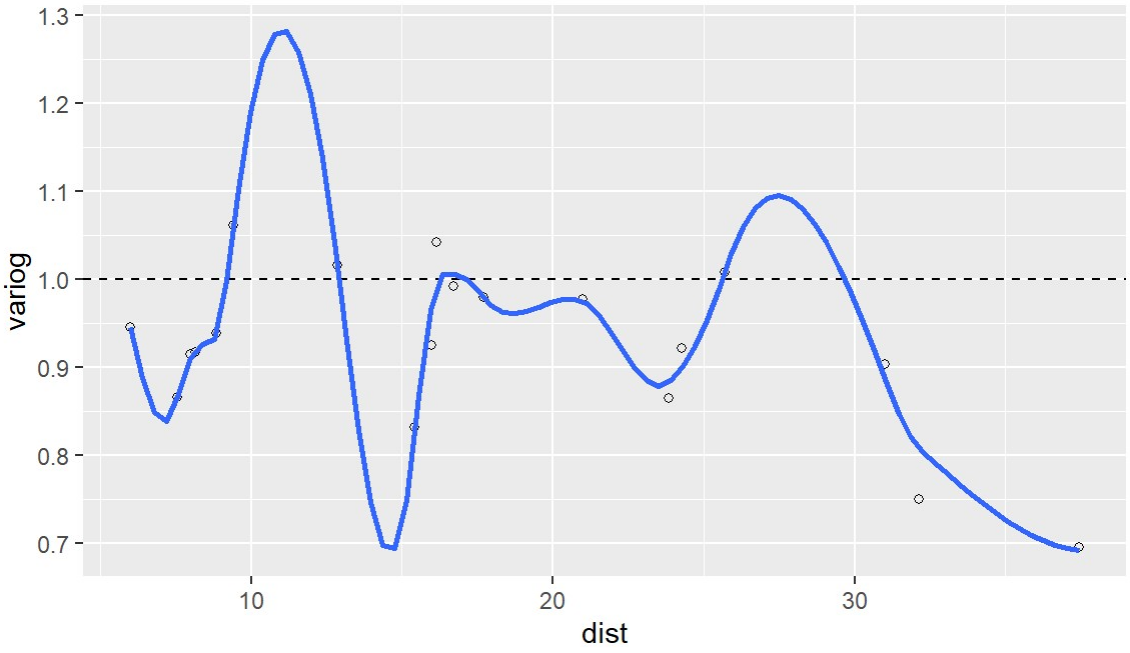
Mahalanobis Distance for LME6 model:

```
## # A tibble: 1,309 x 5
## # Groups:   id [1,309]
##       id data                    df     d     p_value
##    <dbl> <list>                <dbl> <dbl>       <dbl>
##  1   178 <tibble[,1] [5 x 1]>     5  38.9 0.000000252
##  2   692 <tibble[,1] [5 x 1]>     5  34.1 0.00000225
##  3  1118 <tibble[,1] [5 x 1]>     5  33.6 0.00000283
##  4  1193 <tibble[,1] [4 x 1]>     4  28.5 0.00000998
##  5  1207 <tibble[,1] [5 x 1]>     5  30.6 0.0000110
##  6  1110 <tibble[,1] [6 x 1]>     6  30.5 0.0000312
##  7   371 <tibble[,1] [2 x 1]>     2  20.4 0.0000367
##  8   877 <tibble[,1] [6 x 1]>     6  29.8 0.0000430
##  9  1117 <tibble[,1] [5 x 1]>     5  27.3 0.0000498
## 10   626 <tibble[,1] [5 x 1]>     5  26.9 0.0000599
## # ... with 1,299 more rows
```

Based on the MD p-values we can see there are 133 outliers since their p-values are extremely small (smaller than $\alpha = 0.05$). This is to be expected since there is 1313 subjects.

Semi Variogram for LME6 model:

```
## `geom_smooth()` using formula 'y ~ x'
```



Based on the semi-variogram the curve fluctuates randomly near the horizontal line centered at 1. In addition, the curve does not display any systemic trend over time.

GLME Model

Creating Counts

Here we have created another column in the aids dataset which contains the counts. Here we will find the original untransformed counts of cd4 using this transformation: $counts = e^{log(cd4)} - 1$. We will make a new column containing those values called counts.

GLME Model

Now let us create a GLME model as shown below to predict the counts. Here $E(Y_{ij}|X_{ij}) = u_{ij}$.

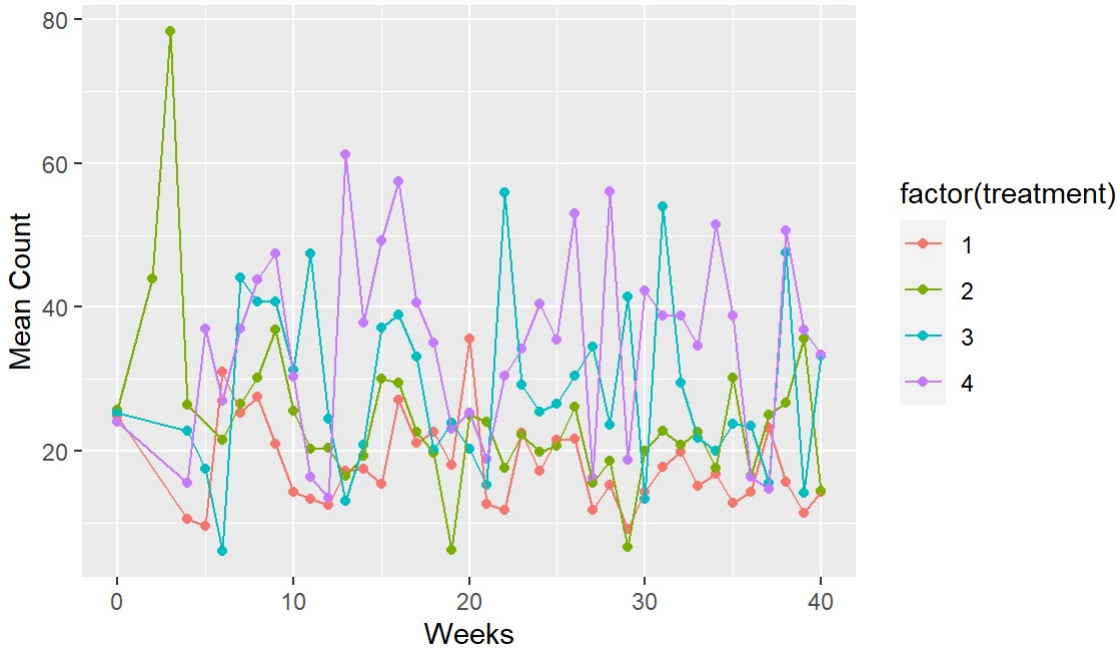
$$\begin{aligned} log(u_{ij}) =& \beta_1 + \beta_2 Week_{ij} + \beta_3 Week_{ij}^2 + \beta_4 Week_{ij}^3 + \beta_5 Age_i + \\ & \beta_6 Treatment_2 Week_{ij} + \beta_7 Treatment_3 Week_{ij} + \beta_8 Treatment_4 Week_{ij} + \\ & \beta_9 Treatment_2 Week_{ij}^2 + \beta_{10} Treatment_3 Week_{ij}^2 + \beta_{11} Treatment_4 Week_{ij}^2 + \\ & \beta_{12} Treatment_2 Week_{ij}^3 + \beta_{13} Treatment_3 Week_{ij}^3 + \beta_{14} Treatment_4 Week_{ij}^3 + \\ & b_{1i} + b_{2i} Week_{ij} + b_{3i} Age_i \end{aligned}$$

The estimates of the GLME model are:

```
##        (Intercept)               week            week_sq            week_cu
##       2.497618e+00       2.079799e-02      -2.091488e-03       2.550590e-05
##                age    week:treatment2    week:treatment3    week:treatment4
##       9.915063e-03       2.761662e-02       5.820975e-02       9.247278e-02
## treatment2:week_sq treatment3:week_sq treatment4:week_sq treatment2:week_cu
##      -1.885380e-03      -3.242295e-03      -4.518624e-03       3.374736e-05
## treatment3:week_cu treatment4:week_cu
##       4.788402e-05       6.674502e-05
```

As the GLME model converged we can conclude this is better than our previous LME models.

Below, for reference, we have a plot of mean counts of the treatment groups versus weeks .



Comparing the LME and GLME Models

GLME Model:

Wald Test statistic for interaction terms on GLME Model

```
##           [,1]
## [1,] 517.7545
```

$H_0 : \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = 0$

H_A : the null is not true

Chi Square Test of the Wald Statistic:

```
##              [,1]
## [1,] 9.05986e-106
```

Since the test returns a value less than $\alpha = 0.05$, we reject the null hypothesis and conclude evidence for the alternative. In other words, the interaction terms between week and treatment, week squared and treatment, and week cubed and treatment are significant predictors of the effect of treatment on the count of CD4 White Blood Cells.

LME Model:

Wald Test statistic for interaction terms on LME Model

```
##          [,1]
## [1,] 93.64141
```

$H_0 : \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = 0$

H_A : the null is not true

Chi Square Test of the Wald Statistic:

```
##              [,1]
## [1,] 3.019154e-16
```

Since the test returns a value less than $\alpha = 0.05$, we reject the null hypothesis and conclude evidence for the alternative. In other words, the interaction terms between week and treatment, week squared and treatment, and week cubed and treatment are significant predictors of the effect of treatment on the count of CD4 White Blood Cells.

As shown by the Wald tests conducted on both our GLME and LME models, we see that the interaction terms between treatment and week, treatment and week squared, and treatment and week cubed are significant predictors in both models. So, while the models go about predicting the outcome in different ways, both models are in agreement that these interaction terms are significant predictors of the results in the aids study.

Now returning to the GLME model, we will conduct the Wald Test for each interaction term individually to double check that every interaction term was significant to the model.

Wald Test statistic for interaction of Week and Treatment in GLME:

```
##          [,1]
## [1,] 351.4927
```

$H_0 : \beta_6 = \beta_7 = \beta_8 = 0$

H_A : the null is not true

Chi Square Test of the Wald Statistic:

```
##              [,1]
## [1,] 7.087055e-76
```

Since the test returns a value less than $\alpha = 0.05$, we reject the null hypothesis and conclude evidence for the alternative. In other words, the interaction term between week and treatment is a significant predictor of the effect of treatment on the count of CD4 White Blood Cells.

Wald Test statistic for interaction of Week Squared and Treatment in GLME:

```
##          [,1]
## [1,] 196.1833
```

$H_0 : \beta_9 = \beta_{10} = \beta_{11} = 0$

H_A : the null is not true

Chi Square Test of the Wald Statistic:

```
##             [,1]
## [1,] 2.81711e-42
```

Since the test returns a value less than $\alpha = 0.05$, we reject the null hypothesis and conclude evidence for the alternative. In other words, the interaction term between week squared and treatment is a significant predictor of the effect of treatment on the count of CD4 White Blood Cells.

Wald Test statistic for interaction of Week Cubed and Treatment in GLME:

```
##          [,1]
## [1,] 115.4969
```

$H_0 : \beta_{12} = \beta_{13} = \beta_{14} = 0$

H_A : the null is not true

Chi Square Test of the Wald Statistic:

```
##             [,1]
## [1,] 7.19638e-25
```

Since the test returns a value less than $\alpha = 0.05$, we reject the null hypothesis and conclude evidence for the alternative. In other words, the interaction term between week cubed and treatment is a significant predictor of the effect of treatment on the count of CD4 White Blood Cells.

Conclusion

After our analyses on the given Aids dataset, we have come to the conclusion that the different treatments have varying effects on the count of CD4 White Blood Cells. It appears that treatment group 4 is the most effective treatment for helping to rebuild the patient's count of CD4 White Blood Cells, while treatment group 1 is the least effective treatment for rebuilding this count of CD4 White Blood cells.

After testing various potential models, we determined that not only was week a good predictor of the success of a treatment group, but the additional terms of week squared and week cubed. These week terms in addition to their interaction terms with each of the treatment groups proved to be good predictors of the count of CD4

White Blood cells in a patient.

Additionally, after looking at both LME and GLME models for the dataset, we determined that while both models estimated differing values for the covariates, both of them are good predictors of the dataset and find significance in the same interaction terms between week and treatment. The Wald test comparison of the interaction terms in the GLME model also confirms that each set of interaction terms with week, week squared, and week cubed are all significant predictors of the effect of each treatment group and worth the degrees of freedom it takes to include them in our model. To recap, the 2 models are as follows:

LME Model:

$$\begin{aligned} Y_{ij} =& \beta_1 + \beta_2 Week_{ij} + \beta_3 Week_{ij}^2 + \beta_4 Week_{ij}^3 + \beta_5 Age_i + \\ & \beta_6 Treatment_2 Week_{ij} + \beta_7 Treatment_3 Week_{ij} + \beta_8 Treatment_4 Week_{ij} + \\ & \beta_9 Treatment_2 Week_{ij}^2 + \beta_{10} Treatment_3 Week_{ij}^2 + \beta_{11} Treatment_4 Week_{ij}^2 + \\ & \beta_{12} Treatment_2 Week_{ij}^3 + \beta_{13} Treatment_3 Week_{ij}^3 + \beta_{14} Treatment_4 Week_{ij}^3 + \\ & b_{1i} + b_{2i} Week_{ij} + b_{3i} Age_i + \epsilon_{ij} \end{aligned}$$

GLME Model: Let $E(Y_{ij}|X_{ij}) = u_{ij}$, then using a log link function we have:

$$\begin{aligned} log(u_{ij}) =& \beta_1 + \beta_2 Week_{ij} + \beta_3 Week_{ij}^2 + \beta_4 Week_{ij}^3 + \beta_5 Age_i + \\ & \beta_6 Treatment_2 Week_{ij} + \beta_7 Treatment_3 Week_{ij} + \beta_8 Treatment_4 Week_{ij} + \\ & \beta_9 Treatment_2 Week_{ij}^2 + \beta_{10} Treatment_3 Week_{ij}^2 + \beta_{11} Treatment_4 Week_{ij}^2 + \\ & \beta_{12} Treatment_2 Week_{ij}^3 + \beta_{13} Treatment_3 Week_{ij}^3 + \beta_{14} Treatment_4 Week_{ij}^3 + \\ & b_{1i} + b_{2i} Week_{ij} + b_{3i} Age_i \end{aligned}$$

Overall, we conclude that treatment group 4 appears to be the most successful treatment, while treatment 1 appears to be the least successful.