Data Analysis of Aids Clinical Trial Group Study

Micah Fadrigo, Christie Yang, Baolong Truong

6/4/2021

```
library(tidyverse)
library(nlme)
library(mgcv)
library(geepack)
library(lme4)
```

Introduction

The data is from a randomized, double-blind, study of AIDS patients with advanced immune suppression and CD4 counts of less than or equal to 50 cells/mm^3. The 1309 patients in this trial were randomized to one of four daily regimens of a medication called Zidovudine. The goal of this project is to compare the effect of treatment types on the changes in log CD4 cell count counts over time.

```
treatment_code <- data.frame(Code = c(1, 2, 3, 4),
                             Treatment = c("zidovudine alternating monthly with 400mg didanosine",
                                           "zidovudine plus 2.25mg of zalcitabine",
                                           "zidovudine plus 400mg of didanosine",
                                           "zidovudine plus 400mg of didanosine plus 400mg of nevirapin

treatment_code
```

```
##   Code                                                  Treatment
## 1    1         zidovudine alternating monthly with 400mg didanosine
## 2    2                        zidovudine plus 2.25mg of zalcitabine
## 3    3                          zidovudine plus 400mg of didanosine
## 4    4 zidovudine plus 400mg of didanosine plus 400mg of nevirapine
```

EDA ON DATASET

Number of Subjects, Covariates, and Summary

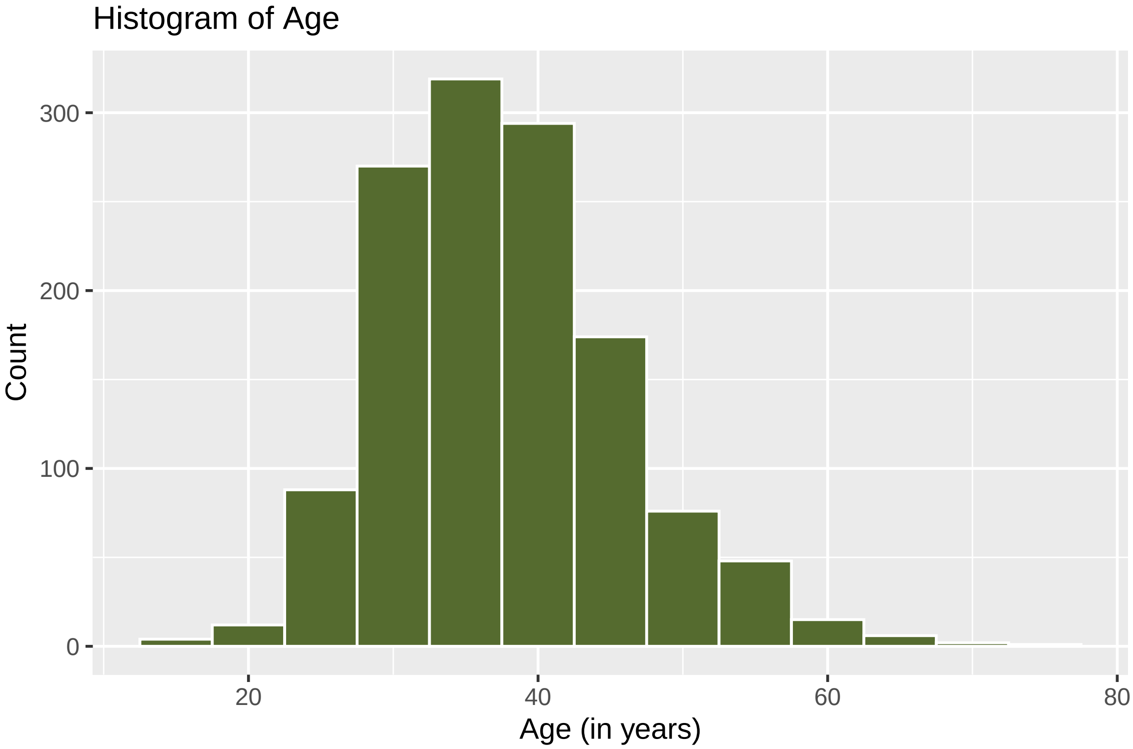
```
##  treatment      age           gender               week          log_cd4
##  1:1239    Min.   :14.90   Length:5036        Min.   : 0.00   Min.   :0.000
##  2:1251    1st Qu.:31.76   Class :character   1st Qu.: 0.00   1st Qu.:2.303
##  3:1254    Median :36.85   Mode  :character   Median :15.86   Median :2.944
##  4:1292    Mean   :37.73                      Mean   :15.46   Mean   :2.872
##            3rd Qu.:42.54                      3rd Qu.:25.00   3rd Qu.:3.570
##            Max.   :74.19                      Max.   :40.00   Max.   :6.297
```

In this study, there are **1313 subjects** and **4 covariates**. Note that the summary for id is not meaningful.

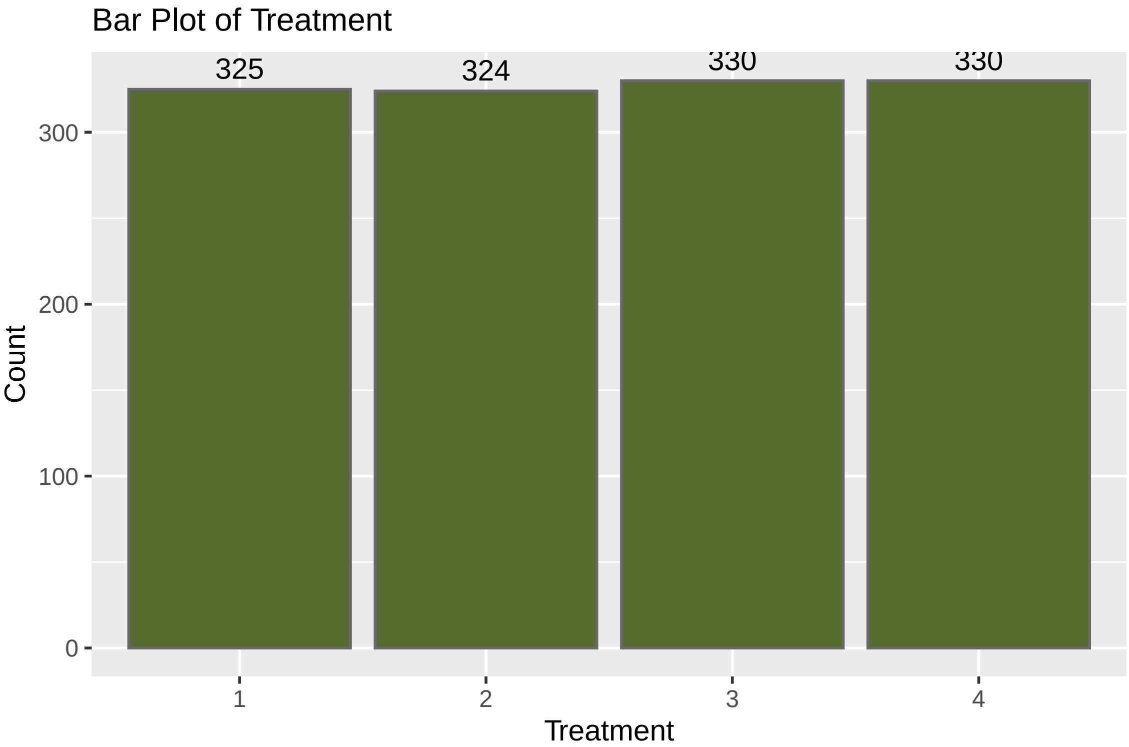
Univariate summaries

Age

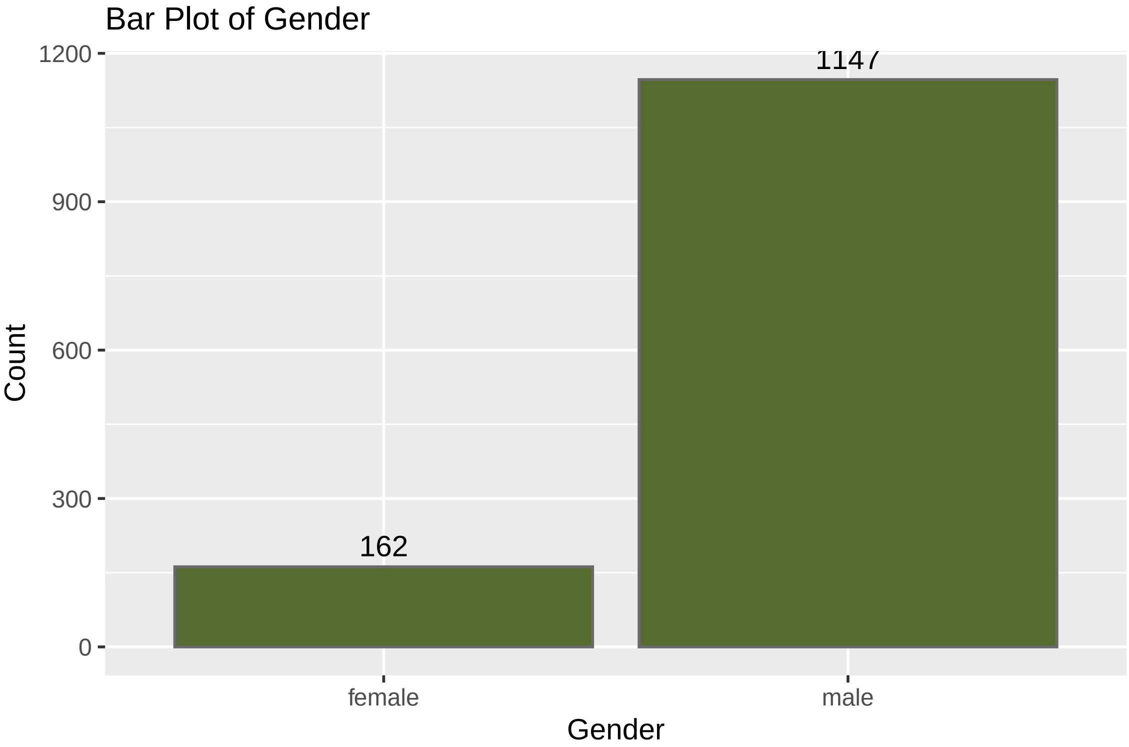
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   14.90   31.80   36.86   37.73   42.41   74.19 
```



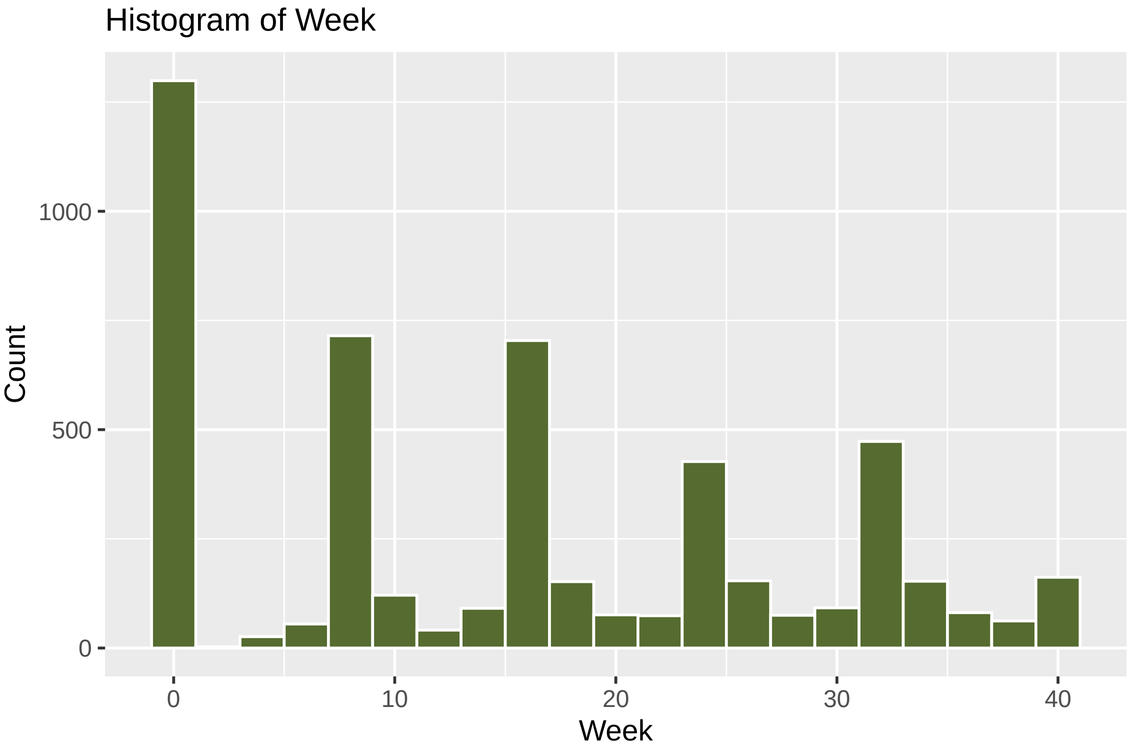
Treatment



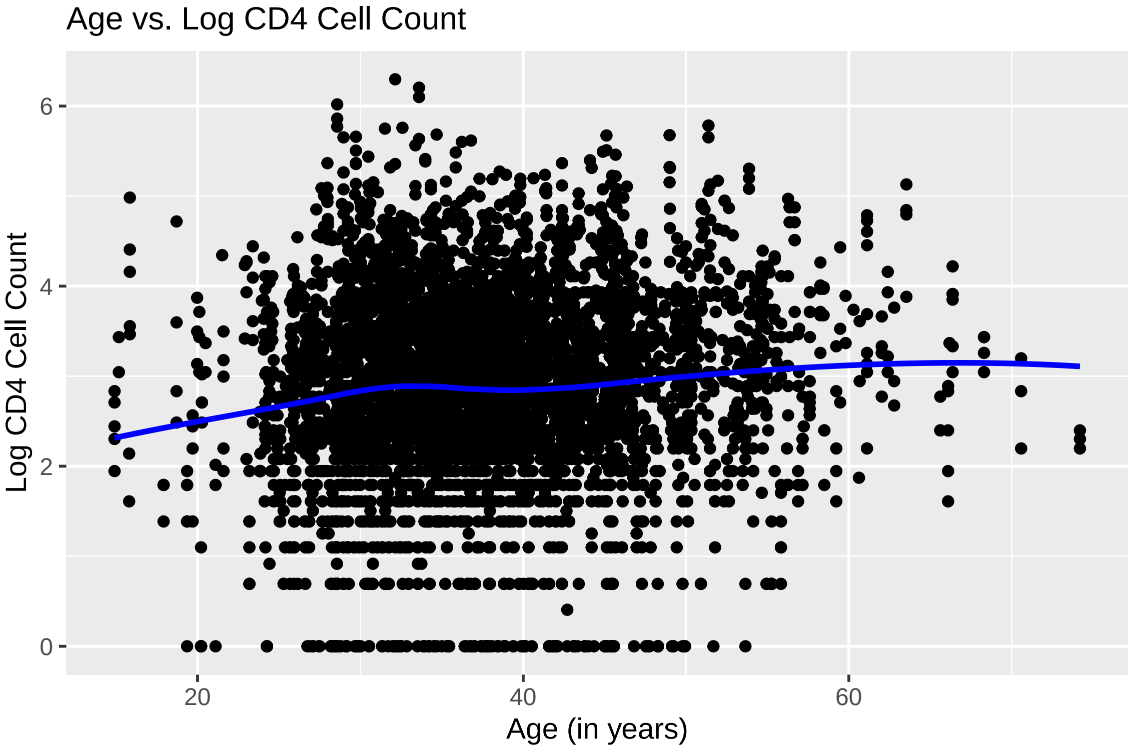
Gender

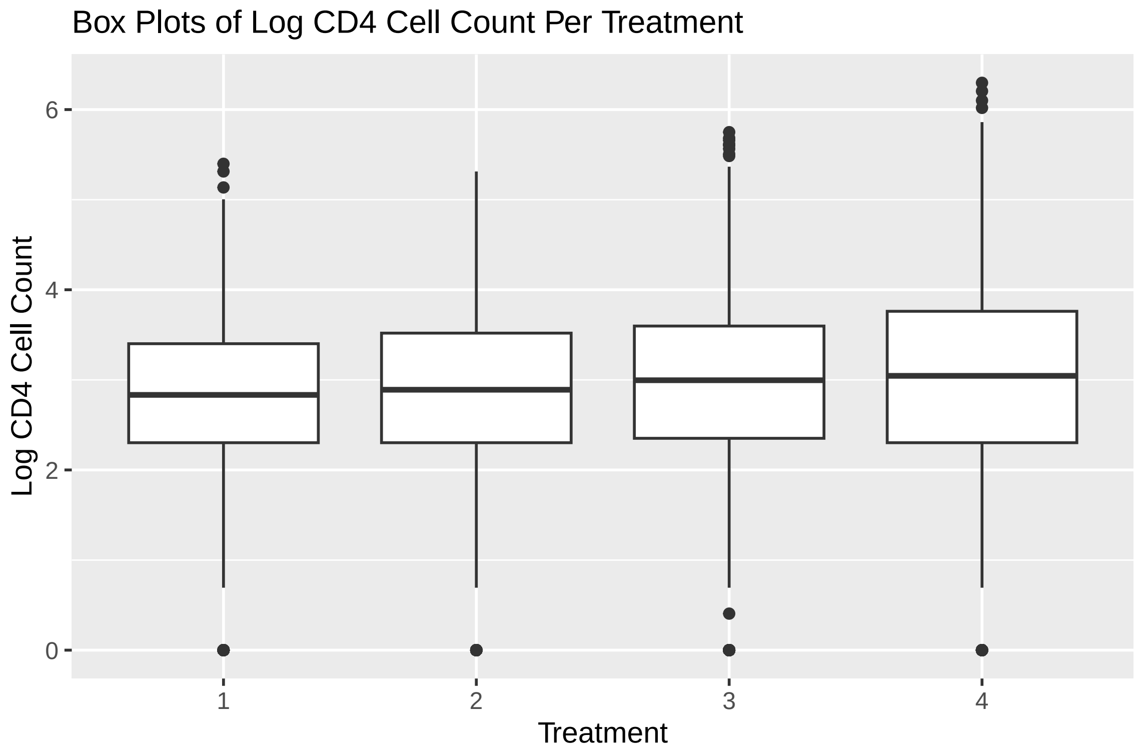


Week

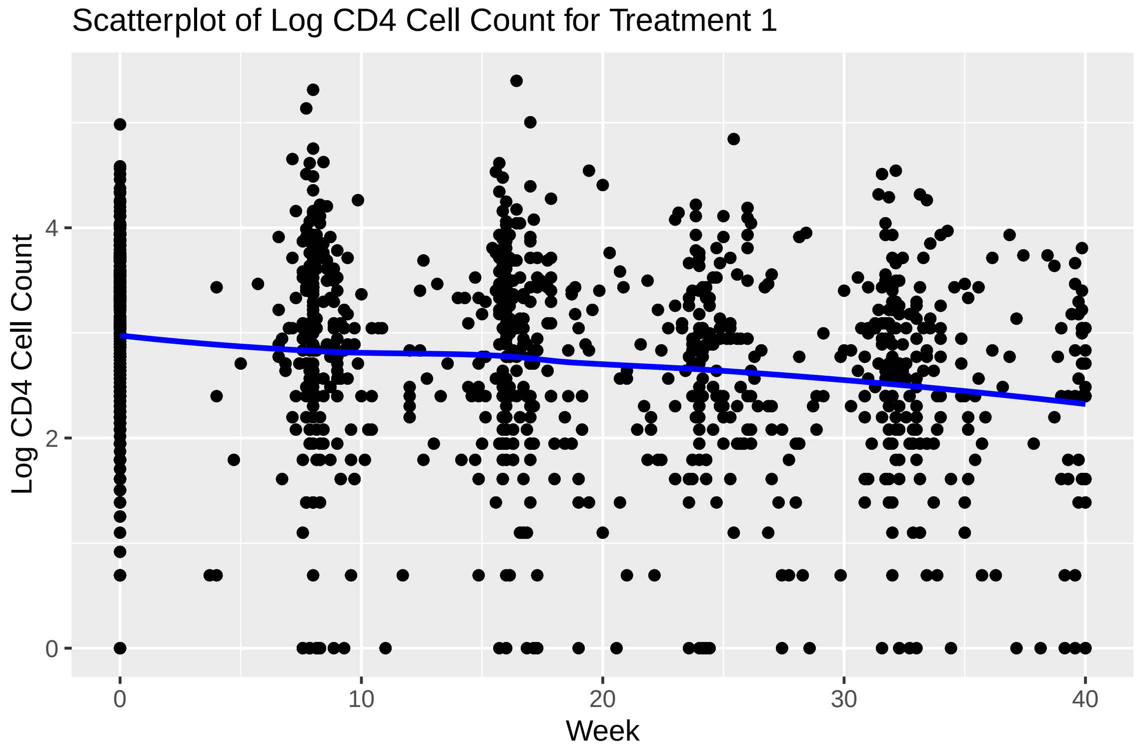


Bivariate summaries

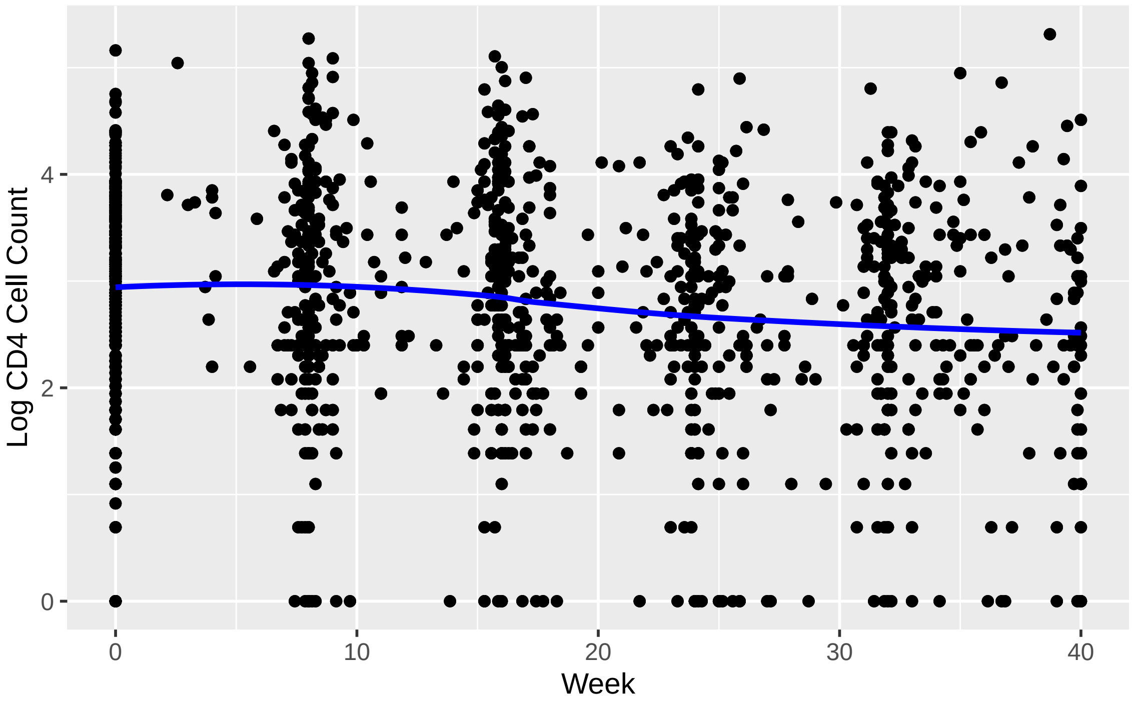




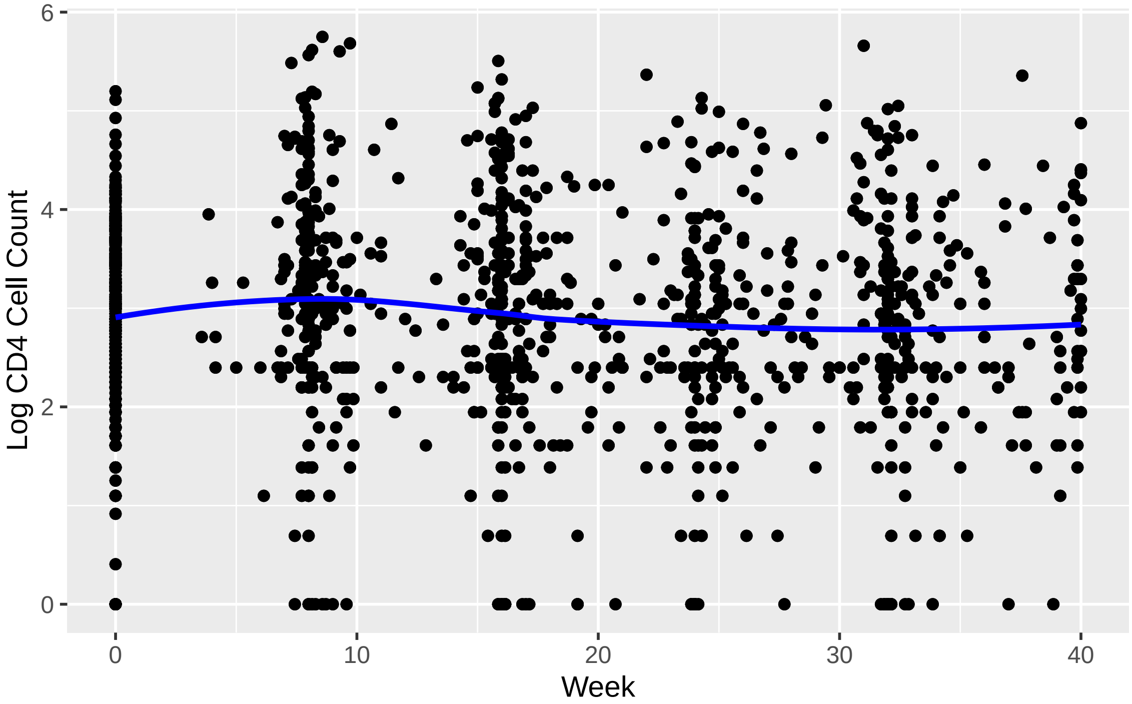
Trend of Log CD4 Cell Count Over Time Across Treatments

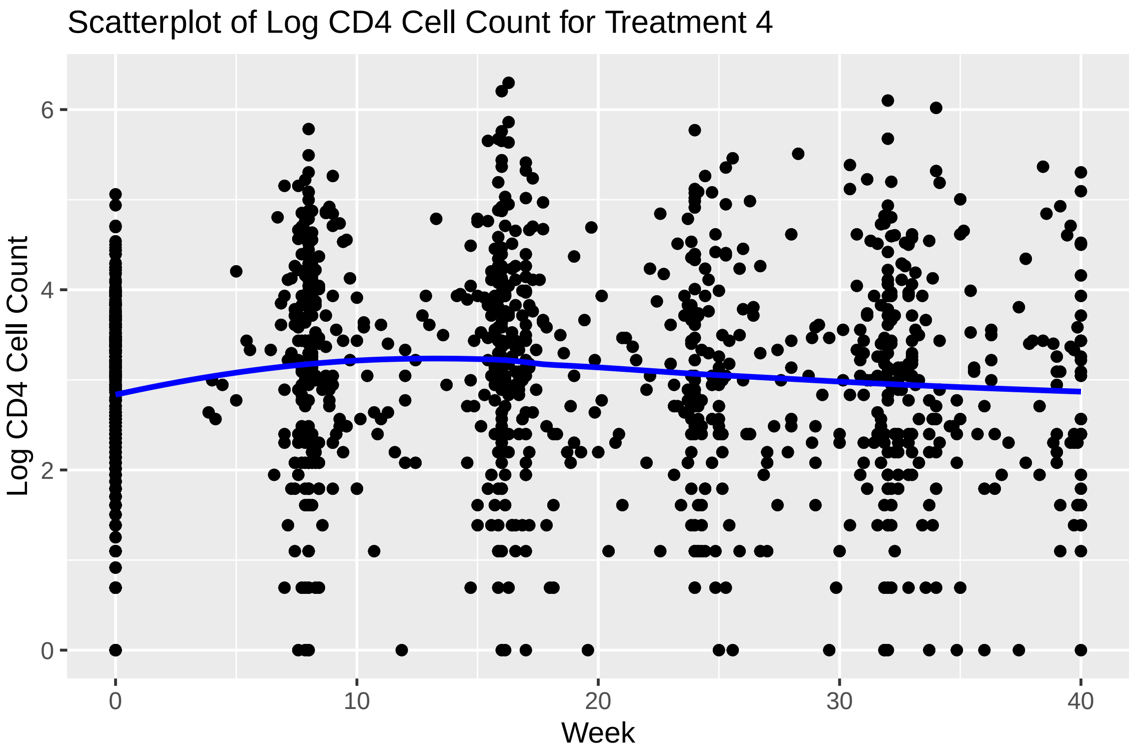


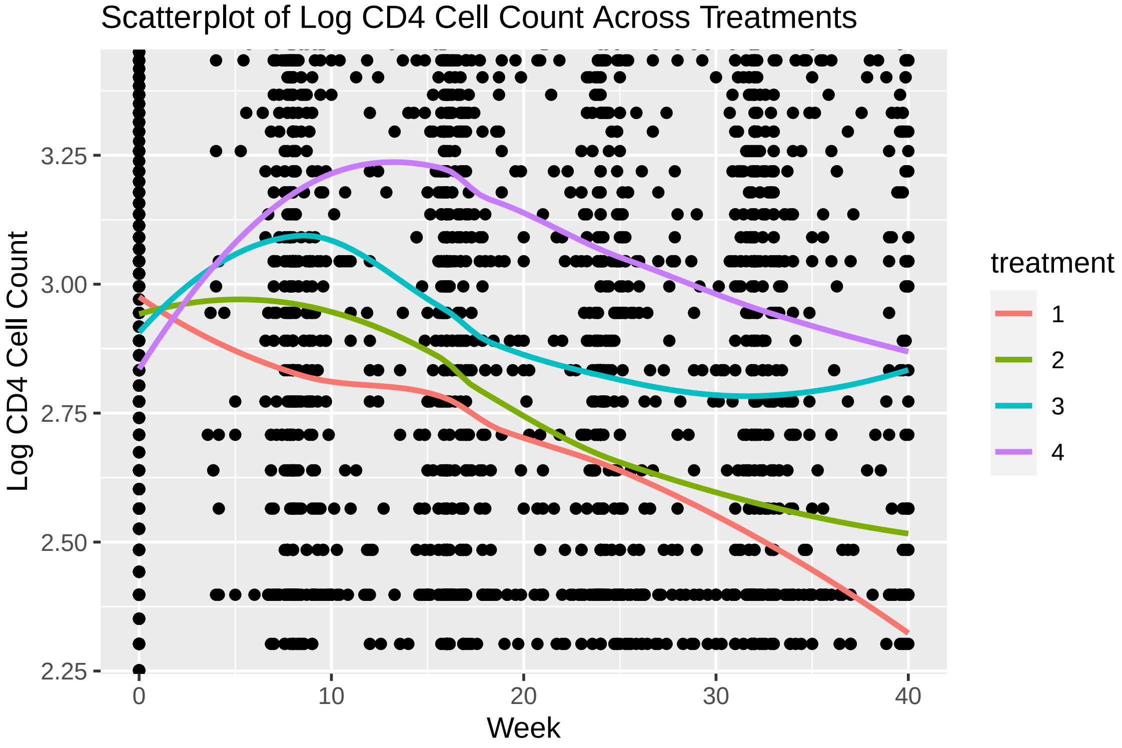
Scatterplot of Log CD4 Cell Count for Treatment 2



Scatterplot of Log CD4 Cell Count for Treatment 3

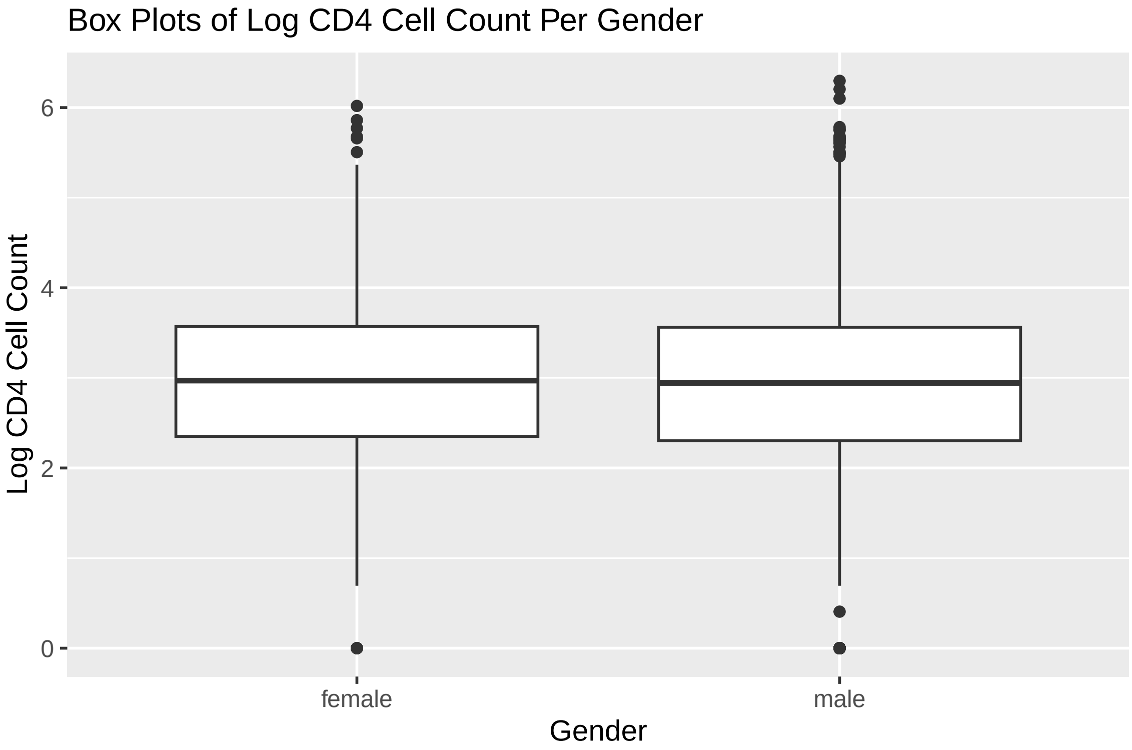




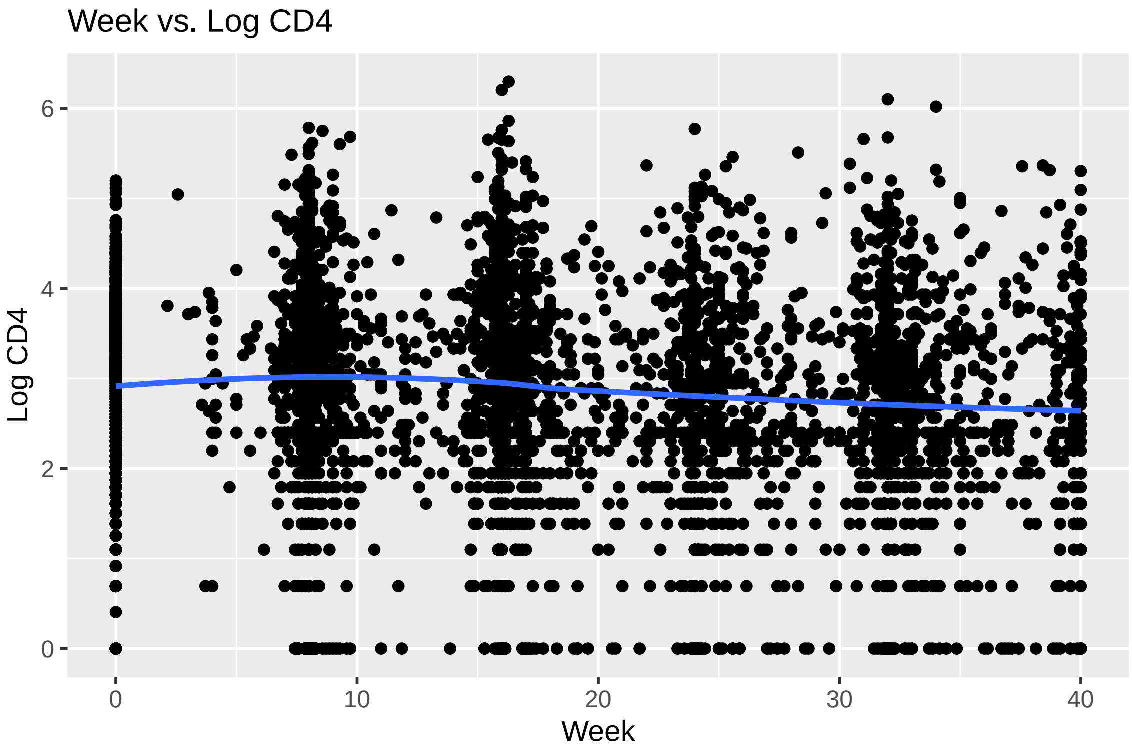


Zooming into the plot, a cubic pattern is identified across all treatments, especially for treatment 3. A cubic term for `Week` variable is necessary in the linear model.

Log CD4 Cell Count Grouped by Gender



Week vs. Log CD4 Over Time



Overall trends

In the scatterplot comparing age and log CD4 cell count, we can see that as age increases over time, log CD4 cell count slightly increases. In the boxplots of log CD4 cell count for each treatment group, we can see that the treatments have similar group means at approximately a log CD4 cell count value of 3. They also have similar sized interquartile ranges and similar sized upper and lower whiskers. Treatment groups 1, 2, and 3 seem to have the most outliers in which treatment group 2 only has one outlier.

In the scatterplot of log CD4 cell count across treatments, groups are similar in the response at baseline and have similar trends over time. However, the treatment 4 group has a slightly higher average of log CD4 cell count over time and the treatment 1 group has a slightly lower average of log CD4 cell count over time (relative to other treatment groups). The smoothed curves suggest that there is a curvilinear relationship between `Week` and the log CD4 cell count, implying that adding a cubic term should be considered.

In the boxplots of log CD4 cell count for each gender, the boxplots between females and males are very similar- in mean, interquartile range, whiskers, and outliers.

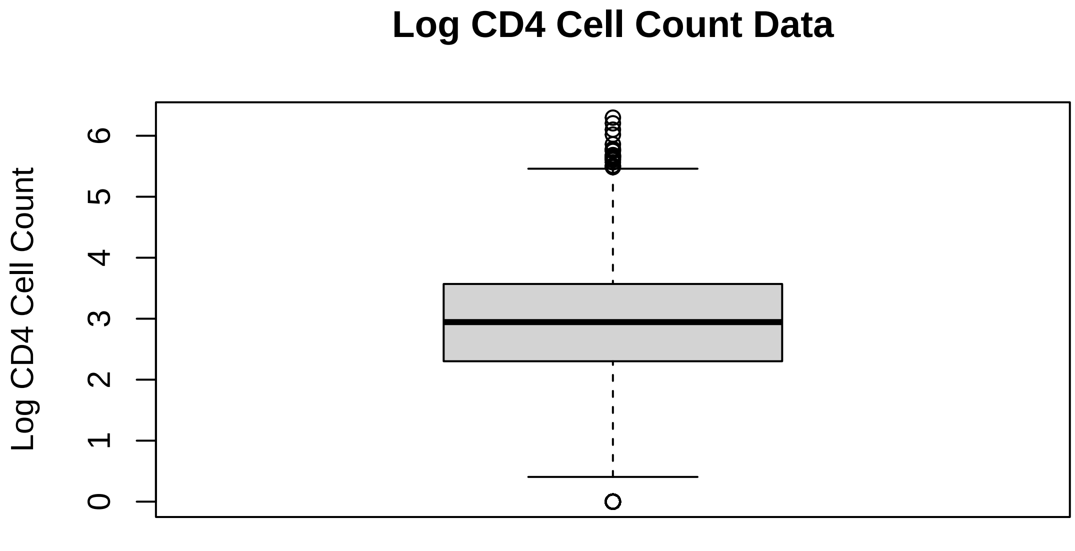
In the scatterplot comparing week and log CD4 cell count, we can see that as weeks progress over time, log CD4 cell count slightly decreases. At baseline, the average log CD4 cell count value was approximately 3 and at the end of the study, the average log CD4 cell count value decreased to approximately 2.75.

Imbalances in the dataset

This is an unbalanced study. Not all subjects in the study had the same number of observations obtained at a common set of occasions. In other words, times of measurement were not common to all subjects, mostly because of mistimed measurements. The subjects were also unbalanced when comparing the different genders. There was a much larger number of males than females in the study.

Outliers in the dataset

Based on the boxplot below, there seems to be a few outliers. There are a few which take on log CD4 cell count values above approximately 5.5 and one outlier which takes on a log CD4 cell count value close to 0.



Generalized Least Squares (GLS) & Linear Mixed Effects (LME) Models

Since this is a randomized experiment, we assume that each treatment group has similar baseline values of log CD4 cell count. In other words, they have similar intercepts, which is why the main treatment effect is not included in any of the following models.

LME Model with No Interaction Term

- model_no_inter: $logCD4_i = \beta_1 + \beta_2 age_i + \beta_3 gender_i + \beta_4 week_i + \beta_5 weekcube_i + b_{1i} + b_{4i} week + b_{5i} weekcube_i$
  - if gender = male then gender = 1, else gender = 0 (female)

LME Model with Interaction Term (week:treatment)

- model_inter: $logCD4_i = \beta_1 + \beta_2 age_i + \beta_3 gender_i + \beta_4 week_i + \beta_5 weekcube_i + \beta_6 week_i * treatment2_i + \beta_7 week_i * treatment3_i + \beta_8 week_i * treatment4_i + \beta_9 weekcube_i * treatment2_i + \beta_{10} weekcube_i * treatment3_i + \beta_{11} weekcube_i * treatment4_i + b_{1i} + b_{4i} week_i + b_{5i} weekcube_i$
  - if gender = male then gender = 1, else gender = 0 (female)

  - treatment_1 is the reference group

Testing Significance of Interaction Term (week:treatment)

```
##                Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model_no_inter     1 12 11993.68 12071.97 -5984.839
## model_inter        2 18 11938.75 12056.19 -5951.377 1 vs 2 66.92427  <.0001
```

- **H_0: Reduced model with all covariates and no week:treatment interaction term.**

- **H_1: Full model with week:treatment interaction term.**

- We compared a reduced model with all covariates and no week:treatment interaction term to a full model with a week:treatment interaction term. Our p-value was <0.0001, and at the 5% significance level, we concluded that the interaction term was significant. The effect of `Week` on log CD4 cell count differs across treatment groups. Since the week:treatment interaction term is significant, we decided to include it in the model.
- In the following anova tests, we will be testing the significance of the age and gender covariates.

LME Model without Age Covariate (including week:treatment interaction)

- model_no_age: $logCD4_i = \beta_1 + \beta_2 gender_i + \beta_3 week_i + \beta_4 weekcube_i + \beta_5 week_i * treatment2_i + \beta_6 week_i * treatment3_i + \beta_7 week_i * treatment4_i + \beta_8 weekcube_i * treatment2_i + \beta_9 weekcube_i * treatment3_i + \beta_{10} weekcube_i * treatment4_i + b_{1i} + b_{3i} week_i + b_{4i} weekcube_i$

LME Model without Gender Covariate (including week:treatment interaction)

- model_no_gender: $logCD4_i = \beta_1 + \beta_2 age_i + \beta_3 week_i + \beta_4 weekcube_i + \beta_5 week_i * treatment2_i + \beta_6 week_i * treatment3_i + \beta_7 week_i * treatment4_i + \beta_8 weekcube_i * treatment2_i + \beta_9 weekcube_i * treatment3_i + \beta_{10} weekcube_i * treatment4_i + b_{1i} + b_{3i} week_i + b_{4i} weekcube_i$

LME without Week Covariate (including week:treatment interaction)

- model_no_week: $logCD4_i = \beta_1+\beta_2 gender_i+\beta_3 week_i*treatment2_i+\beta_4 week_i*treatment3_i+\beta_5 week_i*treatment4_i+\beta_6 weekcube_i*treatment2_i+\beta_7 weekcube_i*treatment3_i+\beta_8 weekcube_i*treatment4_i+b_{1i}$

Testing Significance of Individual Covariates

Testing Significance of Age Covariate

```
##              Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model_no_age     1 17 11947.91 12058.82 -5956.953
## model_inter      2 18 11938.75 12056.19 -5951.377 1 vs 2 11.15197  0.0008
```

- **H_0: Reduced model with no age covariate.**
- **H_1: Full model with week:treatment interaction term.**
- We compared a reduced model with no age covariate and the full model. With a p-value of 0.0008, we conclude that the full model is better than the reduced model and the age covariate is significant.

Testing Significance of Gender Covariate

```
##                 Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model_no_gender     1 17 11937.97 12048.88 -5951.984
## model_inter         2 18 11938.75 12056.19 -5951.377 1 vs 2 1.215017  0.2703
```

- **H_0: Reduced model with no gender covariate.**
- **H_1: Full model with week:treatment interaction term.**
- We compared a reduced model with no gender covariate and the full model. With a p-value of 0.2703, we conclude that the reduced model is good enough and the gender covariate is not significant.

Testing Significance of Week Covariate

```
##               Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model_no_week     1 15 12090.74 12188.60 -6030.368
## model_inter       2 18 11938.75 12056.19 -5951.377 1 vs 2 157.9831  <.0001
```

- **H_0: Reduced model with no week covariate.**
- **H_1: Full model with week:treatment interaction term.**
- We compared a reduced model with no week covariate and the full model. With a p-value < 0.0001, we conclude that the full model is better than the reduced model and the week covariate is significant.

Linear Model

- According the results of the anova tests above, the week:treatment interaction covariate, age covariate, and week covariate are significant and the gender covariate is insignificant. Thus, for our linear model, we decided that the `model_no_gender` was the best fit, which also includes a random intercept and random slope for the week covariate.
- model_no_gender: $logCD4_i = \beta_1+\beta_2 age_i+\beta_3 week_i+\beta_4 weekcube_i+\beta_5 week_i*treatment2_i+\beta_6 week_i*treatment3_i+\beta_7 week_i*treatment4_i+\beta_8 weekcube_i*treatment2_i+\beta_9 weekcube_i*treatment3_i+\beta_{10} weekcube_i*treatment4_i+b_{1i}+b_{3i} week_i+b_{4i} weekcube_i$

Linear Spline Model

- model_spline: $logCD4_i = \beta_1+\beta_2 age_i+\beta_3 week_i+\beta_4 knotterm_i+\beta_5 week_i*treatment2_i+\beta_6 week_i*treatment3_i+\beta_7 week_i*treatment4_i++\beta_8 knotterm_i*treatment2_i+\beta_9 knotterm_i*treatment3_i+\beta_{10} knotterm_i*treatment4_i+b_{1i}+b_{3i} week_i+b_{4i} knotterm.$
    - if week $>$ 16 then knot_term = week, else knot_term = 0.

    - if gender = male then gender = 1, else gender = 0 (female).

```
##                  Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model_no_gender      1 17 11937.97 12048.88 -5951.984                        
## model_spline         2 18 11911.46 12028.89 -5937.728 1 vs 2 28.51294  <.0001
```

- **H_0: Reduced model (linear model) is adequate.**
- **H_1: Reduced model (linear model) is inadequate; we need the full model (linear spline model).**
- From the plots in the Exploratory Data Analysis, patterns of change in log CD4 cell count over time somewhat resemble 2 separate linear trends across all treatment groups. We decided to add a knot term at week 16. We then fit a linear spline model and conducted an anova test comparing the reduced model (linear model) to the full model (linear spline model). With a p-value < 0.0001, we conclude that the full model is better than the reduced model. That is, the linear spline model is a better fit than the linear model.

Quadratic Model

- model_quadratic: $logCD4_i = \beta_1 + \beta_2 age_i + \beta_3 week_i + \beta_4 weeksq_i + \beta_5 week_i * treatment2_i + \beta_6 week_i * treatment3_i + \beta_7 week_i * treatment4_i + +\beta_8 weeksq_i * treatment2_i + \beta_9 weeksq_i * treatment3_i + \beta_{10} weeksq_i * treatment4_i + b_{1i} + b_{3i} week_i + b_{4i} weeksq_i.$

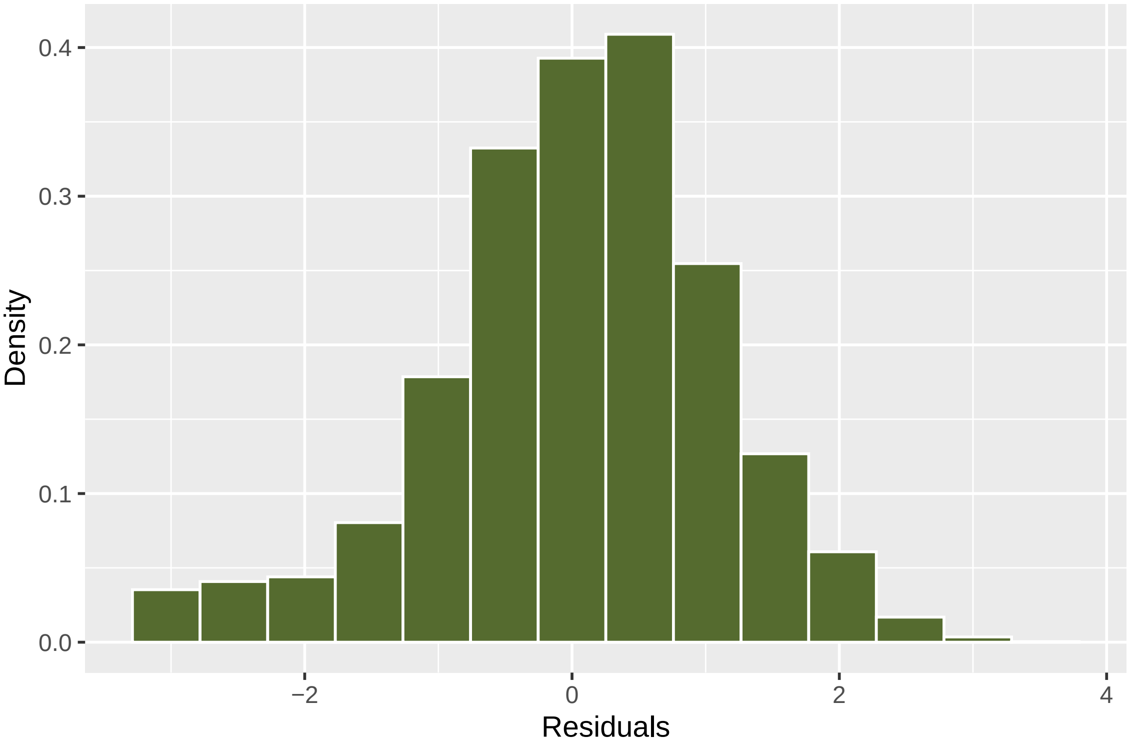
```
##                  Model df      AIC      BIC    logLik
## model_spline         1 18 11911.46 12028.89 -5937.728
## model_quadratic      2 18 11895.49 12012.93 -5929.747
```

```
##                  Model df      AIC      BIC    logLik
## model_spline         1 18 11911.46 12028.89 -5937.728
## model_quadratic      2 18 11895.49 12012.93 -5929.747
```

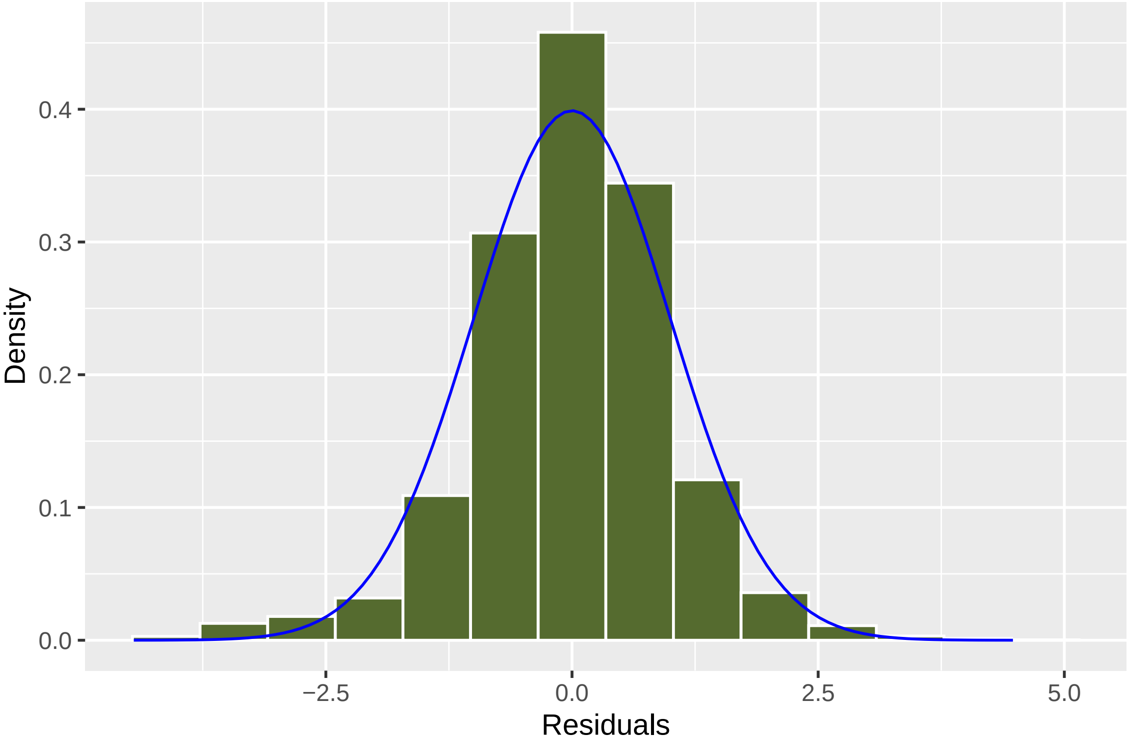
Although the piecewise linear and quadratic trends models are not nested, they both have the same number of parameters and therefore their log-likelihoods can be directly compared. From a comparison of the log-likelihoods ($\hat{l}_{modelquadratic} > \hat{l}_{modelspline}$), it is evident that the qudratic model fits these data better than the linear spline model.

Residual Analysis

Histogram of untransformed residuals



Histogram of transformed residuals



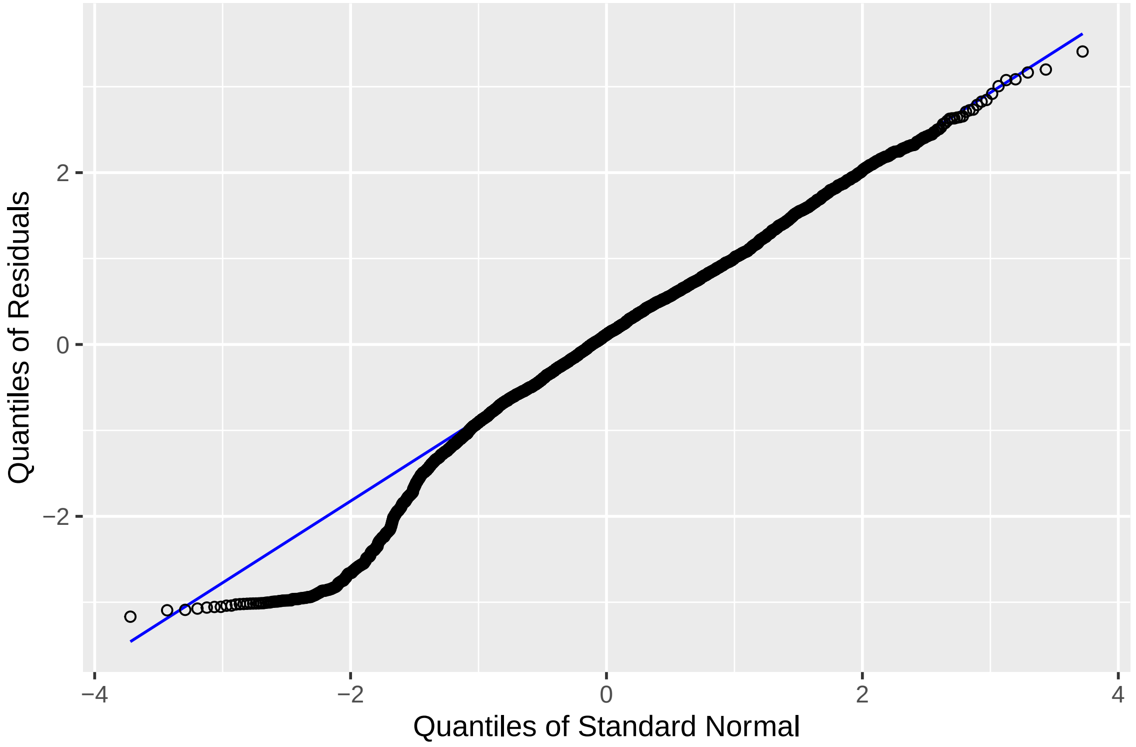
The above figures present the histograms of the transformed and untransformed residuals, and they do not indicate any noticeable skewness.

```
## # A tibble: 1,309 x 5
## # Groups:   id [1,309]
##        id data                df     d      p_value
##     <dbl> <list>           <dbl> <dbl>        <dbl>
##  1   178 <tibble [5 x 1]>     5  39.7 0.000000173
##  2   692 <tibble [5 x 1]>     5  37.5 0.000000486
##  3  1118 <tibble [5 x 1]>     5  33.4 0.00000307
##  4   371 <tibble [2 x 1]>     2  23.9 0.00000642
##  5  1193 <tibble [4 x 1]>     4  28.3 0.0000111
##  6  1100 <tibble [3 x 1]>     3  23.8 0.0000278
##  7  1207 <tibble [5 x 1]>     5  27.9 0.0000382
##  8   362 <tibble [5 x 1]>     5  27.3 0.0000498
##  9   877 <tibble [6 x 1]>     6  29.3 0.0000542
## 10  1110 <tibble [6 x 1]>     6  28.7 0.0000688
## # i 1,299 more rows
```

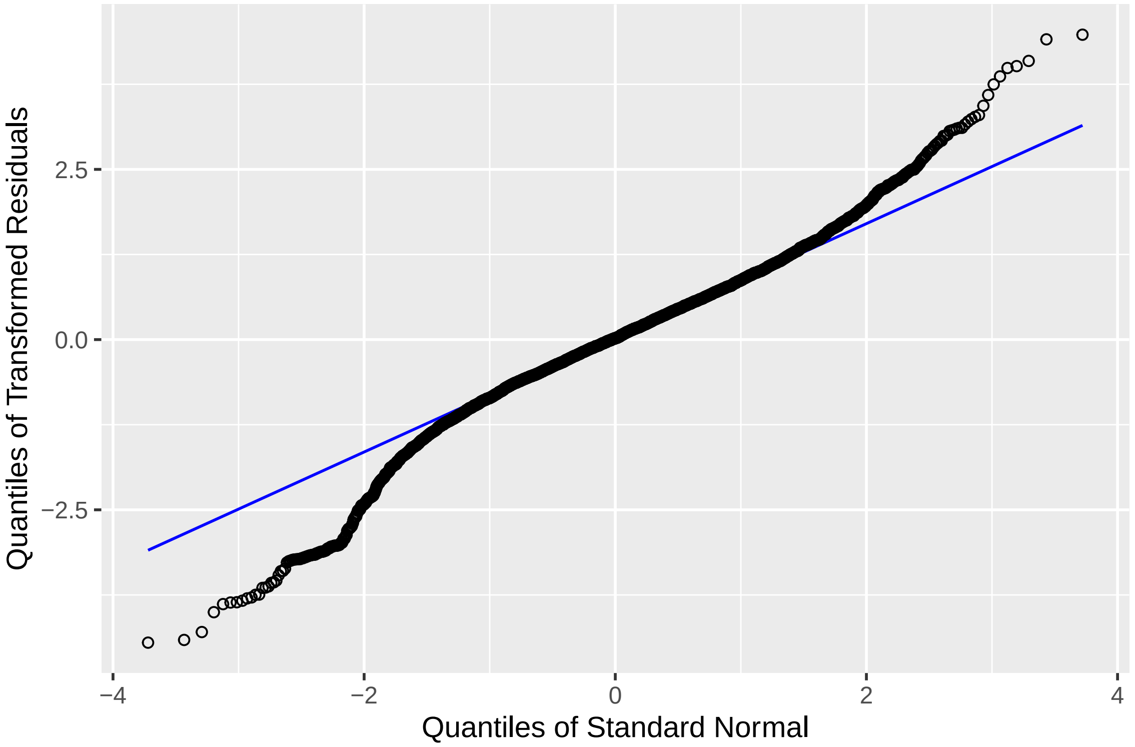
Calculating the associated p-values, there were 137 individuals whose d_i yielded p-values less than 0.05. Given that the sample is comprised of 1313 individuals, we would expect to see only about (1313 * 0.05) = 65.65 individuals with extreme values that happen by random chance. This suggests that the distances of these magnitudes are not expected due to chance alone.

QQ Plot

QQ plot for untransformed residual (untransformed is not specified)



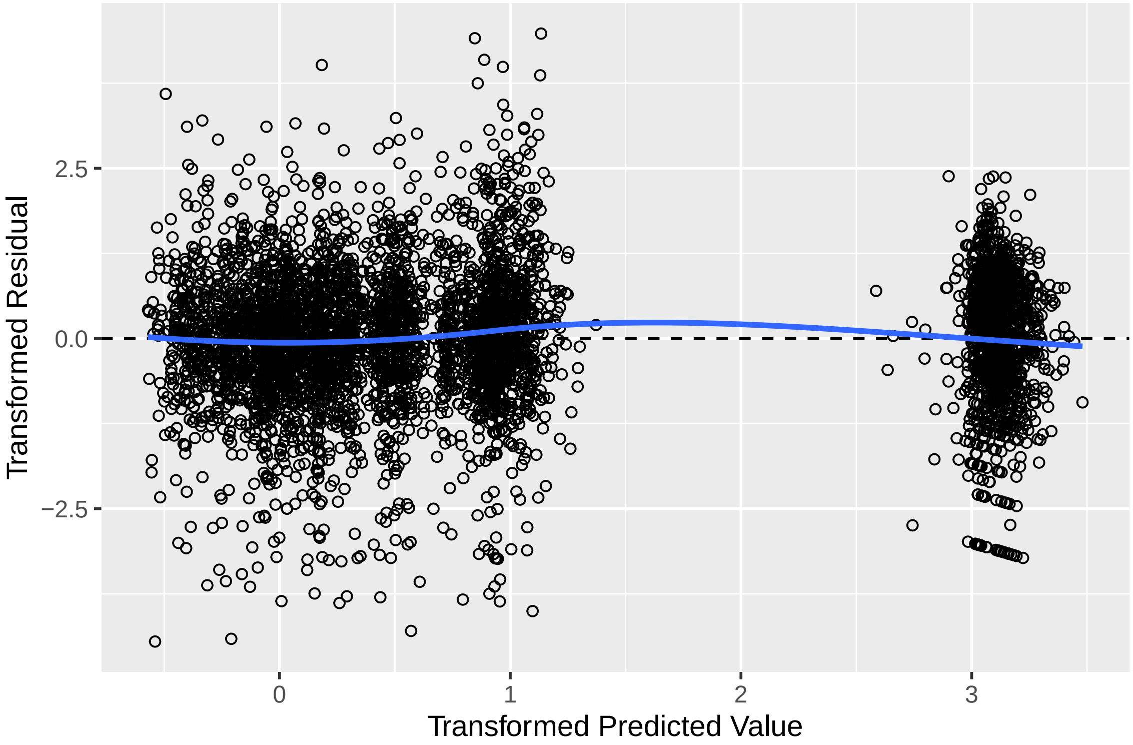
QQ plot for transformed residual (transformed is specified)



The following Q-Q plot of the residuals do display some systematic departures from a straight line around the lower tail.

Scatter Plots (Predicted Values)

Transformed residuals



GEE & GLME

When using the cd4 count and attempting to fit the glmer() model, our ideal model was not converging. We believe that working with the transformed log_cd4 and lme() model is a better option.

Conclusion

Through our analysis of the different models and variables we found that the quadratic model fit the data the best compared to a linear model. However, while this model did contain more outliers than when we expected when we were analyzing, we decided it was the best fit for log_CD4. When testing for different linear models we concluded that gender was not a significant variable in the data. But when we tested for the quadratic model, we found that a full model containing all of the variables was needed. Lastly, when attempting to fit the model using glmer() the quadratic model was not converging, therefore we decided to keep the response variable to be log_CD4 instead of CD4 counts.