Using Longitudinal Data Analysis Methods to Model Different Drug Treatments Among Advanced Stage AIDS Patients in a Randomized, Double-Blind Study

Lynn Gao, Jay Mantuhac, Alec Santiago

STATS 112/203: Statistical Methods for Data Analysis III

Professor Lernik Asserian

Spring 2021

Introduction

Acquired Immunodeficiency Syndrome (AIDS) is a condition known for the gradual destruction of the immune system resulting from a prolonged untreated infection by the Human Immunodeficiency Virus (HIV). Across the globe, in 2019, approximately 690,000 people died from illnesses that are complicated by AIDS, although this has dropped from a peak of 1.1 million AIDS-related deaths in 2010 (HIV.gov 2020b). Upon infection, HIV's mechanism involves attacking the host's CD4 T lymphocytes, also known as CD4 cells, which act as an integral part of the host's immune system. As a result of this mechanism, the method for assessing when an individual has transitioned from an HIV infection to the development of AIDS is the count of CD4 cells in blood reaching a level below 200 cells/mm^3 (HIV.gov 2020a).

Since the AIDS epidemic of the 1980s, developments in the medical field have allowed the perception of AIDS to change. Rather than being known as a guaranteed "death sentence" for all individuals who are infected with HIV, AIDS can now be managed with proper medication prescription and adherence. One such medication used for AIDS treatment is Zidovudine, an orally administered antiretroviral drug that effectively reduces the incidence of opportunistic infections and boosts blood CD4 cell concentrations, which, in turn, prolongs survival among AIDS patients (Langtry and Campoli-Richards 1989). In order to maximize the effectiveness of the medication, however, Zidovudine is typically combined with other medications, such as Didanosine, to further delay disease progression and death (Perry and Noble 1999).

The purpose of this paper is to utilize statistical methods for analyzing longitudinal data to assess for differences in 4 separate different drug treatments that are used in conjunction with Zidovudine in a randomized, double-blind study of AIDS patients with severely compromised immune systems (which correspond to a CD4 count below 50 cells/mm^3 in this study). Throughout the paper, we will refer to the randomized patient treatments as "Treatment 1", "Treatment 2", etc. which correspond to the specific drug treatments seen below.

Treatment 1	Zidovudine alternating monthly with 400 mg Didanosine
Treatment 2	Zidovudine plus 2.25 mg of Zalcitabine
Treatment 3	Zidovudine plus 400 mg of Didanosine
Treatment 4	Zidovudine plus 400 mg of Didanosine plus 400 mg of Nevirapine

Table 1. Types of Treatment Used in the Study

The analysis done from this paper used a provided dataset with the following variables and specifications.

Variable	Data Type	Description
Treatment	Numerical Variable	Values range from 1-4 to represent the 4 treatments of interest
Week	Numerical Variable	Time since baseline (in weeks)
Age	Numerical Variable	Age (in years)
Gender	Character Variable	Contains entries “male” and “female”
log_cd4	Numerical variable	CD4 cell counts, with the log transformation applied

Table 2. Brief description of dataset columns

Exploratory Data Analysis

	Count	Percent
Total	1309	--
Males	1147	87.6%
Females	162	12.4%
Treatment 1	325	24.8%
Treatment 2	324	24.8%
Treatment 3	330	25.2%
Treatment 4	330	25.2%

Table 3. Descriptive summary of Data

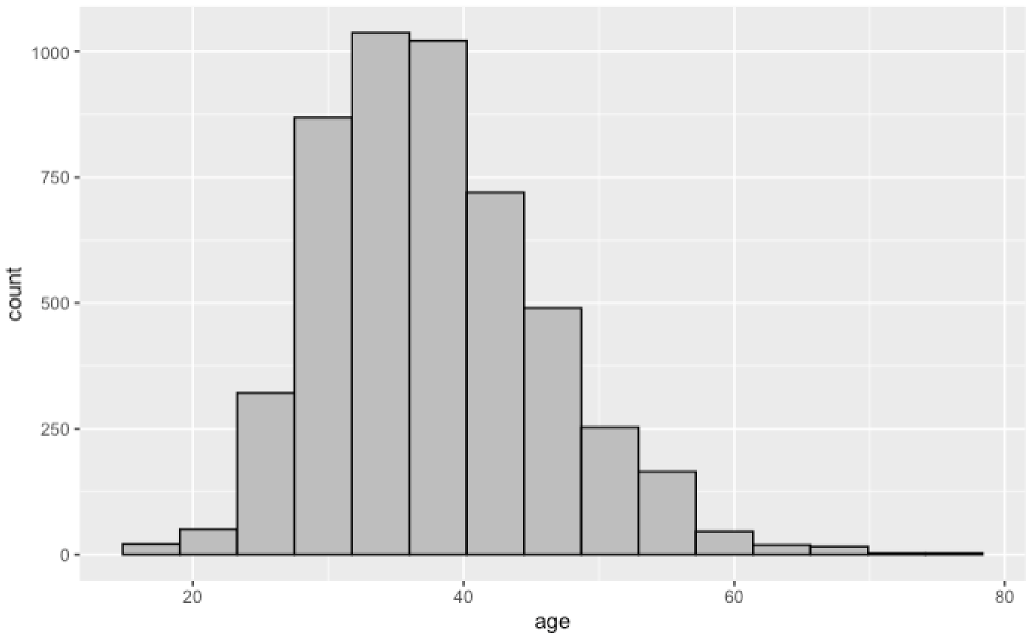


Figure 1. Distribution of Age

The distribution of age shows that it is skewed to the left where most individuals fall between ages 25 to 50. To get a better sense of the age distribution, we provide a five number summary, where we can see that although the maximum age is 74, the median and mean ages are all around 36 and the 3rd quantile is around 42, indicating that 75% of participants are under the age of 42.54.

Min	1st Quant	Median	Mean	3rd Quant	Max
14.90	31.76	36.85	37.73	42.54	74.19

Table 4: Five Number Summary of Age

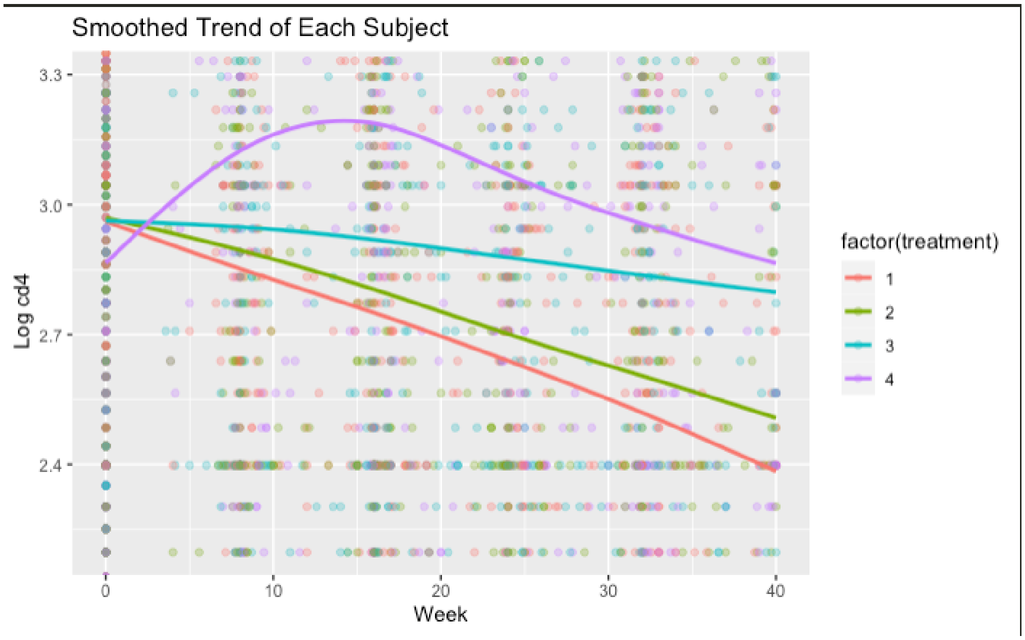


Figure 2. Smoothed spaghetti plot separated by treatment

From the smoothed spaghetti plot, we can see that the treatments are all linear except for treatment 4, which shows a curved relationship. This trend potentially indicates that a linear or quadratic spline model for treatment 4 and a linear model for all other treatments may work best for modeling log(CD4) cell counts.

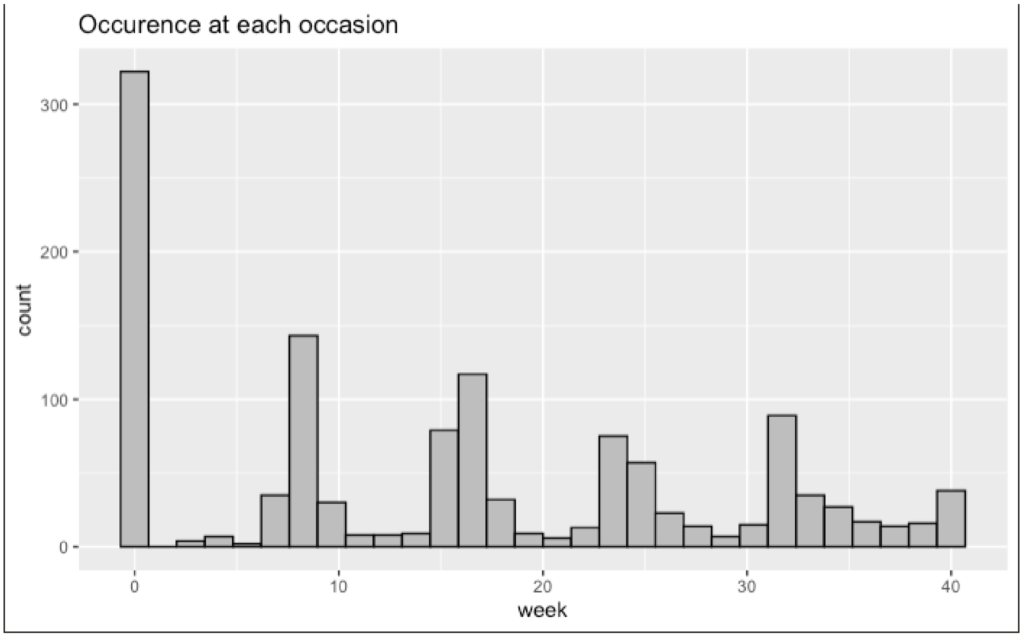
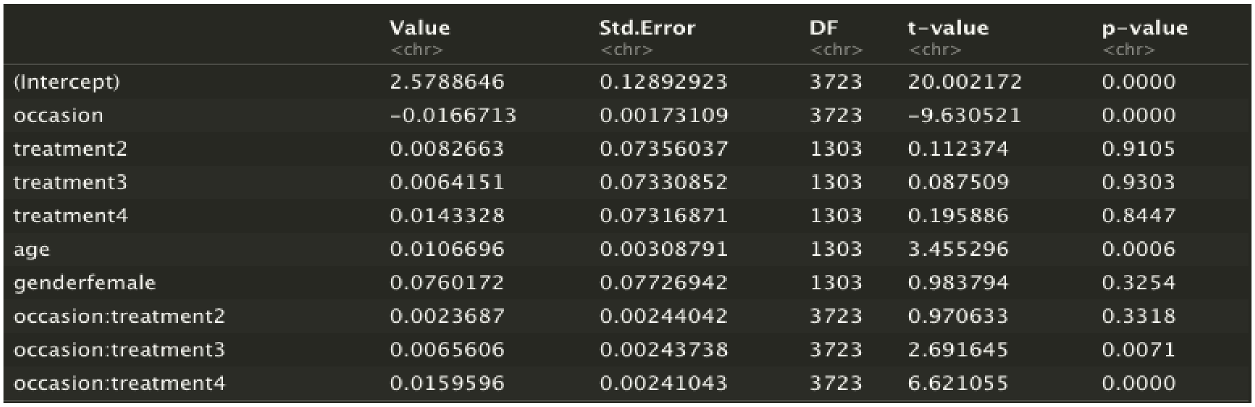


Figure 3. Histogram of data points by the week of study

From this histogram we can see that the data is unbalanced, since at week zero (or baseline) we have about 300+ participants, but as the weeks increase, the number of participants decrease, which could be due to a multitude of factors, including death, participant loss to follow up, and participants voluntarily dropping out of the study.

GLM Modeling



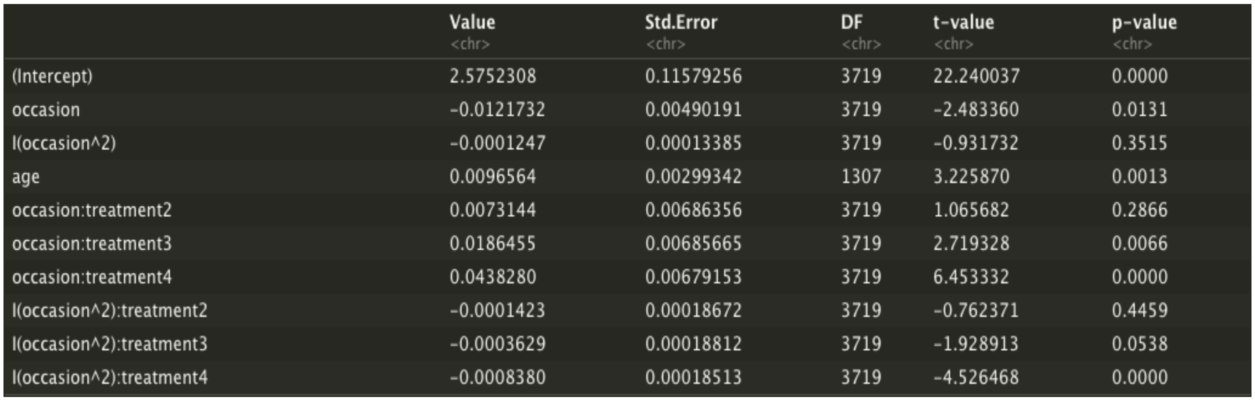
	Value	Std.Error	DF	t-value	p-value
	<chr>	<chr>	<chr>	<chr>	<chr>
(Intercept)	2.5788646	0.12892923	3723	20.002172	0.0000
occasion	-0.0166713	0.00173109	3723	-9.630521	0.0000
treatment2	0.0082663	0.07356037	1303	0.112374	0.9105
treatment3	0.0064151	0.07330852	1303	0.087509	0.9303
treatment4	0.0143328	0.07316871	1303	0.195886	0.8447
age	0.0106696	0.00308791	1303	3.455296	0.0006
genderfemale	0.0760172	0.07726942	1303	0.983794	0.3254
occasion:treatment2	0.0023687	0.00244042	3723	0.970633	0.3318
occasion:treatment3	0.0065606	0.00243738	3723	2.691645	0.0071
occasion:treatment4	0.0159596	0.00241043	3723	6.621055	0.0000

Figure 4. Output of the linear model with all the covariates plus interaction terms

The experiment is randomized and double-blind, so any group effect would be insignificant. This means that the treatment and gender will be insignificant when added to the model. The models that will be constructed are variations of the linear model shown in Figure 4.

Quadratic Model:

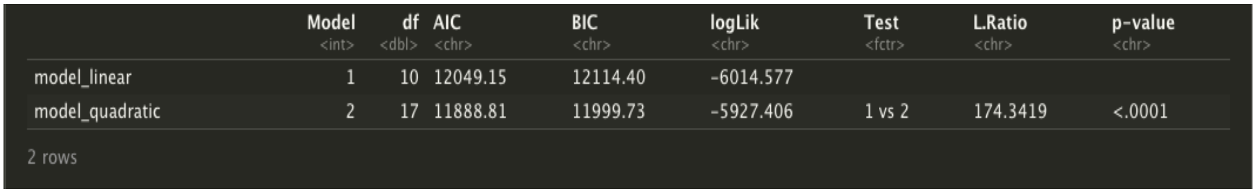
$$E(Y_{ij}|b_i) = \beta_1 + \beta_2 * occasion_{ij} + \beta_3 * occasion_{ij}^2 + \beta_4 * age_{ij} +$$
$$\beta_5 * occasion_{ij} : treatment_2 + \beta_6 * occasion_{ij} : treament_3 +$$
$$\beta_7 * occasion_{ij} : treament_4 + \beta_8 * occasion_{ij}^2 : treament_2 +$$
$$\beta_9 * occasion_{ij}^2 : treament_3 + \beta_{10} * occasion_{ij}^2 : treament_4 +$$
$$b_1 + b_2 * occasion_{ij} + b_3 * occasion_{ij}^2$$



	Value <chr>	Std.Error <chr>	DF <chr>	t-value <chr>	p-value <chr>
(Intercept)	2.5752308	0.11579256	3719	22.240037	0.0000
occasion	-0.0121732	0.00490191	3719	-2.483360	0.0131
I(occasion^2)	-0.0001247	0.00013385	3719	-0.931732	0.3515
age	0.0096564	0.00299342	1307	3.225870	0.0013
occasion:treatment2	0.0073144	0.00686356	3719	1.065682	0.2866
occasion:treatment3	0.0186455	0.00685665	3719	2.719328	0.0066
occasion:treatment4	0.0438280	0.00679153	3719	6.453332	0.0000
I(occasion^2):treatment2	-0.0001423	0.00018672	3719	-0.762371	0.4459
I(occasion^2):treatment3	-0.0003629	0.00018812	3719	-1.928913	0.0538
I(occasion^2):treatment4	-0.0008380	0.00018513	3719	-4.526468	0.0000

Figure 5. Output of the quadratic model with quadratic interaction terms

Figure 5 shows a summary of the quadratic model. Notice that the occasion^2, occasion:treatment2, occasion^2:treatment2, and occasion^2:treatment3 are all not significant. This could be because in the previous plot we see that the treatments 2 to 3 appear to be linear but treatment 4 is quadratic. Perhaps this is why these values are not significant, and a linear spline model where we have the knot term at occasion equals 16 would be better.



	Model <int>	df <dbl>	AIC <chr>	BIC <chr>	logLik <chr>	Test <fctr>	L.Ratio <chr>	p-value <chr>
model_linear	1	10	12049.15	12114.40	-6014.577			
model_quadratic	2	17	11888.81	11999.73	-5927.406	1 vs 2	174.3419	<.0001

2 rows

Figure 6. ANOVA test comparing the linear model to the quadratic model

We compared our proposed linear model and our proposed quadratic model using ANOVA with the following hypotheses:

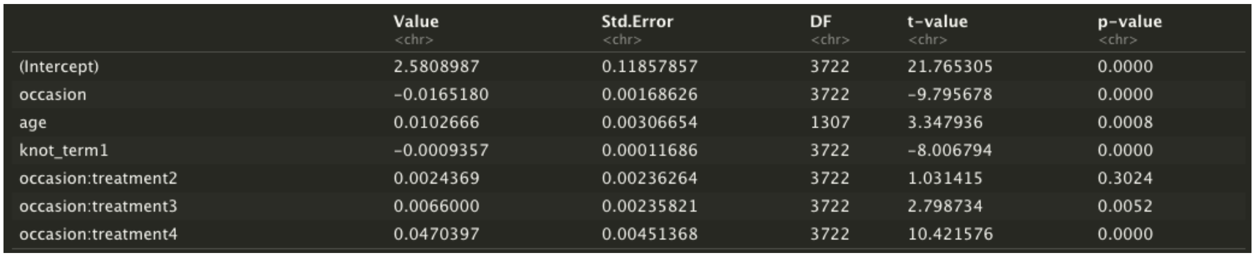
H$_0$: Reduced model (Linear model) is better

H$_a$: Full model (Quadratic model) is better

The p-value is less than 0.0001, which provides evidence that our full model (Quadratic model) is more adequate, and should be used to model the log(CD4) counts.

Linear-Splines Model:

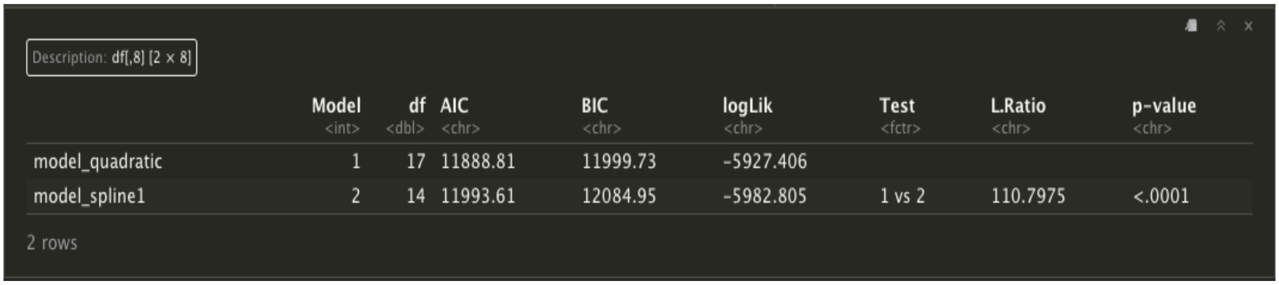
$$E(Y_{ij}|b_i) = \beta_1 + \beta_2 occasion_{ij} + \beta_3 age_{ij} + \beta_4 (occasion_{ij})_+ + \beta_5 occasion_{ij} treatment2 + \beta_6 occasion_{ij} treatment3 + \beta_7 occasion_{ij} treatment4 + b_{1i} + b_{2i} occasion_{ij} + b_{3i} (occasion_{ij})_+$$



	Value	Std.Error	DF	t-value	p-value
	<chr>	<chr>	<chr>	<chr>	<chr>
(Intercept)	2.5808987	0.11857857	3722	21.765305	0.0000
occasion	-0.0165180	0.00168626	3722	-9.795678	0.0000
age	0.0102666	0.00306654	1307	3.347936	0.0008
knot_term1	-0.0009357	0.00011686	3722	-8.006794	0.0000
occasion:treatment2	0.0024369	0.00236264	3722	1.031415	0.3024
occasion:treatment3	0.0066000	0.00235821	3722	2.798734	0.0052
occasion:treatment4	0.0470397	0.00451368	3722	10.421576	0.0000

Figure 7. Output of the linear splines model

The linear spline model summary (Figure 7) shows that almost all of its values are significant, except for treatment2. While in the smoothed spaghetti plot of log(CD4) count disaggregated by treatment (Figure 2), treatment appears to follow a linear trend, this model output potentially says otherwise. This model can potentially perform better in comparison to other models due to the presence of fewer insignificant covariates.



Description: df[,8] [2 × 8]

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
	<int>	<dbl>	<chr>	<chr>	<chr>	<fctr>	<chr>	<chr>
model_quadratic	1	17	11888.81	11999.73	-5927.406			
model_spline1	2	14	11993.61	12084.95	-5982.805	1 vs 2	110.7975	<.0001

2 rows

Figure 8. ANOVA test on the quadratic model vs. the linear spline model

Using ANOVA we assessed for differences between our proposed linear spline model and our proposed quadratic model using the following hypotheses:

H$_0$: The reduced model (Linear spline) is better

H_a: The full model (Quadratic model) is better

The p-value is <.0001, which provides evidence that the quadratic model is a better fit for our data compared to the linear spline model.

This analysis has revealed that the best performing model is the quadratic model out of all the other models. This will be the model that will be used with the residual analysis and creating linear mixed effect models.

Summary of Linear Mixed Random effect models (using REML estimates):

Model	AIC	BIC
Linear spline model (random intercept, occasion, knot term)	12068.92	12160.24
Quadratic model (random intercept, occasion, and occasion^2)	12012.15	12123.03
Linear spline model (random intercept, occasion)	12062.81	12134.57
Linear model with no group effects (random intercept and occasion)	12108.2	12173.44
Linear model with group effects (random intercept)	12129.35	12220.66

Table 5: Performance of each Model

Residual Analysis

For our residual analysis, we first start with a Histogram of Transformed Residuals, which will allow us to assess the normality of the residuals.

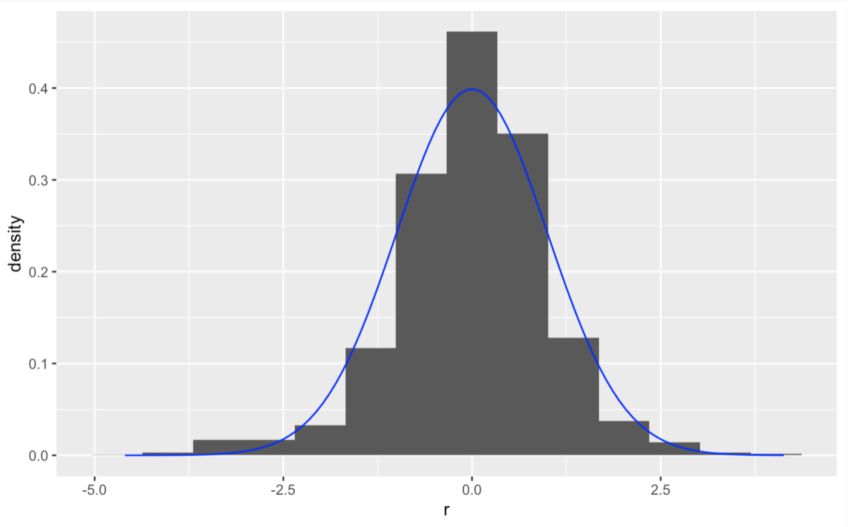


Figure 9. Histogram of transformed residual density

Our histogram of transformed residuals (Figure 9) seem to follow a Normal distribution, or at the very least, a symmetric distribution.

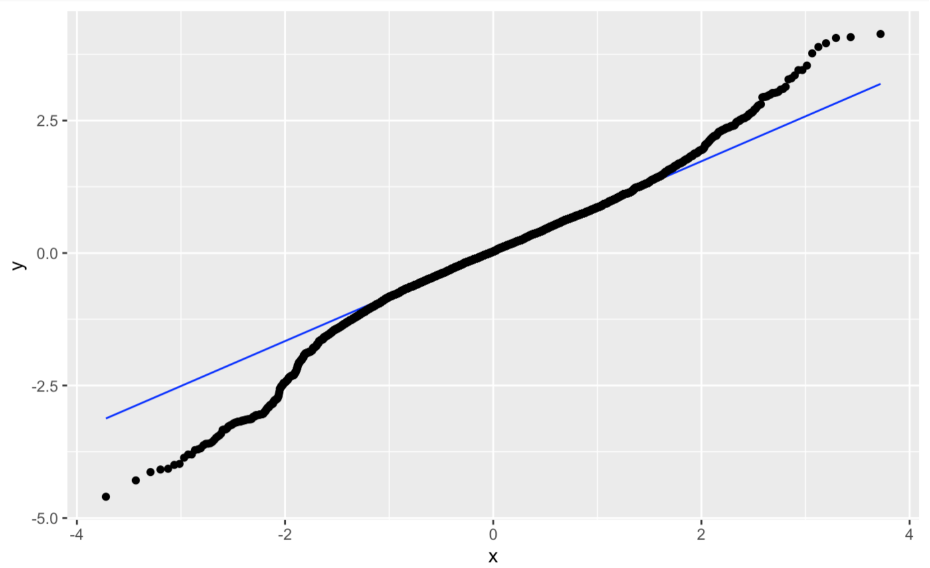


Figure 10. QQ Plot of the transformed residuals

The QQ plot (Figure 10) reveals the tails for the distribution of our residuals are quite heavy. This brings into question whether or not the residuals actually follow a normal distribution. Furthermore, several lingering points at the end of each tail suggest that we may have outliers present in our dataset.

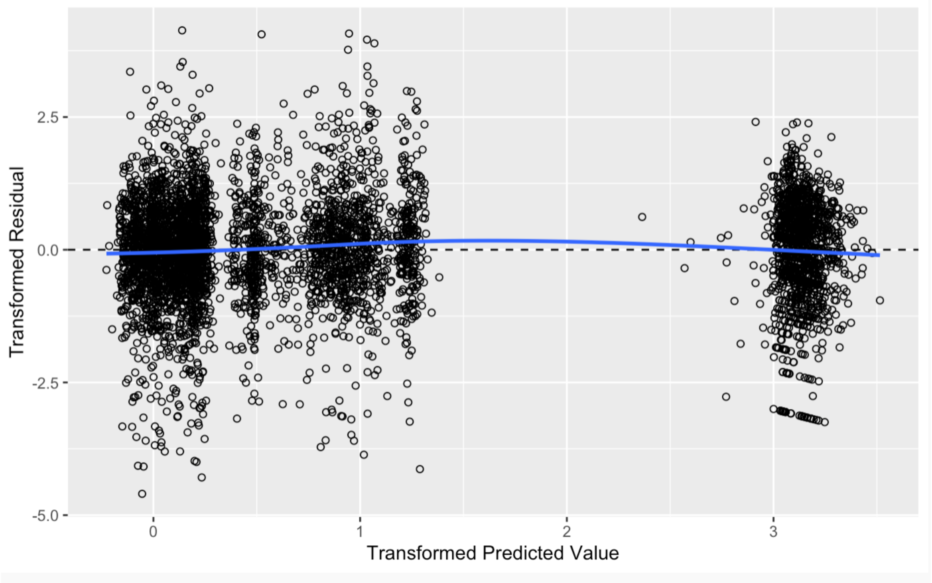


Figure 11. Transformed residual vs. Transformed predicted plot

In addition, we can see in the predicted value vs. residual plot (Figure 11) that there doesn't seem to be any significant curvature in this graph, indicating that the constant variance assumption is correct. However, we can see that there is a large gap between the points, which shows that there seems to be a gap in the observed covariates (x values).

From the Transformed Predicted values vs. Absolute Transformed Residuals graph (Figure 12), using the loess smoothed curve, we can see that there is not a noticeable departure from the straight line centered at approximately 0.8. The smoothed curve is relatively straight and is centered around 0.8. This plot indicates that the model for the variance (and covariance) is adequate.

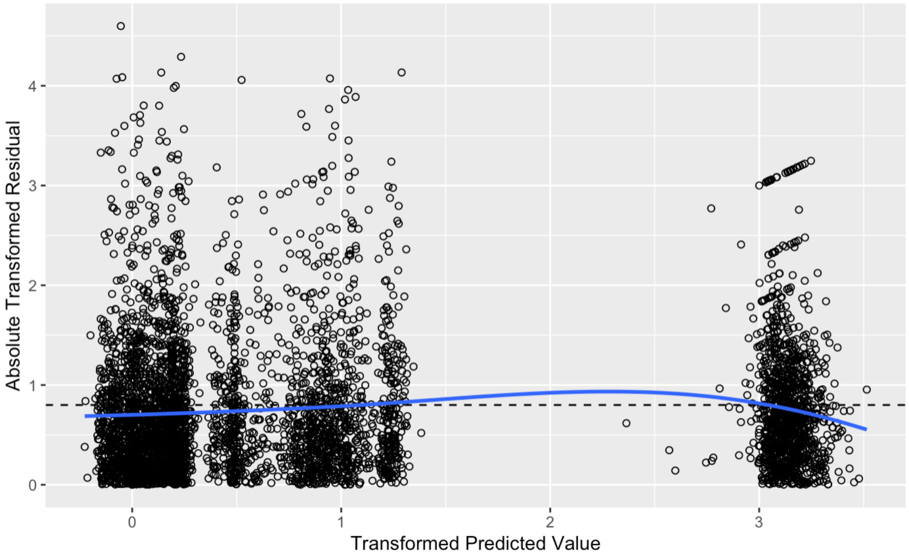


Figure 12. Absolute transformed residual vs. Transformed predicted plot

When we analyzed the Mahalanobis Data, we found that there are approximately 129 outliers in this data set (subjects who have a p-value < 0.05). From the size of our data, we expect that we will have 251.8 outliers. Our actual 129 outliers fall within the range of 251.8, so we do not need to be concerned about the outliers we find here and see in the QQ-plot.

Looking at the semi-variogram in Figure 13 below, we can see that the loess smoothed curve seems to fluctuate randomly around 1.0, but has a general decreasing trend. Although this graph may not fluctuate randomly around the y = 1.0 line, we can still infer that the model's covariance matrix is adequate.

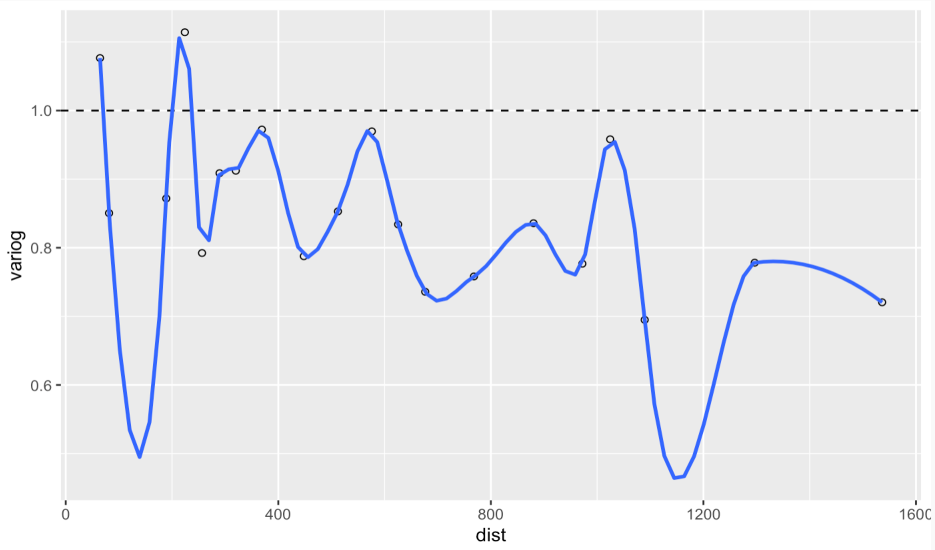


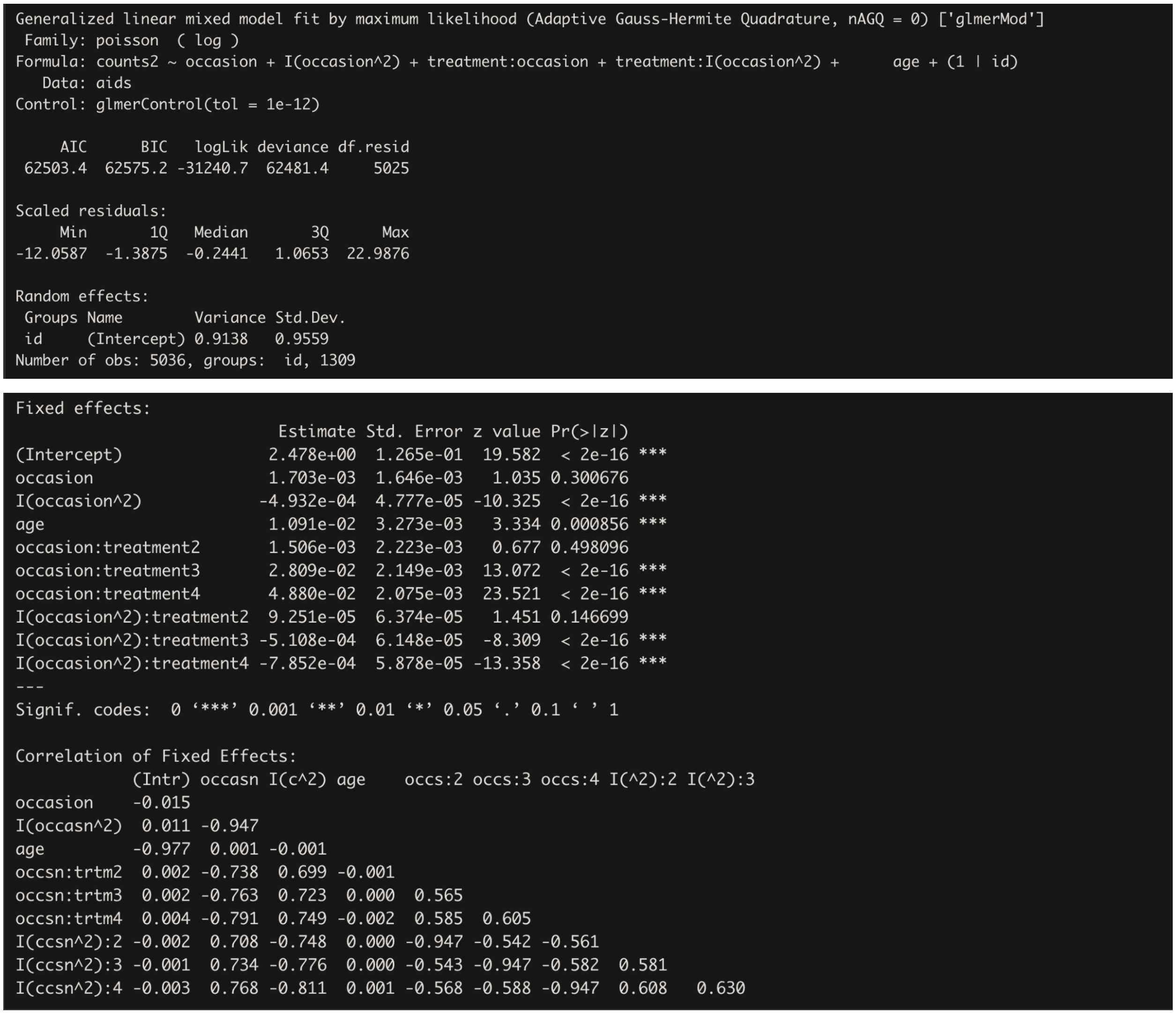
Figure 13. Semi-Variogram

GLME Modeling

We proposed two different Generalized Linear Mixed Effect Poisson models to model the actual count of CD4 cells (which were transformed from the log(CD4) values originally provided in the dataset), one with only a random intercept and another with a random intercept and a random slope. Since we could not add random slopes for quadratic variables, we decided to implement random slopes only for our linear occasion variables.

GLME Model with random slope:

$$log(E(Y_{ij}|b_i)) = \beta_1 + \beta_2 * occasion_{ij} + \beta_3 * occasion_{ij}^2 + \beta_4 * age_{ij} + \beta_5 * occasion_{ij} : treatment_2 + \beta_6 * occasion_{ij} : treament_3 + \beta_7 * occasion_{ij} : treament_4 + \beta_8 * occasion_{ij}^2 : treament_2 + \beta_9 * occasion_{ij}^2 : treament_3 + \beta_{10} * occasion_{ij}^2 : treament_4 + b_1$$



```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature, nAGQ = 0) ['glmerMod']
 Family: poisson  ( log )
Formula: counts2 ~ occasion + I(occasion^2) + treatment:occasion + treatment:I(occasion^2) +      age + (1 | id)
   Data: aids
Control: glmerControl(tol = 1e-12)

     AIC      BIC   logLik deviance df.resid
 62503.4  62575.2 -31240.7  62481.4     5025

Scaled residuals:
     Min       1Q   Median       3Q      Max
-12.0587  -1.3875  -0.2441   1.0653  22.9876

Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 0.9138   0.9559
Number of obs: 5036, groups:  id, 1309
```

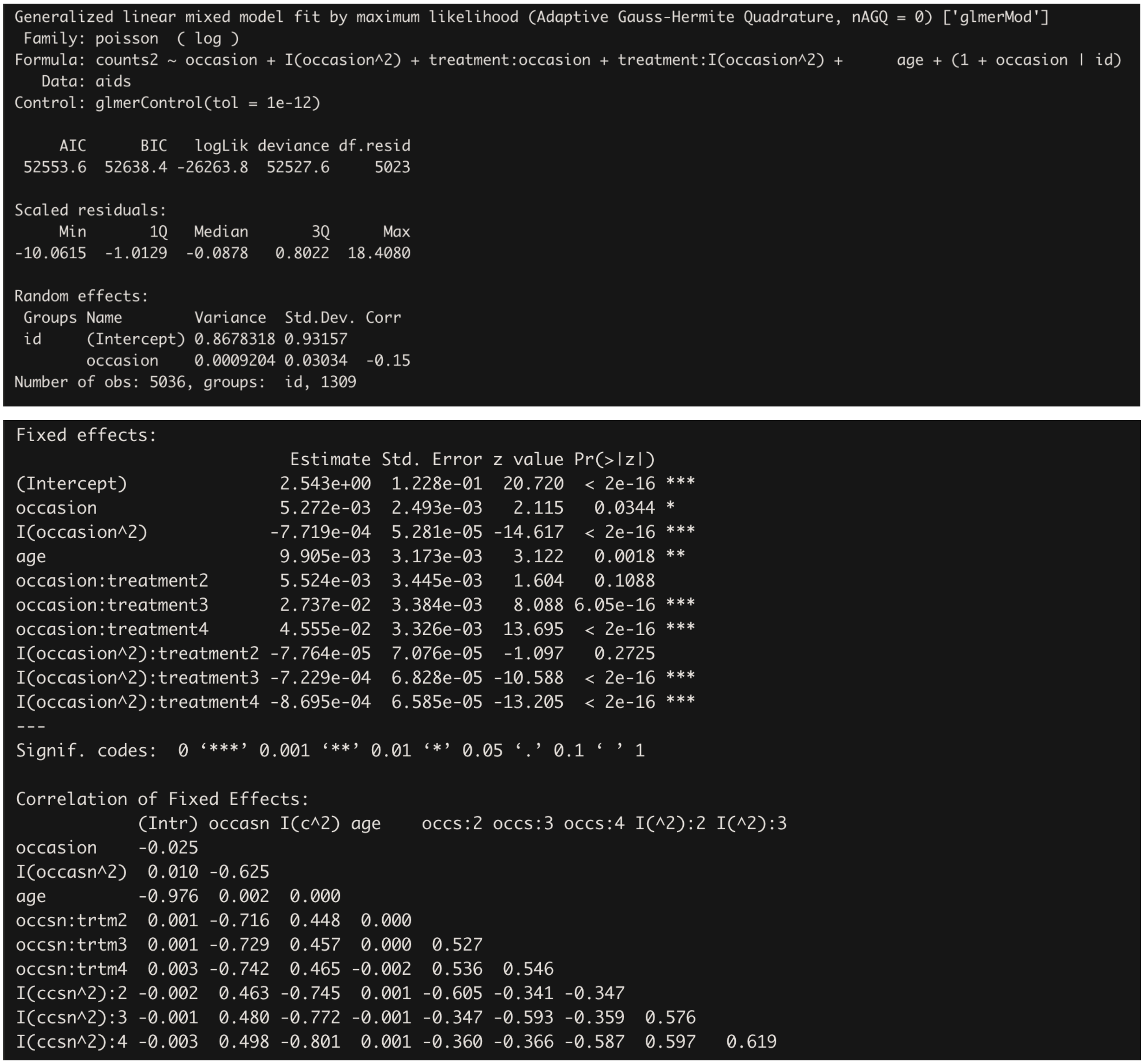
```
Fixed effects:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              2.478e+00  1.265e-01  19.582  < 2e-16 ***
occasion                 1.703e-03  1.646e-03   1.035 0.300676
I(occasion^2)           -4.932e-04  4.777e-05 -10.325  < 2e-16 ***
age                      1.091e-02  3.273e-03   3.334 0.000856 ***
occasion:treatment2      1.506e-03  2.223e-03   0.677 0.498096
occasion:treatment3      2.809e-02  2.149e-03  13.072  < 2e-16 ***
occasion:treatment4      4.880e-02  2.075e-03  23.521  < 2e-16 ***
I(occasion^2):treatment2  9.251e-05  6.374e-05   1.451 0.146699
I(occasion^2):treatment3 -5.108e-04  6.148e-05  -8.309  < 2e-16 ***
I(occasion^2):treatment4 -7.852e-04  5.878e-05 -13.358  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) occasn I(c^2) age    occs:2 occs:3 occs:4 I(^2):2 I(^2):3
occasion    -0.015
I(occasn^2)  0.011 -0.947
age         -0.977  0.001 -0.001
occsn:trtm2  0.002 -0.738  0.699 -0.001
occsn:trtm3  0.002 -0.763  0.723  0.000  0.565
occsn:trtm4  0.004 -0.791  0.749 -0.002  0.585  0.605
I(ccsn^2):2 -0.002  0.708 -0.748  0.000 -0.947 -0.542 -0.561
I(ccsn^2):3 -0.001  0.734 -0.776  0.000 -0.543 -0.947 -0.582  0.581
I(ccsn^2):4 -0.003  0.768 -0.811  0.001 -0.568 -0.588 -0.947  0.608   0.630
```

Figure 14. Output of GLE model with random intercept

GLME Model with Random Intercept and Random Slope:

$$log(E(Y_{ij}|b_i)) = \beta_1 + \beta_2 * occasion_{ij} + \beta_3 * occasion_{ij}^2 + \beta_4 * age_{ij} + \beta_5 * occasion_{ij} : treatment_2 + \beta_6 * occasion_{ij} : treament_3 + \beta_7 * occasion_{ij} : treament_4 + \beta_8 * occasion_{ij}^2 : treament_2 + \beta_9 * occasion_{ij}^2 : treament_3 + \beta_{10} * occasion_{ij}^2 : treament_4 + b_1 + b_2 * occasion_{ij}$$



```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature, nAGQ = 0) ['glmerMod']
 Family: poisson  ( log )
Formula: counts2 ~ occasion + I(occasion^2) + treatment:occasion + treatment:I(occasion^2) +      age + (1 + occasion | id)
   Data: aids
Control: glmerControl(tol = 1e-12)

     AIC      BIC   logLik deviance df.resid
 52553.6  52638.4 -26263.8  52527.6     5023

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.0615  -1.0129  -0.0878   0.8022  18.4080

Random effects:
 Groups Name        Variance  Std.Dev. Corr
 id     (Intercept) 0.8678318 0.93157
        occasion    0.0009204 0.03034  -0.15
Number of obs: 5036, groups:  id, 1309
```

```
Fixed effects:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               2.543e+00  1.228e-01  20.720  < 2e-16 ***
occasion                  5.272e-03  2.493e-03   2.115   0.0344 *
I(occasion^2)            -7.719e-04  5.281e-05 -14.617  < 2e-16 ***
age                       9.905e-03  3.173e-03   3.122   0.0018 **
occasion:treatment2       5.524e-03  3.445e-03   1.604   0.1088
occasion:treatment3       2.737e-02  3.384e-03   8.088 6.05e-16 ***
occasion:treatment4       4.555e-02  3.326e-03  13.695  < 2e-16 ***
I(occasion^2):treatment2 -7.764e-05  7.076e-05  -1.097   0.2725
I(occasion^2):treatment3 -7.229e-04  6.828e-05 -10.588  < 2e-16 ***
I(occasion^2):treatment4 -8.695e-04  6.585e-05 -13.205  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) occasn I(c^2) age    occs:2 occs:3 occs:4 I(^2):2 I(^2):3
occasion    -0.025
I(occasn^2)  0.010 -0.625
age         -0.976  0.002  0.000
occsn:trtm2  0.001 -0.716  0.448  0.000
occsn:trtm3  0.001 -0.729  0.457  0.000  0.527
occsn:trtm4  0.003 -0.742  0.465 -0.002  0.536  0.546
I(ccsn^2):2 -0.002  0.463 -0.745  0.001 -0.605 -0.341 -0.347
I(ccsn^2):3 -0.001  0.480 -0.772 -0.001 -0.347 -0.593 -0.359  0.576
I(ccsn^2):4 -0.003  0.498 -0.801  0.001 -0.360 -0.366 -0.587  0.597   0.619
```

Figure 15. Output of GLME Model with random intercept and slope

When comparing these two models, we can see that the AIC for the model with only random intercept is 62503.4, while the AIC for the model with both random intercept and random slope is 52553.6. Because the AIC for the model with both random intercept and random slope is smaller than the model with only the random intercept, we can conclude that the model with both random intercept and random slope is a better fit for our data compared to the model with only a random slope.

However, both GLME Poisson models have convergence issues where having any number of quadrature will make the model fail to converge despite having optimizers. Even after setting the number of quadrature 0 to make the model converge, the GLME Poisson model AIC is about 52553.6, five times higher than the Quadratic LME model for log(cd4) with random intercept and slope. Therefore the Quadratic model is significantly better compared to modeling the actual count of cd4 using the generalized mixed effect poisson model.

Discussion/Conclusion

Overall, we conducted an analysis of the AIDS data, tested different models based on our analysis, and chose the best model by comparing AICs and ANOVA test results. We found that the LME models that we created (that were not count models) performed better than the GLME models that were especially for counts. As such, overall, in order to best model the effect of treatment on the log(CD4) count, we conclude that the ideal model to use is the linear spline model given a random intercept, occasion term, and knot term:

$$E(Y_{ij}|b_i) = \beta_1 + \beta_2 occasion_{ij} + \beta_3 age_{ij} + \beta_4 (occasion_{ij})_+ + \beta_5 occasion_{ij} treatment2 + \beta_6 occasion_{ij} treatment3 + \beta_7 occasion_{ij} treatment4 + b_{1i} + b_{2i} occasion_{ij} + b_{3i} (occasion_{ij})_+$$

One aspect to note about this model we propose is the lack of specific covariate terms for the main effects of all 4 treatments, which is due to the fact the study is randomized. As a result of this study design, the main effect of each treatment is rendered statistically insignificant, and as such, not necessary to incorporate in the final model.

It is important to note that other factors, such as psychosocial stress (Remor et al. 2007) and the presence of other diseases, such as Syphilis (Buchacz et al. 2004), also affect CD4 cell counts. Future research should look into quantifying the effects of such covariates in order to expand on the model that we proposed.

References

Buchacz, Kate, Pragna Patel, Melanie Taylor, Peter R. Kerndt, Robert H. Byers, Scott D. Holmberg, and Jeffrey D. Klausner. 2004. “Syphilis Increases HIV Viral Load and Decreases CD4 Cell Counts in HIV-Infected Patients with New Syphilis Infections.” *AIDS (London, England)* 18 (15): 2075–79. https://doi.org/10.1097/00002030-200410210-00012.

HIV.gov. 2020a. “What Are HIV and AIDS?” HIV.Gov. June 5, 2020. https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/what-are-hiv-and-aids.

———. 2020b. “Global Statistics.” HIV.Gov. November 25, 2020. https://www.hiv.gov/hiv-basics/overview/data-and-trends/global-statistics.

Langtry, Heather D., and Deborah M. Campoli-Richards. 1989. “Zidovudine.” *Drugs* 37 (4): 408–50. https://doi.org/10.2165/00003495-198937040-00003.

Perry, C. M., and S. Noble. 1999. “Didanosine: An Updated Review of Its Use in HIV Infection.” *Drugs* 58 (6): 1099–1135. https://doi.org/10.2165/00003495-199958060-00009.

Remor, Eduardo, Frank J Penedo, Ji Shen, and Neil Schneiderman. 2007. “Perceived Stress Is Associated with CD4+ Cell Decline in Men and Women Living with HIV/AIDS in Spain.” *AIDS Care* 19 (2): 215–19. https://doi.org/10.1080/09540120600645570.