Final Project

Abstract

The artical is mainly a report for the 2020's data analysis exam which analysed a longitudinal data to estimate whether Zidovudine can lower CD4 level or not.

The first step is wrangling data. The provided data consists 6 variables and except age and CD4 are integer, other variables are all factorial. The next step is to viasualize the data so that we can pre-estimate this data. There are plots showing CD4 against gender, age, treatment and time respectively. Also, there is a shadow plot indicating where the missing values distributed in our data. The author found that there are so many missings in CD4 that can cause great error in estimating the effect of Zidovudine afterwards. Then, based on information from the previous plots, the author built linear regression model for each patients and impute missings by the predicted value according to the model.

Finally, by building linear regression model with interaction term and GEE model, the author draw a conclusion that Zidovudine does have a significant effect on lowering the level of CD4.

Contents

1 Introduction **1**

2 Background **1**

3 Data Wrangling **2**

4 Visualizing Data **2**

5 Imputation of Missing Values **8**

6 Estimatee the Effect of Zidovudine **10**

6.1 Linear Regression Model 10

6.2 Generalized Estimating Equation Model 11

7 Summary **12**

8 Reference **13**

1 Introduction

For this final project, I'm going to study for qualifying exam. The data I'm going to work with is from 2020's data analysis exam, which is a longitudinal data evaluating the effect of Zidovudine on CD4 cell counts. Since the original questions need knowledge that I haven't learnt yet(like GEE and LME), so I will visualize data structure of the given data. Then I will impute missing value in CD4. After that, I will use my own method to analyze Zidovudine effect. At last I will try to learn generalized estimating equations and

Table 1: discriptive data

	id	gender	age	trt	time	cd4
	1160006: 14	0: 210	Min. :12.00	0:2184	1 : 265	Min. : 0.0
	1160089: 14	1:3500	1st Qu.:31.00	1:1526	2 : 265	1st Qu.:105.0
	1160091: 14	NA	Median :36.00	NA	3 : 265	Median :180.0
	1160094: 14	NA	Mean :37.02	NA	4 : 265	Mean :194.6
	1160096: 14	NA	3rd Qu.:42.00	NA	5 : 265	3rd Qu.:269.0
	1160100: 14	NA	Max. :63.00	NA	6 : 265	Max. :893.0
	(Other):3626	NA	NA	NA	(Other):2120	NA's :768

answer the first question(i.e. Using generalized estimating equations (GEE) with an appropriately chosen working correlation structure to quantify the effect of treatment with Zidovudine on the trajectory of CD4 counts over time.) raised in the qualifying exam.

In this project, I self learnt 3 packages. Package `naniar` (Tierney 2021) to visualize structure of missing data, `simputation` to impute missing values, `gee` (Hong and Ottoboni 2017) to estimate effect of Zidovudine.

2 Background

The given data is a longitudinal data comes from a study of Harvard AIDS clinical trial group. They evaluated the treatment effect of Zidovudine on CD4 cell counts, which represents the endpoint of HIV positive individuals-lower CD4 counts means shorter times to progress AIDS. I will work on a subset of this data that includes 265 patients' trail information. All individuals included in this subset had CD4 counts above 50 at baseline and longitudinal data on CD4 counts for 14 time points. Measurements were taken at approximately 1 month intervals. Patients might have intermittent missing or dropout with an overall missing rate of 20.2%.

There are 6 variables in the data set: ID, Gender, Age, TRT, Time and CD4. ID is a unique number for each patient and Gender = M/F for male/female. Age is measured in years. TRT is a factor variable: TRT = 1 means the patient received the Zidovudine treatment while Trt = 0 means the patient received placebo. Time is a discrete variable counting from 1 to 14. CD4 represents the measured number for CD4 at each time point.

3 Data Wrangling

```
## Rows: 3,710
## Columns: 6
## $ ID     <dbl> 1160841, 1160841, 1160841, 1160841, 1160841, 1160841, 1160841, ~
## $ Gender <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M"~
## $ Age    <dbl> 36, 36, 36, 36, 36, 36, 36, 36, 36, 36, 36, 36, 36, 36, 35, 35,~
## $ TRT    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Time   <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1, 2, 3, 4, 5, 6~
## $ CD4    <dbl> 328, 533, 500, 520, 403, 480, 437, 319, 228, 194, NA, 112, 146,~
```

The original data is not in tidyverse style of naming variables. It also set Gender as character and other variables as double. Clearly, it is not appropriate. We need to transform these variables into tidyverse style and change their type manually.

First, we changed all columns name into lower case writing. Then, replace all values in gender by factor indicator 0 and 1. At last, set id, gender, trt, time as factor, age as integer. Then, briefly summarize the data, we have Table 1. There are 210 female patients and 3500 male patients participated in this survey.

4 Visualizing Data

Since we want to impute missing value of CD4, we need to find out what variable is correlated with it. In this part, we will use plot to detect relationship between CD4 and other variables.

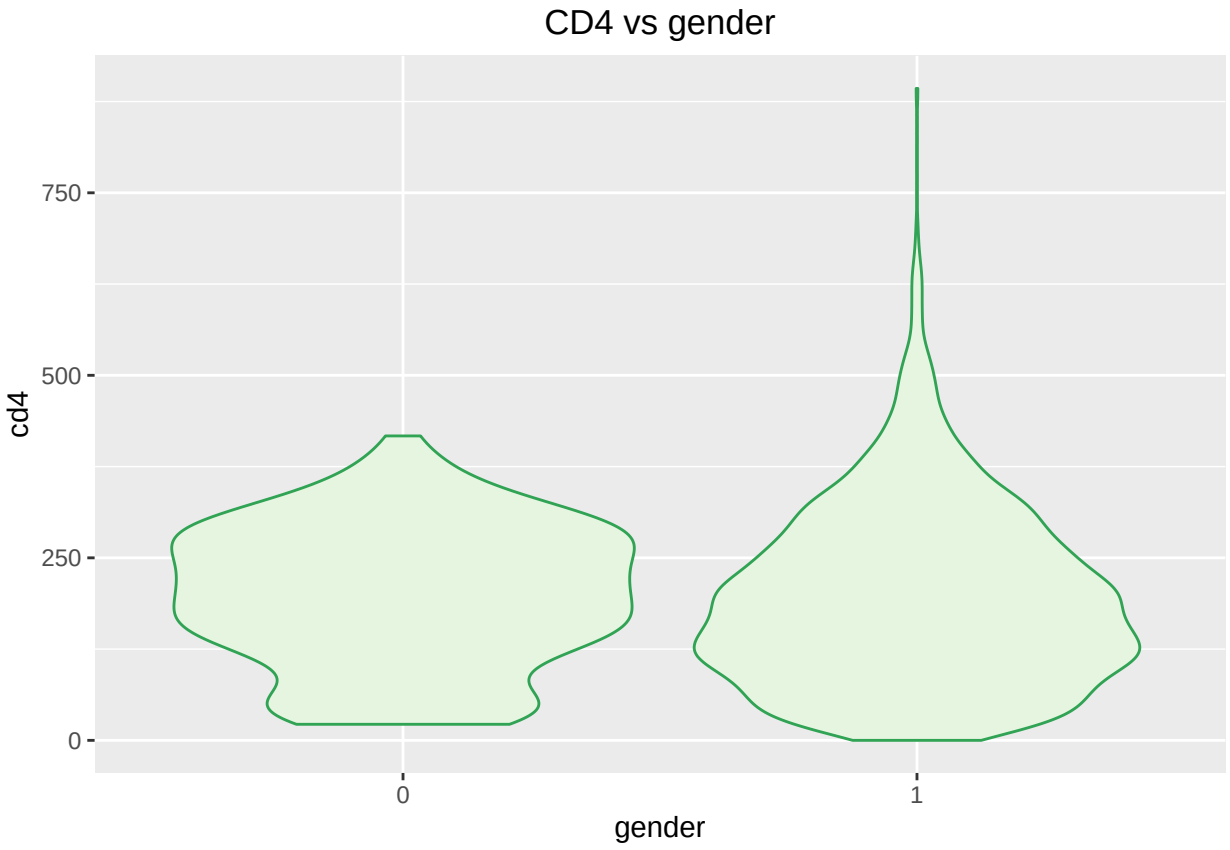

Figure 1: Violin plot on CD4 and gender

Figure 1 shows a violin plot of CD4 level in different gender. From the plot we can see that only men have higher CD4 while women don't. Recall the previous descriptive statistics table, this might because the data is biased: more men participated in the survey than women.

Figure 2 is a scatter plot of age and CD4, which shows that the youngest patient is 12 years old and the eldest patient is 63 years old. Most patients are between 20 to 50 years old. Also, we can see there are extremely high CD4 that were shown in age 38 and 48. Patient under 20 or elder than 50 tend to have lower level of CD4.

The box plot in Figure 3 shows that patients who receive placebo have a slightly higher CD4 in 1st quantile, mean and the 3rd quantile comparing to those who receive treatment. As for outliers, the two groups have almost the same maximum value, but except that, treatment group has a slightly smaller outliers than placebo group.

Figure 4 shows all patients CD4 level at each time point. We can conclude from the plot that there is a decline in CD4 level with time passed. But we need further exploration to figure out reasons for this decline.

Figure 5 is a shadow plot showing where the missings distributed in each variable. Obviously, only CD4 has missing values, and there are many of them. The distribution of missings seems random, no obvious regular pattern. Only cd4 has NAs, and there are many of them, but seems like randomly distributed.

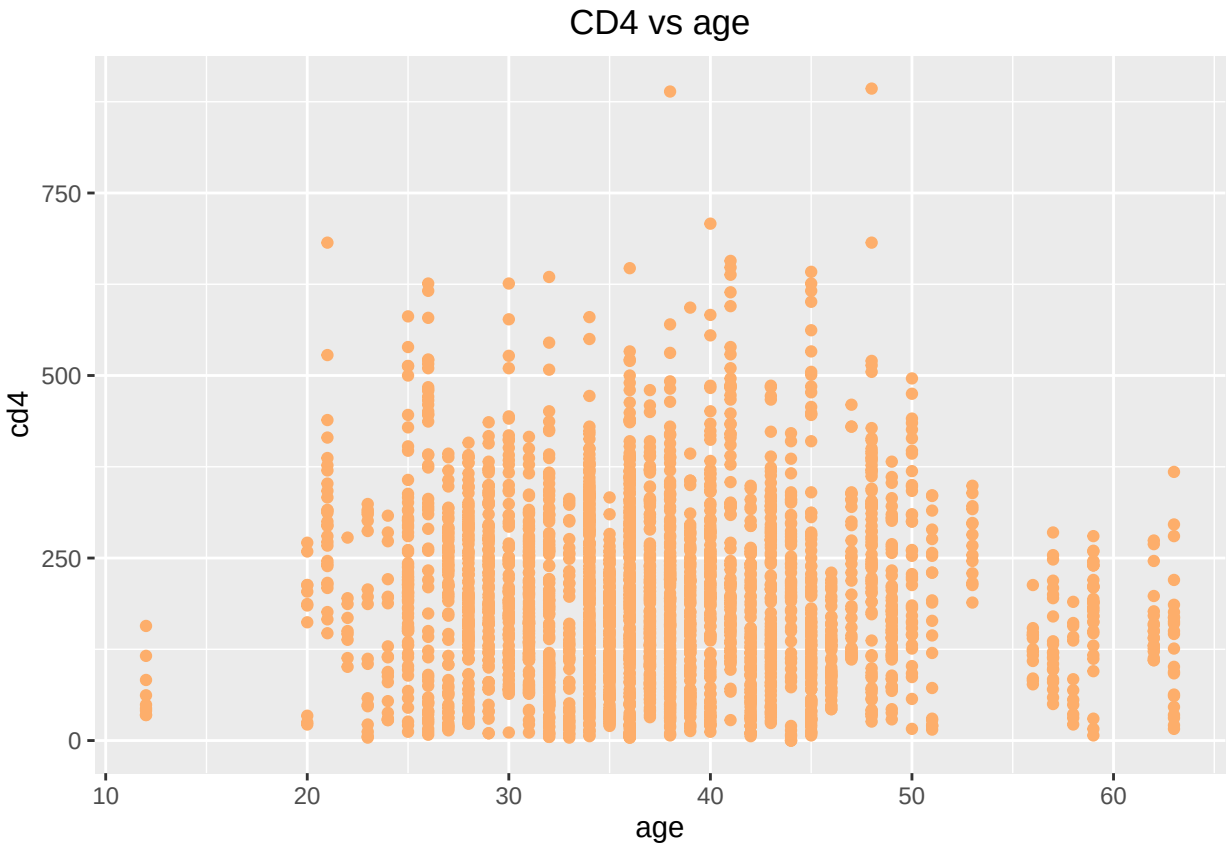

Figure 2: Scatter plot on CD4 and age

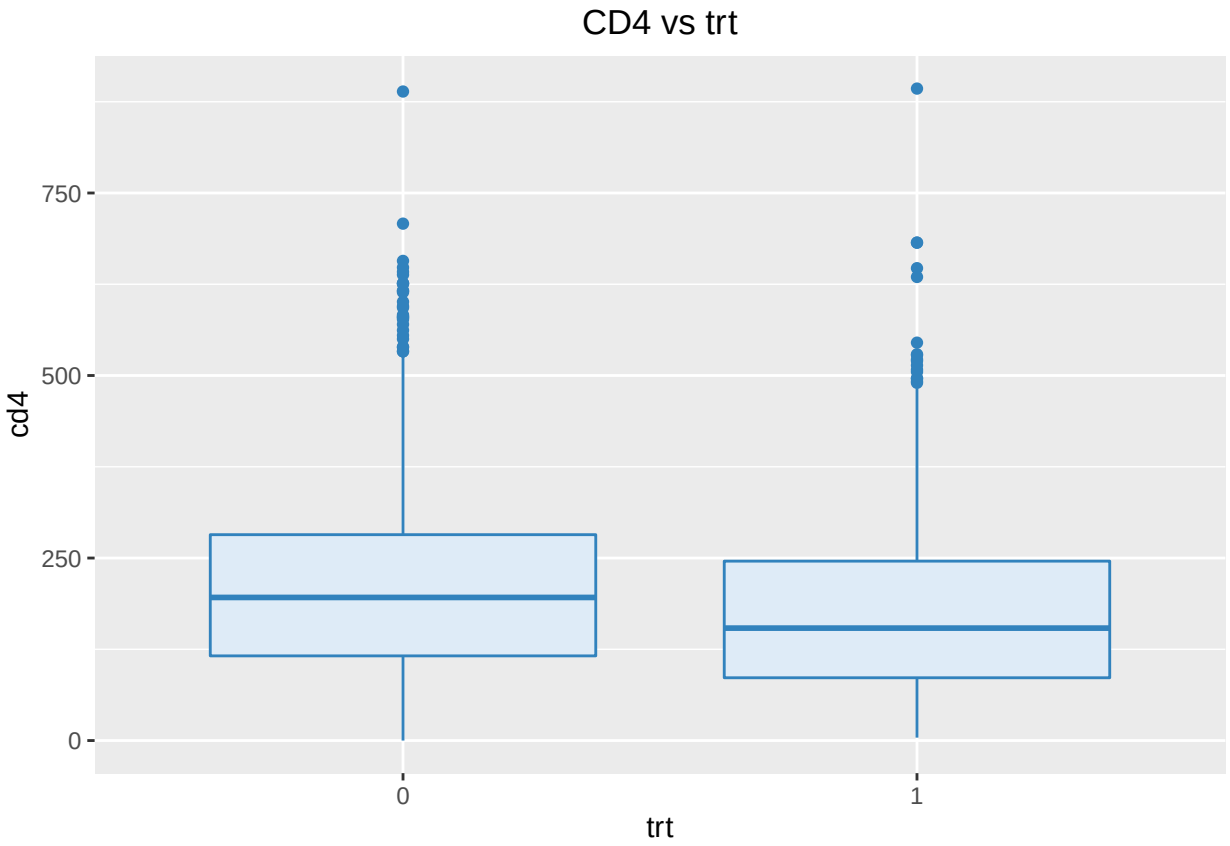

Figure 3: Box plot on CD4 and trt

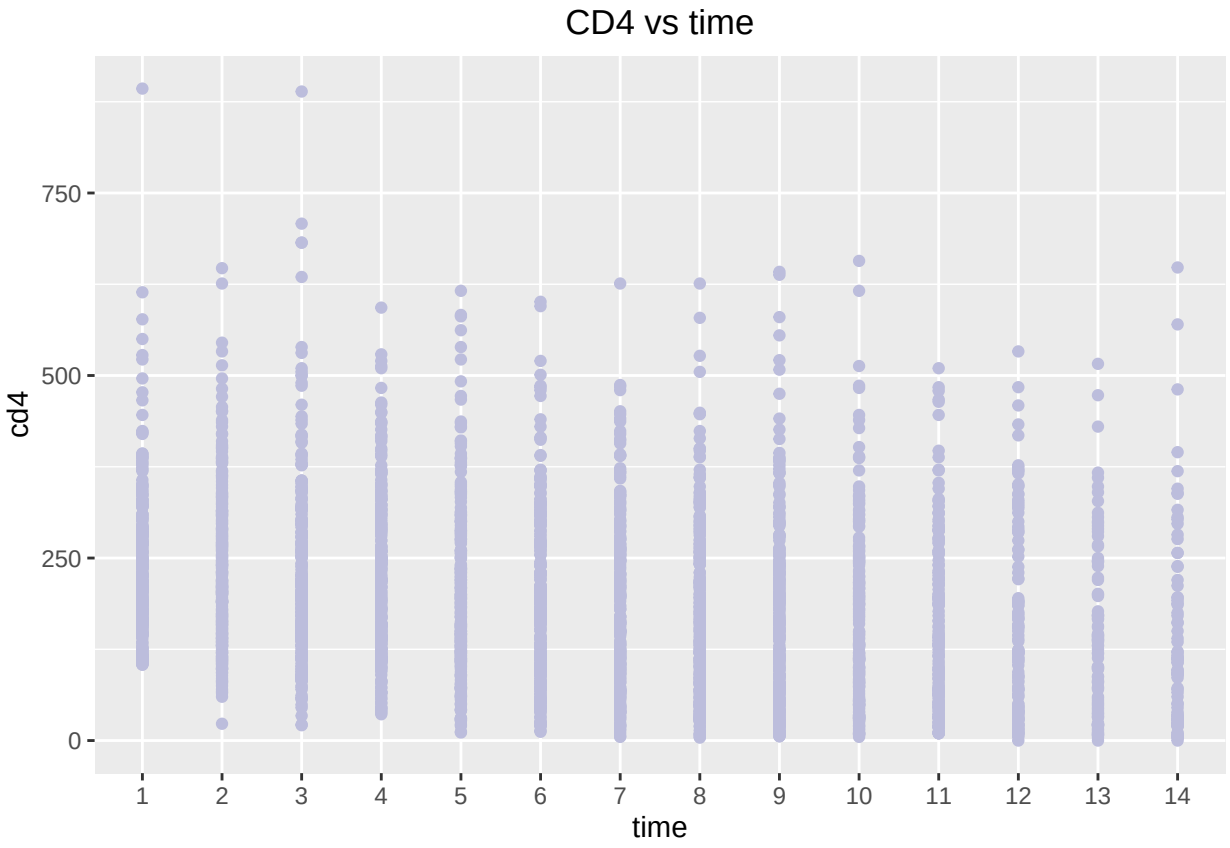

Figure 4: Scatter plot on CD4 and time

Figure 5: Shadow plot of NAs

Since having treatment or not and time has effect on CD4 level, and also missings are at a large amount, we then plot the relationship between these variables.

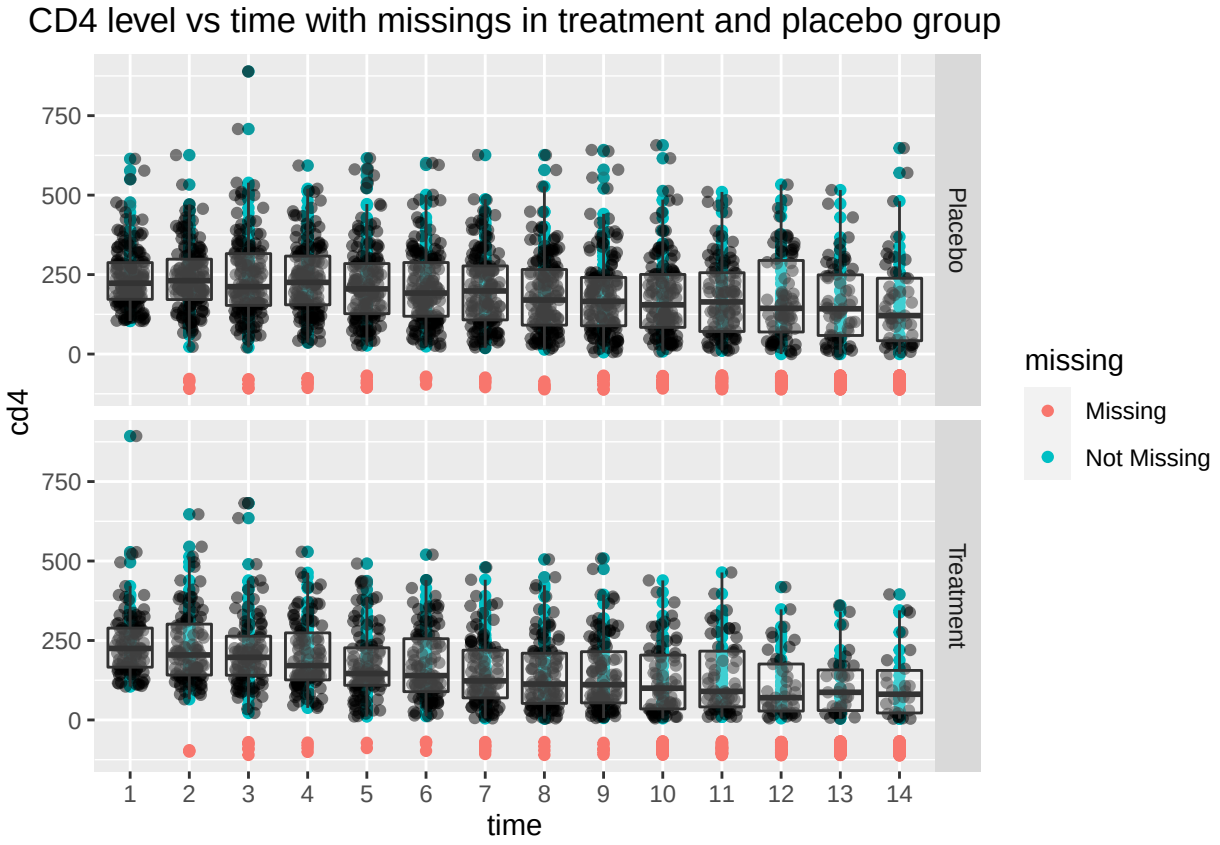

Figure 6: plot of cd4

Figure 6 shows changes in CD4 with time of placebo group and treatment group. Though there is a obvious decrease in CD4 level in treatment group compare with placebo group, however, there is also a significant decline in the number of non-missing CD4 value. It is hard to say the decrease in CD4 of treatment group is the effect of Zidovudine or, instead, is due to those missing values in treatment group.

5 Imputation of Missing Values

We have known from previous analysis that the amount of missing value is large, which can deviate our estimation on the effect of Zidovudine. That is the reason why we are trying to impute them. After visualizing data structure, we found that missing values concentrate in the later period of the study and are more in treatment group than in placebo group. To impute missing CD4, we here use the `impute_lm` in the package `simputation`, which is using linear regression to impute missings. As for predictor in the model, we just put time in it. Then, the idea of imputation is: one patient's CD4 level is only related to himself. So we build liner regression model for every patient and impute their CD4 level separately.

After imputed the missing value, we would like to know the distribution of those imputed values. It is shown in figure 7. We can see from the plot that the decrease in CD4 level of treatment group is significant comparing to the placebo group. None of the patient in treatment group has CD4 higher than 500.

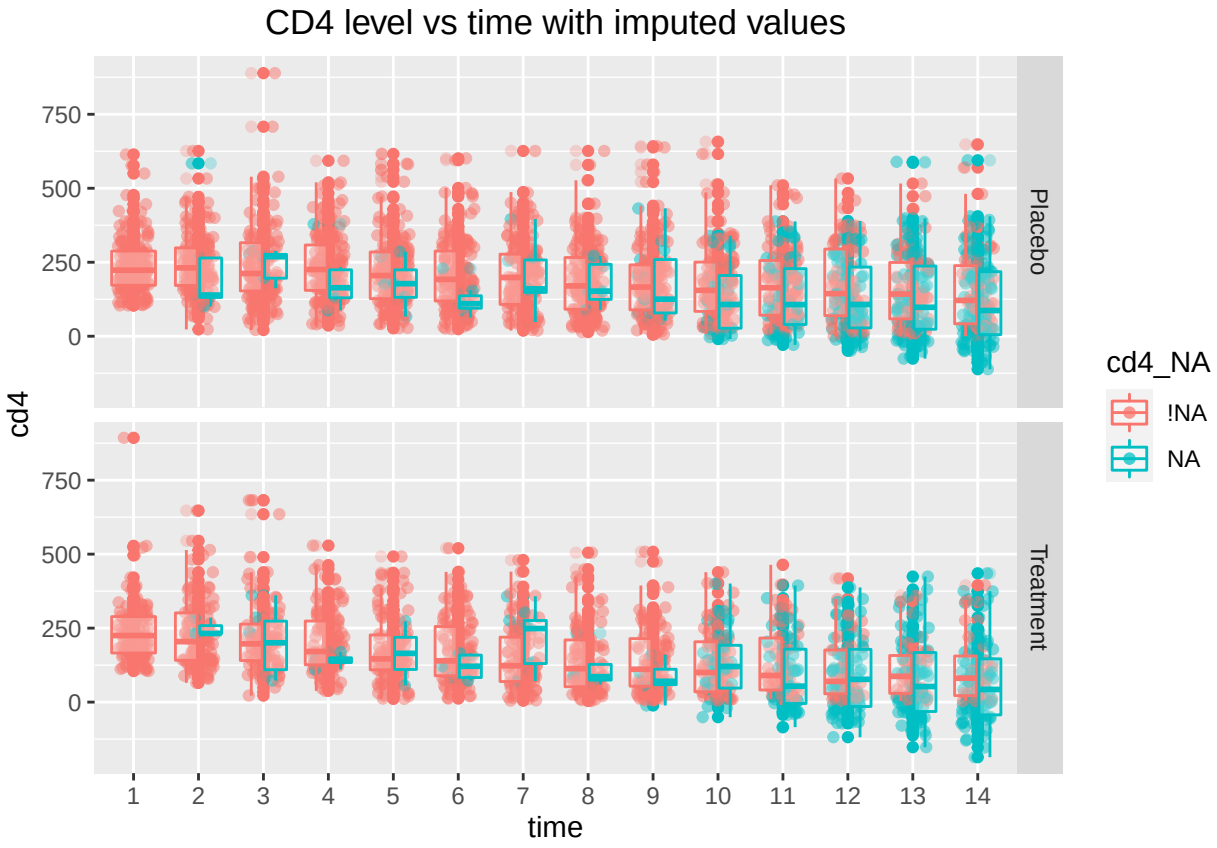

Figure 7: plot of cd4 after imputation

Then, since we are curious about how Zidovudine behave in different age group, we compute the average value of each age group's CD4 level in both placebo group and treatment group. As is shown in figure 8, comparing each age group's mean CD4 line, we can conclude that CD4 has significant effect on lowering CD4 level for patients age below 55. While for patients' age elder than 55, changing of mean CD4 is not obviously different between placebo group and treatment group.

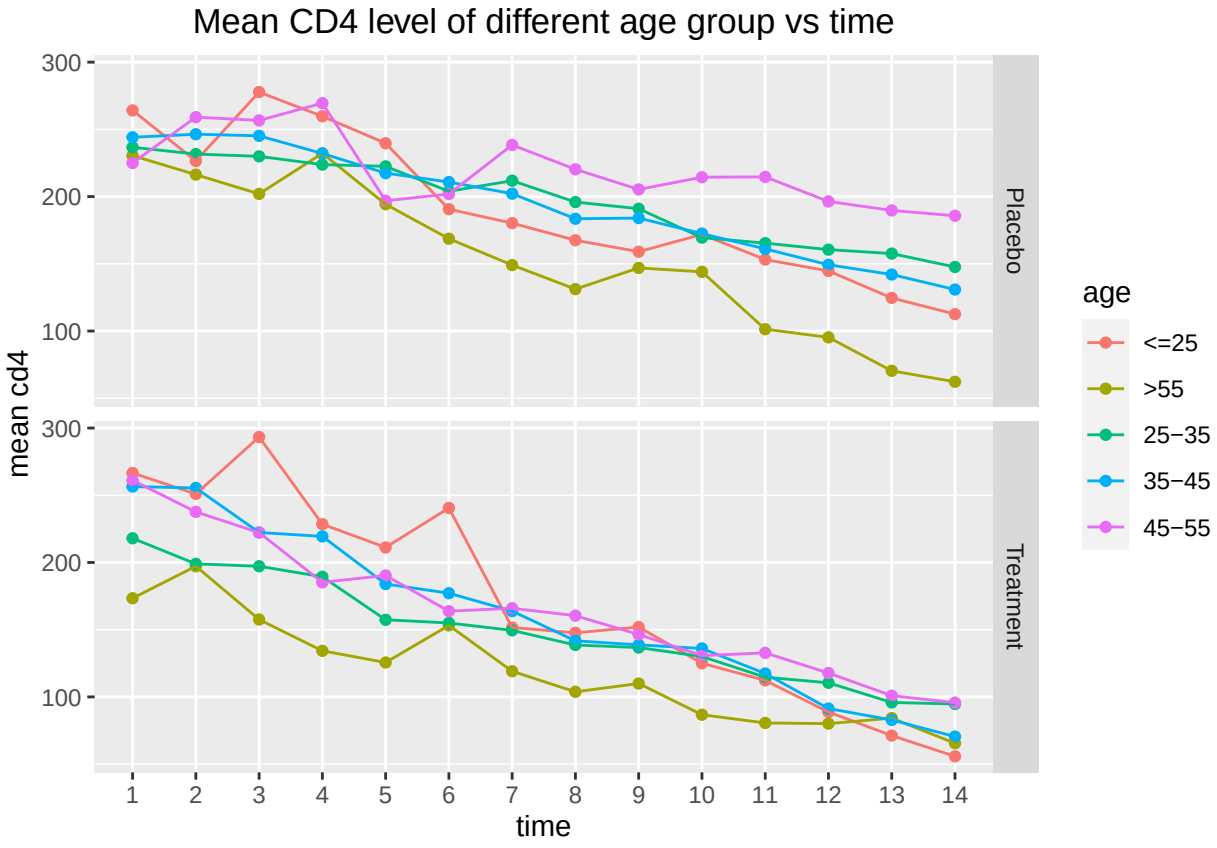

Figure 8: plot of cd4 for different age group

6 Estimatee the Effect of Zidovudine

6.1 Linear Regression Model

Now that we have imputed missings in x, we can begin building model to estimate the drug effect of Zidovudine. As is known from our previous analysis, CD4 level is relevent to treatment and time, so in the sense, by building a linear regression model with interaction term of treatment and time we can get the effect of Zidovudine.

```
##
## Call:
## lm(formula = cd4 ~ time * trt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -262.51  -89.66  -22.07   70.42  656.87
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 259.0624     5.3831  48.125  < 2e-16 ***
## time         -8.4980     0.6322 -13.442  < 2e-16 ***
## trt1        -10.6406     8.3935  -1.268 0.204979
## time:trt1    -3.7902     0.9858  -3.845 0.000123 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 119.1 on 3706 degrees of freedom
## Multiple R-squared:  0.1274, Adjusted R-squared:  0.1267
## F-statistic: 180.4 on 3 and 3706 DF,  p-value: < 2.2e-16
```

We can derive the treatment effect from the output, which is the coefficient of the interaction term -3.7902. The t-statistics of this term shows the treatment effect over time is significant. Based on this, we can conclude that Zidovudine has a significant effect on reducing patients' CD4 level.

6.2 Generalized Estimating Equation Model

The qualifying exam requires using GEE model to estimate the effect of Zidovudine, though I have not learnt this model yet, I searched relative materials and learnt to utilize this model, here is my attempt:

```
## (Intercept)        time        trt1   time:trt1
##  259.062367   -8.497953  -10.640607   -3.790236
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gaussian
##  Correlation Structure:     Exchangeable
##
## Call:
## gee(formula = cd4 ~ time * trt, id = id, data = data_impute,
##     corstr = "exchangeable")
##
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max
## -262.51489  -89.66364  -22.06646   70.42228  656.86643
##
##
## Coefficients:
##               Estimate Naive S.E.     Naive z Robust S.E.    Robust z
## (Intercept) 259.062367  8.5902983  30.1575519   8.1810863  31.6660108
## time         -8.497953  0.3325181 -25.5563610   0.7717447 -11.0113524
## trt1        -10.640607 13.3942356  -0.7944169  13.4090754  -0.7935377
## time:trt1    -3.790236  0.5184716  -7.3104019   1.3127977  -2.8871440
##
```

```
## Estimated Scale Parameter:  14185.28
## Number of Iterations:  1
##
## Working Correlation
##            [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
##  [1,] 1.0000000 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705
##  [2,] 0.7233705 1.0000000 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705
##  [3,] 0.7233705 0.7233705 1.0000000 0.7233705 0.7233705 0.7233705 0.7233705
##  [4,] 0.7233705 0.7233705 0.7233705 1.0000000 0.7233705 0.7233705 0.7233705
##  [5,] 0.7233705 0.7233705 0.7233705 0.7233705 1.0000000 0.7233705 0.7233705
##  [6,] 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 1.0000000 0.7233705
##  [7,] 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 1.0000000
##  [8,] 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705
##  [9,] 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705
## [10,] 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705
## [11,] 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705
## [12,] 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705
## [13,] 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705
## [14,] 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705
##            [,8]      [,9]     [,10]     [,11]     [,12]     [,13]     [,14]
##  [1,] 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705
##  [2,] 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705
##  [3,] 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705
##  [4,] 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705
##  [5,] 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705
##  [6,] 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705
##  [7,] 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705
##  [8,] 1.0000000 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705
##  [9,] 0.7233705 1.0000000 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705
## [10,] 0.7233705 0.7233705 1.0000000 0.7233705 0.7233705 0.7233705 0.7233705
## [11,] 0.7233705 0.7233705 0.7233705 1.0000000 0.7233705 0.7233705 0.7233705
## [12,] 0.7233705 0.7233705 0.7233705 0.7233705 1.0000000 0.7233705 0.7233705
## [13,] 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 1.0000000 0.7233705
## [14,] 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 0.7233705 1.0000000
```

The result is quite similar with that of linear regression, with the same coefficient of interaction term and the robust z is -2.8871400, smaller than the 0.05 threshold, indicating the effect of treatment over time is fairly significant.

7 Summary

The given data is consists of 4 factorial variables, one integer and one numeric variable. But the numeric variable, which is our main target of analyzing drug effect has a large majority of missings. After building linear regression model for each patient, we finally imputed those missing roughly. At last, by using linear regression model and generalized estimating equations, we came to the conclusion that drug effect of Zidovudine is very significant in reducing CD4 level.

8 Reference

Hong, Johnny, and Kellie Ottoboni. 2017. "Generalized Estimating Equations (GEE)." https://rlbarter.github.io/Practical-Statistics/2017/05/10/generalized-estimating-equations-gee/.

Tierney, Nicholas. 2021. "Getting Started with Naniar." https://cran.r-project.org/web/packages/naniar/vignettes/getting-started-w-naniar.html.