PHP2517: Applied Multilevel Data Analysis

Homework 1

Antonella Basso

April 20, 2022

Data

The "**cd4**" dataset includes CD4 percentages for a set of young children with HIV who were measured several times over a period of two years. The dataset also includes the ages of the children at each measurement. The "**cd4**" dataset has information on the following variables:

- **id**: Child ID
- **visit**: Number of visit
- **VDATE**: Date of visit
- **time**: Time (in days) after the first visit
- **CD4PCT**: CD4 percentage
- **visage**: Age (in years) at each visit
- **trt**: Treatment group
- **CD4CNT**: CD4 count
- **baseage**: Age at baseline (first visit)

Question 1:

a. Exploratory Data Analysis (EDA): Explore your data and provide appropriate descriptive statistics and plots for summarizing and presenting the information collected in this study. *Note: Some of the variables included have similar information (e.g., CD4 counts and percentages, visit date, time, and age). When this is the case, select just one of the relevant variables to include in the EDA.

b. Randomly select 10 children from the sample, and graph the outcome $Y =$ CD4 percentage on the square root scale, for each child as a function of time. What do you observe?

c. Each child's data has a time course that can be summarized by a linear fit. Estimate these lines and plot them for the same 10 children that you randomly selected from the sample.

d. Set up a model for the children's slopes (for time) and intercepts as a function of the treatment and age at baseline. Estimate this model using a two-step procedure:

   1. Estimate the intercept and slope separately for each child.
   2. Fit a model describing the between-child differences using the point estimates from the first step.

Solution

a. **Exploratory Data Analysis (EDA)**

   Overview of Data:

   - 1,055 total observations for 245 individuals
   - 1 missing observation for `CD4PCT` (and `CD4CNT`)
   - 95 missing observations for `CD4CNT`
   - 126 individuals on treatment 1

- 119 individuals on treatment 2
- individual ages range between 0.2-12.5 years
- observations are taken no less than 1 and no more than 709 days after the first visit
- each individual has between 1-7 observations

Descriptive Statistics:

Table 1: Descriptive Statistics of Primary Outcome by Treatment Group

Treatment Group	Individuals	Mean CD4 %	Variance CD4 %
1	126	22.714	190.640
2	119	25.192	164.846

Table 2: Descriptive Statistics of Primary Outcome by Observation Count

Observation Count	Individuals	Mean CD4 %	Variance CD4 %
1	26	24.327	197.139
2	36	17.793	174.297
3	27	30.837	162.078
4	34	20.703	141.377
5	39	21.755	180.959
6	37	24.918	161.376
7	46	25.284	186.386

Plots:

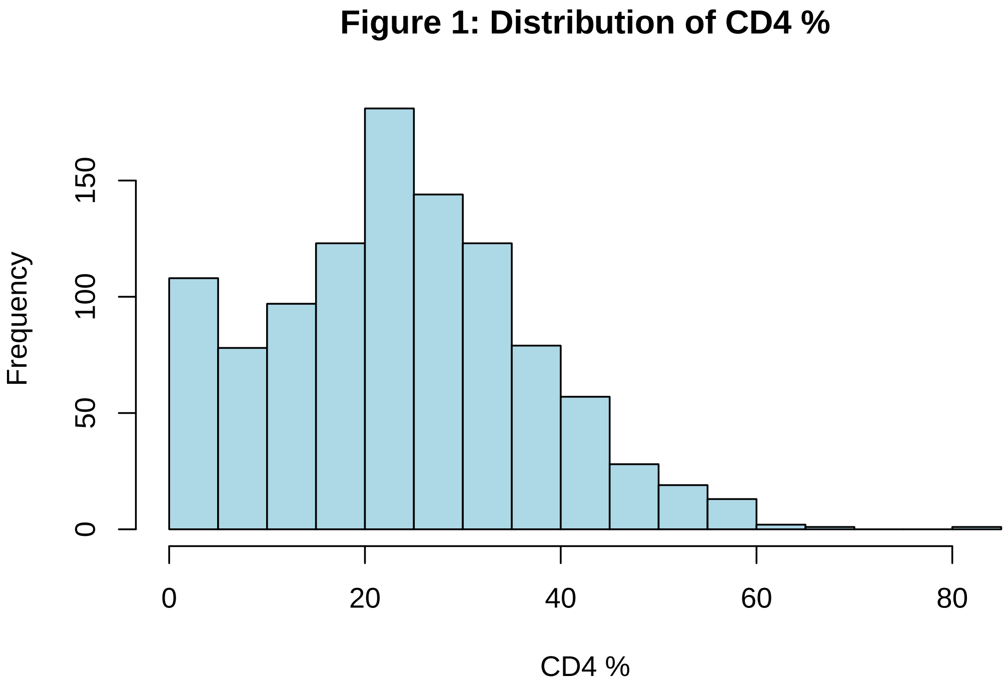


Figure 2: CD4 Percentages by Treatment Group

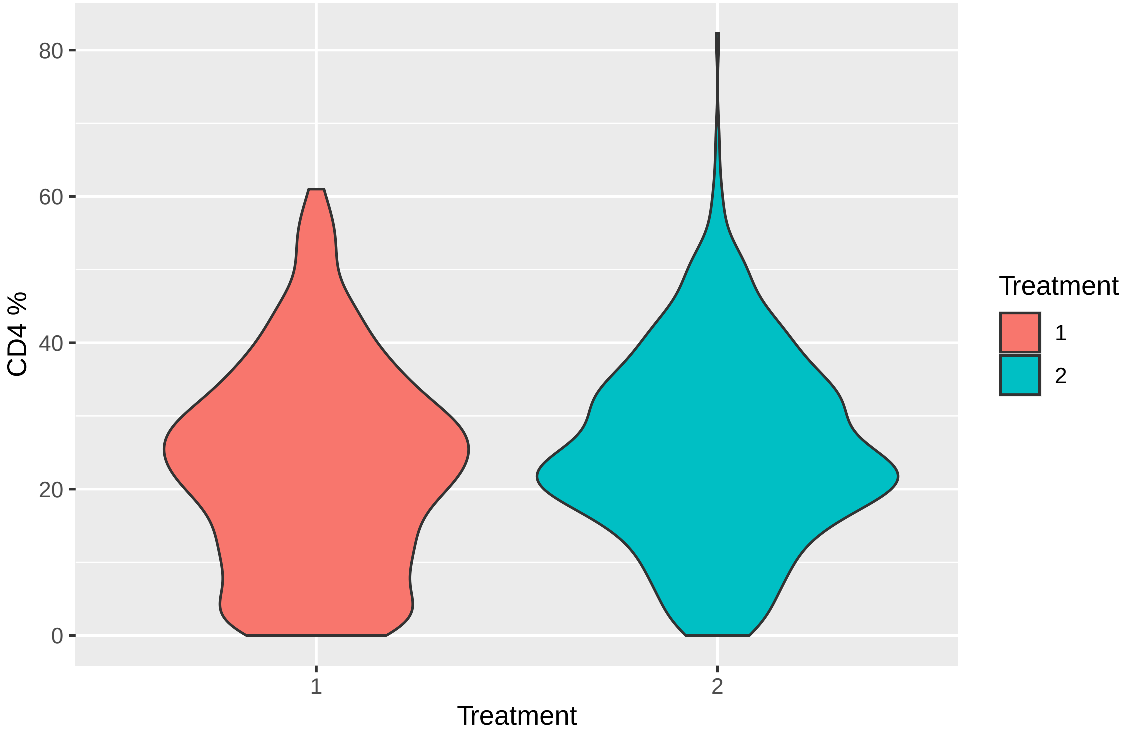


Figure 3: Mean CD4 Percentage by Age and Treatment Group

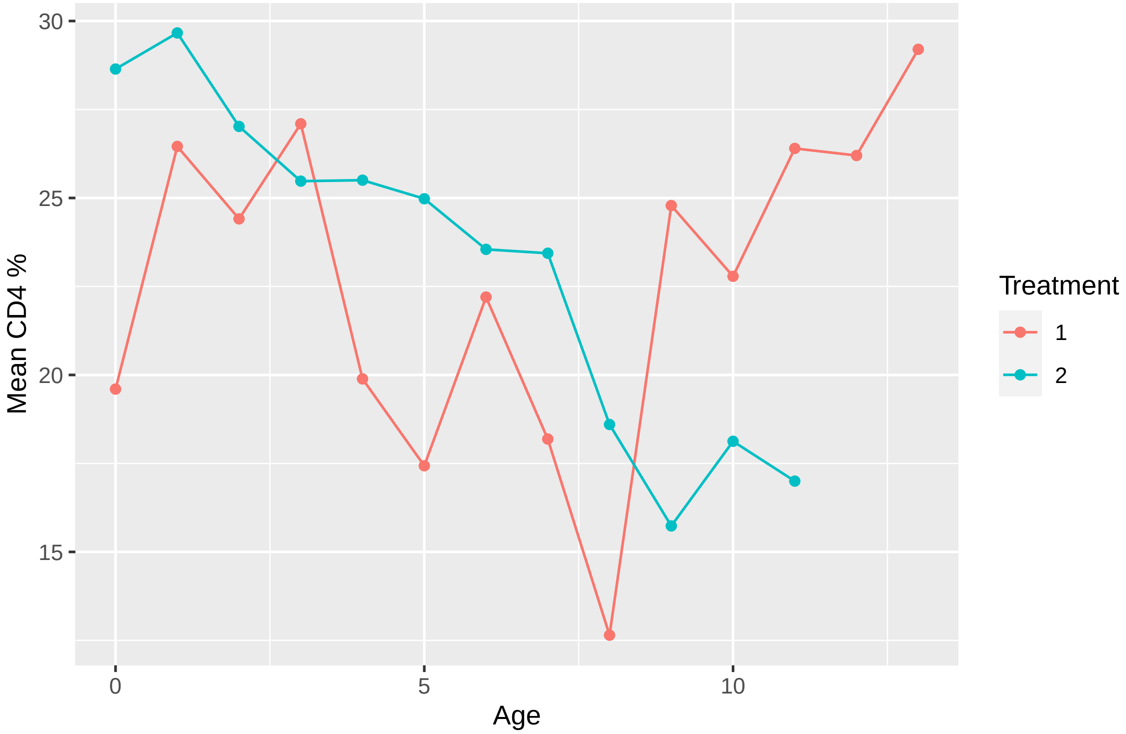


Figure 4: CD4 Percentages by Observation Count

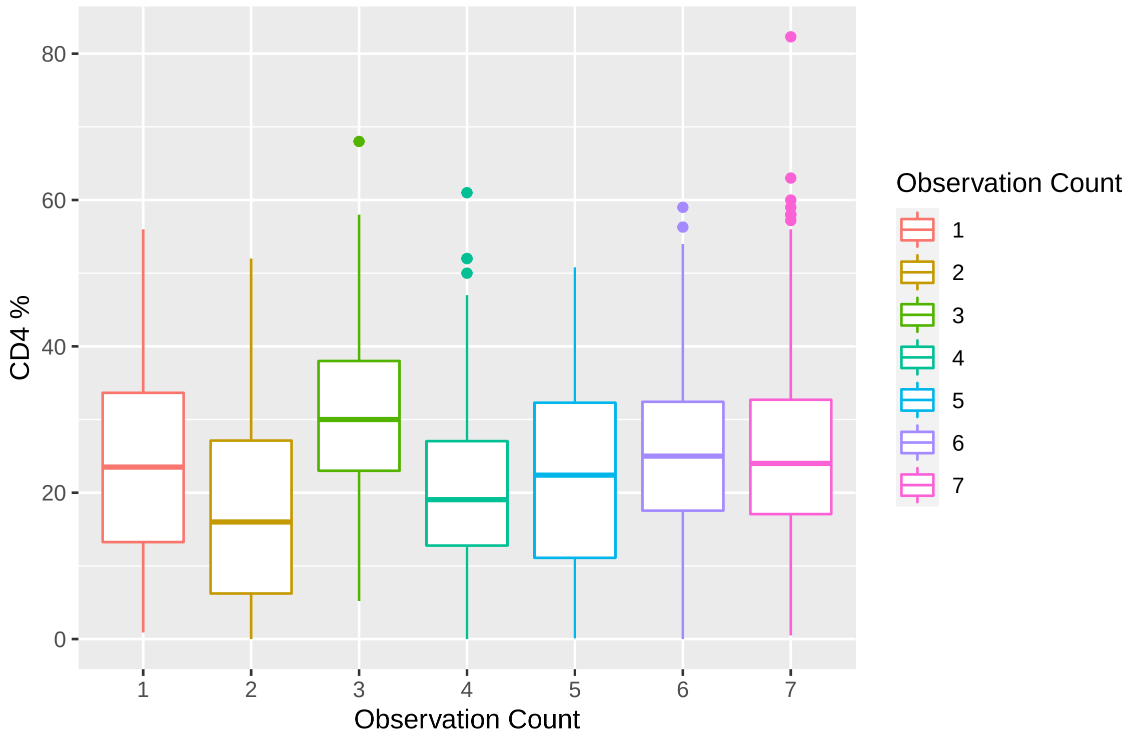
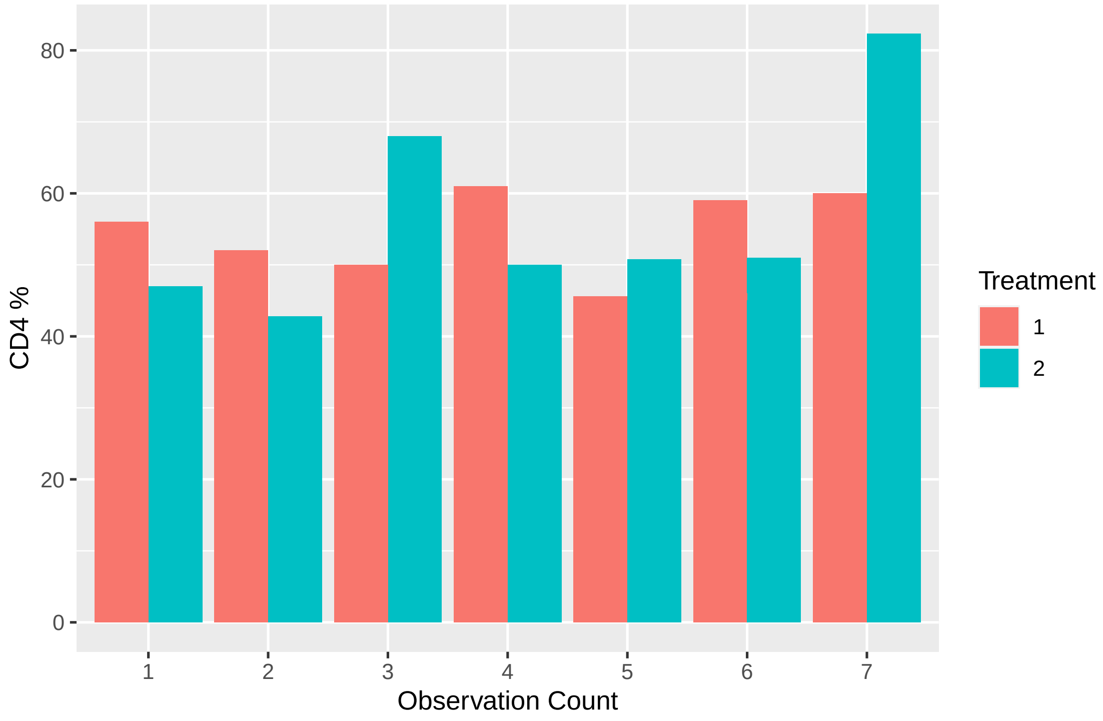
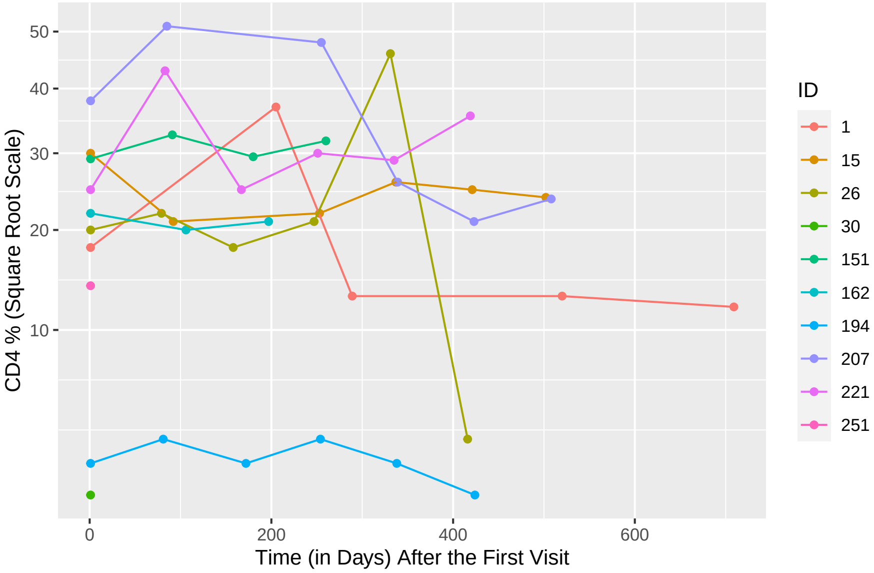


Figure 5: Mean CD4 Percentage by Observation Count and Treatment Grou



b. Individual Trajectories of CD4 % Over Time

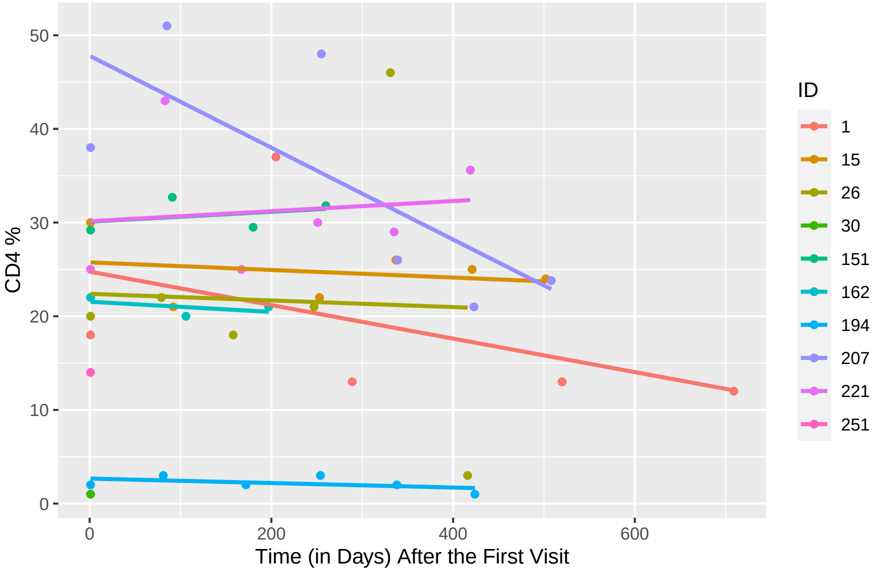
Figure 6: Individual CD4 Percentages (Square Root Scale) Over Time



The 10 sampled individual trajectories of CD4 percentages (on the square root scale), given by Figure 6 above, suggest that individuals have wide ranging responses with respect to time. While some individuals maintain a rather constant percentage of CD4, others (like those with ID's of 26, 207, 221, and 251) show radically fluctuating CD4 percentage. Particualarly, such individuals, for whom (generally) more observations were recorded and displayed somewhat higher or spiked CD4 percentages at earlier times (a few days after the first visit), showed relatively declining or more stabilizing trends in CD4 percentage later on.

c. Individual (Linear) Trajectories of CD4 % Over Time

Figure 7: Estimated Individual Linear Trajectories of CD4 Percentage



d. Describing between-child differences with respect to time (intercept and slope) as a function of treatment and baseline age.

   Model Estimation (two-step procedure):

   1. Estimating the intercept and slope separately for each child.

```
for (i in id){
  b0 <- coef(lm(CD4PCT ~ time, data=cd4_group[which(cd4_group$id==i), ]))[1]
  bt <- coef(lm(CD4PCT ~ time, data=cd4_group[which(cd4_group$id==i), ]))[2]
}
```

Table 3: Individual Time Intercept and Slope Estimates

id	intercept	slope_time
1	24.7644	-0.0179
2	1.0090	-0.0090
3	29.8664	0.0060
4	29.6999	-0.0032
5	16.0000	NA
6	24.5128	0.0064
7	27.1212	0.0060
10	29.4754	0.0014
11	9.2843	-0.0160
12	17.7471	-0.0032

Note:

Only shown for first 10 individuals.

   2. Fitting a model to describe between-child differences using point estimates from (1).

      Variation between individual time intercepts:

```
lm(b0 ~ trt + baseage, data=cd4_m1)
```

      Variation between individual time slopes:

```
lm(bt ~ trt + baseage, data=cd4_m1)
```

Table 4: Variation Between Individual Time (Intercept and Slope) Estimates

	Intercept	Slope
(Intercept)	28.02834	-0.00927
trt2	2.23579	-0.00012
baseage	-0.93499	0.00015

Estimated individual (linear) trajectories in response (CD4 %) over time (days since initial visit) vary in their slopes (CD4 % change over time) and intercepts (starting CD4 %). Here, we observe the extent to which these coefficients vary between individuals with respect to treatment and age at baseline. Looking at the estimates in Table 4 above, we see that the average intercept and slope across individuals, adjusting for other covariates in the model, are approximately 28.03 and -0.009, respectively. Additionally, we may notice that, on average, an individual's intercept will be 2.236 units greater for individuals on treatment 2 compared to those on treatment 1 (reference group). Similarly, those on treatment 2 have slopes that are generally

0.000119 lower than those for treatment 1. This suggests that although CD4 % decreases (by very little) over time for all individuals, it decreases slightly more rapidly for individuals on treatment 2 (adjusting for other covariates and not accounting for additional predictors). Moreover, we see that for each unit increase in baseline age, individual intercepts will decrease by a factor of 0.935 and slopes will increase by a factor of 0.000148. These findings validate the observations made from the plotted sample of individual trajectories from parts (b) and (c).

Question 2:

a. Write down a model using multilevel notation for predicting CD4 percentage as a function of time with varying intercepts across children. Fit the model (if working in R use the `lmer()` function) and interpret the coefficient for time.
b. Extend the model in (a) to include treatment and age at baseline as predictors. Write down the model using multilevel notation. Fit the model and interpret the coefficients on time, treatment, and age at baseline.
c. Investigate the change in partial pooling from (a) to (b) both graphically and numerically. Compare the results in (b) to those obtained in part (c).

Solution

a. Predicting CD4 percentage as a function of time with varying intercepts across children.

   Multilevel Model 1:

   ```
   lmer(CD4PCT ~ time + (1|id), data=cd4_group)
   ```

   $$Y_{ij} = \beta_{0j} + \beta_1 T_{ij} + \epsilon_{ij}$$
   $$Y_{ij} \sim \mathrm{N}(\beta_{0j} + \beta_1 T_{ij}, \hat{\sigma}^2)$$
   $$\beta_{0j} \sim \mathrm{N}(\hat{\mu}, \hat{\sigma}_\beta^2)$$

   Where Y_{ij} is the predicted outcome for the j^{th} individual on their i^{th} visit/observation; β_{0j} is the random (subject-specific) intercept (initial CD4 percentage); and β_1 is fixed slope for time (T_{ij}) accross individuals.

   $$\beta_1 = -0.00824$$
   $$\hat{\mu} = 24.98,\ \hat{\sigma}^2 = 131.37,\ \hat{\sigma}_\beta^2 = 53.95$$

   Interpretation: A coefficient of -0.00824 for the time predictor, indicates that CD4 percentage declines at an average rate of 0.00824 per unit increase in time across individuals. Thus, according to this model, we can expect individuals to display (on average) slowly decreasing linear trends in CD4 percentage over time (in days since their first visit).

b. Predicting CD4 percentage as a function of time, treatment, and baseline age, with varying intercepts across children.

Multilevel Model 2:

```
lmer(CD4PCT ~ time + trt + baseage + (1|id), data=cd4_group)
```

$$Y_{ij} = \beta_{0j} + \beta_1 T_{ij} + \beta_2 X_{ij}^{\text{trt=2}} + \beta_3 X_{ij}^{\text{b.age}}$$

$$Y_{ij} \sim \text{N}(\beta_{0j} + \beta_1 T_{ij} + \beta_2 X_{ij}^{\text{trt=2}} + \beta_3 X_{ij}^{\text{b.age}}, \hat{\sigma}^2)$$

$$\beta_{0j} \sim \text{N}(\hat{\mu}, \hat{\sigma}_\beta^2)$$

Where β_2 and β_3 are the fixed slope coefficients for treatment ($X_{ij}^{\text{trt=2}}$) and baseline age ($X_{ij}^{\text{b.age}}$), respectively, and all else is as previously stated. Note that the binary coefficient for treatment assumes treatment 1 is the reference group.

$$\beta_1 = -0.00814,\ \beta_2 = 1.4066,\ \beta_3 = -0.9488$$

$$\hat{\mu} = 27.54,\ \hat{\sigma}^2 = 127.23,\ \hat{\sigma}_\beta^2 = 53.97$$

Interpretation: Both β_2 and β_3, like β_1 for time, can be seen independently as the average effects of treatment and age at baseline on the primary outcome (CD4 percentage), adjusting for other model covariates. That is, according to this model, the average effect of treatment on CD4 percentage is an additional 1.4066 units higher for individuals on treatment 2 compared to those on treatment 1, adjusting for time and age at baseline. Similarly, the average effect of baseline age on the response, adjusting for time and treatment, is an approximate decrease of 0.9488 units for each unit increase in baseline age.

c. Comparing changes in partial pooling from multilevel model 1 (in part (a)) to 2 (in part (b)).

Numerically:

Table 5: Individual Random Intercepts, Residuals, and Standard Errors (SE)

	Multilevel Model 1			Multilevel Model 2		
ID	Intercept	Residual	SE	Intercept	Residual	SE
1	21.7087	-3.2722	3.1576	25.3043	-2.2314	3.1544
2	5.0679	-19.9129	4.7306	7.2533	-20.2823	4.7185
3	33.7530	8.7722	2.6981	39.3322	11.7965	2.6963
4	30.4321	5.4512	2.9009	32.6090	5.0733	2.8985
5	18.6202	-6.3607	6.1841	19.8775	-7.6582	6.1560
6	28.0680	3.0872	2.9009	29.4784	1.9428	2.8985
7	30.5691	5.5883	2.6981	35.1170	7.5813	2.6963
8	31.4998	6.5190	2.9009	39.2148	11.6791	2.8985
9	8.2800	-16.7008	2.6981	9.8420	-17.6936	2.6963
10	19.3087	-5.6722	2.9009	23.2388	-4.2968	2.8985

Table 6: Fixed Effects

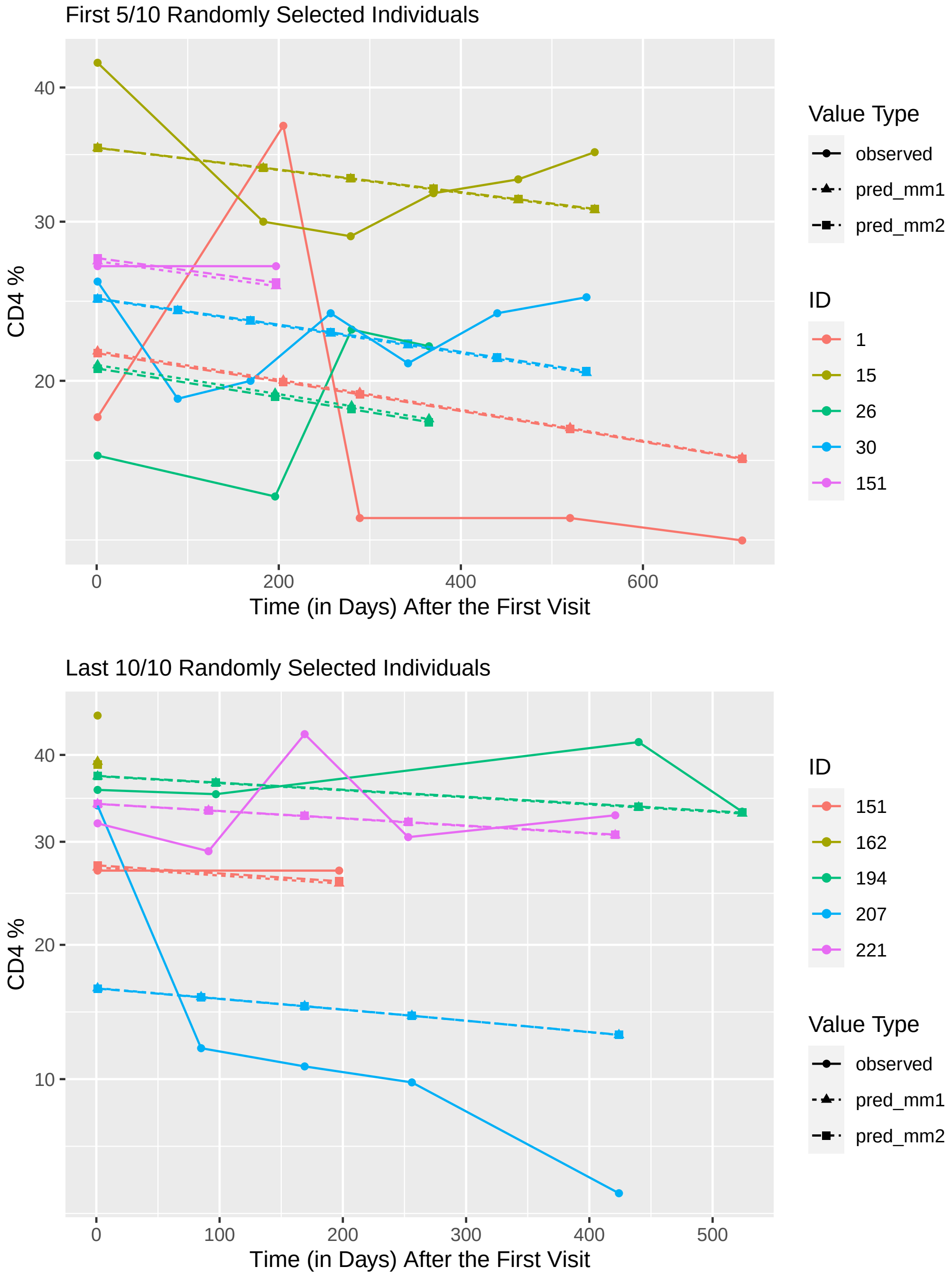
Multilevel Model	Intercept	Intercept SE	Time Slope	Time Slope SE
1	24.9808	0.8191	-0.0082	0.0014
2	27.5356	1.5890	-0.0081	0.0014

Table 7: Group-Level and Individual-Level Variabilities

Multilevel Model	Group-Level Variability	Individual-Level Variability
1	131.37	53.95
2	127.23	53.97

Graphically:

Figure 8: Observed vs. Predicted Trajectories of CD4 %


First 5/10 Randomly Selected Individuals

Last 10/10 Randomly Selected Individuals

Evidently, there is very little change in partial pooling from multilevel model 1 to multilevel model 2. However, given the decreased group-level variance in the second model, it is safe to assume that age and treatment explain at least some of the variation between individuals in the data, and hence, this model provides a slightly better fit.

Question 3:

a. Use the model fit from Question 2(b) to simulate predicted CD4 percentages for each child in the dataset at a hypothetical next time point.

b. Use the same model fit to simulate predicted CD4 percentages at each of the time periods for a new child who was 4 years old at baseline.

Solution

a. Given the variability in time between individuals, we appeal to the following metric to generate appropriate “next” time points.

   Metric:

   1. For individuals with $n = 1, 2, ..., 7$ different time points (observations), calculate the mean time (in days since the first visit) for the first, second, third, etc. observations.

      ```
      n=1: 1
      n=2: 1 139
      n=3: 1 132 287
      n=4: 1 130 261 394
      n=5: 1 119 218 330 461
      n=6: 1 104 198 287 379 486
      n=7: 1  87 174 262 351 439 525
      ```

   2. Aside from the first case ($n = 1$), find the average distance accross time points.

      ```
      n=2: 138                          -> mean=138
      n=3: 131 155                      -> mean=143
      n=4: 129 131 133                  -> mean=138
      n=5: 118  99 112 131              -> mean=138
      n=6: 103  94  89  92 107          -> mean=138
      n=7:  86  87  88  89  88  86      -> mean=138
      ```

   3. While individuals with a single observation will have a “next” time point equal to the sum of their original time and 140 (to give 141), the remaining individuals’ “next” time value will be the sum of their last (and maximum) time and their group’s (in terms of number of observations) corresponding mean time point difference.

   Having implemented this metric to obtain a “next” time point for each individual, we compile the data frame given by Table 8 (only showing the first 10 individuals/rows), from which we apply the fit from multilevel model 2 to get the predicted and simulated CD4 percentages shown in the last two columns, as follows.

   ```
   fixed effects:
      time: -0.008135
      trt2: 1.406590
      baseage: -0.948838
   ```

```
        group-level variability:    individual-level variability:
sigma2:  127.23                      53.97
sigma:    11.28                       7.347
```

Model Predictions:

```
for (i in 1:length(id)){
  pred[i] <- rand_int[i] - 0.008135*new_time[i] + 1.406590*trt2[i] - 0.948838*baseage[i]
}
```

Simulated Model Predictions:

```
for (i in 1:length(id)){
  sim_pred[i] <- mean(rnorm(1000,
                            rand_int[i] -
                              0.008135*new_time[i] +
                              1.406590*trt2[i] -
                              0.948838*baseage[i],
                            sigma.hat(mm2)$sigma$data))
 }
```

Table 8: Individual New Time Points, Covariate Values, and Predicted/Simulated Responses

id	rand_int	new_time	trt2	baseage	pred	sim_pred
1	25.3043	824	0	3.9100	14.8911	15.0962
2	7.2533	217	1	3.5650	3.5120	3.0055
3	39.3322	718	0	6.1242	27.6804	27.3153
4	32.6090	532	0	2.3025	26.0965	25.9364
5	19.8775	141	0	0.6542	18.1097	17.9709
6	29.4784	658	1	2.9183	22.7632	22.8452
7	35.1170	620	1	6.4425	25.3670	25.2635
8	39.2148	658	0	8.5583	25.7415	25.4635
9	9.8420	606	1	3.0583	3.4169	3.7943
10	23.2388	574	1	5.7383	14.5312	14.3891

Figure 9.1: Individual Trajectories with Original Time Points

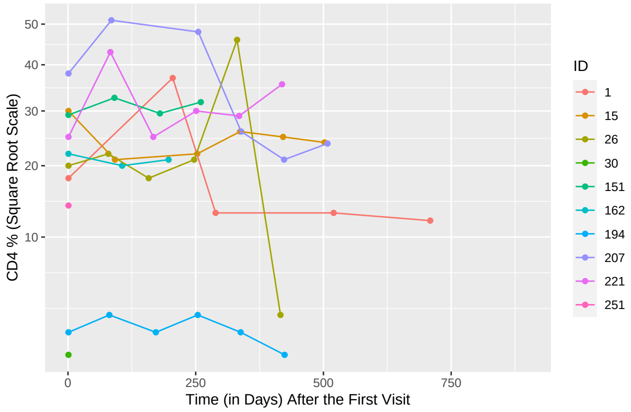


Figure 9.2: Individual Trajectories with New Time Points

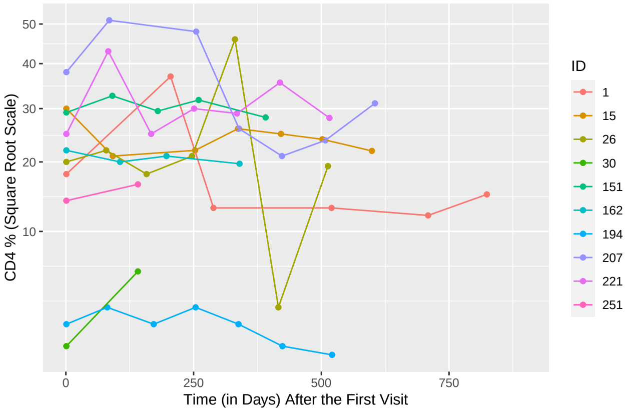
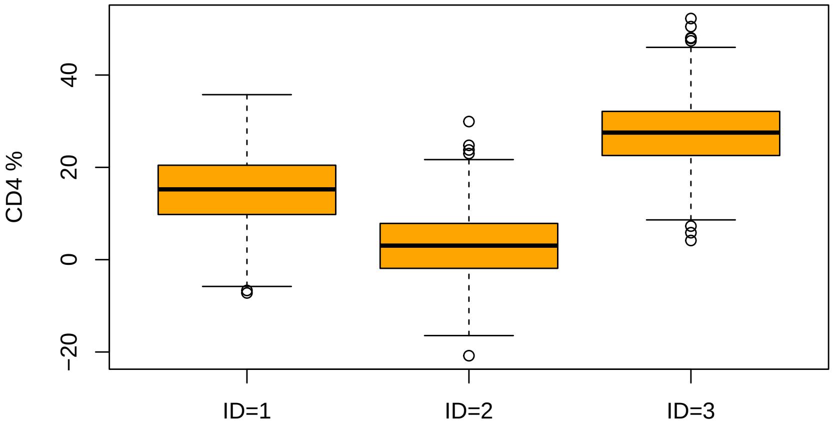


Figure 10: Prediction for First 3 Individuals with New Time Point



b. Under the same model (multilevel model 2), we simulate predicted CD4 percentages at 7 different time periods for a new child whose baseline age was set to 4 years. Using the mean times for those with 7 observations, we obtain the following:

```
time points: 1 87 174 262 351 439 525
```

```
## Simulating Predictions for New Individual with 7 Time Points

beta0 <- fixef(mm2)["(Intercept)"] #int_coef <- 27.53565
beta1 <- fixef(mm2)["time"] # time_coef <- -0.008135
beta2 <- fixef(mm2)["trt2"] # trt2_coef <- 1.406590
beta3 <- fixef(mm2)["baseage"] # baseage_coef <- -0.948838

sigma_beta0j <- sigma.hat(mm2)$sigma$id # group-level standard deviation
sigma_y_hat <- sigma.hat(mm2)$sigma$data # individual-level standard deviation

# random intercepts for new individual
beta0j <- rnorm(1000, beta0, sigma_beta0j)

# matrix with random intercepts & fixed slopes
rand_coefs_mat <- as.matrix(cbind(beta0+beta0j, beta1, beta2, beta3))

y_trt1_list <- c() # treatment 1
y_trt2_list <- c() # treatment 2
for (i in id9$time){
  x_trt1 <- c(1, i, 0, 4)
  x_trt2 <- c(1, i, 1, 4)
  y_trt1 <- rnorm(1000, rand_coefs_mat %*% x_trt1, sigma_y_hat)
  y_trt2 <- rnorm(1000, rand_coefs_mat %*% x_trt2, sigma_y_hat)
  y_trt1_list <- c(y_trt1_list, y_trt1)
  y_trt2_list <- c(y_trt2_list, y_trt2)
}
```

Figure 11.1: Predictions for New Individual with 7 Time Points

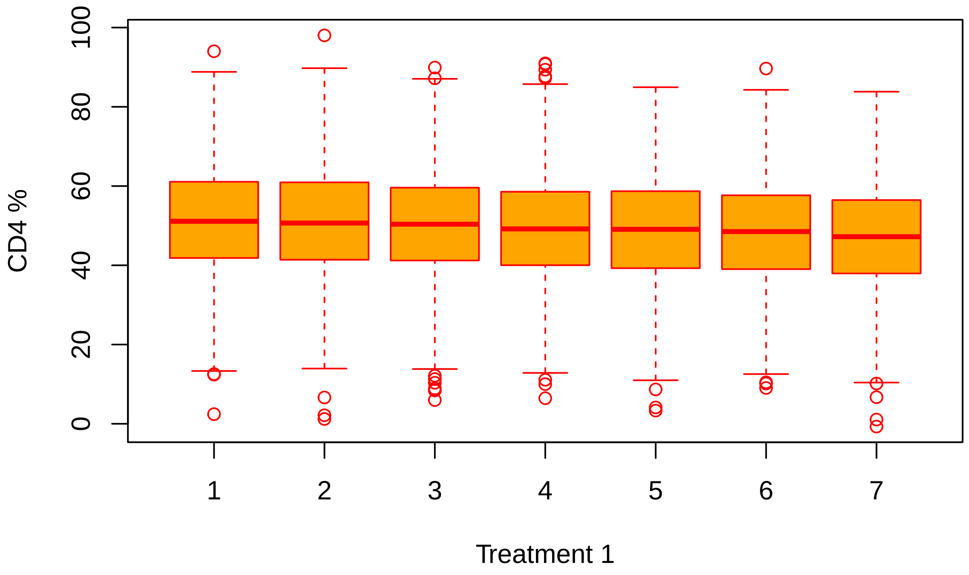
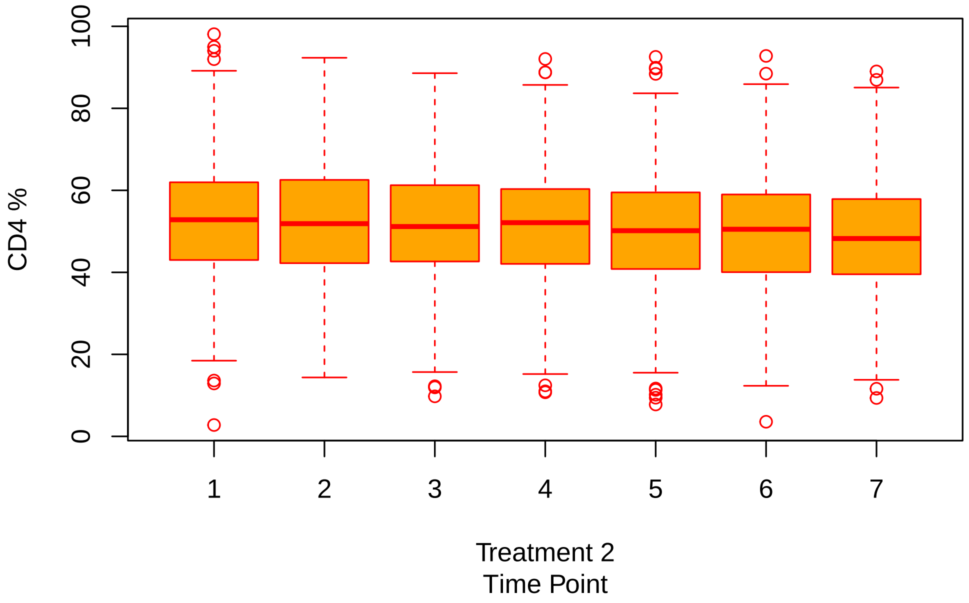


Figure 11.2: Predictions for New Individual with 7 Time Points



Code

```
# Importing "cd4" Data
cd4 <- read.csv("/Users/antonellabasso/Desktop/PHP2517/DATA/cd4.csv")
```

```
## Descriptive Statistics

nrow(cd4) # 1055 observations
apply(cd4, 2, function(x) length(unique(x))) # 245 individuals

# missing data
cd4_na <- apply(cd4, 2, function(x) sum(is.na(x))) # for all columns
cd4_na[cd4_na != 0] # 1 missing CD4 %, 95 missing CD$ counts
cd4[which(is.na(cd4$CD4PCT)), ] # 1 individual missing outcome of interest (id=83)
cd4[which(cd4$id==83), ]

# removing observation with missing outcome of interest
cd4 <- cd4[-348, ]

# id numbers not in range 1-254
not_id <- c()
for(i in 1:254){
  if (i %in% unique(cd4$id)==FALSE){
    not_id <- c(not_id, i)
  }
}

# number of observations for each individual
obs_id <- as.data.frame(table(cd4$id)) %>% rename("id"=Var1, "observations"=Freq)
unique(obs_id$count) # each had 1, 2, 3, 4, 5, 6, or 7 visits/observations

# factorizing categorical variables
cd4$id <- as.factor(cd4$id)
cd4$trt <- as.factor(cd4$trt)

# grouping by observation count
cd4_group <- full_join(cd4, obs_id)

# descriptive statistics wrt CD4 % by visit count (group)
obs_ds <- cd4_group %>%
  group_by(observations) %>%
  summarise(count=n(),
            mean=mean(CD4PCT),
            var=var(CD4PCT)) %>%
  mutate(num_individuals=count/observations)

# descriptive statistics wrt CD4 % by treatment group
trt_ind <- cd4_group %>% group_by(id, trt) %>% summarise(n()) %>%
  group_by(trt) %>% summarise(count=n()) #trt1=126, trt2=119
trt_ds <- cd4_group %>% group_by(trt) %>% summarise(count=n(),
                                                    mean=mean(CD4PCT),
                                                    var=var(CD4PCT))
trt_ds$num_individuals <- trt_ind$count
```

```r
## EDA Tables

# descriptive statistics of primary outcome by treatment group
dstats1 <- data.frame(trt=trt_ds$trt, num_individuals=trt_ds$num_individuals,
                      mean=trt_ds$mean, var=trt_ds$var) %>%
  rename("Treatment Group"=trt, "Individuals"=num_individuals,
         "Mean CD4 %"=mean, "Variance CD4 %"=var)
# descriptive statistics of primary outcome by observation count
dstats2 <- data.frame(observations=obs_ds$observations, num_individuals=obs_ds$num_individuals,
                      mean=obs_ds$mean, var=obs_ds$var) %>%
  rename("Observation Count"=observations, "Individuals"=num_individuals,
         "Mean CD4 %"=mean, "Variance CD4 %"=var)
```

```r
## EDA Plots

# distribution of CD4 %
hist(cd4$CD4PCT,
     main="Figure 1: Distribution of CD4 %",
     xlab="CD4 %",
     col="lightblue")

# individual averages
mean_ind <- cd4 %>% group_by(id) %>% summarise(mean=mean(CD4PCT), .groups="keep")
scatterplot <- ggplot(mean_ind, aes(x=as.numeric(id), y=mean)) +
  geom_point(color="purple") +
  labs(title="Mean Individual CD4 Percentages",
       x="ID",
       y="Mean CD4 %")

# CD4 % by treatment group
violinplot <- ggplot(cd4, aes(x=trt, y=CD4PCT, fill=trt)) +
  geom_violin() +
  labs(title="Figure 2: CD4 Percentages by Treatment Group",
       x="Treatment",
       y="CD4 %",
       fill="Treatment")

# CD4 % by age (rounded)
mean_age <- cd4 %>% group_by(age=round(visage), trt) %>% summarise(mean=mean(CD4PCT), .groups="keep")
lineplot <- ggplot(mean_age, aes(x=age, y=mean, color=trt)) +
  geom_point() +
  geom_line() +
  labs(title="Figure 3: Mean CD4 Percentage by Age and Treatment Group",
       x="Age",
       y="Mean CD4 %",
       color="Treatment")

# CD4 % by number of observations
boxplot <- ggplot(cd4_group, aes(x=as.factor(observations), y=CD4PCT, color=as.factor(observations))) +
  geom_boxplot() +
  labs(title="Figure 4: CD4 Percentages by Observation Count",
       x="Observation Count",
       y="CD4 %",
       color="Observation Count")
```

```
# mean CD4 % by observation count and treatment
mean_obs_trt <- cd4_group %>% group_by(observations, trt) %>% summarise(mean=mean(CD4PCT), .groups="kee
barplot <- ggplot(cd4_group, aes(x=as.factor(observations), y=CD4PCT, fill=trt)) +
  geom_bar(position="dodge", stat="identity") +
  labs(title="Figure 5: Mean CD4 Percentage by Observation Count and Treatment Group",
       x="Observation Count",
       y="CD4 %",
       fill="Treatment")
```

```
## Individual Trajectories of CD4 % Over Time

# 10 randomly selected individuals
set.seed(47)
randsample_10 <- sample(unique(cd4_group$id), 10, replace=FALSE)
randsample_10_df <- cd4_group[which(cd4_group$id %in% randsample_10), ]

# individual trajectories of CD4 % (on the square root scale)
# set y=sqrt(CD4PCT) to change values to sqrt()
randsample_10_trajectories <- ggplot(randsample_10_df,
                                     aes(x=time, y=CD4PCT, color=id)) +
  geom_point() +
  geom_line() +
  scale_y_sqrt() +
  labs(title="Figure 6: Individual CD4 Percentages (Square Root Scale) Over Time",
       x="Time (in Days) After the First Visit",
       y="CD4 % (Square Root Scale)",
       color="ID")

# individual trajectories of CD4 % (NOT sqrt scale)
ggplot(randsample_10_df, aes(x=time, y=CD4PCT, color=id)) +
  geom_point() +
  geom_line() +
  labs(title="Individual CD4 Percentages Over Time",
       x="Time (in Days) After the First Visit",
       y="CD4 %",
       color="ID")
```

```
## Individual Linear Trajectories of CD4 % Over Time

# individual linear trajectories of CD4 % (lm)
randsample_10_lms <- ggplot(randsample_10_df, aes(x=time, y=CD4PCT, color=id)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE)+
  labs(title="Figure 7: Estimated Individual Linear Trajectories of CD4 Percentage",
       x="Time (in Days) After the First Visit",
       y="CD4 %",
       color="ID")
```

```
## Linear Model:

# Step 1: estimating intercepts and time slopes separately for each individual
id <- unique(cd4_group$id) # unique ids
individuals_lm <- as.data.frame(id) # data frame for id & corresponding coefs
intercept <- c()
```

```
slope_time <- c()
for (i in id){
  b0 <- coef(lm(CD4PCT ~ time, data=cd4_group[which(cd4_group$id==i), ]))[1] # intercept estimates
  bt <- coef(lm(CD4PCT ~ time, data=cd4_group[which(cd4_group$id==i), ]))[2] # time slope estimates
  intercept <- c(intercept, b0)
  slope_time <- c(slope_time, bt)
}
individuals_lm$intercept <- as.vector(intercept)
individuals_lm$slope_time <- as.vector(slope_time)
individuals_lm %>% rename("ID"=id, "Intercept"=intercept, "Slope"=slope_time)

# Step 2: estimating variation between individual time trends given treatment and baseline age
cd4_m1 <- full_join(cd4_group, individuals_lm)
lm(intercept ~ trt + baseage, data=cd4_m1) # variation between individual time intercepts
lm(slope_time ~ trt + baseage, data=cd4_m1) # variation between individual time slopes
```

```
## Multilevel Model 1

mm1 <- lmer(CD4PCT ~ time + (1|id), data=cd4_group) # partial pooling
summary(mm1)
#fixef(mm1) # fixed effects (time)
#ranef(mm1) # random effect estimates (residuals) <- ranef(mm1)=resid(mm1)
#coef(mm1) # coefficients (random intercepts and fixed slope for time)

mean_cd4pct <- c() # individual CD4 % means
for (i in id){
  mean <- coef(lm(CD4PCT ~ 1, data=cd4_group[which(cd4_group$id==i), ])) # individual means (no pooling
  mean_cd4pct <- c(mean_cd4pct, mean)
}
individuals_mm1 <- as.data.frame(cbind(id=individuals_lm$id, # individual ids
                                       mean_cd4pct=as.vector(mean_cd4pct), # individual means
                                       intercept_lm=individuals_lm$intercept, # pp intercept - lm
                                       time_lm=individuals_lm$slope_time, # pp slope - lm
                                       intercept_mm=coef(mm1)$id[,"(Intercept)"], # pp intercept - lmer
                                       time_mm=coef(mm1)$id[,"time"], # pp slope - lmer
                                       intercept_resid=ranef(mm1)$id[,"(Intercept)"])) # residuals
```

```
## Multilevel Model 2

mm2 <- lmer(CD4PCT ~ time + trt + baseage + (1|id), data=cd4_group) # partial pooling
summary(mm2) # - residual variance for individuals has decreased
#fixef(mm2) # fixed effects (time, trt, baseage)
#ranef(mm2) # random effect estimates (residuals)
#coef(mm2) # coefficients (random intercepts and fixed slopes)

# getting coefficient estimates
mm2_coefs <- as.data.frame(cbind(id=individuals_lm$id, # individual ids
                                 mean_cd4pct=as.vector(mean_cd4pct), # individual means
                                 intercept=coef(mm2)$id[,"(Intercept)"], # random intercepts
                                 residual=ranef(mm2)$id[,"(Intercept)"], # residuals
                                 time=coef(mm2)$id[,"time"], # fixed time slope
                                 treatment=coef(mm2)$id[,"trt2"], # fixed treatment slope
                                 baseage=coef(mm2)$id[,"baseage"])) # fixed baseage slope
```

```
## Tables (Comparing Multilevel Models 1 and 2)

# comparing residual, (time) slope, and intercept estimates to those obtained in part (a)
individuals_mm2 <- individuals_mm1[, -c(2, 3, 4)] %>%
  mutate(intercept_mm2=coef(mm2)$id[, "(Intercept)"],
         time_mm2=coef(mm2)$id[, "time"],
         intercept_resid2=ranef(mm2)$id[, "(Intercept)"])

# comparing only residual and intercept estimates
mm1vmm2_randef <- individuals_mm2[, -c(3, 6)] %>%
  rename("ID"=id,
         "Intercept (Model 1)"=intercept_mm,
         "Residual (Model 1)"=intercept_resid,
         "Intercept (Model 2)"=intercept_mm2,
         "Residual (Model 2)"=intercept_resid2)

# model 1 random intercepts, residuals, and SE
mm1_randef <- mm1vmm2_randef[, -c(4, 5)] %>%
  mutate(se=se.ranef(mm1)$id[, "(Intercept)"]) %>%
  rename("Intercept"=2, "Residual"=3, "SE"=se)

# model 2 random intercepts, residuals, and SE
mm2_randef <- mm1vmm2_randef[, -c(2, 3)] %>%
  mutate(se=se.ranef(mm2)$id[, "(Intercept)"]) %>%
  rename("Intercept"=2, "Residual"=3, "SE"=se)

# comparing fixed effects (slopes for time)
mms <- c(1, 2)
ints <- c(24.980845, 27.535647)
ints_se <- c(0.819118, 1.588992)
slopes <- c(-0.008236, -0.008135)
slopes_se <- c(0.001409, 0.001409)
mm1vmm2_fixedef <- as.data.frame(cbind(mms, ints, ints_se, slopes, slopes_se)) %>%
  rename("Multilevel Model"=mms,
         "Intercept"=ints,
         "Intercept SE"=ints_se,
         "Time Slope"=slopes,
         "Time Slope SE"=slopes_se)

# comparing group-level and individual-level variabilities
#sigma.hat(mm2)$sigma$id # group-level standard deviation
#sigma.hat(mm2)$sigma$data # individual-level standard deviation
bet_ind <- c(131.37, 127.23)
with_ind <- c(53.95, 53.97)
mm1vmm2_vars <- as.data.frame(cbind(mms, bet_ind, with_ind)) %>%
  rename("Multilevel Model"=mms,
         "Group-Level Variability"=bet_ind,
         "Individual-Level Variability"=with_ind)
```

```
## Observed vs. Predicted/Fitted Values

pred_mm1 <- as.vector(fitted(mm1)) # model 1
pred_mm2 <- as.vector(fitted(mm2)) # model 2
obsvpred <- as.data.frame(cbind(id=cd4_group$id,
                                time=cd4_group$time,
                                observed=cd4_group$CD4PCT,
                                pred_mm1=pred_mm1,
                                pred_mm2=pred_mm2))

# observed vs. predicted/fitted values for 10 individuals
randsample_10_obsvpred <- obsvpred[which(obsvpred$id %in% randsample_10), ]
randsample_10_obsvpred2 <-  pivot_longer(randsample_10_obsvpred, c(3, 4, 5), names_to="value_type")

# plot for first 5 individuals
obsvpred_plot1 <- ggplot(randsample_10_obsvpred2[which(randsample_10_obsvpred2$id<=151), ],
       aes(x=time, y=value, color=as.factor(id))) +
  geom_point(aes(shape=value_type)) +
  geom_line(aes(lty=value_type)) +
  scale_y_sqrt() +
  labs(title="Figure 8: Observed vs. Predicted Trajectories of CD4 %",
       subtitle="First 5/10 Randomly Selected Individuals",
       x="Time (in Days) After the First Visit",
       y="CD4 %",
       color="ID",
       shape="Value Type",
       lty="Value Type")

# plot for remaining 5 individuals
obsvpred_plot2 <- ggplot(randsample_10_obsvpred2[which(randsample_10_obsvpred2$id>=151), ],
       aes(x=time, y=value, color=as.factor(id))) +
  geom_point(aes(shape=value_type)) +
  geom_line(aes(lty=value_type)) +
  scale_y_sqrt() +
  labs(title=" ",
       subtitle="Last 10/10 Randomly Selected Individuals",
       x="Time (in Days) After the First Visit",
       y="CD4 %",
       color="ID",
       shape="Value Type",
       lty="Value Type")
```

```
## Metric for New Time Points

# unique times in order
time_vec <- as.vector(unique(cd4_group$time))
time_ord <- c(1, rep(0, length(time_vec)-1))
time_df <- data.frame(time_vec, time_ord)
for (i in 2:length(time_vec)){
  time_df$time_ord[i] <- min(time_df[which(time_df$time_vec>time_df$time_ord[i-1]), 1])
}

# average times across individuals with the same number of observations
for (i in 1:7){
  obs <- cd4_group[which(cd4_group$observations==i),]
  row.names(obs) <- 1:nrow(obs)

  obs_avgs <- c()
  for (j in 1:i){
    pattern <- seq(j, nrow(obs), i)
    avg <- round(mean(obs[which(as.numeric(row.names(obs)) %in% pattern), 4]))
    obs_avgs <- c(obs_avgs, avg)
  }
  print(obs_avgs)
}

# time increases (changes between time points) for each group individuals with the same number of obser
diff2 <- c(138) # mean=138
diff3 <- c(131, 155) # mean=143
diff4 <- c(129, 131, 133) # mean=131
diff5 <- c(118, 99, 112, 131) # mean=115
diff6 <- c(103, 94, 89, 92, 107) # mean=97
diff7 <- c(86, 87, 88, 89, 88, 86)  # mean=87

# metric:
# new added times = individual's max time value + mean time increase for corresponding group
mean(diff2) # 138
mean(diff3) # 143
mean(diff4) # 131
mean(diff5) # 115
mean(diff6) # 97
mean(diff7) # 87
# new time point for individuals with 1 observation = 1 + 140 = 141
```

```
## Getting New Times (based on metric)

id <- individuals_lm$id
max_times <- rep(0, length(id))
new_times <- rep(0, length(id))
new_times_df <- data.frame(id, max_times, new_times)

for (i in id){
  new_times_df[which(new_times_df$id==i), ]$max_times <- max(cd4_group[which(cd4_group$id==i), "time"])
}

for (i in 1:length(id)){
  if (cd4_group[which(cd4_group$id==new_times_df[i, "id"]), "observations"]==1){
    new_times_df$new_times[i] <- new_times_df$max_times[i] + 140
  } else if (cd4_group[which(cd4_group$id==new_times_df[i, "id"]), "observations"]==2){
    new_times_df$new_times[i] <- new_times_df$max_times[i] + 138
  } else if (cd4_group[which(cd4_group$id==new_times_df[i, "id"]), "observations"]==3){
    new_times_df$new_times[i] <- new_times_df$max_times[i] + 143
  } else if (cd4_group[which(cd4_group$id==new_times_df[i, "id"]), "observations"]==4){
    new_times_df$new_times[i] <- new_times_df$max_times[i] + 131
  } else if (cd4_group[which(cd4_group$id==new_times_df[i, "id"]), "observations"]==5){
    new_times_df$new_times[i] <- new_times_df$max_times[i] + 115
  } else if (cd4_group[which(cd4_group$id==new_times_df[i, "id"]), "observations"]==6){
    new_times_df$new_times[i] <- new_times_df$max_times[i] + 97
  } else if (cd4_group[which(cd4_group$id==new_times_df[i, "id"]), "observations"]==7){
    new_times_df$new_times[i] <- new_times_df$max_times[i] + 87
  }
}
```

```
## Obtaining Predicted Values for New Time Points (based on multilevel model 2 fit)

# covariate values for each individual
trt <- rep(0, length(id))
baseage <- rep(0, length(id))
mm2_new_preds <- new_times_df[, c(1, 3)] %>%
  mutate(trt=trt,
         baseage=baseage,
         rand_int=coef(mm2)$id[, "(Intercept)"])

for (i in id){
  mm2_new_preds[which(mm2_new_preds$id==i), ]$trt <-
    mean(as.numeric(cd4_group[which(cd4_group$id==i), "trt"]))
  mm2_new_preds[which(mm2_new_preds$id==i), ]$baseage <-
    mean(cd4_group[which(cd4_group$id==i), "baseage"])
}

mm2_new_preds$trt2 <- (mm2_new_preds$trt)-1

# predicted CD4 % values for new time points based on model 2 fit
time_coef <- -0.008135
trt2_coef <- 1.406590
baseage_coef <- -0.948838
# beta1 <- fixef(mm2)["time"]
# beta2 <- fixef(mm2)["trt2"]
# beta3 <- fixef(mm2)["baseage"]
pred <- rep(0, length(id))
mm2_new_preds <- mm2_new_preds %>% mutate(pred=pred)

for (i in 1:length(id)){
  mm2_new_preds$pred[i] <-
    mm2_new_preds$rand_int[i] +
    mm2_new_preds$new_times[i]*time_coef +
    mm2_new_preds$trt2[i]*trt2_coef +
    mm2_new_preds$baseage[i]*baseage_coef
}
```

```
## Simulated Predictions for New Time Points
set.seed(47)

sim_preds <- c()
for (i in 1:nrow(mm2_new_preds)){
  preds <- rnorm(1000,
                 mm2_new_preds$rand_int[i] -
                   0.008135*mm2_new_preds$new_time[i] +
                   1.406590*mm2_new_preds$trt2[i] -
                   0.948838*mm2_new_preds$baseage[i],
                 sigma.hat(mm2)$sigma$data)
  sim_preds[i] <- mean(preds)
}
```

```
## Comparing Individual Trajectories with Old/New Time Points

randsample_10_preds <- randsample_10_df[, c(1, 4, 5)]
```

```r
for (i in randsample_10){
  new_row <- c(i,
               mm2_new_preds[which(mm2_new_preds$id==i), ]$new_times,
               round(mm2_new_preds[which(mm2_new_preds$id==i), ]$pred, 1))
  randsample_10_preds <- rbind(randsample_10_preds, new_row)
}

randsample_10_preds <- randsample_10_preds %>%
  group_by(id) %>%
  arrange(.by_group=TRUE)
randsample_10_preds

rand_10_new_preds1 <- ggplot(randsample_10_df, aes(x=time, y=CD4PCT, color=id)) +
  geom_point() +
  geom_line() +
  scale_y_sqrt() +
  xlim(0, 900) +
  labs(title="Figure 9.1: Individual Trajectories with Original Time Points",
       x="Time (in Days) After the First Visit",
       y="CD4 % (Square Root Scale)",
       color="ID")

rand_10_new_preds2 <- ggplot(randsample_10_preds,
                             aes(x=as.numeric(time), y=as.numeric(CD4PCT), color=id)) +
  geom_point() +
  geom_line() +
  scale_y_sqrt() +
  xlim(0, 900) +
  labs(title="Figure 9.2: Individual Trajectories with New Time Points",
       x="Time (in Days) After the First Visit",
       y="CD4 % (Square Root Scale)",
       color="ID")
```

```r
## Plotting Prediction Distributions for First 3 Individuals
set.seed(47)

sim_preds_id1 <- rnorm(1000,
                       mm2_new_preds$rand_int[1] -
                         0.008135*mm2_new_preds$new_time[1] +
                         1.406590*mm2_new_preds$trt2[1] -
                         0.948838*mm2_new_preds$baseage[1],
                       sigma.hat(mm2)$sigma$data)
sim_preds_id2 <- rnorm(1000,
                       mm2_new_preds$rand_int[2] -
                         0.008135*mm2_new_preds$new_time[2] +
                         1.406590*mm2_new_preds$trt2[2] -
                         0.948838*mm2_new_preds$baseage[2],
                       sigma.hat(mm2)$sigma$data)
sim_preds_id3 <- rnorm(1000,
                       mm2_new_preds$rand_int[3] -
                         0.008135*mm2_new_preds$new_time[3] +
                         1.406590*mm2_new_preds$trt2[3] -
                         0.948838*mm2_new_preds$baseage[3],
                       sigma.hat(mm2)$sigma$data)

sim_preds_ids <- boxplot(sim_preds_id1, sim_preds_id2, sim_preds_id3,
                         names = c("ID=1", "ID=2", "ID=3"),
                         ylab="CD4 %",
                         main="Figure 10: Prediction for First 3 Individuals with New Time Point",
                         col="orange")

#sim_preds_ids$stats (boxplot stats)
```

```r
## Simulating Predictions for New Individual with 7 Time Points
set.seed(47)

id9 <- as.data.frame(cbind(id=rep(9, 7),
                           time=c(1, 87, 174, 262, 351, 439, 525),
                           trt2=rep(1, 7),
                           baseage=rep(4, 7)))

beta0 <- fixef(mm2)["(Intercept)"] #int_coef <- 27.53565
beta1 <- fixef(mm2)["time"] # time_coef <- -0.008135
beta2 <- fixef(mm2)["trt2"] # trt2_coef <- 1.406590
beta3 <- fixef(mm2)["baseage"] # baseage_coef <- -0.948838

sigma_beta0j <- sigma.hat(mm2)$sigma$id # group-level standard deviation
sigma_y_hat <- sigma.hat(mm2)$sigma$data # individual-level standard deviation

# random intercepts for new individual
beta0j <- rnorm(1000, beta0, sigma_beta0j)

# matrix with random intercepts & fixed slopes
rand_coefs_mat <- as.matrix(cbind(beta0+beta0j, beta1, beta2, beta3))

y_trt1_list <- c() # treatment 1
y_trt2_list <- c() # treatment 2
for (i in id9$time){
  x_trt1 <- c(1, i, 0, 4)
  x_trt2 <- c(1, i, 1, 4)
  y_trt1 <- rnorm(1000, rand_coefs_mat %*% x_trt1, sigma_y_hat)
  y_trt2 <- rnorm(1000, rand_coefs_mat %*% x_trt2, sigma_y_hat)
  y_trt1_list <- c(y_trt1_list, y_trt1)
  y_trt2_list <- c(y_trt2_list, y_trt2)
}
```

```r
## Plotting Distributions for Simulated Predictions for New Individual (with 7 Time Points)

# distributions for each time point for treatment 1
boxplot(y_trt1_list[1:1000],
        y_trt1_list[1001:2000],
        y_trt1_list[2001:3000],
        y_trt1_list[3001:4000],
        y_trt1_list[4001:5000],
        y_trt1_list[5001:6000],
        y_trt1_list[6001:7000],
        names=c("1", "2", "3", "4", "5", "6", "7"),
        xlab="Treatment 1",
        ylab="CD4 %",
        main="Figure 11.1: Predictions for New Individual with 7 Time Points",
        sub=" ",
        border="red",
        col="orange")

# distributions for each time point for treatment 2
boxplot(y_trt2_list[1:1000],
        y_trt2_list[1001:2000],
        y_trt2_list[2001:3000],
        y_trt2_list[3001:4000],
        y_trt2_list[4001:5000],
        y_trt2_list[5001:6000],
        y_trt2_list[6001:7000],
        names=c("1", "2", "3", "4", "5", "6", "7"),
        xlab="Treatment 2",
        ylab="CD4 %",
        main="Figure 11.2: Predictions for New Individual with 7 Time Points",
        sub="Time Point",
        border="red",
        col="orange")
```