exercises_week13

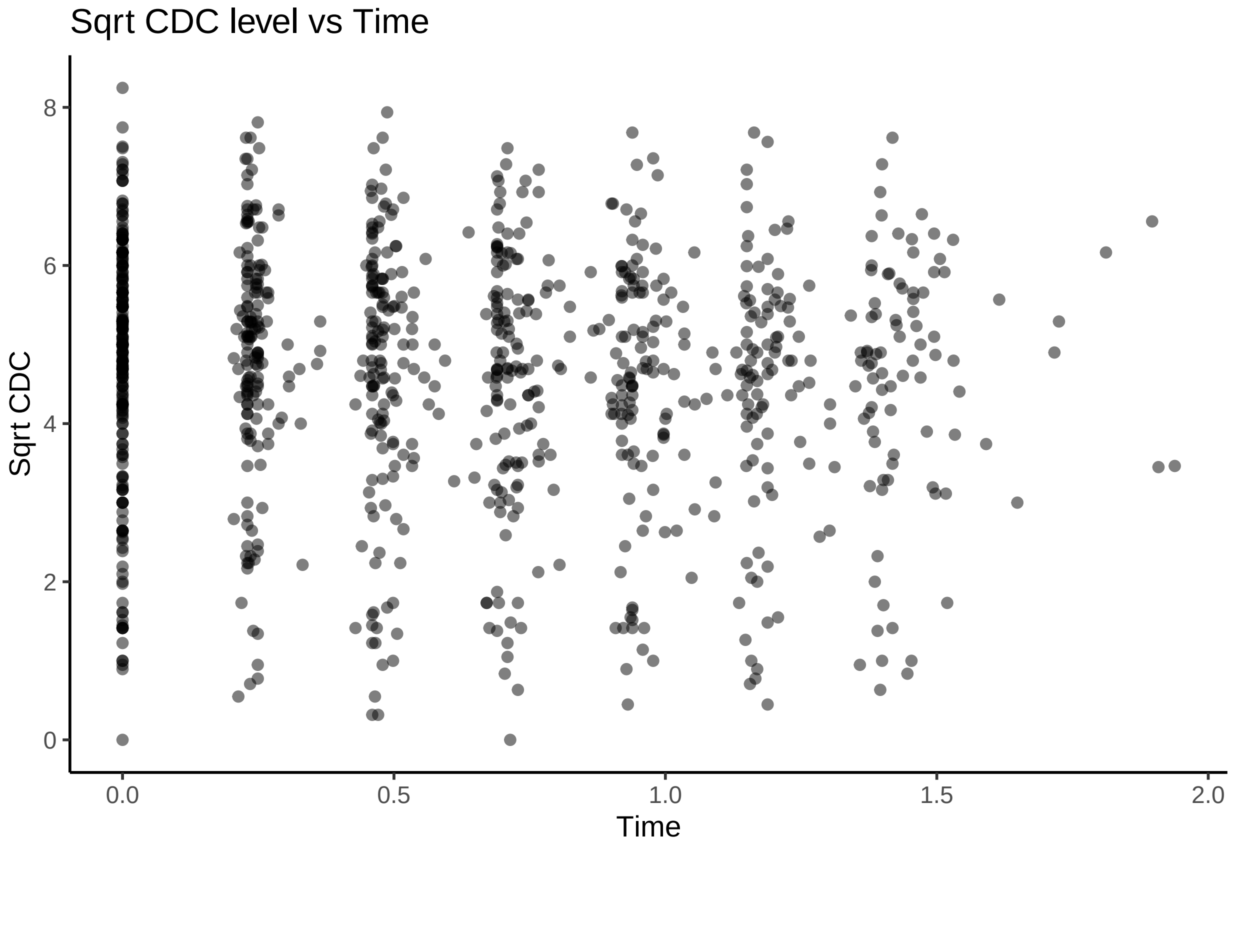
Lindsey Greenhill

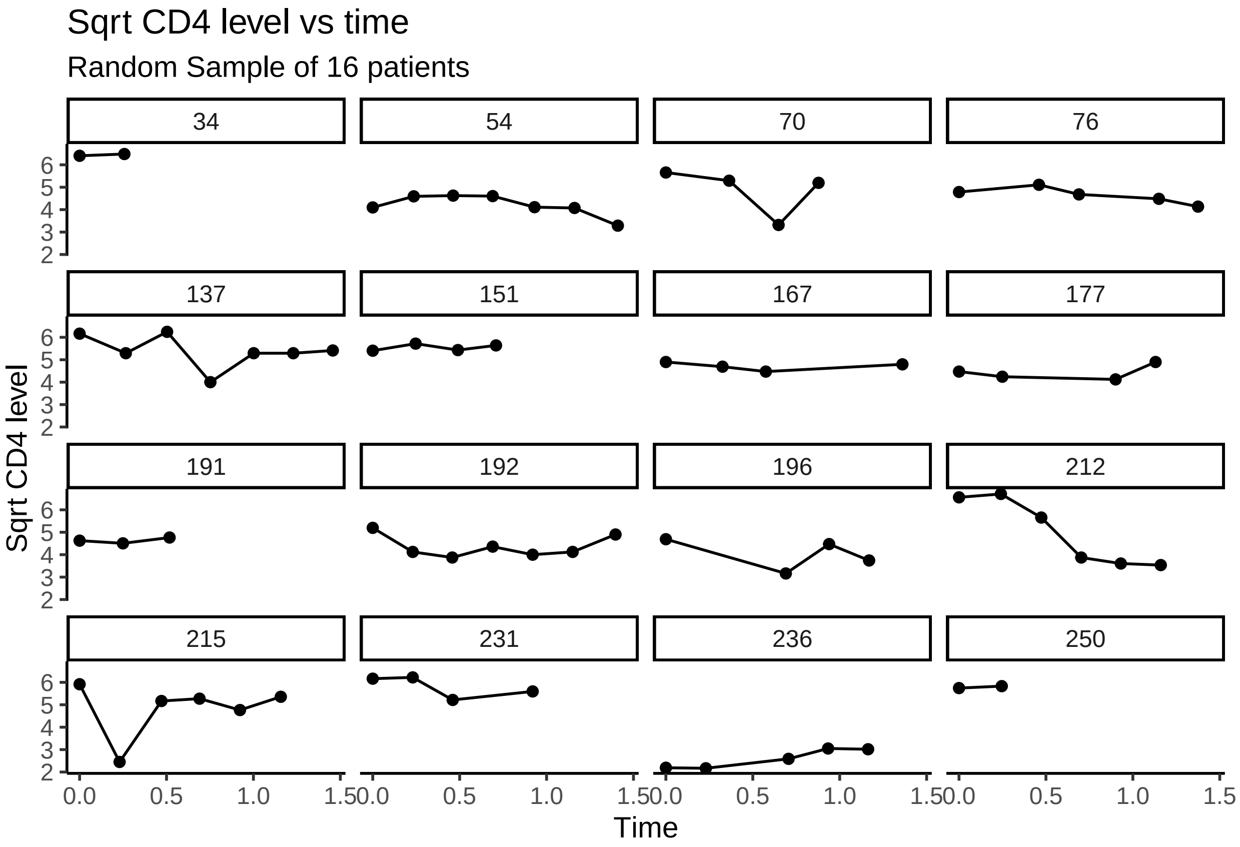
4/28/2021

Question 11.4

Part a

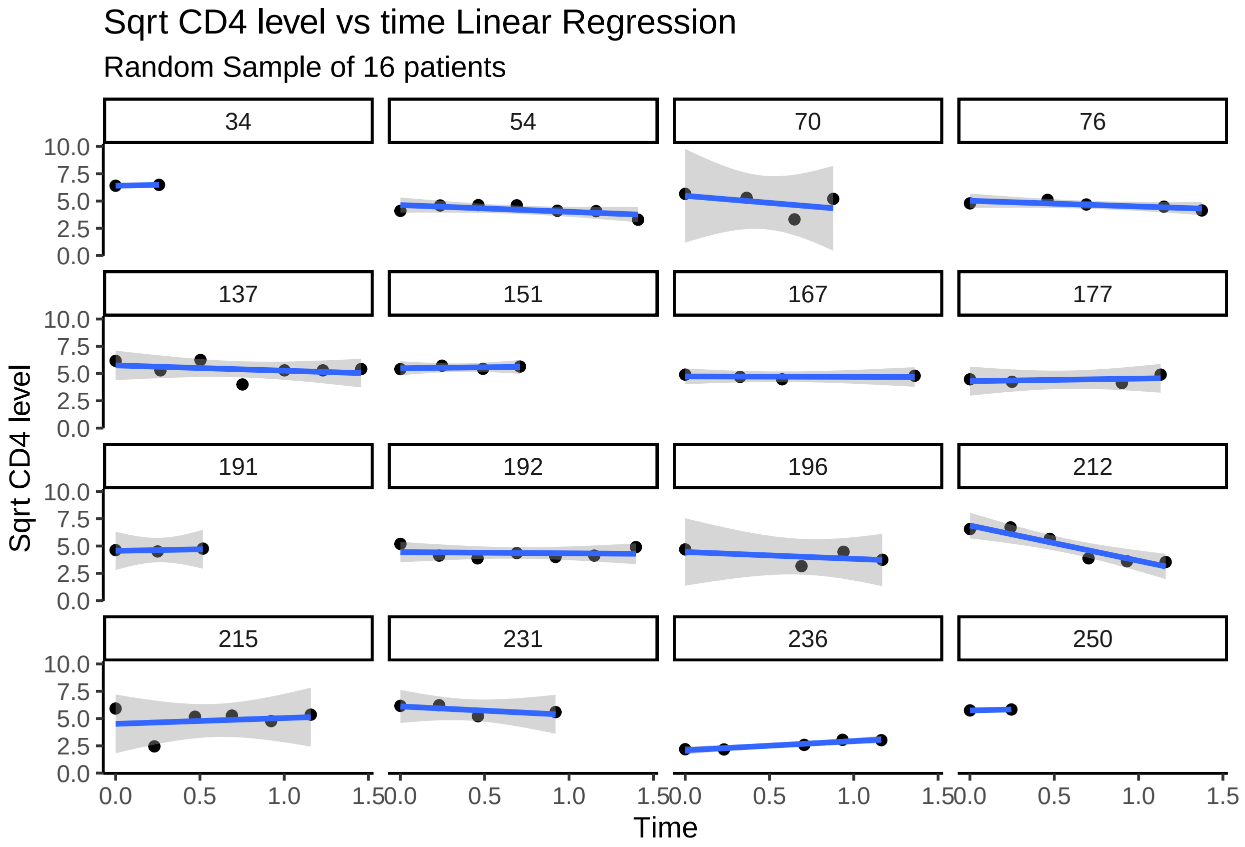
In the code below, I graph the square root CD4 percentage against time for all patients and for 16 randomly sampled patients. Both graphs show that there doesn't appear to be a consistent pattern amongst the patients. Some patients seems to have a negative correlation between CD4 levels and time while other appear to have no correlation or positive correlation.





Part b

In the code below, I show the linear fit of sqrt CD4 level vs time for the 16 randomly sampled individuals from part a. As noted in part a, there does not seem to be a consistent relationship between CD4 levels and time amongst these 16 patients.



Part c

In the code below, I first calculate the slopes and intercepts of CD4 level for each patient in the data. In the first step, I group the data by newpid to generate intercepts and slopes specific to each indiviudal patient. In the next step, I regressed the intercept and slopes from the individual models on baseage and treatment to get a

```
# creating regression coefficients grouped by newpid

c_mod <- data %>%
  group_by(newpid, baseage, treatmnt) %>%
  summarise(intercept = coef(lm(sqrtpct ~ time))[1],
         slope = coef(lm(sqrtpct ~ time))[2])

# creating models to explain the differences in between the children

c_mod_1 <- lm(intercept ~ baseage + treatmnt, data = c_mod)
c_mod_2 <- lm(slope ~ baseage + treatmnt, data = c_mod)

summary(c_mod_1)
```

```
##
## Call:
## lm(formula = intercept ~ baseage + treatmnt, data = c_mod)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.0355 -0.7107  0.1900  1.0598  2.9416 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  4.83501    0.33819  14.297  < 2e-16 ***
## baseage     -0.12388    0.04367  -2.837  0.00497 ** 
## treatmnt     0.25301    0.19835   1.276  0.20344    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.487 on 223 degrees of freedom
## Multiple R-squared:  0.04151,    Adjusted R-squared:  0.03291 
## F-statistic: 4.829 on 2 and 223 DF,  p-value: 0.008852
```

```r
summary(c_mod_2)
```

```
## 
## Call:
## lm(formula = slope ~ baseage + treatmnt, data = c_mod)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -12.8064  -0.4692   0.1309   0.6836   6.0357 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.25675    0.43854  -0.585    0.559
## baseage     -0.02134    0.05862  -0.364    0.716
## treatmnt    -0.06038    0.25898  -0.233    0.816
## 
## Residual standard error: 1.829 on 199 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:  0.0009493,  Adjusted R-squared:  -0.009091 
## F-statistic: 0.09454 on 2 and 199 DF,  p-value: 0.9098
```

The first summary shows the regression output for the intercept model. The coefficient for baseage means that the model predicts that on average for every 1 year increase in baseage there is a .12 point decrease in the intercept for in between children holding all else constatn. The model predicts that that treated children will on average have an intercept .25 points higher compared to non treated children, holding all else constant.

The second summary shows the regression output for the slope model. The coefficient for baseage means that the model predicts that on average for every 1 year increase in baseage a child's slope will be .02 points lower, holding all else constant. The model predicts that that treated children will on average have a slope .06 points lower compared to non treated children, holding all else constant.

Question 12.2

Part a

In the code below, I create a model that predicts CD4 levels as a function of time, varying intercepts across children.

```
mod_12_a <- lmer(sqrtpct ~ time + (1 | newpid),
                 data = data)

# can’t use stargzer with this model

summary(mod_12_a)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: sqrtpct ~ time + (1 | newpid)
##    Data: data
##
## REML criterion at convergence: 2844.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.7925 -0.4403 -0.0007  0.4433  5.0666
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  newpid   (Intercept) 1.9471   1.3954
##  Residual             0.5836   0.7639
## Number of obs: 978, groups:  newpid, 226
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  4.80737    0.10108  47.559
## time        -0.38564    0.05586  -6.904
##
## Correlation of Fixed Effects:
##      (Intr)
## time -0.278
```

- Coefficient of time: -.38. For every year following the initial visit, the patient’s square root CD4 level is predicted to decrease by .38 on average. This coefficient is statistically significant.
- Additionally, it seems like grouping the model by patient improves the model because the sd of its random effects is larger than that of the residual.

Part b

In the code below, I extend the model in part a to include more predictors, specifically treatment and age at baseline.

```
mod_12_b <- lmer(sqrtpct ~ time + treatmnt + baseage + (1 | newpid),
                   data = data)

summary(mod_12_b)
```

```
## Linear mixed model fit by REML [’lmerMod’]
## Formula: sqrtpct ~ time + treatmnt + baseage + (1 | newpid)
##    Data: data
##
## REML criterion at convergence: 2840.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.8055 -0.4468  0.0096  0.4402  5.0816
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  newpid   (Intercept) 1.8688   1.3670
##  Residual             0.5839   0.7641
## Number of obs: 978, groups:  newpid, 226
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  4.76394    0.32715  14.562
## time        -0.38216    0.05586  -6.841
## treatmnt     0.29442    0.19114   1.540
## baseage     -0.11827    0.04225  -2.799
##
## Correlation of Fixed Effects:
##          (Intr) time   trtmnt
## time     -0.082
## treatmnt -0.854  0.004
## baseage  -0.421 -0.016 -0.002
```

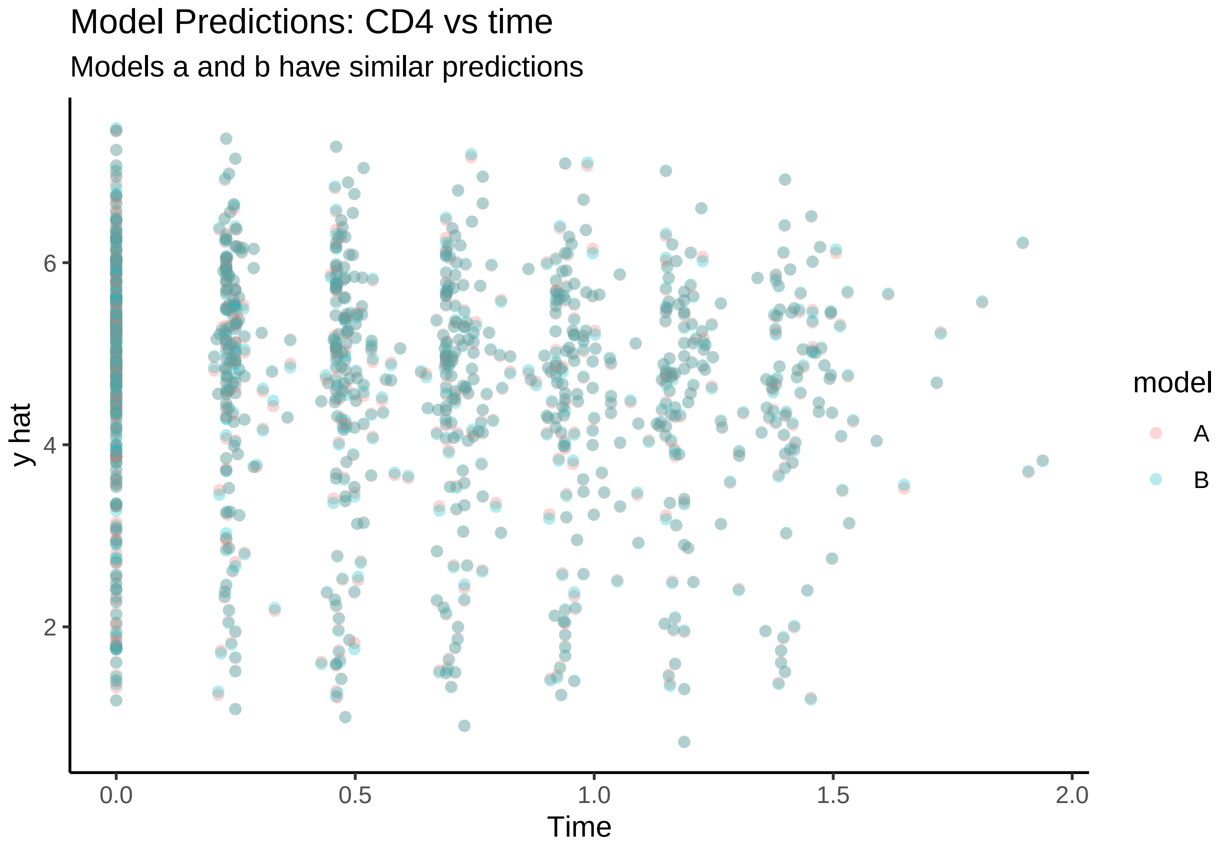
Coefficient interpretations

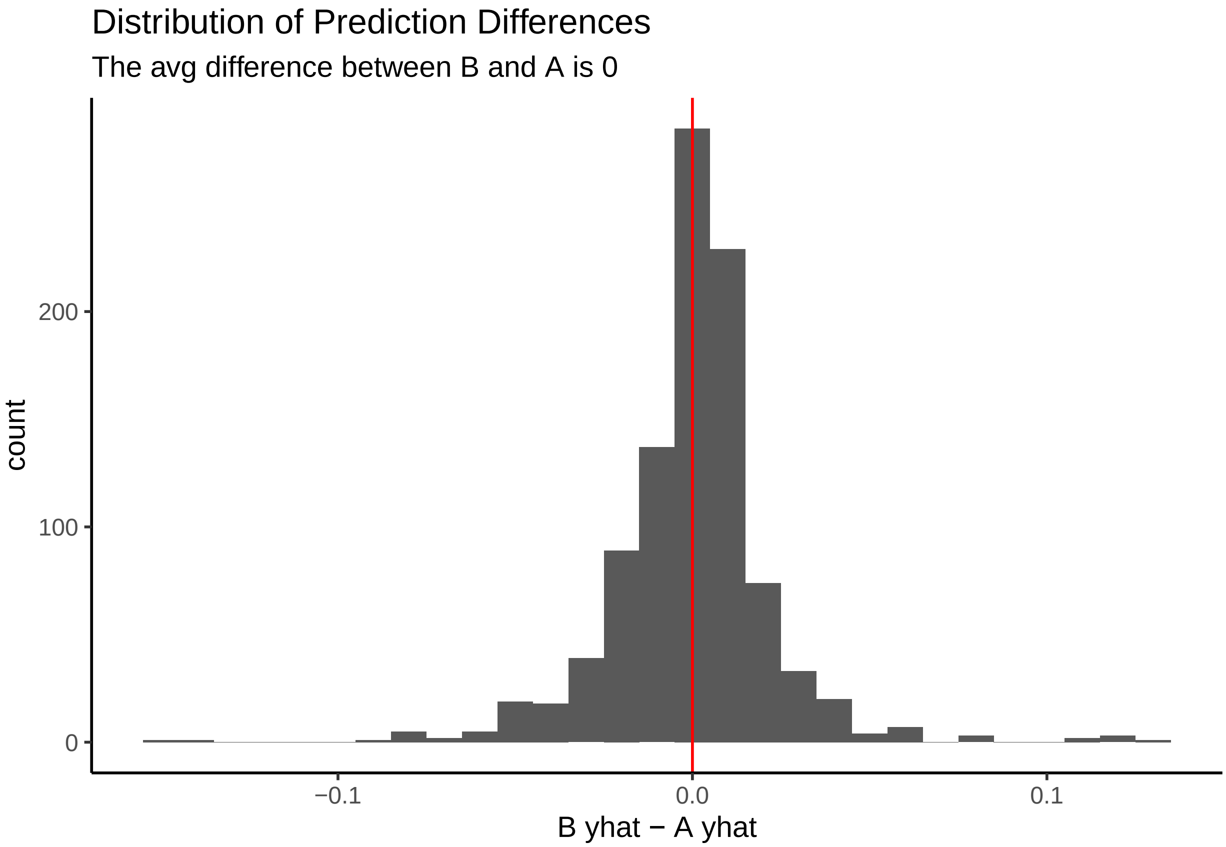
- Coefficient of time: -.38. For every year following the initial visit, the patient’s square root CD4 level is predicted to decrease by .38 on average, holding all else constant. This coefficient is statistically significant.
- Coefficient of base age: -.118. for every increase in the base age of a patient, the model predicts that the square root CD4 level will decrease by .11 on average, holding all else constant. This coefficient is statistically significant.
- Coefficient of treatment: .294. The model predicts that the patients who received treatment have on average .294 higher square root CD4 levels than those who did not receive the treatment, holding all else constant. The coefficient is not statistically significant.

Part c and d

In the code below, I look at the y hats for the models made in part a and part b above. I then plot these y hats against the only shared predictor in the models, time since baseline visit. The predictions from model a and the predictions from model b appear to be more or less the same. I confirmed this by looking at the distribution of the differences of predictions between b and a in the histogram below. The average difference in predictions is 0.

Additionally, we can compare the results from part a and b by looking at the different sigma and alpha values. For the first model, sigma = 1.39 and alpha = .76, meaning that the variation between groups is more than the variation within groups. For the second model, sigma = 1.36, which is less than sigma in the first model. This decrease suggests that adding the two predictors of baseage and treatment reduces variation between groups.




Distribution of Prediction Differences
The avg difference between B and A is 0
count
200
100
0
−0.1
0.0
0.1
B yhat − A yhat