StatR 502 Homework 8 (Last HW!)

Rebecca Hadi

Due Thursday, March 8, 2018, 6:30 pm

Contents

Submission guidelines: please submit a knitted PDF or Word document, and optionally your `.Rmd` file. As always, ask in the discussion forum if you're having trouble!

Remember that this homework covers two weeks of class. Do try to get started early, but some topics won't be covered in the first lecture—the lab and next lecture will help.

All of these problems rely on the CD4 data available at http://www.stat.columbia.edu/~gelman/arm/examples/cd4/allvar.csv. As a response, you should use the square root of the CD4 percentage (`CD4PCT`). (For some context, CD4 cells are white blood cells that get infected by HIV. The count of CD4 cells in a blood sample is often used as a measure of progression to AIDS, with lower counts indicating a weaker immune system.) Each subject is uniquely identified by the `newpid` column. As the main time variable (and variable whose coefficient we will call "slope"), use the time since treatment began in years, that is, the difference `visage - baseage`.

The subjects are divided into two treatment groups, with `treatmnt = 1` as the control and `treatmnt = 2` as the experimental group. I'm not sure what the treatment being evaluated was, so we'll follow G&H's lead in ignoring it (though, optionally, you could add in a treatment covariate and assess whether or not you think the treatment is promising. If we assume the treatment was randomly assigned at the first visit, it *shouldn't* have any effect on the intercept. Thus, this is a case where we would only include the treatment interacted with the time variable, leaving its main effect out as that would fit adjust the intercept.)

Do:

1. Final Project Question

- **1:** In your *final project data*, describe your response variable and any categorical predictor variables you have. Which of the categorical variables do you think might work as groupings for random effects (e.g., do they have enough levels, are they a sample of a population, etc.)? Is there nesting? Do you think mixed methods might improve your modeling approach?

Response Variable: Binary indicator of whether or not movie was profitable (i.e. revenue > budget)
Categorical Predictor Variables:
* Genre
* Original Language

I think the variable "genre" will be a good candidate for multi-level modeling given that there are a large number of factor levels and it would be difficult to interpret each one on its own. I will certainly evaluate this approach in my final project.

Upon inspection of my data set, the data are highly skewed toward movies released in English, so much so that I have been considering sub-setting the data to only include movies released in English to increase

the validity of the model. I think there is a different level of resourcing and advertising for movies that are released in English (and assumed to be from the US) opposed to other small countries that do not have such a large film industry.

G & H 11.4

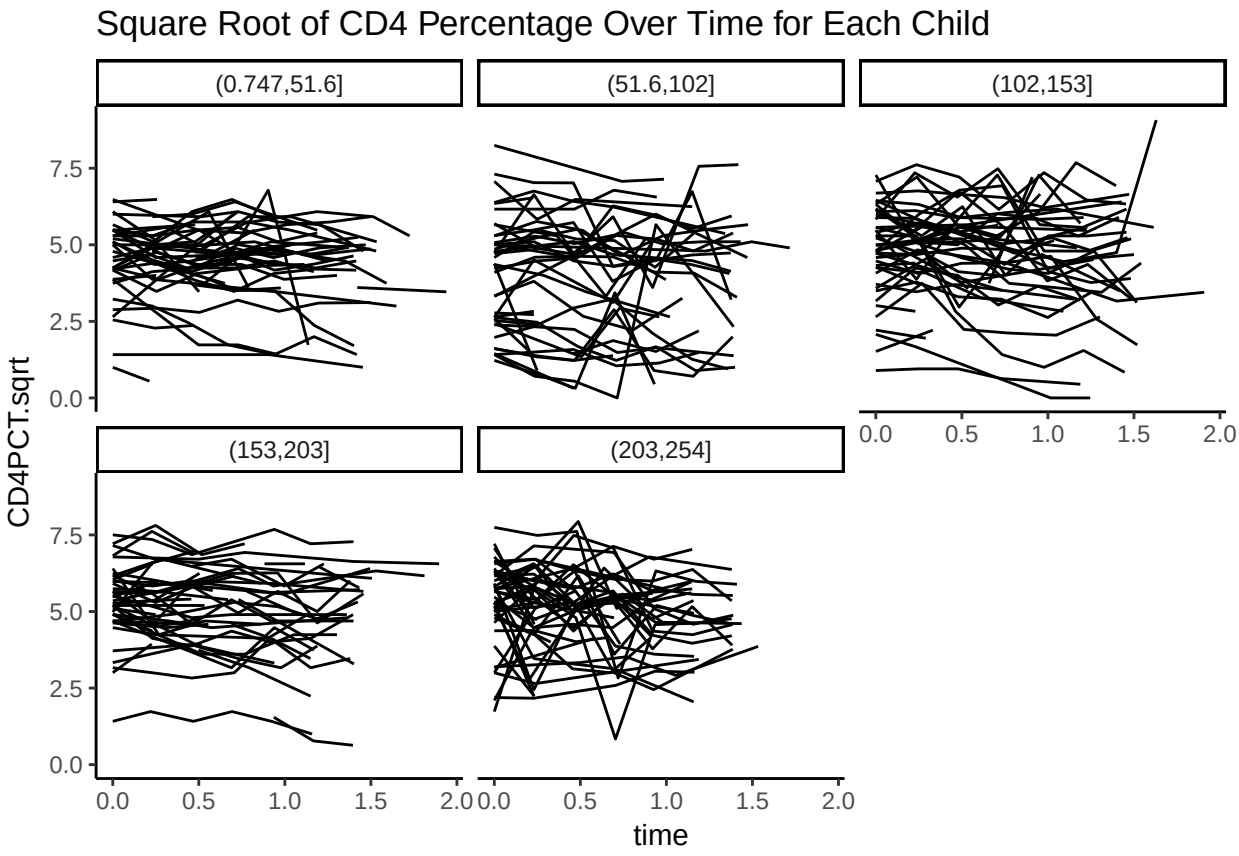
- **G&H 11.4:** Chapter 11, #4 (page 249)

```
#load data
cd4 <- read.csv("http://www.stat.columbia.edu/~gelman/arm/examples/cd4/allvar.csv")
```

The folder cd4 has CD4 percentages for a set of young children with HIV who were measured several times over a period of two years. The data set also includes the ages of the children at each measurement.

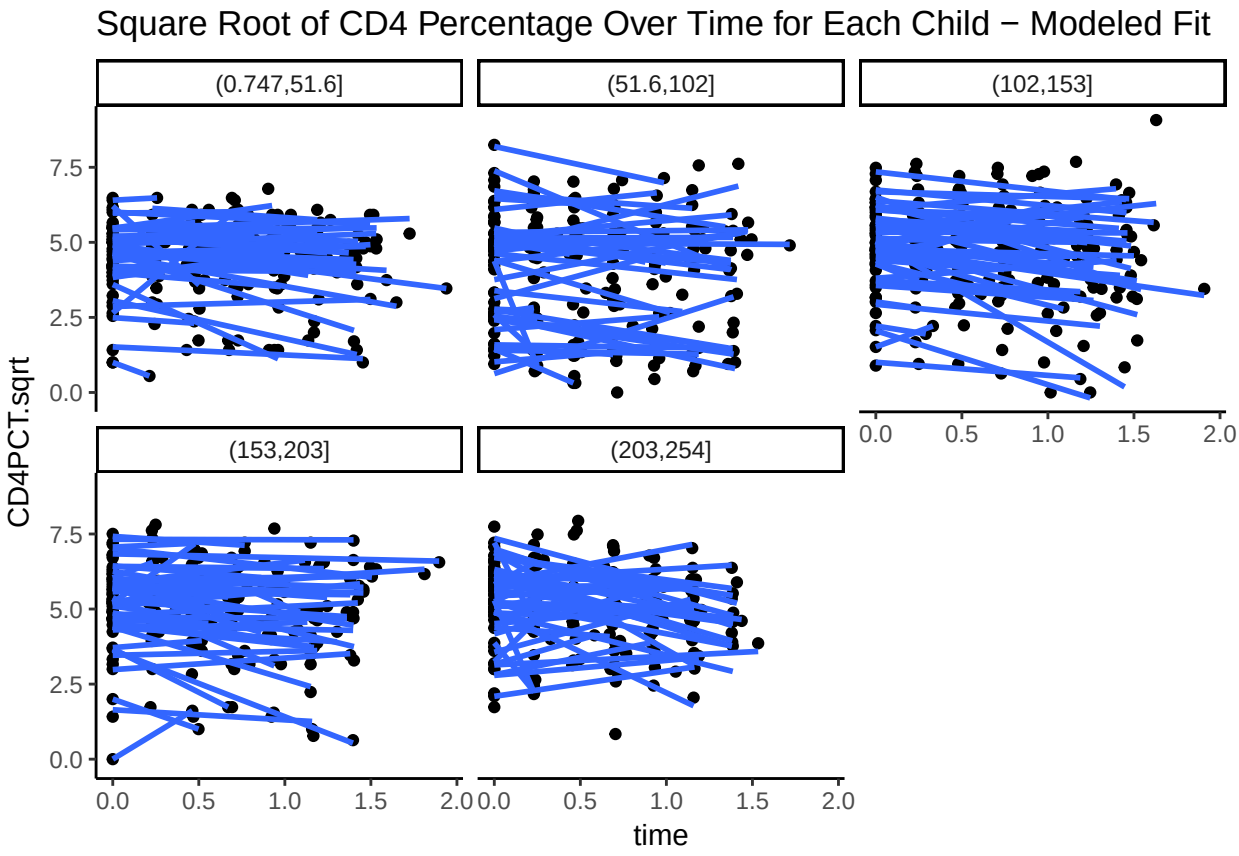
(a) Graph the outcome (the CD4 percentage, on the square root scale) for each child as a function of time.

```
#load ggplot2
library(ggplot2)

#add sqrt variable
cd4$CD4PCT.sqrt <- sqrt(cd4$CD4PCT)

#create time variable
cd4$time <- cd4$visage - cd4$baseage

#cut newpid into groups for wrapping
cd4$newpid.group <- cut(cd4$newpid, breaks = 5)

#create plot
ggplot(data = cd4, aes(x = time, y = CD4PCT.sqrt, group = newpid), na.rm = T) +
      geom_line() +
      theme_classic() +
      facet_wrap (~newpid.group) +
      ggtitle("Square Root of CD4 Percentage Over Time for Each Child")
```



(b) Each child's data has a time course that can be summarized by a linear fit. Estimate these lines and plot them for all the children.

```
#create plot with estimated line
ggplot(data = cd4, aes(x = time, y = CD4PCT.sqrt, group = newpid), na.rm = T) +
      geom_point() +
      geom_smooth(method = "lm", alpha = 0.25, se = F, aes(group = newpid)) +
      theme_classic() +
      facet_wrap (~newpid.group) +
      ggtitle("Square Root of CD4 Percentage Over Time for Each Child - Modeled Fit")
```



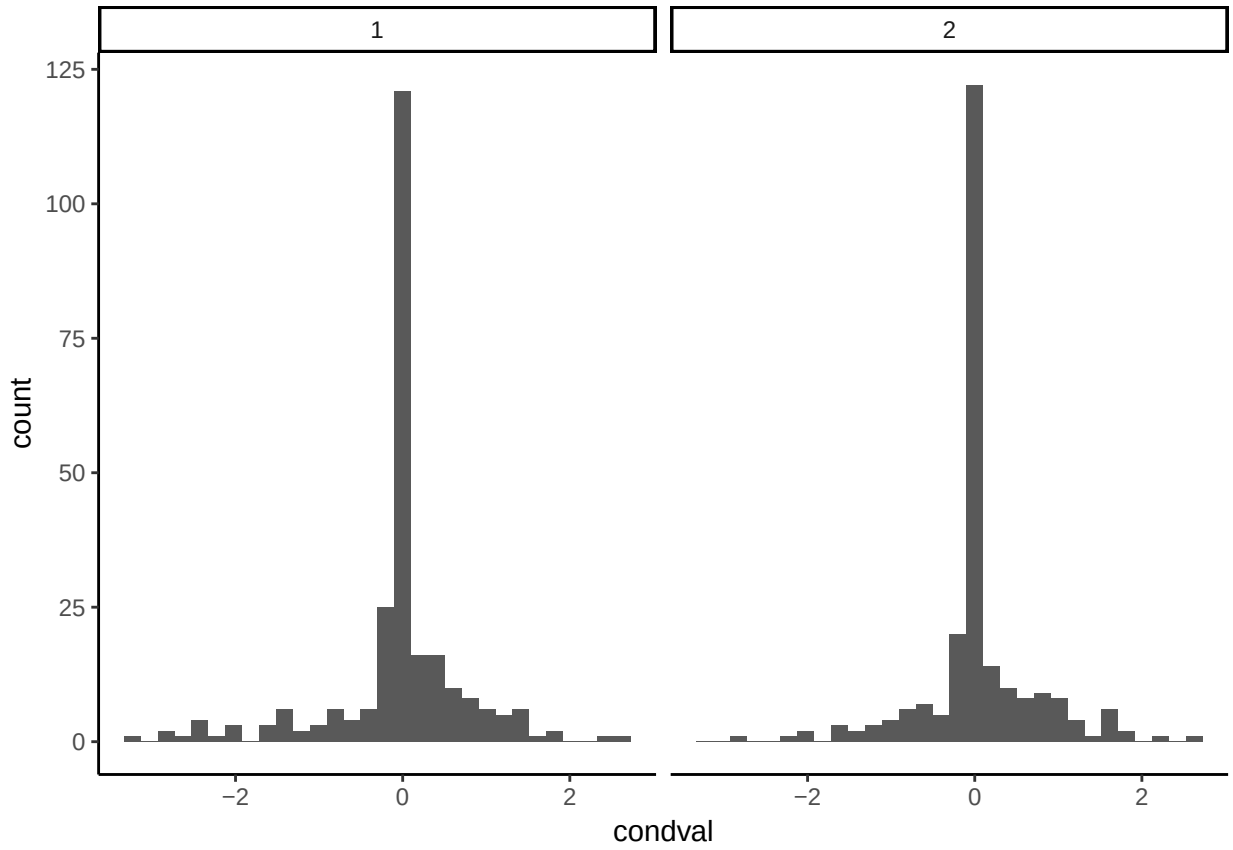
(c) Set up a model for the children's slopes and intercepts as a function of the treatment and age at baseline. Estimate this model using the two-step procedure. First estimate the intercept and slope separately for each child, then fit the between-child models using the point estimates from the first step.

```
#load packages
library(lme4)
library(dplyr)

#fit multi level model allowing intercept and slope to vary per child
cd4.mod <- lmer(data = cd4, CD4PCT.sqrt ~ (1 + baseage | newpid))


#extract random effects into a data frame
cd4.mod.ranef.df <- as.data.frame(ranef(cd4.mod))


#get unique treatment and patient id
trt <- as.data.frame(cbind(cd4$newpid, cd4$treatmnt))  %>%  unique()
colnames(trt) <- c("newpid","trtmt") #add column names
trt$newpid <- as.factor(trt$newpid) #convert to factor for join

#add treatment to existing random effects
cd4.mod.ranef.df.trt <- as.data.frame(inner_join(cd4.mod.ranef.df, trt, by = c("grp" = "newpid")))


#fit linear model for each treatment group based out estimate from MLM (summarizes coefficients from ML
trt.mod <- lm(data = cd4.mod.ranef.df.trt,  condval ~ as.factor(trtmt))
```

```r
#summary output
summary(trt.mod)
```

```
##
## Call:
## lm(formula = condval ~ as.factor(trtmt), data = cd4.mod.ranef.df.trt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2370 -0.1090  0.0126  0.1594  2.6149
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.03832    0.04635  -0.827    0.409
## as.factor(trtmt)2  0.08970    0.06690   1.341    0.181
##
## Residual standard error: 0.7473 on 498 degrees of freedom
## Multiple R-squared:  0.003598,   Adjusted R-squared:  0.001597
## F-statistic: 1.798 on 1 and 498 DF,  p-value: 0.1806
```

```r
#create plot for each group based on coefficients
ggplot(data = cd4.mod.ranef.df.trt, aes(x = condval, group = trtmt)) +
                  geom_histogram() +
                  theme_classic() +
                  facet_wrap (~ trtmt) #separate out by treatment
```



- **G&H 12.2, 12.3, 12.4:** Chapter 12, #2, #3, #4 (page 277)

G & H 12.2

Continuing with the analysis of the CD4 data from Exercise 11.4: (a) Write a model predicting CD4 percentage as a function of time with varying intercepts across children. Fit using lmer() and interpret the coefficient for time.

```
cd4.mod.time <- lmer(data = cd4, CD4PCT.sqrt ~ time + (1 | newpid))

summary(cd4.mod.time)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: CD4PCT.sqrt ~ time + (1 | newpid)
##    Data: cd4
##
## REML criterion at convergence: 3140.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.7379 -0.4379  0.0024  0.4324  5.0017
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  newpid   (Intercept) 1.9569   1.3989
##  Residual             0.5968   0.7725
## Number of obs: 1072, groups:  newpid, 250
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  4.76341    0.09648   49.37
## time        -0.36609    0.05399   -6.78
##
## Correlation of Fixed Effects:
##      (Intr)
## time -0.278
```

For the average child, for every 1 unit increase in time, we expect a -0.366 decrease in the square root of CD4.

(b) Extend the model in (a) to include child-level predictors (that is, group-level predictors) for treatment and age at baseline. Fit using lmer() and interpret the coefficients on time, treatment, and age at baseline.

```
#load package
library(arm)

#construct model. Allow time to vary within treatment group and age at baseline
cd4.mod.time.child <- lmer(data = cd4, CD4PCT.sqrt ~ time + (1 | newpid) + (1 | treatmnt) + (1 | baseag

#model output
display(cd4.mod.time.child)
```

```
## lmer(formula = CD4PCT.sqrt ~ time + (1 | newpid) + (1 | treatmnt) +
##     (1 | baseage), data = cd4)
##             coef.est coef.se
## (Intercept)  4.76     0.10
## time        -0.37     0.05
##
```

```
## Error terms:
##  Groups   Name        Std.Dev.
##  newpid   (Intercept) 1.24
##  baseage  (Intercept) 0.65
##  treatmnt (Intercept) 0.00
##  Residual             0.77
## ---
## number of obs: 1072, groups: newpid, 250; baseage, 239; treatmnt, 2
## AIC = 3152.6, DIC = 3126.7
## deviance = 3133.7
```

Interpreting coefficients * Time: For the average child, treatment, and baseage, for every 1 unit increase in time, we expect a -0.366 decrease in the square root of CD4.
* Treatment: The estimated variation across treatment groups is 0 (meaning there is little to no variation across treatment groups) * Age at Baseline: The estimated variation across age at the start of treatment is 0.65.

(c) Investigate the change in partial pooling from (a) to (b) both graphically and numerically.

```
#First model
ranef.1 <- ranef(cd4.mod.time) %>% unlist %>% as.numeric

#summary stats
mean(ranef.1, na.rm = T)
```

```
## [1] 4.246887e-14
```

```
sd(ranef.1)
```

```
## [1] 1.335525
```

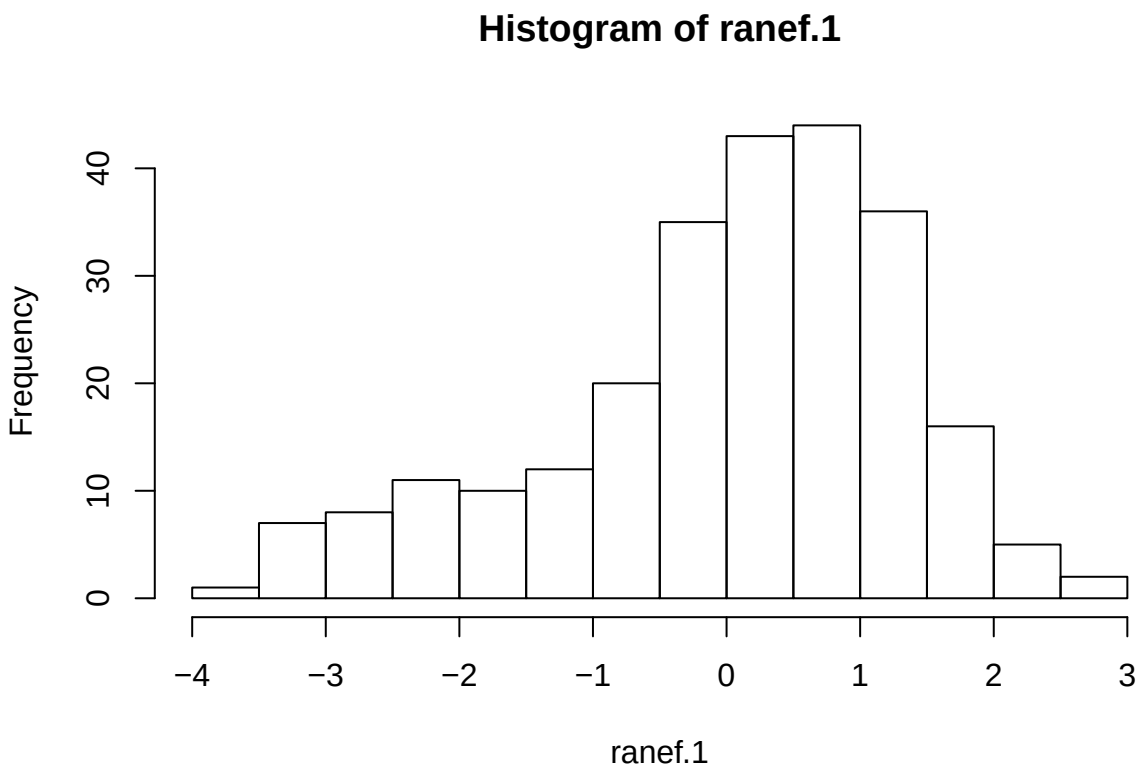
```
#Second model
ranef.2 <- ranef(cd4.mod.time.child) %>% unlist %>% as.numeric

#summary stats
mean(ranef.2)
```

```
## [1] -3.363384e-14
```

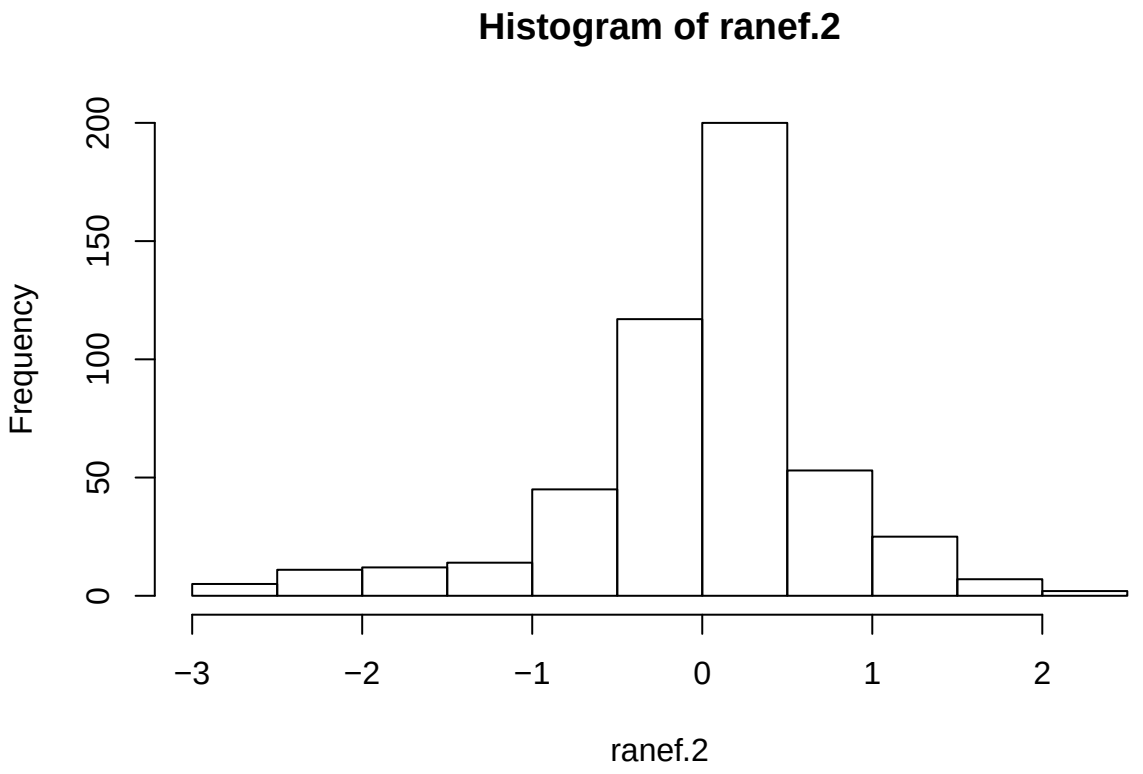
```
sd(ranef.2)
```

```
## [1] 0.7746706
```

```
#Compare graphically
#model 1
hist(ranef.1)
```




```
#model2
hist(ranef.2)
```



(d) Compare results in (b) to those obtained in part (c).

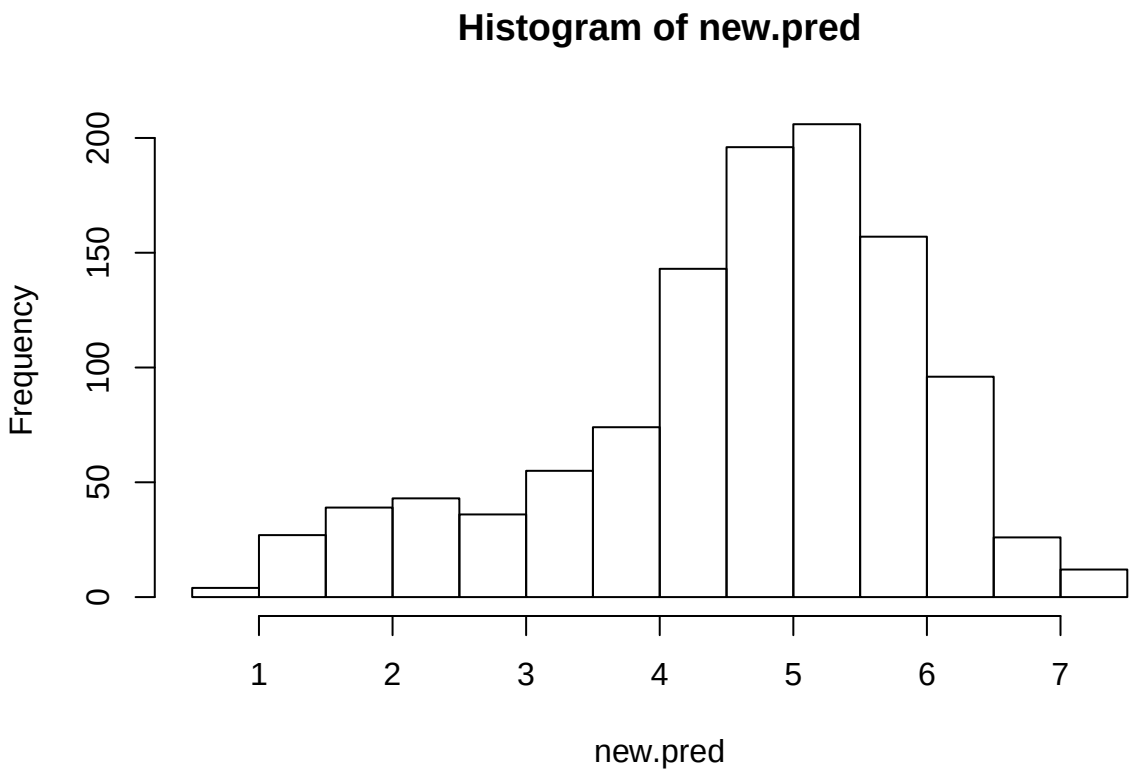
The model that did not include treatment and base age as random effects appears to have a larger standard

error than the model that *does* include these predictors as random effects. They are both centered between 0 and 1, but the second model has a tighter distribution.

G & H 12.3

Predictions for new observations and new groups: (a) Use the model fit from Exercise 12.2(b) to generate simulation of predicted CD4 percentages for each child in the data set at a hypothetical next time point.

```
#create a new data frame at a "next" time point
new.data <- subset(cd4, !is.na(treatmnt) & !is.na(baseage))
#pick an arbitrary date (same for all children)
new.data$VDATE <- as.Date("1999-01-01")

#let's feed the new data into the model
new.pred <- predict(cd4.mod.time.child, newdata = new.data)

#plot the new prediction
hist(new.pred)
```



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

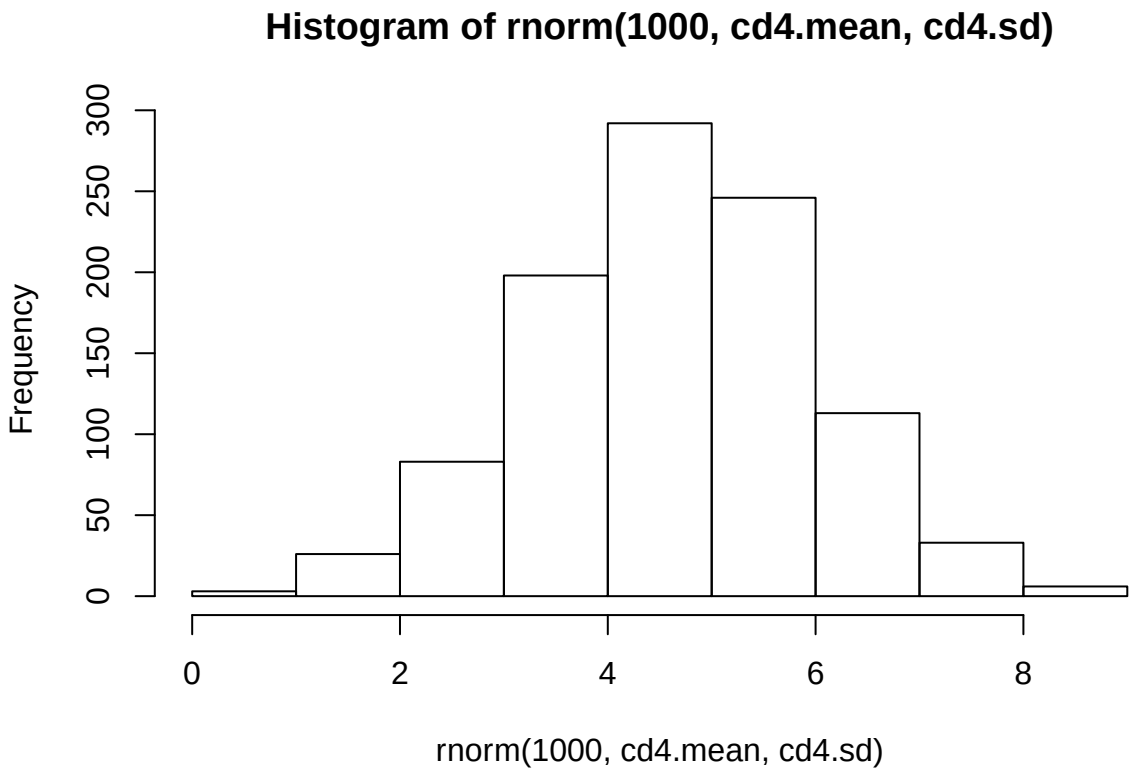
```
#create data for a child who is 4 years old at the baseline
new.data.4yr <- subset(cd4, baseage  == 4.0025000) #close to 4

#feed into model
new.pred.4yr <- predict(cd4.mod.time.child, newdata = new.data.4yr)

new.pred.4yr
```

```
##     1130     1131     1132     1133     1134     1135
## 5.309228 5.226057 5.142886 5.057886 4.981722 4.890630
```

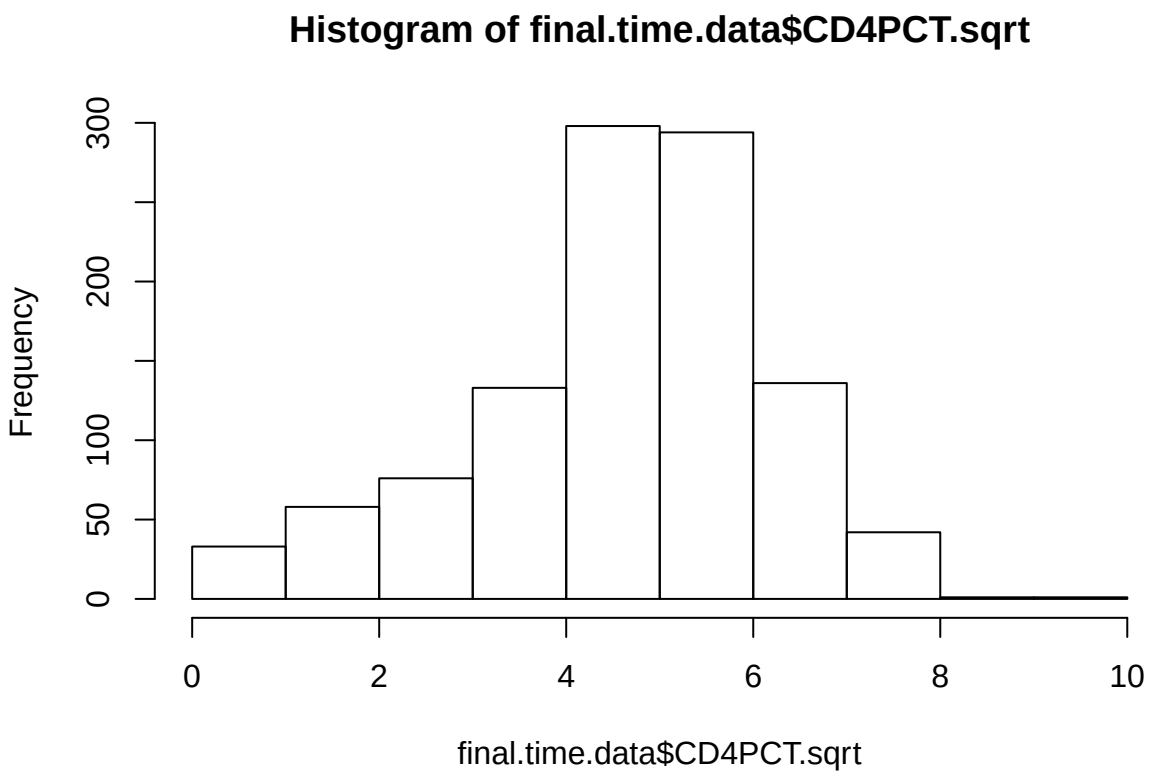
G & H 12.4

Posterior predictive checking: continuing the previous exercise, use the fitted model from Exercise 12.2(b) to simulate a new data set of CD4 percentages (with the same sample size and ages of the original data set) for the final time point of the study, and record the average CD4 percentage in this sample. Repeat this process 1000 times and compare the simulated distribution to the observed CD4 percentage at the final time point for the actual data.

```r
#create data
final.time.data <- subset(cd4, !is.na(treatmnt) & !is.na(baseage))

#force time to be last date "final" (and remove omissions)
final.time.data$VDATE <- max(as.Date(final.time.data$VDATE,format="%m-%d-%Y"), na.rm = T)

#calculate mean and sd from new data
cd4.mean <- mean(predict(cd4.mod.time.child, newdata = final.time.data), na.rm = T) #mean
cd4.sd <- sd(predict(cd4.mod.time.child, newdata = final.time.data), na.rm = T) #standard deviation

#simulate 1000 times
set.seed(100)
hist(rnorm(1000, cd4.mean, cd4.sd))
```




```r
#Compare distribution at final time point for actual data
hist(final.time.data$CD4PCT.sqrt)
```



The predicted values appear to be less closely distributed around their mean than the observed values. The mean of the predicted values also appear to be slightly lower than the observed values.

- **G&H 13.5 (a):** Chapter 13, #5 part (a) (page 299)

G & H 13.5 a

Extend the model in Exercise 12.2 to allow for varying slopes for the time predictor.

```
cd4.mod.time.var <- lmer(data = cd4, CD4PCT.sqrt ~ time + (1 + time | newpid))

display(cd4.mod.time.var)
```

```
## lmer(formula = CD4PCT.sqrt ~ time + (1 + time | newpid), data = cd4)
##             coef.est coef.se
## (Intercept)  4.76     0.09
## time        -0.36     0.07
##
## Error terms:
##  Groups   Name        Std.Dev. Corr
##  newpid   (Intercept) 1.39
##           time        0.58     -0.05
##  Residual             0.72
## ---
## number of obs: 1072, groups: newpid, 250
## AIC = 3123.2, DIC = 3098.2
## deviance = 3104.7
```