homework 07

Shiyu Zhang

November 1, 2018

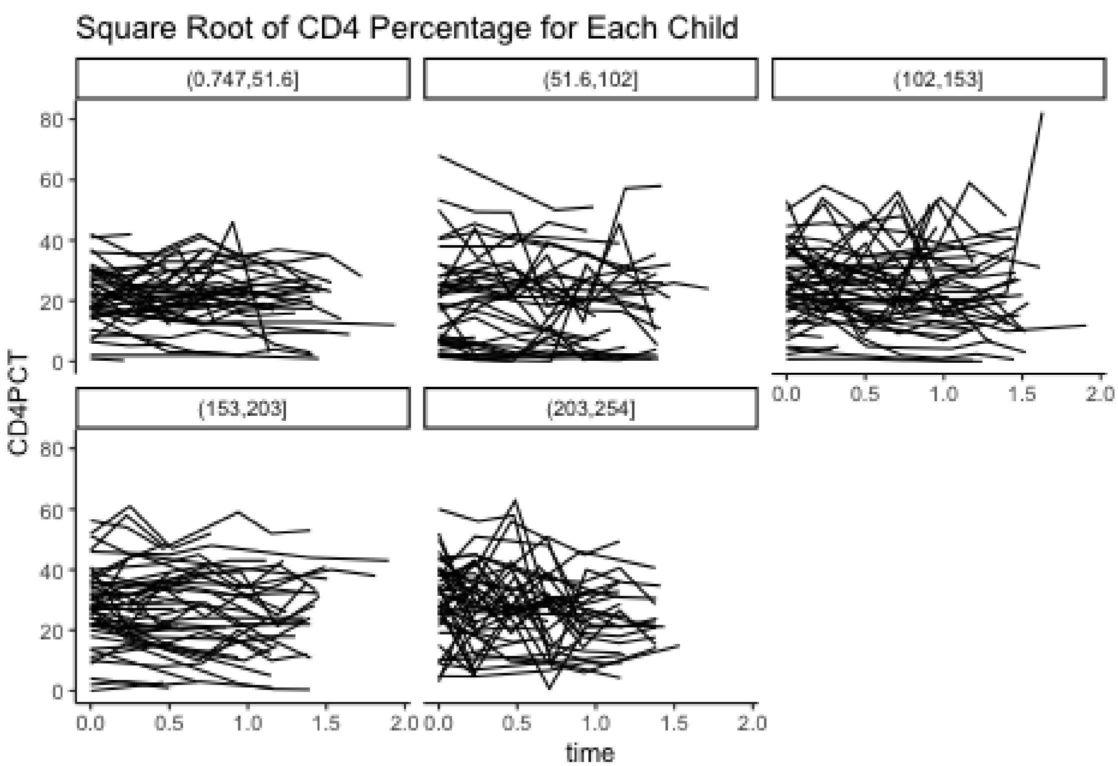
Data analysis

CD4 percentages for HIV infected kids

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 dataset also includes the ages of the children at each measurement.

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

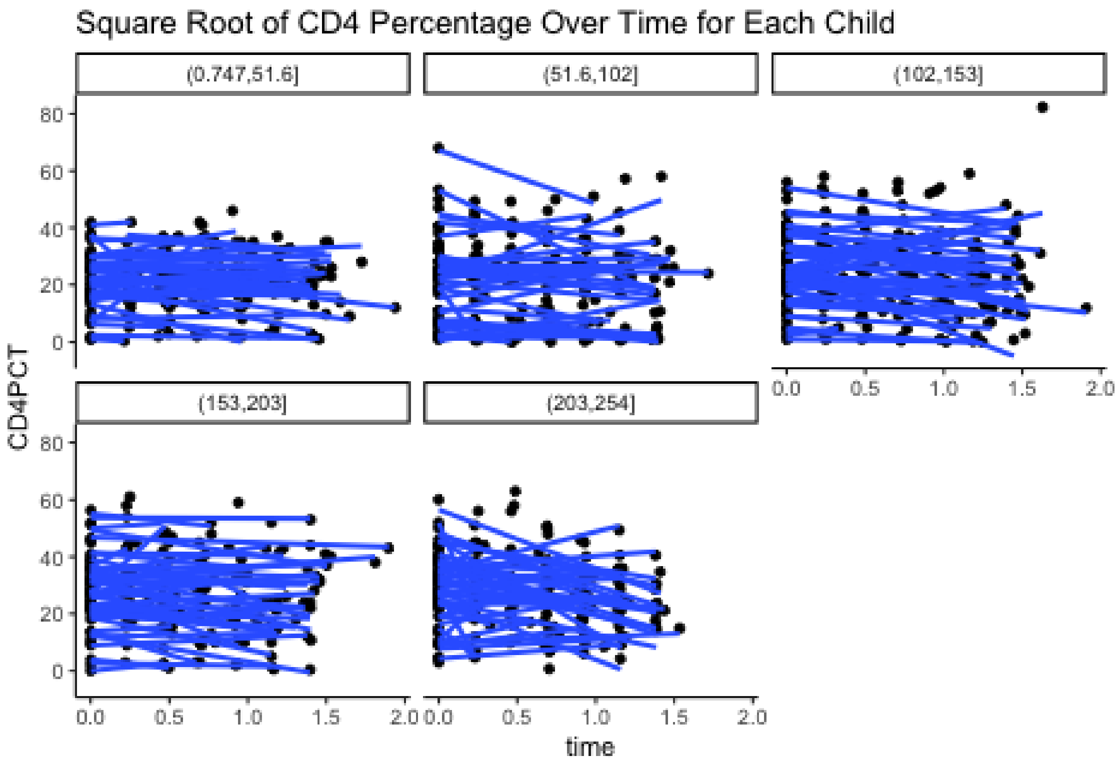
```
#cut newpid into groups for wrapping
hiv.data$newpid.group <- cut(hiv.data$newpid, breaks = 5)

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



2. 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.

```
ggplot(data = hiv.data, aes(x = time, y = CD4PCT, group = newpid), na.rm = T) +
      geom_point() +
```

```r
    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")
```



3. 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.

```r
model_child<- matrix(0,nrow=254,ncol = 3)
colnames(model_child) <- c("newpid","intercept","slope")
for (i in unique(hiv.data$newpid)){
  model_lm <- lm(y ~ time, hiv.data[newpid == i,c("y","time")])
  model_child[i,1] <- i
  model_child[i,2] <- coef(model_lm)[1]
  model_child[i,3] <- coef(model_lm)[2]
}
hiv.data.new <- hiv.data[,list(age.baseline=unique(age.baseline),treatment=unique(treatment)), by=newpid
hiv.data.new <- merge(model_child,hiv.data.new,by="newpid")
lm(intercept~ age.baseline+factor(treatment),data = hiv.data.new)
```

```
##
## Call:
## lm(formula = intercept ~ age.baseline + factor(treatment), data = hiv.data.new)
##
## Coefficients:
##       (Intercept)       age.baseline  factor(treatment)2
##            5.1179            -0.1210              0.1236
```

```r
lm(slope~ age.baseline+factor(treatment),data=hiv.data.new)
```

```
##
## Call:
## lm(formula = slope ~ age.baseline + factor(treatment), data = hiv.data.new)
##
## Coefficients:
##       (Intercept)        age.baseline  factor(treatment)2
##          -0.26568            -0.04223            -0.13926
```

4. 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.

```r
model <- lmer(data = hiv.data, CD4PCT ~ time + (1 | newpid))
summary(model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: CD4PCT ~ time + (1 | newpid)
##    Data: hiv.data
##
## REML criterion at convergence: 7879.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.4823 -0.4582 -0.0581  0.3795  6.8099
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  newpid   (Intercept) 129.36   11.374
##  Residual              53.25    7.297
## Number of obs: 1072, groups:  newpid, 250
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  25.0394     0.8058  31.076
## time         -3.0019     0.5083  -5.905
##
## Correlation of Fixed Effects:
##      (Intr)
## time -0.316
```

5. Extend the model in (4) 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.

```r
model.child <- lmer(y~time+factor(treatment)+age.baseline+(1|newpid), data = hiv.data)
display(model.child)
```

```
## lmer(formula = y ~ time + factor(treatment) + age.baseline +
##     (1 | newpid), data = hiv.data)
##                    coef.est coef.se
## (Intercept)         5.09     0.19
## time               -0.36     0.05
## factor(treatment)2  0.18     0.18
## age.baseline       -0.12     0.04
##
```

```
## Error terms:
##  Groups   Name        Std.Dev.
##  newpid   (Intercept) 1.37
##  Residual             0.77
## ---
## number of obs: 1072, groups: newpid, 250
## AIC = 3149.2, DIC = 3110.9
## deviance = 3124.1
```

```
"Time: For the average child, treatment, and baseage, for every 1 unit increase in time,
we expect a 0.3 decrease in the square root of CD4.

Treatment: The estimated variation for treatment is 0

Age: The estimated variation across age is 5.30. "
```

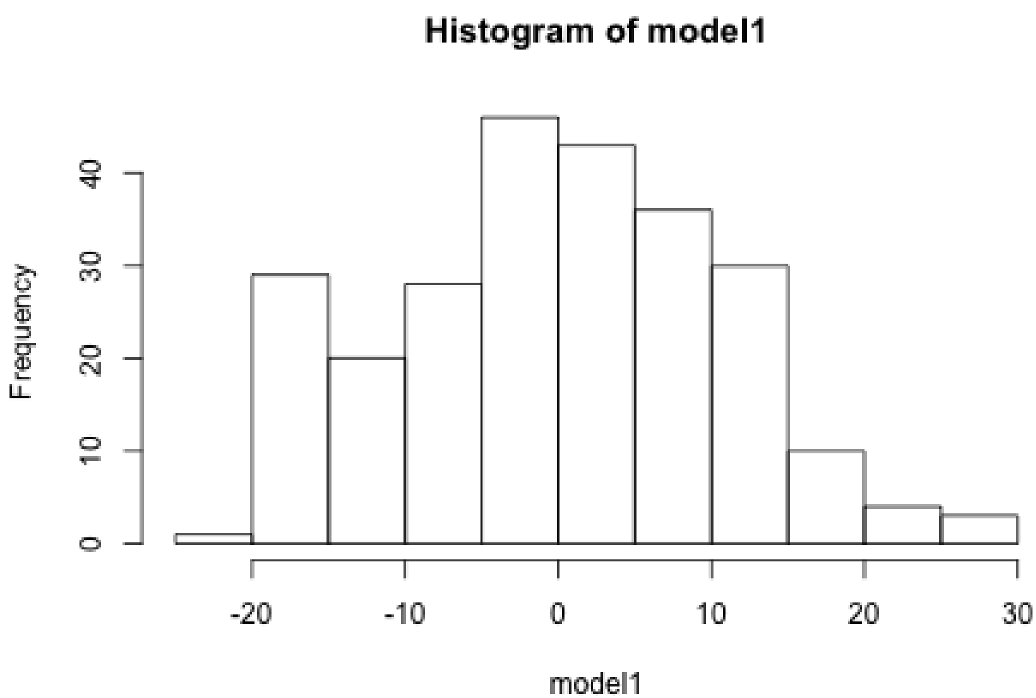
```
## [1] "Time: For the average child, treatment, and baseage, for every 1 unit increase in time, \nwe exp
```

6. Investigate the change in partial pooling from (4) to (5) both graphically and numerically.

```
model1 <- ranef(model) %>% unlist %>% as.numeric
summary(model1)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -21.07627  -7.16531   0.05595   0.00000   7.07075  29.03885
```

```
hist(model1)
```

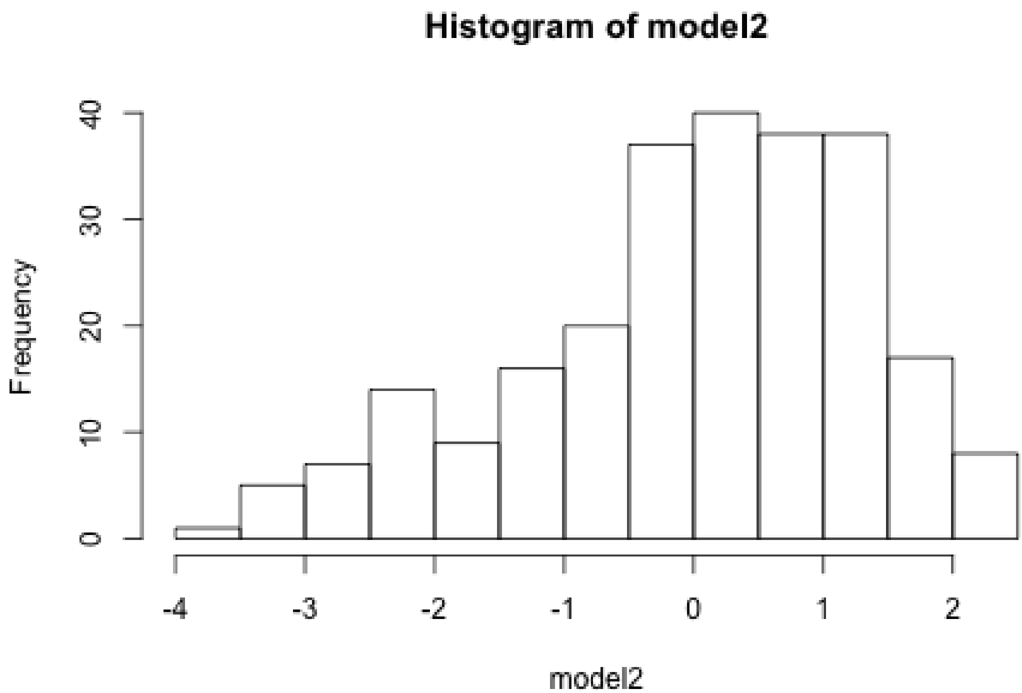



```
model2 <- ranef(model.child) %>% unlist %>% as.numeric
summary(model2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
## -3.5801 -0.6751  0.1820  0.0000  0.9956  2.4470
```

```r
hist(model2)
```



7. Use the model fit from (5) to generate simulation of predicted CD4 percentages for each child in the dataset at a hypothetical next time point.

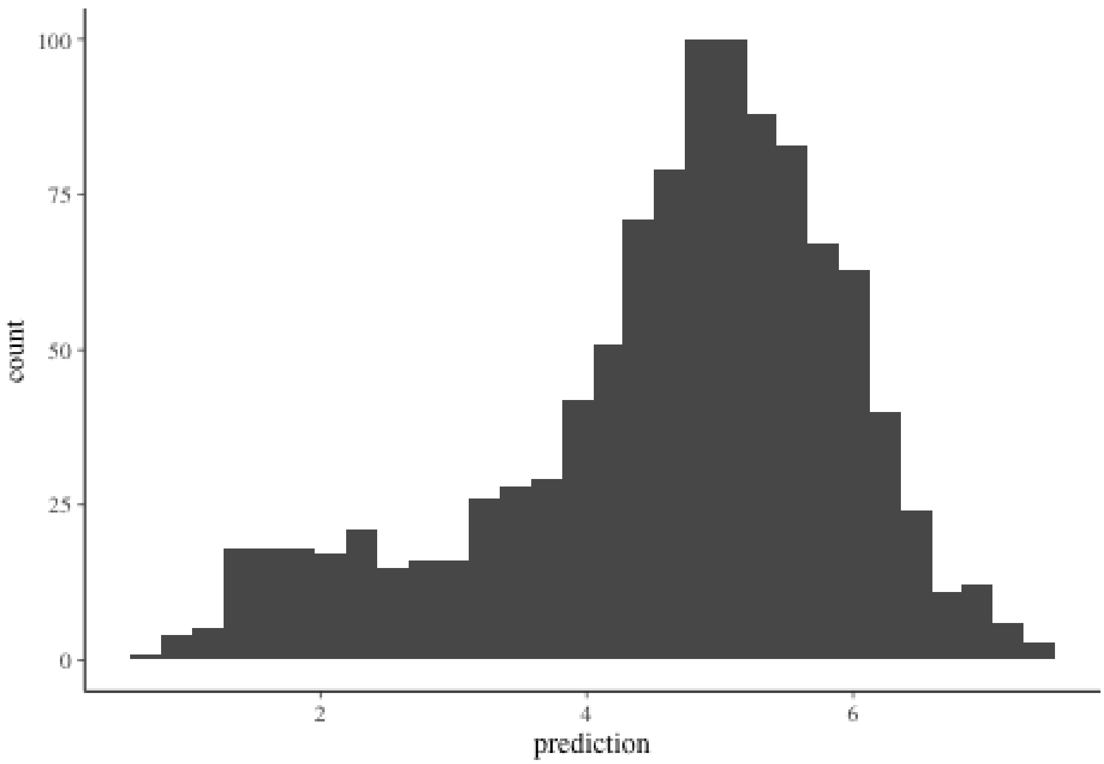
```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
```

```
##
##     intersect, setdiff, setequal, union
```

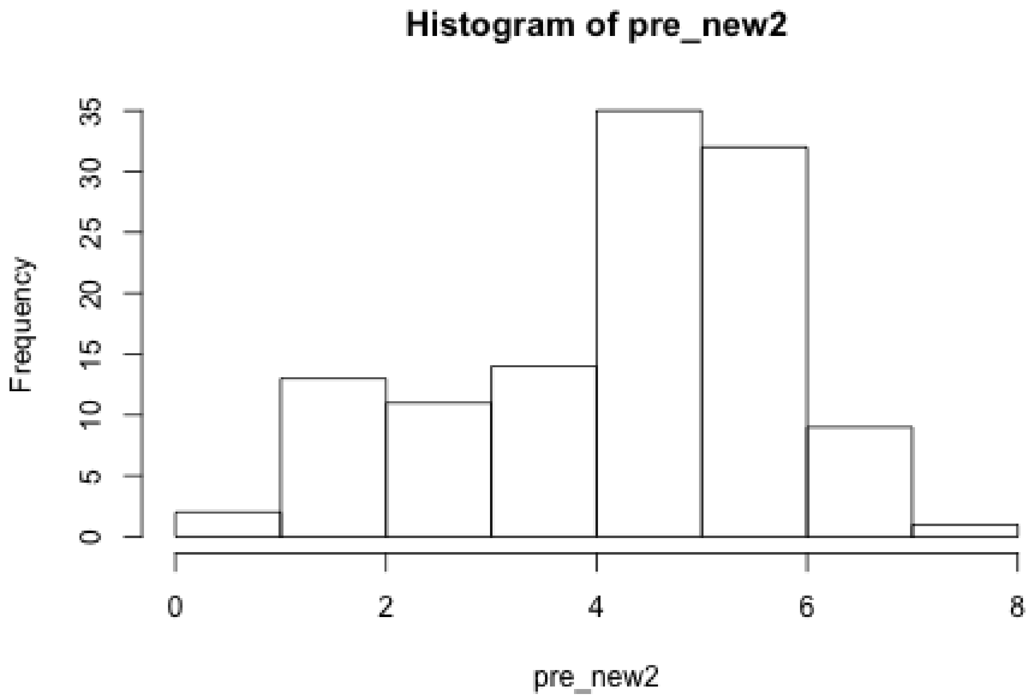
```
pre_data <- subset(hiv.data, !is.na(hiv.data$treatment) & !is.na(age.baseline))
pre_new <- predict(model.child,newdata=pre_data)
pre_com <- cbind(pre_new,pre_data)
colnames(pre_com)[1] <- c("prediction")
ggplot(pre_com,aes(x=prediction))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



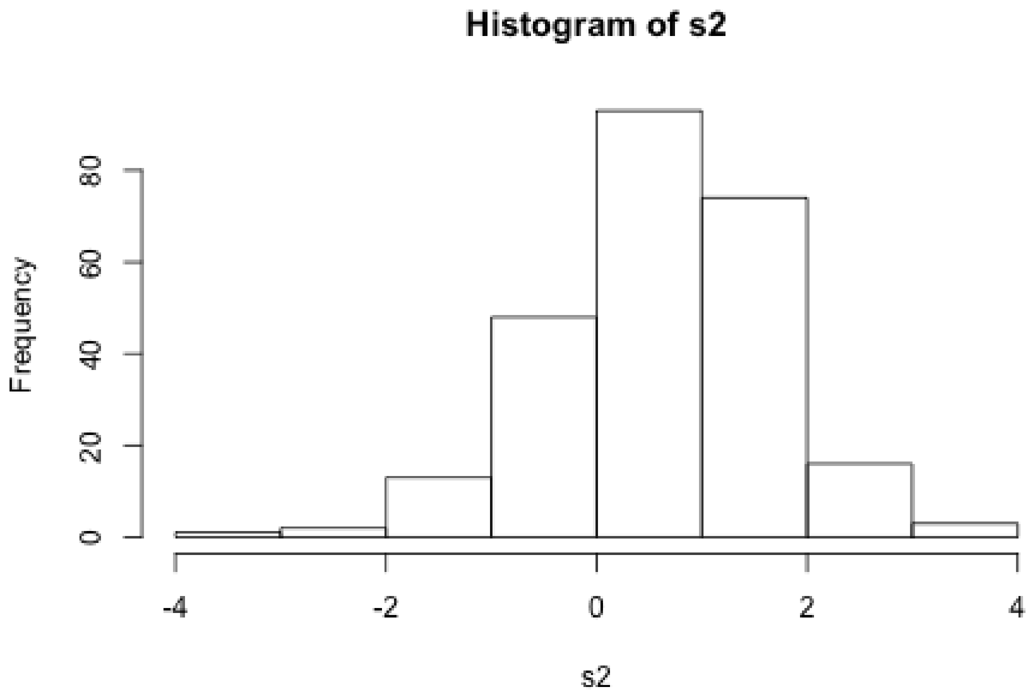
8. Use the same model fit to generate simulations of CD4 percentages at each of the time periods for a new child who was 4 years old at baseline.

```
pre_data2 <- subset(hiv.data, !is.na(hiv.data$treatment) & !is.na(age.baseline))
pre_data2 <- pre_data2[, -c(1, 4, 5, 6, 8)]
pre_data2 <- pre_data2[which(round(pre_data2$age.baseline) == 4 ),]
pre_new2 <- predict(model.child, newdata=pre_data2)
hist(pre_new2)
```



9. Posterior predictive checking: continuing the previous exercise, use the fitted model from (5) to simulate a new dataset of CD4 percentages (with the same sample size and ages of the original dataset) 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.

```
pre_new3 <- hiv.data[,list(time=max(time),age.baseline=unique(age.baseline),
                       treatment=unique(treatment)),by =newpid]
cm3<-coef(model.child)$newpid
s1<-sigma.hat(model.child)$sigma$data
p1<-cm3[,1]+cm3[,2]*pre_new3$time+cm3[,3]*pre_new3$age.baseline+cm3[,4]*(pre_new3$treatment-1)
avg<-NULL
s2<-matrix(NA,nrow(pre_new3),1000)
for (i in 1:1000){
  yti<-rnorm(p1,s1)
  s2[,1]<-yti
}
hist(s2)
```



10. Extend the modelto allow for varying slopes for the time predictor.

```
hiv_slope<- lmer(y~time+factor(treatment)+age.baseline+(1+time|newpid), data = hiv.data)
summary(hiv_slope)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ time + factor(treatment) + age.baseline + (1 + time | newpid)
##    Data: hiv.data
##
## REML criterion at convergence: 3107
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.0998 -0.4057  0.0174  0.4030  5.0157
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  newpid   (Intercept) 1.8464   1.3588
##           time        0.3374   0.5808   -0.04
##  Residual             0.5145   0.7173
## Number of obs: 1072, groups:  newpid, 250
##
## Fixed effects:
##                      Estimate Std. Error t value
## (Intercept)           5.10850    0.18594  27.474
## time                 -0.35258    0.06763  -5.214
## factor(treatment)2    0.15952    0.18137   0.880
## age.baseline         -0.12423    0.03971  -3.128
##
## Correlation of Fixed Effects:
```

```
##              (Intr) time   fct()2
## time         -0.114
## fctr(trtm)2 -0.463  0.010
## age.baselin -0.729 -0.013 -0.004
```

11. Next fit a model that does not allow for varying slopes but does allow for different coefficients for each time point (rather than fitting the linear trend).

```r
hiv_reg <- lmer(y~factor(time)+(1|newpid), data = hiv.data)
```

12. Compare the results of these models both numerically and graphically.

```r
anova(hiv_reg,hiv_slope,model.child,model)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: hiv.data
## Models:
## model: CD4PCT ~ time + (1 | newpid)
## model.child: y ~ time + factor(treatment) + age.baseline + (1 | newpid)
## hiv_slope: y ~ time + factor(treatment) + age.baseline + (1 + time | newpid)
## hiv_reg: y ~ factor(time) + (1 | newpid)
##              Df    AIC    BIC  logLik deviance    Chisq Chi Df Pr(>Chisq)
## model         4 7889.0 7908.9 -3940.5   7881.0
## model.child   6 3136.1 3165.9 -1562.0   3124.1 4756.957      2  < 2.2e-16
## hiv_slope     8 3110.3 3150.1 -1547.1   3094.3   29.789      2  3.399e-07
## hiv_reg     405 3244.5 5260.3 -1217.3   2434.5  659.753    397  2.261e-15
##
## model
## model.child ***
## hiv_slope   ***
## hiv_reg     ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure skate in the 1932 Winter Olympics

The folder olympics has seven judges' ratings of seven figure skaters (on two criteria: "technical merit" and "artistic impression") from the 1932 Winter Olympics. Take a look at http://www.stat.columbia.edu/~gelman/arm/examples/olympics/olympics1932.txt

1. Construct a $7 \times 7 \times 2$ array of the data (ordered by skater, judge, and judging criterion).

```r
library(reshape)
```

```
##
## Attaching package: 'reshape'
```

```
## The following object is masked from 'package:dplyr':
##
##     rename
```

```
## The following object is masked from 'package:data.table':
##
##     melt
```

```
## The following object is masked from 'package:Matrix':
##
```

```
##     expand
```

```
arry1<-melt(data = olympics1932,id.vars=c("pair","criterion"),
              measure.vars=c(colnames(olympics1932)[3:9]))
arry1
```

```
##    pair   criterion variable value
## 1     1     Program  judge_1   5.6
## 2     1 Performance  judge_1   5.6
## 3     2     Program  judge_1   5.5
## 4     2 Performance  judge_1   5.5
## 5     3     Program  judge_1   6.0
## 6     3 Performance  judge_1   6.0
## 7     4     Program  judge_1   5.6
## 8     4 Performance  judge_1   5.6
## 9     5     Program  judge_1   5.4
## 10    5 Performance  judge_1   4.8
## 11    6     Program  judge_1   5.2
## 12    6 Performance  judge_1   4.8
## 13    7     Program  judge_1   4.8
## 14    7 Performance  judge_1   4.3
## 15    1     Program  judge_2   5.5
## 16    1 Performance  judge_2   5.5
## 17    2     Program  judge_2   5.2
## 18    2 Performance  judge_2   5.7
## 19    3     Program  judge_2   5.3
## 20    3 Performance  judge_2   5.5
## 21    4     Program  judge_2   5.3
## 22    4 Performance  judge_2   5.3
## 23    5     Program  judge_2   4.5
## 24    5 Performance  judge_2   4.8
## 25    6     Program  judge_2   5.1
## 26    6 Performance  judge_2   5.6
## 27    7     Program  judge_2   4.0
## 28    7 Performance  judge_2   4.6
## 29    1     Program  judge_3   5.8
## 30    1 Performance  judge_3   5.8
## 31    2     Program  judge_3   5.8
## 32    2 Performance  judge_3   5.6
## 33    3     Program  judge_3   5.8
## 34    3 Performance  judge_3   5.7
## 35    4     Program  judge_3   5.8
## 36    4 Performance  judge_3   5.8
## 37    5     Program  judge_3   5.8
## 38    5 Performance  judge_3   5.5
## 39    6     Program  judge_3   5.3
## 40    6 Performance  judge_3   5.0
## 41    7     Program  judge_3   4.7
## 42    7 Performance  judge_3   4.5
## 43    1     Program  judge_4   5.3
## 44    1 Performance  judge_4   4.7
## 45    2     Program  judge_4   5.8
## 46    2 Performance  judge_4   5.4
## 47    3     Program  judge_4   5.0
## 48    3 Performance  judge_4   4.9
```

```
## 49    4     Program  judge_4   4.4
## 50    4 Performance  judge_4   4.8
## 51    5     Program  judge_4   4.0
## 52    5 Performance  judge_4   4.4
## 53    6     Program  judge_4   5.4
## 54    6 Performance  judge_4   4.7
## 55    7     Program  judge_4   4.0
## 56    7 Performance  judge_4   4.0
## 57    1     Program  judge_5   5.6
## 58    1 Performance  judge_5   5.7
## 59    2     Program  judge_5   5.6
## 60    2 Performance  judge_5   5.5
## 61    3     Program  judge_5   5.4
## 62    3 Performance  judge_5   5.5
## 63    4     Program  judge_5   4.5
## 64    4 Performance  judge_5   4.5
## 65    5     Program  judge_5   5.5
## 66    5 Performance  judge_5   4.6
## 67    6     Program  judge_5   4.5
## 68    6 Performance  judge_5   4.0
## 69    7     Program  judge_5   3.7
## 70    7 Performance  judge_5   3.6
## 71    1     Program  judge_6   5.2
## 72    1 Performance  judge_6   5.3
## 73    2     Program  judge_6   5.1
## 74    2 Performance  judge_6   5.3
## 75    3     Program  judge_6   5.1
## 76    3 Performance  judge_6   5.2
## 77    4     Program  judge_6   5.0
## 78    4 Performance  judge_6   5.0
## 79    5     Program  judge_6   4.8
## 80    5 Performance  judge_6   4.8
## 81    6     Program  judge_6   4.5
## 82    6 Performance  judge_6   4.6
## 83    7     Program  judge_6   4.0
## 84    7 Performance  judge_6   4.0
## 85    1     Program  judge_7   5.7
## 86    1 Performance  judge_7   5.4
## 87    2     Program  judge_7   5.8
## 88    2 Performance  judge_7   5.7
## 89    3     Program  judge_7   5.3
## 90    3 Performance  judge_7   5.7
## 91    4     Program  judge_7   5.1
## 92    4 Performance  judge_7   5.5
## 93    5     Program  judge_7   5.5
## 94    5 Performance  judge_7   5.2
## 95    6     Program  judge_7   5.0
## 96    6 Performance  judge_7   5.2
## 97    7     Program  judge_7   4.8
## 98    7 Performance  judge_7   4.8
```

2. Reformulate the data as a 98×4 array (similar to the top table in Figure 11.7), where the first two columns are the technical merit and artistic impression scores, the third column is a skater ID, and the fourth column is a judge ID.

```r
arry2 <- rename(arry1, c("pair"="skater_ID", "variable"="judge_ID"))
arry2 <- arry2[order(arry2$judge_ID),]
arry2 <- arry2[c("criterion", "value", "skater_ID", "judge_ID")]
```

3. Add another column to this matrix representing an indicator variable that equals 1 if the skater and judge are from the same country, or 0 otherwise.

```r
arry2$SameCountry <-ifelse(arry2[,3] == " 1"&arry2[,4] == "judge_5",1,
  ifelse(arry2[,3] == " 2"&arry2[,4] == "judge_7",1,
  ifelse(arry2[,3] == " 3"&arry2[,4] == "judge_1",1,
  ifelse(arry2[,3] == " 4"&arry2[,4] == "judge_1",1,
  ifelse(arry2[,3] == " 7"&arry2[,4] == "judge_7",1,0
  )))))
```

4. Write the notation for a non-nested multilevel model (varying across skaters and judges) for the technical merit ratings and fit using lmer().

```r
data3 <- arry2 %>%
  filter(criterion=="Program")
data4 <- arry2 %>%
  filter(criterion=="Performance")
reg <- lmer(value ~ 1 + (1|skater_ID) + (1|judge_ID),data=data3)
display(reg)
```

```
## lmer(formula = value ~ 1 + (1 | skater_ID) + (1 | judge_ID),
##     data = data3)
## coef.est  coef.se
##     5.13     0.20
##
## Error terms:
##  Groups    Name        Std.Dev.
##  skater_ID (Intercept) 0.42
##  judge_ID  (Intercept) 0.28
##  Residual              0.33
## ---
## number of obs: 49, groups: skater_ID, 7; judge_ID, 7
## AIC = 68, DIC = 57
## deviance = 58.5
```

5. Fit the model in (4) using the artistic impression ratings.

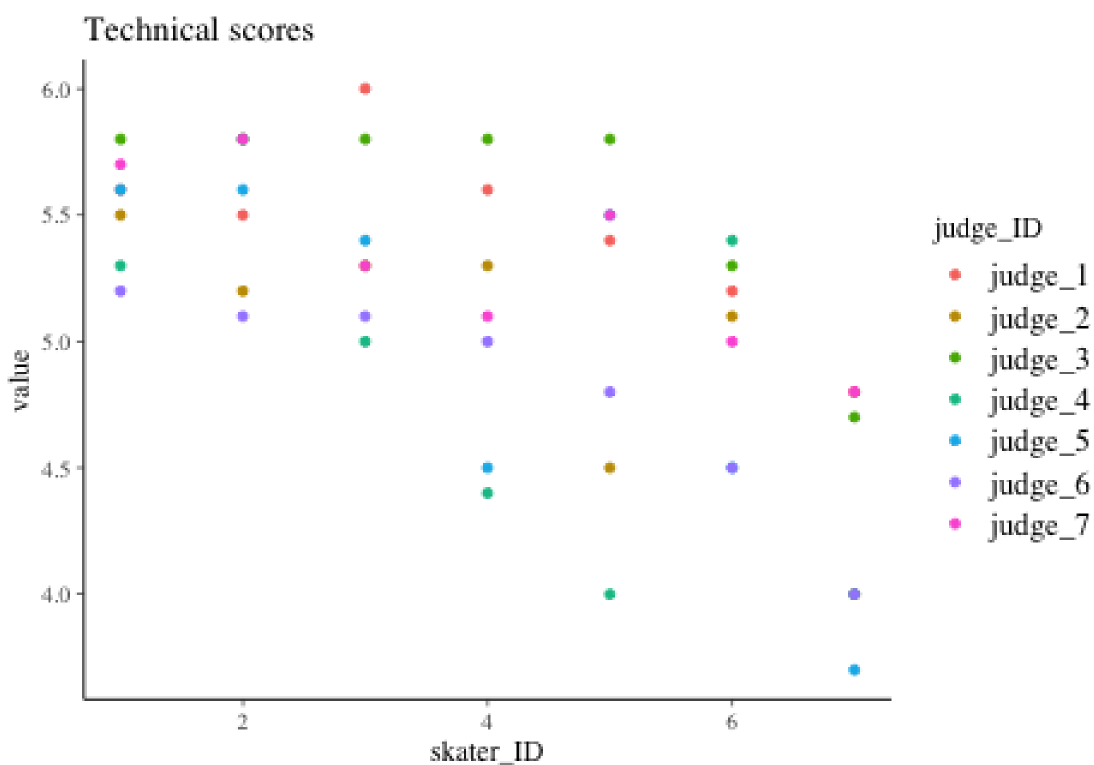
```r
reg2 <- lmer(value ~ 1 + (1|skater_ID) + (1|judge_ID),data=data4)
display(reg2)
```

```
## lmer(formula = value ~ 1 + (1 | skater_ID) + (1 | judge_ID),
##     data = data4)
## coef.est  coef.se
##     5.09     0.20
##
## Error terms:
##  Groups    Name        Std.Dev.
##  skater_ID (Intercept) 0.45
##  judge_ID  (Intercept) 0.28
##  Residual              0.27
## ---
## number of obs: 49, groups: skater_ID, 7; judge_ID, 7
```

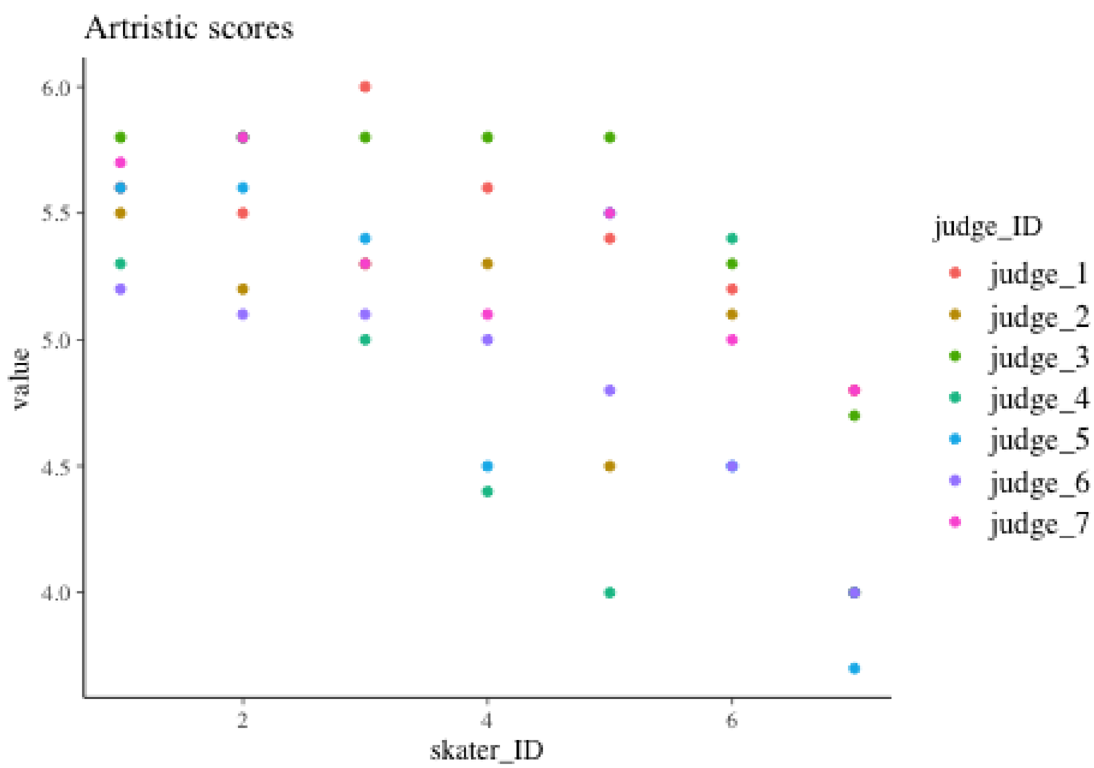
```
## AIC = 54.2, DIC = 43.4
## deviance = 44.8
```

6. Display your results for both outcomes graphically.

```
ggplot(data3,aes(x=skater_ID,y=value,color=judge_ID))+geom_point()+
  ggtitle("Technical scores")
```

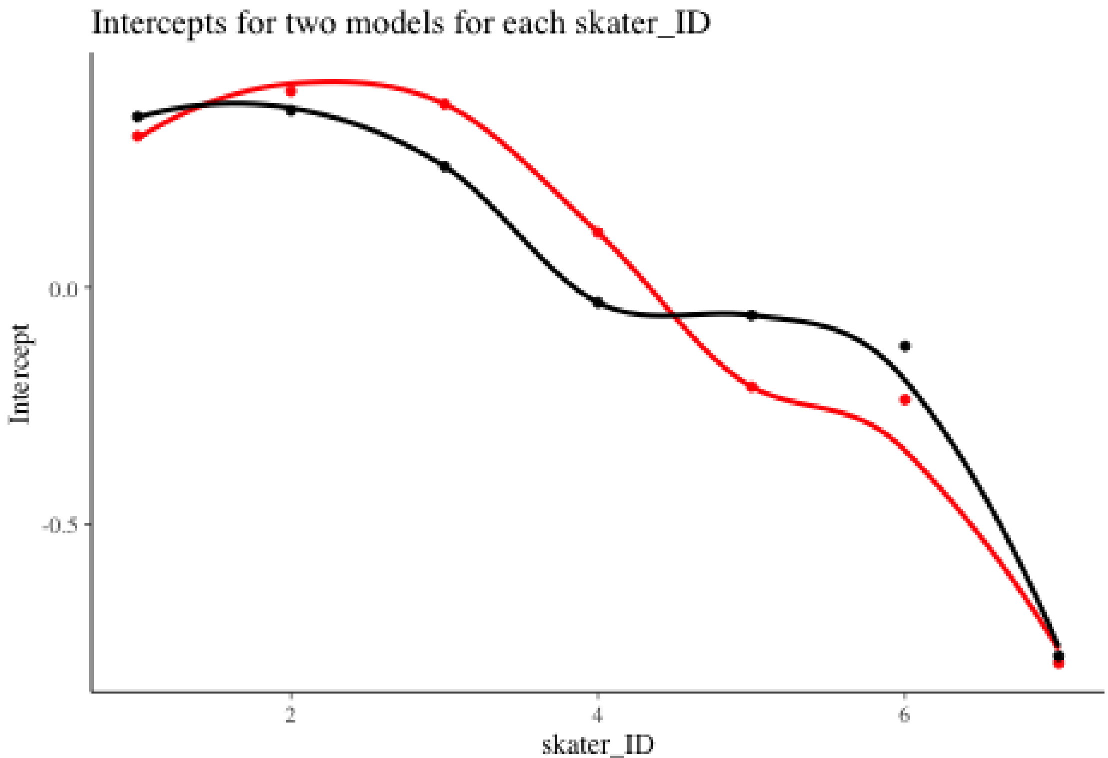



```
ggplot(data3,aes(x=skater_ID,y=value,color=judge_ID))+geom_point()+
  ggtitle("Artristic scores")
```



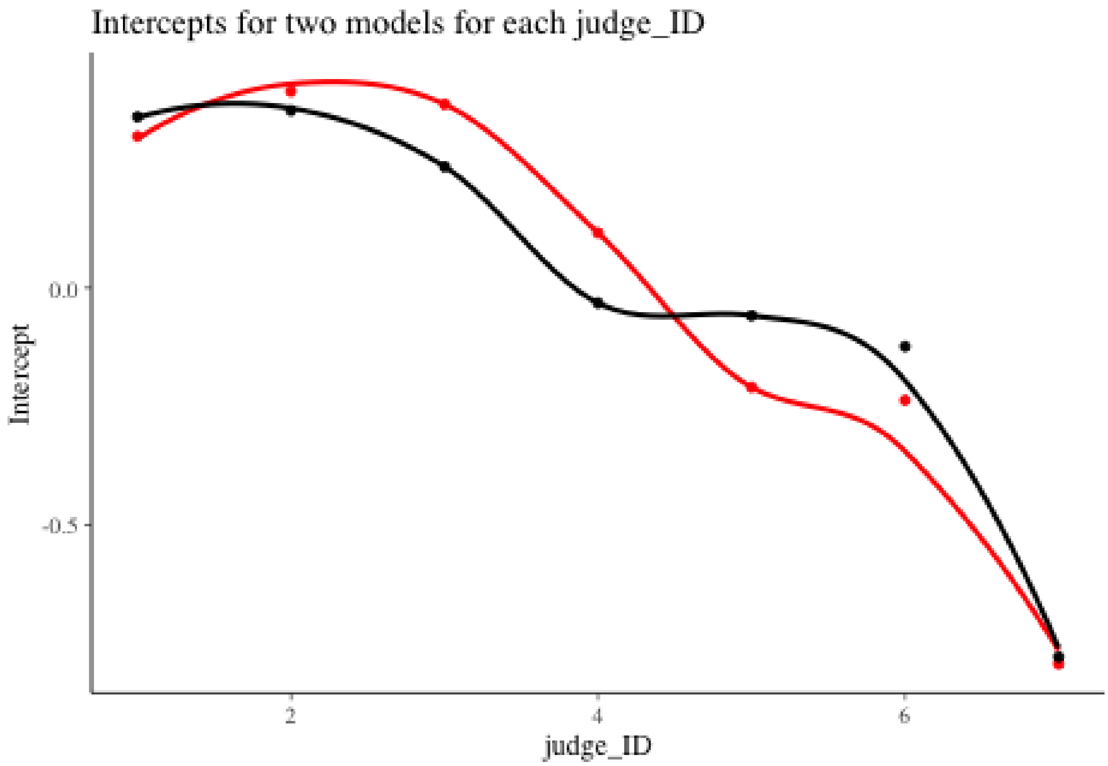

```
skate <- as.data.frame(cbind(unlist(ranef(reg2))[1:7],unlist(ranef(reg))[1:7]))
skate$skater_ID <-c(1:7)
ggplot(data=skate)+
  geom_point(col="red",aes(x=skater_ID,y=V1))+geom_smooth(col="red",aes(x=skater_ID,y=V1),se=FALSE)+
  geom_point(col="black",aes(x=skater_ID,y=V2))+geom_smooth(col="black",aes(x=skater_ID,y=V2),se=FALSE)+
  ggtitle("Intercepts for two models for each skater_ID")+
  ylab("Intercept")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```r
# the same method as the previous one
judge <- as.data.frame(cbind(unlist(ranef(reg2))[1:7],unlist(ranef(reg))[1:7]))
judge$judge_ID <-c(1:7)
ggplot(data=judge)+
  geom_point(col="red",aes(x=judge_ID,y=V1))+geom_smooth(col="red",aes(x=judge_ID,y=V1),se=FALSE)+
  geom_point(col="black",aes(x=judge_ID,y=V2))+geom_smooth(col="black",aes(x=judge_ID,y=V2),se=FALSE)+
  ggtitle("Intercepts for two models for each judge_ID")+
  ylab("Intercept")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



7. (optional) Use posterior predictive checks to investigate model fit in (4) and (5).

Different ways to write the model:

Using any data that are appropriate for a multilevel model, write the model in the five ways discussed in Section 12.5 of Gelman and Hill.

The fixed effects part of the model: $y = \alpha_{j[i]} + \beta_{time} X_{itime} + \beta_{treatment} X_{itreatment} + \beta_{age.base} X_{iage.base} + \epsilon_i$

1:

$y = 4.91 + X_{itime} * (-0.36) + X_{itreatment} * (-0.12) + X_{iage.base} * 0.18 + 0.77 \; \alpha_j \sim \; \mathrm{N}(0, 1.37^2)$

2:

$y \sim \; N(4.91 + X_{itime} * (-0.36) + X_{itreatment} * (-0.12) + X_{iage.base} * (0.18), 0.77^2)$

$\alpha_j \sim \; \mathrm{N}(RandomIntercept, 1.37^2)$

3:

$y_i \sim \; N(4.91 + X_{itime} * (-0.36) + X_{itreatment} * (-0.12) + X_{iage.base} * (0.18), 0.77^2) \; \beta_j \sim \; N(0, 1.37^2)$

4:

$$y_i \sim \ N(4.91 + X_{itime} * (-0.36) + X_{itreatment} * (-0.12) + X_{iage.base} * (0.18) + 1.37^2, 0.77^2)$$

5:

$$y_i \sim \ N(4.91 + X_{itime} * (-0.36) + X_{itreatment} * (-0.12) + X_{iage.base} * (0.18), 1.37^2 + 0.77^2)$$

Models for adjusting individual ratings:

A committee of 10 persons is evaluating 100 job applications. Each person on the committee reads 30 applications (structured so that each application is read by three people) and gives each a numerical rating between 1 and 10.

1. It would be natural to rate the applications based on their combined scores; however, there is a worry that different raters use different standards, and we would like to correct for this. Set up a model for the ratings (with parameters for the applicants and the raters).

lmer(rating_scores~applicants_ID+raters_ID+(1|raters_ID))

2. It is possible that some persons on the committee show more variation than others in their ratings. Expand your model to allow for this.

lmer(rating~applicants+raters+(1+raters|raters))