Homework 7

Yifu Dong

November 12, 2018

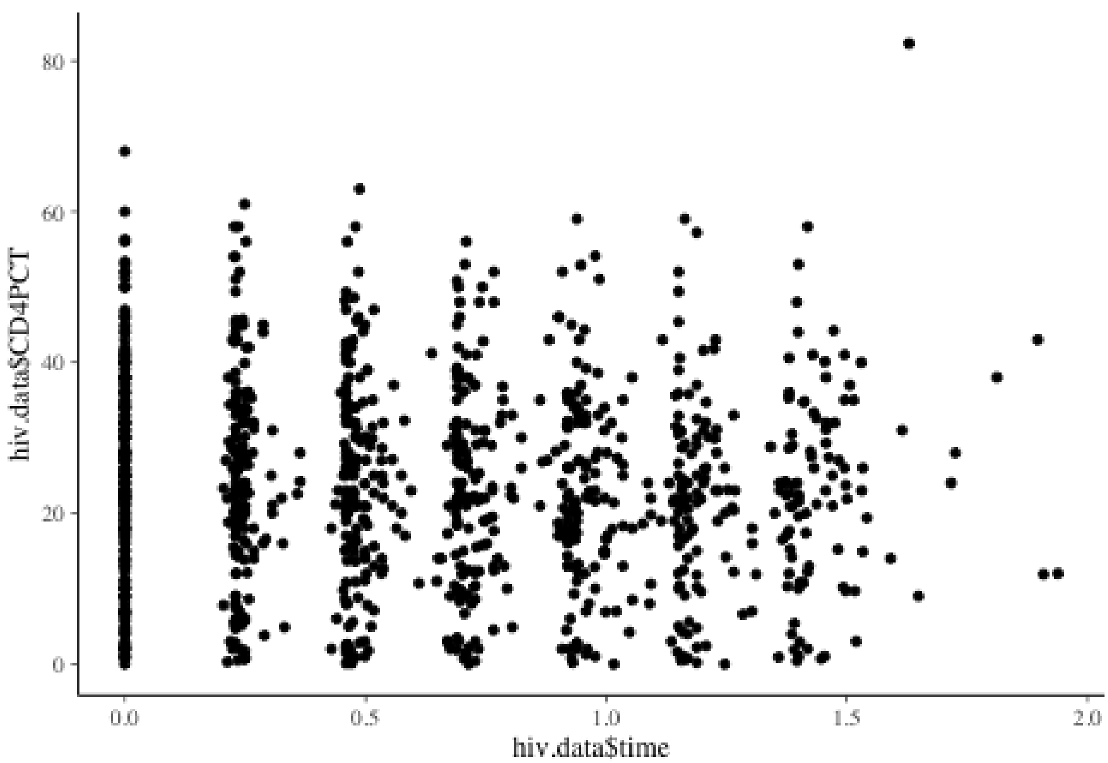
Data analysis

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.

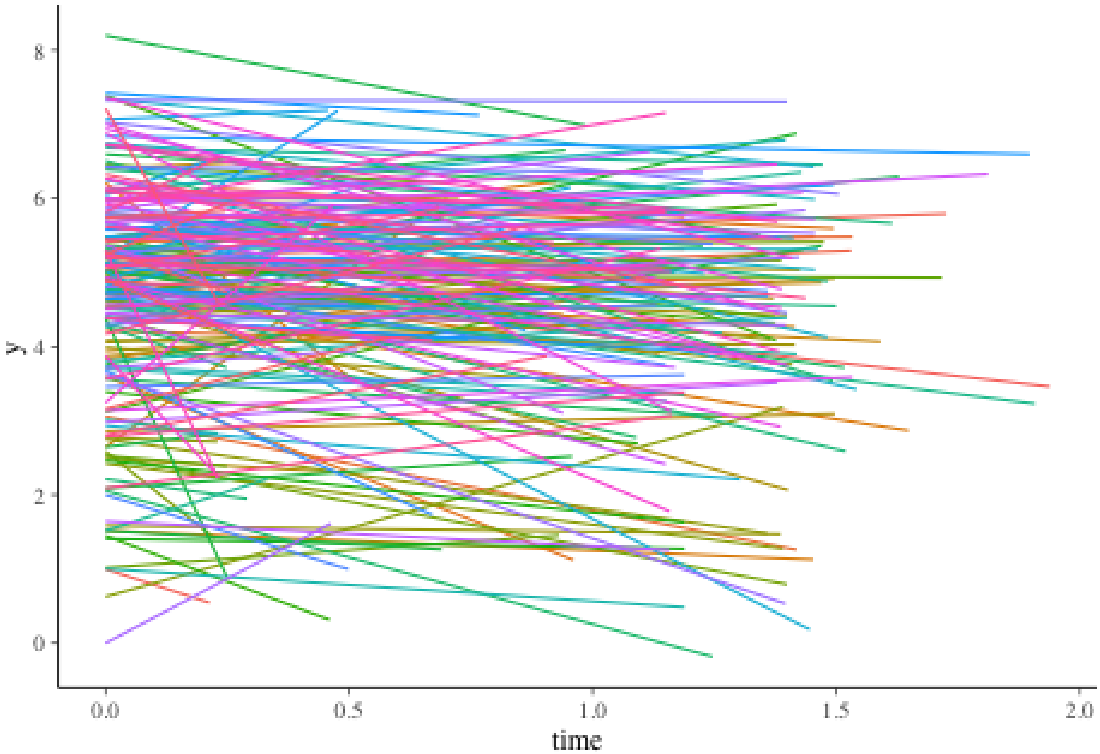
1. Graph the outcome (the CD4 percentage, on the square root scale) for each child as a function of time.

```
ggplot(data=hiv.data,mapping = aes(x=hiv.data$time,y=hiv.data$CD4PCT))+
  geom_point()
```



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.

```
m1_2 <- lm (y ~ time+factor(newpid)-1, data = hiv.data) #Varying Intercepts
# display(m1)
coef_m1_2 <- data.frame(coef(m1_2))
ggplot(data = hiv.data,aes(x=time, y=y,color=factor(newpid))) +
  geom_smooth(method="lm",se=FALSE,size=0.5) +
  theme(legend.position="none")
```



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.

```
##
## Call:
## lm(formula = coef.id ~ baseage + factor(treatment), data = r1.coef)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1594 -0.7039  0.2265  1.1215  2.7256
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.10627    0.18728  27.265  < 2e-16 ***
## baseage            -0.12088    0.04023  -3.005  0.00293 **
## factor(treatment)2  0.14558    0.18421   0.790  0.43012
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.455 on 247 degrees of freedom
## Multiple R-squared:  0.03753,    Adjusted R-squared:  0.02974
## F-statistic: 4.816 on 2 and 247 DF,  p-value: 0.008875
```

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.

```
q4m=lmer(y~time+(1|newpid),data=hiv.data)
summary(q4m)
```

```
## Linear mixed model fit by REML ['lmerMod']
```

```
## Formula: y ~ time + (1 | newpid)
##    Data: hiv.data
##
## 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  49.372
## time        -0.36609    0.05399  -6.781
##
## Correlation of Fixed Effects:
##      (Intr)
## time -0.278
```

A single global estimate for the effect (slope) of variable "time". We should expect a decrease of about 0.366 in CD4 each year, in any given child.

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
q5m=lmer(y~time+treatment+baseage+(1|newpid),data=hiv.data)
summary(q5m)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ time + treatment + baseage + (1 | newpid)
##    Data: hiv.data
##
## REML criterion at convergence: 3137.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.7490 -0.4392  0.0097  0.4282  5.0141
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  newpid   (Intercept) 1.8897   1.3747
##  Residual             0.5969   0.7726
## Number of obs: 1072, groups:  newpid, 250
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  4.90606    0.31684  15.485
## time        -0.36216    0.05399  -6.708
## treatment    0.18008    0.18262   0.986
## baseage     -0.11945    0.04000  -2.986
```

```
##
## Correlation of Fixed Effects:
##           (Intr) time   trtmnt
## time      -0.086
## treatment -0.850  0.010
## baseage   -0.430 -0.017 -0.003
```

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

```
## ensuring that the two models are fitted to exactly the same data sets
reduced.data <- hiv.data[with(hiv.data, !is.na(time+age.baseline+treatment)),]

summary(fit.a <- lmer(y ~ time + (1|newpid), data = reduced.data))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ time + (1 | newpid)
##    Data: reduced.data
##
## 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  49.372
## time        -0.36609    0.05399  -6.781
##
## Correlation of Fixed Effects:
##      (Intr)
## time -0.278
```

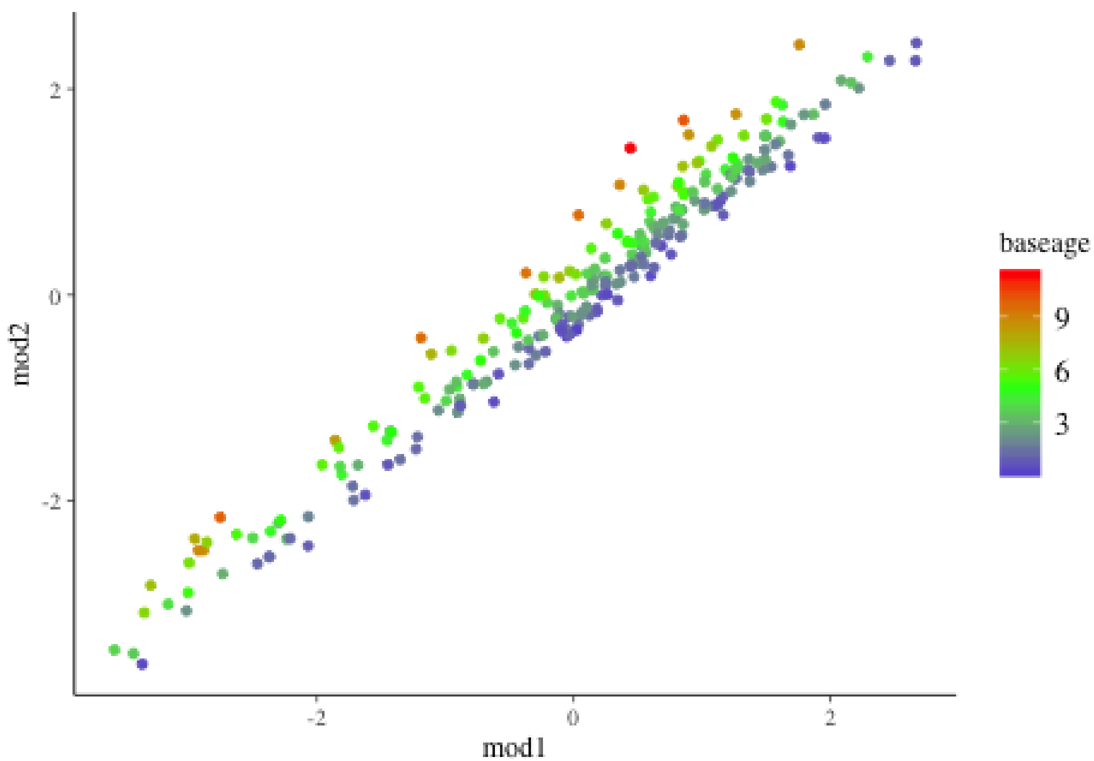
```
summary(fit.b <- lmer(y ~ time + baseage + treatment + (1|newpid),data = reduced.data))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ time + baseage + treatment + (1 | newpid)
##    Data: reduced.data
##
## REML criterion at convergence: 3137.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.7490 -0.4392  0.0097  0.4282  5.0141
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  newpid   (Intercept) 1.8897   1.3747
##  Residual             0.5969   0.7726
## Number of obs: 1072, groups:  newpid, 250
```

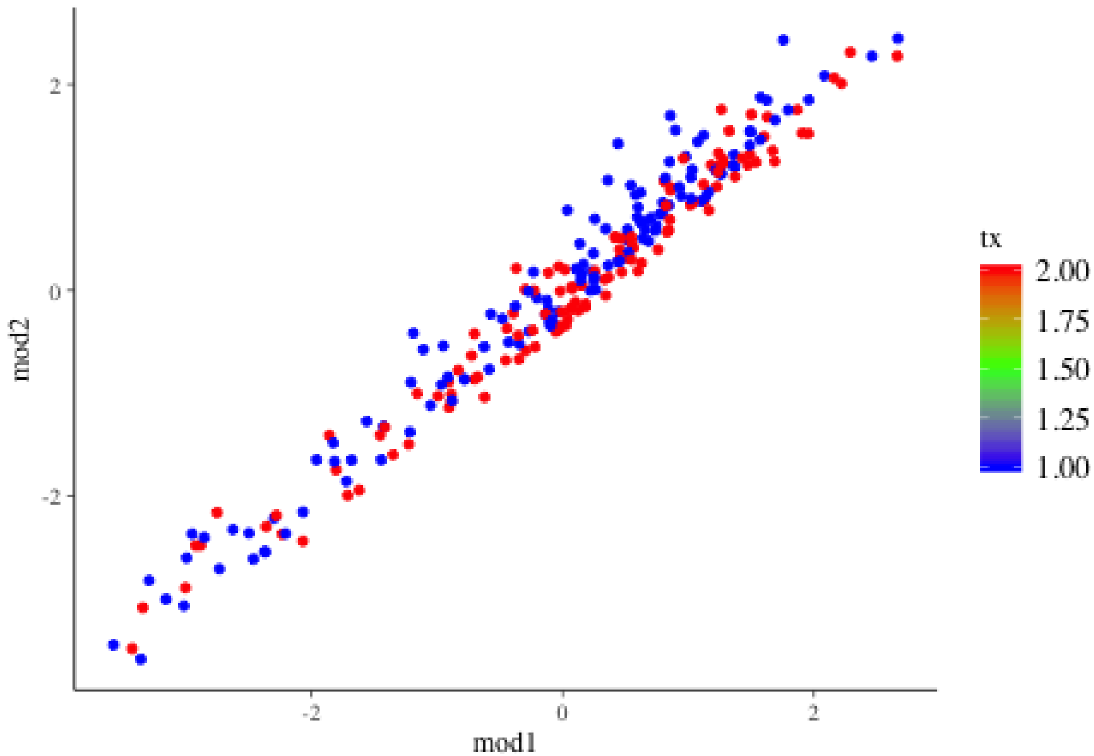
```
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  4.90606    0.31684  15.485
## time        -0.36216    0.05399  -6.708
## baseage     -0.11945    0.04000  -2.986
## treatment    0.18008    0.18262   0.986
##
## Correlation of Fixed Effects:
##           (Intr) time   baseag
## time      -0.086
## baseage   -0.430 -0.017
## treatment -0.850  0.010 -0.003
```

```
plotdata <- data.frame(mod1=ranef(fit.a)[[1]][,1],mod2=ranef(fit.b)[[1]][,1],
                       count=as.vector(table(reduced.data$newpid)),
                       baseage=sapply(split(reduced.data$baseage,
                                            reduced.data$newpid),
                                      function(x) x[1]),
                       tx=sapply(split(reduced.data$treatment,
                                       reduced.data$newpid),
                                 function(x) x[1]))

#By base age.
AGE <- ggplot(plotdata,aes(x=mod1,y=mod2)) +
  geom_point(aes(color=baseage))
AGE+scale_color_gradient2(midpoint=5, low="blue", mid="green",
                      high="red", space ="Lab" )
```

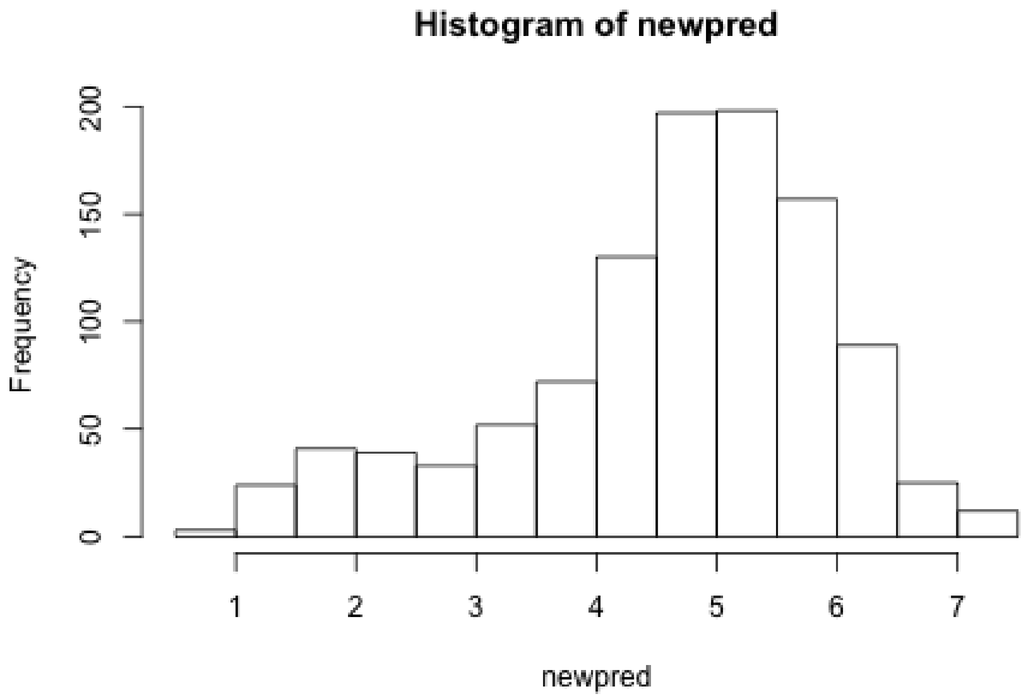


```r
# By treatment status.
TX <- ggplot(plotdata,aes(x=mod1,y=mod2)) +
  geom_point(aes(color=tx))
TX +scale_color_gradient2(midpoint=1.5, low="blue", mid="green",
                          high="red", space ="Lab" )
```



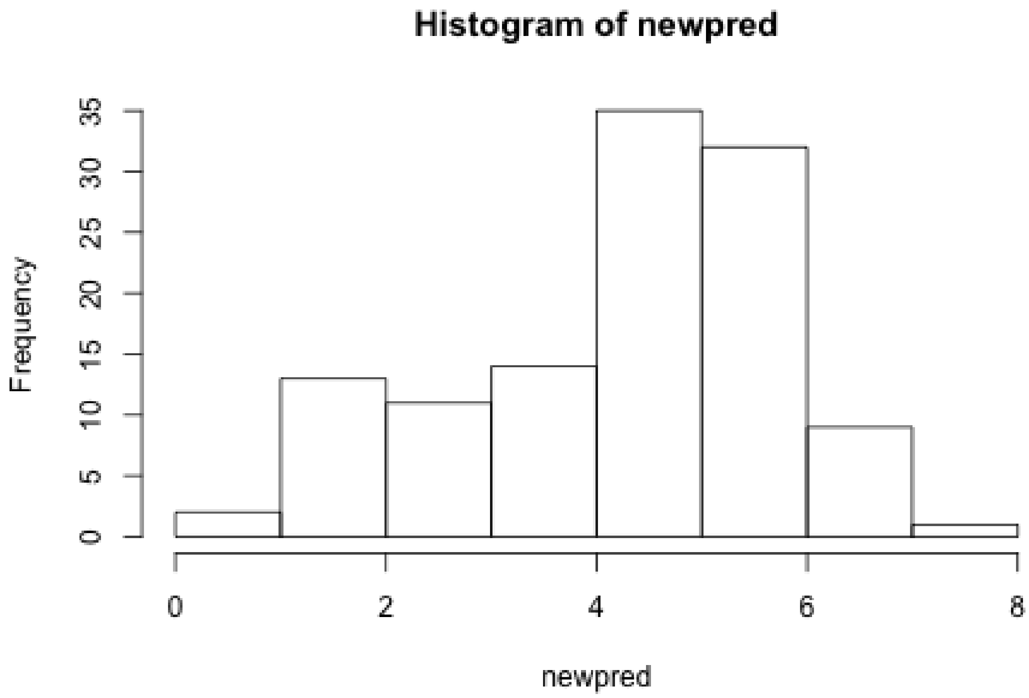
7. Use the model fit from (5) to generate simulation of predicted CD4 percentages for each child in the dataset at a hypothetical next time point.

```r
pred_data <- subset(hiv.data, !is.na(treatment) & !is.na(baseage))
pred_data <- pred_data[, -c(1, 4, 5, 6, 8)]
newpred <- predict(q5m, newdata = pred_data)
hist(newpred)
```



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.

```
pred_data <- pred_data[which(round(pred_data$baseage) == 4 ),]
newpred <- predict(q5m, newdata = pred_data)
hist(newpred)
```

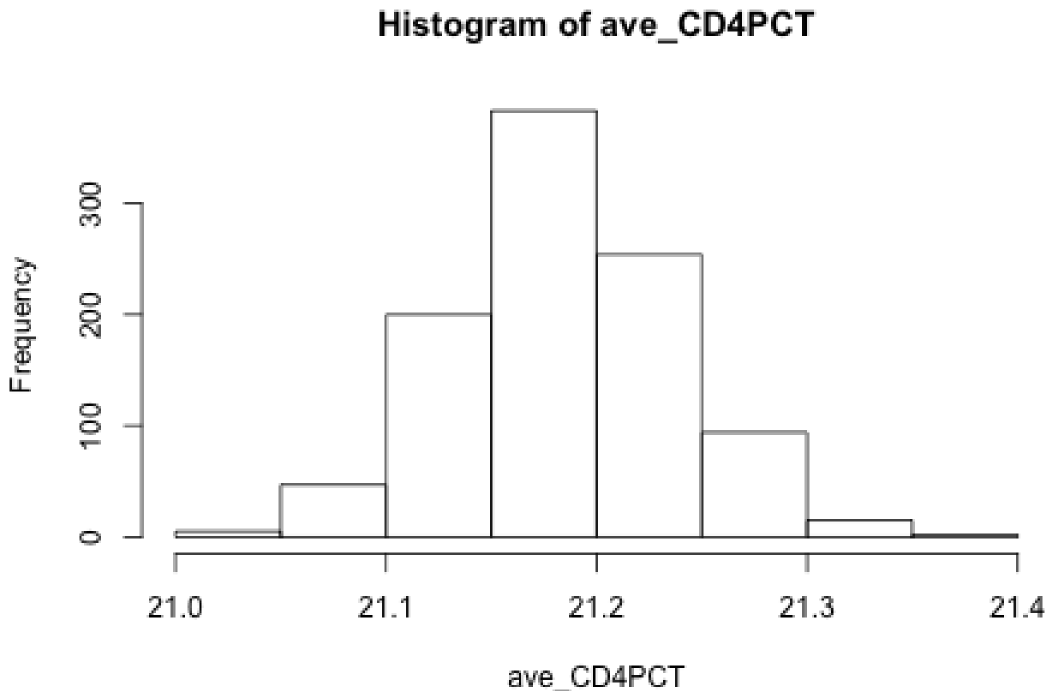


9. Posterior predictive checking: continuing the previous exercise, use the fitted model from (5) to simulate a new dataset of CD4 percentages (with the same sample size and ages of the original dataset) for the final time point of the study, and record the average CD4 percentage in this sample. Repeat this process 1000 times and compare the simulated distribution to the observed CD4 percentage at the final time point for the actual data.

```
newdata<-hiv.data %>%
  group_by(newpid) %>%
  arrange(desc(time),.by_group=TRUE) %>%
  filter(row_number() == 1) %>%
  select(newpid,treatment,time,age.baseline,CD4PCT)
newdata_original_mean<-mean(newdata$CD4PCT)

for(i in 1:1000) {
  newdata$treatment<-purrr::rbernoulli(dim(newdata)[1], p = sum(hiv.data$treatment==1)/dim(hiv.data)[1]
  newdata$treatment[newdata$treatment==0]<-2
  model_sim<-lmer(data = hiv.data, sqrt(CD4PCT)~(1|newpid) + time + treatment + age.baseline)
  re<-predict(model_sim,newdata=newdata)^2
  if(i==1)
    result<-re
  else
    result<-cbind(result,re)
}

ave_CD4PCT<-apply(result,2,mean)
hist(ave_CD4PCT)
```



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

```
m_10<-lmer(y~time+(1+time|newpid),data=hiv.data)
display(m_10)
```

```
## lmer(formula = y ~ time + (1 + time | newpid), data = hiv.data)
##             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
```

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).

```
m1_11<-lmer(y ~ factor(time) + (1 | newpid),data=hiv.data)
```

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

```
par(mfrow=c(2,2))
plot(m1_11)
```

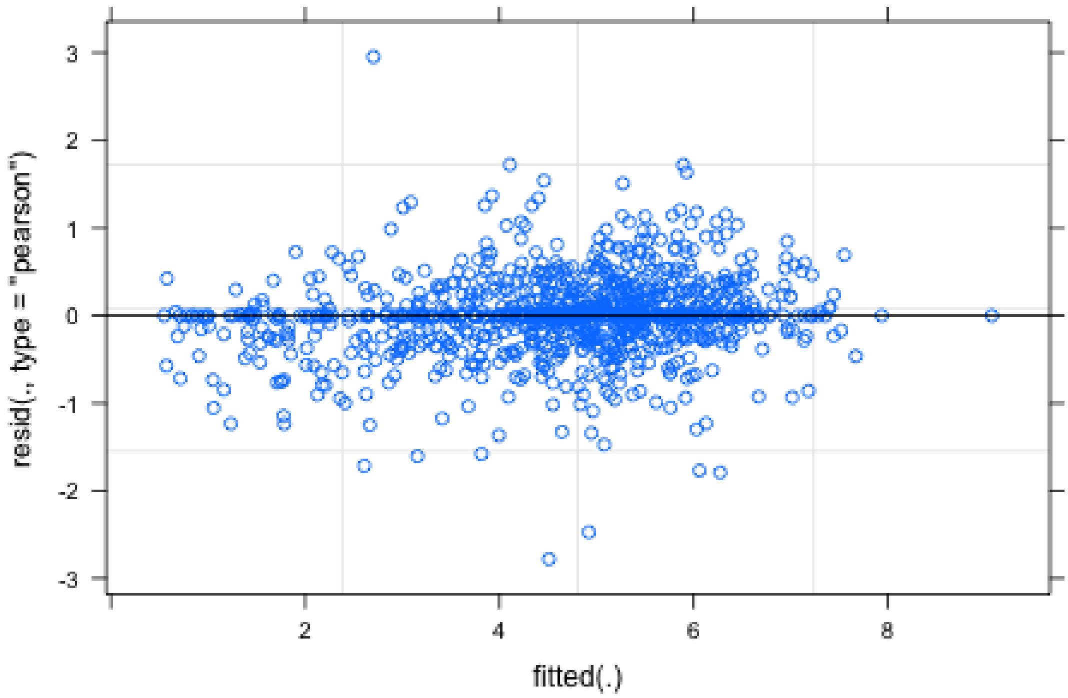


Figure skate in the 1932 Winter Olympics

The folder olympics has seven judges' ratings of seven figure skaters (on two criteria: "technical merit" and "artistic impression") from the 1932 Winter Olympics.

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

```
library(reshape)
```

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

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

## The following objects are masked from 'package:tidyr':
##
##     expand, smiths

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

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

```
m2_1<-melt(data = olympics1932,id.vars=c("pair","criterion"),measure.vars=c(colnames(olympics1932)[3:9])
m2_1
```

```
##    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.

```
m2_2 <- rename(m2_1, c("pair"="skater_ID", "variable"="judge_ID"))
m2_2 <- m2_2 [order(m2_2 $judge_ID),]
m2_2 <- m2_2 [c("criterion", "value", "skater_ID", "judge_ID")]
summary(m2_2)
```

```
##   criterion             value         skater_ID    judge_ID
##  Length:98          Min.   :3.600   Min.   :1   judge_1:14
##  Class :character   1st Qu.:4.800   1st Qu.:2   judge_2:14
##  Mode  :character   Median :5.250   Median :4   judge_3:14
##                     Mean   :5.113   Mean   :4   judge_4:14
##                     3rd Qu.:5.575   3rd Qu.:6   judge_5:14
##                     Max.   :6.000   Max.   :7   judge_6:14
##                                                 judge_7:14
```

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
m2_2$SameCountry <-ifelse(m2_2[,3] == " 1"&m2_2[,4] == "judge_5",1,
  ifelse(m2_2[,3] == " 2"&m2_2[,4] == "judge_7",1,
  ifelse(m2_2[,3] == " 3"&m2_2[,4] == "judge_1",1,
  ifelse(m2_2[,3] == " 4"&m2_2[,4] == "judge_1",1,
  ifelse(m2_2[,3] == " 7"&m2_2[,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
data_tech <- m2_2 %>%
  dplyr::filter(criterion=="Program")
data_art <- m2_2 %>%
  dplyr::filter(criterion=="Performance")

reg_tech <- lmer(value ~ 1 + (1|skater_ID) + (1|judge_ID),data=data_tech)
summary(reg_tech)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: value ~ 1 + (1 | skater_ID) + (1 | judge_ID)
##    Data: data_tech
##
## REML criterion at convergence: 60
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.51025 -0.45646 -0.05459  0.63866  1.89709
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  skater_ID (Intercept) 0.17488  0.4182
##  judge_ID  (Intercept) 0.07664  0.2768
##  Residual              0.11057  0.3325
## Number of obs: 49, groups:  skater_ID, 7; judge_ID, 7
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   5.1347     0.1954   26.28
```

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

```r
reg_art <- lmer(value ~ 1 + (1|skater_ID) + (1|judge_ID),data=data_art)
summary(reg_tech)
```

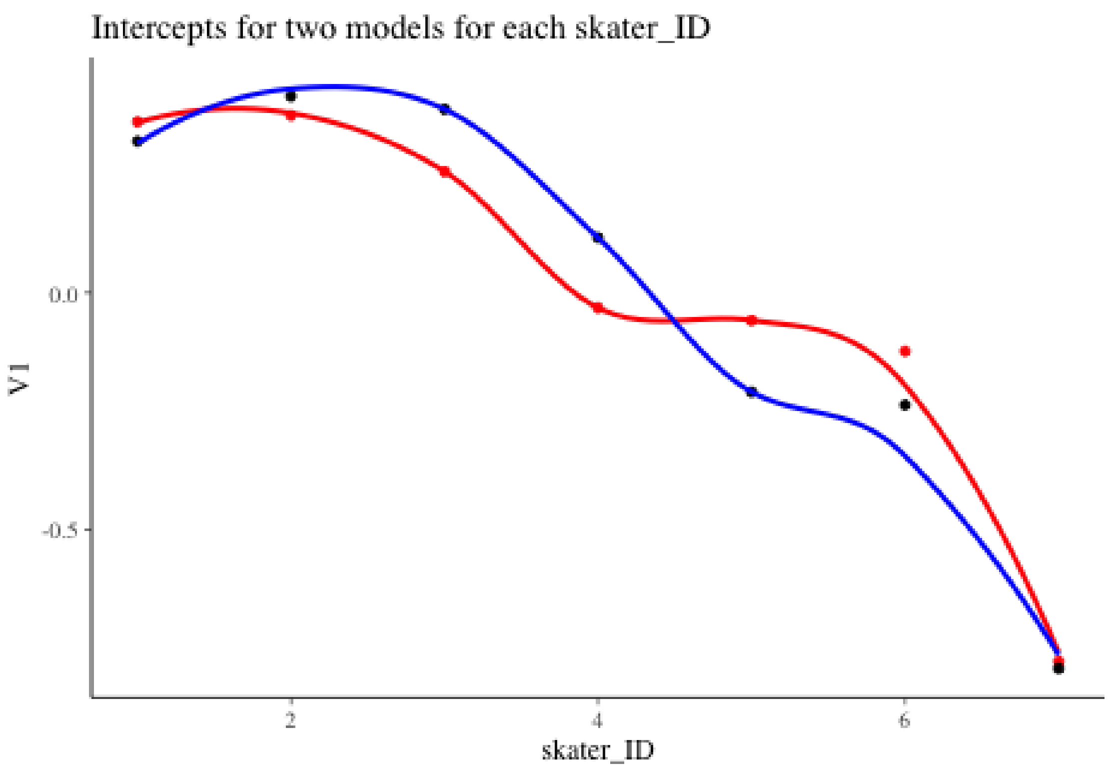
```
## Linear mixed model fit by REML ['lmerMod']
```

```
## Formula: value ~ 1 + (1 | skater_ID) + (1 | judge_ID)
##    Data: data_tech
##
## REML criterion at convergence: 60
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.51025 -0.45646 -0.05459  0.63866  1.89709
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  skater_ID (Intercept) 0.17488  0.4182
##  judge_ID  (Intercept) 0.07664  0.2768
##  Residual              0.11057  0.3325
## Number of obs: 49, groups:  skater_ID, 7; judge_ID, 7
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   5.1347     0.1954   26.28
```

6. Display your results for both outcomes graphically.

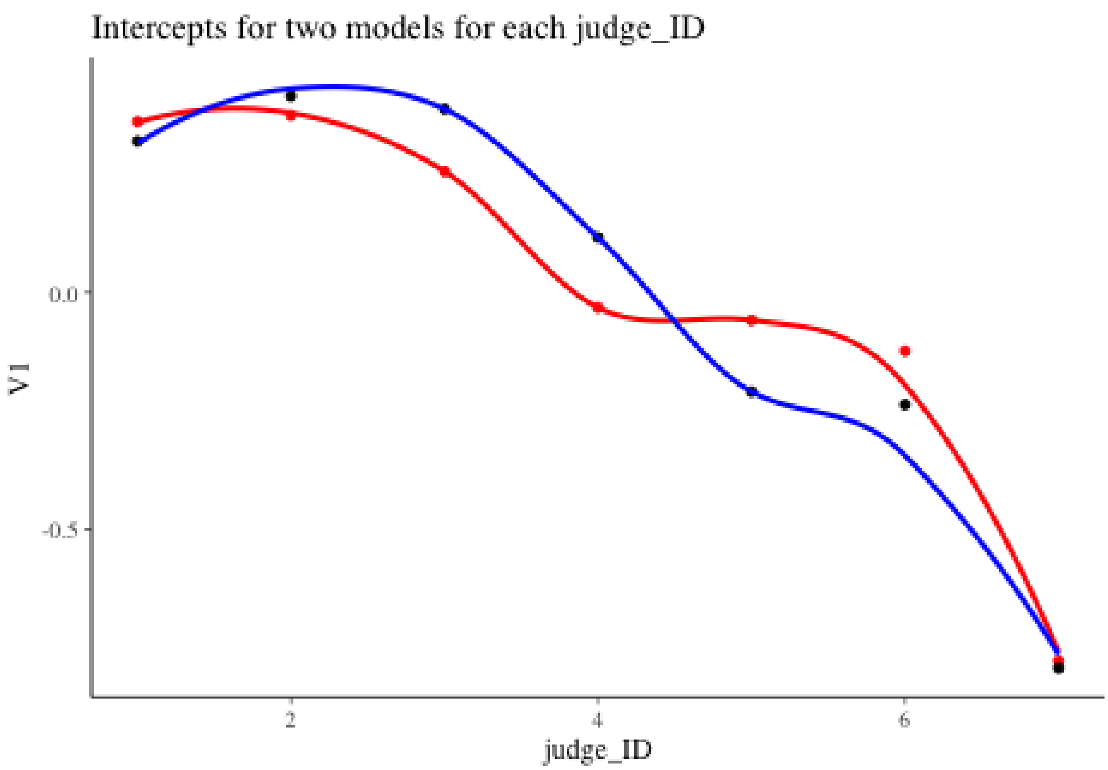
```
inter_skate <- as.data.frame(cbind(unlist(ranef(reg_tech))[1:7],unlist(ranef(reg_art))[1:7]))
inter_skate$skater_ID <-c(1:7)
ggplot(data=inter_skate)+
  geom_point(col="red",aes(x=skater_ID,y=V1))+geom_smooth(col="red",aes(x=skater_ID,y=V1),se=FALSE)+
  geom_point(col="black",aes(x=skater_ID,y=V2))+geom_smooth(col="blue",aes(x=skater_ID,y=V2),se=FALSE)+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

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




```
inter_judge <- as.data.frame(cbind(unlist(ranef(reg_tech))[1:7],unlist(ranef(reg_art))[1:7]))
inter_judge$judge_ID <-c(1:7)
ggplot(data=inter_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="blue",aes(x=judge_ID,y=V2),se=FALSE)+
  ggtitle("Intercepts for two models for each judge_ID")
```

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



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. For this question, we use the hiv data, which is the model of question1. First we fit the model using "lmer"

```
lmer(formula=hiv.data$y~hiv.data$time+hiv.data$age.baseline+hiv.data$treatment+(1|hiv.data$newpid))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## hiv.data$y ~ hiv.data$time + hiv.data$age.baseline + hiv.data$treatment +
##     (1 | hiv.data$newpid)
## REML criterion at convergence: 3137.209
## Random effects:
##  Groups          Name        Std.Dev.
##  hiv.data$newpid (Intercept) 1.3747
##  Residual                    0.7726
## Number of obs: 1072, groups:  hiv.data$newpid, 250
## Fixed Effects:
##           (Intercept)          hiv.data$time  hiv.data$age.baseline
##                4.9061                -0.3622                -0.1195
##    hiv.data$treatment
##                0.1801
```

$$X1 = time, \;\; X2 = age.baseline, \;\; X3 = treatment$$

.

$$y = 4.91 + X_{i1} * (-0.36) + X_{i2} * (-0.12) + X_{i3} * 0.18 + 0.77\ ,\ \ for\ i = 1, ..., n_{250}$$

$$y \sim\ N(4.91 + X_{i1} * (-0.36) + X_{i2} * (-0.12) + X_{i3} * 0.18,\ 0.77^2)\ ,\ \ for\ i = 1, ..., n_{250}$$

$$y_i \sim\ N(4.91 + X_{i1} * (-0.36) + X_{i2} * (-0.12) + X_{i3} * 0.18,\ 0.77^2)$$

$$y_i \sim\ N(4.91 + X_{i1} * (-0.36) + X_{i2} * (-0.12) + X_{i3} * 0.18 + 1.37^2,\ 0.77^2)$$

$$y_i \sim\ N(4.91 + X_{i1} * (-0.36) + X_{i2} * (-0.12) + X_{i3} * 0.18,\ 1.37^2 + 0.77^2)$$

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).

Since we don't know what exactly the data is, so we put the code here: lmer(rating_scores~applicants_ID+raters_ID+(1|raters_

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_scores~applicants_ID+raters_ID+(1+raters_ID|raters_ID))