homework 07

Xuan Zhu

November 17, 2018

Data analysis

CD4 percentages for HIV infected kids

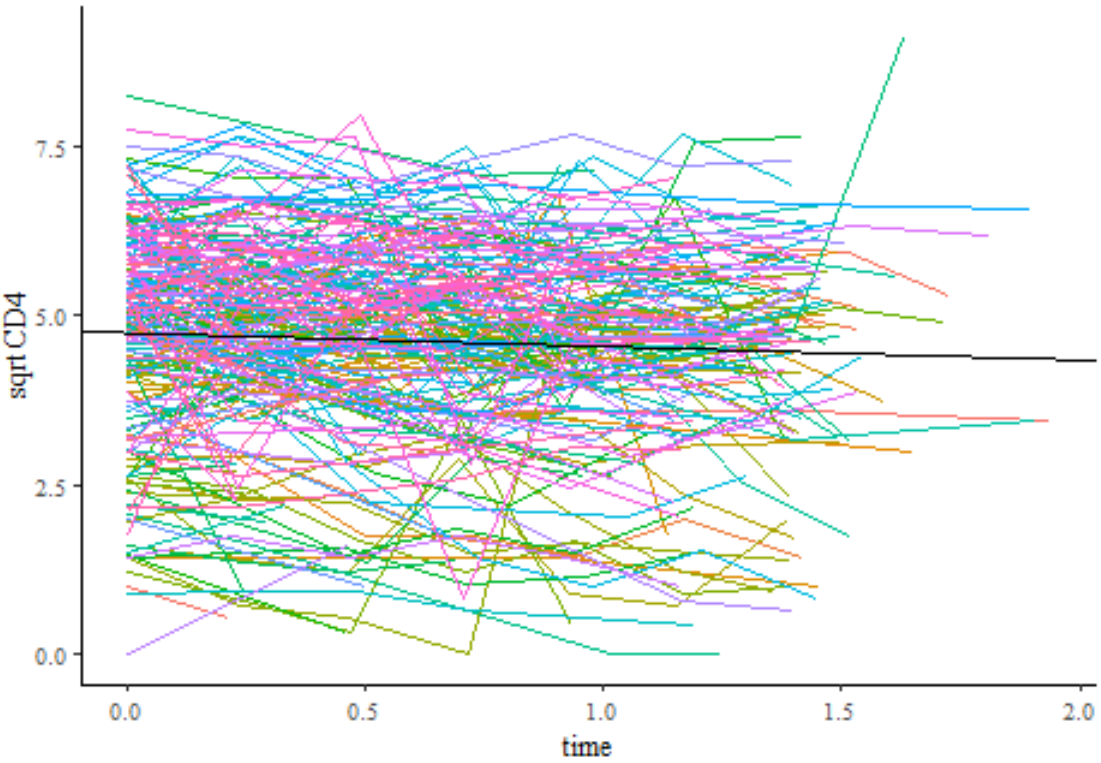
The folder `cd4` has CD4 percentages for a set of young children with HIV who were measured several times over a period of two years. The dataset also includes the ages of the children at each measurement.

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

```
lm <- lm(y~time,data=hiv.data)
display(lm)
```

```
## lm(formula = y ~ time, data = hiv.data)
##             coef.est coef.se
## (Intercept)  4.75     0.08
## time        -0.20     0.10
## ---
## n = 1072, k = 2
## residual sd = 1.56, R-Squared = 0.00
```

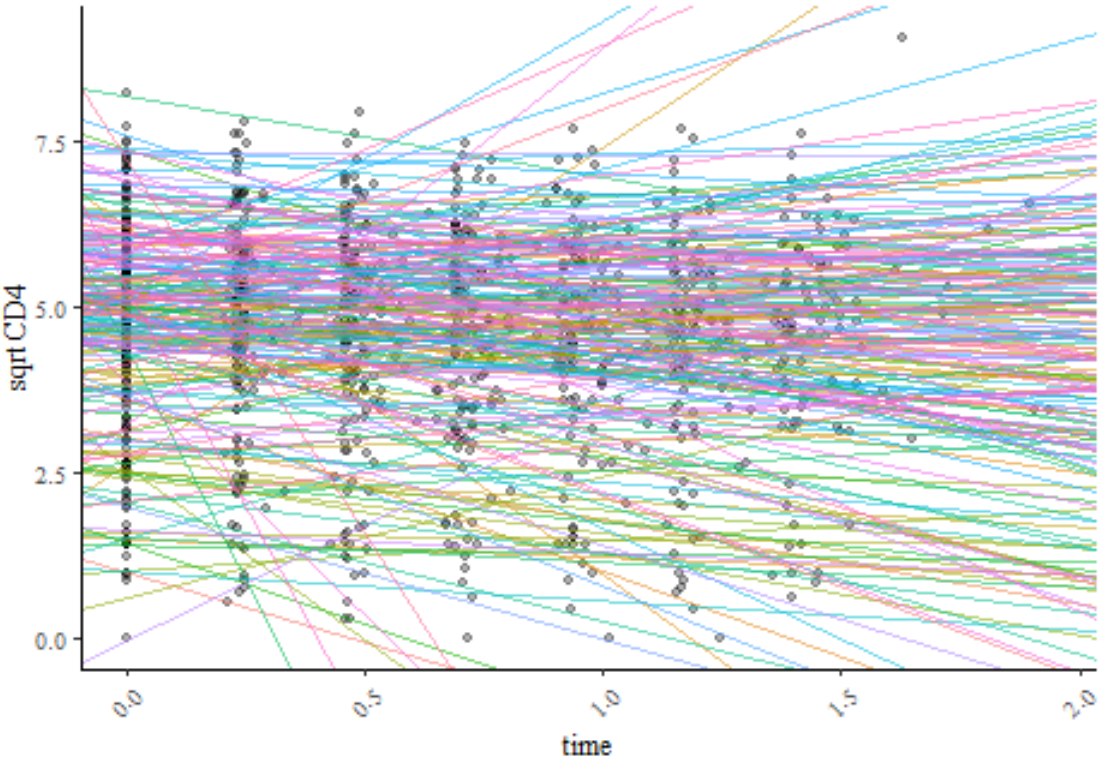
```
ggplot(hiv.data)+aes(x=time, y=y,color=factor(newpid))+geom_line()+  theme(legend.position="none")+ylab
```



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.

```
no_pooling<-hiv.data[,list(alpha=coef(lm(y~time))[1],
              beta=coef(lm(y~time))[2]),by=newpid]
ggplot(hiv.data)+aes(x=time, y=y)+
  geom_jitter(alpha=0.3)+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  theme(legend.position="none")+
  ylab("sqrt CD4")+geom_abline( data=no_pooling,aes(slope=beta, intercept=alpha,color=factor(newpid),al
```

```
## Warning: Removed 26 rows containing missing values (geom_abline).
```



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 a. first estimate the intercept and slope separately for each child, then fit the between-child models using the point estimates from the first step.

```
intercept <- matrix()
slope <- matrix()
for (i in 1:254){
  if (nrow(hiv.data[newpid==i]) != 0 ){
     partlm <-lm(y~time,data=hiv.data[newpid==i])  #no-pooling
     intercept[i] <- coef(partlm)[1]
     slope[i] <- coef(partlm)[2]
  }else{
     intercept[i] <- NA
     slope[i] <- NA
  }
}
hiv.data<-cbind.data.frame(hiv.data,intercept,slope)
```

```
## Warning in data.table::data.table(...): Item 2 is of size 254 but maximum
## size is 1072 (recycled leaving remainder of 56 items)
```

```
## Warning in data.table::data.table(...): Item 3 is of size 254 but maximum
## size is 1072 (recycled leaving remainder of 56 items)
```

```
regression21 <-lm(slope~treatment+age.baseline,data=hiv.data)
display(regression21)
```

```
## lm(formula = slope ~ treatment + age.baseline, data = hiv.data)
##              coef.est coef.se
## (Intercept)  -0.54     0.22
## treatment     0.09     0.13
## age.baseline -0.02     0.03
## ---
## n = 944, k = 3
## residual sd = 1.97, R-Squared = 0.00
```

```
regression22 <-lm(intercept~treatment+age.baseline,data=hiv.data)
display(regression22)
```

```
## lm(formula = intercept ~ treatment + age.baseline, data = hiv.data)
##              coef.est coef.se
## (Intercept)   4.66     0.16
## treatment     0.07     0.09
## age.baseline -0.01     0.02
## ---
## n = 1055, k = 3
## residual sd = 1.49, R-Squared = 0.00
```

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.

```
lmer_vi <- lmer(y ~ time + (1 | newpid), hiv.data)
summary(lmer_vi)
```

```
## 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
```

For an average level of subject, one unit of change in time leads to 0.36609 decrease in CD4 percentage.

The average line is y=4.76-0.366time

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.

```
lmer_vi2 <-lmer(data=hiv.data,y ~ time+treatment+age.baseline+(1|newpid))
summary(lmer_vi2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ time + treatment + age.baseline + (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
## age.baseline -0.11945    0.04000  -2.986
##
## Correlation of Fixed Effects:
##             (Intr) time   trtmnt
## time        -0.086
## treatment   -0.850  0.010
## age.baselin -0.430 -0.017 -0.003
```

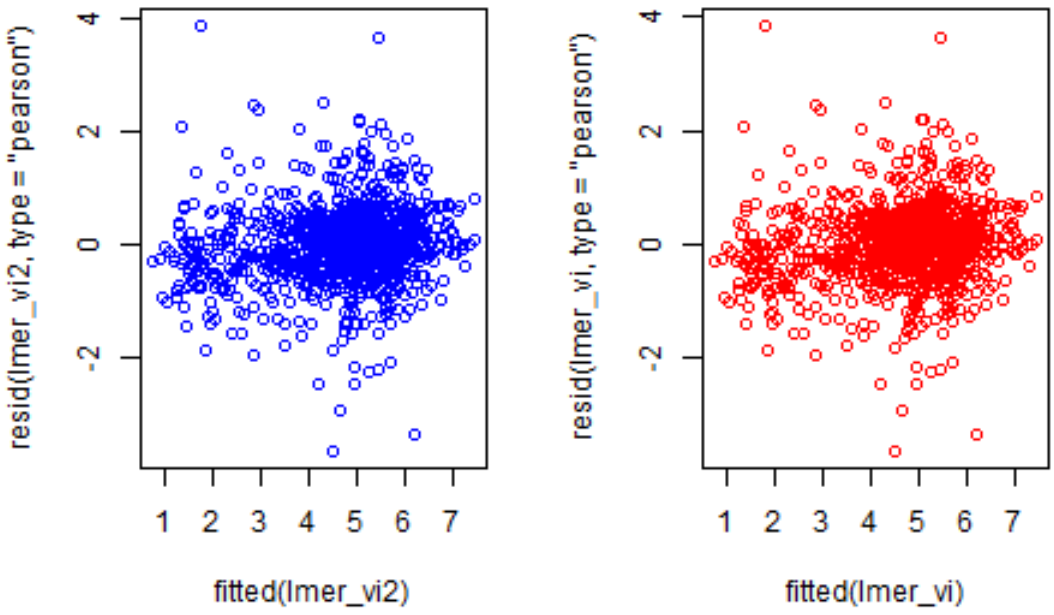
The average line is y=4.9-0.36time+0.18treatment-0.12age.baseline

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

(4) var(alpha)=1.9569 var(y)=0.5968 var(a)/var(y)=3.28

(5) var(alpha)=1.8897 var(y)=0.5969 var(a)/var(y)=3.16 The variance of alpha decreases.

```
par(mfrow=c(1,2))
plot(fitted(lmer_vi2),resid(lmer_vi2,type="pearson"),col="blue")
plot(fitted(lmer_vi),resid(lmer_vi,type="pearson"),col="red")
```



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.
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.
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.
10. Extend the model to allow for varying slopes for the time predictor.
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).
12. Compare the results of these models both numerically and graphically.

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).

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------- tidyverse 1.2.1 --
## v tibble  1.4.2     v purrr   0.2.5
```

```
## v tidyr   0.8.1     v dplyr   0.7.6
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------- tidyverse_conflicts() --
## x dplyr::between()   masks data.table::between()
## x dplyr::combine()   masks gridExtra::combine()
## x tidyr::expand()    masks Matrix::expand()
## x tidyr::extract()   masks rstan::extract()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x dplyr::recode()    masks car::recode()
## x dplyr::select()    masks MASS::select()
## x purrr::some()      masks car::some()
## x purrr::transpose() masks data.table::transpose()
```

```
performance <- olympics1932 %>% filter(criterion=="Performance")
program <- olympics1932 %>% filter(criterion=="Program")
a <- list()
a[[1]]<-performance
a[[2]]<-program
```

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.

```
tidydata <- matrix(NA,nrow=98,ncol=4)
colnames(tidydata)<- c("merit","impression","skaterID","judgeID")
tidydata[1:14,4]<- 1
tidydata[15:28,4]<- 2
tidydata[29:42,4]<- 3
tidydata[43:56,4]<- 4
tidydata[57:70,4]<- 5
tidydata[71:84,4]<- 6
tidydata[85:98,4]<- 7
for (i in 1:14){
  tidydata[(7*i-6):(7*i),3]<- c(1:7)
}
```

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.
4. Write the notation for a non-nested multilevel model (varying across skaters and judges) for the technical merit ratings and fit using lmer().
5. Fit the model in (4) using the artistic impression ratings.
6. Display your results for both outcomes graphically.
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.

We use the hiv.data from the first problem

1. Allowing regression coefficeints to vary accross groups $y_i \sim N(\alpha_{j[i]} - 0.36X_{time}, 0.59), \sigma_\alpha^2 = 1.95$
2. Combining separate local regressions $y \sim N(\alpha_j - 0.36X_{time}, 0.59), for i = 1, 2, ...n_j, \alpha_j \sim N(\gamma_0 + \gamma_1 u_j, 1.95)$

3.Modeling the coefficients of a large regression model

$y_i \sim N(4.91 - 0.36X_{time}, 0.59, for = 1, ...n)\ \beta_j \sim N(\mu_\alpha, 1.95) for j = 4, ...j + 3$

4.Regression with multiple error terms $y_i \sim N(-0.36X_{time} + \eta_{j[i]}, 0.59, \eta_j \sim N(0, 1.95)$

5.Large regression with correlated errors

$y_i = -0.36X_{time} + E_i, E_i \sim N(0, V)$

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).

The final combined scores are affected by two things: the quality of the applicants themselves, and the random effect from each rater.

so lmer(data=. . . , score~quality+(1|grader))

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.

Now we apply varying slope+varying intercept model.

lmer(data=. . . ,score~quality+{1+quality|grader})