homework 07

Kerui Cao

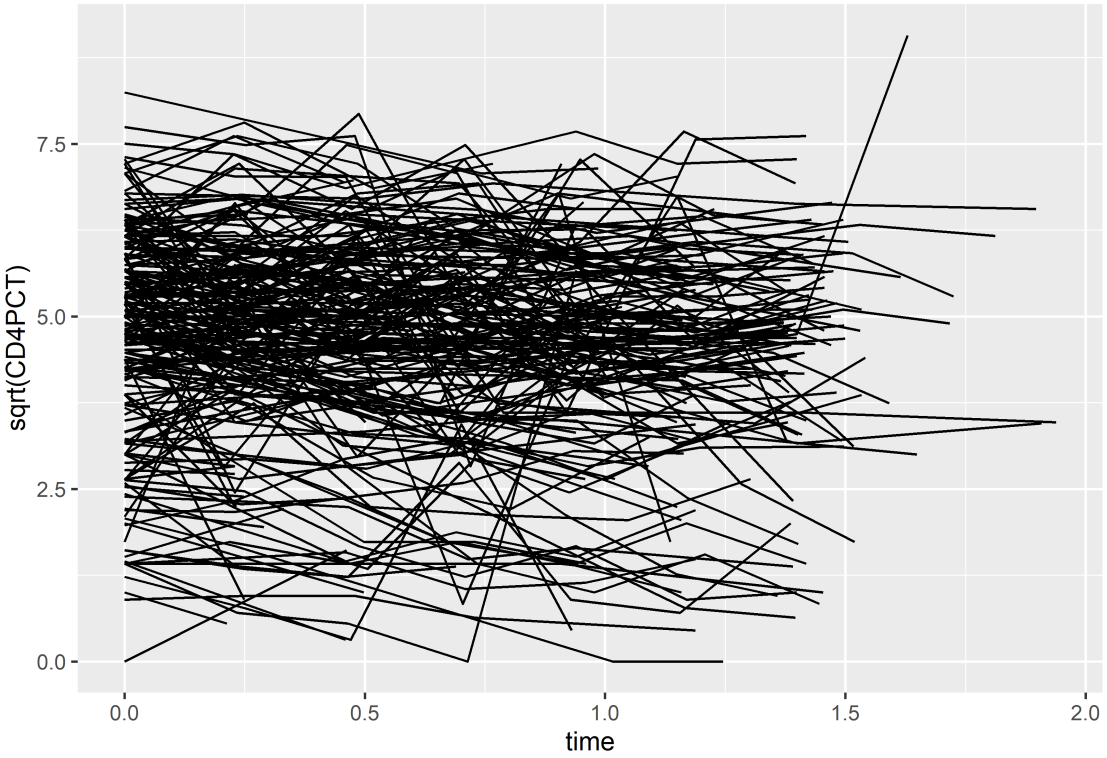
November 3, 2019

Data analysis

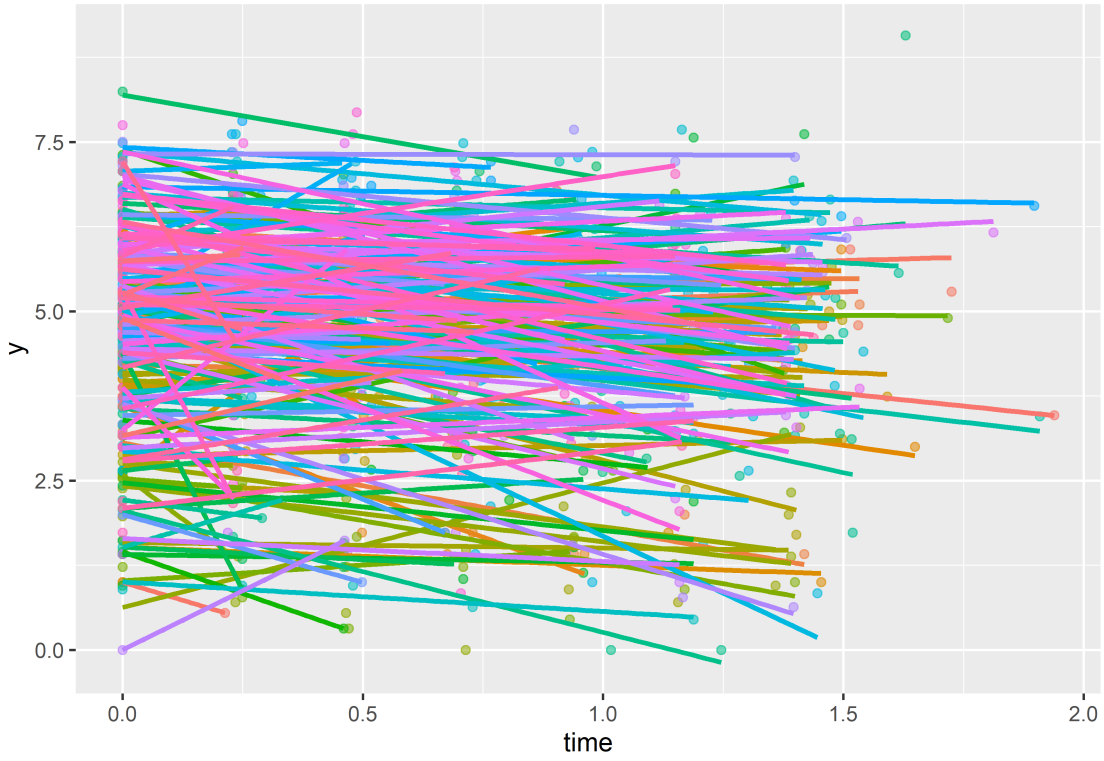
CD4 percentages for HIV infected kids

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.



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.



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 ~ age.baseline + factor(treatment), data = fit1.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 ***
## age.baseline       -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.

```
## 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
```

The coeffcient for time is the same for all children. With every unit increase in time, the square root of CO4 percentage is expected to reduce by 0.3661

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.

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ time + factor(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)          5.08614    0.18793  27.064
## time                -0.36216    0.05399  -6.708
## factor(treatment)2   0.18008    0.18262   0.986
## age.baseline        -0.11945    0.04000  -2.986
##
## Correlation of Fixed Effects:
##             (Intr) time   fct()2
## time        -0.135
```

```
## fctr(trtm)2 -0.462  0.010
## age.baselin -0.727 -0.017 -0.003
```

Coefficient for time: with every unit increase in time, the square root of CO4 percentage is expected to reduce by 0.135 for children with the same treatment, baseline age.

Coefficient for treatment: the average difference of CO4 percentage square root is 0.18 for childern in treatment 2 and treatment 1, if they have the same baseline age and at the same time point.

Coefficient for baseline age: with every unit increase in aseline age, the square root of CO4 percentage is expected to reduce by 0.727 for children with the same treatment and at the same time point.

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

	Model in (4)	Model in (5)
(Intercept)	4.7634086	5.0861380
time	-0.3660932	-0.3621573
factor(treatment)2	NA	0.1800822
age.baseline	NA	-0.1194538

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.

y	newpid	treatment	age.baseline	time
3.8212979	1	1	3.9100000	2
0.6375377	2	2	3.5650000	2
5.1373054	3	1	6.1241667	2
4.8355881	4	1	2.3025000	2
3.5176450	5	1	0.6541667	2
4.6065669	6	2	2.9183333	2
4.8295027	7	2	6.4425000	2
4.3581317	8	1	5.0266667	2
5.3982441	9	1	1.4975000	2
4.8953757	10	1	8.5583333	2

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.

y	time	treatment	age.baseline
4.788405	0.0000000	2	4
4.714163	0.2050000	2	4
4.714163	0.2050000	2	4
4.712352	0.2100000	2	4
4.711145	0.2133333	2	4
4.711145	0.2133333	2	4
4.710240	0.2158333	2	4
4.710240	0.2158333	2	4
4.709938	0.2166667	2	4
4.709334	0.2183333	2	4

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.

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ factor(treatment) + age.baseline + time + (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
## factor(treatment)2   0.15952    0.18137   0.880
## age.baseline        -0.12423    0.03971  -3.128
## time                -0.35258    0.06763  -5.214
##
## Correlation of Fixed Effects:
##             (Intr) fct()2 ag.bsl
## fctr(trtm)2 -0.463
## age.baselin -0.729 -0.004
## time        -0.114  0.010 -0.013
```

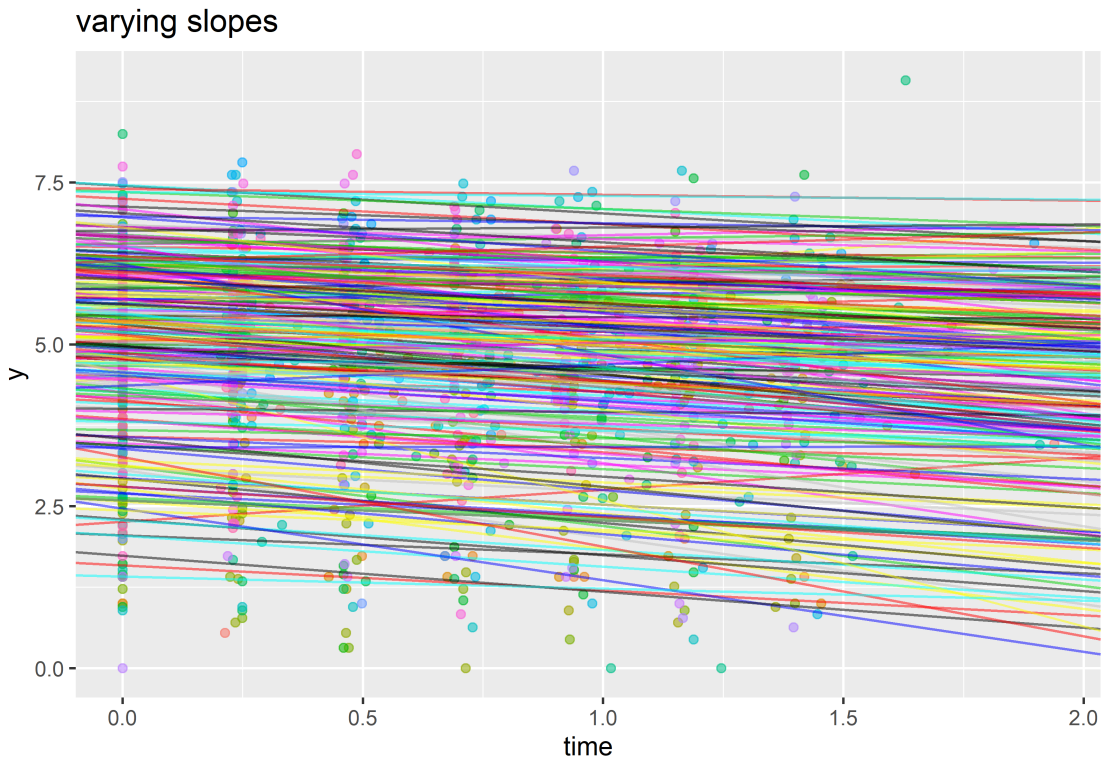
11. Next fit a model that does not allow for varying slopes but does allow for different coefficients for each time point (rather than fitting the linear trend).

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ factor(treatment) + age.baseline + time + (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
## factor(treatment)2  0.15952    0.18137   0.880
## age.baseline       -0.12423    0.03971  -3.128
## time               -0.35258    0.06763  -5.214
##
## Correlation of Fixed Effects:
##             (Intr) fct()2 ag.bsl
## fctr(trtm)2 -0.463
## age.baselin -0.729 -0.004
## time        -0.114  0.010 -0.013
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ factor(treatment) + age.baseline + time + (time - 1 | newpid)
##    Data: hiv.data
##
## REML criterion at convergence: 3713.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.2749 -0.4197  0.0510  0.5128  2.5362
##
## Random effects:
##  Groups   Name Variance Std.Dev.
##  newpid   time 1.957    1.399
##  Residual      1.393    1.180
## Number of obs: 1072, groups:  newpid, 250
##
## Fixed effects:
##                     Estimate Std. Error t value
## (Intercept)         5.09838    0.10765  47.363
## factor(treatment)2  0.24232    0.10107   2.398
## age.baseline       -0.12347    0.02234  -5.526
## time               -0.30568    0.13310  -2.297
##
## Correlation of Fixed Effects:
##             (Intr) fct()2 ag.bsl
## fctr(trtm)2 -0.438
## age.baselin -0.715 -0.004
## time        -0.269  0.013 -0.017
```



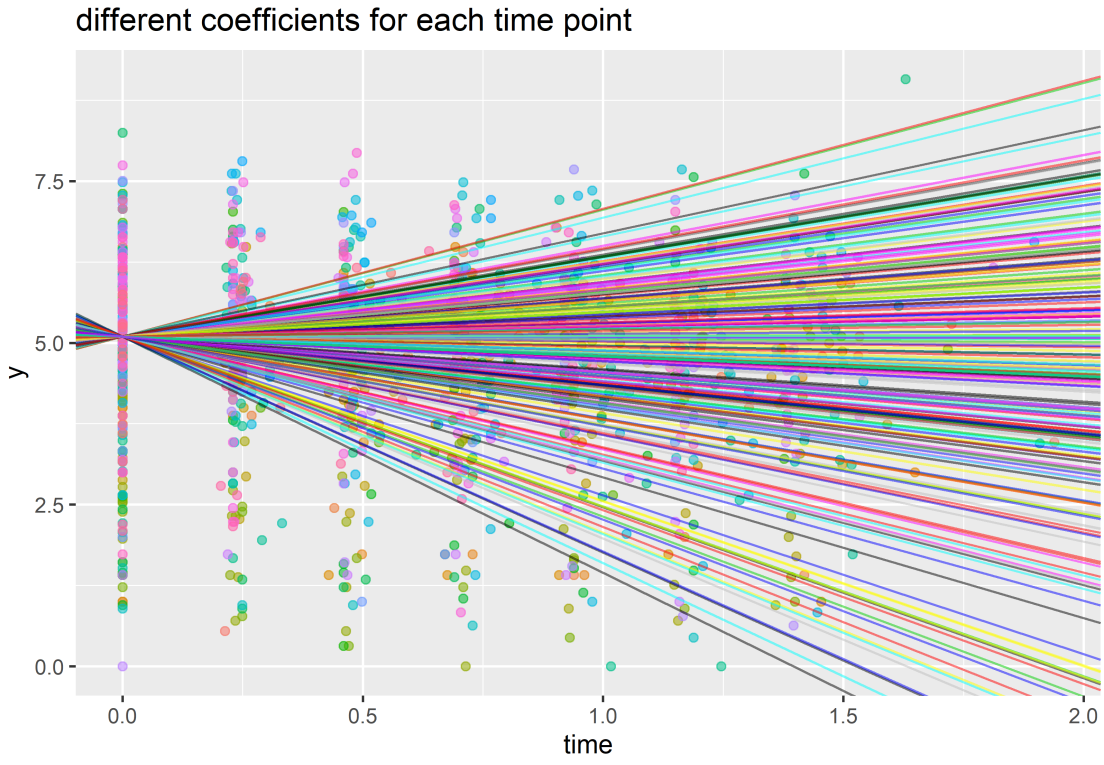


Figure skate in the 1932 Winter Olympics

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).

```
## , , 1
##
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## [1,]  5.6  5.5  5.8  5.3  5.6  5.2  5.7
## [2,]  5.5  5.2  5.8  5.8  5.6  5.1  5.8
## [3,]  6.0  5.3  5.8  5.0  5.4  5.1  5.3
## [4,]  5.6  5.3  5.8  4.4  4.5  5.0  5.1
## [5,]  5.4  4.5  5.8  4.0  5.5  4.8  5.5
## [6,]  5.2  5.1  5.3  5.4  4.5  4.5  5.0
## [7,]  4.8  4.0  4.7  4.0  3.7  4.0  4.8
##
## , , 2
##
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## [1,]  5.6  5.5  5.8  4.7  5.7  5.3  5.4
## [2,]  5.5  5.7  5.6  5.4  5.5  5.3  5.7
## [3,]  6.0  5.5  5.7  4.9  5.5  5.2  5.7
## [4,]  5.6  5.3  5.8  4.8  4.5  5.0  5.5
## [5,]  4.8  4.8  5.5  4.4  4.6  4.8  5.2
## [6,]  4.8  5.6  5.0  4.7  4.0  4.6  5.2
## [7,]  4.3  4.6  4.5  4.0  3.6  4.0  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.

Program	Performance	pair	judge
5.6	5.6	1	1
5.5	5.5	1	2
5.8	5.8	1	3
5.3	4.7	1	4
5.6	5.7	1	5
5.2	5.3	1	6
5.7	5.4	1	7
5.5	5.5	2	1
5.2	5.7	2	2
5.8	5.6	2	3

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().

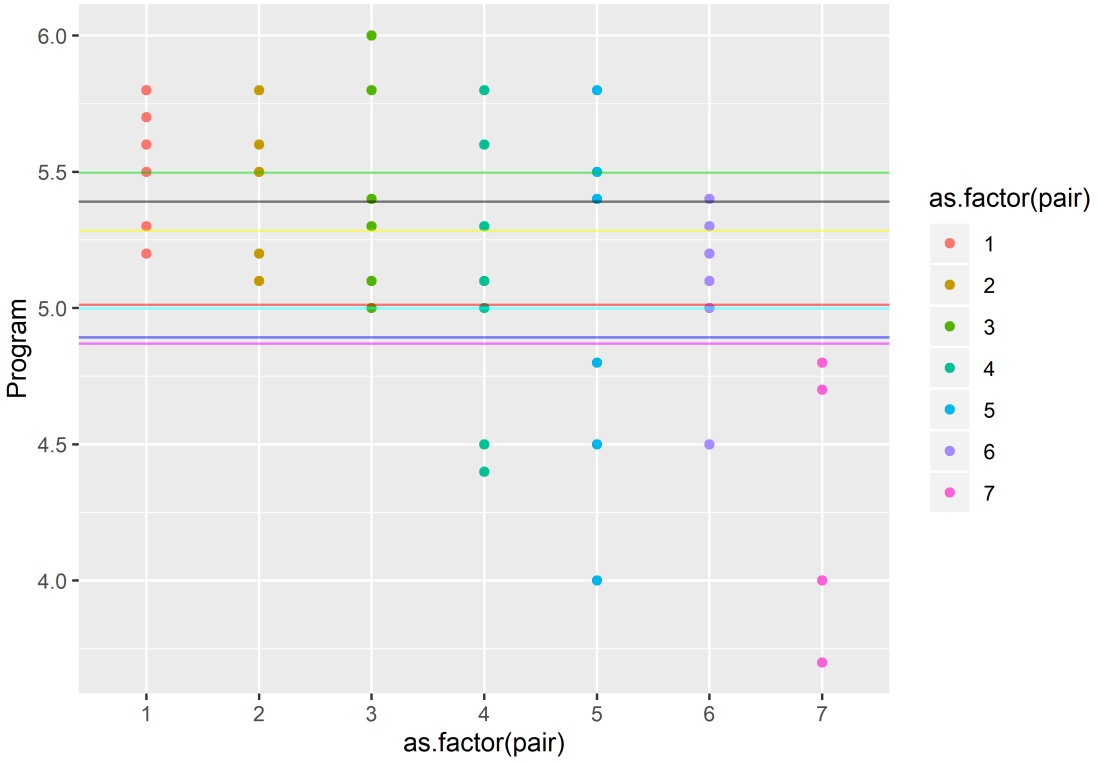
```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Program ~ (1 | pair) + (1 | judge)
##    Data: data
##
## REML criterion at convergence: 60
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
```

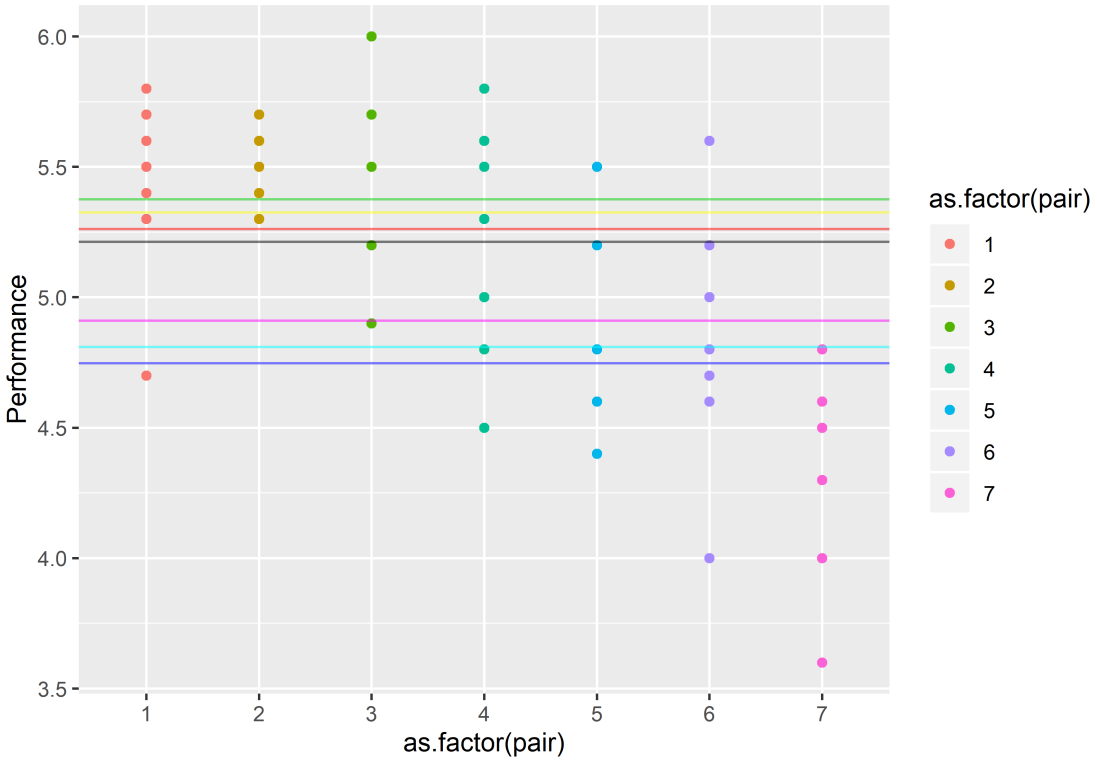
```
## -2.51032 -0.45651 -0.05458  0.63874  1.89726
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  pair     (Intercept) 0.17495  0.4183
##  judge    (Intercept) 0.07667  0.2769
##  Residual             0.11056  0.3325
## Number of obs: 49, groups:  pair, 7; judge, 7
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   5.1347     0.1955   26.27
```

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Performance ~ (1 | pair) + (1 | judge)
##    Data: data
##
## REML criterion at convergence: 46.2
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.10128 -0.50469 -0.09884  0.40875  2.10489
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  pair     (Intercept) 0.20486  0.4526
##  judge    (Intercept) 0.07759  0.2785
##  Residual             0.07446  0.2729
## Number of obs: 49, groups:  pair, 7; judge, 7
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   5.0918     0.2046   24.88
```

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.

Use the Olympic data as example and we set the technical merit as response variable, and with no predictors, treat each athelete as a group with score from each judge, the score of each athelete is y_{ij} means the score of athelete i from judge j.

- model in way 1: $y_{ij} = N(\alpha_i, \sigma^2), \alpha_i = \rho + \rho_1 u_i + \varepsilon, \varepsilon = N(0, \sigma_\alpha^2)$, here the α_i is the mean of each athlete, the score of each athlete comply with normal distribution, at the same time, the mean of each athlete also comply with normal distribution, but the mean are different cross athletes;
- model in way 2: $y_{ij} = N(\alpha_i, \sigma^2), \alpha_i = N(\rho + \rho_1 u_i, \sigma_\alpha^2)$, this model is exactly the same with way 1, just in way 1, the variability of α_i is introduced by ε_i, here we directly represent α_i as a random variable.

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).
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.