homework 07

Zara Waheed

November 9th, 2021

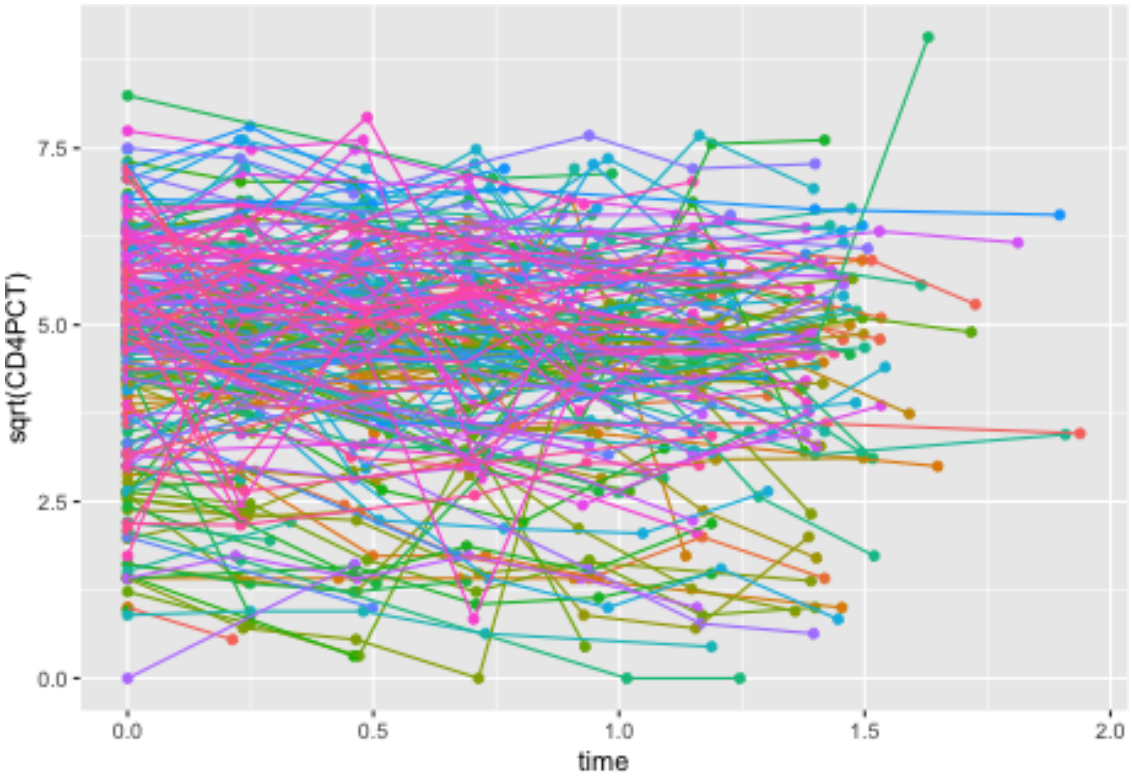
Data analysis

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(aes(x=time, y=sqrt(CD4PCT), color=factor(newpid)), data=hiv.data) +
  geom_point() +
  geom_line() +
  theme(legend.position = "none")
```

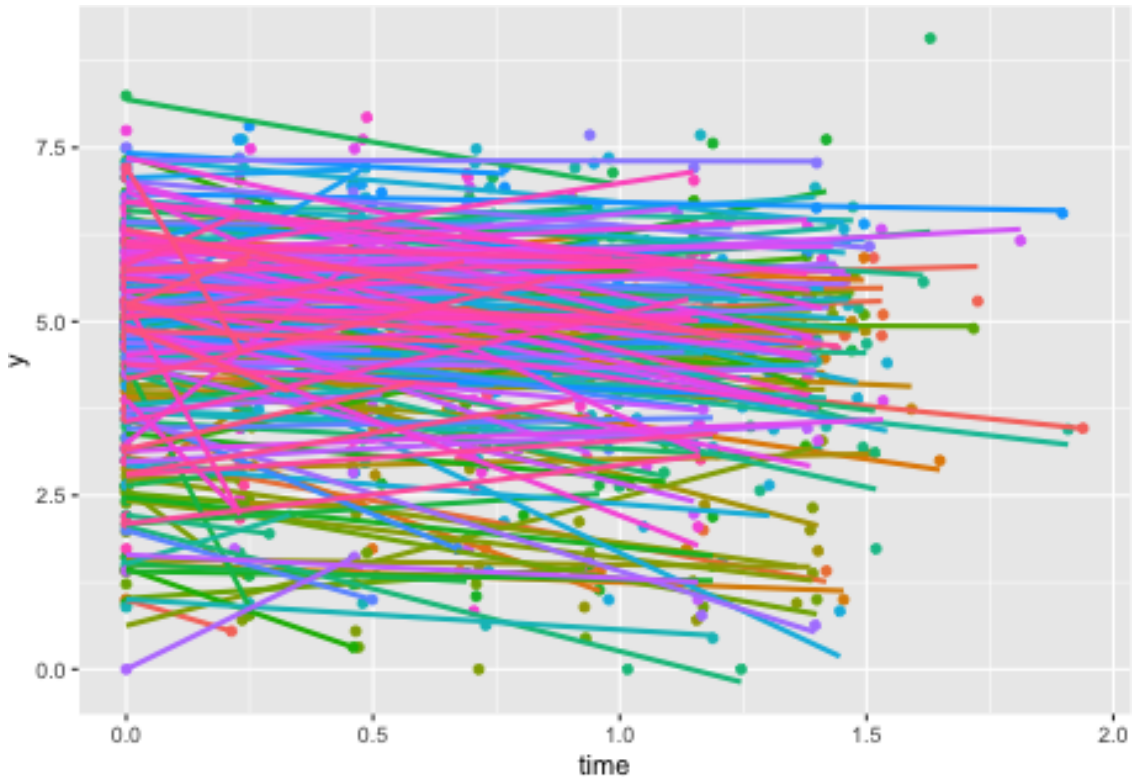


2. Each child's data has a time course that can be summarized by a linear fit. Estimate these lines and plot them for all the children.

```
fit <-stan_glm(y~factor(newpid) -1 + time, data = hiv.data, refresh = 0)
```

```
ggplot(data = hiv.data,aes(x=time,y=y,col=factor(newpid))) +
  geom_point() +
  geom_smooth(se=F,method = "lm", linetype=1) +
  theme(legend.position = "none")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



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.

```
coef <- coef(fit)
coef_matrix <- matrix(0,nrow = length(coef)-1, ncol=5)
colnames(coef_matrix) <- c("newpid","intercept","slope","treatment","age")
newpid <- unique(hiv.data$newpid)
coef_matrix[,1] <- factor(newpid)
coef_matrix[,2] <- coef[-length(coef)]
coef_matrix[,3] <- rep(coef[length(coef)],length(coef)-1)
data <- hiv.data %>%
  group_by(factor(newpid)) %>%
  filter(row_number() == 1) %>%
  ungroup %>%
  dplyr::select(newpid,treatment,age.baseline)
coef <- merge(coef_matrix[,1:3],data,by="newpid")
model_intercept <- lm(intercept~treatment + age.baseline,data=coef)
model_slope <- lm(slope~treatment + age.baseline,data=coef)
summary(model_intercept)
```

```
##
## Call:
```

```
## lm(formula = intercept ~ treatment + age.baseline, data = coef)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0167 -0.7881  0.2775  1.0279  2.9296
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.82985    0.32419  14.898   <2e-16 ***
## treatment    -0.11647    0.18726  -0.622    0.535
## age.baseline  0.03782    0.04082   0.927    0.355
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.467 on 243 degrees of freedom
## Multiple R-squared:  0.005087,   Adjusted R-squared:  -0.003102
## F-statistic: 0.6212 on 2 and 243 DF,  p-value: 0.5382
```

```
summary(model_slope)
```

```
##
## Call:
## lm(formula = slope ~ treatment + age.baseline, data = coef)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3.282e-14 -1.000e-18  2.240e-16  2.520e-16  3.540e-16
##
## Coefficients:
##                Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  -3.852e-01  4.671e-16 -8.246e+14   <2e-16 ***
## treatment     2.586e-16  2.698e-16  9.590e-01    0.339
## age.baseline -1.203e-17  5.880e-17 -2.050e-01    0.838
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.114e-15 on 243 degrees of freedom
## Multiple R-squared:  0.4998, Adjusted R-squared:  0.4956
## F-statistic: 121.4 on 2 and 243 DF,  p-value: < 2.2e-16
```

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.

```
set.seed(1)
fit2 <-lmer(sqrt(CD4PCT)~(1|newpid) + time, data = hiv.data)
summary(fit2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: sqrt(CD4PCT) ~ (1 | newpid) + time
##    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
```

y= 4.76-0.37*time

A 1% increase in time will decrease CD4 percentage by 0.37% on the square root scale

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.

```
set.seed(1)
fit3 <- lmer(sqrt(CD4PCT)~(1|newpid) + time + treatment + age.baseline, data = hiv.data)
summary(fit3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: sqrt(CD4PCT) ~ (1 | newpid) + time + treatment + age.baseline
##    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
```

y=4.91-0.36time+0.18treatment-0.12age.baseline

Time and age have a negative effect on CD4 while treatment has a positive effect. Random effects should be considered when calculkating values.

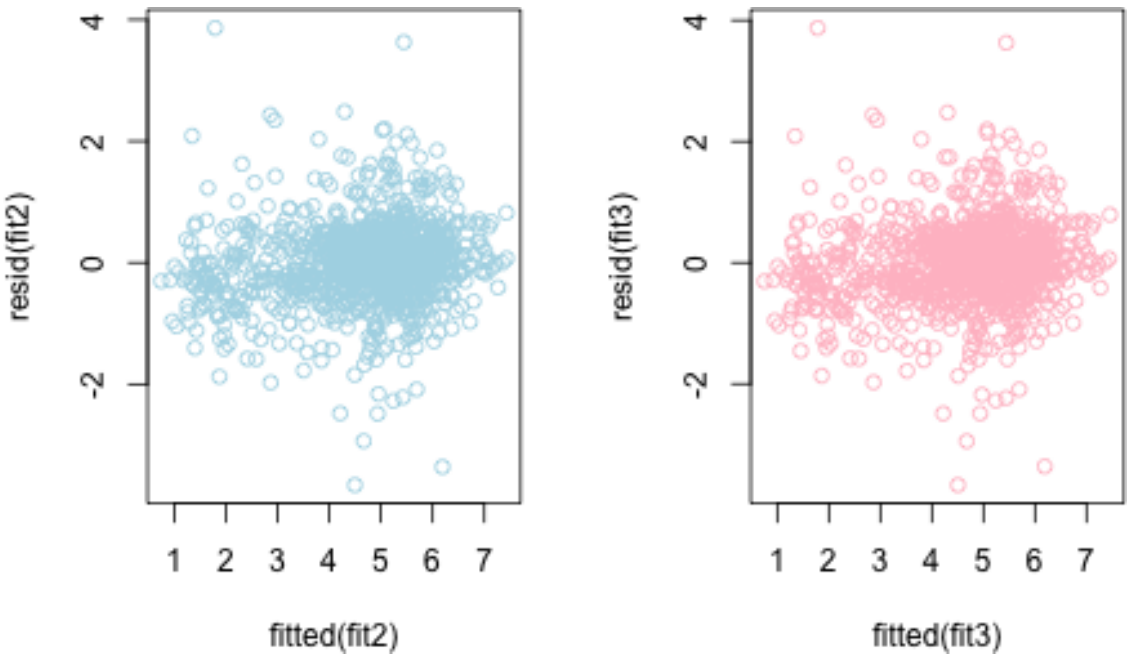
6. Investigate the change in partial pooling from (4) to (5) both graphically and numerically.

```
anova(fit2, fit3)
```

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

```
## Data: hiv.data
## Models:
## fit2: sqrt(CD4PCT) ~ (1 | newpid) + time
## fit3: sqrt(CD4PCT) ~ (1 | newpid) + time + treatment + age.baseline
##      npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fit2    4 3141.9 3161.8 -1566.9   3133.9
## fit3    6 3136.1 3165.9 -1562.0   3124.1 9.7956  2   0.007463 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

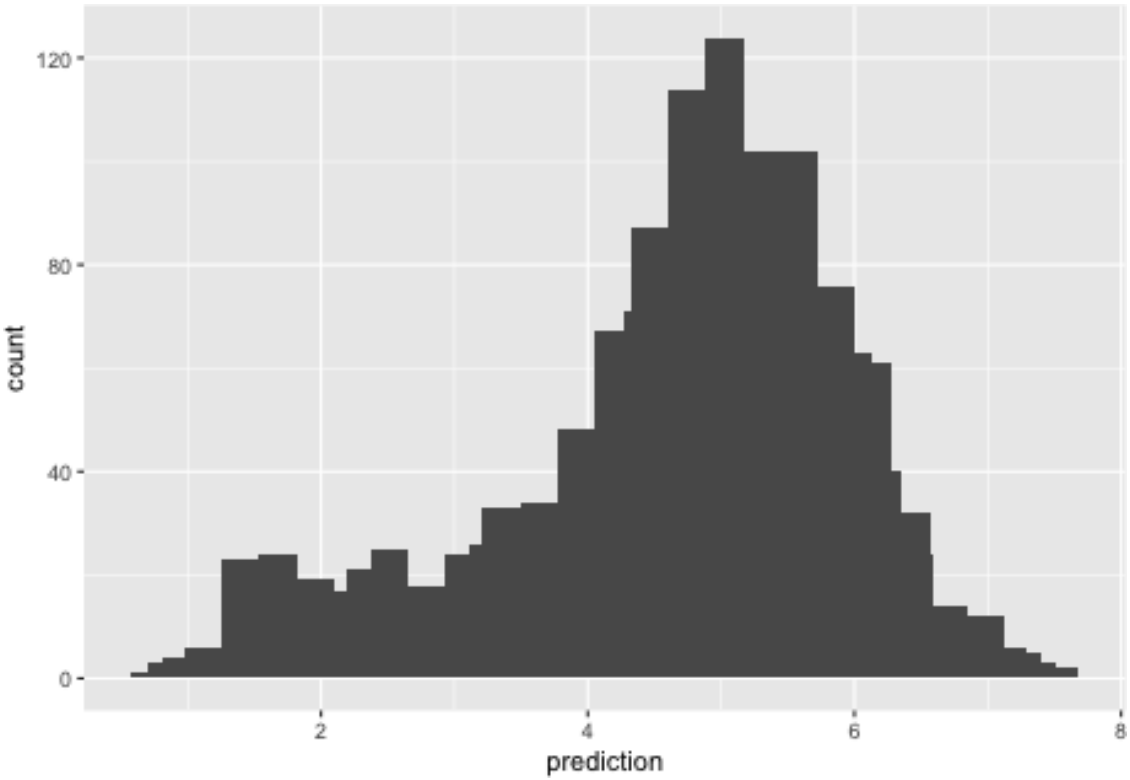
```
par(mfrow=c(1,2))

plot(fitted(fit2),resid(fit2),col="light blue")
plot(fitted(fit3),resid(fit3),col="pink")
```



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.

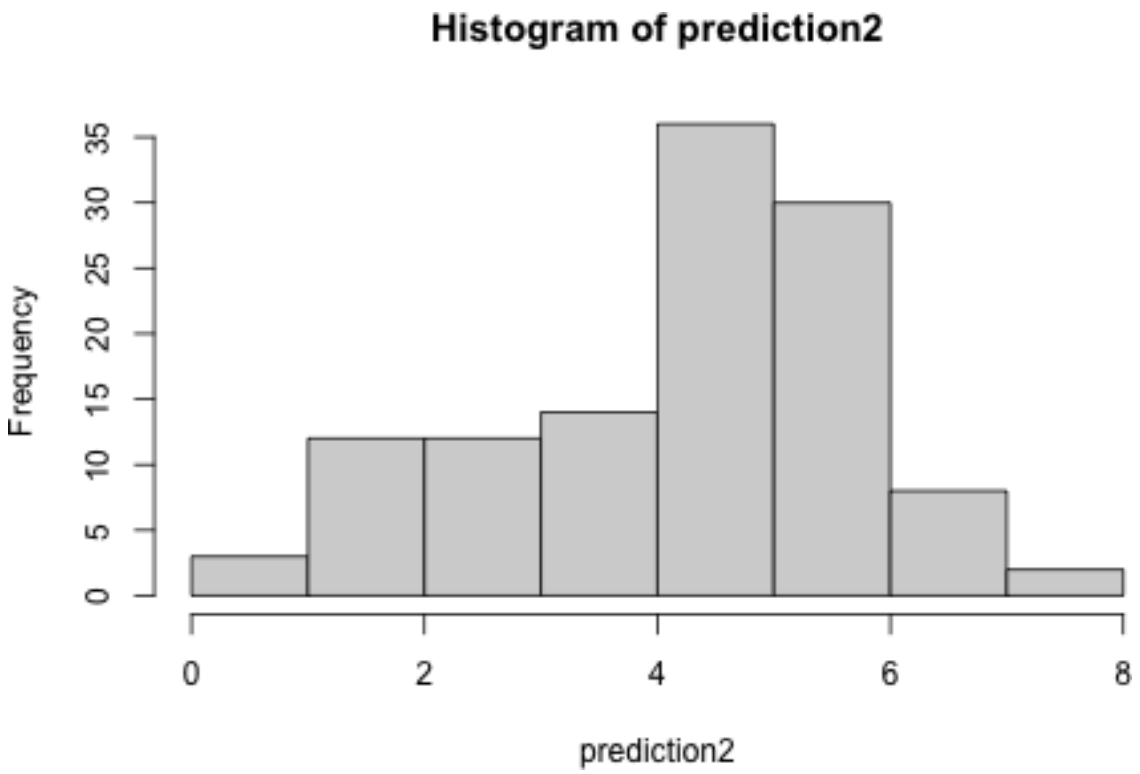
```
actual <- subset(hiv.data, !is.na(hiv.data$treatment) & !is.na(age.baseline))
prediction <- predict(fit3, newdata=actual)
data <- cbind(prediction, actual)
colnames(data)[1] <- c("prediction")
ggplot(data,aes(x=prediction))+geom_histogram() + stat_bin(bins = '25')
```

```
## `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.

```
data2 <- actual[, -c(1, 4, 5, 6, 8)]
data2$age.baseline <- round(data2$age.baseline, digits = 0)
data2 <- data2 %>% filter(age.baseline == 4)
prediction2 <- predict(fit3, newdata=data2)
hist(prediction2)
```



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.

```
fit4<-lmer(y~time+(1+time|newpid), data=hiv.data)
```

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).

```
fit5<-lmer(y ~ factor(time) + treatment + (1|newpid), data=hiv.data)
```

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

```
anova(fit4, fit5)
```

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

```
## Data: hiv.data
## Models:
## fit4: y ~ time + (1 + time | newpid)
## fit5: y ~ factor(time) + treatment + (1 | newpid)
##      npar    AIC    BIC  logLik deviance  Chisq  Df Pr(>Chisq)
## fit4    6 3116.7 3146.6 -1552.4   3104.7
## fit5  406 3245.3 5266.1 -1216.7   2433.3 671.43 400  4.562e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(1,2))
```

```
plot(fitted(fit4),resid(fit4),col="light blue")
plot(fitted(fit5),resid(fit5),col="pink")
```

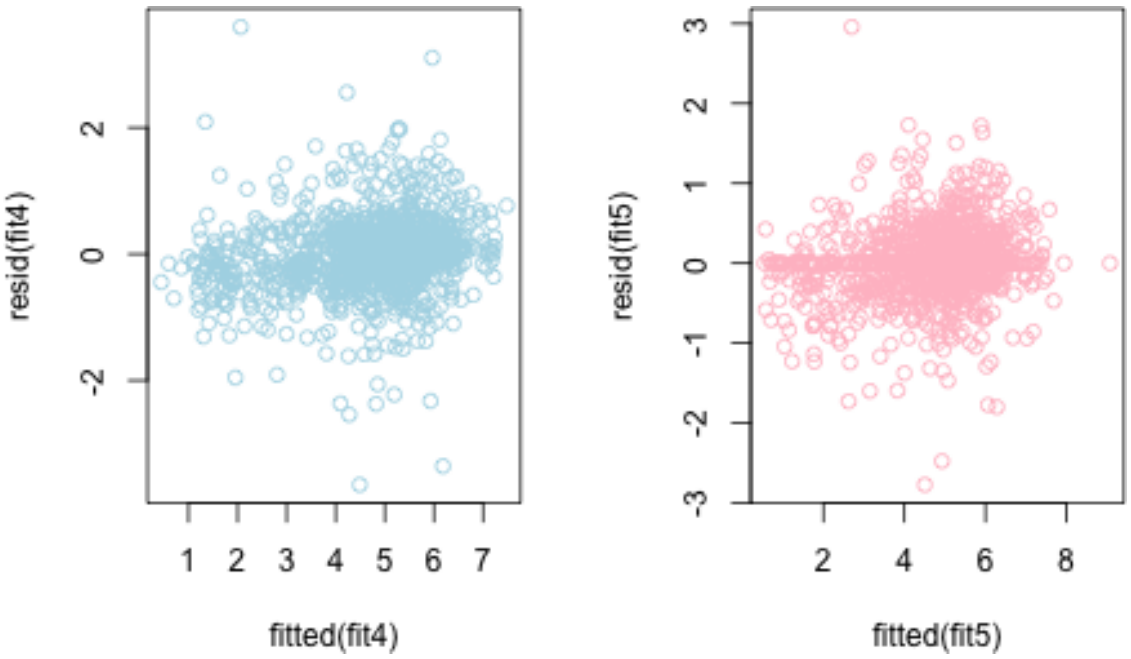


Figure skate in the 1932 Winter Olympics

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).

```
olympics<-read.fwf(filename,widths=c(2,14,9,9,9,9,9,9,9),,skip=21, header = FALSE)
colnames(olympics)<- c("pair", "criterion", "judge_1",  "judge_2",  "judge_3", "judge_4",  "judge_5" ,

olympics<-na.locf(olympics)
olympics$criterion<-str_trim(olympics$criterion)
```

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.
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).