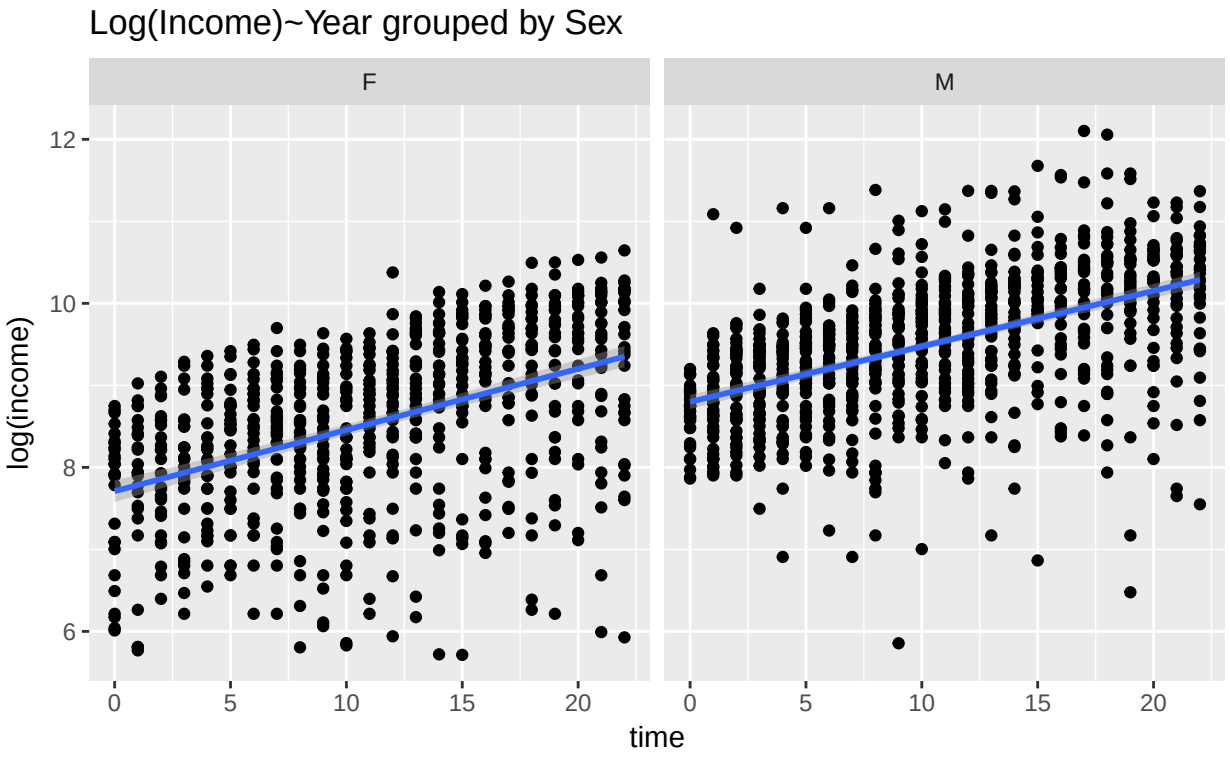
Hw5

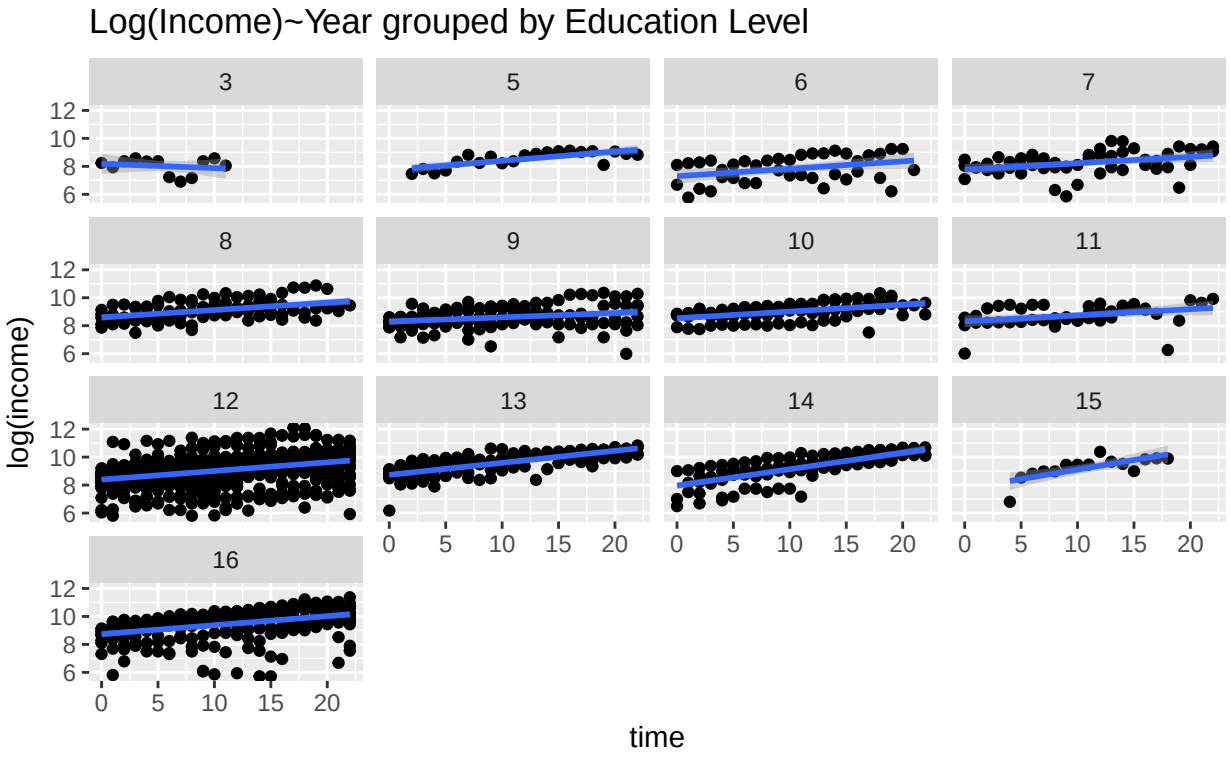
Trevor Freeland

April 18, 2018

42

From some initial data exploration, I thought it would be best to look at income on the log() scale since there was such a large range on incomes. When looking at the relationship between the log(income) and some of the explanatory variables, I saw that there appears to be different intercepts and possibly different slopes for income~time based on sex. There also looks like age in 1968 and education could be related to different slopes and intercepts for the income~time relationship. There could definitely be interactions between time and some of the other explanatory variables like age, gender and education.





43

After running the unconditional means model I wanted to investigate the intraclass correlation. The computed intraclass correlation was .522. This indicates that about 52% of the variation in the data can be explained by grouping individuals responses together. This left about 48% of the variation in the data up to random noise with just the unconditional means model.

We then ran a unconditional growth model. Our unconditional growth model decreased our residual variation by about 47% ((.52 - .27)/(.52)). With the unconditional growth model, our residual variance now only accounts for about 2% of the variation in our data.

```
income.lmerM <- lmer(log(income)~(1|person), data = income)
summary(income.lmerM, cor = F)
mean.resid <- .5145
intraclass.cor <- .5612/(.5145 + .5612)
income.lmerG <- lmer(log(income)~time + (time|person), data = income)
summary(income.lmerG, cor = F)
growth.resid <- .2708
decrease.in.resid <- (mean.resid - growth.resid) / mean.resid
growth.resid.percent <- .271 / (.217+.0021 + 12.9)
```

44

To determine which random effects we are going to keep I added all of the fixed effects into the model by themselves. Then after determining which random effects we need, I will narrow down any of the fixed effects and possibly add interactions where needed.

$log(Income)_{ij} = a_i + b_i(time) + \epsilon_{ij}$

$a_i = \alpha_0 + \alpha_1(age) + \alpha_2(educ) + \alpha_3(male) + \mu_i$ $b_i = \beta_0 + \beta_1(age) + \beta_2(educ) + \beta_3(male) + v_i$

Composite $= log(Income)_{ij} = \alpha_0 + \alpha_1(age) + \alpha_2(educ) + \alpha_3(male) + \beta_0(time) + \beta_1(age : time) + \beta_2(educ : time) + \beta_3(male : time) + \mu_i + v_i(time) + \epsilon_{ij}$

When running the model with random slope and random intercept against the model with just the random intercept and conducting the likelihood ratio test, we get a test statistic of D = 220.2. When calculating the p-value from a combination of χ^2 distributions we get a p-value of essentially 0, indicating that we need to stick with the model that has the random slope and the random intercept.

```
income.lmer1 <- lmer(log(income)~time*(age+educ+sex) + (1|person), data = income)
income.lmer2 <- lmer(log(income)~time*(age+educ+sex) + (time|person), data = income)
l0 <- logLik(income.lmer1)
l1 <- logLik(income.lmer2)
D <- 2*(l1-l0)
.5*(1-pchisq(D,2)) + .5*(1-pchisq(D,1))
```

```
'log Lik.' 0 (df=12)
```

45

Now that we have decided which random effects to keep we need to determine what fixed effects are neccesary. Judging on the summary table of our model, we might be able to get rid of the age variable. Let's test this. Making sure to set REML=FALSE for both models, we conducted an anova test to determine if the model without age would suffice. Our anova gave us a p-value of .73, indicating that we do not lose sufficient information by dropping the age variable so we will use the model without age. From this models summary it looks like we might be able to drop the interactions between time and education or time and sex, so let's test that next. Running an anova test against dropping the interactions between time and education and time and sex we get p-values around .06 if we drop one of the interaction terms. If we drop both interaction terms we get a p-value of .03. This indicates we definitely want one of the interaction terms, but since the model doesn't appear to favor dropping one or the other very much, I think it would be best to keep the interactions in the model, especially since dropping them from the model gives a p-value of .06, which is right on the border of statistical significance.

With this in mind, our final model is:

$$log(Income)_{ij} = \alpha_0 + \alpha_1(educ) + \alpha_2(male) + \beta_0(time) + \beta_1(educ : time) + \beta_2(male : time) + \mu_i + v_i(time) + \epsilon_{ij}$$

Some Interpretations:

sexM = 1.23 -> Holding everything else constant, on average a male in 1968 had 123% more income than a female.

educ = .07 -> Holding everything else constant, on average, an increase in education by 1 level was associated with a 7% increase in income.

time:sexM = -.02 -> For every additional year, a man with similar education to a woman is expected to have their income increase by 2% less than a woman with the same education levels.

```
income.lmer2 <- lmer(log(income)~time*(age+educ+sex) + (time|person), data = income, REML = F)
summary(income.lmer2, cor = F)
income.lmer3 <- lmer(log(income)~time*(educ+sex) + (time|person), data = income, REML = F)
anova(income.lmer2, income.lmer3)
summary(income.lmer3, cor = F)
income.lmer4 <- lmer(log(income)~time * educ + sex  + (time|person), data = income, REML = F)
income.lmer5 <- lmer(log(income)~time * sex + educ  + (time|person), data = income, REML = F)
```

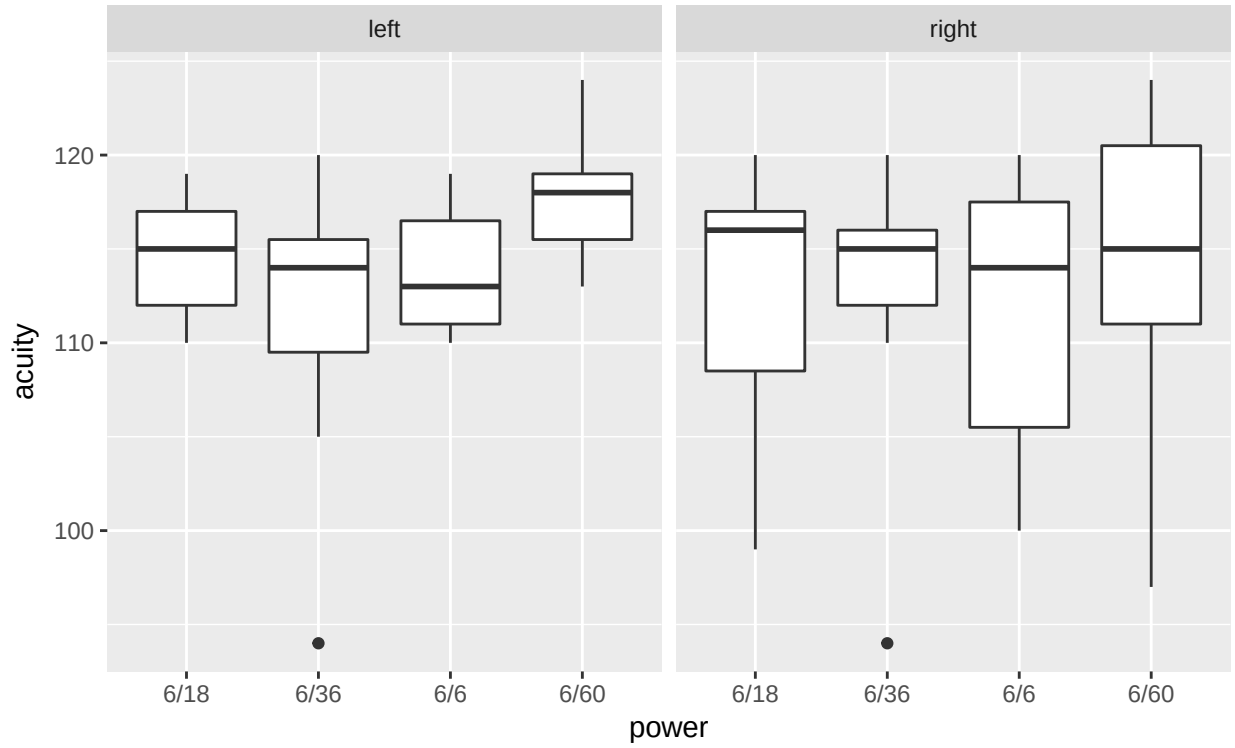
```
anova(income.lmer3, income.lmer5)
anova(income.lmer3, income.lmer4)
fixef(income.lmer3)
```

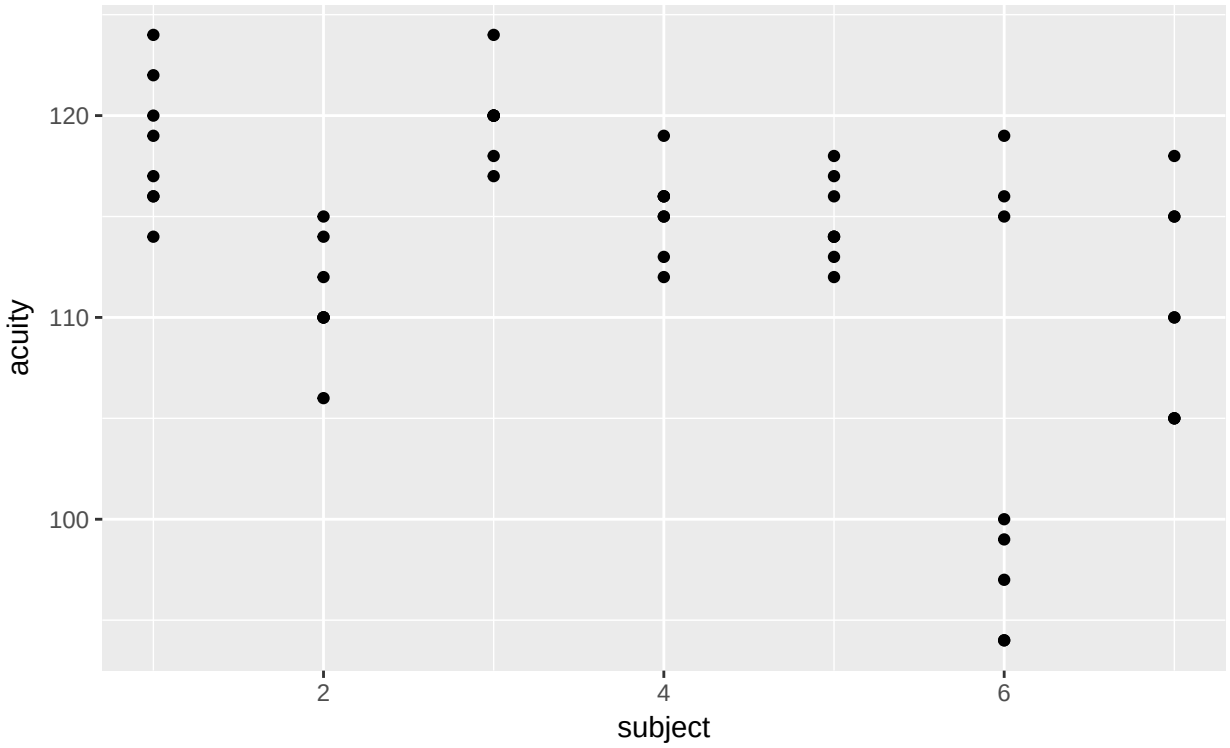
46

In our graphs below we can see the signs of some trends for our data. In our first graph, we look at the effects of power on acuity, and grouped by which eye was being tested. In this graph we can see that different power levels have different levels of acuity, but we can also see that over the left and right eye appear to have relatively the same means, but the variability in the right eye seems to be larger than that in the left. We will take note of that for our future models.

Our second graph is grouping acuity based on each subject/person. We can see that each subject appears to have different means and possibly different levels of variability. This means that we should most likely use a random intercepts model to best fit our data.

Our unconditional means model gives us an intraclass correlation of 51%, which means that about 51% of the variation in our data can be attributed to the difference between subjects and the other 49% of the variation, with an uncondintional means model, is attributed to random noise.






```
vision.lmer <- lmer(acuity~(1|subject), data = vision)
summary(vision.lmer, cor = F)
class.cor2 <-25.7/(25.7 + 24.33)
```

47

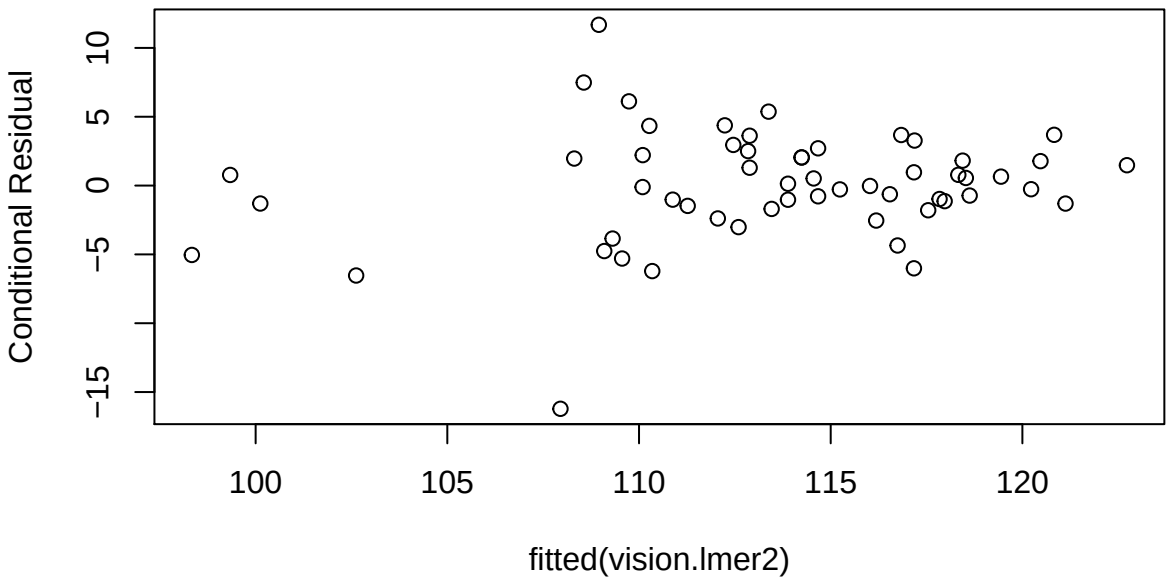
When testing for whether or not we needed to add eye nested in subject I used the log likelihood method. When doing this method we get p-value of .007 which implies that we need eye nested in subject. So our final model looks like:

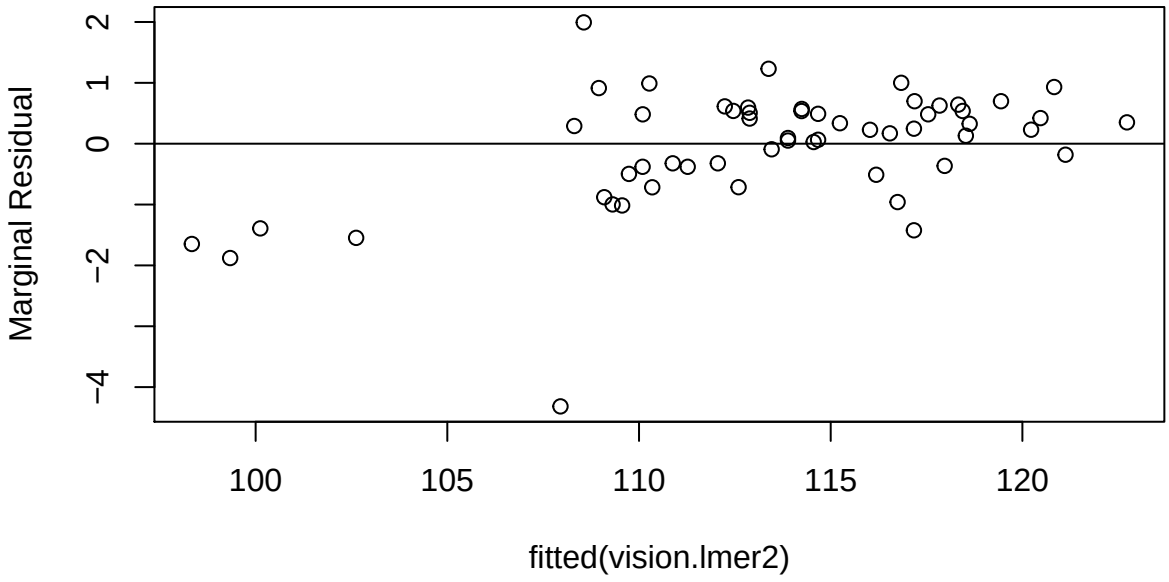
$$Acuity_{ijk} = \alpha_0 + \alpha_1(Power6/36) + \alpha_2(power6/6) + \alpha_3(power6/60) + \mu_i + \mu_{ij} + \epsilon_{ijk}$$

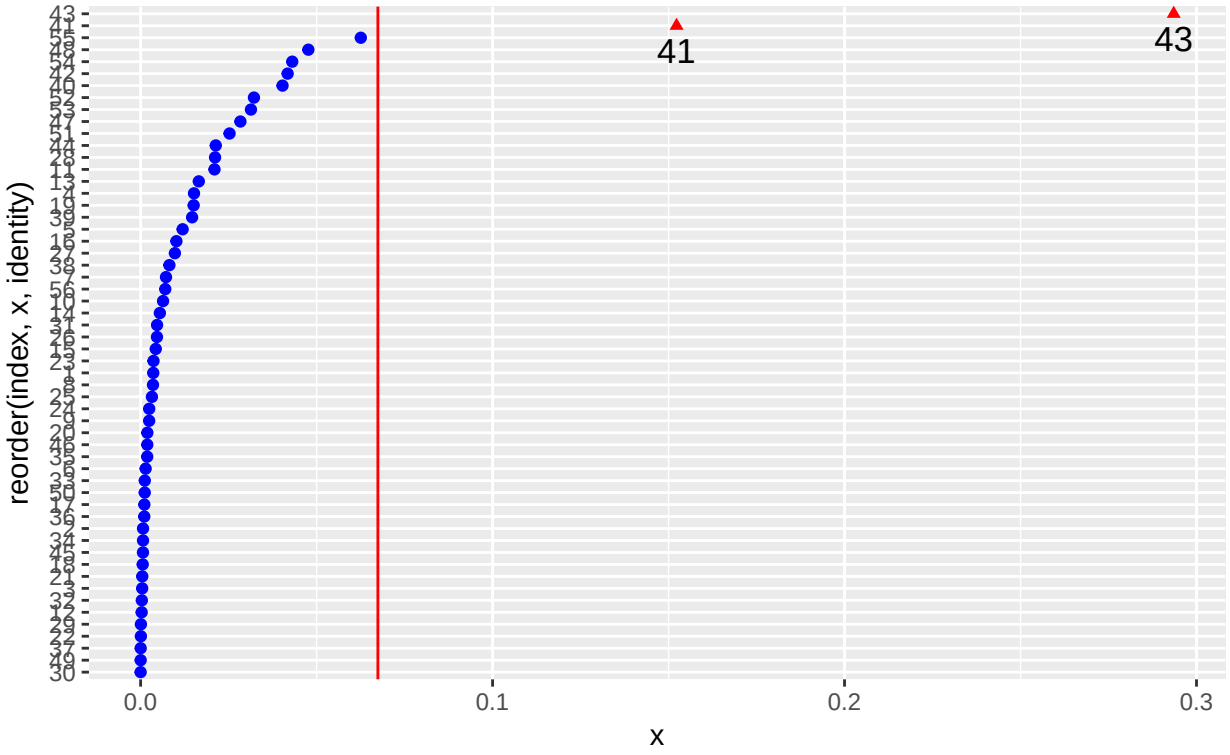
```
vision.lmer1 <- lmer(acuity~power + (1|subject), data = vision)
summary(vision.lmer1, cor = F)
vision.lmer2 <- lmer(acuity~power + (1|subject/eye), data = vision)
summary(vision.lmer2, cor = F)
l0 <- logLik(vision.lmer1)
l1 <- logLik(vision.lmer2)
D <- 2*as.numeric(l1-l0)
.5*(1-pchisq(D,1))
```

48

In the residuals plots below and the cooks distance plots there appear to be a few outliers. When I refit the model without the few outliers, nothing changed significance. The model estimates did not change significantly either so I decided to leave them in and leave my model as it was in #47.


Conditional Residual
10
5
0
−5
−15
100
105
110
115
120
fitted(vision.lmer2)


Marginal Residual
2
1
0
−2
−4
100
105
110
115
120
fitted(vision.lmer2)



49

The idea of boundary constraints is that there are some basic statistical facts that we cannot violate, for example the variation of a term cannot be negative. This means that when conducting MLE, we could possibly have a combination where the algorithm computes the max likelihood to be when the variance of a paramenter is -2, but since we can't do that, we go to the smallest that the parameter can be, which in this case is 0. This can be a potential issue in multilevel models because the number of parameters needed to be estimated can very quickly get extremely large and out of hand, and this might lead to us running into more and more of the boundary constraints.

This is why it is best to try and make adjustments to our model so that we do not run into this boundary constraints issue.

50

We could remove some of the random effects which is what we later did (removing the random slope). We can also check on the scales of our variables to see if some of our numeric variables are on vastly different scales which could be causing issues in some of the models. These are the two first approaches to the issue of boundary slopes. If neither of these things fixed the problems we could do what we initialy did with model C and assume the correlation is 0, but it is best to try and rescale or re-evaluate the need for random effects before assuming things are independent.

51

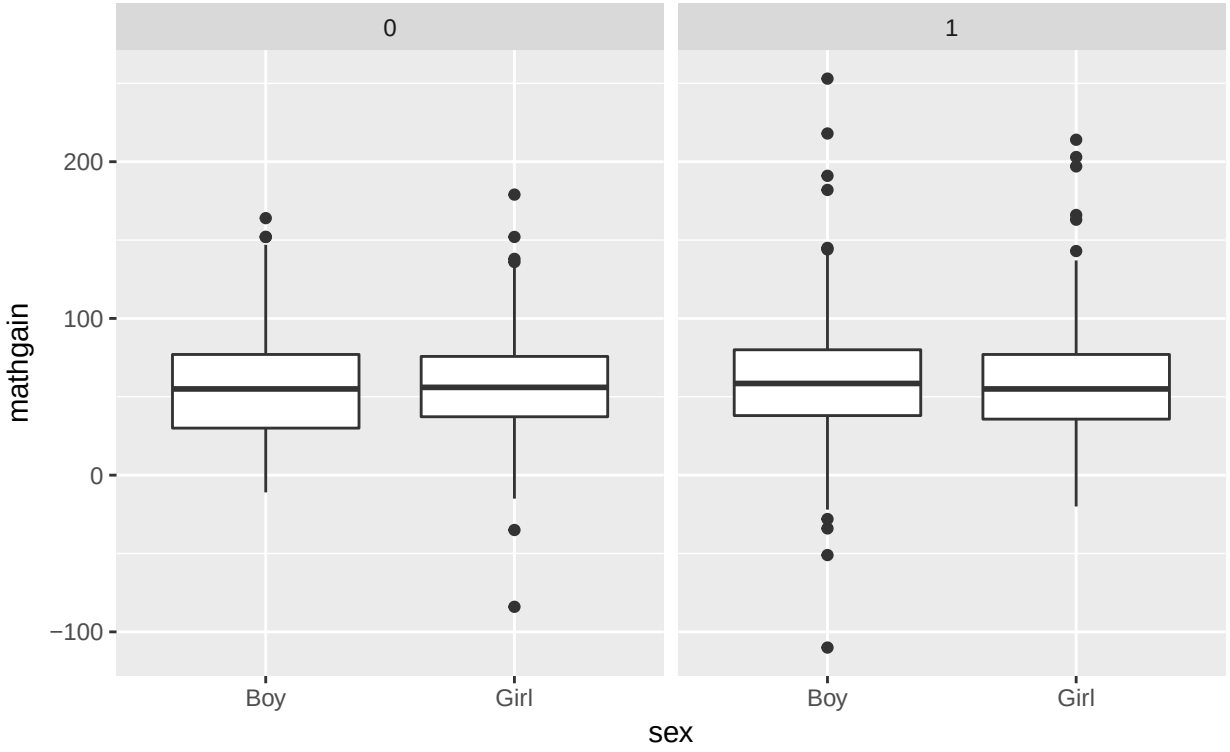
Level 1: Children:, childid, mathgain, gender, minority, ses, mathkind

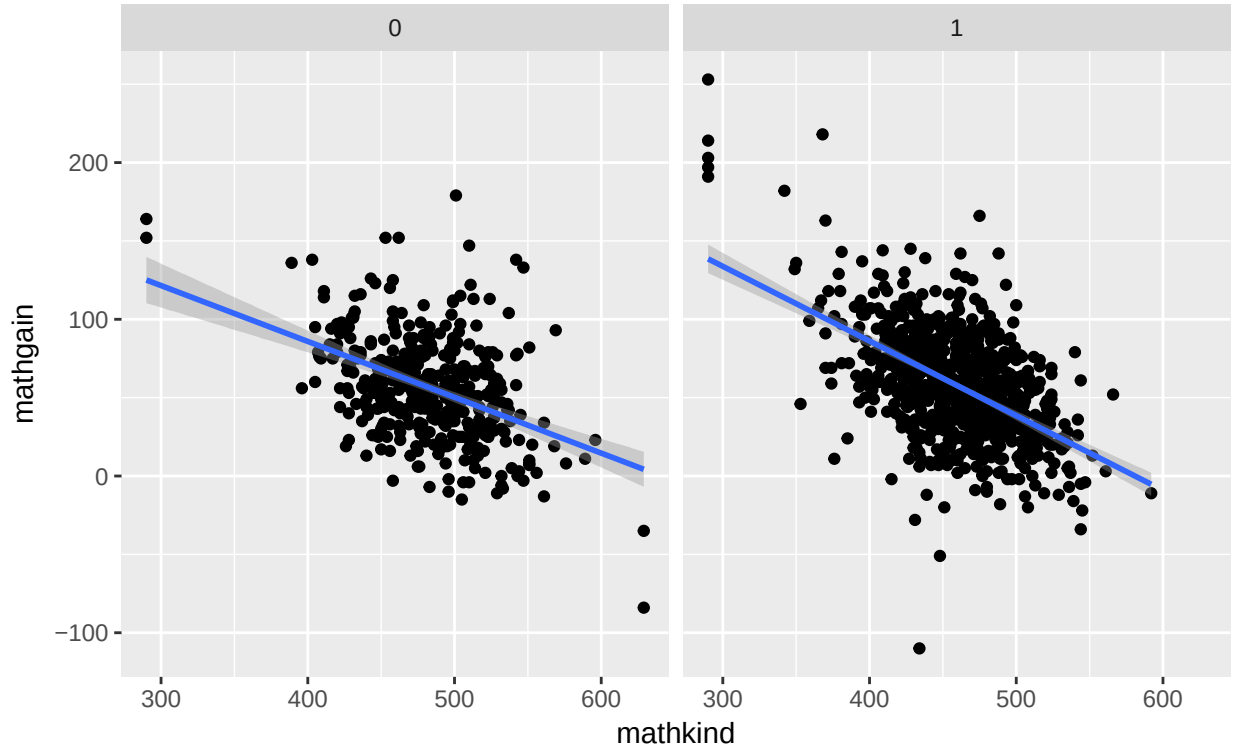
Level 2: Classrooms: classid, teachExp, teachPrep

Level 3: Schools: schoolid, housepov

In the first graph below I plot the mathgain between Boy and Girl while also splitting into minority vs non-minority. There doesn't appear to be a large difference between boy and girl in non-minorities, maybe a slight one between the sexes for minority students but it looks fairly small. I then looked at the Mathkind against mathgain for minorities vs non-minorities. There definitely appears to be different slopes in the line, and also the negative trend here makes sense because though that scored high on math grades in Kindergarden have smaller room for imporvement compared to a student who scored very poorly on the first test.

When I ran my unconditional means model we can see that the vast majority of the variation is at the student level, not at the school or class level. In fact, about 85% of the variation in our data is at the student level compared to the class or the school level.






```
class.lmer <- lmer(mathgain~(1|schoolid/classid), data = classroom)
summary(class.lmer, cor = F)
student.resid <- 1028.23/(1028.23+99.23+77.47)
```

52

Level 1: $Y_{ijk} = a_{ij} + b_{ij}(ses) + c_{ij}(male) + d_{ij}(minorYes) + \epsilon_{ijk}$

Level 2: $a_{ij} = a_i + \mu_{ij}$

$b_{ij} = b_i$

$c_{ij} = c_i$

$d_{ij} = d_i$

Level 3: $a_i = \alpha_0 + \mu_i$

$b_i = \beta_0 + v_i$

$c_i = \gamma_0 + c_i$

$d_i = \phi_0 + w_i$

For this 3 level model we would have to estimate 16 different paramaters. We have the 4 random effects at the school level, + the 6 correlations between those 4 random effects, + 1 random intercept at the class level, + 4 fixed effects, + 1 residual standard deviation = 16 total parameters.

Using simulation to test to see if I can get by without the random slopes for ses and gender I had 33% of my simulations have a test statistic larger than our original test statistic, which means we can get rid of the random slopes for gender and ses.

```
set.seed(546378)
classroom$gender <- as.factor(classroom$gender)
classroom$minority <- as.factor(classroom$minority)
class.lmer1 <- lmer(mathgain~ses + gender + minority + (1|schoolid/classid) + (0+ses+gender+minority|sch
class.lmer2 <-lmer(mathgain~ses + gender + minority + (1|schoolid/classid) + (0+minority|schoolid), data

d<- 2*(logLik(class.lmer1) - logLik(class.lmer2))
N <- 100
Dsim <- numeric(N)
nullY <- simulate(class.lmer2, nsim=N)
nullY[1:5, 1:5]

for (i in 1:N){
  print(i)
  null.lmer <- refit(class.lmer2, nullY[,i])
  alt.lmer <- refit(class.lmer1, nullY[,i])
  Dsim[i] <- 2*(logLik(alt.lmer) - logLik(null.lmer))

}
mean(Dsim>d)
```

53

Now that we have our random effects we want to see what fixed effects are neccesary. When looking at the summary table for our model with the random effects and ses, gender and minority in, the t-values for all three fixed effects seem small, so I decided to test to see if we needed any of them. Holding the random effects constant I used the anova test and I got a p value of .61, which implies none of the fixed effects are needed. This leaves us with a final model of:

$$mathgain_{ijk} = \alpha_0 + w_i(minorityNo) + v_i(minorityYes) + \mu_{ij} + \mu_i + \epsilon_{ijk}$$

```
class.lmer1 <- lmer(mathgain~ses + gender + minority + (1|schoolid/classid) + (0+minority|schoolid), dat
class.lmer2 <- lmer(mathgain~ (1|schoolid/classid) + (0+minority|schoolid), data = classroom, REML = F)
anova(class.lmer1, class.lmer2)
summary(class.lmer2, cor = F)
```