Lecture 6 - Mixed effects models

Ahmed Nadeem

2024-10-23

Simulate some predictor variables

We're going to simulate three predictor variables and one random effect (nesting variable). The three predictor variables are all going to affect our response variable. The random effect is just going to mess with this a bit

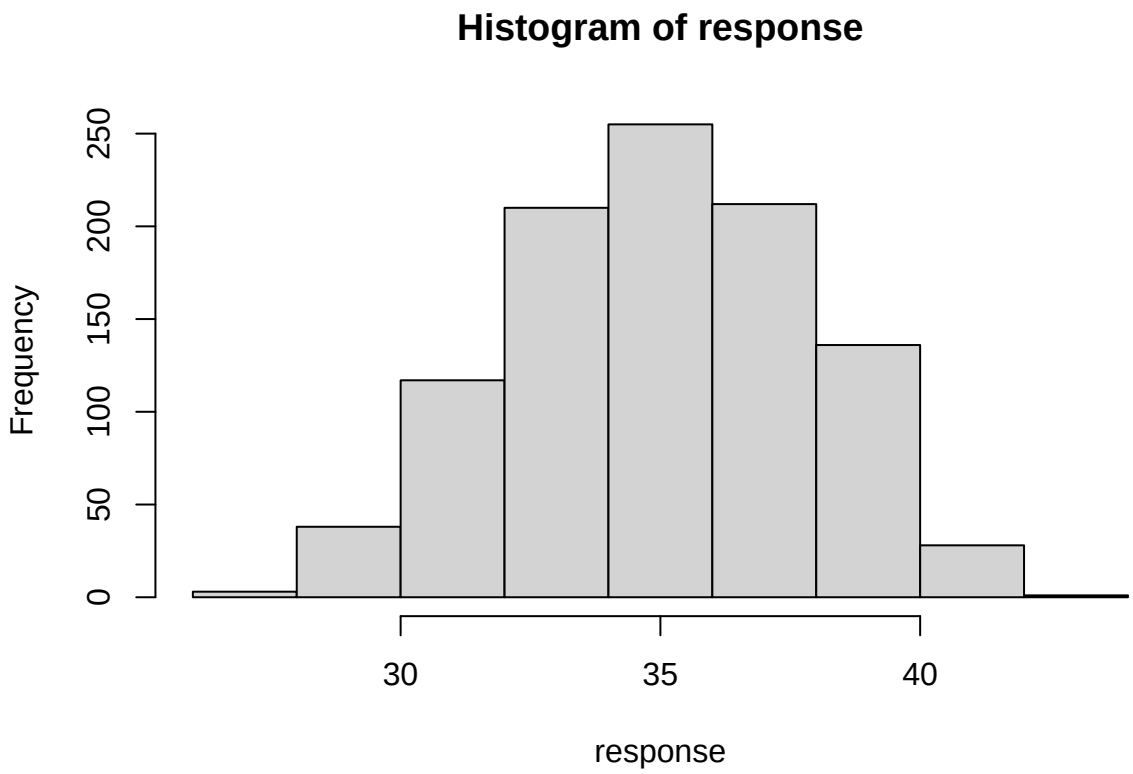
This nesting variable is going to 1) change the intercept for individuals within that group, and 2) change the error depending on who's in that group

```
a<-rnorm(n = 50, mean = 30, sd =2) ### change these numbers, change
###   these terms to something meaningful!
b<-rnorm(50, 20, 4)
c<-rnorm(50, 40, 4)   ## i changed the numbers to be more in line with what id expect with my perceptio
### large mean values (cuz we're measuring distances and large SDs)

g<-sample(1:3, 1000, replace=TRUE)
intercept<-ifelse(g==1, 2, ifelse(g==2, -2, 0))
error<-ifelse(g==1, 0.1, ifelse(g==2, 0.5, 1.5))
g<-as.factor(g)
response<-intercept+0.9*a+0.2*b+0.1*c+rnorm(1000, mean=0, sd=error)
data<-cbind.data.frame(response, a, b, c, g)
```

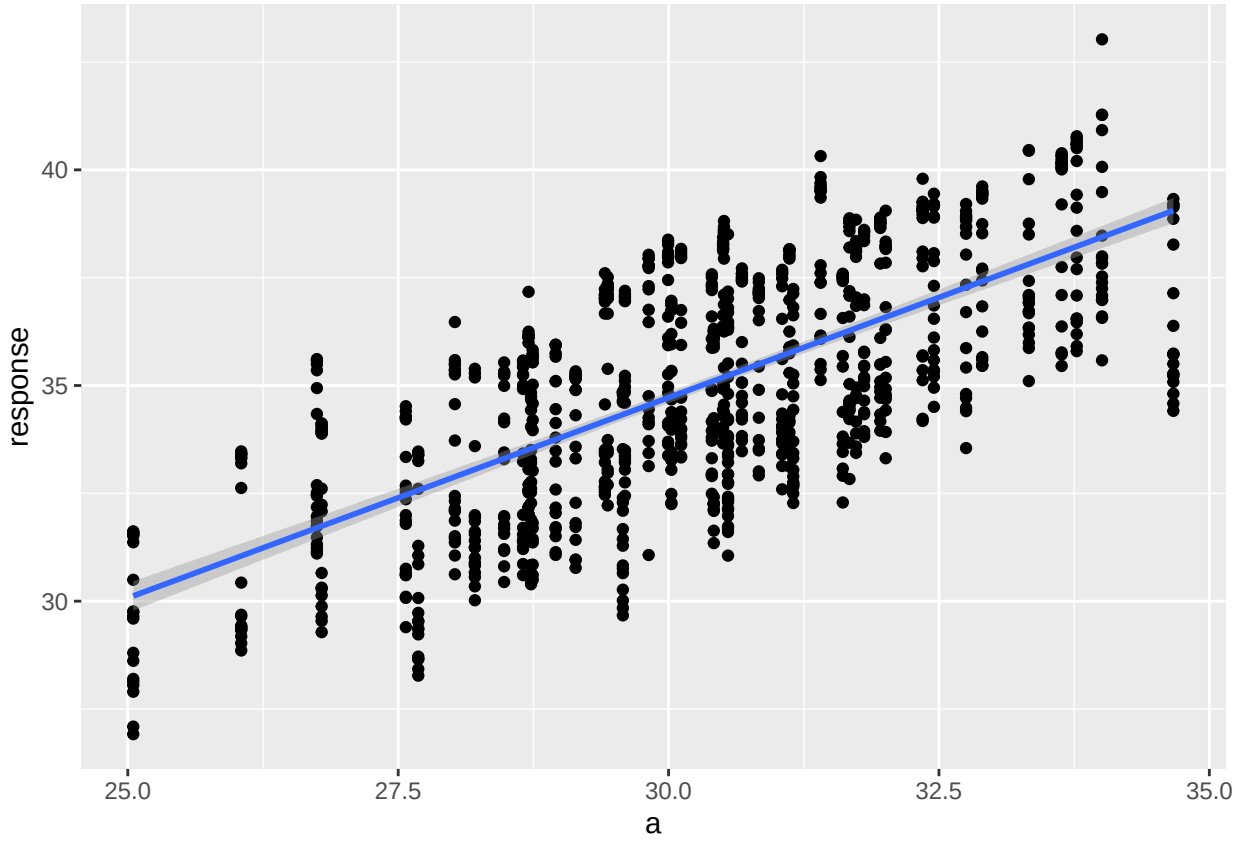
Let's look at the plots of the data, and then the data against (at least one) of the predictor variables. Feel free to plot separately against each predictor variable, or, plot partial residuals (remember Lecture 4)

```
hist(response) ### lovely and normally distributed
```



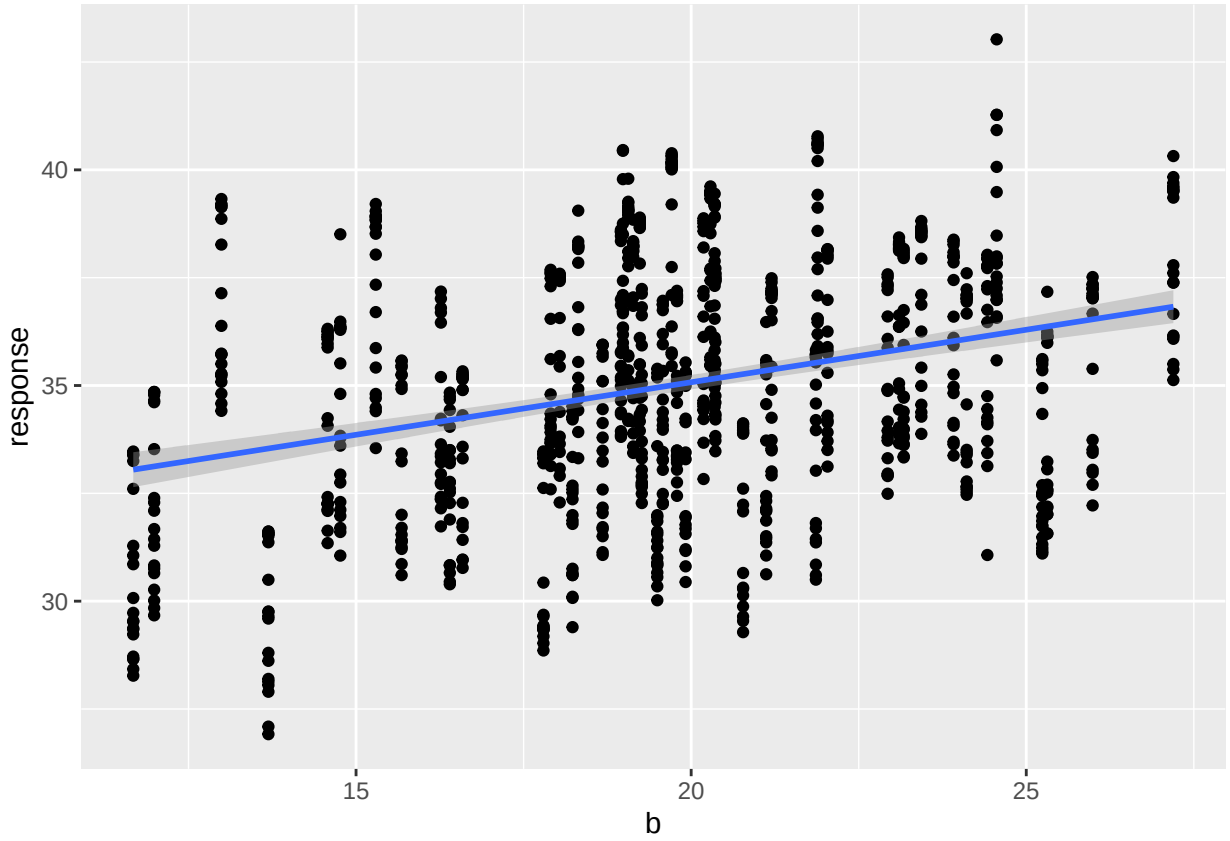

```
ggplot(data, aes(y=response, x=a))+geom_point()+geom_smooth(method=lm)
```

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



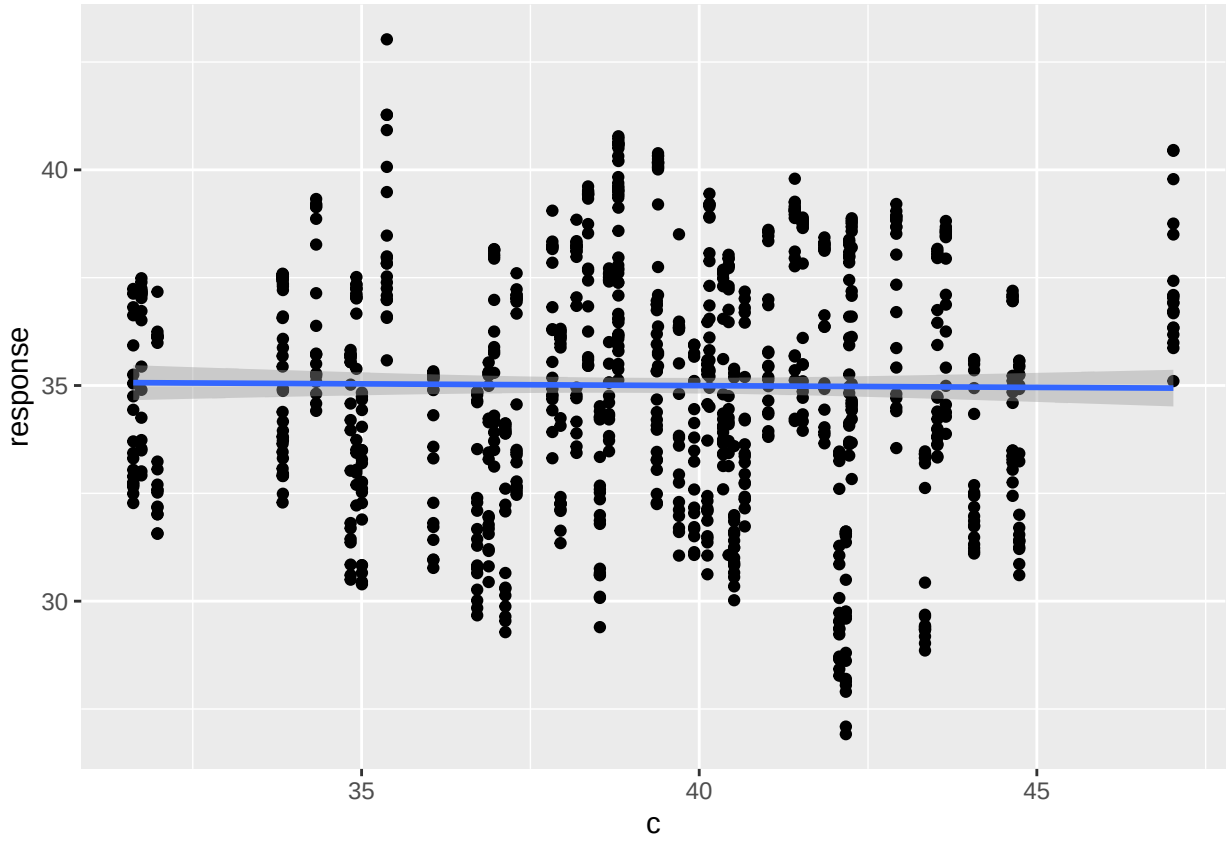

```
ggplot(data, aes(y=response, x=b))+geom_point()+geom_smooth(method=lm)
```

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



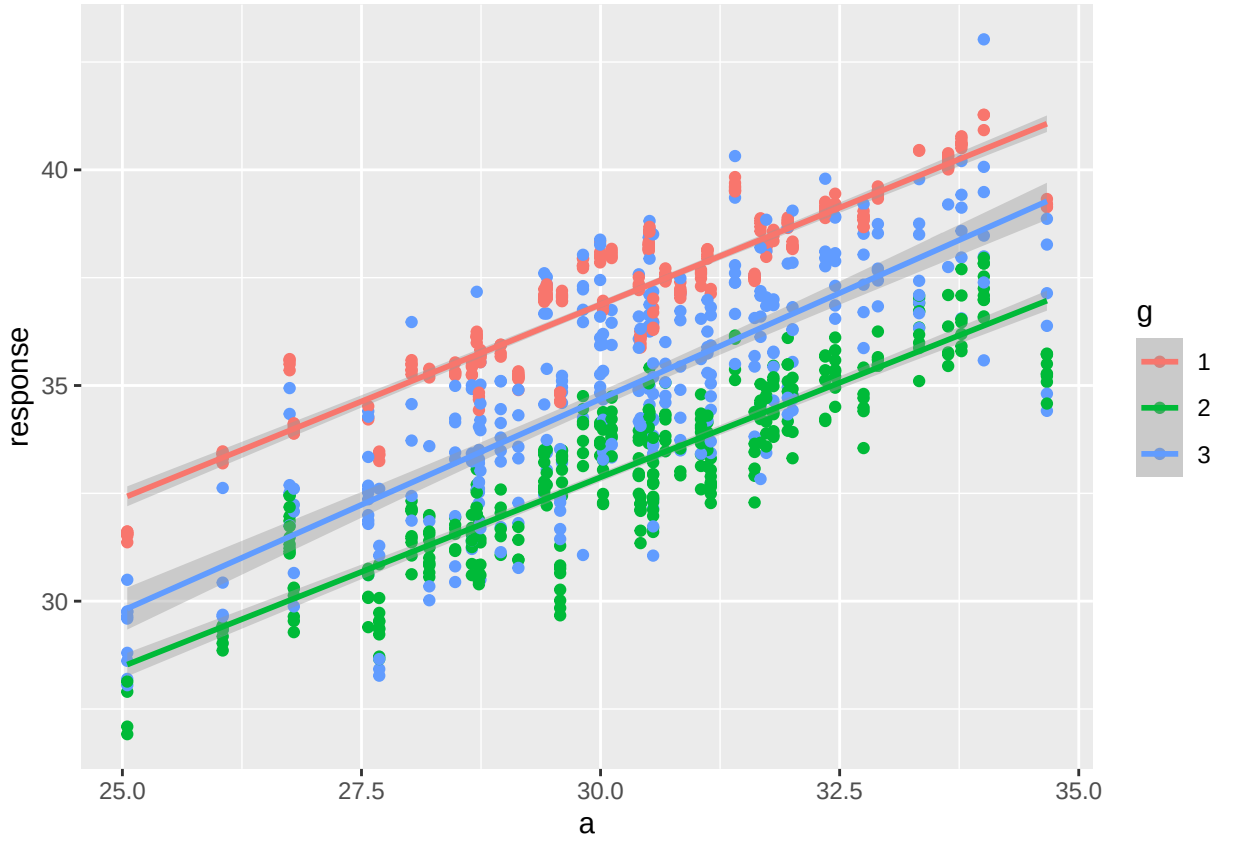

```
ggplot(data, aes(y=response, x=c))+geom_point()+geom_smooth(method=lm)
```

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



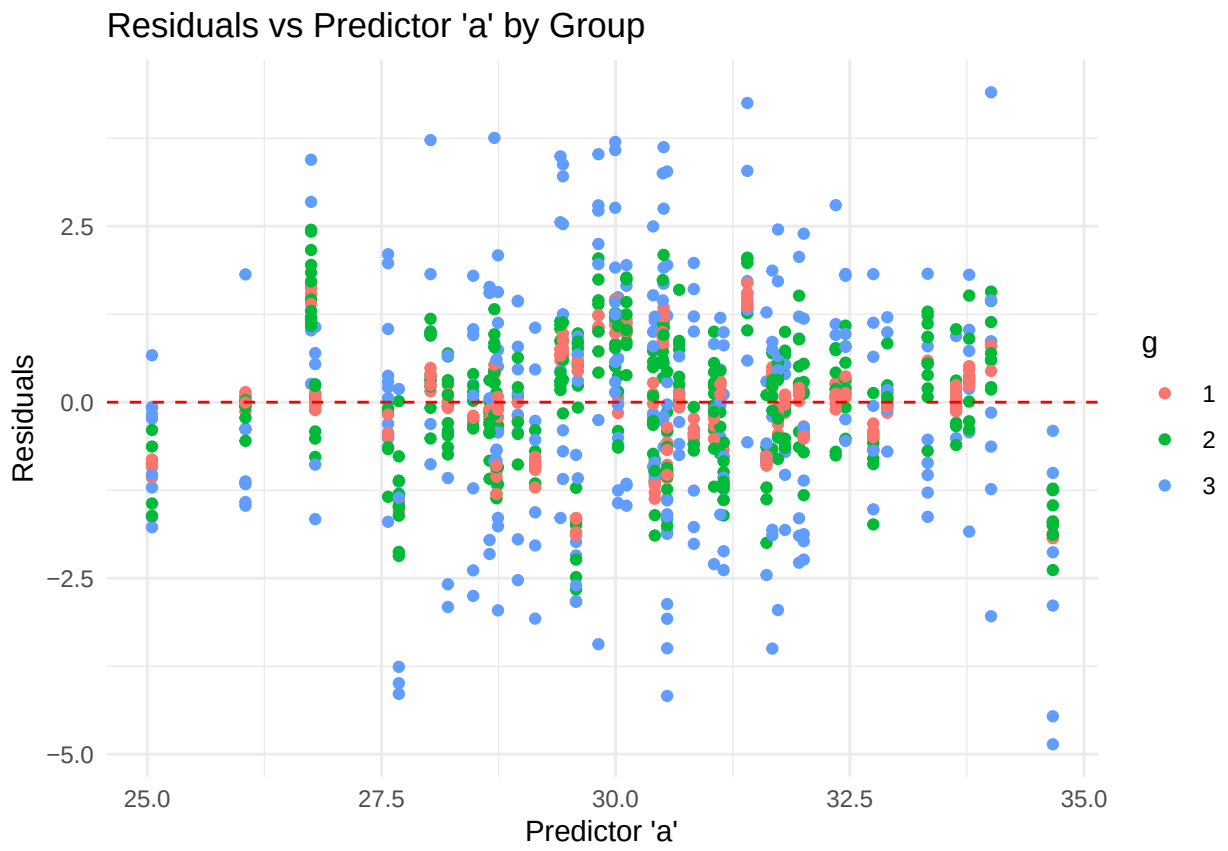

```
ggplot(data, aes(y=response, x=a, colour=g))+geom_point()+geom_smooth(method=lm)
```

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




```
model_grouped <- lm(response ~ a * g, data = data)  # Fit a model with group interaction
data$residuals_grouped <- residuals(model_grouped)

ggplot(data, aes(x = a, y = residuals_grouped, color = g)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals vs Predictor 'a' by Group",
       x = "Predictor 'a'",
       y = "Residuals") +
  theme_minimal()
```



QUESTION: Describe the difference between the two scatter plots. What does adding the different colour lines show you? ANSWER: The differnce between the two scatter plots is that the inital one, which I've made 3 copies of for the 3 different predictor variables, a, b, and c is that those only show that variable's effect on the response. Whereas the grouped colored plot shows the effect of the predictor variable per catergory. Showing me how does variablity influence my predictor's spread/closessness to the line

QUESTION: Describe the residuals in both plots. ANSWER:The residuals in both plots are technically the same since they both look at predictor a, but in the group plot, it's clear that category 1 has lower residuals than b or c, because of the lower variablity from it's normal distribution

Let's do some analyses on this to take a look at the differences.

```
##
## Call:
## lm(formula = response ~ a + b + c, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6165 -1.7264 -0.1033  2.0213  4.3231
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.04724    1.18492  -0.040    0.968
## a            0.92055    0.02841  32.404  < 2e-16 ***
## b            0.22027    0.01643  13.406  < 2e-16 ***
## c            0.07192    0.01669   4.309  1.8e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.864 on 996 degrees of freedom
## Multiple R-squared:  0.5611, Adjusted R-squared:  0.5598
## F-statistic: 424.5 on 3 and 996 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: response
##            Df Sum Sq Mean Sq F value  Pr(>F)
## a           1 3767.1  3767.1 1084.39 < 2e-16 ***
## b           1  592.1   592.1  170.43 < 2e-16 ***
## c           1   64.5    64.5   18.57 1.8e-05 ***
## Residuals 996 3460.1     3.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: response ~ a + b + c + (1 | g)
##    Data: data
##
## REML criterion at convergence: 2674.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.0780 -0.3047  0.0009  0.3231  4.7226
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  g        (Intercept) 3.9796   1.995
##  Residual             0.8172   0.904
## Number of obs: 1000, groups:  g, 3
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  -0.079020   1.287355   3.117452  -0.061    0.955
## a             0.912826   0.013785 994.001324  66.218   <2e-16 ***
## b             0.210325   0.007971 994.000719  26.386   <2e-16 ***
## c             0.085961   0.008100 994.001834  10.612   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##   (Intr) a      b
## a -0.344
## b -0.130 -0.046
## c -0.291  0.102  0.095
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##   Sum Sq Mean Sq NumDF DenDF F value    Pr(>F)
## a 3583.1  3583.1     1   994 4384.77 < 2.2e-16 ***
## b  568.9   568.9     1   994  696.21 < 2.2e-16 ***
## c   92.0    92.0     1   994  112.62 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
## R2 for each model
summary(model1)$adj.r.squared
```

```
## [1] 0.5597933
```

```r
r.squaredGLMM(model2) ### this gives two R2, what are they both?
```

```
##            R2m       R2c
## [1,] 0.4729305 0.9102085
```

```r
##intercepts for each model
 summary(model1)$coef[1,1]
```

```
## [1] -0.04723918
```

```r
 ranef(model2)$g
```

```
##   (Intercept)
## 1  2.03732315
## 2 -1.94823585
## 3 -0.08908731
```

```r
 ### coefficents for each model
model1$coefficients
```

```
## (Intercept)           a           b           c
## -0.04723918  0.92054615  0.22027352  0.07191707
```

```r
summary(model2)$coefficients[, "Estimate"]
```

```
## (Intercept)           a           b           c
## -0.07901964  0.91282642  0.21032510  0.08596125
```

QUESTION: Compare and contrast the R2, intercepts and estimates from each model. Which is closer to what you simulated? Why is this? ANSWER: the mixed effects model is closer to what was simulated, the linear model does a good job (r2 = 0.44) but the linear mixed model with the conditional fixed and random effects is a lot better at modelling the data