Hw4

Trevor Freeland

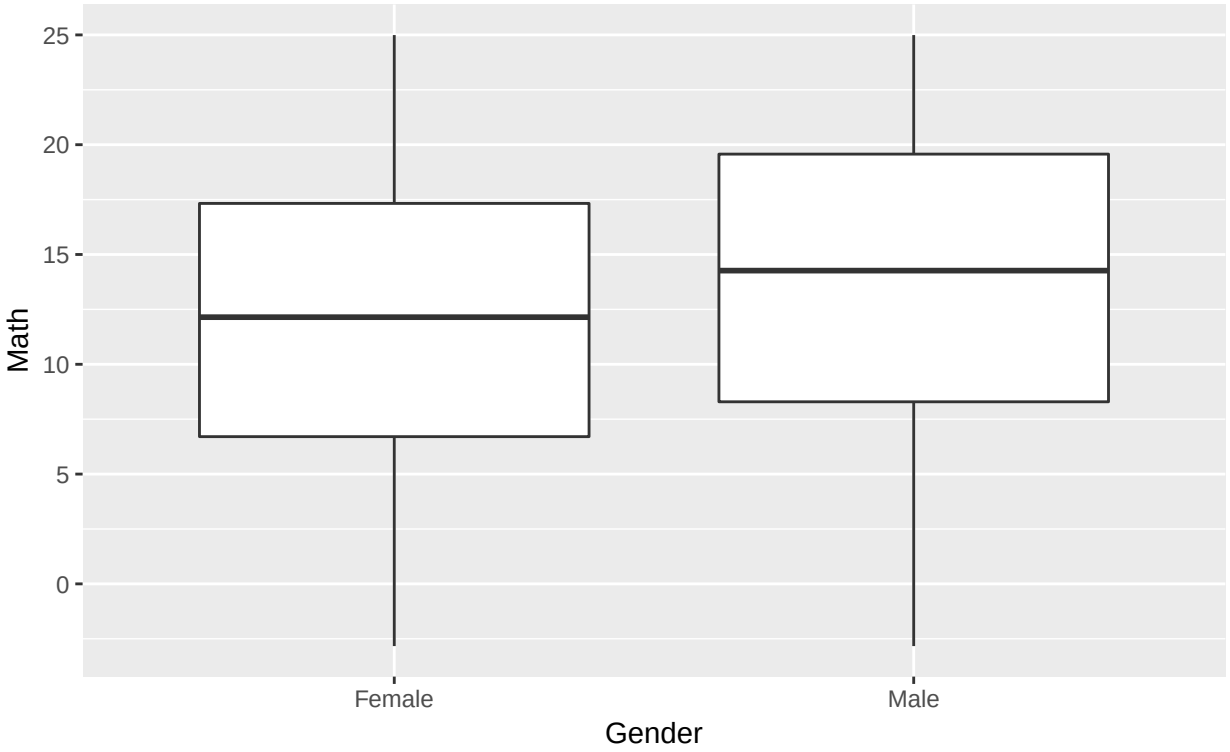
April 13, 2018

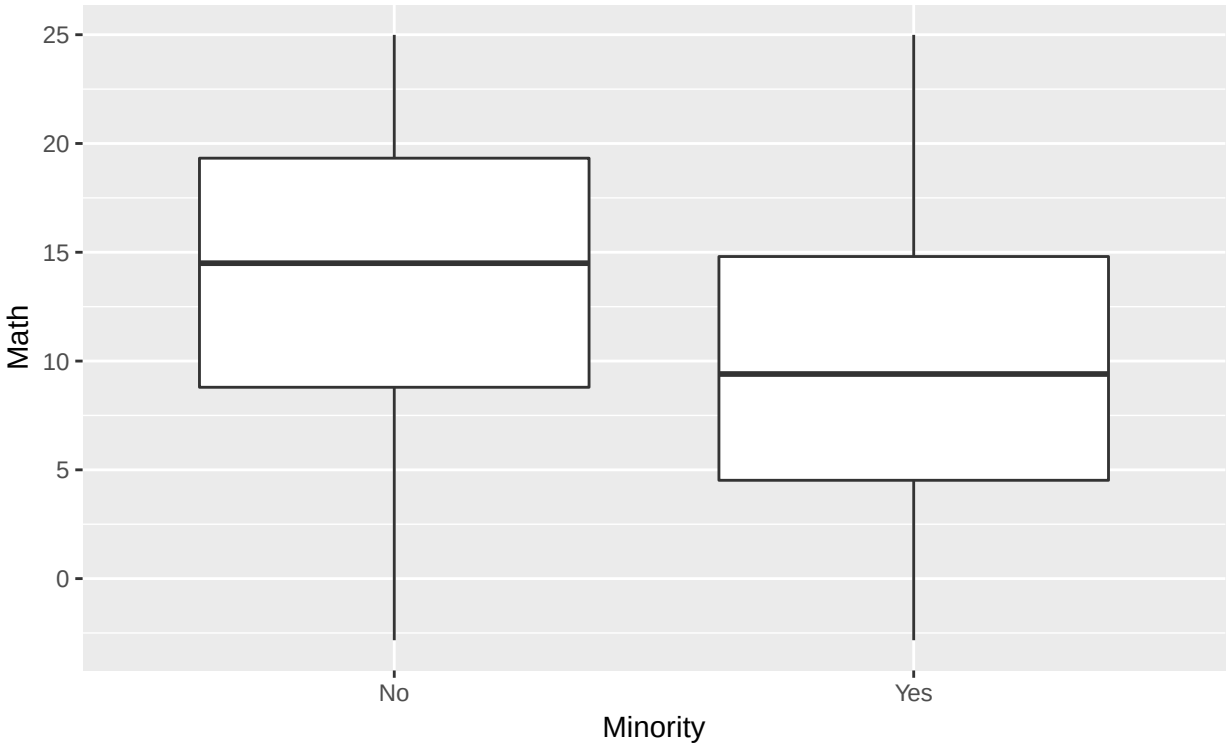
31

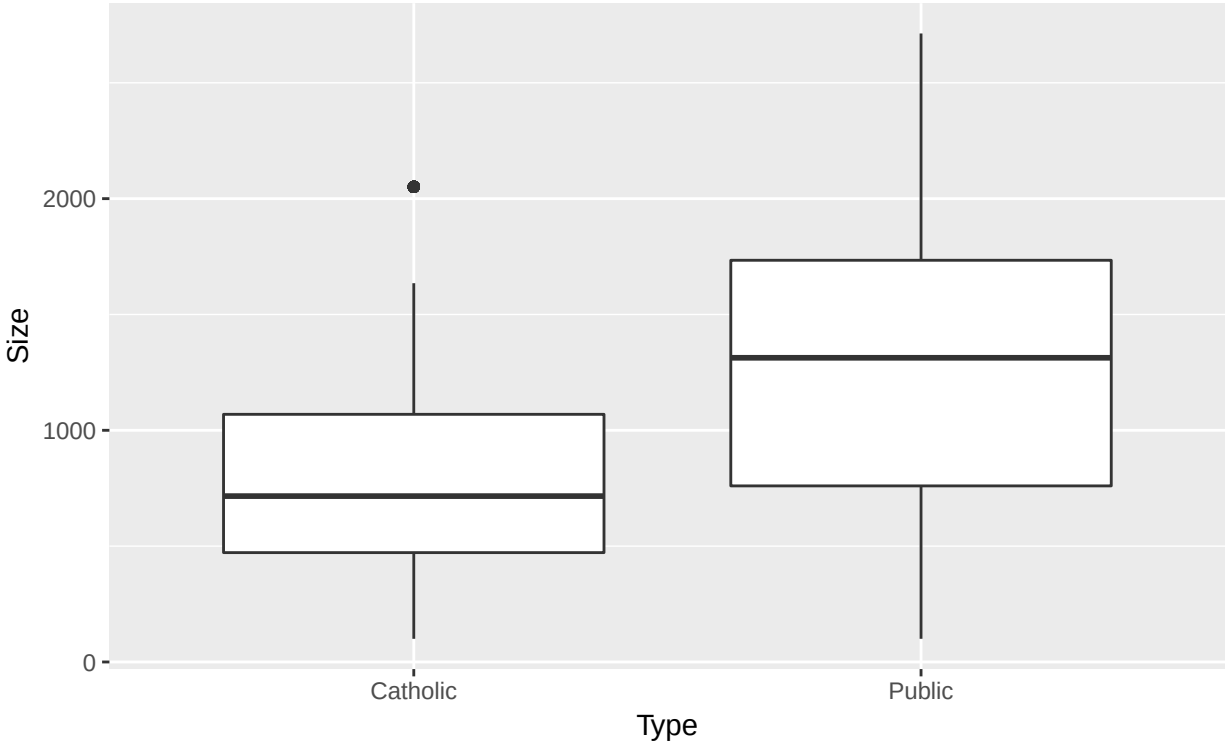
After running the simulations multiple times with different σ and σ_b values it is clear that there is a significant difference between the 3 different P-values. The half and half p-value always lies in between the two full values. While the p-value from the χ^2_q gives on average smaller p-values, and the χ^2_{q+1} gives on average larger p-values. This simulation backs up the concepts that we talked about in class that the combination of both of the distributions seem to be the best fit for doing this kind of simulation testing.

32

Below I have some exploratoy plots and a table with some summary statistics of the data. Through my initial exploratory analysis It appears that Gender may play a small role in the Math scores, it looks like Males on average have about a point or two higher than the the average Female. Whether or not the student was a minority seemed to have a larger effect on the Math scores, with non-minorities averaging about 5 points higher than the average student who is a minority. When looking at some of the School variables I noticed that the Type of school seems to be related to size of the school, with Catholic schools being smaller on average than public schools.







Ave.Size	Ave.Academic	Ave.AveSES	Ave.Math	Ave.SES
1098	0.5139	-0.0001875	12.75	0.0001434

33

(a)

The anova command on the model indicates that the school variable is definitely significant. The anova command gives us a p-value of essentially 0. This all implies that which school students are at definitely has an effect on their Math Scores, which indicates we might need to use a random effects model.

(b)

We get a intraclass corralation of about 20%. This indicates that there is a decent of relationship between students who are all in the same school, again signifying that we might need to use the random effects model.

34

(a)

Level 1: $Y_{ij} = a_i + b_i(Male) + c_i(SES) + d_i(MinorityYes) + \epsilon_{ij}$

Level 2: $a_i = \alpha_0 + \alpha_2(Size) + \alpha_3(Type) + \alpha_4(Academic) + \alpha_5(AveSES) + \mu_i$

$b_i = \beta_0 + w_i$

$c_i = \gamma_0 + v_i$

$d_i = \phi_0 + p_i$

Composite Form

$$Y_{ij} = \alpha_0+\alpha_2(Size)+\alpha_3(Type)+\alpha_4(Academic)+\alpha_5(AveSES)+\beta_0(Male)+\gamma_0(SES)+\phi_0(MinorityYes)+w_i(Male)+v_i(SES)$$

(b)

The model gives us a warning saying that some predictor variables are on very different scales, consider rescaling and we get a model failed to converge: degenerate Hessian with 3 negative eigenvalues warning.

(c)

The range of the numeric variables are very small expect for the range of the size of the schools, which is incredibly large when compared to the other ones. See table below for the specific ranges.

	Min	Max	Range
Size	100.000	2713.000	2613.000
Academic	0.000	1.000	1.000
AveSES	-1.188	0.831	2.019
SES	-3.758	2.692	6.450

(d)

Using the given code in the assignment we rescaled the Size variable.

(e)

After rerunning our model with the scaled version of size it does appear to have resolved the problem since we no longer got the warning about the predictor variables being on very different scales but we still get a warning about Hessian with 1 negative eigenvalues.

35

(a)

Our simulation gave a p-value of .15, indicating that we should keep the reduced model and we won't lose significant information. So we will go with the model that only has the Minority random slope.

(b)

We get a small p-value $< .05$ which indicates that we need to stick with the full model, so we will stay with the model that has the Minority random slope.

36

(a)

It appears that all of the fixed effects are significant if we are assuming t-values > 2.5 to be significant.

(b)

We can use an anova test for this because we are just checking fixed effects. Once we make sure to have REML=F in both of the models an anova test gives us an incredibly small p-value which indicates that we want to keep the full model, which in this case means Model 5 is preferred.

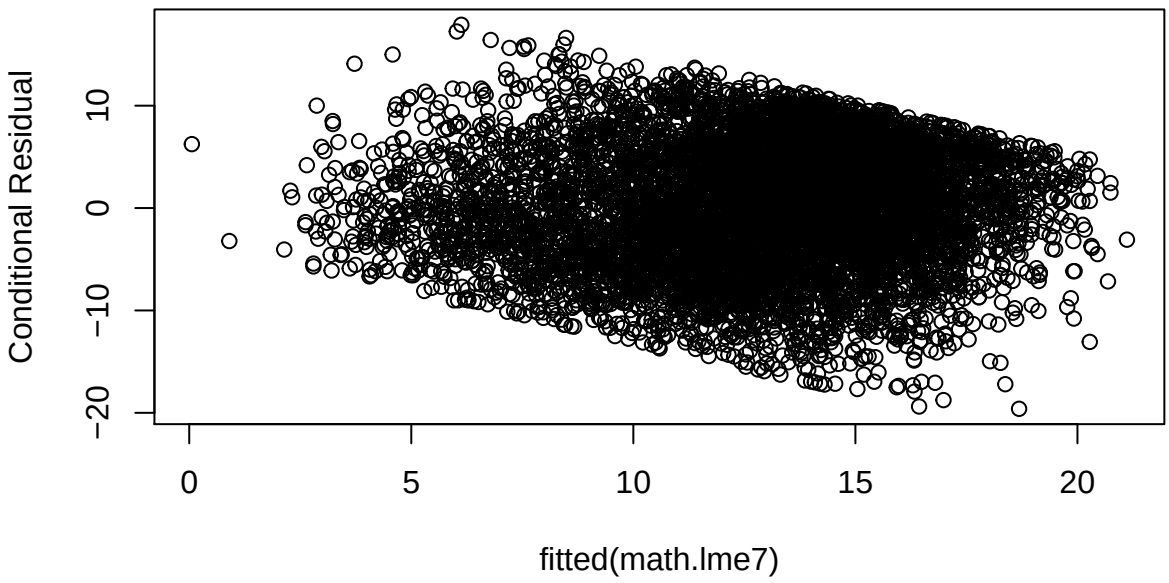
(c)

I was not satisfied with having both of the gender and school type and the minority and school type interactions in the model because it doesn't appear that the gender and school type interaction is significant. I took that interaction out and using the anova command I did not see a reason to keep it in the model so my final model has the interaction with Type and Minority but not Type and Gender.

37

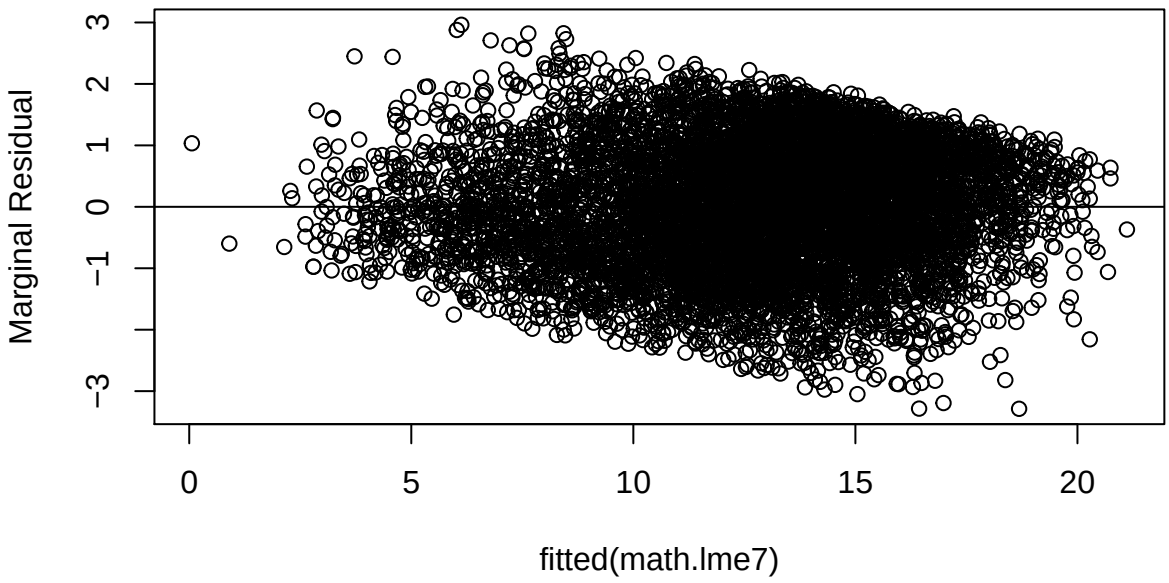
(a)

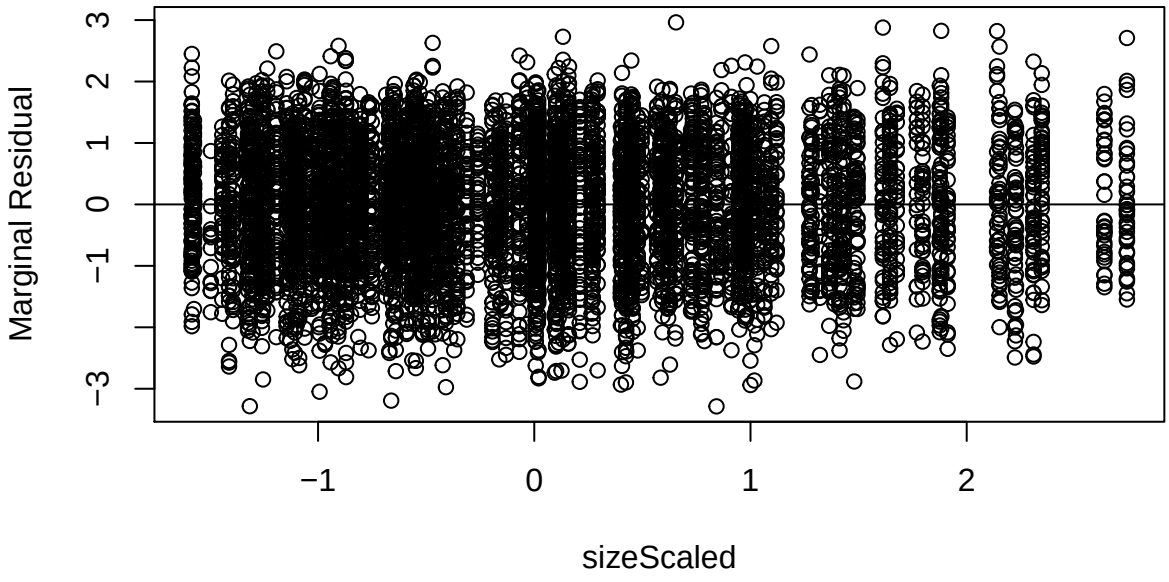
It seems like we have a fairly constant variance based on the plot below.

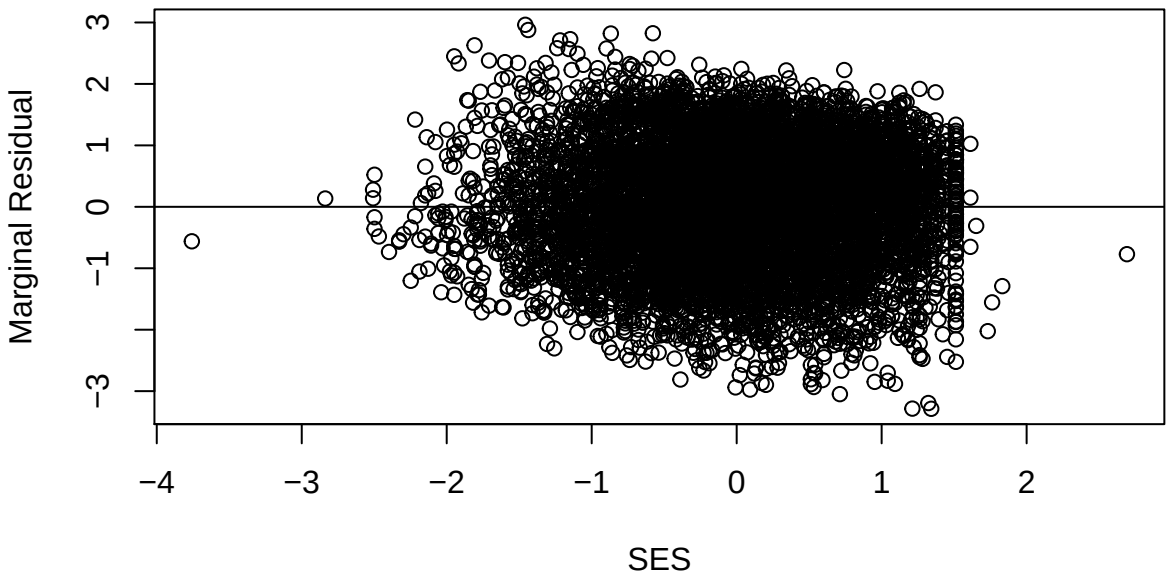


(b)

There doesn't appear to be any curvature or outliers in the plot against the fitted values or against the numeric variables.

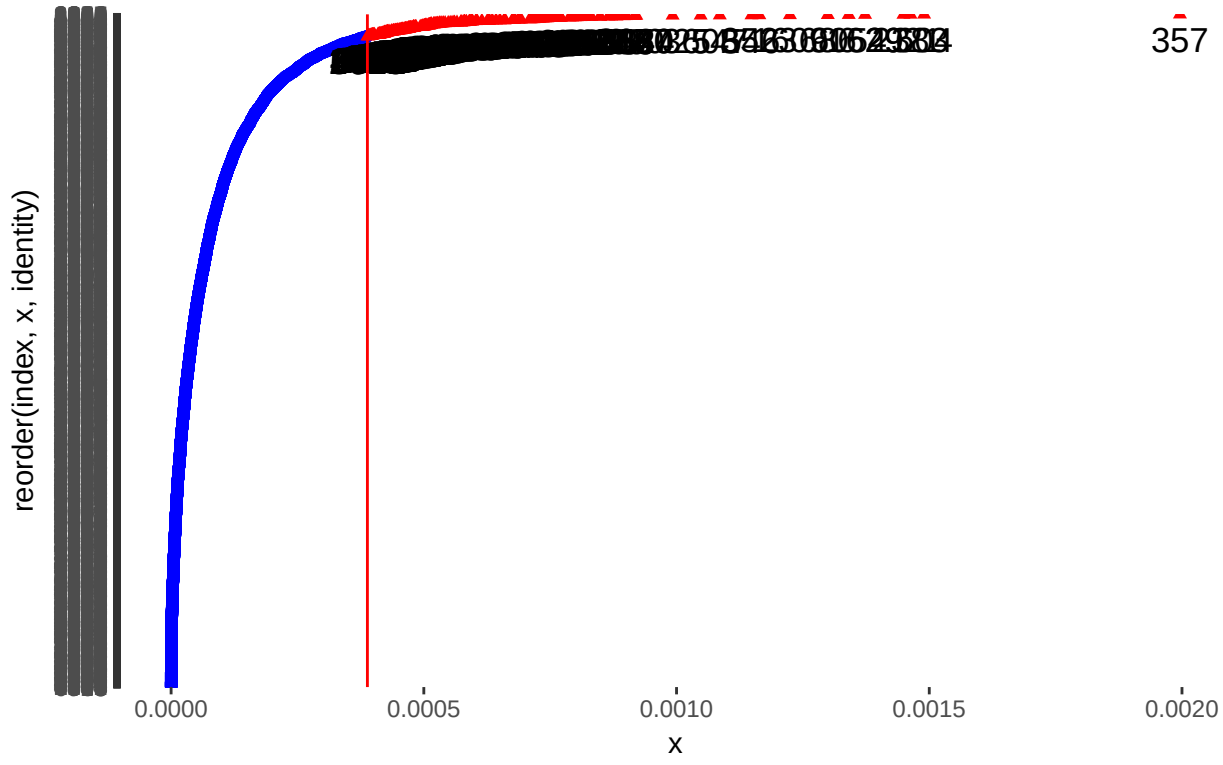

Marginal Residual
3
2
1
0
−1
−3
0
5
10
15
20
fitted(math.lme7)


Marginal Residual
3
2
1
0
−1
−3
−1
0
1
2
sizeScaled



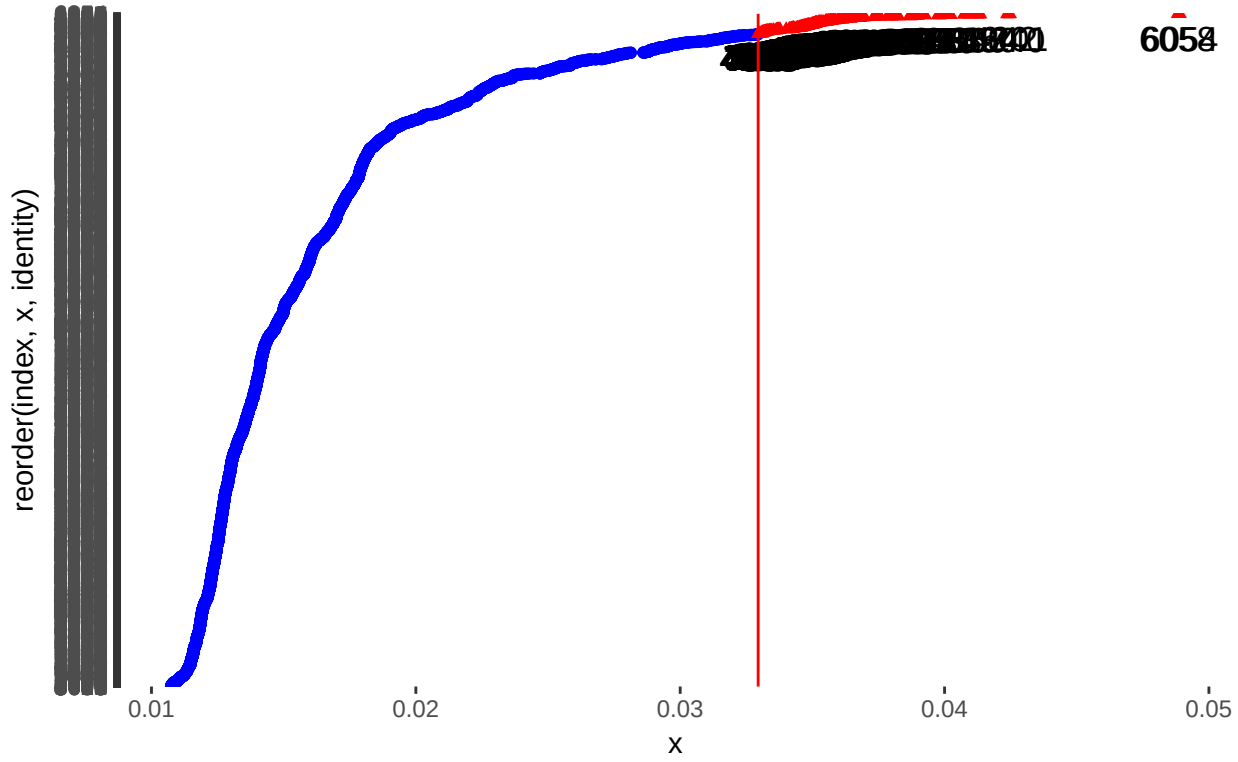
(c)

Observation 357 should be inspected based on the cooks distance plot below.



(d)

The two students who stand out the most in the leverage plot have very low Math scores, both of them ahve scores < 1 which seems very unsual.



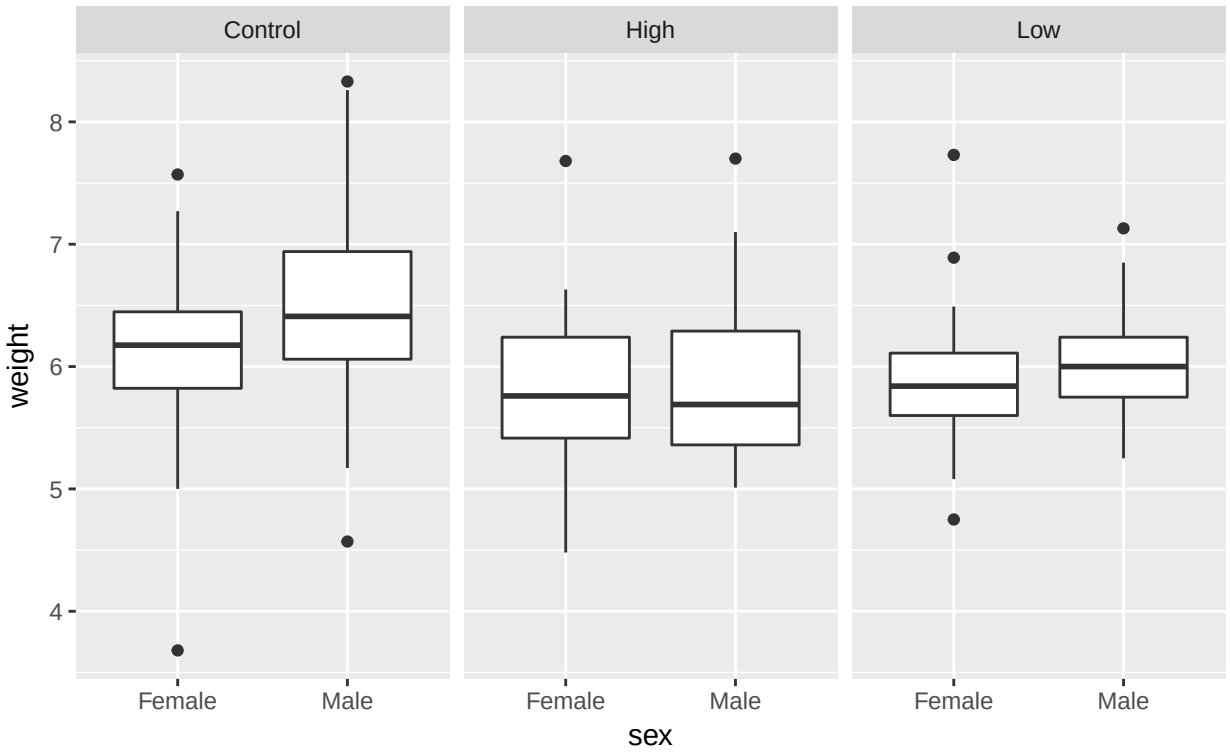

```
     School Size   Type Academic AveSES Student   Math Gender Minority
6054   8367  153 Public        0  0.032    6054 -2.832   Male      Yes
6058   8367  153 Public        0  0.032    6058  0.836   Male      Yes
        SES sizeScaled
6054  0.532  -1.496033
6058 -1.028  -1.496033
```

38

(a)

The plot below shows that there appears to be a relationship betweentreatment and rat weight. There also appears to be a relationship between sex of the rat and the wight, and possibly a interaction between treatment and sex of the rats.

```
  Mean.Weight Mean.LitSize
1    6.080963     13.32919
```



(b)

With the unconditional means model we get an intraclass correlation of .60, meaning that we will want to use a random effects model because a lot of the variation in our data can be explained by the different litters the rats are in.

39

(a)

Level 1: $Y_{ij} = a_i + b_i(Male) + \epsilon_{ij}$

Level 2: $a_i = \alpha_0 + \alpha_1(litsize) + \alpha_2(High) + \alpha_3(Low) + \mu_i$

$b_i = \beta_0 + w_i$

(b)

Model 1: random intercept and random slope

Model 2: random intercept only

Model1 vs Model2

D = 4.21 w/df=9, P-value of .081, so I went with model 2.

Model 3: no random effects

Model2 vs Model3

D = 90.5 df = 7, p-value = 0, sticking with model 2.

So we are sticking with the mdoel with the random effects of no random slope just random intercept.

40

(a)

Model 4: Interactions with Sex and Treatment

Model2 vs Model4

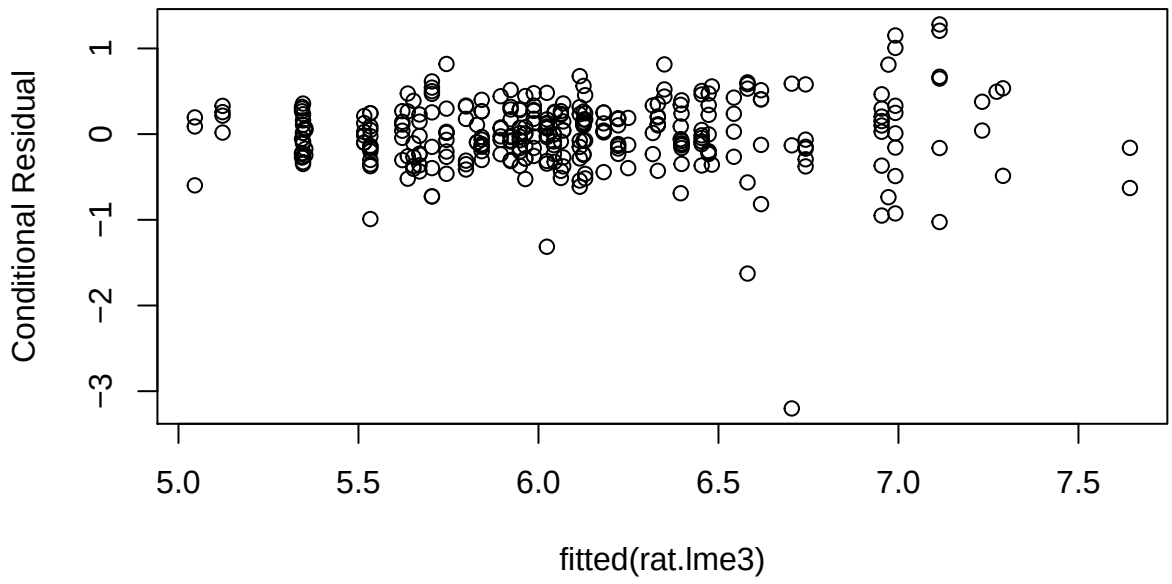
Anova test gives p-value of .615, Sticking with larger model, Model4

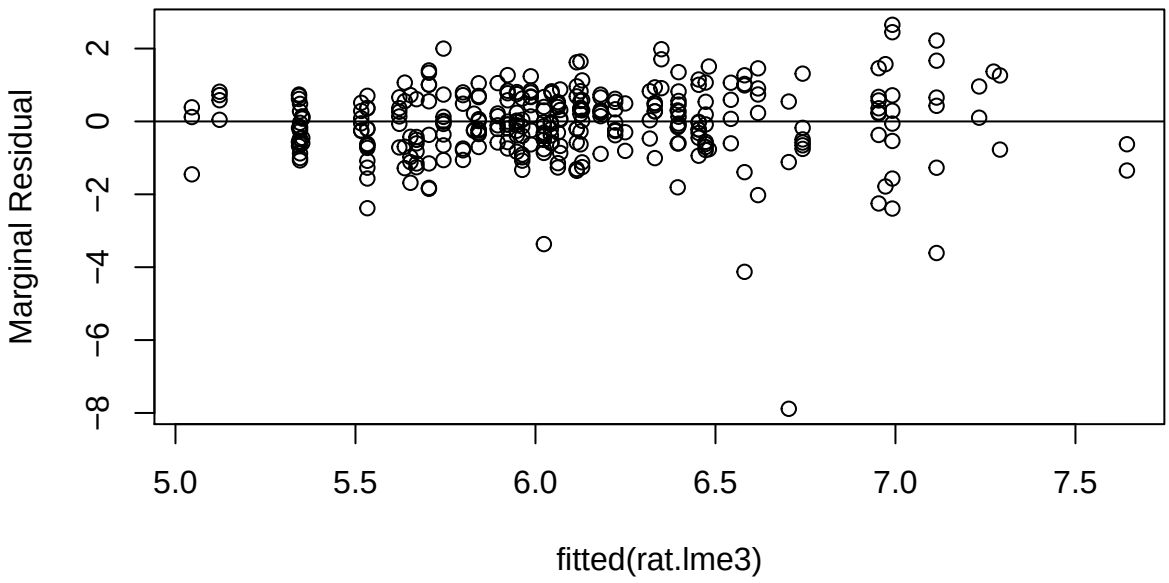
I am happy with Model 4 and so that will be my final model, with all of the fixed effects by themselves plus an interaction between sex and treatment level.

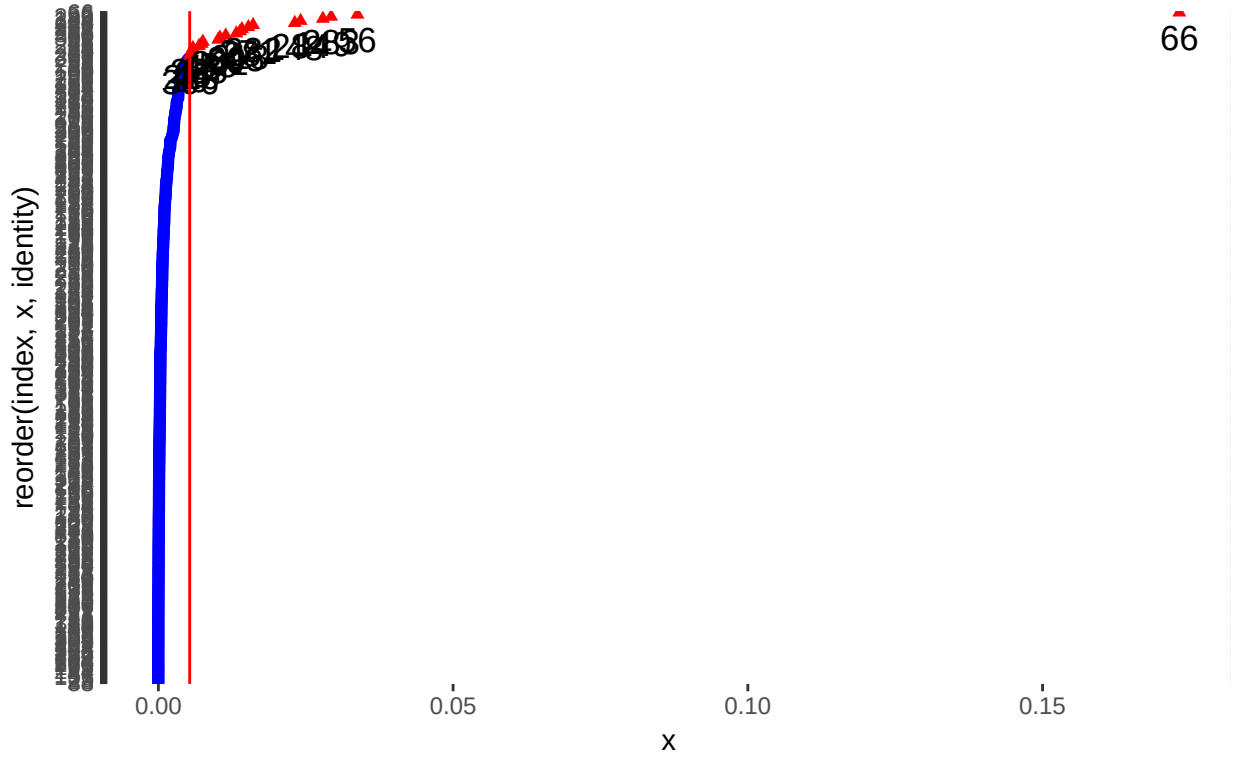
$$Y_{ij} = \alpha_0 + \alpha_1(litsize) + \alpha_2(High) + \alpha_3(Low) + \beta_0(Male) + \beta_1(MaleHigh) + \beta_2(MaleLow) + \mu_i + \epsilon_{ij}$$

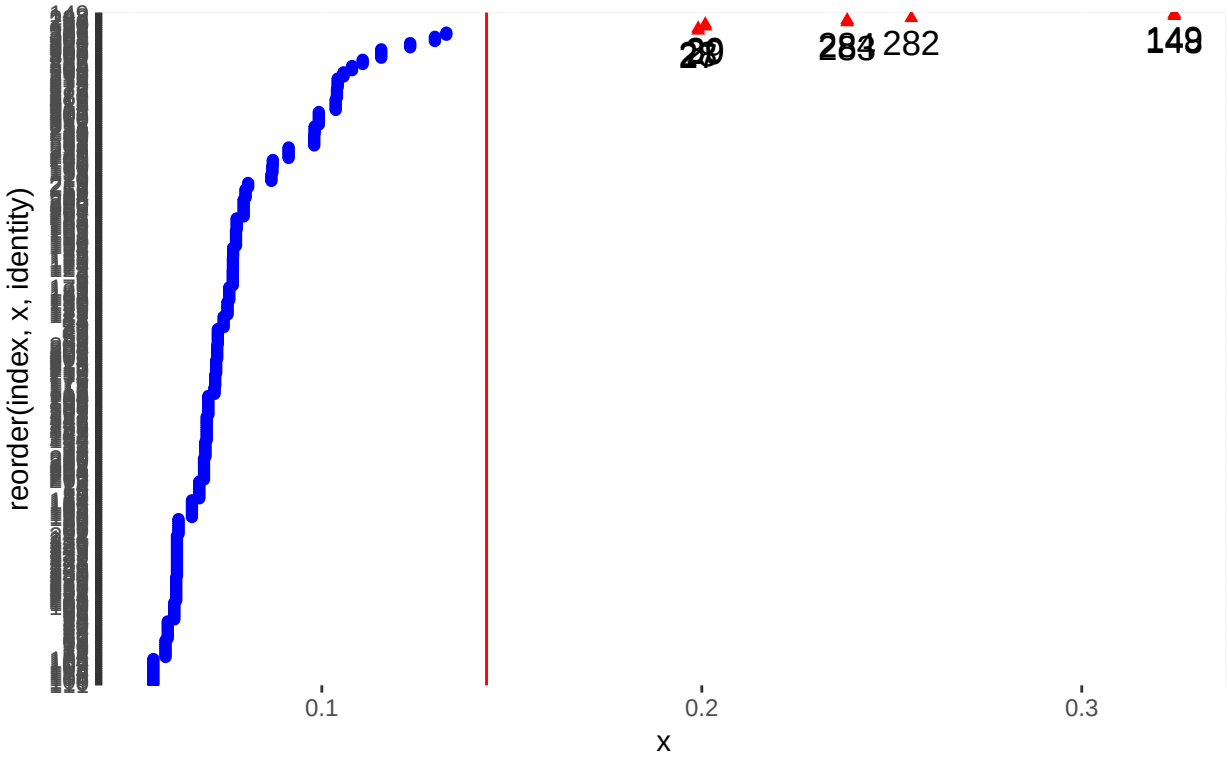
(b)

In our Marginal residual plot we can definitely see a few outliers in our model. In the COnditional Residuals there does appear to be some fanning out, so our model may not be the best fit, we may be missing something. In both the Cooks distance and leverage there are a few major outliers so again their may be more going on with our data then our model can accurately explain right now.




Marginal Residual
2
0
−2
−4
−6
−8
5.0
5.5
6.0
6.5
7.0
7.5
fitted(rat.lme3)


reorder(index, x, identity)
66
0.00
0.05
0.10
0.15
x



41

(1)

(a)

Level 1: Information about their Gang Activity and time of year

Level 2: Infromation about their parents, and their ethnic and cultural heritage

(b)

Level 1: Time of year, gang activity

Level 2: Race, Socioeconomic status

(2)

A wide format could have the explanatory variable as a seperate variable for each time the data was gathered. So a variable for March a variable for May, the observations would be just the 300 9th graders.

Long format the rows would be a specific measurement of a 9th grader at a given time. So each 9th grader would have 8 different rows where they would have some data in common and other things would change between those rows.

(3)

Lattice plots help you look at the relationship for each individual case or group that you are looking at, while spagetti plots allow you to more easily look at the overall trends across all of your data since you may be able to see a pattern having all of the spagetti lines laid on top of one another.