Group Project 2

Clare Clingain

April 26, 2018

Part 2: Clare

Running initial model

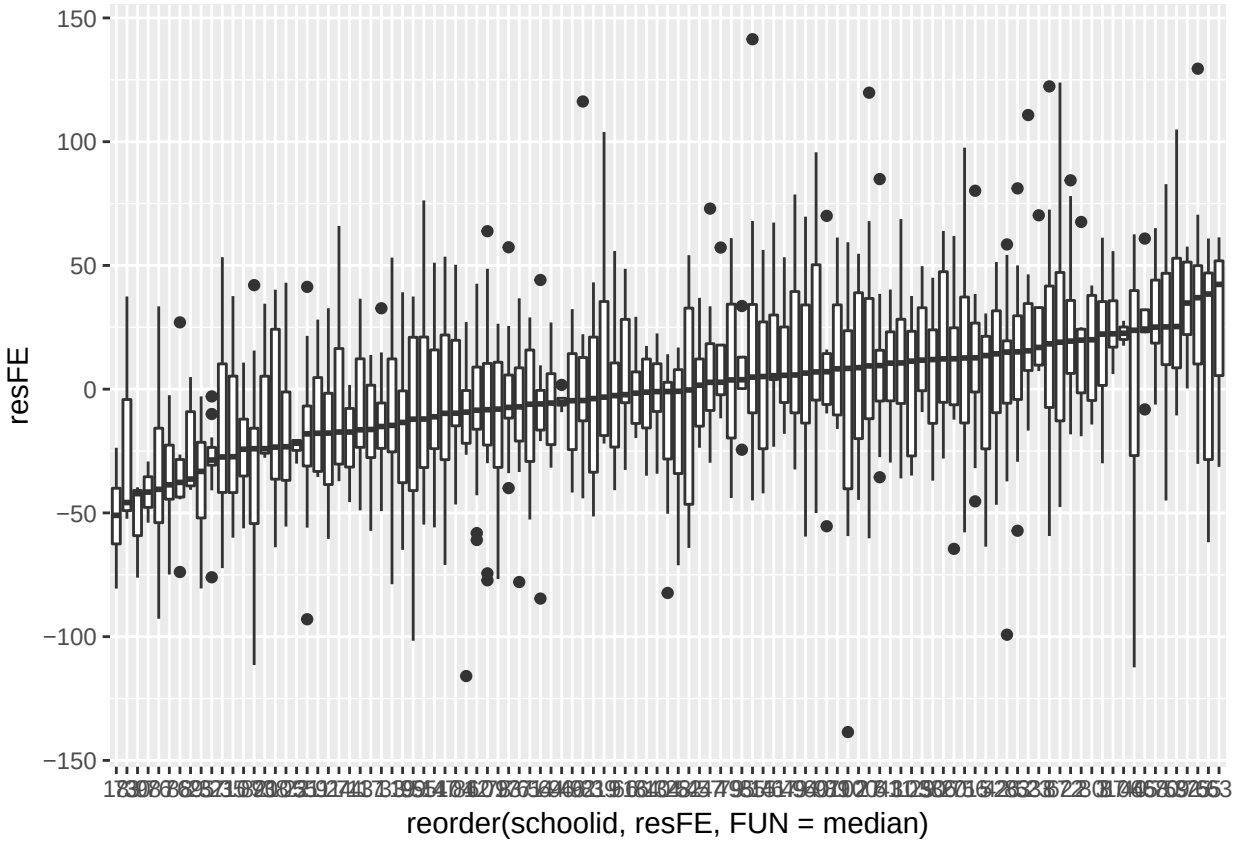
The initial model was run on a smaller dataset with 1081 observations due to missing data. School-level and classroom-level random intercepts are included in the model.

```
#remove missing data -- not ideal, but have to do it for this analysis
classroom <- classroom %>% mutate(Math1st = mathkind + mathgain)
classroom2 <- na.omit(classroom)
#model
new1 <-lmer(Math1st~housepov+mathknow+yearstea+mathprep+sex
            +minority+ses+(1|schoolid)+(1|classid),data=classroom2)
```

Residual that removes only the "fixed effects"

Below we calculate the residuals that removes only the fixed effects. The boxplot of the residuals shows that there is great variation within schools and that there is a steady linear trend to the residuals, suggesting dependence.

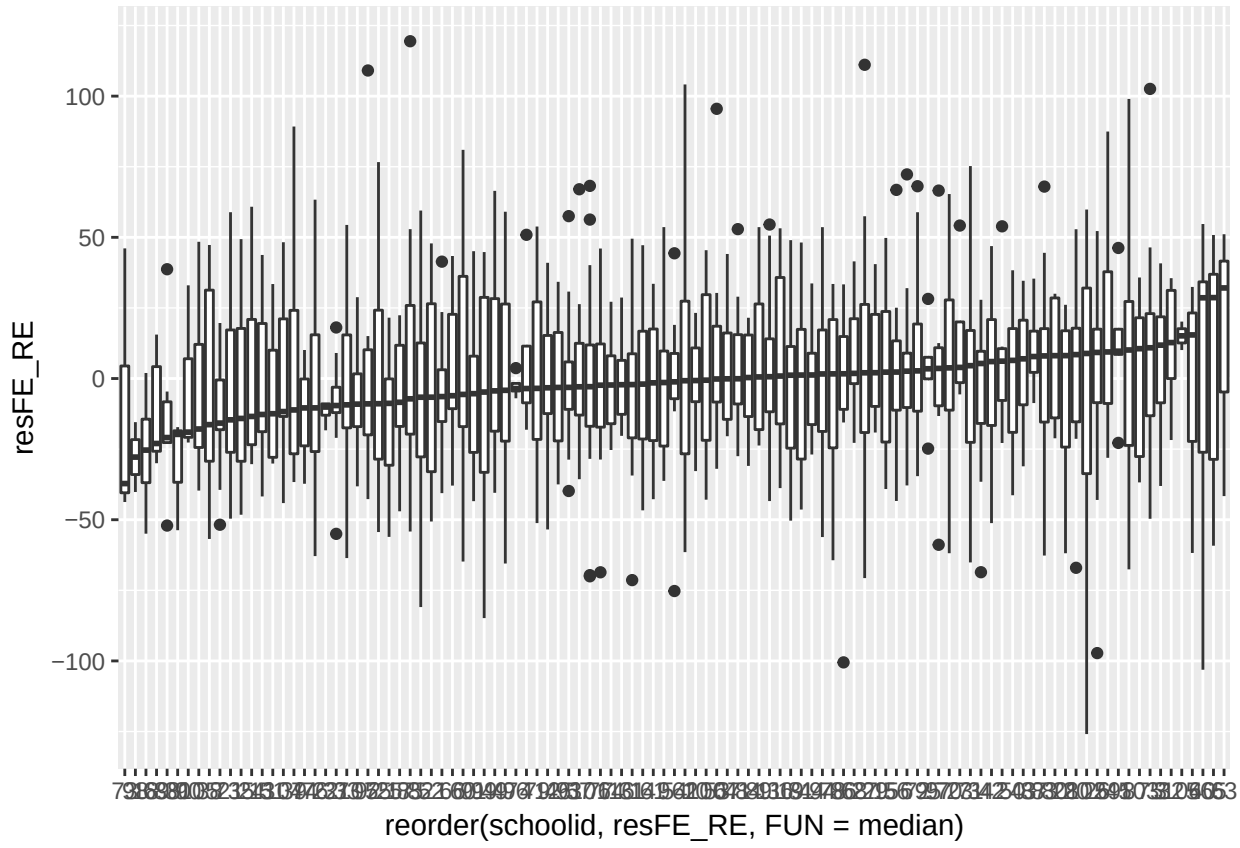
```
#predicted scores
pred.yhat <- predict(new1,re.form=~0)

#residual
resFE <- classroom2$Math1st-pred.yhat

#show that it's not independent
if (vanillaR) {
ord <- order(unlist(tapply(resFE, classroom2$schoolid, median)))
boxplot(split(resFE, classroom2$schoolid)[ord])
} else {
ggplot(classroom2, aes(x = reorder(schoolid, resFE, FUN = median), y = resFE)) +
geom_boxplot()
}
```



Residuals for BLUPs random effects

The residuals for the BLUPs random effects are calculated below. The boxplot reveals a similar dependency to the previous plot, though not as pronounced. There doesn't seem to be as high a correlation as there is in the other residuals plot.

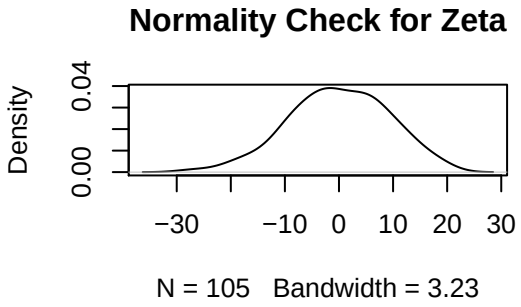
```
#getting predicted zeta_0 and eta_0
ranefs <- ranef(new1)
zeta0 <- ranefs$schoolid[,1]
eta0 <- ranefs$classid[,1]
#indexing
idx.sch <- match(classroom2$schoolid, sort(unique(classroom2$schoolid)))
idx.cls <- match(classroom2$classid, sort(unique(classroom2$classid)))
classroom2$zeta0 <- zeta0[idx.sch]
classroom2$eta0 <- eta0[idx.cls]
#now subtract all from outcome
resFE_RE <- classroom2$Math1st-pred.yhat-classroom2$zeta0-classroom2$eta0
#show that it's not independent, but much less correlated than resFE
if (vanillaR) {
ord <- order(unlist(tapply(resFE_RE, classroom2$schoolid, median)))
boxplot(split(resFE_RE, classroom2$schoolid)[ord])
}else{
ggplot(classroom2, aes(x = reorder(schoolid, resFE_RE, FUN = median), y = resFE_RE)) +
geom_boxplot()
}
```

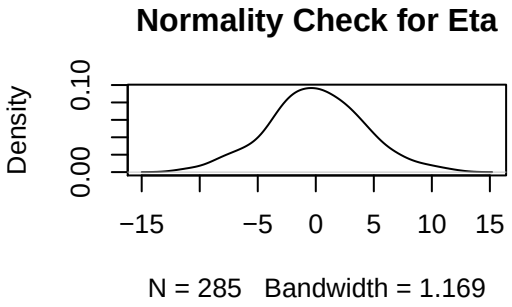


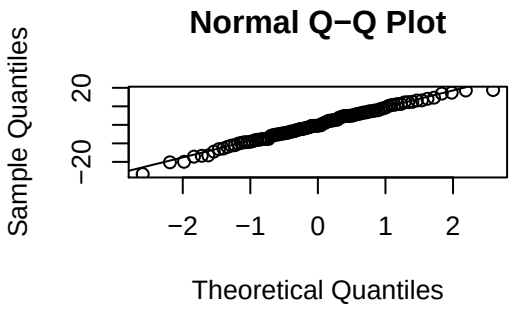
Examining BLUPs for normality

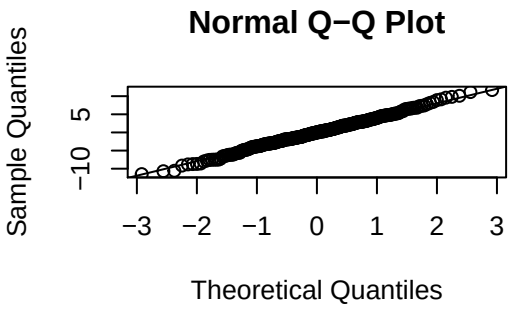
To examine the BLUPs for mormality, density plots and Q-Q plots were constructed. Both $zeta_0$ and eta_0 appear to be normal, with a few poossible outliers near the tails.

```
par(mfrow=c(2,2))
plot(density(zeta0), main ="Normality Check for Zeta")
plot(density(eta0), main = "Normality Check for Eta")
#looking good
qqnorm(zeta0);qqline(zeta0)
qqnorm(eta0);qqline(eta0)
```







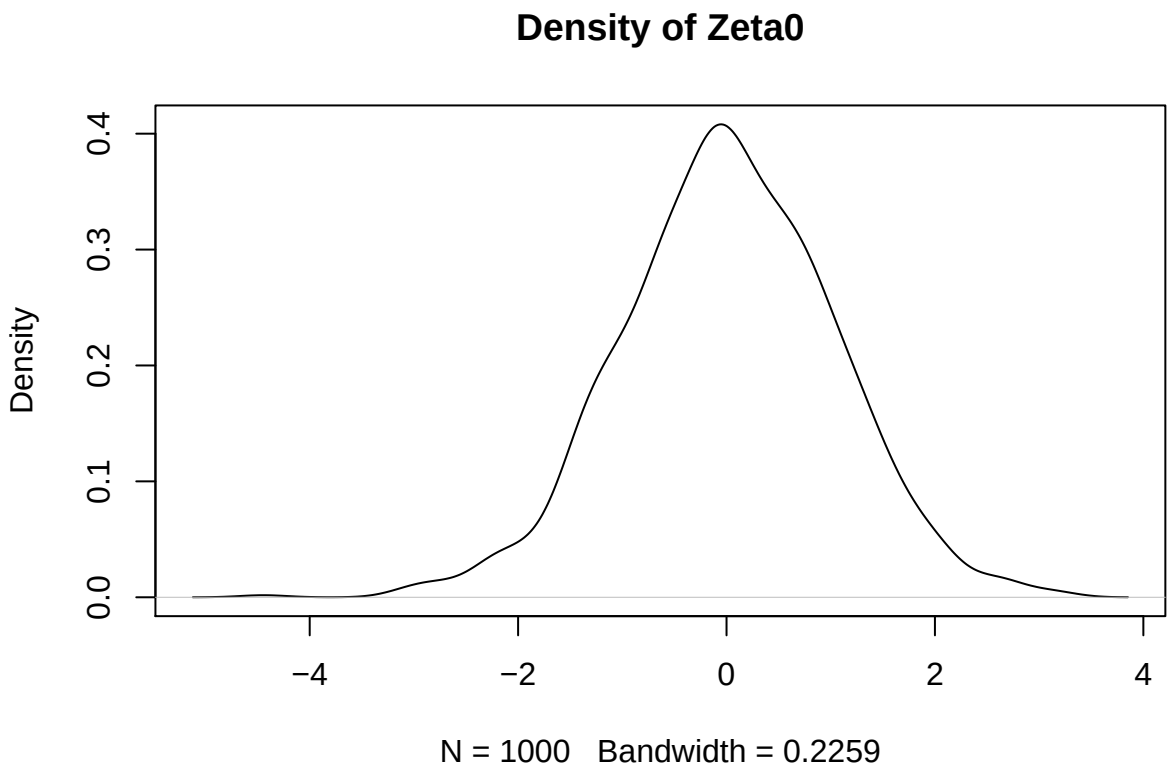



```
#looking good
```

Simulation

Below is a simulation based on the H0 being true, and a $\sigma_\epsilon = 1$. We find that the potential estimate is very close to 0, which we would expect since our $\sigma^2_{\zeta_0}$ has a "true" value of 0.

```
set.seed(10314)
school.sim <- matrix(1,10,100)
for (i in 1:100){
school.sim[,i] <- rnorm(10,mean=0, sd=1)
}
plot(density(school.sim), main = "Density of Zeta0")
```




```
paste("A potential estimate of sigma_{zeta_0} is ",mean(school.sim))
```

```
## [1] "A potential estimate of sigma_{zeta_0} is  0.0142117878263361"
```

New Complex Model

We now include a correlated random slope at the school-level for minority.

```
classroom <- read.csv("classroom.csv")
classroom <- classroom %>% mutate(Math1st = mathkind+mathgain)
classroom2 <- na.omit(classroom)
newcomplex <-lmer(Math1st~housepov+mathknow+yearstea+mathprep+sex+minority+ses+
                    (minority|schoolid)+(1|classid),data=classroom2)
summary(newcomplex)
```

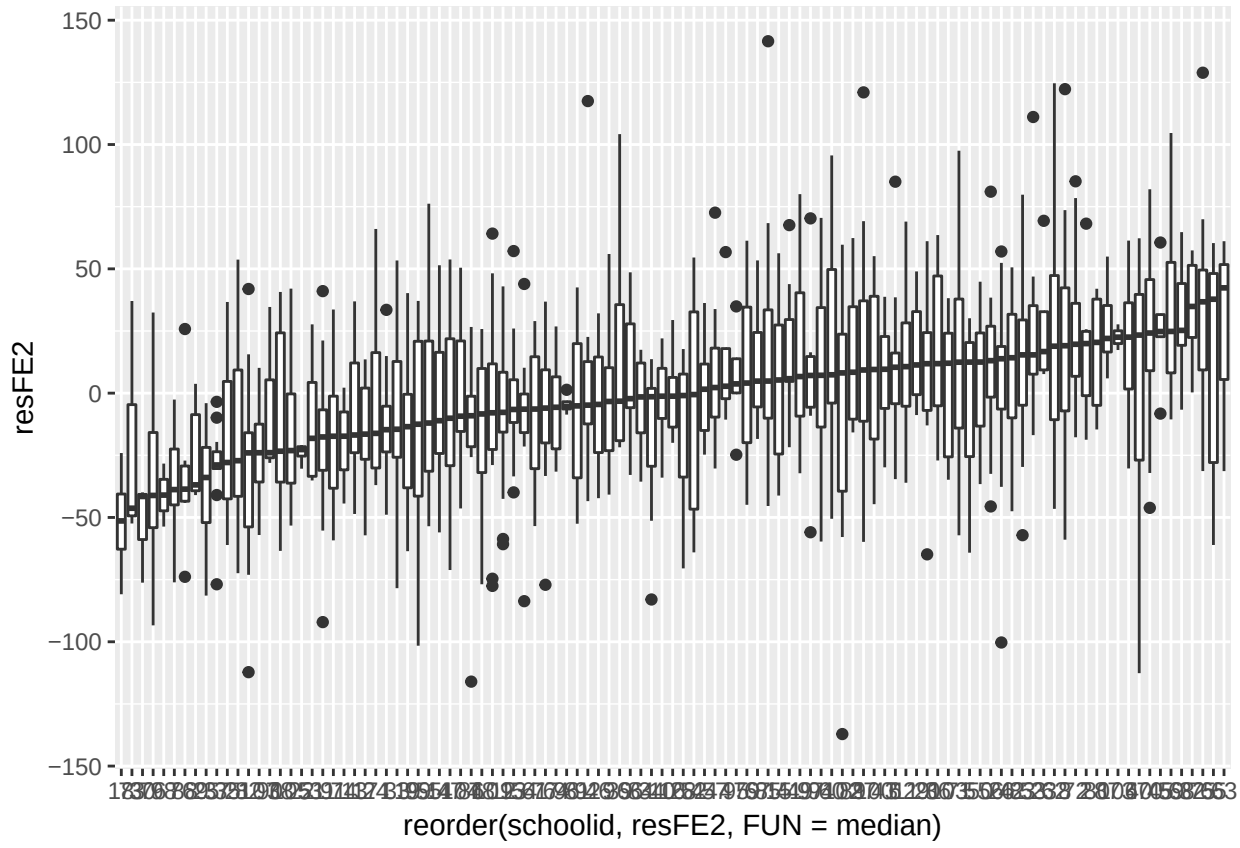
```
## Linear mixed model fit by REML t-tests use Satterthwaite approximations
##   to degrees of freedom [lmerMod]
## Formula:
## Math1st ~ housepov + mathknow + yearstea + mathprep + sex + minority +
##     ses + (minority | schoolid) + (1 | classid)
##    Data: classroom2
##
## REML criterion at convergence: 10717.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3.8952 -0.6358 -0.0345  0.6129  3.6444
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  classid  (Intercept)   86.69   9.311
##  schoolid (Intercept)  381.20  19.524
##           minority     343.13  18.524   -0.83
##  Residual             1039.39  32.240
## Number of obs: 1081, groups:  classid, 285; schoolid, 105
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  5.395e+02  5.655e+00  1.731e+02  95.399  < 2e-16 ***
## housepov    -1.606e+01  1.257e+01  1.000e+02  -1.277    0.204
## mathknow     1.632e+00  1.359e+00  2.248e+02   1.201    0.231
## yearstea    -4.368e-03  1.376e-01  2.172e+02  -0.032    0.975
## mathprep    -2.918e-01  1.335e+00  1.981e+02  -0.218    0.827
## sex         -8.628e-01  2.084e+00  1.022e+03  -0.414    0.679
## minority    -1.638e+01  3.896e+00  5.820e+01  -4.203 9.17e-05 ***
## ses          9.431e+00  1.543e+00  1.063e+03   6.111 1.39e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) houspv mthknw yearst mthprp sex    minrty
## housepov -0.394
## mathknow -0.078  0.061
## yearstea -0.253  0.091  0.024
## mathprep -0.576  0.037 -0.002 -0.167
## sex      -0.172 -0.013  0.010  0.014 -0.005
## minority -0.494 -0.157  0.099  0.027 -0.002 -0.014
## ses      -0.105  0.089 -0.005 -0.021  0.052  0.024  0.113
```

Manually calculate residuals for fixed effects

In the new model, we see a similar pattern of dependency. There is a general positive, linear trend to the residuals, and there is heterogeneity of variance across and within schools. These findings all suggest dependence.

```
#predicted scores
pred.yhat2 <- predict(newcomplex,re.form=~0)

#residual
resFE2 <- classroom2$Math1st-pred.yhat2

#show that it's not independent
if (vanillaR) {
ord <- order(unlist(tapply(resFE2, classroom2$schoolid, median)))
boxplot(split(resFE2, classroom2$schoolid)[ord])
} else {
ggplot(classroom2, aes(x = reorder(schoolid, resFE2, FUN = median), y = resFE2)) +
geom_boxplot()
}
```

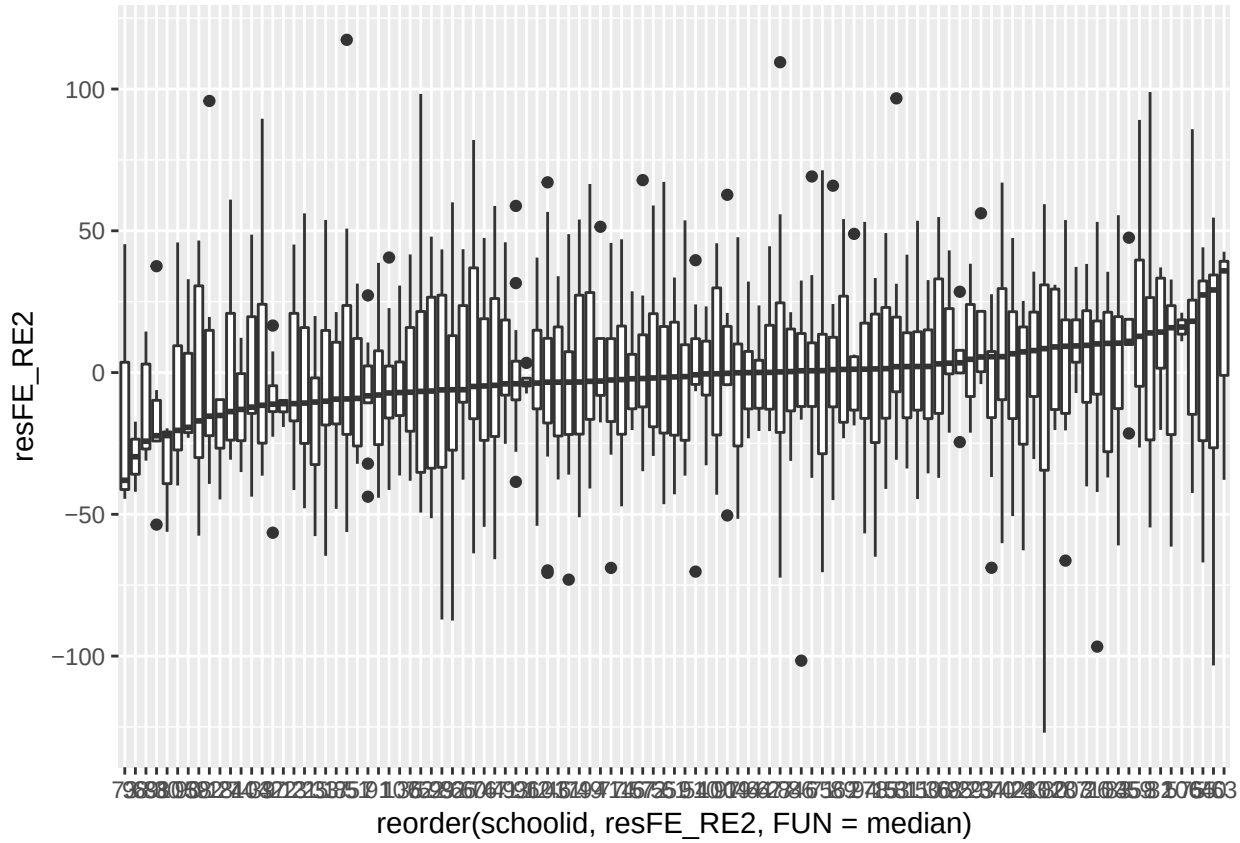


Residuals from BLUPs random effects

The residuals from the BLUPs random effects are calculated below. The boxplot of the residuals appears to be only slightly correlated, partly due to the uptake near the final set of schools on the x-axis. Although the correlation of the residuals is probably near 0, there is still enough variation within schools, and enough of a correlation in the data to suggest dependence.

```
#getting predicted zeta_0 and eta_0
ranefs2 <- ranef(newcomplex)
zeta0c <- ranefs2$schoolid[,1]
eta0c <- ranefs2$classid[,1]
zeta1c <- ranefs2$schoolid[,2]
#indexing
idx.sch <- match(classroom2$schoolid, sort(unique(classroom2$schoolid)))
idx.cls <- match(classroom2$classid, sort(unique(classroom2$classid)))
classroom2$zeta0c <- zeta0c[idx.sch]
classroom2$eta0c <- eta0c[idx.cls]
classroom2$zeta1c <- zeta1c[idx.sch]
#now subtract all from outcome
resFE_RE2 <- classroom2$Math1st-pred.yhat-classroom2$zeta0c-classroom2$eta0c-(classroom2$minority*classr
#show that it's not independent, but much less correlated than resFE
if (vanillaR) {
ord <- order(unlist(tapply(resFE_RE2, classroom2$schoolid, median)))
boxplot(split(resFE_RE2, classroom2$schoolid)[ord])
}else{
```

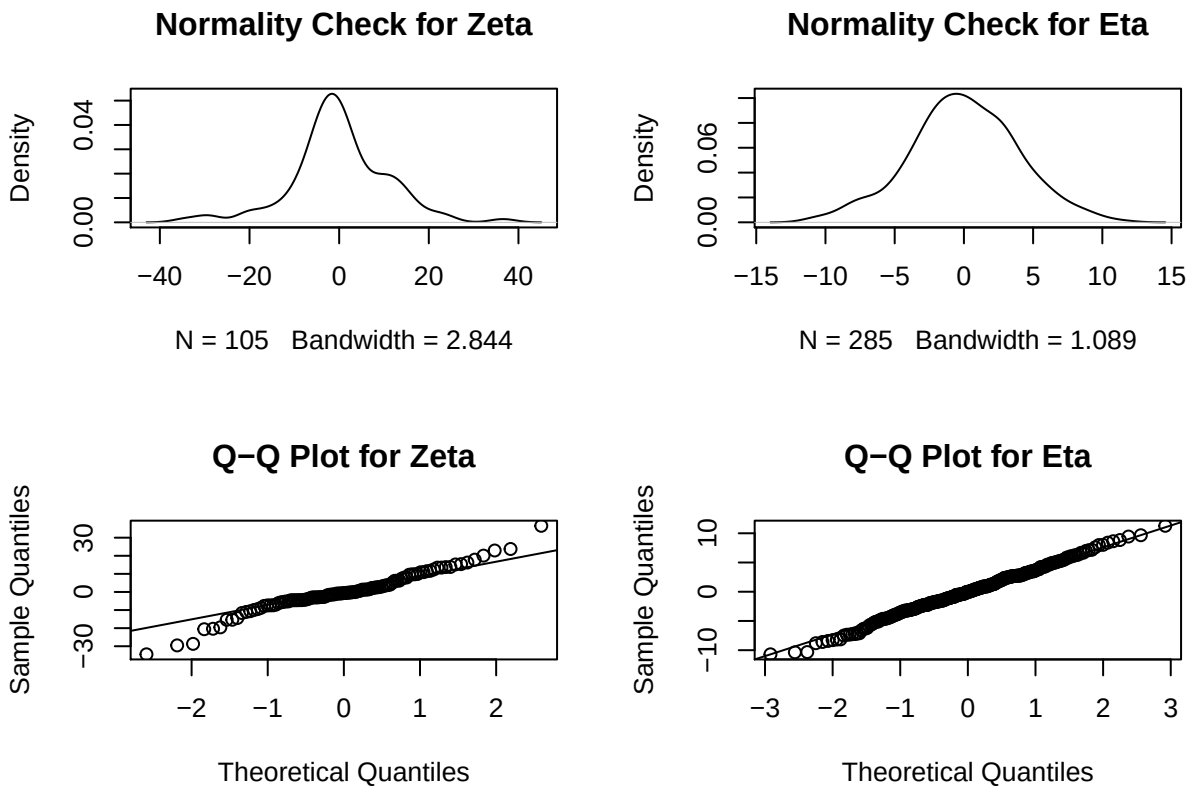
```
ggplot(classroom2, aes(x = reorder(schoolid, resFE_RE2, FUN = median), y = resFE_RE2)) +
geom_boxplot()
}
```



Examining Normality of BLUPs

Below we examine the normality of ζ_0 and η_0. The density and Q-Q plots for η_0 suggest normality, with a possibility of a few outliers near the tails. The normality of ζ_0 is more questionable. The tails do not appear to fit a normal distribution.

```
par(mfrow=c(2,2))
plot(density(zeta0c), main ="Normality Check for Zeta")
plot(density(eta0c), main = "Normality Check for Eta")
# eta looks pretty normal
#zeta not so much
qqnorm(zeta0c, main = "Q-Q Plot for Zeta");qqline(zeta0c)
qqnorm(eta0c, main = "Q-Q Plot for Eta");qqline(eta0c)
```



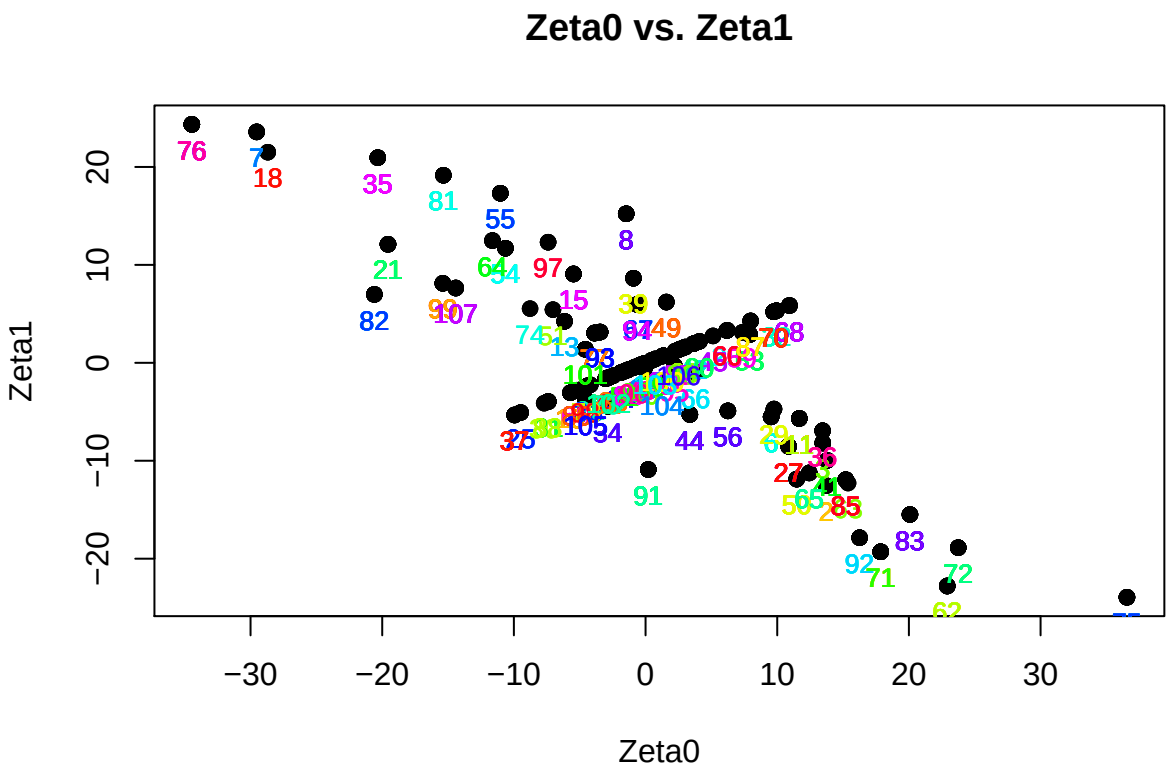

```
#zeta looking iffy, but with a few possible outliers
#eta good too, with few outliers.
```

Plotting ζ_0 versus ζ_1

The correlation between ζ_0 and ζ_1 in the output is -0.83. The graph below suggests a moderate negative trend, but there are some outliers that do not support this trend. Rather, they seem to be positively related.

*Note: the labels were put in rainbow in order to better discern their locations.

```
plot(classroom2$zeta0c,classroom2$zeta1c, main = "Zeta0 vs. Zeta1",
     ylab = "Zeta1",xlab = "Zeta0", pch=19)
 text(classroom2$zeta0c,classroom2$zeta1c, labels = classroom2$schoolid,
      cex = 0.8, col = rainbow(100), pos = 1)
```



Tracking down outliers

The outliers from the plots above can be tracked down by examining the data points via their IDs.

```
classroom2$zeta0c[classroom2$schoolid==45][[1]]/classroom2$zeta1c[classroom2$schoolid==45][[1]]
```

```
## [1] 1.868107
```

```
classroom2$zeta0c[classroom2$schoolid==68][[1]]/classroom2$zeta1c[classroom2$schoolid==68][[1]]
```

```
## [1] 1.868107
```

```
classroom2$zeta0c[classroom2$schoolid==30][[1]]/classroom2$zeta1c[classroom2$schoolid==30][[1]]
```

```
## [1] 1.868107
```

```
#there seems to be a trend here that the zeta0/zeta1 ratio is > 3, so let's filter it out
outliers <- classroom2 %>% filter(round(zeta0c/zeta1c,6)==1.868107) %>% select(zeta0c,zeta1c,schoolid,m
#now let's make sure the IDs from the plot are showing up here
unique(outliers$schoolid)
```

```
##  [1]   1   4   5   9  10  12  14  16  17  19  20  22  23  24  25  26  28
## [18]  30  31  32  33  37  38  42  43  45  46  47  52  57  60  61  68  69
## [35]  70  73  78  79  80  84  86  87  88  89  90  96  98 100 102 103 106
```

```
#They are! Now what's going on with minority?
table(outliers$minority)
```

```
##
```

```
##   1
## 455
```

```
tapply(outliers$minority, INDEX = outliers$schoolid, FUN = sum)
```

```
##   1   4   5   9  10  12  14  16  17  19  20  22  23  24  25  26  28  30
##   8   4   6   6  10  24  15   6   8  11  12   4   5   8   7  15  10   3
##  31  32  33  37  38  42  43  45  46  47  52  57  60  61  68  69  70  73
##  16   9  18  11   6  18   4   5  10  10   9   8  13  11  16   5  19   3
##  78  79  80  84  86  87  88  89  90  96  98 100 102 103 106
##  12  12   7   8   7  10   6   4   3   4   2   6  11   8   2
```

```
#The values match -- all students are minorities!
```

It seems like the (perfectly) positive trend in the data is being driven by schools in which all the students are minorities. That is, in schools in which there are only minority students, all other factors held equal, there is a boost in math scores in 1st grade for minority students. In a way, being in a totally minority school is a "protective" factor for minority students.