MLM Final Project Part 2ab

May 13 2020

Team Members and division of work:

Question 1

Refit the model in Part 1 that has all fixed effects as well as random intercepts (in schools and classrooms). Recall that `math1st = mathkind + mathgain` is the outcome. The model is `math1st ~ housepov + yearstea + mathprep + mathknow + ses + sex + minority + (1|schoolid/classid), REML = T)`

```
lm1 <- lmerTest::lmer(math1st ~ housepov + yearstea + mathprep + mathknow +
                   ses + sex + minority + (1|schoolid/classid), REML = T, data = classroom)
summary(lm1)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula:
## math1st ~ housepov + yearstea + mathprep + mathknow + ses + sex +
##     minority + (1 | schoolid/classid)
##    Data: classroom
##
## REML criterion at convergence: 10729.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.8581 -0.6134 -0.0321  0.5971  3.6598
##
## Random effects:
##  Groups           Name        Variance Std.Dev.
##  classid:schoolid (Intercept)   93.89   9.689
##  schoolid         (Intercept)  169.45  13.017
##  Residual                     1064.96  32.634
## Number of obs: 1081, groups:  classid:schoolid, 285; schoolid, 105
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  539.63041    5.31209  275.39009 101.585  < 2e-16 ***
## housepov     -17.64850   13.21755  113.87814  -1.335    0.184
## yearstea       0.01129    0.14141  226.80861   0.080    0.936
## mathprep      -0.27705    1.37583  205.27111  -0.201    0.841
## mathknow       1.35004    1.39168  234.49768   0.970    0.333
## ses           10.05076    1.54485 1066.56211   6.506 1.18e-10 ***
## sex           -1.21419    2.09483 1022.42110  -0.580    0.562
## minority     -16.18676    3.02605  704.47787  -5.349 1.20e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) houspv yearst mthprp mthknw ses    sex
## housepov -0.451
## yearstea -0.259  0.071
## mathprep -0.631  0.038 -0.172
```

```
## mathknow -0.083  0.058  0.029  0.004
## ses      -0.121  0.082 -0.028  0.053 -0.007
## sex      -0.190 -0.007  0.016 -0.006  0.007  0.020
## minority -0.320 -0.178  0.024  0.001  0.115  0.162 -0.011
```

a. Construct the residual that removes only the 'fixed effects' then subtract it from the outcome; call this residual resFE
   i. R hint 1: `predict` has an option to generate the prediction based on the fixed effects only.
   ii. R hint 2: If you decide to add a column to your data frame with resFE, note that predict only generates predictions for cases uses in the model *after listwise deletion.*

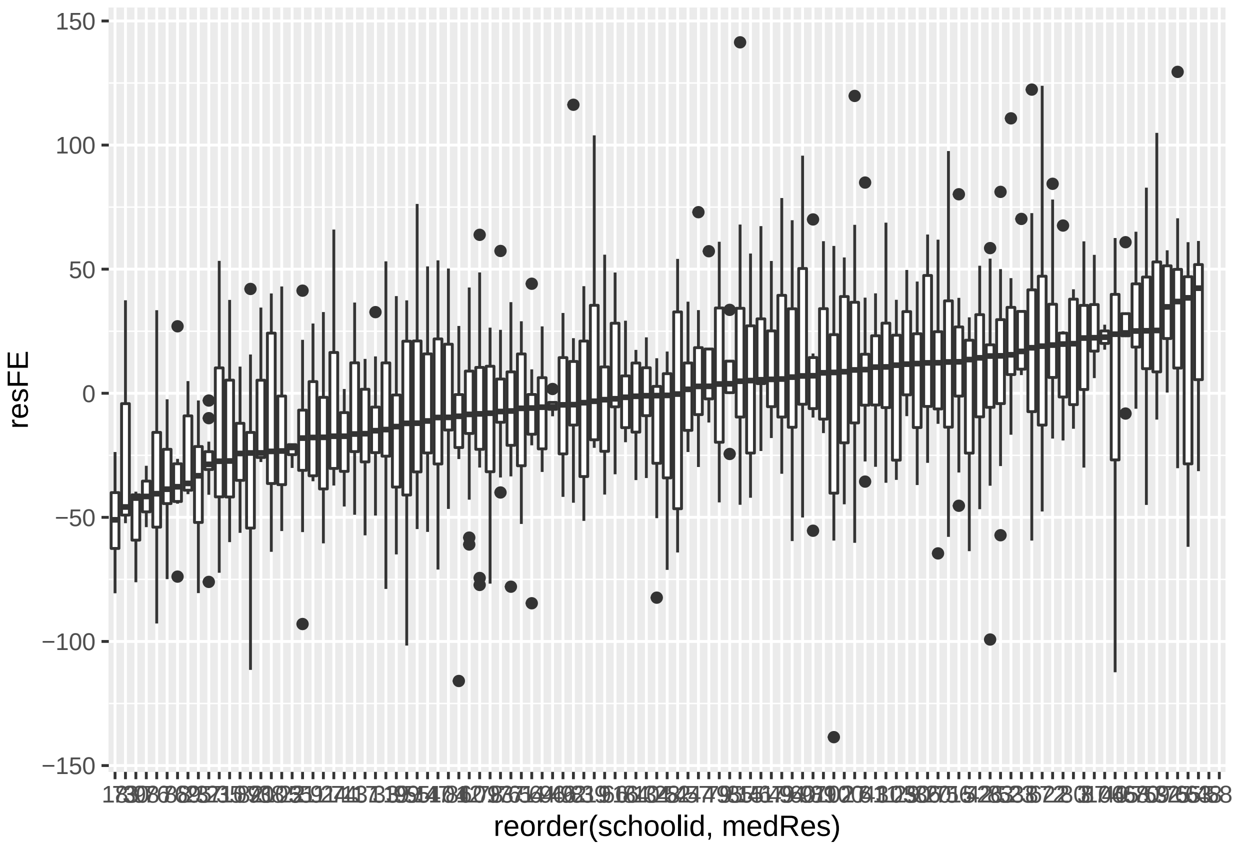
```
# Calculate predictions using fixed effects only:
predsFE <- predict(lm1, re.form = ~0)

# Calculate residual and add to dataframe:
resFE <- classroom[complete.cases(classroom),"math1st"] - predsFE
classroom[complete.cases(classroom), "resFE"] = resFE
```

Question 2

Show that the residual is not indepedent within schools in some manner.

```
# Insert code to show that the residual, resFE, is not indepedent within schools
classroom %>% group_by(schoolid) %>% mutate(medRes = median(resFE, na.rm = T)) %>% ggplot(., aes(x = re
```

```
## Warning: Removed 109 rows containing non-finite values (stat_boxplot).
```



- The boxplots of residuals show evidence of a relationship of scores within schools. After exlcluding random effects due to schools, the variation between each school is no longer accounted for, and the plot shows that some schools have residuals below the overall average, and some are above, indicative of heterogeneity.

Question 3

a. Construct the residual that utilizes the BLUPs for the random effects using the R command `residuals`.

i. Call the new residual resFE_RE

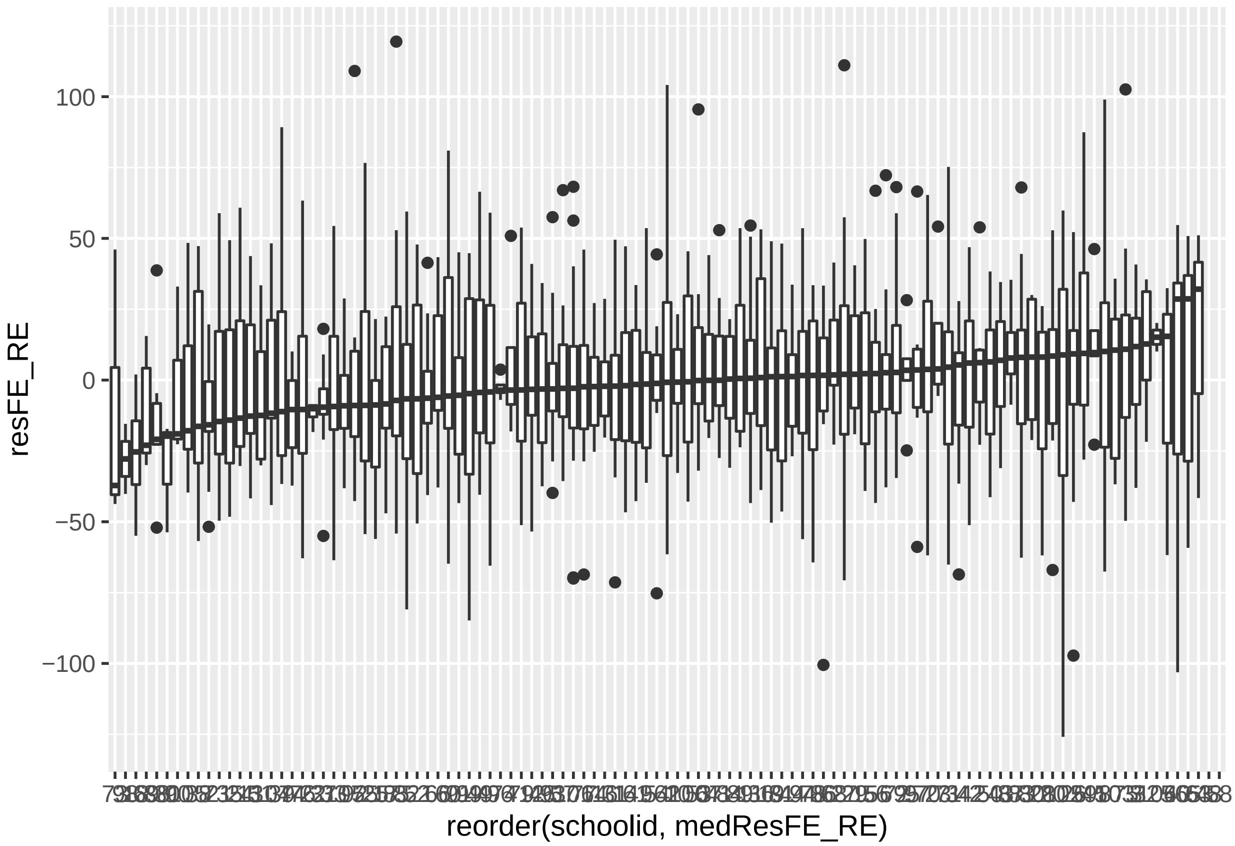
```
# Insert code to construct the residual
resFE_RE <- residuals(lm1)
classroom[complete.cases(classroom), "resFE_RE"] <- resFE_RE
```

Question 4

a. Show that these new residuals, resFE_RE are MUCH LESS (if not completely un-) correlated within schools, using the same method as before (boxplot?) (you should comment)

```
classroom %>% group_by(schoolid) %>% mutate(medResFE_RE = median(resFE_RE, na.rm = T)) %>% ggplot(., aes
```

```
## Warning: Removed 109 rows containing non-finite values (stat_boxplot).
```



`Response:`

The relationship within schools appears to be much less than before. The mean residuals for each school

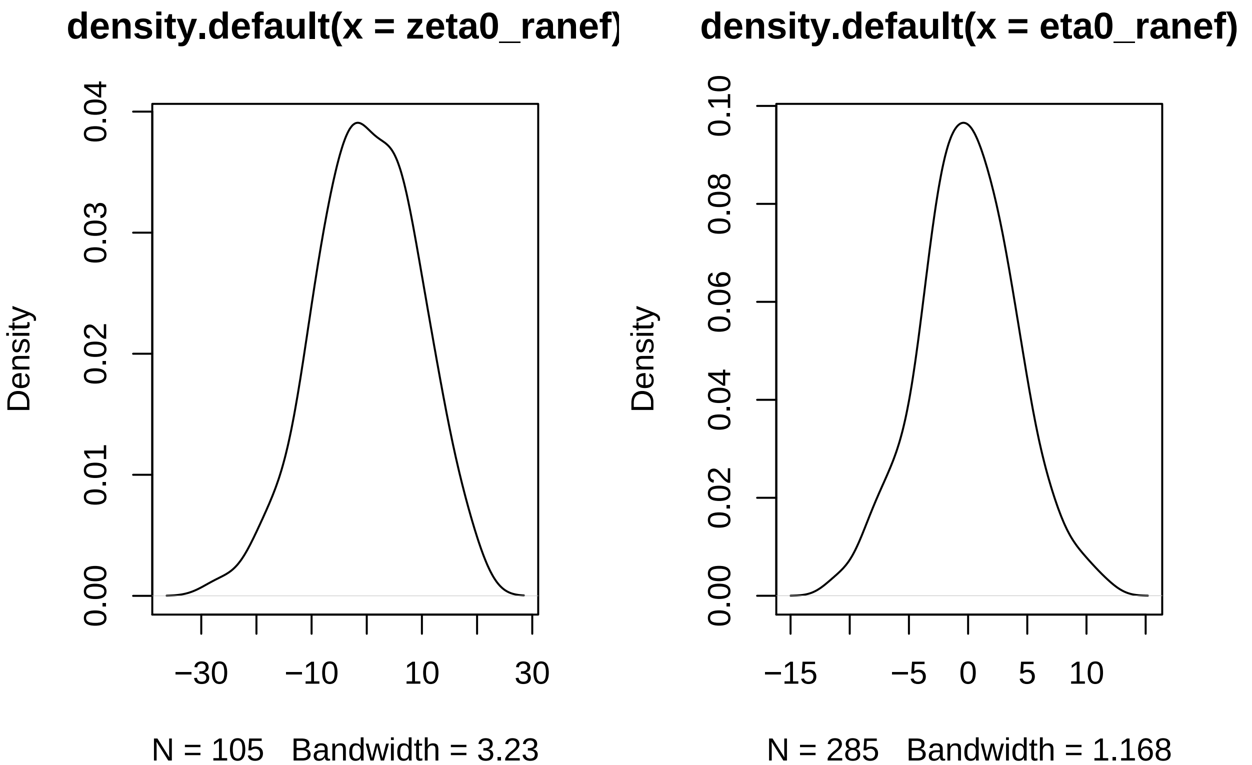
Question 5

a. Generate the two sets of BLUPs (for random effects zeta0 and eta0)

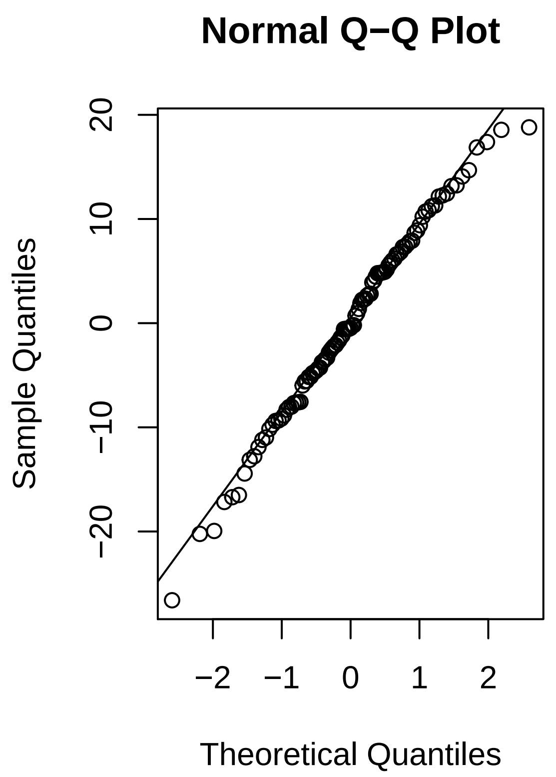
```
# Insert code to generate the two sets of BLUPS (zeta0 and eta0)
ranefs <- ranef(lm1)
zeta0_ranef <- ranefs$schoolid[,1]
eta0_ranef <- ranefs$classid[,1]
```

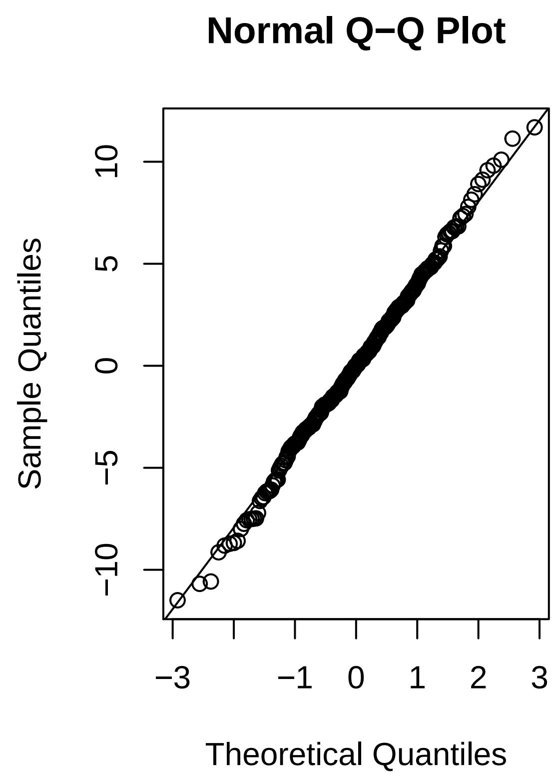
b. Examine these for normality (include evidence), and comment.

```
# Insert code to examine BLUPs for normality
# par(mfrow=c(1,2)) produces palette for one row of plots with two columns

par(mfrow = c(1,2))
plot(density(zeta0_ranef))
plot(density(eta0_ranef))
```




```
qqnorm(zeta0_ranef)
qqline(zeta0_ranef)

qqnorm(eta0_ranef)
qqline(eta0_ranef)
```






```
Response:
```

The density and QQ-plots of the random effects for schools and classrooms (ζ_0 and η_0) show evidence of normality.

Question 6

a. Fit a slightly more complicated model with the same fixed effects, but now add a random slope for minority, correlated with the random intercept, at the school level (keep the classroom level random intercept).

```
# Insert code to fit the slightly more complicated model and print the summary
lm2 <- lmerTest::lmer(math1st ~ housepov + yearstea + mathprep + mathknow +
                  ses + sex + minority + (minority | schoolid) + (1 | classid), REML = T, data = class
summary(lm2)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula:
## math1st ~ housepov + yearstea + mathprep + mathknow + ses + sex +
##     minority + (minority | schoolid) + (1 | classid)
##    Data: classroom
##
## 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:
```

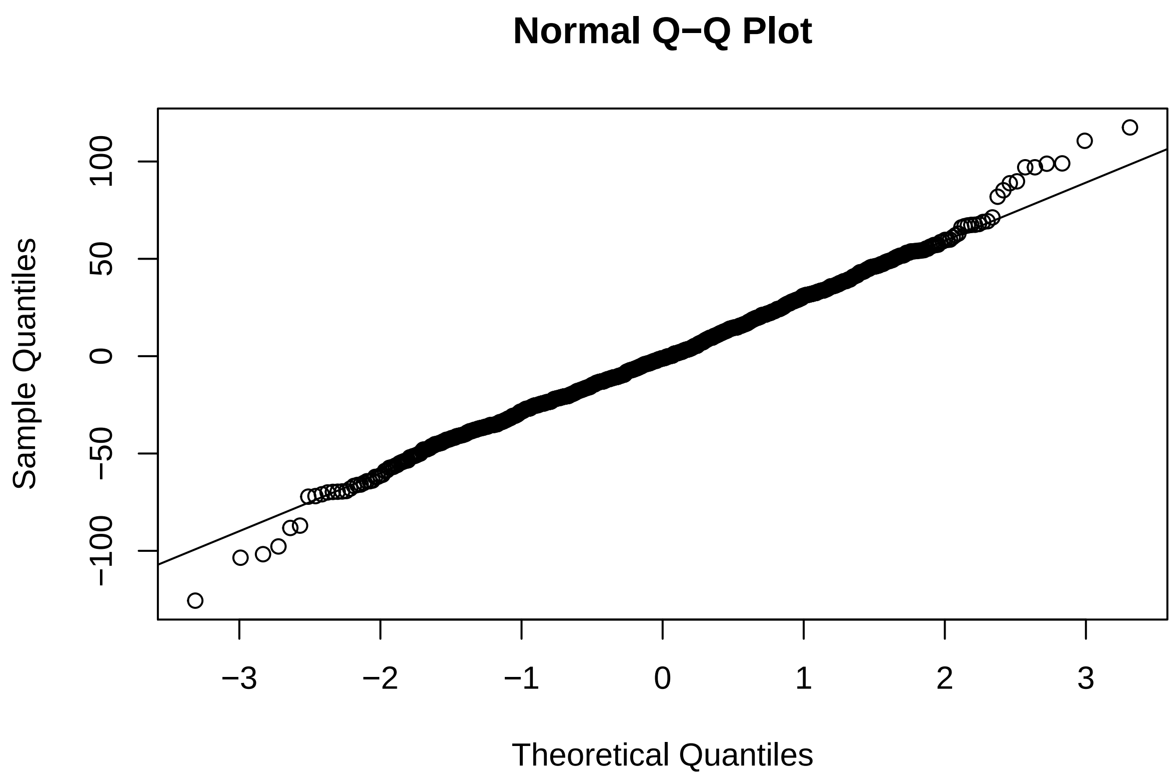
```
##  Groups   Name        Variance Std.Dev. Corr
##  classid  (Intercept)   86.7    9.311
##  schoolid (Intercept)  381.2   19.524
##           minority     343.2   18.525   -0.83
##  Residual             1039.4   32.240
## Number of obs: 1081, groups:  classid, 285; schoolid, 105
##
## Fixed effects:
##                Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  539.49369    5.65513  173.09178  95.399  < 2e-16 ***
## housepov     -16.06251   12.57477   99.99134  -1.277    0.204
## yearstea      -0.00437    0.13765  217.17884  -0.032    0.975
## mathprep      -0.29178    1.33537  198.06922  -0.218    0.827
## mathknow       1.63216    1.35929  224.78144   1.201    0.231
## ses            9.43095    1.54335 1063.13485   6.111 1.39e-09 ***
## sex           -0.86278    2.08382 1021.81437  -0.414    0.679
## minority     -16.37547    3.89604   58.24604  -4.203 9.17e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) houspv yearst mthprp mthknw ses    sex
## housepov -0.394
## yearstea -0.253  0.091
## mathprep -0.576  0.037 -0.167
## mathknow -0.078  0.061  0.024 -0.002
## ses      -0.105  0.089 -0.021  0.052 -0.005
## sex      -0.172 -0.013  0.014 -0.005  0.010  0.024
## minority -0.494 -0.157  0.027 -0.002  0.099  0.113 -0.014
```

b. Construct the residual (individual, level 1) and the BLUPs for the remaining random effects. Call the new residual resFE_RE as before.

```r
# Insert code to construct residual and BLUPs
resFE_RE <- residuals(lm2)
```

c. Examine all error estimates (individual level residuals, BLUPs (school and classroom level) for normality (and comment)).

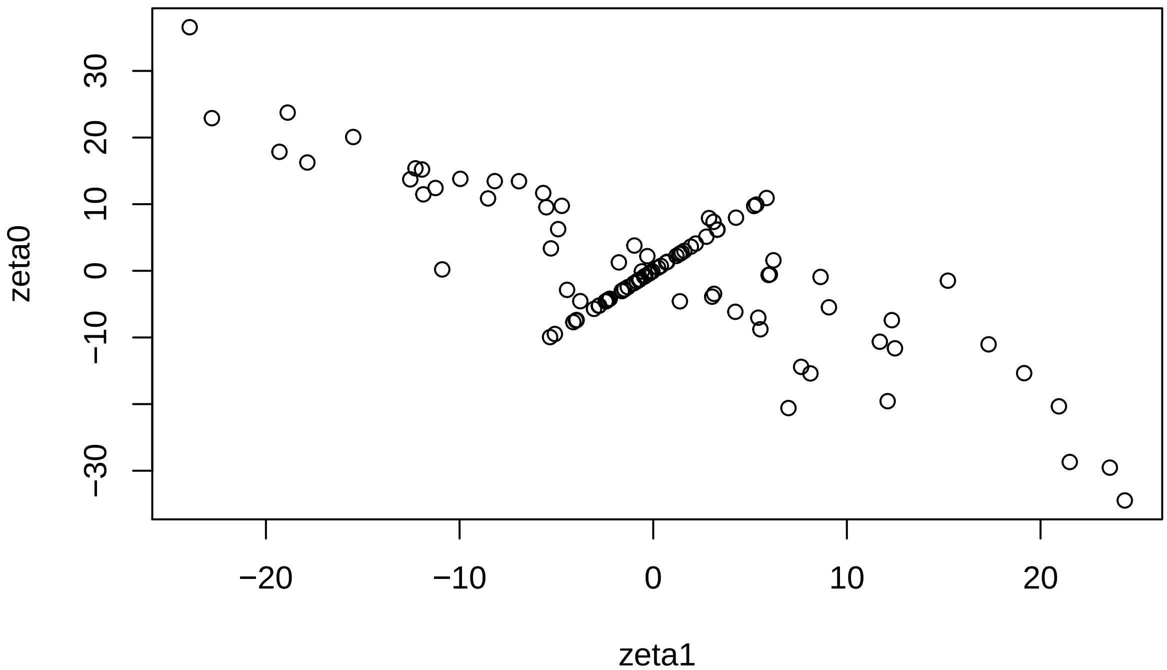
```r
qqnorm(resFE_RE)
qqline(resFE_RE)
```



Response:

d. Plot zeta0 vs. zeta1 to see whether the estimated correlation is consistent with the observed. Briefly comment.

```
zeta0 <- ranef(lm2)$schoolid[,1]
zeta1 <- ranef(lm2)$schoolid[,2]

plot(x = zeta1, y = zeta0)
```




```
Response:
```

e. Track down those odd points in the scatterplot. What schools are they? Do they have anything in common? (You should comment)

```r
# Insert code if you want to examine odd points

# Identify which schools are odd on the scatterplot:
test_df <- data.frame(zeta0 = zeta0, zeta1 = zeta1, z0z1 = zeta0*zeta1)
which(test_df$z0z1 > 0)
```

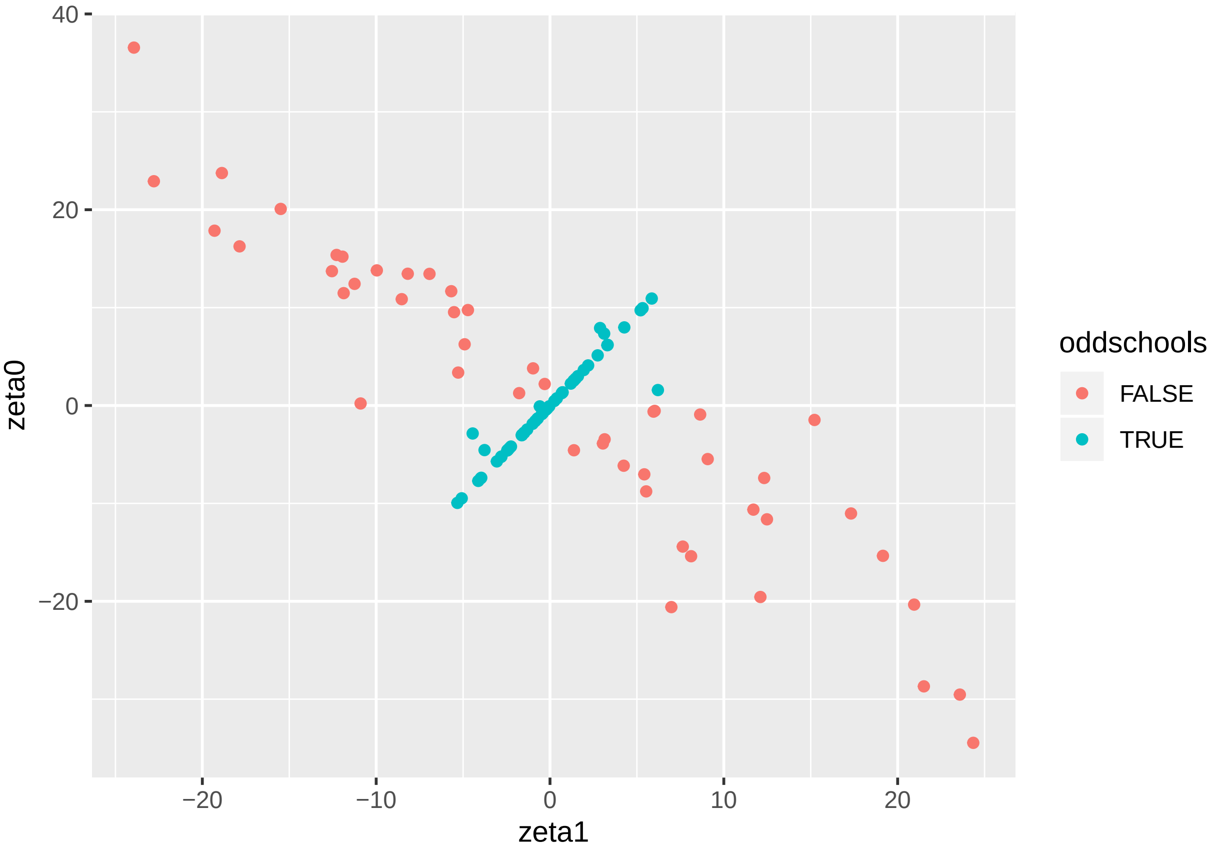
```
##  [1]   1   4   5   9  10  12  14  16  17  19  20  22  23  24  25  26  28
## [18]  30  31  32  33  34  37  38  40  42  43  45  46  47  48  51  52  56
## [35]  57  58  59  66  67  68  71  76  77  78  82  84  85  86  87  88  94
## [52]  96  98 100 101 103 104
```

```r
# Add "oddschools" indicator to dataset:
classroom$oddschools <- classroom$schoolid %in% which(test_df$z0z1 > 0)

# Calculate percentage of minority students in each odd school:
classroom %>% group_by(oddschools) %>% summarize(minority_avg = mean(minority))
```

```
## # A tibble: 2 x 2
##   oddschools minority_avg
##   <lgl>             <dbl>
## 1 FALSE             0.569
## 2 TRUE              0.772
```

```r
# Show odd schools in plot of zeta1 v. zeta0:
distinct(test_df %>%
           mutate(schoolid = row_number()) %>%
           left_join(classroom[, c("schoolid", "oddschools")])) %>%
            ggplot(., aes(x = zeta1, y = zeta0, color = oddschools)) + geom_point()
```

```
## Joining, by = "schoolid"
```




```
Response:
```

The "odd" schools in the scatterplot are those schools that have mostly minority populations. This makes it difficult to estimate a random slope for these schools because there is little variation in minority (i.e. slope estimates close to 0).

Question 7

Make a *person-period* file with math score (Kindergarten and First grade). That is, `math0 <- mathkind; math1 <- mathkind + mathgain` (you have to make this work in the dataframe). Using `reshape` in R, you have to be careful to specify the name of the math variable (`math0` and `math1`) as *varying*.

```r
# Insert code to create the variables math0 and math1 and to reshape data
personperiod <- classroom %>% mutate(math0 = mathkind, math1 = mathkind + mathgain)

class_pp <- reshape(personperiod, varying = c("math0", "math1"), v.names = "math", timevar = "year",
times = c(0, 1), direction = "long")
```

Question 8

We ignore classrooms in this analysis, but keep it in the notation.

a. Fit a model with `math` as outcome, and fixed effect for time trend (`year`), and random intercepts for schools.

```r
lm3 <- lmerTest:: lmer(math ~ year + (1 | schoolid), data = class_pp)
summary(lm3)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: math ~ year + (1 | schoolid)
##    Data: class_pp
##
## REML criterion at convergence: 23951.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.2833 -0.6084  0.0037  0.6329  3.7761
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  schoolid (Intercept)  348.7   18.67
##  Residual             1268.4   35.62
## Number of obs: 2380, groups:  schoolid, 107
##
## Fixed effects:
##             Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)  464.932      2.116  132.154  219.73   <2e-16 ***
## year          57.566      1.460 2270.855   39.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr)
## year -0.345
```

b. Write down the model

Equation:

$$MATH_{tijk} = b_0 + \zeta_{0k} + b_1 TIME_{tijk} + \epsilon_{tijk}$$
$$\text{with } \zeta_{0k} \sim N(0, \sigma^2_{\zeta_0}),\ \epsilon_{tijk} \sim N(0, \sigma^2_{\epsilon}), \text{ independent of each other}$$

c. Add random intercepts for child

```r
# Insert code to fit new model and print summary output
lm4 <- lmerTest:: lmer(math ~ year + (1 | schoolid/childid), data = class_pp)
summary(lm4)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: math ~ year + (1 | schoolid/childid)
##    Data: class_pp
##
## REML criterion at convergence: 23554.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4.7492 -0.4811  0.0085  0.4881  3.4957
##
## Random effects:
##  Groups           Name        Variance Std.Dev.
##  childid:schoolid (Intercept) 702.0    26.50
##  schoolid         (Intercept) 307.5    17.54
##  Residual                     599.1    24.48
## Number of obs: 2380, groups:  childid:schoolid, 1190; schoolid, 107
##
## Fixed effects:
##             Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)  465.118      2.042  117.023  227.74   <2e-16 ***
## year          57.566      1.003 1189.000   57.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr)
## year -0.246
```

d. Write down the model

`Equation:`

$$MATH_{tijk} = b_0 + \delta_{0ijk} + \zeta_{0k} + b_1 TIME_{tijk} + \epsilon_{tijk}$$
$$\text{with } \zeta_{0k} \sim N(0, \sigma^2_{\zeta_0}),\ \delta_{0ijk} \sim N(0, \sigma^2_{\delta_0}),\ \epsilon_{tijk} \sim N(0, \sigma^2_{\epsilon}), \text{ independent of each other}$$

Question 9

Report original and new variance estimates of $\sigma^2_{\zeta_0}$ (between schools) and σ^2_{ε} (within schools):

$\sigma^2_{\zeta_0}$:

- Original 348.7
- New: 307.5

  σ^2_{ε} :

- Original: 1268.4
- New: 599.1

a. Compute a pseudo R^2 relating the between school variation and ignoring between students in the same school. In other words, what fraction of the between-school variance in the first model is 'explained' by the addition of a student random effect?

```
# Insert code to compute psuedo R^2 or do this inline
(rsq_b <- (348.7 - 307.5)/(348.7))
```

```
## [1] 0.1181531
```

The Psuedo-R^2 is `0.1181531` which means that approximately 15% of between-school variance in the first model is explained by the addition of the student random effect in the second model.

b. Does the total variation stay about the same (adding between children within schools variance as well, to the second model results) (you should comment)?

`Response:`

The total variation is approximately the same between both models (1619.9 in the first model, and 1608.6 in the second model).

Question 10

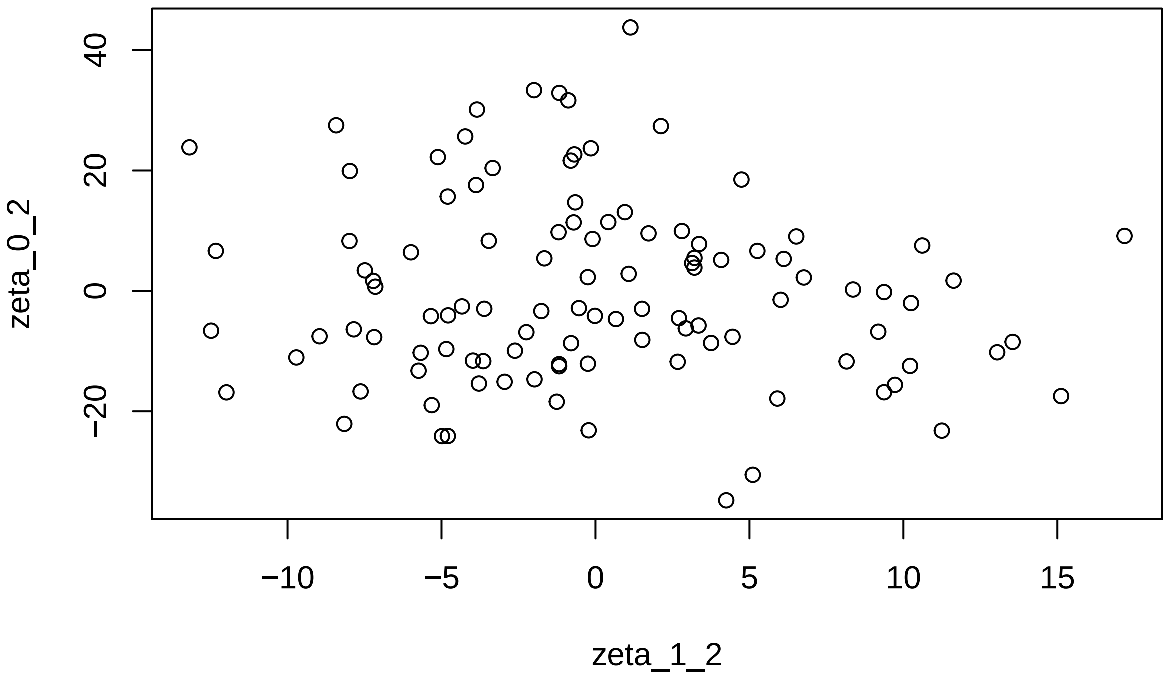
Add a random slope (ζ_1) for the trend (year) within schools (uncorrelated with random intercept (ζ_0))

```
lm5 <- lmerTest:: lmer(math ~ year + (0 + year | schoolid) + (1 | schoolid/childid), data = class_pp)
summary(lm5)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: math ~ year + (0 + year | schoolid) + (1 | schoolid/childid)
##    Data: class_pp
##
## REML criterion at convergence: 23529.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.7665 -0.4721  0.0139  0.4686  3.6080
##
## Random effects:
##  Groups           Name        Variance Std.Dev.
##  childid.schoolid (Intercept) 725.12   26.928
##  schoolid         (Intercept) 324.81   18.023
##  schoolid.1       year         88.67    9.417
##  Residual                     552.20   23.499
## Number of obs: 2380, groups:  childid:schoolid, 1190; schoolid, 107
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  465.087      2.081 109.946  223.44   <2e-16 ***
## year          57.499      1.370  99.916   41.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr)
## year -0.178
```

a. Generate the BLUPs for the random effects and examine whether the independence between zeta_0 and zeta_1 is reflected in a scatterplot of these two sets of effects. (you should comment)

```
# Insert code to generate BLUPs
zeta_0_2 <- ranef(lm5)$schoolid[,1]
zeta_1_2 <- ranef(lm5)$schoolid[,2]
delta_0 <- ranef(lm5)$childid[,1]

plot(zeta_1_2, zeta_0_2)
```



Response:

The plot of the random effects of the intercept and slope for schools shows some evidence of independence. The points are approximately randomly scattered across different values of of the random slope, showing little to no correlation between the two.

b. Compute V_S(year = 0) and V_S (year = 1). Since there are only two years, this is a form of heteroscedasticity in the random effects.

- V_S(year = 0) $= \sigma^2_{\zeta_0} + 0^2\sigma^2_{\zeta_1} = \sigma^2_{\zeta_0} = 324.81$
- V_S(year = 1) $= \sigma^2_{\zeta_0} + 1^2\sigma^2_{\zeta_1} = \sigma^2_{\zeta_0} + \sigma^2_{\zeta_1} = 324.81 + 88.67 = 413.48$

i. In which year is there more between school variation, net of all else, **(you should comment)**?

   Response: In year 1 there is more between school variation.

Question 11

If you ran the model BY YEAR, and removed the year trend from the model, would you get the same estimates for the variances between schools? **(you should comment)* *

```
# Insert code to fit the two models by year and print out the summary
```

Response:

Question 12

Rerun the last nested longitudinal model, allowing correlation between intercept and slope.

a. Is the correlation significant? (you should comment)

```
lm6 <- lmerTest:: lmer(math ~ year + (year | schoolid) + (1 | childid), data = class_pp)
summary(lm6)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
```

```
## lmerModLmerTest]
## Formula: math ~ year + (year | schoolid) + (1 | childid)
##    Data: class_pp
##
## REML criterion at convergence: 23520.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.7030 -0.4686  0.0066  0.4669  3.5142
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  childid  (Intercept) 728.0    26.98
##  schoolid (Intercept) 370.6    19.25
##           year        109.1    10.44    -0.45
##  Residual             547.0    23.39
## Number of obs: 2380, groups:  childid, 1190; schoolid, 107
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  465.099      2.188 102.918  212.60   <2e-16 ***
## year          57.668      1.440  94.572   40.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr)
## year -0.439
```

```r
anova(lm5, lm6, refit = F)
```

```
## Data: class_pp
## Models:
## lm5: math ~ year + (0 + year | schoolid) + (1 | schoolid/childid)
## lm6: math ~ year + (year | schoolid) + (1 | childid)
##     Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## lm5  6 23541 23576 -11764    23529
## lm6  7 23534 23575 -11760    23520 8.8241      1   0.002973 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

`Response:`

The correlation is significant, suggesting that we need to add the correlation between the random slope and intercept for year varying by schools.

b. Compute V_S (year = 0) and V_S(year = 1) for this new model (your formula should include covariance terms).

- V_S(year = 0) $= \sigma^2_{\zeta_0} + 0^2\sigma^2_{\zeta_1} + 2 \cdot 0 \cdot Cov(\sigma^2_{\zeta_0}, \sigma^2_{\zeta_1}) = \sigma^2_{\zeta_0} = 370.6$
- V_S(year = 1) $=$

$$\sigma^2_{\zeta_0} + 1^2 \cdot \sigma^2_{\zeta_1} + 2 \cdot 1 \cdot Cov(\sigma^2_{\zeta_0}, \sigma^2_{\zeta_1}) =$$
$$\sigma^2_{\zeta_0} + \sigma^2_{\zeta_1} + 2 \cdot \sigma_{\zeta_0} \cdot \sigma_{\zeta_1} \cdot \rho(\sigma^2_{\zeta_0}, \sigma^2_{\zeta_1}) =$$
$$370.6 + 109.1 + 2 * (19.25) * (10.44) * (-0.45) = 298.827$$

i. Is this result (and thus model) more consistent with the separate grade analysis? You are implicity testing model fit here. **(you should comment)**

Response: