Stat 414 - Day 15
Longitudinal Models (Ch. 15)

Previously

Have data nested within groups. Want to include the grouping variable in the model. Including it as random intercepts gives us a multilevel model, which has advantages including

- Allows separation of within group and between group variation
- Allows for inclusion of Level 1 and Level 2 variables
- Induces/Estimates within group correlation
- Including random slopes models heterogeneous responses (Level 2)
- Including cross-level interactions can explain variation in slopes (Level 2 equation)
- Does not require equal group sizes/handles missing values well

Multilevel models are especially helpful for "longitudinal data" (e.g., repeat observations on the same individual over time). Typically with longitudinal data we want to focus on changes over time and the effect of Level 2 variables. (We've actually already been looking at repeated measures data, but you will see some different terminology come up.)

Example 1: Minnesota schools

Data were collected by the Minnesota Department of Education for all Minnesota schools during the years 2008-2010 to compare charter and non-charter schools. School performance is measured by the mean score on the math portion of the Minnesota Comprehensive Assessment (MCA-II) data for the 6th grade students enrolled in 618 different Minnesota schools during the years 2008, 2009, and 2010. (MCA test scores for sixth graders are scaled to fall between 600 and 700, where scores above 650 for individual students indicate "meeting standards." Thus, schools with averages below 650 will often have increased incentive to improve their scores the following year.)

Explore the data

(a) Identify the Level 1 and Level 2 units.

Level 1 units = measurement occassions; Lever 2 units = schools

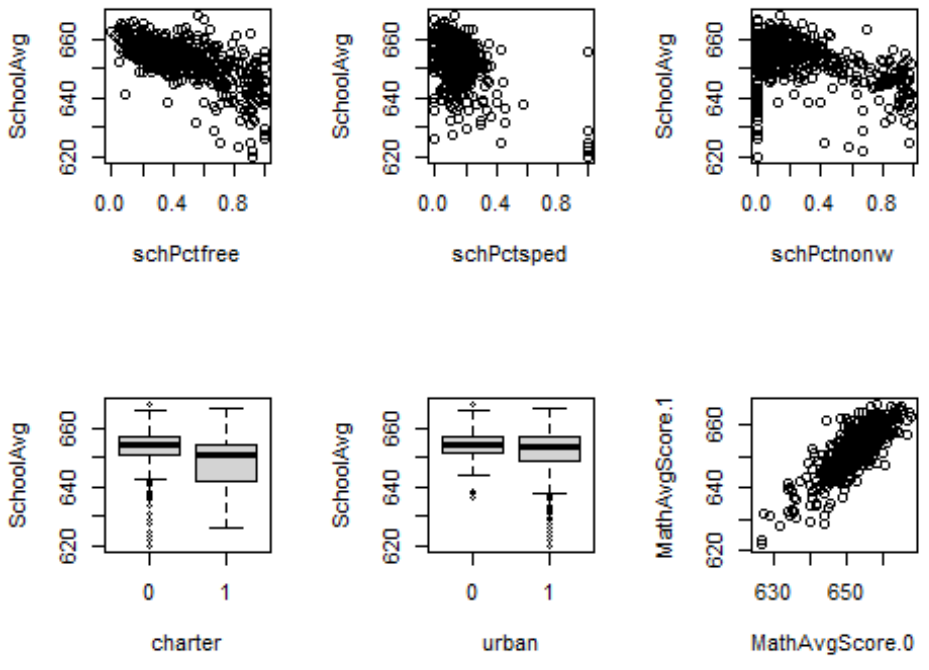
First we want to explore how MCA math test scores relate to important Level 2 variables. This can be done using the data values for all three years or by averaging the data values for the three years into one number, or by using the 2010 values.

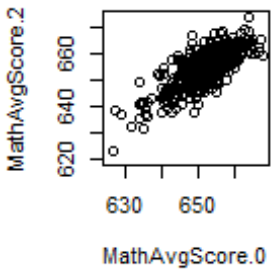
(b) What assumption is made by these last 2 approaches? Reasonable?

That year to year variation in the values is small enough to ignore them. Probabiilty a reasonable assumption. In general, we will want to classify variables as time varying (level 1) or time invariant (level 2).

For the second approach, open the "wide format" of the data (chart.wide.txt, this includes three columns for the three time points for each school) and use the SchoolAvg variable as the response. Examine the associations of these variable with several of the Level 2 variables.

```
  schoolID                   schoolid                          schoolName urban
1     1141    Dtype 1 Dnum 704 Snum 2        A.I. JEDLICKA MIDDLE SCHOOL     1
2     1636  Dtype 7 Dnum 4073 Snum 10 ACADEMIA CESAR CHAVEZ CHARTER SCH.     1
3      362 Dtype 1 Dnum 2396 Snum 150     ACGC ELEMENTARY GRADES 5 AND 6     0
4     1582  Dtype 7 Dnum 4018 Snum 10           ACHIEVE LANGUAGE ACADEMY     1
5      959  Dtype 1 Dnum 625 Snum 410            ADAMS MAGNET ELEMENTARY     1
6      821   Dtype 1 Dnum 511 Snum 15                      ADRIAN MIDDLE     0
  charter schPctnonw schPctsped schPctfree MathAvgScore.0 MathAvgScore.1
1       0    0.01600     0.1040     0.2320          651.1          650.3
2       1    0.00000     0.1429     0.9286          634.5          640.1
3       0    0.01667     0.1333     0.4833          652.3          647.5
4       1    0.91111     0.2000     0.8889          646.4          649.3
5       0    0.67105     0.1053     0.4474          654.0          651.5
6       0    0.05556     0.1389     0.4444          649.7          651.0
  MathAvgScore.2 SchoolAvg
1          653.9     651.8
2          640.3     638.3
3          655.4     651.7
4          650.6     648.8
5          650.0     651.8
6          658.2     653.0
                 MathAvgScore.0 MathAvgScore.1 MathAvgScore.2
MathAvgScore.0           1.0000         0.8064         0.7727
MathAvgScore.1           0.8064         1.0000         0.8331
MathAvgScore.2           0.7727         0.8331         1.0000
```





(c) Which variable(s) seem(s) most useful in predicting the average math score?

percentage free lunch, maybe percentage nonwhite (but check out the zeros virst), maybe percentage special ed (but check out the zeros and ones first), charter vs. public

Now open the "long format" of the data.

```
 [1] "obsNum"       "distschNum"   "year08"       "districtType" "MathAvgScore"
 [6] "MonPerChild"  "schoolName"   "schPctnonw"   "schPctsped"   "schPctfree"
[11] "city"         "urban"        "charter"      "schoolnum"
  obsNum              distschNum year08 districtType MathAvgScore MonPerChild
1      1   Dtype 1 Dnum 1 Snum 2      0            1        652.8        8000
2      2   Dtype 1 Dnum 1 Snum 2      1            1        656.6        8266
3      3   Dtype 1 Dnum 1 Snum 2      2            1        652.6        8119
4      4 Dtype 1 Dnum 100 Snum 1      0            1        646.9        7682
5      5 Dtype 1 Dnum 100 Snum 1      1            1        645.3        8511
6      6 Dtype 1 Dnum 100 Snum 1      2            1        651.9        8357
             schoolName schPctnonw schPctsped schPctfree      city urban
1 RIPPLESIDE ELEMENTARY     0.0000     0.1176     0.3627    Aitkin     0
2 RIPPLESIDE ELEMENTARY     0.0000     0.1176     0.3627    Aitkin     0
3 RIPPLESIDE ELEMENTARY     0.0000     0.1176     0.3627    Aitkin     0
4  WRENSHALL ELEMENTARY     0.0303     0.1515     0.4242 Wrenshall     0
5  WRENSHALL ELEMENTARY     0.0303     0.1515     0.4242 Wrenshall     0
6  WRENSHALL ELEMENTARY     0.0303     0.1515     0.4242 Wrenshall     0
  charter schoolnum
1       0         1
2       0         1
3       0         1
4       0         2
5       0         2
6       0         2
```

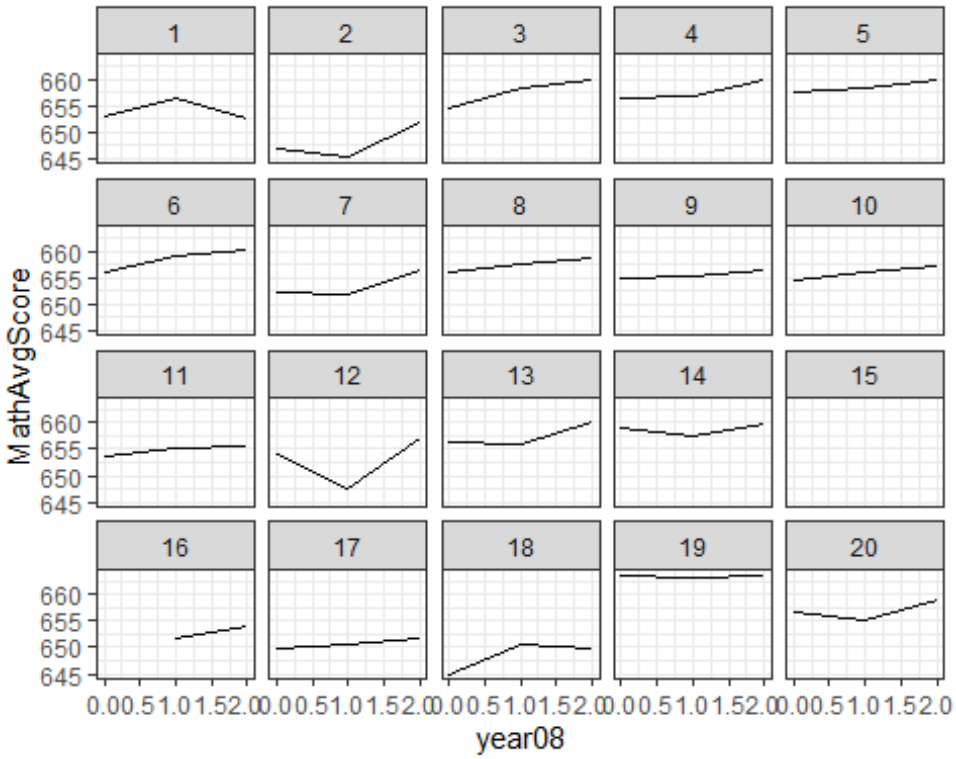
(d) Explain what year08 represents.

The number of years since 2008

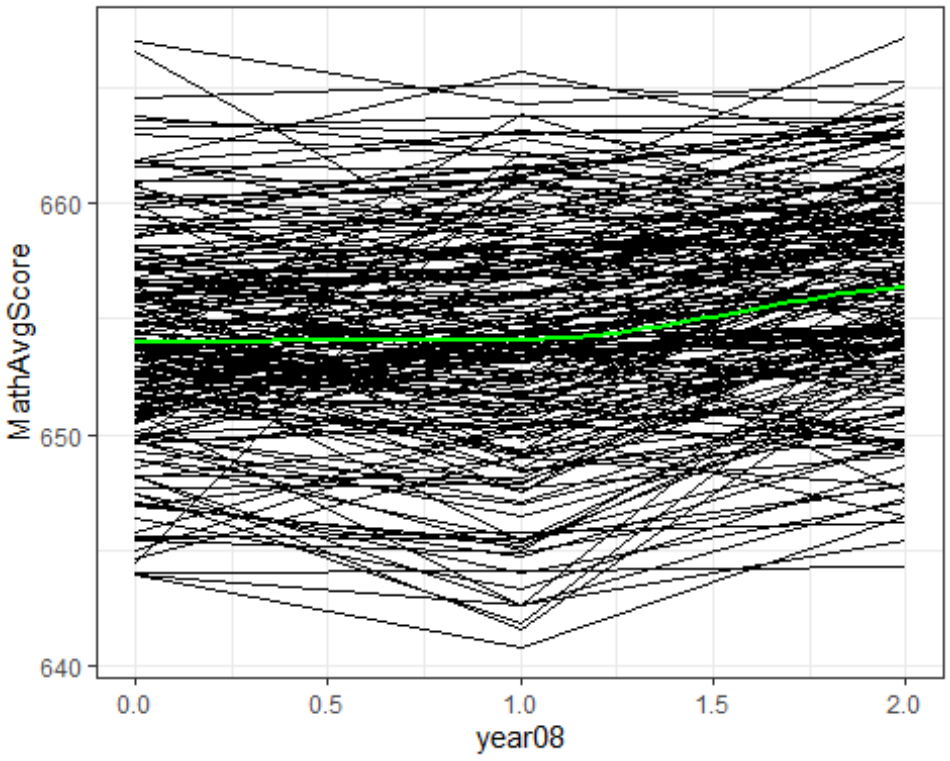
Create two visual representations of math scores vs. time for the first 20 schools:

- separate graphs for each school
- connecting lines or smoothers for each school overlaid on same graph (i.e., "spaghetti plot")

```
#   separate graphs for each school (first 20 schools = 60 observations)
ggplot(data = chart_long[1:57,], aes(y = MathAvgScore, x = year08)) +
  geom_line() +
  facet_wrap(~schoolnum) +
  theme_bw()
```




```
#   connecting lines or smoothers for each school (1st 200) overlaid on same graph
(i.e., "spaghetti plot")
ggplot(data = chart_long[1:600,], aes(y = MathAvgScore, x = year08, group = schooln
um)) +
         geom_line() +
        geom_smooth(aes(group=1),color="green",size=1, se=F) +
         theme_bw()
```

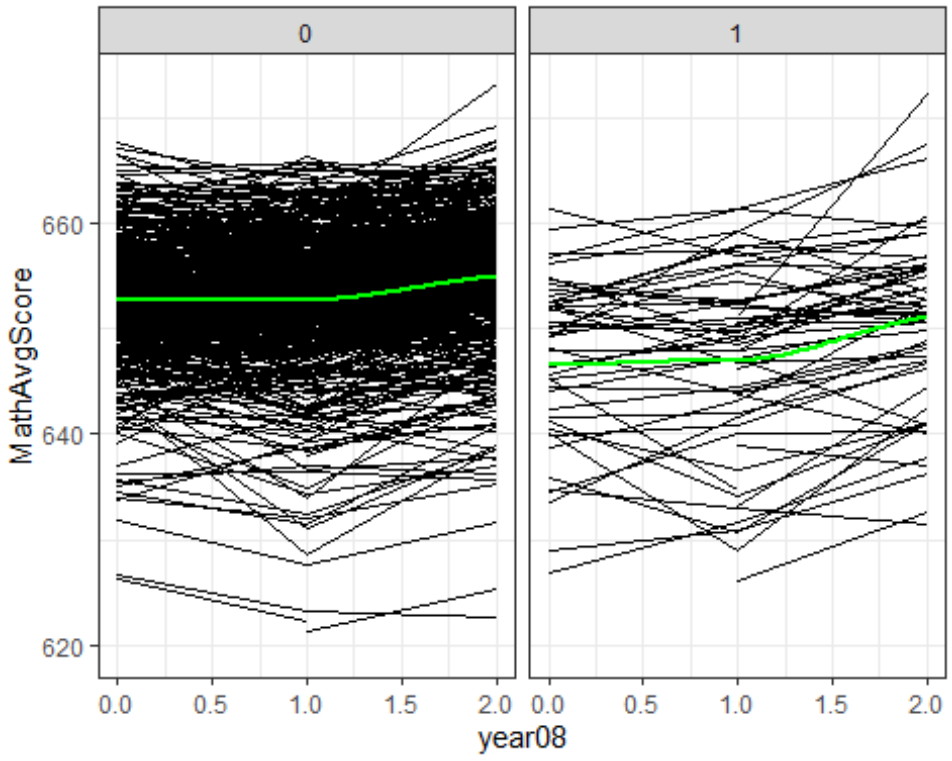


(e) Does the change over time in the average math score look linear? Does it look like we will want to include random intercepts? (Meaning?) Does it look like we will want to include random slopes? (Meaning?)

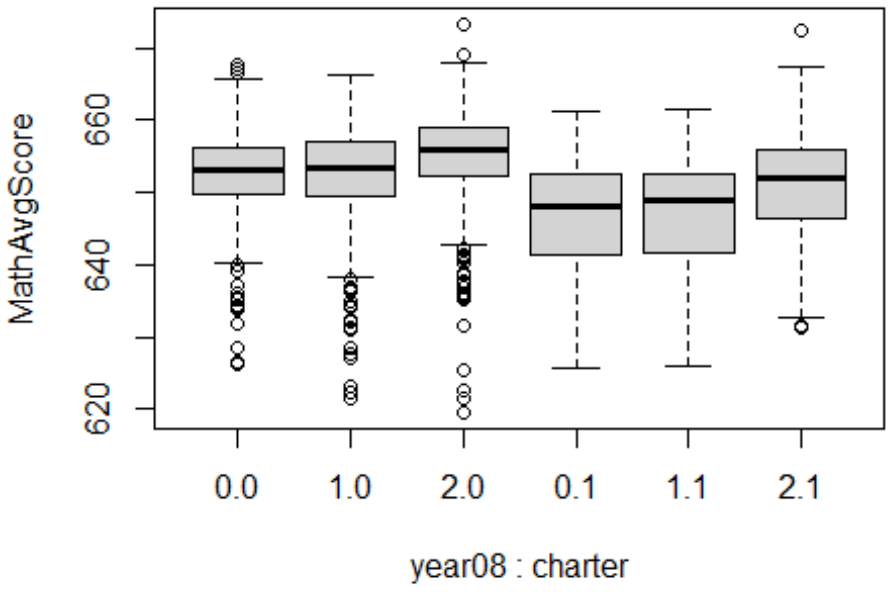
Yes to random intercepts: the school averages vary a lot in 2008. Maybe not random slopes as the rate of change from year to year doesn't seem to change as much as school to school.

Produce a graph of the Math scores vs. year, separated by the charter (charter = 1) vs. public (charter = 0) schools.

```
#charter = 1, public = 0
ggplot(data = chart_long, aes(y = MathAvgScore, x = year08)) +
  geom_line(aes(group=schoolnum)) +
  facet_wrap(.~charter) +
  geom_smooth(aes(group=1),color="green",size=1, method="loess", se = F) +
  theme_bw()
```




```
#SEE ALSO
boxplot(MathAvgScore ~ year08*charter, data = chart_long)
```



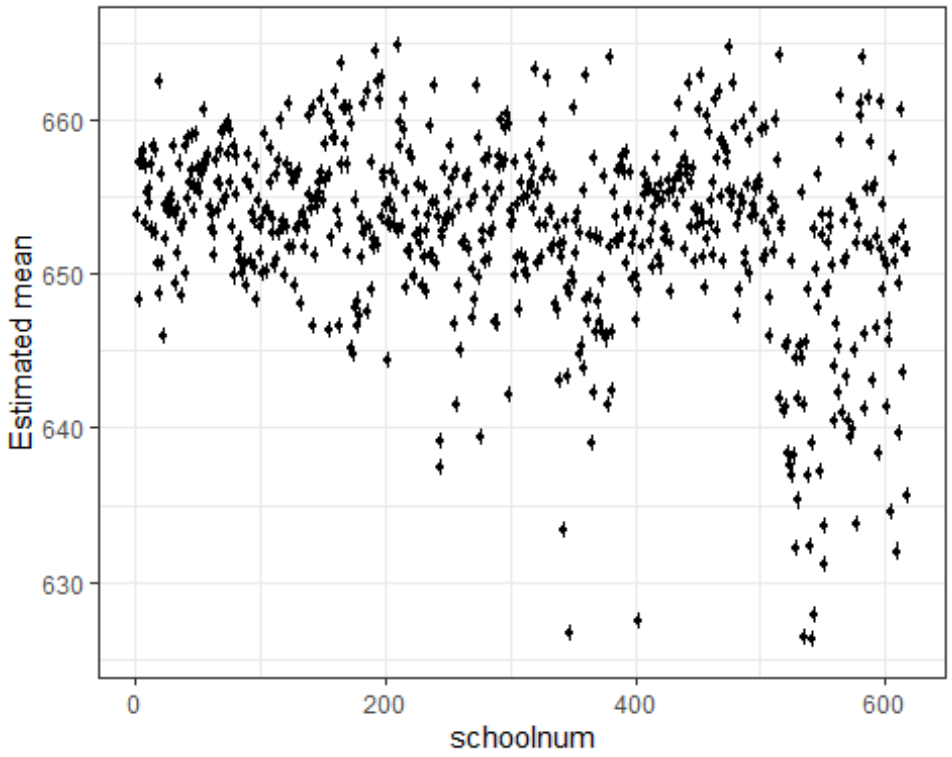
(f) What do you learn?

Lower 2008 averae scores for charter schools but maybe steeper slopes (year to year growth) than for public schools.

Modelling

Fit the null model.

```
#null model - using lmer
model0 = lmer(MathAvgScore ~ 1 + (1 | schoolnum), data = chart_long); summary(model
0)
Linear mixed model fit by REML ['lmerMod']
Formula: MathAvgScore ~ 1 + (1 | schoolnum)
   Data: chart_long

REML criterion at convergence: 10530

Scaled residuals:
   Min     1Q Median     3Q    Max
-3.157 -0.495  0.014  0.513  3.569

Random effects:
 Groups    Name        Variance Std.Dev.
 schoolnum (Intercept) 41.9     6.47
 Residual              10.6     3.25
Number of obs: 1733, groups:  schoolnum, 618

Fixed effects:
            Estimate Std. Error t value
(Intercept)  652.746      0.273    2395
performance::icc(model0)
# Intraclass Correlation Coefficient

    Adjusted ICC: 0.798
  Unadjusted ICC: 0.798
pred <- ggpredict(model0, terms = "schoolnum [all]", type = "random")
pred_df <- as.data.frame(pred)
ggplot(pred_df, aes(x = x, y = predicted)) +
  geom_point(size = 1) +
  geom_errorbar(aes(ymin = conf.low, ymax = conf.high), width = 0) +
  labs(y = "Estimated mean", x = "schoolnum") +
  theme_bw()
```



(g) What is the ICC for these data? What does this tell you? Does this model adequately capture the behavior of our longitudinal data?

ICC = 0.798, so about 80% of variation in 2008 in math scores is between schools rather than within, and the correlation of two observations (measurement occassions) for the same school is 0.80.

```
#Changing to lme to get the variance-covariance matrix of the (predicted) responses
#library(nlme)
model0b = lme(MathAvgScore ~ 1, random = ~ 1 | schoolnum, data = chart_long); summa
ry(model0)
Linear mixed model fit by REML ['lmerMod']
Formula: MathAvgScore ~ 1 + (1 | schoolnum)
   Data: chart_long

REML criterion at convergence: 10530

Scaled residuals:
   Min     1Q Median     3Q    Max
-3.157 -0.495  0.014  0.513  3.569

Random effects:
 Groups    Name        Variance Std.Dev.
 schoolnum (Intercept) 41.9     6.47
 Residual              10.6     3.25
Number of obs: 1733, groups:  schoolnum, 618
```

```
Fixed effects:
            Estimate Std. Error t value
(Intercept)  652.746      0.273    2395
getVarCov(model0b, type = "conditional")
schoolnum 1
Conditional variance covariance matrix
      1     2     3
1 10.57  0.00  0.00
2  0.00 10.57  0.00
3  0.00  0.00 10.57
  Standard Deviations: 3.251 3.251 3.251
getVarCov(model0b, type = "marginal")
schoolnum 1
Marginal variance covariance matrix
      1     2     3
1 52.44 41.87 41.87
2 41.87 52.44 41.87
3 41.87 41.87 52.44
  Standard Deviations: 7.242 7.242 7.242
cov2cor(getVarCov(model0b, type = "marginal")[[1]])
       1      2      3
1 1.0000 0.7984 0.7984
2 0.7984 1.0000 0.7984
3 0.7984 0.7984 1.0000
```

Key Idea

The “exchangeability assumption” assumes the correlation between any two observations in the same cluster are the same. This is often not an appropriate assumption with longitudinal data (measures over time).

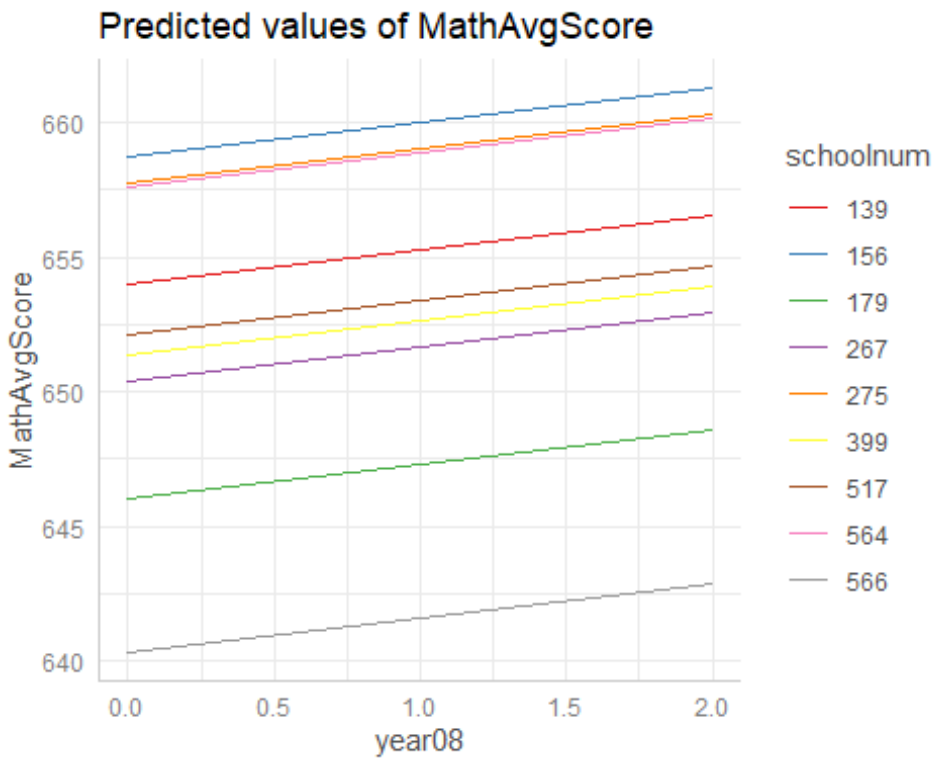
(h) How do we get variances and correlations to change over time?

Random slopes

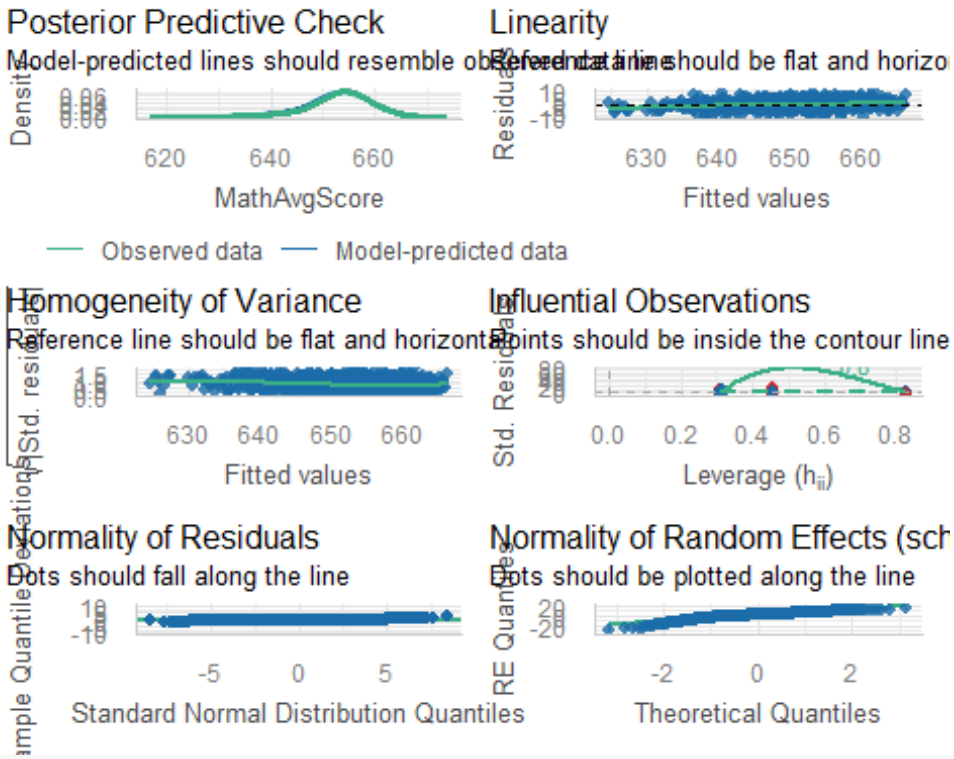
Add Time

```
#using lmer for ggpredict
model1 = lmer(MathAvgScore ~ 1 + year08 + (1 | schoolnum), data = chart_long)
summary(model1, corr=FALSE)
Linear mixed model fit by REML ['lmerMod']
Formula: MathAvgScore ~ 1 + year08 + (1 | schoolnum)
   Data: chart_long

REML criterion at convergence: 10346

Scaled residuals:
   Min     1Q Median     3Q    Max
-3.168 -0.461  0.014  0.471  3.602

Random effects:
```

```
 Groups    Name        Variance Std.Dev.
 schoolnum (Intercept) 42.86    6.55
 Residual               8.91    2.98
Number of obs: 1733, groups:  schoolnum, 618

Fixed effects:
            Estimate Std. Error t value
(Intercept) 651.3898     0.2895  2250.1
year08        1.2799     0.0895    14.3
performance::r2(model1, by_group = TRUE)
# Explained Variance by Level

Level     |     R2
------------------
Level 1   |  0.158
schoolnum | -0.024
#within school variance decreased from 10.57 to 8.906 (10.57 - 8.906)/10.57 = 0.157
#library(ggeffects)
plot(ggpredict(model1, terms = c("year08", "schoolnum [sample = 9]"), type = "rando
m"), show_ci=FALSE)
```




```
performance::check_model(model1)
```




```
#usine lme for variance-covariance matrix
model1b = lme(MathAvgScore ~ 1 + year08, random = ~ 1  | schoolnum, data = chart_lo
ng); summary(model1b)
Linear mixed-effects model fit by REML
  Data: chart_long
    AIC   BIC logLik
  10354 10375  -5173

Random effects:
 Formula: ~1 | schoolnum
        (Intercept) Residual
StdDev:       6.547    2.984

Fixed effects:  MathAvgScore ~ 1 + year08
            Value Std.Error   DF t-value p-value
(Intercept) 651.4    0.2895 1114  2250.1       0
year08        1.3    0.0895 1114    14.3       0
 Correlation:
       (Intr)
year08 -0.326

Standardized Within-Group Residuals:
     Min       Q1      Med       Q3      Max
-3.16755 -0.46088  0.01414  0.47079  3.60218

Number of Observations: 1733
Number of Groups: 618
```

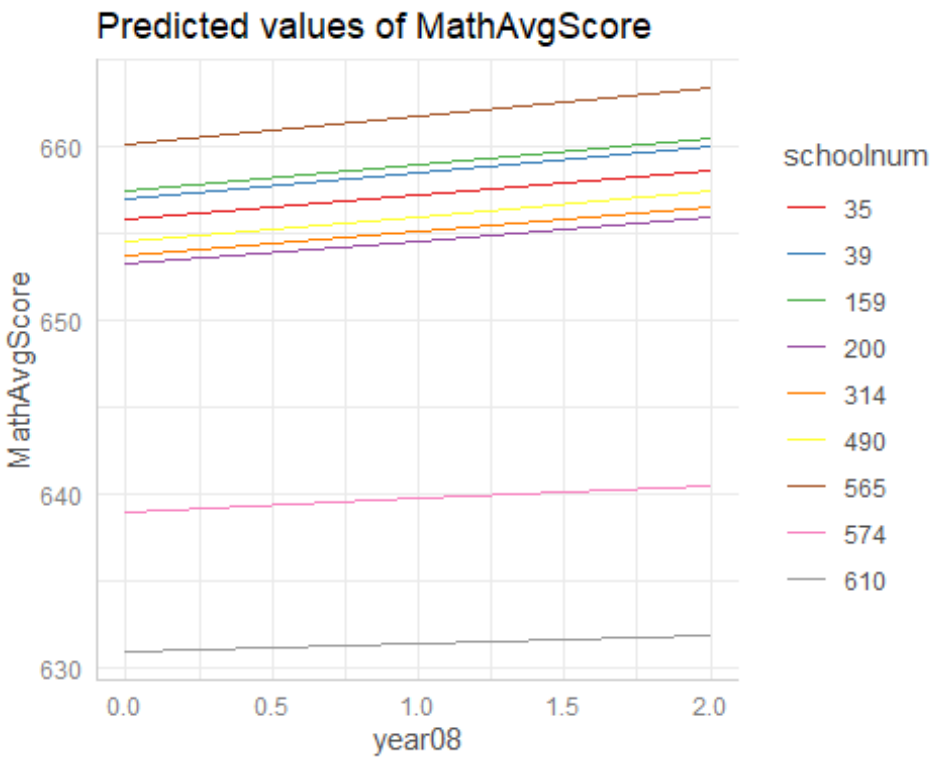
```
cov2cor(getVarCov(model1b, type = "marginal")[[1]])
      1     2     3
1 1.000 0.828 0.828
2 0.828 1.000 0.828
3 0.828 0.828 1.000
```

Unconditional growth model

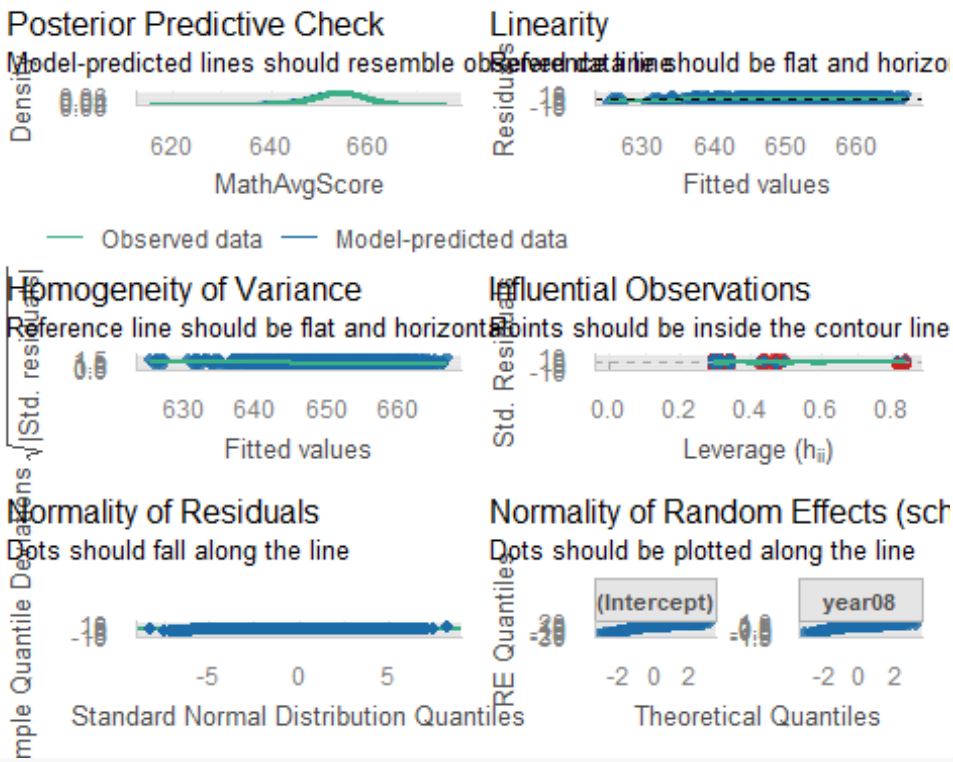
Adding time is not enough to model heterogeneity, need random slopes… With longitudinal data, we usually we start with the "unconditional growth model" (time is only Level 1 variable, we haven't "conditioned" or "controlled" for any other possible covariates): multilevel model with year08, random intercepts, and random slopes. (Be sure to use schoolnum, which are unique, not school name):

$mathscore_{ij} = \beta_{0j} + \beta_{1j} year08_{ij} + \epsilon_{ij}$ where $\epsilon_{ij} \sim N(0, \sigma^2)$

```
#using lmer for ggpredict
#replacing earlier model1
model1 = lmer(MathAvgScore ~ 1 + year08 + (1 + year08| schoolnum), data = chart_lon
g)
summary(model1, corr=FALSE)
Linear mixed model fit by REML ['lmerMod']
Formula: MathAvgScore ~ 1 + year08 + (1 + year08 | schoolnum)
   Data: chart_long

REML criterion at convergence: 10340

Scaled residuals:
   Min     1Q Median     3Q    Max
-3.158 -0.467  0.018  0.460  3.497

Random effects:
 Groups    Name        Variance Std.Dev. Corr
 schoolnum (Intercept) 39.441   6.280
           year08       0.111   0.332    0.72
 Residual               8.820   2.970
Number of obs: 1733, groups:  schoolnum, 618

Fixed effects:
            Estimate Std. Error t value
(Intercept)  651.408      0.279  2332.0
year08         1.265      0.090    14.1
#within school variance decreased from 10.57 to 8.906 (10.57 - 8.82)/10.57 = 0.166

#library(ggeffects)
plot(ggpredict(model1, terms = c("year08", "schoolnum [sample = 9]"), type = "rando
m"), show_ci=FALSE)
```




```
performance::check_model(model1)
```




```
#usine lme for variance-covariance matrix
model1b = lme(MathAvgScore ~ 1 + year08, random = ~ 1 + year08 | schoolnum, data =
chart_long); summary(model1b)
```

```
Linear mixed-effects model fit by REML
  Data: chart_long
    AIC   BIC logLik
  10352 10384  -5170

Random effects:
 Formula: ~1 + year08 | schoolnum
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev Corr
(Intercept) 6.280  (Intr)
year08      0.332  0.724
Residual    2.970

Fixed effects:  MathAvgScore ~ 1 + year08
            Value Std.Error   DF t-value p-value
(Intercept) 651.4   0.27934 1114  2331.9       0
year08        1.3   0.08997 1114    14.1       0
 Correlation:
       (Intr)
year08 -0.234

Standardized Within-Group Residuals:
     Min       Q1      Med       Q3      Max
-3.15786 -0.46676  0.01796  0.45956  3.49726

Number of Observations: 1733
Number of Groups: 618
cov2cor(getVarCov(model1b, type = "marginal")[[1]])
       1      2      3
1 1.0000 0.8223 0.8261
2 0.8223 1.0000 0.8332
3 0.8261 0.8332 1.0000
```

(i) Describe what this model is doing. What assumptions does this model make about the "occasion-specific" residuals? Does that seem like a reasonable assumption in this context? Interpret the variance components (and covariance). What can you tell me about the populations of intercepts and slopes? How would you determine the percentage of within-school variation explained by the linear increase over time? How else can we evaluate the model?

The model assumes that once you account for the multiple time points the observations are now independent ('conditional indepencen'). We note there is much mroe school to school variation in intercepts than in slopes. We assume the population distribution of intercepts across schools is normally distribution with mean 651.41 and variance 39.44. We assume the population distribution of slopes across schools is normally distribution with mean 1.265 and variance 0.1105. Positive covariance between intercepts and slopes: schools with higher avg scores in 2008 tend to have larger increases in avg score year to year than schools with lower 2008 avg scores. If we compare this model to the null model, $\hat{\sigma}^2$ has decreased from 10.57 to 8.82 (16%). The residual plots look ok though a few influential observations to check out. Could also look more into the normality of the random effects.

Adding a Level 2 Variable

Include charter (charter schools = 1, public schools = 0) as a Level 2 variable (for both level equations, i.e., fixed effect and cross-level interaction with *year08*).

```
model2 = lmer(MathAvgScore ~ year08 + charter + charter:year08 + (year08 | schoolnu
m), data = chart_long);summary(model2, corr=FALSE)
Linear mixed model fit by REML ['lmerMod']
Formula: MathAvgScore ~ year08 + charter + charter:year08 + (year08 |
    schoolnum)
   Data: chart_long

REML criterion at convergence: 10292

Scaled residuals:
   Min     1Q Median     3Q    Max
-3.193 -0.471  0.013  0.466  3.459

Random effects:
 Groups    Name        Variance Std.Dev. Corr
 schoolnum (Intercept) 35.832   5.986
           year08       0.131   0.362    0.88
 Residual               8.784   2.964
Number of obs: 1733, groups:  schoolnum, 618

Fixed effects:
               Estimate Std. Error t value
(Intercept)    652.0584     0.2845 2292.00
year08           1.1971     0.0943   12.70
charter         -6.0184     0.8656   -6.95
year08:charter   0.8557     0.3143    2.72
```

(j) Summarize the charter effect on the intercepts and the charter effect on the slopes. (Consistent with the graphs above?) Is either statistically significant? (Be very clear how you are deciding.) How much school-to-school variation in the intercepts has been explained by the charter school variable? What about the slopes?

In 2008, charter schools averaged lower average scores (by 6.018) than the non-charter schools (consistent with the lower intercept of the greem model in the graph). Charter schools have a larger rate of increase in average scores over the 3 years than public schools, on average. The difference between charter and public schools decreases from 2008 to 2019 (on average). About 9% of the variation in intercepts is explained, but variation in slopes actually increases. Total variance goes from 39.44 + .11 + 8.82 = 48.37 to (35.83 + .13 + 8.78) = 44.74)

Relaxing the linearity assumption

The graphs of average scores over time indicated that there appeared to potentially be a nonlinear trend. There are many ways to relax the linearity assumption but first we will just consider a quadratic effect of time.

(k) If we plan to use $time$ and $time^2$, do we need to center time first?

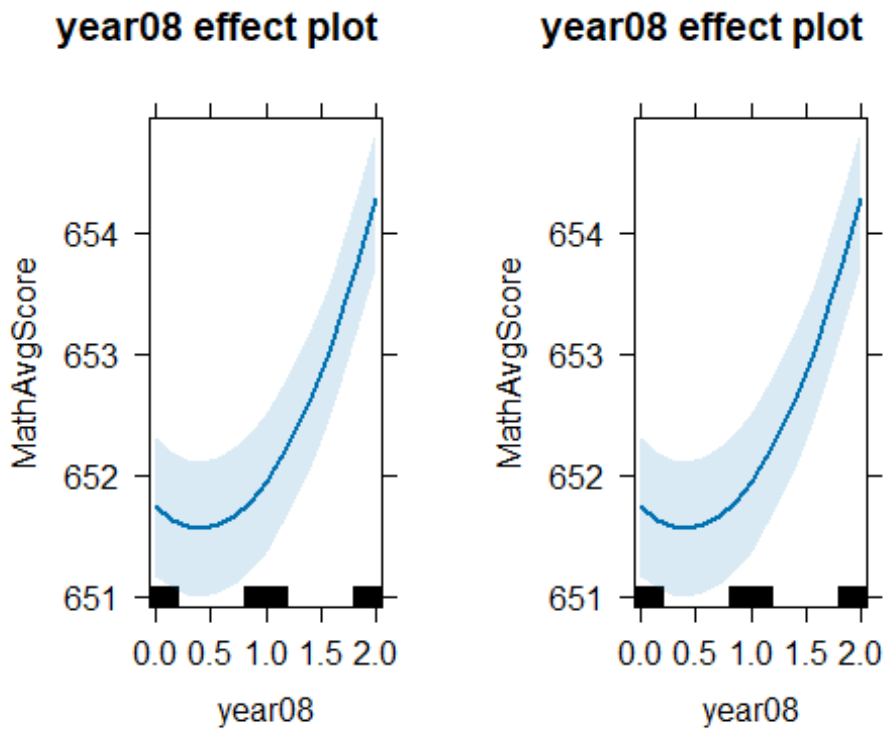
Not necessary here, is pretty much close to centered already.

(l) Write out the Level 1 and Level 2 equations that include $time$ and $time^2$, but only allow the intercepts to vary. What assumptions is this "unconditional quadratic growth model" imposing on the time trends?

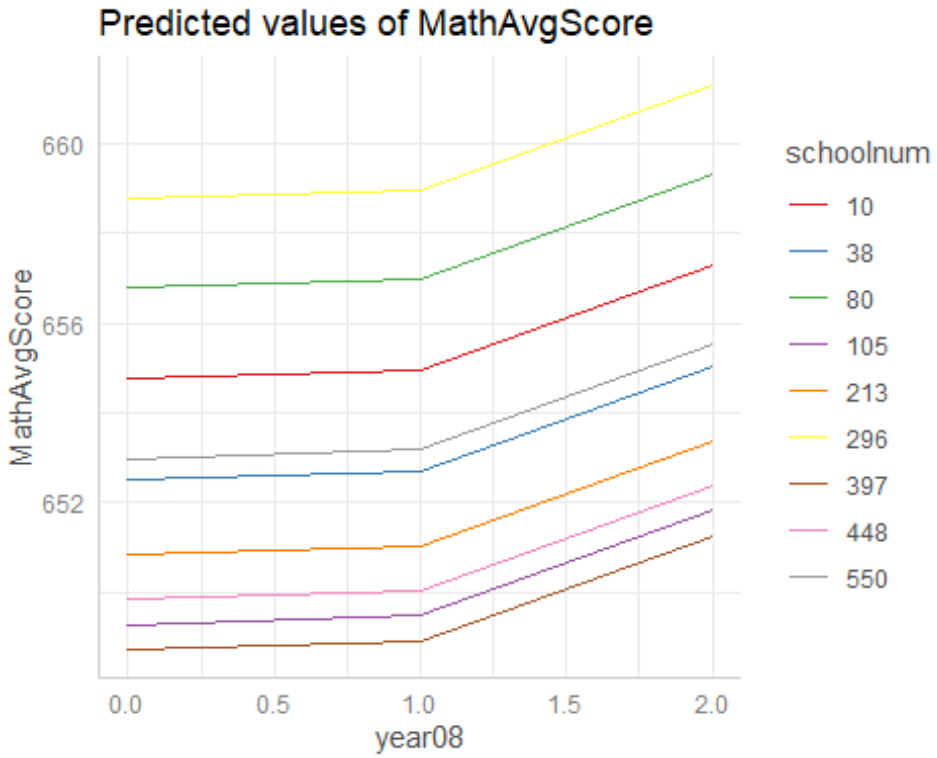
$y_{ij} = \beta + 0j + \beta_1 time_{ij} + \beta_2 time_{ij}^2 + \epsilon_{ij}$ and $\beta_{0j} = \beta_{00} + u_{0j}$. We are assuming each school has the same quadratic trend, just different intercepts.

Fit and interpret the model specified in (l).

```
summary(quadmodel <- lmer(MathAvgScore ~ 1 + year08 + I(year08^2) + (1 | schoolnum)
, data=chart_long), corr=F)
Linear mixed model fit by REML ['lmerMod']
Formula: MathAvgScore ~ 1 + year08 + I(year08^2) + (1 | schoolnum)
   Data: chart_long

REML criterion at convergence: 10298

Scaled residuals:
   Min     1Q Median     3Q    Max
-3.213 -0.476  0.009  0.469  3.495

Random effects:
 Groups    Name        Variance Std.Dev.
 schoolnum (Intercept) 43.05    6.56
 Residual               8.52    2.92
Number of obs: 1733, groups:  schoolnum, 618

Fixed effects:
            Estimate Std. Error t value
(Intercept)  651.741      0.293 2222.75
year08        -0.867      0.315   -2.75
I(year08^2)    1.068      0.150    7.10
logLik(quadmodel)
'log Lik.' -5149 (df=5)
library(effects)
plot(allEffects(quadmodel))
```




```
plot(ggpredict(quadmodel, terms = c("year08", "schoolnum [sample = 9]"), type = "ra
ndom"), show_ci=FALSE)
```



(m) Is the quadratic effect statistically significant? How do you interpret the sign of the coefficient of this term?

Yes, t = 7.101. The coefficients are negative then positive, indicatign a dip from 2008 to 2009 and then an increase in 2010 (like an interaction with time, the improvement over time increases with time).

Consider the Level 1 and Level 2 equations that include $time$ and $time^2$, allowing the slopes and intercepts to vary, but with no Level 2 covariates.

```
#summary(testmodel <- lmer(MathAvgScore ~ 1 + year08 + I(year08^2) + (1 + year08 +
I(year08^2) | schoolnum), data=chart_long))
```

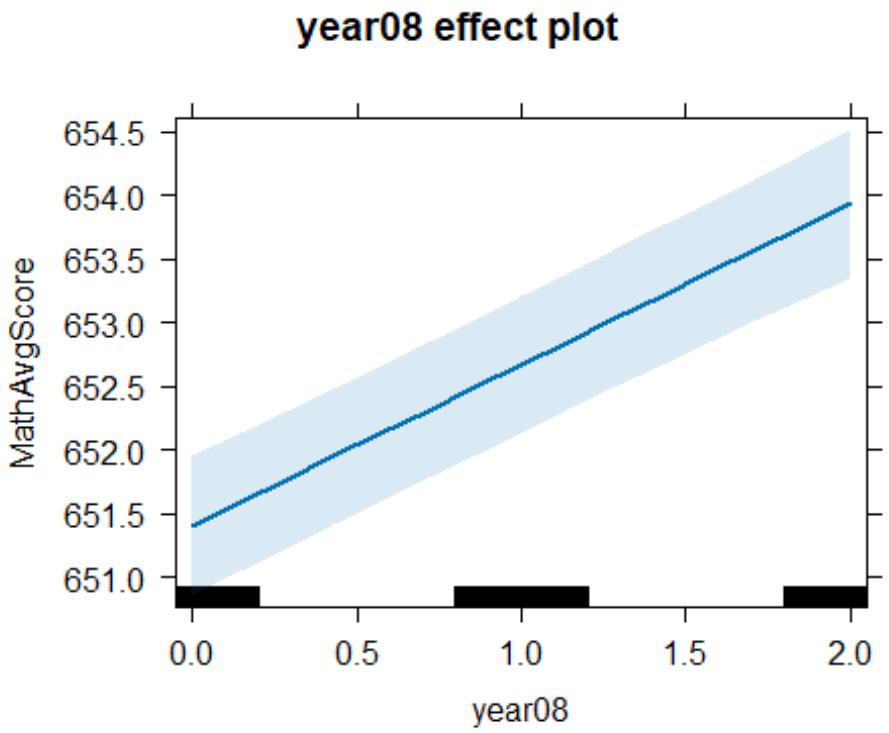
(n) Can we fit the suggested model?

No because we only have 3 data values per school, so we don't have enough degrees of freedom to fit a different quadratic model per school

Computer problem 15 - due Monday, 7am

Compare the quadratic model to the model that is only linear in time, but with random slopes.

```
summary(linearmodel <- lmer(MathAvgScore ~ 1 + year08  + (1 + year08 | schoolnum),
data=chart_long))
Linear mixed model fit by REML ['lmerMod']
Formula: MathAvgScore ~ 1 + year08 + (1 + year08 | schoolnum)
   Data: chart_long

REML criterion at convergence: 10340

Scaled residuals:
   Min     1Q Median     3Q    Max
-3.158 -0.467  0.018  0.460  3.497

Random effects:
 Groups    Name        Variance Std.Dev. Corr
 schoolnum (Intercept) 39.441   6.280
           year08       0.111   0.332    0.72
 Residual               8.820   2.970
Number of obs: 1733, groups:  schoolnum, 618

Fixed effects:
            Estimate Std. Error t value
(Intercept)  651.408      0.279  2332.0
year08         1.265      0.090    14.1

Correlation of Fixed Effects:
       (Intr)
year08 -0.234
library(effects)
plot(allEffects(linearmodel), lines = T)
```




```
anova(quadmodel, linearmodel)
Data: chart_long
Models:
quadmodel: MathAvgScore ~ 1 + year08 + I(year08^2) + (1 | schoolnum)
linearmodel: MathAvgScore ~ 1 + year08 + (1 + year08 | schoolnum)
            npar   AIC   BIC logLik -2*log(L) Chisq Df Pr(>Chisq)
quadmodel      5 10302 10330  -5146     10292
linearmodel    6 10348 10381  -5168     10336     0  1          1
```

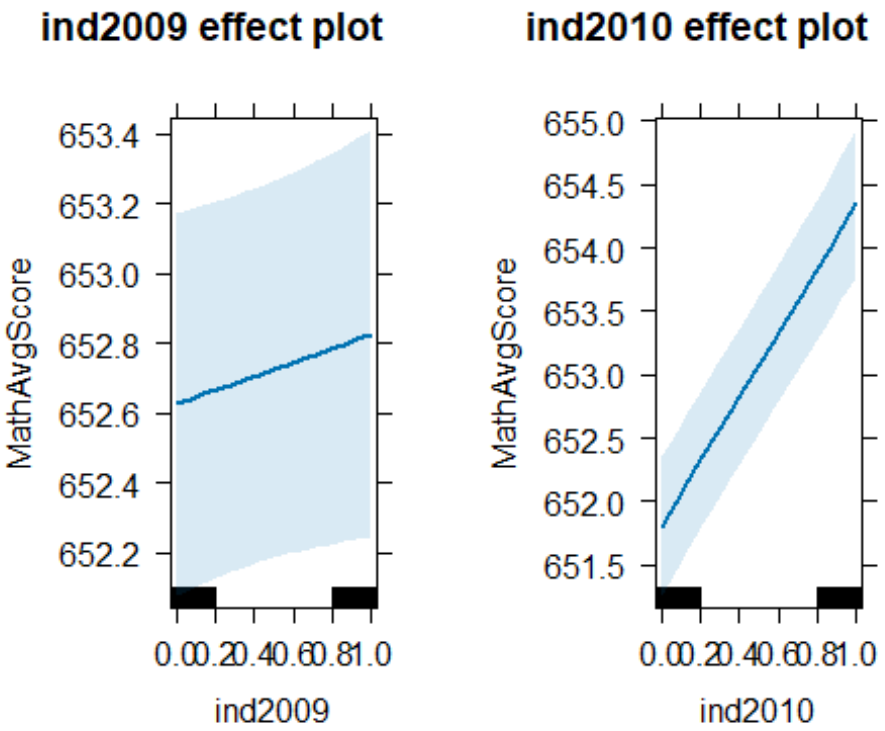
(a) Is it ok to do a likelihood ratio test here? How many parameters are estimated by each model? How do the AIC/BIC values compare? Which model do you recommend?

Another option is a *piecewise function.* With three time points this means we allow one slope from 2008 to 2009 and a different slope from 2009 to 2010. Create an indicator variable for 2009 and another for 2010. Include these two indicator variables (but not year08) in the model, with random intercepts (only).

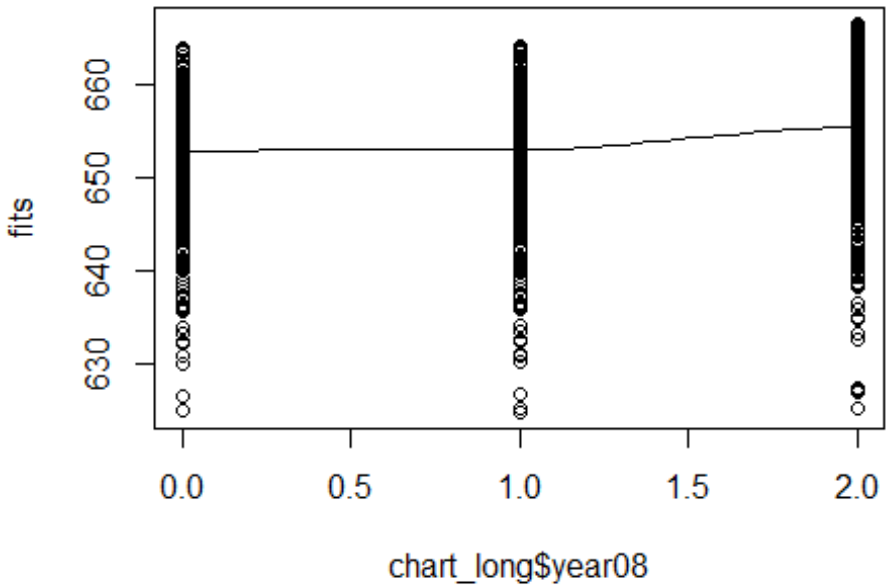
```
head(chart_long$year08)
[1] 0 1 2 0 1 2
chart_long$ind2009 = as.numeric(chart_long$year08 == 1)
head(chart_long$ind2009)
[1] 0 1 0 0 1 0
chart_long$ind2010 = as.numeric(chart_long$year08 == 2)
head(chart_long$ind2010)
[1] 0 0 1 0 0 1
```

(b) Why do the previous commands work?

```
piecemodel = lmer(MathAvgScore ~ ind2009 + ind2010 + (1 | schoolnum), data = chart_
long)
summary(piecemodel, corr=F)
Linear mixed model fit by REML ['lmerMod']
Formula: MathAvgScore ~ ind2009 + ind2010 + (1 | schoolnum)
   Data: chart_long

REML criterion at convergence: 10297

Scaled residuals:
   Min     1Q Median     3Q    Max
-3.213 -0.476  0.009  0.469  3.495

Random effects:
 Groups    Name        Variance Std.Dev.
 schoolnum (Intercept) 43.05    6.56
 Residual               8.52    2.92
Number of obs: 1733, groups:  schoolnum, 618

Fixed effects:
            Estimate Std. Error t value
(Intercept)  651.741      0.293 2222.75
ind2009        0.202      0.175    1.15
ind2010        2.540      0.175   14.50
#library(effects)
plot(allEffects(piecemodel))
```



```
fits = fitted.values(piecemodel)
scatter.smooth(fits ~ chart_long$year08)
```




```
AIC(linearmodel, quadmodel, piecemodel)
            df   AIC
linearmodel  6 10352
quadmodel    5 10308
piecemodel   5 10307
BIC(linearmodel, quadmodel, piecemodel)
            df   BIC
linearmodel  6 10384
quadmodel    5 10335
piecemodel   5 10334
```

(c) How do you interpret the coefficient of ind2010? Compare this model to the quadratic model – does it describe a similar time trend? How so? How do the AIC/BIC values compare?

(d) Give a "modelling" reason to prefer the linear model to the quadratic or piecewise linear models.

Notes:

- Keep in mind the importance of the interpretability of your model, especially to non-statisticians.
- You can also consider functions that allow for "exponential growth"
- Also consider how well your model can extrapolate. It is definitely riskier to extrapolate with quadratic models.

- From Finch and Bolin (2017): Modeling longitudinal data in a multilevel framework has a number of advantages over more traditional methods of longitudinal analysis (e.g. ANOVA designs). For example, using a multilevel approach allows for the simultaneous modeling of both intraindividual change (how an individual changes over time), as well as interindividual change (differences in this temporal change across individuals). A particularly serious problem that afflicts many longitudinal studies is high attrition within the sample. Quite often, it is difficult for researchers to keep track of members of the sample over time, especially over a lengthy period of time. When using traditional techniques for longitudinal data analysis such as repeated measures ANOVA, only complete data cases can be analyzed. Thus, when there is a great deal of missing data, either a sophisticated missing data replacement method (e.g. multiple imputation) must be employed, or the researcher must work with a greatly reduced sample size. In contrast, multilevel models are able to use the available data from incomplete observations, thereby not reducing sample size as dramatically as do other approaches for modeling longitudinal data, nor requiring special missing data methods.