Stat 414 - Day 9
Random Intercepts Models (Ch. 4)

Last Time

- Fixed vs. Random effects: If the categories/labels themselves aren't of much interest, but want to consider the levels of the "grouping variable" as a random sample from some larger population, can treat as random effects.
- $Y_{ij} = \beta_0 + u_j + \epsilon_{ij}$ where we are assuming $\epsilon_{ij} \sim N(0, \sigma^2)$ and $u_j \sim N(0, \tau^2)$ and $cov(\epsilon_{ij}, u_j) = 0$
- Benefits include fewer parameters to estimate and generalizability to larger population of units.
- Also induces a non-zero correlation between two observations from the same Level 2 units (this allows us to model dependence within the groups)
- Results in "partial pooling": An estimated group mean is a weighted average of the observed sample mean and the "overall mean." The degree of shrinkage to the overall mean will depend on the amount of within group variation, between group variation, and sample sizes. This allows the model to "borrow strength" from all the groups, e.g., when estimating a group mean that had a small sample size.
- Can demonstrate that partial pooling tends to give more accurate predictions overall than no pooling or complete pooling

Example 1: Netherlands Language Scores

The Netherlands Language dataset examines language test scores (langPOST) in Grade 8 students (~ age 11) for elementary schools in the Netherlands. (See p. 50 for more information about this dataset.) Students (Level 1) are nested within a random sample of schools (Level 2). Because the schools are a random sample from a larger population, it seems natural to treat them as random effects.

```
neth = read.table("https://www.rossmanchance.com/stat414F20/data/NetherlandsLanguag
e.txt", "\t", header=TRUE)
head(neth)
  schoolnr pupilNR_new langPOST    ses IQ_verb sex Minority denomina sch_ses
1        1           3       46  -4.73    3.13   0        0        1  -14.04
2        1           4       45 -17.73    2.63   0        1        1  -14.04
3        1           5       33 -12.73   -2.37   0        0        1  -14.04
4        1           6       46  -4.73   -0.87   0        0        1  -14.04
5        1           7       20 -17.73   -3.87   0        0        1  -14.04
6        1           8       30 -17.73   -2.37   0        1        1  -14.04
  sch_iqv sch_min
1  -1.404    0.63
2  -1.404    0.63
3  -1.404    0.63
4  -1.404    0.63
5  -1.404    0.63
6  -1.404    0.63
load(url("https://www.rossmanchance.com/iscam4/ISCAM.RData"))
```

Create the null model (using lmer for graph below)

```
#install.packages(lme4)
library(lme4)
nullmodel = lmer(langPOST ~ 1 + (1|schoolnr), data = neth, REML = FALSE)
#using ML to better match the output in the text
summary(nullmodel)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: langPOST ~ 1 + (1 | schoolnr)
   Data: neth

      AIC       BIC    logLik -2*log(L)  df.resid
    26601     26620    -13298     26595      3755

Scaled residuals:
   Min     1Q Median     3Q    Max
-4.185 -0.642  0.091  0.723  2.528

Random effects:
 Groups   Name        Variance Std.Dev.
 schoolnr (Intercept) 18.1     4.26
 Residual             62.9     7.93
Number of obs: 3758, groups:  schoolnr, 211

Fixed effects:
            Estimate Std. Error t value
(Intercept)   41.005      0.325     126
performance::icc(nullmodel)
# Intraclass Correlation Coefficient

    Adjusted ICC: 0.224
  Unadjusted ICC: 0.224
```

(a) Based on the above output, how many students are in the data set? How many schools are in the dataset?

3758 students across 211 schools

(b) What do you learn from the ICC? Which is larger the within group or between group variation?

In the null model, 22.4% of the variation in language scores is across schools, but most is within schools.

Key Idea

To decide whether there is statistically significant group to group variation, you have several options

- Use the traditional fixed-effects ANOVA
- Use a LRT (using gls) to compare the model with and without the grouping variable
- Examine confidence intervals for τ

(c) State appropriate null and alternative hypotheses for this question

$H_o: \tau^2 = 0$ vs. $H_a: \tau^2 > 0$ Should probably consider a one-sided test because variance can never be negative

(d) Carry out all three approaches and summarize your results

Code

```
anova(lm(langPOST ~ 1 + factor(schoolnr), data = neth))
Analysis of Variance Table

Response: langPOST
                   Df Sum Sq Mean Sq F value Pr(>F)
factor(schoolnr)  210  74802     356    5.68 <2e-16 ***
Residuals        3547 222325      63
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Code

```
library(nlme)
model0 <- gls(langPOST ~ 1, data = neth, method = "ML")
nullmodel2 = lme(langPOST ~ 1, random = ~1 | schoolnr, data = neth, method = "ML
")
anova(model0, nullmodel2)
           Model df   AIC   BIC logLik   Test L.Ratio p-value
model0         1  2 27092 27105 -13544
nullmodel2     2  3 26601 26620 -13298 1 vs 2   492.9  <.0001
```

Code

```
confint(nullmodel)
             2.5 % 97.5 %
.sig01       3.773  4.813
.sigma       7.746  8.116
(Intercept) 40.361 41.642
intervals(nullmodel2)
Approximate 95% confidence intervals

 Fixed effects:
            lower est. upper
(Intercept) 40.37   41 41.64

 Random Effects:
  Level: schoolnr
                lower  est. upper
sd((Intercept))  3.77 4.257 4.807

 Within-group standard error:
lower  est. upper
7.745 7.928 8.115
```

We have convincing evidence (very small p-value) the variability in mean language scores across schools did not occur by random sampling alone

Note: Even if this variation was not statistically significant, we might argue to still include schoolnr in the model because that was the structure of our data!

(e) Using the null model, what do you predict for the language score of a randomly selected student? Is this the same as the mean language score in the dataset? Why or why not?

$\hat{\beta}_0$ =41.00. This isn't exactly the same as the sample mean $\bar{y}$ = 41.41 because we have unequal group sizes. .

(f) What is the estimated standard deviation of the language scores? Is this the same as the standard deviation of all the language scores in the sample? Why or why not?

The model estimates the variability in language scores in the population as $\sqrt{\hat{\sigma}^2 + \hat{\tau}^2} = \sqrt{18.13 + 62.85} \sim 9$ compared to s_y = 8.89.

Consider the first estimated random effect

```
ranef(nullmodel)$schoolnr[1,]
[1] -4.044
```

(g) How do you interpret this value?

School 1's average language score is about 4 points below average

and its standard error

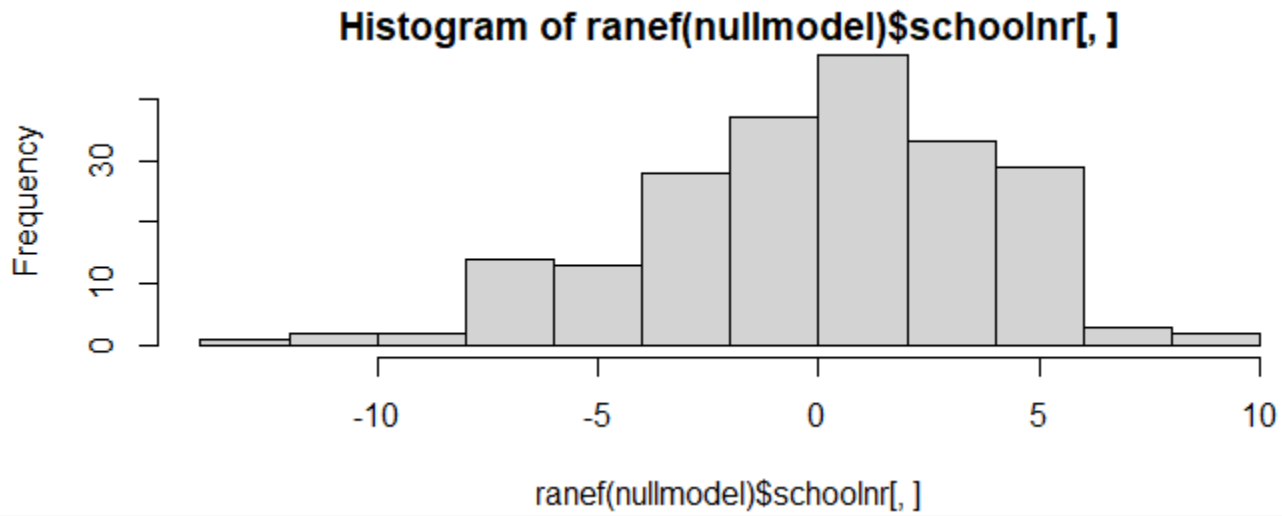
```
rand_ints <- as.data.frame(ranef(nullmodel, condVar = TRUE))
rand_ints[1,]
    grpvar        term grp condval condsd
1 schoolnr (Intercept)   1  -4.044  1.486
```

(h) Verify the calculation of this value from the model output. How would you interpret this value? How would you use it in a confidence interval?

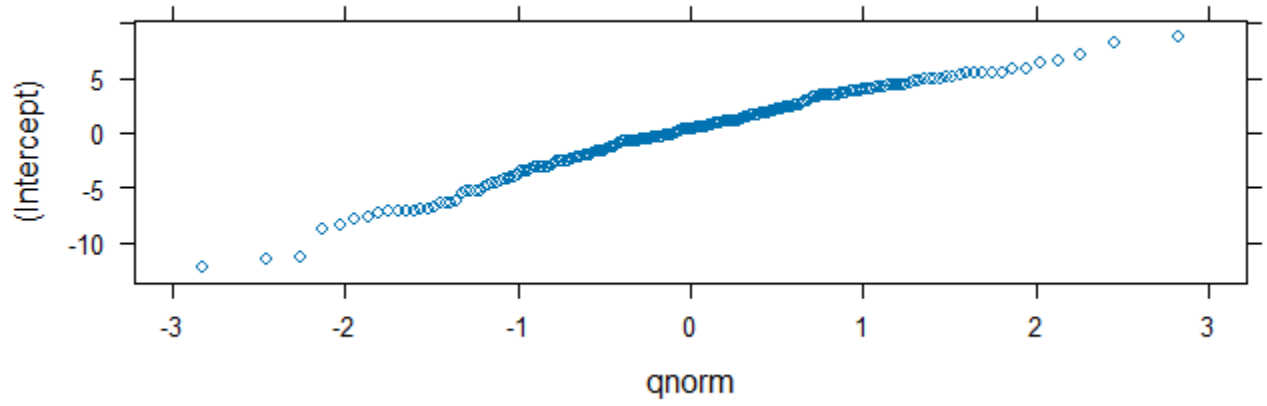
$\sqrt{(1/(1/18.13 + 25/62.85))}$ = 1.486. If we were to take other random samples and find $\hat{u}_1$ each time, then the sample to sample variation in that estimate is about 1.5. In other words, we think -4.04 is about 1.5 away from the true effect for that school in the population. So a rough 95% confidence interval for the school effect is -4.04 $\pm$ 2 x 1.486.

Examine the distribution of estimated random effects

```
hist(ranef(nullmodel)$schoolnr[,])
```




```
plot(ranef(nullmodel))
$schoolnr
```



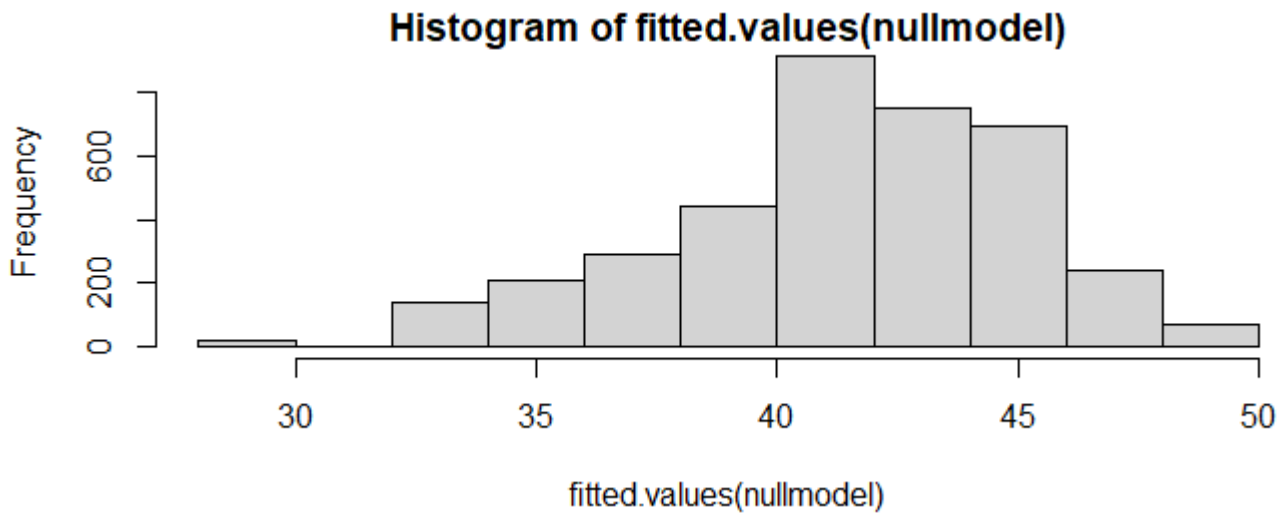
```
#qqnorm(rand_ints[,4])
#qqline(rand_ints[,4])

#qqnorm(residuals(nullmodel))
#qqline(residuals(nullmodel))

#performance::check_model(nullmodel)
#wonky residual plot?
```

(i) Would you say school 1 had an extreme or a typical random *effect*?

Minus 4 is in about the lower 20 percent based on the histogram?

Code

```
hist(fitted.values(nullmodel))
```



(i) Do the school effects appear to follow a roughly normal distribution?

yes

Note: So we have checked our additional model assumption. But why is it a little sketch to use these estimated effects to check the model assumption?

Because we created the $\hat{u}$ values assuming normality so the fact that they now look pretty normal is somewhat expected vs. an "independent chance" of what that was a reasonable assumption to begin with.

(j) What is the estimated standard deviation of this normal distribution? (Check the histogram to make sure you answer is reasonable)

That's $\hat{\tau}$ = 4.275 Does look like about 2/3 of our estimated school effects are within 4 or 5 of the mean of zero.

(k) What are the largest and smallest school *means* we expect to see according to this model?

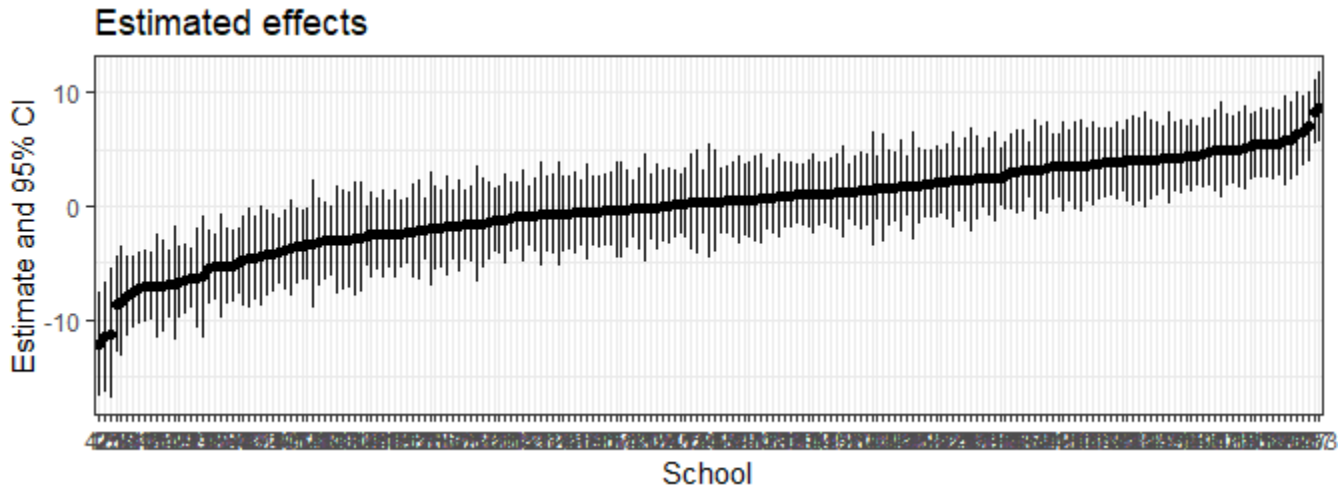
The overall mean language score is estimated to be 41.0046, but we expect the SD of school means about that overall mean to be about 4.257. So roughly 95% of schools should have a mean language score bewteen 41 - 2*4.26 = 32.48 and 41 + 2*4.26 = 49.52. (Or 3 SDs for 99.7%)

(l) What do you predict for the average language score for a school in the 84th percentile? (Hint: What is special about the 84th percentile in a normal distribution?)

The 84th percentile corresponds to one SD above the mean in a normal distribution. overall mean + 1 SD = 41.00 + 4.275 = 45.275. This is the predicted mean language score for a school in the 84th percentile.

A "Caterpillar plot" is a nice visual for sorting and visualizing the estimated effects.

```
rand_ints <- as.data.frame(ranef(nullmodel, condVar = TRUE))
ggplot(rand_ints, aes(y = condval, x = grp)) +
geom_point() +
geom_errorbar(aes(ymin = condval - 1.96*condsd,
max = condval + 1.96*condsd), width = 0) +
labs(title = "Estimated effects",
x = "School",
y = "Estimate and 95% CI") +
theme_bw()
```




```
#Also try?
#merTools::plotREsim( merTools::REsim( nullmodel ) )
```

Notes

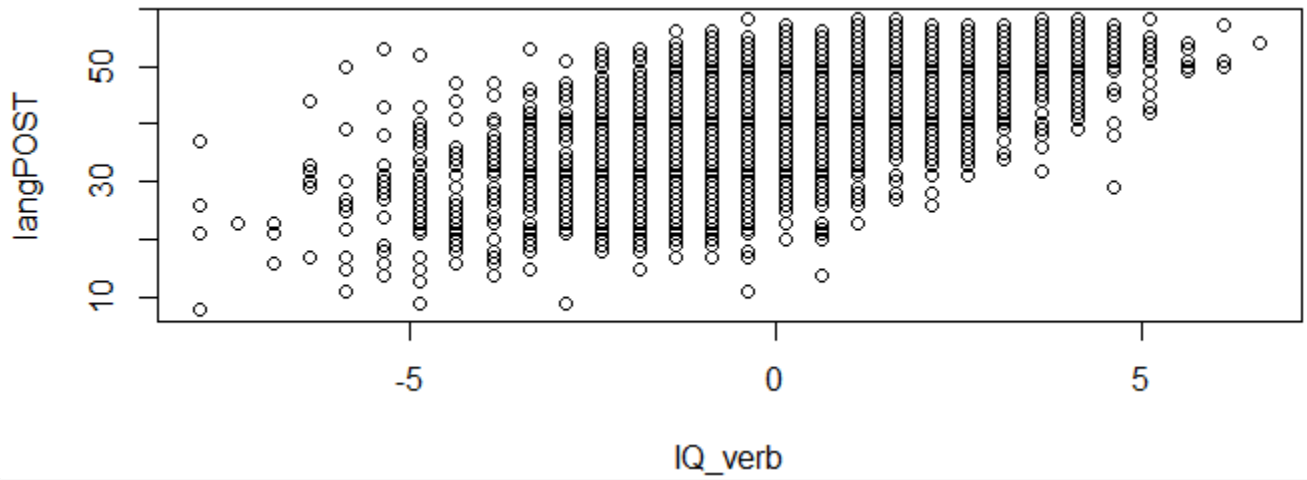
- The confidence intervals for the variance components are a little more "controversial" and different packages may approach these methods a little differently. All of them are aiming to test $H_0: \tau^2 = 0 \; vs. \; H_a: \tau^2 > 0$. The fact the variance can never be zero can occasionally lead to "boundary conditions" but usually something you don't have to worry about. You could also cut the p-value in half to reflect the one-sided alternative.

- Treating school as a fixed effect would "fail to reflect uncertainty resulting from variation among schools." That's why the standard errors tend to be smaller. With random effects we are able to make inferences about the population of schools, not just the ones in the study, a more difficult task.
- Three is a lot more controversy to the idea of picking out the schools with the largest positive (negative) effects and conclude=ing they are doing something better (worse) than the other schools? These are sometimes referred to as "value added models" http://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf

Example 2: More Netherland school analysis

Include the IQ variable (which has been centered (though before students with missing values were removed)) in the model.

```
plot(langPOST ~ IQ_verb, data = neth)
```




```
model1 = lmer(langPOST ~ 1 + IQ_verb + (1|schoolnr), data = neth, REML=F)
summary(model1)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: langPOST ~ 1 + IQ_verb + (1 | schoolnr)
   Data: neth

      AIC       BIC    logLik -2*log(L)  df.resid
    24920     24945    -12456     24912      3754

Scaled residuals:
   Min     1Q Median     3Q    Max
-4.196 -0.639  0.066  0.710  3.214

Random effects:
 Groups   Name        Variance Std.Dev.
 schoolnr (Intercept)  9.85    3.14
 Residual             40.47    6.36
Number of obs: 3758, groups:  schoolnr, 211
```

```
Fixed effects:
            Estimate Std. Error t value
(Intercept)  41.0549     0.2434   168.7
IQ_verb       2.5074     0.0544    46.1

Correlation of Fixed Effects:
        (Intr)
IQ_verb 0.003
confint(model1)
             2.5 % 97.5 %
.sig01       2.775  3.553
.sigma       6.216  6.513
(Intercept) 40.573 41.533
IQ_verb      2.400  2.614
```

(a) Provide interpretations of the estimated slope and intercept. (Hint: Remember lessons learned!)

the predicted (average) language score for a (all) student(s) with 'average IQ' in the 'average school' is 41.05. The predicted increase in the language score associated with a one-unit increase in IQ, after adjusting for school, is 2.51.

(b) Is IQ-verb statistically significant? How are you deciding?

the t-statistic for verbal IQ is quite large (46.11 > 2). This also gives VERY strong evidence that IQ_verb is helpful in predicting language scores, after adjusting for school.The confidence interval is (2.40, 2.61) which does not include zero. You could also do a likelihood ratio test for the models with and without verbal IQ.

```
Data: neth
Models:
nullmodel: langPOST ~ 1 + (1 | schoolnr)
model1: langPOST ~ 1 + IQ_verb + (1 | schoolnr)
          npar   AIC   BIC logLik -2*log(L) Chisq Df Pr(>Chisq)
nullmodel    3 26601 26620 -13298     26595
model1       4 24920 24945 -12456     24912  1683  1     <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(c) What is the estimated variation in responses for a particular value of IQ_verb?

This is the new $\hat{\sigma}$ = 6.362, for a particular school (conditional on verbal IQ and the school effect)

(d) What percentage of the Level 1 variance was explained by verbal IQ?

The new within-school residual variance estimate is 40.469, which is down from 62.85: (62.85 - 40.469)/62.85) = .356. So 35.6% of the within school variability in language scores was explained by verbal IQ.

(e) What percentage of the school-to-school variability in average language scores was explained by verbal IQ?

How much has τ decreased? Was 18.13 and is now 9.845, so (18.13-9.845)/18.13 = 0.457, so 45.7% of variation in language scores across the schools is accounted for by differences in the average verbal IQ scores across the schools.

(f) Which has changed more, the estimated within-group variation or the estimated between-group variation? Does this make sense in context? Is it possible for both of them to decrease? What does that mean?

Student verbal IQ scores actually explains more school-to-school variation in average language scores. This tells us that the verbal IQ scores vary across the schools as well as within the schools.

(g) What percentage of the total variance was explained by verbal IQ?

(18.13 + 62.85 - (9.845 + 40.469))/(18.13 + 62.85) = 0.379, so 37.9%

(h) What is the new value of the ICC? How do you interpret this? What would it mean for this value to be super close to zero?

The new ICC is 9.845/(9.845 + 40.469) = 0.195. This tells us how correlated the language scores of between pairs of students in the same school with the same verbal IQ. Can also interpret as the proportion of variability that is at the school level after accounting for the verbal IQ scores; would be zero if the verbal IQ scores explained all of the school to school variation in language scores.

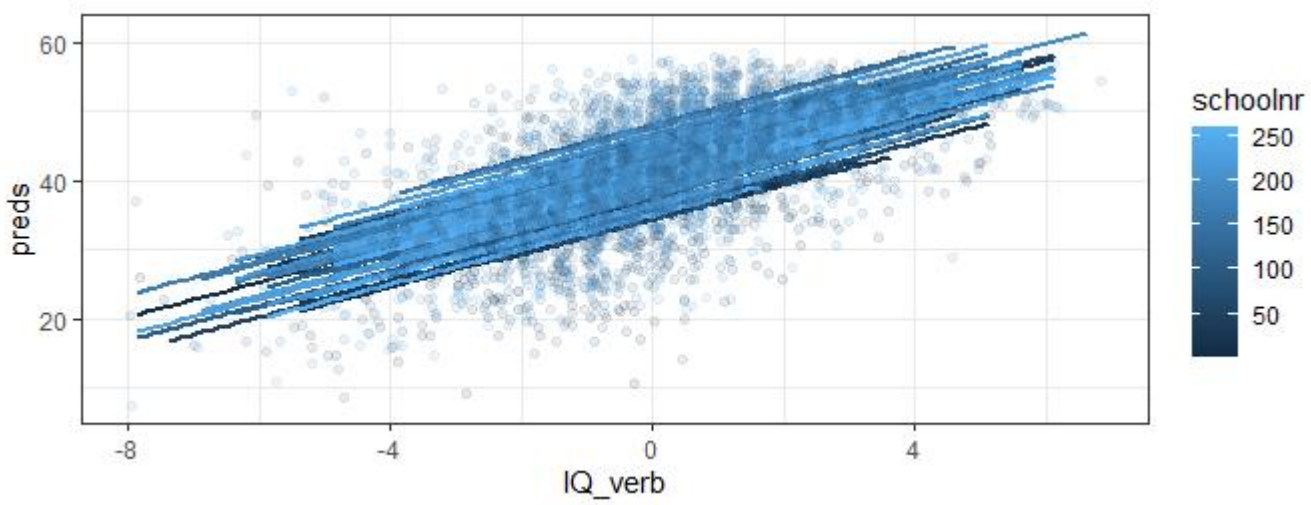
(i) What would a graph of this model look like?

Lots of parallel lines with the same slope but different intercepts (school effects)

A neat graph showing the fitted lines:

ⓘ Code

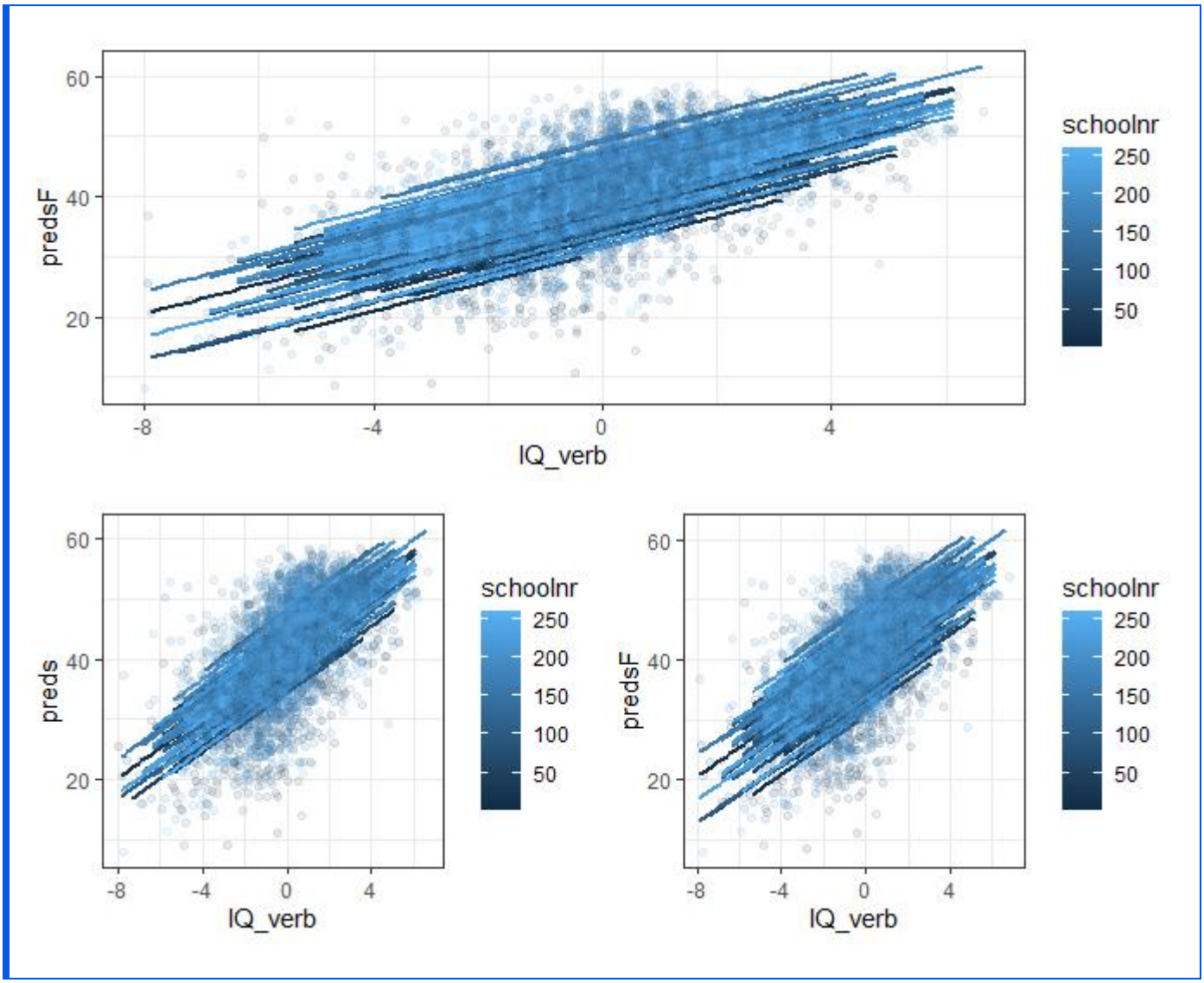
```
preds = predict(model1, newdata = neth)

ggplot(neth, aes(x = IQ_verb , y = preds , group = schoolnr, color = schoolnr )) +
geom_smooth(method = "lm", alpha = .5, se = FALSE) +
geom_jitter(data = neth, aes(y = langPOST), alpha = .1) +
  theme_bw()
```



(j) What if we had treated the schools as fixed effects

Same picture but lines will be more spread out vertically, no 'shrinkage'.

ⓘ Code



(k) What if we had ignored the grouping by schools?

```
summary(lmmodel <- lm(neth$langPOST~neth$IQ_verb))

Call:
lm(formula = neth$langPOST ~ neth$IQ_verb)

Residuals:
    Min      1Q  Median      3Q     Max
-29.315  -4.355   0.662   5.034  25.941

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   41.2958     0.1152     359   <2e-16 ***
neth$IQ_verb   2.6513     0.0564      47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 7.06 on 3756 degrees of freedom
Multiple R-squared:  0.37,  Adjusted R-squared:  0.37
F-statistic: 2.21e+03 on 1 and 3756 DF,  p-value: <2e-16
```

The slope looks similar but is actually different by about 3 SEs and notice the intercept SE is about twice as large (think of it like adjusting for effective sample size)

Optional

The Performance package reports many of the numbers we just calculated but also some other ones that can be hard to understand. See handout in canvas for more details.

```
performance::model_performance(model1)
# Indices of model performance

AIC     |    AICc |     BIC | R2 (cond.) | R2 (marg.) |   ICC |  RMSE | Sigma
-------------------------------------------------------------------------------
24925.1 | 24925.2 | 24950.1 |      0.471 |      0.342 | 0.196 | 6.219 | 6.362
performance::icc(model1)
# Intraclass Correlation Coefficient

    Adjusted ICC: 0.196
  Unadjusted ICC: 0.129
performance:::r2(model1)
# R2 for Mixed Models

  Conditional R2: 0.471
     Marginal R2: 0.342
var(model.matrix(model1) %*% fixef(model1))  #26.18, variance explained by fixed ef
fects
      [,1]
[1,] 26.18
performance::r2(model1, by_group = TRUE)
# Explained Variance by Level

Level    |    R2
---------------
Level 1  | 0.356
schoolnr | 0.457
```

Notes:

- We can think of ICC as the proportion of total variance explained by the grouping variable in the null model and R^2 as the proportion explained by the fixed effects, but now 3 sources of variation: sigma, tau, fixed effects
- The difference between adjusted/unadjusted ICC is whether you take into account the "variance explained by the fixed effects" in the denominator as well (the change in unexplained variation when the fixed effect is added to the model)

- This is actually an area of current research ("the literature does not seem to have converged on this topic") in how to calculate R^2 values for these models as the formulas provided here can actually turn out to be negative!

Reference: Nakagawa S, Johnson P, Schielzeth H (2017) The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisted and expanded. J. R. Soc. Interface 14. doi: 10.1098/rsif.2017.0213