

STA 610L: MODULE 2.7

RANDOM EFFECTS ANCOVA (INTRODUCTION)

DR. OLANREWaju MICHAEL AKANDE

INTRODUCTION

- ANOVA, ANCOVA, MANOVA: what is the difference?
- **ANOVA (Analysis of Variance)**: continuous outcome, categorical predictor(s)
 - one-way ANOVA: one categorical predictor
 - two-way ANOVA: two categorical predictors
 - two-way ANOVA with interaction: you get the picture!
- **ANCOVA (Analysis of Covariance)**: continuous outcome, categorical predictor(s), at least one continuous predictor that is generally considered a "nuisance".
- **MANOVA (Multivariate ANOVA)**: multiple continuous outcomes, categorical predictor(s).

Historically these names had implications regarding the estimation methods used, but that is no longer always the case.

MOTIVATING EXAMPLE: NELS

Hoff considers a subset of the National Educational Longitudinal Study of Education (NELS) data containing information on math scores for a random sample of 10th graders from a national sample of 685 large urban public schools.

```
load('data/nels.Rdata')  
head(nels)
```

```
##   school enroll flp public urbanicity hwh   ses mscore  
## 1   1011     5   3     1     urban    2 -0.23  52.11  
## 2   1011     5   3     1     urban    0  0.69  57.65  
## 3   1011     5   3     1     urban    4 -0.68  66.44  
## 4   1011     5   3     1     urban    5 -0.89  44.68  
## 5   1011     5   3     1     urban    3 -1.28  40.57  
## 6   1011     5   3     1     urban    5 -0.93  35.04
```

MOTIVATING EXAMPLE: NELS

```
str(nels)
```

```
## 'data.frame':    12974 obs. of  8 variables:
## $ school      : Factor w/ 684 levels "1011","1012",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ enroll      : num  5 5 5 5 5 5 5 5 5 5 ...
## $ flp         : num  3 3 3 3 3 3 3 3 3 3 ...
## $ public      : num  1 1 1 1 1 1 1 1 1 1 ...
## $ urbanicity: Factor w/ 3 levels "rural","suburban",...: 3 3 3 3 3 3 3 3 3 3 ...
## $ hwh         : num  2 0 4 5 3 5 1 4 8 2 ...
## $ ses         : num  -0.23 0.69 -0.68 -0.89 -1.28 -0.93 0.36 -0.24 -1.07 -0.1 ...
## $ mscore      : num  52.1 57.6 66.4 44.7 40.6 ...
```

```
summary(nels)
```

```
##      school      enroll      flp      public
## 4401   :   50   Min.   :0.000   Min.   :1.000   Min.   :0.0000
## 4271   :   48   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.0000
## 4341   :   46   Median :2.000   Median :2.000   Median :1.0000
## 4522   :   39   Mean    :2.314   Mean    :1.933   Mean    :0.7564
## 1792   :   36   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:1.0000
## 4101   :   34   Max.    :5.000   Max.    :3.000   Max.    :1.0000
## (Other):12721
##      urbanicity      hwh      ses      mscore
## rural   :2349   Min.    : 0.000   Min.    : -2.11000   Min.    :19.38
## suburban:6114   1st Qu.: 1.000   1st Qu.: -0.49000   1st Qu.:44.62
## urban   :4511   Median : 2.000   Median : 0.02000   Median :51.38
##          Mean    : 2.755   Mean    : 0.06735   Mean    :51.17
##          3rd Qu.: 4.000   3rd Qu.: 0.60000   3rd Qu.:58.21
##          Max.    :21.000   Max.    : 1.98000   Max.    :86.68
##
```

NELS EXAMPLE

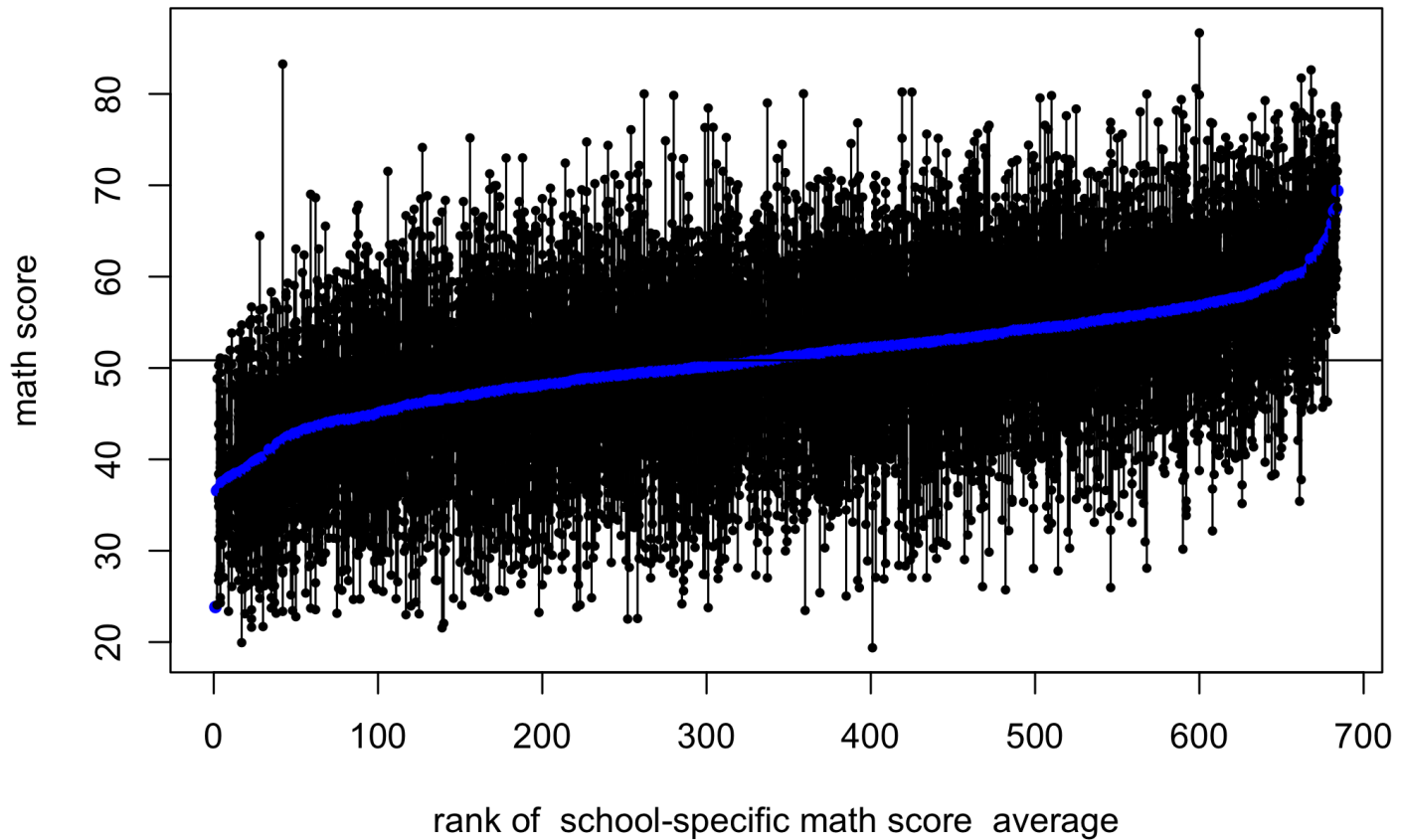
```
avmscore.schools<-tapply(nels$mscore,nels$school,mean,na.rm=TRUE)
id.schools<-names(avmscore.schools)
m<-length(id.schools)

plot(c(1,m),range(nels$mscore), type="n",ylab="math score",
      xlab="rank of school-specific math score average",cex=.7)

for(school in id.schools[order(avmscore.schools)[seq(1,length(avmscore.schools),
                                                    by=1)]]])
{
  y<-nels$mscore[nels$school==school]
  x<-rank(avmscore.schools)[ id.schools==school]
  points( rep(x,length(y)), y,pch=16,cex=.6 )
  points(x, mean(y),col="blue",pch=16,cex=.8)
  segments( x,min(y),x,max(y))
}

abline(h=mean(avmscore.schools))
```

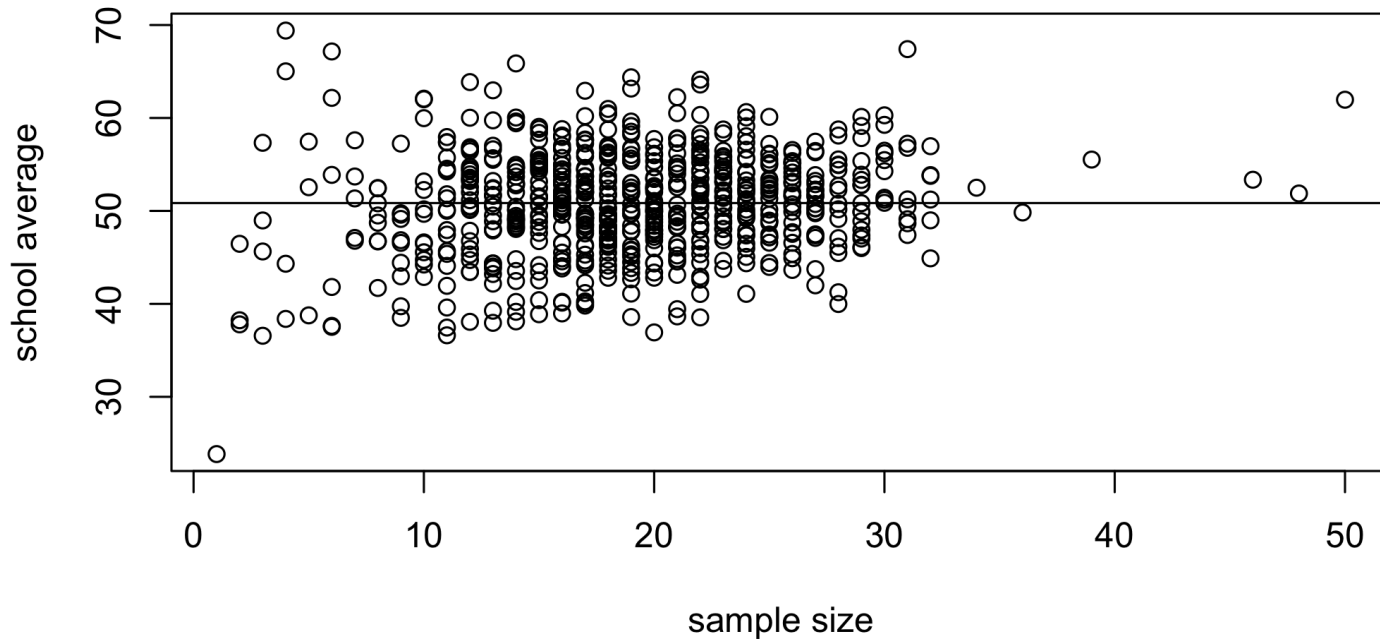
NELS EXAMPLE



NELS EXAMPLE

- The school-specific averages range from 24 to 69, with 51 the average of all 685 school averages (weighting each school equally).
- The school-specific variances range from 3 to 187 -- quite a wide range!
- The school with the highest average has a very small sample size ($n_j = 4$).

WHICH SCHOOL IS BEST?



Do we really have strong evidence that the true mean in this school is substantially larger than that in other schools in the sample?

ANOVA

We could start by fitting a "standard" ANOVA model:

```
m1 <- lm(mscore~school,data=nels)
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: mscore
##           Df Sum Sq Mean Sq F value    Pr(>F)
## school      683 342385   501.30   6.8118 < 2.2e-16
## Residuals 12290 904450    73.59
```

```
summary(aov(mscore~school,data=nels))
```

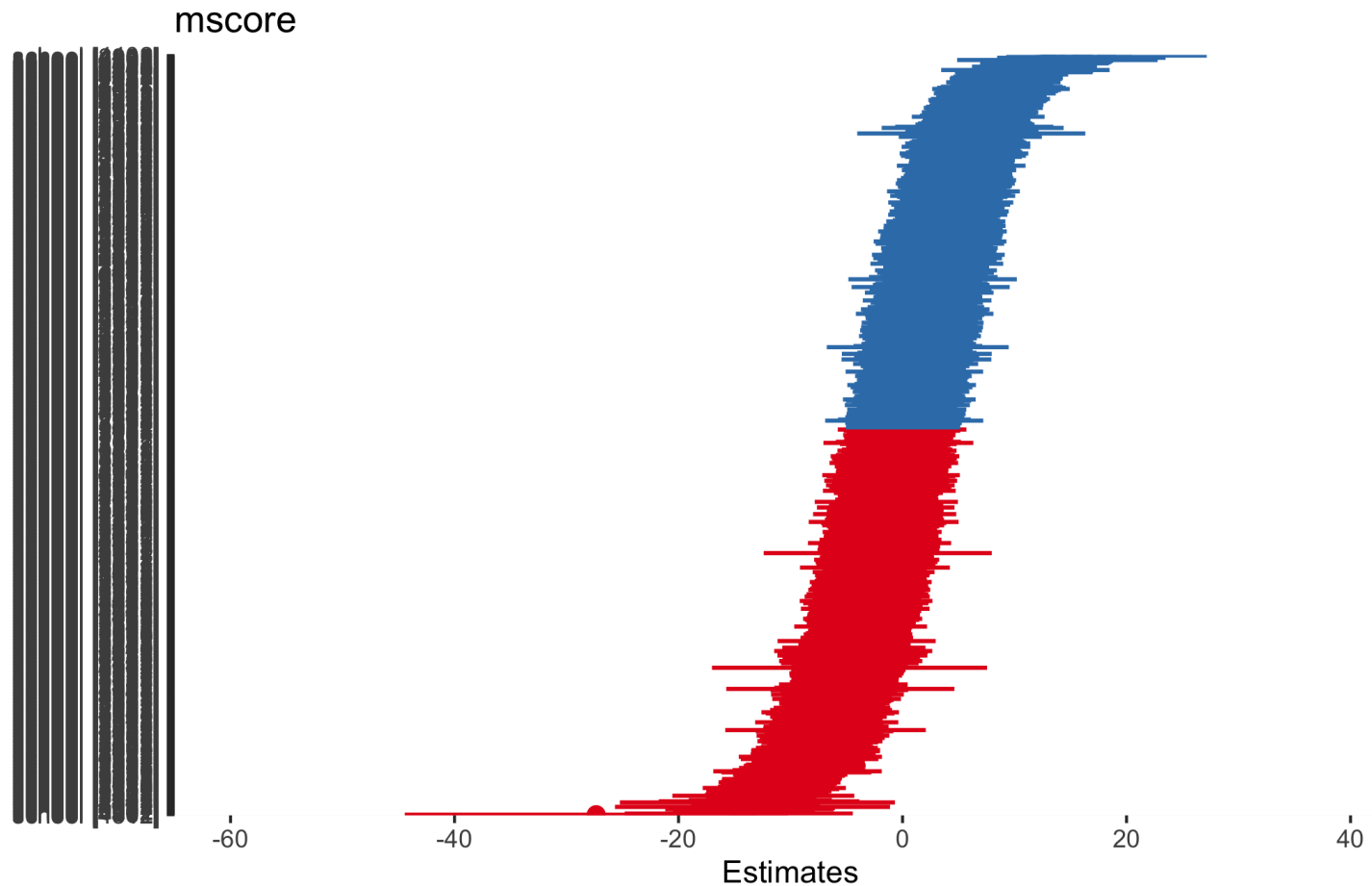
```
##           Df Sum Sq Mean Sq F value Pr(>F)
## school      683 342385   501.3    6.812 <2e-16
## Residuals 12290 904450    73.6
```

Here we see clear evidence of heterogeneity in math scores across schools.

ANOVA RESULTS

```
library(sjPlot)  
plot_model(m1, sort.est=TRUE)
```

ANOVA RESULTS



Messy plot but here we might even try to draw conclusions about the schools that are most different.

RANDOM EFFECTS ANOVA

We may wish to use shrinkage estimation in order to stabilize the estimates, particularly for schools in which few students provide data, as we have done a few times already.

We can fit the usual random effects ANOVA model given by

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij},$$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$ and $\alpha_j \sim N(0, \tau^2)$.

```
library(lme4)
m2 <- lmer(mscore~(1|school),data=nels)
summary(m2)

#library(sjstats)
#icc(m2)
```

RANDOM EFFECTS ANOVA

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: mscore ~ (1 | school)
## Data: nels
##
## REML criterion at convergence: 93914.6
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.8113 -0.6534  0.0094  0.6732  4.7000
##
## Random effects:
##   Groups    Name      Variance Std.Dev.
##   school  (Intercept) 23.68     4.866
##   Residual                73.71     8.585
## Number of obs: 12974, groups:  school, 684
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  50.9390     0.2028   251.2
```

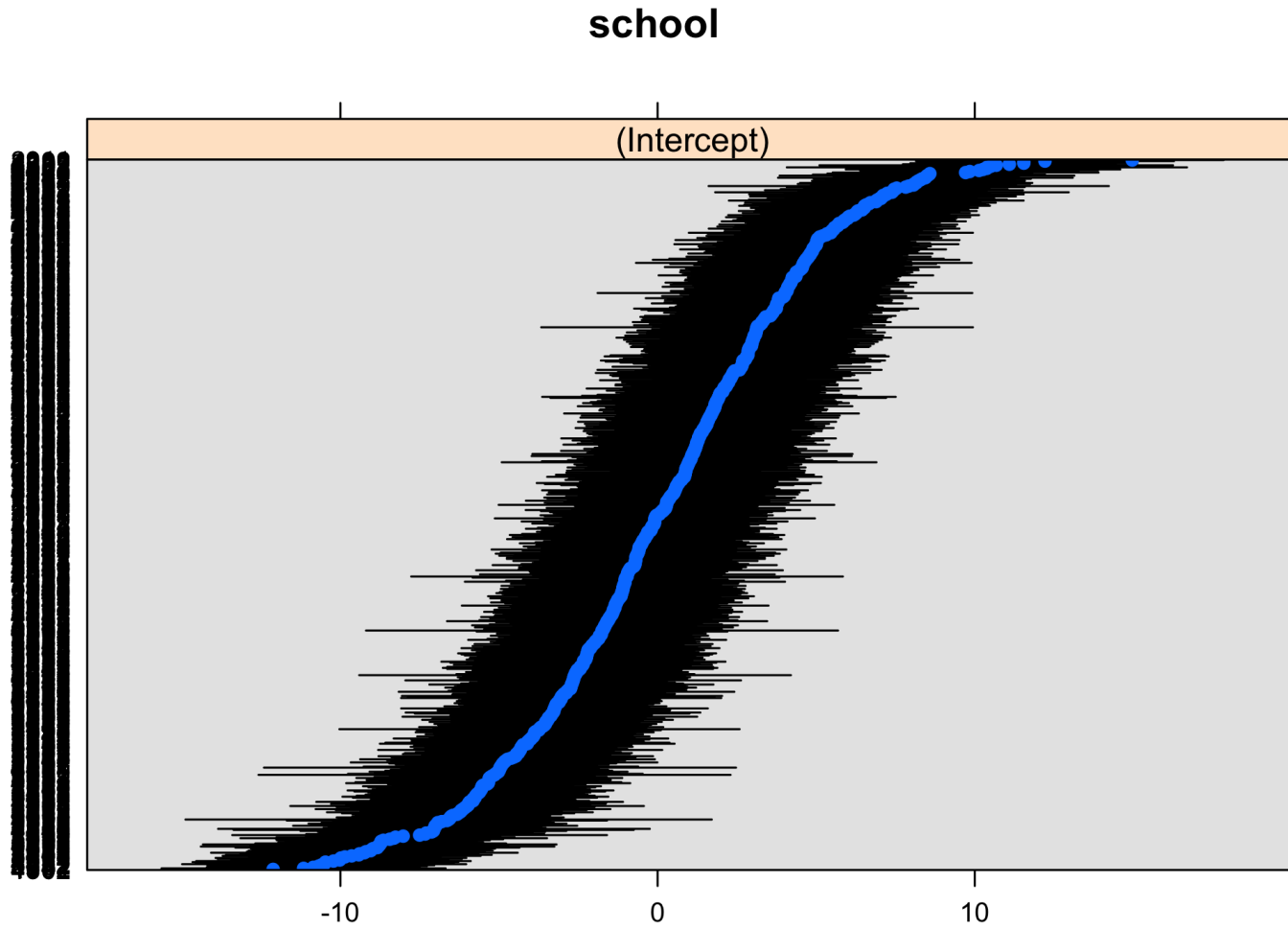
The ICC is 0.243 so that we do have some correlation between students in the same school but not so strong. You should check this yourself.

RANDOM EFFECTS ANOVA

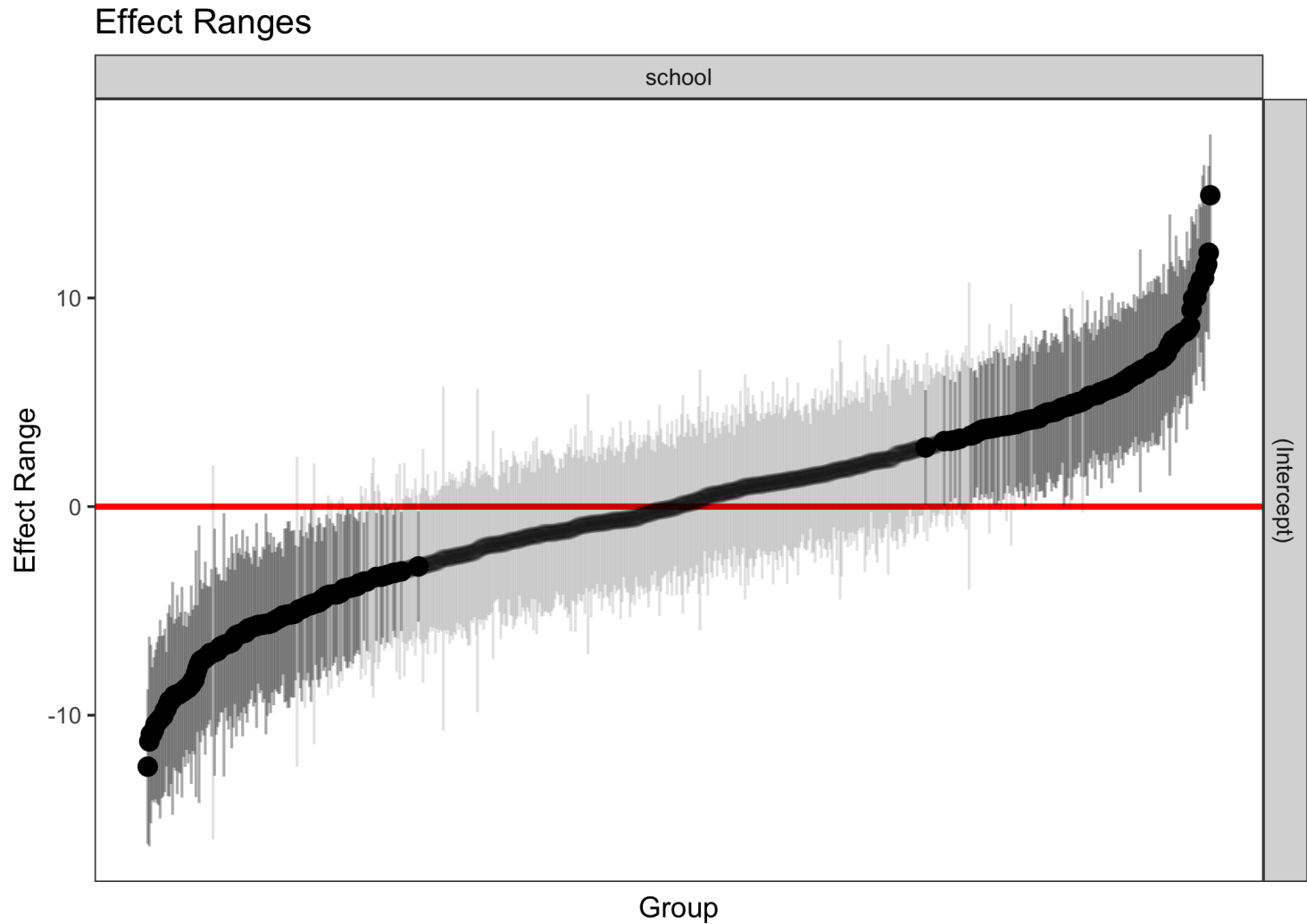
Next, we examine the distribution of random effects.

```
library(lattice)
dotplot(ranef(m2, condVar=TRUE))
#OR
library(merTools)
plotREsim(REsim(m2, n.sims=100), stat='median', sd=TRUE)
```

RANDOM EFFECTS ANOVA



RANDOM EFFECTS ANOVA



RANDOM EFFECTS ANOVA

How do we conduct a formal test of heterogeneity in this random effects setting?

Well, this is a bit more complicated than in the **fixed effects setting**, where we had an F-test.

In particular, no heterogeneity corresponds to the case in which

$$\tau^2 = 0 \iff \alpha_1 = \dots = \alpha_J = 0,$$

so saying something about the single parameter τ^2 has implications about the J parameters α_j .

A second problem is that τ^2 cannot be < 0 , and we wish to test the null hypothesis $H_0 : \tau^2 = 0$, so we are conducting a hypothesis test at the boundary of the parameter space.

RANDOM EFFECTS ANOVA

As shown in Stram and Lee (1994), the approximate asymptotic null distribution for $H_0 : \tau^2 = 0$ using a likelihood ratio test comparing our model

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}; \quad \alpha_j \sim N(0, \tau^2)$$

to a model without random effects

$$(y_{ij} = \mu + \varepsilon_{ij})$$

in this case is a 50-50 mixture of a χ_0^2 (point mass on 0) and a χ_1^2 distribution.

In general, if we wish to compare a model with $q + 1$ random effects (calculated as terms that have a random effect, not the number of groups) to a nested model with q random effects, the asymptotic null distribution is a 50-50 mixture of χ_{q+1}^2 and χ_q^2 distributions.

RANDOM EFFECTS ANOVA

Letting LR denote twice the difference in maximized log-likelihoods in the model with and without a single random effect, you can obtain the null density in R using

$$0.5 * (\text{dchisq}(x, q + 1) + \text{dchisq}(x, q))$$

and the p-value via

$$0.5 * (1 - \text{pchisq}(LR, q + 1) + 1 - \text{pchisq}(LR, q)).$$

RANDOM EFFECTS ANOVA

For the NELS data fit using a frequentist random effects model, we would calculate this as follows.

```
m3 <- lmer(mscore~(1|school),data=nels,REML=FALSE) #ML estimation
m4 <- lm(mscore~1,data=nels)
LR <- 2*(logLik(m3)-logLik(m4))
LR
```

```
## 'log Lik.' 2137.067 (df=3)
```

```
0.5*(1-pchisq(LR[1],1)+1-pchisq(LR[1],0))
```

```
## [1] 0
```

```
anova(m3,m4)
```

```
## Data: nels
## Models:
## m4: mscore ~ 1
## m3: mscore ~ (1 | school)
##      npar    AIC    BIC logLik deviance  Chisq Df Pr(>Chisq)
## m4      2 96054 96069 -48025    96050
## m3      3 93919 93942 -46957    93913 2137.1  1 < 2.2e-16
```

We conclude that there is significant heterogeneity across schools in the mean math scores.

BRINGING SES INTO THE MIX

NELS contains a measure of socioeconomic status (SES) for each student.

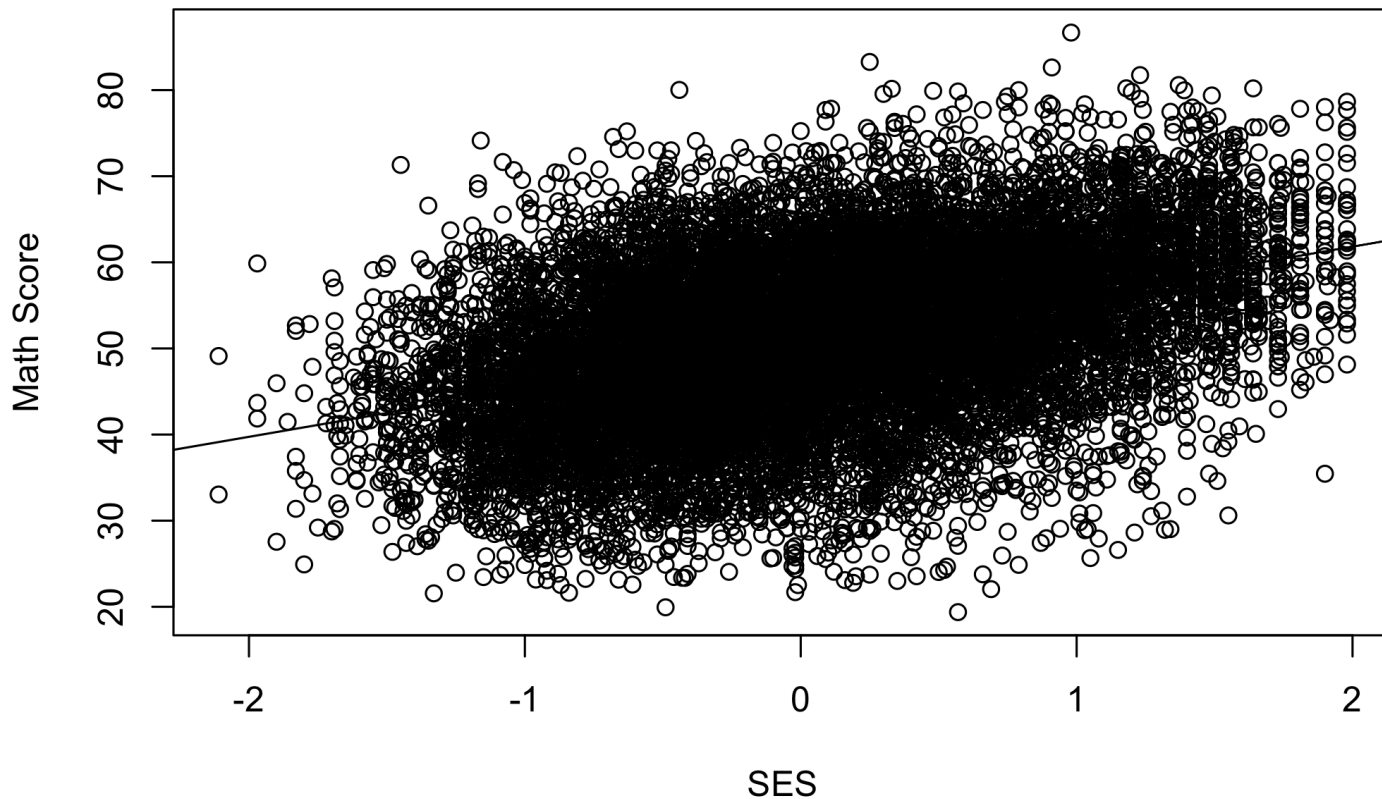
We generally expect some degree of correlation between SES and math score (people who are good at math often can get good jobs, and then have kids who may inherit math talents; rich parents may have more time and resources to devote to their kids).

Of course the relationship is not deterministic (there are plenty of math whizzes who did not have rich parents -- Gauss!, and there are plenty of rich parents who have kids who do not make good math scores -- college admissions scandal!).

Also, the SES score itself is a summary score and not particularly interesting to interpret as is. We might want to standardize it to help with that.

BRINGING SES INTO THE MIX

Let's look overall at the association between SES and math score in NELS.



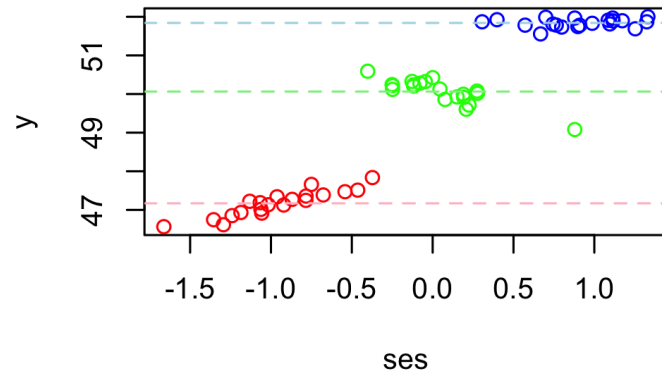
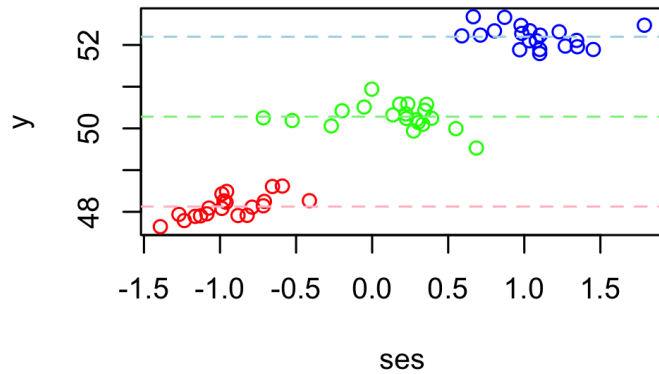
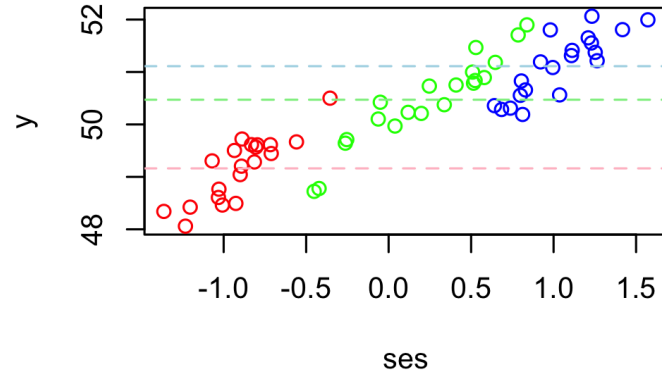
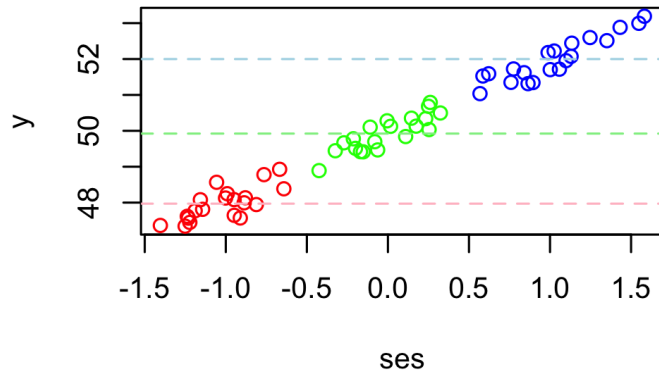
BIG PICTURE

Consider "simulated" data on schools, which we represent using red, green, and blue points on the plot on the next slide, respectively.

The schools we illustrate include one low SES school, one middle SES school, and one high SES school.

Let's consider multiple ways in which we could obtain the marginal association between SES and math score on the previous slide.

BIG PICTURE



RANDOM EFFECTS ANCOVA

We want our model to be able to help us understand how SES (x_{ij}) and math scores are related in schools.

In the framework of the model, and switching notations a bit to linear regression notation,

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij},$$

what values of β_j are consistent with these figures?

One way to assess how SES is related to math score is to start with an ANCOVA model, allowing school-specific intercepts while including SES as a "fixed" covariate x_{ij} :

$$y_{ij} = \beta_{0j} + \beta_1x_{ij} + \varepsilon_{ij}.$$

In this model, we estimate the same effect of SES for each school.

We can then compare the two models.

RANDOM EFFECTS ANCOVA

First, the second model.

```
m5 <- lmer(mscore~ses+(1|school),data=nels)
summary(m5)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: mscore ~ ses + (1 | school)
## Data: nels
##
## REML criterion at convergence: 92558.1
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.8753 -0.6428  0.0165  0.6693  4.4322
##
## Random effects:
## Groups   Name      Variance Std.Dev.
## school  (Intercept) 12.22    3.495
## Residual                68.03    8.248
## Number of obs: 12974, groups: school, 684
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  50.7175    0.1542  328.99
## ses          4.3766    0.1123   38.98
##
## Correlation of Fixed Effects:
##      (Intr)
## ses -0.042
```

As mentioned earlier, SES is not particularly interesting to interpret as is.

RANDOM EFFECTS ANCOVA

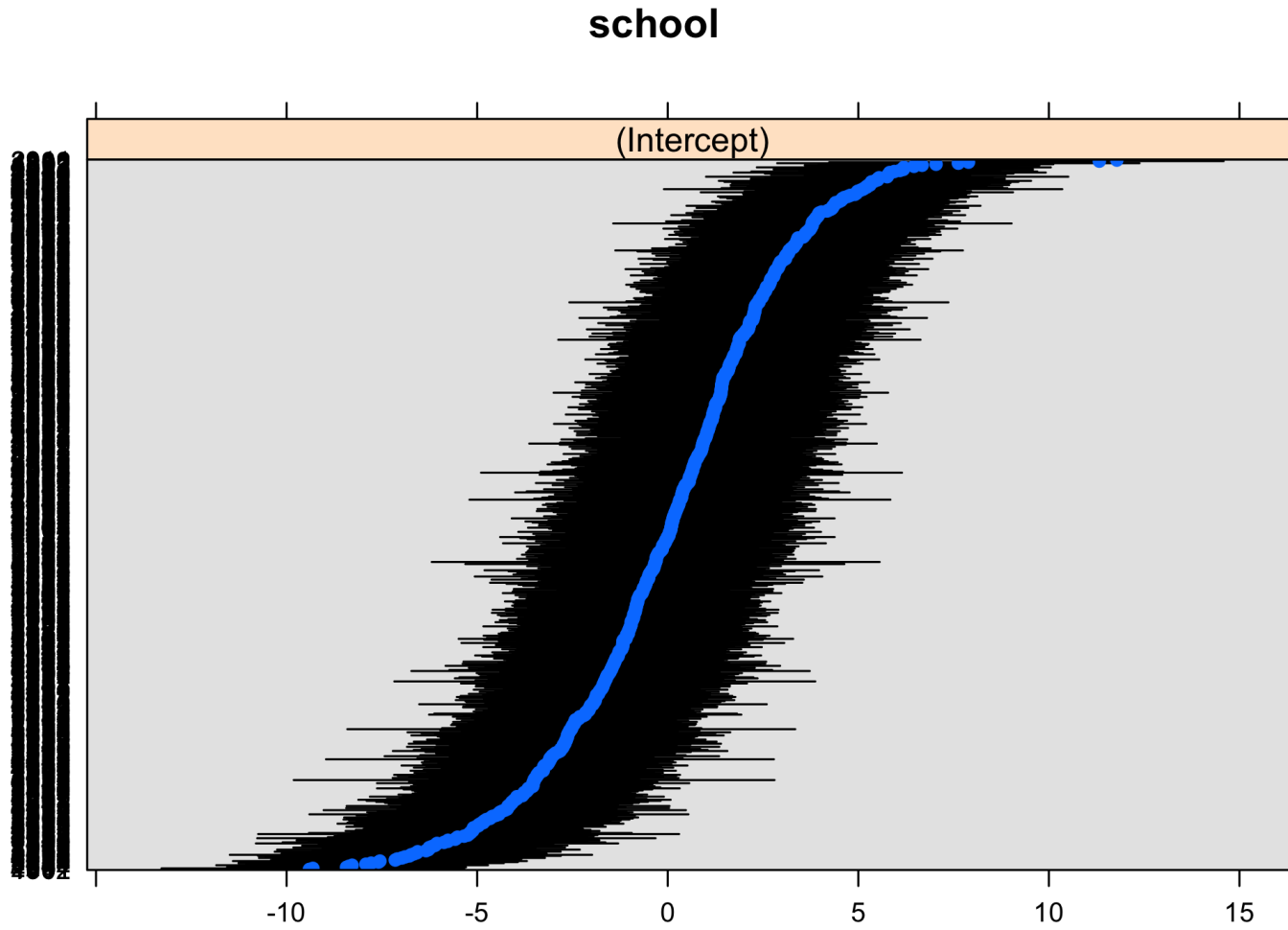
We can standardize the variable to get a different kind of interpretation.

```
nels$sesstd <- nels$ses/sd(nels$ses)
m5 <- lmer(mscore~sesstd+(1|school),data=nels)
summary(m5)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: mscore ~ sesstd + (1 | school)
## Data: nels
##
## REML criterion at convergence: 92558.7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.8753 -0.6428  0.0165  0.6693  4.4322
##
## Random effects:
## Groups Name Variance Std.Dev.
## school (Intercept) 12.22  3.495
## Residual 68.03  8.248
## Number of obs: 12974, groups: school, 684
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  50.7175     0.1542  328.99
## sesstd       3.2900     0.0844   38.98
##
## Correlation of Fixed Effects:
##      (Intr)
## sesstd -0.042
```

Pretty big effect of SES -- a 1 SD increase in SES is associated with a 3.3 point increase in math score on average.

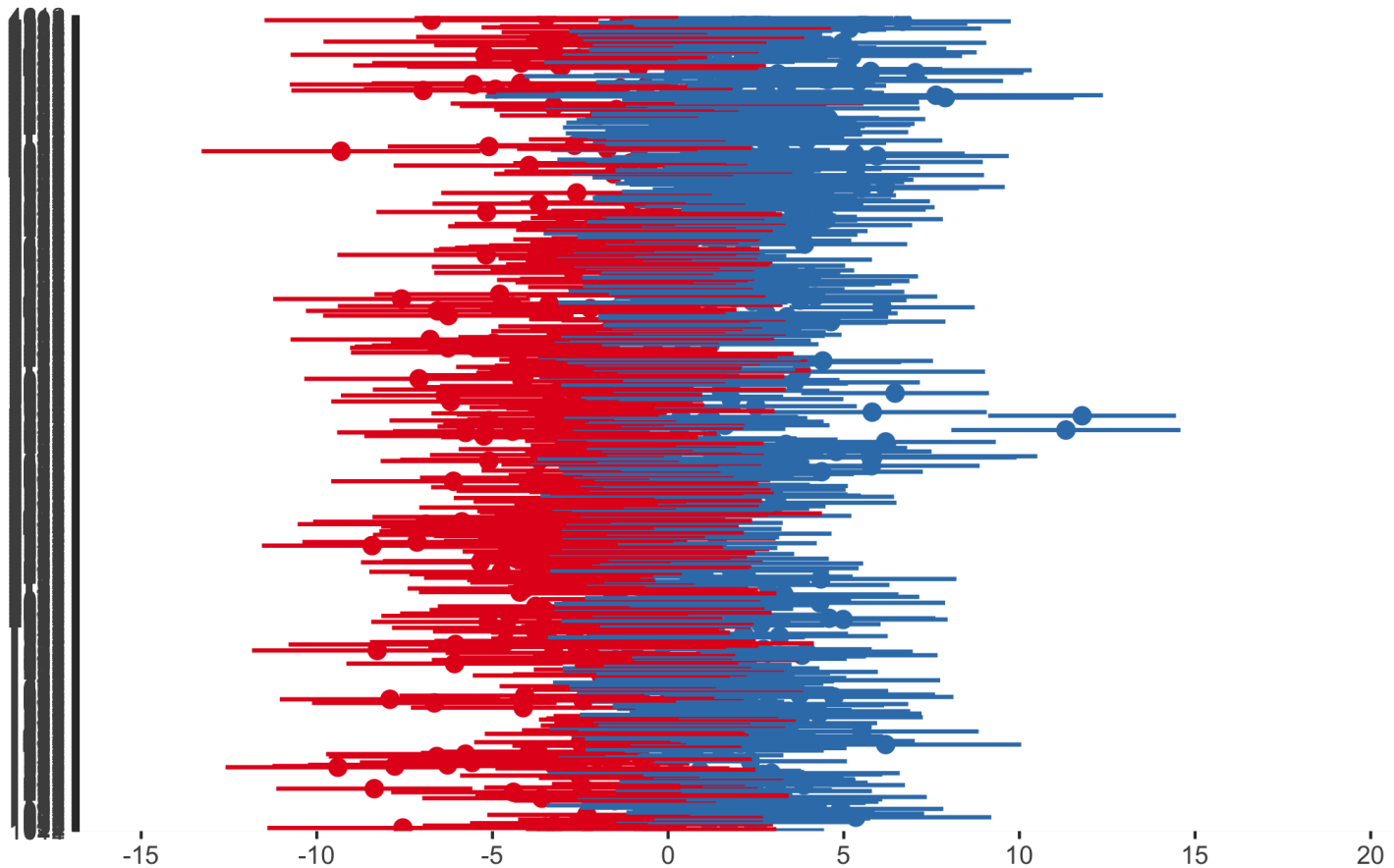
RANDOM EFFECTS ANCOVA



RANDOM EFFECTS ANCOVA

```
plot_model(m5,type='re')
```

Random effects

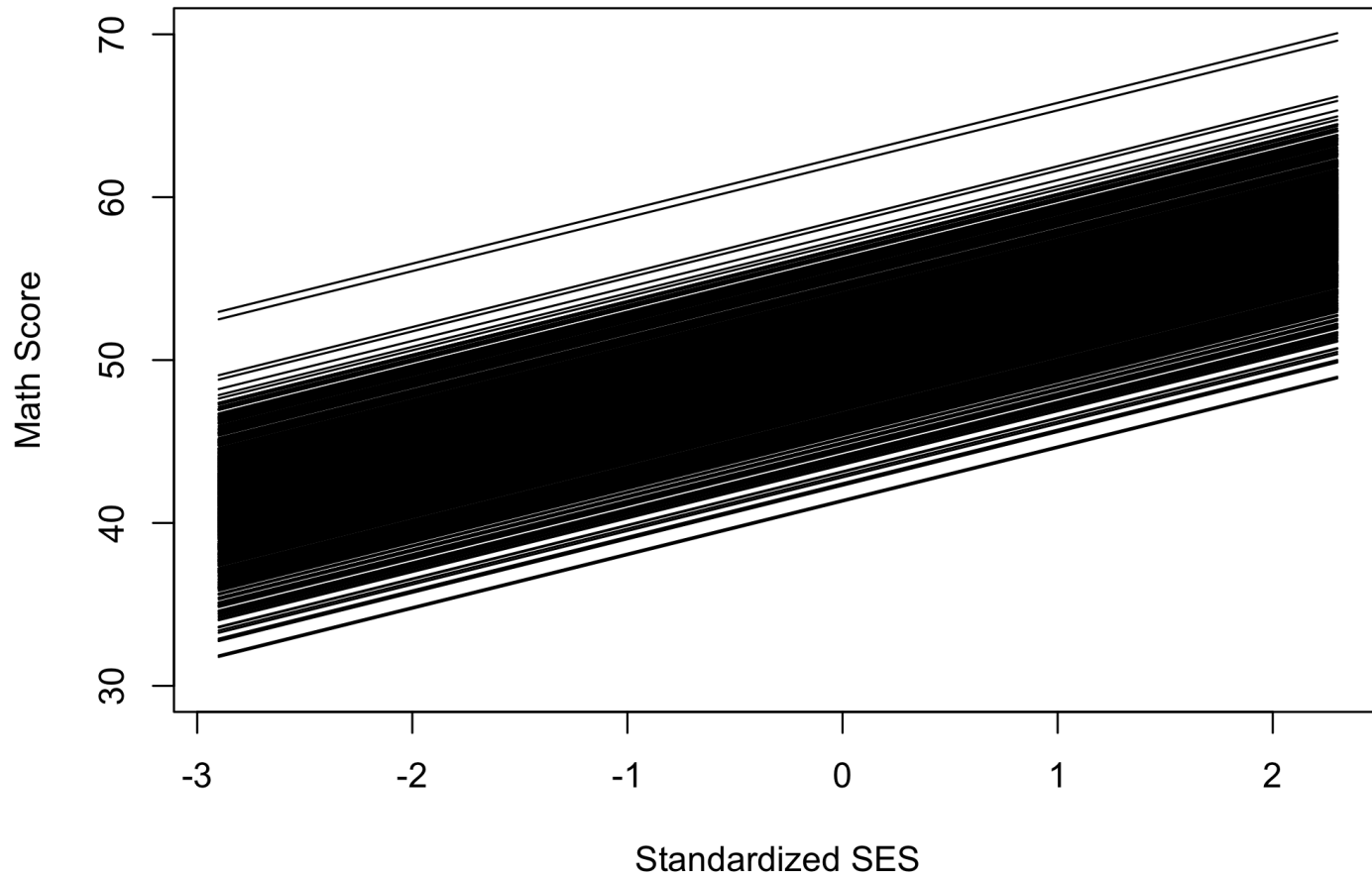


RANDOM EFFECTS ANCOVA

Let's plot the estimated school-specific lines from the random intercept model.

```
xplot=seq(-2.9,2.3,by=.1)
yplot=rep(60,length(xplot))
plot(xplot,yplot,type="n",ylim=c(30,70),xlab="Standardized SES",ylab="Math Score")
for(school in 1:length(id.schools))
{
  yplot=coef(m5)$school[school,1]+coef(m5)$school[school,2]*xplot
  lines(xplot,yplot)
}
```

RANDOM EFFECTS ANCOVA



RANDOM EFFECTS ANCOVA

This model allows separate intercepts for each school but assumes a common slope.

As we have already discussed, one concern is whether SES has the same relationship with math scores in all schools.

For example, some schools may have less of a disparity in scores across levels of SES than others.

So, next we would want to try random slopes of SES as discussed earlier and test the two nested models.

We'll do this in the next module.

WHAT'S NEXT?

MOVE ON TO THE READINGS FOR THE NEXT MODULE!