S-043/Stat-151
Analysis for Clustered and Longitudinal Data
(Multilevel & Longitudinal Models)

Unit 2, Lecture 2
Random Intercept Models, Continued
(Primarily Level 2 Covariates.)

Instructor: Prof. Luke W. Miratrix

lmiratrix@g.harvard.edu

Larsen 603

Acknowledgement of unknown terms

If I am using mathematical, coding, or statistical vocabulary you are unfamiliar with please do any or all of the following:

1) Ask in class for clarification

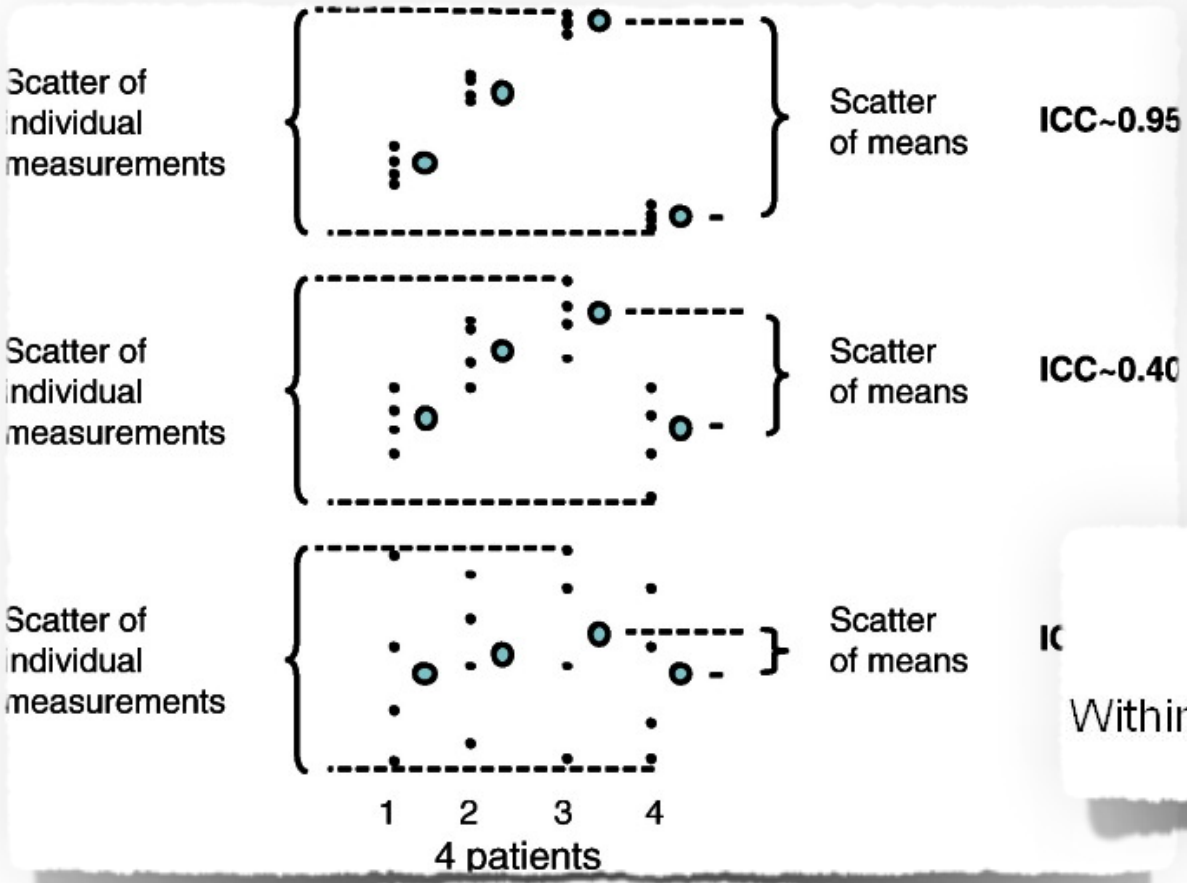
2) Ask in the quiet questions for clarification

2) Email a TF with the term

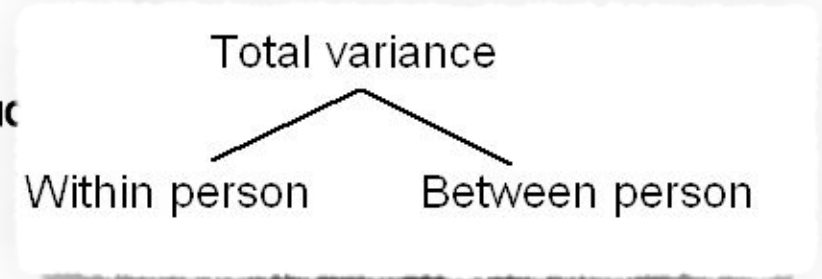
Lecture Goals

- ★ Discuss the Interclass Correlation Coefficient (ICC)
- ★ Discuss covariates at Level 1 and Level 2 in Random Intercept models
- ★ Show how one can represent models mathematically in a variety of ways.
- ★ Talk about how multilevel modeling handles data with non-independent observations.
- ★ Connecting R notation and output to math notation and models.

The Interclass Correlation Coefficient



$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$$



We use MLM to learn about variation

Multilevel Models make assumptions about how the schools and the individuals in those schools relate to each other.

Based on those assumptions (to be discussed more in future) we can separate out *individual variation* and *school variation*

Research Question: Variation across schools

RQ1: How much variation is there across schools in school mean math achievement?

RQ2: Is this a lot of variation or a little?

BACK TO SCHOOL:
Our canonical representation of our model for HS&B

We have students *i* nested in schools *j*.

Student *i* is in school *j[i]*.

We then have:

$$Math_i = \alpha_{j[i]} + \beta SES_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma_y^2)$$

$$\alpha_j \sim N(\mu, \sigma_\alpha^2)$$

Variance decomposition

Student residual $\epsilon_i \sim N(0, \sigma_y^2)$ Unexplained student variation

School random intercept $\alpha_j \sim N(\mu, \sigma_\alpha^2)$ Unexplained school variation

Once we have fit our model we have:

- ★ How much variation there is in students within schools
- ★ How much variation there is in the school averages

This gives an estimate of *total variation*

$$\sigma^2_{total} = \sigma^2_\alpha + \sigma^2_y$$ Variance decomposition

The total variation is the variation of outcomes. $\sigma^2_{total} \approx Var(Y_{ij})$

The above equation is a *variance decomposition*

- ★ We have divided up our variation into variation in schools and students

A rule of statistics:

Variances add, NOT standard deviations

The Intraclass Correlation Coefficient (ICC)

$$ICC = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_y^2}$$

ICC will be in [0, 1]
"Percent of variation due to school random effect"

The ICC uses a baseline (overall variation) to measure the extent of school variation.

It asks; What proportion of the variation is due to school effects?

Why the bottom terms? Because

$$Y_i = \alpha_{j[i]} + \epsilon_i$$

so

$$\mathrm{Var}(Y_i) = \mathrm{Var}\left(\alpha_{j[i]}\right) + \mathrm{Var}(\epsilon_i) = \sigma_\alpha^2 + \sigma_y^2$$

Calculating the ICC

```
> display( M0 )
lmer(formula = mathach ~ 1 + (1 | id), data = dat)
coef.est  coef.se
   12.64     0.24

Error terms:
 Groups   Name        Std.Dev.
 id       (Intercept) 2.93
 Residual             6.26
---
number of obs: 7185, groups: id, 160
AIC = 47122.8, DIC = 47114.8
deviance = 47115.8
```

Our simple MLM!

```
> 2.93^2 / (2.93^2 + 6.26^2)
[1] 0.1797038
```

Hand-calc the ICC!

```
> sigma.alpha = sigma.hat( M0 )$sigma$id
> sigma.y = sigma( M0 )
> ICC =  sigma.alpha^2 / (sigma.alpha^2 + sigma.y^2 )
> ICC
(Intercept)
  0.1803518
```

This number is more precise. In the above, we rounded and thus our final answer was a bit off.

We can also calculate using R commands by extracting the numbers from our model and doing the math that way.

Random intercept models with Level-2 Predictors

A Motivating Research Question: More variation across schools

Q: How is the variation of mean math achievement across schools associated with school type (public vs. catholic)?

One way: Use the individual adjusted means (the Empirical Bayes estimates) and see how they vary by sector id.

Let's try!


{See in class R coding.}


Reminder: our original random-intercept model

$$Math_i = \alpha_{j[i]} + \varepsilon_i$$
$$\varepsilon_i \sim N\left(0, \sigma_y^2\right)$$
$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$$

We can rewrite this as

$$Math_i = \alpha_{j[i]} + \varepsilon_i$$
$$\varepsilon_i \sim N\left(0, \sigma_y^2\right)$$
$$\alpha_j = \mu_\alpha + u_j$$
$$u_j \sim N(0, \sigma_\alpha^2)$$

This notation will make some of our models a lot easier to write.

u_j is now a school-level residual from the grand mean, μ_α

Extend to including a group-level predictor

$$y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i$$ **Level 1 Model**

$$\epsilon_i \sim N(0, \sigma_y^2)$$

$$\alpha_j = \gamma_0 + \gamma_1 s_j + u_j$$ **Level 2 Model**

$$u_j \sim N(0, \sigma_\alpha^2)$$

Our level-2 predictor

$$cov(u_j, r_{i[j]}) = 0$$

The big idea: We write down a new linear model for the random intercepts. This is our "level 2 model." Multilevel modeling is regression at level 1 and level 2.

How R thinks of this model (The model in reduced form.)

$$y_i = \alpha_{j[i]} + \varepsilon_i$$ **Level 1 Model**

$$\alpha_j = \gamma_0 + \gamma_1 sector_j + u_j$$ **Level 2 Model**

So

Plug our level 2 model in to the level 1 model

$$y_i = \gamma_0 + \gamma_1 sector_{j[i]} + u_{j[i]} + \varepsilon_i$$

Our regression equation. **Our two residuals.**

This is how R thinks of the model.

We are left with a reduced form, single model, with both level-1 and level-2 terms in it.

Impact of the Level-2 Predictor

mathach ~ 1 + (1 | id)

```
coef.est  coef.se
   12.64     0.24

Error terms:
 Groups   Name        Std.Dev.
 id       (Intercept) 2.93
 Residual             6.26
---
number of obs: 7185, groups: id, 160
AIC = 47122.8, DIC = 47114.8
deviance = 47115.8
```

mathach ~ 1 + sector + (1 | id)

```
                coef.est coef.se
(Intercept)     11.39     0.29
sectorCatholic   2.80     0.44

Error terms:
 Groups   Name        Std.Dev.
 id       (Intercept) 2.58
 Residual             6.26
---
number of obs: 7185, groups: id, 160
AIC = 47088.1, DIC = 47078.1
deviance = 47079.1
```

How has the meaning of the intercept changed?

The school-level variation has declined, but is still substantial.

Why did student residual variability not change?

Writing out the fitted model

$$MATHACH_i = a_{j[i]} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma_y^2) = N(0, 6.26^2)$$

$$a_j = \gamma_0 + \gamma_1 CATHOLIC_j + u_j$$

$$= 11.39 + 2.80 \cdot CATHOLIC_j + u_j$$

$$u_j \sim N(0, \sigma_a^2) = N(0, 2.58^2)$$

Changed the variable name to CATHOLIC to make the meaning more clear.

Or, reduced form:

$$MATHACH_i = 11.39 + 2.80 \cdot CATHOLIC_j + \varepsilon_i$$

$$\varepsilon_i \sim N(0, 6.26^2)$$

$$u_j = simN(0, 2.58^2)$$

mathach ~ 1 + sector + (1 | id)

```
                coef.est coef.se
(Intercept)     11.39     0.29
sectorCatholic   2.80     0.44

Error terms:
 Groups   Name        Std.Dev.
 id       (Intercept) 2.58
 Residual             6.26
```

A taste of inference: confidence intervals

```
> display( M2 )
lmer(formula = mathach ~ 1 + sector + (1 | id), data =
dat)
             coef.est coef.se
(Intercept) 11.39     0.29
sector       2.80     0.44
…
```

These are the standard errors for the fixed effects

```
> confint( M2 )
Computing profile confidence intervals ...
                  2.5 %    97.5 %
.sig01         2.265614  2.918854
.sigma         6.155051  6.362058
(Intercept) 10.819305 11.966760
sector         1.944660  3.665259
```

These "profile confidence intervals" give nominal confidence intervals for all our parameters, including the variance parameters.

Do not trust the variance parameter intervals too, too much.

Conceptual questions

How can we interpret the grand mean (11.72)?

In an average public school, the mean value of MATHACH for a student is 11.72

How can we interpret the coefficient of CATHOLIC (2.80)?

On average students at Catholic schools tend to score 2.80 points higher than students at public schools.

Suppose we had a null-hypothesis that there was no school-level variability in MATHACH, once we adjusted for sector. What would that mean?

Every public school has the same mean value of MATHACH, and similarly for every Catholic school

Commentary

Aggregating at the school level gives you estimates for the schools without regard to your school-level covariates. You can then look for how they vary by school-level covariate. (We did this in our R demo.)

This could make it easier to protect yourself against claims of a "built in relationship" due to model misspecification.

But, if you are using the MLM framework, you usually should use it completely. Drink ***all*** the Kool-aid.

There and back again: Translating models to R & identifying R output

In-class exercise

Write down a mathematical model (in two-level notation) for regressing math achievement onto SES (at level 1), and sector (at level 2).

Make it an R command. Fit it.

Identify the estimated values for all parameters in the original model.


{Even more in class R coding.}


Getting the estimates: fixef(), ranef() and coef()

```
> fixef( M2 )
(Intercept)         ses      sector
  11.718908    2.374711    2.100837

> length( ranef( M2 ) )
[1] 1

> head( ranef( M2 )$id )
     (Intercept)
1224  -0.8007023
1288   1.0721120
1296  -2.5405496

> head( coef( M2 )$id )
     (Intercept)      ses    sector
1224   10.918205 2.374711 2.100837
1288   12.791020 2.374711 2.100837
1296    9.178358 2.374711 2.100837
```

These are offsets; they tell us how much each school's intercept differs from the grand mean

Note we have an entire data frame for each grouping variable. In this example, there is only school id.

These are school-level equations (including slope)

Plotting all the school lines…

```
> coefs = coef( M3 )$id
> head( coefs )
      (Intercept)      ses   sector
1224    10.918206 2.374711 2.100837
1288    12.791019 2.374711 2.100837
1296     9.178358 2.374711 2.100837
…
```

The raw output from our lmer() call

```
> coefs$id = rownames( coefs )
```

We put the IDs back in

Also name our variables better

```
> coefs = rename( coefs, alpha = `(Intercept)`, beta=`ses`, gamma1 = `sector`  )
```

```
> coefs = merge( coefs, sdat, by="id", all=TRUE )
```

Add in our school level covariates

```
> head( coefs )
```

Our data! Each row is a school.

```
    id     alpha     beta   gamma1 size sector pracad disclim himinty meanses
1 1224 10.918206 2.374711 2.100837  842      0   0.35   1.597       0  -0.428
2 1288 12.791019 2.374711 2.100837 1855      0   0.27   0.174       0   0.128
3 1296  9.178358 2.374711 2.100837 1719      0   0.32  -0.137       1  -0.420
…
```

Plotting all the school lines (continued)

KEY CONCEPT LINE

gamma1 is the catholic school effect

Sector is 0/1 for each school

```
> coefs$alpha = coefs$alpha + coefs$gamma1 * coefs$sector
```

We need to adjust our catholic schools by the impact of catholic school.
Each school intercept needs to get a bump from this school level covariate.

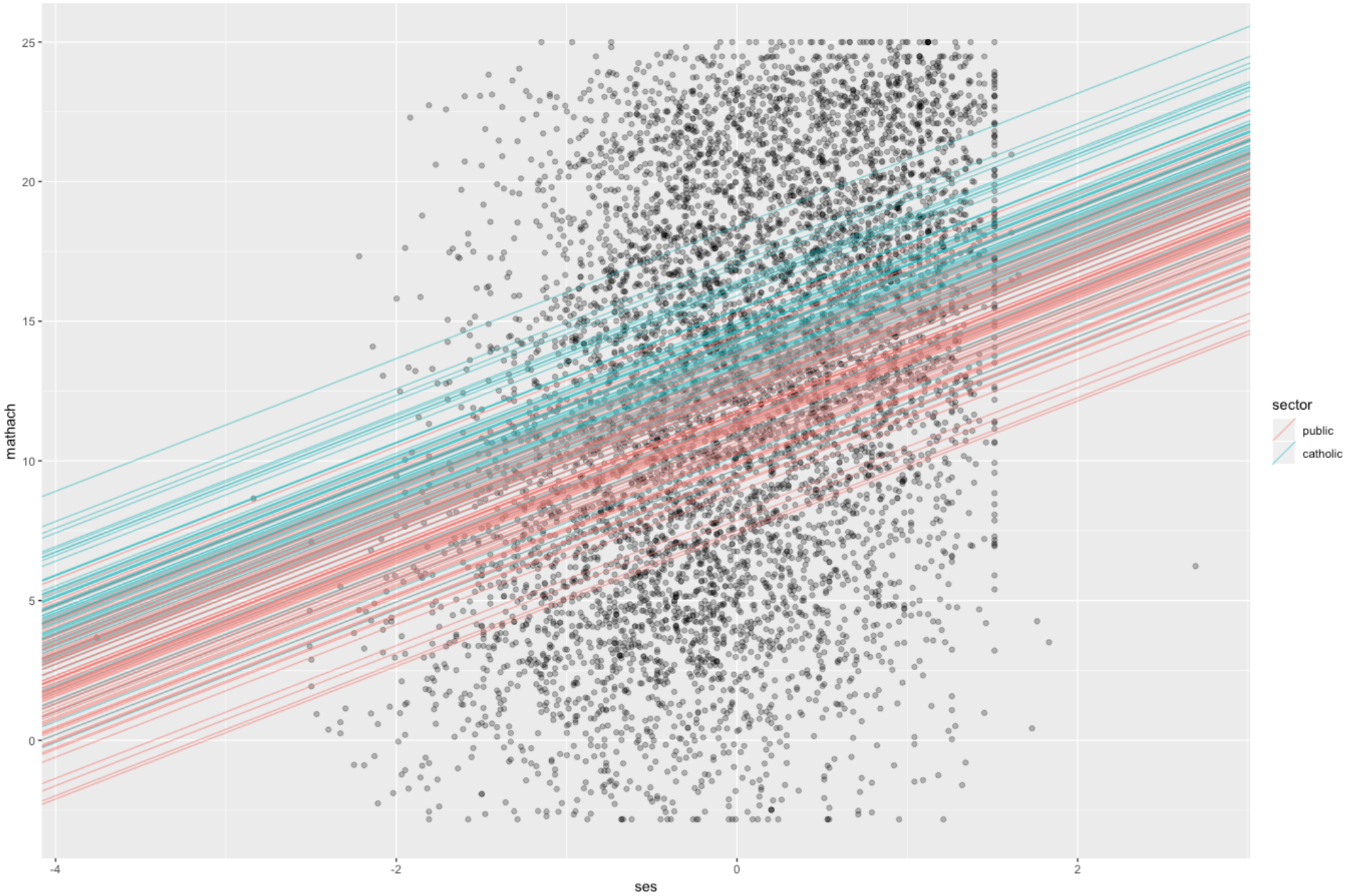
Make sector a factor so R gives us nice labels

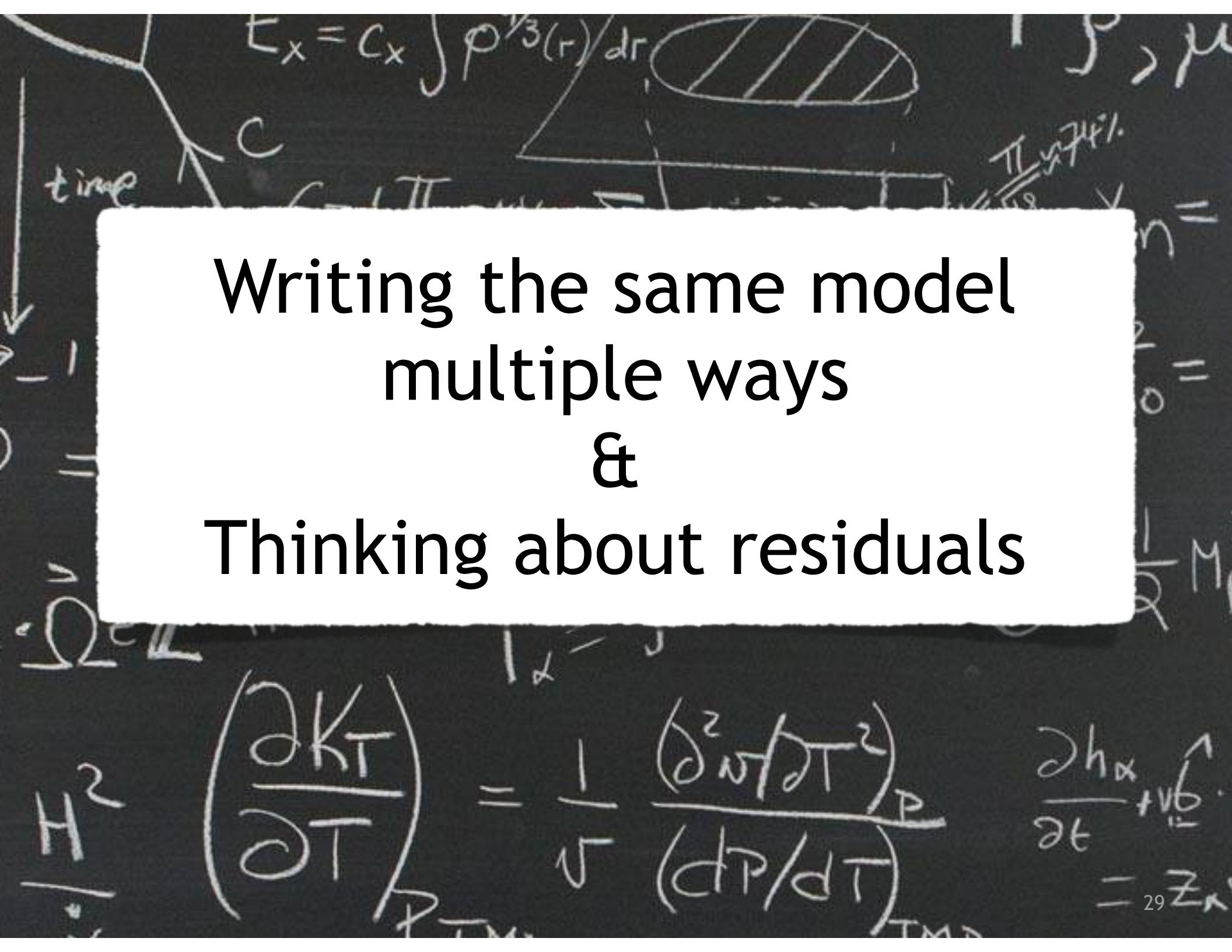
```
> coefs = mutate( coefs,
+                 sector = factor( sector,
+                                  levels=c(0,1),
+                                  labels=c("public","catholic" ) ) )
```

Make our plot! (students and the lines)

```
> ggplot( all.dat, aes( ses, mathach ) ) +
+   geom_point( alpha=0.3 ) +
+   geom_abline( data = coefs, aes( intercept=alpha, slope=beta, col=sector ),
                                  alpha=0.5 )
```

Catholic schools vs Public Schools





Writing the same model multiple ways
&
Thinking about residuals

Sanity Check: Why is this WRONG???

```
> M.ols = lm( mathach ~ 1 + ses + sector,
                    data=dat )
> display( M.ols )
lm(formula = mathach ~ 1 + ses + sector,
data = dat)
            coef.est coef.se
(Intercept) 11.79     0.11
ses          2.95     0.10
sector       1.94     0.15
---
n = 7185, k = 3
residual sd = 6.35, R-Squared = 0.15
```

Consider our school model...

$$y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i$$

Level 1 Model

$$\epsilon_i \sim N(0, \sigma_y^2)$$

$$\alpha_j = \gamma_0 + \gamma_1 s_j + u_j$$

Level 2 Model

$$u_j \sim N(0, \sigma_\alpha^2)$$

Implied Assumptions:
Level 1 and Level 2 error terms are not correlated with each other or themselves. Knowing one tells you nothing about another.
The expected value of an error given the covariates is 0.
The variance of the error terms does not change with the covariate.

A working definition of independence

Two random variables (values) are independent if the knowledge of the value of one of your variables (in addition to overall knowledge of averages and so forth) tells you nothing about the other.

An apparent conundrum (that isn't)

Take two students in the same school

- ★ Knowing that one student is above average across all schools gives us a hint that the other student is likely above average also.
- ★ Knowing one student is above school average does not tell us if the other is, however.

This translates into:

- ★ The overall (combined) residuals of these two students is correlated.
- ★ The individual component of their residuals are independent.
- ★ The MLM framework models the dependencies in our data.

We often see models as double-indexed

For student i in school j we have

$$
\begin{aligned}
y_{ij} &= \alpha_j + \beta x_{ij} + \epsilon_{ij} \\
&= \gamma_0 + \gamma_1 s_j + \beta x_{ij} + u_j + \epsilon_{ij} \\
\epsilon_{ij} &\sim N(0, \sigma_y^2) \\
u_j &\sim N(0, \sigma_\alpha^2)
\end{aligned}
$$

...collapsed into a single regression

$$
\begin{aligned}
y_i &= \alpha_{j[i]} + \beta x_i + \epsilon_i \\
&= (\gamma_0 + \gamma_1 s_j + u_j) + \beta x_i \\
&= \gamma_0 + \gamma_1 s_j + u_j + \beta x_i + \epsilon_i \\
&= \gamma_0 + \gamma_1 s_j + \beta x_i + u_j + \epsilon_i
\end{aligned}
$$

This is the gateway to the econometric view

The u_j are the "school-level random effects"

The overall error term. These are NOT independent

What just happened?

Recap

Check-In
http://cs179.org/lec22

- ★ The Intraclass Correlation Coefficient (ICC) is a way of measuring how much variation is at the 2nd level vs 1st level
- ★ Multilevel modeling is actually fitting two regression models, one at level 1 and one at level 2.
- ★ R works with collapsed forms of these models, with the level 2 stuff substituted into level 1.
- ★ There are different notational conventions for writing models, but the models are the same.