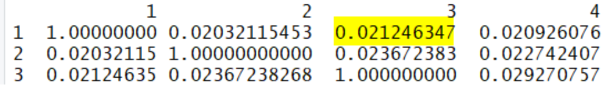
Stat 414 – Day 12
Adding Level 2 Variables (5.2)

Last Time: Random slopes, Centering

- Random slopes model the Var(Y_{ij}) and Cov(Y_{aj}, Y_{bj}) changing with x
    - Correlation between two language scores depending on their IQ scores
    - getVarCov(model, type = "marginal")



```
           1              2             3             4
1  1.00000000  0.02032115453  0.021246347  0.020926076
2  0.02032115  1.00000000000  0.023672383  0.022742407
3  0.02124635  0.02367238268  1.000000000  0.029270757
```

- Grand mean centering ($y_{ij} - \bar{y}$) vs. Group mean centering ($y_{ij} - \bar{y}_j$)
    - Not equivalent (e.g., can change fitted values)
    - Probably good idea to grand mean center all explanatory variables, especially with random coefficient models
    - Variance components now = expected variation for "average subject"
- The "group mean" (e.g., average IQ in the school) can be a very important "contextual" variable to include

Example 1: Continue Example 3 from Day 11
We now have a model that includes random intercepts (average IQ varies among schools), random slopes ("effect" of IQ varies among schools), IQ, mean IQ (explained some of the variation in intercepts). 2.47 effect IQ-verb for avg school (school w/ avg mean IQ)
−.18 decrease in effect of IQ-verb w/ each 1 unit increase in school mean IQ

(a) What are the implications/interpretation of adding the interaction between IQ and mean IQ to the model? How do we interpret the interaction for these centered variables?

$\beta_{0j} = \beta_{00} + \beta_{01}\,\text{meanIQ} + u_{0j}$ Level1: $Y_{ij} = \beta_{0j} + \beta_{1j}\,IQ + \epsilon_{ij}$

$\beta_{1j} = \beta_{10} + \beta_{11}\,\text{meanIQ} + u_{1j}$

composite equation:

$Y_{ij} = (\beta_{00} + \beta_{01}\,\text{meanIQ} + u_{0j}) + (\beta_{10} + \beta_{11}\,\text{meanIQ} + u_{1j})IQ + \epsilon_{ij}$

$+ \beta_{10}\,IQ + \beta_{11}\,\underline{\text{meanIQ} * IQ} + u_{1j}IQ + \cdots$

(b) Can you create a model that adds SES, mean SES, and all six interactions to the model? What might you first suggest to simplify the model? How do you do that in R?

remove corr between slopes IQ & slopes SES

(IQ + ses | schoolnr) (IQ | schoolnr) ~~(SES | schoolnr)~~ (−1 + SES | schoolnr)

(c) What might you next suggest to simplify the model?

treat SES w/ fixed slopes
+ remove the cross-level interactions
(test w/ LRT)

(d) Interpret the coefficients

$\hat{\tau}_0^2 = 8.45$ – school to school variation in intercepts (avg lang score)

$\hat{\tau}_1^2 = .165$ – school to school variance in slopes of IQ (after adjusting)

$\hat{\sigma}^2 = 37.38$ – variance in students within schools

cov – $\hat{\tau}_{01}$ −.79 – correlation between IQ slopes & intercepts

$\hat{\beta}_{00} = 41.62$ – expected lang score for student w/ mean ses, mean IQ, at school w/ avg IQ & avg SES

2.23 – expected increase in lang w/ 1 unit ↑ IQ adj for ses (ses = 0, avg ses), school avg IQ & avg SES

.709 – when at school w/ mean SES, w/ 1 unit ↑ in school mean IQ...

−.115 – if mean IQ at school is large then smaller slope mean SES school

$$Y_j = \beta_{00} + u_{0j} + \beta_1 x + u_{ij} x + \varepsilon_{ij}$$
$$= X\beta + Z\gamma$$

Model diagnostics

The assumptions we make in a multilevel model include:

- Linearity of the response with the explanatory variables
- Normality of the level 1 errors with constant variance (homogeneity)
- Normality of the (adjusted) level 2 random effects with constant variance (multivariate normal with constant covariance matrix)
- Independence of errors across levels

To check these conditions we can look at

$Y_{ij} - X\beta - Z\gamma$

- Level 1 ***conditional residuals*** vs. Level 1 "conditional" fitted values for equal variance
  - Conditional residuals are distances from observation to prediction for its group
  - These are what R returns with residuals(model)
  - Can also plot vs. other variables, use smoothers
- Normal probability plots/histograms of Level 1 residuals

β_1 vs $\beta_1 + u_j$

- Distributions of Level 2 random effects to check for unusual observations
  - Check for normality but doesn't always guarantee real effects follow normal distribution, check for outliers
  - This is what R returns with ranef(model)
  - Useful to plot vs. Level 2 units, other variables
  - ***Random effect residuals*** = response – fixed effects – conditional residuals

$Y - X\beta - \varepsilon = Z\gamma$

- Distribution of ***marginal residuals*** to check for unusual observations
  - Marginal residuals are distances from observation to overall prediction
  - In R: response - model.matrix(model) %*% fixef(model
  - Accounts for (confounds) both random effects and random error
  - Can be informative to plot these across the groups (probably differ)

$Y - X\beta$

Example 2: Have a hypothetical dataset with 10 subjects with 4 temporal-based observations (one every year). Each person has data for age, sex, average number of cigarettes smoked each week, level of nicotine dependence from the Fagerstrom Test of Nicotine Dependence (FTND), ratings of depressive symptoms from the Beck Depression Inventory (BDI), and a count variable for the total number of lifetime major depressive episodes suffered up to that point of data collection.

(a) Fit a model predicting cigarette use based on time and self-reported depression (BDI), including their interaction, and FTND score, with random intercepts for subjects (subjectID) and separate random slopes for FTND (uncorrelated with the intercepts).

R tip: Use (1 | subjectID) + (0 + x | subjectID)

| FTND

Y = Time + BDI + Time*BDI + (1|ID) + FTND + (0+ FTND |ID)

(b) Based on your model predict the first subject's Cig usage at time = 1.

$= -9\,58 + 4.6(1) + .60(7) + -.21(1)(7) + 1.42(6) = 6.3195$

(c) Determine the first subject's random FTD slope effect. Now what do you predict?

ranef(model)[1]

$6.3195 + 0 + .0336(6) = 6.539$

(d) Compute the marginal and conditional residuals for this subject.

observed 10

residual = 10 - 6.539 conditional = 3.46

10 - 6.3195 marginal = 3.68

(e) Plot the conditional residuals vs. the predicted values. Look for equal variance across the fitted values, outliers. Also check for normality.

```
plot(residuals(model) ~ fitted.values(model)); qqnorm(residuals(model))
```

(f) A fancier check of the equal variance assumption that can also point to remedies is to plot the squared residuals vs. explanatory variables (again use smoothers) and to run an ANOVA on the squared residuals vs. subjectID (ala Levene's Test).

```
anova(lm(squaredresids~as.factor(subjectID)))
```

p-value = .1729, flat smoother => no evidence of changes in var subject to subject

(g) Examine a graph of the marginal residuals vs. the marginal fitted values

```
fits = model.matrix(model) %*% fixef(model)
margresids = Cigs -fits; plot(margresids~ fits)
```

obs 28 still unusual

(h) Look at graphs of the random slopes. Do they seem normally distributed?

```
R: qqnorm(ranef(model)[[1]]$FTND)
```

intercepts & slopes each look ok (linear qq plot)

Example 3: Have data on 38 schools in London. The response is an end-of-year test and possible explanatory variables include gender, verbal reasoning level (high, medium, low) and the LRT (London Reading Test), school gender (all boy, all girl, mixed), and school denomination (ther, CofE, RomCath, State). school-frame.txt

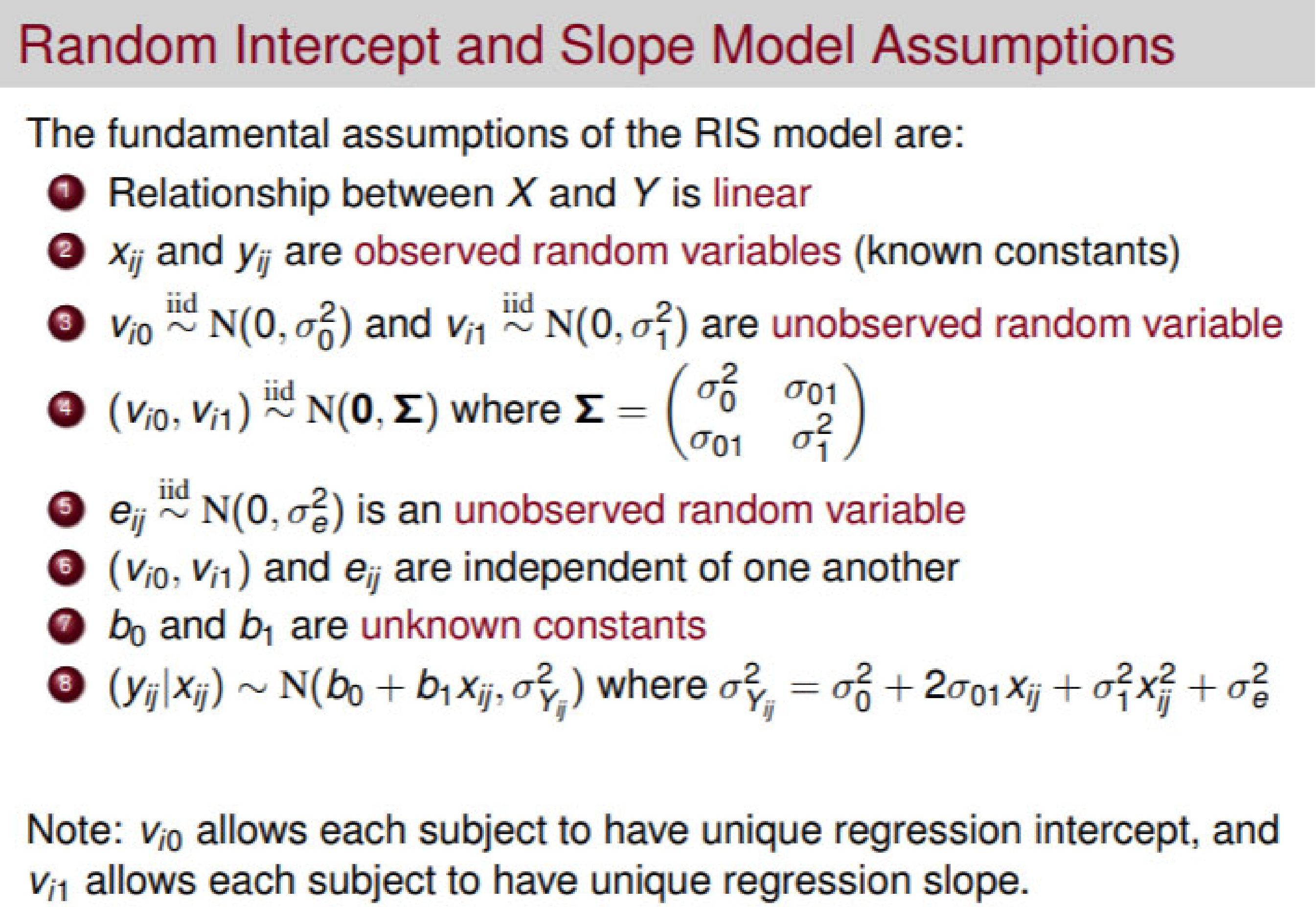
(a) Fit a model with random intercepts and random slopes for LRT.

(b) Look at plots of the conditional residuals across school.

```
ggplot(data=londondata, aes(x=index, y = residuals(model))) +geom_point(pch=1,color="Blue") +
facet_wrap(~school) +  geom_hline(yintercept=0)
```

(c) For this model, the random effects residuals depend on both the random intercepts and the random slopes. Compute these residuals and examine the graph vs. subject and then by index for each school school.

```
R: raneffresids = Test - model.matrix(model) %*% fixef(model)  - residuals(model)
ggplot(data=london,aes(x=index,y= raneffresids)) +   facet_wrap( ~ school, as.table=F)
+     geom_point(pch=1,color="Blue") +   geom_hline(yintercept=0)
```



Random Intercept and Slope Model Assumptions

The fundamental assumptions of the RIS model are:

1. Relationship between X and Y is linear
2. x_{ij} and y_{ij} are observed random variables (known constants)
3. $v_{i0} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_0^2)$ and $v_{i1} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_1^2)$ are unobserved random variable
4. $(v_{i0}, v_{i1}) \overset{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma} = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}$
5. $e_{ij} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_e^2)$ is an unobserved random variable
6. (v_{i0}, v_{i1}) and e_{ij} are independent of one another
7. b_0 and b_1 are unknown constants
8. $(y_{ij}|x_{ij}) \sim \mathrm{N}(b_0 + b_1 x_{ij}, \sigma^2_{Y_{ij}})$ where $\sigma^2_{Y_{ij}} = \sigma_0^2 + 2\sigma_{01} x_{ij} + \sigma_1^2 x_{ij}^2 + \sigma_e^2$

Note: v_{i0} allows each subject to have unique regression intercept, and v_{i1} allows each subject to have unique regression slope.

http://users.stat.umn.edu/~helwig/notes/lmer-Notes.pdf