Value Added Modeling

Dr. J. Kyle Roberts

Southern Methodist University
Simmons School of Education and Human Development
Department of Teaching and Learning
Background for VAMs

- Recall from previous lectures that we can define a growth model as:

\[
Y_{ti} = \pi_0 + \pi_1 t + e_{ti} \\
\pi_0 = \beta_0 + u_{0i} \\
\pi_1 = \beta_1 + u_{1i}
\]

- This is a two-level model with random intercepts and a random component for the time variable \( t \).

- This can also be viewed as a full model where:

\[
Y_{ti} = (\beta_0 + u_{0i}) + (\beta_1 + u_{1i}) \cdot t + e_{ti} \\
= [\beta_0 + \beta_1 \cdot t] + [u_{0i} + u_{1i} \cdot t + e_{ti}]
\]

\[\delta_i \sim N(0, \Omega), \quad \epsilon_{ti} \sim N(0, \sigma^2 I)\]
Setting Up The Model

- Before fitting the model we use two diagnostic plots to examine the data.
- First, we examine the functional form of the data using within-subject linear fits using OLS. This is accomplished using the \texttt{lmList} function.
- Second, we examine the variability of the intercepts and slopes to decide which random effects will be permitted to vary. This is accomplished using \texttt{confint}.
- The dataset we will be using is the \texttt{star} dataset in the \texttt{mlmRev} package and can be loaded simply by running \texttt{library(mlmRev)}.
Within-Subject Linear Fits

- Before fitting a linear model it is useful to examine the functional form of the data
- Instead of plotting the entire dataset we take a random sample of 50 students (because our dataset is so large)

```r
> samp <- sample(names(which(table(subset(star, + !is.na(math)))$id) > 1)), 50)
> datsamp <- subset(star, id %in% samp)
> child.lm <- lmList(math ~ yrs | id, datsamp, na.action = na.omit)
```
- The first two lines take the subset
- The last command fits the OLS regression and stores them in the object `child.lm`. 
Defining and fitting VAMs

Basic Linear Growth Model with Random Effects

Checking Assumptions

```r
> xyplot(math ~ yrs | id, datsamp, type = c("g", "r", "p"), aspect = "xy", layout = c(17, 3))
```
> xyplot(..., index.cond = function(x, y) predict(lm(y ~ x), list(x = -2.5)))
Within-Subject Confidence Intervals

• In some multilevel packages one typically fits the unconditional growth model and examines the chi-square statistic associated with the variance components. If the p-value is statistically significant, the random effects are retained.

• In R, this is accomplished a little differently. First, we examine the data graphically and we subsequently use a likelihood ratio test.

• Creating the visual displays requires use of the `confint` command.

• We have already created an object with the within-subject linear fits. We can now use the `plot` function to plot the intervals.
Defining and fitting VAMs

Basic Linear Growth Model with Random Effects

Checking Assumptions

> plot(confint(child.lm))
Fitting the VAM

- First, we will fit linear growth model with random effects for time.

\[ Y_{ti} = [\mu + \beta_1(year)] + [\delta_0 + \delta_1(year) + \epsilon_{ti}] \]

```r
> nograd <- list(gradient = FALSE, niterEM = 0)
> fm1 <- lmer(math ~ yrs + (yrs | id), star, control = nograd)
```

- The portion of code after the `control` statement turns off the analytic gradient and tells R to use no EM iterations. This is used to speed up computations when the number of grouping factors becomes exceedingly large.
Model Results

> fm1

Linear mixed model fit by REML
Formula: math ~ yrs + (yrs | id)
   Data: star

    AIC   BIC logLik deviance REMLdev
247325 247373  -123656  247311  247313

Random effects:
   Groups   Name    Variance  Std.Dev.  Corr
      id  (Intercept)  1506.605  38.8150
        yrs            68.418   8.2715  -0.405
       Residual       600.937  24.5140
Number of obs: 24613, groups: id, 10767

Fixed effects:
    Estimate  Std. Error   t value
(Intercept)  484.2708      0.4947  978.9
         yrs    43.2031      0.1962   220.2

Correlation of Fixed Effects:
   (Intr)  yrs
         yrs -0.612
Model Modifications

- We can now test our original random effects model against a model in which we do not allow for the time variable \( \text{(yrs)} \) to have random effects

\[
\text{> fm2 <- lmer(math ~ yrs + (1 | id), star, control = nograd)}
\]

\[
\text{> anova(fm1, fm2)}
\]

Data: star
Models:
\[\text{fm2: math ~ yrs + (1 | id)}\]  
\[\text{fm1: math ~ yrs + (yrs | id)}\]

\[
\begin{array}{llllllll}
\text{Df} & \text{AIC} & \text{BIC} & \text{logLik} & \text{Chisq} & \text{Chi Df} & \text{Pr(>Chisq)} \\
\text{fm2} & 4 & 247579 & 247612 & -123786 & \\
\text{fm1} & 6 & 247323 & 247372 & -123656 & 259.8 & 2 & < 2.2e-16
\end{array}
\]

- Testing these models against each other reveals that, in deed, the random effect for \( \text{yrs} \) is warranted.
Checking Assumptions

- The distributional assumptions underlying \( fm1 \) are that the within-group residuals are distributed as \( \epsilon_{ti} \sim \mathcal{N}(0, \sigma^2 I) \) and the random effects are distributed as \( \delta_i \sim \mathcal{N}(0, \Omega) \).
- Both of these assumptions must be examined.
> with(fm1, xyplot(resid(.) ~ fitted(.) | gr, type = c("g", + "p")))
```r
> qqmath(~resid(fm1))
```
> qqmath(~ranef(fm1)$id$'(Intercept)'), cex = 0.7)
> qqmath(~ranef(fm1)$id$yrs, cex = 0.7)
> plot(ranef(fm1))
$id$
Extending the Model

- The **Project Star** dataframe includes unique school IDs so we can extend this to a three-level model and examine the school effects.

- We previously noted how the structure of the random effects can be modified to handle additional “levels”. R can handle an arbitrary number of levels when fitting models with nested random effects.

- The portion of code `(yrs|sch:id)` denotes that students are nested within schools. We could also run this portion of the code as `(yrs|id)+(yrs|sch)`, such that we have defined `yrs` to have random effects for both individuals and schools.
Extending the Model (cont.)

- We specify the model as:

\[ Y_{ti} = [\mu + \beta_1(time)] + [\theta_{0j(i)} + \theta_{1j(i)}(time) + \delta_0 + \delta_1(time) + \epsilon_{ti}] \]

\[ \theta_j \sim \mathcal{N}(0, \Phi), \quad \delta_i \sim \mathcal{N}(0, \Omega), \quad \epsilon_{ti} \sim \mathcal{N}(0, \sigma^2 I) \]
Examining the Variance Components

- As we did before, we use the `lmList` function to produce within-school confidence intervals.
- The only difference here is that we have 80 schools, so we do not take a subset.

```r
> school.lm <- lmList(math ~ yrs | sch, data = star)
```
Defining and fitting VAMs

Basic Linear Growth Model with Random Effects

Checking Assumptions

> plot(confint(school.lm))
Fitting the VAM

- The trellis plots suggest sufficient variability to proceed with random intercepts and random slopes at the school level

```r
> sch.lmer <- lmer(math ~ yrs + (yrs | id) + (yrs | sch), star, control = nograd)

Linear mixed model fit by REML
Formula: math ~ yrs + (yrs | id) + (yrs | sch)
Data: star
AIC  BIC  logLik deviance REMLdev
244619 244692 -122301 244606 244601
Random effects:
  Groups   Name     Variance Std.Dev. Corr
       id  (Intercept) 1134.316  33.6796
        yrs          19.026   4.3619 -0.339
     sch  (Intercept) 341.751  18.4865
        yrs          58.019   7.6170 -0.614
  Residual            600.161  24.4982
Number of obs: 24613, groups: id, 10767; sch, 80
Fixed effects:
```

```r
> sch.lmer

Linear mixed model fit by REML
Formula: math ~ yrs + (yrs | id) + (yrs | sch)
Data: star
AIC  BIC  logLik deviance REMLdev
244619 244692 -122301 244606 244601
Random effects:
  Groups   Name     Variance Std.Dev. Corr
       id  (Intercept) 1134.316  33.6796
        yrs          19.026   4.3619 -0.339
     sch  (Intercept) 341.751  18.4865
        yrs          58.019   7.6170 -0.614
  Residual            600.161  24.4982
Number of obs: 24613, groups: id, 10767; sch, 80
Fixed effects:
```
Model Testing

- Suppose we wanted to modify the structure of the random effects such that we did not model random slopes for `yrs`.

\[ Y_{ti} = [\mu + \beta_1 \cdot time] + [\theta_{0j(i)} + \theta_{1j(i)} + \delta_0i + \epsilon_{ti}] \]

```r
> sch2.lmer <- lmer(math ~ yrs + (1 | id) + (1 | sch), star, control = nograd)
> anova(sch.lmer, sch2.lmer)
```

<table>
<thead>
<tr>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Chisq</th>
<th>Chi Df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sch2.lmer</td>
<td>5</td>
<td>246225</td>
<td>246265</td>
<td>-123107</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sch.lmer</td>
<td>9</td>
<td>244624</td>
<td>244697</td>
<td>-122303</td>
<td>1609.2</td>
<td>&lt; 2.2e-16</td>
</tr>
</tbody>
</table>
Looking at Effects

> head(ranef(sch.lmer)$sch)

(Intercept) yrs
1 11.003624 -4.254397
2 -30.828722 4.528853
3 16.818051 -5.660053
4 -9.848673 9.203151
5 -17.785526 0.764398
6 -15.086948 3.816181

> head(ranef(sch.lmer)$id)

(Intercept) yrs
100017 93.5158906 -4.107405381
100028 -40.9552884 1.798838369
100045 10.4070394 -0.014233537
100064 -0.1676666 0.007364254
100070 49.1877278 -1.999841478
100096 30.7611035 0.212732219