Longitudinal Data Analysis

Adam J Sullivan, PhD

04/18/2018

Two Stage Models

Two-Stage (Two-Level) Formulation

- We will proceed with Linear Mixed effects models.
- They are very useful in longitudinal as well as other hierarchical aspects.
- The basic idea of the model is that we assume
  1. **Stage 1**: A straight line (or more generally a "growth" curve) fits the observed responses for each subject.
  2. **Stage 2**: A Regression model relating the mean of the individual intercepts and slopes to the subject specific effects.

Stage 1

- In the first stage we assume that all subjects have their own unique trajectory.
- So for subject i:

$$Y_{ij} = Z_{ij}\beta_i + \varepsilon_{ij}, \qquad j = 1, \ldots, n_i$$

- where β_i is a vector of subject-specific regression parameters, the errors are typically considered independent within a subject.

Stage 1: Subject Specific Effects

- Many times we use a model with subject specific intercepts and slope:

$$Y_{ij} = \beta_{1i} + \beta_{2i} t_{ij} + e_{ij}$$

- So in stage 1 each subject has their own unique regression model.
  - Basically we allow each subject to have their own line.
  - We restrict the covariates in these models to be ones that vary over time.
- Any covariates that do not vary over time or refer to between-subject changes (sex, gender, treatment group, exposure group,…) are not included at this stage.

Stage 2

- In this stage we assume that the β_i's (subject-specific effects) are random and come from some distribution (IE. normal or some other).
- We then model the mean and covariance of the β_i's in the population.

$$\beta_i = A_i\beta + b_i, \text{ where } b_i \sim N(0, G)$$

Stage 2

- Where
  - A_i are the between subject covariates
  - $b_i = \begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix}$ are the random effects for individuals
  - $G = \begin{bmatrix} var(b_{1i}) & cov(b_{1i}, b_{2i}) \\ cov(b_{1i}, b_{2i}) & var(b_{2i}) \end{bmatrix}$ is the covariance matrix for the subject specific effects.

Quick Example

- Consider a treatment vs control setting where we have subject specific intercept, β_{1i}, and slope β_{2i}.
- Then we would model the subject specific effects with a group effect:

$$E(\beta_{1i}) = \beta_1 + \beta_2 \text{GROUP}_i$$
$$E(\beta_{1i}) = \beta_3 + \beta_4 \text{GROUP}_i$$

Quick Example

- Where GROUP_i is an indicator variable for treatment.
- Then in this example we would have the following models for means:

Quick Example

- For the control group:

$$E(\beta_{1i}) = \beta_1$$
$$E(\beta_{1i}) = \beta_3$$

Quick Example

- for the treatment group:

$$E(\beta_{1i}) = \beta_1 + \beta_2$$
$$E(\beta_{1i}) = \beta_3 + \beta_4$$

How do we fit these models:

- One approach has been coined as the "NIH Method" since it was popularized by statisticians working at the NIH.
- What they did was:
  1. Fit a regression to the response data for each subject.
  2. Regress the estimates of the individual intercepts and slopes on subject specific covariates.
- This method was very easy to perform because it did not require any special form of regression software.
- This works very well with balanced data.

Mixed Effects Models

Mixed Effects Models

- In contrast what we tend to do now is consider a model that contains the 2 stages but fits everything all at once:

$$\begin{aligned} Y_{ij} &= Z_{ij}\beta_i + \varepsilon_{ij} \\ &= Z_{ij}(A_i\beta + b_i) + \varepsilon_{ij} \\ &= Z_{ij}A_i\beta + Z_{ij}b_i + \varepsilon_{ij} \\ &= X_{ij}\beta + Z_{ij}b_i + \varepsilon_{ij} \end{aligned}$$

Mixed Effects Models

- We then have:
  - $X_{ij}\beta$ fixed effects (population)
  - $Z_{ij}b_i$ random effects (individual)

An Example

- To illustrate this we consider a study done on orthodontic measurement.
- Investigators at the University of North Carolina Dental School followed the growth of 27 children (16 males, 11 females) from age 8 until age 14.
- Every two years they measured the distance between the pituitary and the pterygomaxillary fissure, two points that are easily identified on x-ray exposures of the side of the head.

An Example

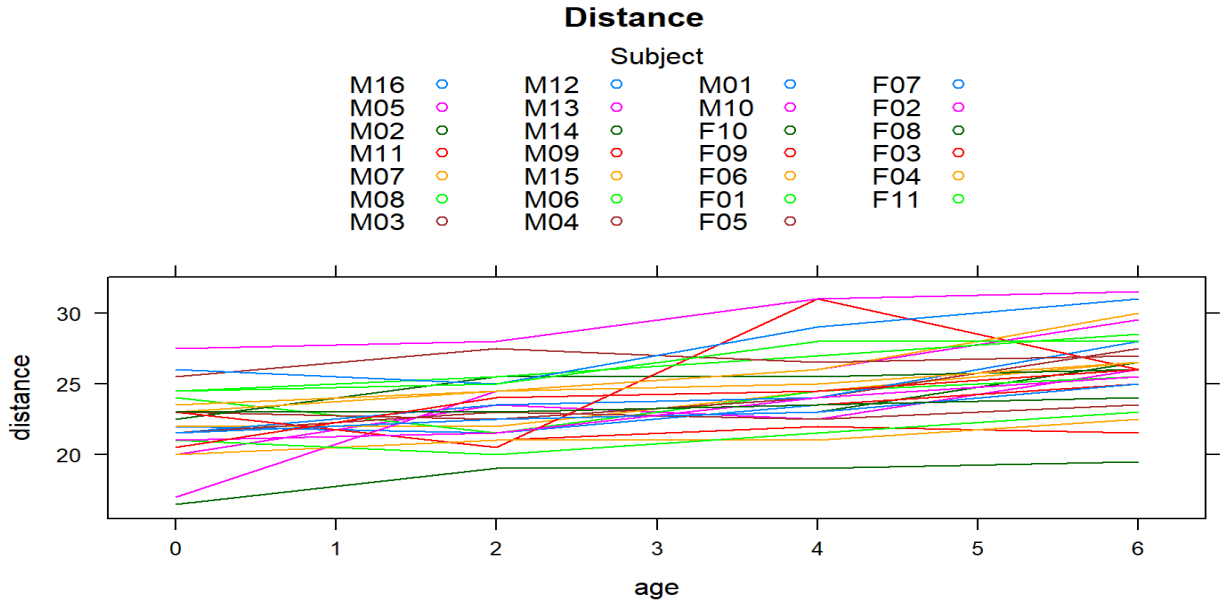
```
library(nlme)
head(Orthodont)
Orthodont$age <- Orthodont$age - 8
```

```
## Grouped Data: distance ~ age | Subject
##   distance age Subject  Sex
## 1     26.0   8     M01 Male
## 2     25.0  10     M01 Male
## 3     29.0  12     M01 Male
## 4     31.0  14     M01 Male
## 5     21.5   8     M02 Male
## 6     22.5  10     M02 Male
```

Example: Another Spaghetti Plot

```
library(lattice)
xyplot( distance ~ age , data= Orthodont, groups=Subject, type='l', auto.key=list(space="top", columns=4
                      title="Subject", cex.title=1), main="Distance")
```

Example: Another Spaghetti Plot



What do you see?

2 Stage Approach

- Now in the 2 stage approach we first would model the change in distance for each individual.

```
library(nlme)
reg.list <- lmList(distance ~ age, data=Orthodont)
summary(reg.list)
```

2 Stage Approach

```
Call:
  Model: distance ~ age | Subject
   Data: Orthodont

Coefficients:
   (Intercept)
    Estimate Std. Error t value Pr(>|t|)
M16     21.4        1.1    19.5 4.36e-26
M05     20.4        1.1    18.7 3.32e-25
M02     21.1        1.1    19.2 8.51e-26
M11     22.7        1.1    20.7 2.60e-27
M07     21.4        1.1    19.5 4.36e-26
M08     22.8        1.1    20.8 2.10e-27
M03     22.0        1.1    20.1 1.05e-26
M12     21.2        1.1    19.4 5.44e-26
M13     18.4        1.1    16.8 4.37e-23
M14     23.3        1.1    21.3 6.64e-28
M09     22.2        1.1    20.3 6.79e-27
M15     22.5        1.1    20.5 3.57e-27
M06     24.4        1.1    22.2 7.81e-29
M04     26.1        1.1    23.8 2.59e-30
M01     24.9        1.1    22.7 2.62e-29
M10     27.2        1.1    24.9 3.05e-31
```

2 Stage Approach

	age			
	Estimate	Std. Error	t value	Pr(>|t|)
M16	0.550	0.293	1.878	6.58e-02
M05	0.850	0.293	2.902	5.36e-03
M02	0.775	0.293	2.646	1.07e-02
M11	0.325	0.293	1.109	2.72e-01
M07	0.800	0.293	2.731	8.51e-03
M08	0.375	0.293	1.280	2.06e-01
M03	0.750	0.293	2.560	1.33e-02
M12	1.000	0.293	3.414	1.22e-03
M13	1.950	0.293	6.657	1.49e-08
M14	0.525	0.293	1.792	7.87e-02
M09	0.975	0.293	3.328	1.58e-03
M15	1.125	0.293	3.840	3.25e-04
M06	0.675	0.293	2.304	2.51e-02
M04	0.175	0.293	0.597	5.53e-01
M01	0.950	0.293	3.243	2.03e-03
M10	0.750	0.293	2.560	1.33e-02
F10	0.450	0.293	1.536	1.30e-01
F09	0.275	0.293	0.939	3.52e-01
F06	0.375	0.293	1.280	2.06e-01
F01	0.375	0.293	1.280	2.06e-01
F05	0.275	0.293	0.939	3.52e-01

Abstract Coefficients

- We can then abstract the estimated model coefficients and the variance-covariance matrices:

```
b <- lapply(reg.list, coef)
b
V <- lapply(reg.list, vcov)
V
```

Abstract Coefficients

```
## Error in lapply(reg.list, coef): object 'reg.list' not found
```

```
## Error in eval(expr, envir, enclos): object 'b' not found
```

```
## Error in lapply(reg.list, vcov): object 'reg.list' not found
```

```
## Error in eval(expr, envir, enclos): object 'V' not found
```

Abstract Coefficients

- An indicator variable of the estimate type (alternating intercept and slope) and a subject id variable are also needed, which can be created with:

```
estm <- rep(c("intercept","slope"), length(b))
estm
subj <- rep(names(b), each=2)
subj
```

Abstract Coefficients

```
## Error in eval(expr, envir, enclos): object 'b' not found
```

```
## Error in eval(expr, envir, enclos): object 'estm' not found
```

```
## Error in eval(expr, envir, enclos): object 'b' not found
```

```
## Error in eval(expr, envir, enclos): object 'subj' not found
```

Variance Covariance

- Next, we create one long vector with the model coefficients and the corresponding block-diagonal variance-covariance matrix with (the metafor package needs to be loaded for the bldiag() function):

```
library(metafor)
b <- unlist(b)
V <- bldiag(V)
```

Variance Covariance

```
## Error in unlist(b): object 'b' not found
```

```
## Error in bldiag(V): object 'V' not found
```

```
## Error in eval(expr, envir, enclos): object 'b' not found
```

Variance Covariance

```
## Error in eval(expr, envir, enclos): object 'V' not found
```

Final Model

- Finally, we conduct a multivariate meta-analysis with the model coefficients (since we have two correlated coefficients per subject). -The V matrix contains the variances and covariances of the sampling errors.
- We also allow for heterogeneity in the true outcomes (i.e., coefficients) and allow them to be correlated (by using an unstructured variance-covariance matrix for the true outcomes).

Final Model

- The model can be fitted with:

```
res2 <- rma.mv(b ~ estm-1, V, random = ~ estm | subj, struct="UN")
summary(res2)
```

Final Model

```
Multivariate Meta-Analysis Model (k = 54; method: REML)

  logLik  Deviance       AIC       BIC      AICc
-64.4574  128.9148  138.9148  148.6710  140.2192

Variance Components:

outer factor: subj (nlvls = 27)
inner factor: estm (nlvls = 2)

            estim    sqrt  k.lvl  fixed      level
tau^2.1    8.3710  2.8933     27     no  intercept
tau^2.2    0.0478  0.2187     27     no      slope
```

Final Model

```
           rho.intr  rho.slop    intr  slop
intercept         1    0.7394       -    no
slope        0.7394         1      27     -

Test for Residual Heterogeneity:
QE(df = 52) = 1611.6315, p-val < .0001

Test of Moderators (coefficient(s) 1:2):
QM(df = 2) = 3080.0214, p-val < .0001
```

Final Model

Model Results:

	estimate	se	zval	pval	ci.lb	ci.ub	
estmintercept	26.8868	0.5980	44.9609	<.0001	25.7148	28.0589	***
estmslope	0.5762	0.0555	10.3868	<.0001	0.4675	0.6850	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

What do we have?

We have:

- We have an estimated average intercept of $b_0 = 22.28$ (SE=0.410)
- an estimated average slope of $b_1 = 0.58$ (SE=0.056)
- estimated standard deviations of the underlying true intercepts and slopes equal to SD(b_{0i})=1.987 and SD(b_{1i})=0.219 , respectively.
- A correlation between the underlying true intercepts and slopes equal to $\hat{\rho} = 0.20$ (no residual standard deviation is given, since that source of variability is already incorporated into the V matrix).

Mixed Effects Model

Alternative with a Mixed Effects Model

- Alternatively we could have fit this with a mixed model:

```
reg.mix <- lme(distance ~ age, random = ~ age | Subject, data=Orthodont)
summary(reg.mix)
```

Alternative with a Mixed Effects Model

```
Linear mixed-effects model fit by REML
 Data: Orthodont
  AIC BIC logLik
  455 471   -221

Random effects:
 Formula: ~age | Subject
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev Corr
(Intercept) 2.875  (Intr)
age         0.226  0.767
Residual    1.310
```

Alternative with a Mixed Effects Model

```
Fixed effects: distance ~ age
            Value Std.Error DF t-value p-value
(Intercept) 27.32     0.634 80    43.1       0
age          0.66     0.071 80     9.3       0
 Correlation:
    (Intr)
age 0.762

Standardized Within-Group Residuals:
     Min       Q1      Med       Q3      Max
-3.22311 -0.49376  0.00732  0.47215  3.91603

Number of Observations: 108
Number of Groups: 27
```

What do we see?

- The estimated average distance at age 8 is $b_0 = 22.04$ millimeters (SE=0.420).
- For each year, the distance is estimated to increase on average by $b_1 = 0.66$ millimeters (SE=.071). - However, there is variability in the intercepts and slopes, as reflected by their estimated standard deviations (SD($_{0i}$)=1.887 and SD(b_{1i})=0.226, respectively). Also, intercepts and slopes appear to be somewhat correlated ($\hat{\rho} = -0.21$).
- Finally, residual variability remains (reflecting deviations of the measurements from the subject-specific regression lines), as given by the residual standard deviation of $\hat{\sigma} = 1.310$.

How did this model compare?

- Notice that when we fit this with one model we have smaller standard errors.
- With this approach we are using all of the data at the same time and fitting them together.
- When the model is correctly specified the mixed model approach is preferred.

Adjusting for Sex

- At the same time, it is much easier for us to consider also adjusting for sex.
- This would not be done at stage one but stage 2.
- So in the case of a mixed model we would consider this to be part of the fixed effects but not the random effects:

```
reg.mix2 <- lme(distance ~ age + Sex, random = ~ age | Subject, data=Orthodont)
summary(reg.mix2)
```

Adjusting for Sex

```
Linear mixed-effects model fit by REML
 Data: Orthodont
  AIC BIC logLik
  449 468   -218

Random effects:
 Formula: ~age | Subject
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev Corr
(Intercept) 2.330  (Intr)
age         0.226  0.636
Residual    1.310
```

Adjusting for Sex

```
Fixed effects: distance ~ age + Sex
               Value Std.Error DF t-value p-value
(Intercept) 28.20     0.626 80    45.1   0.000
age          0.66     0.071 80     9.3   0.000
SexFemale   -2.15     0.757 25    -2.8   0.009
 Correlation:
            (Intr) age
age          0.635
SexFemale -0.493  0.000

Standardized Within-Group Residuals:
    Min      Q1     Med      Q3     Max
-3.0814 -0.4568  0.0155  0.4470  3.8944

Number of Observations: 108
Number of Groups: 27
```

What can we see?

- We can see that there does not appear to be a large change in the outcomes by adding sex even though it was significant.
- What we can see that that for Females at the mean age of 8, there is on average a 2.15 mm smaller distance than that of Males who are the same age.

Linear Mixed Effects Models

- We have the general framework

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i$$

- What we have here is the ability to model population characteristics (fixed effects) and allow for subject specific effects (random effects).
- This allows us to understand population information as well as allow for individuals to vary differently.

Random Intercept Model

- One approach that is often used to handle the covariance among repeated measures is to assume that is comes from a random subject effect.
- This would mean that each subject has an underlying difference in change that is constant over all measurements.
- We model this as:

$$Y_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots \beta_p X_{ijp} + b_i + \varepsilon_{ij}$$

Random Intercept Model

- Typically we have $X_{ij1} = 1$ for all subjects so that our model is now:

$$Y_{ij} = (\beta_1 + b_i) + \beta_2 X_{ij2} + \cdots \beta_p X_{ijp} + \varepsilon_{ij}$$

Random Intercept Models

- This means that we have population attributes for all $\beta_2 - \beta_p$ but we allow for individual intercepts $|beta_1 + b_i$.
- This is why we call it the random intercept model.
- This means that the population mean is:

$$E(Y_{ij}) = \beta_1 + \beta_2 X_{ij2} + \cdots \beta_p X_{ijp}$$

- because

$$b_i \sim N(0, \sigma_b^2) \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

Random Intercept Models

- It is not necessary for the errors to be normal but this is the case for a linear regression.
- For example if we considered a simple case of a trend over time

$$Y_{ij} = (\beta_1 + b_i) + \beta_2 t_{ij} + \varepsilon_{ij}$$

What the random effects tell us

- Consider two subjects:
  1. This subject responds at a higher level than the population so they would have a $b_i > 0$.
  2. This subject responds at a lower level than the population so they would have a $b_i < 0$.
- What we would see on a plot is that Subject 1 would have a line parallel to that of the population however it would be higher.
- Subject 2 would be the opposite.

Covariance and Correlation Structure

- We said before that we have two "random" terms here:

$$b_i \sim N(0, \sigma_b^2) \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

- This would imply that

$$Var(Y_{ij} = Var(b_j) + Var(\varepsilon_{ij}) = \sigma_b^2 + \sigma^2$$

Covariance and Correlation Structure

- Then by introducing a subject specific effect of b_i we have induced correlation among repeated measures.
- For example consider subject i at times points j and k:

$$Cov(Y_{ij}, Y_{ik}) = \sigma_b^2 \Rightarrow Corr(Y_{ij}, Y_{ik}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$$

Variance Matrix

- We then have the following Covariance Matrix:

$$\begin{bmatrix} \sigma_b^2 + \sigma^2 & \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma^2 & \cdots & \sigma_b^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma^2 & \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 + \sigma^2 \end{bmatrix}$$

Compound Symmetry Matrix

- We refer to this as a structure.
  - : Variances and correlations are constant across time occasions.
  - : Allow for heterogeneity in trends across times.

Random Intercept and Slope Model

- Instead of suggesting that individuals have a persistent constant difference across time points, we can allow their difference to change based on time.

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + b_{1i} + b_{2i} t_{ij} + \varepsilon_{ij} \qquad j = 1, \ldots, n_i$$

Random Intercept and Slope Model Example

- Consider if we have a treatment vs control study

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + \beta_3 trt_i \beta_4 t_{ij} \times trt_i + b_{1i} + b_{2i} t_{ij} + \varepsilon_{ij} \qquad j = 1, \ldots, n_i$$

- Then for the control group:

$$Y_{ij} = (\beta_1 + b_{1i}) + (\beta_2 + b_{2i}) t_{ij} + \varepsilon_{ij}$$

The Treated Group

- For the treated group we would have:

$$Y_{ij} = (\beta_1 + \beta_3 + b_{1i}) + (\beta_2 + \beta_4 b_{2i}) t_{ij} + \varepsilon_{ij}$$

Covariance and Correlation Structure

- Then we would have:
  - $b_{1i} \sim N(0, \sigma^2_{b_1})$
  - $b_{2i} \sim N(0, \sigma^2_{b_2})$
  - $Cov(b_{1i}, b_{2i}) = \sigma_{b_1, b_2}$
  - $\varepsilon_{ij} \sim N(0, \sigma^2)$

Covariance and Correlation Structure

- Thus we have that

$$\begin{aligned} Var(Y_{ij}) &= Var(b_{1i} + b_{2i}t_{ij}\varepsilon_{ij}) \\ &= Var(b_1 i) + 2t_{ij}Cov(b_{1i}, b_{2i}) + t_{ij}^2 Var(b_{2i}) + Var(\varepsilon_{ij}) \\ &= \sigma_{b_1}^2 + 2t_{ij}\sigma_{b_1,b_2} + t_{ij}^2\sigma_{b_2}^2 + \sigma^2 \end{aligned}$$

Covariance and Correlation Structure

We can also show that

$$Cov(Y_{ij}, Y_{ik}) = \sigma^2_{b_1} + (t_{ij} + t_{ik})\sigma_{b_1,b_2} + t_{ij}t_{ik}\sigma^2_{b_2} + \sigma^2$$

Estimation Techniques

- We already saw that linear regression used least squares estimation unfortunately there is no simple expression for the maximum likelihood estimator of the covariance components that we have.
- We rely on an iterative algorithm instead.

Maximizing the Likelihood

- Then what happens with this is that if we maximize the likelihood we end up with the same least squares estimator that we have before however our estimate of σ^2 is

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(y_i - X_i\beta)^2}{N}$$

- where N is the sample size, we could then show that

$$E(\hat{\sigma}^2) = \left(\frac{N-p}{N}\right)\sigma^2$$

Residual Maximum Likelihood

- This means that if we have small samples of size N than we underestimate the σ^2.
- Due to this we use what is called

Residual Maximum Likelihood

- Basically the bias we saw came from the fact that we do not know the true β so we replace this term with a $\hat{\beta}$ however this is an estimated value.
- We saw a similar problem when we tried to originally estimate variance and had to adjust for sample size.
- We will find that R, SAS and Stata use this as the default method.