Lecture 12 Handout: Models for Clustered Data

Elizabeth Colantuoni

3/2/2021

I. Objectives

Upon completion of this session, you will be able to do the following:

- define marginal models, conditional models and linear mixed models
- describe how longitudinal growth data is generated via subject specific growth rates
- implement a linear mixed model in R
- interpret key elements of linear mixed models applied to growth curves that are relevant for public health researchers

II. Two-approaches to modeling longitudinal data

- Descriptive: Marginal model, goal is to describe and make inference for the mean model. Have to account for the variance/correlation structure to get valid inferences but we don't necessary care about describing that structure.
- Etiologic: Conditional models, we are specifically interested in describing where the correlation comes from. E.g. the current observation may depend on the prior observation (transition model) OR each subject may be distinguished by latent variables/random effects which separate their data from other subjects data. Our goal is to describe the population level patterns (similar to marginal models) but we can also quantify heterogeneity across subjects in features of the data that are very important for public health researchers, e.g. variation in child specific growth rates.

III. Transition models

Here past values of the outcome cause future values of the outcomes. Namely, a transition model where the current value of Y_{ij} depends on the p past observations can be expressed as:

$$E(Y_{ij}|Y_{ij-1},...,Y_{ij-p},X_{ij}) = X_{ij}^{\prime}\beta^{c} + \sum_{k=1}^{p}\alpha_k Y_{ij-k}$$

where X_{ij} is the column vector of 1 plus covariates for the j^{th} observation for subject i.

The special case of the AR-1 model is where $p = 1$.

$$E(Y_{ij}|Y_{ij-1},X_{ij}) = X_{ij}^{\prime}\beta^{c} + \alpha Y_{ij-1}$$

Note that the models above make a strong assumption: the relationship between the mean of Y_{ij} and X_{ij} is the same regardless of the past values of Y. This assumption can be made flexible by including interaction terms of components of X_{ij} and past values of Y.

IV. Subject-specific models / Random effects models

A. Motivation based on subject-specific framework

Consider the data generating structure within the NEPAL1 and NEPAL2 simulated datasets:

- Children are enrolled between 1 and 5 months of age
- Children are followed over time and growth in weight is recorded every 4 months for a total of 5 assessments (enrollment + 4 follow-ups)

For each child, we can think of the child's growth:

$$Y_{ij} = \beta_{0i} + \beta_{1i} age_{ij} + \beta_{2i}(age_{ij} - 6)^{+} + e_{ij}, e_{ij} \sim N(0, \sigma^2)$$

where

- Y_{ij} is the weight for child i $(i = 1, ..., m)$ at assessment j $(j = 1, 2, 3, 4, 5)$
- β_{0i}, β_{1i} and β_{2i} define the child i specific expected birthweight, expected monthly growth rate while the child is less than 6 months of age, and the change in the monthly growth rate once the child is over 6 months of age, respectively
- $e_i j$ represent residuals that are specific to child i's data; this is natural heterogeneity/fluctuations in Y_{ij}. In the simplest version of this model, we assume $Corr(e_{ij}, e_{ik}) = 0$.

The parameters β_i describe characteristcs of the specific children and we assume that these characteristics can vary from child to child, specifically,

$$\begin{bmatrix} \beta_{0i} \\ \beta_{1i} \\ \beta_{2i} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{bmatrix}$$

$$\beta_i = \beta + b_i, b_i \sim MVN(0, D), D = \begin{bmatrix} \tau_0^2 & \tau_{01} & \tau_{02} \\ \tau_{01} & \tau_1^2 & \tau_{12} \\ \tau_{02} & \tau_{12} & \tau_2^2 \end{bmatrix}$$

- We refer to β as **fixed effects**, these parameters represent the population average birthweight, population average monthly growth rate in the first 6 months of age, and the population average change in monthly growth rate comparing growth after 6 months to growth prior to 6 months of age, respectively.
- We refer to b_i as **random effects**, these are latent/unobserved variables/residuals that tell us how different an individual child is from the population mean. For example, $\beta_{0i} = \beta_0 + b_{0i}$ such that b_{0i} represents the difference in expected birthweight for child i compared to the population average birthweight.
- The variances τ_0^2, τ_1^2 and τ_2^2 quantify the heterogeneity in children with respect to the population average. E.g. τ_0^2 is the variation in child birthweights and τ_1^2 is variation in child growth rates during the first 6 months of a life. NOTE: These variance estimates contain very useful information about heterogeneity across children in the population and we don't estimate these in the marginal model setting.

B. General framework

We can rewrite the model above as:

$$Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})age_{ij} + (\beta_2 + b_{2i})(age_{ij} - 6)^+ + e_{ij}$$

In vector notation,

$$Y_{ij} = \begin{bmatrix} 1 \\ age_{ij} \\ (age_{ij} - 6)^+ \end{bmatrix}' \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 1 \\ age_{ij} \\ (age_{ij} - 6)^+ \end{bmatrix}' \begin{bmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{bmatrix} + e_{ij}$$

Even more generally,

$$Y_{ij} = X_{ij}'\beta + Z_{ij}'b_i + e_{ij}$$

where $b_i \sim MVN(0, D)$, e_{ij} iid $N(0, \sigma^2)$ **and** b_i and e_{ij} are independent!

- In the random effects model, we express the mean function for an individual subject as:

$$E(Y_{ij}|X_{ij}, b_i) = X_{ij}'\beta + Z_{ij}b_i$$

- We can express the population mean (i.e. the average over all subjects) as:

$$E(Y_{ij}|X_{ij}) = E[E(Y_{ij}|X_{ij}, b_i)] = E[X_{ij}'\beta + Z_{ij}'b_i] = X_{ij}'\beta$$

- We can derive the variance of Y_{ij} as

$$\begin{aligned} Var(Y_{ij}|X_{ij}) &= E_{b_i}[Var(Y_{ij}|X_{ij}, b_i)] + Var_{b_i}[E(Y_{ij}|X_{ij}, b_i)] \\ &= E_{b_i}[\sigma^2] + Var_{b_i}[X_{ij}'\beta + Z_{ij}'b_i] \\ &= \sigma^2 + Z_{ij}'DZ_{ij} \end{aligned}$$

C. Correlation Structures

Consider a random interectp model, i.e. let $Z_{ij} = 1$ be column vector of 1s. So the model is:

$$Y_{ij} = (\beta_0 + b_{0i}) + \beta_1 age_{ij} + \beta_2(age_{ij} - 6)^+ + e_{ij}$$

where e_{ij} iid $N(0, \sigma^2)$ and $b_{0i} \sim N(0, \tau^2)$.

- What is $Corr(Y_{ij}, Y_{ik})$?

$$\begin{aligned} Cov(Y_{ij}, Y_{ik}) &= Cov((\beta_0 + b_{0i}) + \beta_1 age_{ij} + \beta_2(age_{ij} - 6)^+ + e_{ij}, (\beta_0 + b_{0i}) + \beta_1 age_{ik} + \beta_2(age_{ik} - 6)^+ + e_{ik}) \\ &= Cov(b_{0i} + e_{ij}, b_{0i} + e_{ik}) \\ &= Cov(b_{0i}, b_{0i}) \\ &= \tau_0^2 \end{aligned}$$

$$Corr(Y_{ij}, Y_{ik}) = \frac{Cov(Y_{ij}, Y_{ik})}{\sqrt{(\tau_0^2 + \sigma^2)(\tau_0^2 + \sigma^2)}} = \frac{\tau_0^2}{\tau_0^2 + \sigma^2}$$

Does this make sense to you?

D. You do:

You derive for $X_{ij} = (1, age_{ij}, (age_{ij} - 6)^+)$, $Z_{ij} = (1, age_{ij})$

1. $E(Y_{ij}|X_{ij})$
2. $Corr(Y_{ij}, Y_{ik})$

E. Link to Marginal Models

We can re-express our linear mixed model as:

$$E(Y_i|X_i) = X_i\beta$$
$$Var(Y_i|X_i) = \sigma^2 I + Z_i D Z_i'$$

The random effects models define a marginal model with a specific covariance model.

F. Example: NEPAL1

Let's fit a random intercept and random slope for age model to the NEPAL1 dataset:

$$Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})age_{ij} + \beta_2(age_{ij} - 6)^+ + e_{ij}$$

where e_{ij} iid $N(0, \sigma^2)$ and $b_i \sim MVN(0, D_{2\times 2})$.

```
load("nepal_simulated.rda")
fit = lmer(wt~age+age_sp6+(1+age|id),data=nepal1,control = lmerControl(optimizer ="Nelder_Mead"))
summary(fit)$coefficients
```

```
##                Estimate Std. Error   t value
## (Intercept)  4.9777731 0.15426618  32.26743
## age          0.4984283 0.01867078  26.69563
## age_sp6     -0.3497761 0.01802296 -19.40725
```

```
summary(fit)$varcor
```

```
##  Groups   Name        Std.Dev. Corr
##  id       (Intercept) 1.045047
##           age         0.082252 -0.345
##  Residual             0.281274
```

```
est = fixef(fit)
```

1. What is the estimate of the population mean birthweight?

Here we want to provide the estimate of $\hat{\beta}_0$:
The estimated population average/mean birthweight is 5 kg, with 95% CI: 4.7 to 5.3.

2. What is the estimate of the population average growth rate during the first 6 months of life?

Here we want to provide the estimate of $\hat{\beta}_1$:
The estimated population average/mean monthly growth per month is 0.5 with 95% CI: 0.46 to 0.54

3. What is the estimated difference in the population average growth rates comparing children older than 6 months to children less than 6 months?

Here we want to provide the estimate of $\hat{\beta}_2$:
The estimated population average/mean monthly growth per month is -0.35 with 95% CI: -0.39 to -0.31

4. For a given child at a specific age, how much does the observed/measured weights differ from the child's average weight at that age?

Here we are providing the estimate of σ^2: 0.28

5. Construct an interval that contains 95% of birthweights for Nepali children.

The population average/mean birthweight is given by $\hat{\beta}_0$.
The $\sqrt{var(b_{0i})} = \sqrt{\tau_0^2} = \tau_0$ and $\hat{\tau}_0 = 1.05$ represents the standard deviation of nepali children birthweights.

So we would expect that 95% of birthweights would fall within 4.98 $\pm 2\times$ 1.05 kg, which is: 2.89 to 7.07 kg.

6. Construct an interval that contains 95% of growth rates for Nepali children under 6 months of age.

The population average growth rate for Nepali children under 6 months of age is given by $\hat{\beta}_1$.
The $\sqrt{var(b_{1i})} = \sqrt{\tau_1^2} = \tau_1$ and $\hat{\tau}_1 = 0.08$ represents the standard deviation of nepali children's growth rates for the first 6 months of life.

So we would expect that 95% of children to have monthly growth rates during the first 6 months of life to fall within

0.5 $\pm 2\times$ 0.08 kg, which is: 0.33 to 0.66 kg.

7. Construct an interval that contains 95% of growth rates for Nepali children over 6 momths of age.

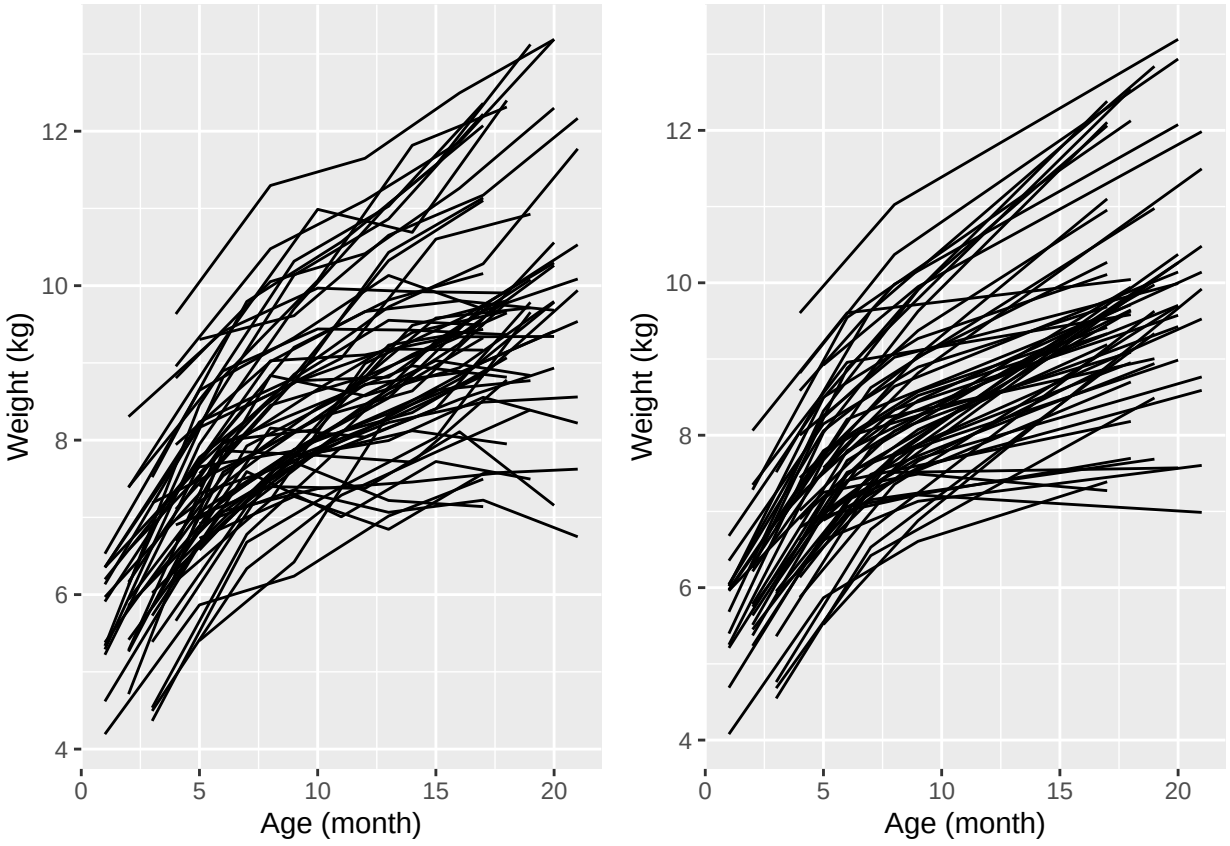
The population average growth rate for Nepali children over 6 months of age is given by $\hat{\beta}_1 + \hat{\beta}_2$.
The $\sqrt{var(b_{1i})} = \sqrt{\tau_1^2} = \tau_1$ and $\hat{\tau}_1 = 0.08$ represents the standard deviation of nepali children's growth rates after 6 months of age.
NOTE: For an individual child, the growth rate after 6 months is given by $\beta_1+\beta_2+b_{1i}$ and $Var(\beta_1+\beta_2+b_{1i}) = Var(b_{1i}) = \tau_1^2$.

So we would expect that 95% of children to have monthly growth rates after the first 6 months of life to fall within

0.15 $\pm 2\times$ 0.08 kg, which is: -0.02 to 0.31 kg.

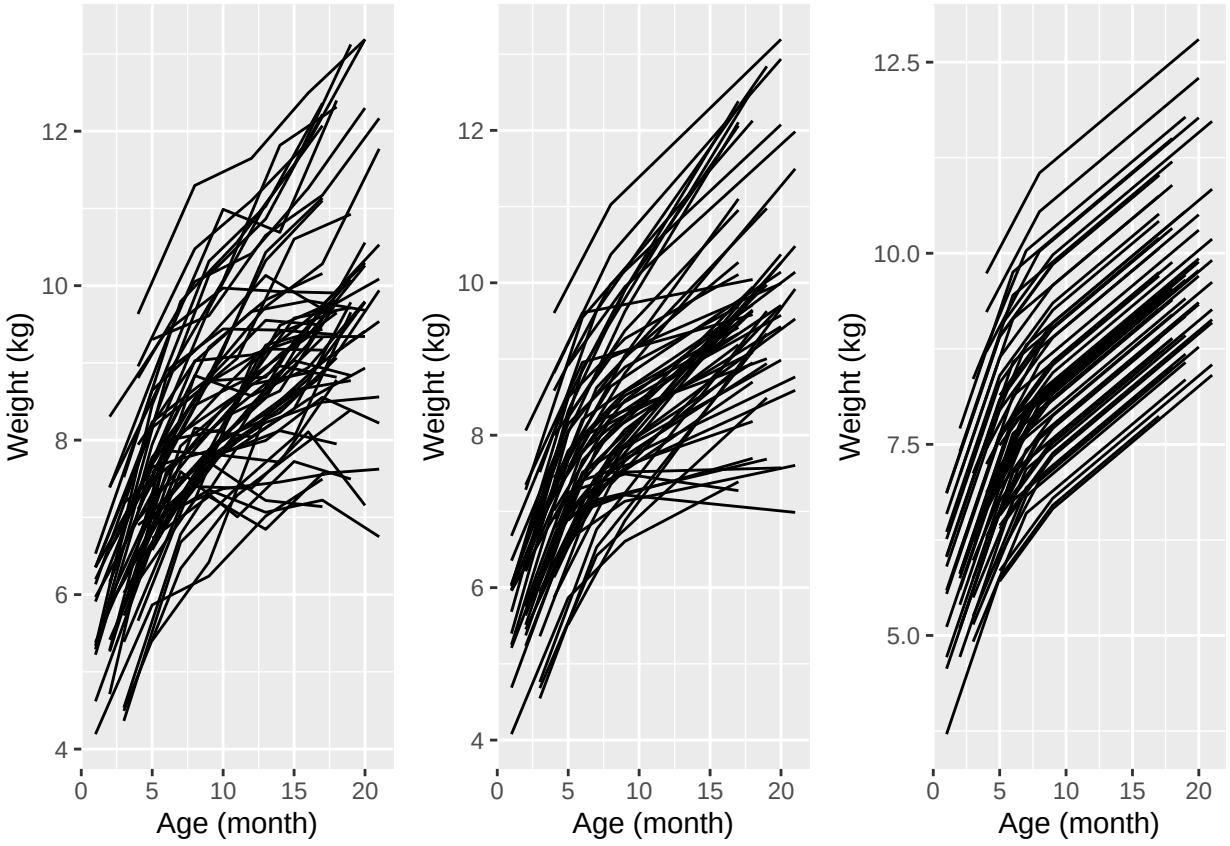
8. Compare the observed data to the predicted growth for children based on the random effects model.

```
nepal1$fitted = fitted(fit)
plot.data = ggplot(data = nepal1) +
  geom_line(aes(age,wt,group = id)) +
  xlab("Age (month)") +
  ylab("Weight (kg)") +
  theme(legend.position='bottom', legend.box='horizontal')

plot.slope = ggplot(data = nepal1) +
  geom_line(aes(age,fitted,group = id)) +
  xlab("Age (month)") +
  ylab("Weight (kg)") +
  theme(legend.position='bottom', legend.box='horizontal')

grid.arrange(plot.data, plot.slope, ncol=2)
```



9. What if we had fit a random intercept only model?

```
fit.int = lmer(wt~age+age_sp6+(1|id),data=nepal1)
nepal1$fit.int = fitted(fit.int)
plot.int = ggplot(data = nepal1) +
  geom_line(aes(age,fit.int,group = id)) +
  xlab("Age (month)") +
  ylab("Weight (kg)") +
  theme(legend.position='bottom', legend.box='horizontal')

grid.arrange(plot.data, plot.slope, plot.int, ncol=3)
```



10. Compare the fits from the *gls* models and random intercept and slope models, using AIC.

```
AIC(fit,fit.int)
```

```
##         df      AIC
## fit      7 526.8088
## fit.int  5 729.3473
```

NOTE: The data was generated under the random intercept and random slope for age model!