Lecture 13 Handout: Practice with linear mixed models

Elizabeth Colantuoni

3/8/2021

I. Objectives

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

- 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

In this lecture, we will quickly review the analysis of "NEPAL1" dataset and walk more slowly through a second analysis where the goal of each analysis is to:

- Describe / estimate the monthly increase in weight as a function of age
- Quantify variation in child specific growth patterns, i.e. variation in birthweights or variation in growth rates

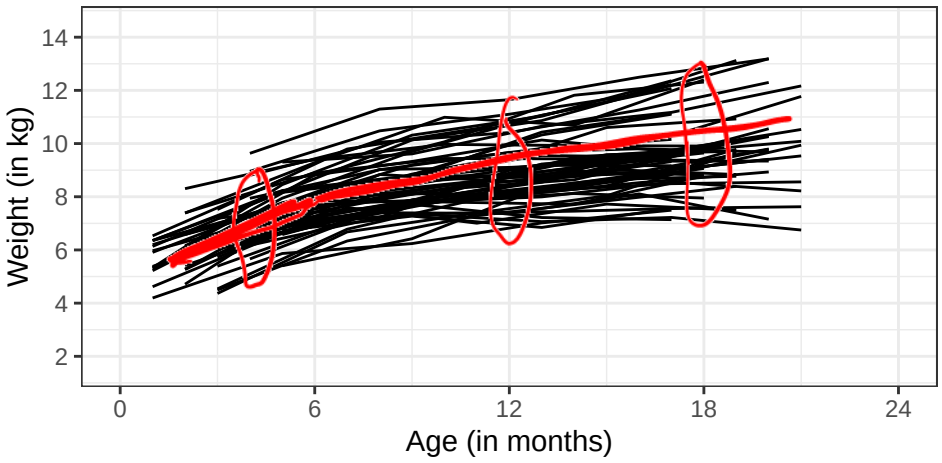
II. Analysis of NEPAL1

A. Summary of Exploratory Analysis

From the spaghetti plot below, we found:

- On average, growth in weight is more steep at younger ages compared to older ages
- There is variation in individual children's growth rates as noted by fanning out of and crossover of individual child trajectories
    - The variation in growth rates is linked to the increasing variation in weights as children age

```
load("nepal_simulated.rda")
ggplot(data = nepal1, aes(x = age, y = wt, group = factor(id))) +
  geom_line() + theme_bw() +
  labs(y="Weight (in kg)",x="Age (in months)") +
  scale_y_continuous(breaks=seq(2,14,2),limits=c(1.5,14.5)) +
  scale_x_continuous(breaks=seq(0,24,6),limits=c(0,24))
```



Next, we explored the correlation structure in the data by computing $Corr(r_{ij}, r_{ik})$ where r are residuals from a linear spline model assuming a single knot at 6 months of age.

Based on the correlation matrix below, we noted stronger correlation between residuals that were measured closer in time compared to father apart in time.

```
## Here you need to get the set of residuals and then look at the correlation between residuals at the
nepal1$residuals = residuals(lm(wt~age+age_sp6,data=nepal1))
nepal1_wide = nepal1 %>% select(id, fuvisit, residuals) %>% spread(fuvisit,residuals)
cor(nepal1_wide[,-1])
```

```
##           0         1         2         3         4
## 0 1.0000000 0.8785910 0.7595446 0.6685220 0.4958610
## 1 0.8785910 1.0000000 0.8814021 0.8246947 0.7181845
## 2 0.7595446 0.8814021 1.0000000 0.9350597 0.8890514
## 3 0.6685220 0.8246947 0.9350597 1.0000000 0.9389962
## 4 0.4958610 0.7181845 0.8890514 0.9389962 1.0000000
```

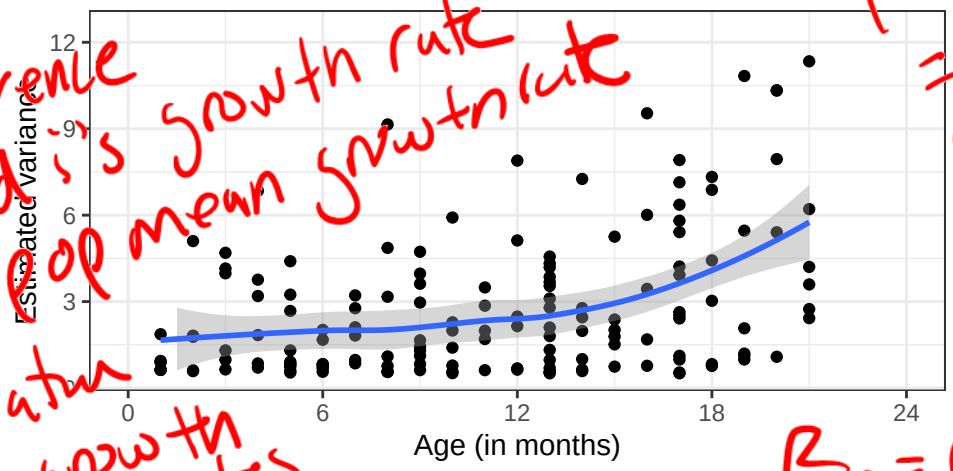
Lastly, we explored the variance in the residuals and found that variance increases as a function of age; consistent with our observation from the spaghetti plot.

```
ggplot(nepal1,aes(x=age,y=residuals^2)) +
    geom_point() + geom_smooth() + theme_bw() +
    labs(y="Estimated variance",x="Age (in months)") +
    scale_y_continuous(breaks=seq(0,12,3),limits=c(0.5,12.5)) +
  scale_x_continuous(breaks=seq(0,24,6),limits=c(0,24))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 146 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 146 rows containing missing values (geom_point).
```



B. Fit and interpretation of linear mixed model

We 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 $b_i \sim MVN(0, D_{2\times 2})$.

$$D = \begin{bmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{bmatrix}$$

```
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
```

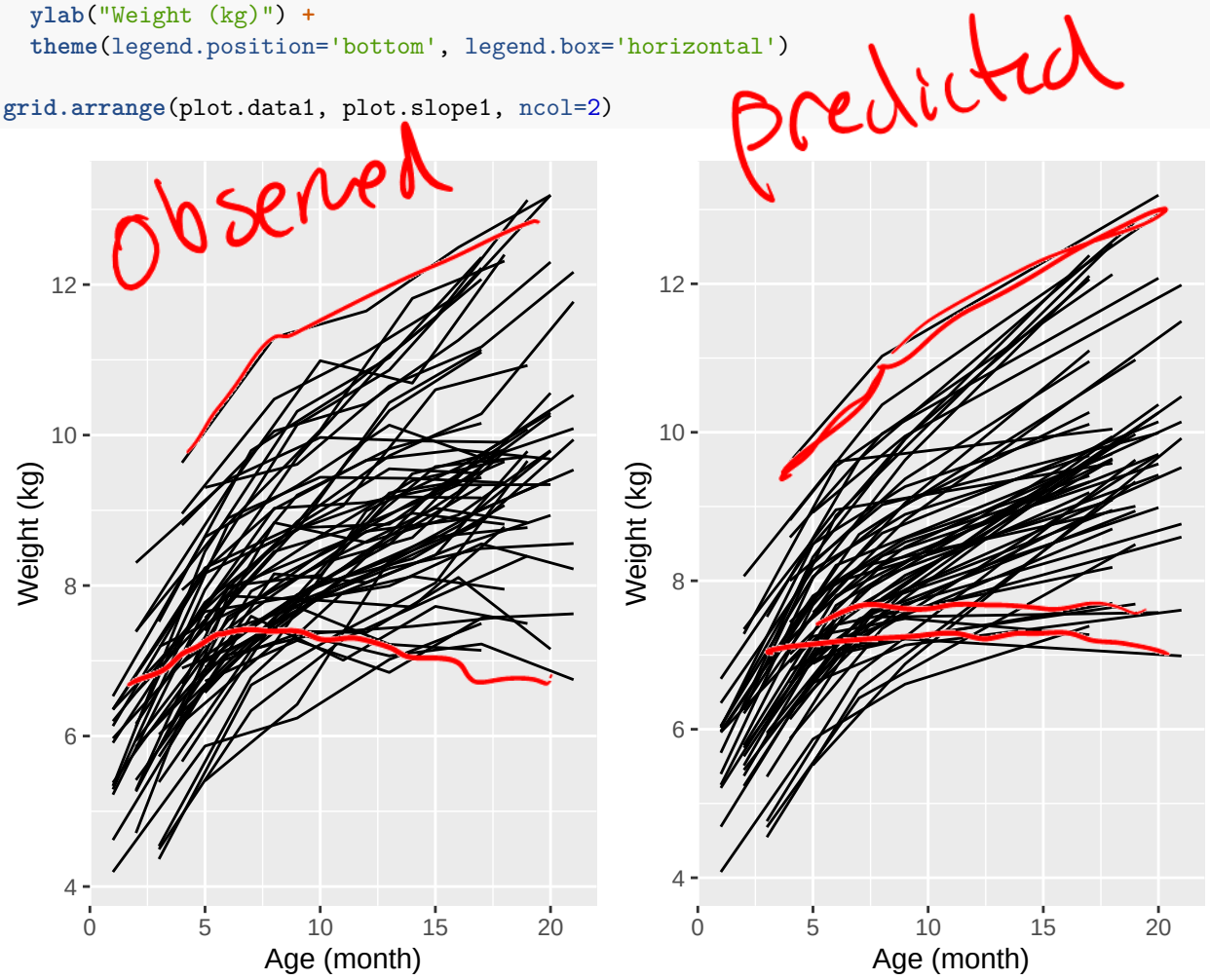
```
est = fixef(fit)
```

1. What does β_0 represent? What does b_{0i} represent? What does τ_0^2 represent?
2. What does β_1 represent? What does b_{1i} represent? What does τ_1^2 represent?
3. What does $\beta_1 + \beta_2$ represent? How does the model define the child specific growth rate when children are over 6 months of age?
4. What does σ^2 represent?
5. We compared the observed data to the predicted growth for children based on the random effects model. How well do you think the model fits?

```
nepal1$fitted = fitted(fit)
plot.data1 = ggplot(data = nepal1) +
  geom_line(aes(age,wt,group = id)) +
  xlab("Age (month)") +
  ylab("Weight (kg)") +
  theme(legend.position='bottom', legend.box='horizontal')

plot.slope1 = 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.data1, plot.slope1, ncol=2)
```



II. Example 2

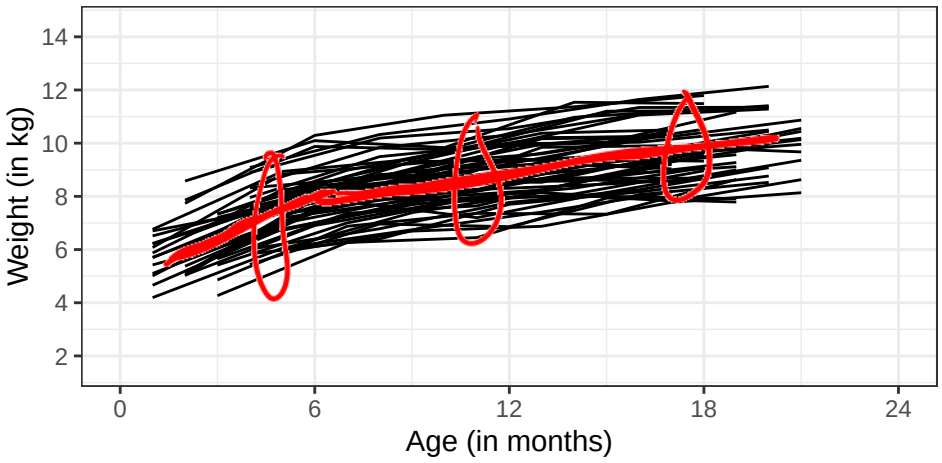
Now, you conduct a similar analysis for NEPAL2

A. Exploratory analysis

1. Exploration of the mean model

Below you will find a spaghetti plot of the NEPAL2 data. What do you notice about the data? Can you describe some patterns you observe?

```
ggplot(data = nepal2, aes(x = age, y = wt, group = factor(id))) +
  geom_line() + theme_bw() +
  labs(y="Weight (in kg)",x="Age (in months)") +
  scale_y_continuous(breaks=seq(2,14,2),limits=c(1.5,14.5)) +
  scale_x_continuous(breaks=seq(0,24,6),limits=c(0,24))
```

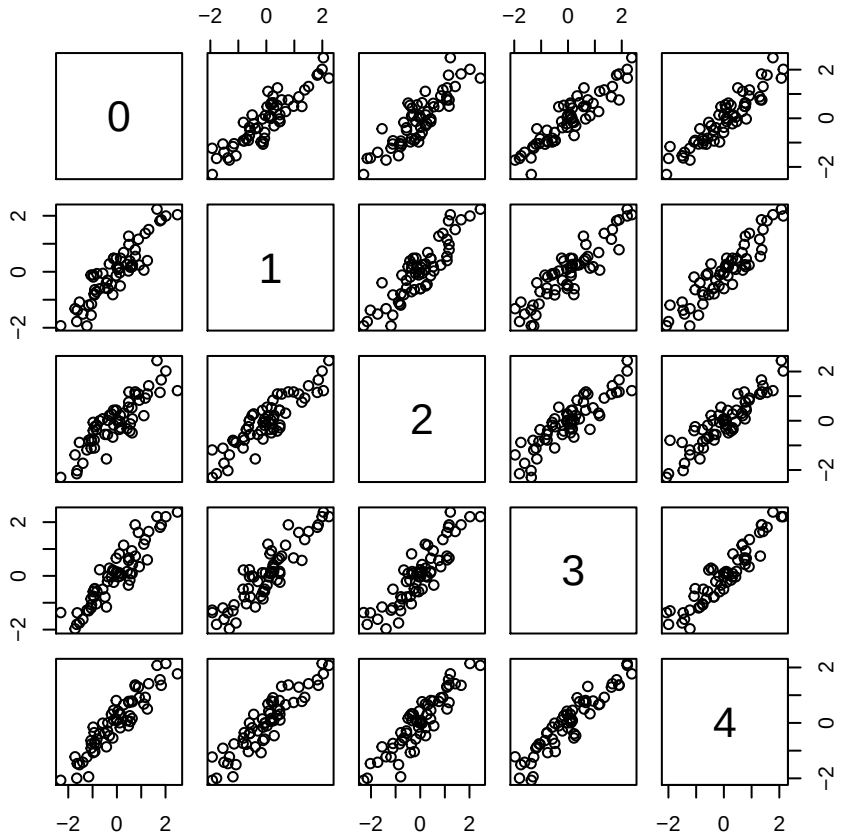


Constant variance

2. Explore the correlation structure

Next, explore the correlation structure in the data by computing $Corr(r_{ij}, r_{ik})$ where r are residuals from a linear spline model assuming a single knot at 6 months of age.

```
## Here you need to get the set of residuals and then look at the correlation between residuals at the
nepal2$residuals = residuals(lm(wt~age+age_sp6,data=nepal2))
nepal2_wide = nepal2 %>% select(id, fuvisit, residuals) %>% spread(fuvisit,residuals)
pairs(nepal2_wide[,-1])
```




```
cor(nepal2_wide[,-1])
```

```
##           0         1         2         3         4
## 0 1.0000000 0.9020224 0.8690523 0.9210077 0.9273866
## 1 0.9020224 1.0000000 0.8928467 0.9026381 0.9190152
## 2 0.8690523 0.8928467 1.0000000 0.8884843 0.9119631
## 3 0.9210077 0.9026381 0.8884843 1.0000000 0.9356709
## 4 0.9273866 0.9190152 0.9119631 0.9356709 1.0000000
```

.9

Can you look at the table of correlation estimates and provide a rough estimate for the autocorrelation function?

exchangeable

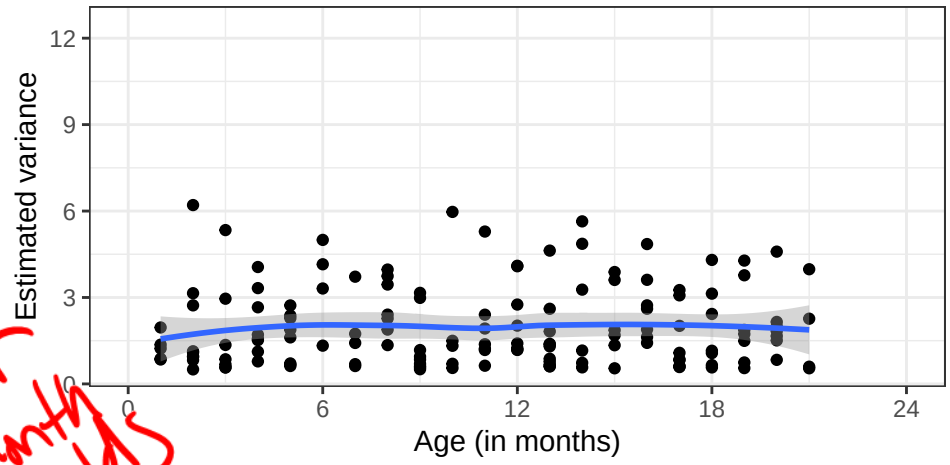
3. Explore the variance structure

In addition to understanding the correlation structure, we need understand if the variance in the residuals is the same at all ages or the variance of the residuals changes over age.

```
ggplot(nepal2,aes(x=age,y=residuals^2)) +
    geom_point() + geom_smooth() + theme_bw() +
    labs(y="Estimated variance",x="Age (in months)") +
    scale_y_continuous(breaks=seq(0,12,3),limits=c(0.5,12.5)) +
  scale_x_continuous(breaks=seq(0,24,6),limits=c(0,24))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 154 rows containing non-finite values (stat_smooth).
## Warning: Removed 154 rows containing missing values (geom_point).
```



4. Summary of the exploratory analysis

From the exploratory analysis, can you specify a linear mixed model that you think is consistent with the data?

B. Fit and interpretation of linear mixed model

Fit a random intercept model to the NEPAL2 dataset:

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

where e_{ij} iid $N(0, \sigma^2)$ and $b_i \sim N(0, \tau_0^2)$.

```
fit = lmer(wt~age+age_sp6+(1|id),data=nepal2)
summary(fit)$coefficients
```

```
##               Estimate Std. Error   t value
## (Intercept)  5.0112341 0.14639985  34.22978
## age          0.5064909 0.01549233  32.69301
## age_sp6     -0.3653625 0.01810495 -20.18025
```

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

```
##  Groups   Name        Std.Dev.
##  id       (Intercept) 0.97226
##  Residual             0.31244
```

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

1. What does β_0 represent? What does b_{0i} represent? What does τ_0^2 represent?
2. What does β_1 represent? Does our model allow the individual child growth rates to vary?
3. What does $\beta_0 + 3\beta_1$ represent?
4. Can you provide an interval that contains 95% of weights for 3-month old children?

5. What does $\beta_0 + 12\beta_1 + (12-6)\beta_2$ represent?

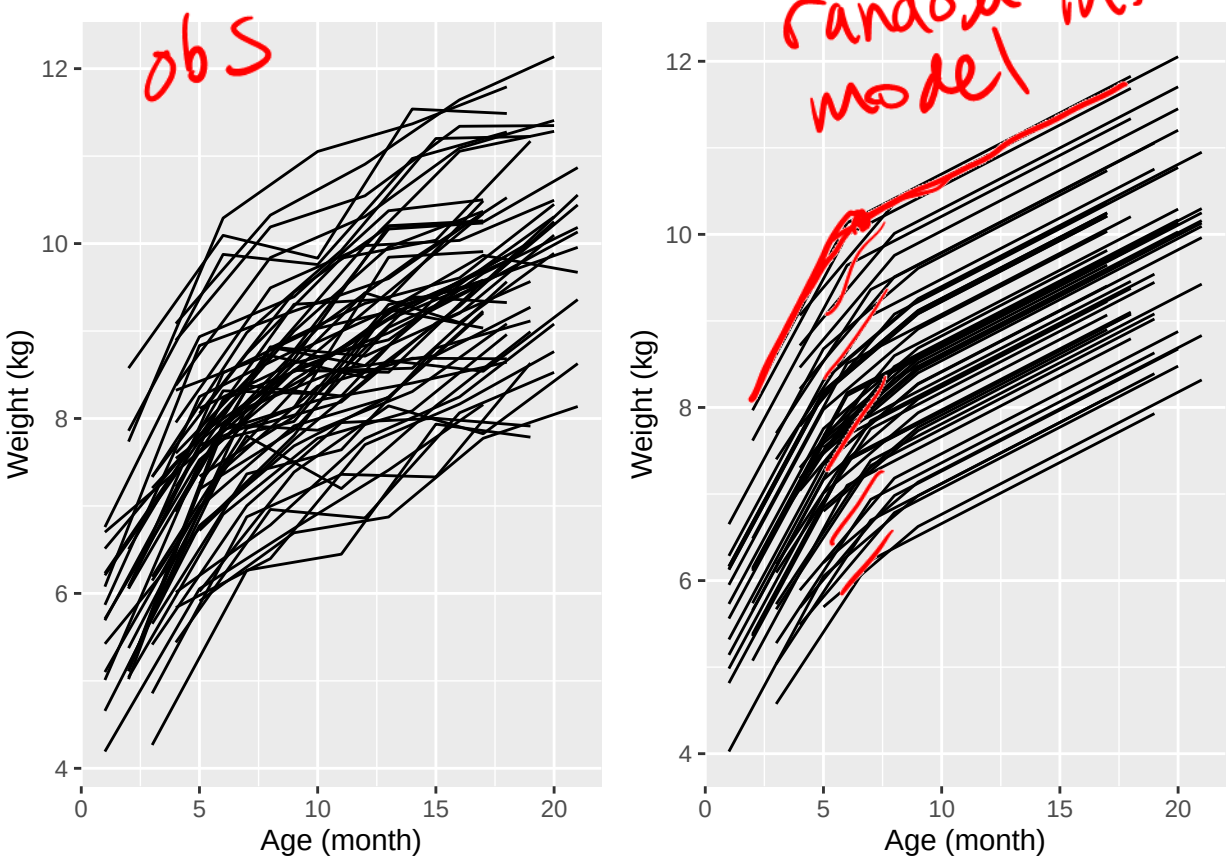
6. Can you provide an interval that contains 95% of weights for 12-month old children?

7. We compared the observed data to the predicted growth for children based on the random effects model. How well do you think the model fits?

```
nepal2$fitted = fitted(fit)
plot.data2 = ggplot(data = nepal2) +
  geom_line(aes(age,wt,group = id)) +
  xlab("Age (month)") +
  ylab("Weight (kg)") +
  theme(legend.position='bottom', legend.box='horizontal')

plot.int2 = ggplot(data = nepal2) +
  geom_line(aes(age,fitted,group = id)) +
  xlab("Age (month)") +
  ylab("Weight (kg)") +
  theme(legend.position='bottom', legend.box='horizontal')

grid.arrange(plot.data2, plot.int2, ncol=2)
```



8. Try to add a random slope for age to your linear mixed model. Does this work?

```
fit.slope = lmer(wt~age+age_sp6+(1+age|id),data=nepal2)

## boundary (singular) fit: see ?isSingular
```

```
isSingular(fit.slope)
```

```
## [1] TRUE
```

```
summary(fit.slope)
```

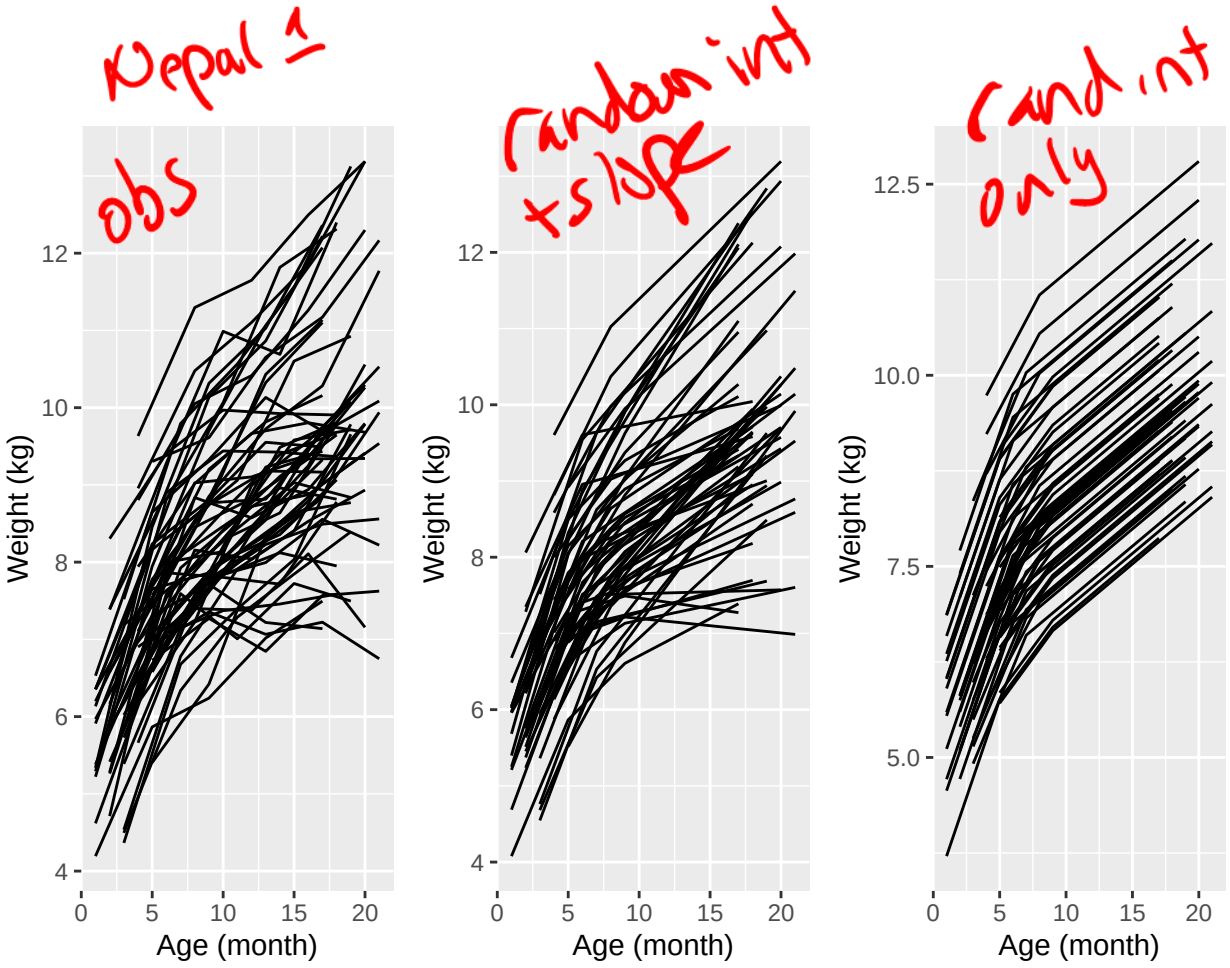
```
## Linear mixed model fit by REML ['lmerMod']
## Formula: wt ~ age + age_sp6 + (1 + age | id)
##    Data: nepal2
##
## REML criterion at convergence: 402.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3144 -0.6368 -0.0177  0.6124  2.2396
##
## Random effects:
##  Groups   Name        Variance  Std.Dev. Corr
##  id       (Intercept) 9.119e-01 0.95496
##           age         2.433e-06 0.00156  1.00
##  Residual             9.754e-02 0.31231
## Number of obs: 300, groups:  id, 60
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  5.01292    0.14444   34.71
## age          0.50614    0.01548   32.70
## age_sp6     -0.36494    0.01809  -20.18
##
## Correlation of Fixed Effects:
##         (Intr) age
## age     -0.474
## age_sp6  0.446 -0.979
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

III. Comparison of analyses

For NEPAL1, we noted variation in growth rates across children, so we allowed our linear mixed model to include a random intercept plus a random slope for age.

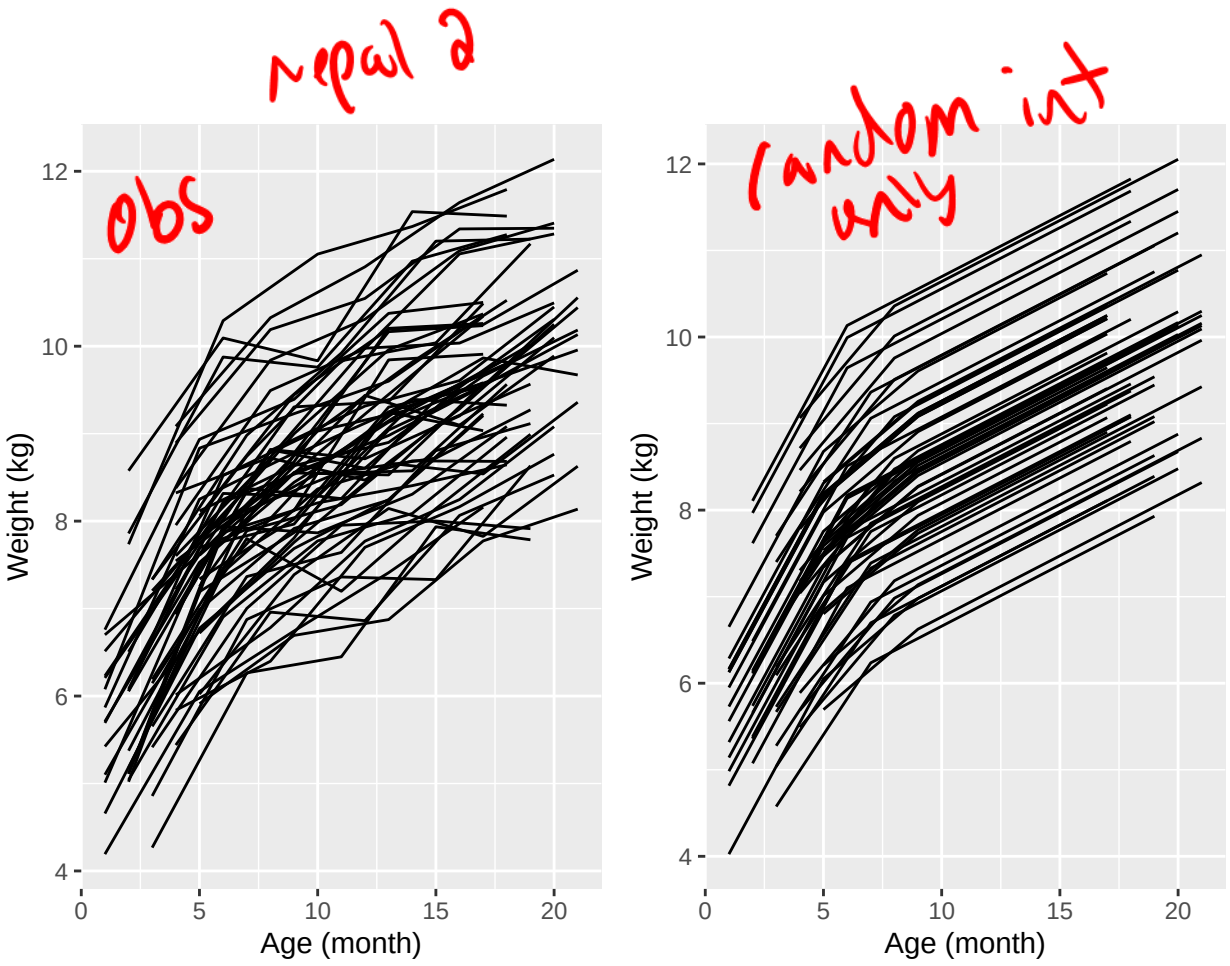
```
fit.int = lmer(wt~age+age_sp6+(1|id),data=nepal1)
nepal1$fit.int = fitted(fit.int)
plot.int1 = ggplot(data = nepal1) +
  geom_line(aes(age,fit.int,group = id)) +
  xlab("Age (month)") +
  ylab("Weight (kg)") +
  theme(legend.position='bottom', legend.box='horizontal')

grid.arrange(plot.data1, plot.slope1, plot.int1, ncol=3)
```



For Nepal2, we noted that the child's growth rates were similar! We were not able to estimate the random slope for age variance (essentially estimated to be 0)!

```
grid.arrange(plot.data2,plot.int2,ncol=2)
```


obs
random int only
Weight (kg)
Age (month)
12
10
8
6
4
0
5
10
15
20