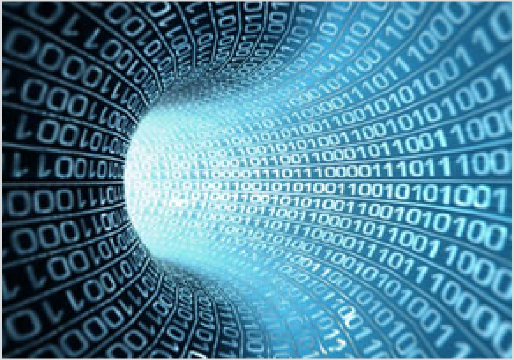
# StatLab Workshop

## Mixed Effects Modeling with R

W. D. Brinda

Yale University

05/19/2020



# Outline

# Outline

1. Theory
2. Examples
   - Repeated Measurements
   - Clustered data from random lines
   - Clustering and repeated measurements
   - Longitudinal growth

# Theory

# Linear model (with only *fixed* effects)

The general formulation of the multiple linear model with iid Normal errors:

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_d x_i^{(d)} + \epsilon_i$$

with $\epsilon_1, \ldots, \epsilon_n$ iid Normal.

# Linear model (with only *fixed* effects)

The general formulation of the multiple linear model with iid Normal errors:

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_d x_i^{(d)} + \epsilon_i$$

with $\epsilon_1, \ldots, \epsilon_n$ iid Normal.

The coefficients are considered *fixed*; that is, the model doesn't include any assumption that they're drawn from some particular distribution.

The general formulation of the multiple linear model with iid Normal errors:

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_d x_i^{(d)} + \epsilon_i$$

with $\epsilon_1, \ldots, \epsilon_n$ iid Normal.

The coefficients are considered *fixed*; that is, the model doesn't include any assumption that they're drawn from some particular distribution.

The least-squares procedure finds the coefficient values that minimize the sum of squared residuals. (This is also the maximum likelihood estimator.)

# Linear model (with only *fixed* effects)

The general formulation of the multiple linear model with iid Normal errors:

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_d x_i^{(d)} + \epsilon_i$$

with $\epsilon_1, \ldots, \epsilon_n$ iid Normal.

The coefficients are considered *fixed*; that is, the model doesn't include any assumption that they're drawn from some particular distribution.

The least-squares procedure finds the coefficient values that minimize the sum of squared residuals. (This is also the maximum likelihood estimator.)

Inference tasks based on the least-squared estimators are straightforward.

# Mixed linear model (with both *fixed* and *random* effects)

The general formulation of the *mixed* linear model with iid Normal errors:

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_d x_i^{(d)} + b_1 z_i^{(1)} + \ldots + b_m z_i^{(m)} + \epsilon_i$$

with $\epsilon_1, \ldots, \epsilon_n$ iid Normal and the *random effects* $(b_1, \ldots, b_m)$ are multivariate Normal. (Note $x_1, \ldots, x_d$ and $z_1, \ldots, z_m$ are all explanatory variables.)

The general formulation of the *mixed* linear model with iid Normal errors:

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_d x_i^{(d)} + b_1 z_i^{(1)} + \ldots + b_m z_i^{(m)} + \epsilon_i$$

with $\epsilon_1, \ldots, \epsilon_n$ iid Normal and the *random effects* $(b_1, \ldots, b_m)$ are multivariate Normal. (Note $x_1, \ldots, x_d$ and $z_1, \ldots, z_m$ are all explanatory variables.)

Maximum likelihood estimation (of the fixed effects) is much more complicated in this case. It doesn't have a closed-form solution in general but iterative algorithms such as EM are used.

# Mixed linear model (with both *fixed* and *random* effects)

The general formulation of the *mixed* linear model with iid Normal errors:

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_d x_i^{(d)} + b_1 z_i^{(1)} + \ldots + b_m z_i^{(m)} + \epsilon_i$$

with $\epsilon_1, \ldots, \epsilon_n$ iid Normal and the *random effects* $(b_1, \ldots, b_m)$ are multivariate Normal. (Note $x_1, \ldots, x_d$ and $z_1, \ldots, z_m$ are all explanatory variables.)

Maximum likelihood estimation (of the fixed effects) is much more complicated in this case. It doesn't have a closed-form solution in general but iterative algorithms such as EM are used.

Inference tasks are also much more complicated, but approximations (asymptotic distributions) are used.

The general formulation of the *mixed* linear model with iid Normal errors:

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_d x_i^{(d)} + b_1 z_i^{(1)} + \ldots + b_m z_i^{(m)} + \epsilon_i$$

with $\epsilon_1, \ldots, \epsilon_n$ iid Normal and the *random effects* $(b_1, \ldots, b_m)$ are multivariate Normal. (Note $x_1, \ldots, x_d$ and $z_1, \ldots, z_m$ are all explanatory variables.)

Maximum likelihood estimation (of the fixed effects) is much more complicated in this case. It doesn't have a closed-form solution in general but iterative algorithms such as EM are used.

Inference tasks are also much more complicated, but approximations (asymptotic distributions) are used.

The errors don't actually have to be independent of each other. Our final example will demonstrate this.

# Examples

Suppose that $n_f$ females and $n_m$ males each had the blood pressures measured. Assume that the distributions of female and male blood pressures are both Normal with the same standard deviation $\sigma$. Let their expectations be denoted $\mu_f$ and $\mu_m$ respectively. With $x_1, \ldots, x_n$ representing sex (the only explanatory variable) and $Y_1, \ldots, Y_N$ representing blood pressures, we can express this scenario with the linear model

$$Y_i = \mu_f + (\mu_m - \mu_f)\mathbb{I}(x_i = \text{"}M\text{"}) + \epsilon_i.$$

Suppose that $n_f$ females and $n_m$ males each had the blood pressures measured. Assume that the distributions of female and male blood pressures are both Normal with the same standard deviation $\sigma$. Let their expectations be denoted $\mu_f$ and $\mu_m$ respectively. With $x_1, \ldots, x_n$ representing sex (the only explanatory variable) and $Y_1, \ldots, Y_N$ representing blood pressures, we can express this scenario with the linear model

$$Y_i = \mu_f + (\mu_m - \mu_f)\mathbb{I}(x_i = \text{"}M\text{"}) + \epsilon_i.$$

Inference for the means fits into the usual linear modeling

framework. (In this case, it simplifies to an ordinary two-sample $t$-test.)

# Repeated blood pressure measurements

Suppose alternatively that each of the subjects’ blood pressures were measured up to four times, at random moments over the course of a week. The measurements of a single person’s blood pressure can vary quite a bit when measured at different times. How could we adapt our model to capture this?

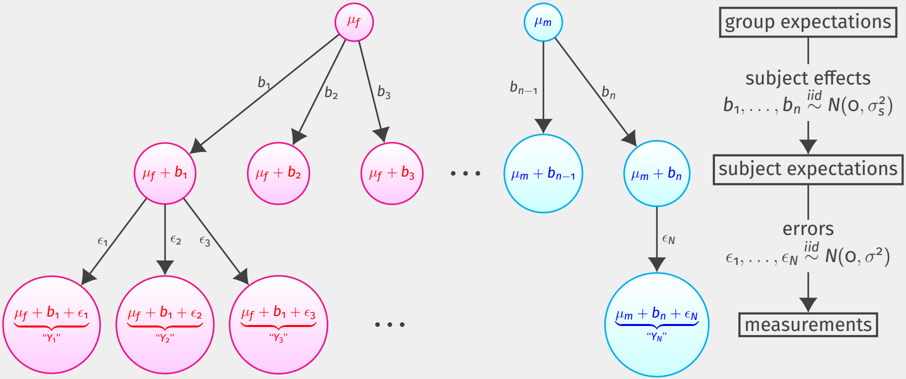
# Repeated blood pressure measurements

Suppose alternatively that each of the subjects' blood pressures were measured up to four times, at random moments over the course of a week. The measurements of a single person's blood pressure can vary quite a bit when measured at different times. How could we adapt our model to capture this?

We might assume that each subject has an *expected* blood pressure and model the specific measurements as iid Normal draws centered at that expectation. Assume further that each subject's distribution of blood pressure measurements shares the same standard deviation $\sigma$. Let $b_i$ denote the deviation of subject $i$ from his or her group's mean $\mu_{x_i}$. Assume that these deviations are iid Normal with mean zero and standard deviation $\sigma_S$.

Let $n := n_f + n_m$ be the total number of subjects. Our assumptions fit into the mixed effects modeling framework if you define $z_i^{(1)}, \ldots, z_i^{(m)}$ to be the indicator variables that tell us which subject observation $j$ came from. Specifically, $b_i^{(j)}$ is 1 if measurement $i$ is from subject $j$ and it's 0 otherwise. To see how the model equation simplifies, suppose that observation $i$ is from subject $j$ who is a male. Then

$$\begin{aligned} Y_i &= \mu_f + (\mu_m - \mu_f)\mathbb{I}(x_i = \text{“}M\text{”}) + b_1 z_i^{(1)} + \ldots + b_m z_i^{(m)} + \epsilon_i \\ &= \mu_m + b_j + \epsilon_i. \end{aligned}$$


$\mu_f$
$b_1$
$b_2$
$b_3$
$\mu_f + b_1$
$\mu_f + b_2$
$\mu_f + b_3$
$\epsilon_1$
$\epsilon_2$
$\epsilon_3$
$\underbrace{\mu_f + b_1 + \epsilon_1}_{\text{"}Y_1\text{"}}$
$\underbrace{\mu_f + b_1 + \epsilon_2}_{\text{"}Y_2\text{"}}$
$\underbrace{\mu_f + b_1 + \epsilon_3}_{\text{"}Y_3\text{"}}$
$\mu_m$
$b_{n-1}$
$b_n$
$\mu_m + b_{n-1}$
$\mu_m + b_n$
$\epsilon_N$
$\underbrace{\mu_m + b_n + \epsilon_N}_{\text{"}Y_N\text{"}}$
group expectations
subject effects
$b_1, \ldots, b_n \overset{iid}{\sim} N(0, \sigma_s^2)$
subject expectations
errors
$\epsilon_1, \ldots, \epsilon_N \overset{iid}{\sim} N(0, \sigma^2)$
measurements

# R script Example 1

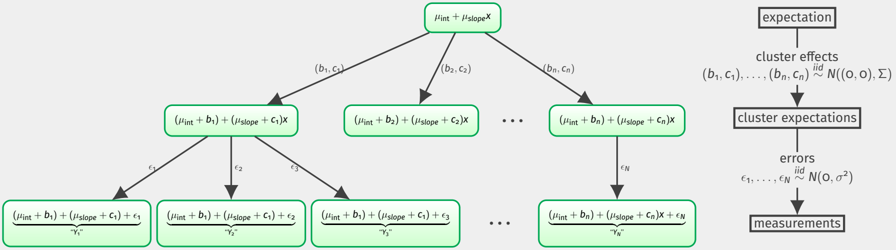
Now, let's turn to the first portion of *mixed-effects.R*, where we'll simulate blood pressure data according to the repeated measurements mechanism just described. Then we'll use the *lme4* package to fit the corresponding mixed effects model, and compare the true values to the resulting estimates. Finally, we'll use the *lmertest* package for inference and model selection.

Suppose that, at any given school, there's a line relating parents' log income (explanatory variable) and expected SAT score (response variable). The students' actual SAT scores deviate from the line by Normal errors with a common variance $\sigma^2$. Suppose, however, that each school has its own line, that is, its own intercept and slope. Let's assume that the schools' lines are iid draws from a bivariate Normal distribution centered at $(\mu_{\text{int}}, \mu_{\text{slope}})$.

This scenario fits into the linear mixed effects modeling framework described above. Let $Y_i$ and $x_i$ represent the $i$th student's SAT score and parents' log income. Let $z_i^{(j)}$ be the indicator that student $i$ attends school $j$. Letting student $i$ attend school $k$ (i.e. $z_i^{(k)} = 1$ and the others are 0), and letting $m$ denote the number of schools in the dataset,

$$\begin{aligned} Y_i &= \mu_{\text{int}} + \mu_{\text{slope}} x_i + b_1 z_i^{(1)} + \ldots + b_m z_i^{(m)} + c_1 z_i^{(1)} x_i + \ldots + c_m z_i^{(m)} x_i + \epsilon_i \\ &= \underbrace{(\mu_{\text{int}} + b_k)}_{\text{school } k\text{'s intercept}} + \underbrace{(\mu_{\text{slope}} + c_k)}_{\text{school } k\text{'s slope}} x_i + \epsilon_i \end{aligned}$$

where $\epsilon_1, \ldots, \epsilon_n$ are iid $N(0, \sigma^2)$. The Normal deviations $(b_j, c_j)$ of the $j$th school's intercept and slope from the expectations shouldn't be assumed to be independent of each other.


$\mu_{\text{int}} + \mu_{slope}x$
$(b_1, c_1)$
$(b_2, c_2)$
$(b_n, c_n)$
$(\mu_{\text{int}} + b_1) + (\mu_{slope} + c_1)x$
$(\mu_{\text{int}} + b_2) + (\mu_{slope} + c_2)x$
$\cdots$
$(\mu_{\text{int}} + b_n) + (\mu_{slope} + c_n)x$
$\epsilon_1$
$\epsilon_2$
$\epsilon_3$
$\epsilon_N$
$\underbrace{(\mu_{\text{int}} + b_1) + (\mu_{slope} + c_1) + \epsilon_1}_{\text{“}Y_1\text{”}}$
$\underbrace{(\mu_{\text{int}} + b_1) + (\mu_{slope} + c_1) + \epsilon_2}_{\text{“}Y_2\text{”}}$
$\underbrace{(\mu_{\text{int}} + b_1) + (\mu_{slope} + c_1) + \epsilon_3}_{\text{“}Y_3\text{”}}$
$\cdots$
$\underbrace{(\mu_{\text{int}} + b_n) + (\mu_{slope} + c_n)x + \epsilon_N}_{\text{“}Y_N\text{”}}$
expectation
cluster effects
$(b_1, c_1), \ldots, (b_n, c_n) \overset{iid}{\sim} N((0,0), \Sigma)$
cluster expectations
errors
$\epsilon_1, \ldots, \epsilon_N \overset{iid}{\sim} N(0, \sigma^2)$
measurements

# R script Example 2

Now, let's turn to the second portion of *mixed-effects.R*, where we'll simulate SAT scores according to the random lines mechanism just described. Then we'll again fit the corresponding mixed effects model, compare the true values to the resulting estimates, and engage in inference and model selection.

# Repeated SAT scores by sex at various schools

Suppose that the average female SAT score is $\mu_f = 1055$ and the average male SAT score is $\mu_m = 1042$.

# Repeated SAT scores by sex at various schools

Suppose that the average female SAT score is $\mu_f = 1055$ and the average male SAT score is $\mu_m = 1042$.

Suppose that each school has its own effect on SAT score and that the school's expected males' scores is the same as the effect on female scores. In other words, if a school's effect is $b_j$, then it's expected female score is $\mu_f + b_j$ and its expected male score is $\mu_m + b_j$. Assume that the effects $b_1, b_2, \ldots$ are iid $N(0, \sigma^2_{\text{school}})$.

Suppose that the average female SAT score is $\mu_f = 1055$ and the average male SAT score is $\mu_m = 1042$.

Suppose that each school has its own effect on SAT score and that the school's expected males' scores is the same as the effect on female scores. In other words, if a school's effect is $b_j$, then it's expected female score is $\mu_f + b_j$ and its expected male score is $\mu_m + b_j$. Assume that the effects $b_1, b_2, \ldots$ are iid $N(0, \sigma^2_{\text{school}})$.

There are also student effects $c_1, c_2, \ldots$. Suppose that student $k$ is a male who attends school $j$. His expected SAT score is $\mu_m + b_j + c_k$. Assume the student effects are iid $N(0, \sigma^2_{\text{student}})$.

Suppose that the average female SAT score is $\mu_f = 1055$ and the average male SAT score is $\mu_m = 1042$.

Suppose that each school has its own effect on SAT score and that the school's expected males' scores is the same as the effect on female scores. In other words, if a school's effect is $b_j$, then it's expected female score is $\mu_f + b_j$ and its expected male score is $\mu_m + b_j$. Assume that the effects $b_1, b_2, \ldots$ are iid $N(0, \sigma^2_{\text{school}})$.

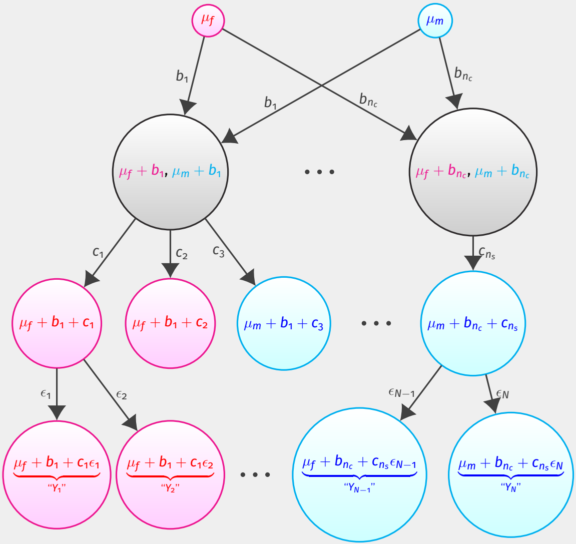
There are also student effects $c_1, c_2, \ldots$. Suppose that student $k$ is a male who attends school $j$. His expected SAT score is $\mu_m + b_j + c_k$. Assume the student effects are iid $N(0, \sigma^2_{\text{student}})$. Each student takes the SAT up to 3 times, and their observed

scores differ from the expectation by a Normal error. Assume the errors are iid $N(0, \sigma^2)$.

We can represent this scenario as a mixed effects model with $Y_i$ representing the $i$th SAT score, $z_i^{(1)}, z_i^{(2)}, \ldots$ indicators representing the student tested, $w_i^{(1)}, w_i^{(2)}, \ldots$ representing that student's school, and $x_i$ representing that student's sex. If $Y_i$ is a test score of student $k$ who is a male at school $j$,

$$\begin{aligned} Y_i &= \mu_f + (\mu_m - \mu_f)\mathbb{I}(x_i = \text{“}M\text{”}) + b_1 w_i^{(1)} + b_2 w_i^{(2)} + \ldots + c_1 z_i^{(1)} + c_2 z_i^{(2)} + \\ &= \mu_m + b_j + c_k + \epsilon_i. \end{aligned}$$

# Diagram




expectations

cluster effects

$b_1, \ldots, b_n \overset{iid}{\sim} N(0, \sigma_c^2)$

cluster expectations

subject effects

$c_1, \ldots, c_n \overset{iid}{\sim} N(0, \sigma_s^2)$

subject expectations

errors

$\epsilon_1, \ldots, \epsilon_N \overset{iid}{\sim} N(0, \sigma^2)$

measurements

# R script Example 3

Now, we'll work through the third part of *mixed-effects.R*, where we'll simulate SAT scores according to the mechanism just described: first, generate schools, then generate students who attend those schools, then generate SAT test scores for those students. Then we'll again fit the corresponding mixed effects model, compare the true values to the resulting estimates, and engage in inference and model selection.

Pretend that three different fertilizers are being studied. For each fertilizer, a number of trees will be planted and their heights will be measured after 1 year, 2 years, 3 years, and 4 years. Assume that for each fertilizer, the expected amount growth (increased height) is the same every year for the first four years. A tree with fertilizer A is expected to grow $\mu_A = 2.7$ feet per year, with B it's expected to grow 3 feet per year, and with C it's expected to grow 3.6 feet per year.

# Measurements of trees over time with different fertilizers

Pretend that three different fertilizers are being studied. For each fertilizer, a number of trees will be planted and their heights will be measured after 1 year, 2 years, 3 years, and 4 years. Assume that for each fertilizer, the expected amount growth (increased height) is the same every year for the first four years. A tree with fertilizer A is expected to grow $\mu_A = 2.7$ feet per year, with B it's expected to grow 3 feet per year, and with C it's expected to grow 3.6 feet per year.

Any particular tree will have its own effect. In particular, suppose it's expected growth per year deviates from the treatment group's expectation by some Normal draw. Furthermore, during any particular year, this tree's growth will differ from its expected growth by a Normal draw. This is a special type of repeated measurments scenario in which the timing of the measurements is important; it's called "longitudinal data."

# Measurements of trees over time with different fertilizers

Pretend that three different fertilizers are being studied. For each fertilizer, a number of trees will be planted and their heights will be measured after 1 year, 2 years, 3 years, and 4 years. Assume that for each fertilizer, the expected amount growth (increased height) is the same every year for the first four years. A tree with fertilizer A is expected to grow $\mu_A = 2.7$ feet per year, with B it's expected to grow 3 feet per year, and with C it's expected to grow 3.6 feet per year.

Any particular tree will have its own effect. In particular, suppose it's expected growth per year deviates from the treatment group's expectation by some Normal draw. Furthermore, during any particular year, this tree's growth will differ from its expected growth by a Normal draw. This is a special type of repeated measurments scenario in which the timing of the measurements is important; it's called “longitudinal data."

Why does that matter? Each tree has a sequence of deviations from its expected line. If a tree grows more than it was expected to during the first year, then most likely it will still be taller than its expectation after the second year. With longitudinal data, we typically don't assume that an individual's repeated errors (deviations from the expectation) are independent of each other; dependence among neighboring errors is called *autocorrelation*.

We can represent this scenario as a mixed effects model; this time we'll use double subscripts to index the observations. Let $Y_{i,j}$ represent the $j$th measurement of tree $i$ (which is the height of tree $i$ after $j$ years), $z_i^{(1)}, z_i^{(2)}, \ldots$ indicators representing the tree tested, and $b_i$ representing the $i$th tree's deviation from the expected growth rate of its treatment group. Assume $b_1, b_2, \ldots$ are iid $N(0, \sigma_t^2)$. If, for example, tree $i$ got fertilizer B, then

$$\begin{aligned} Y_{i,j} &= \eta_A j + (\eta_B - \eta_A)\mathbb{I}(x_i = \text{``}B\text{''})j + (\mu_C - \mu_A)\mathbb{I}(x_i = \text{``}C\text{''})j \\ &\quad + b_1 z_i^{(1)} j + b_2 z_i^{(2)} j + \ldots + \epsilon_{i,j} \\ &= \underbrace{(\mu_B + b_i)}_{\text{slope of the tree } i} \; j + \epsilon_{i,j}. \end{aligned}$$

Recall that $\epsilon_{i,1}, \epsilon_{i,2}, \ldots$ aren't independent of each other. Rather each year there is an independent Normal deviation from the expected amount of growth, and $\epsilon_{i,j}$ is the sum of draw 1 through draw $j$.

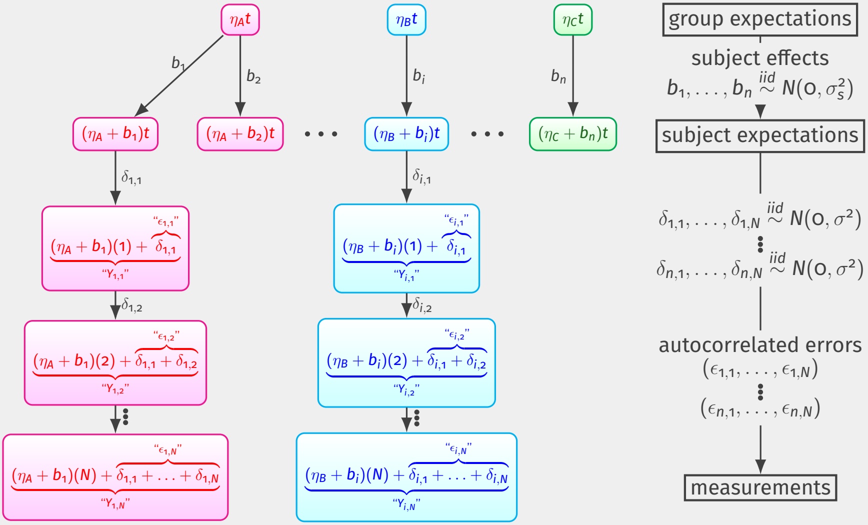
We can represent this scenario as a mixed effects model; this time we'll use double subscripts to index the observations. Let $Y_{i,j}$ represent the $j$th measurement of tree $i$ (which is the height of tree $i$ after $j$ years), $z_i^{(1)}, z_i^{(2)}, \ldots$ indicators representing the tree tested, and $b_i$ representing the $i$th tree's deviation from the expected growth rate of its treatment group. Assume $b_1, b_2, \ldots$ are iid $N(0, \sigma_t^2)$. If, for example, tree $i$ got fertilizer B, then

$$\begin{aligned} Y_{i,j} &= \eta_A j + (\eta_B - \eta_A)\mathbb{I}(x_i = \text{“}B\text{”})j + (\mu_C - \mu_A)\mathbb{I}(x_i = \text{“}C\text{”})j \\ &\quad + b_1 z_i^{(1)} j + b_2 z_i^{(2)} j + \ldots + \epsilon_{i,j} \\ &= \underbrace{(\mu_B + b_i)}_{\text{slope of the tree } i} \; j + \epsilon_{i,j}. \end{aligned}$$

Recall that $\epsilon_{i,1}, \epsilon_{i,2}, \ldots$ aren't independent of each other. Rather each year there is an independent Normal deviation from the expected amount of growth, and $\epsilon_{i,j}$ is the sum of draw 1 through draw $j$.

(We know that all trees had a height of zero at time zero, so I've built that into the model by not including an intercept term.)

# Diagram


$\eta_A t$
$\eta_B t$
$\eta_C t$
$b_1$
$b_2$
$b_i$
$b_n$
$(\eta_A + b_1)t$
$(\eta_A + b_2)t$
$\cdots$
$(\eta_B + b_i)t$
$\cdots$
$(\eta_C + b_n)t$
$\delta_{1,1}$
$\delta_{i,1}$
$\underbrace{(\eta_A + b_1)(1) + \overbrace{\delta_{1,1}}^{\text{“}\epsilon_{1,1}\text{”}}}_{\text{“}Y_{1,1}\text{”}}$
$\underbrace{(\eta_B + b_i)(1) + \overbrace{\delta_{i,1}}^{\text{“}\epsilon_{i,1}\text{”}}}_{\text{“}Y_{i,1}\text{”}}$
$\delta_{1,2}$
$\delta_{i,2}$
$\underbrace{(\eta_A + b_1)(2) + \overbrace{\delta_{1,1} + \delta_{1,2}}^{\text{“}\epsilon_{1,2}\text{”}}}_{\text{“}Y_{1,2}\text{”}}$
$\underbrace{(\eta_B + b_i)(2) + \overbrace{\delta_{i,1} + \delta_{i,2}}^{\text{“}\epsilon_{i,2}\text{”}}}_{\text{“}Y_{i,2}\text{”}}$
$\underbrace{(\eta_A + b_1)(N) + \overbrace{\delta_{1,1} + \ldots + \delta_{1,N}}^{\text{“}\epsilon_{1,N}\text{”}}}_{\text{“}Y_{1,N}\text{”}}$
$\underbrace{(\eta_B + b_i)(N) + \overbrace{\delta_{i,1} + \ldots + \delta_{i,N}}^{\text{“}\epsilon_{i,N}\text{”}}}_{\text{“}Y_{i,N}\text{”}}$
group expectations
subject effects
$b_1, \ldots, b_n \overset{iid}{\sim} N(0, \sigma_s^2)$
subject expectations
$\delta_{1,1}, \ldots, \delta_{1,N} \overset{iid}{\sim} N(0, \sigma^2)$
$\vdots$
$\delta_{n,1}, \ldots, \delta_{n,N} \overset{iid}{\sim} N(0, \sigma^2)$
autocorrelated errors
$(\epsilon_{1,1}, \ldots, \epsilon_{1,N})$
$\vdots$
$(\epsilon_{n,1}, \ldots, \epsilon_{n,N})$
measurements

Now, we'll work through the fourth and final part of *mixed-effects.R*, where we'll simulate SAT scores according to the mechanism just described: first, generate trees, then generate a sequence of heights for those trees. We'll again fit the corresponding mixed effects model, compare the true values to the resulting estimates, and engage in inference and model selection. Following this example, you'll find an exercise to attempt on your own. My solution is provided in *mixed-effects-solutions.R*.