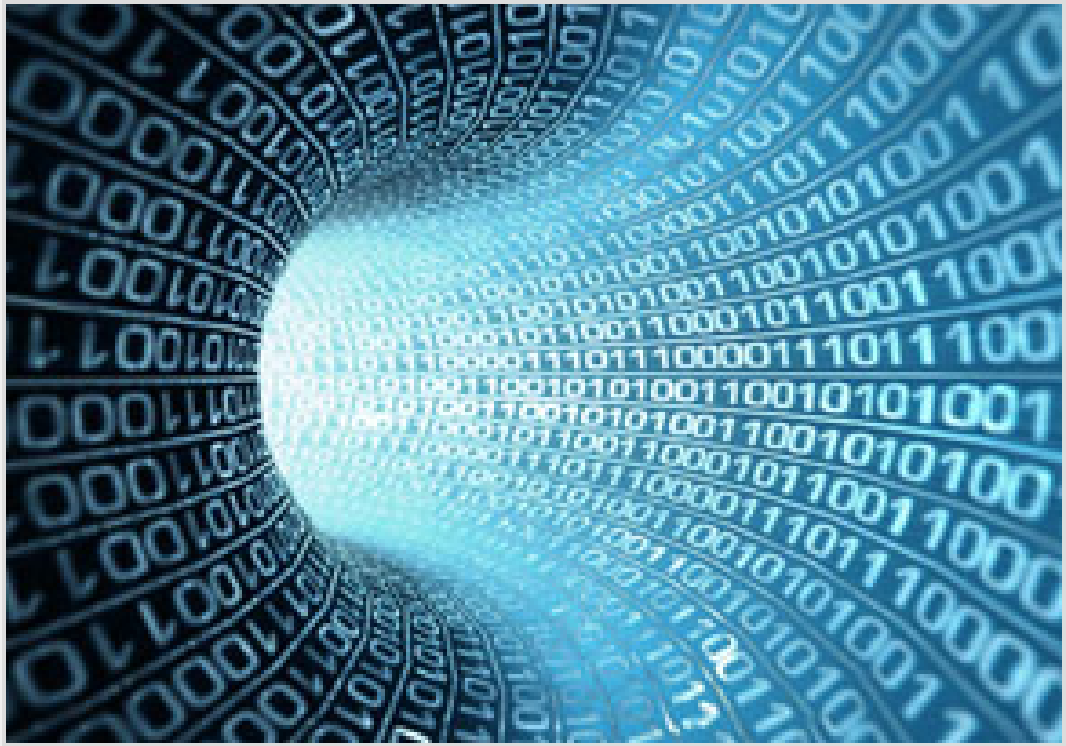
Generalized Additive Modeling and Mixed Effects

with Applications

W. D. Brinda

SageSure Insurance Managers LLC

6 July 2023



Outline

Outline

1. Generalized Additive Models
   - Concept
   - Simulated Examples

2. Mixed Effects Modeling
   - Concept
   - Simulated Examples
     - Repeated Measurements
     - Clustered data from random lines
     - Clustering and repeated measurements
     - Longitudinal growth

Generalized Additive Models

GLM and GAM

The *generalized linear model* framework generalizes two aspects of the linear model.

- The expectation of the response is related to the linear combination of parameters by some specified *link function*:

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

- The distribution of the response variable doesn't have to be Normal.

GLM and GAM

The *generalized linear model* framework generalizes two aspects of the linear model.

- The expectation of the response is related to the linear combination of parameters by some specified *link function*:

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

- The distribution of the response variable doesn't have to be Normal.

Now, consider an "even more general" formulation by allowing transformations of the original explanatory variables:

$$g(\mathbb{E}Y_i) = \beta_0 + \beta_1 f_1(x_i^{(1)}, \ldots, x_i^{(d)}) + \ldots + \beta_m f_m(x_i^{(1)}, \ldots, x_i^{(d)})$$

GLM and GAM

The *generalized linear model* framework generalizes two aspects of the linear model.

- The expectation of the response is related to the linear combination of parameters by some specified *link function*:

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

- The distribution of the response variable doesn't have to be Normal.

Now, consider an "even more general" formulation by allowing transformations of the original explanatory variables:

$$g(\mathbb{E}Y_i) = \beta_0 + \beta_1 f_1(x_i^{(1)}, \ldots, x_i^{(d)}) + \ldots + \beta_m f_m(x_i^{(1)}, \ldots, x_i^{(d)})$$

This reduces to ordinary GLM and thus isn't actually any more general, but it's indicative of a more general modeling strategy:

GLM and GAM

The *generalized linear model* framework generalizes two aspects of the linear model.

- The expectation of the response is related to the linear combination of parameters by some specified *link function*:

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

- The distribution of the response variable doesn't have to be Normal.

Now, consider an "even more general" formulation by allowing transformations of the original explanatory variables:

$$g(\mathbb{E}Y_i) = \beta_0 + \beta_1 f_1(x_i^{(1)}, \ldots, x_i^{(d)}) + \ldots + \beta_m f_m(x_i^{(1)}, \ldots, x_i^{(d)})$$

This reduces to ordinary GLM and thus isn't actually any more general, but it's indicative of a more general modeling strategy: ***search over a large space of functions as part of the fitting process***.

GLM and GAM

The *generalized linear model* framework generalizes two aspects of the linear model.

- The expectation of the response is related to the linear combination of parameters by some specified *link function*:

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

- The distribution of the response variable doesn't have to be Normal.

Now, consider an "even more general" formulation by allowing transformations of the original explanatory variables:

$$g(\mathbb{E}Y_i) = \beta_0 + \beta_1 f_1(x_i^{(1)}, \ldots, x_i^{(d)}) + \ldots + \beta_m f_m(x_i^{(1)}, \ldots, x_i^{(d)})$$

This reduces to ordinary GLM and thus isn't actually any more general, but it's indicative of a more general modeling strategy: ***search over a large space of functions as part of the fitting process***.

This approach is called **generalized additive modeling**.

Adaptive Estimation

It's wise to *prefer* simple models. It's unwise to *limit yourself* to simple models even when abundant data is available.

Adaptive Estimation

It's wise to *prefer* simple models. It's unwise to *limit yourself* to simple models even when abundant data is available.

> Some rough categories are suggested for statisticians and engineers. Firstly, there are those who limit themselves to the simplest and most thoroughly understood models... Such models have low complexity (underfit)... Many researchers are in this first category because they lack the creativity to invent new laws, or the persistence to search for models that fit the data, or the willingness to let a computer aid in the search.
>
> Secondly, there are those who recognize the need for accurate nonparametric fits (consistency), but who routinely employ techniques... which provide little understanding of the data. Typically, the estimates have limited practical usefulness because all of the data is retained – rather than summarized. The estimates are overfit.
>
> The third category of scientists are those who employ sufficient insight and computational resources to consider a rich variety of conceivable distributions and find the law which best explains the data... Diligence is rewarded with discovery.

Adaptive Estimation

It's wise to *prefer* simple models. It's unwise to *limit yourself* to simple models even when abundant data is available.

> Some rough categories are suggested for statisticians and engineers. Firstly, there are those who limit themselves to the simplest and most thoroughly understood models... Such models have low complexity (underfit)... Many researchers are in this first category because they lack the creativity to invent new laws, or the persistence to search for models that fit the data, or the willingness to let a computer aid in the search.
>
> Secondly, there are those who recognize the need for accurate nonparametric fits (consistency), but who routinely employ techniques... which provide little understanding of the data. Typically, the estimates have limited practical usefulness because all of the data is retained – rather than summarized. The estimates are overfit.
>
> The third category of scientists are those who employ sufficient insight and computational resources to consider a rich variety of conceivable distributions and find the law which best explains the data... Diligence is rewarded with discovery.
>
> - "*Three kinds of statisticians*", excerpt from Prof Andrew Barron's PhD Thesis

R Script

Now, I'll turn to an R script to simulate wind speed and insurance claim severity data and compare the performance of GLM and GAM in estimating the expected severity.

If you'd like, you can download the file and execute the code along with me: *quantitations.com/static/gam.r*

Mixed Effects Modeling

Linear model vs linear mixed model

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

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

where $\epsilon_1, \ldots, \epsilon_n$ are iid Normal with mean zero.

Linear model vs linear mixed model

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

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

where $\epsilon_1, \ldots, \epsilon_n$ are iid Normal with mean zero.

The coefficients are considered *fixed*; that is, the model makes *absolutely no assumptions* about them.

Linear model vs linear mixed model

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

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

where $\epsilon_1, \ldots, \epsilon_n$ are iid Normal with mean zero.

The coefficients are considered *fixed*; that is, the model makes *absolutely no assumptions* about them.

In contrast, the general formulation of the linear *mixed* model with iid Normal errors:

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

with $\epsilon_1, \ldots, \epsilon_n$ iid Normal and the "*random effects*" $b_1, \ldots, b_m$ are assumed to be drawn from a multivariate Normal distribution.

Example: Blood pressure measurements

Suppose that n_f females and n_m males each had their [systolic] blood pressures measured. Assume that the distributions of female and male blood pressures are both Normal with the same standard deviation σ_w. Let their expectations be denoted μ_f and μ_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 represent this scenario with the linear model

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

Example: Blood pressure measurements

Suppose that n_f females and n_m males each had their [systolic] blood pressures measured. Assume that the distributions of female and male blood pressures are both Normal with the same standard deviation σ_w. Let their expectations be denoted μ_f and μ_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 represent this scenario with the linear model

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

Now suppose that each of the subjects' blood pressure was measured up to four times, over the course of a year. The measurements of a single person's blood pressure can vary quite a bit when measured at different times. How could we adapt our previous blood pressure model to capture this?

Example: Blood pressure measurements

Suppose that n_f females and n_m males each had their [systolic] blood pressures measured. Assume that the distributions of female and male blood pressures are both Normal with the same standard deviation σ_w. Let their expectations be denoted μ_f and μ_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 represent this scenario with the linear model

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

Now suppose that each of the subjects' blood pressure was measured up to four times, over the course of a year. The measurements of a single person's blood pressure can vary quite a bit when measured at different times. How could we adapt our previous blood pressure model to capture this?

Define $z_i^{(j)}$ to be to be the indicator variable for whether observation i came from subject j. Then, we might want to modify our model to

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

with iid Normal *random effects* $b_1, \ldots, b_m$.

Example: Blood pressure measurements

Suppose that n_f females and n_m males each had their [systolic] blood pressures measured. Assume that the distributions of female and male blood pressures are both Normal with the same standard deviation σ_w. Let their expectations be denoted μ_f and μ_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 represent this scenario with the linear model

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

Now suppose that each of the subjects' blood pressure was measured up to four times, over the course of a year. The measurements of a single person's blood pressure can vary quite a bit when measured at different times. How could we adapt our previous blood pressure model to capture this?

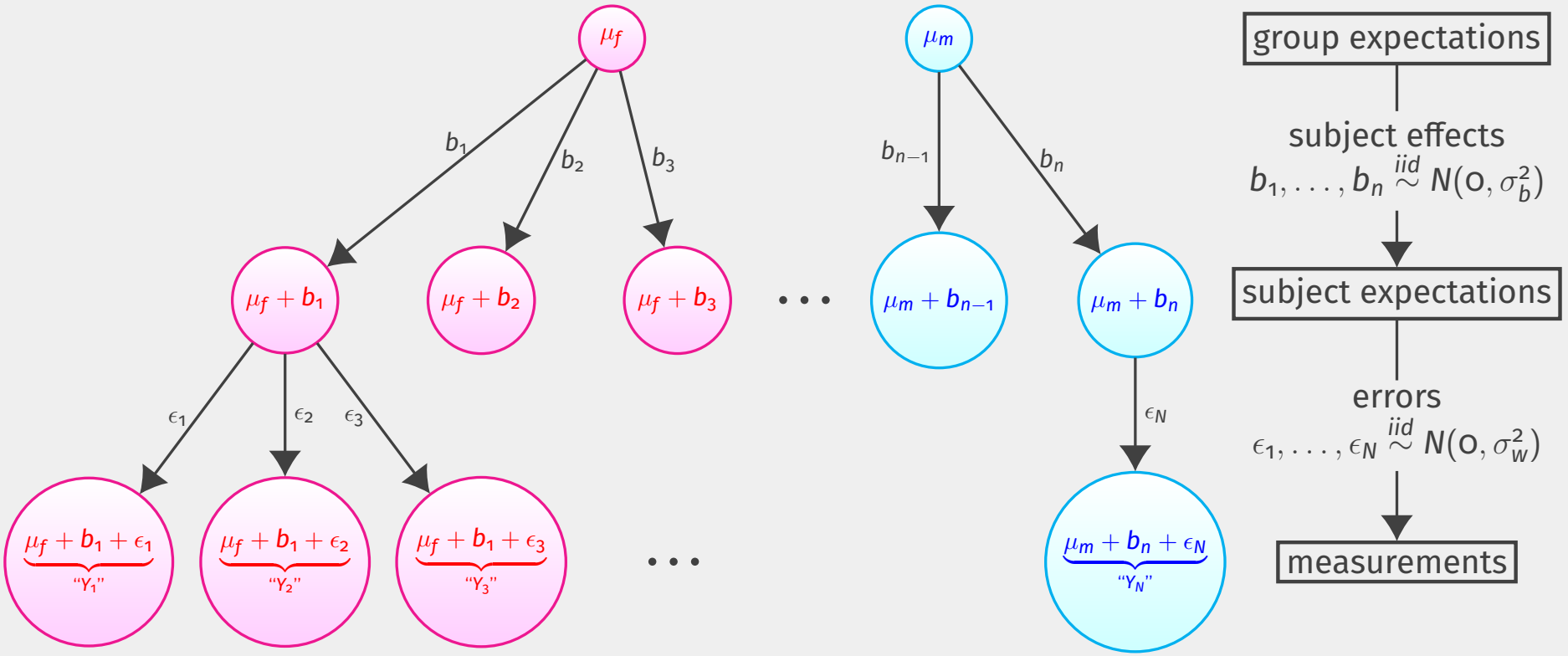
Define $z_i^{(j)}$ to be to be the indicator variable for whether observation i came from subject j. Then, we might want to modify our model to

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

with iid Normal *random effects* $b_1, \ldots, b_m$.
To better understand what this model means, suppose observation 12 came from subject 5 who is a female: the corresponding equation simplifies to

$$Y_{12} = \mu_f + b_5 + \epsilon_{12}$$

Model Diagram



R Script Example 1

Now, I'll turn to an R script to simulate blood pressure data according to the mechanism just described and draw some plots to visualize it.

If you'd like, you can download the file and execute the code along with me: *quantitations.com/static/mixed-effects.r*

SAT scores at various schools

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 σ^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 $(\mu_{\text{int}}, \mu_{\text{slope}})$.

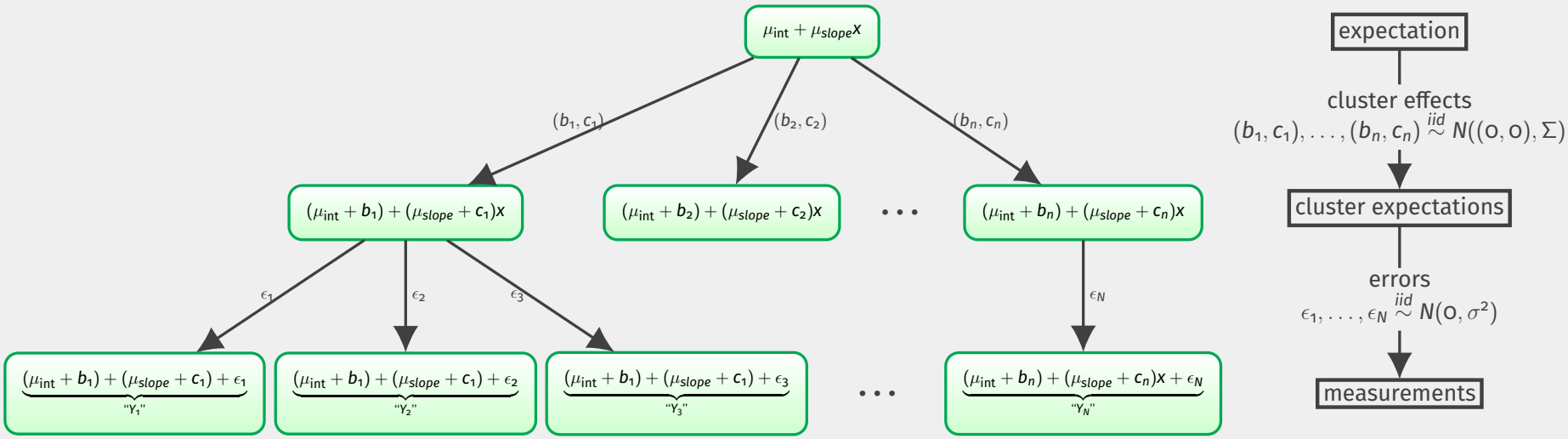
Model

This scenario fits into the linear mixed effects modeling framework described above. Let Y_i and x_i represent the ith 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 jth school's intercept and slope from the expectations shouldn't be assumed to be independent of each other.

Diagram


$\mu_{\text{int}} + \mu_{slope}x$
(b_1, c_1)
(b_2, c_2)
(b_n, c_n)
$(\mu_{\text{int}} + b_1) + (\mu_{slope} + c_1)x$
$(\mu_{\text{int}} + b_2) + (\mu_{slope} + c_2)x$
$(\mu_{\text{int}} + b_n) + (\mu_{slope} + c_n)x$
ϵ_1
ϵ_2
ϵ_3
ϵ_N
$\underbrace{(\mu_{\text{int}} + b_1) + (\mu_{slope} + c_1) + \epsilon_1}_{\text{“}Y_1\text{”}}$
$\underbrace{(\mu_{\text{int}} + b_1) + (\mu_{slope} + c_1) + \epsilon_2}_{\text{“}Y_2\text{”}}$
$\underbrace{(\mu_{\text{int}} + b_1) + (\mu_{slope} + c_1) + \epsilon_3}_{\text{“}Y_3\text{”}}$
$\underbrace{(\mu_{\text{int}} + b_n) + (\mu_{slope} + c_n)x + \epsilon_N}_{\text{“}Y_N\text{”}}$
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 $N(0, \sigma^2_{\text{school}})$.

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 $N(0, \sigma^2_{\text{school}})$.

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 $N(0, \sigma^2_{\text{student}})$.

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 $N(0, \sigma^2_{\text{school}})$.

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 $N(0, \sigma^2_{\text{student}})$.

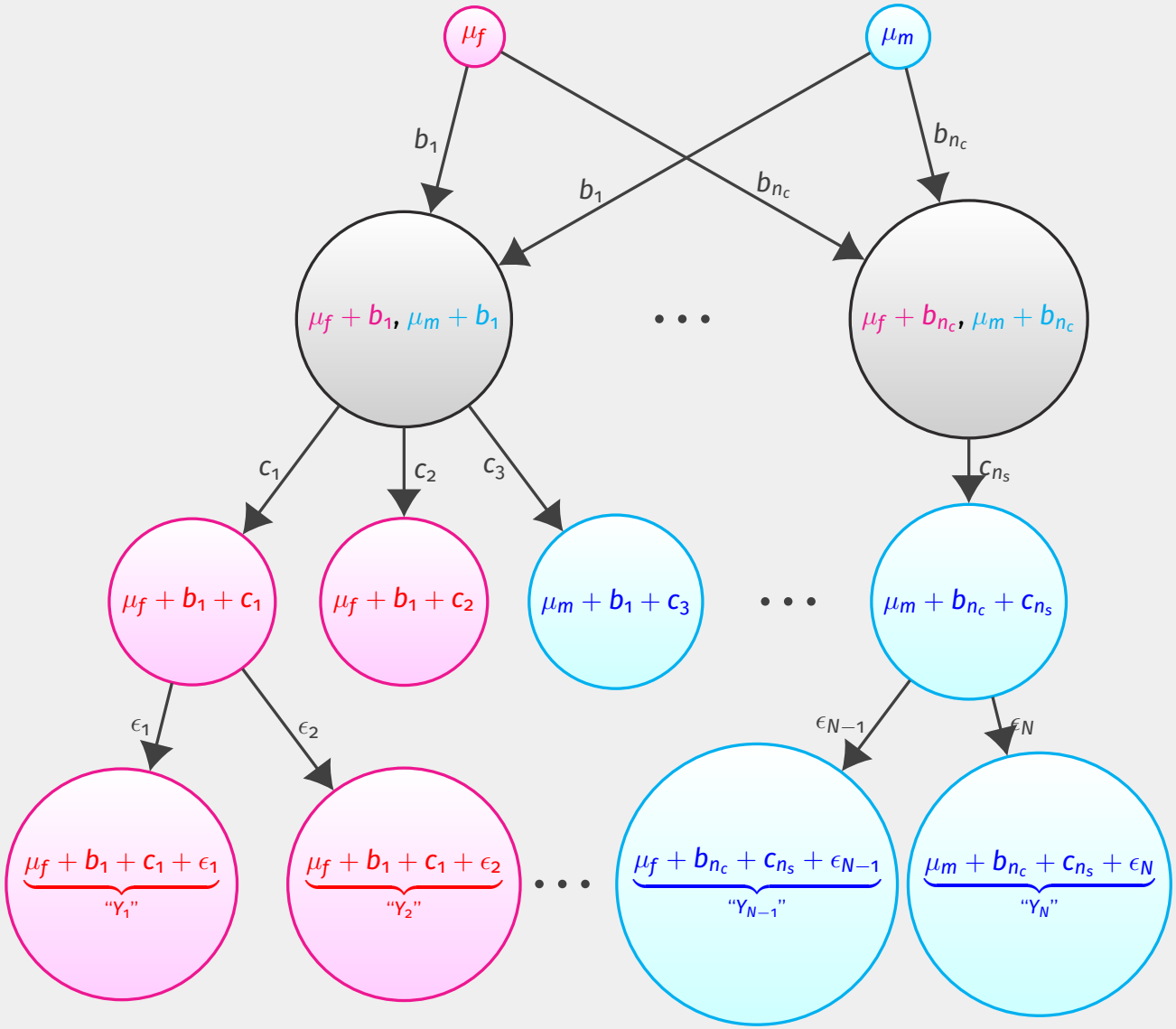
Each student takes the SAT up to 3 times, and their observed scores differ from the expectation by a Normal error. Assume the errors are iid $N(0, \sigma^2)$.

Model

We can represent this scenario as a mixed effects model with Y_i representing the ith 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,

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

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.

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 of 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 of 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 its 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 measurements 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 of 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 its 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 measurements 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*.

Model

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 jth 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 ith 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.

Model

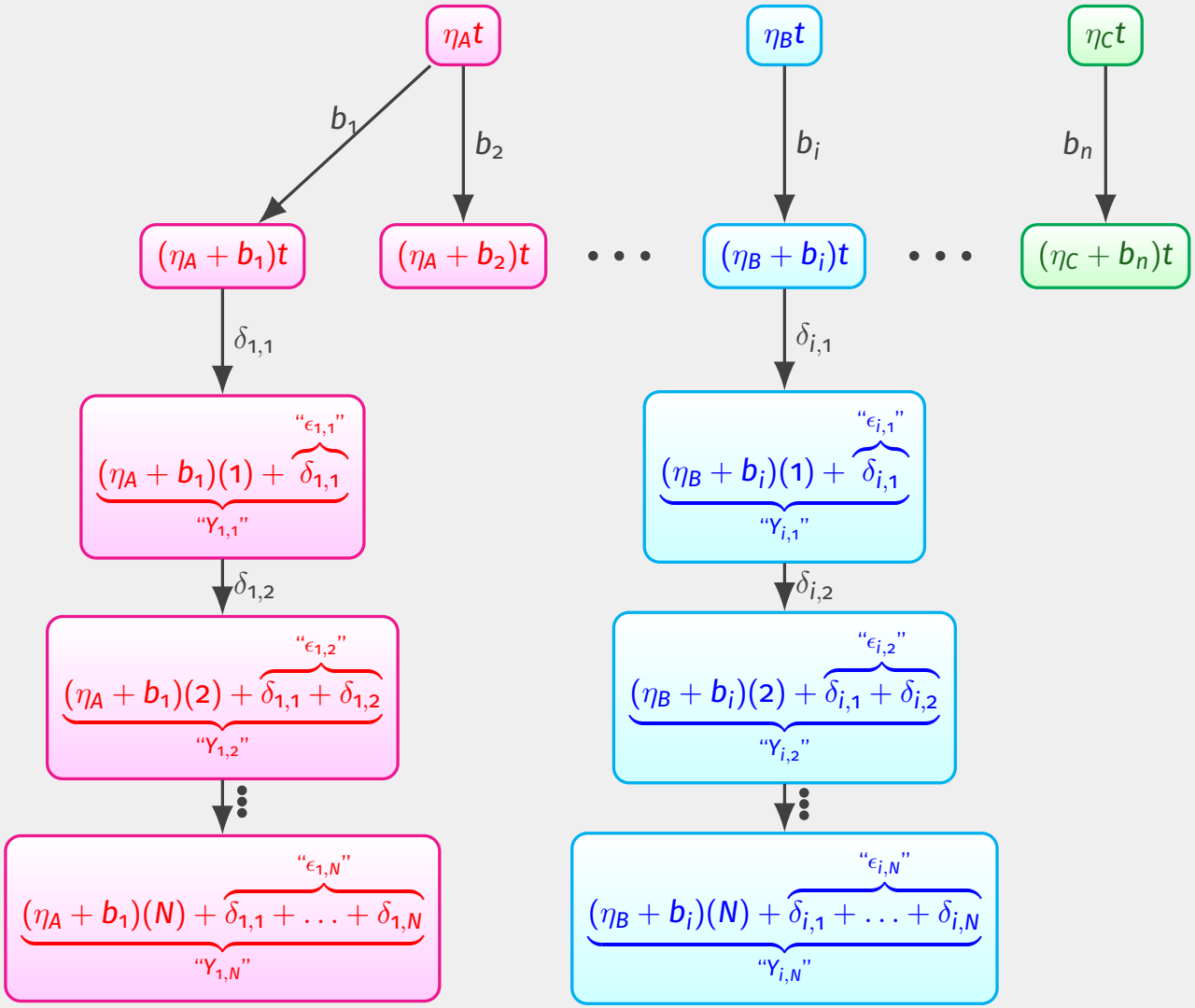
We can represent this scenario as a mixed effects model; this time we'll use double subscripts to index the observations. Let $Y_{i,j}$ represent the jth measurement of tree i (which is the height of tree i after j years), $z_i^{(1)}, z_i^{(2)}, \ldots$ indicators representing the tree tested, and b_i representing the ith 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

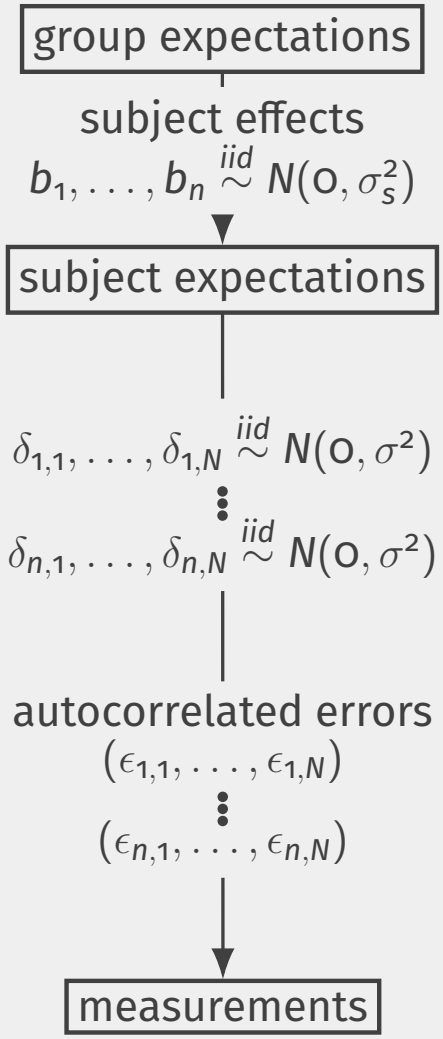
$$\begin{aligned} Y_{i,j} &= \eta_A j + (\eta_B - \eta_A)\mathbb{I}(x_i = \text{“}B\text{”})j + (\mu_C - \mu_A)\mathbb{I}(x_i = \text{“}C\text{”})j \\ &\quad + b_1 z_i^{(1)} j + b_2 z_i^{(2)} j + \ldots + \epsilon_{i,j} \\ &= \underbrace{(\mu_B + b_i)}_{\text{slope of the tree } i} j + \epsilon_{i,j}.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

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
$(\eta_A + b_1)t$
$(\eta_A + b_2)t$
$(\eta_B + b_i)t$
$(\eta_C + b_n)t$
$\delta_{1,1}$
$\delta_{i,1}$
"$\epsilon_{1,1}$"
$(\eta_A + b_1)(1) + \delta_{1,1}$
"$Y_{1,1}$"
"$\epsilon_{i,1}$"
$(\eta_B + b_i)(1) + \delta_{i,1}$
"$Y_{i,1}$"
$\delta_{1,2}$
$\delta_{i,2}$
"$\epsilon_{1,2}$"
$(\eta_A + b_1)(2) + \delta_{1,1} + \delta_{1,2}$
"$Y_{1,2}$"
"$\epsilon_{i,2}$"
$(\eta_B + b_i)(2) + \delta_{i,1} + \delta_{i,2}$
"$Y_{i,2}$"
"$\epsilon_{1,N}$"
$(\eta_A + b_1)(N) + \delta_{1,1} + \ldots + \delta_{1,N}$
"$Y_{1,N}$"
"$\epsilon_{i,N}$"
$(\eta_B + b_i)(N) + \delta_{i,1} + \ldots + \delta_{i,N}$
"$Y_{i,N}$"


group expectations
subject effects
$b_1, \ldots, b_n \overset{iid}{\sim} N(0, \sigma_s^2)$
subject expectations
$\delta_{1,1}, \ldots, \delta_{1,N} \overset{iid}{\sim} N(0, \sigma^2)$
$\delta_{n,1}, \ldots, \delta_{n,N} \overset{iid}{\sim} N(0, \sigma^2)$
autocorrelated errors
$(\epsilon_{1,1}, \ldots, \epsilon_{1,N})$
$(\epsilon_{n,1}, \ldots, \epsilon_{n,N})$
measurements

R script Example 4

Now, we'll work through the fourth and final part of *mixed-effects.r*, where we'll simulate tree growth 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.