Khairil Anwar Notodiputro

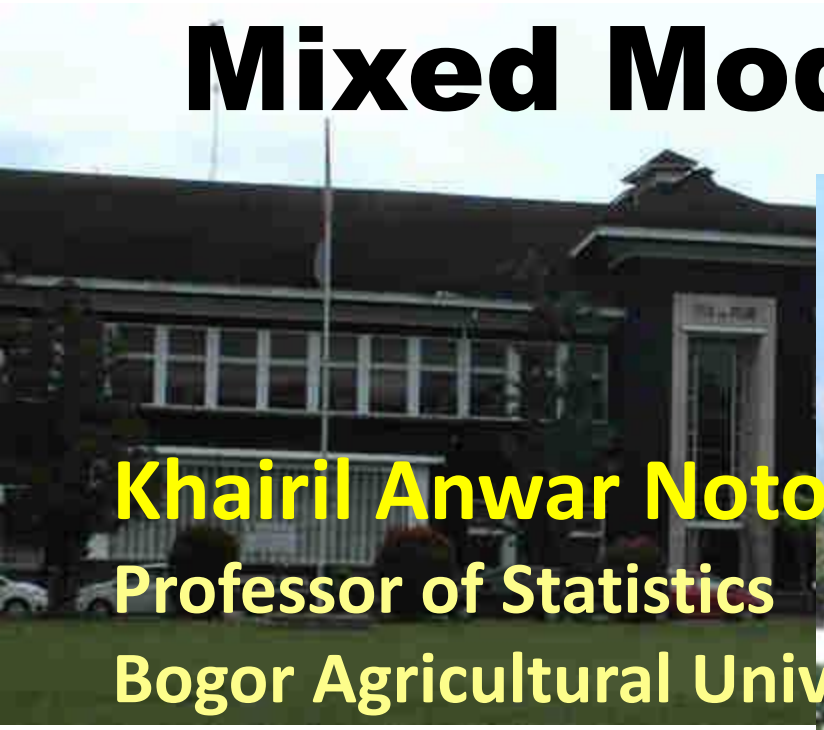
Earned BS and Master degrees in statistics from Bogor Agricultural University and Ph.D in statistics from Macquarie University, Australia.

Now he is a **Profesor of Statistics at Bogor Agricultural University** and a **Visiting Professor at Prince of Songkla University, Thailand.**

He is a member of ISI (International Statistical Institute) and a former Chairman of Indonesian Statistical Association. He is actively doing competitive research as well as supervising undergraduate and postgraduate students. His research interests include **mixed models, small area estimation** and **time series analysis.**

During the last five years he has taught the following subjects: **Time Series Analysis and Forecasting, Advanced Data Analysis, Categorical Data Analysis** and **Statistical Analysis.**

Mixed Models in Practice



Khairil Anwar Notodiputro
Professor of Statistics
Bogor Agricultural University

khairilnotodiputro@gmail.com

INVITATION LETTER

March 21, 2016

Professor Khairil Anwar Notodiputro, Ph.D.
Department of Statistics
Faculty of Mathematics and Sciences
Bogor Agricultural University
Indonesia

Dear Professor Khairil Anwar Notodiputro,

The Department of Statistics invites you to visit Kasetsart University during May 2016. We expect that your visit will be on May 13th 2016.

We would like to invite you to give a special lecture on the topic of "Mixed Models in Practice" for graduate students. Furthermore, we would like you to collaborate on research projects that will focus on applied sciences in the future.

We would like to thank you in advance for visiting department of statistics at Kasetsart University, Thailand.

Sincerely yours,

Assist.Prof. Boonorm Chomtee
Department Head

DEPARTMENT OF STATISTICS,
FACULTY OF SCIENCE
KASETSART UNIVERSITY
P. O. BOX 1086 , BANGKOK 10903 , THAILAND
TEL: 66-0-2562-5444
e-mail: fsciboc@ku.ac.th

Mixed Models in Practice

Khairil Anwar Notodiputro
Department of Statistics, Bogor Agricultural University
e-mail: khairilnotodiputro@gmail.com

Abstract. *Over the past decade there has been an explosion of developments in mixed effects models and their applications. The models have been applied in many areas such as agriculture, medicine, biology, biostatistics, education, social, economic and management sciences. The mixed models have also been intensively used in small area estimation, an area of research where proper statistical methods are needed to obtain estimates with high precision in a situation where the available samples are very limited. In this presentation, the mixed models have been discussed. We started by discussing the limitation of linear models and then introduced the linear mixed models. The importance of the linear mixed models has become evident if the predictor variables consisted of both fixed effects and random effects. These two types of effects have often been observed in practice. However, the linear mixed models are based on normal distributions of the response variables as well as the random effects whereas in practice these variables may not be normally distributed. Hence, the generalized linear mixed models (GLMM) need to be employed for non-normal distributions of the response variables, especially the exponential family of distributions. Moreover, since the models are very general, then basically the models can be used to handle various distributions of the response variables as well as the random effect distributions. Finally, to demonstrate the practice of GLMM, we have shown our works on mixed models. Several of the works have been related to the Na Thap river project, an IPB - PSU Collaborative Research project, RPM ID16287.*

Keywords: *Bias Reduction, Generalized Linear Mixed Models Hierarchical Models,, Linear Models, Mixed Models, Multilocation Eksperiment, Na Thap River*

- Introduction
- Types of Model Effects
- Why Mixed Models?
- Estimation and Inference
- Mixed Model in Practice
- Concluding Remarks

Introduction

Introduction

Consider a **linear model** $\boldsymbol{y = Xb+\varepsilon}$, where $\boldsymbol{\varepsilon}$ is assumed to have a Gaussian disribution.

What if $\boldsymbol{\varepsilon}$ is non-Gaussian?

Pair	Treatment_0	Treatment_1
1	98	94
2	95	36
3	93	85
4	94	88
5	99	91
6	61	82
7	84	43
8	92	71

Imagine a study to compare two treatments, "control" treatment vs "test" treatment. The response variable is the incidence of favourable outcomes. A paired comparison is used by observing 100 individuals on each treatment on each pair. Assuming independent observations, the number of favorable outcomes for each pair-treatment combination has a binomial distribution with n=100 and probability π_{ij} for the i^{th} treatment and j^{th} pair.

How should we model the data?

Introduction (cont..)

We can start with the normal approximation and the model can be described as follows:

Pair	Treatment_0	Treatment_1
1	98	94
2	95	36
3	93	85
4	94	88
5	99	91
6	61	82
7	84	43
8	92	71

- The sample proportions are 0.738 and 0.895
- T-test: p-value = 0.1132

- Response variable: Let $p_{ij} = y_{ij}/100$ denotes the sample proportion, where y_{ij} is the number of favorable outcomes out of 100 individuals.
- The model is $p_{ij} = \mu + \tau_i + \rho_j + \varepsilon_{ij}$ where μ denotes an overall mean, τ_i denotes i^{th} treatment effect, ρ_j denotes j^{th} pair effect, and ε_{ij} is the sampling error assumed i.i.d $N(0,\sigma^2)$

Introduction (cont..)

One obvious problem with the previous analysis is as follows

- Normal approximation not-withstanding, the response variable remains binomial, meaning the variance must depend on π.
- Assuming π changes from treatment to treatment (or among pairs) the assumption of equal variance does not hold.
- Transformation is required → the **arc sine square root** transformation.

At this point in this example, we have reached the limits of the linear model.

Introduction (cont..)

The limitation of the Linear Model (LM) can be overcome by the Generalized Linear Mixed Model (GLMM) which involves 3 components:

- The **distribution** of the observations:

$$y_{ij} \mid \rho_j \sim Binomial\ (100, \pi_{ij}\ ).$$

- The **linear predictor**: $\eta_{ij} = \eta + \tau_i + \rho_j$, where η is the intercept and asume the pair effects ρ_j are i.i.d. $N(0, \sigma^2{}_\rho)$.
- The **link function**: with non-normal data, the canonical parameter of the log likelihood is typically a better candidate for fitting a linear model than the mean itself. For the binomial, the canonical link function is $\eta_{ij} = log[\pi_{ij} /(1-\pi_{ij}\ )]$.

Types of Model Effects

Types of Model Effects

- The types of model effects can be distinguished based on how the levels of an effect are chosen and the scope of the intended inference.
- The model of the previous example:

$$p_{ij} = \mu + \tau_i + \rho_j + \varepsilon_{ij}$$

- Treatment effects τ have been fixed with two conditions (either "control" or "test")
- Pair effects ρ depend on the individuals chosen as a sample.
- So τ are fixed effects and ρ are random effects

Types of Model Effects (cont..)

In a **fixed-effects model** for an experiment, all the factors in the experiment have a predetermined set of levels and the only inferences are for the levels of the factors actually used in the experiment.

In a **random effects model** for an experiment, the levels of factors used in the experiment are randomly selected from a population of possible levels. The inferences from the data in the experiment are for all levels of the factors in the population from which the levels were selected and not only the levels used in the experiment.

Types of Model Effects (cont..)

In a **mixed-effects model** (or simply **mixed model**) for an experiment, the levels of some of the factors used in the experiment are randomly selected from a population of possible levels, whereas the levels of the other factors in the experiment are predetermined.

The inferences concerning factors with fixed levels are only for the levels of the factors used in the experiment, whereas inferences concerning factors with randomly selected levels are for all levels of the factors in the population from which the levels were selected.

Why Mixed Models?

Why Mixed Models?

Consider a hypothetical data (Stroup, 2013):

Binomial regression data

Obs	X	N	Fav
1	0	11	0
2	1	7	0
3	2	9	2
4	3	11	2
5	4	12	2
6	5	15	5
7	6	11	7
8	7	15	12
9	8	11	10
10	9	16	16
11	10	10	9

One observation per level of X

Multi-batch data, two responses (Y and Fav)

X	Batch 1			Batch2			Batch 3			Batch 4		
	Y	Fav	N	Y	Fav	N	Y	Fav	N	Y	Fav	N
0	95.6	15	21	96.6	18	21	96.6	16	19	96.6	18	21
3	96.9	13	17	96.7	14	16	96.5	13	19	96.8	18	21
6	98.5	19	23	96.3	14	17	97.7	17	22	96.4	14	20
9	99.0	14	17	96.3	17	20	98.3	23	27	96.6	14	19
12	100.2	18	23	96.5	15	20	99.1	16	21	96.8	14	19
18	101.9	19	27	96.4	14	22	100.5	11	16	96.9	11	20
24	104.1	15	20	95.7	12	22	101.2	13	18	97.1	11	16
36	107.8	14	21	95.2	11	25	103.3	10	17	97.4	10	21
48	111.5	13	18	94.6	5	26	105.9	6	16	97.4	5	19

Levels of X are observed for multiple batches

Why Mixed Models? (cont..)

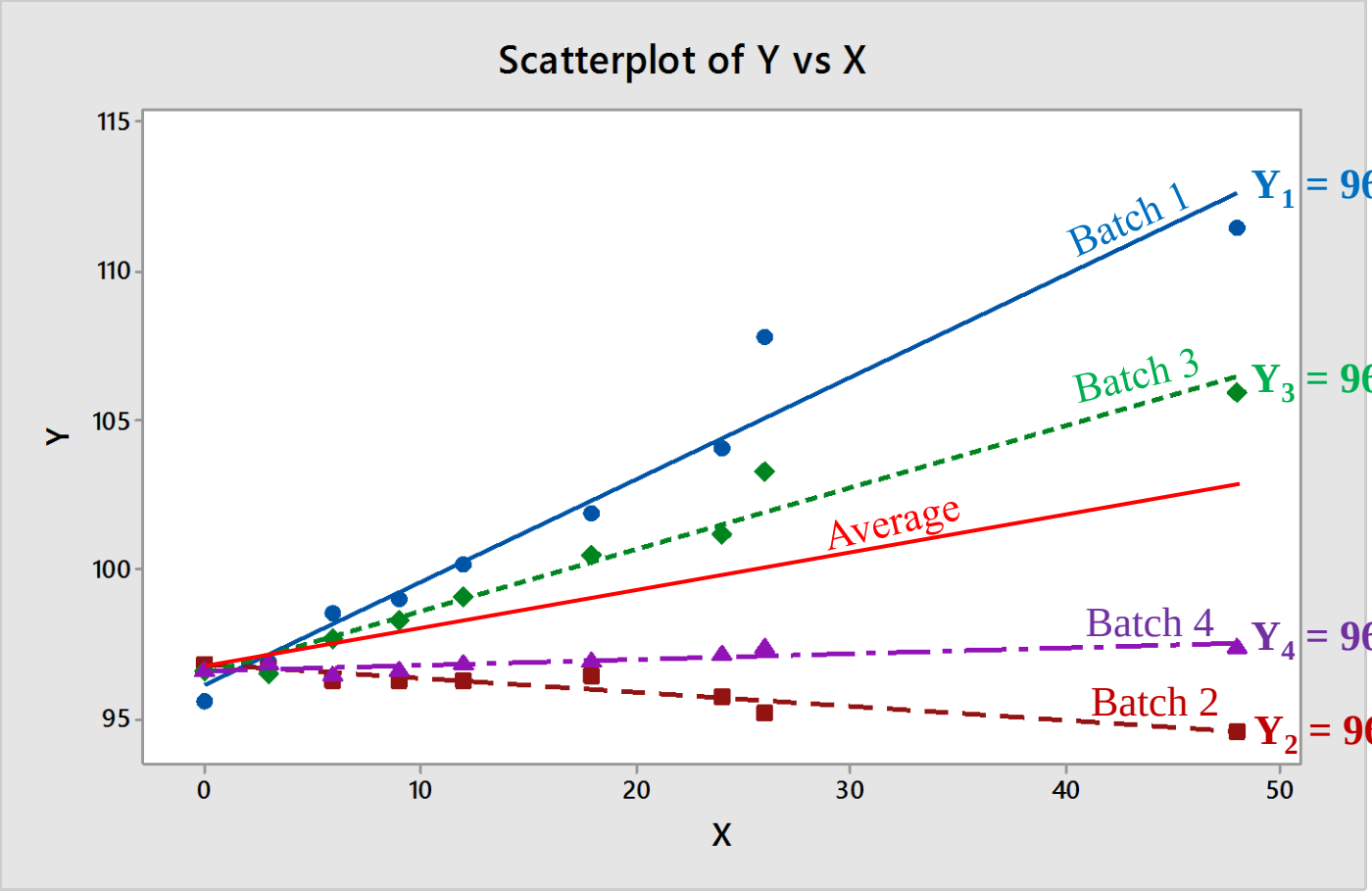
Multi-batch data

X	Batch 1			Batch2			Batch 3			Batch 4		
	Y	Fav	N	Y	Fav	N	Y	Fav	N	Y	Fav	N
0	95.6	15	21	96.6	18	21	96.6	16	19	96.6	18	21
3	96.9	13	17	96.7	14	16	96.5	13	19	96.8	18	21
6	98.5	19	23	96.3	14	17	97.7	17	22	96.4	14	20
9	99.0	14	17	96.3	17	20	98.3	23	27	96.6	14	19
12	100.2	18	23	96.5	15	20	99.1	16	21	96.8	14	19
18	101.9	19	27	96.4	14	22	100.5	11	16	96.9	11	20
24	104.1	15	20	95.7	12	22	101.2	13	18	97.1	11	16
36	107.8	14	21	95.2	11	25	103.3	10	17	97.4	10	21
48	111.5	13	18	94.6	5	26	105.9	6	16	97.4	5	19

This table shows 9 levels of *X* varying from 0 to 48. For each of four batches, a continuous variable (*Y*) and a binomial variable (*N* = number of Bernoulli trials and *Fav* = number of "successes") are observed at each level of *X*.

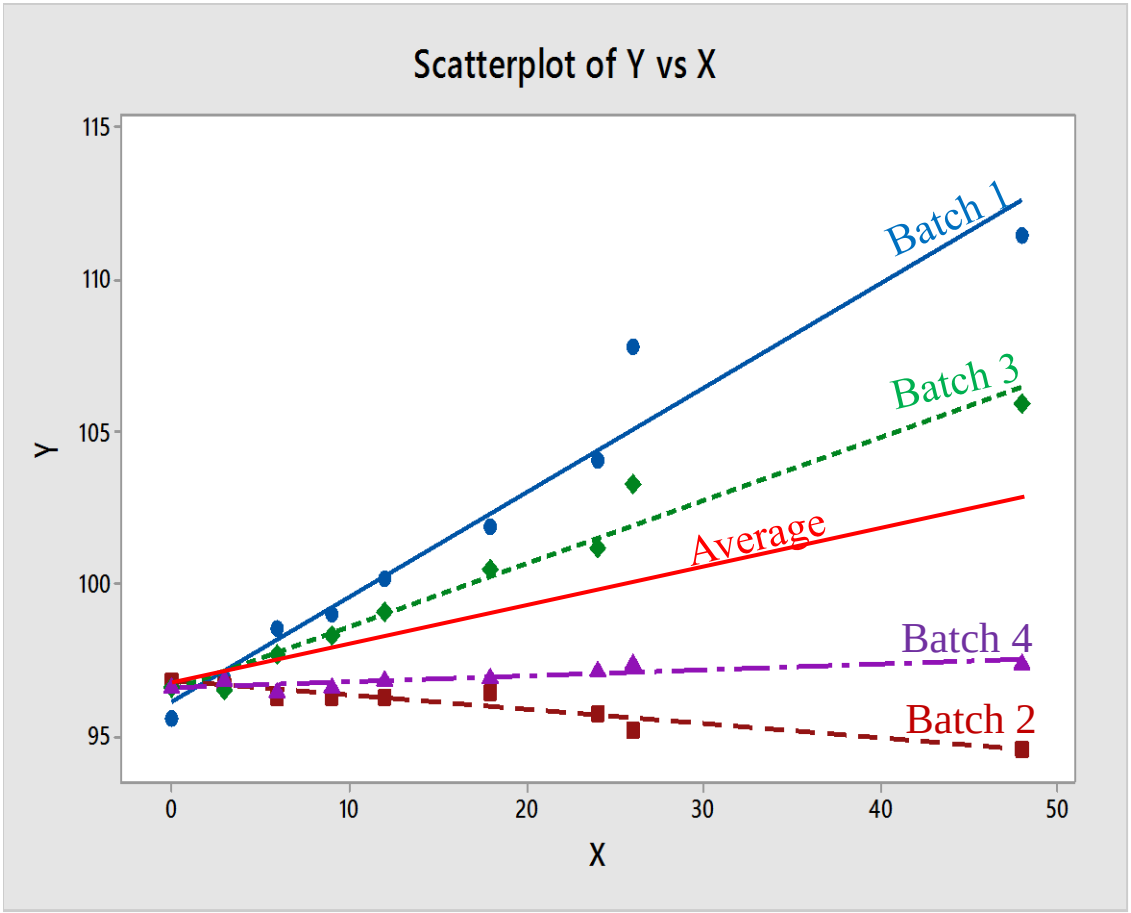
How should these data be modeled ?

Why Mixed Models? (cont..)



Look at the response Y. The middle line shows the average linear regression over all batches; the other lines show the regressions for each individual batch. Inspecting the graph suggests that assuming linear regression is reasonable but a common regression for all batches may not be justified.

Why Mixed Models? (cont..)



Separate linear regressions by batch yields the linear predictor $\eta_{ij} = \beta_{0i} + \beta_{1i} X_{ij}$, where β_{0i} and β_{1i} are, respectively, the intercept and slope for the i^{th} batch.

$\beta_0 + b_{0i}$

$(\beta_1 + b_{1i})$

Or $\eta_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})X_{ij}$,

β_0, β_1 overall intercept and slope

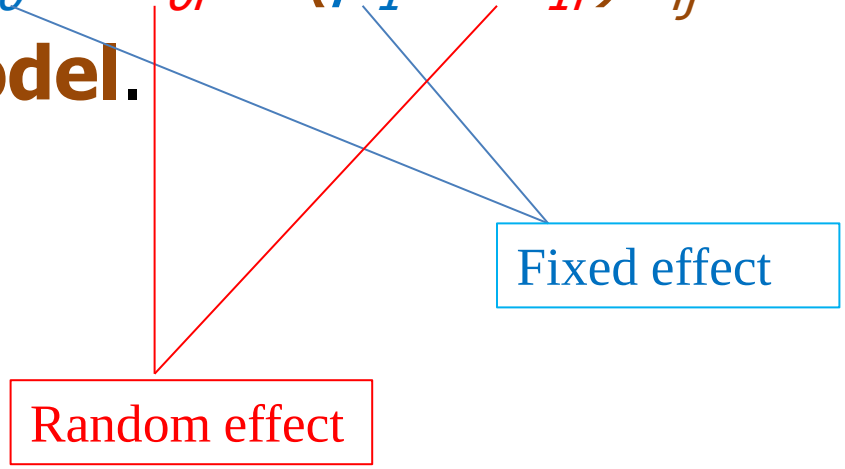
b_{0i}, b_{1i} batch-specific deviations from β_0 and β_1.

Why Mixed Models? (cont..)

- Let us assume that the batches represent a sample of a larger population of batches so that b_{0i} and b_{1i} are random variables and, thus, have probability distributions.
- Hence, the above model, i.e.

$$\eta_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})X_{ij}$$

is a **mixed model**.



Why Mixed Models? (cont..)

- To summarize, the multi-batch regression example gives rise to the linear predictor:

$$\eta_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) X_{ij}$$

- (b_{0i}, b_{1i}) pairs are assumed independent and within each batch, the pairs are bivariate normal, that is,

$$\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix} \right)$$

where

σ_0^2 and σ_1^2 are the variances of b_{0i} and b_{1i}, respectively

σ_{01} is the covariance between b_{0i} and b_{1i}.

Why Mixed Models? (cont..)

- If the observed data is continuous (Y) and have a Gaussian distribution, then the linear predictor is an estimate of the data's expected value, conditional on the random effects b_{0i} and b_{1i} as follows:

1. Observations: $(y_{ij}|b_{0i},b_{1i}) \sim N(\mu_{ij}|b_{0i},b_{1i},\sigma^2)$
2. Model focus: $E(y_{ij}|b_{0i},b_{1i}) = \mu_{ij}|b_{0i},b_{1i}$
3. Linear predictor: $\eta_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})\, X_{ij}$
4. Assumptions about b_{0i} and b_{1i} (if random), for example, as shown previously
5. Link function: identity, that is, $\mu_{ij}|b_{0i},b_{1i}$ modeled by $\beta_0 + b_{0i} + (\beta_1 + b_{1i})\, X_{ij}$

Why Mixed Models? (cont..)

- On the other hands, if you model the variable *Fav* in the table, then you adjust the distribution of the observations and the link function, accordingly.

1. Observations: $(Fav_{ij}|b_{0i}, b_{1i}) \sim Binomial[N_{ij}, (\pi_{ij}|b_{0i}, b_{1i})]$
2. Model focus: $E(Fav_{ij}|b_{0i}, b_{1i}) = \pi_{ij}|b_{0i}, b_{1i}$
3. Linear predictor: $\eta_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})\, X_{ij}$
4. Assumptions about b_{0i} and b_{1i} (if random)
5. Link function: logit for example, $log[\pi_{ij}/(1 - \pi_{ij})]$ modeled by $\beta_0 + b_{0i} + (\beta_1 + b_{1i})\, X_{ij}$, where π_{ij} is used here as a shorthand for $\pi_{ij}|b_{0i}, b_{1i}$

Why Mixed Models? (cont..)

- Back to the linear predictor:

$$\eta_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})\, X_{ij}$$

- This predictor consists of **fixed** effects and **random** effects.
- The fixed effects component of the linear predictor is $(\beta_0 + \beta_1 X_{ij})$ whereas the random effects component is $(b_{0i} + b_{1i} X_{ij})$.
- In matrix form the fixed effects are denoted by $\boldsymbol{X\beta}$ whereas the random effects are denoted by $\boldsymbol{Zb}$.
- Hence, the linear predictor: $\boldsymbol{\eta = X\beta + Zb}$
- Vector $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{G})$

Why Mixed Models? (cont..)

$$\boldsymbol{X\beta} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ \dots \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} 0 \\ 3 \\ 6 \\ 9 \\ 12 \\ 18 \\ 24 \\ 36 \\ 48 \\ 0 \\ \dots \\ 48 \\ 0 \\ 3 \\ 6 \\ 9 \\ 12 \\ 18 \\ 24 \\ 36 \\ 48 \end{bmatrix} \beta_1$$

$$\boldsymbol{Zb} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_{01} \\ b_{02} \\ b_{03} \\ b_{04} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 \\ 9 & 0 & 0 & 0 \\ 12 & 0 & 0 & 0 \\ 18 & 0 & 0 & 0 \\ 24 & 0 & 0 & 0 \\ 36 & 0 & 0 & 0 \\ 48 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 48 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & 9 \\ 0 & 0 & 0 & 12 \\ 0 & 0 & 0 & 18 \\ 0 & 0 & 0 & 24 \\ 0 & 0 & 0 & 36 \\ 0 & 0 & 0 & 48 \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{12} \\ b_{13} \\ b_{14} \end{bmatrix}$$

Estimation and Inference

Estimation and Inference

- Recall that a fully specified GLMM is given by:
  - A linear predictor $\boldsymbol{X\beta + Zb}$ (or simply $\boldsymbol{X\beta}$ if there are no random effects)
  - The distribution of $\boldsymbol{y|b}$ — the observations, conditional on the random effects if there are random effects
  - The link function, $\boldsymbol{\eta = g(\mu|b)}$, where $\boldsymbol{\mu|b = E(y|b)}$
  - The random effects distribution, $\boldsymbol{b \sim N(0,G)}$

- For models with Gaussian data dan random effects:
  - Linear predictor: $\boldsymbol{\eta = X\beta + Zb}$
  - Distributions: $\boldsymbol{y|b \sim N(\mu|b, R); b \sim N(0,G)}$
  - Link: $\boldsymbol{\eta = \mu|b}$

Estimation and Inference

- The log-likelihood equations for $\boldsymbol{y}|\boldsymbol{b}$ and $\boldsymbol{b}$, respectively, are

$$\ell(y \mid b) = -\left(\frac{n}{2}\right)\log(2\pi) - \left(\frac{1}{2}\right)\log(|R|) - \left(\frac{1}{2}\right)(y - X\beta - Zb)' R^{-1}(y - X\beta - Zb)$$

and

$$\ell(b) = -\left(\frac{b}{2}\right)\log(2\pi) - \left(\frac{1}{2}\right)\log(|G|) - \left(\frac{1}{2}\right)b'G^{-1}b$$

- Focusing on the "quasi-likelihood" part:

$$\ell(y, b) = -\left(\frac{1}{2}\right)(y - X\beta - Zb)' R^{-1}(y - X\beta - Zb) - \left(\frac{1}{2}\right)b'G^{-1}b$$

Estimation and Inference

- MLE can be obtained by setting $\partial \ell(\mathbf{y},\mathbf{b})/\partial \boldsymbol{\beta}'$ and $\partial \ell(\mathbf{y},\mathbf{b})/\partial \mathbf{b}'$ equal to zero and solving the resulting set of equations for $\boldsymbol{\beta}$ and $\mathbf{b}$:

$$\frac{\partial[\ell(\mathbf{y},\mathbf{b})]}{\partial \boldsymbol{\beta}'} = \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} - \mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\mathbf{b}$$

$$\frac{\partial[\ell(\mathbf{y},\mathbf{b})]}{\partial \mathbf{b}'} = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} - \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{\beta} - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{b} - \mathbf{G}^{-1}\mathbf{b}$$

- The resulting mixed model equations:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

This is called the Linear Mixed Model, i.e. the GLMM with Gaussian response and random effects.

Estimation and Inference

- If the distribution of response variable is not normal then we use the GLMM with the essential features:

  - Linear predictor: $\eta = X\beta + Zb$
  - Distribution: $b \sim N(0,G)$
  - Distribution or quasi-likelihood: $E(y|b) = \mu|b$;

    $Var(y|b) = V_\mu^{1/2} A V_\mu^{1/2}$, where

    $V_\mu^{1/2} = \text{diag}\,[(\partial^2 b(\theta)/\partial\theta^2)^{1/2}]$ and $A = diag\,[1/a(\varphi)]$, and $y|b$ either has a distribution that belongs to the exponential family or a quasi-likelihood
  - Link: $\eta = g(\mu|b)$, or alternatively, inverse link:

    $X\beta + Zb = h(\eta)$. Typically, $h(\cdot) = g^{-1}(\cdot)$

Estimation and Inference

- The quasi-likelihood of the observation conditional on the random effects is

$$ql(\boldsymbol{y}|\boldsymbol{b}) = \boldsymbol{y'A\theta - 1'Ab(\theta)}$$

- The log-likelihood of the random effects is

$$\ell(\boldsymbol{b}) = -(\boldsymbol{b/2})\, log(2\pi) - (1/2)log(|\boldsymbol{G}|) - (1/2)\boldsymbol{b'G^{-1}b}$$

- The joint log(quasi)-likelihood is thus $\ell(\boldsymbol{b}) + ql(\boldsymbol{y}|\boldsymbol{b})$
- The marginal (quasi-) likelihood is

$$\int_{\mathbf{b}}\int \left[ql(\mathbf{y}\,|\,\mathbf{b}) + \ell(\mathbf{b})\right] d\mathbf{b} = ql(\mathbf{y})$$

$$= \int_{\mathbf{b}}\int \left[\mathbf{y}'\mathbf{A}\boldsymbol{\theta} - 1'\mathbf{A}b(\boldsymbol{\theta}) - \left(\frac{b}{2}\right)\log(2\pi) - \left(\frac{1}{2}\right)\log(|\mathbf{G}|) - \left(\frac{1}{2}\right)\mathbf{b}'\mathbf{G}^{-1}\mathbf{b}\right] d\mathbf{b}$$

Approximation is required → *pseudo-likelihood* method or Laplace approximation or Gauss Hermite quadrature.

Estimation and Inference

- The idea of ***pseudo-likelihood*** is to approximate the inverse link function by Taylor series expansion evaluated at $\tilde{\eta}$
- The Taylor series expansion: $h(\eta) \cong h(\tilde{\eta}) + \left.\frac{\partial h(\eta)}{\partial \eta}\right|_{\eta=\tilde{\eta}} (\eta - \tilde{\eta})$

- If $\boldsymbol{D} = diag\,[(\partial g(\mu|\boldsymbol{b})/\partial \mu)]$ then (Jiang, 2007)

$$h(\eta) \approx h(\tilde{\eta}) + \tilde{\boldsymbol{D}}^{-1}(\boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} - \boldsymbol{X}\tilde{\beta} - \boldsymbol{Z}\tilde{\boldsymbol{b}})$$

where $\tilde{D}$ denotes D evaluated at $\tilde{\eta} = \boldsymbol{X}\tilde{\beta} - \boldsymbol{Z}\tilde{\boldsymbol{b}}$.

- Let $\boldsymbol{y}^* = \tilde{\eta} + \tilde{\boldsymbol{D}}^{-1}[\boldsymbol{y} - (\tilde{\mu}|\tilde{\boldsymbol{b}})] = \boldsymbol{X}\tilde{\beta} + \boldsymbol{Z}\tilde{\boldsymbol{b}} + \tilde{\boldsymbol{D}}^{-1}[\boldsymbol{y} - h(\tilde{\eta})]$
- Then $E(\boldsymbol{y}^*|\boldsymbol{b}) = \tilde{\boldsymbol{D}}[h(\eta) - h(\tilde{\eta})] + \boldsymbol{X}\tilde{\beta} + \boldsymbol{Z}\tilde{\boldsymbol{b}} \approx \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b}$

$$Var(\boldsymbol{y}^*|\boldsymbol{b}) = \boldsymbol{D}\boldsymbol{V}_{\mu}^{1/2}\boldsymbol{A}\boldsymbol{V}_{\mu}^{1/2}\boldsymbol{D}$$

Estimation and Inference

- Let us fit a binomial mixed model to the multibatch data:

X	Batch 1			Batch2			Batch 3			Batch 4		
	Y	Fav	N	Y	Fav	N	Y	Fav	N	Y	Fav	N
0	95.6	15	21	96.6	18	21	96.6	16	19	96.6	18	21
3	96.9	13	17	96.7	14	16	96.5	13	19	96.8	18	21
6	98.5	19	23	96.3	14	17	97.7	17	22	96.4	14	20
9	99.0	14	17	96.3	17	20	98.3	23	27	96.6	14	19
12	100.2	18	23	96.5	15	20	99.1	16	21	96.8	14	19
18	101.9	19	27	96.4	14	22	100.5	11	16	96.9	11	20
24	104.1	15	20	95.7	12	22	101.2	13	18	97.1	11	16
36	107.8	14	21	95.2	11	25	103.3	10	17	97.4	10	21
48	111.5	13	18	94.6	5	26	105.9	6	16	97.4	5	19

1. Observations: $(Fav_{ij}|b_{0i}, b_{1i}) \sim B\,[N_{ij}, (\pi_{ij}|b_{0i}, b_{1i})]$
2. Model focus: $E(Fav_{ij}|b_{0i}, b_{1i}) = \pi_{ij}|b_{0i}, b_{1i}$
3. Linear predictor: $\eta_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})\, X_{ij}$
4. Assumptions about b_{0i} and b_{1i} (if random)
5. Logit link: $log[\pi_{ij}/(1 - \pi_{ij})] = \beta_0 + b_{0i} + (\beta_1 + b_{1i})\, X_{ij}$,

Fixed effects: X

Random effects : Batch

$$\beta_0 + b_{0i} + (\beta_1 + b_{1i})\, X_{ij}$$

Estimation and Inference

```
setwd("d:/Prof.Kan/mixmodel_Kan_1D")
read.csv("multibatchdata.csv",h=T,as.is=T) -> dt
library(lme4)
Batch<- as.character(dt$Batch)
gm1 <- glmer(cbind(dt$Fav, dt$N- dt$Fav) ~ dt$X+ (1 | Batch), family = binomial)
summary(gm1)
```

Binomial reponse

```
Generalized linear mixed model fit by maximum likelihood (Laplace  Approximation) [glmerMod]
Family: binomial  ( logit )
Formula: cbind(dt$Fav, dt$N - dt$Fav) ~ dt$X + (1 | Batch)

    AIC      BIC           logLik        deviance     df.resid
  151.8    156.5           -72.9          145.8           33
Scaled residuals:
    Min        1Q        Median     3Q            Max
-1.6315    -0.6042    0.1118       0.4402       3.0004

Random effects:
 Groups Name              Variance     Std.Dev.
 Batch  (Intercept)      0.01395      0.1181
Number of obs: 36, groups:  Batch, 4

Fixed effects:
                                Estimate     Std. Error    z value       Pr(>|z|)
(Intercept)   1.584756   0.149651      10.590       < 2e-16 ***
dt$X                            -0.045388   0.005963  -7.612          2.7e-14 ***
```

Estimation and Inference

- Let us fit a normal mixed model to the multibatch data:

X	Batch 1			Batch2			Batch 3			Batch 4		
	Y	Fav	N	Y	Fav	N	Y	Fav	N	Y	Fav	N
0	95.6	15	21	96.6	18	21	96.6	16	19	96.6	18	21
3	96.9	13	17	96.7	14	16	96.5	13	19	96.8	18	21
6	98.5	19	23	96.3	14	17	97.7	17	22	96.4	14	20
9	99.0	14	17	96.3	17	20	98.3	23	27	96.6	14	19
12	100.2	18	23	96.5	15	20	99.1	16	21	96.8	14	19
18	101.9	19	27	96.4	14	22	100.5	11	16	96.9	11	20
24	104.1	15	20	95.7	12	22	101.2	13	18	97.1	11	16
36	107.8	14	21	95.2	11	25	103.3	10	17	97.4	10	21
48	111.5	13	18	94.6	5	26	105.9	6	16	97.4	5	19

1. Observations: $(y_{ij}|b_{0i}, b_{1i}) \sim N(\mu_{ij}|b_{0i}, b_{1i}, \sigma^2)$
2. Model focus: $E(y_{ij}|b_{0i}, b_{1i}) = \mu_{ij}|b_{0i}, b_{1i}$
3. Linear predictor: $\eta_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) X_{ij}$
4. Assumptions about b_{0i} and $b_{1i} \sim N(0, \boldsymbol{G})$
5. Identity link: $(\mu_{ij}|b_{0i}, b_{1i}) = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) X_{ij}$

Fixed effects: X

Random effects : Batch

$\beta_0 + b_{0i} + (\beta_1 + b_{1i}) X_{ij}$

Estimation and Inference

Normal reponse

```
setwd("d:/Prof.Kan/mixmodel_Kan_1D")
read.csv("multibatchdata.csv",h=T,as.is=T) -> dt
library(lme4)
Batch<- as.character(dt$Batch)
gm2 <- glmer(dt$Y ~ dt$X+ (1 | Batch), family =gaussian)
summary(gm2)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: dt$Y ~ dt$X + (1 | Batch)
 REML criterion at convergence: 176.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.4164 -0.4797  0.0253  0.5767  2.4351

Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept)  6.332   2.516
 Residual              5.857   2.420
Number of obs: 36, groups:  Batch, 4

Fixed effects:
            Estimate Std. Error t value
(Intercept) 96.50627    1.40072   68.90
dt$X         0.13129    0.02867    4.58
```

Linear model

Source	DF	Adj SS	Adj MS	F-Value	P-Value
X	1	122.838	122.838	264.83	0.000
Batch	3	0.782	0.261	0.56	0.644
X*Batch	3	168.579	56.193	121.15	0.000
Error	28	12.987	0.464		
Total	35	492.943			

Estimation and Inference

Model diagnostics

- Literatures on diagnostics of mixed models involving random effects is not extensive.
- Limited methodology is available, mostly regarding assessing the distribution of the random effects and errors.
- The methods may be classified as diagnostic plots and goodness-of-fit tests.

Diagnostic plots. Lange and Ryan (1989) proposed to examine a Q–Q plot of some standardized linear combinations

$$z_i = \frac{c'\tilde{\alpha}_i}{\{c'\mathrm{Var}(\tilde{\alpha}_i)c\}^{1/2}}, \qquad i = 1, \ldots, m,$$

where c is a known vector. They argued that, through appropriate choices of c, the plot can be made sensitive to different types of model departures.

Estimation and Inference

Goodness of fit tests. Few authors have developed tests for checking distributional assumptions involved in linear mixed models. Consider a mixed ANOVA model

$$y = X\beta + Z\alpha + \varepsilon$$

where for $1 \le i \le s$, $\alpha_i = (\alpha_{ij})_{1 \le j \le mi}$, where the α_{ij}s are *i.i.d.* $\sim (0, \sigma^2)$, and continuous distribution $Fi = Fi(\cdot\,|\sigma_i)$; and $\varepsilon = (\varepsilon_\theta)_{1 \le j \le N}$, where the e_js are *i.i.d.* $\sim (0, \tau^2)$, and continuous distribution $G = G(\cdot\,|\tau)$; and $\alpha_1, \ldots, \alpha_s, \varepsilon$ are independent.

We are interested in testing the following hypothesis,

$$H_0 : F_i(\cdot|\sigma_i) = F_{0i}(\cdot|\sigma_i) \quad , 1 \le i \le s,$$
$$\text{and } G(\cdot|\tau) = G_0(\cdot|\tau);$$

that is, the distributions of the random effects and errors, up to a set of unknown variance components $\sigma_1^2, \ldots, \sigma_1^2, \tau^2$, are as assumed. The model does not fit if we reject this hipothesis.

Mixed Models in Practice

Mixed Models in Practice

- Mixed model is very popular
- **Google search** (April 25th, 2016):
  - "mixed model" → 629,000 results (0.46 seconds)
  - "robust statistics" → 197,000 results (0.40 seconds)
  - "AMMI" → 11,900 results (0.45 seconds)
    (Additive Main effect Multiplicative Interaction)
- The mixed models have been in many areas such as:
  - Biology, Education, Social and economics, Humanity, Physics and astronomy, Environment
- Next is several works in Bogor on mixed models which are promising but still in early stages.

Mixed Models in Practice (cont..)

- **Example 1.** (Notodiputro and Yahya, 2016)
- "The Gamma Mixed Model for Analyzing Length of Stay of Tuberculosis Patients at Gorontalo Hospital"
- Gorontalo is a province in eastern part of Indonesia and is considered having high incidence of tuberculosis.
- The length of stay of a patient is the response variable or the outcomes (Y measured in days).
- Usually distribution of duration (length of stay, Y) is skewed, and we may approximate using gamma distribution.
- The data was collected from a hospital in Gorontalo province and it consists of many predictors, namely room type, age category, sex, secondary diseases, rural or urban, doctor.
- The sample size is 612 patients, these are the patients during 2015.

Mixed Models in Practice (cont..)

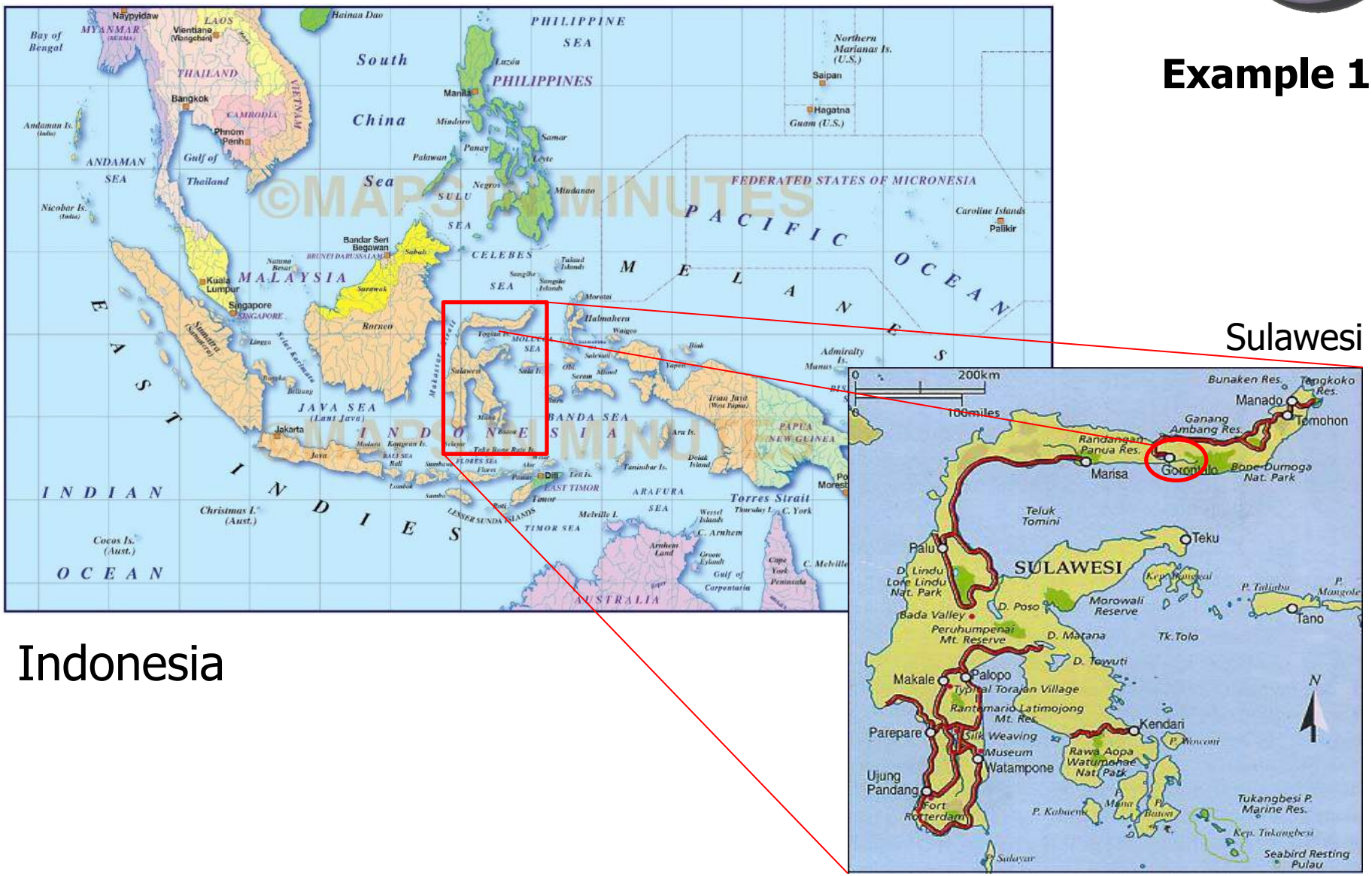
- **Example 1.** (Notodiputro and Yahya, 2016)
- "The Gamma Mixed Model for Analyzing Length of Stay of Tuberculosis Patients at Gorontalo Hospital"
- Response var.: length of stay in the hospital (y) assumed to have (conditional on random effects) gamma distributions.
- Explanatory var.: room type (p), age category (u), sex (s), secondary diseases (k), rural or urban (a), doctor (d)
- The linear predictor:

$$\eta_{ijklmno} = \mu + p_i + u_j + s_k + k_l + a_m + (pu)_{ij} + (ps)_{ik} + (pk)_{il} + (pa)_{im} + (us)_{jk} + (uk)_{jl} + (ua)_{jm} + (sk)_{kl} + (sa)_{km} + (ka)_{lm} + d_o$$

where the distribution of random effects $d \sim N(0, \sigma_\delta^2)$

Mixed Models in Practice (cont..)

Example 1.



Mixed Models in Practice (cont..)

- **Example 1.** (Notodiputro and Yahya, 2016)
- The anova table

Source	DF	F	p-value
Room type (*p*)	1	5.51	0.019
Sex (*s*)	1	0.01	0.908
Age category (*u*)	2	0.03	0.972
Secondary disease (*k*)	1	3.58	0.059
Rural or urban (*a*)	1	8.22	0.004
*p*s*	1	0.30	0.582
*p*u*	2	0.95	0.388
*p*k*	1	13.91	0.000
*p*a*	1	8.78	0.003
*s*u*	2	0.54	0.581
*s*k*	1	0.01	0.925
*s*a*	1	0.06	0.809
*u*k*	2	2.06	0.129
*u*a*	2	0.11	0.900
*k*a*	1	1.39	0.239

n = 612 patients

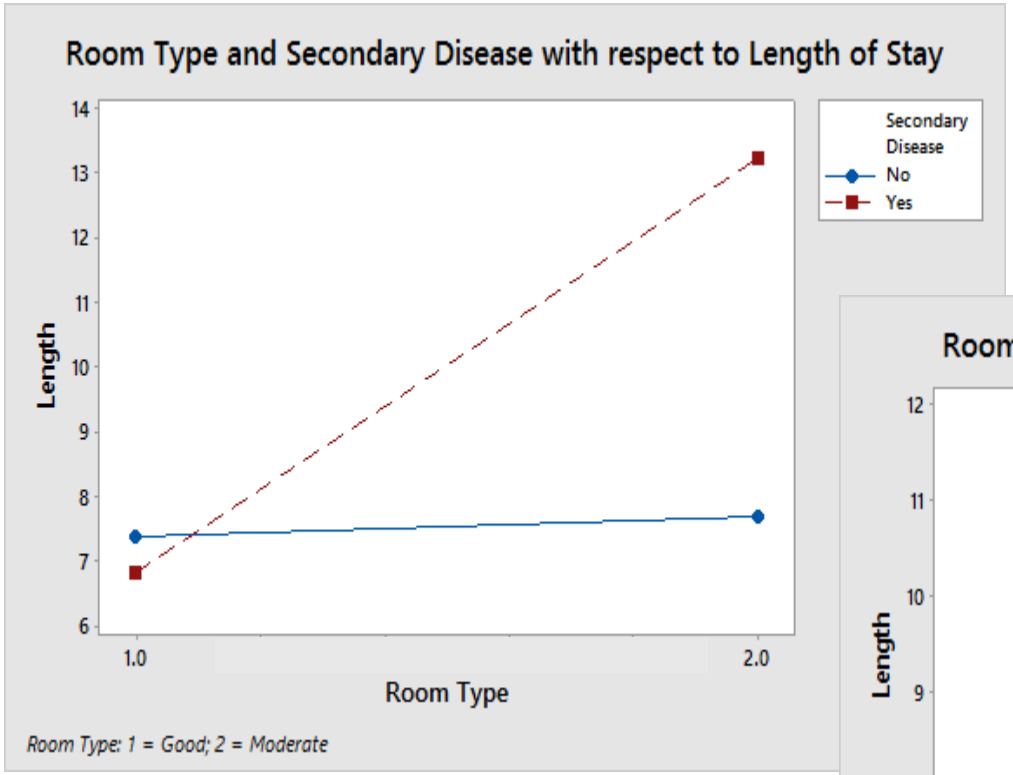
$s_\delta^2 = 0.00679$

$s_\varepsilon^2 = 2.52600$

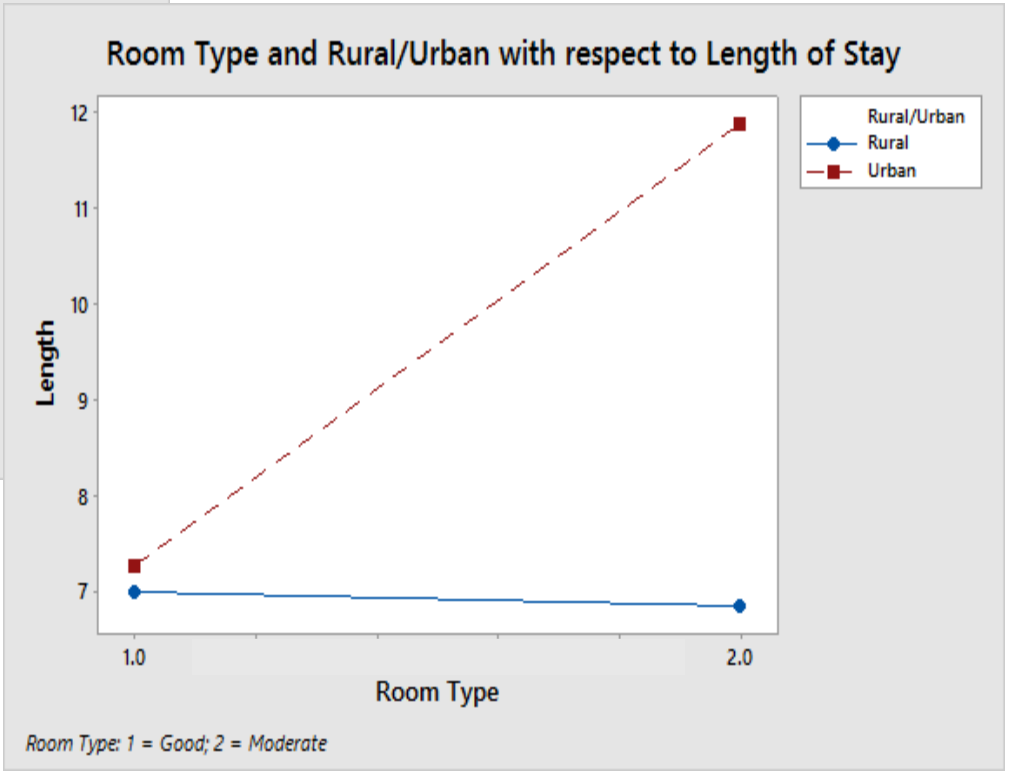
- Variation within doctor is smaller than variation between doctors
- Room type is significant: the patients in better room stay longer
- Interaction p*k and p*a are significant → further exploration is required

Mixed Models in Practice (cont..)

- **Example 1.** (Notodiputro and Yahya, 2016)



Interactions



Mixed Models in Practice (cont..)

- **Example 2.** (Notodiputro and Adabiyah, 2016)
- "Nested Linear Mixed Model and Parametric Stability Analysis for Multilocation Experiments of Shorgum Genotypes"

- 10 shorghum genotypes were evaluated based on their yields
- 3 sorghum were used as control
- Randomized block experiments were carried out in two seasons and 5 different locations within the seasons.

- Eventually, we want to know the performance of these genotypes and which of them that can produce stable yields.

Mixed Models in Practice (cont..)

- **Example 2.** (Notodiputro and Adabiyah, 2016)
- "Nested Linear Mixed Model and Parametric Stability Analysis for Multilocation Experiments of Shorgum Genotypes"
- Response var.: Shorgum yield (y) assumed to have (conditional on random effects) normal distributions
- Explanatory var.: genotypes (G), seasons (S), locations (L), blocks (B)
- The linear predictor:

$$\eta_{ijkl} = \mu + G_i + S_j + L_{k(j)} + (GS)_{ij} + (GL)_{ik(j)} + B_{l(jk)} + \varepsilon_{ijkl}$$

where the distribution of random effects $B \sim N(0, \sigma_B^2)$; $L \sim N(0, \sigma_L^2)$; and $GL \sim N(0, \sigma_{GL}^2)$

Mixed Models in Practice (cont..)

Indonesia

Locations:

- Yogyakarta
- Depok
- Wonosari
- Telukbetung
- Mataram
- Bogor
- Boyolali
- Tj.Karang
- Pekanbaru

Mixed Models in Practice (cont..)

- **Example 2.** (Notodiputro and Adabiyah, 2016)
- The anova table

13x3x5x2 = 390 experimental units

Random effects

Random Effects	Var.est	Std eror	Z	p-value
Location (Season)	0.2248	0.1203	1.8700	0.0309
Genotype*Location (Season)	0.0882	0.0316	2.7900	0.0026
Block (Season*Location)	0.0000	.	.	.
Error	0.3554			

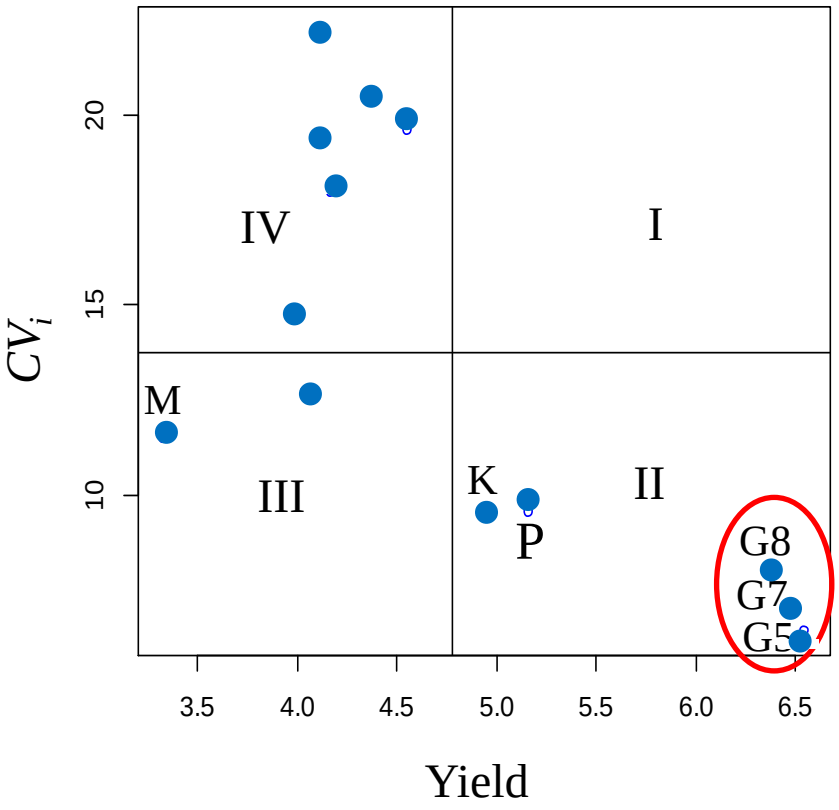
Fixed effects

Fixed Effects	DF for numerator	DF for denominator	F	P-value
Genotype	12	96	54.4700	<.0001
Season	1	8	0.3100	0.5923
Genotype*Season	12	96	1.3500	0.2027

Mixed Models in Practice (cont..)

- **Example 2.** (Notodiputro and Adabiyah, 2016)

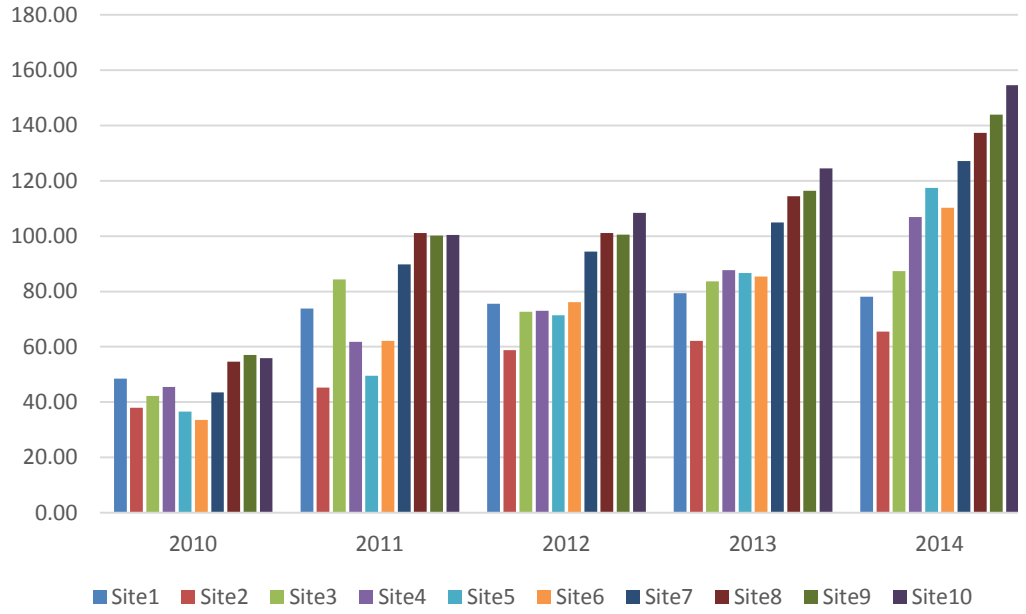
Genotypes	Yield		
	CV_i	W_i^2	d_i^2
G1	19.5814	21.3965	10.5619
G2	20.4037	21.3650	8.8558
G3	22.2235	25.1734	12.2653
G4	17.9484	23.1228	7.1587
G5	6.4129	14.4406	4.1602
G6	14.6978	24.7314	4.6860
G7	6.9182	11.8964	3.1412
G8	8.0645	12.1867	4.5061
G9	19.2566	19.7342	7.4759
G10	12.7184	29.7786	5.3025
Kawali (K)	9.5080	15.3093	4.6290
Mandau (M)	11.4640	27.5873	2.0947
Pahat (P)	9.5229	14.9245	3.4744

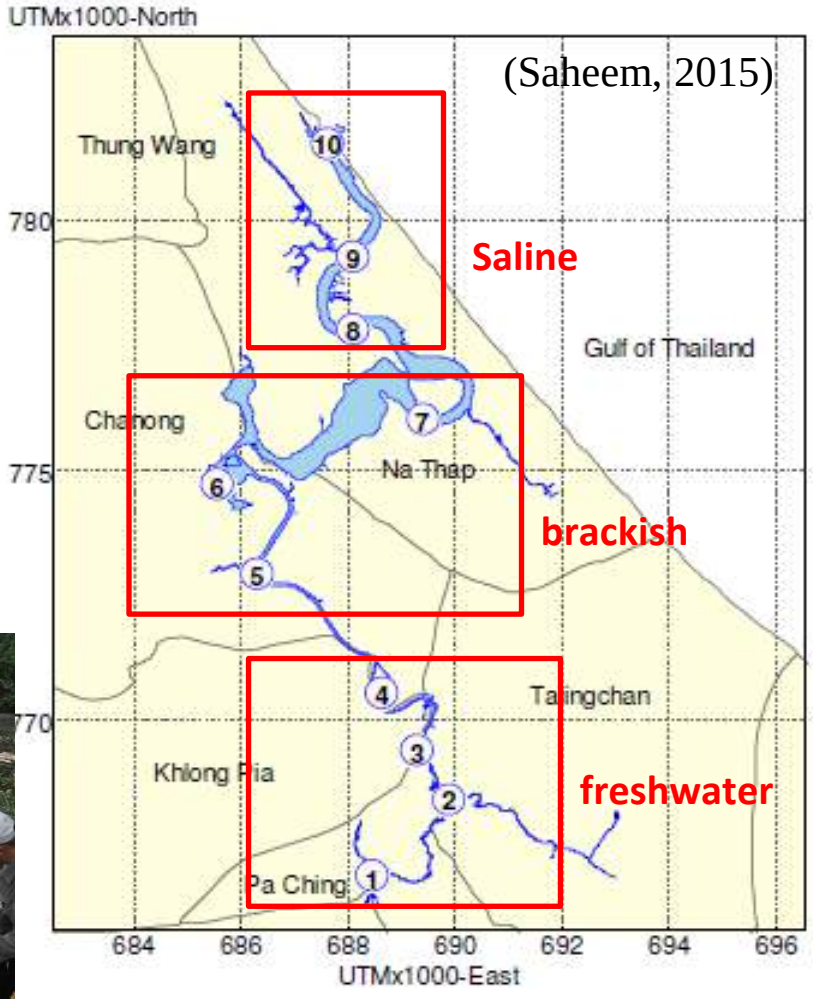


CV_i is coefficient of variation; W_i^2 is ecovalence; and d_i^2 is genotypic stability

Mixed Models in Practice (cont..)

- **Example 3.** (Angraini and Notodiputro, 2016)





IPB - PSU Collaborative Research RPM ID16287; The Na Thap river Project

Mixed Models in Practice (cont..)

- **Example 3.** (Angraini and Notodiputro, 2016)
- "*Generalized Linear Mixed Models for Analyzing Fish Stock at Na Thap River*"
- Response var.: Fish stock (*y*) assumed to have (conditional on random effects) Gamma distributions
- Explanatory var.: Zone (*Z*), WDEPTH (*W*), SAL (*L*), DO (*D*) BOD (*B*), and Site (*S*).
- The linear predictor:

$$\eta_{ij} = \mu + \beta_1 Z_i + \beta_2 W_i + \beta_3 L_i + \beta_4 D_i + \beta_5 B_i + S_j$$

where the distribution of random effects $S_i \sim N(0, \sigma_S^2)$;

- Or, $\boldsymbol{\eta = X\beta + Zs}$ and $\boldsymbol{s \sim N(0,G)}$
- *Link function: log*

Mixed Models in Practice (cont..)

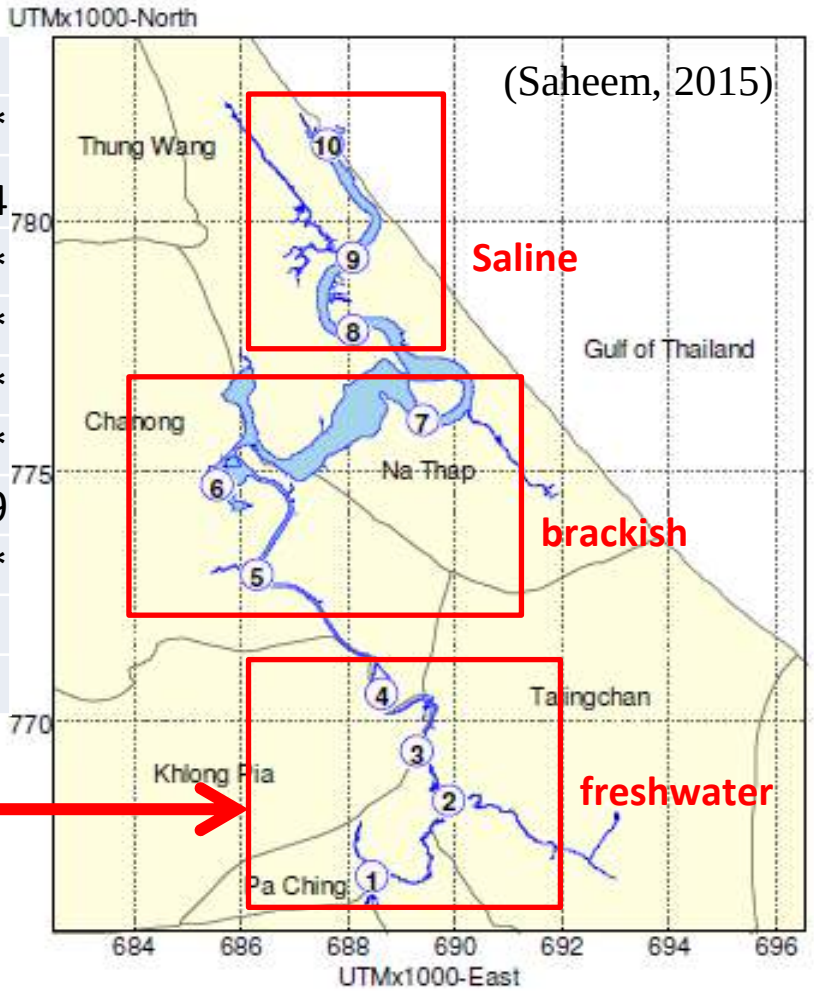
- **Example 3.** (Angraini and Notodiputro, 2016)

	Estimates	Std.error	t	Pr(>|t|)
Intersep	4.549	0.164	27.740	0.000*
Zone2	-0.008	0.030	-0.287	0.774
Zone3	0.282	0.034	8.224	0.000*
WTEMP	-0.026	0.005	-5.381	0.000*
WDEPTH	-0.039	0.006	-6.300	0.000*
SAL	0.018	0.001	20.552	0.000*
DO	0.015	0.009	1.603	0.109
BOD	0.028	0.008	3.267	0.001*
σ^2 site	0.320			
σ^2 error	0.084			

Variation among sites was evident

???

n = 3741



Mixed Models in Practice (cont..)

- **Example 3.** (Angraini and Notodiputro, 2016)
- The pairwise confidence intervals:

Site	Substracted from Site		
	2	3	4
1	(-7909,4741)	(-774,11876)	(-8336,4315)
2		(810,13460)	(-6751,5899)
3			(-13886,-1236)

- The power plant is in site 3
- Site 2 vs 3 is significant (site 2 > site 3)
- Site 3 vs 4 is siginificant (site 4 > site 3)

Mixed Models in Practice (cont..)

- **Example 4.** (Arisanti and Notodiputro, 2016)
- "*Bias Reduction in Estimating Variance Components of Phytoplankton Abundance at Na Thap River based on Logistic Linear Mixed Models*"
- Response var.: Phytoplankton abundance (y) assumed to have (conditional on random effects) binomial distributions
- Explanatory var.: SAL (L), DO (D) BOD (B), and Site (S).
- The linear predictor:

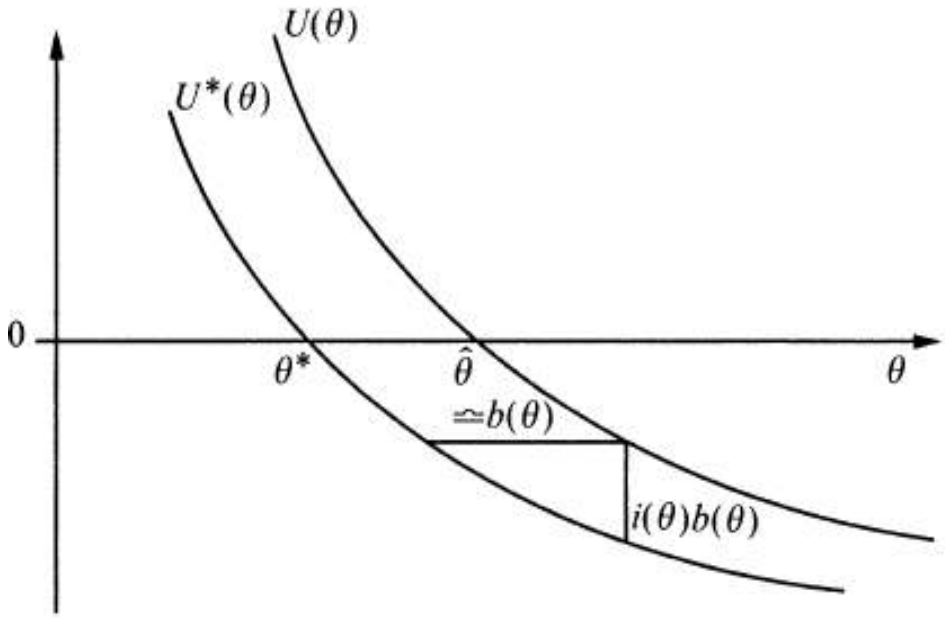
$$\eta_{ij} = \mu + \beta_3 L_i + \beta_4 D_i + \beta_5 B_i + S_j$$

where the distribution of random effects $S_i \sim N(0, \sigma_S^2)$;

- Or, $\boldsymbol{\eta = X\beta + Zs}$ and $\boldsymbol{s \sim N(0,G)}$
- Link function: logistics

Mixed Models in Practice (cont..)

- **Example 4.** (Arisanti and Notodiputro, 2016)
- It is well known that variance estimates of MLE are biased
- We follow the idea of Firth (1993) to adjust variance components in GLMM with logistic link function



If $U(\theta)$ is the score function in ML estimation, and $\hat{\theta}$ is subject to a positive bias $b(\theta)$, the score function is shifted downward at each point θ by amount $i(\theta)b(\theta)$, where $-i(\theta)=U'(\theta)$ is the local gradient, then the adjusted score function:

$$U^*(\theta) = U(\theta) - i(\theta)b(\theta)$$

or

$$\boldsymbol{U^*(\theta) = U(\theta) + A(\theta)}$$

Mixed Models in Practice (cont..)

- **Example 4.** (Arisanti and Notodiputro, 2016)
- The adjusted score function:

$$U^*(\theta) = U(\theta) - i(\theta)b(\theta) \text{ or } \boldsymbol{U^*(\theta)} = \boldsymbol{U(\theta)} + \boldsymbol{A(\theta)}$$

- The solution to this adjusted score function is $\boldsymbol{\theta^*}$
- Firth adjustment is based on **F** and **H** are Fisher's information and Hessian matrix

$$\boldsymbol{A_{\theta j}} = -\frac{1}{2} tr\left(\boldsymbol{F}^{-1} E\left[U_{\theta j}(\boldsymbol{H} - \boldsymbol{UU'})\right]\right)$$

Mixed Models in Practice (cont..)

- **Example 4.** (Arisanti and Notodiputro, 2016)
- The ML estimate of variance component of site:

$$\hat{\sigma}^2\text{site} = 0.986$$

n = 466 (Synedra)

- The adjusted estimate is
- $$\tilde{\sigma}^2\text{site} = 0.735$$
- The variance component has been reduced by 26%
- The analysis of variance table:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.0288	0.3682	5.511	0.0000
DO	-0.3153	0.0775	-4.067	0.0000
BOD	0.1325	0.0998	1.327	0.1840
Salinitas	-0.0108	0.0106	-1.021	0.3070

Mixed Models in Practice (cont..)

- **Example 5.** (Angraini and Notodiputro, 2016)
- *"A Hierarchical Approach to Generalized Linear Mixed Model for Analyzing Fish Species Abundance at Na Thap River"*
- Response var.: Fish species abundance (y) assumed to have (conditional on random effects) Poisson (λ) distributions.
- Explanatory var.: SAL (L), DO (D) BOD (B), and Site (S).
- The linear predictor:

$$\eta_{ijk} = \mu + \beta_1 Z_i + \beta_2 W_j + \beta_3 L_j + \beta_4 D_j + \beta_5 B_j + S_k$$

where the distribution of random effects $S_k \sim Gamma(\alpha)$

- Link function: log
- The hierarchical likelihood

$$L(y|\lambda)+L(\alpha;u)= \sum(y_{ijk} ln\lambda_{ij} - \lambda_{ij}) + \sum\{\alpha\, lnu_{ij} + \alpha\, ln\,\alpha - \alpha u_{ij} - \alpha\, ln\,\Gamma(\alpha)\}$$

Mixed Models in Practice (cont..)

- **Example 5.** (Angraini and Notodiputro, 2016)

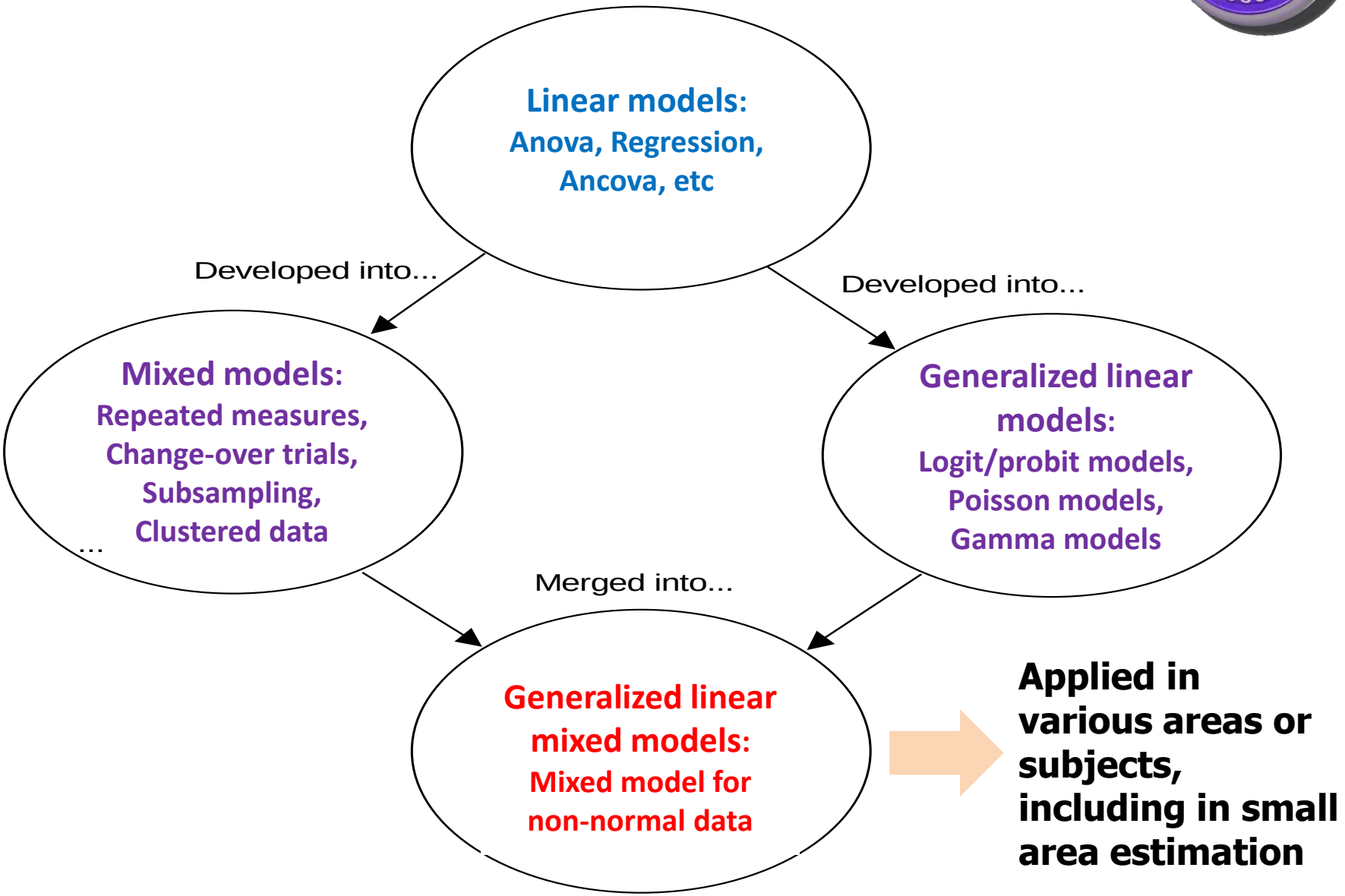
	Hierarchical Model			
	Estimate	Std.error	T	Pr(>|t|)
Intersep	3.021	0.180	16.826	0.000
Zonefreshwater	-0.242	0.156	-1.558	0.120
Zonesaline	0.344	0.167	2.056	0.040
WTEMP	0.002	0.004	0.552	0.581
WDEPTH	-0.010	0.010	-0.945	0.345
SAL	0.006	0.001	8.004	0.000
DO	0.026	0.007	3.507	0.000
BOD	0.025	0.007	3.619	0.000
σ^2Site	0.709			
σ^2Sisaan	0.041			

n = 531

Variation among sites was evident

Concluding Remarks

Concluding Remarks



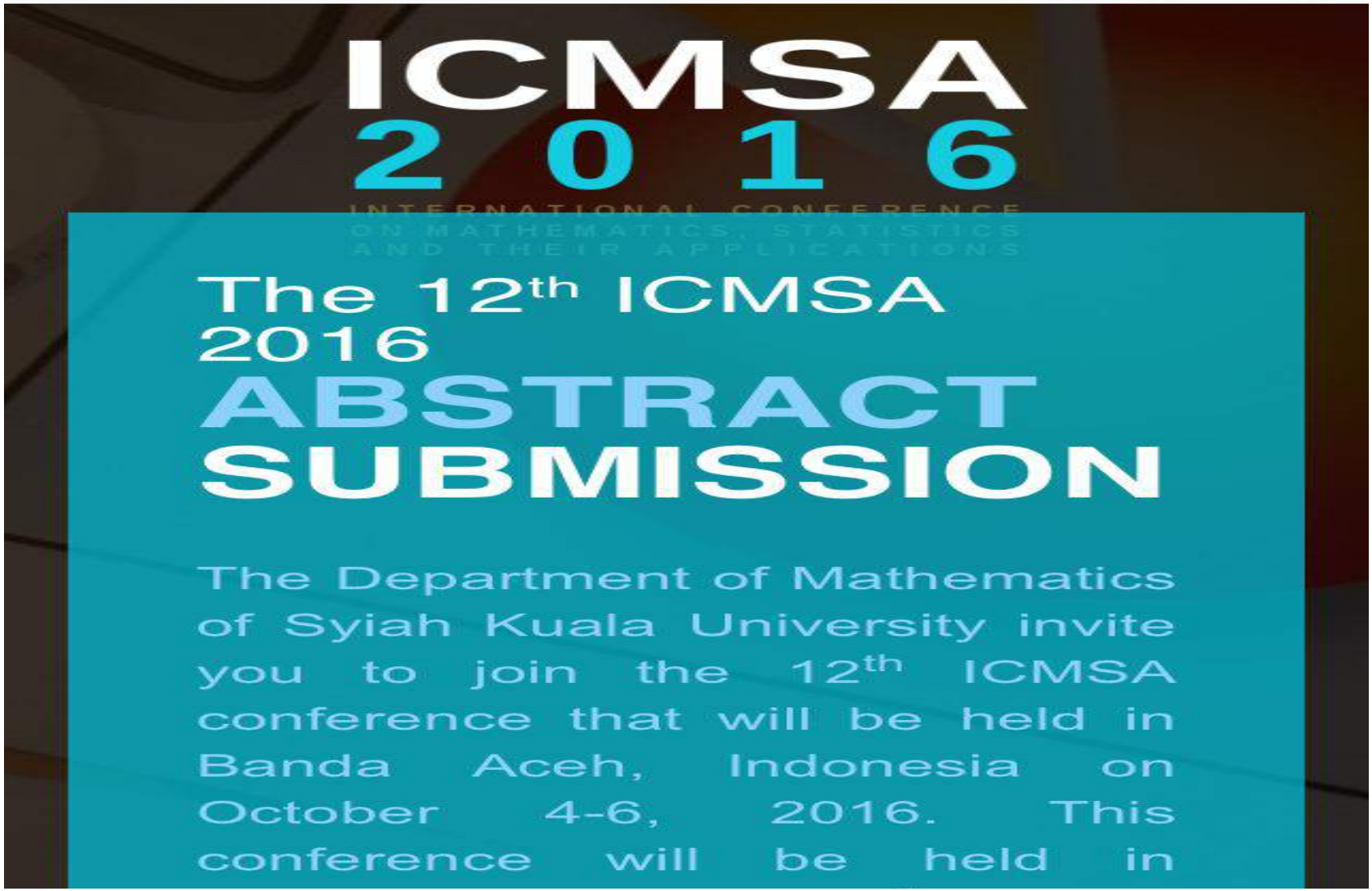
References

1. Firth, D. 1993. *Bias reduction of maximum likelihood estimation. Biometrika,* Vol. 80 No. 1. pp 27-38
2. Jiang, J. 2007. *Linear and Generalized Linear Mixed Models and Their Applications*. Springer.
3. Lee, Y. and Nelder. J.A. 1996. Hierarchical generalized linear models. *Journal of the Royal Statistical Society*. Series B, (Methodological), Vol. 58, No. 4(1996), pp. 619-678
4. Mehari M., Alamerew S., Lakew B., Yirga H., Tesfay M. 2014. Parametric stability analysis of malt barley genotypes for grain yield in Tigray, Ethiopia. *World Journal of Agricultural Sciences*. 10(5): 210-215
5. Saheem, N. (2015). *Statistical Modeling of Aquatic Animal Abundance in the Na Thap River*. Thesis. Prince of Songkla Univ.
6. Stroup, W. W. 2013. *Generalized Linear Mixed Models: Modern Concepts, Methods, and Applications.* Chapman & Hall



Thank You!


Syiah Kuala University conference in Aceh



http://icmsa.unsyiah.ac.id