Fitting linear mixed models under misspecification

Jin Yoon* Alan Welsh†

Statistical Consulting Unit*
The Australian National University

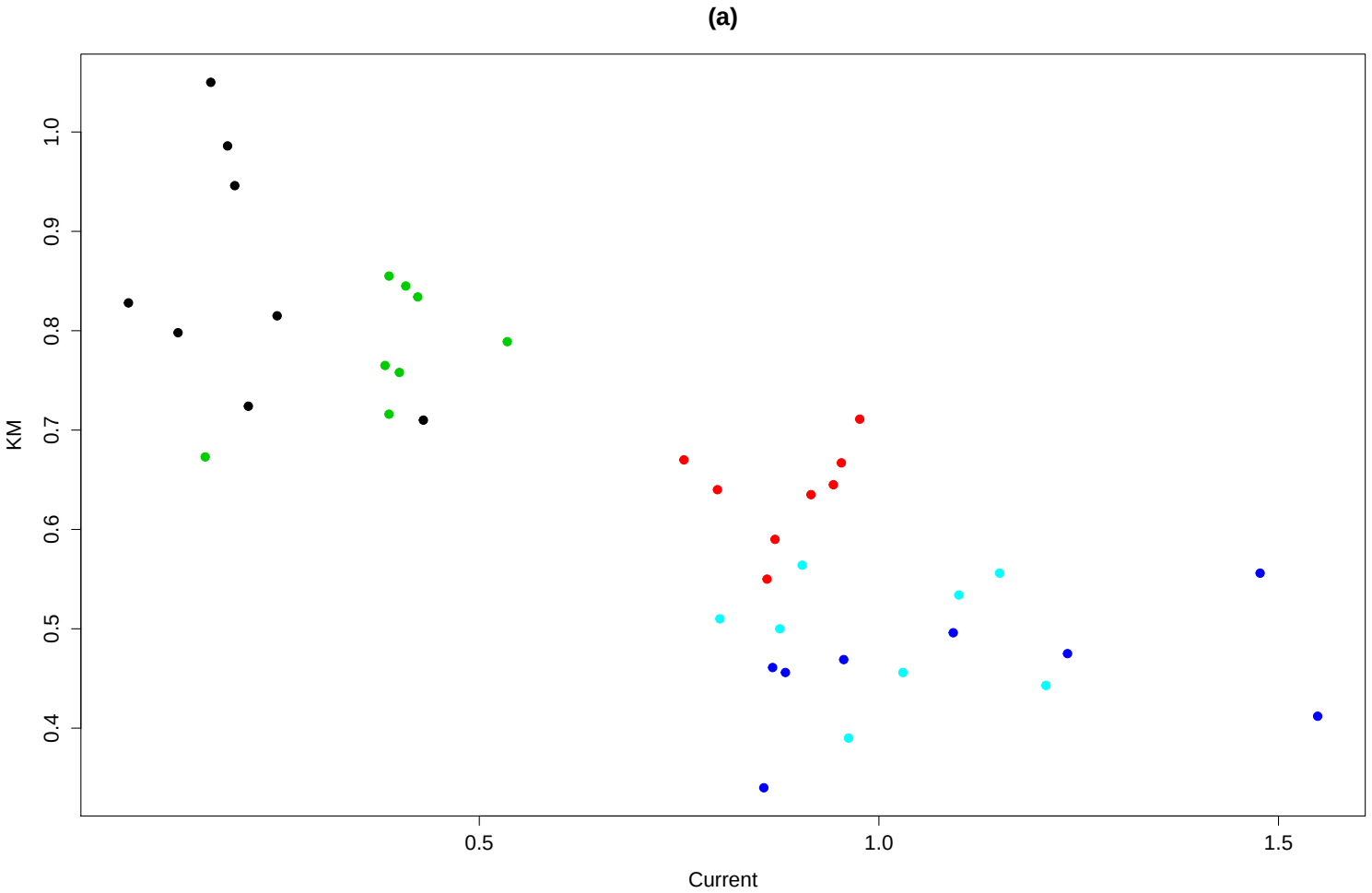
Mathematical Sciences Institute†
The Australian National University

December 2015

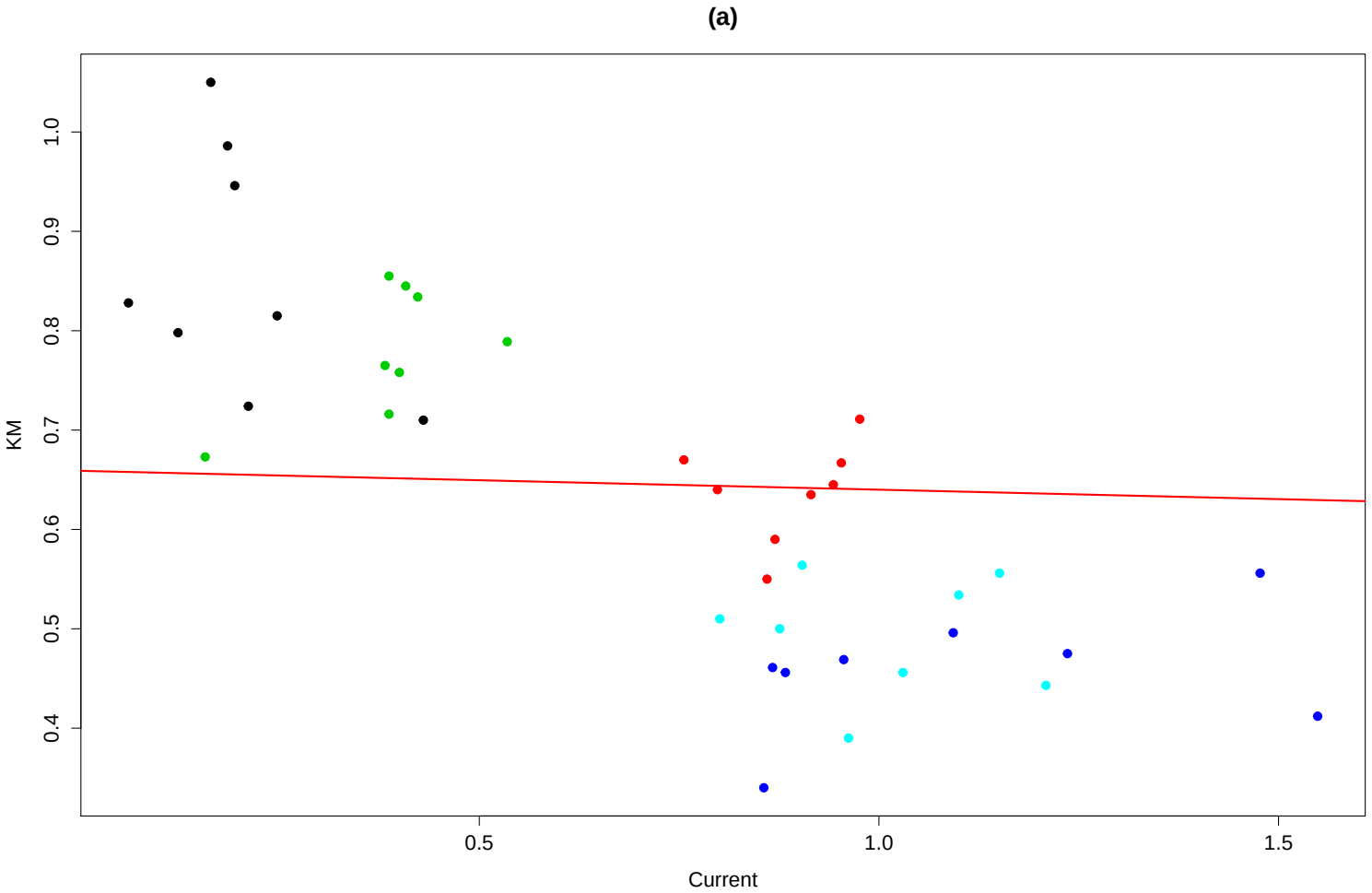
How much believe in LMM ?

- Linear mixed model (LMM) is one of the most popular statistical methods and being used without a doubt
- What if LMM does not give right estimates ?

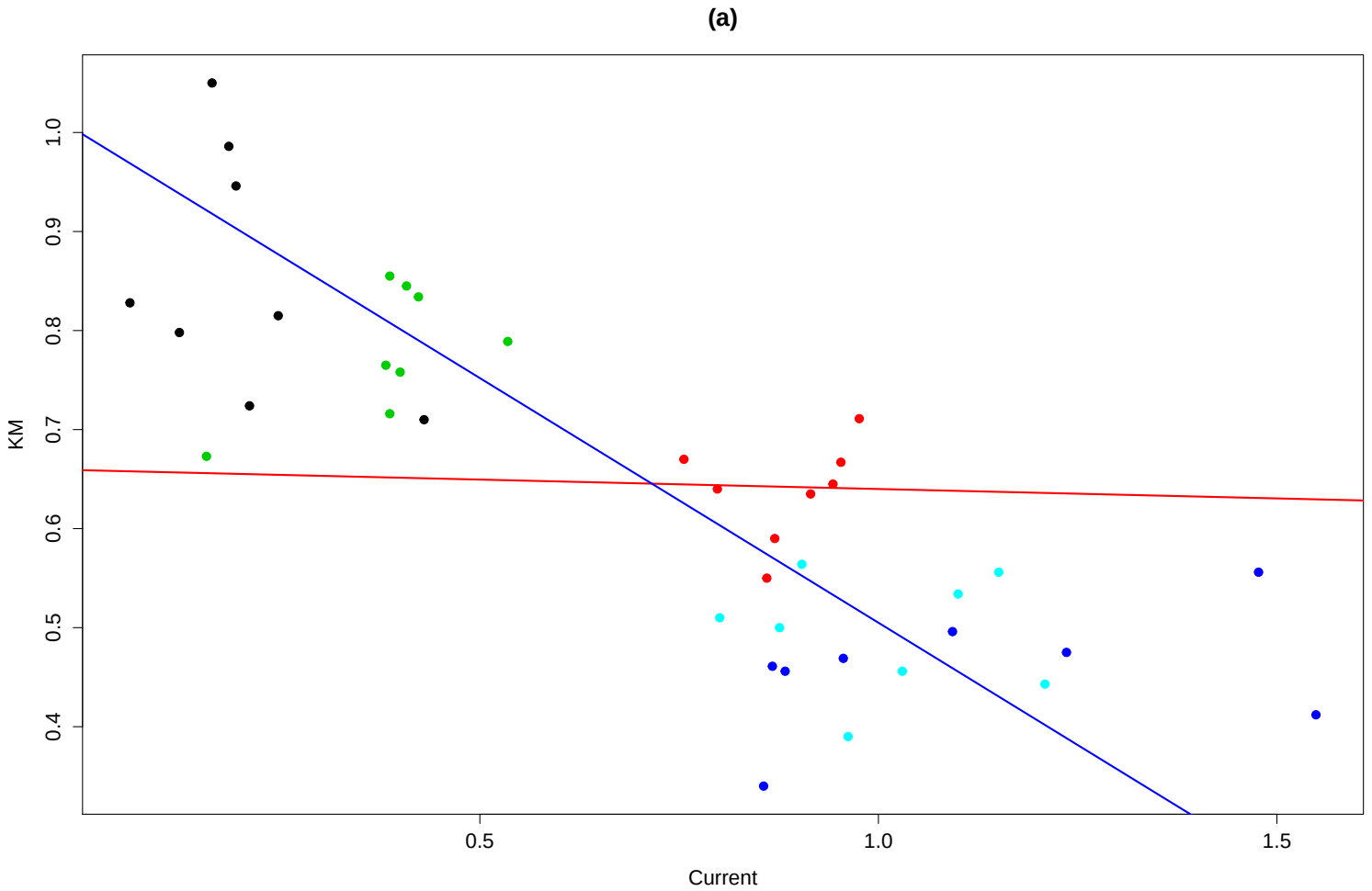
Examples



Examples



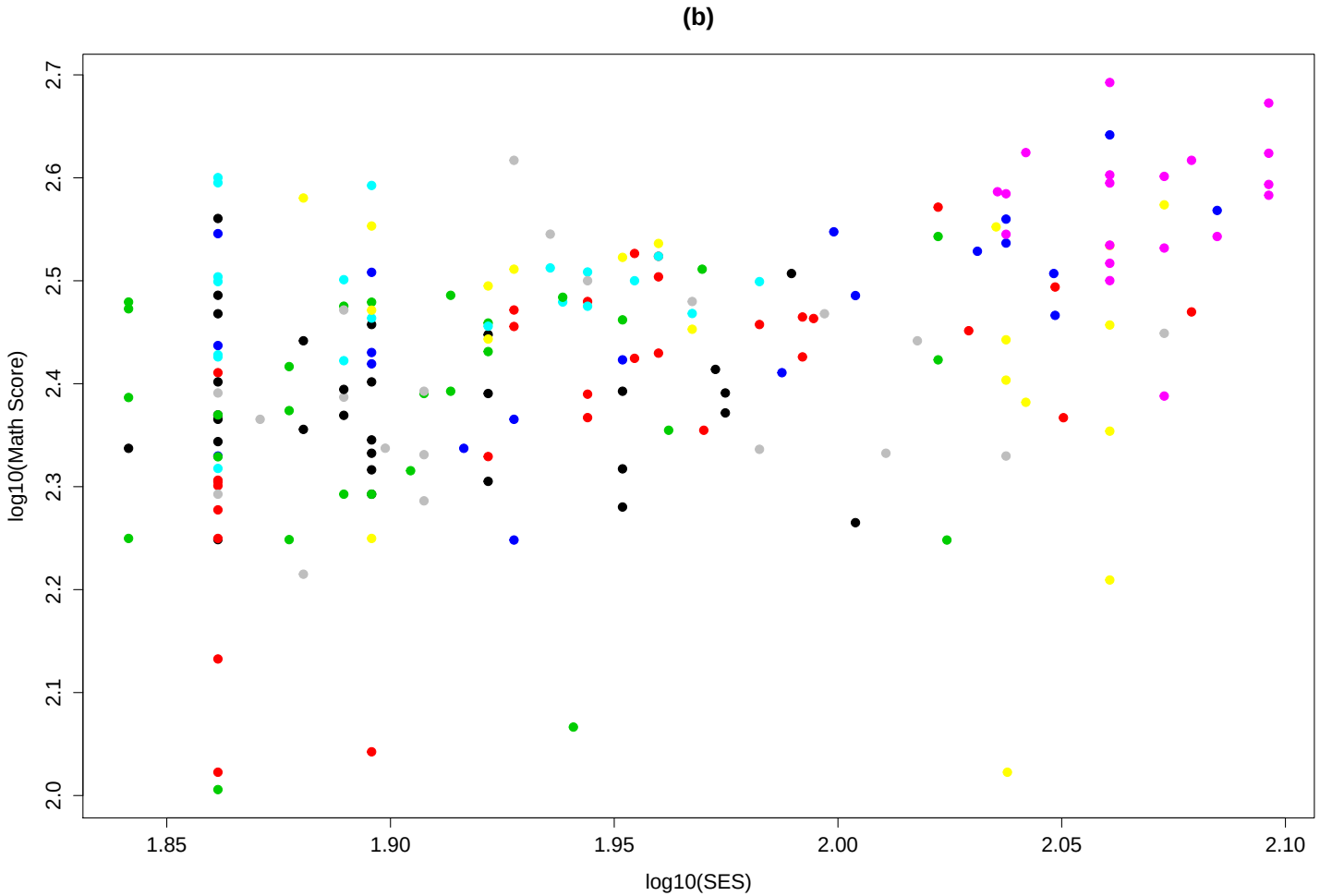
Examples



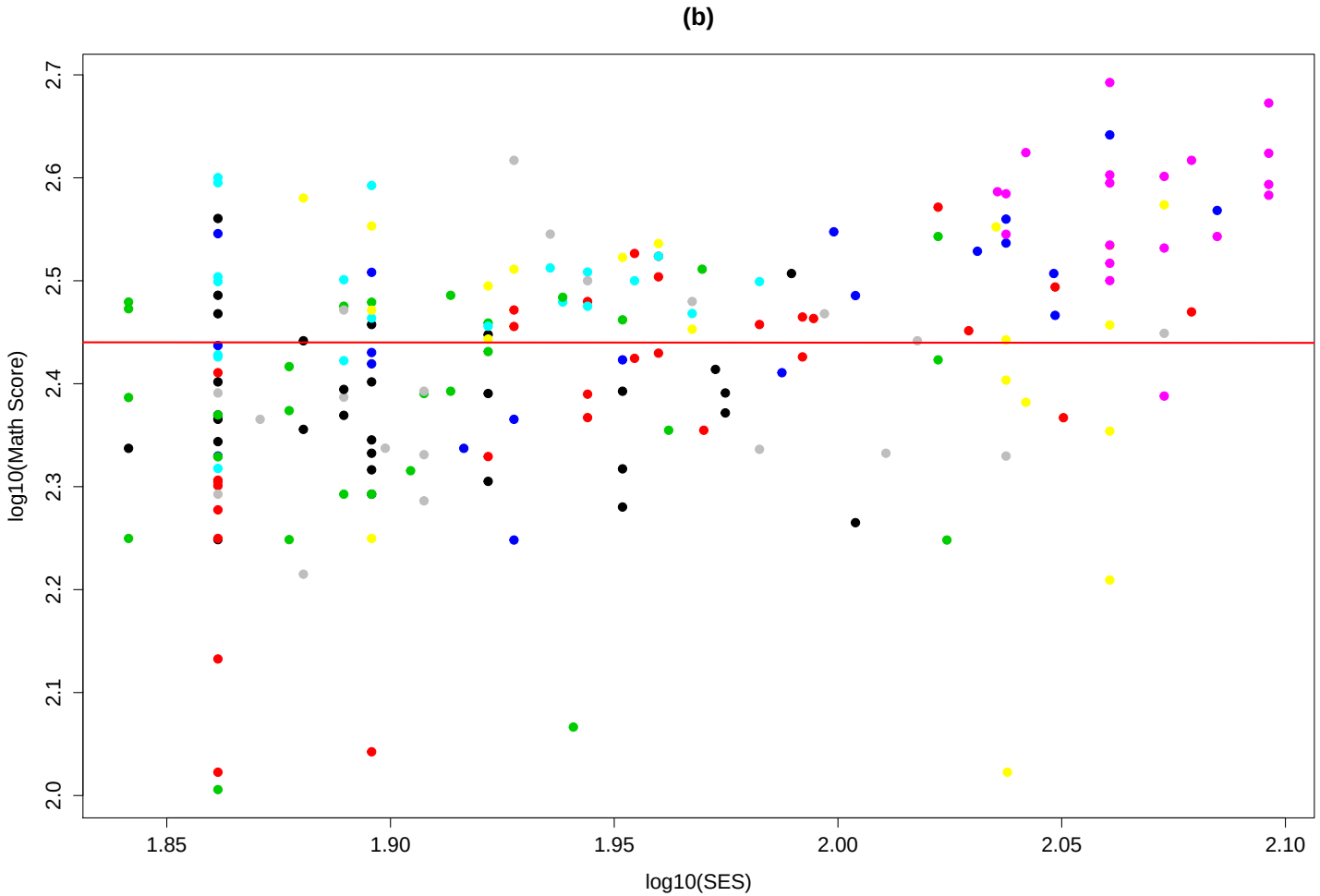
Examples (a)

- Enzyme kinetics data
- $n = 8$ measurements on four mutant mice and one wild type mouse ($m = 5$) of kinetics (Y) and the current (X)
- REML based on lme4 in R and REML on our method
- Between cluster variance of the current (X) is $\hat{\tau}_x = 0.397^2$

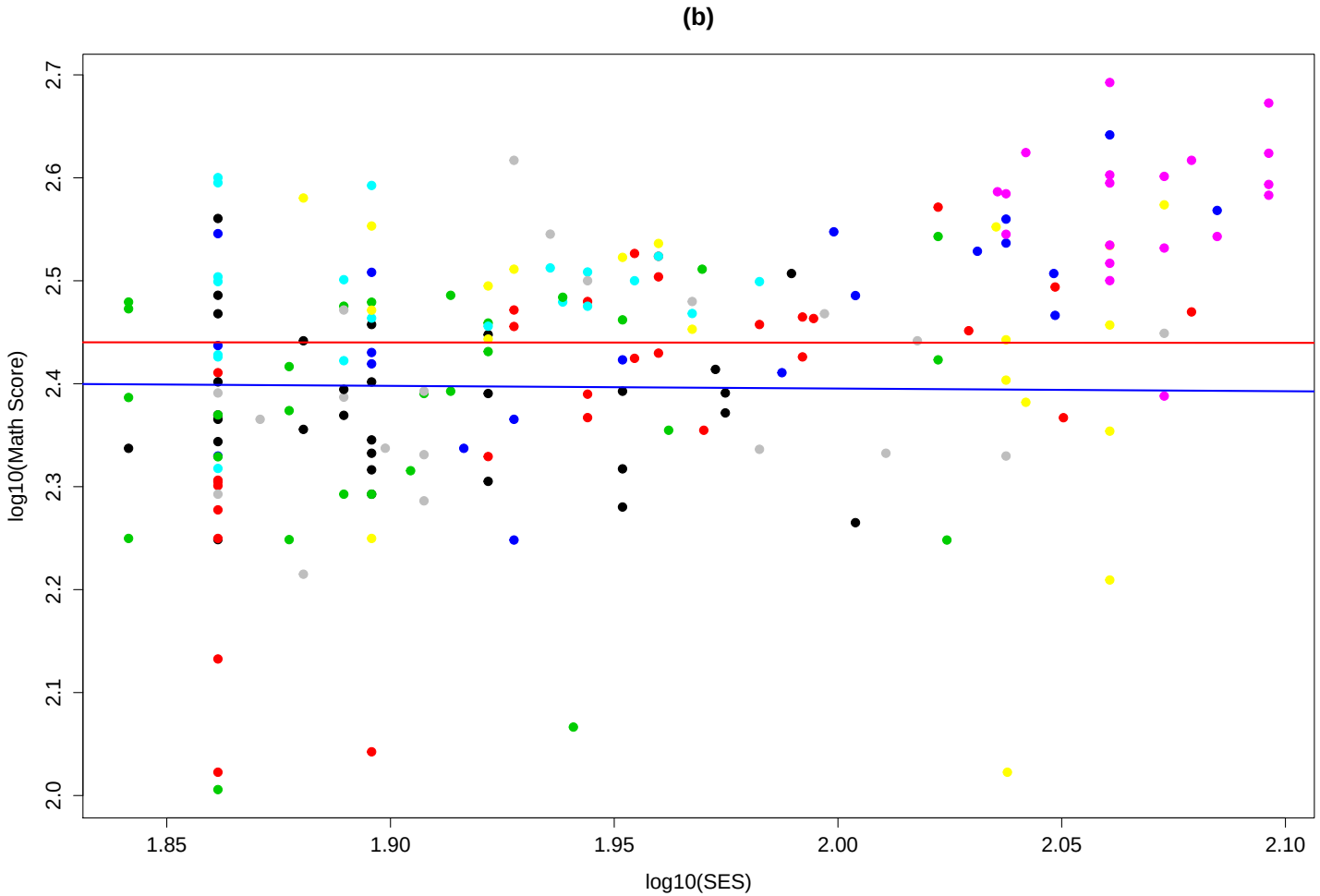
Examples


(b)
log10(Math Score)
2.0
2.1
2.2
2.3
2.4
2.5
2.6
2.7
1.85
1.90
1.95
2.00
2.05
2.10
log10(SES)

Examples


(b)
log10(Math Score)
2.0
2.1
2.2
2.3
2.4
2.5
2.6
2.7
1.85
1.90
1.95
2.00
2.05
2.10
log10(SES)

Examples


(b)
log10(Math Score)
2.0
2.1
2.2
2.3
2.4
2.5
2.6
2.7
1.85
1.90
1.95
2.00
2.05
2.10
log10(SES)

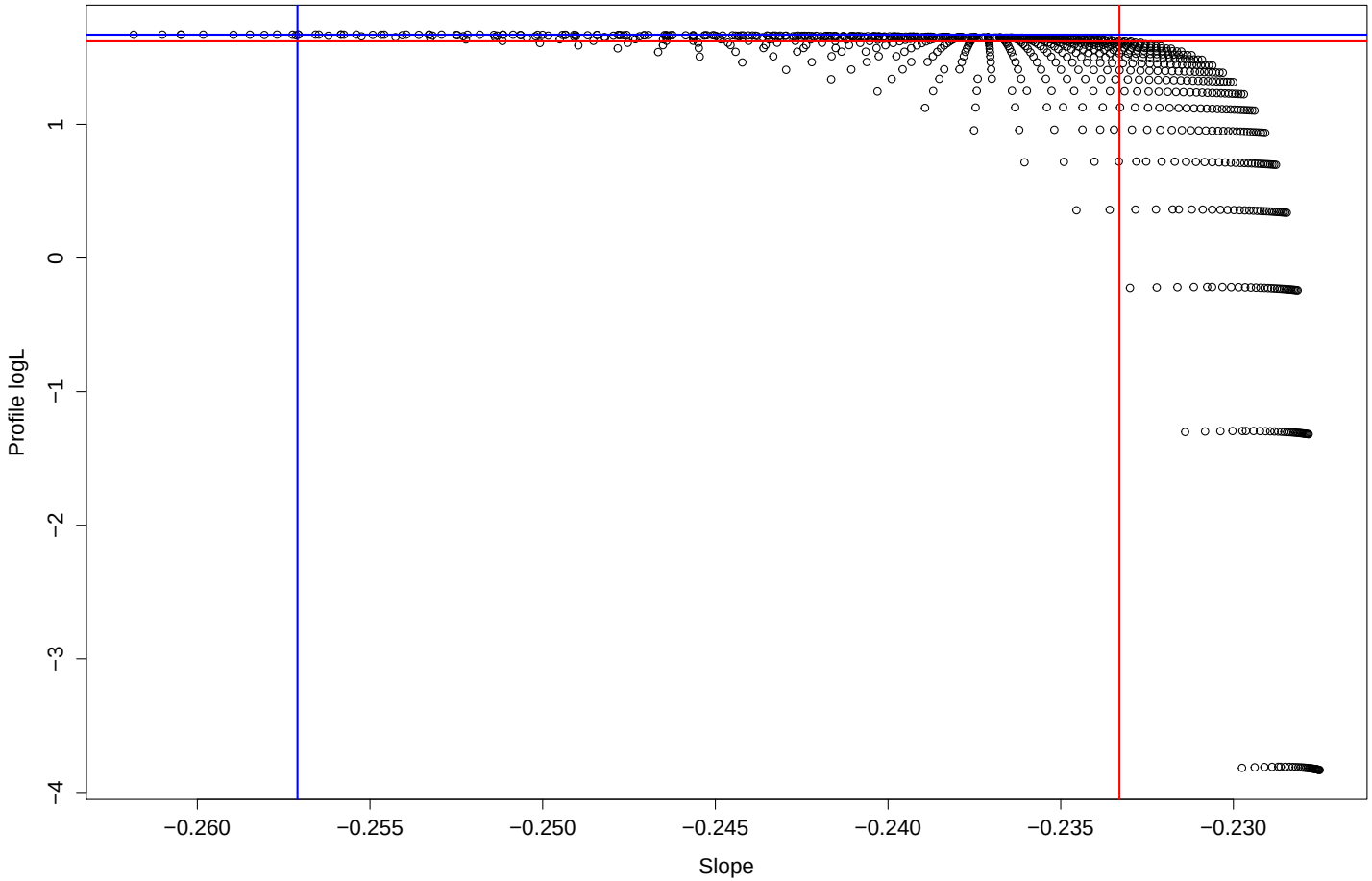
Examples (b)

- Sixth Grade data
- $n = 10$ measurements on $m = 19$ different schools of mathematics score (Y) and the socio-economic status (X)
- REML based on lme4 in R and REML on our method
- Between cluster variance of the socio-economic status (X) is $\hat{\tau}_x = 14.645^2$

Examples cont.

- Why ? $\rightarrow$ fit misspecified model: unintentionally, because the true model is unknown and intentionally in either informal model exploration or formal model selection

Profile log-likelihood (a)



Linear mixed model

- Classic LMM

$$Y_{ij} = \beta_0 + \beta_c X_{ij} + \delta_i + \epsilon_{ij},$$

where $(\beta_0,\ \beta_c)$ the unknown regression parameters and $(\delta_i,\ \epsilon_{ij})$ an unobserved random intercept and random error.

LMM cont.

- Contextual LMM ($\tilde{X}_{ij} = X_{ij} - \bar{X}_i$)

$$Y_{ij} = \beta_0^* + \beta_w \tilde{X}_{ij} + \beta_b \bar{X}_i + \delta_i^* + \epsilon_{ij}^*,$$

where (β_w, β_b) the within- and between-group regression parameters and (δ_i^*, ϵ_{ij}^*) an unobserved random intercept and random error.

Misspecification

- When $\beta_b - \beta_w \neq 0$, LMM is missepecified due to omitting the group level confounder $\bar{x}_i$ that is, clustering in x_i is ignored.
- Affects the estimates of the variance components not just, the estimates of the regression parameters

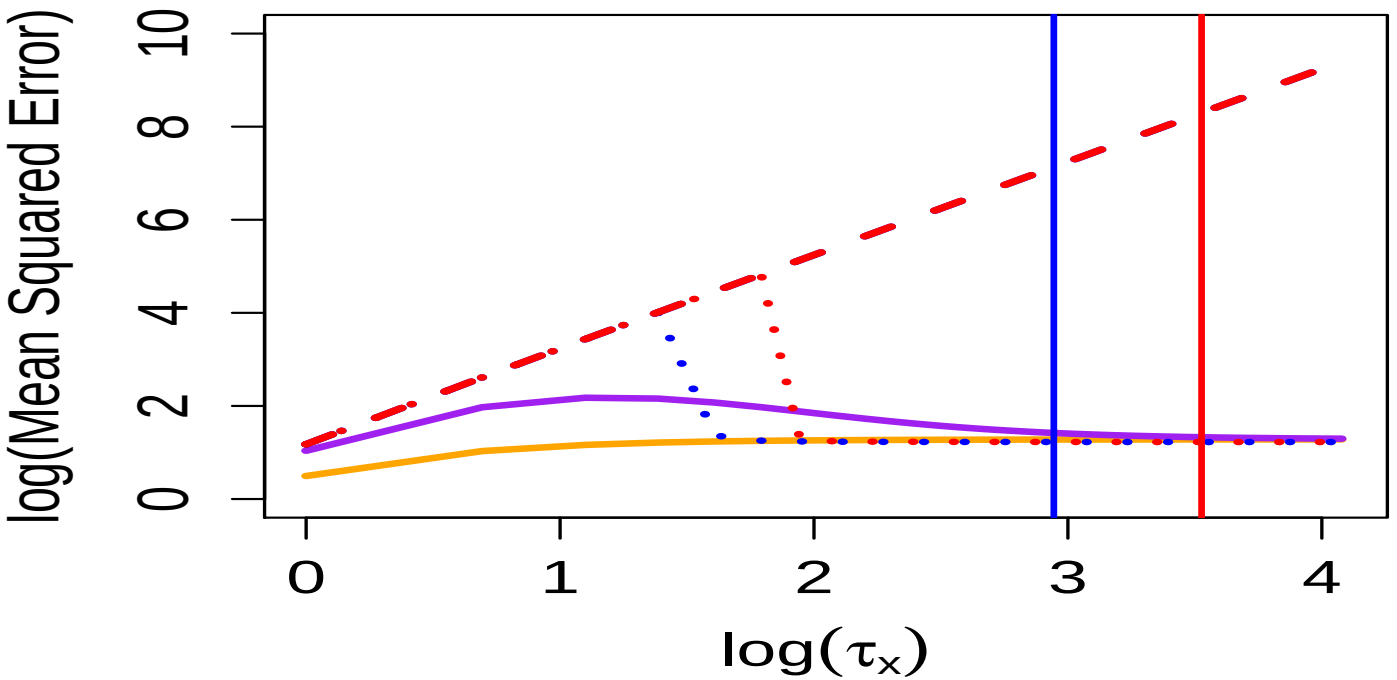
Result

- Fitting LMM when CLMM is true model can lead to very misleading assessments of the association between Y and X
- Change LS and WLS estimators of the regression parameters smoothly as changing between cluster variance of X, τ_X
- but jump ML and REML estimators of the regression and variance components.

Result cont.

- The reason for the jumps in ML and REML estimators: as $\tau_x \uparrow$, the ML and REML criterion develop two distinct local maxima and which of these is the global maximum changes at the jump point.

Mean squared errors



Thank you !