Simon Says Stats
"First we take Manhattan, then we take Berlin"
Adjusting for site-effects (rejoinder)

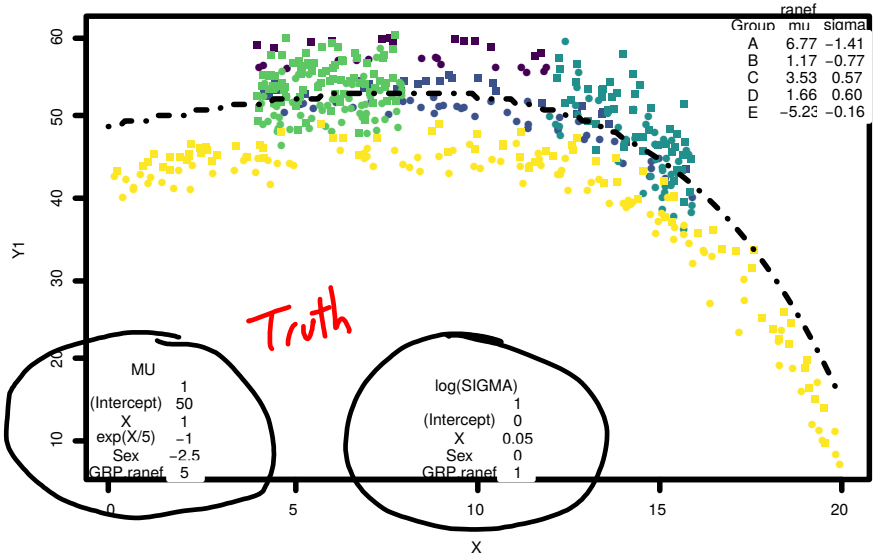
Simon White

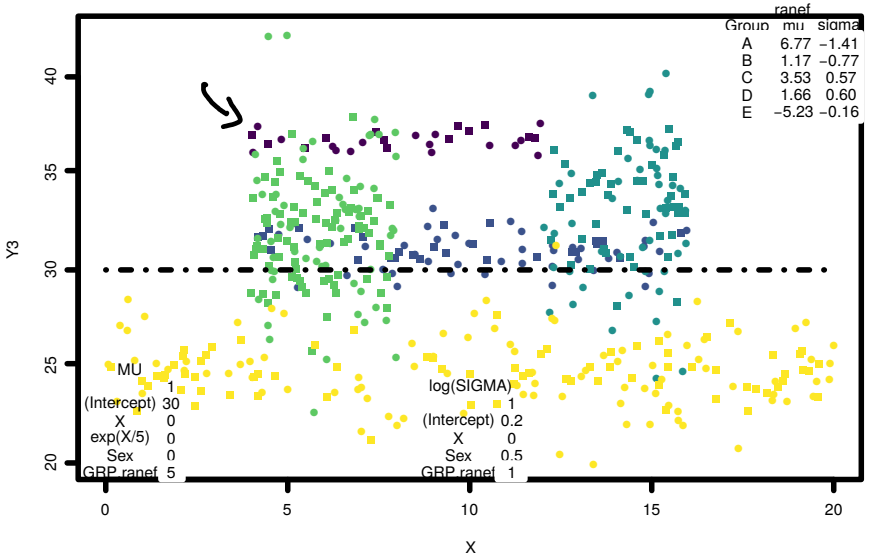
Department of Psychiatry, University of Cambridge
MRC Biostatistics Unit, University of Cambridge

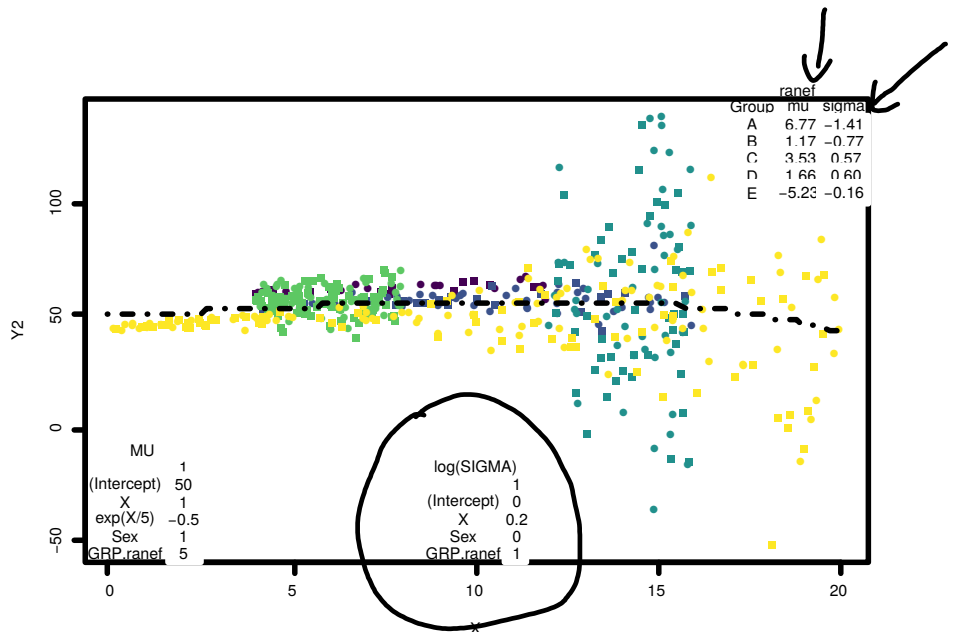
2020/Jul/21
Ed Group Meeting

Our example with three ROIs (truth via GAMLSS)

Data (with true female–population level curve plotted)





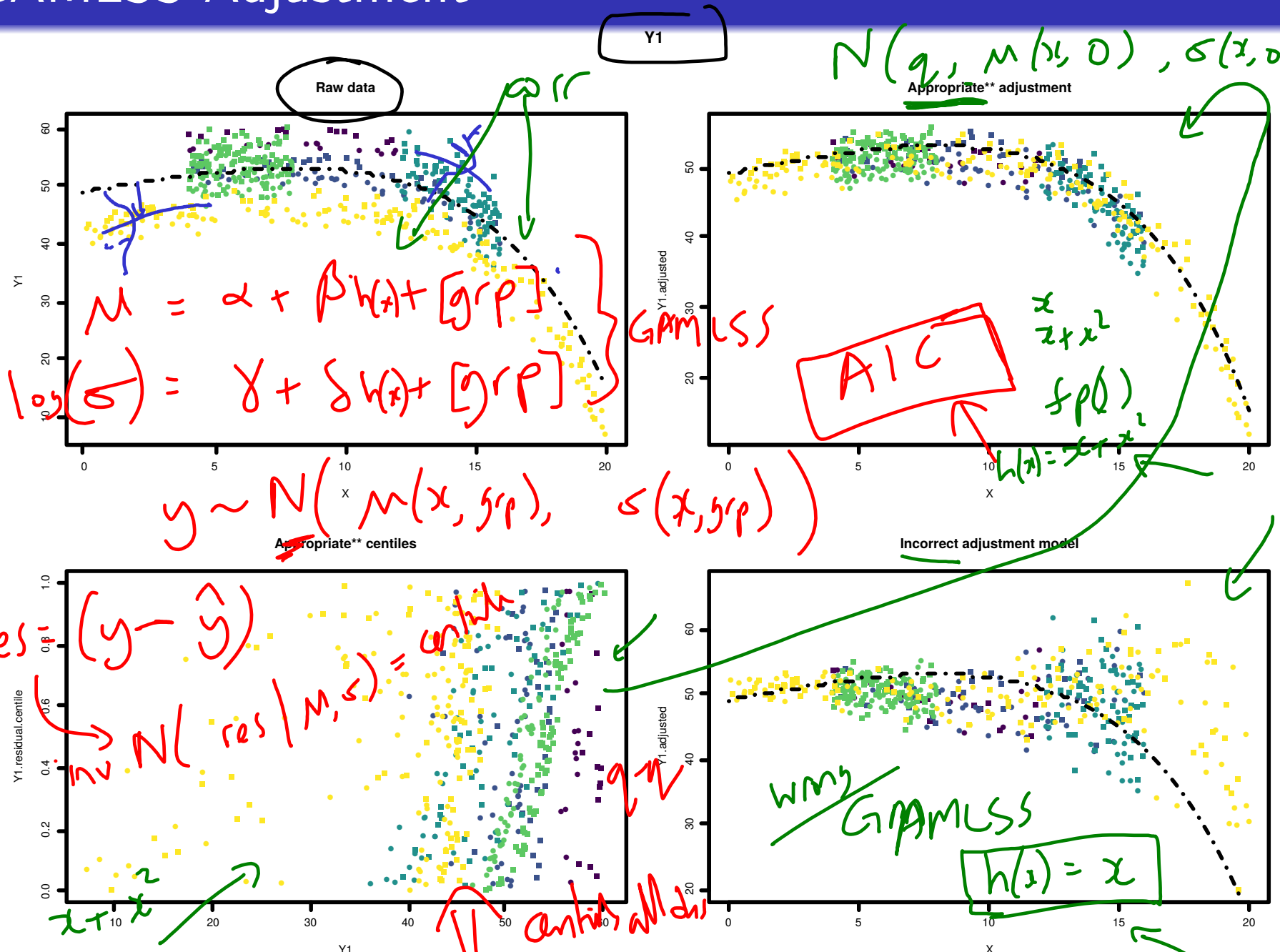


3 ROIs

$\mu \sim 1 + X + \exp(\frac{X}{5}) + Sex + (1|Grp)$

$\log(\sigma) \sim 1 + X + Sex + (1|Grp)$

GAMLSS Adjustment



Obtaining "adjusted" y from GAMLSS

① fit GAMLSS model (for each ROI independently)

↳ likely need some model selection (AIC?)

→ obtain fit, eg. $\hat{\mu} = \hat{\alpha} + \hat{\beta} X + \hat{\gamma} \mathbb{1}_{sex} + \begin{Bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_K \end{Bmatrix}$ grp coef

$\widehat{\log(\sigma)} = \hat{a} + \hat{b} X + \hat{c} \mathbb{1}_{sex} + \begin{Bmatrix} d_1 \\ d_2 \\ \vdots \\ d_K \end{Bmatrix}$ grp coef

under GAMLSS we use a link distribution

$\delta_i \sim N(0, \sigma_\mu^2)$

$d_i \sim N(0, \sigma_\sigma^2)$

$$\hat{y} \sim \underbrace{f(\hat{\mu}, \hat{\sigma})}_{\text{really}}$$

$f(\hat{\mu}(x, \ldots), \hat{\sigma}(x, \ldots))$ [or even $\hat{y} = f(\hat{\mu}, \hat{\sigma}, \hat{\tau})$]

where $f()$ is a probability distribution

(2) So we can obtain GAMLSS-specific quantiles for the output

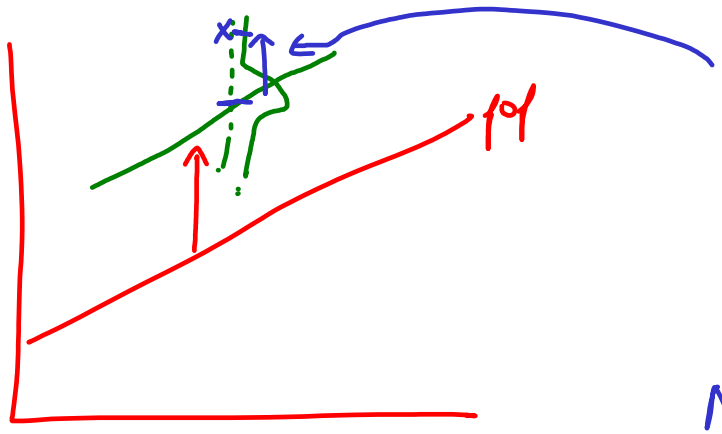
$f(x, \ldots.) : [-\infty, \infty) \mapsto [0,1]$ $\quad f^{-1} : [0,1] \mapsto (-\infty, \infty)$

the covariates

We "undo" the GAMLSS model

$y_{obs} - \hat{y} =$ residual

↳ residual specific to group random effect



looking for quantile of observation specific to group-level

NOTE using quantiles since GAMLSS link doesn't have to be normal dist (if normal, then we z-scores)

So $f^{-1}\left(y-\hat{y} \mid M(x_{\ldots}, grp), \sigma(x_{\ldots}, grp) \right)$

→ obtain quantile for each observation

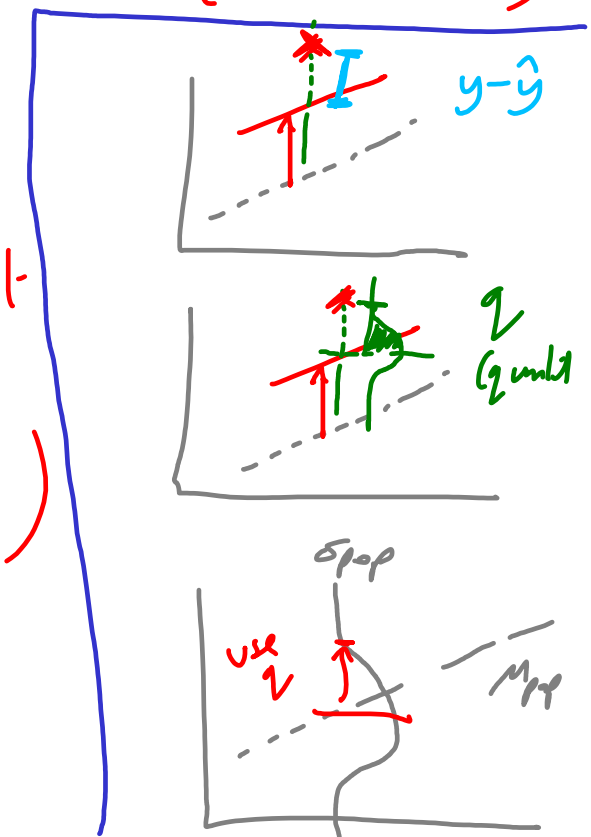
⟹ having removed grp effects (on μ & σ)

(3) redo GAMLSS link

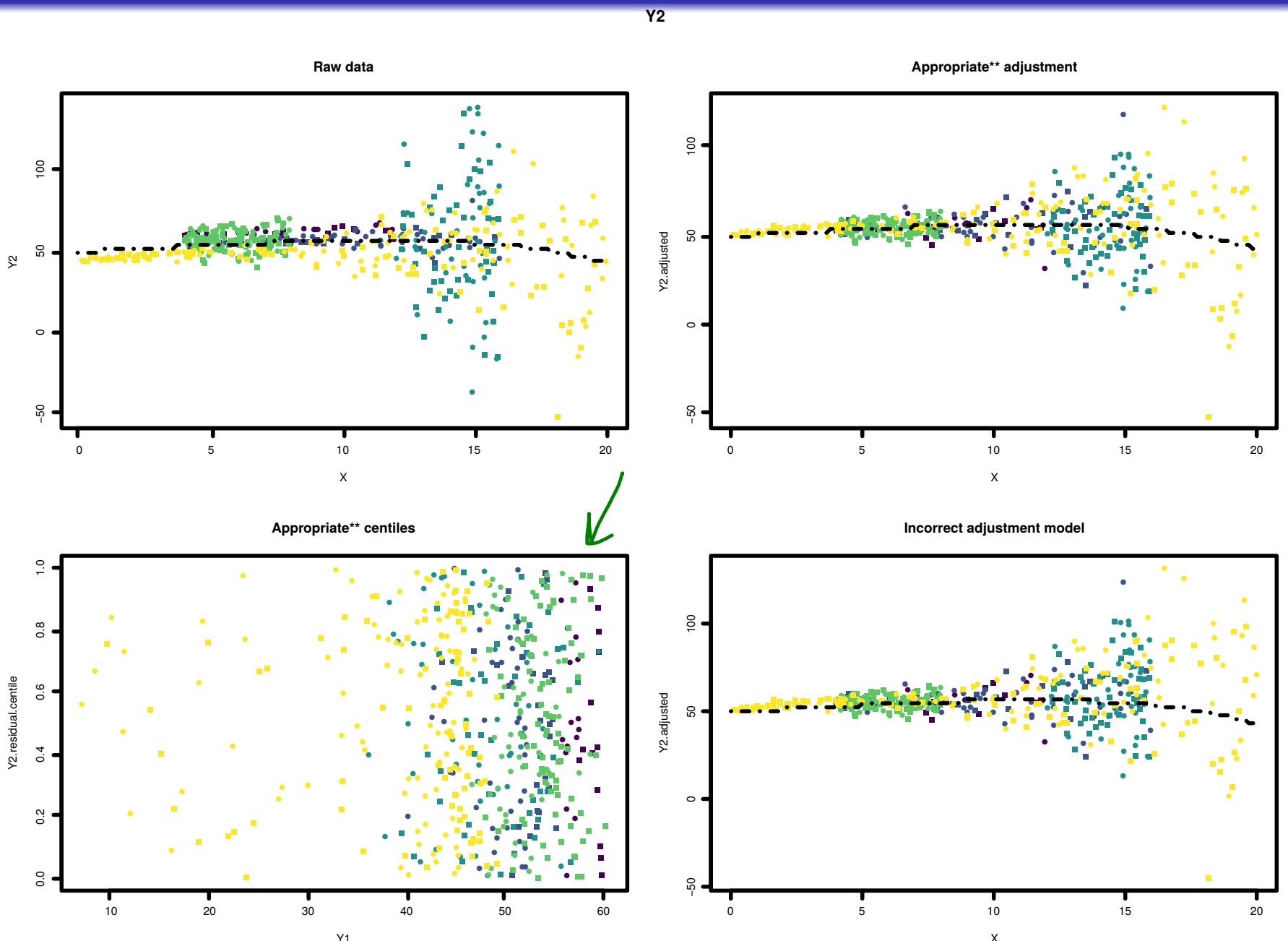
$F(x) = \int_{-\infty}^{x} f(x)$ — ie cumulative dist-

$F\left(M(x_{\ldots}, 0), \sigma(x_{\ldots}, 0) \right)$

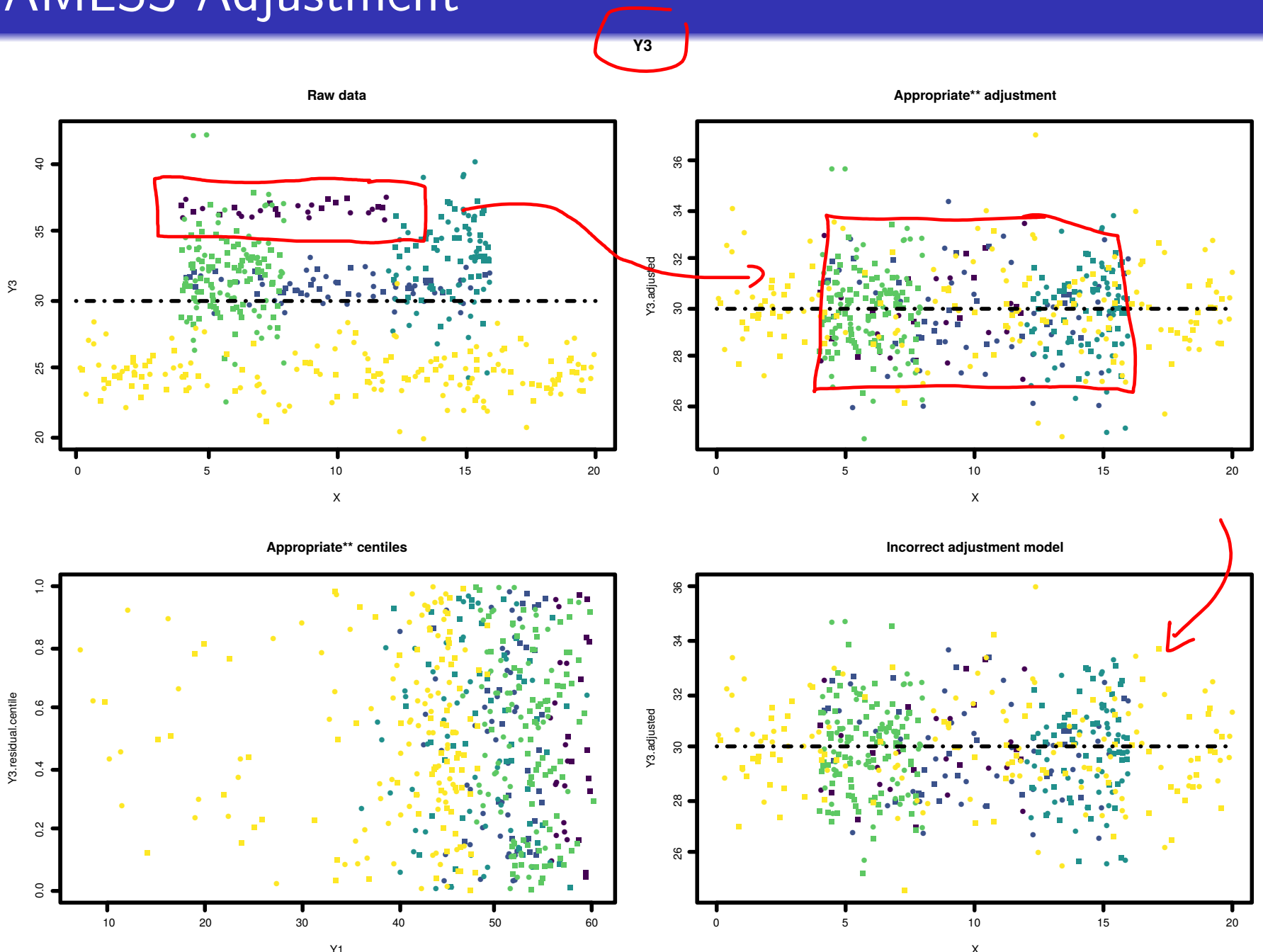
set grp effect to zero



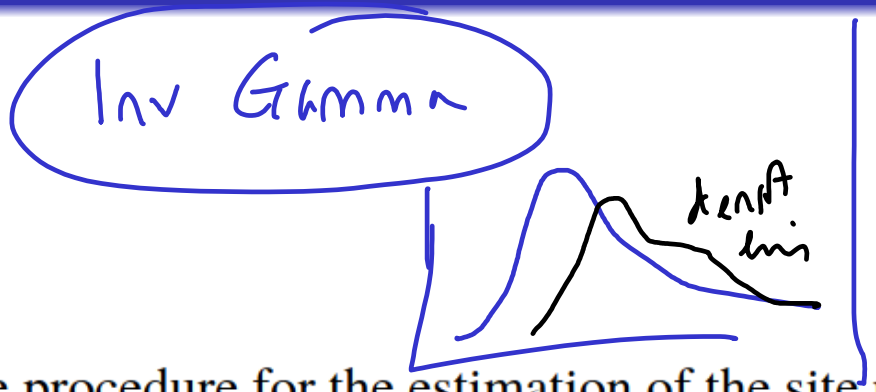
GAMLSS Adjustment



GAMLSS Adjustment



ComBAT



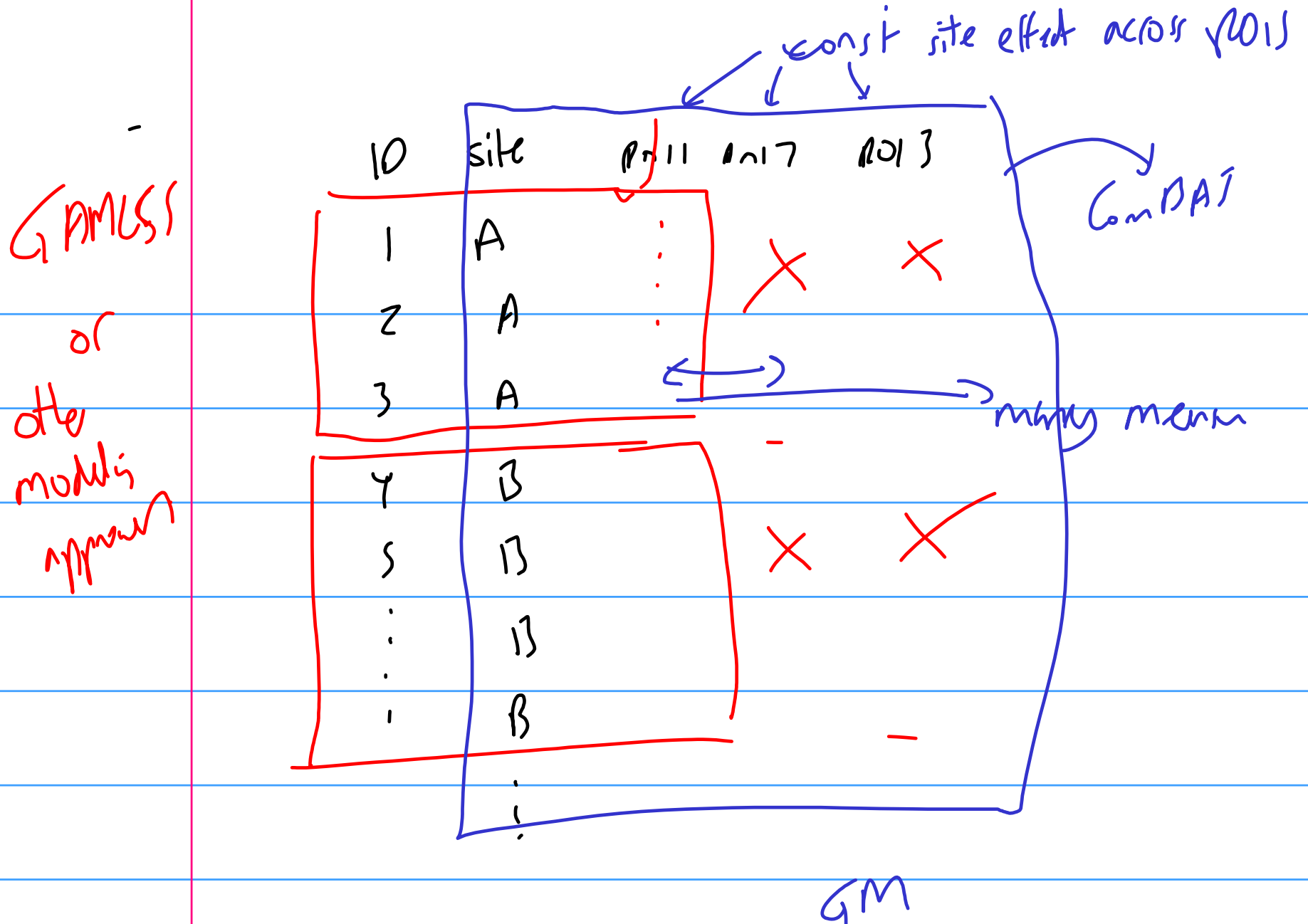
Small sample sizes within each study/batch

The procedure for the estimation of the site parameters γ_{iv} and δ_{iv} uses Empirical Bayes, and is described in [Johnson et al. 2007] and Fortin et al. [2017]. The final ComBat-harmonized cortical thickness measurements are defined as

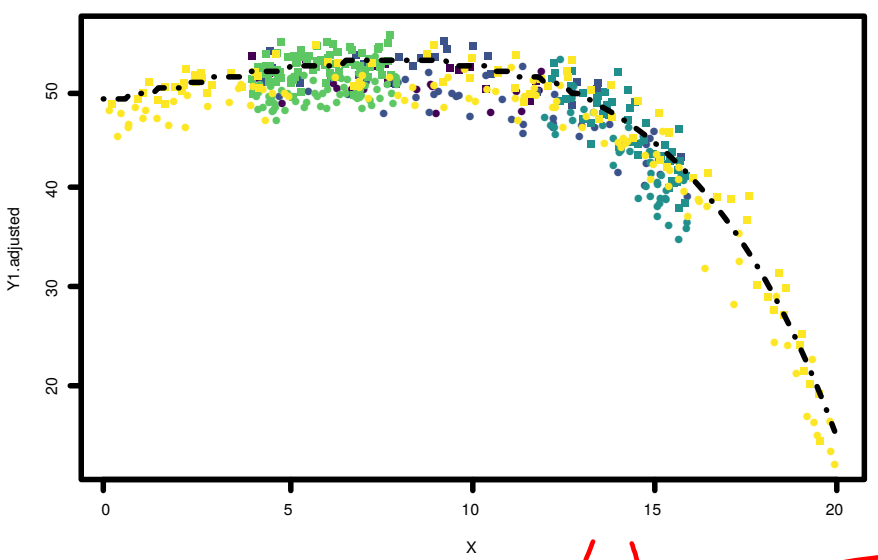
$$y_{ijv}^{\text{ComBat}} = \frac{y_{ijv} - \hat{\alpha}_v - \mathbf{X}_{ij}\hat{\beta}_v - \gamma_{iv}^*}{\delta_{iv}^*} + \hat{\alpha}_v + \mathbf{X}_{ij}\hat{\beta}_v$$

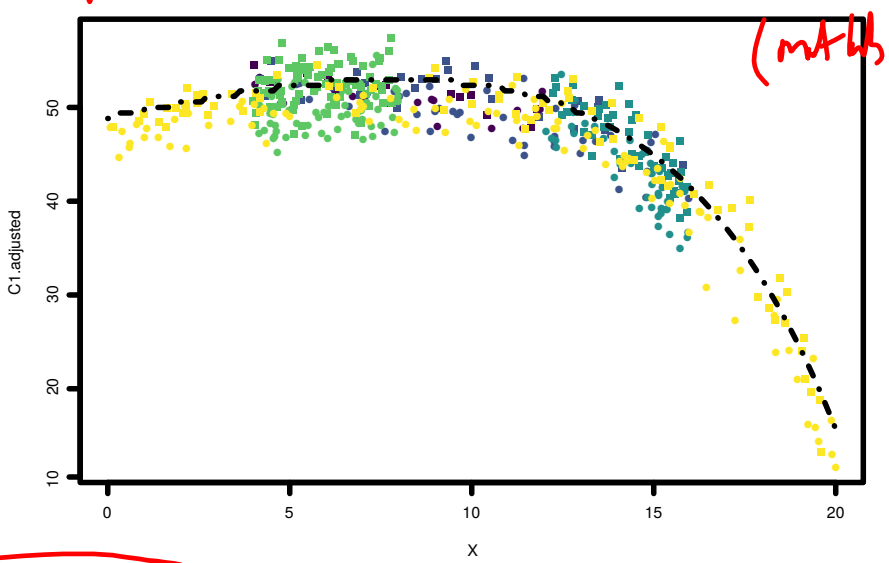
ComBat $y \sim N(\mu(x, grp), \sigma(x, grp))$

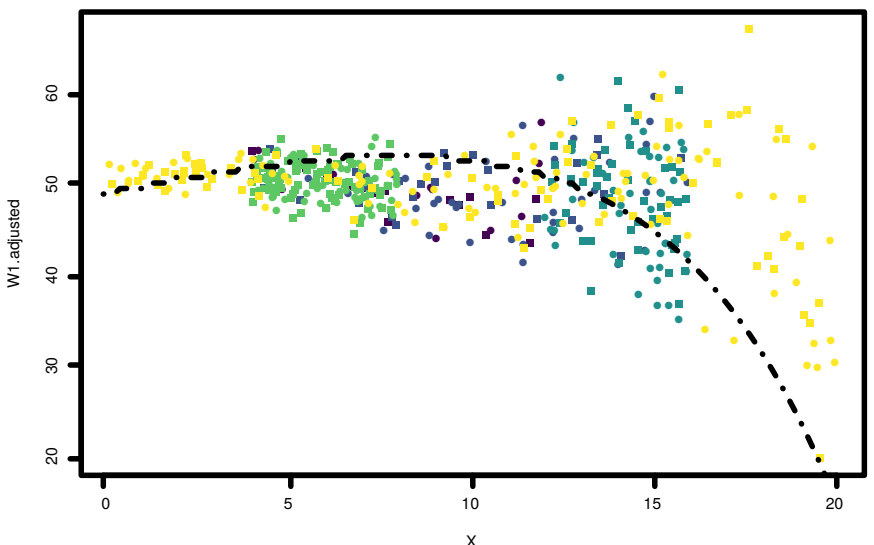
GAMLSS $y \sim f(\mu(\cdot), \sigma(\cdot), \tau(\cdot))$

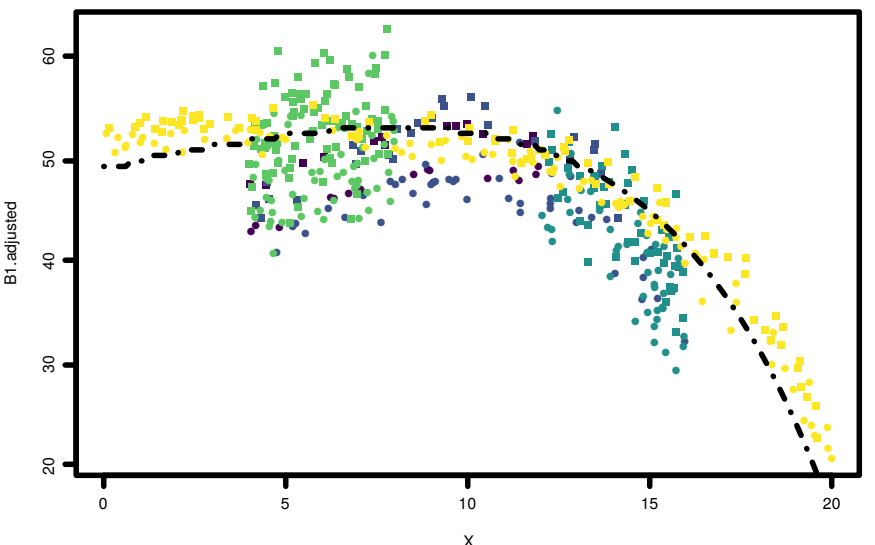

const site effect across ROIs
ID
site
ROI 1
ROI 2
ROI 3
1
2
3
A
A
A
4
5
B
B
B
B
ComBAT
many more
GM

GAMLSS and ComBat Adjustment

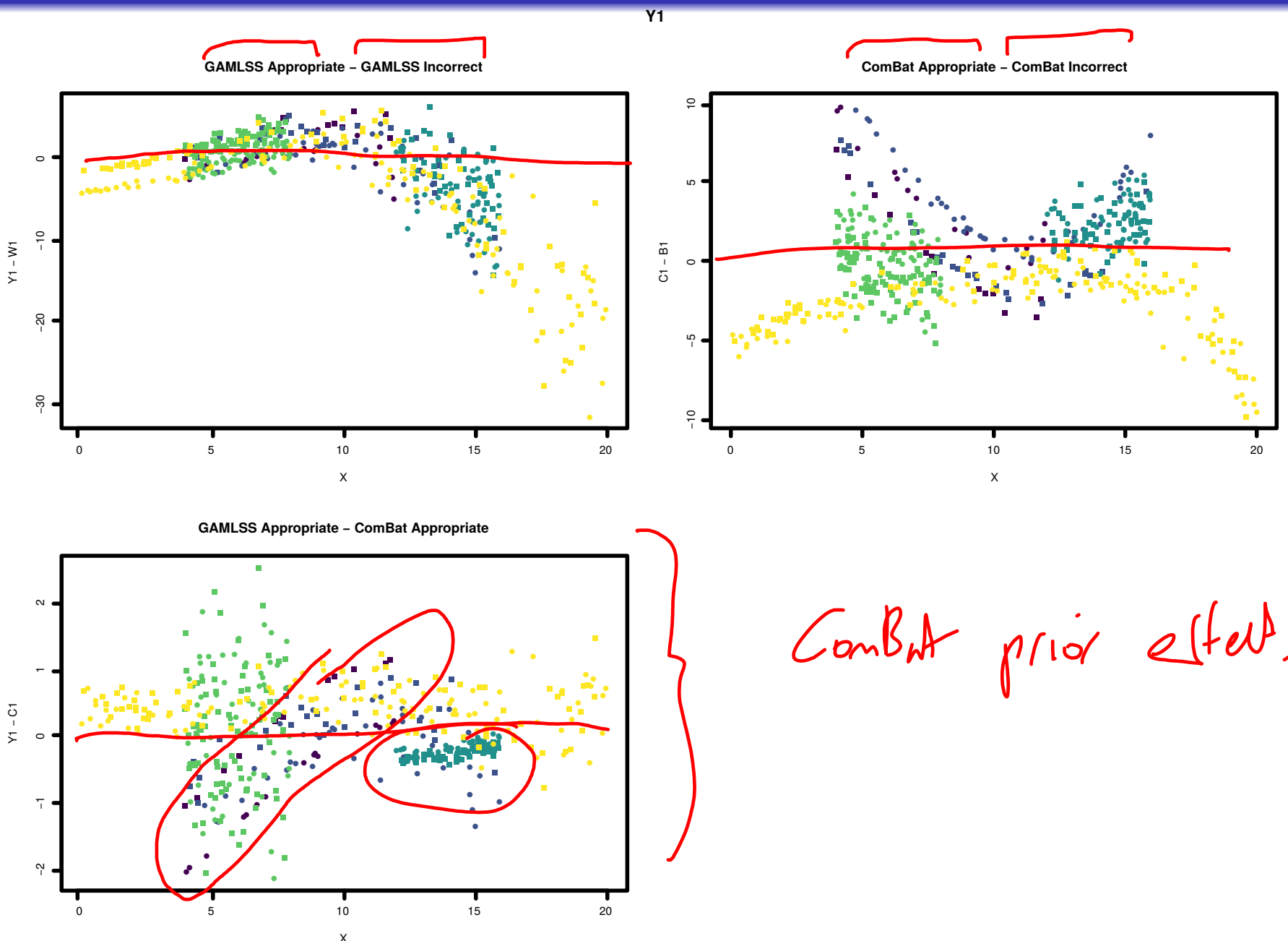




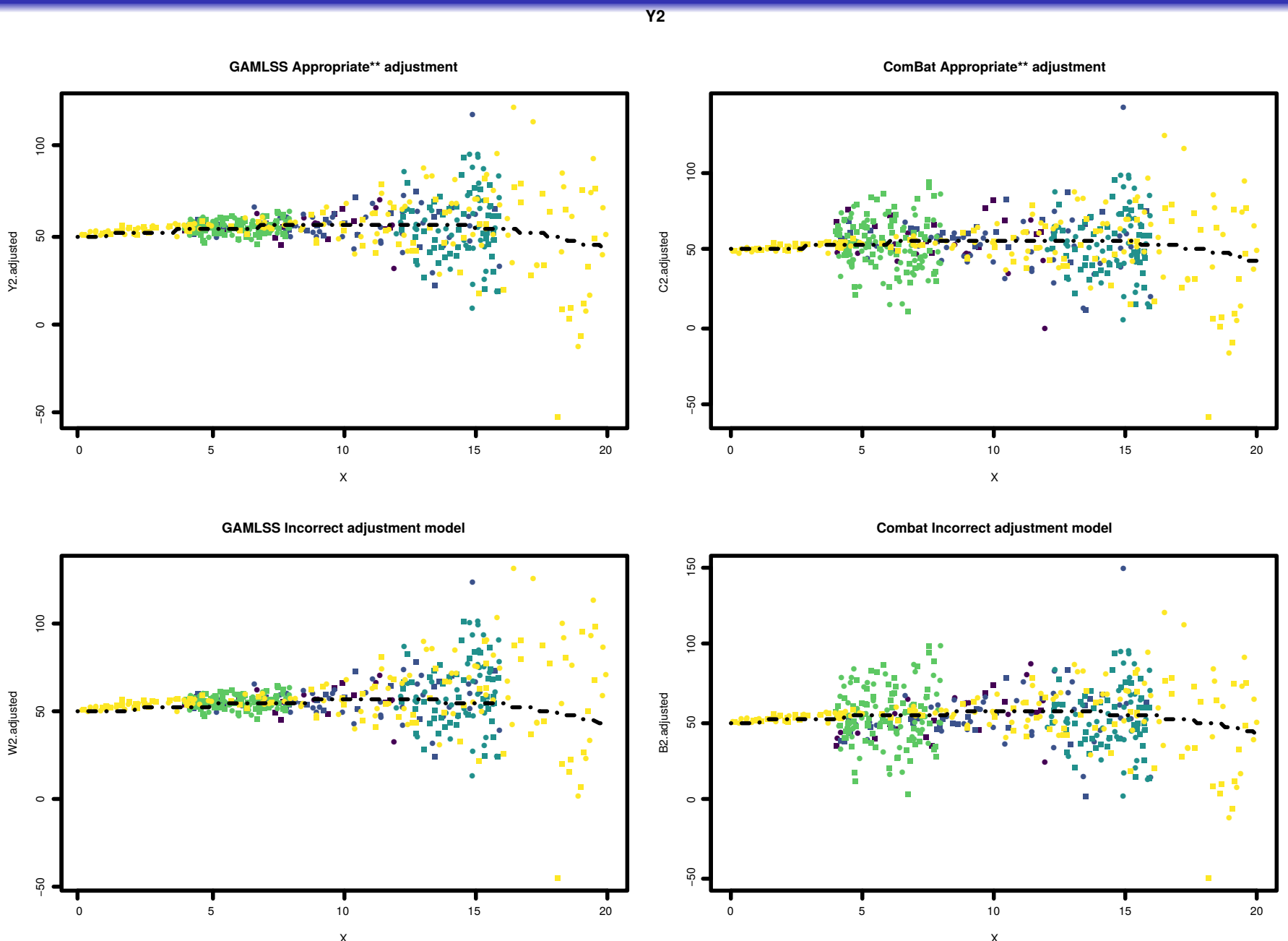




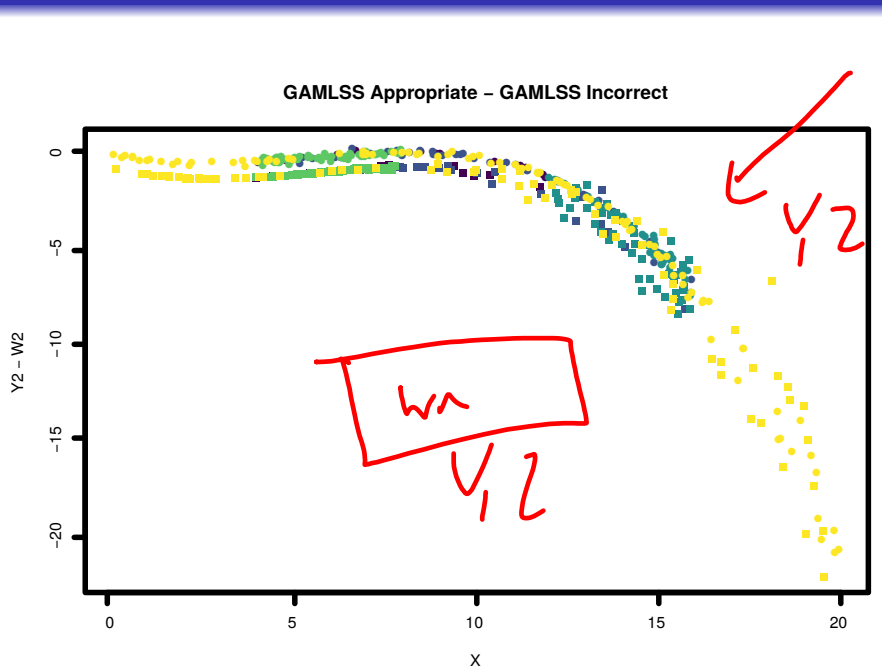
GAMLSS and ComBat Adjustment

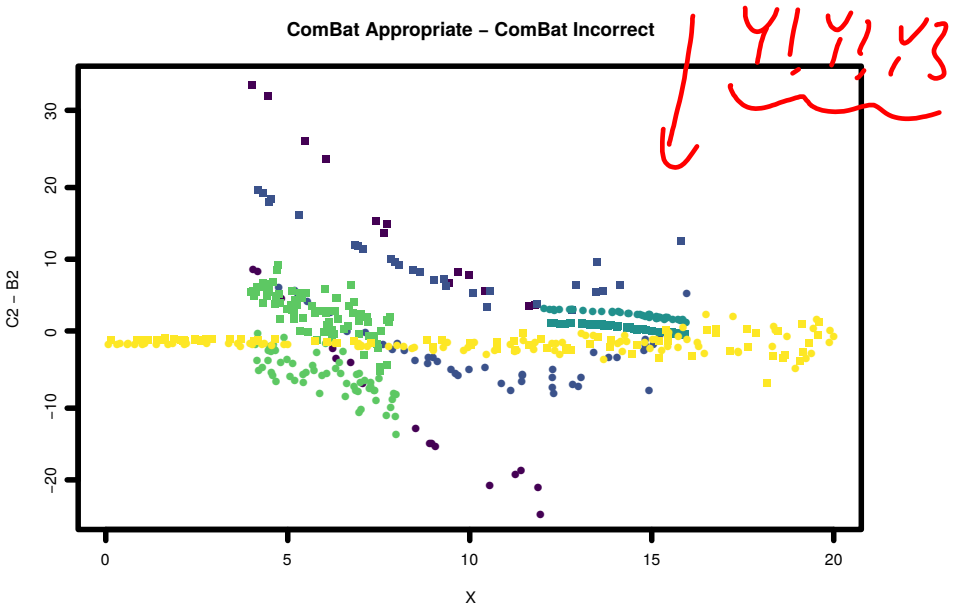


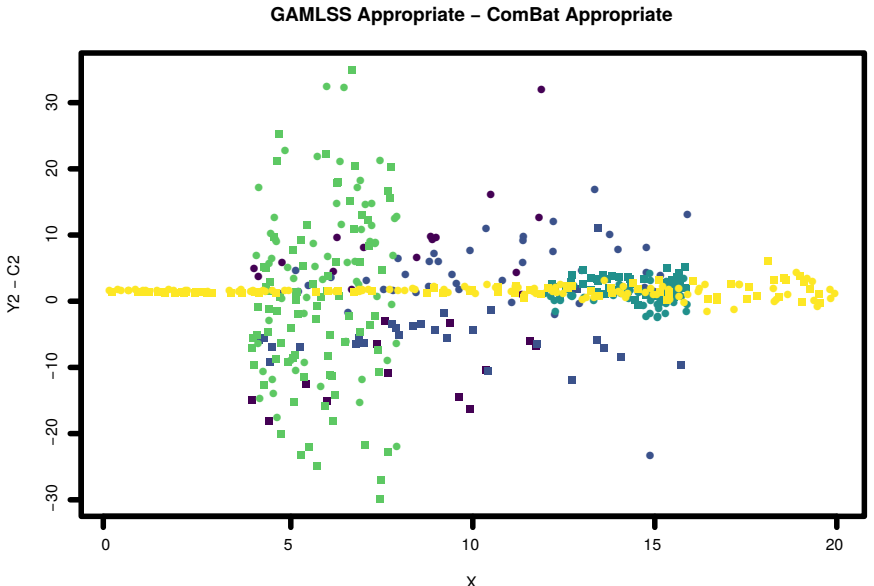
GAMLSS and ComBat Adjustment



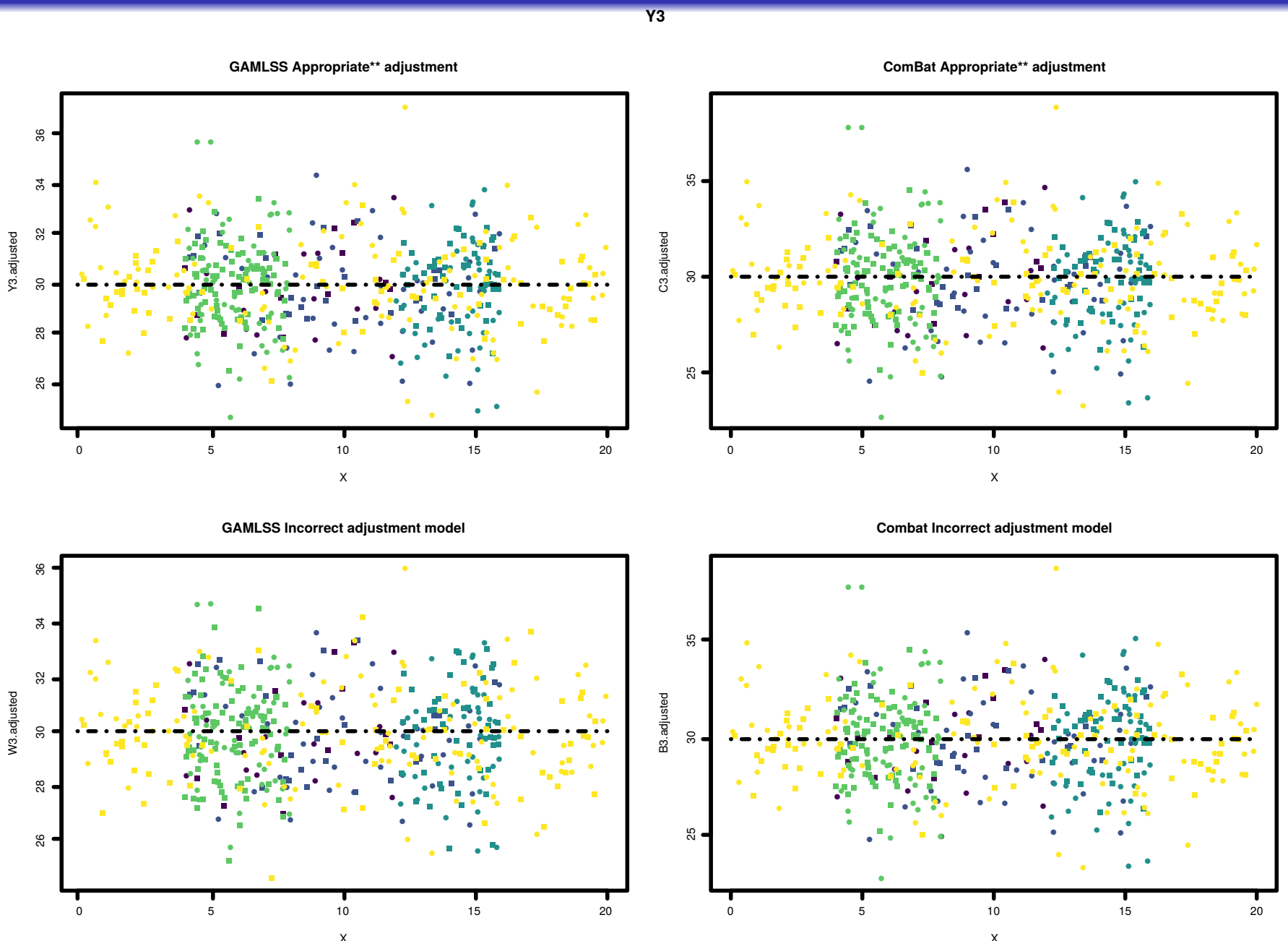
GAMLSS and ComBat Adjustment



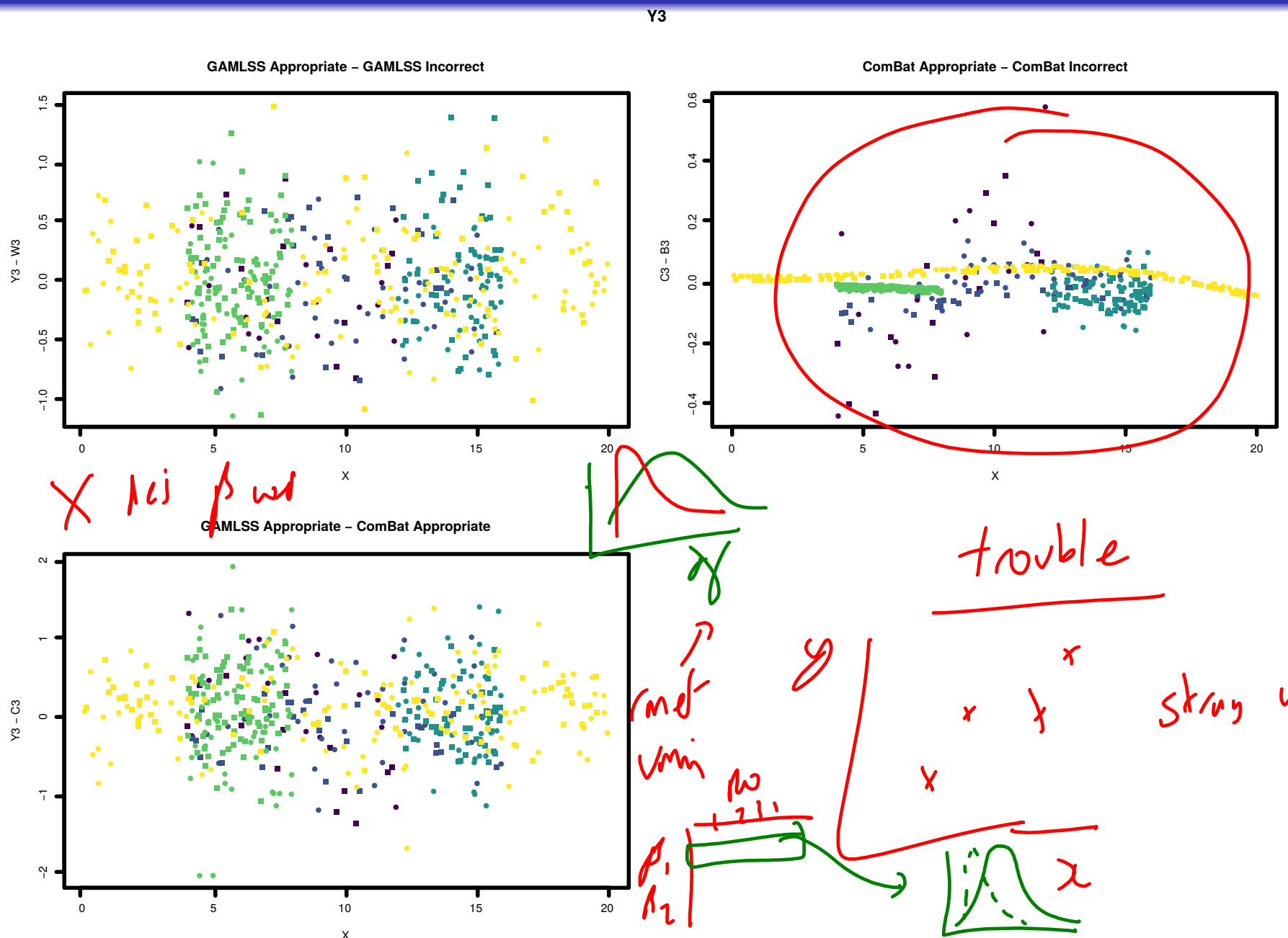




GAMLSS and ComBat Adjustment



GAMLSS and ComBat Adjustment



Summary

- ComBat uses Empirical Bayes to account for small number of samples per batch/group/study/site
- As number of samples per group increases, ComBat will converge to LME with heteroscedasticity
- GAMLSS more flexible. . . but requires bigger sample sizes!

Summary

- ComBat uses Empirical Bayes to account for small number of samples per batch/group/study/site
- As number of samples per group increases, ComBat will converge to LME with heteroscedasticity
- GAMLSS more flexible. . . but requires bigger sample sizes!

- Incorrect adjustment is a problem for all these methods
  - perhaps see Jacob, Gagnon-Bartsch and Speed (2016) *Biostatistcs*
  - Setting with unknown factor of interest (but this is hard)

Summary

- ComBat uses Empirical Bayes to account for small number of samples per batch/group/study/site
- As number of samples per group increases, ComBat will converge to LME with heteroscedasticity
- GAMLSS more flexible. . . but requires bigger sample sizes!

- Incorrect adjustment is a problem for all these methods
  - perhaps see Jacob, Gagnon-Bartsch and Speed (2016) *Biostatistcs*
  - Setting with unknown factor of interest (but this is hard)
- ComBat requires multiple measures per subject (i.e. ROIs), but assumes constant batch affect across all measures (strong assumption?)

Summary

- ComBat uses Empirical Bayes to account for small number of samples per batch/group/study/site
- As number of samples per group increases, ComBat will converge to LME with heteroscedasticity
- GAMLSS more flexible. . . but requires bigger sample sizes!

- Incorrect adjustment is a problem for all these methods
  - perhaps see Jacob, Gagnon-Bartsch and Speed (2016) *Biostatistcs*
  - Setting with unknown factor of interest (but this is hard)
- ComBat requires multiple measures per subject (i.e. ROIs), but assumes constant batch affect across all measures (strong assumption?)
- GAMLSS does not share batch random-effect estimates across measures (lossing information?)