



Stat/Biostat 571
Statistical Methodology
Regression Models for Dependent Data

Ken Rice

UW Biostatistics

March 22, 2016

CHAPTER 0: OVERVIEW

- Welcome
- This course is about regression when outcomes are **d**ependent; it builds on 570's analyses of independent outcomes
- Prerequisites, learning outcomes, syllabus, grading, and the place of this course in the degree program(s) are all discussed **on the course site**

Please read the site carefully – particularly if you did not take 570 last quarter

- **Office Hours:** Ken's office, Tuesday 10.30–noon. Or by appointment (kenrice@uw)
- **TAs:** Austin (aeschuma@uw), Yali (yaliwan@uw) and David (davidw19@uw) also have office hours, grade HWs, write keys, and run discussion sections – which are optional
- **Questions?** Come and ask – or email **all** of us

CHAPTER 0: OVERVIEW

- Lectures and Homework (and keys) will appear on the class site in due course. I will try to post/email any corrections/clarifications in a timely manner
- You are expected to read upcoming material **ahead of class** – say 20-30 slides ahead. The slides are intentionally wordy, to help with this. Reading from The Book (Wakefield) may also be assigned
- Starred (*) material is non-examinable; ***= 
- Unfortunately, I have to miss some classes; there will be guest lecturers/guest topics
- Please interrupt!

Any questions? ...any questions the class site doesn't answer?

Course info: about the instructor

- Ken Rice

Associate Professor in Biostatistics

12th year on UW faculty

Prior to that, worked in Cambridge, UK



- Collaborative research in cardiovascular disease, much of it in genetic/genomic epidemiology (categorical data) often clustered (e.g. family data)
- Chair, Cohorts for Heart and Aging Research in Genomic Epidemiology (**CHARGE**) consortium analysis committee, also **TOPMed**
- Methods research arising from these; also cracking heads between Bayesian and frequentist statisticians (e.g. p-values, sandwiches, shrinkage, fancy likelihoods)

Course info: about the TAs

And also;



Austin
Schumacher
aeschuma@uw



Yali
Wan
yaliwan@uw



David
Whitney
davidw19@uw

- Click the links for *their* CVs, research interests, etc
- They will handle grading, writing keys, discussion sections, and their office hours

Course info: textbooks

A familiar face from 570 (and his book)



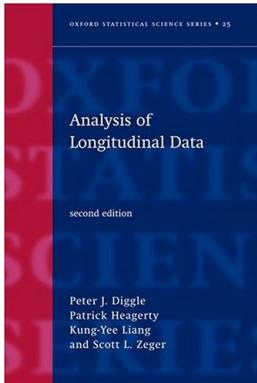
Wakefield, J.C. (2013)
Bayesian and Frequentist Regression Analysis
Selected Chapters

Chapters 8&9 cover GEE/mixed modeling, for dependent outcomes. (Chapters 10–12 discuss highly flexible models, e.g. splines)

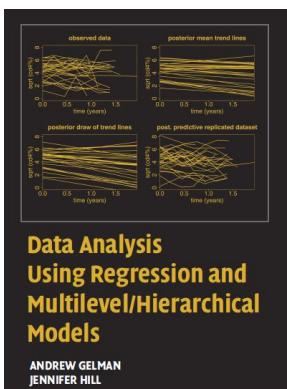
Other notes will be provided, including book chapters when I can make these available. Research papers will be featured, for more recent topics – 571 is more ‘cutting edge’ than some other UW courses.

Course info: textbooks

Also worth your time/money (but not required);



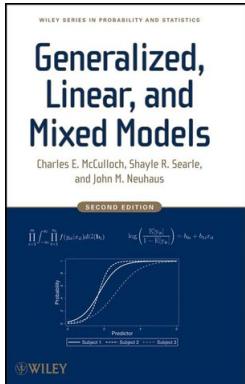
Most authoritative text for GEE, marginal models, also good on other frequentist methods. Concentrates on longitudinal data



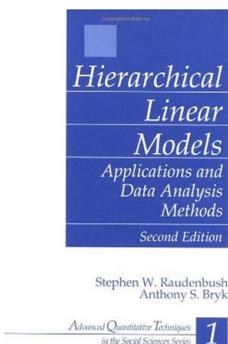
Well-written coverage of regression, in general frequentist and Bayesian settings. Chaps 1–10 cover 570 material, the rest is useful in 571

See also the '**Song Book**' for mathematical details/proofs.

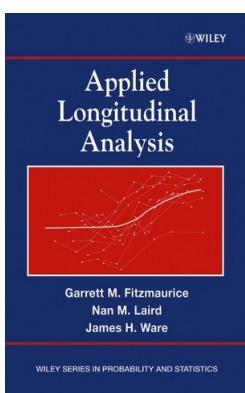
Course info: supplemental texts



Extensive coverage of model-based approaches to 570 and 571 material. Like an updated McCullagh & Nelder. Not much GEE.



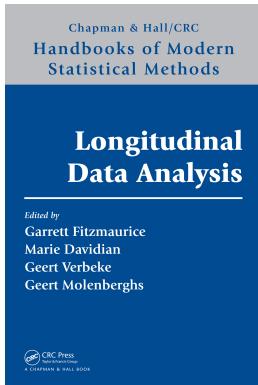
Only covers linear models, but well-written and with extensive examples. Predates GEE



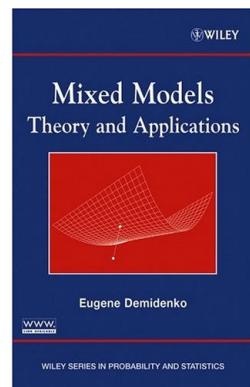
Similar to DHLZ, but a little more introductory – also more on testing, a topic I may not stress

Course info: supplemental texts

Relevant, but perhaps most useful post-571;



Terrific starting point for further research in this area – provides a comprehensive and recent overview



'This book is for people who want to know mixed models inside-out'... or just fit them inside-in. Not much Bayesian material.

Note there are several introductory texts available on this material (for e.g. BOST 540). Few are terrible, but they may not give PhD-level descriptions/motivation/justifications you may need.

Course info: topics

Some key topics; (more on this later)

- Review of 570 material: key topics, and how they might be extended for dependent data settings
- Marginal models: extension of 570 to generalized estimating equations (GEE). Parameter interpretation, robustness properties, heuristics on where the relevant asymptotics/efficiency may break down,
- Conditional models: distinguishing marginal from conditional inference, particularly for hierarchical data. Parameter interpretation, sensitivity, comparison with GEE with finite samples
- Bayes for hierarchical models. Exchangeability as a fundamental concept, shrinkage. Numerical fitting procedures (beyond those from 570)

Methods for dealing with missing data will be introduced.

Course info: topics

Time permitting, we *may* discuss some of these;

- Meta-analysis – and its marginal/conditional varieties
- Errors in variables, missing data – the impact of these, and methods that address them
- G -estimation – answering some causal questions that regression methods cannot address, often in longitudinal settings
- Still-more advanced MCMC
- Transition/multi-state models

Please be aware that data-based examples (given throughout) are usually presented primarily to *illustrate* methods; “real-life” data analysis require more time and use much more application-specific background knowledge. For more practice with that, see e.g. Biost 579 (data analysis) and 590 (Consulting).

Course info: topics

571 is a ‘methods’ course - i.e. it’s not a theory or data analysis course. If these terms are unclear...

- **Methods:** *Biometrics, Annals of Applied Statistics, Biostatistics, Statistics in Medicine*
- **Theory:** *Annals of Statistics, JRSSB, Statistica Sinica*
- **Data Analysis:** *JASA CS&A, JRSSA, ‘applied’ non-statistical journals in many fields*

Several journals cover more than one of these areas; *JASA Theory & Methods, Biometrika, Statistical Science, American Statistician* – and modern methods papers use simulation studies to illustrate statistical properties.

As well as training you to understand & explain advanced methods, 570/571 illustrate material that’s similar to *most* PhD theses. Understanding papers on this material is also good preparation for 572.

Course info: homework & exams

Homework is **very** important, both for understanding methods **and** practice explaining them to others

- Discussion is encouraged, but code it and write it up **on your own** – because writing is an important skill
- It's assumed you are adept with R help pages... and know not to sit stumped with code for hours
- Where specified, answer applied questions as you would in an applied paper; give valid statistical statements AND *interpret* these results as ‘real-world’ conclusions – in full sentences, in paragraphs. R code appears in appendices **only**
- Methods is not pure math. There may be no 100% right answer – ability to *discuss* pros/cons is essential. If unsure what HW is asking, check with Instructor/TAs

While I will try to minimize them, typos can be expected.

Course info: homework & exams

I intend to follow 570, and have;

- A take-home midterm exam, in place of HW5 (± 1)
- A take-home final exam, in place of HW10

All topics up until the exam begins are examinable, and work must be completed **entirely on your own** – just like research, and your future career.

However, mindful of what happened in 570, if I doubt that **everyone** is giving good-faith efforts on regular HW, to get an honest assessment of everyone's understanding, it will be;

- An in-class midterm exam
- An in-class final exam
- Regular HWs (10 in all)

Course info: Fair warning

As in 570, we will get ‘under the hood’ of methods – which is a lot of work. In 571, the methods are recent; two ‘classic’ manuscripts are;

- Liang and Zeger (1986) Longitudinal data analysis using generalized linear models, *Biometrika* 73(1): 13-22
- Pepe and Anderson (1994) A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data

Jon’s book is excellent and up to date – but a lot of material out there is not. E.g. McCullagh and Nelder’s book (from 1983/1990) will contain *at best* preliminary descriptions of the newer 571 material. Statistical methods – and our understanding of them, and what’s computable – have all changed substantially in the last 25 years. This evolution will continue.

CHAPTER 1: REVIEW

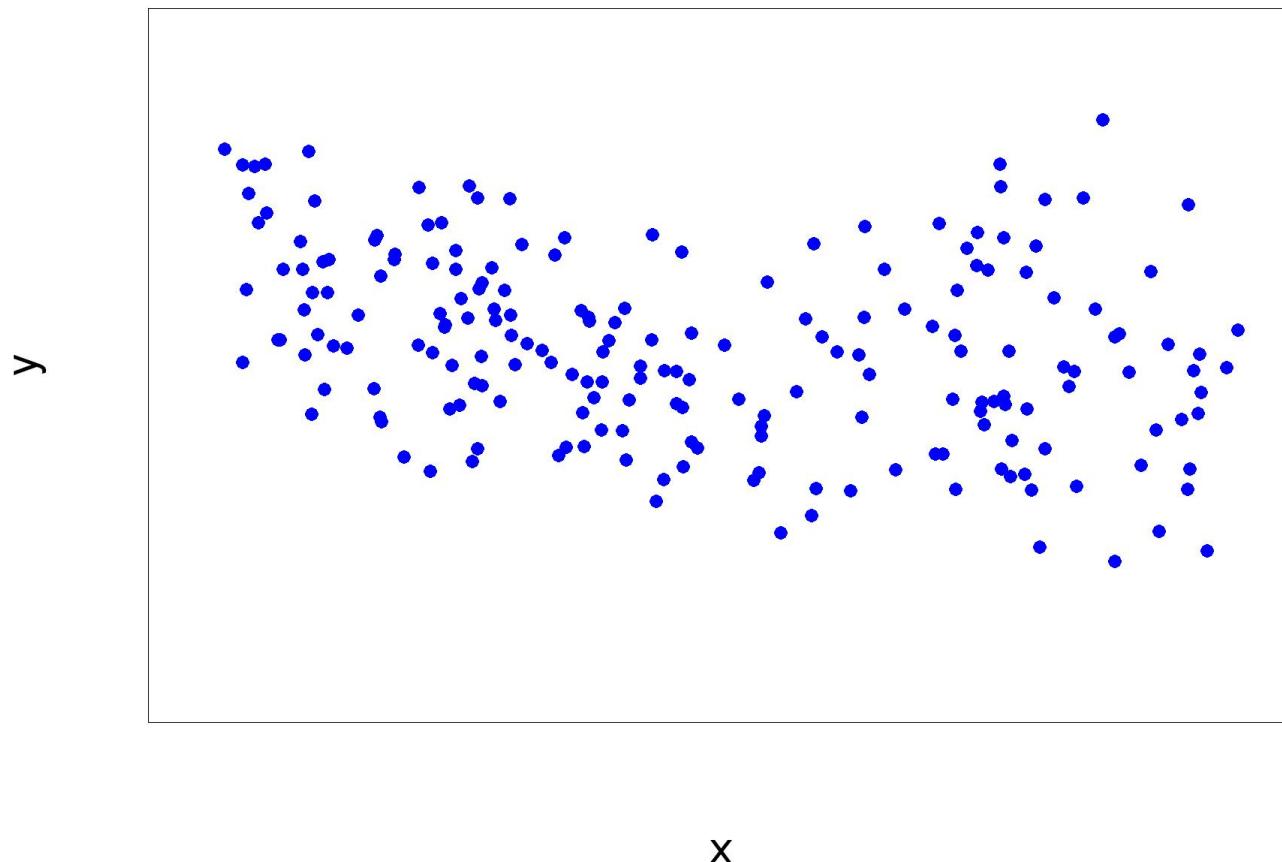


Where no plan is laid, where the disposal of time is surrendered merely to the chance of incidents, all things lie huddled together in one chaos, which admits neither of distribution nor review.

Hugh Blair (1718–1800)
Presbyterian author and preacher,
participant in the Scottish Enlightenment

Defining parameters: the big picture(s)

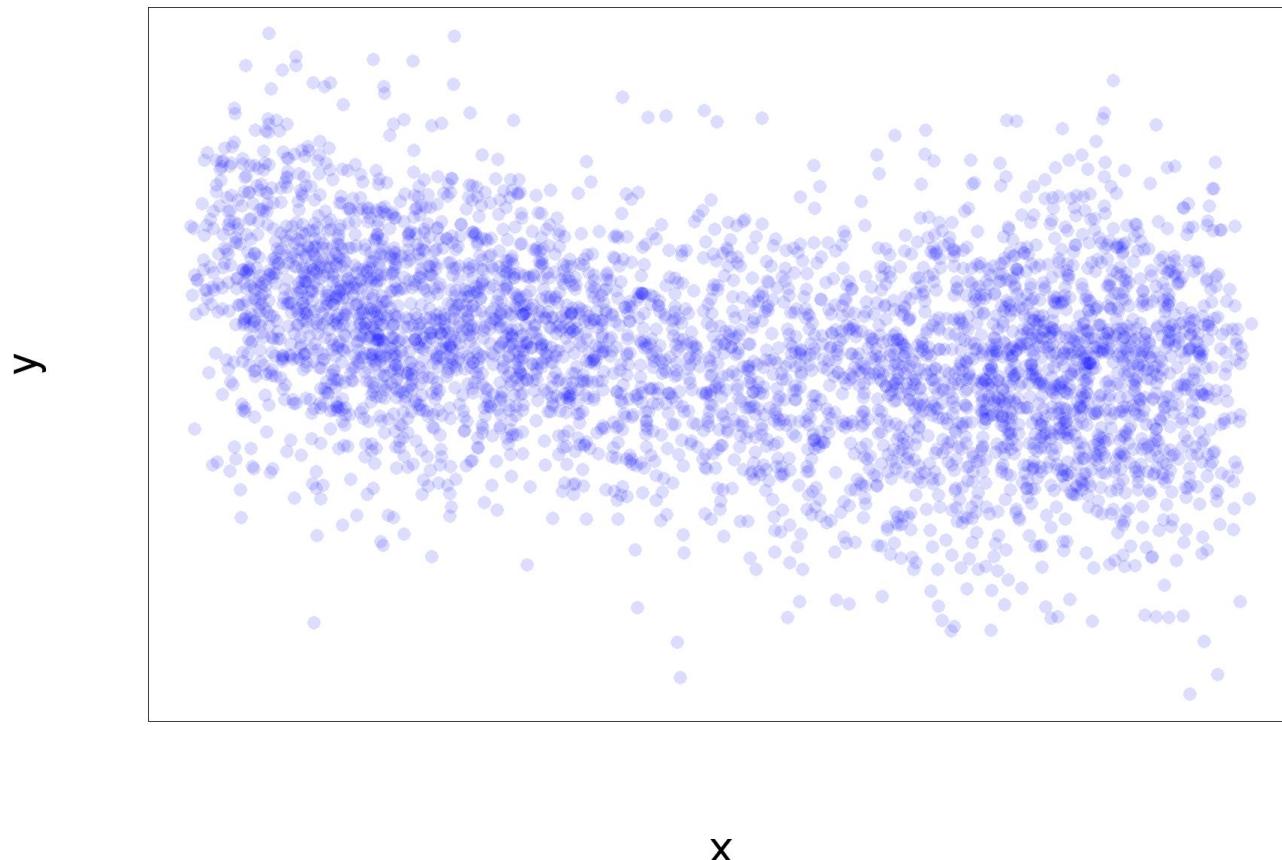
We first huddle together key ideas from 570 – first, the ‘big picture’ on parameters. Here’s a dataset;



... which you could manipulate in several ways.

Defining parameters: the big picture(s)

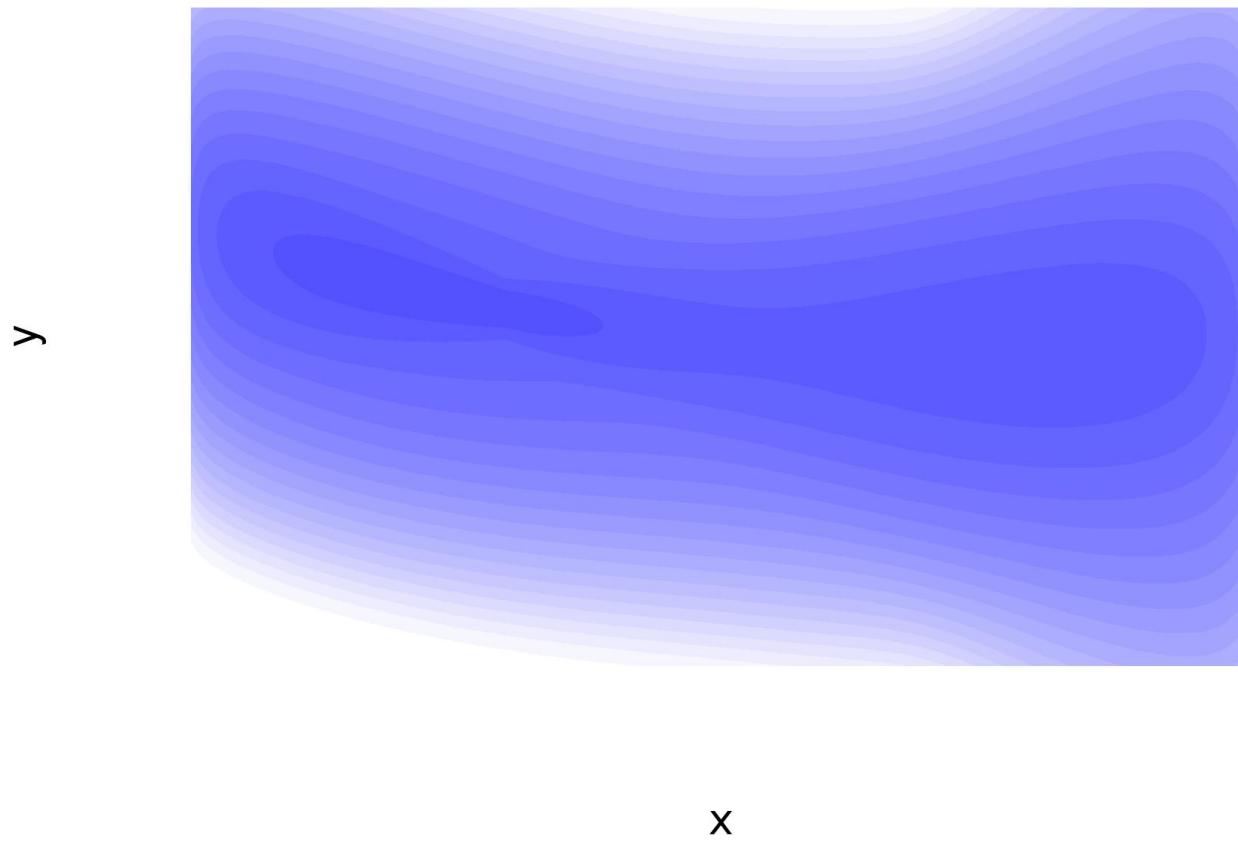
Here's a larger sample (plot uses transparency);



Of course, this sample can also be manipulated.

Defining parameters: the big picture(s)

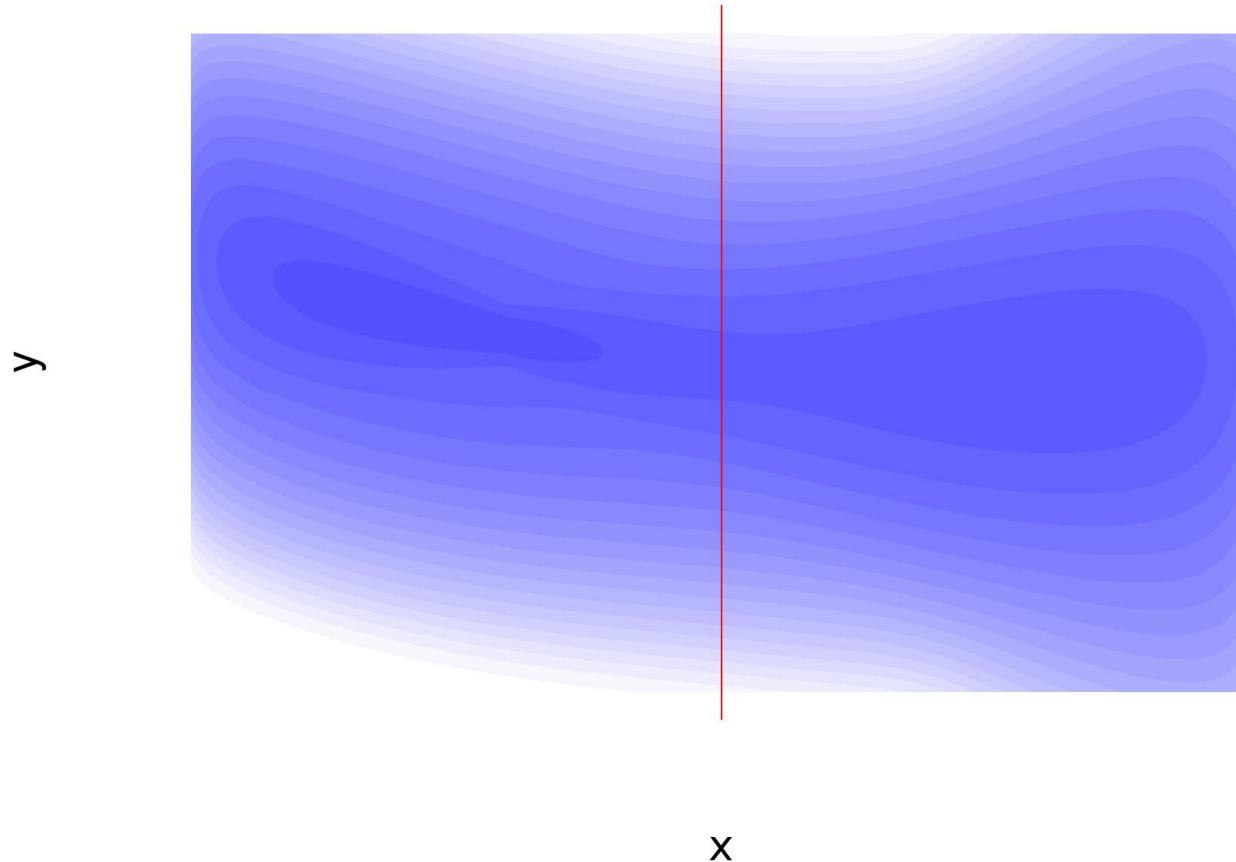
To define parameters, we consider an **infinite** ‘super’-population;



... and manipulations of it.

Defining parameters: the big picture(s)

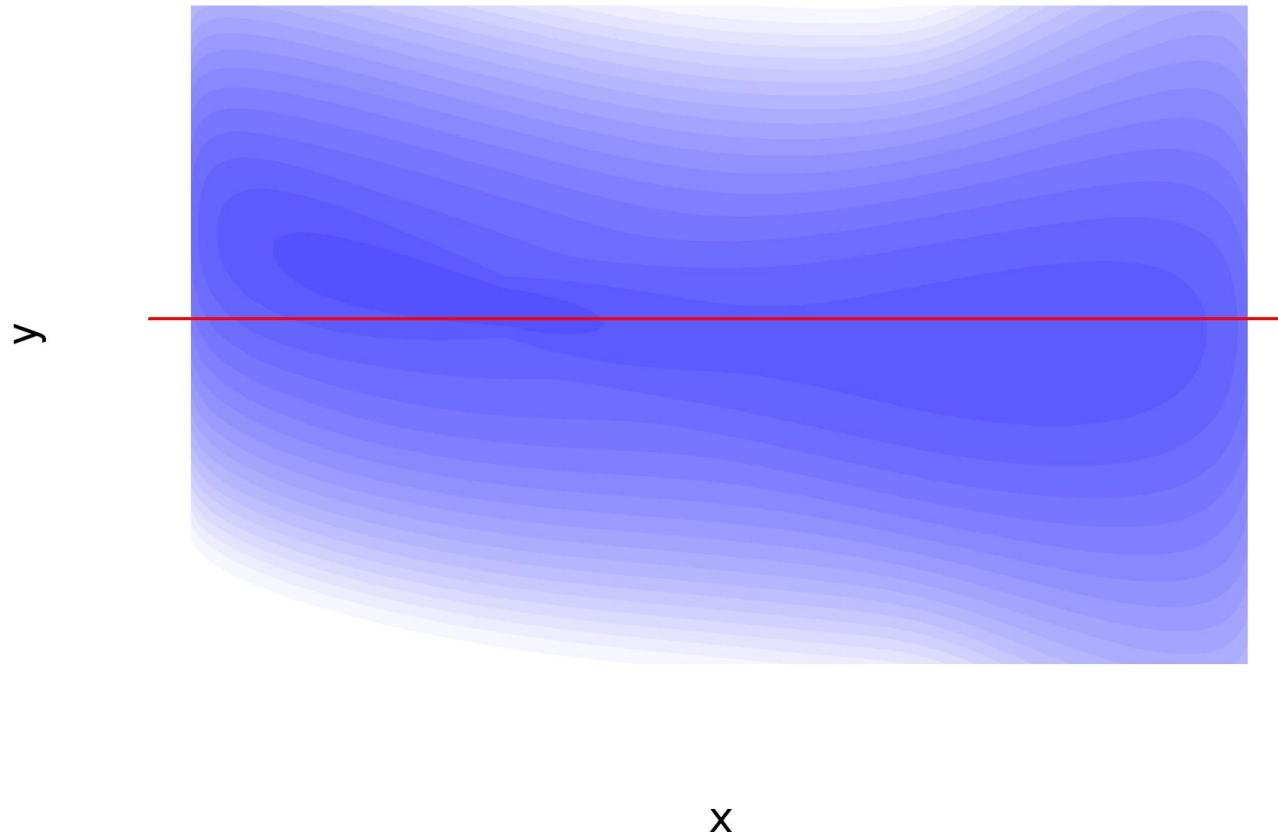
The mean of X :



(note: requires finite moments of X to be well-defined)

Defining parameters: the big picture(s)

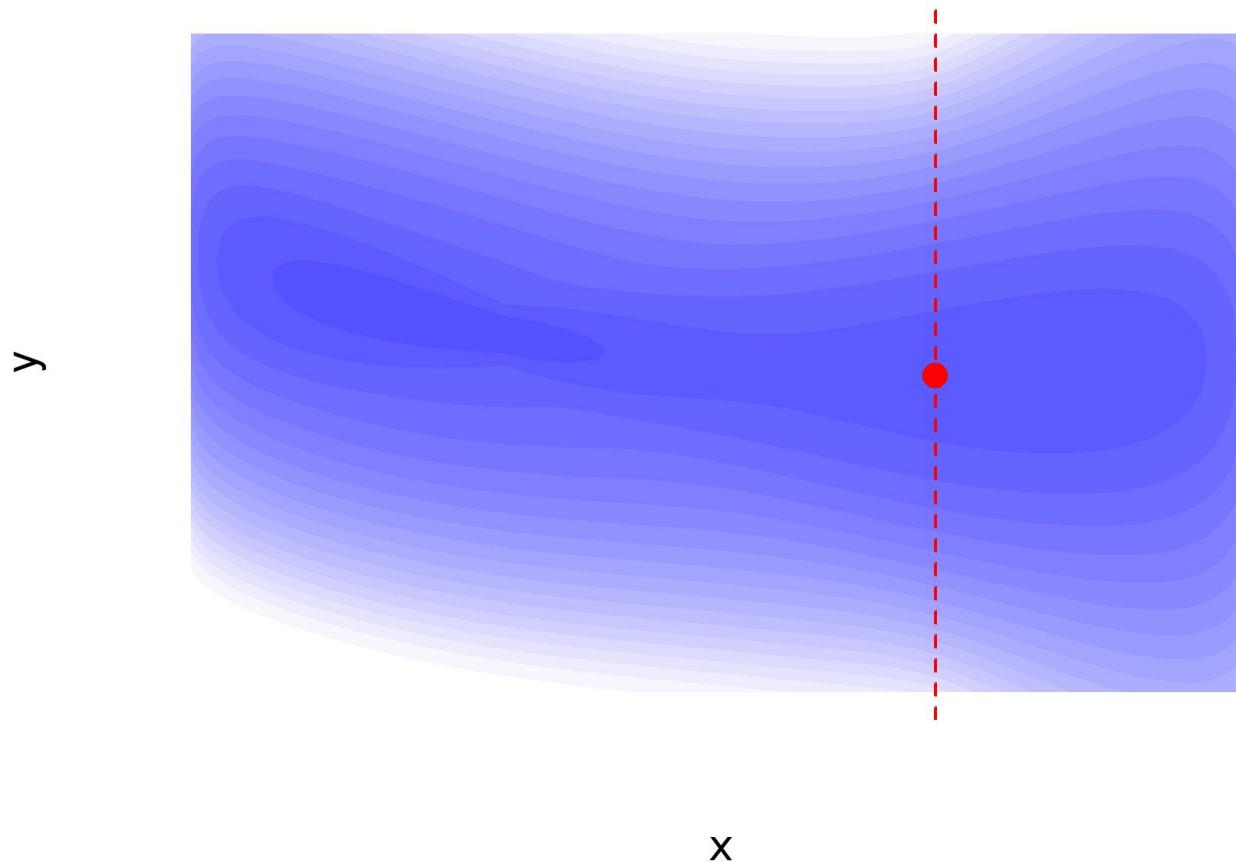
The mean of Y :



... *mild* regularity conditions also apply

Defining parameters: the big picture(s)

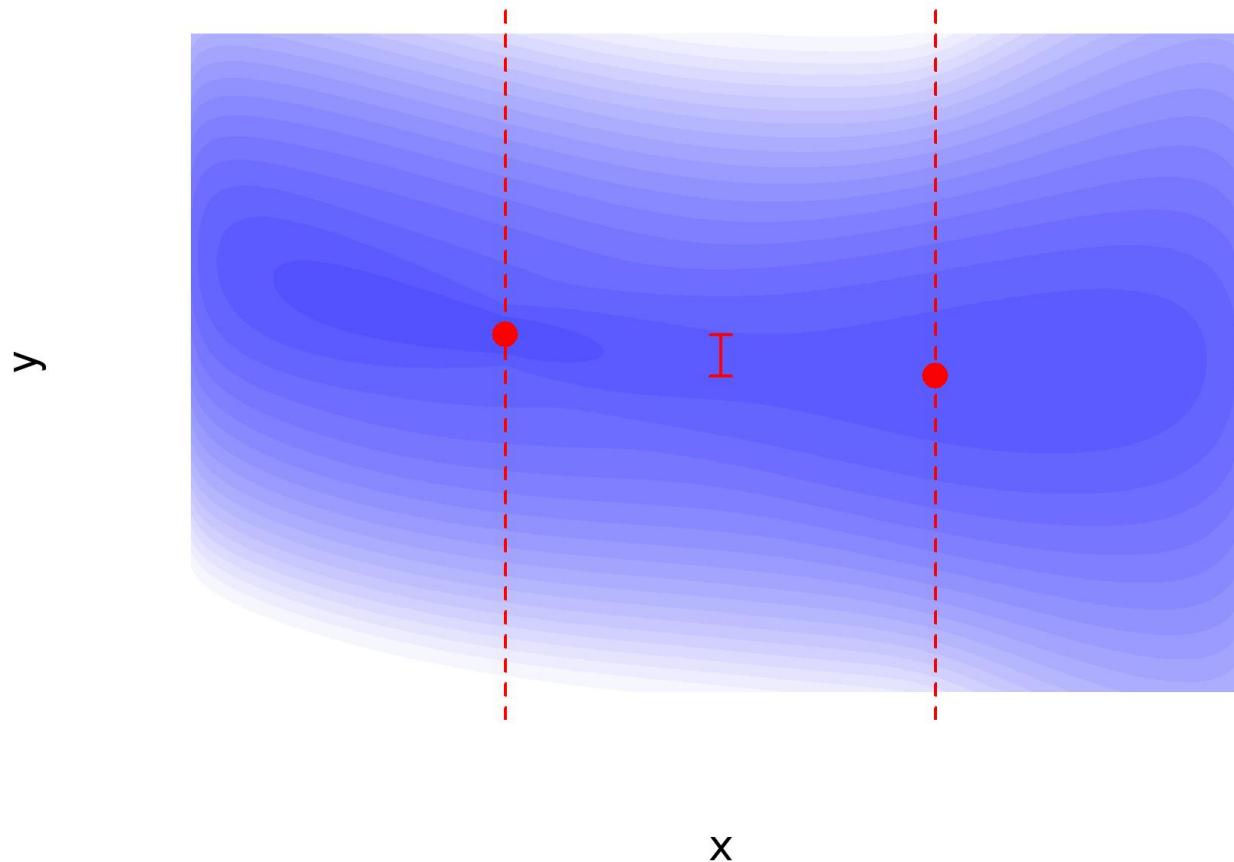
The mean of Y at a given value of X



... only useful if you know the given value of X (!)

Defining parameters: the big picture(s)

Difference in mean of Y , between two values of X ;



... which is unchanged, if $Y \rightarrow Y + c$

Defining parameters: definitions

Formally, a *parameter* is an operation on a population, mapping it to a ‘parameter space’ Θ , such as \mathbb{R} , or \mathbb{R}^p , or $\{0, 1\}$.

The parameter *value* (typically denoted β or θ) is the result of this operation*.

- ‘Inference’ means making one or more conclusions about the parameter value, i.e. about the population
- These could be estimates, intervals, or binary (Yes/No) decisions
- ‘*Statistical* inference’ means drawing conclusions **without** the full population’s data, i.e. in the face of uncertainty. Parameter values themselves are fixed unknowns; they are not ‘uncertain’ or ‘random’ in any stochastic sense.

Choice of the target parameter is subjective, and should be based on scientific context.

* Other names are the ‘true state of Nature’ or just ‘the truth’

Defining parameters: definitions

As you saw in 570, a general definition gives θ as the solution to

$$\mathbb{E}[G(\theta, Y, \mathbf{X})] = 0,$$

where the data are sampled independently from the relevant population. Useful special cases include;

- Means, other moments
- Variances and other centralized moments (as a term in some $\theta \in \mathbb{R}^2$ or higher)
- Differences in means
- Essentially all parameters in standard regressions
- Instrumental variable approaches
- Outlier-robust approaches

Keen people: the framework has to be extended *slightly* to include quantiles, e.g. $\theta = \text{median}(Y)$ – see e.g. the review by Stefanski & Boos (2002).

Defining parameters: which ones?

Why choose any particular $G()$? For very simple situations we can argue directly;

- In a simple randomized trial, where treatment X is assigned randomly, i.e. $X \sim \text{Bern}(0.5)$,

$$\theta = \mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0]$$

tells us about the (average) causal effect of treatment on outcome Y

- To summarize how Y changes at different values of X , define a weighted ‘average slope’ between pairs of observations;

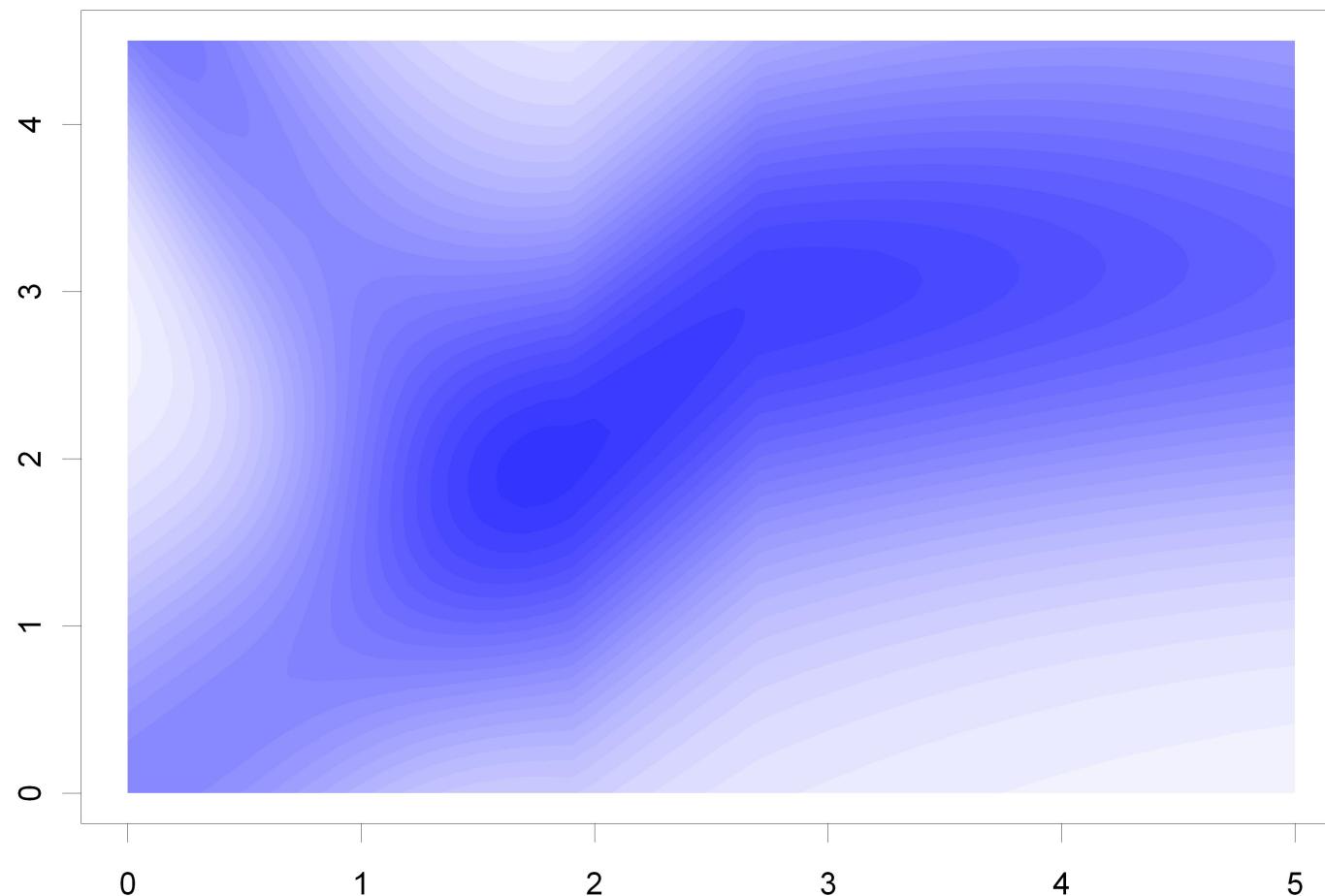
$$\begin{aligned}\beta &= \frac{\mathbb{E}[(X - X')^2 \frac{Y - Y'}{X - X'}]}{\mathbb{E}[(X - X')^2]} \\ ... &= \text{Var}[X]^{-1} \mathbb{E}[(X - \mathbb{E}[X])Y],\end{aligned}$$

where ‘... =’ is due to Jacobi (1841). You may be more familiar with another definition, as the β_1 slope term here;

$$\boldsymbol{\beta} = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \mathbb{E}[(Y - \beta_0 - \beta_1 X)^2]$$

Defining parameters: which ones?

This ‘least-squares slope’ concept is probably easiest to visualize; roughly what is its value here?



Defining parameters: which ones?

Writing these definitions without math but with sneaky caveats;

- β is the average weighted slope defined by pairs of observations ...where the weights are proportional to between-X distances²
- β is the ratio of the joint centralized second moment of X and Y to that of X ...where your reader knows what this means
- β is the slope of the ‘best-fitting’ straight line drawn through the population ...where ‘best’ means least mean squared distance on the Y axis
- β is a trend saying how Y (or mean $Y|X$) varies with X , averaging over the whole population ...where ‘trend’ isn’t actually defined at all!

Like many parameters defined by some $G()$, inference for β is available – see next slide. But without further assumptions it’s often very challenging to connect β (or other parameters) with scientific context. The problem only gets worse when $\mathbf{X} \in \mathbb{R}^2$ or higher, and/or when non-linear forms of regression are used.

Inference: with minimal assumptions

From 570, with independent observations and under regularity conditions (not stated here) we in general estimate using

$$\hat{\beta} \text{ such that } \frac{1}{n} \sum_{i=1}^n \mathbf{G}(Y_i, \mathbf{X}_i, \hat{\beta}) = 0,$$

i.e. an *empirical* or *plug-in* estimate, based on β 's definition.

For inference; (formally *non-parametric* inference)

- The ‘sandwich’ $\hat{\mathbf{A}}^{-1} \hat{\mathbf{B}} \hat{\mathbf{A}}^{-1} / n$ estimates asymptotic variance $\text{Var}[\hat{\beta}] = \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1} / n$, and hence gives confidence intervals
- Bootstrapping estimates similarly, & performs similarly
- In small samples either approach can work badly – depending on n , $\mathbf{G}()$, and the joint distribution of $\{Y, \mathbf{X}\}$

For *some* other jobs other tools may be preferable – e.g. permutation tests – for particular inferential goals.

Note: Bayesian inference with these assumptions (*NP-Bayes*) is beyond the scope of 570/571 – and is not used much in practice.

Defining parameters: more assumptions

One key assumption helps connect statistics with science;

$$\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] = g(\mathbf{x}^T \boldsymbol{\beta}), \text{ or } g^{-1}(\mathbb{E}[Y|\mathbf{X} = \mathbf{x}]) = \mathbf{x}^T \boldsymbol{\beta}$$

...a.k.a. assuming a *mean model*, as well as assuming independent observations. This is a *semi-parametric* approach.

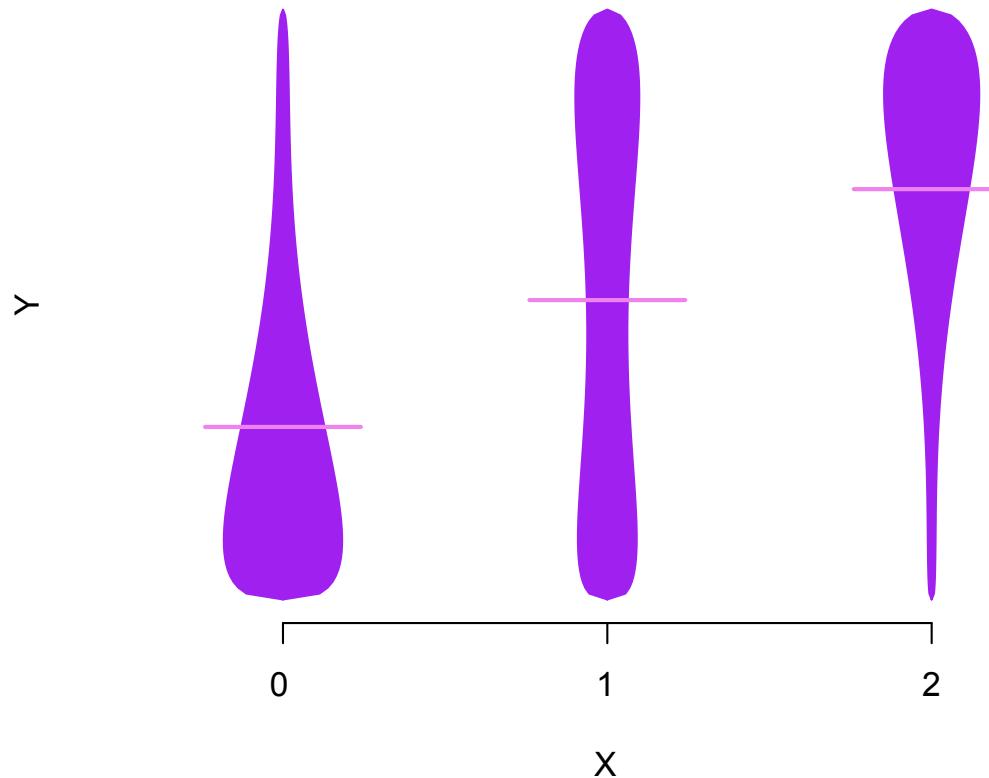
- If \mathbf{X} contains an intercept term (all 1s), its coefficient is the transformed mean – i.e. $g^{-1}(\mathbb{E}[Y|\mathbf{X}])$ – when all non-intercept terms are zero
- For $\mathbf{X} = \{X_1, X_2, \dots, X_k\}$, and $\boldsymbol{\beta} = \{\beta_1, \beta_2, \dots, \beta_k\}$, β_j is then the difference in $g^{-1}(\mathbb{E}[Y|\mathbf{X}])$, per 1-unit difference in X_j , in observations with all other covariates identical

Isolating the ‘effect’ on Y due to X_j alone is key, when we want to avoid the meaning of this association being *confounded* by other covariates.

Including all the other covariates is a.k.a. ‘adjusting’ for them.

Defining parameters: more assumptions

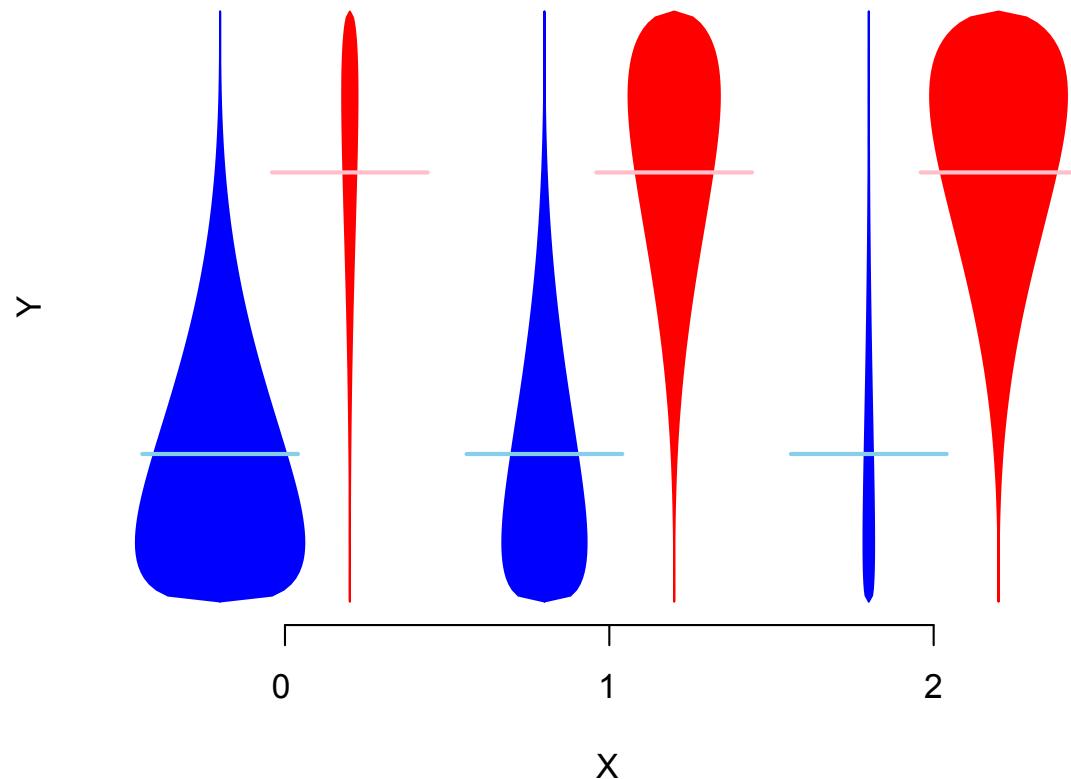
A quick example, from my applied work;



- Picture shows the population – X (genotype) takes only values 0/1/2 ... the number of copies of the minor allele of a genetic variant. The lines show $\mathbb{E}[Y|X = x]$
- Outcome Y is associated with genotype X

Defining parameters: more assumptions

A quick example, from my applied work;



- Outcome Y *is* associated with genotype X ... but only because Y and X are both affected by race/ethnicity Z
- Adjusting or not, we can get valid statistical inference. But only the adjusted result is of scientific interest

Defining parameters: more assumptions

Other key points;

- Adjusting only stops confounding of estimate if you:
 - Measured and adjusted for all relevant confounders (!)
 - Specified the mean model correctly, including the link function $g^{-1}()$ & the representation of all adjustment variables, including non-linear terms and/or interactions
- Flexible representations (e.g. splines) of confounders help get the model right (or right-enough). So use them!
- For testing, need right model **under the null**
- Diagnostics can help spot terrible model-misspecification, but do not rely on them having enough power to spot all problems. Expect background scientific knowledge to be crucial for choosing a mean model

Keen people: formal theory on the inability of data to tell you the ‘right’ model is a recent **research topic**.

Defining parameters: more assumptions

Assuming a mean model permits more flexibility in estimating function. In particular, if

$$\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] = g(\mathbf{x}^T \boldsymbol{\beta}), \forall \mathbf{x} \in \mathbb{R}^p,$$

$$\text{then } \mathbb{E}[Y - g(\mathbf{x}^T \boldsymbol{\beta})|\mathbf{X} = \mathbf{x}] = 0,$$

$$\text{and so } \mathbb{E}[h(\mathbf{x})(Y - g(\mathbf{x}^T \boldsymbol{\beta}))|\mathbf{X} = \mathbf{x}] = 0$$

for any (sane) function $h()$ mapping \mathbb{R}^p to \mathbb{R}^p . We define the same parameter $\boldsymbol{\beta}$ regardless of the ‘weights’ defined by $h()$.

As we are allowed *any* X values, we can restrict to those seen in the actual data – i.e. define $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta} \text{ such that } \frac{1}{n} \sum_{i=1}^n \mathbb{E}[h(\mathbf{x}_i)(Y_i - g(\mathbf{x}_i^T \boldsymbol{\beta}))|\mathbf{X} = \mathbf{x}_i] = 0$$

... and again, this is always the same parameter.

The mean model tells us that what holds for some \mathbf{x} values holds for all of them. So with a mean model, we can (justifiably) ignore uncertainty due to sampling the covariates.

Defining parameters: more assumptions

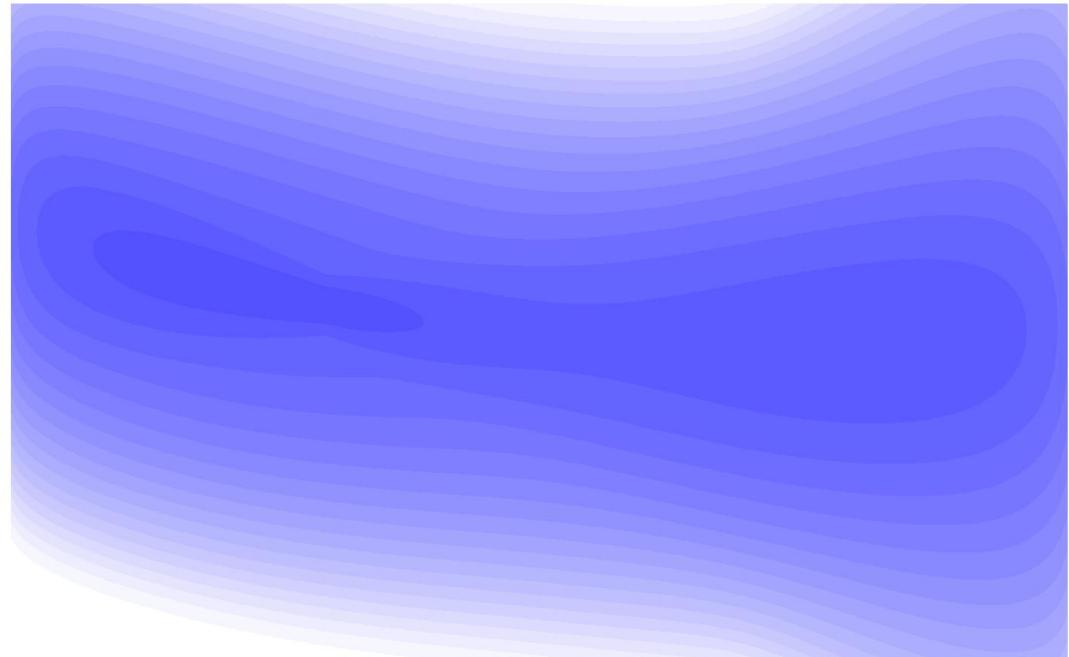
If this isn't obvious, consider using linear regression to define a quadratic,

$$\boldsymbol{\beta} = \underset{\beta_0, \beta_1, \beta_2}{\operatorname{argmin}} \mathbb{E}[(Y - \beta_0 - \beta_1 X - \beta_2 X^2)^2]$$

What happens
to the curve if
(right) we only
sample...

y

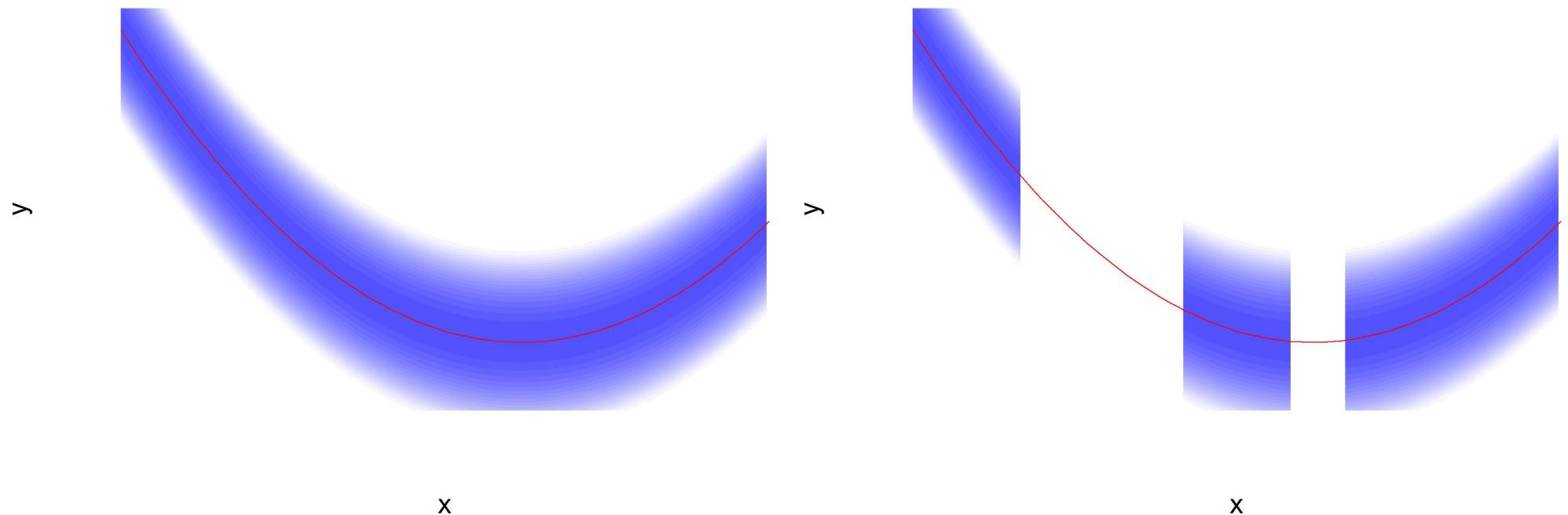
- small X ?
- large X
- just three
values of
 X ?



x

Defining parameters: more assumptions

Assuming a mean model, which X we sample matters somewhat less – because the same β governs $\mathbb{E}[Y|X]$ everywhere;



Q If you can fix X by design, what choices will increase the precision of the resulting $\hat{\beta}$?

Inference: with more assumptions

The same empirical ‘plug-in’ as before motivates the estimate

$$\hat{\beta} \text{ such that } \frac{1}{n} \sum_{i=1}^n h(x_i)(Y_i - g(\hat{x}_i^T \beta)) = 0$$

- The same sandwich and/or bootstrap can be used as before
 - but estimates of $\text{Var}[\hat{\beta}]$ may be (slightly) too large, as these allow for random \mathbf{X}
- ‘Fixed- \mathbf{X} ’ sandwiches do exist (see e.g. Szpiro et al, 2010), but require coding – usually for very minor benefit, if any
- Can bootstrap residuals $e_i = Y_i - g(x_i^T \hat{\beta})$, but this also assumes e_i are i.i.d. – stronger than just a mean model
- Which $h()$ to use? Godambe & Heyde (1987) showed that

$$h(\mathbf{X}) \propto \frac{\partial}{\partial \beta^T} g(x_i^T \beta) \text{Var}[Y_i | \mathbf{X} = x_i]^{-1}$$

gives optimal estimates $\hat{\beta}$ (in a sense). This motivates thinking about what $\text{Var}[Y | \mathbf{X} = x]$ might be...

Inference: quasi-likelihood

Adding on more assumptions, quasi-likelihood methods assume that the variance is a function of the mean – often a function with free parameters. For example, assuming that

$$\begin{aligned}\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] &= e^{\mathbf{x}^T \boldsymbol{\beta}}, \text{ defined as } \mu_{\mathbf{x}} \\ \text{Var}[Y|\mathbf{X} = \mathbf{x}] &= \alpha \mu_{\mathbf{x}}\end{aligned}$$

leads to *quasi-Poisson* regression – with a canonical, log link function.

- QL uses efficient estimates, as per Godambe-Heyde
- The ‘scale’ parameter α has to be estimated, using *ad hoc* methods. If $\hat{\boldsymbol{\beta}}$ depends on α , iterate until convergence
- Under correct (enough) assumptions, QL outperforms the sandwich: as ‘bread’ $\mathbf{A} =$ ‘meat’ \mathbf{B} , asymptotic variance $\text{Var}[\hat{\boldsymbol{\beta}}]$ simplifies to just \mathbf{A}^{-1}/n . Basically, QL has less to estimate, so it works better than the sandwich, with small n
- Can force $\alpha = 1$ if confident in assumptions (or if Y is binary)

Note: Bayesian semi-parametric methods exist, but are beyond the scope of 570/571, and are not used much in practice.

Inference: quasi-likelihood

Two very important special cases of QL:

$$\begin{aligned}\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] &= \mathbf{x}^T \boldsymbol{\beta} \\ \text{Var}[Y|\mathbf{X} = \mathbf{x}] &= \sigma^2\end{aligned}$$

#1 *Linear regression*, assuming constant variance (*homoskedasticity*) – it gives same results as likelihood-based use of classical linear models. Hence, *heteroskedasticity* can be an issue for linear regression but – despite many authors claiming the opposite – Normality is not*.

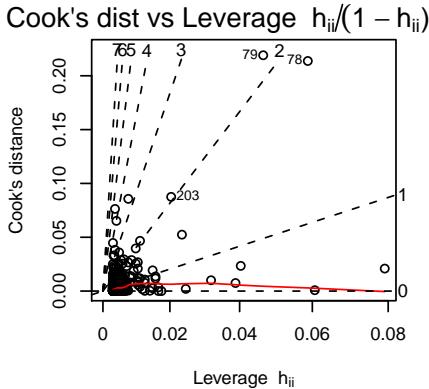
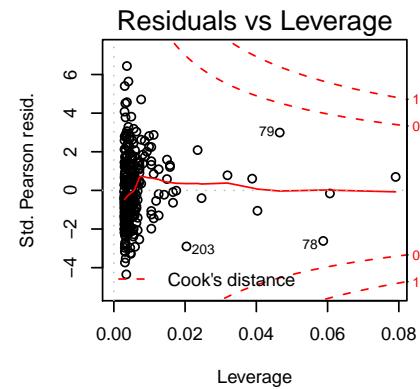
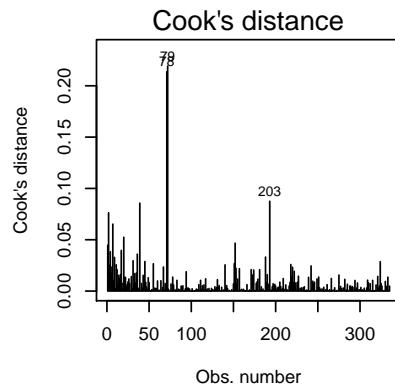
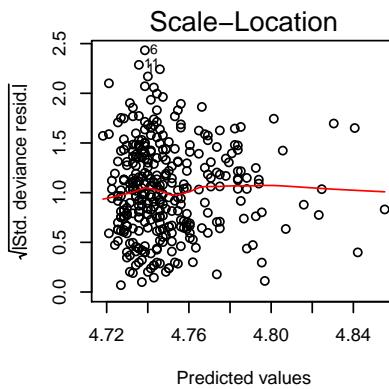
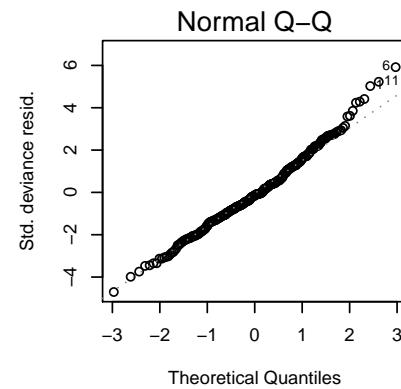
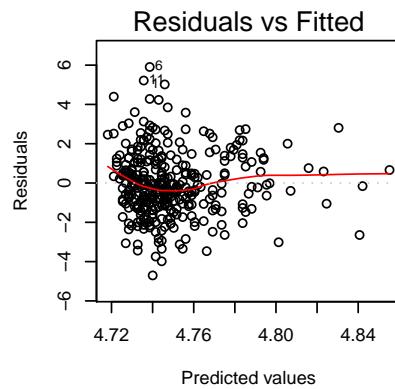
$$\begin{aligned}Y &\in \{0, 1\} \\ \mathbb{E}[Y|\mathbf{X} = \mathbf{x}] &= \text{expit}(\mathbf{x}^T \boldsymbol{\beta}) = \mu_{\mathbf{x}} \\ \text{Var}[Y|\mathbf{X} = \mathbf{x}] &= \mu_{\mathbf{x}}(1 - \mu_{\mathbf{x}})\end{aligned}$$

#2 *Logistic regression* – for binary Y , the mean-variance relationship is **known**, so it **should** be assumed.

*...except *perhaps* with small samples, where calibration of standard error estimates/ p -values may be an issue. As small samples don't give you power to detect non-Normality, diagnostics are not a reliable 'fix' for this.

Inference: quasi-likelihood

Diagnostics can spot **large** deviations from assumptions – the same basic residual plots are standard for checking first and second moment assumptions. Q. Which plots are they?



NB: don't forget `iter=0` using `lowess()`, if making your own.

Defining parameters: with full models

The basic steps here should be familiar from Stat 512/513;

- You assume that $Y_i|\mathbf{X}_i = x_i \sim F_i$, with density $f(y_i|x_i; \boldsymbol{\theta})$, for unknown parameter $\boldsymbol{\theta}$
- Finding

$$\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} \prod_{i=1}^n f(y_i|x_i; \boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^n \log f(y_i|x_i; \boldsymbol{\theta})$$

defines the *maximum likelihood estimate*

- MLEs consistent, & efficient with large n (\pm regularity)
- Estimated Fisher information \mathcal{I} estimates $\operatorname{Var}[\hat{\boldsymbol{\theta}}]$ and hence provides confidence intervals, in large samples. Or parametric bootstrap, generating from $f(y_i|x_i, \hat{\boldsymbol{\theta}})$

If your assumptions – i.e. the form of F – were correct (or correct-enough) this *parametric* approach will generally outperform QL & non-parametric approaches.

Note: the model **still** has to be chosen based on scientific interest, i.e. **you** must choose to model $Y|X$ or $Y|X, Z$.

Full models: GLMs

For *generalized linear models*, some blurring of the lines between parametric, semi-parametric and non-parametric frequentist inference occurs.

Recall: GLMs assume distributions of exponential family form;

$$f(Y_i; \theta_i, \alpha) = \exp\left(\frac{Y_i \theta_i - b(\theta_i)}{\alpha} + c(Y_i, \alpha)\right)$$

for $\theta_i, \alpha \in \mathbb{R}$, and functions $b(\cdot), c(\cdot)$ which map to \mathbb{R} . Covariates are introduced by assuming $b'(\theta_i) = g(x_i^T \beta)$, usually denoted μ_i , i.e. through a mean model assumption.

By the chain rule, the Score function for a GLM is

$$S(\beta) = \sum_{i=1}^n \frac{\partial}{\partial \beta} \log f(Y_i) = \sum_{i=1}^n \frac{\partial}{\partial \beta} l_i = \sum_{i=1}^n \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta}$$

Stochastic elements (specifically $Y_i - \mu_i$) only appear in the first term; the other terms tell us how the ‘canonical parameters’ depend on the mean, and through the mean on β .

Full models: GLMs

Because the estimating function depends on data only through $Y_i - \mu_i$, likelihood approaches for GLMs have some robustness – i.e. they don't need the entire model to be correct;

- If the GLM's mean model is specified correctly, the MLE is consistent for β
- If the GLM's mean-variance relationship is also correct, the MLE is an efficient choice (by Godambe-Heyde)

For *canonical link* GLMs, where $\theta_i = \mathbf{x}_i^T \boldsymbol{\beta}$, $S(\boldsymbol{\beta})$ simplifies to

$$S(\boldsymbol{\beta}) = \alpha^{-1} \sum_{i=1}^n \mathbf{x}_i^T (Y_i - g(\mathbf{x}_i^T \boldsymbol{\beta})),$$

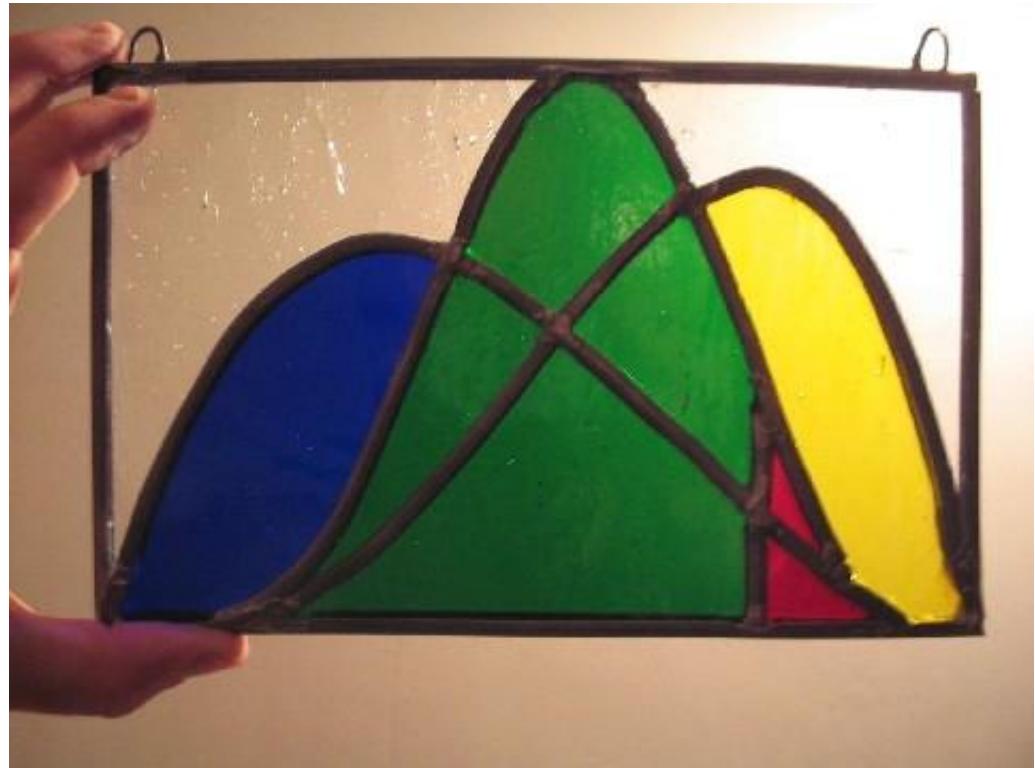
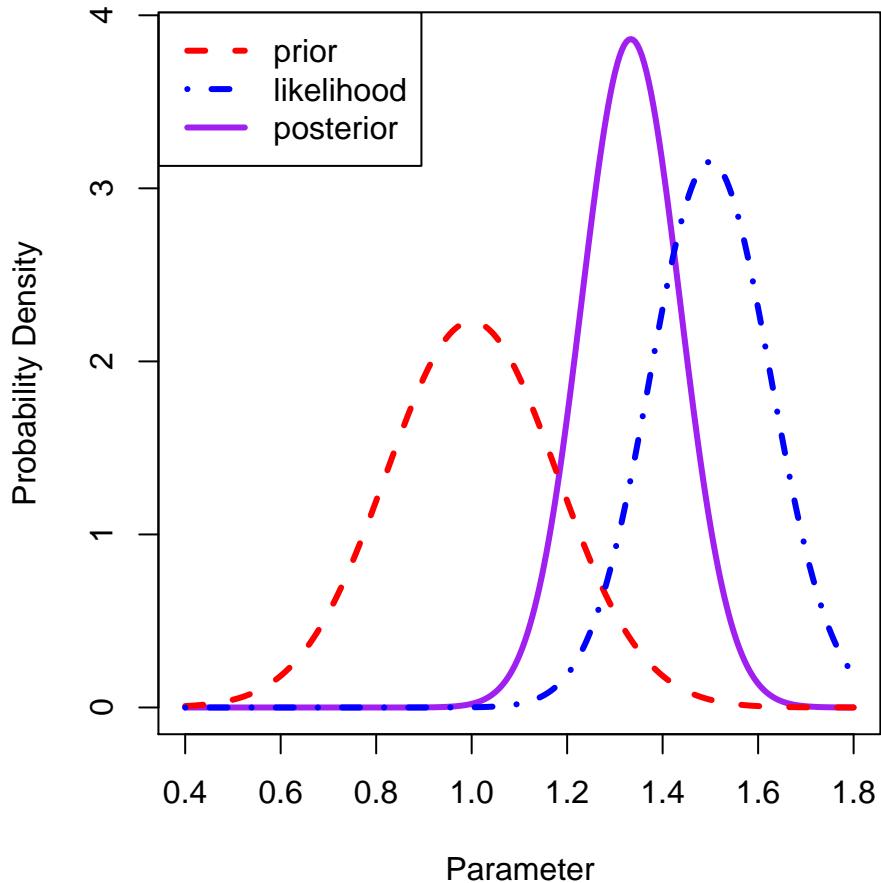
which has a unique root when no aliasing* occurs. In this case;

- If the GLM's mean and variance are correct, likelihood-based inference is valid in large samples
- Standard non-parametric sandwich approach coincides with versions that assume mean model (e.g. `sandwich` package)

* See Jon's Book, §5.5.2

Full models: Bayes

Bayes' Theorem (for univariate θ) made transparent;



The logic is elegant – “Common sense reduced to computation” ?

Bayes: paradigm

Just as in model-based frequentist work, an **assumed** model relates outcome Y to covariates X . Formally;

- What You know about the value of θ from **other sources** is expressed as Your *prior distribution*, $\pi(\theta)$
- Via Your choice of model, the *likelihood* $L(\theta) \equiv p(Y|\theta)$ is how the data inform You
- Putting these together, what You **know from both sources** is expressed as a *posterior distribution* $p(\theta|Y) \propto p(Y|\theta)\pi(\theta)$

Bayes uses the ‘language of probability’ to describe **belief**; non-negative, integrate to 1, and $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$. The actual parameters are (still!) fixed unknowns; belief in them is described in the same way as we describe random variables.

Note: Prior and posterior are standard names, but note no ordering in time is actually required.

Bayes: paradigm

To Bayesian purists, that's it! (essentially)

- No asymptotics: computing the posterior tells you what you want to know (or ‘should believe’) for the data you have
- Hypothetical replications not required: Get Your prior, do experiment(s), do inference using Your posterior
- Posterior for θ_1 is **always** given by integrating posterior with respect to nuisance parameters $\theta_2, \theta_3, \dots$; ‘average over what you don’t know’ – compare this with MLE
- The constructive aspect is great, but...
 - What summary of the posterior?
 - What prior? (including choice of model)
 - Why should anyone else care what You believe?

Also, computing the posterior can be a computational challenge.

Bayesian inference: deeper reasons (*)

As motivated here, Bayesian arguments follow from using probabilities to describe beliefs. They can also be motivated as (essentially) the only way to guarantee inference that is;

- Coherent: So long as the prior states Your belief, the posterior summarizes Your belief, for any parameter, and no self-contradiction can occur. An example of non-Bayesian *incoherence* is given by (optimal!) unbiased estimates of probabilities that take negative values.
- Admissible: For a stated measure of utility* the corresponding Bayes rule cannot be ‘beaten’, in expectation – see 581 for more.

In practice Bayesian analyses may sacrifice coherence/admissibility for other properties, e.g. robustness, simplicity, computability – or just what a field is familiar with.

* i.e. how good a given answer is, for possible values of the truth

Bayes: convincing others

Asymptotics (when you have large n) may help you convince others that Your prior is not driving everything. With fixed p , if the assumed model is correct...

- **Doob's Theorem:** Excepting sets of Θ with zero prior belief, Bayesian point estimates are consistent for the true θ
- **Bernstein-von Mises Theorem:** For large n , ‘efficient’ 95% confidence intervals are $\approx 95\%$ credible intervals, and vice versa

In short, for fixed p , eventually everyone agrees – and gets the right answer (subject to regularity conditions).

More pragmatically, for finite n , with a reasonable model, Bayesian methods give not-terrible frequentist coverage. It's therefore justified (and not uncommon) to use Bayesian intervals as frequentist ones, either as a spin-off of a Bayesian method – or where no frequentist tool works well.

Bayes: convincing others

With small n , if you want Your Bayesian analysis to convince others who may not (exactly) share Your prior, expect to do sensitivity analyses;

- Change the prior(s) – make it enthusiastic/skeptical
- Make light-tailed priors heavy-tailed, or vice versa
- Make components of θ correlated, *a priori*
- Include other nuisance parameters, and check whether inference on θ_1 is stable
- Try a different form of model (i.e. likelihood)
- Try model-averaging (i.e. a larger model with various sub-models) ... but then defend your prior probabilities for each sub-model

Unfortunately, which sensitivity analyses ‘work’ well is highly application-specific, and audience-specific.

Alternatively, argue strongly for Your choice of prior.

Bayes: computation

A different issue is calculation of the posterior. You should be familiar with;

- Direct evaluation (using conjugate priors, typically in GLMs of some form)
- Rejection sampling (direct sampling)
- Importance sampling
- Markov Chain Monte Carlo, particularly the Metropolis-Hastings and Gibbs samplers
- Gaussian quadrature and INLA

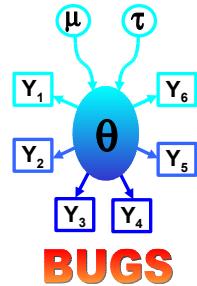
Direct coding of most of these can be error prone: as well as using sampler diagnostics (e.g. trace plots) and long runs, it's good practice, where possible, to compare your Bayesian output against related forms of analysis, e.g. MLEs.

For GLMs there are close connections between the posterior mean and parametric bootstrap output. See Efron ([2012](#), [2015](#))

Bayes: computation

Also note most MCMC calculations are already implemented in some R package... though it may not be obvious which.

Some general-purpose software, appropriate for 570 & 571;



WinBUGS – for Bayesian analysis Using Gibbs Sampling. (Though its sampling is not just Gibbs.) It is now ‘legacy’, but is widely-used & documented **BUGS**

JAGS – Just Another Gibbs Sampler. It does everything WinBUGS can do, but in C so it’s (somewhat) easier to adapt and extend



Stan – for Stanislaw Ulam. New, and implements more sophisticated samplers, and non-Bayesian methods.

All have R interfaces... and provide no warnings that you may be fitting a silly model, or guarantees that their MCMC will succeed.

Bayes: what summary?

Why might we be interested in the posterior mean $\mathbb{E}[\theta|\mathbf{Y}]$?

- Decision theory: loss function $(\theta - d)^2$...and others
- Computability – particularly INLA
- Interpretability

If $\mathbb{E}[\theta|\mathbf{Y}]$ differs from posterior median/mode – the other standard point estimates – say why, and perhaps compare with MLE. But major issues are rare. For testing life is less simple;

- Posterior probabilities give absolute support for e.g. $\mathcal{M} = 0$ and $\mathcal{M} = 1$
- Bayes Factors indicate the change in odds of support for $\mathcal{M} = 1$ over $\mathcal{M} = 0$, going from prior to posterior
- p -values need not behave like either of these. In regression, they *can* be interpreted as transformed approximations to Bayesian measures of signal-to-noise. ([Rice 2010](#))

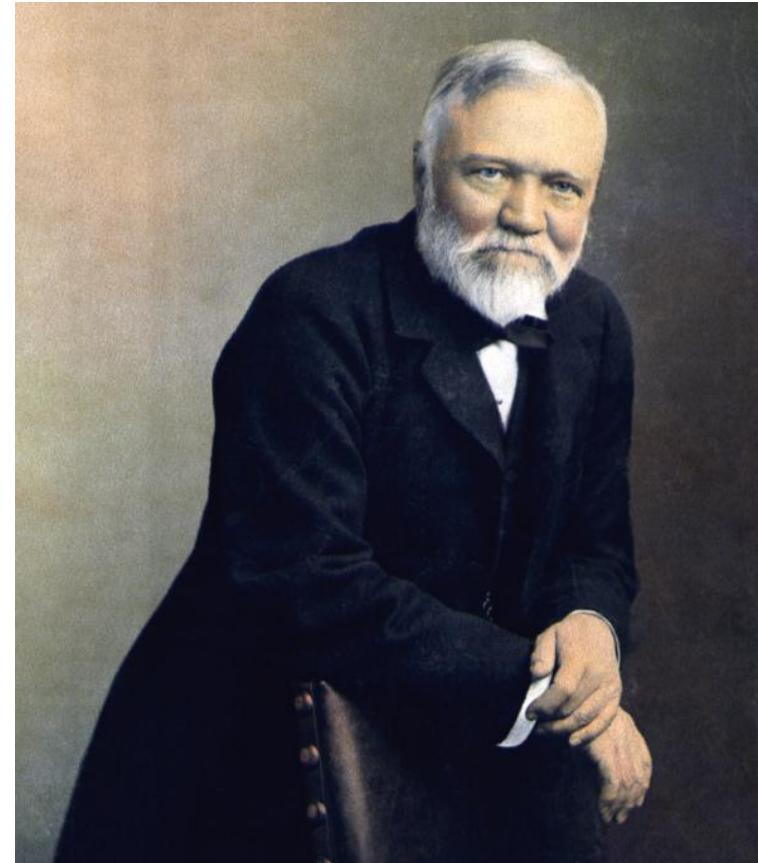
There is no single ‘right’ choice. Less worryingly, it’s often helpful – and not ‘wrong’ – to present multiple summaries.

Summary

- Inference means learning about populations
- We can do this with few assumptions, or with more. Efficiency and interpretability may be better if more assumptions are made – but at the cost of validity, if those assumptions are badly wrong
- Frequentist statements – while somewhat convoluted – provide inference without prior assumptions on the value of parameter θ
- Bayesian statements are (arguably) closer to most intuition, but do require a prior on θ . With limited n , expect the prior to matter
- Context matters, in all approaches; performing a reasonable analysis to answer a scientific question is only possible if you understand that question. There is no uniquely ‘right’ analysis for any given dataset – say what you did, and why

CHAPTER 1: MOTIVATION

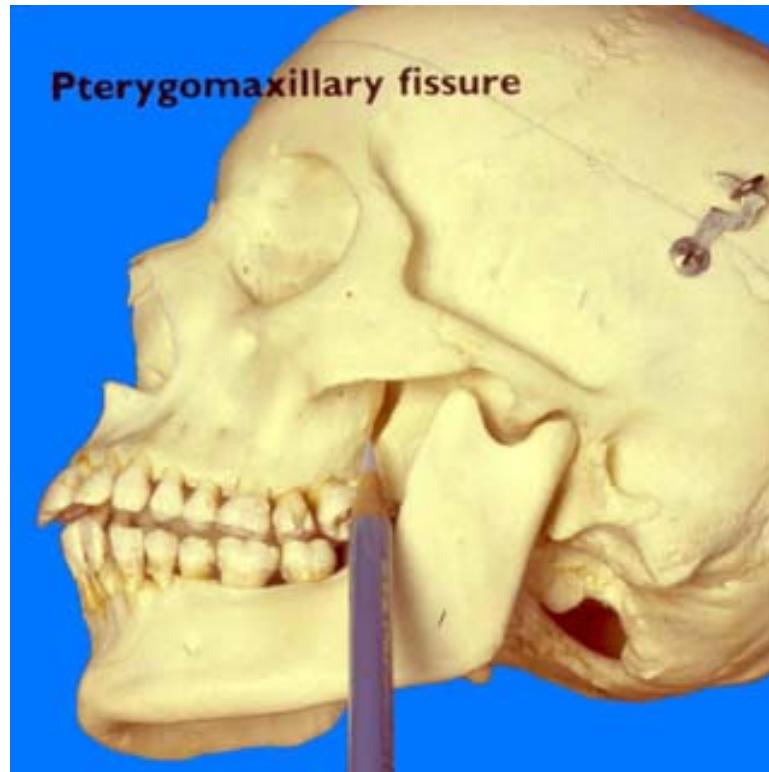
People who are unable to motivate themselves must be content with mediocrity, no matter how impressive their other talents



Andrew Carnegie (1835–1919)
Scottish-born industrialist and philanthropist

Motivation: Dental Growth

An example of content mediocrity? – the distance from the pterygomaxillary fissure to the center of the pituitary gland is obtained from x-rays;



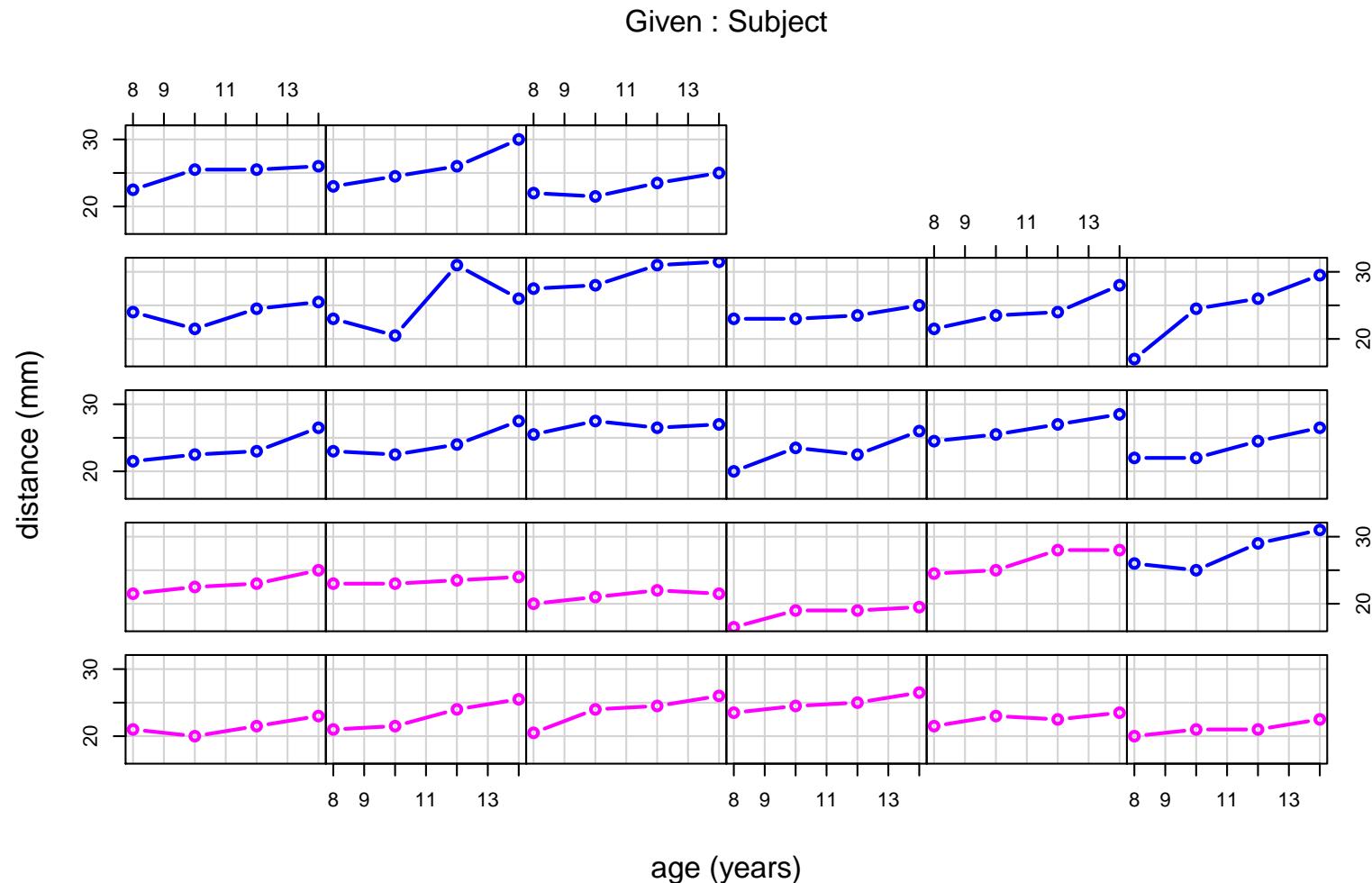
Motivation: Dental Growth

It's a measure of growth, useful for e.g. orthodontists. Here is the data, in mm, for $n=11$ girls and 16 boys;

Girls				Boys			
8 yrs	10 yrs	12 yrs	14 yrs	8 yrs	10 yrs	12 yrs	14 yrs
21	20	21.5	23	26	25	29	31
21	21.5	24	25.5	21.5	22.5	23	26.5
20.5	24	24.5	26	23	22.5	24	27.5
23.5	24.5	25	26.5	25.5	27.5	26.5	27
21.5	23	22.5	23.5	20	23.5	22.5	26
20	21	21	22.5	24.5	25.5	27	28.5
21.5	22.5	23	25	22	22	24.5	26.5
23	23	23.5	24	24	21.5	24.5	25.5
20	21	22	21.5	23	20.5	31	26
16.5	19	19	19.5	27.5	28	31	31.5
24.5	25	28	28	23	23	23.5	25
				21.5	23.5	24	28
				17	24.5	26	29.5
				22.5	25.5	25.5	26
				23	24.5	26	30
				22	21.5	23.5	25

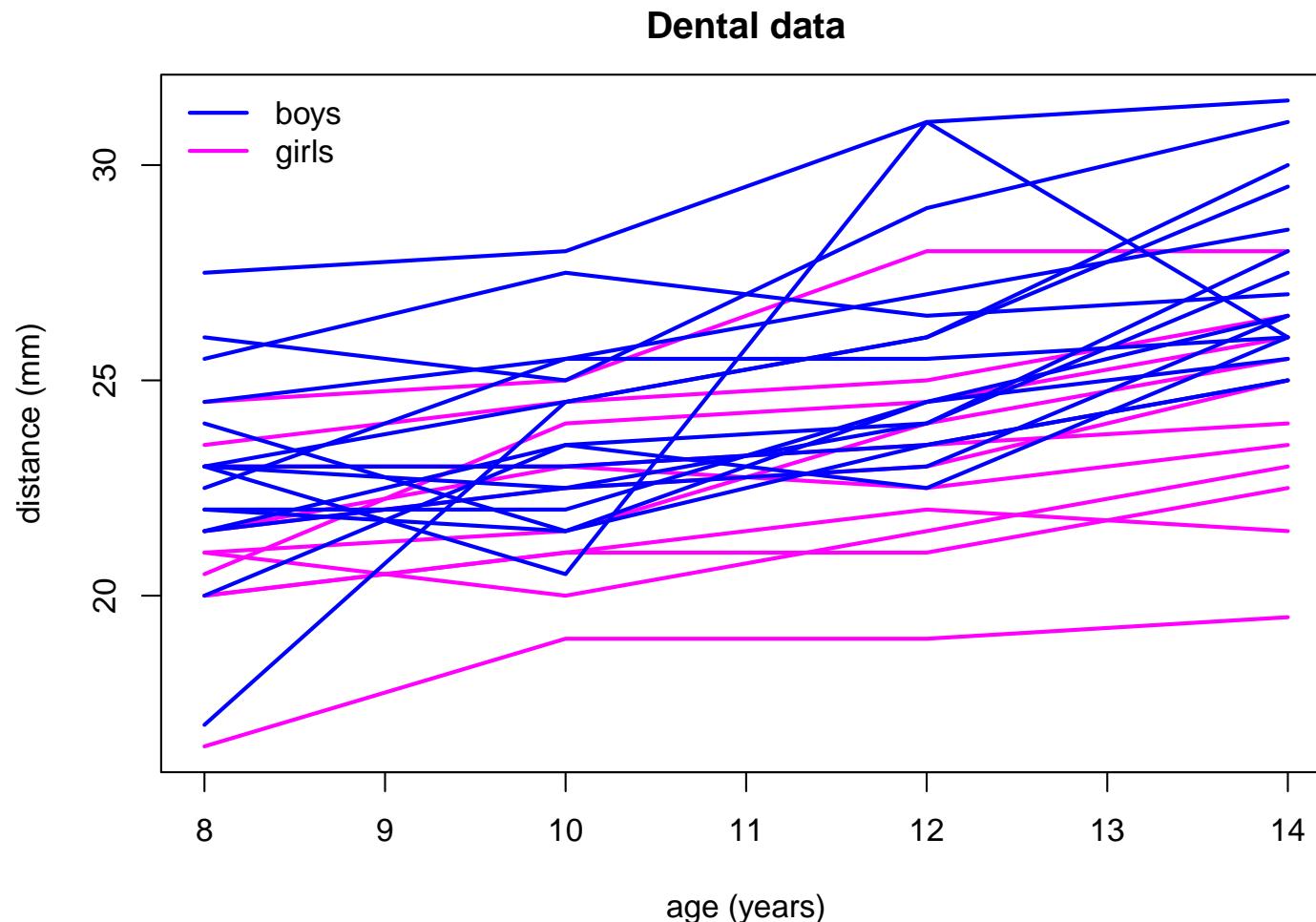
Motivation: Dental Growth

Displaying the data with coplot – `coplot(distance~age | Subject)`



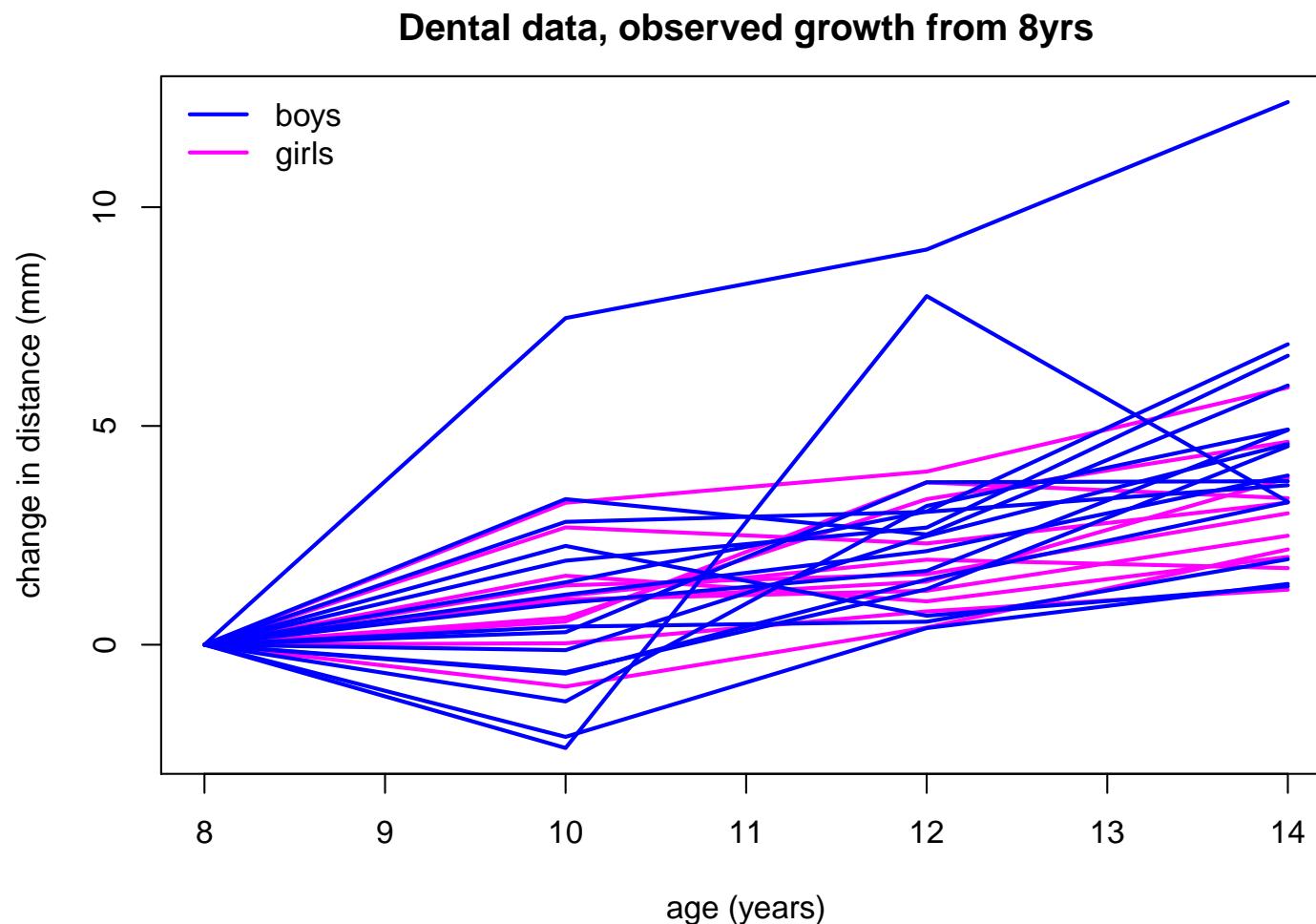
Motivation: Dental Growth

`matplotlib()` makes 'spaghetti plots' – for big n use transparent cols



Motivation: Dental Growth

Growth from 8 years old (slightly jittered, at 10+ years old);



Motivation: Dental Growth

Here are some scientific questions – match them to the graphs;

- How much ‘longer’ are distances, comparing e.g. 12 year-olds and 10 year-olds?
- What shape is the ‘average’ curve?
- Do boys or girls grow faster from age 8?
- Do boys or girls grow faster?
- What distance do you expect your patient (a boy, age 10, currently with $d=23\text{mm}$) to attain by age 14?
- Is variability in d better explained by within-child or between-child differences?

All of these are reasonable questions – but they require different analyses, even given the same data.*

* Recall e.g. causal inference in 570, there is no isomorphism between a dataset and its ‘correct’ analysis, context **must** be considered

Motivation: Dental Growth

We distinguish two importantly-different types of question;

- What is the mean difference in length, per 1-year difference in *age at observation*? **[marginal]**
- For a *specific child*, what is the growth curve? **[conditional]**

The former is a question about whole populations; the latter is about particular subjects (or types of subjects).

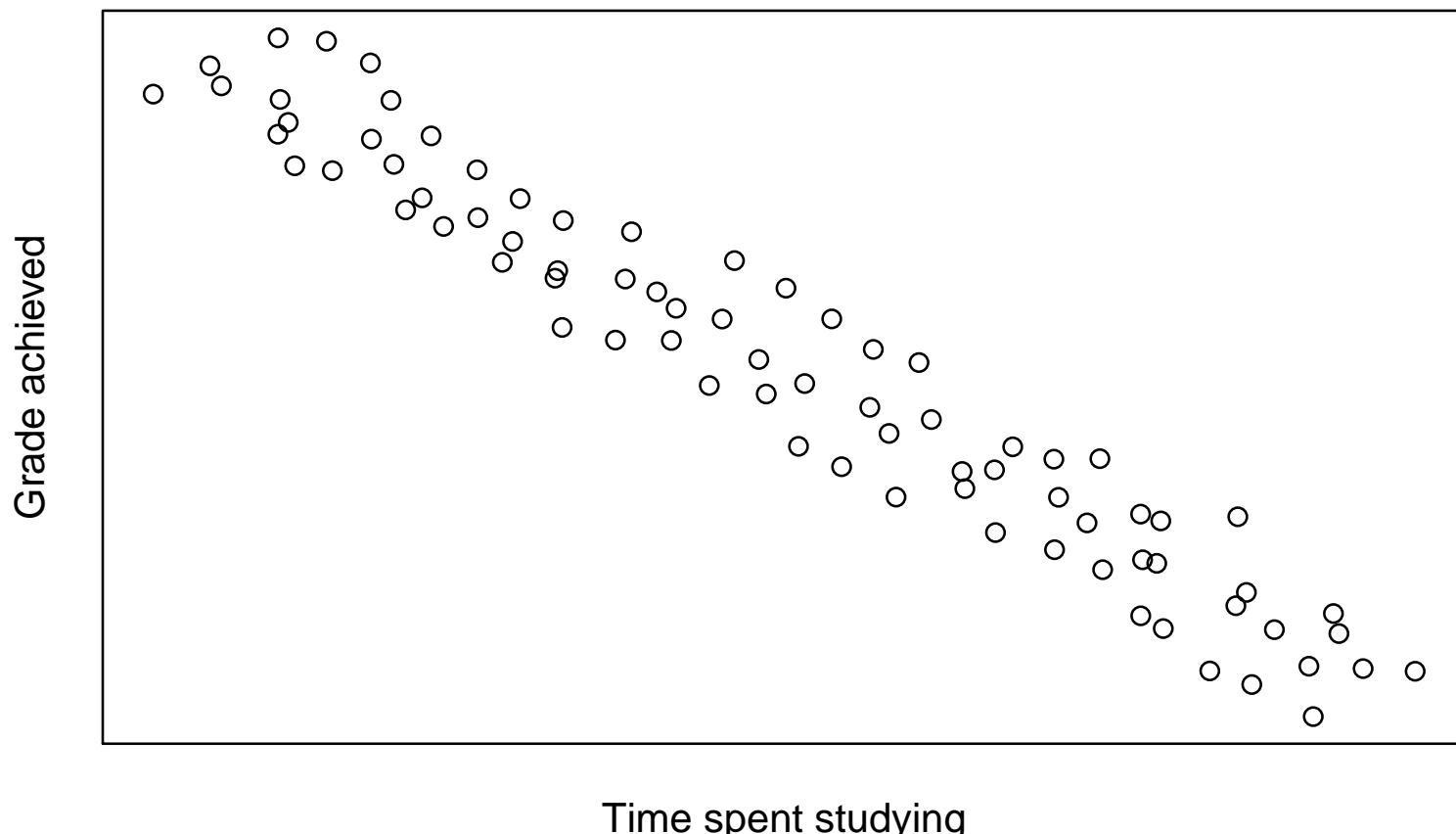
- Even for marginal inferences, it will still be invalid to treat the data as $27 \times 4 = 108$ independent outcomes; we actually have 27 sets of 4 correlated outcomes – **your** value of d_8 and d_{14} will be closer to each other than to **my** d_8, d_{14} .*
- This holds even if you and I are the same sex – though in e.g. males only, our data might be ‘closer’ than otherwise
- For either question, naively assuming independence can easily lead to *underestimates* of standard errors (anti-conservative) though over-estimation is also possible (loss of power)

* Blame your parents – or mine

Motivation: Conditional and marginal

Some made-up data, on a topic that concerns you;

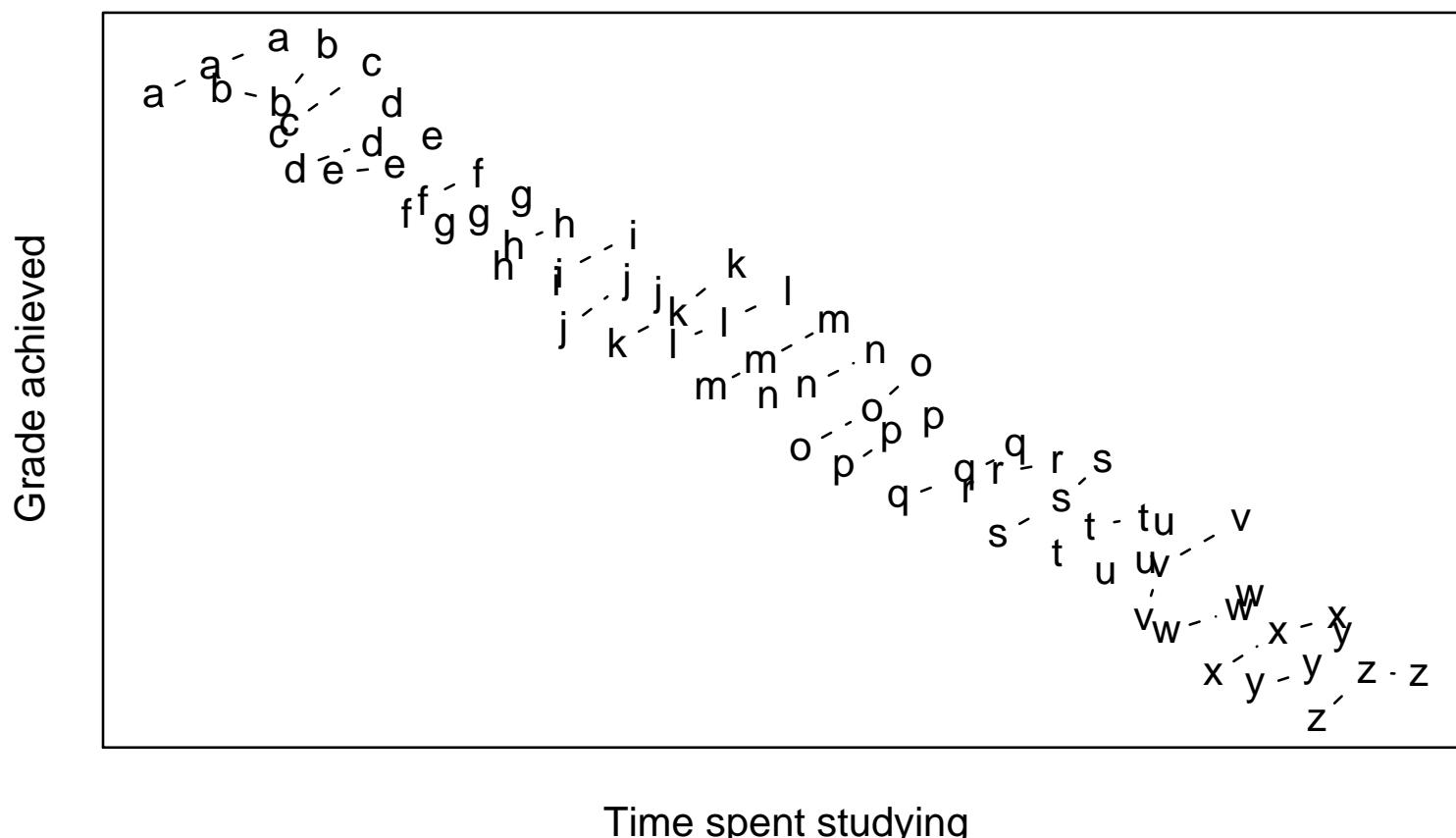
Does studying look helpful?



Motivation: Conditional and marginal

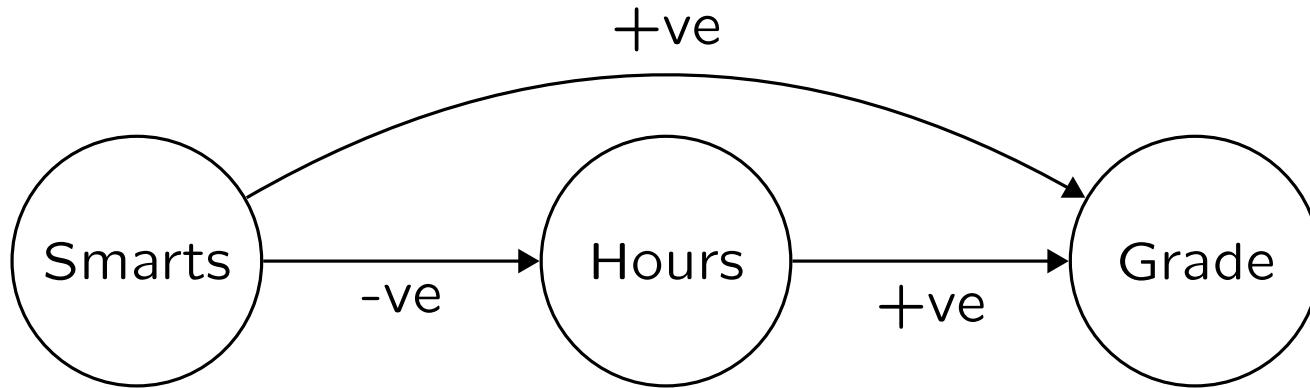
Some made-up data, on a topic that concerns you;

Does studying look helpful?



Motivation: Conditional and marginal

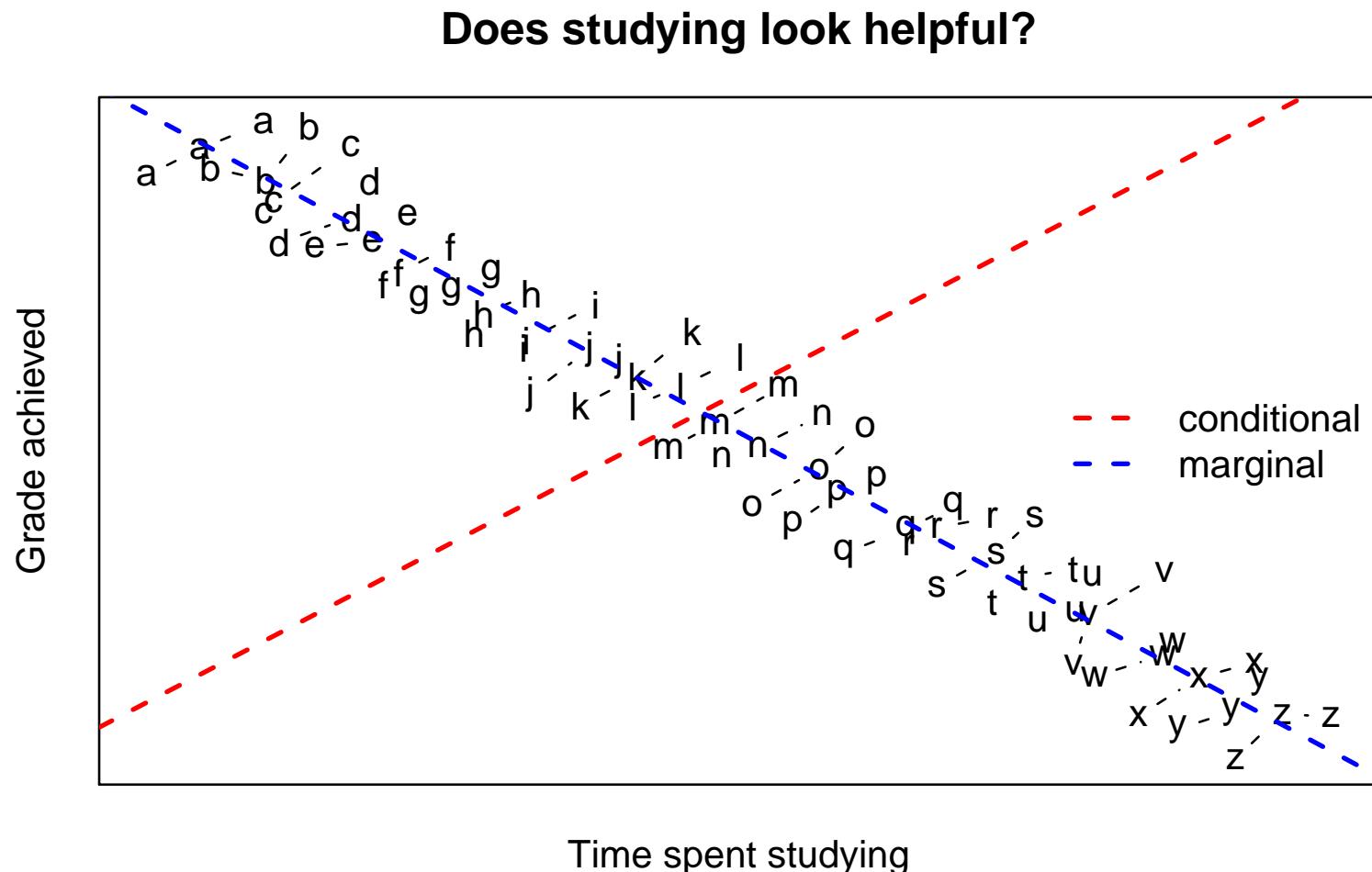
Plausibly, here's what's going on; arrows indicate causality



- On its own, Hours **is** negatively associated with Grade – note I only said ‘associated’
- The causal version of this statement (Hours decreases Grade) is wrong – a.k.a. confounding
- Keeping person fixed, Hours **is** positively associated with Grade
- Averaging over the whole population, Hours **is** negatively associated with Grade

Motivation: Conditional and marginal

Two lines illustrating different parameters;



Motivation: Conditional and marginal

- Marginal = ‘averaging over everything else’
- Conditional = ‘keeping everything else fixed’

Marginal and conditional statements are **not the same**.

- This is true regardless of whether you infer causality or not (so mistaking one for the other is not confounding)
- Both associations can be useful. The data will not tell you which is ‘right’, because ‘right’ depends on what you want to know. You must pick.
- Recall non-collapsibility; for parameters defined via linear operations, strata-specific and population-averaged versions may be numerically identical
- Expect debate: *Our experience over more than 20 years of modelling longitudinal data in social science and medical applications is that [...]the...] conditional formulation has a scientific supremacy* – Crouchley & Davies, JRSSA 2001

Motivation: Correlation structure

A subtler question we can *only* ask with correlated data;

*Q. How does the correlation change, as a function of time? **

We define the **semi-variogram function**

$$\delta(u) = \frac{1}{2} \text{Var}[Y(t) - Y(t - u)].$$

For a **stationary** process $\mathbb{E}[Y(t)] = \mathbb{E}[Y(u)]$, with constant or stationary variance $\text{Var}[Y(t)] = \text{Var}[Y(t - u)] = \sigma^2$ we get

$$\delta(u) = \frac{1}{2} (\sigma^2 + \sigma^2 - 2\text{Cov}[Y(t), Y(t - u)]) = \sigma^2 (1 - \rho(u))$$

where $\rho(u)$ is the **autocorrelation** of the process, at **lag** u .

We can estimate $\delta(u)$ empirically, using the sum of squared differences between all observations at lag u .

*... or other measure of distance between observations

Motivation: Correlation structure

Before an example, what to look for? Suppose we have residuals $Y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta}$ from

$$\mathbf{Y}_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \gamma_{0i} + W_i(t_{ij}) + \epsilon_{ij},$$

where we allow for three independent sources of variation;

$$\begin{aligned}\gamma_{0i} &\sim (0, \sigma_\gamma^2) && \text{Random intercept} \\ \epsilon_{ij} &\sim (0, \sigma_\epsilon^2) && \text{Measurement error}\end{aligned}$$

$$\text{Cov}[W_i(t_{ij}), W_i(t_{ij'})] = \sigma_W^2 \times \rho^{|t_{ij} - t_{ij'}|} \quad \text{Serial dependence}$$

- the serial dependence is one form of autoregressive (AR) model; correlation decays with ρ raised to the power u .

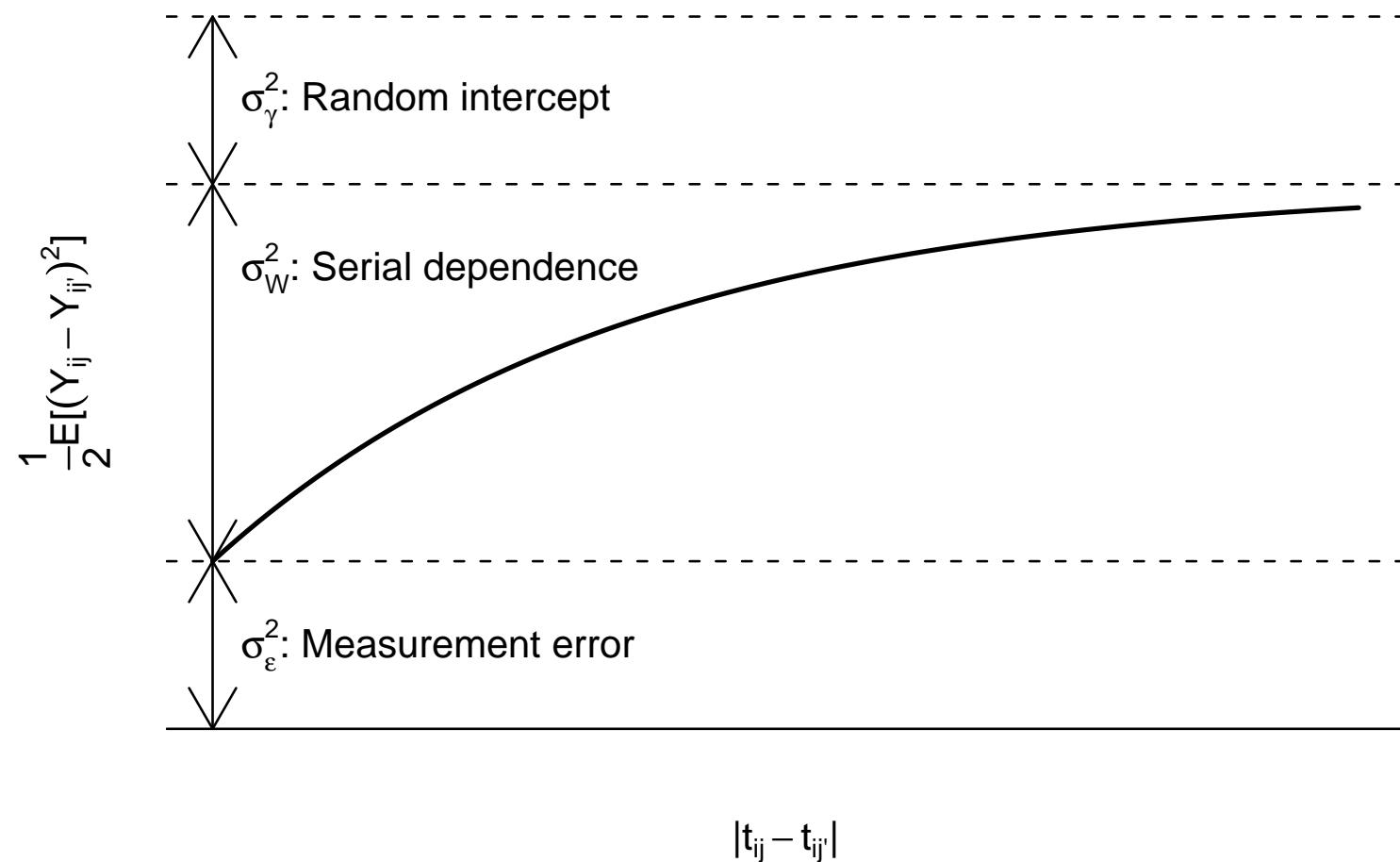
It can be shown [exercise] that

$$\text{Var}[Y_{ij}|\boldsymbol{\beta}] = \sigma_\gamma^2 + \sigma_W^2 + \sigma_\epsilon^2,$$

i.e. that the marginal variance is constant

Motivation: Correlation structure

The semi-variogram function, for this example;



Motivation: Correlation structure

- For this model, the semivariogram is

$$\delta(|t_{ij} - t_{ij'}|) = \sigma_W^2 \left(1 - \rho^{|t_{ij} - t_{ij'}|}\right) + \sigma_\epsilon^2$$

- All comparisons are *within-person* – so σ_γ^2 doesn't appear above
- We can still estimate σ_γ^2 – perhaps by subtraction from the marginal variance, σ^2
- For one implementation, see the `variogram()` function on the class site. There are many others, e.g. for a single time series default R has `acf()`, and several packages for analysis of spatial data look at how (co)variance decays with distance.
- Note the calculation involves a double loop, over lags u and people i .

Motivation: Correlation structure

For each lag u observed in the dataset, we can estimate $\delta(u)$ using fewer assumptions;

$$\hat{\delta}(u) = \frac{1}{2} \text{mean} \left\{ (Y_{ij} - Y_{ij'})^2 : |j - j'| = u \right\},$$

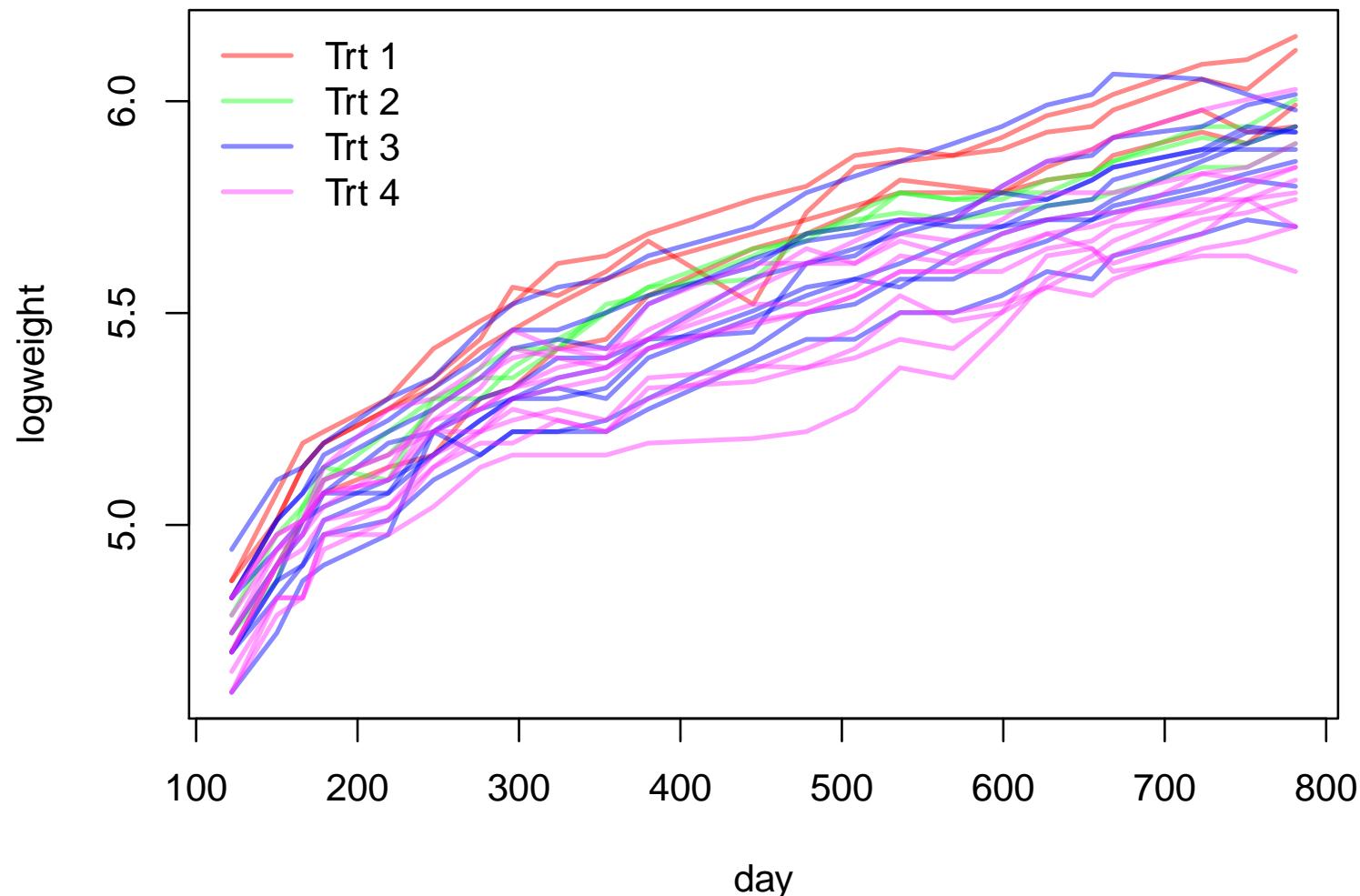
or in other words, half the mean of the squared differences of pairs of observations, where the pairs are within-cluster (same i) and are at lag u .

In most data sets this will be a ‘noisy’ curve; a moving-average smoother such as `lowess(, iter=0)` can be used to smooth out the empirical $\hat{\delta}(u)$.

Note: taking the mean at each lag u and then using `lowess()` can be ‘collapsed’ into one step; just plot all within-person squared differences against lag, and put a smoother through them.

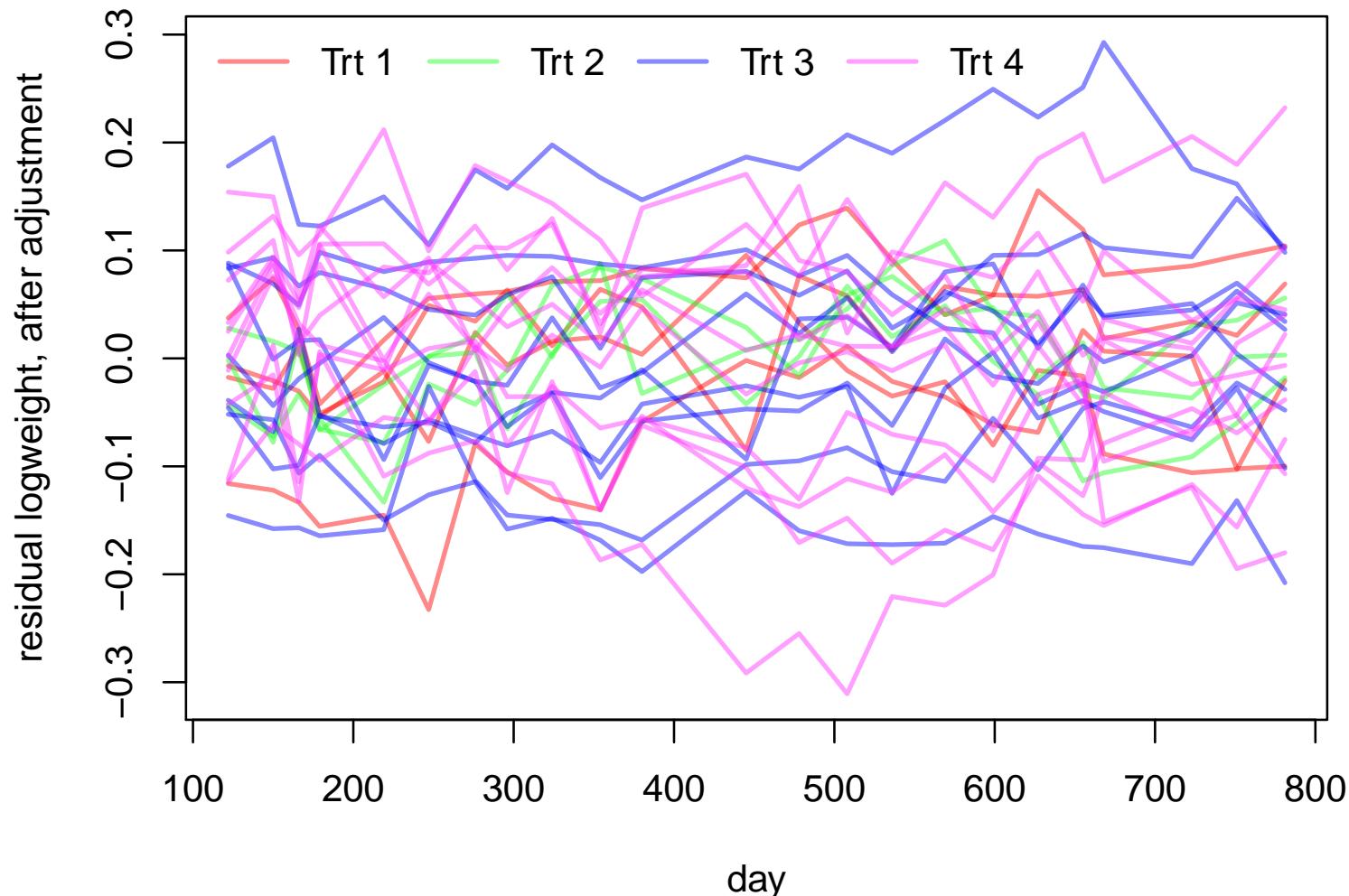
Motivation: Correlation structure

An udder example; the cows data, from DHLZ.



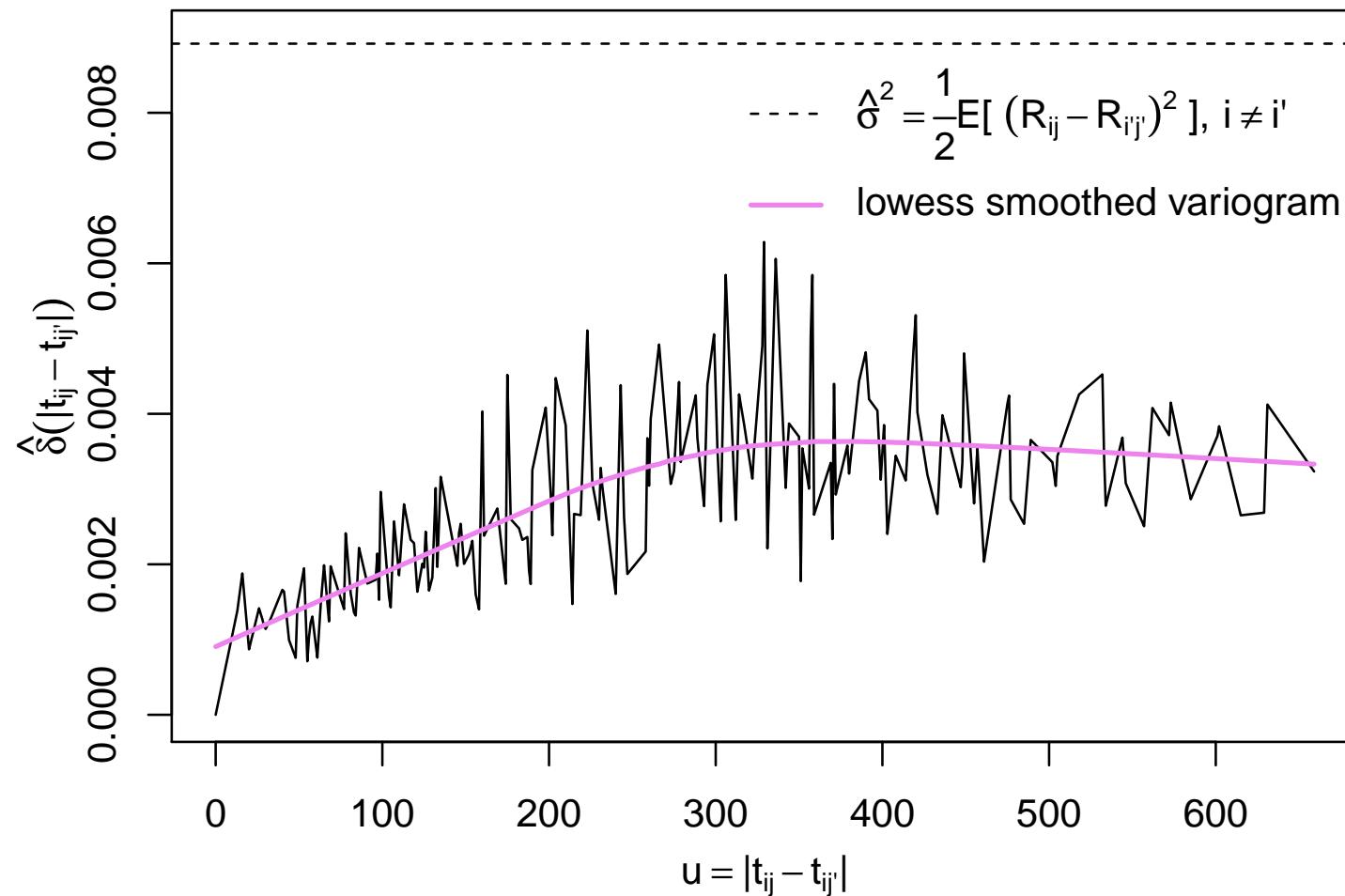
Motivation: Correlation structure

Subtract out means per day, and per treatment



Motivation: Correlation structure

The empirical variogram;



Motivation: Correlation structure

Some caution in all of this;

- This technique is **exploratory**. Beginners tend to over-interpret these plots (also other plots)
- The RH of the plot, i.e. the longest lag, has the fewest contributing data points
- If neither mean nor variance is stationary, the variogram becomes extremely hard to interpret; it may never ‘tail off’
- Non-linear transformation of the Y variable will give different answers
- The variogram ‘lag’ idea extends to e.g. longitude and latitude – but we run out of data fast, as dimension increases
- Contributions to variance at each lag (or in neighborhoods of lag values) will not be independent; this makes estimating uncertainty more difficult than in 570...

Motivation: Accounting for correlation

An only-slightly-stylized example; subjects in a trial receive treatment ($X=1$) or placebo ($X=0$) at two time points. At each time point, we measure some outcome Y . Interest lies in the difference in the mean value of Y comparing treatment to placebo.

Notation and assumptions;

- i indexes subjects, $1 \leq i \leq 2n$
- t indexes time, $t = 1, 2$
- Covariates are X_{it}
- Outcomes are Y_{it}
- All outcomes have same variance; $\text{Var}[Y_{it}] = \sigma^2$
- Outcomes in different subjects are independent
- Within-person correlation, $\text{Corr}[Y_{i1}, Y_{i2}] = \rho$

Motivation: Accounting for correlation

Here's a sane design; $\mathbf{X}_i = (0, 0)$ for n subjects, $\mathbf{X}_i = (1, 1)$ for the other n

i	Placebo		Treatment	
	$t = 1$	$t = 2$	$t = 1$	$t = 2$
1	$Y_{1,1}$	$Y_{1,2}$		
2	$Y_{2,1}$	$Y_{2,2}$		
:	:	:		
n	$Y_{n,1}$	$Y_{n,2}$		
$n + 1$			$Y_{n+1,1}$	$Y_{n+1,2}$
$n + 2$			$Y_{n+2,1}$	$Y_{n+2,2}$
:			:	:
$2n$			$Y_{2n,1}$	$Y_{2n,2}$

We can work out the average outcome under Treatment, under Placebo, and then calculate their difference

Motivation: Accounting for correlation

Here's the algebra;

$$\hat{\beta} = \underbrace{\frac{1}{2n} \sum_{i=n+1}^{2n} \sum_{t=1}^2 Y_{it}}_{\text{Treatment}} - \underbrace{\frac{1}{2n} \sum_{i=1}^n \sum_{t=1}^2 Y_{it}}_{\text{Placebo}}$$

Now recall our assumptions;

$$\text{Var}[Y_{i1} + Y_{i2}] = 2\sigma^2(1 + \rho)$$

... which means that

$$\text{Var}[\hat{\beta}] = \frac{n}{4n^2} 2\sigma^2(1 + \rho) + \frac{n}{4n^2} 2\sigma^2(1 + \rho) = \frac{\sigma^2}{n}(1 + \rho).$$

Naïvely treating all the data as independent, we would have said this variance was

$$\frac{\sigma^2}{2n} + \frac{\sigma^2}{2n} = \frac{\sigma^2}{n}$$

If $\rho > 0$ we'd **overstate** precision; if $\rho < 0$ it'd be **understated**...
of course, neither option is good.

Motivation: Accounting for correlation

Now another sane design; n subjects get placebo **then** treatment – $\mathbf{X}_i = (0, 1)$ – and n get the reverse, i.e. $\mathbf{X}_i = (1, 0)$

i	Placebo		Treatment	
	$t = 1$	$t = 2$	$t = 1$	$t = 2$
1	$Y_{1,1}$			$Y_{1,2}$
2	$Y_{2,1}$			$Y_{2,2}$
:	:			:
n	$Y_{n,1}$			$Y_{n,2}$
$n + 1$		$Y_{n+1,2}$	$Y_{n+1,1}$	
$n + 2$		$Y_{n+2,2}$	$Y_{n+2,1}$	
:		:		:
$2n$		$Y_{2n,2}$	$Y_{2n,1}$	

In the top group, take the later values from the former – reverse this in the lower group.

Motivation: Accounting for correlation

Again, here's the algebra;

$$\hat{\beta} = \frac{1}{2} \left(\underbrace{\frac{1}{n} \sum_{i=1}^n Y_{i,2} - Y_{i,1}}_{\text{top group}} + \underbrace{\frac{1}{n} \sum_{i=n+1}^{2n} Y_{i,1} - Y_{i,2}}_{\text{bottom group}} \right).$$

The variance of each summand is $2(1 - \rho)\sigma^2$. This means that

$$\text{Var}[\hat{\beta}] = \frac{\sigma^2}{n}(1 - \rho).$$

- The dependence on ρ is now reversed; positive ρ now means naïve approaches underestimate precision
- Whether correlation is ‘good’ or ‘bad’ – relative to just getting $4n$ independent observations – depends on the design
- As a consequence, even if methods which ignore the sign of ρ end up being valid, e.g. having Type I error rate ≤ 0.05 , they **must** be inefficient, compared to using the sign of ρ .

Motivation: Being likeli-hoodwinked

When tackling problems with nuisance parameters, naïve application of 513 ideas suggests adding a nuisance parameter for every cluster, a.k.a. ‘strata’.

An example*; you have two grades ($j = 1, 2$) from $1 \leq i \leq 26$ stata – each is a student. Let’s assume

$$Y_{ij} | \mu_i \stackrel{ind}{\sim} N(\mu_i, \sigma^2).$$

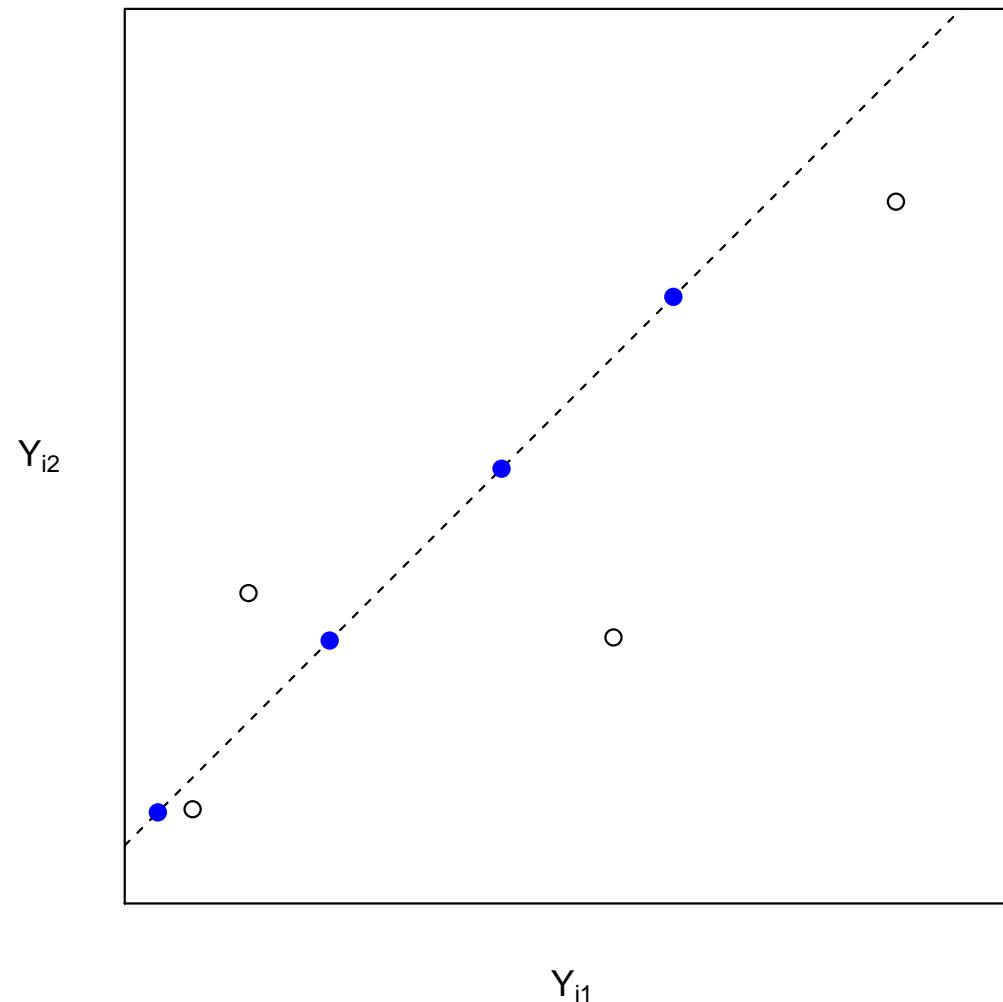
Keen to know the amount of random noise in the scores, you estimate σ , by maximizing w.r.t. σ and *all* the μ_i .

- As you might guess based on earlier warnings (570, 512/513) this gives *badly* inconsistent estimation of σ
- But why?

* Example due to Neyman and Scott (1946) who used an astronomy example. Fisher overlooked this problem with MLEs

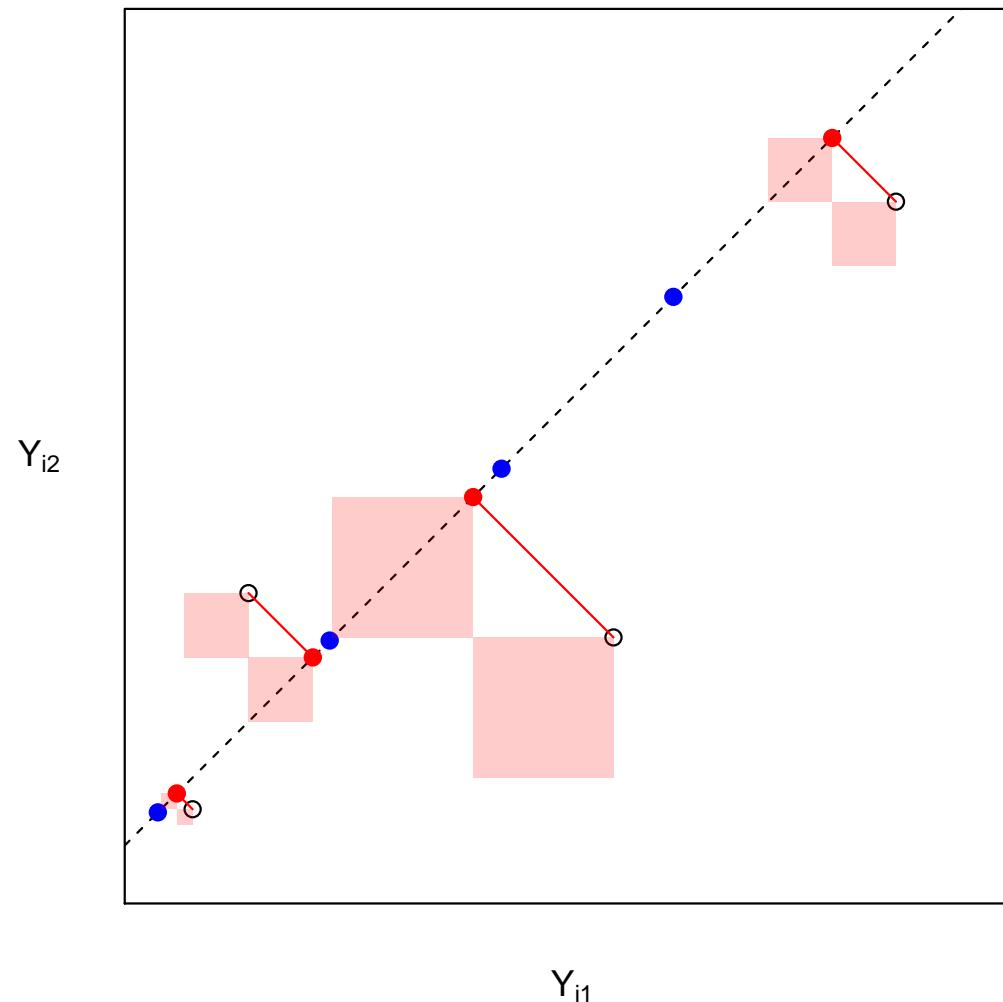
Motivation: Being likeli-hoodwinked

Four pairs of points, each pair has a common mean;



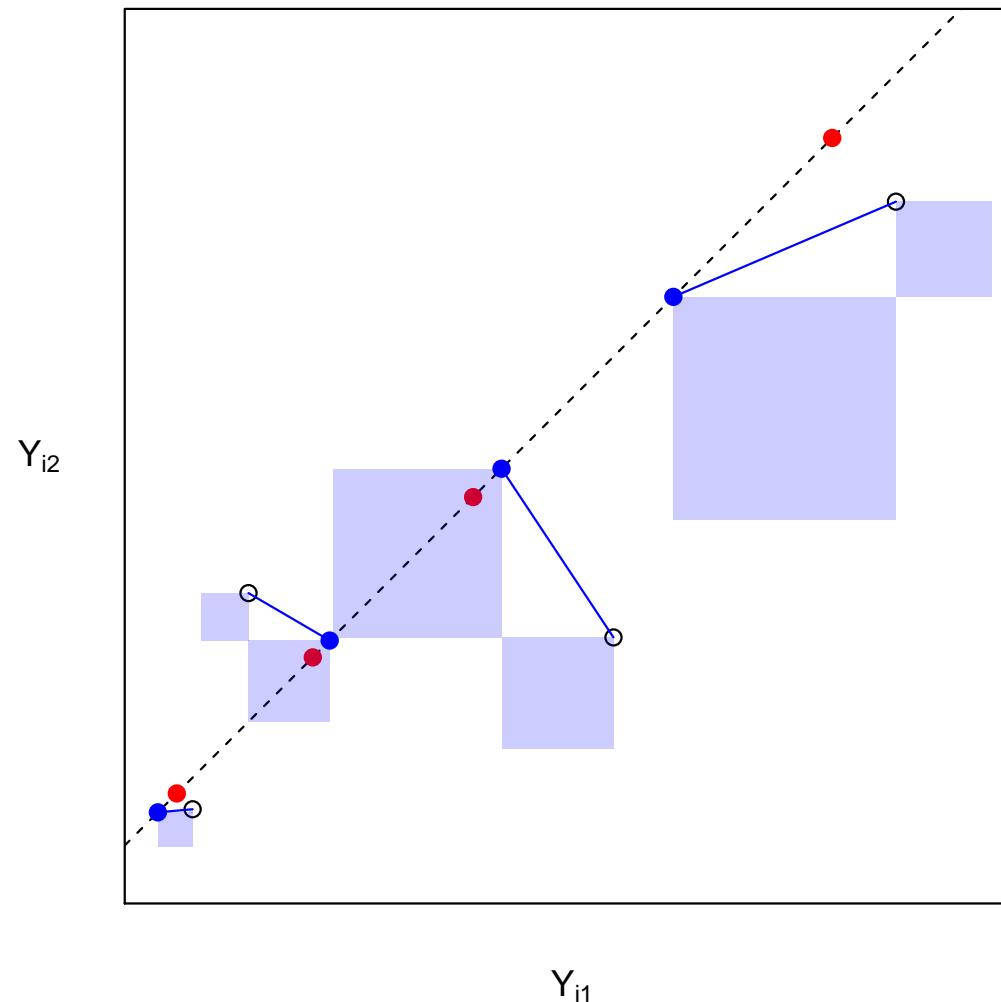
Motivation: Being likeli-hoodwinked

Squared deviations around pair-specific mean observations;



Motivation: Being likeli-hoodwinked

Squared deviations around pair-specific true means;



Motivation: Being likeli-hoodwinked

The pattern we see holds generally;

- We saw 50% less red ink (around observation means) than blue ink (around true means) – even with large n , so having more pairs of data does *not* fix this
- Fitting pair-specific means μ_i – one is estimated for each pair
- Average blue square area is a valid estimate of σ^2 – but the MLE estimate is the average red square area (!)
- i.e. sample variability underestimates true variability
- ... underestimates it a *lot*, if there are many small clusters, giving way-too-small (i.e. **wrong**) standard errors

In other situations (that we'll see later) this severe over-fitting also affects estimates of mean parameters – even though the model is 100% correct.

Also note this phenomenon doesn't depend on Normality, at all; pairs of i.i.d. data points tend to be closer to their pairwise average than to their true mean.

Motiv'n: Institutional Comparisons (*)

A calm, reasoned headline*;

Heart surgery death leagues bring new fears for patients

DAILY EXPRESS 

- Each hospital's performance was/is recorded, published – and acted on; 'naming and shaming' is part of it
- The press (and the public, and hospital administrators, and politicians) *love* this stuff
- Outcomes within a hospital are correlated – they share e.g. surgery and nursing practices, but also many other factors

* ... by the standards of the British press

Motiv'n: Institutional Comparisons (*)

NHS to run death rate leagues

Whitehall cautious as Scots start table

or individual surgeons' mortality rates on the grounds that people would fail to appreciate the complexities involved, particularly differences in the health of patients on admission.

Death rates among patients having gall bladder removal

the guardian

- Schools, hospitals, heart surgeons (in NY), gonnorhea clinics, GPs/mass murderers...
- This is now a **serious business!**

Schools in 'worst borough' improve



THE INDEPENDENT
Wednesday 10 March 1999

Motiv'n: Institutional Comparisons (*)

An understandably high-profile example;



Babies may have died needlessly

ONE hundred babies may have died needlessly following heart surgery at the Bristol Royal Infirmary, it was claimed today.

By PA News reporter

submission that in itself is justification for this inquiry."

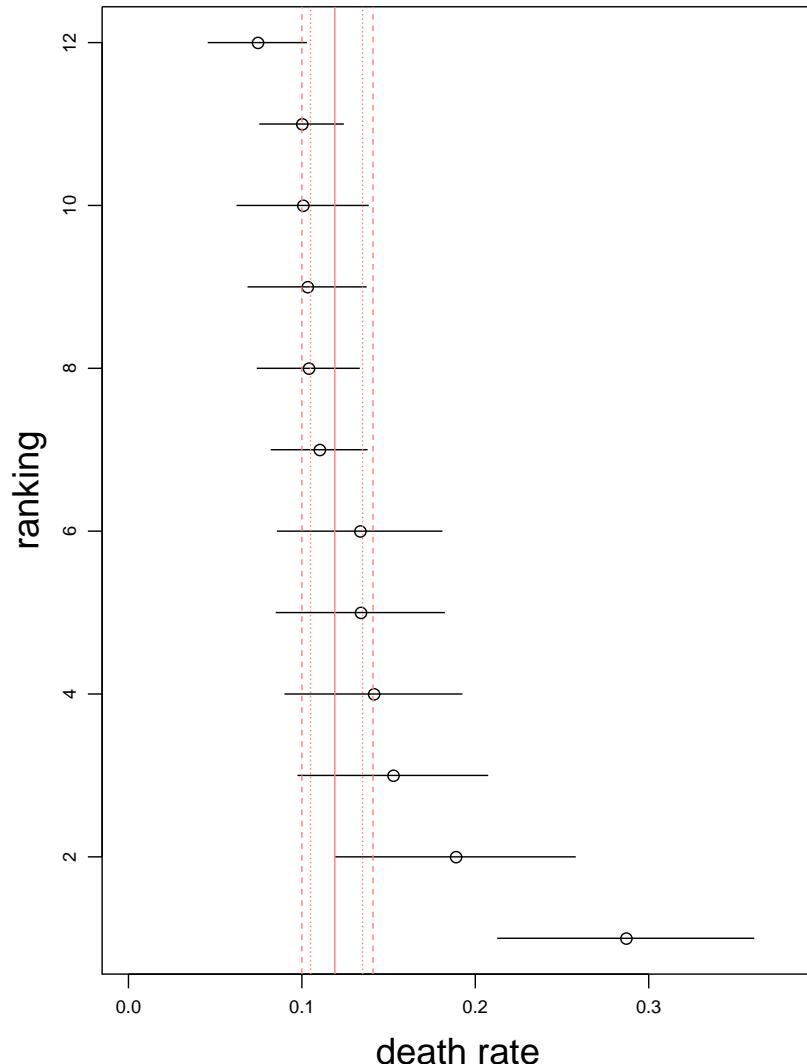
He added: "The events this inquiry is

investigation looked at operations on 53 children of whom 29 died and four were left brain damaged. This inquiry has now examined around 2,000 cases.

In his half-hour long submission he said Bristol failed in a number of areas

- Bristol Royal Infirmary 1984-1995; heart ops on under 1's
- "...more children died than might have been expected..."
[Public Inquiry]
- But *someone* has to be worst – with $p \approx 0.05$, if $n \approx 20$

Motiv'n: Institutional Comparisons (*)

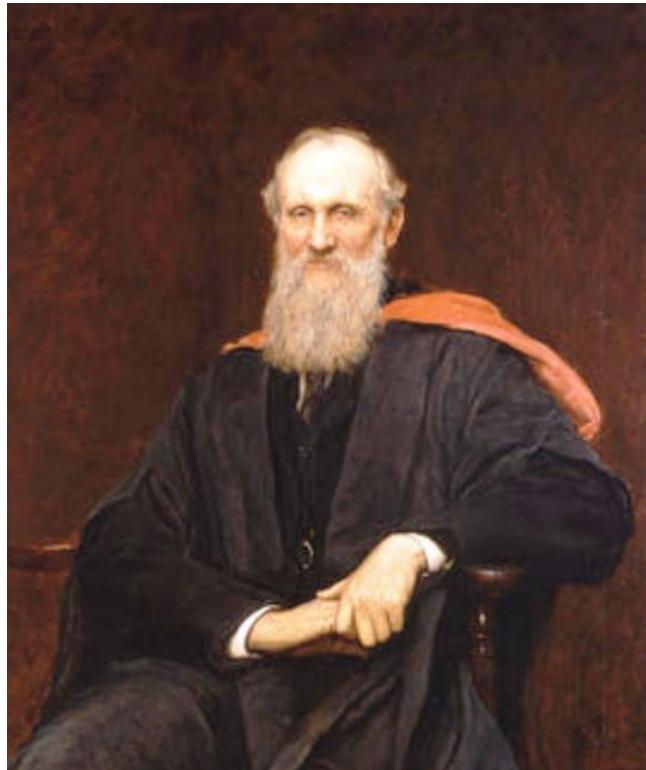


- Point estimates incorporate subject-specific covariates; ‘case-mix adjustment’
- Assuming all **hospitals** perform equally, get $\approx 70\%$ ‘outliers’, even with Bonferroni
- A sane model for hospital variation/unmeasured variables might be $N(0, \sigma^2)$
- ... even allowing for this, Bristol’s data is still way beyond chance

Summary

- Data with correlated outcomes are common, useful – and interesting. Just looking at the data is more complex than in 570
- Analyses **must** account for correlation, somehow. Approaches used for ‘fixed- p large- n ’ may not work
- There is considerable flexibility in what we *can* estimate, when there are correlated outcomes;
 - Good: more nuanced scientific understanding
 - Bad: statisticians have to explain the options
- Different levels of assumptions can be motivated/required; non-parametric, semi-parametric and parametric methods are available. Bayesian methods provide some useful ideas (exchangeability) and tools (MCMC)

CHAPTER 2: VECTOR OUTCOMES



Symmetrical equations are good in their place, but ‘vector’ is a useless survival, or offshoot from quaternions, and has never been of the slightest use to any creature.

William Thomson a.k.a. Lord Kelvin (1824–1907)
Scottish Physicist, pioneer of thermodynamics
(writing to G. F. Fitzgerald, in 1896)

CHAPTER 2: VECTOR OUTCOMES

As we have been discussing, non-parametrically;

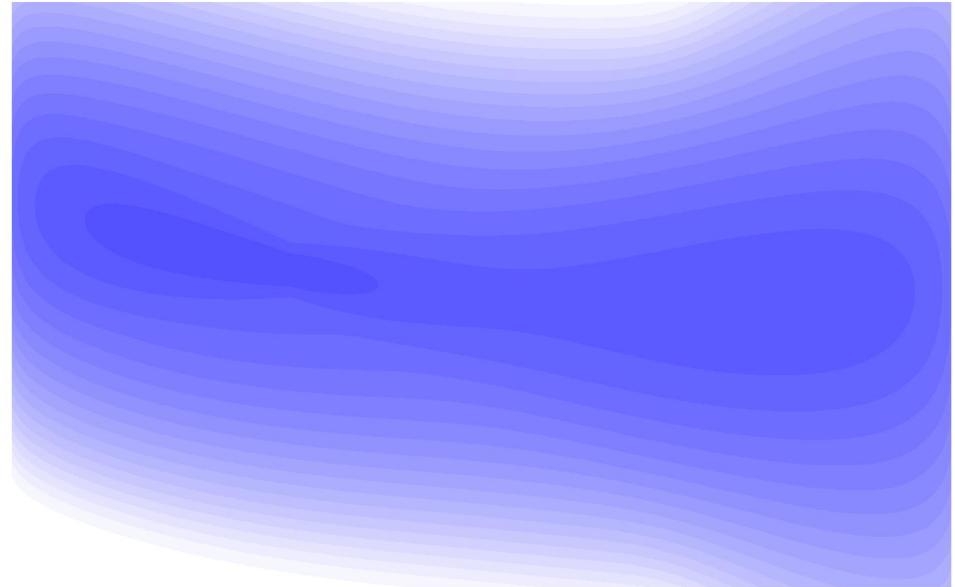
- Parameter θ maps superpopulation (right) to \mathbb{R}^p
- Define θ as the solution to

$$\mathbb{E}[G(Y, \mathbf{X}, \theta)] = 0$$

- Estimate θ via

$$\frac{1}{n} \sum_{i=1}^n G(Y, \mathbf{X}, \hat{\theta}) = 0,$$

i.e. empirically



x

- Asymptotic variance of $\hat{\theta}$ given by $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}/n$, where $\mathbf{A} = \mathbb{E}[\frac{\partial}{\partial \theta}G(Y, \mathbf{X}, \theta)]$ and $\mathbf{B} = \mathbb{E}[G(Y, \mathbf{X}, \theta)G(Y, \mathbf{X}, \theta)^T]$ are also estimated empirically

CHAPTER 2: VECTOR OUTCOMES

In 570, the independent samples were of the form $\{Y_i, \mathbf{X}_i\}$, with univariate Y_i . But we can apply **exactly** the same ideas if we sample independent *clusters* (see Chap 1: Motivation), that each contribute ≥ 1 outcome to the analysis.

It should be clear we can do *some* inference; boiling down each clusters's data to one observation $\{Y_i, \mathbf{X}_i\}$, we could apply familiar non-parametric EE approaches.

But we can do much more than this. Results which hold (asymptotically) for i.i.d. observations $\{Y_i, X_{i1}, X_{i2}, X_{i3}\}$ will also hold (asymptotically) for i.i.d. observations $\{Y_{i1}, Y_{i2}, X_{i1}, X_{i2}\}$ – provided the mild regularity conditions are met.

Mathematically, there is nothing particular to prove; we are just re-labeling from Y to X , then using familiar machinery.

Vector outcomes: notation

However, with multiple outcomes per cluster, each with multiple covariates, we must add subscripts to our notation;

Exp'n	Meaning	Contents
n	Number of clusters	
n_i	Number of outcomes in cluster i *	
\mathbf{Y}_i	n_i -vector of outcomes in cluster i	$(Y_{i1}, Y_{i2}, \dots, Y_{in_i})^T$
\mathbf{X}_i	$n_i \times p$ covariate matrix for cluster i	$(\mathbf{X}_{i1}, \mathbf{X}_{i2}, \dots, \mathbf{X}_{in_i})^T$
\mathbf{X}_{ij}	p -vector of covariates at observation j in cluster i	$(X_{ij1}, X_{ij2}, \dots, X_{ijp})^T$
X_{ijk}	Value of covariate k , at observation j , in cluster i	
β	p -vector of parameters	$(\beta_1, \beta_2, \dots, \beta_p)^T$

When it's used, the $(\sum n_i) \times p$ matrix obtained by 'stacking' all the \mathbf{X}_i will be denoted just \mathbf{X} . But more often we just write estimating equations using sums, where the summands (from each cluster) involve matrix-valued expressions.

* ... so, yes, n_n is the size of the last cluster. Some use m for # clusters

Vector outcomes: notation

We define parameters of the population of clusters using estimating equations. As a first example, define β as solving

$$\mathbb{E}_F[\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\beta)] = 0,$$

where F denotes i.i.d sampling of $\{\mathbf{Y}_{n_i \times 1}, \mathbf{X}_{n_i \times p}\}$, i.e. simple random sampling from the population of clusters. This parameter can also be defined directly as

$$\beta = \mathbb{E}_F[\mathbf{X}^T \mathbf{X}]^{-1} \mathbb{E}_F[\mathbf{X}^T \mathbf{Y}].$$

Assuming* that all n_i are identical, from 570 we know β is estimated consistently by solving

$$\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T (\mathbf{Y}_i - \mathbf{X}_i \hat{\beta}) = \mathbf{0}_p,$$

$$\text{i.e. by calculating } \hat{\beta} = \left(\sum_{i=1}^n \mathbf{X}_i^T \mathbf{X}_i \right)^{-1} \sum_{i=1}^n \mathbf{X}_i^T \mathbf{Y}_i.$$

*for simplicity; similar results if the n_i vary little, or are random

Vector outcomes: notation

What does this parameter mean? It may help to re-write $\hat{\beta}$:

$$\hat{\beta} = \left(\sum_{i=1}^n \mathbf{X}_i^T \mathbf{X}_i \right)^{-1} \sum_{i=1}^n \mathbf{X}_i^T \mathbf{Y}_i = \left(\sum_i^n \sum_{j=1}^{n_i} \mathbf{X}_{ij} \mathbf{X}_{ij}^T \right)^{-1} \sum_i^n \sum_{j=1}^{n_i} \mathbf{X}_{ij} Y_{ij},$$

i.e. it's the estimate from 'plain vanilla' linear regression, of the 'stacked' outcomes on the 'stacked' covariates.

So, we can interpret $\hat{\beta}$ as the intercept & slope(s) of the least-squares best-fitting line $Y = \mathbf{X}^T \hat{\beta}$, using **all the observations in all the clusters**.

Similarly, β defines the least-squares best-fitting line, where we stress that 'best' takes into account **all observations in all clusters**. Note this treats all observations identically **whether they are in the same cluster or not**.

(Keen people: adapt the other definitions of linear regression β from slide 1.13, to acknowledge clustering.)

Vector outcomes: notation

As you have seen (e.g. HW1) even though 570-style ‘plain vanilla’ linear regression may provide estimates of a useful quantity, standard error estimates* that assume each outcome $\{Y_{ij}, \mathbf{X}_{ij}\}$ is independent do not give valid inference here – only the clusters $\{\mathbf{Y}_i, \mathbf{X}_i\}$ are independent.

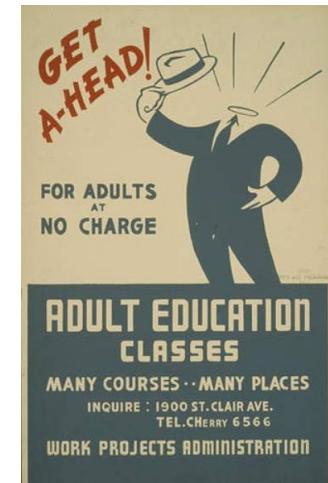
The simplest modification is obtained using the bootstrap; we resample n **whole clusters** $\{\mathbf{Y}_i^*, \mathbf{X}_i^*\}$ with replacement, and then re-run the regression. (A.k.a. *sampling from the empirical distribution*.) Doing this B times over, the set of $\hat{\beta}^*$ generated can be used for standard bootstrap inference – in particular, to compute approximate confidence intervals around the original $\hat{\beta}$.

The *essential* bootstrap idea still holds – empirical distribution $\tilde{F} \approx F$, and we estimate properties of F by those of \tilde{F} .

* robust or model-based; they both rely on independence

Vector outcomes: education!

We will apply the bootstrap to a *cluster randomized trial*, in adult education, and specifically attendance at adult literacy classes. There are $n = 28$ classes, with n_i between 2 and 9 adults in each. Each adult attended between 0 and 15 sessions (more is better).



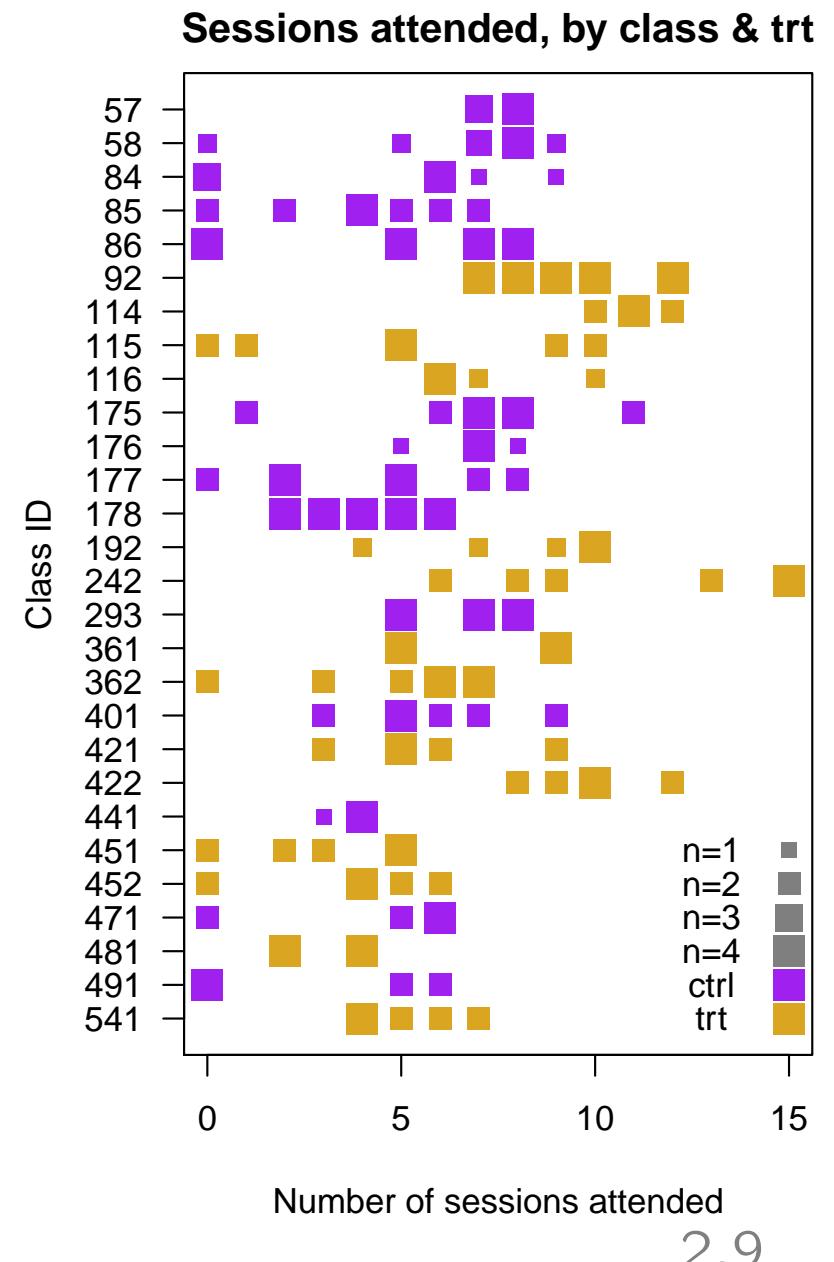
Because of the impossibility of randomizing adults blindly, the whole class is randomized to receive no incentives ($X = 0, 14$ classes) to attend, or to receive incentives ($X = 1, 14$ classes).

Of primary interest is β , the difference in mean number of sessions attended, comparing adults who get $X = 1$ to those with $X = 0$. Secondary interest may lie in β_0 , the mean number of sessions attended by adults with $X = 0$. Both are defined by averages over the population of classes.

Vector outcomes: education!

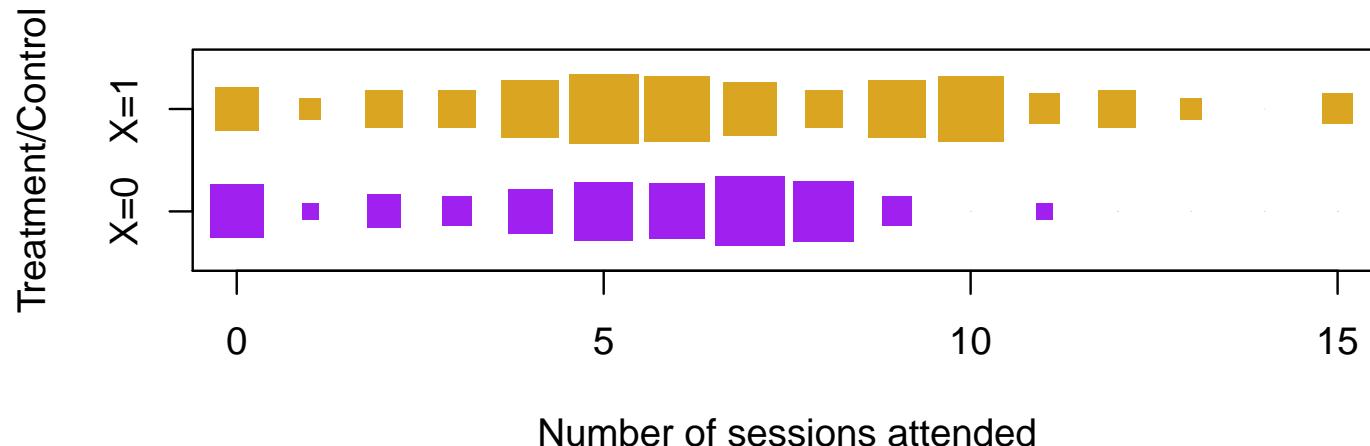
The full dataset (right)

- n_i (class size) between 2 and 9
- $n = 28$ classes
- $\sum_{i=1}^n n_i = 152$ observations
- $Y_{ij} = \#$ attendances, for student j in class i ; between 0 and 15 attendances
- $X_{ij} = X_i$, is 1/0 for treatment/control
- $\bar{X} = 0.46$ – pretty even



Vector outcomes: education!

The outcomes, broken down just by treatment;



- Mean number of sessions in the $X = 0$ (Control) group was 5.28 sessions
- Mean number of sessions in the $X = 1$ (Treatment) group was 6.69 sessions
- Difference in mean number of sessions, comparing observations on treatment to observations on control, is $\hat{\beta} = 6.69 - 5.28 = 1.41$ sessions

As this is a trial, estimate $\hat{\beta}$ of the treatment effect is not confounded. But how precise is it?

Vector outcomes: education!

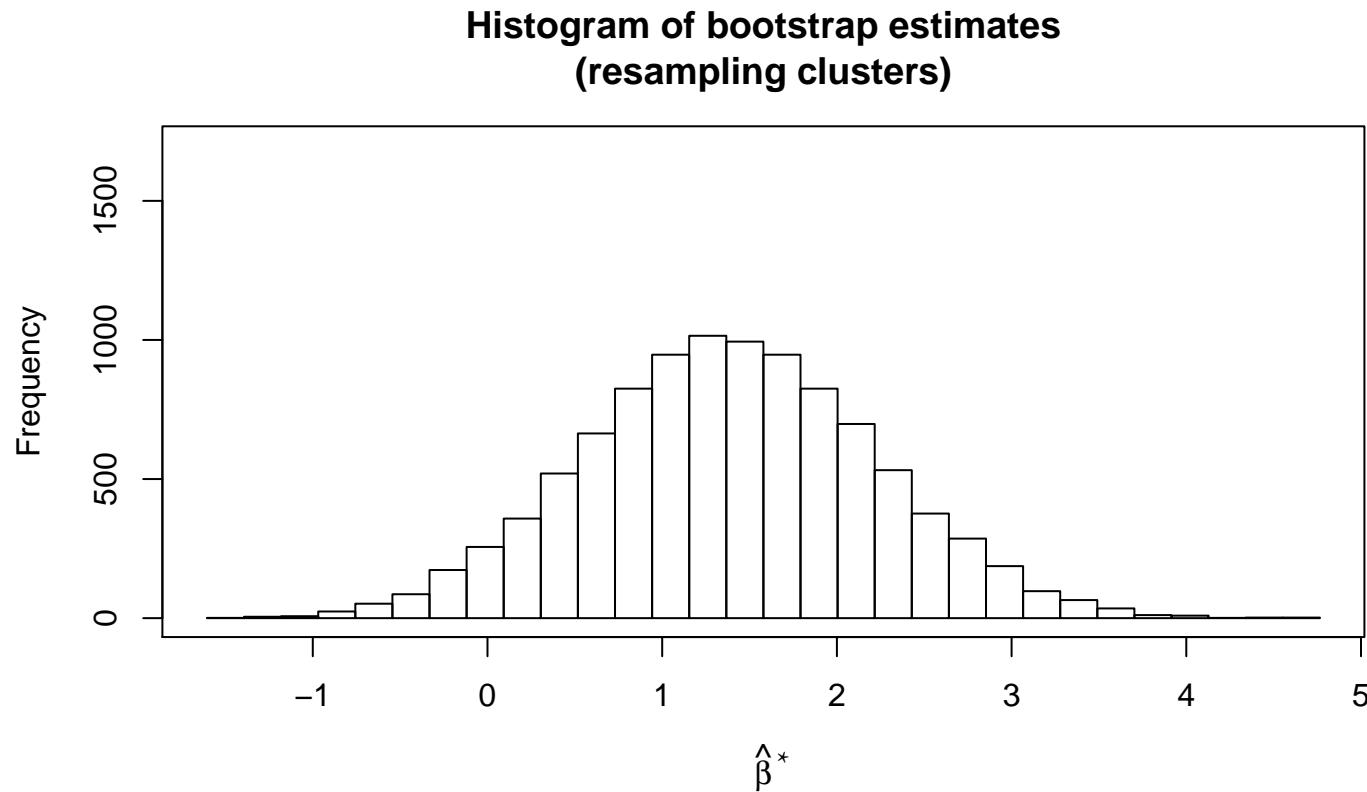
To answer this, we use the nonparametric bootstrap, resampling clusters. As with bootstrapping in 570, we do the same calculation on the original data, and the resampled version;

```
# original data
m0 <- with(subset(educ, group==0), mean(sessions))
m1 <- with(subset(educ, group==1), mean(sessions))
m1-m0

classes <- unique(educ$class)
do.one <- function(){
  # do the resampling of clusters;
  classes.star <- sample(classes, replace=TRUE)
  indexes <- sapply(classes.star, function(ccc){ which(educ$class==ccc) })
  educ.star <- educ[unlist(indexes),]
  # with these resampled clusters, do same calculation as above
  m0.star <- with(subset(educ.star, group==0), mean(sessions))
  m1.star <- with(subset(educ.star, group==1), mean(sessions))
  m1.star-m0.star
}
set.seed(4)
many.betahat.star <- replicate(10000, do.one() )
```

Vector outcomes: education!

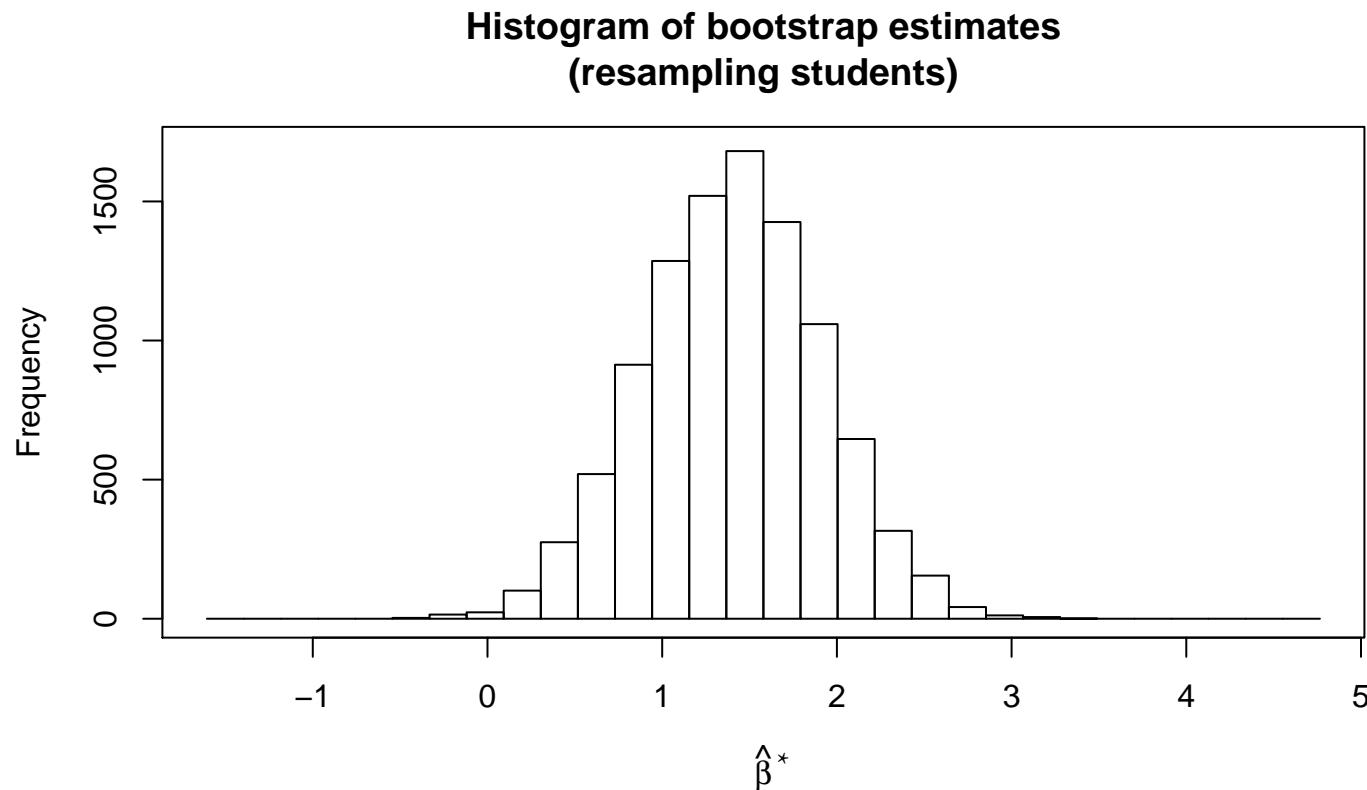
We get a (good) estimate of the distribution of $\hat{\beta}$ from the distribution of the $\hat{\beta}^*$ – given here up to Monte Carlo error from 10,000 samples;



We get a $\approx 95\%$ CI from 2.5% and 97.5% percentiles, i.e .(-0.23, 3.02). Note this is wide compared to the range of Y .

Vector outcomes: education!

If we (erroneously!) treat all outcomes as independent;



Now the interval is $(0.41, 2.39)$ – a reduction of $\approx 40\%$, and not overlapping zero

Vector outcomes: education!

Further notes;

- Clearly, the correlation matters for inference. Ignoring it means overstating precision (\propto sample size) by $1.4^2 \approx 2$
- Using a more complex $\hat{\beta}$ doesn't affect the method, we still just re-sample clusters, and compute many $\hat{\beta}^*$
- Theory tells us the bootstrap will work on asymptotically-Normal $\hat{\beta}$. The histogram is *some* re-assurance of this, but not much – what if class #29 had 200 students?
- No attempt made to model within-cluster covariance. This is potentially wasteful, if specific assumptions could be justified
- Coding might be easier with data in ‘wide’ format (each class on one row) – see e.g. the `reshape()` function. But as class sizes differ, some rows will have NAs

Vector outcomes: which bootstrap? (*)

It's important to note that we are resampling whole clusters – not individuals within them. For the purposes of non-parametric inference, the superpopulation is one of clusters, not of (say) the individual subjects these clusters comprise.

In many applications, the population of subjects seems more intuitive to many users, who would like to further resample the observations within each cluster.

We therefore compare two competing bootstrap ‘strategies’;

- Resample clusters with replacement, and within them sample observations $\{Y_{\cdot j}, \mathbf{X}_{\cdot j}\}$ **without** replacement (S1)
- Resample clusters with replacement, and within each cluster sample observations $\{Y_{\cdot j}, \mathbf{X}_{\cdot j}\}$ **with** replacement (S2)

Assuming the order of the data doesn't matter, S1's second step of sampling n_i observations without replacement affects nothing.

Vector outcomes: which bootstrap? (*)

We* examine the properties of S1 and S2, in a tractable situation. The data generating mechanism is that

$$Y_{ij} = b_i + Z_{ij}, \quad 1 \leq i \leq n, \quad 1 \leq j \leq n_i$$

where n_i is constant for all i , and all b_i and Z_{ij} are independent, with

$$\begin{array}{ll} \mathbb{E}[b_i] = 0 & \mathbb{E}[Z_{ij}] = 0 \\ \text{Var}[b_i] = \sigma_b^2 & \text{Var}[Z_{ij}] = \sigma_Z^2 \end{array}$$

It follows that

$$\begin{aligned} \text{Var}[Y_{ij}] &= \sigma_b^2 + \sigma_Z^2 \\ \text{Cov}[Y_{ij}, Y_{ij'}] &= \text{Cov}[b_i + Z_{ij}, b_i + Z_{ij'}] \\ &= \text{Var}[b_i] + \text{Cov}[Z_{ij}, Z_{ij'}] = \sigma_b^2 \end{aligned}$$

NB the **intraclass correlation** is $\frac{\sigma_b^2}{\sigma_b^2 + \sigma_Z^2}$; it can be interpreted as the correlation between two outcomes in the same cluster, or as the proportion of total variance that is ‘between clusters’

* following Davison and Hinkley, pg 100–102

Vector outcomes: which bootstrap? (*)

Denoting the bootstrap distribution by a star, and denoting the random cluster number by I^* , under either S1 or S2 we get

$$\begin{aligned}\mathbb{E}_{F^*}[Y_{I^*j}|I^* = i] &= n_i^{-1} \sum_l Y_{il} \equiv \bar{Y}_i \\ \mathbb{E}_{F^*}[Y_{I^*j}^2|I^* = i] &= n_i^{-1} \sum_l Y_{il}^2,\end{aligned}$$

and consequently

$$\begin{aligned}\mathbb{E}_{F^*}[Y_{ij}] &= n^{-1} \sum_{i,j} n_i^{-1} Y_{ij} \equiv \bar{Y} \\ \mathbb{E}_{F^*}[Y_{ij}^2] &= n^{-1} \sum_i (\bar{Y}_i - \bar{Y})^2 + n^{-1} \sum_i n_i^{-1} (Y_{ij} - \bar{Y}_i)^2.\end{aligned}$$

We see immediately that the expectation of the resampled outcomes is unbiased for $\mathbb{E}_F[Y_{ij}]$. Slightly more work gives

$$\mathbb{E} [\text{Var}_{F^*}[Y_{ij}]] = \frac{n-1}{n} \sigma_b^2 + \frac{nn_i - 1}{nn_i} \sigma_Z^2$$

Vector outcomes: which bootstrap? (*)

But for the cross terms with $j \neq k$ we get

$$\mathbb{E}_{F^*}[Y_{I^*j} Y_{I^*k} | I^* = i] = \begin{cases} \frac{1}{n_i(n_i-1)} \sum_{l \neq m} Y_{il} Y_{im}, & S1 \\ \frac{1}{n_i^2} \sum_{l \neq m} Y_{il} Y_{im}, & S2 \end{cases}$$

Given a particular dataset, the bootstrap(s) give

$$\text{Cov}_{F^*}[Y_{ij} Y_{ik}] = \begin{cases} n^{-1} \sum_i (\bar{Y}_i - \bar{Y})^2 - \frac{1}{nn_i(n_i-1)} \sum_{i,j} (Y_{ij} - \bar{Y}_i)^2, & S1 \\ \frac{1}{n} \sum_i (\bar{Y}_i - \bar{Y})^2, & S2 \end{cases}$$

and this leads to

$$\mathbb{E} [\text{Cov}_{F^*}[Y_{ij} Y_{ik}]] = \begin{cases} \frac{n-1}{n} \sigma_b^2 - \frac{1}{nn_i} \sigma_Z^2, & S1 \\ \frac{n-1}{n} \sigma_b^2 - \frac{n-1}{nn_i} \sigma_Z^2, & S2 \end{cases}$$

Recall that this covariance is, in truth, just σ_b^2 . **Q.** Which formula above gets closest to that value?

Vector outcomes: which bootstrap? (*)

Notes on this;

- Resampling whole clusters **really does** lead to better reconstruction of the first two moments of F
- Given that we didn't specify a within-cluster correlation structure, naively using a bootstrap that treats observations as i.i.d within clusters seems unwise. Q unless what?
- If n_i is very large or σ_Z^2 is tiny, resampling within clusters won't be terrible
- Where the bootstrap is used, resampling whole clusters is standard; in STATA using `cluster(id)` and `vce(boot)` gives this automatically
- See the class site for recent methods work that **does** permit S2, under some conditions

Vector outcomes: bootstrap summary

- Resampling clusters mimics the stated independence assumptions, and thus provides the best (or at least most reliable) way to get an asymptotically-valid bootstrap interval
- The non-parametric bootstrap works just as in 570; resampling clusters not individuals is the only difference
- As in 570, bootstrapping can be slow, and the size/importance of Monte Carlo error should be considered
- Also as in 570, need to decide what to do with rare-but-possible ‘weird’ resamples, e.g. only resampling classes with $X = 1$, in the education example

Quicker and Monte-Carlo-free non-parametric inference is available using the sandwich...

Vector outcomes: sandwiches

Given estimating equations with a summand from each independent cluster, the same theory holds as in 570. Under mild regularity conditions, solving the EEs defines an estimator $\hat{\beta}$ consistent for parameter β with asymptotic distribution

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{D}} N\left(\mathbf{0}_p, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^T - 1\right).$$

To keep the algebra simple(r), here we only consider estimating equations of the form

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial g(\mathbf{X}_i \beta)}{\partial \beta^T} (\mathbf{Y}_i - g(\mathbf{X}_i \beta)) = \mathbf{0}_p,$$

where, by convention, $g(\cdot)$ is applied component-wise. We are finding least-squares fits of the line $y = g(\mathbf{x}^T \beta)$, over all observations in all clusters. The matrix of derivatives is $p \times n_i$, and it multiplies an $n_i \times 1$ vector of residuals.

Vector outcomes: sandwiches

In this situation, the asymptotic variance is $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^T - 1/n$, where

$$\begin{aligned}\mathbf{A} &= \mathbb{E}_F \left[\frac{\partial}{\partial \boldsymbol{\beta}} \left(\frac{\partial g(\mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} (\mathbf{Y} - g(\mathbf{X}\boldsymbol{\beta})) \right) \right] \\ \mathbf{B} &= \mathbb{E}_F \left[\left(\frac{\partial g(\mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} (\mathbf{Y} - g(\mathbf{X}\boldsymbol{\beta})) \right) \left(\frac{\partial g(\mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} (\mathbf{Y} - g(\mathbf{X}\boldsymbol{\beta})) \right)^T \right],\end{aligned}$$

and the expectation is over i.i.d clusters $\{\mathbf{Y}_i, \mathbf{X}_i\}$.

- The formulae are generalizations of those seen in 570. But when the elements of cluster-specific \mathbf{Y} are not assumed independent, further simplification is not generally possible
- Weighted versions of the EEs also give tractable forms – particularly if the weights cancel parts of the derivative terms. Recall use of canonical link functions did this, for GLMs

Vector outcomes: sandwiches

The standard empirical estimators of the bread and meat are;

$$\begin{aligned}\hat{\mathbf{A}} &= \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \beta} \left(\frac{\partial g(\mathbf{X}_i \hat{\beta})}{\partial \beta^T} (\mathbf{Y}_i - g(\mathbf{X}_i \hat{\beta})) \right) \\ \hat{\mathbf{B}} &= \frac{1}{n} \sum_{i=1}^n \left(\frac{\partial g(\mathbf{X}_i \hat{\beta})}{\partial \beta^T} (\mathbf{Y}_i - g(\mathbf{X}_i \hat{\beta})) \right) \left(\frac{\partial g(\mathbf{X}_i \hat{\beta})}{\partial \beta^T} (\mathbf{Y}_i - g(\mathbf{X}_i \hat{\beta})) \right)^T\end{aligned}$$

where the derivatives are evaluated at the point estimate $\beta = \hat{\beta}$.

These estimates are combined in the usual way;

$$\widehat{\text{Var}}(\hat{\beta}) = \frac{1}{n} (\hat{\mathbf{A}})^{-1} \hat{\mathbf{B}} (\hat{\mathbf{A}}^T)^{-1}$$

to give an asymptotically-justified estimate of the covariance of $\hat{\beta}$; approximate 95% confidence intervals, and Wald tests follow exactly as in 570.

(Also as in 570, under replications where the \mathbf{X}_i are actually fixed, the sandwich is asymptotically conservative at worst – usually very mildly.)

Vector outcomes: sandwiches

To show this working, we use an example with simulated data. Consider the situation where the true data-generating mechanism is;

$$a_i \sim N(0, 1)$$

$$X_{ij} \sim N(0, 1)$$

$$Y_{ij}|X_{ij} = x, a_i = a \sim N(a + \gamma_0 + \gamma_1 x + \gamma_2 x^2, 0.5^2)$$

for $1 \leq j \leq n_i$, and $1 \leq i \leq n$.

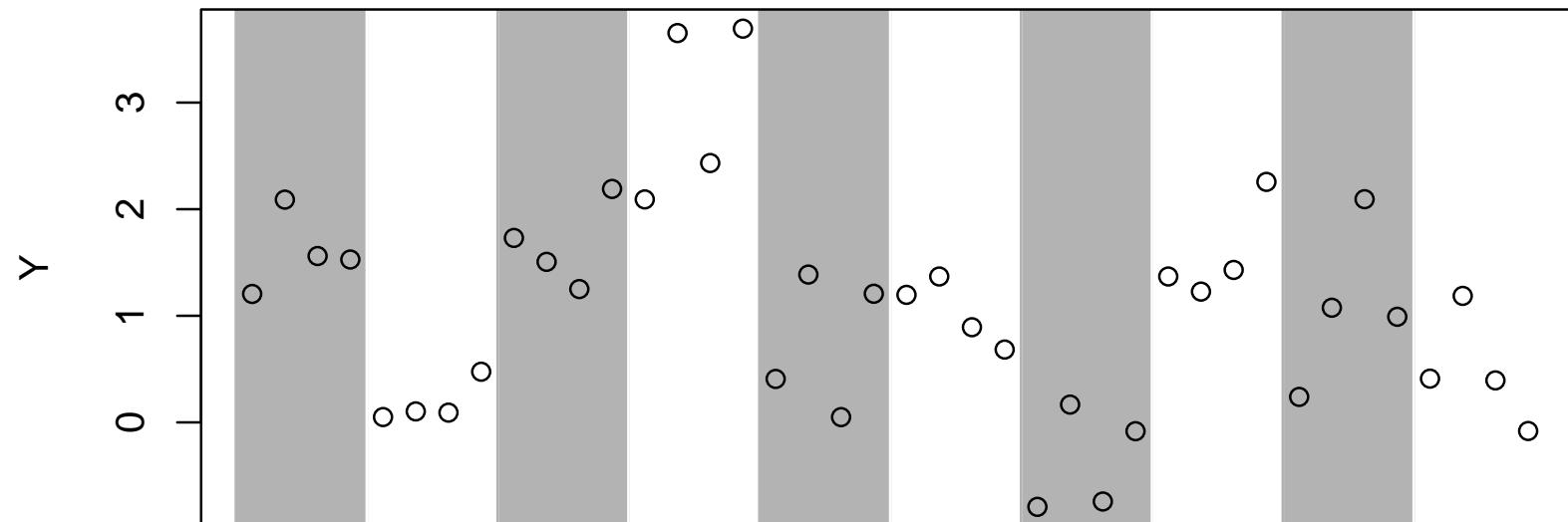
Here we use $n_i = 4$ for all clusters, and $\gamma_0 = 1, \gamma_1 = 0.4, \gamma_2 = 0.1$

This data-generating mechanism F is an example of a *linear mixed model*. However, our analysis will *not* assume we know the parametric form of sampling distribution F , except independence of clusters.

Vector outcomes: sandwiches

In our setup, *within-cluster* dependence is induced by the $\{a_i\}$, which are known as *random effects*;

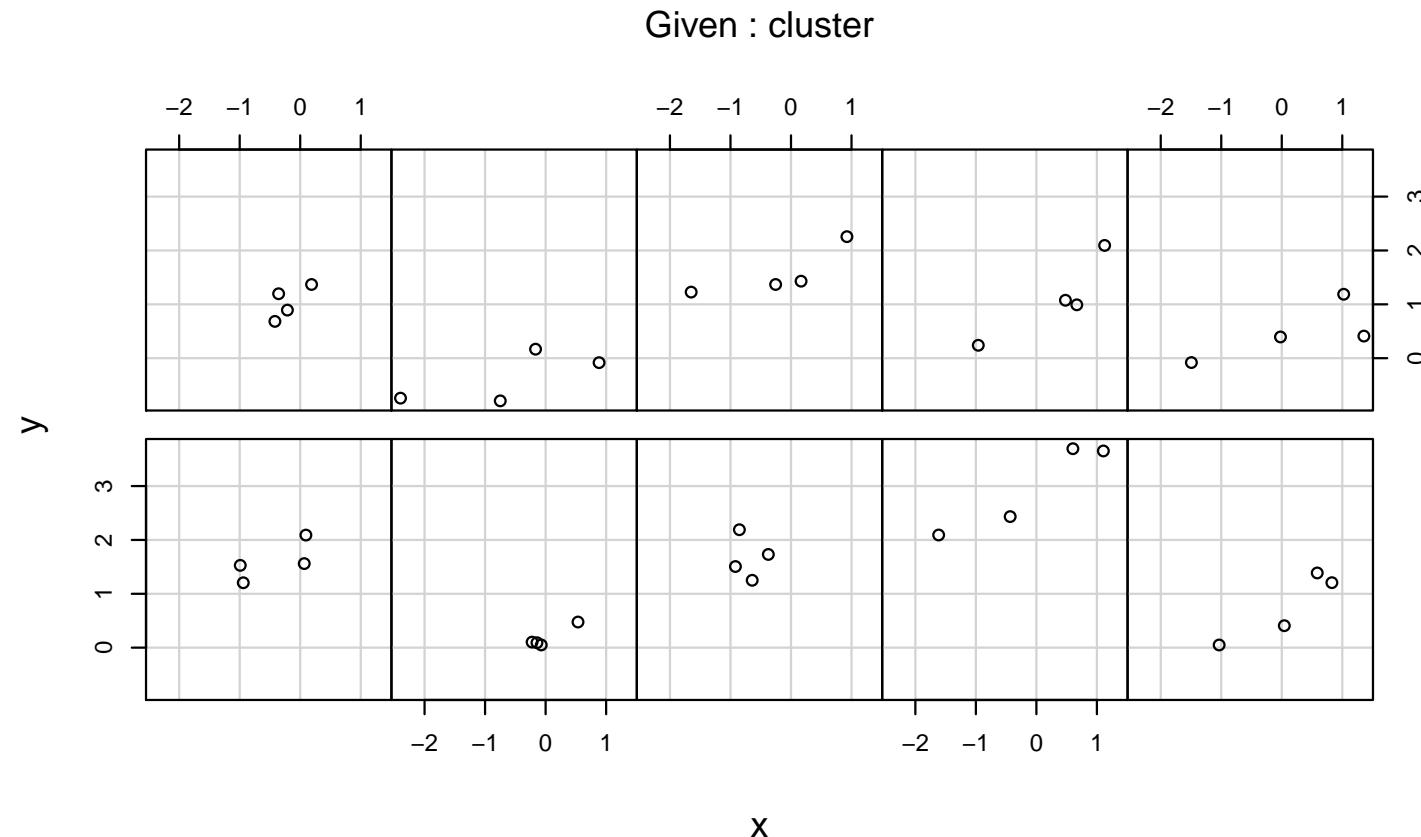
Typical data, from n=10 clusters



Note that outcomes within a cluster correlate more closely with each other than with outcomes in other clusters.

Vector outcomes: sandwiches

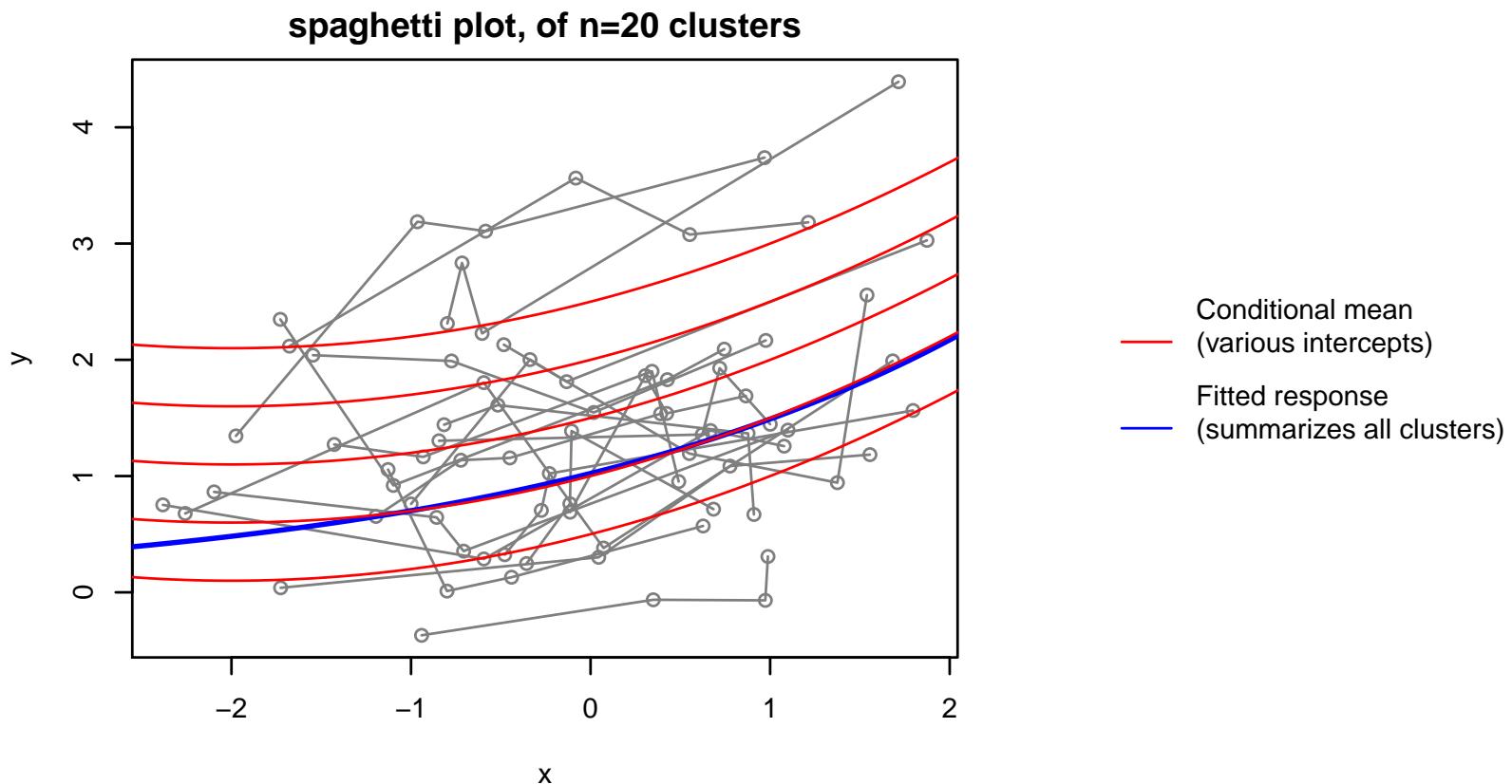
Within each cluster, the mean of $Y|X = x$ is truly quadratic in x – but with only four data points per cluster, this is not obvious by inspection;



Vector outcomes: sandwiches

In our regression, using all observations, we fit the line;

$$y = \exp(\beta_0 + \beta_1 x)$$



– i.e. we estimate a (sane) log-linear summary of the clusters.

Vector outcomes: sandwiches

The corresponding estimating equations are

$$\sum_{i=1}^n \left[\frac{\partial e^{\mathbf{X}_i \boldsymbol{\beta}}}{\partial \boldsymbol{\beta}} \right]_{p \times n_i}^T \left(\mathbf{Y}_i - e^{\mathbf{X}_i \boldsymbol{\beta}} \right)_{n_i \times 1} = \mathbf{0}_{p \times 1},$$

where subscripts denote dimensions. This can also be written as

$$\sum_{i,j} \mathbf{G}(Y_{ij}, \mathbf{X}_{ij}) \equiv \sum_{i,j} \frac{\partial e^{\mathbf{X}_{ij}^T \boldsymbol{\beta}}}{\partial \boldsymbol{\beta}} \left(Y_{ij} - e^{\mathbf{X}_{ij}^T \boldsymbol{\beta}} \right) = \mathbf{0}_{p \times 1},$$

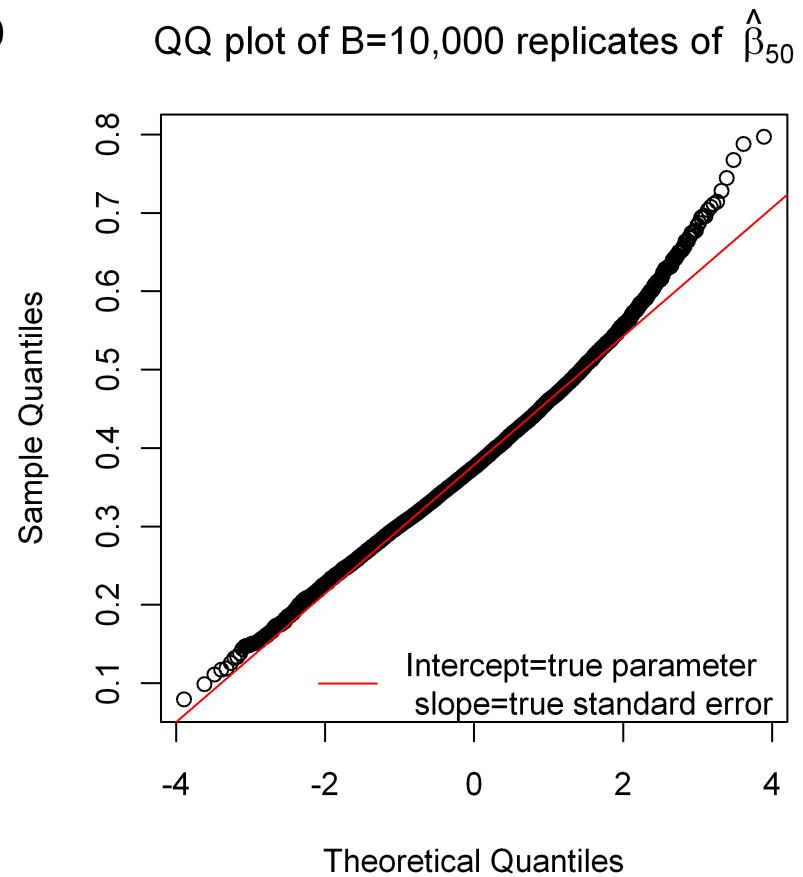
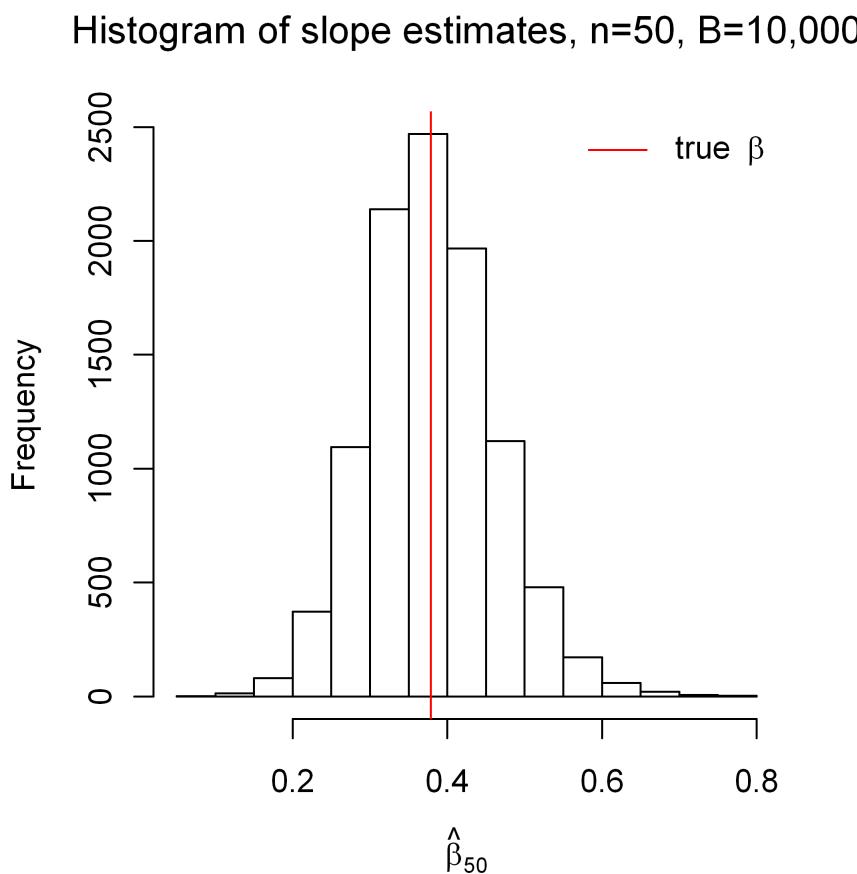
where the terms in the summands are $p \times 1$ and 1×1 . Note that using a non-diagonal weighting matrix between the derivative and residual would not permit this use of a common $\mathbf{G}()$, making the estimating function harder to code up.

Q. To what use of `glm()` does these EEs correspond? (i.e. what family and link functions?)

Note this use of `glm()` is just a computational ‘trick’, to solve equations; it does **not** mean we assumed more about F .

Vector outcomes: sandwiches

Any way you solve them*, here's $\hat{\beta}$ from $B = 10,000$ example datasets; (x -axis of RH plot uses `qnorm(ppoints(B))`)



* The EEs solve p smooth functions; Newton-Raphson works too

Vector outcomes: sandwiches

We see that, despite not having a closed form, or being based on i.i.d. $\{Y_i \in \mathbb{R}, \mathbf{X}_i \in \mathbb{R}^p\}$, the estimate $\hat{\beta}$ really *is* approximately Normal – it is slightly heavy-tailed in the RH tail, light-tailed in the LH.

- The Central Limit Theorem still applies; we sum n terms each involving $\mathbf{Y}_i \in \mathbb{R}^{n_i}$ and $n_i \times p$ matrix \mathbf{X}_i , but we still have p estimating functions
- In this example, with $n = 50$ bias is not a major problem (although for rare datasets a finite root may not exist)
- For this population, the true log-linear slope parameter is $\beta = 0.378$ *; note this is not equal to any of the γ parameters, nor is it a simple function of them.
- The standard error of $\hat{\beta}_{50}$ is 0.082 *

* calculated using numeric integration (not shown)

Vector outcomes: sandwiches

Particularly when n_i is not constant, coding up these expressions can require considerable care;

- If possible, check your code on some data where $n_i = 1$; you should get the same estimates from e.g. `glm()`
- Break the job into small chunks you understand
- Become adept with `by()`, `aggregate()`, `lapply()` etc

With different n_i it may be convenient to store each subjects' data as an element in a `list` object;

- Elements of a `list` can be of any type (including other `lists`), and any dimension
- Extract the element you want using `my.list[[1]]`, `my.list[[2]]`, etc. Obtain sublists using e.g. `my.list[1:5]`, `my.list[c(1,1,3,4)]`
- `by(data, INDICES, FUN)` makes subsets of data according to the value of `INDICES`, then applies `FUN` to each subset; the output is returned in a `list`

Vector outcomes: sandwiches

The generality of `list` objects means there is no generic `sum()` method for them. But if you *know* each element is a $p \times p$ matrix, you can write your own, and use it;

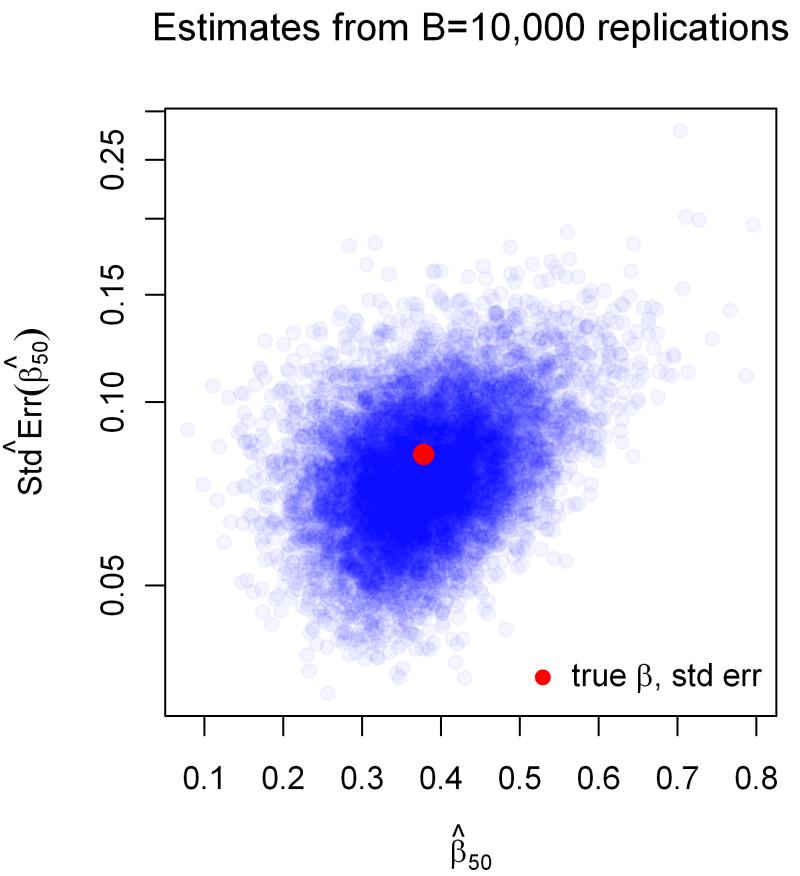
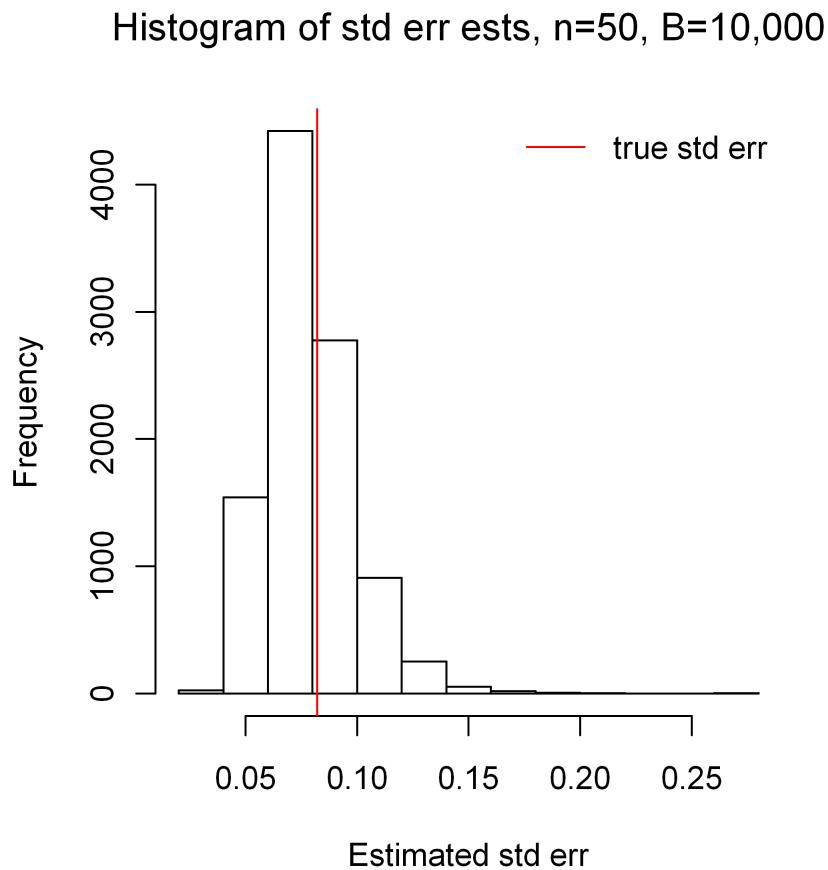
```
# first a helper function, that adds up elements of a list, using a for() loop
sum.list <- function(myL){ n <- length(myL); out <- myL[[1]]
  for(i in 2:n){ out <- out + myL[[i]] }
  out}

# generate data, calculate betahat
m1 <- make.one(100,4,c(0.4,0.1)) # makes data frame with "int"=1, "x", "y", "id"
b1 <- fit.one(m1)$coef          # calculates beta.hat - using glm()
# evaluate the A matrix's summands;
allAi <- by(m1, m1$id, function(cluster){      # look up ?by if it's new to you
  Xi      <- as.matrix(cluster[,c("int","x")])
  Yi      <- cluster$y
  mui     <- as.vector(exp(Xi %*% b1))        # fitted values
  resid   <- Yi - mui                          # residuals
  dgdbeta <- Xi * mui                         # derivatives
  t(Xi) %*% diag(mui * resid) %*% Xi - t(dgdbeta) %*% dgdbeta
})
Ahat <- sum.list(allAi)
```

... see class site for `make.one()`, `fit.one()` and $\widehat{\mathbf{B}}$. Doing calculations with ‘stacked’ $n \times (p^2)$ matrices can be a little faster – but is harder to code/debug. See also the `Matrix` package.

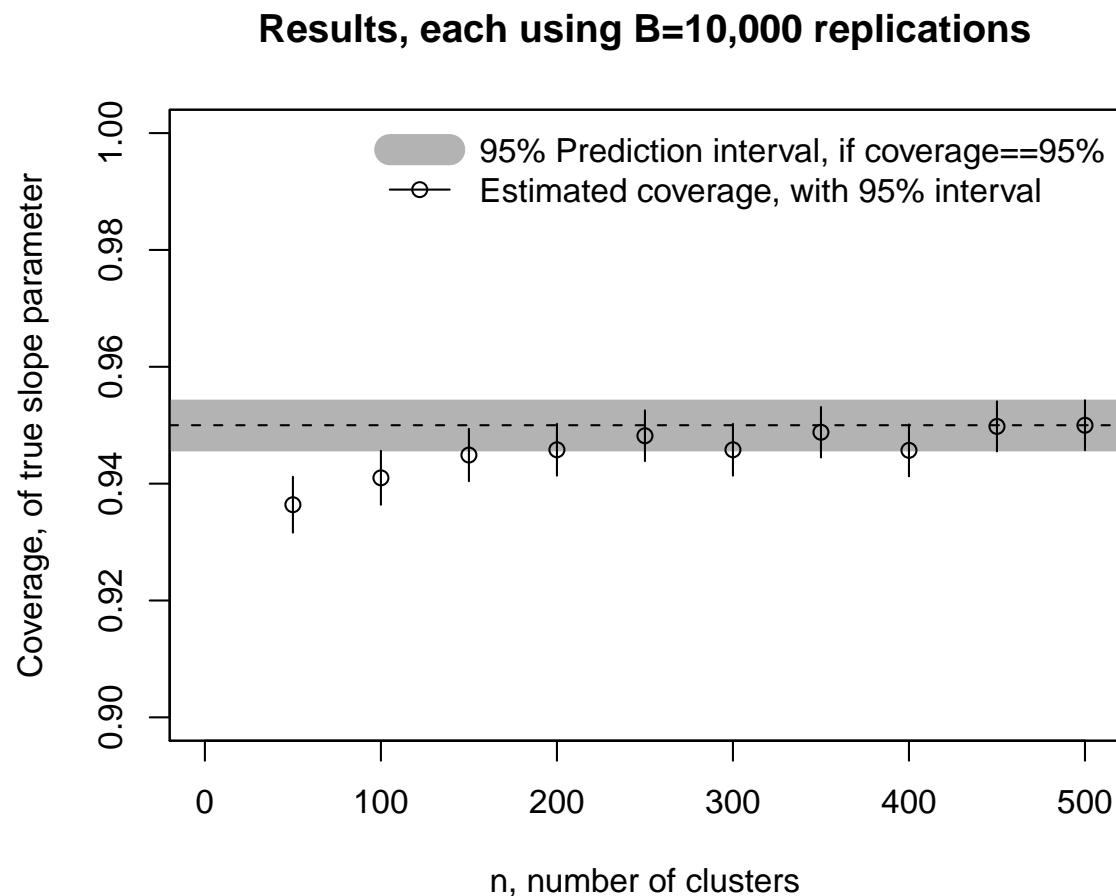
Vector outcomes: sandwiches

Does it work? For $n = 50$, here are point estimates and standard error estimates, estimating ‘slope’ parameter β_1



Vector outcomes: sandwiches

Varying n , the usual $\hat{\beta} \pm 1.96 \times \widehat{Std.Err}$ approach yields nominal 95% intervals with this coverage;



Vector outcomes: sandwiches

- The approach is asymptotic in n , the number of clusters. For small n , it's not perfect, but it does work eventually. Expect practitioners used to very small n to distrust this approach
- The variability in the estimated standard error is *much* smaller than that of $\hat{\beta}$ (compare the histograms) – which is important, for accurate asymptotic approximations*
- The estimated standard error tends to be ‘too big’ when $\hat{\beta}$ is also ‘too big’. This makes the asymptotics work faster than we’d see for independent $\hat{\beta}$ and $\widehat{Std.Err}$, keeping the same marginal distributions. If the inaccuracies worked against each other, the asymptotics would work more slowly.
- The estimated standard error is not Normal – nor do we need it to be.

Q. What type of clusters make the biggest sandwich contributions?

* Recall from 533 that this ‘extra’ variability results in using t_{df} , not $N(0, 1)$ – very similar, if df is large

Vector outcomes: summary

In summary, while the computation may get somewhat trickier, there is nothing *in principle* to stop us using the same results from 570 to provide asymptotically justified inference here.

- Both the sandwich and bootstrap are valid non-parametric methods*. The estimated standard errors and confidence intervals allow for within-cluster correlation – which is important!
- Inference is *entirely* on marginal parameters, e.g. describing what happens comparing observations from across the entire population, **not** within a given cluster
- Asymptotics approximations may be poor for the available n . (Some diagnostics will follow, later)

We have given short examples, but the results seen throughout also hold for much more complex EEs.

* subject to regularity conditions

Vector outcomes: summary

A major drawback to the nonparametric approach (bootstrap and sandwich) is that interpreting β is often difficult.

- Comparison of means at distinct levels of X is an exception
- ‘Least squares lines’ are not too hard to describe, but why minimize least-squares around all observations? If we suspect the variance of each Y_{ij} is different, weighting them all equally seems inefficient. Also, if we know Y_{i1} and Y_{i2} give the same information, shouldn’t that alter what we minimize – perhaps not a simple summation of observation-specific terms?
- With binary or count outcomes, least squares (of any form) is unconvincing; describing log- or logit-linear ‘trends’ that summarize the observations is an unattractive alternative

As in Chap 1 (and 570) the ability to get valid inference on θ is no guarantee of θ ’s scientific relevance. This problem eases when we make more assumptions...

Generalized Estimating Equations



I saw Deep Purple live once and I paid money for it and I thought, ‘Geez, this is ridiculous.’ ... I never liked those Deep Purples or those sorts of things ... I always thought it was a poor man’s Led Zeppelin

Angus Young (1955–)

Scottish-born Australian guitarist

Co-founder, lead guitarist, and songwriter for AC/DC

Generalized Estimating Equations

We move to semiparametric inference for vector outcomes, and in particular, **Generalized Estimating Equations**.

- In a deep (but not purple) paper, Liang and Zeger (1986) developed GEE as a generalization of earlier quasi-likelihood techniques. GEE and QL have similar robustness properties
- However, we will make assumptions about the mean model, and *working* assumptions about the covariance of the outcomes (notably within-cluster correlations)
- $\hat{\mathbf{A}}^{-1} \hat{\mathbf{B}} \hat{\mathbf{A}}^{T-1} / n$ sandwiches are typically used to provide estimates of standard error – **unlike** use of $\hat{\mathbf{A}}^{-1} / n$ in 570's QL

In fields in which I work, GEE is the ‘industry standard’ method for clustered data, notably longitudinal data. Statisticians and non-statisticians are expected to use it, appropriately, or to carefully explain why it’s not appropriate. (See, e.g. Biost 540)

GEE: assumed marginal mean model

In addition to the between-cluster independence seen before, in GEE we assume;

$$\mathbb{E}[Y_{ij}|\mathbf{X}_{ij} = \mathbf{x}] = g(\mathbf{x}^T \boldsymbol{\beta})$$

– with the usual $g^{-1}(\cdot)$ link function. To save space we often denote $\mu_{ij} = \mathbb{E}[Y_{ij}|\mathbf{X}_{ij} = \mathbf{x}_{ij}]$.

There are two parts to this assumption;

- The stated mean is linear in \mathbf{X}_{ij} , on the scale $g(\cdot)$ indicates
- We are conditioning the distribution of each Y_{ij} on the value of its corresponding \mathbf{X}_{ij} – *and nothing else*

Note that the parameters $\boldsymbol{\beta}$ refers to expectations $\mathbb{E}[Y_{ij}|\mathbf{X}_{ij} = \mathbf{x}]$ over **all** observations from **all** clusters in the population, that have the stated covariate values – so this is a *marginal mean model*. Statements based on it **must** average over all other factors, i.e. variables not in the model, and over the whole population of clusters.

GEE: assumed marginal mean model

The first assumption should be familiar from 570. The second may require more attention;

- If X_{ij} is the same for all j (e.g. treatment/placebo in a cluster-randomized trial) describe regression parameters as contrasting averages over all clusters. Without more assumptions, no within-cluster contrast can be described – the data tell us nothing about such contrasts
- When X_{ij} varies within a cluster, can interpret parameters as comparing the conditional mean of outcomes – conditioning on two different values of X_{ij} **and nothing else**. E.g. compare outcomes in male vs female students j , **whether or not** they are in the same school i
- In longitudinal settings it may seem strange to condition on current drug dosage X_{ijk} and **not** earlier values – does X_{ijk} indicate medium dosage because the previous low dose didn't work, or because the previous high dose induced side-effects? Parameters defined using averages over earlier outcomes may not be scientifically useful, even if μ_{ij} is correctly specified

GEE: working assumptions

The estimating equations in GEE generalize what we saw before;

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial g(\mathbf{X}_i \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \mathbf{V}_i^{-1} (\mathbf{Y}_i - g(\mathbf{X}_i \boldsymbol{\beta})) = \mathbf{0}_p$$

To evaluate these, we must specify some choice of $\{\mathbf{V}_i\}$, the inverses of which weights contribution within each cluster.

The standard choice of *working covariance matrix* specifies;

$$\mathbf{V}_i = \phi \text{diag}\{S(\mu_{ij})^{1/2}\} \mathbf{R}_i \text{diag}\{S(\mu_{ij})^{1/2}\}.$$

The three components are;

- $\phi > 0$, a positive scale parameter
- $S(\cdot)$, a positive function of the mean; each entry in \mathbf{R}_i is multiplied by its corresponding $S(\mu_{ij})^{1/2}$, $S(\mu_{ij'})^{1/2}$ and ϕ
- \mathbf{R}_i , known as the *working correlation matrix*

To complete the working covariance matrix, we require a working choice of working correlation matrices \mathbf{R}_i .

GEE: working assumptions

Some commonly-used correlation structures for \mathbf{R}_i are;

Independence:

$$\mathbf{R}_i = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Exchangeable:

$$\mathbf{R}_i = \begin{bmatrix} 1 & \alpha & \alpha & \cdots & \alpha \\ \alpha & 1 & \alpha & \cdots & \alpha \\ \alpha & \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \alpha & \cdots & 1 \end{bmatrix}$$

GEE: working assumptions

Auto-regressive (AR1):

$$\mathbf{R}_i = \begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{n_i-1} \\ \alpha & 1 & \alpha & \cdots & \alpha^{n_i-2} \\ \alpha^2 & \alpha & 1 & \cdots & \alpha^{n_i-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^{n_i-1} & \alpha^{n_i-2} & \alpha^{n_i-3} & \cdots & 1 \end{bmatrix}$$

Unstructured: (symmetric, so $\alpha_{12} = \alpha_{21}$ etc)

$$\mathbf{R}_i = \begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} & \cdots & \alpha_{1n_i} \\ \alpha_{21} & 1 & \alpha_{23} & \cdots & \alpha_{2n_i} \\ \alpha_{31} & \alpha_{32} & 1 & \cdots & \alpha_{3n_i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{n_i1} & \alpha_{n_i2} & \alpha_{n_i3} & \cdots & 1 \end{bmatrix}$$

Note: apart from n_i , the working correlation is identical across clusters; this working assumption may not reflect the truth.

GEE: estimating equations

For now, we assume that ϕ and/or any α parameters are known. In particular, this lets us evaluate the estimating equations, and consequently to find $\hat{\beta}$, the root of the estimating equations.

(The EEs we used for nonparametric work are equivalent to setting $\mathbf{V}_i = \sigma^2 I_{n_i}$ for arbitrary unknown σ . This choice is actually common in applications of GEE – providing a non-parametric interpretation to the semi-parametric procedures here. The NP version doesn't rely on the assumed mean model)

We will show that, under the GEE assumption

- The estimate $\hat{\beta}$ is consistent and asymptotically Normal ($n \rightarrow \infty$) under mild regularity conditions (robustness)
- Reasonably simple sandwich formulae apply
- We get better $\hat{\beta}$ when choosing $\mathbf{V}_i \propto \text{Cov}[\mathbf{Y}_i | \mathbf{X}_i]$ (efficiency)

GEE: sandwiches

In what follows we denote

$$\mathbf{D}_i = \frac{\partial g(\mathbf{x}_i \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

i.e. the $n_i \times p$ matrix of derivatives of the n_i means, with respect to the p elements of $\boldsymbol{\beta}$. This lets us write the estimating function as simply

$$\frac{1}{n} \sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i).$$

Taking expectations with \mathbf{X} fixed, this is the sum of n functions which have expectation zero – and finite higher moments, according to regularity conditions. Hence its assumed-unique root $\hat{\boldsymbol{\beta}}$ is consistent for marginal mean model parameter $\boldsymbol{\beta}$.

As this is the mean of n similarly-distributed independent variables, asymptotic Normality of $\hat{\boldsymbol{\beta}}$ also holds*. The asymptotic variance may be written $\mathbf{A}^{-1} \mathbf{B} \mathbf{A}^T / n$, as usual.

* under slightly-less-mild regularity conditions

GEE: sandwiches

The ‘bread’ matrix \mathbf{A} is the expectation of the EE’s first derivative, i.e. the limiting value of

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\left(\frac{\partial \mathbf{D}_i^T}{\partial \beta} \right) \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) + \mathbf{D}_i^T \left(\frac{\partial \mathbf{V}_i^{-1}}{\partial \beta} \right) (\mathbf{Y}_i - \boldsymbol{\mu}_i) - \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i^T \right]$$

as $n \rightarrow \infty$.

Using the assumption of the marginal mean model, only the right hand term here is nonzero, i.e. the limiting value of

$$\mathbf{A} = -\frac{1}{n} \sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i,$$

where the expectation is redundant, because we are considering \mathbf{X}_{ij} fixed. Also, the minus sign is irrelevant to all sandwich calculations, so is omitted.

GEE: sandwiches

Similarly, the ‘meat’ matrix is the limiting value of

$$\mathbf{B} = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\mathbf{D}_i^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) (\mathbf{Y}_i - \boldsymbol{\mu}_i)^T \mathbf{V}_i^{-1} \mathbf{D}_i \right]$$

Recall that \mathbf{V}_i was only a *working* covariance matrix, so no further cancellation happens – unlike using $\hat{\mathbf{A}}^{-1}/n$ alone in 570’s QL.

Corresponding empirical estimates should look very familiar;

$$\hat{\mathbf{A}} = \frac{1}{n} \sum_i \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i$$

$$\hat{\mathbf{B}} = \frac{1}{n} \sum_i \mathbf{D}_i^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i) (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)^T \mathbf{V}_i^{-1} \mathbf{D}_i,$$

where everything is evaluated at the point estimates for all relevant parameters, and for inference we use

$$\widehat{\text{Cov}}[\hat{\boldsymbol{\beta}}] = \hat{\mathbf{A}}^{-1} \hat{\mathbf{B}} \hat{\mathbf{A}}^{T-1} / n$$

to provide confidence intervals and Wald tests.

GEE: sandwiches

It's worth re-iterating that GEE is robust to;

- Clustering (assuming the clusters are all independent — so if in doubt use larger clusters)
- Mis-specification of the working covariance matrix

GEE is **not**, in general, robust to mis-specification of the mean model — some authors call this being ‘semi-robust’ or ‘semi-Huber’, see e.g. Newson (1999).

However, if the canceled terms in \hat{A} were **zero anyway** (e.g. under canonical links with independence working assumptions) GEE is **also** robust to;

- Mis-specification of the mean model

... i.e. we're back to the non-parametric ‘full’ robust, a.k.a. ‘full Huber’ approach we saw before GEE. The estimates may be hard to interpret, but inference based on them is large-sample valid.

GEE: measles!

Sherman and Le Cessie (1997, on the class site) studied number of cases of measles in preschool children, in 15 counties in the US, between 1985 and 1991.

For each county the annual number of preschoolers with measles was recorded, as well as factors possibly related to measles incidence, like immunization rate, and density of preschoolers per county.

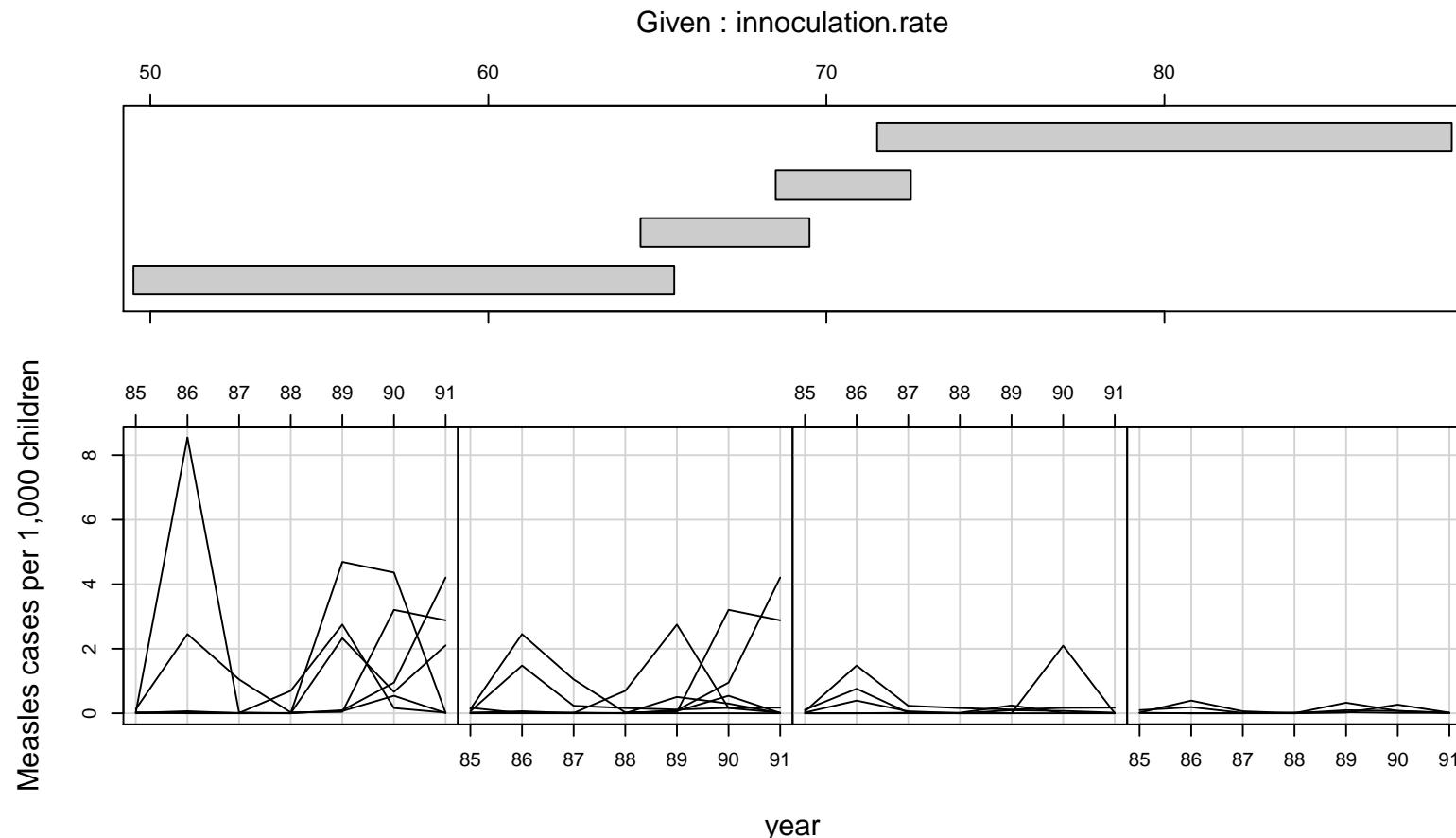
The data are annual for each county. We know;

- Number of cases of measles
- Immunization rate (percentage, fixed over time)
- Total number of preschoolers

We want to know the association between immunization rate and measles incidence.

GEE: measles!

Conditioning the measles data (rates per 1,000 children) on inoculation rate, via `coplot()`. What trend do you notice?



GEE: measles!

Using GEE, the key assumption is a mean model;

$$\mathbb{E}[Y_{ij} | \mathbf{X}_{ij} = x_{ij}] = \text{n.children}_{ij} \times e^{\beta_0 + \beta_1 x_{ij}}$$

This model states that, comparing observations where x_{ij} differs by $\Delta(x)$, the expected rates of measles cases differ by a factor of $\exp(\beta_1)\Delta(x)$, for any $\Delta(x)$. In particular, the expected rate of measles cases differs by $\exp(\beta_1)$ for each one-unit difference x_{ij} . (Also, the expected rate at $x = 0$ is unconstrained)

We also assume $S(\mu_{ij}) = \mu_{ij}$, i.e. the same mean-variance relationship as for Poisson outcomes – though we also have scale factor ϕ . For simplicity, we will use the independence working assumption here.

The estimating equations are

$$\sum_{i=1}^n \mathbf{X}_i^T (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}_2,$$

$$\text{where } \boldsymbol{\mu}_{ij} = \exp(\log(\text{n.children}_{ij}) + \beta_0 + \beta_1 X_{ij}).$$

GEE: measles!

The EEs can be solved using `glm()`, followed by some ‘by hand’ code for the other steps;

```
measles <- read.table("measlesdata.txt", header=T)
measles$county <- as.character(measles$county)

glm1 <- glm(cases~rate + offset(log(children)), data=measles, family=poisson )
measles$muhat <- fitted(glm1)
# i.e. the point estimates - these include the offset

# same helper function as before
sum.list <- function(l1){
  l <- length(l1);   out <- l1[[1]]
  for(i in 2:l){out <- out + l1[[i]]}
  out}

allAi <- by(measles, measles$county, simplify=F, function(data){
  ni <- dim(data)[1]
  Xi <- cbind(rep(1 ,ni), data$rate)
  Di <- Xi * data$muhat
  Vi <- diag(data$muhat)      # could use crossprod()
  t(Di) %*% solve(Vi) %*% Di
  })
Ahat <- sum.list(allAi)/n      # where n=length(allAi)
```

GEE: measles!

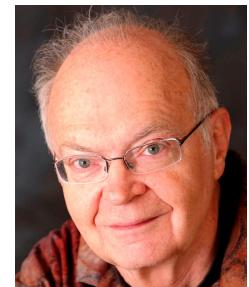
Similar code provides $\hat{\mathbf{B}}$, and the sandwich estimate;

```
allBi <- by(measles, measles$county, simplify=F, function(data){  
  ni <- dim(data)[1]  
  Xi <- cbind(rep(1 ,ni), data$rate)  
  Di <- Xi * data$muhat  
  Vi <- diag(data$muhat)  
  resid <- data$cases - data$muhat  
  t(Di) %*% solve(Vi) %*% resid %*% t(resid) %*% solve(Vi) %*% Di  
})  
Bhat <- sum.list(allBi)/n  
sand1 <- solve(Ahat) %*% Bhat %*% solve(Ahat)/n
```

- Where did ϕ go? (Hint: count the \mathbf{V} terms in the sandwich)
- Within each cluster, these calculations should be familiar from 570
- `crossprod()` would save time, `sum.list()` is clunky but works

Premature optimization is the root of all evil

Donald Knuth, comp sci and T_EX guru



GEE: measles!

Finally, the answers;

```
> coef(glm1)
(Intercept)      rate
-0.4581488 -0.1081273
> sqrt(diag(sand1))
[1] 1.01237231 0.01589221
> coef(glm1) + sqrt(diag(sand1)) %o% qnorm(c(0.025, 0.975))
     [,1]      [,2]
[1,] -2.4423621 1.52606448 # approx 95% CI for Intercept
[2,] -0.1392755 -0.07697913 # ... and rate
```

- Comparing measles incidence where the `rate` differs by one percentage point, the estimated log-rate ratio is -0.11 (-0.14, -0.08). The estimates RR is 0.90 (0.87, 0.93)
- A naïve standard error estimate, assuming independence, is 0.002 (Poisson model-based) or 0.021 (sandwich)
- Non-parametric bootstrapping gives estimated standard error 0.046, which might concern you.
- **Q.** What problem(s) might GEE face in this example?

GEE: parameter interpretation

As we saw in the measles example, for marginal means $g(\mathbf{x}_{ij}^T \boldsymbol{\beta})$, we can always interpret ‘slope’ β_k as a measure of between-observation difference in mean Y , per unit difference in x_{ijk} , among observations that have the same value of all other $x_{ijk'}$.

In longitudinal settings, we index the data from subject i as $\{\mathbf{Y}_{it}, \mathbf{x}_{it}\}$ where t is time – and where one of the covariates frequently *is* time, or a function of it.

For example, in the dental growth data, with observations at age 8,10,12,14 we might fit

$$\mathbb{E}[Y_{it} | \mathbf{X}_{it}] = \beta_0 + \beta_1(t - 8) + \beta_2 \text{Male}_i + \beta_3(t - 8) \times \text{Male}_i$$

... i.e. $\mathbf{X}_{it} = \{1, t - 8, \text{Male}_i, (t - 8) \times \text{Male}_i\}$, for all subjects and observations

GEE: parameter interpretation

Assuming this mean model is true, we interpret;

- β_0 as the mean distance* in 8-year old girls (i.e. not Male)
- β_1 as the difference in mean distance per additional year of age, in girls (i.e. the age effect in girls)
- β_2 as the difference in mean distance comparing 8-year olds boys to 8-year old girls
- β_3 as the difference in the age effect, comparing boys to girls

Exactly the same underlying model could be re-interpreted with the ‘baseline’ at age 10, or age 0 – or the average ‘maleness’.

If different relationships are expected at different ages, it makes sense to include terms in both $(t - 8), (t - 10)^+$, i.e. a linear spline. **Q** Why is including quadratic terms t^2 problematic?

* from pterygomaxillary fissure to pituitary gland, in mm

GEE: parameter interpretation

The dental growth study (all subjects from one age going forward, at the same time in the same setting) is a simple design – which makes for reasonably ‘clean’ comparisons.

Now imagine the same study done with subjects drawn over a vastly longer time-frame - say over 100 years. Do we want to learn about

- The effect of living 1 year later, and thus having access to better nutrition, medical care etc? This is a *cohort effect*
- The effect of being 1 year older, with one more year's food and growth hormones? This is a *longitudinal effect*

Using just $t = \text{year}$, we may learn about cohort effects, but are not adjusting for longitudinal effects. Using only $t = \text{age}$ we'd learn about longitudinal effects unadjusted for cohort effects. Neither seems very appealing.

GEE: parameter interpretation

With subjects entering at different times, we can fit e.g.

$$\mathbb{E}[Y_{it}|\mathbf{X}_{it}] = \beta_0 + \beta_C t + \beta_L(t - t_{i0})$$

where t_{i0} denotes times at entry; note that $(t - t_{i0})$ increases with age. β_C and β_L are cohort and longitudinal effects, respectively.

Now consider a design where everyone enters the study at $t = 0$, and is followed for a fixed 4 year period. Due at least in part to cohort effects, we might expect different mean outcomes in those who enter whilst young, versus old – better outcomes, usually.

To adjust for these cohort effects in this design, we can fit e.g.

$$\mathbb{E}[Y_{it}|\mathbf{X}_{it}] = \beta_0 + \beta_C \text{age}_{i0} + \beta_L t \equiv \beta_0 + \beta_C \text{age}_{i0} + \beta_L (\text{age}_{it} - \text{age}_{i0})$$

where age_{it} denotes age of subject i at time t .

GEE: parameter interpretation

In cross-sectional studies with only one observation, we cannot untangle cohort and longitudinal effects. $t = 0$ for every observation, meaning $(\text{age}_{it} - \text{age}_{i0}) = 0$ for all i, t , and consequently that the design matrix for $\{\beta_0, \beta_C, \beta_L\}$ is not of full rank. (This is *intrinsic aliasing*, see Jon's Book, §5.5.2.)

A natural alternative analysis fits

$$\mathbb{E}[Y_{it}|\mathbf{X}_{it}] = \beta_0 + \beta_1 \text{age}_{it}$$

which can also be written

$$\mathbb{E}[Y_{it}|\mathbf{X}_{it}] = \beta_0 + \beta_1 \text{age}_{it} + \beta_1 (\text{age}_{it} - \text{age}_{i0}),$$

motivating the interpretations (Jon's Book, §8.9) that this approach assumes $\beta_C = \beta_L$, or that β_1 estimates some form of average between β_C and β_L .

If cohort effects are plausible, neither of these is very convincing in practice. Obtaining longitudinal observations can therefore be scientifically very valuable.

GEE: parameter interpretation

Which regression to pick?

- “One that reflects your scientific question of interest”
- Adjust for confounding; which types of observations do you want to compare? What would you ideally keep fixed, to avoid being misled?
- Consider flexible representations of variables, e.g. splines, and interactions when effects of X differ by levels of Z
- Take particular care distinguishing $\Delta(\text{age})$ and $\Delta(\text{time})$; what exactly are you estimating? (For interactions between cohort and age see Jon’s Book §8.9)

Recall that, if your mean model is wrong, you **can** also interpret the estimates as those of a nonparametric ‘trend’ – with standard errors from e.g. the bootstrap, or a sandwich that does **not** simplify **A** as per slide 2.47. Also note that diagnostics for mean models follow later.

GEE: estimating α

Whatever our marginal mean model, unknown α must be estimated (except using $R_i = I_{n_i}$). Fortunately, as with ‘plug in’ and ‘profile’ operations in 570, it turns out that* any consistent estimator $\hat{\alpha}$ can be ‘plugged in’ for α , and **in large samples** the behavior of β is as if α were actually known. The basic argument (fleshed out in the 580s) is that;

- Using $R_i(\alpha)$ for any known α , $\hat{\beta}_\alpha$ is consistent for β
- Estimating α in any (sane) way will give a $\hat{\alpha}$ that is consistent for **some** value α'
- If $\hat{\alpha} \rightarrow \alpha'$ sufficiently quickly, $\hat{\beta}_{\hat{\alpha}}$ is consistent for β .

NB when precision is low, having consistency alone is weak;

*Consistency is a puny requirement
of dubious importance in samples of moderate size*
Drum & McCullagh (1993)
Statistical Science 8:3 300–301

* basically! ... mild regularity conditions apply

GEE: estimating α

The consistency result motivates finding straightforward estimates of α that we can incorporate into a fitting procedure for both $\hat{\beta}$ and $\hat{\alpha}$.

For some now-popular choices of R, Liang and Zeger (1986) proposed the following moment-based estimators of α , making use of observation-specific Pearson residuals;

$$e_{ij} = \frac{Y_{ij} - g(\mathbf{X}_{ij}^T \hat{\beta})}{\sqrt{\phi S(g(\mathbf{X}_{ij}^T \hat{\beta}))}}$$

Exchangeable: (with a degrees of freedom correction)

$$\hat{\alpha} = \frac{1}{(\sum_i \frac{1}{2}n_i(n_i - 1)) - p} \sum_{i,j < j'} e_{ij} e_{ij'},$$

i.e. averaging the products of all pairs of residuals, where we get $n_i(n_i - 1)/2$ pairs from within each cluster.

GEE: estimating α

Autoregressive (AR1):

$$\hat{\alpha} = \frac{1}{(\sum_i(n_i - 1)) - p} \sum_{\{i,j \leq n_i - 1\}} e_{ij}e_{i,j+1}$$

i.e. averaging the products of all pairs of residuals, where the pairs are ‘neighbors’ within cluster. (Not what Liang & Zeger proposed, but what SAS’ PROC GENMOD does)

Unstructured:

$$\hat{\alpha}_{jj'} = \frac{1}{n} \sum_{i=1}^n e_{ij}e_{ij'},$$

i.e. averaging the products of all pairs of residuals at positions j and j' , over all clusters. (Adjust the denominator if clusters have different indexes j)

Together with *Independence*, which has no α to estimate, expect these methods (or minor variations of them) to be built into any decent GEE software. For others... roll up your sleeves.

GEE: estimating α

The GEE fitting algorithm:

1. Start with initial $\hat{\beta}^{(0)}$, obtained from e.g. a univariate analysis
 - typically, by using `glm()`
- 2.(a) Estimate scale parameter $\hat{\phi} = \frac{1}{(\sum_i n_i) - p} \sum_{i,j} \frac{(Y_{ij} - \hat{\mu}_{ij})^2}{S(\hat{\mu}_{ij})}$
 - (b) Calculate Pearson residuals (using current $\hat{\phi}, \hat{\beta}$)
 - (c) Calculate $\hat{\alpha}$, using Pearson residuals
3. Update $\hat{\beta}$, by plugging-in current $\hat{\alpha}, \hat{\phi}, \hat{\beta}$ and doing one Fisher-scoring step – i.e. by setting

$$\hat{\beta}^{(s+1)} = \hat{\beta}^{(s)} + \left(\sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i(\hat{\alpha})^{-1} \mathbf{D}_i \right)^{-1} \left(\sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i(\hat{\alpha})^{-1} (\mathbf{Y}_i - \hat{\mu}_i) \right)$$

Repeat 2 and 3 until convergence. This algorithm generalizes those seen in 570; it is one (good) way to solve simultaneously for $\hat{\alpha}, \hat{\phi}, \hat{\beta}$. Step 2(a) is not needed in linear regressions.

GEE: off-the-shelf software

A non-trivial factor in the success of GEE is the availability of free software that implements it, using syntax that is reasonably familiar to most users. We consider R's `gee package` – `geeM` is very similar, and uses `Matrix` structures.

- data stored in data frame, with one row per observation
- As well as a `formula`, an `id` argument is also specified, indicating cluster membership. Contiguous rows with matching `id` values form the clusters
- Like `glm()`, the familiar $y \sim x + z$ syntax specifies the linear predictor, which by default includes an intercept
- The assumed $S(\cdot)$ follows the `family` specification, exactly as with `glm()` in 570. Link functions are specified within the `family` argument
- The working correlation `R` must be specified, by `corstr`

GEE: dental growth!

Recall the dental growth data from Chap 2; we have $n = 27$ clusters (children), with measurements at $n_i = 4$ timepoints (8, 10, 12, 14 yrs old).

Coding the covariates as;

$$\mathbf{X}_{ij} = \{1, (\text{Age}_{ij} - 8), \text{Male}_{ij}, (\text{Age}_{ij} - 8) \times \text{Male}_{ij}\}$$

we fit the mean model

$$\mathbb{E}[Y_{ij} | \mathbf{X}_{ij}] = \beta_0 + \beta_1(\text{Age}_{ij} - 8) + \beta_2 \text{Male}_{ij} + \beta_3(\text{Age}_{ij} - 8) \times \text{Male}_{ij}$$

We will not use weights in the EE, i.e. we will use a constant mean-variance relationship as our working assumption. In formulae, this means we will set $S(\mu) = 1$, or equivalently

$$\mathbf{V}_i = \phi \mathbf{R}_i$$

GEE: dental growth!

Using `family=gaussian` is `gee()`'s default. So, after carefully setting up the data so contiguous rows represent contiguous observations within a cluster...

```
data(Orthodont, package="lme4")
d4      <- Orthodont           # using shorter names
d4$id   <- d4$Subject
d4$male <- d4$Sex=="Male"
```

... we investigate different choices of **R** as follows;

```
library("gee")

gee1 <- gee(distance~I(age-8)*male, id=id, data=d4, corstr="independence")
gee2 <- gee(distance~I(age-8)*male, id=id, data=d4, corstr="exchangeable")
gee3 <- gee(distance~I(age-8)*male, id=id, data=d4, corstr="AR-M", Mv=1)
gee4 <- gee(distance~I(age-8)*male, id=id, data=d4, corstr="unstructured")
```

The `summary()` command has a method for these objects, giving estimated standard errors (and much else)

GEE: dental growth!

The relevant parts of the results;

	$\hat{\beta}_0$ (SE)	$\hat{\beta}_1$ (SE)	$\hat{\beta}_2$ (SE)	$\hat{\beta}_3$ (SE)
1. Indept	21.2 (0.56)	0.48 (0.06)	1.41 (0.77)	0.30 (0.12)
2. Exchg	21.2 (0.56)	0.48 (0.06)	1.41 (0.77)	0.30 (0.12)
3. AR1	21.2 (0.59)	0.48 (0.06)	1.56 (0.82)	0.29 (0.12)
4. Unstr	21.2 (0.55)	0.48 (0.06)	1.41 (0.76)	0.31 (0.12)
Naïve <code>lm()</code>	21.2 (0.50)	0.48 (0.15)	1.41 (0.70)	0.30 (0.19)

- Independence and `lm()` give same estimates – as seen before
- In **this** linear regression, Independence and Exchangeable give identical results (due to balanced design and complete data – does not hold in general)
- Inference involving $\hat{\beta}_2$ changes somewhat – but likely not enough to alter conclusions, i.e. these results are not **sensitive** to these choices of R
- Naïve approach SEs for Age effects **are** notably off (and wrong)

GEE: dental growth!

The working.correlation from each gee.object;

$$2. \text{ Exchangeable } \mathbf{R}_i(\hat{\alpha}) = \begin{bmatrix} 1.00 & 0.61 & 0.61 & 0.61 \\ & 1.00 & 0.61 & 0.61 \\ & & 1.00 & 0.61 \\ & & & 1.00 \end{bmatrix}$$
$$3. \text{ AR-1 } \mathbf{R}_i(\hat{\alpha}) = \begin{bmatrix} 1.00 & 0.61 & 0.38 & 0.23 \\ & 1.00 & 0.61 & 0.38 \\ & & 1.00 & 0.61 \\ & & & 1.00 \end{bmatrix}$$
$$4. \text{ Unstructured } \mathbf{R}_i(\hat{\alpha}) = \begin{bmatrix} 1.00 & 0.50 & 0.74 & 0.51 \\ & 1.00 & 0.56 & 0.62 \\ & & 1.00 & 0.78 \\ & & & 1.00 \end{bmatrix}$$

GEE: dental growth!

Key bits of the output, with Independence Working Correlation:

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	21.2090909	0.5693435	37.251833	0.5604314	37.844221
I(age - 8)	0.4795455	0.1521635	3.151515	0.0631326	7.595845
male	1.4065341	0.7395990	1.901752	0.7737993	1.817699
I(age - 8):male	0.3048295	0.1976661	1.542143	0.1168673	2.608339
Estimated Scale Parameter:	5.093818				

- Naïve SE and Z statistics are the model-based variety that only use $\hat{\mathbf{A}}^{T-1}/n$, not $\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}/n$
- They are provided here **for reference only**. In practice you should use the Robust results **for GEE**, i.e. the sandwich
- The scale parameter $\equiv \phi$ – not really required, but useful for thinking about model-checking, and/or coding snafus. Also, for e.g. roughly-Poisson data, you may perhaps want to know how far ‘off’ you are. ϕ is estimated using a bias-corrected Pearson estimator (see 570 QL notes)

GEE: dental growth!

Key bits of the output, with AR-1 Working Correlation;

Coefficients:

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	21.1914047	0.6691392	31.669651	0.58665343	36.122528
I(age - 8)	0.4837647	0.1406311	3.439956	0.06291492	7.689189
male	1.5588576	0.8692373	1.793363	0.81581249	1.910804
I(age - 8):male	0.2856922	0.1826851	1.563850	0.12238030	2.334462
Estimated Scale Parameter:	5.099523				

- Naïve SE is the model-based variety, with weighting determined by $\mathbf{R}(\hat{\alpha})$ – i.e. an example of WLS as seen in 533. If the full variance specification is right, these **are** valid (see Chap 3) but usual motivation for GEE is not making binding assumptions about the covariance
- Again, the Robust output is from the sandwich, and uses our working covariance assumptions
- Could use Naïve/Robust differences as an (informal) diagnostic, **for the model-based inference**

GEE: more on off-the-shelf software

- `gee()` supports a wide range of `family` options
- Observations with relevant missing values are omitted. Comparing $\hat{\beta}$, `Est.std.err`($\hat{\beta}$) when we do/don't adjust for some Z , this may mislead
- If something relevant to β affects the size of n_i (e.g. people dropping out early due to treatment working badly, a.k.a. *informative missing*) expect bias, compared to having full data where every n_i is equal
- Get rid of the irritating `Cgee` message by making a copy of the `gee()` function, and removing its `message()` call
- Too-small limits on cluster size apply ($n_i \leq 32$); to fix this hack the underlying C code – or code by hand, or use `geeM`

`gee()` is somewhat creaky, but used in most reference material, which is why it's presented here.

GEE: more on off-the-shelf software

Also be aware of the `geepack` package, which features a `geeglm()` function.

- Has essentially the same syntax as `gee()`
- It calls function `geese()`, which in turn calls `geese.fit()`
- Slightly different internal calculations, e.g. dispersion parameters, sandwich estimate, which are not (quite) those given here; see the documentation for details
- Output is designed to mimic that of `glm()`, which is good if you are intricately familiar with that setup
- Permits use of `anova()`... should you want that
- Messes around with `options("digits")`, on my machine
- Can crash R (!) – see the documentation

GEE: working correlation

Calling the inverse-weighting matrix \mathbf{V}_i the ‘working’ covariance matrix designation emphasizes that exact knowledge of the true covariance is not required. We get consistency, asymptotic Normality, and valid sandwich-based intervals asymptotically, **regardless** of the chosen \mathbf{V}_i .

(If we actually knew the variance structure of the outcomes, simpler and more stable sandwich estimates are available – as per the cancellation seen in 570 when $\mathbf{A} = \mathbf{B}$ for QL. But this is not standard; robustness to variance assumptions is alluring.)

But just as in 570’s study of efficient EEs, in GEE we get **efficient** estimates by choosing the ‘working’ \mathbf{V}_i as close as possible to the true variance-covariance matrix of elements in each \mathbf{Y}_i – see Song’s book (pg 91) for a formal statement.

Practically, how much is efficiency affected by choice of \mathbf{R} ?

GEE: choice of working correlation

Asymptotic Relative Efficiency, relative to the ‘right’ choice) for $n_i = 10^*$; (Liang and Zeger 1986, Table 1)

True Corr'n	α	Working correlation		
		Indep	Exch	AR1
Exch	0.3	0.99	1.00	0.95
	0.7	0.99	1.00	0.72
AR	0.3	0.97	0.97	1.00
	0.7	0.88	0.88	1.00

- Minor correlation \Rightarrow choice doesn’t matter
- Large correlation \Rightarrow large efficiency loss possible

Note that the choice of working correlation structure should be based on external information – if we ‘try some and just pick the best’ we induce the ‘wiggle room’ problem discussed in 570, which invalidates confidence statements, Type I error rates, etc.

* $n_i = 10$, $\mu_{it} = 1 + t$, $\text{Var}[Y_{it}|t] = \sigma^2$

GEE: choice of working correlation

ARE also depends on covariate design (Fitzmaurice, 1995)

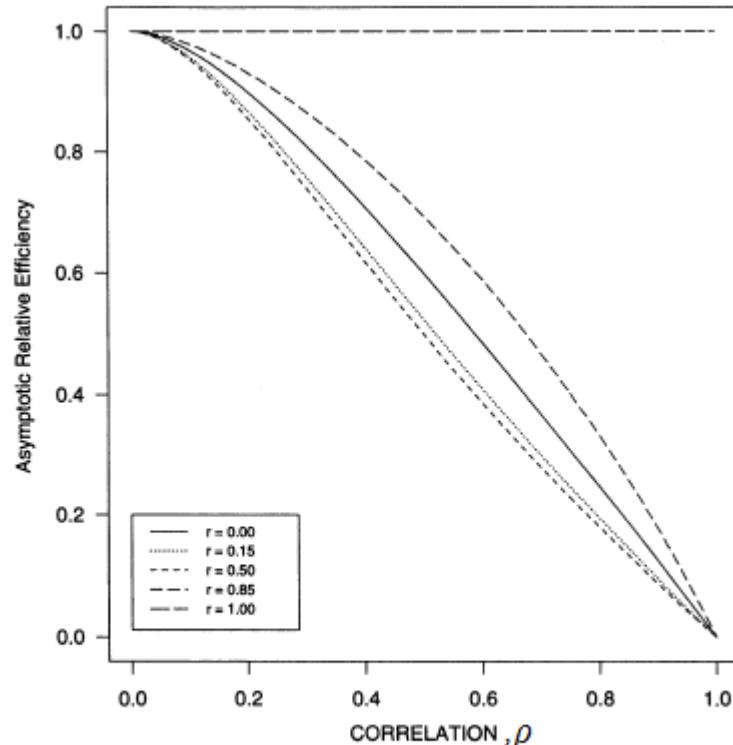


Figure 2. Asymptotic efficiency of the Independence GEE estimator, relative to the Exchangeable estimator, for selected values of the intra-cluster correlation for the covariate (and when $T = 5$).

$$T = n_i = 5$$
$$\text{logit}(\mu_{ij}) = \beta x_{ij}$$

$$\rho = \text{Corr}[Y_{ij}, Y_{ij'}], j \neq j'$$

(i.e. usual α)

$$r = \text{Corr}[X_{ij}, X_{ij'}], j \neq j'$$

(with \mathbf{X} exchangeable)

Large $\alpha \Rightarrow$ big efficiency loss, but not if every x_{ij} is identical

GEE: choice of working correlation

Cluster size also matters (Mancl and Leroux 1996, $\rho_y = \alpha$)

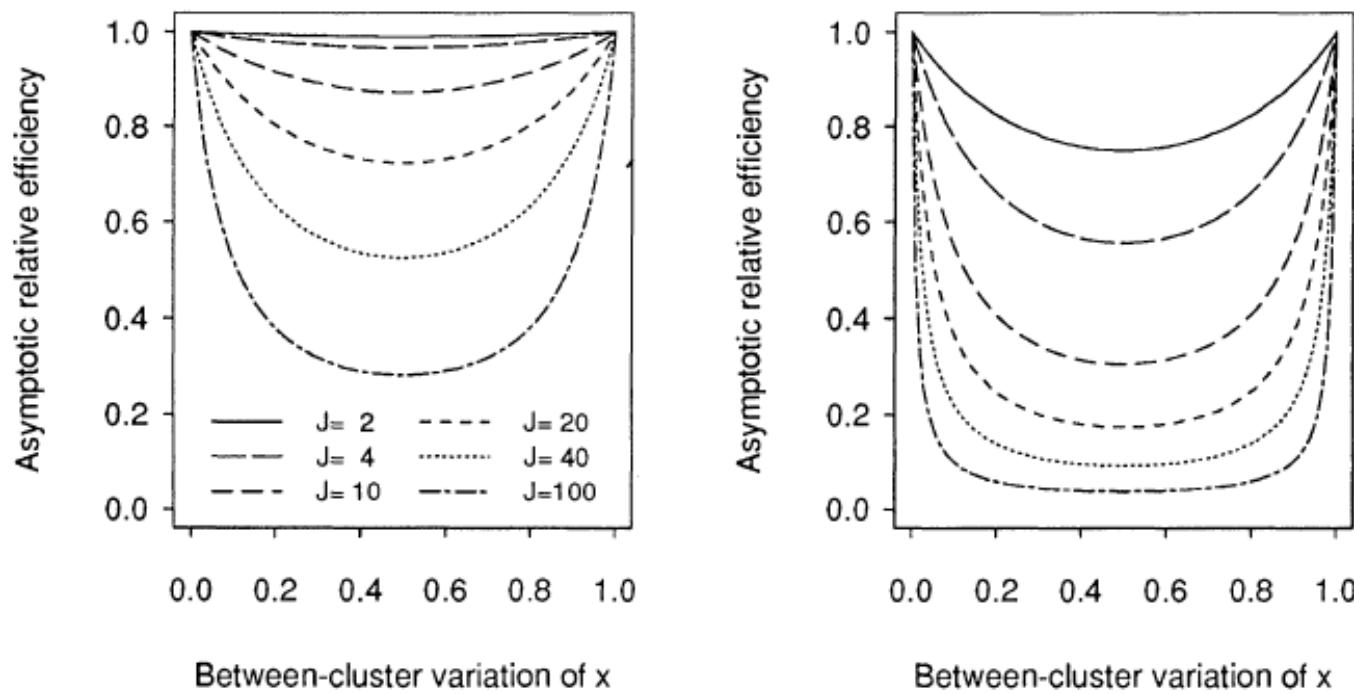
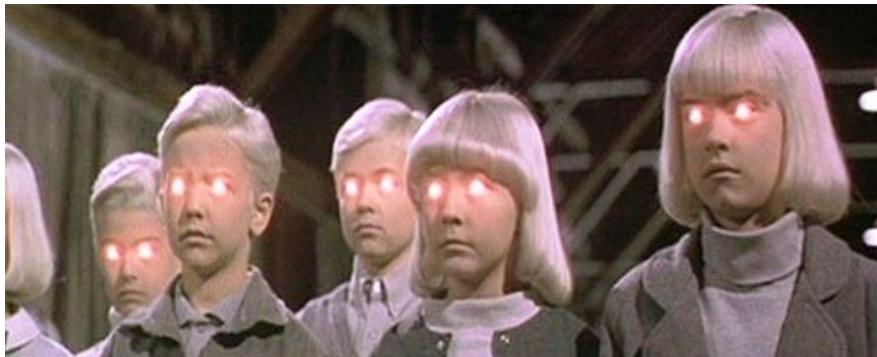


Figure 1. Asymptotic relative efficiency of independence to exchangeable for the case of constant weights, equal cluster sizes (J), and common pairwise correlation between responses of (a) $\rho_y = 0.1$ and (b) $\rho_y = 0.5$.

- Small clusters \Rightarrow possibly minor efficiency loss
- Large clusters \Rightarrow possibly major efficiency loss

GEE: choice of working correlation

Q. If clusters=classes, would ρ be large/small in these? How would working R as exchangeable compare to independent?



Village of The Damned
a.k.a.
The Midwich Cuckoos



The Prime of Miss Jean Brodie



The Breakfast Club

GEE: choice of working correlation

A **very** much more trivial reason to choose exchangeable, independence and unstructured working correlation matrices over AR- M is that they avoid this problem;

```
> d5 <- d4[ order(d4$id, as.vector(replicate(27, sample(1:4)))) ], ]  
# shuffling the observations, within each cluster  
> summary( gee(distance~I(age-8)*male, id=id, data=d5, corstr="AR-M", Mv=1) )  
              Estimate Naive S.E.  Naive z Robust S.E.  Robust z  
(Intercept)    21.1515956  0.60135890 35.172998  0.57868199 36.551329  
I(age - 8)      0.4733017  0.09138209  5.179370  0.06031453  7.847225  
male            1.6414661  0.78349024  2.095069  0.76382814  2.148999  
I(age - 8):male 0.3214420  0.12109656  2.654427  0.10085284  3.187238
```

- Makes enough difference to matter (and is just wrong)
- For *exchangeable* working correlations, we can **exchange** the order of the observations and not change the estimating equations. Independence is a special case of exchangeable, so the same property holds

GEE: choice of working correlation

Does choice of R matter? It depends who you ask...

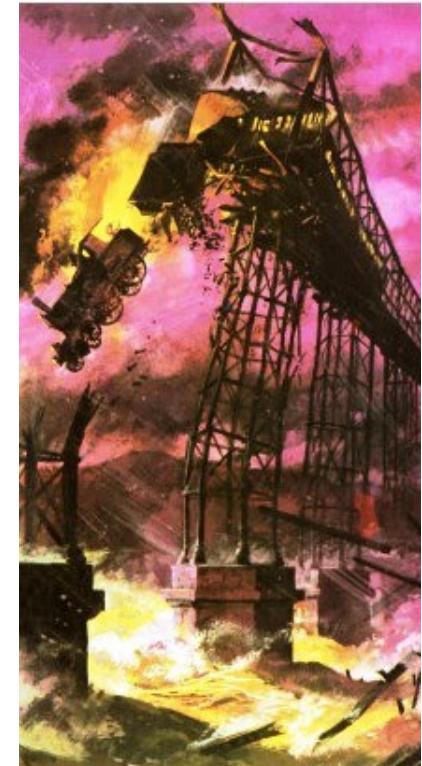
- Liang and Zeger (1986): ... *little difference when correlation is moderate.*
- McDonald (1993): The independence working assumption *may be recommended for practical purposes*
- Zhao, Prentice and Self (1992): assuming independence *can lead to important losses of efficiency*
- Fitzmaurice, Laird and Rotnitzky (1993): ... *important to obtain a close approximation to Cov[\mathbf{Y}_i] in order to achieve high efficiency*

... in fairness, they'd all also tell you it depends on the applied context. Whether efficiency matters **in your context** will depend on how much precision you (plausibly) have; with high/low precision, the truth will be obvious/obviously absent. The true correlation **and** $\{n, \mathbf{X}, \boldsymbol{\beta}, \phi\}$ all determine this precision.

GEE: diagnostics

*Beautiful Railway Bridge of the Silv'ry Tay!
Alas! I am very sorry to say
That ninety lives have been taken away
On the last Sabbath day of 1879,
Which will be remember'd for a very long time.*

*As soon as the catastrophe came to be known
The alarm from mouth to mouth was blown,
And the cry rang out all o'er the town,
Good Heavens! the Tay Bridge is blown down*



From 'The Tay Bridge Disaster' (1880)
by William McGonagall (1825–1902)
Scottish weaver and actor
...also acclaimed as the **worst poet in history**

GEE: diagnostics

Before we develop diagnostics for GEE analyses, note we are assuming that **essential** sanity checks have already been implemented;

- You have checked the data, including plotting it, to ensure that there are no coding errors, or other errors of data input. (Obviously, contextual knowledge is required to do this well)
- In regression, you have chosen to adjust for confounders (i.e. include them in the mean model) in a way that reflects plausible causal assumptions
- The link function produces models that are scientifically plausible and/or relevant – this one is typically not difficult in practice
- The working correlation matrix is also at least plausible

After this, think of diagnostics as **some** checks that your GEE model is not **catastrophically** bad. Note that diagnostics do **not** confirm your model is perfect (it's not) and do **not** reliably identify all problems that may be serious.

GEE: diagnostics

Independence of cluster outcomes $\{\mathbf{Y}_i | \mathbf{X}_i = \mathbf{x}_i\}$ is usually clear from context. We'll focus on the other assumptions in GEE;

- n is big enough so that asymptotic approximations to the distribution of $\hat{\beta}$ work well; re-running the analysis after 'leaving-one-out' should not affect answers much*
- GEE inference relies** on mean model assumptions being correct – so error terms are mean zero, for all covariate values

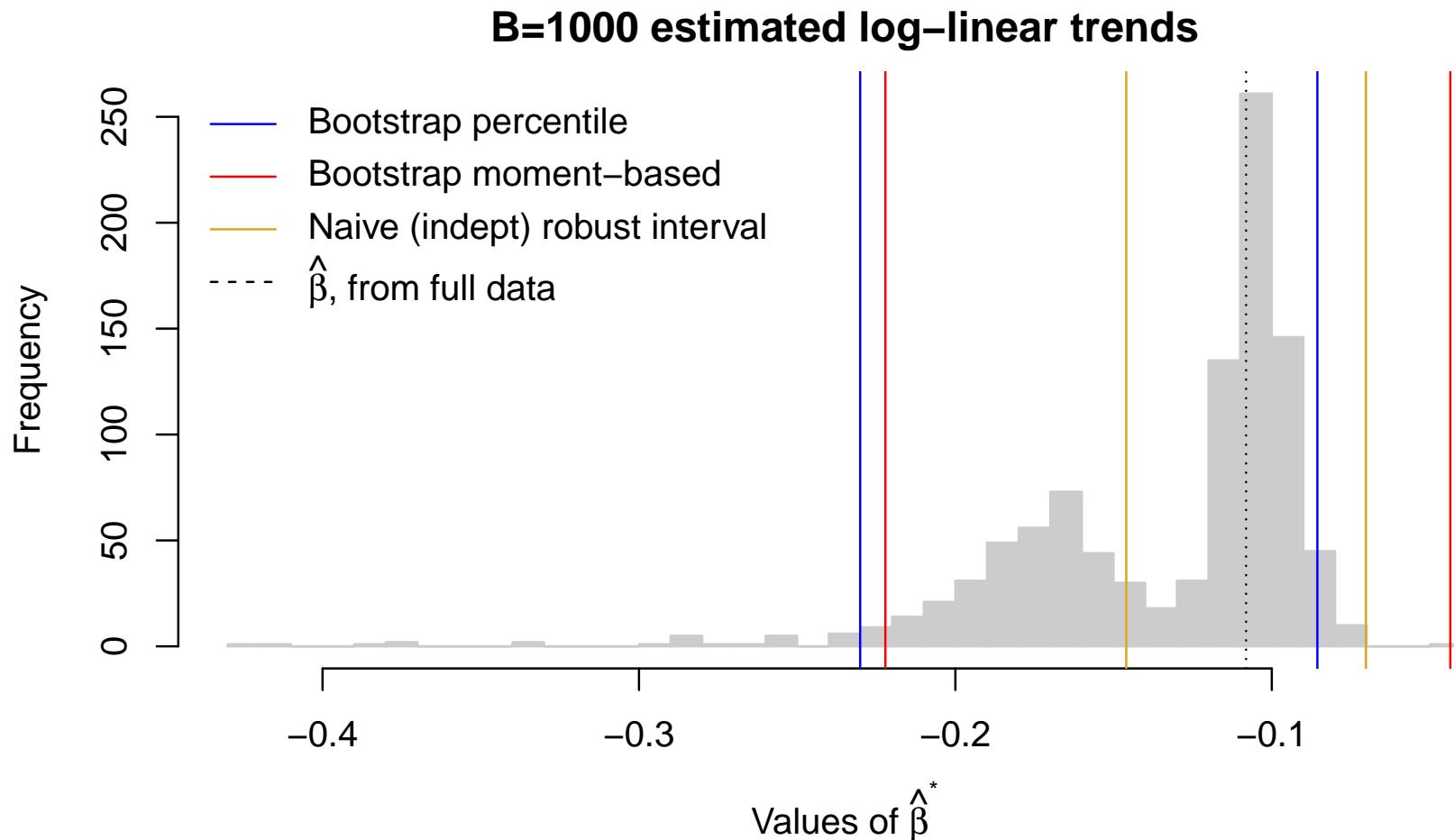
We'll use leave-one-out diagnostics to assess the first assumption, and plots of residuals assess the second. Both methods can spot egregious errors – which is better than nothing – but otherwise expect power to be “modest”.

* Even more pragmatically, if one cluster largely determines the inference, the inference won't be very convincing, and it's useful to know that. This is also a concern in nonparametric approaches

** ...in general! But even being nonparametric, if a ‘fitted line’ doesn't summarize the mean response well, it's *arguably* not a useful summary

Diagnostics: leave one out

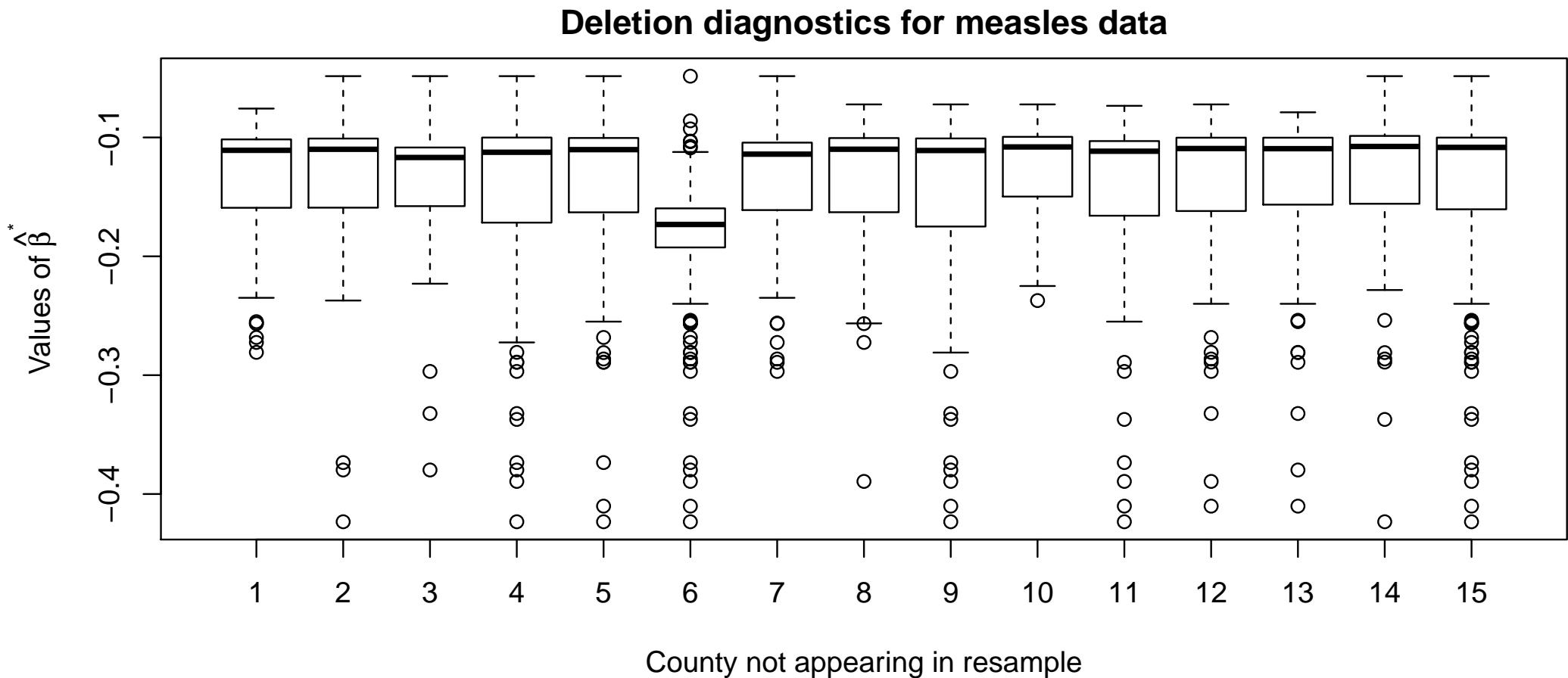
'Slopes' $\hat{\beta}^*$ bootstrapping the measles example ($n = 15, \mathbf{R}_i = I_{n_i}$)



Q. Do you think $\hat{\beta} \sim N(\cdot, \cdot)$ under \tilde{F} ? Under F ?

Diagnostics: leave one out

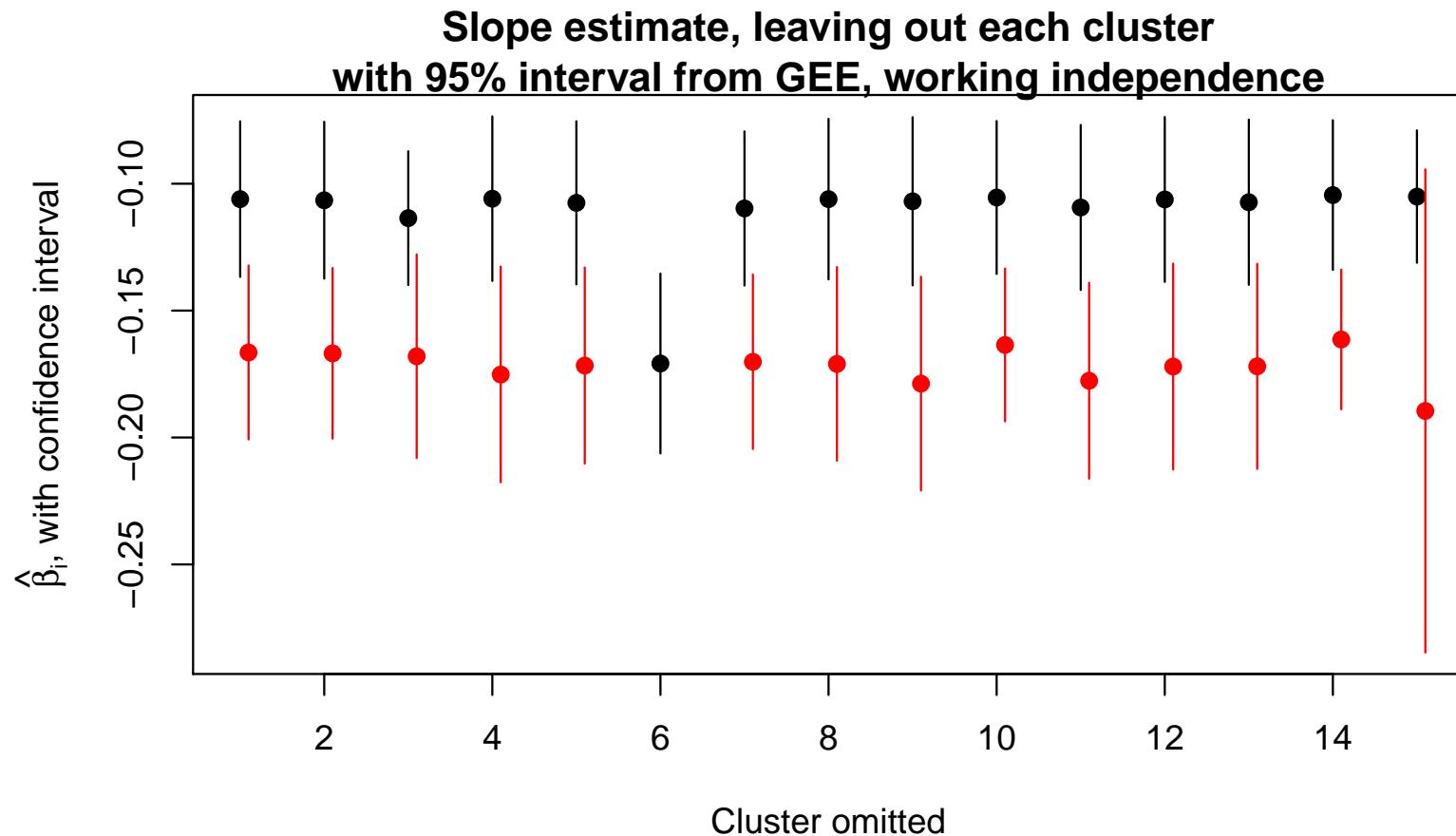
Look at $\hat{\beta}^*$ where each observed cluster i is ‘left out’;



Which $\{Y_i, X_i\}$ affects $\hat{\beta}$ most? What does this cluster do to $\hat{\beta}$?

Diagnostics: leave one out

A close (& fast) analog, using GEE; (Could also use full robust sandwiches in the same way)



Red indicates omitting cluster 6 and one other cluster.

Diagnostics: leave one out (*)

Preisser & Qaqish (1996) show how to linearly approximate $DFBETA_i = \hat{\beta} - \hat{\beta}_{[i]}$, in GEE.

- The calculations are exact, using linear links
- Honestly, the approximation is nice but seldom critical; when each call to `gee()` takes ≤ 1 second, 1000 of them take ≤ 17 mins. Go have a (quick) coffee instead.
- If the variance is also assumed correct, we can normalize the approximate $DFBETA_i$ by their expected values, to obtain a generalized Cook's Distance. (Preisser & Qaqish again)
- ‘One step’ approximations of $\hat{\beta}$ (one Fisher scoring step) are consistent for β , and in practice *almost* as good as $\hat{\beta}$. This is useful in high-throughput work, doing e.g. millions of GEEs

Keen people: I have R code for these $DFBETAs$ for GEE, awaiting release. PROC GENMOD in SAS will do them, Google will show you some XLISP-STAT code too.

Diagnostics: plotting residuals

Knowing which cluster has greatest influence does not tell you *why* that cluster is so influential. Explanations may be some combination of the following;

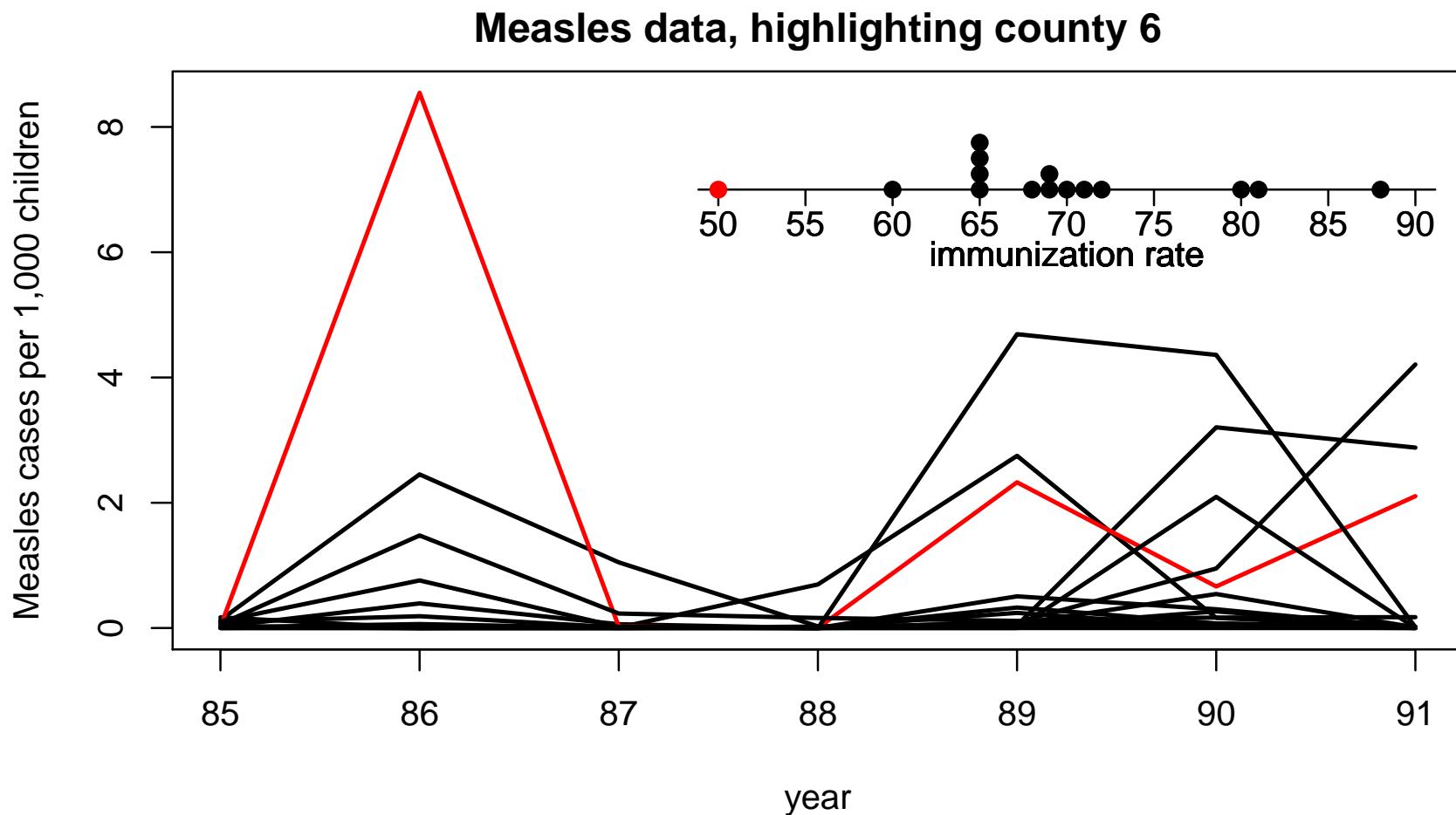
- Extreme covariate values – which are hard to ‘fix’, except by fitting a different regression
- Wrong mean model leading to large residuals (positive or negative)
- Inaccurate variance assumption leading to large standardized residuals (in absolute value)
- Extreme $Y_i - \mu_i$ – perhaps not ‘wrong’, just due to chance?

Plotting residuals against covariate values, or fitted values may enable us to diagnose what’s going ‘wrong’. But beware over-interpretation; **some** data point **has** to look worst.

As in 570, picking models for data **based on that data** can also throw off e.g. 95% coverage. Also, if resources are limited, it may not be feasible to code up ‘bespoke’ mean-variance relationships or $R(\alpha)$, that might better capture the truth.

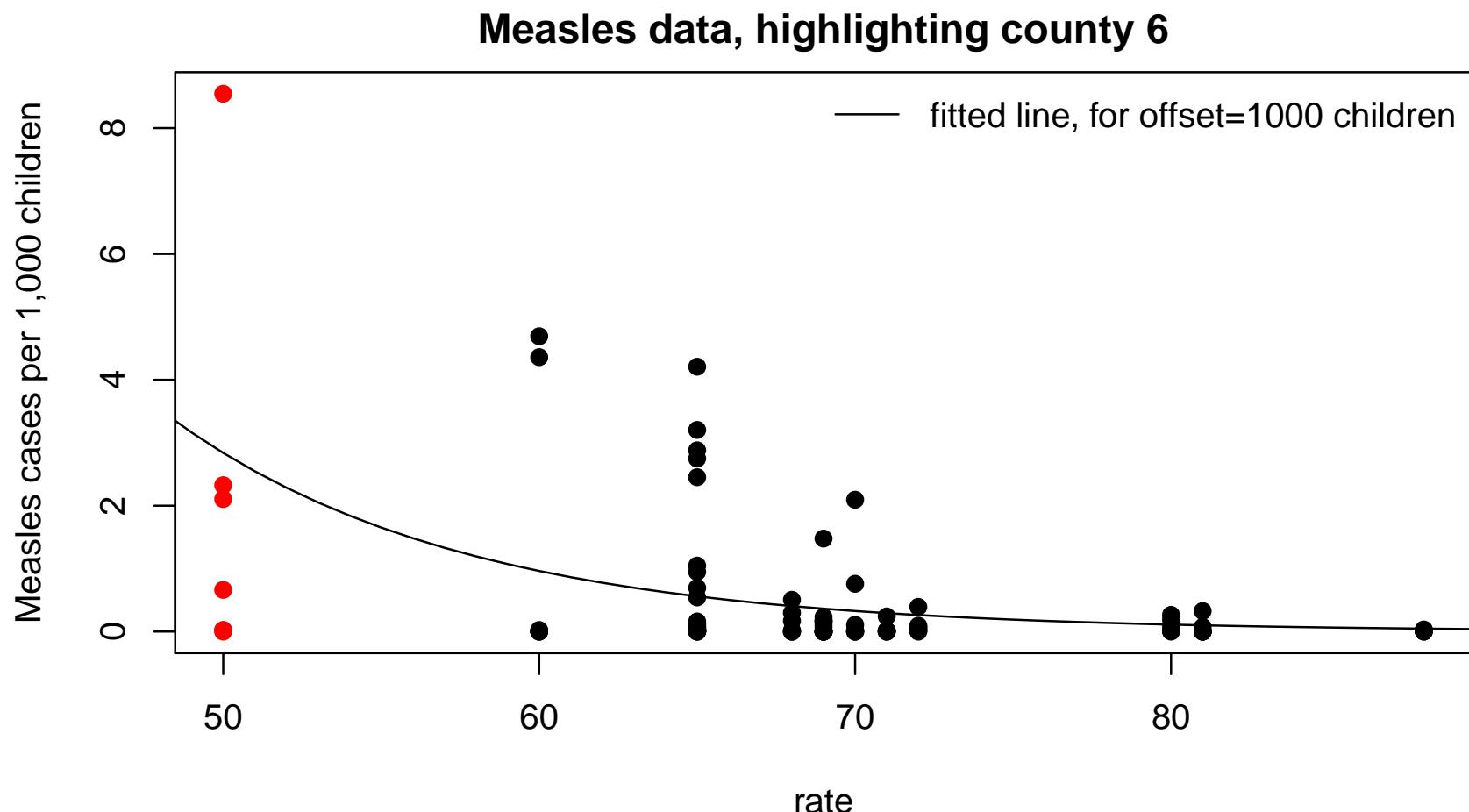
Diagnostics: plotting residuals

Plotting the data, try to explain why omitting cluster 6 reduced $\hat{\beta}$, the association between measles and immunization rates;



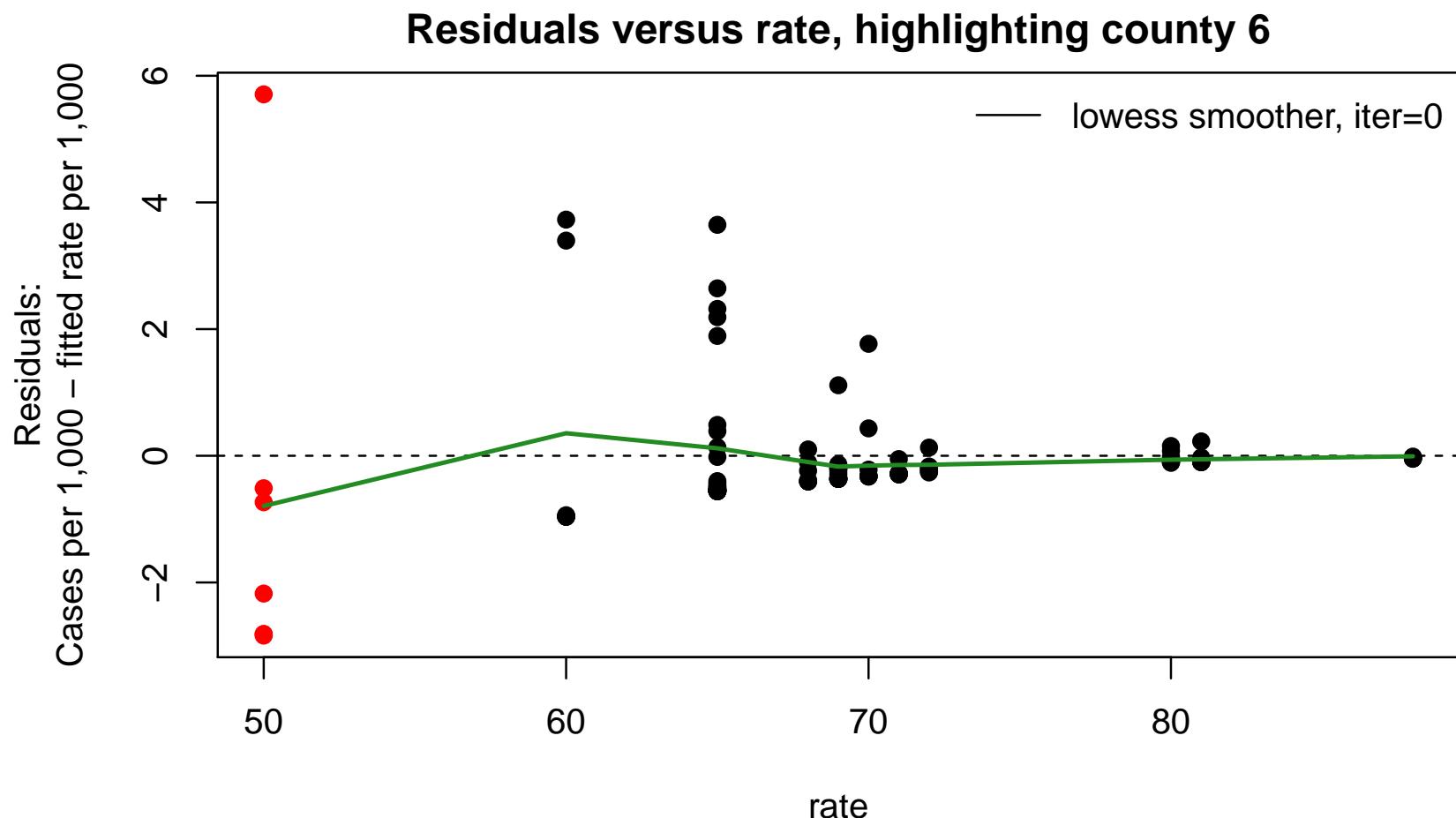
Diagnostics: plotting residuals

Without the time axis, we see rates *below* the fitted values, in cluster 6;



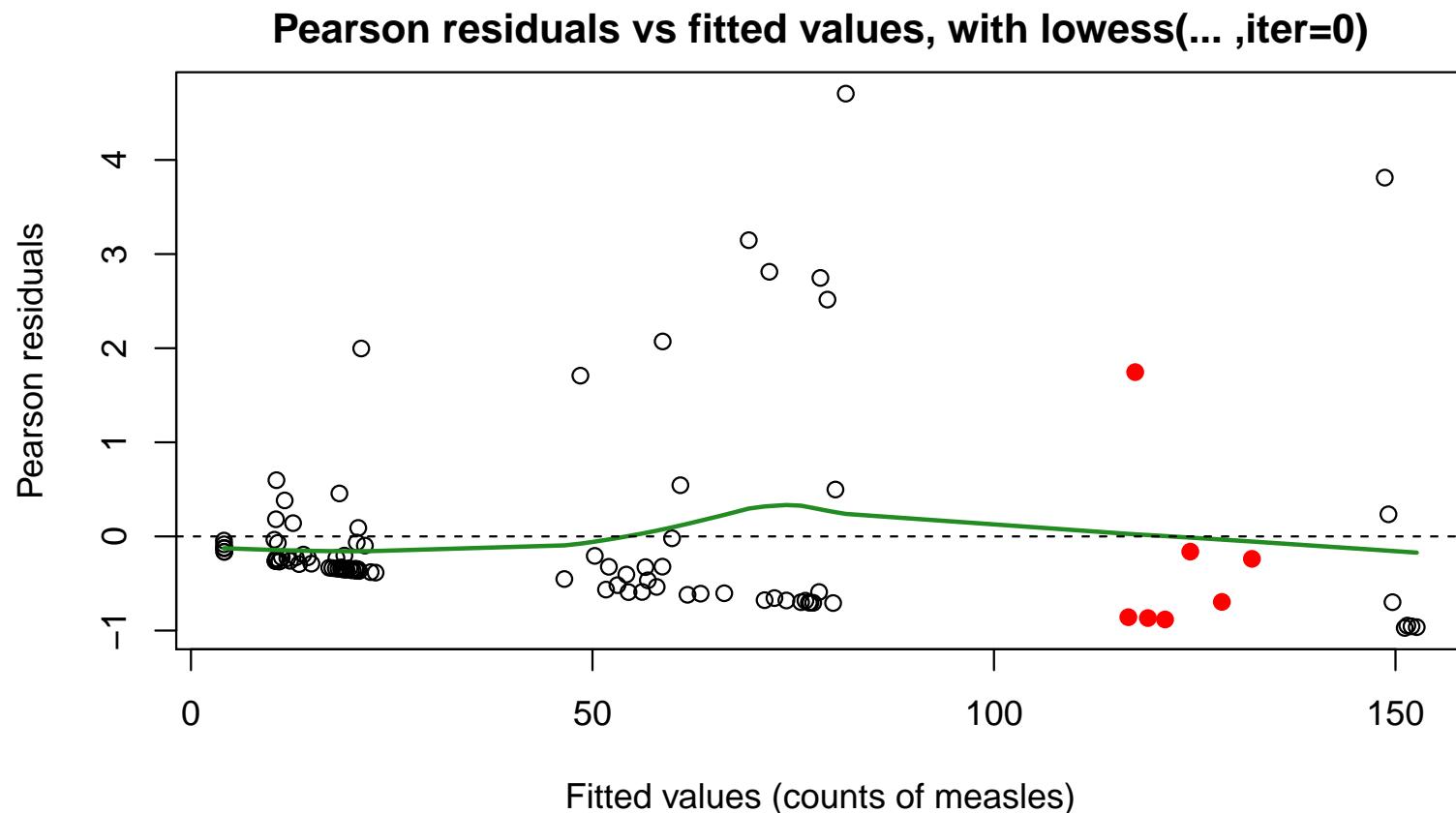
Diagnostics: plotting residuals

Residuals versus rate (and a lowess(... iter=0) curve) shows this better;



Diagnostics: plotting residuals

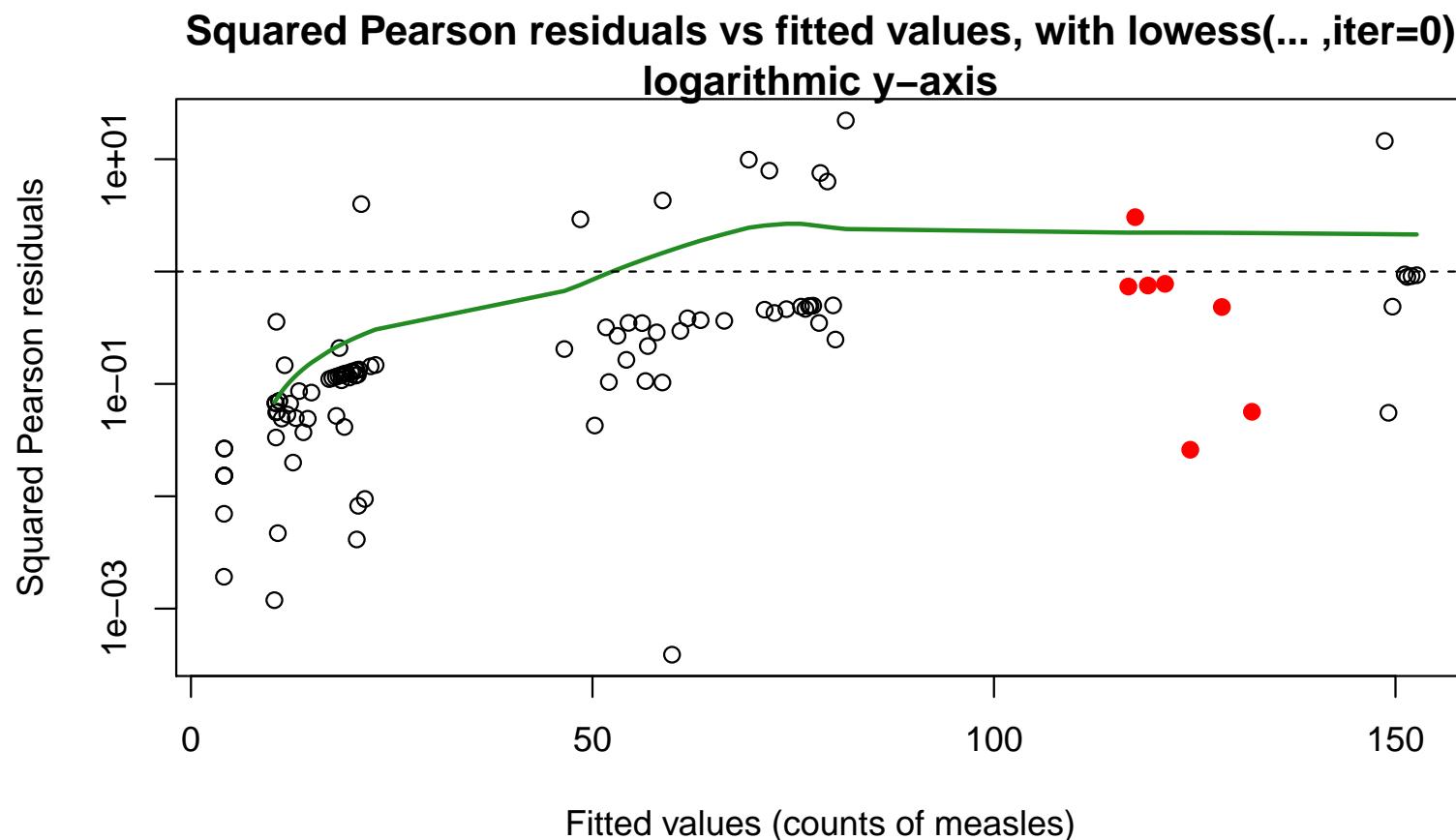
Using Pearson residuals (see $\hat{\alpha}$ estimates) we expect residuals to have mean zero at all fitted values, e.g.



The fitted mean looks okay – note this plot doesn't show extreme rate in cluster 6.

Diagnostics: plotting residuals

If the variance assumption $S(\cdot)$ is *additionally* correct, residuals² should have mean 1 everywhere;



... but remember this only affects efficiency, not large- n validity

Diagnostics: plotting residuals

Observations so far; (in descending order of reliability)

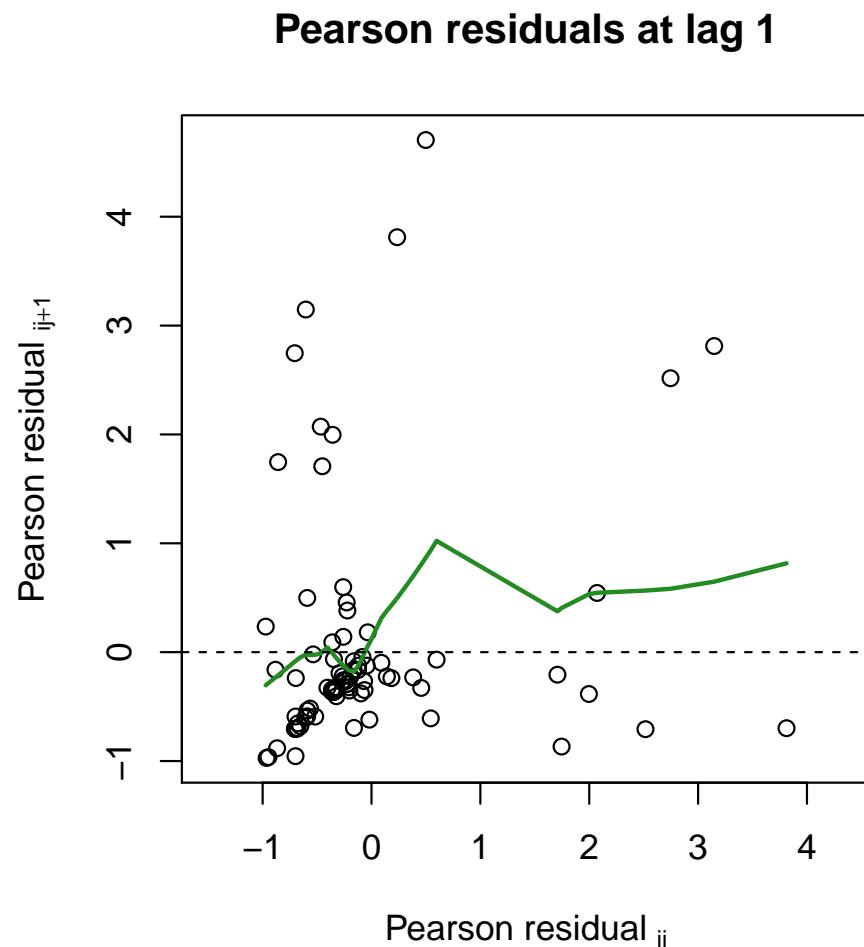
- Cluster 6 has high influence
- In part, this seems to be due to it having high leverage, i.e. X values that are far from the mean
- The fitted mean model doesn't appear to be grossly wrong (but at Cluster 6 we wouldn't expect that, as it has high leverage)
- There's also some suggestion that the working variance assumption is violated (but this doesn't affect validity, assuming large-sample properties)

With more covariates, leverage should be assessed on multiple X axes – e.g. each covariate, or combination of covariates. In practice this is hard – though context may indicate which observations ‘drive’ the estimate of interest.

Diagnostics: others

Checks of the correlation matrix are similarly for efficiency.

In the measles example, we used working independence. If true, Pearson residuals from the same i but differing by one time-period j should be uncorrelated;



Diagnostics: others

For a more general form of plot addressing the working correlation, first recall that

$$\text{Var}[Z_1 - Z_2] = \text{Var}[Z_1] + \text{Var}[Z_2] - 2\text{Cov}[Z_1, Z_2]$$

Therefore, for pairs of observations where we believe the correlation actually is ρ , we should find that

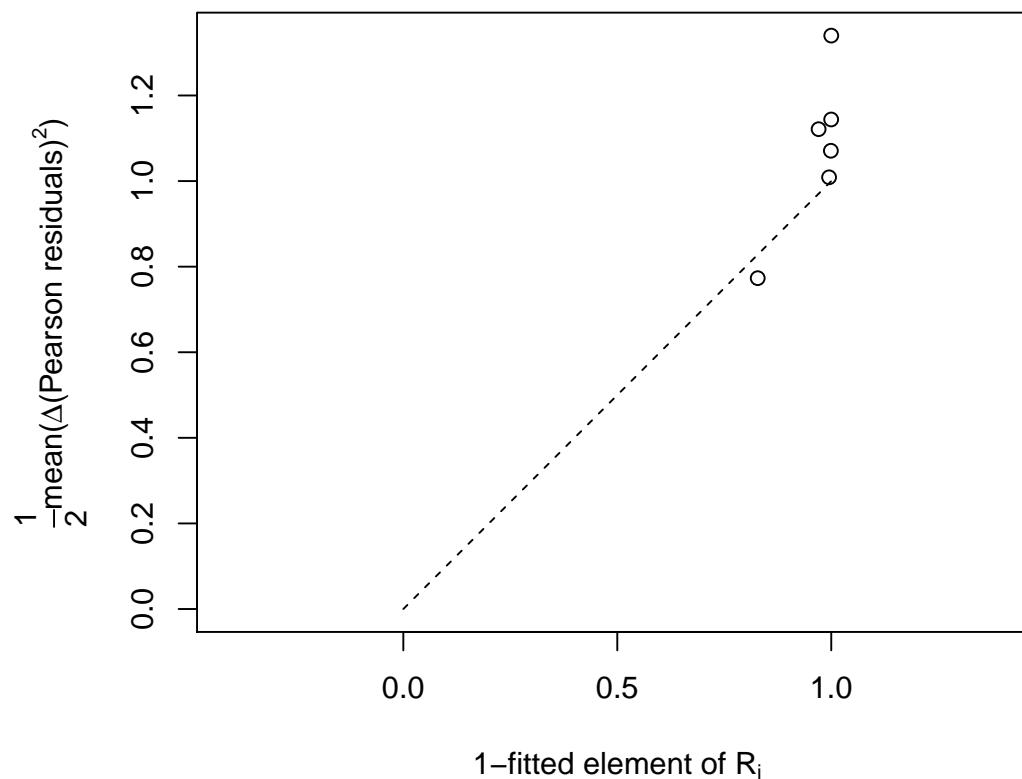
$$\frac{1}{2}\text{Mean}_{\text{pairs } j,j'} \left\{ (\text{Pearson residual}_j - \text{Pearson residual}_{j'})^2 \right\} \approx 1 - \rho$$

- Averaging this mean for nearby values of ρ (i.e. using a smoother) is equivalent to using the semi-variogram (see Chapter 1)
- Note the justification *does* rely on the Pearson residuals all having mean ≈ 0 (or at least mean difference ≈ 0) and variance ≈ 1 . These properties may not hold.

Diagnostics: others

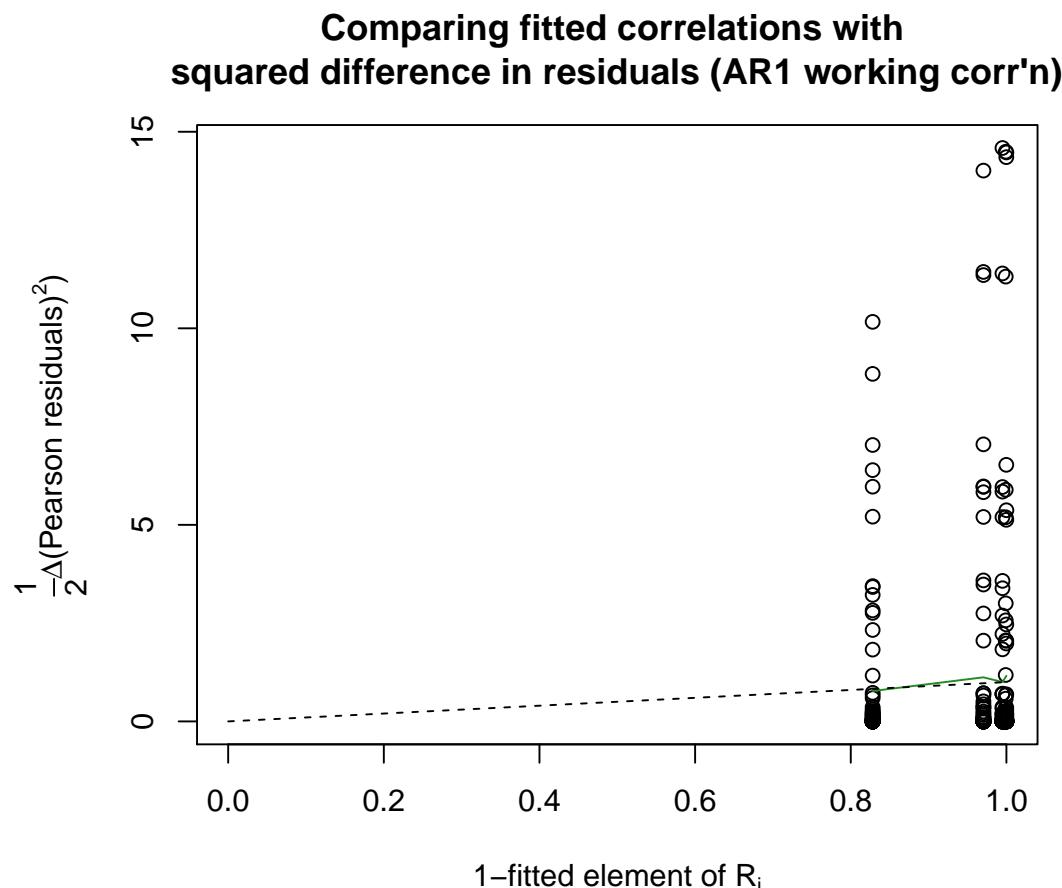
For the measles example, working independence gave $\alpha \equiv 0$. So instead we illustrate the plot using the fit with an AR-1 working assumption, in which (it turns out) $R_i(\hat{\alpha})_{jj'} = 0.171^{|j-j'|}$;

Comparing fitted to correlations to
squared difference in residuals (AR1 working corr'n)



Diagnostics: others

Plotting the squared difference in residuals against the fitted correlation and smoothing (green line) tells the same story – although perhaps less clearly;



Diagnostics: summary

Song pg 102 has more discussion of diagnostics. It's important to keep diagnostics in perspective; their power is generally low, and GEE only requires that the mean model be correct.

Some pragmatic advice;

When regression coefficients are the scientific focus, as in the examples here, one should invest the lion's share of time in modeling the mean structure, while using a reasonable approximation to the covariance. The robustness of the inference about β can be checked by fitting a final model using different covariance assumptions and comparing the two sets of estimates and their robust standard errors. If they differ substantially, a more careful treatment of the covariance model may be necessary.

Diagnostics: summary

To summarize, while good models are grounded in ‘the science’;

- Diagnostics *may* stop you fitting egregiously wrong mean models, and this can be helpful
- ‘Inaccurate’ working variance assumptions (e.g. choice of $S(\cdot)$) may also be caught, but power to check this second-order property is likely (even) lower
- Checking complex correlation structures is also hard. (But note that a reasonable working \mathbf{R} may be obvious under your design, particularly when correlations are weak)
- Look for patterns in the whole plot, not just the extremes. ‘Trends’ in only a few data points are easily overinterpreted – and doing so is a sign of being statistically naïve
- Neither working variance nor correlation *need* be right – but this doesn’t make choice of \mathbf{V} ignorable; precision matters

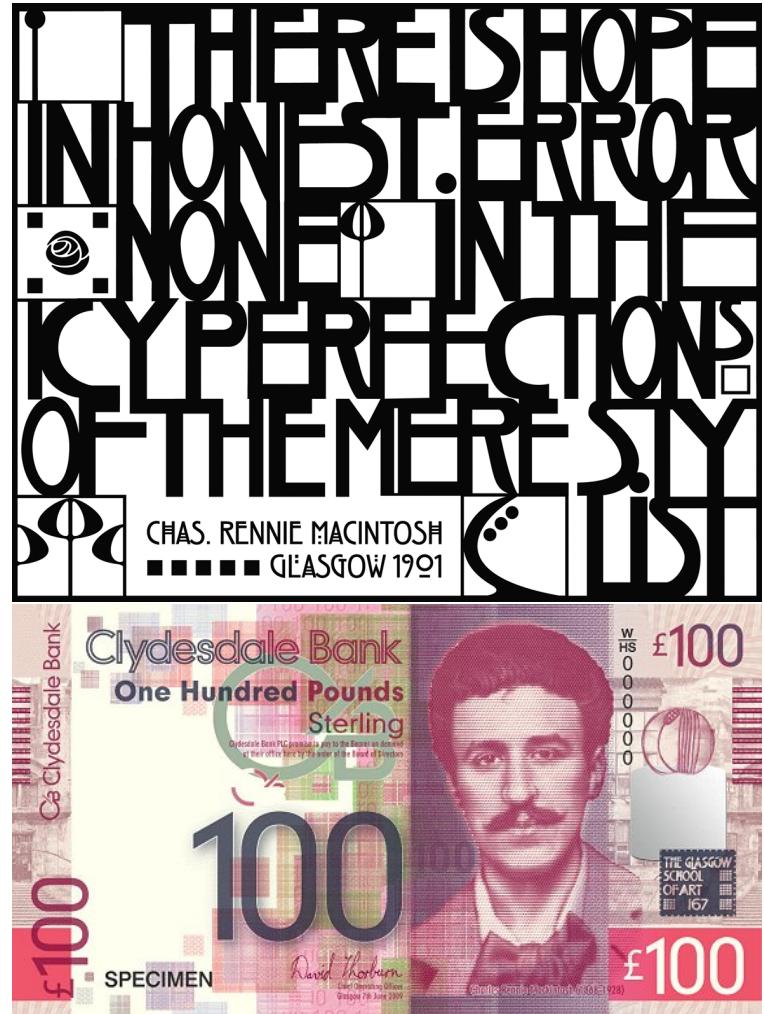
Expect to have to code diagnostic plots by hand, when using R.

GEE: limitations

*There is hope in honest error;
None in the icy perfections of the
mere stylist*

Charles Rennie Mackintosh

(1868–1928)



GEE: limitations

GEE is a standard method, and is recommended as a default, but it's not icy perfection, and honest appraisals also note its shortcomings.

In addition to the issues already seen (i.e. small-sample behavior, correct mean-model assumptions, always learning about marginal parameters) we will discuss;

- **Moment restrictions:** GEE separates the mean model from issues of covariance – but this ignores the fact that **in some situations** the mean **must** be informative about the true covariance within a cluster
- **Conditioning on ‘current’ covariate values:** not only can this take us away from questions of interest (see slide 2.41) but it can also lead to very misleading inference
- **Missing data:** and GEE’s sensitivity to it

GEE: moment restrictions

Data from an apparently *very* easy regression problem; you are interested in the proportion of males ($Y = 1$) in a population. To assess this data, you have sampled households (by physician visit) and counted the number of men, and women;

		$\sum_{j=1}^{n_i} Y_{ij}$							
		0	1	2	3	4	5	6	7
n_i	1	1	0						
	2	0	4	0					
	3	0	7	5	0				
	4	0	1	3	5	0			
	5	0	1	0	1	0	0		
	6	0	0	0	1	0	0	0	
	7	0	0	0	1	0	0	0	0

- $N = 30$ clusters altogether
- Mean outcome in this sample is 0.51, i.e. 51%
- ... equivalently $\widehat{\text{Odds[Male]}} = \log(0.51/0.49) = 0.038$

GEE: moment restrictions

But how to get a confidence interval for this log odds estimate? It seems reasonable to treat it as the intercept, in a GEE logistic regression with no other covariates;

```
> dat <- read.table("nocovarsgee.txt",header=TRUE)
> gee( y ~ 1, data=dat, id=id, family=binomial, corstr="exchangeable",
+ scale.fix=TRUE) # forces phi=1, which is sensible here
Coefficients:
(Intercept)
1.124833
Estimated Scale Parameter: 1.691984
Number of Iterations: 25
Working Correlation[1:4,1:4]
 [,1]      [,2]      [,3]      [,4]
[1,] 1.0000000 -0.1737041 -0.1737041 -0.1737041
[2,] -0.1737041 1.0000000 -0.1737041 -0.1737041
[3,] -0.1737041 -0.1737041 1.0000000 -0.1737041
[4,] -0.1737041 -0.1737041 -0.1737041 1.0000000
Warning messages:
Maximum number of iterations consumed
Convergence not achieved; results suspect
Gee had an error (code= 104). Results suspect.
Working correlation estimate not positive definite
```

GEE: moment restrictions

This is not an R-specific problem – here's Stata's version*;

```
. insheet using nocovarsgee.csv
. xtset id
. xtgee y, family(binomial) // defaults to exchangeable working corr'n
(resetting alpha to -0.1650)
Iteration 1: tolerance = .23848905
(resetting alpha to -0.1650)
Iteration 2: tolerance = .00056247
(resetting alpha to -0.1650)
Iteration 3: tolerance = 1.995e-08

Scale parameter:                               1      Prob > chi2      =
-----+-----+-----+-----+-----+-----+
y |     Coef.    Std. Err.      z      P>|z|      [95% Conf. Interval]
-----+-----+-----+-----+-----+-----+
_cons |  -.2098767   .0670603    -3.13    0.002    -.3413124    -.078441
convergence not achieved
```

Note we get a different estimate — but also a warning.

* You are not expected to use Stata in this course

GEE: moment restrictions

Some things that are *not* going wrong here;

- The mean model *cannot* be wrong; the proportion of males in the population *is* some number – and there are no covariates to worry about
- The stated variance cannot be wrong either. As you know from e.g. Stat 512, if $\mathbb{E}[Y] = p$ and Y is binary, then

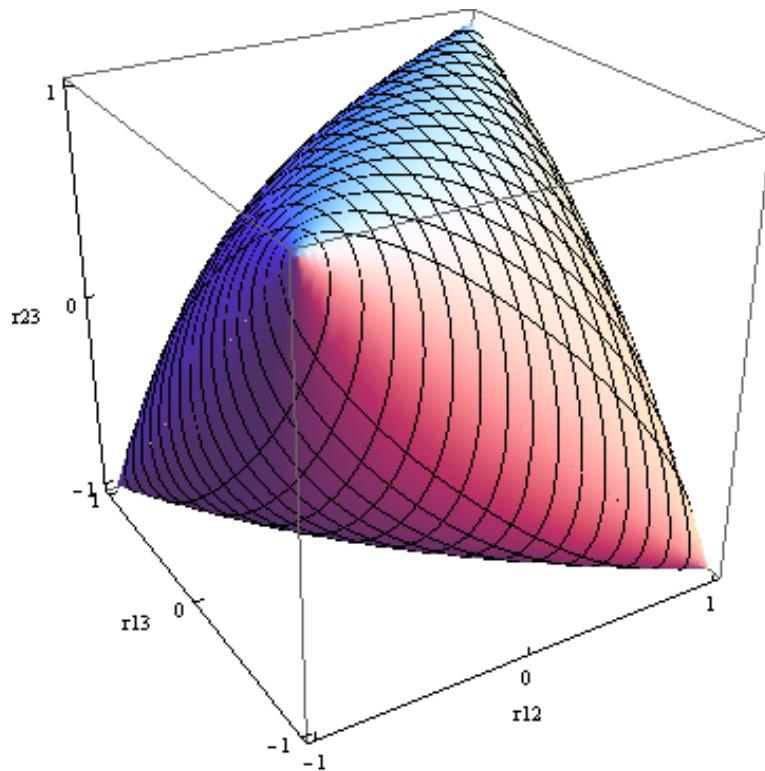
$$\begin{aligned}\text{Var}[Y] &= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \\ &= p - p^2 \\ &= p(1 - p)\end{aligned}$$

- In R we forced `scale.fix=TRUE`, which forces $\hat{\phi} = 1$. Stata does this by default, so the working assumption that the variance is $\phi V(\mu) = \phi\mu(1 - \mu) = p(1 - p)$ is **correct**
- Independence of clusters is plausible here – and even if it weren't, we'd still expect to get **some** answer, not just error messages

The only suspect left is the working correlation matrix...

GEE: moment restrictions

We know that correlation matrices must be symmetric. But they must also be positive semi-definite;



For example, three variables cannot all have correlation -1 with each other.

GEE: moment restrictions

Our exchangeable working assumption uses $\rho_{12} = \rho_{13} = \rho_{23}$.

How low can ρ be, inside the teabag? an elliptical tetrahedron!

$$n_i = 2$$

$$\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

$$n_i = 3$$

$$\begin{pmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1 & -1/2 \\ -1/2 & -1/2 & 1 \end{pmatrix}$$

$$n_i = 4$$

$$\begin{pmatrix} 1 & -1/3 & -1/3 & -1/3 \\ -1/3 & 1 & -1/3 & -1/3 \\ -1/3 & -1/3 & 1 & -1/3 \\ -1/3 & -1/3 & -1/3 & 1 \end{pmatrix}$$

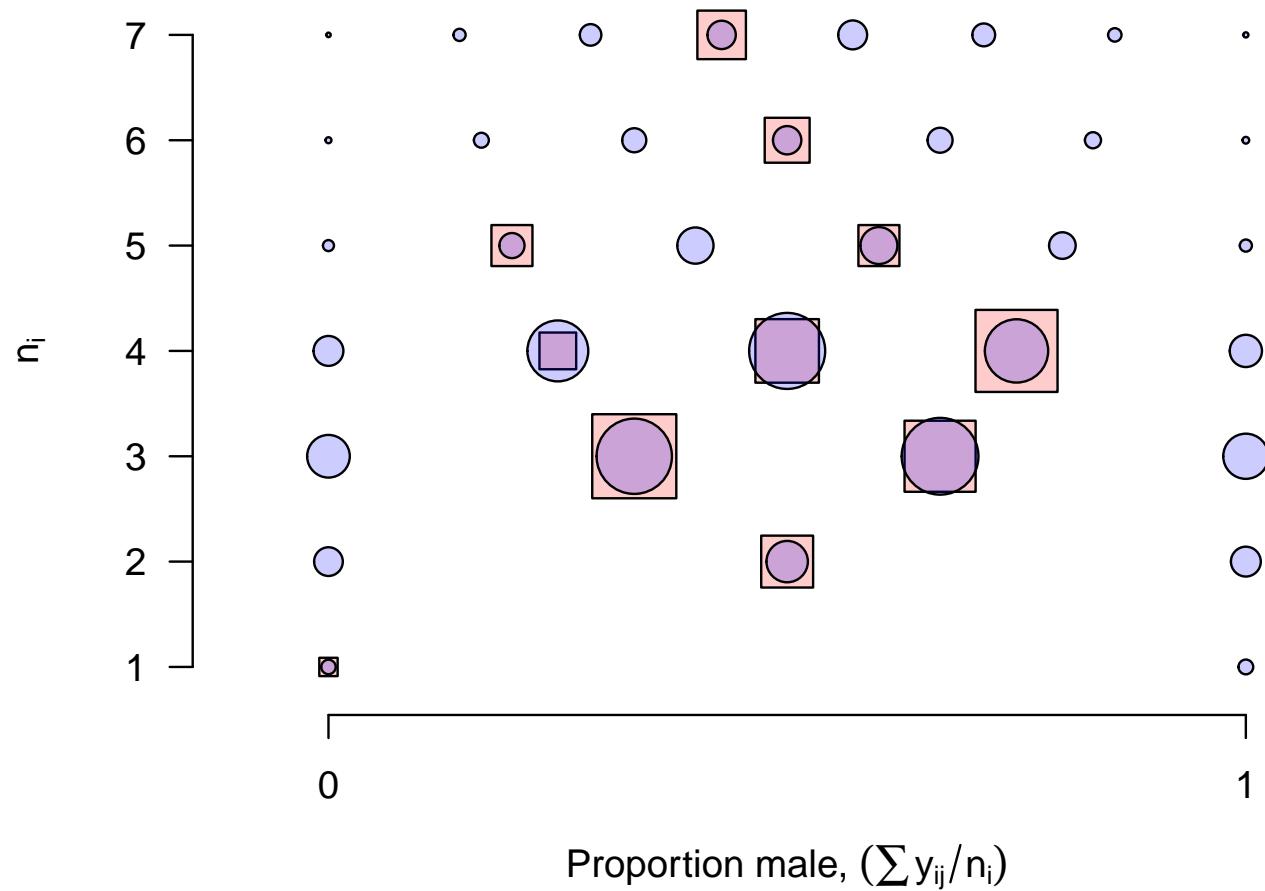
$$n_i = 5$$

$$\begin{pmatrix} 1 & -1/4 & -1/4 & -1/4 & -1/4 \\ -1/4 & 1 & -1/4 & -1/4 & -1/4 \\ -1/4 & -1/4 & 1 & -1/4 & -1/4 \\ -1/4 & -1/4 & -1/4 & 1 & -1/4 \\ -1/4 & -1/4 & -1/4 & -1/4 & 1 \end{pmatrix}$$

So data from small clusters may suggest strong negative correlations – but large clusters **cannot**. (The example's largest is $n_i = 7, 1/6 = 0.1667$). But the working correlation assumption is identical across all clusters, large and small...

GEE: moment restrictions

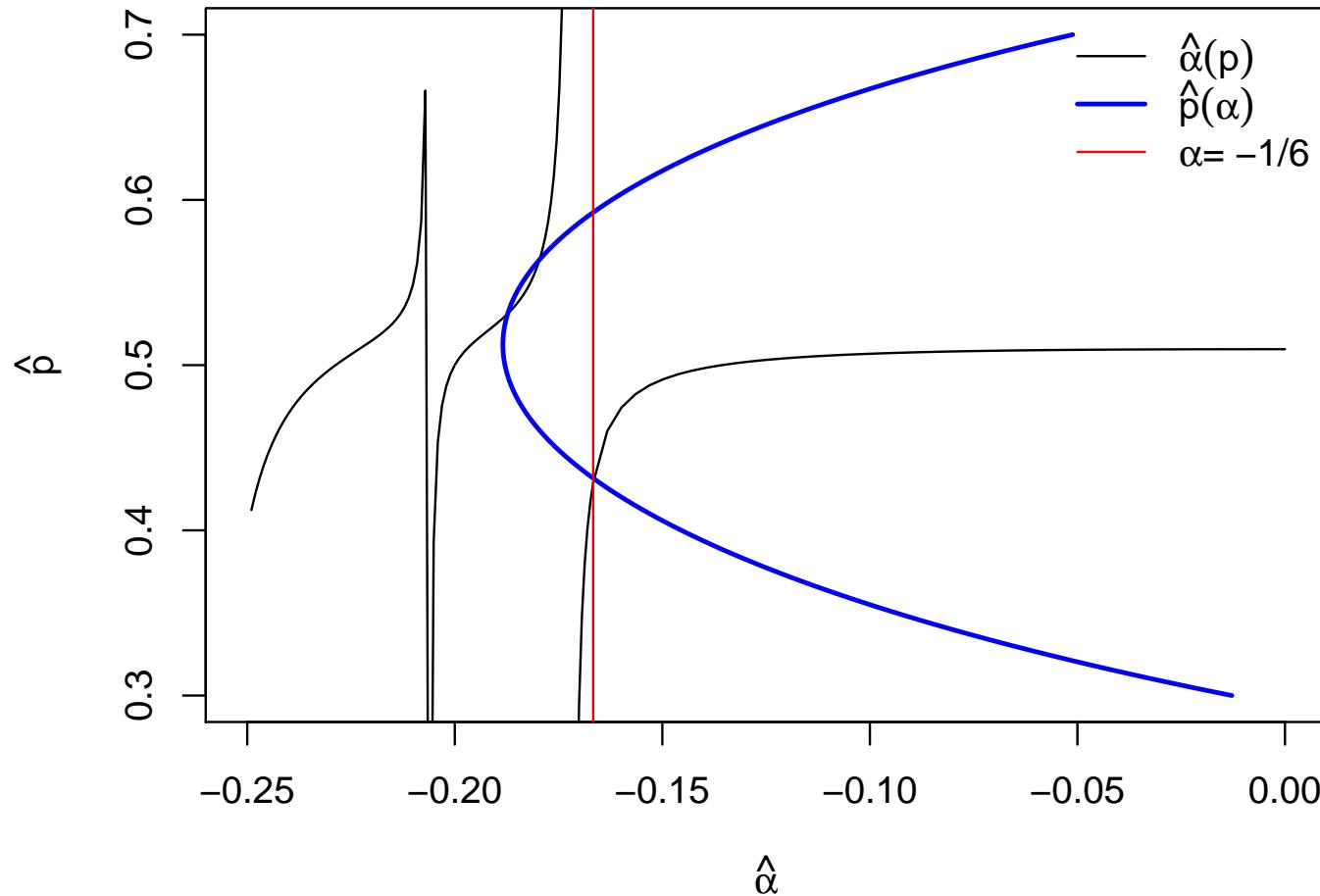
Another way to look at the data (squares) and the spread expected, if correlations within household were zero; (circles)



Q. Is the correlation positive or negative? Why might this be?

GEE: moment restrictions

Here's what happens when we update $\hat{\alpha}$ and $\hat{p} = \frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}}$



The EEs do have roots, but not stable ones. Ideally there would be only one stable root, at least near sensible initial $\hat{\beta}^{(0)}$.

GEE: moment restrictions

If we instead use independence working correlation;

```
> gee2 <- gee( y ~ 1, data=dat, id=id, family=binomial, corstr="independence",
+ scale.fix=TRUE)

> # confidence interval on the "beta", i.e. log-odds scale;
> coef(gee2) + qnorm(c(0.5,0.025, 0.975))*coef(summary(gee2))[4]
[1] 0.03846628 -0.22233443 0.29926699

> # confidence interval on the "p", i.e. probability scale;
> expit <- function(x){exp(x)/(1+exp(x))}
> expit(coef(gee2) + qnorm(c(0.5,0.025, 0.975))*coef(summary(gee2))[4])
[1] 0.5096154 0.4446442 0.5742633
```

Hence, we can get confidence intervals around the simple unweighted estimates of the log odds of being male, or the probability of being male – though note this is letting software concerns choose the precise analysis.

GEE: moment restrictions

Notes from this example;

- **Watch out for warnings!** Any suggestion of lack of convergence should be investigated – e.g. note the flaky original approach gives a ‘confidence interval’ $\sim 1/2$ the width of the working independence version. Ignoring this **is dangerous**
- Warnings are more likely when negative correlations present
- Nothing special about binary data (here)
- The independence working correlation assumption works, but may not be an efficient estimate – these clusters will tend to contain both men and women
- As well as not being fit correctly, it doesn’t look like exchangeable working correlation *with common α at all cluster sizes* is an efficient choice – it’s reasonable, here, to think ρ is stronger in smaller clusters (**Q. Why?**)

GEE: more moment restrictions

The constraint that \mathbf{R} must be positive-definite forces us to live inside the teabag. But particularly for binary Y_{ij} , with no further assumptions, we can say much more about the patterns of covariance and correlation that can occur, i.e. that are *feasible*.

For example, exploiting the fact that $\mu_{ij} \in [0, 1]$ and that $\text{Var}[Y_{ij}] = \mu_{ij}(1 - \mu_{ij})$;

$$\mathbb{E}[Y_{ij}] = \mu_{ij}$$

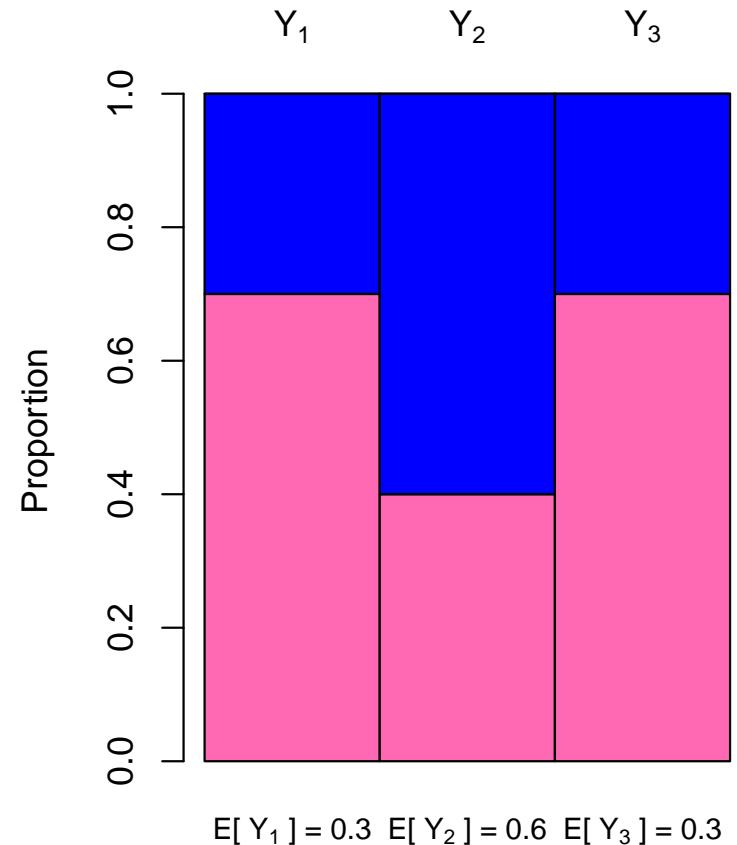
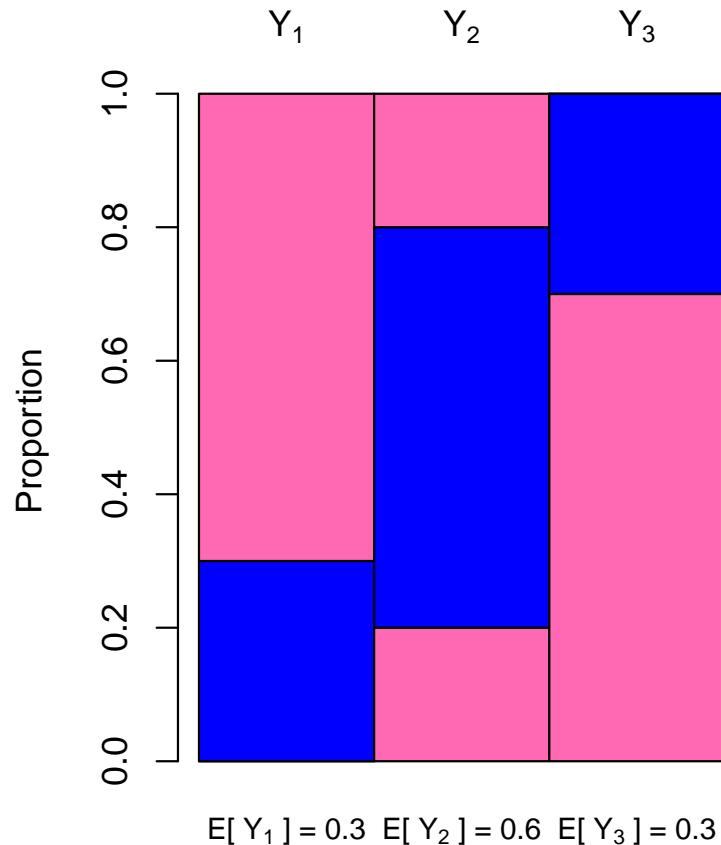
$$\text{and } \mathbb{E}[Y_{ij'}] = \mu_{ij'}$$

$$\Rightarrow \mathbb{E}[Y_{ij}Y_{ij'}] \leq \min\{\mu_{ij}, \mu_{ij'}\}$$

$$\Rightarrow \rho_{ijj'} \leq \min \left\{ \left(\frac{\mu_{ij}/(1 - \mu_{ij})}{\mu_{ij'}/(1 - \mu_{ij'})} \right)^{1/2}, \left(\frac{\mu_{ij'}/(1 - \mu_{ij'})}{\mu_{ij}/(1 - \mu_{ij})} \right)^{1/2} \right\}$$

GEE: more moment restrictions

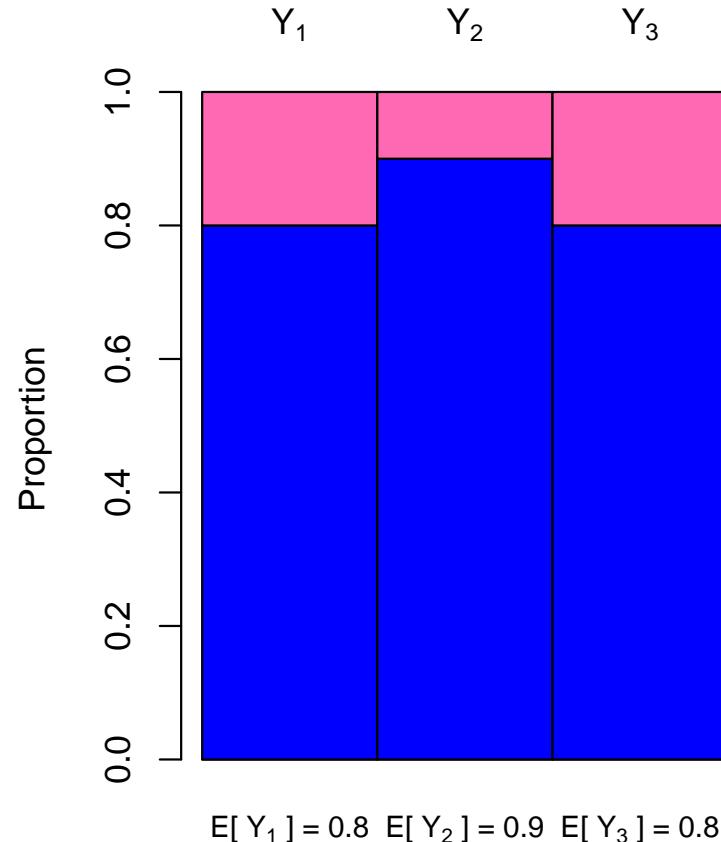
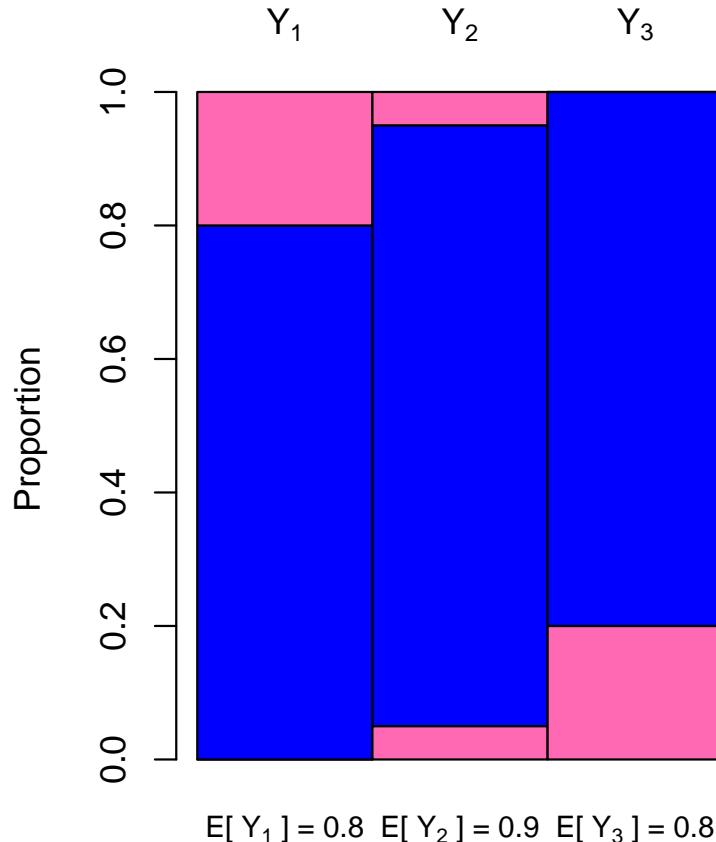
Perhaps somewhat more intuitively, for trios ($n_i = 3$)



$$\rho_{12} = -0.36, \rho_{13} = -0.42, \rho_{23} = -0.35 \quad \rho_{12} = 0.53, \rho_{13} = 1, \rho_{23} = 0.53$$

GEE: more moment restrictions

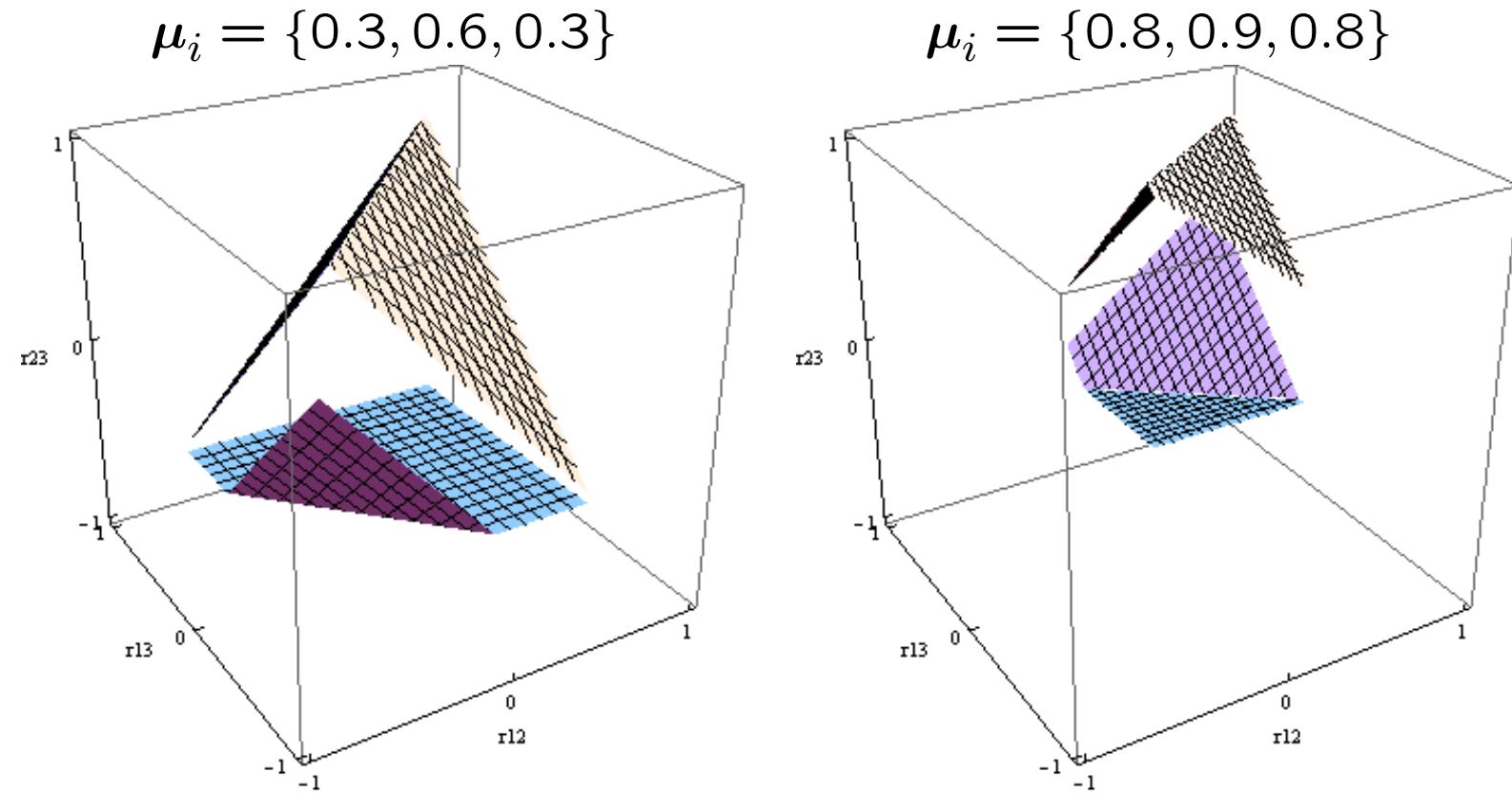
Same thing with more extreme μ_{ij} – and less flexibility;



$$\rho_{12} = 0.25, \rho_{13} = -0.25, \rho_{23} = 0.25 \quad \rho_{12} = 0.67, \rho_{13} = 1, \rho_{23} = 0.67$$

GEE: more moment restrictions

The *Fréchet bounds* determine feasible $\text{Corr}[\mathbf{Y}_i]$. For binary \mathbf{Y}_i , these depend on $\mu_i \dots$ and are ugly;



Q. What one point can you guarantee is always feasible?

GEE: more moment restrictions

Consequences of the teabag problem (which is widely ignored);

- When estimating α , asymptotically, you end up with $R_i(\hat{\alpha})$ as close to the true correlation as your working correlation structure will allow
- ‘Nice’ consistency, asymptotic Normality of $\hat{\beta}$ follows...
- ... unless $\hat{\alpha}$ does not converge or is boundary-valued. This is rare in practice – and also easy to spot if e.g. $\alpha \in [-1, 1]$
- The class site also has several papers by Chaganty and co-authors – who are experts in checking feasibility
- For some F , GEE with binary \mathbf{Y} is *never* fully efficient; but see Song Ch 5 and DHLZ Ch 8 for extensions to GEE that regain efficiency, also Lipsitz & Fitzmaurice (on class site)
- Fully-parametric models don’t face this difficulty; their corresponding correlation matrices *must* be feasible – because even if it’s not the truth, $F(\hat{\theta})$ is a model under which data has the fitted \mathbf{R}

GEE: random covariates

In considering GEE, we have assumed that parameters of interest can be described by

$$\mathbb{E}[Y_{ij} | \mathbf{X}_{ij} = x] = g(x^T \boldsymbol{\beta})$$

where;

- Frequentist replications of interest have the same design
- The mean for each observation is conditioned on its covariates *and nothing else*, i.e. $\boldsymbol{\beta}$ tells us about contrasts in the mean outcomes, averaged across all observations in all clusters

Both appear sub-optimal when we have time-varying covariates, particularly when interim values of Y_{ij} affect $X_{i,j+1}$.

GEE: random covariates

Solving the GEE estimating equations with random \mathbf{X}_i , for consistency we would need

$$\mathbb{E} \left[\sum_{j=1}^{n_i} \left(\frac{\partial g(\mathbf{X}_i \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \mathbf{V}_i^{-1} \right)_{kj} \mathbb{E}[Y_{ij} - g(\mathbf{x}_{ij}^T \boldsymbol{\beta}) | \mathbf{X}_i = \mathbf{x}_i] \right] = 0$$

where the outer expectation is only over \mathbf{X} , and k indexes the individual components of the p estimating equations.

However, there may be a difference between the regression parameters conditioning on observation-specific \mathbf{X}_{ij} may not be the same as those conditioning on cluster-specific \mathbf{X}_i . Do we have

$$\mathbb{E}[Y_{ij} | \mathbf{X}_{ij} = \mathbf{x}] = \mathbb{E}[Y_{ij} | \mathbf{X}_i = \mathbf{x}]?$$

...where, implicitly, the j th row of $\mathbf{x} = \mathbf{x}$.

GEE: random covariates

An example; within each cluster i , generate $\{\mathbf{Y}, \mathbf{X}_i\}$ as follows;

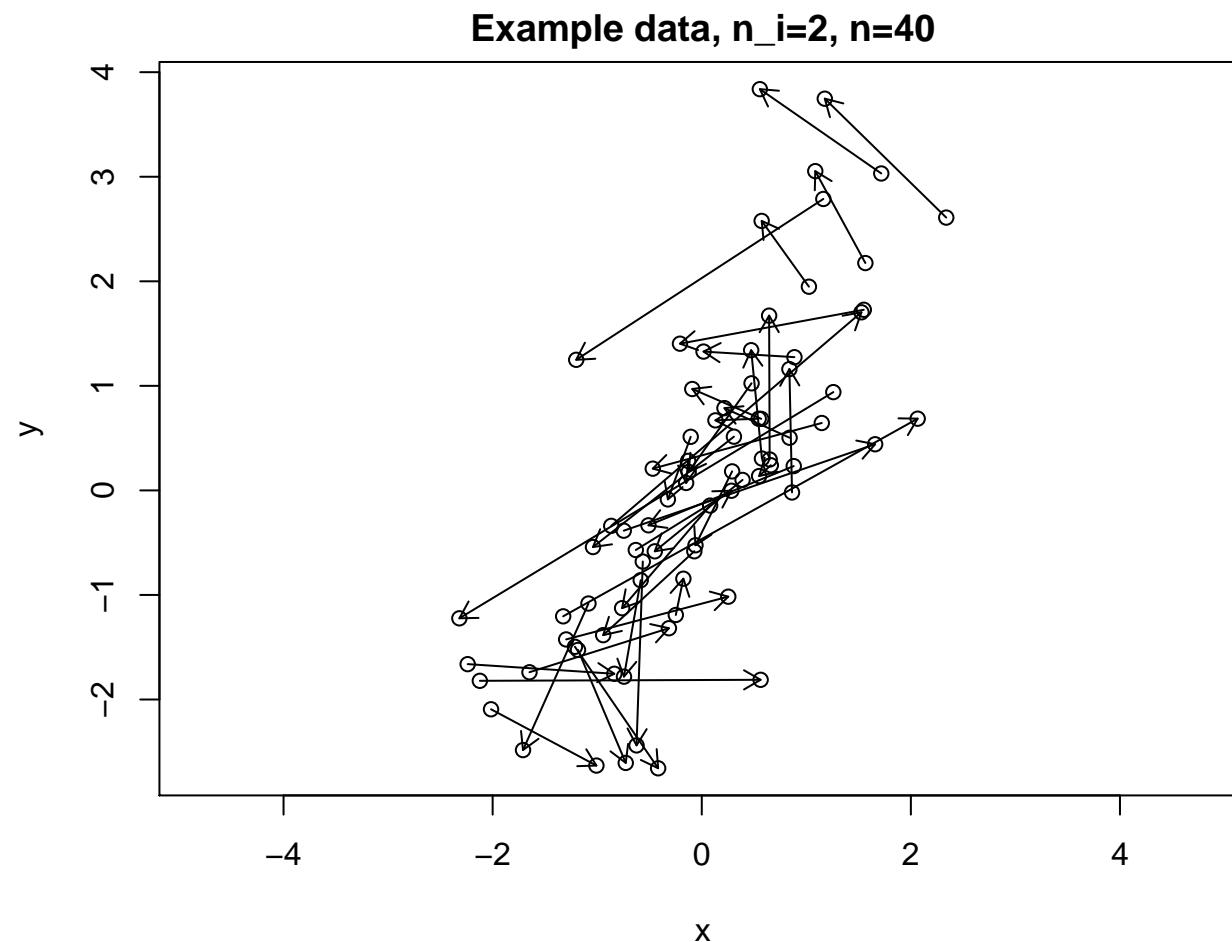
$$\begin{aligned} X_{ij} &\stackrel{i.i.d.}{\sim} N(0, \sigma^2), \\ Y_{i1} &\stackrel{indep}{\sim} N(X_{i1}\beta, \tau^2), \\ Y_{ij}|Y_{i,j-1}, X_{ij} &\stackrel{indep}{\sim} N(Y_{i,j-1} + X_{ij}\beta, \tau^2) \end{aligned}$$

- The plausible parameter of interest is β
- Outcome today depends directly on outcome yesterday, as well as today's covariate – which makes sense, in e.g. pharmacokinetics
- A difficult question; how does γ relate to the various β , here?

$$\begin{aligned} \mathbb{E}[Y_{ij}|X_{i1}, X_{i2}, \dots X_{in_i}] &= \sum_{k=1}^j X_{ik}\beta \\ \mathbb{E}[Y_{ij}|X_{ij}] &= X_{ij}\gamma \end{aligned}$$

GEE: random covariates

Some example data, for $\beta = 1$. Arrows join $j = 1 \rightarrow j = 2$



GEE: random covariates

With $n_i = 2$ and $\beta = 1$, it's immediate that

$$\begin{aligned}\mathbb{E}[Y_{i1}|X_{i1} = x] &= x \\ \mathbb{E}[Y_{i1}|\mathbf{X}_i = \{x, x'\}] &= x\end{aligned}$$

With more work, we find that;

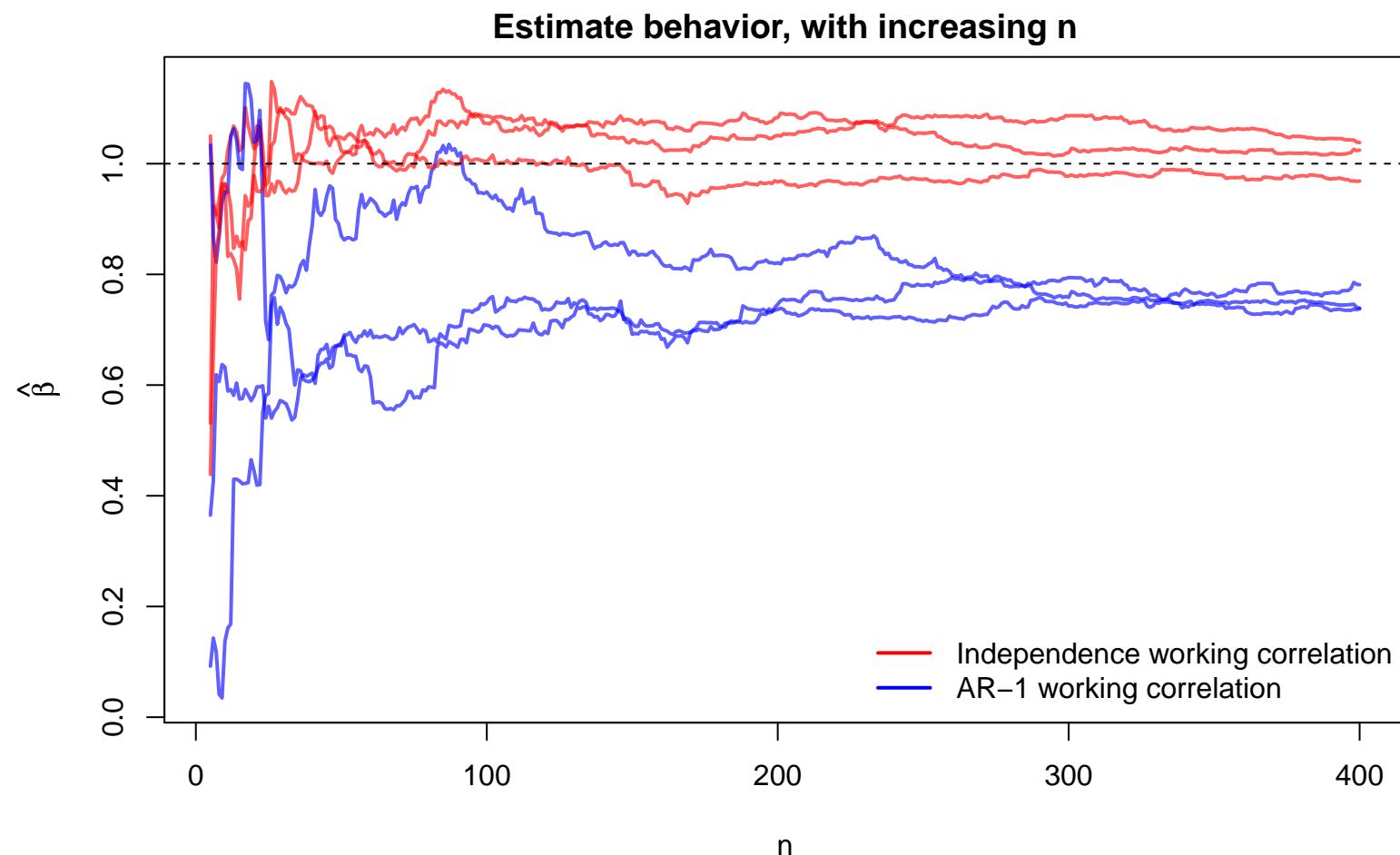
$$\begin{aligned}\mathbb{E}[Y_{i2}|X_{i2} = x'] &= \mathbb{E}\left[\mathbb{E}[Y_{i2}|X_{i1} = x, Y_{i1} = y, X_{i2} = x']|X_{i2} = x'\right] \\ &= \mathbb{E}\left[Y_{i1} + X_{i2}|X_{i2} = x'\right] \\ &= \mathbb{E}[Y_{i1}|X_{i2} = x'] + x' \\ &= \mathbb{E}[Y_{i1}] + x' \\ &= x'.\end{aligned}$$

In other words, $\mathbb{E}[Y_{ij}|X_{ij} = x] = x$, for $j = 1, 2$. However...

$$\begin{aligned}\mathbb{E}[Y_{i2}|\mathbf{X}_i = \{x, x'\}] &= \mathbb{E}\left[\mathbb{E}[Y_{i2}|X_{i1} = x, Y_{i1} = y, X_{i2} = x']|X_{i1} = x, X_{i2} = x'\right] \\ &= \mathbb{E}\left[Y_{i1} + X_{i2}|X_{i1} = x, X_{i2} = x'\right] \\ &= \mathbb{E}[Y_{i1}|X_{i1} = x, X_{i2} = x'] + x' \\ &= x + x' \neq \mathbb{E}[Y_{i2}|X_{i2} = x']\end{aligned}$$

GEE: random covariates

What happens using `gee(y~x, id=cluster)` ?



GEE: random covariates

- Using Independence working correlation we get it right, i.e. we consistently estimate $\beta = 1$.
- The AR-1 approach is inconsistent for β – and so is any other GEE approach that uses $\alpha \neq 0$

Unlike when we considered \mathbf{X} fixed, the two approaches are not consistent with each other. Choice of working correlation matrix *really does* matter, here; we cannot dismiss it as just a matter of efficiency.

This problem was often overlooked in early uses of GEE. Margaret Pepe and Garnet Anderson (1994) – on the class site – brought the matter to somewhat wider attention, see also Pan et al (2001).



GEE: random covariates

Pepe and Anderson (1994) showed a remarkably strong result;

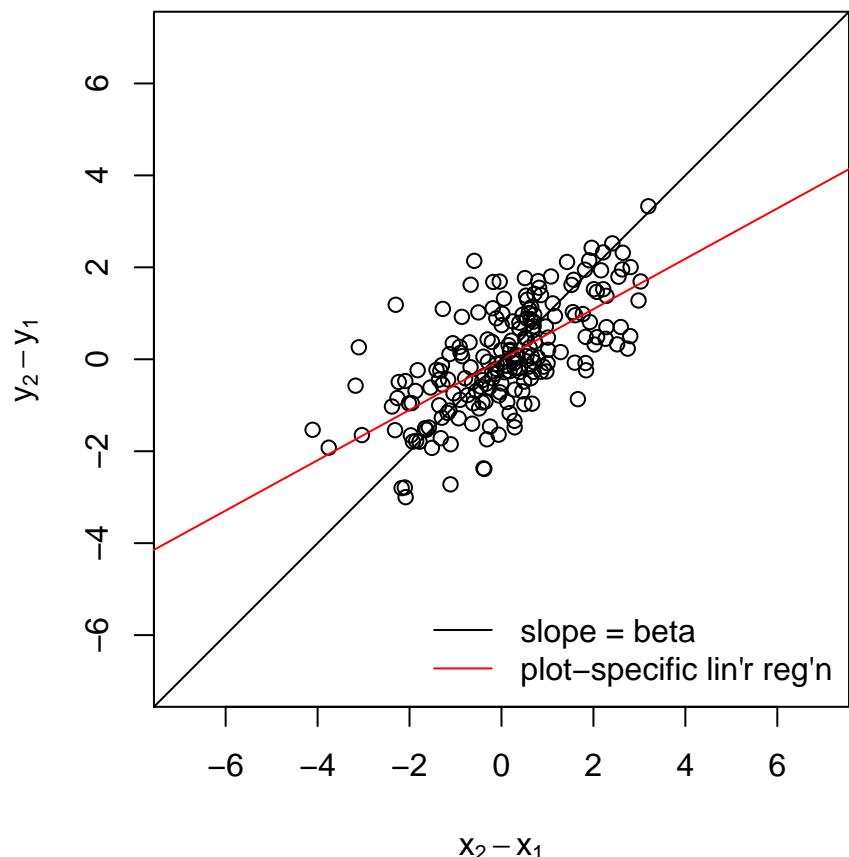
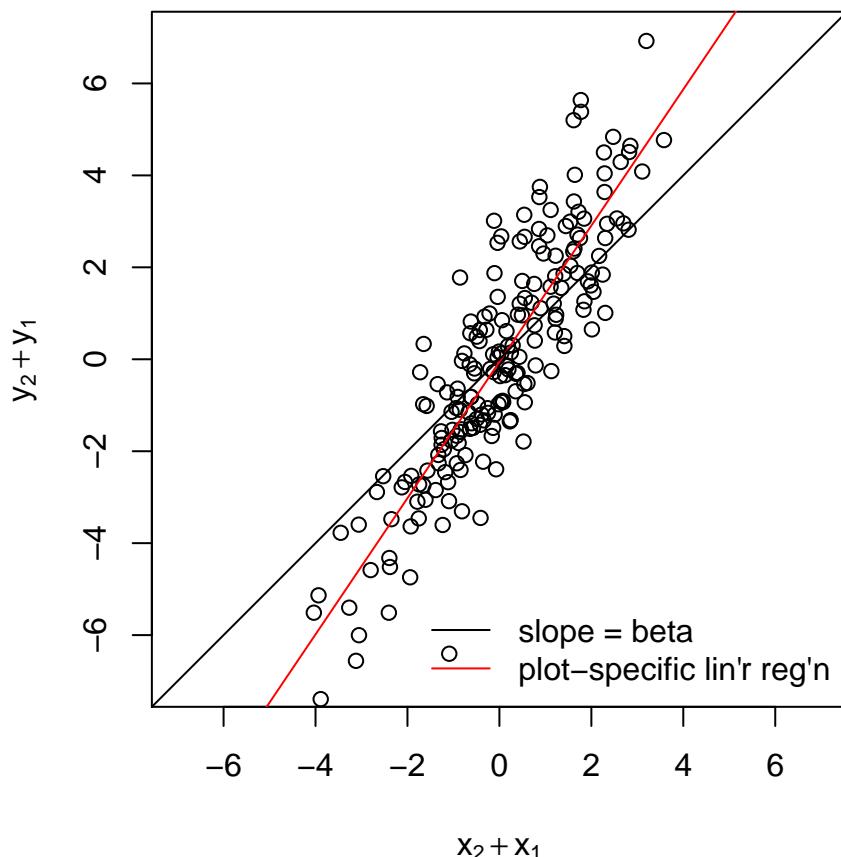
When $\mathbb{E}[Y_{ij}|\mathbf{X}_{ij} = \mathbf{x}] = g(\mathbf{x}^T \boldsymbol{\beta})$ but
 $\mathbb{E}[Y_{ij}|\mathbf{X}_{ij} = \mathbf{x}] \neq \mathbb{E}[Y_{ij}|\mathbf{X}_i = \mathbf{x}]$,
then $\hat{\boldsymbol{\beta}}$ from GEE is *not* consistent
unless independence working assumptions are used.

- This matches what we saw in the example (in which $\text{Corr}[Y_{i1}, Y_{i2}] \neq 0$, but working independence gets $\hat{\boldsymbol{\beta}}$ ‘right’)
- Using working independence may be badly inefficient; the easy ‘fix’ may not be a good one
- Checking that $\mathbb{E}[Y_{ij}|\mathbf{X}_i]$ is just $\mathbb{E}[Y_{ij}|\mathbf{X}_{ij}]$ is another ‘fix’, but the data may provide little capacity to identify discrepancies

If \mathbf{X}_{ij} is identical for all j , or \mathbf{X}_i is fixed by design, or the distribution of \mathbf{X}_i is not informative about $\boldsymbol{\beta}$, inference with fixed \mathbf{X} is appropriate, and free of these difficulties.

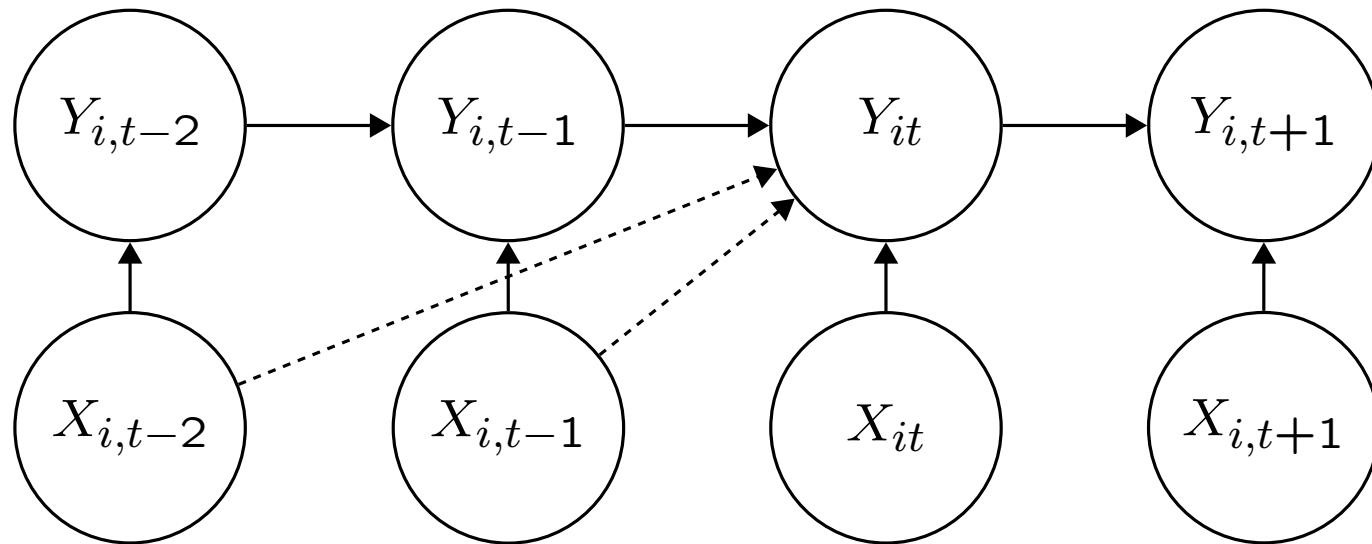
GEE: random covariates

To check $\mathbb{E}[Y_{ij}|\mathbf{X}_i]$, plotting \mathbf{Y} against \mathbf{X} is not feasible; plotting linear combinations of their elements *may* help;



GEE: random covariates

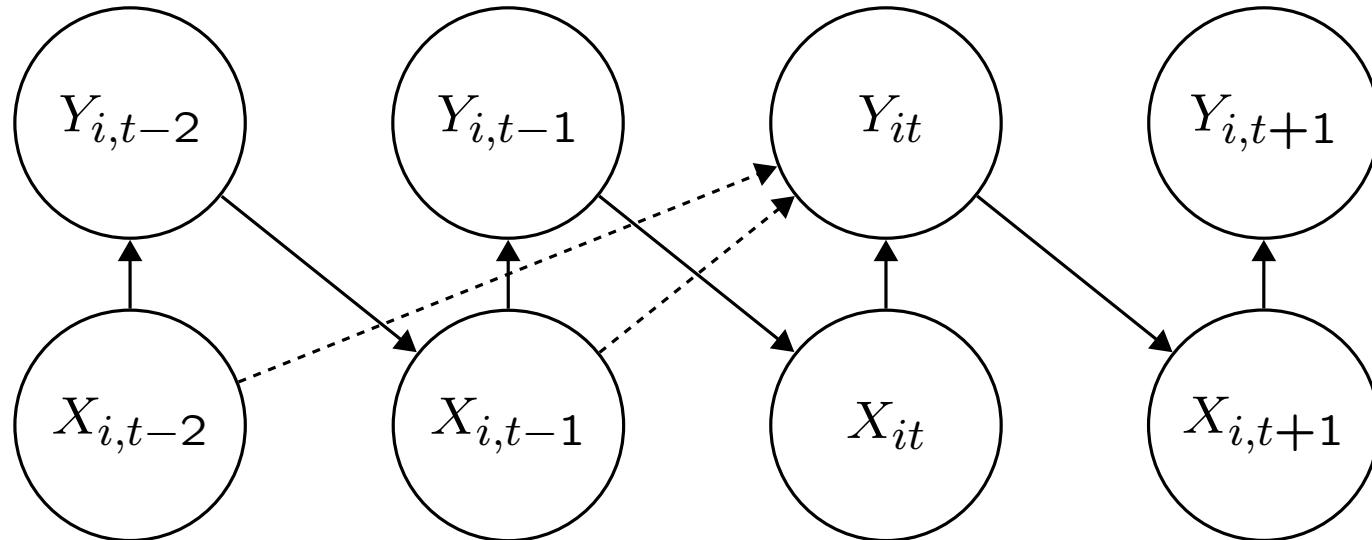
The Pepe and Anderson problem as viewed by causal analysts;



- Previous X affect both previous Y and current Y
- Treating all time points as independent, no bias occurs (as we saw) but analysis may not be efficient
- Using GEE, we could adjust for previous variables; this requires careful coding, and possibly many nuisance parameters

GEE: random covariates

A problem that even clever GEE will not address;



- e.g. previous outcomes affects future dosage
- Adjusting the X variables works; this is done by G -estimation, in which each (solid) arrow gets its own EE contribution. See Robins et al, on the class site.
- Sophisticated causal inference is a current research area

GEE (and more): missing data

Missing data is standard in real-world **clustered and unclustered** settings. Data gets lost, measurements are not taken, assays fail because lab techs over-sleep, patients don't show for appointments, students refuse to answer certain questions, etc.

```
> head(airquality, 10) # Ozone, sun, wind & temp in NYC, 1973
```

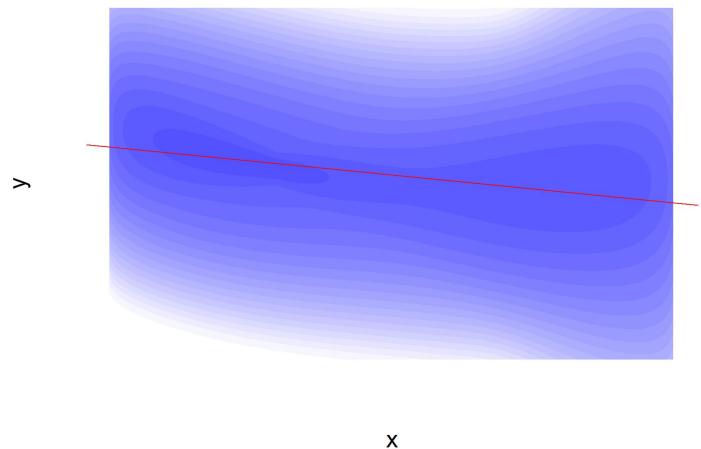
	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6
7	23	299	8.6	65	5	7
8	19	99	13.8	59	5	8
9	8	19	20.1	61	5	9
10	NA	194	8.6	69	5	10

- In R, read NA as 'I don't know'. What is `mean(airquality$Ozone)`? TRUE & NA? FALSE & NA? Also note `is.na()`, not e.g. `42==NA`
- Regression functions default to 'complete case' analysis **with no warning** – so always check n in the output

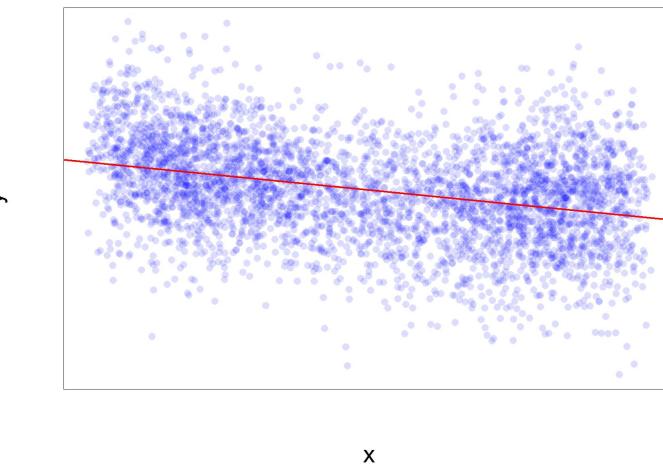
GEE (and more): missing data

Missingness **may or may not** invalidate inference;

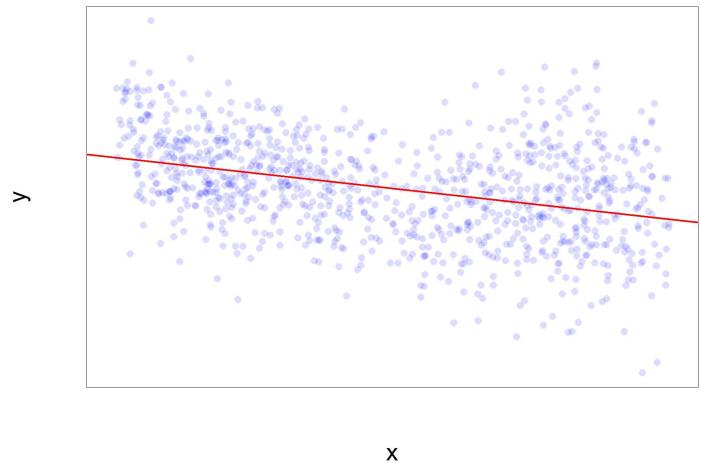
Full population ($n = \infty$)



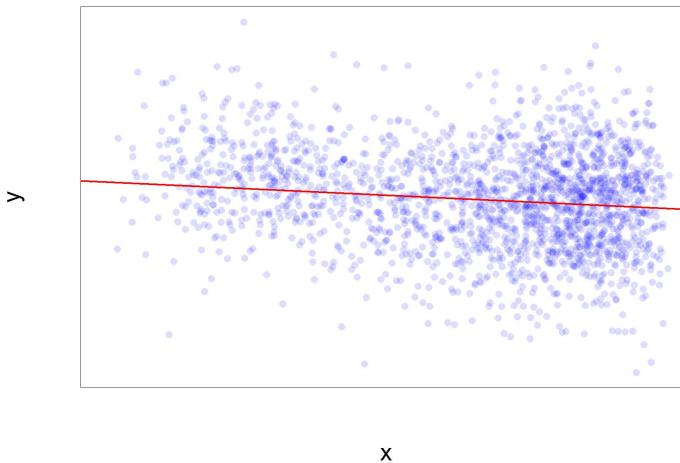
Simple sample ($n = 4000$)



Missing $\perp X, Y$ ($n = 2000$)



Missing $\propto X$ ($n = 2000$)



Missing Data: what can we do?

*“The best way to deal with missing data
is to not have any.”*

Tom Fleming, UW Biostat



- Great idea... **if** this solution or something very close to it is actually possible. (But minimizing NA's does help!)
- If you can ensure a **really** well-run, ‘clean’ clinical trial or lab experiment, you *might* possibly be able to get close to this ideal
- If not, zero tolerance of missing data is not a helpful strategy. Inference that acknowledges limitations due to missing data (e.g. modest inconsistency) is almost always the best you can do – and this can still be useful

Just as we need contextual knowledge to choose a regression model, we need* to know something about **why** data is missing when deciding what (if anything) to do about it.

* The untestability of relevant assumptions can actually be proven

Missing Data: what can we do?

- **Complete-case analyses**
 - Easy to implement; may be acceptable with small amounts of missing data
 - Otherwise may lead to serious bias and loss of efficiency
- **Imputation procedures** to ‘fill-in’ any missing data
 - Includes Hot-deck, mean, and multiple imputation
 - Allows use of standard(ish) methods on ‘filled-in’ data
- **Weighting procedures** to adjust for non-response as if part of design. (Developed from sample-survey methods for non-response weighting)
- **Model-based procedures** where missingness is modeled
 - Selection, pattern-mixture, and random-effects models
 - Can evaluate assumptions underlying the fitted models
- **Others** that should rarely, if ever, be used
 - Last observation carried forward
 - ‘Worst-case’ imputation

GEE: complete-case analyses

As slide 2.130 indicates, complete-case analysis is valid under some patterns of missingness – or patterns of observation, if you prefer. To formalize this, here we consider only missing outcomes*, and introduce some new notation;

- | | |
|----------------|--|
| $M_{ij} = 1/0$ | Indicator of Y_{ij} being missing/observed |
| \mathbf{M}_i | n_i -vector of missingness indicators |
| \mathbf{Y}^o | The set of observed outcomes |
| \mathbf{Y}^m | the set of missing outcomes |

So, the full set of outcomes \mathbf{Y} can be partitioned into $\{\mathbf{Y}^o, \mathbf{Y}^m\}$.
Also recall the definition of β and its full-data estimate;

$$\begin{aligned}\mathbb{E}[\mathbf{Y}_i - g(\mathbf{X}_i\beta)] &= \mathbf{0}_{n_i} \\ \frac{1}{n} \sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) &= \mathbf{0}_p\end{aligned}$$

... where for simplicity we are ignoring estimation of α, ϕ .

* ... assuming a mean model and that conditioning on \mathbf{X}_{ij} is appropriate, i.e.
the usual GEE assumptions

GEE: complete-case analyses

Complete-cases analysis solves these EEs;

$$\frac{1}{n} \sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} \text{diag} \left\{ 1 - M_{ij} \right\} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}_p,$$

and hence is consistent for the large-sample limit solution of

$$\frac{1}{n} \sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} \text{diag} \left\{ \mathbb{P}[M_{ij} = 0 | Y_{ij}, \mathbf{X}_{ij}] \right\} \mathbb{E}[Y_{ij} - \mu_{ij} | \mathbf{X}_{ij}] = \mathbf{0}_p.$$

If missingness M_{ij} is independent of outcome, i.e.

$$\mathbb{P}[M_{ij} | Y_{ij}, \mathbf{X}_{ij} = \mathbf{x}_{ij}] = \mathbb{P}[M_{ij} | \mathbf{X}_{ij} = \mathbf{x}_{ij}]$$

– also known as *Missing Completely At Random* (MCAR) – then the EEs are consistent for the limiting solution of

$$\frac{1}{n} \sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} \text{diag} \left\{ \mathbb{P}[M_{ij} = 0 | \mathbf{X}_{ij} = \mathbf{x}_{ij}] \right\} \mathbb{E}[\mathbf{Y}_i - \boldsymbol{\mu}_i | \mathbf{X}_{ij} = \mathbf{x}_{ij}] = \mathbf{0}_p.$$

The $\mathbb{E}[\mathbf{Y}_i - \boldsymbol{\mu}_i]$ term means complete-case GEE gives consistent estimates of the β parameters and valid standard error estimates, in large samples, regardless of working \mathbf{V} .

GEE: complete-case analyses

Other notes on MCAR:

- Missingness is said to be *ignorable* if complete-case analysis is valid (but this overlooks issues of efficiency – in particular, efficiency may be reduced compared to full-data GEE)
- MCAR is often written as a statement that

$$f(\mathbf{M}|\mathbf{Y}, \mathbf{X} = \mathbf{x}) = f(\mathbf{M}|\mathbf{X} = \mathbf{x})$$

- Some examples of MCAR;
 - Assay results lost in the mail
 - Patient i skipped doctor visit j due to e.g. weather
 - Administrative censoring (i.e. the study ends before final observation) ... **assuming** $\mathbf{Y}|\mathbf{X}$ doesn't differ systematically between early/late entrants
- ‘Completely’ can be a misleading term – may have MCAR where \mathbf{M} depends on variables unrelated to the analysis
- Careful choice of working \mathbf{V} may enhance efficiency – e.g. upweighting rarely-observed observations – but this would require specifying $\mathbb{P}[M_{ij}|\mathbf{X}_{ij} = \mathbf{x}_{ij}]$

GEE: is MCAR required?

Having MCAR is clearly helpful, but skepticism about MCAR reflecting the truth is widespread. One recent article (Aloisio et al 2014) notes that “*the assumption of MCAR is very restrictive in a world where reasons for missingness are generally more complex than just being due to chance*” and that MCAR is “*often implausible*”.

Rhetorical Q. So why should you (or anyone) care about MCAR?

A. MCAR is the **most general condition** under which complete-case GEE* gives valid inference.

In other words, if we’re not getting ‘close’ to MCAR, then we should not rely on complete-case GEE to have robustness to working \mathbf{V} .

* Note: MCAR also make complete-case analysis valid using most other regression tools – for example sandwich estimates for independent outcomes

GEE: is MCAR required?

We illustrate MCAR mattering with a simulation-based example;

$$\begin{aligned} Y_{it}|G_i = g_i &\sim N(\beta_0 + b_{0i} + \beta_1 t + b_{1i}t + \beta_2 g_i + \beta_3 g_i t, \sigma_Y^2) \\ \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} &\sim N\left(\mathbf{0}, \begin{pmatrix} \sigma_{b0}^2 & \rho\sigma_{b0}\sigma_{b1} \\ \rho\sigma_{b0}\sigma_{b1} & \sigma_{b1}^2 \end{pmatrix}\right) \\ \{\beta_0, \beta_1, \beta_2, \beta_3\} &= \{25, -1, 0, -1\} \\ \sigma_Y &= 2 \\ \{\sigma_{b0}, \sigma_{b1}, \rho\} &= \{2, 0.5, -0.1\} \end{aligned}$$

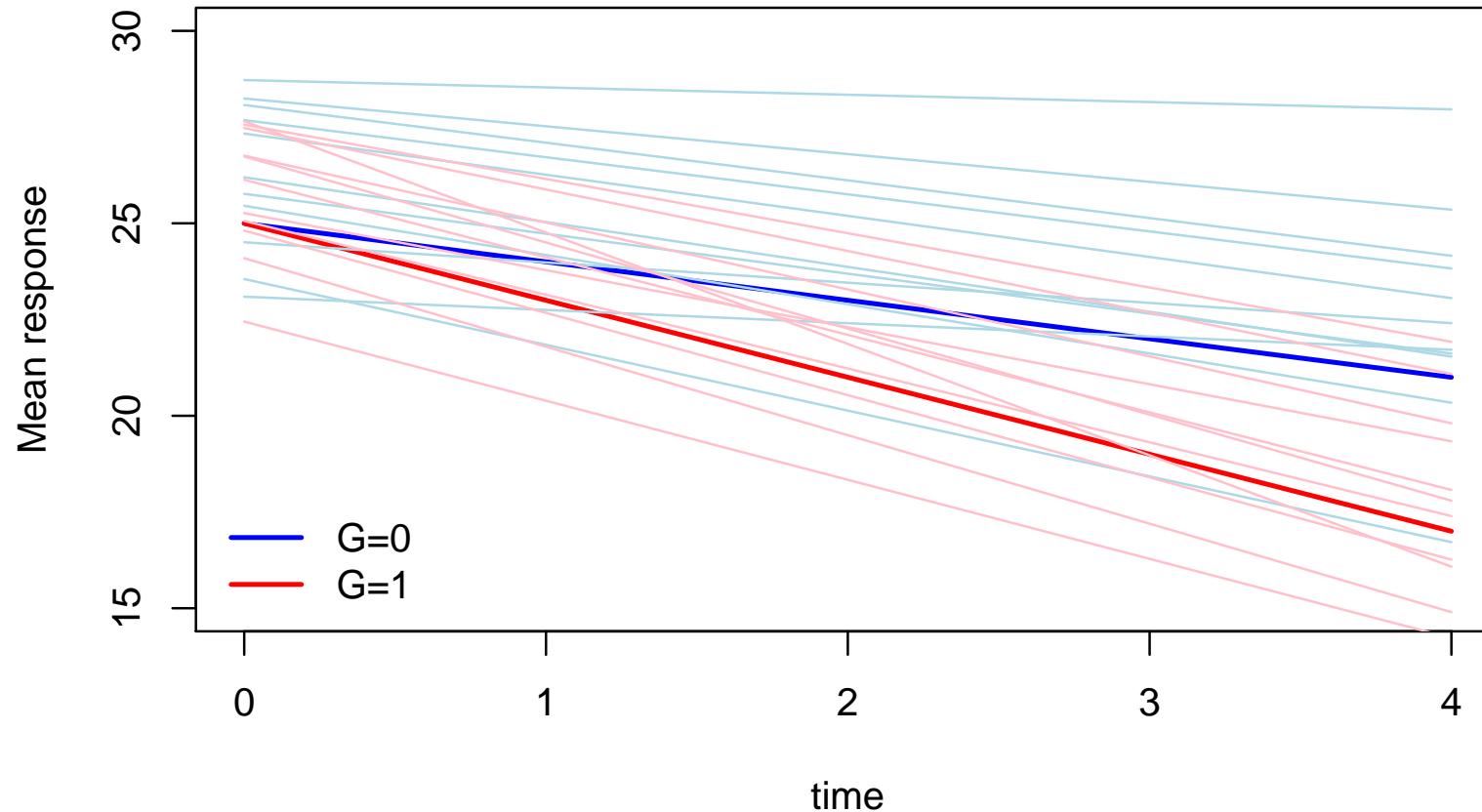
We use $n_i = 5$ equally-spaced observations, from $t = 0$ to $t = 4$; covariate G_i is 0/1, split 50:50. G has no main effect ($\beta_2 = 0$) but does interact with time ($\beta_3 = -1$).

Across clusters, have large variability in random intercepts ($\sigma_{b0} = 2$, comparable to e.g. $\sigma_Y = 2$) and some random variability in slope ($\sigma_{b1} = 0.5$); these are weakly negatively correlated.

As we'll see in Chapter 3, the β_k here have both marginal and conditional interpretations – i.e. GEE linear regression with no missing data gives $\hat{\beta}$ consistent for $\{25, -1, 0, -1\}$.

GEE: is MCAR required?

In a picture;



Solid lines indicate marginal mean; pale lines are subject-specific means. Data (not shown) would be noisier still.

GEE: is MCAR required?

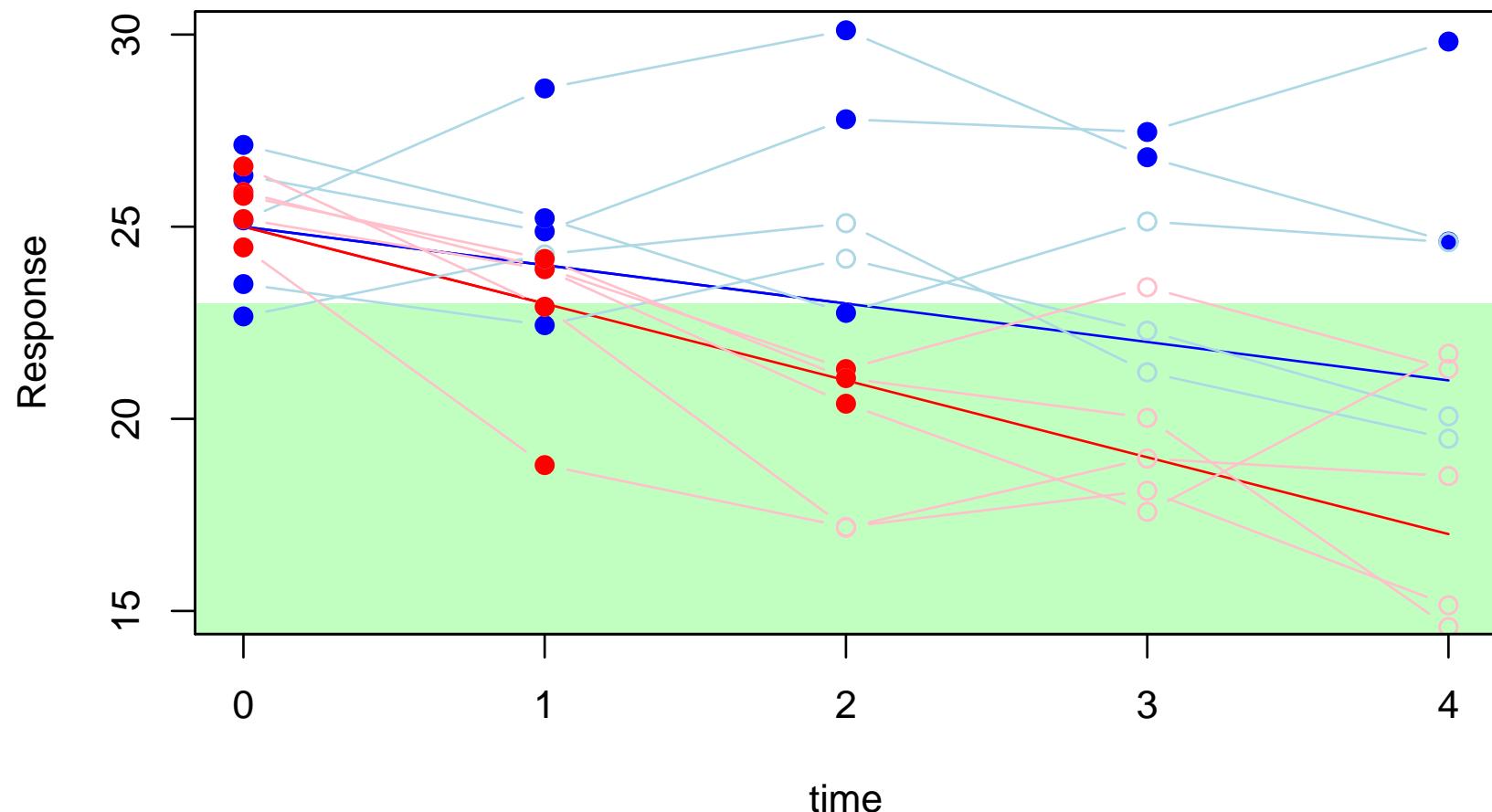
Now consider 3 patterns of missingness;

- **M1:** 50% entirely random missingness; each M_{ij} is an independent coin toss
- **M2:** At $t = 0, 1, 2, 3, 4$ probability of ‘dropping out’ is 0, 0.25, 0.5, 0.75, 0.85 respectively. This gives roughly 50% missing overall. Once subjects have ‘dropped out’ they do not return
- **M3:** If $Y_{it} < 23$ for any observation, omit all **subsequent** observations for that subject – perhaps reflecting withdrawal due to adverse events (again, roughly 50% missingness overall)

Q. Are these MCAR? Is $f(\mathbf{M}|\mathbf{Y}, \mathbf{X}) = f(\mathbf{M}|\mathbf{X})$?

GEE: is MCAR required?

A picture of M3; note M2 wouldn't look dramatically different;



GEE: is MCAR required?

To illustrate consistency properties, we show results for each setting from a single large dataset ($n = 5,000$, independence working correlation). This is not rigorous – but it is a sane way to **get started** answering questions;

Term	Intercept		t		G		$G \times t$	
	$\hat{\beta}_0$	\widehat{SE}	$\hat{\beta}_1$	\widehat{SE}	$\hat{\beta}_2$	\widehat{SE}	$\hat{\beta}_3$	\widehat{SE}
Truth	25.0		-1.00		0.00		-1.00	
M1	25.1	0.16	-1.02	0.05	-0.03	0.22	-1.04	0.07
M2	25.0	0.11	-0.91	0.09	0.04	0.17	-1.19	0.13
M3	25.0	0.11	-0.39	0.06	-0.12	0.16	-0.79	0.10

- For M1 and M2, dropout only depends on covariates and so we get the truth back (\pm random noise)
- For non-ignorable dropout M3 the GEE point inference doesn't cover the truth (particularly for effects of time) even allowing for chance, or small-sample weirdness

Q. What would have happened if $\sigma_{b0} = \sigma_{b1} = 0$, i.e. no clustering?

GEE: diagnostics for MCAR

Assessing MCAR is fundamentally difficult; we want to check

$$f(\mathbf{M}|\mathbf{Y}, \mathbf{X}) = f(\mathbf{M}|\mathbf{X})$$

but haven't observed all of \mathbf{Y} . Two informal ideas;

- Plot M_{ij} against covariates, add a smoother, and look for trends. Knowing that \mathbf{M} depends on some \mathbf{X} is often useful; if Y is also (plausibly) related to M you can **suggest** reasons for missingness. See e.g. Ch4 of Millsap & Allison (2009)
- Using Bayes' Theorem and then assumed MCAR;

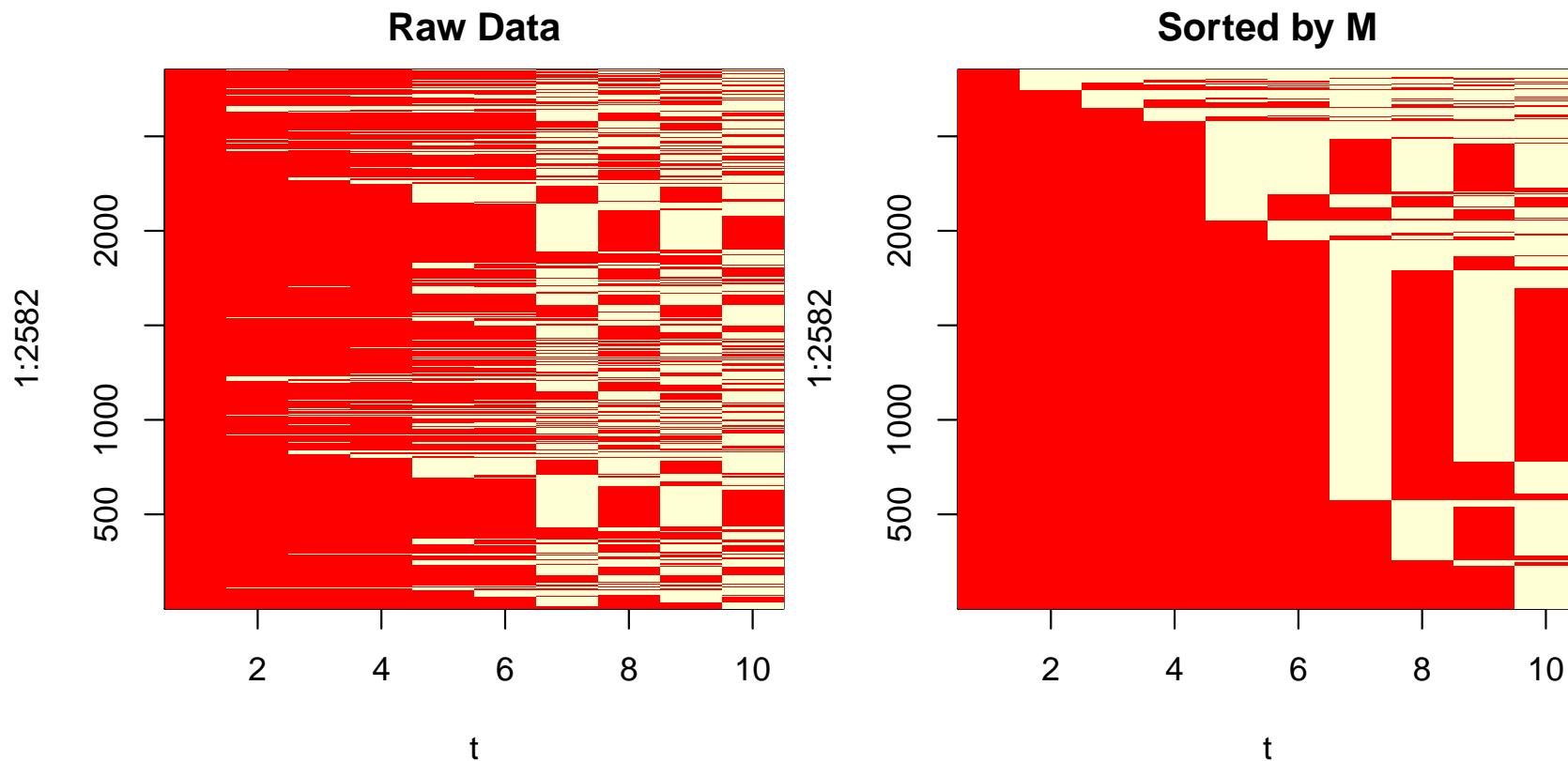
$$f(\mathbf{Y}|\mathbf{M}, \mathbf{x}) = \frac{f(\mathbf{Y}|\mathbf{x})}{f(\mathbf{M}|\mathbf{x})} f(\mathbf{M}|\mathbf{Y}, \mathbf{x}) = f(\mathbf{Y}|\mathbf{x}) \frac{f(\mathbf{M}|\mathbf{Y}, \mathbf{x})}{f(\mathbf{M}|\mathbf{x})} = f(\mathbf{Y}|\mathbf{x}),$$

i.e. the distribution (or mean) of $\mathbf{Y}|\mathbf{x}$ is identical under all patterns of missingness \mathbf{M} . Hence we might compare GEE results (i.e. $\hat{\beta}$) from subsets of clusters with specific patterns of \mathbf{M} . If they differ* it's reasonable to suggest MCAR is violated.

* and we assume everything else in the GEE analysis is okay!

GEE: diagnostics for MCAR

Missingness in cognitive impairment (i.e. dementia) at 10 time points – every 6 months for 5 years – for $n = 2,852$ elderly participants. (Biostat applied exam 2015... confidential data)



- Q. What's happening at the end of the study?
- Q. Why might $M|t$ depend on \mathbf{Y} here?

GEE: diagnostics for MCAR

Chen & Little (1999) turn the intuition about comparing GEE's $\hat{\beta}$ by patterns of M (given covariates, or stratified covariates) into a formal statistical test. Note: I know of no code for it.

- In large samples, and with a correct mean model they produce a valid test of the null hypothesis that (exact) MCAR holds
- It's a diagnostic, so expect it to have low power – unless the Tay Bridge really has collapsed
- The reference distribution is χ^2_{df} where df is \approx the number of different M_i seen. This is typically large (see previous slide) in which case the asymptotic approximate will be poor unless n is huge
- Comparing outcomes Y across large strata (defined by M, X) may be informative, even without a formal test

As usual, don't let a p -value (alone) tell you what to do – context matters, and using it works better than trying to 'test your way' to a decent analysis.

GEE: multiple imputation



Though no man of middling good sense can derive much pleasure from the imputation of a laudable action which he never performed, yet a wise man may suffer great pain from the serious imputation of a crime which he never committed
from The Theory of Moral Sentiments (1759)
by Adam Smith (1723 – 1790)

Scottish moral philosopher, pioneer of political economy

GEE: multiple imputation

So far we've considered when complete-case GEE is a laudable action, i.e. gives valid inference. But again note under MCAR,

$$f(\mathbf{Y}|\mathbf{M}, \mathbf{x}) = \frac{f(\mathbf{Y}|\mathbf{x})}{f(\mathbf{M}|\mathbf{x})} f(\mathbf{M}|\mathbf{Y}, \mathbf{x}) = f(\mathbf{Y}|\mathbf{x}) \frac{f(\mathbf{M}|\mathbf{Y}, \mathbf{x})}{f(\mathbf{M}|\mathbf{x})} = f(\mathbf{Y}|\mathbf{x}).$$

This tells us that what we know about $\mathbf{Y}|\mathbf{x}$ for the complete data, where $M = 0$ **also** tells us about $\mathbf{Y}|\mathbf{x}$ when \mathbf{Y} is missing, i.e. where $M = 1$.

This suggests the method of *multiple imputation* (MI);

0. Fit a model that predicts how \mathbf{Y} behave (in particular their conditional mean given \mathbf{X}) using the observed data, i.e. \mathbf{Y}^o
1. Randomly generate new \mathbf{Y}^{m*} for the missing outcomes
2. Fit a model to the full data, obtaining $\hat{\beta}^*$ and corresponding standard error estimates
3. Repeat steps 1&2 many times, and finally average the results

GEE: multiple imputation

Y_{i0}	Y_{i1}	Y_{i2}	X_i	Z_i
23	31.2	.	6.4	1
24.5	26.7	3.4	5.5	0
22	.	4.6	4.2	0
.	21	.	5.5	1

0. Data as observed:
(No analysis performed)

1. Do K multiple
imputations:

Y_{i0}	Y_{i1}	Y_{i2}	X_i	Z_i
23	31.2	5.5	6.4	1
24.5	26.7	3.4	5.5	0
22	24.5	4.6	4.2	0
23.1	21	6.3	5.5	1

Y_{i0}	Y_{i1}	Y_{i2}	X_i	Z_i
23	31.2	4.8	6.4	1
24.5	26.7	3.4	5.5	0
22	25.1	4.6	4.2	0
22.9	21	6.5	5.5	1

Y_{i0}	Y_{i1}	Y_{i2}	X_i	Z_i
23	31.2	5.3	6.4	1
24.5	26.7	3.4	5.5	0
22	25.5	4.6	4.2	0
22.7	21	6.0	5.5	1

2. Analyze each:

$$\hat{\beta}_1, \hat{s}_1$$

$$\hat{\beta}_{MI} = \frac{1}{K} \sum_{k=1}^K \hat{\beta}_k,$$

$$\hat{s}_{MI}^2 = \frac{1}{K} \sum_{k=1}^K \hat{s}_k^2 + \frac{K+1}{K(K-1)} \sum_{k=1}^K (\hat{\beta}_k - \hat{\beta}_{MI})^2$$

3. Final results:
Use Rubin's rules,
on all K analyses

GEE: multiple imputation

These follow from MI's quasi-Bayesian flavor; for a univariate parameter suppose we have $\{\hat{\beta}_k^*, \hat{\sigma}_k^2\}, k = 1, \dots, K$;

$$\hat{\beta}_{MI} = \frac{1}{K} \sum_{k=1}^K \hat{\beta}_k^*$$

$$\hat{\sigma}_{MI}^2 = \frac{1}{K} \sum_{k=1}^K \hat{\sigma}_k^2 + \left(1 + \frac{1}{K}\right) \frac{1}{K-1} \sum_{k=1}^K (\hat{\beta}_k^* - \hat{\beta}_{MI})^2$$

... note the overall variance can be written informally as $\mathbb{E}[\text{Var}|k] + \text{Var}[\mathbb{E}|k]$, i.e. the variance conditional on what we don't know (k) averaged over what we don't know, plus the spread of what we don't know.

- These are known as 'Rubin's rules'; natural extensions hold for multivariate parameters
- Compare $\hat{\beta}_{MI}/\hat{\sigma}_{MI}$ to $N(0, 1)$ for inference; a t_{df} version is also available, but $df \approx K$... so just make K big

GEE: multiple imputation (*)

Actually MI works under a weaker condition, known as *missing at random* (MAR);

$$f(\mathbf{M}|\mathbf{Y}, \mathbf{X}) = f(\mathbf{M}|\mathbf{Y}^o, \mathbf{Y}^m, \mathbf{X}) = f(\mathbf{M}|\mathbf{Y}^o, \mathbf{X})$$

i.e. missingness only depends on observed outcomes. For example, taking blood pressures, we first take two measurements. If they are not too far apart, use their average – otherwise take another measurement. The third measurement is MAR.

Why does MI work here? Again being informal, under MAR,

$$\begin{aligned} f(\mathbf{Y}|\mathbf{M}, \mathbf{X}) &= \frac{f(\mathbf{Y}|\mathbf{X})}{f(\mathbf{M}|\mathbf{X})} f(\mathbf{M}|\mathbf{Y}, \mathbf{X}) = \frac{f(\mathbf{Y}|\mathbf{X})}{f(\mathbf{M}|\mathbf{X})} f(\mathbf{M}|\mathbf{Y}^o, \mathbf{X}) \\ &= f(\mathbf{Y}^m|\mathbf{Y}^o, \mathbf{X}) f(\mathbf{Y}^o|\mathbf{M}, \mathbf{X}). \end{aligned}$$

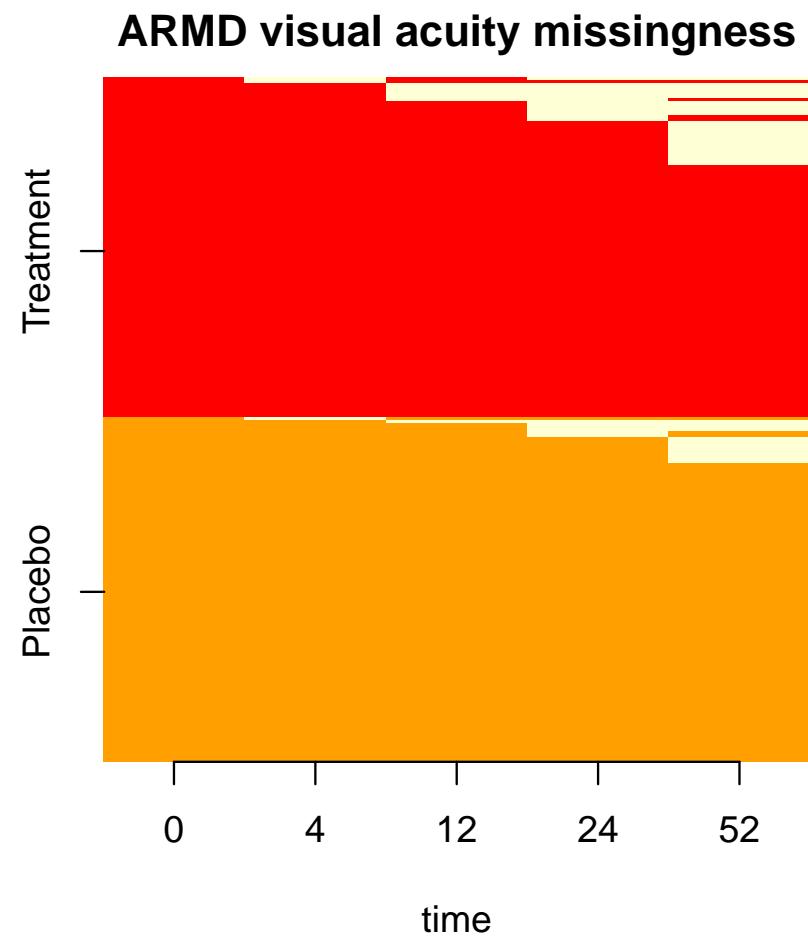
Hence, information about $\mathbf{Y}|\mathbf{M}, \mathbf{X}$ is proportional to information about $\mathbf{Y}^o|\mathbf{M}, \mathbf{X}$, i.e. what we learn about from the observed outcomes (i.e. the \mathbf{Y}^o , where we know $M = 0$) tells us about the general pattern.

GEE: macular degeneration

We consider an example from a clinical trial of a drug for age-related macular degeneration (ARMD) – an eye disease. The active drug ($\text{trt}=1$) is compared to placebo ($\text{trt}=0$)

Visual acuity (i.e. eyesight) is measured at $t = 4, 12, 24$ and 52 weeks; a binary outcome of interest is whether it has improved compared to baseline, i.e. $t = 0$.

Visual acuity is missing in a small percentage of observations. Most appear to be ‘dropouts’ (The other patterns are called *non-monotone* missingness.)



GEE: macular degeneration

The analysis of substantive interest is GEE logistic regression;

$$\text{logit}(\mathbb{P}[Y_{it} | \mathbf{X}_{it}]) = \beta_0 + \beta_1 \text{trt}_i + \gamma_t + \delta_t \text{trt}_i$$

where;

- $t = 4, 12, 24, 52$ indexes time; $t = 0$ indicates baseline
- Y_{it} is an indicator (1/0) of improvement from baseline, i.e. $\text{Visual acuity}_t > \text{Visual acuity}_0$
- We constrain $\gamma_4 = \delta_4 = 0$
- β_0 is the log odds of improvement at $t = 4$
- β_1 is the log odds ratio of improvement in those on treatment versus placebo, at $t = 4$
- γ_t is the log odds ratio of improvement at $t \neq 4$, compared to $t = 4$, in those on placebo
- δ_t is the difference in log odds ratio of improvement at $t \neq 4$ compared to $t = 4$ in those on treatment, between those on active treatment versus placebo

We will use exchangeable working correlation, which is unlikely to be a terrible choice with complete nearly-balanced data.

GEE: macular degeneration

For the imputation model, we work directly with the ‘raw’ outcomes. The goal here is to *predict* the missing outcomes as well as possible, based on *all* the other data.

We use time-specific linear regression, i.e. for $t = 4, 12, 24, 52$

$$Y_{it} \sim N(\alpha_{t0} + \alpha_{t1} \text{trt}_i + \alpha_{t2} \text{visual}_0, \sigma_t^2).$$

To obtain relevant $\hat{\alpha}_{tk}$ and $\hat{\sigma}_t^2$, we fit four completely distinct regressions – note all parameters are time-specific.

Doing this in R;

```
mod4 <- lm(visual~treat + visual0, data=subset(arnd3, time==4))
mod12 <- lm(visual~treat + visual0, data=subset(arnd3, time==12))
mod24 <- lm(visual~treat + visual0, data=subset(arnd3, time==24))
mod52 <- lm(visual~treat + visual0, data=subset(arnd3, time==52))
```

... where visual_{0j} is the baseline visual acuity for each individual, i.e. a copy of visual_{i0} .

GEE: macular degeneration

Imputing the data is then tedious but not hard; first set up which observations to impute and how to impute, for each time point;

```
miss4  <- is.na(armd3$visual) & armd3$time==4
miss12 <- is.na(armd3$visual) & armd3$time==12
miss24 <- is.na(armd3$visual) & armd3$time==24
miss52 <- is.na(armd3$visual) & armd3$time==52

mu4   <- predict( mod4, newdata=armd3[ miss4,]); sig4   <- summary( mod4)$sigma
mu12  <- predict(mod12, newdata=armd3[miss12,]); sig12  <- summary(mod12)$sigma
mu24  <- predict(mod24, newdata=armd3[miss24,]); sig24  <- summary(mod24)$sigma
mu52  <- predict(mod52, newdata=armd3[miss52,]); sig52  <- summary(mod52)$sigma
```

Finally, writing a function to do the imputation, and analyze the imputed dataset;

```
do.one <- function(){
  armd.i <- armd3
  armd.i[ miss4, "visual"] <- rnorm( sum( miss4), mu4, sig4 )
  armd.i[miss12, "visual"] <- rnorm( sum(miss12), mu12, sig12 )
  armd.i[miss24, "visual"] <- rnorm( sum(miss24), mu24, sig24 )
  armd.i[miss52, "visual"] <- rnorm( sum(miss52), mu52, sig52 )
  gee.i <- geem( visual>visual0 ~ factor(treat)*factor(time), id="subject",
  data=subset(armd.i, time>0), corstr="exchangeable", family=binomial)
  gee.i }
```

GEE: macular degeneration

It's not hard to code up Rubin's rules yourself, but...

```
set.seed(4)
my.mi <- replicate(100, do.one(), simplify=FALSE)
library("MItools")
# pull out betahats and their variance estimates from the 100 analyses;
betas <- MIextract(my.mi, fun=coef)
vars  <- MIextract(my.mi, fun=function(x){as.matrix(x$var)})
MIcombine(betas,vars)
```

Comparing with complete-case GEE;

Term	Trt		Trt: $t = 12$		Trt: $t = 24$		Trt: $t = 52$	
	$\hat{\beta}_1$	\widehat{SE}	$\hat{\delta}_{12}$	\widehat{SE}	$\hat{\delta}_{24}$	\widehat{SE}	$\hat{\delta}_{52}$	\widehat{SE}
CC	-0.27	0.26	-0.39	0.32	0.05	0.35	-0.46	0.39
MI	-0.29	0.26	-0.39	0.32	0.09	0.35	-0.32	0.38
% miss	1		2		3		8	

The estimated *rate of missing information* compares the precision of complete case versus imputed analyses. It takes into account leverage of imputed observations – so is similar to counting NAs, but more sophisticated.

GEE: macular degeneration

Computationally, what work do we have to do?

- Most of the work lies in doing the imputation, in particular thinking about how to impute data with different patterns of missingness
- There are several packages that try to remove this, i.e. `mice` and `Amelia`. The defaults use simple models that impute Y_{ij} based on X_{ij} . For clustered data you may need to reshape the data to wide format, i.e. have $Y_{i0}, Y_{i4}, Y_{i12} \dots$ all on one row. Dealing with lags (i.e. impute later times based on earlier data alone) is more work
- Running the code need not take long; unless missingness is very high the Monte Carlo error reduces quickly
- Some large studies keep on record 5–10 ‘fully imputed’ datasets, so that future analyses don’t have to set up imputation models – and so that old & new analyses agree

GEE: macular degeneration

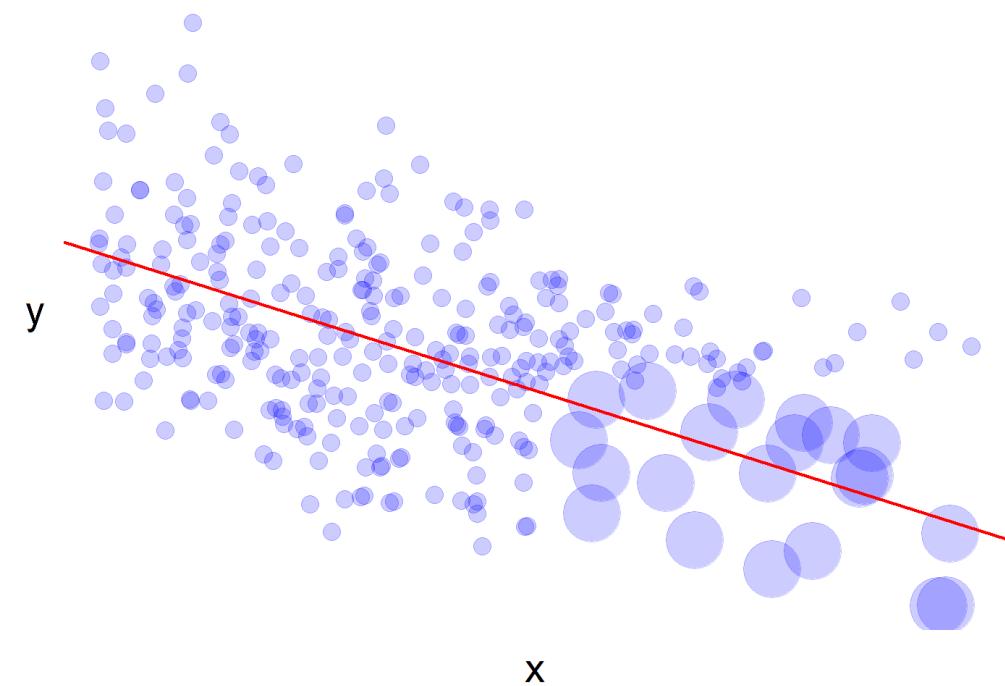
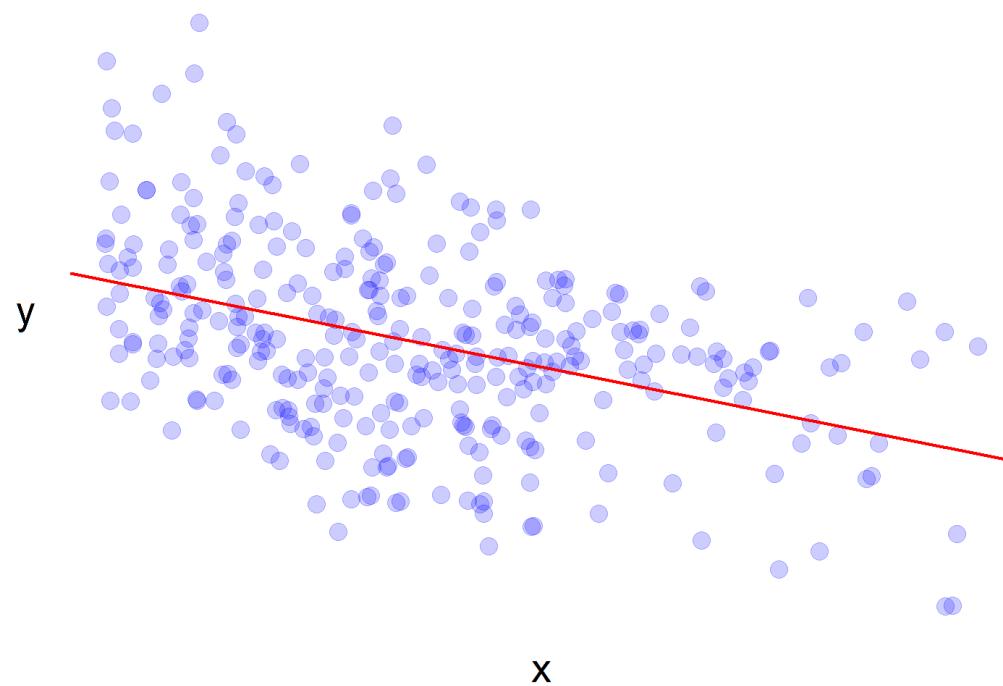
Statistically, what's doing the work?

- The imputation model should correctly reflect aspects of the data that make your analysis ‘work’. Hence, for GEE we need to impute data with the correct $\mathbb{E}[Y_{ij}|\mathbf{X}_{ij}]$ – and the right variance if we want efficiency. This ‘agreement’ is known as *congeniality*. (It wasn’t widely understood until Meng (1994) – see the discussion of ‘controversy’)
- As we don’t know the truth, it’s prudent to use a flexible imputation model, to minimize the risk of uncongeniality
- To fully reflect uncertainty about the imputation model parameters (μ_t, σ_t) ‘proper’ imputation would first sample these from e.g. $N(\hat{\mu}_t, \widehat{\text{Var}}[\hat{\mu}_t])$. But this is not always done, and it only matters if this uncertainty is large
- As noted earlier, the imputation is (just) a prediction method. Using observed Y_{ij} to predict missing Y_{ij} is fine – unlike using regression for inference

As ever, using diagnostics on the imputation model is not silly, but don’t expect diagnostics to tell you how to impute.

GEE: weighted for missingness

MI is can be used very widely. The other general approach to missing data uses *weights* to compensate for the rate of missingness;



Here, only 10% observed in lower RH, so upweight these by $\times 10$ to get consistent estimates.

GEE: weighted for missingness

The GEE version of this weighting solves estimating equations

$$\frac{1}{n} \sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} \text{diag} \left\{ (1 - M_{ij}) w_{ij} \right\} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}_p,$$

where

$$w_{ij} = \begin{cases} \frac{1}{\widehat{\mathbb{P}}[M_{ij}=0|\mathbf{X}_{ij}]}, & M_{ij} = 0 \\ 0, (\text{by default}) & M_{ij} = 1. \end{cases}$$

Assuming these weights can be estimated well enough, this weighting counteracts the $\mathbb{P}[M_{ij} = 0]$ terms seen on slide 2.135, and we obtain consistent estimates.

- Known as *inverse-probability* weighting
- Under MAR, what we know about $\mathbb{P}[M|\mathbf{Y}^o, \mathbf{X}]$ tells us about $\mathbb{P}[M|\mathbf{Y}, \mathbf{X}]$ and hence $\mathbb{P}[M|\mathbf{Y}^m, \mathbf{X}]$ – so it's okay to ‘fit weights’ using complete-case data. Under MCAR can just fit $\mathbb{P}[M|\mathbf{X}]$, ignoring \mathbf{Y} .

GEE: weighted for missingness

The best-known version of this method is *weighted GEE* (WGEE) – developed by Robins & Rotnitzky (1995, see also 1994 and 1995b).

- Despite the name, it only deals with dropout – though this is useful in practice
- Valid under MAR even if the correlation model is incorrectly specified, **provided** the model for M_{ij} is correct
- As with GEE, use of the robust variance estimator in WGEE provides robustness to misspecification of the correlation structure, but it does require a correct model for the weights
- As with GEE, choice of the working correlation matrix affects efficiency – closer to the truth is better
- Requires correct specification for $\mu = g(\mathbf{X}\boldsymbol{\beta})$ and sufficiently large n (as always)

GEE: weighted for missingness

For the model for M_{ij} , denote

$$\begin{aligned}\pi_{ij} &= \mathbb{P}[M_{ij} = 0 | M_{i(j-1)}, \mathbf{Y}_i, \mathbf{X}_i] \\ &= \mathbb{P}[M_{ij} = 0 | M_{i(j-1)}, \mathbf{Y}_i^o, \mathbf{X}_i]\end{aligned}$$

and we fit this model via logistic regression, i.e.

$$\text{logit}(\pi_{ij}) = \mathbf{Z}_{ij}\boldsymbol{\theta}$$

where covariates \mathbf{Z} in this model may include \mathbf{Y}, \mathbf{X} , and other variables.

- Fit and check this model as you would any other
- Flexible models are a good starting place – as then you have a better chance of being close to the truth. However, more flexibility give less precision, i.e. poorer finite-sample behavior
- Dealing only with dropout makes specification of the model somewhat easier

GEE: weighted for missingness

Finally, weight w_{ij} is the inverse of the unconditional probability that subject i is observed at time j , and is estimated by the inverse of the cumulative product of the estimated conditional probabilities, e.g., fitted values

$$w_{ij} = \frac{1}{\hat{\pi}_{i1} \times \hat{\pi}_{i2} \times \dots \times \hat{\pi}_{ij}}.$$

Observations with low probability of being observed receive a large weight; observations with high probability of being observed receive a small weight.

- In R, `geem()` permits weights, `gee()` does not. Stata **only permits cluster-specific weights**
- The sandwich calculations should contain both $\hat{\beta}$ and $\hat{\theta}$ – so are tedious, but nothing new
- Do look at the weights! Observations with $w_{ij} = 10$ have $\times 10$ more influence than in unweighted analysis; extreme sensitivity can occur

GEE: weighted for missingness

Other notes on this approach;

- Weighting doesn't use partially-observed observations (except in $\hat{\mathbb{P}}[M]$) so may be inefficient compared to multiple imputation
- Compared to knowing the true $\mathbb{P}[M_{ij}]$, using estimated weights can be **more** efficient. (Broadly, your data may indicate weights that re-weight your data closer to the population, versus weights that consider all possible datasets.)
- While not considered here, *doubly robust* methods provide valid inference when **at least one** of the mean model and missingness model is correctly specified

However...

Arguably, it is a tall order to fit a parametric drop-out model, for which the data necessarily provide relatively sparse information, in circumstances where the analysts are reluctant to commit themselves to a parametric model for the covariance structure.

GEE: going beyond MCAR and MAR

Just like MCAR, MAR can be unrealistic in practice;

- Missingness for sensitive outcomes may depend on those outcomes, even with the same \mathbf{X} and observed outcomes.
For example, weight, drug adherence, sexual activity
- Any pattern where $\mathbb{P}[M|\mathbf{Y}^m, \mathbf{X}] \neq \mathbb{P}[M|\mathbf{Y}^o, \mathbf{X}]$ is called *Missing Not At Random* (MNAR) or *Not Missing At Random* (NMAR)
- Both are **terrible** names! But standard...

With MNAR, the data **cannot** tell you how to impute or weight.
Instead, the best you can do is sensitivity analysis; e.g. if

$$\text{logit}(\mathbb{P}[M|\mathbf{Y}^m = \mathbf{y}, \mathbf{X} = \mathbf{x}]) = \text{logit}(\mathbb{P}[M|\mathbf{Y}^o = \mathbf{y}, \mathbf{X} = \mathbf{x}]) + \boldsymbol{\gamma}^T \mathbf{y} + \boldsymbol{\delta}^T \mathbf{x}$$

for some $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$ you specify, how does the WGEE answer change? If the answer is ‘not importantly’ for **plausible** MNAR mechanisms, this provides some reassurance that the original analysis was not grossly wrong.

GEE: other names for missing

While the literature on ‘dropout’ in longitudinal studies is large, the basic same problem is also called *informative cluster size*;

- Oral health: We might expect to see more observations from more teeth in those who have lower rates of cavities (Y)
- Fertility clinics: couples may be observed to have more pregnancies if they have fewer successful births (Y)
- Education: as good schools attract students, we may see more students in schools that achieve better grades (Y)

Methods do exist to upweight small clusters to account for this form of missingness – see the class site – but to use them, interest must lie in a possibly-hypothetical population where all clusters are equal-sized;

- Everyone has all their teeth
- Every couple has the same number of pregnancies
- School size is identical (or at least unaffected by grades)

GEE: missing covariates

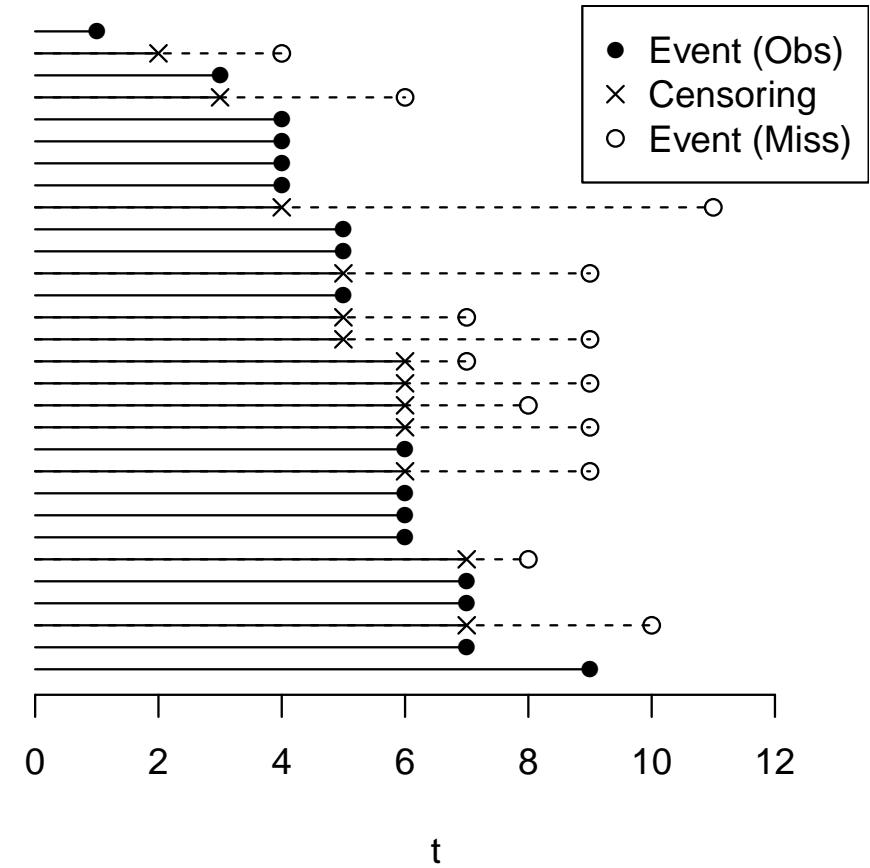
We have only discussed missing outcomes;

- Relying on the ability to do inference conditional on the observed \mathbf{X} , this is valid
- With random \mathbf{X} , this conditioning may throw away useful information – recall the setting of the Pepe & Anderson result
- Treating \mathbf{X} as random, missing values can be imputed or weighted, with natural generalizations of the methods we have discussed
- Obviously, bigger missingness models involve more assumptions to think about, fit, and check
- In addition, definitions of MCAR/MAR become more complex, as do issues of congeniality – see the class site for some related literature
- ‘Chained equation’ approaches (that closely resemble MCMC) are becoming standard, see e.g. the `mice` package

GEE: connection to survival analysis

Instead of doing inference on $\mathbb{E}[Y_{it}]$, an alternative is to consider how time-to-event (e.g. first $Y_{it} = 1$, or first $Y_{it} > c$) depends on covariates. Some subjects are *censored* before they have the event of interest, e.g. dropout before observed death;

- Analysis relies on *non-informative censoring*, a form of MAR. Given \mathbf{X} , later risk is identical for those censored/not censored at any given t .
- Only one ‘outcome’ per person, so not really clustered
- Considers *hazard*, a non-standard function of $\mathbb{E}[Y_{it}]$, with methods tailored to this. See BIOC 537 or 576



GEE: missing data summary

Methods for missing data;

- Are important! Careful thinking about them can be a key contribution by statisticians
- Are important for clustered settings; the difference between MCAR and MAR is particularly important when you do see **some** other outcomes for each subject
- Complete-case analysis (where justified), imputation and weighting approaches are all reasonable solutions
- Directly modeling the missingness is also sensible, but hard to reconcile with GEE's semi-parametric flavor
- Beware of anything else! – even if well-intentioned

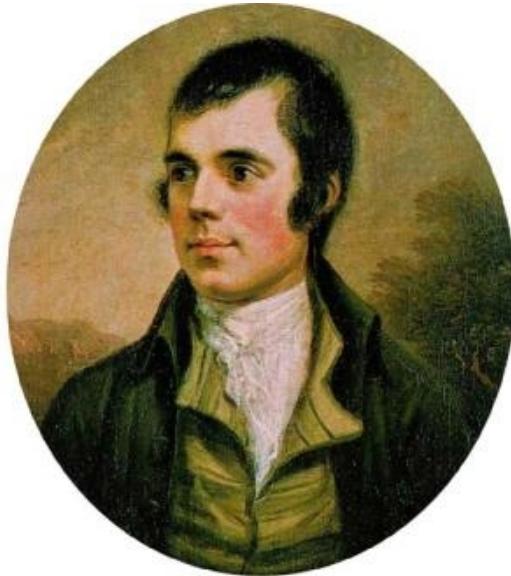
For those taking 572, note this area has lots of recent work, some by local faculty.

Chapter 2: summary

Overall;

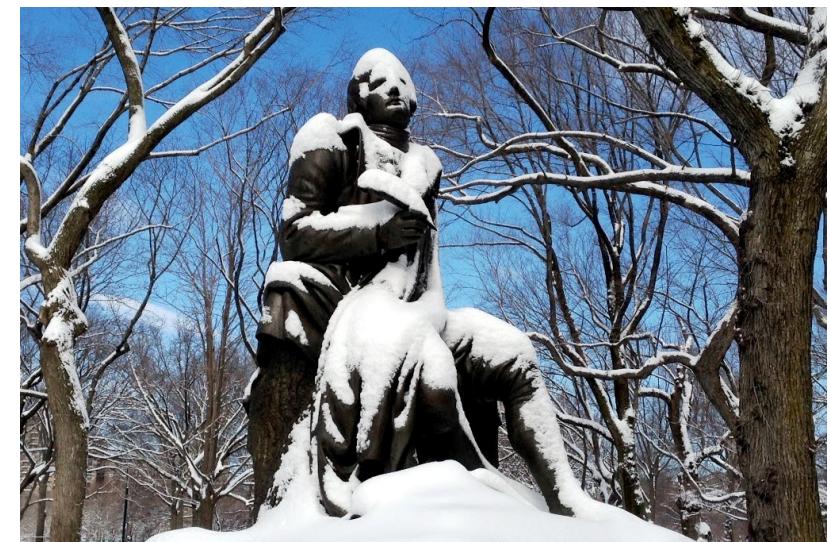
- Treating clusters as vector outcomes, we get inference on marginal parameters
- Direct use of EEs/bootstrap is possible, and well-calibrated, with large n , but mean models aid parameter interpretation
- GEE is a standard tool, that competent statisticians are expected to understand
- Using GEE, efficiency is obtained with working covariances that mimic the truth. With restricted \mathbf{Y} (i.e. binary \mathbf{Y}) this is more challenging
- GEE does not answer every question you could ask
- Complete-case GEE requires the strong assumption of MCAR ... or more work/more assumptions to accommodate MAR missingness

CHAPTER 3: MIXED MODELS



*I waive the quantum o' the sin,
The hazard of concealing
But, och! it hardens a' within,
And petrifies the feeling!*

from 'Epistle to a Young Friend'
by Robert Burns (1759–1796)
Scottish poet, farmer, and lyricist



CHAPTER 3: MIXED MODELS

The big attraction of Chapter 2's robust methods is that several potential 'sins' (i.e. misspecified within-cluster variances and covariances) are waived – when we have many clusters.

A big drawback is that, without assumptions of this sort, we only obtain inference on *marginal* parameters, i.e. averaging over all observations in all clusters.

To say what happens within given clusters, i.e. to make inference on *conditional* parameters, we have to make assumptions – about how individual clusters, and their observations, differ from the rest of the population.

Saying how different each cluster is from the average enables us to 'adjust for cluster' – just like saying how different men and women are (e.g. a single additive term) enables us to 'adjust for sex'.

Mixed models: motivation

Suppose you're studying test scores, in students in some schools;



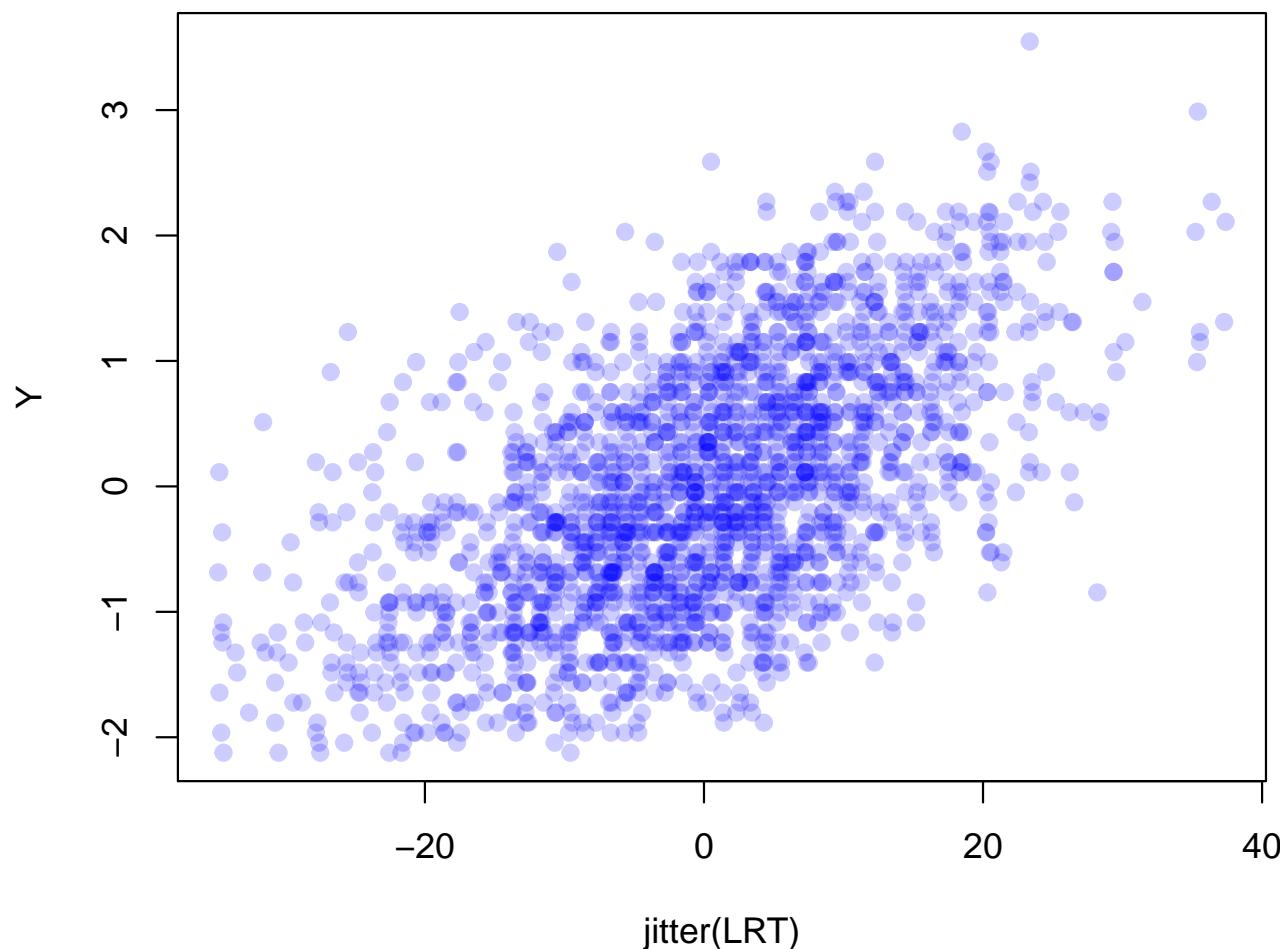
Mixed models: motivation

You sampled schools, but they were a random sample from a much bigger *population* of schools;



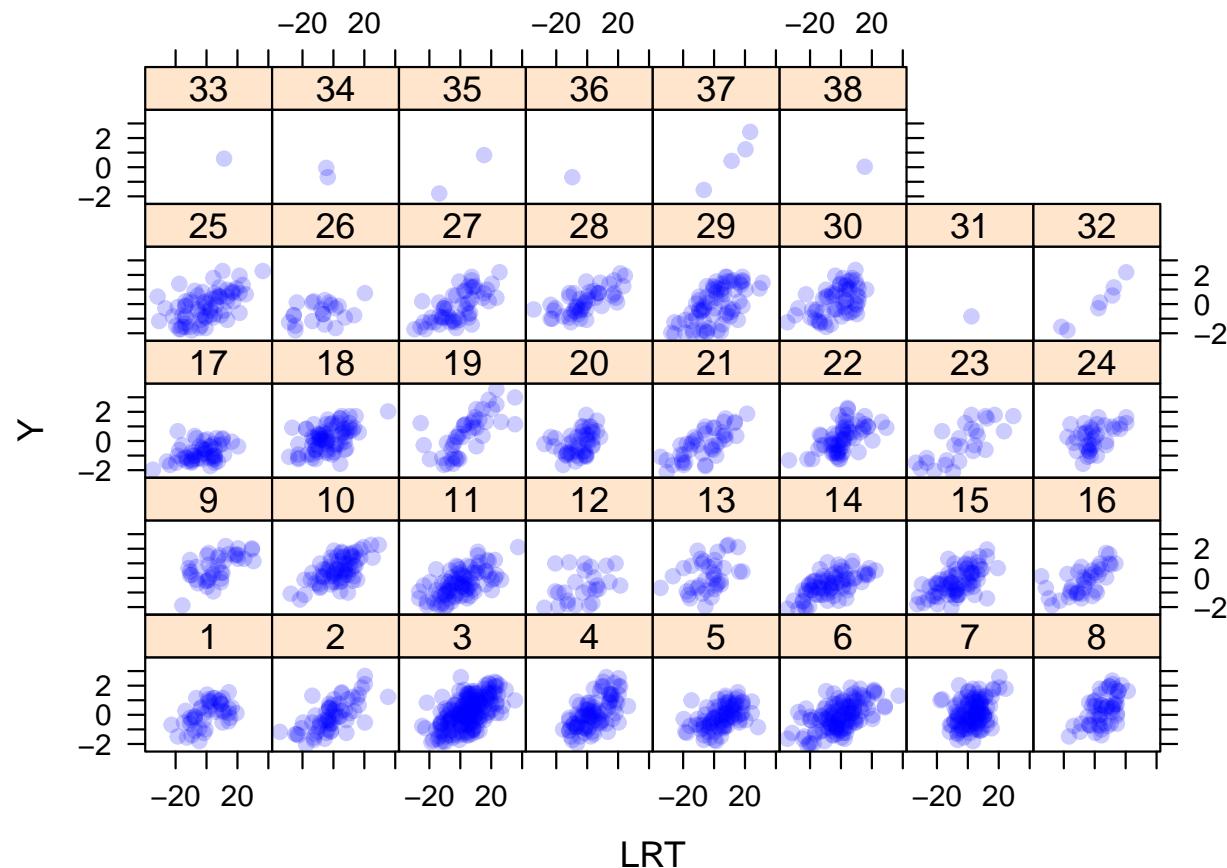
Mixed models: motivation

End-of-year exam scores (Y) versus earlier London Reading Test scores ($X = LRT$) for 1978 students in 38 London schools;



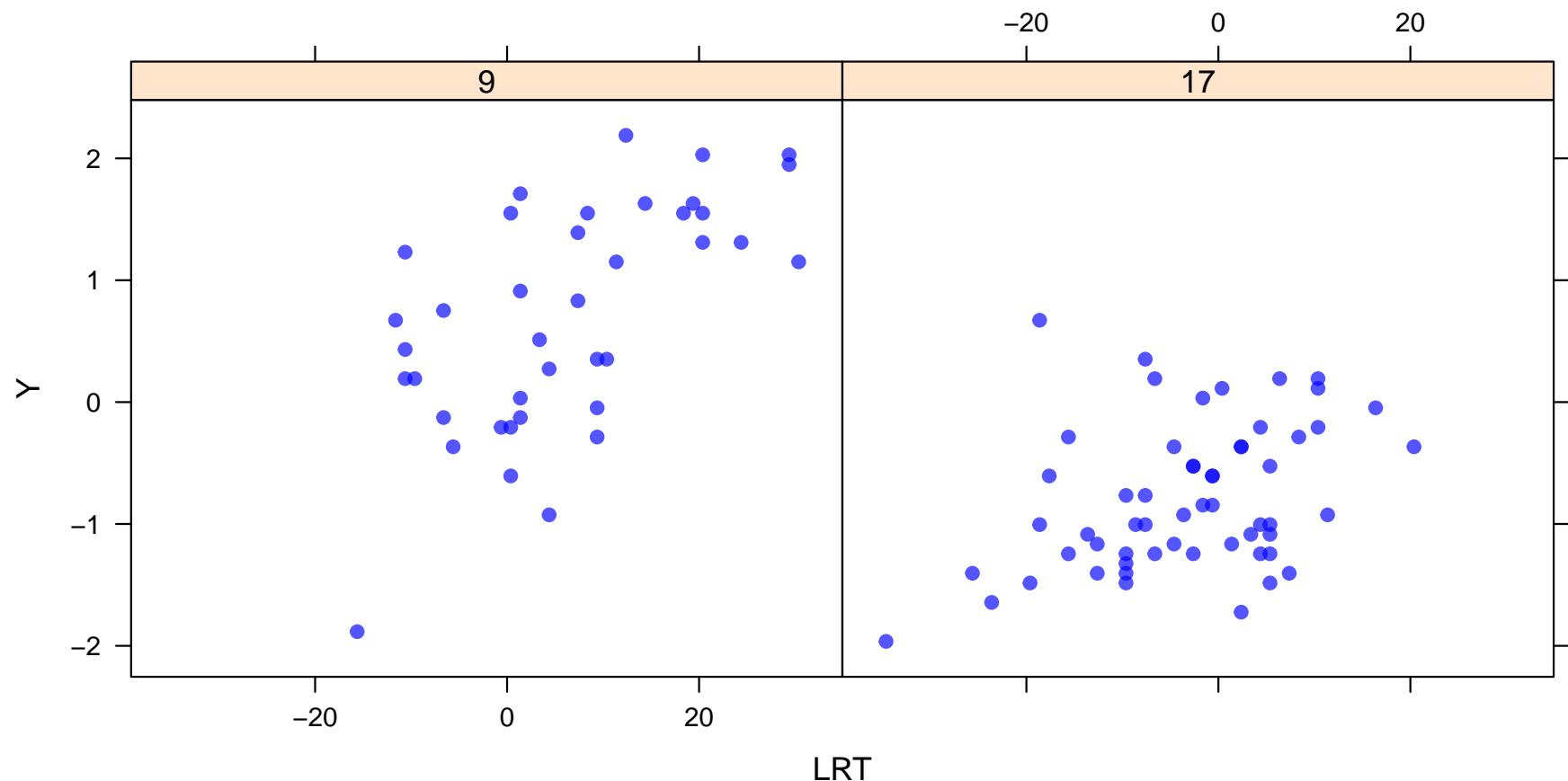
Mixed models: motivation

End-of-year exam scores (Y) versus earlier London Reading Test scores ($X = LRT$) for 1978 students in 38 London schools;



Mixed models: motivation

Same plot for schools 9 and 17; clearly, these schools are different *in some way*.



Mixed effects models: notation

Using the same Y_{ij}, X_{ij} notation as before, a model for the data assumes the randomly-sampled schools all have same X coefficient, but differ (only) through a school effect b_i in the mean, conditional on which the Y_{ij} are independent with:

$$\begin{aligned}\mathbb{E}[Y_{ij}|X_{ij}, b_i] &= \beta_0 + b_i + \beta_1 X_{ij} && \text{Conditional mean} \\ \text{Var}[Y_{ij}|X_{ij}, b_i] &= \sigma_Y^2 && \text{Conditional variance} \\ \mathbb{E}[b_i|\mathbf{X}_i] &= 0 && \text{Mean school effect} \\ \text{Var}[b_i|\mathbf{X}_i] &= \sigma_b^2 && \text{Var of school effects}\end{aligned}$$

- We assume **everything** ‘different’ about how students perform in school i (conditional on X_{ij}) is summarized by b_i
- ... so an ‘average’ school would have $b_i = 0$, i.e. mean score β_0 , for students with $X_{ij} = 0$
- The schools were sampled at random, so the b_i ’s are also a random i.i.d. sample; they are known as *random effects*
- β_0 and β_1 are not random, and are known as *fixed effects*. When both appear, we have a *mixed [effects] model*

Mixed models: mean & (co)variance

Q. What does this model say about outcomes in observations with particular X_{ij} ?

A1. The marginal mean, $\mathbb{E}[Y_{ij}|X_{ij}] = \beta_0 + \beta_1 X_{ij}$ – i.e. it tells us about **conditional β**

A2. In different schools, outcomes are independent – because each b_i is only involved in mean scores in school i . So we know $\text{Cov}[Y_{ij}, Y_{i'j'}|X_{ij}, X_{i'j'}] = \text{Corr}[Y_{ij}, Y_{i'j'}|X_{ij}, X_{i'j'}] = 0$.

A2. Within schools, **for this model** it's helpful to write

$$Y_{ij} = \beta_0 + b_i + \beta_1 X_{ij} + \epsilon_{ij}$$

... where ϵ_{ij} are known as the *error terms*. Then

$$\begin{aligned}\text{Var}[Y_{ij}|X_{ij}] &= \text{Var}[\beta_0 + b_i + \beta_1 X_{ij} + \epsilon_{ij}|X_{ij}] \\ &= \text{Var}[b_i + \epsilon_{ij}|X_{ij}] \\ &= \text{Var}[b_i] + \text{Var}[\epsilon_{ij}] \\ &= \sigma_b^2 + \sigma_Y^2.\end{aligned}$$

Mixed models: mean & (co)variance

A4. In two different observations in one school;

$$\begin{aligned}\text{Cov}[Y_{ij}, Y_{ik}|X_{ij}, X_{ik}] &= \text{Cov}[\beta_0 + b_i + \beta_1 X_{ij} + \epsilon_{ij}, \beta_0 + b_i + \beta_1 X_{ik} + \epsilon_{ik}|X_{ij}, X_{ik}] \\ &= \text{Cov}[b_i + \epsilon_{ij}, b_i + \epsilon_{ik}|X_{ij}, X_{ik}] \\ &= \text{Cov}[b_i, b_i] + \text{Cov}[\epsilon_{ij}, b_i] + \text{Cov}[b_i, \epsilon_{ik}] + \text{Cov}[\epsilon_{ij}, \epsilon_{ik}] \\ &= \text{Var}[b_i|X_{ij}, X_{ik}] + 0 + 0 + 0 \\ &= \sigma_b^2.\end{aligned}$$

The *intra-class correlation* or *intra-cluster correlation* (see Chapter 2) is therefore;

$$\text{Corr}[Y_{ij}, Y_{ik}|X_{ij}, X_{ik}] = \frac{\sigma_b^2}{\sqrt{\sigma_b^2 + \sigma_Y^2} \sqrt{\sigma_b^2 + \sigma_Y^2}} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_Y^2}$$

- Within each cluster, this is the same for all pairs j, k – as we saw with exchangeable correlation matrices, in Chapter 2. Another name is *compound symmetric* matrices
- This ICC is positive – so we're well inside the teabag!

Mixed models: inference on β

How might we estimate β ? Using answers A1–A4 we have established the *marginal* mean and variance of all the observations.

Assuming we knew the variance terms (i.e. $\mathbf{V}_i = \sigma_Y^2 I_{n_i} + \sigma_b^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T$) the Gauss-Markov theorem would tell us the minimum-variance linear unbiased estimator;

$$\hat{\beta} = \left(\sum_{i=1}^n \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^n \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{Y}_i \right),$$

in the usual notation, so \mathbf{X}_i is the design matrix of intercept and covariates and \mathbf{Y}_i the vector of outcomes, in each cluster.

- Actually Gauss & Markov only proved it for univariate outcomes; the vector version is due to **Aitken** (1935)
- As \mathbf{V}_i acts only as a weight, the ICC $\sigma_b^2 / (\sigma_b^2 + \sigma_Y^2)$ is sufficient to calculate $\hat{\beta}$ – no scale term is needed

Mixed models: inference on β

Under the modeling assumptions this estimate's variance is

$$\begin{aligned}\text{Var}[\hat{\beta}|\mathbf{X}] &= \left(\sum_{i=1}^n \mathbf{x}_i^T \mathbf{V}_i^{-1} \mathbf{x}_i \right)^{-1} \left(\sum_{i=1}^n \mathbf{x}_i^T \mathbf{V}_i^{-1} \mathbf{V}_i \mathbf{V}_i^{-1} \mathbf{x}_i \right) \left(\sum_{i=1}^n \mathbf{x}_i^T \mathbf{V}_i^{-1} \mathbf{x}_i \right)^{-1} \\ &= \left(\sum_{i=1}^n \mathbf{x}_i^T \mathbf{V}_i^{-1} \mathbf{x}_i \right)^{-1} \left(\sum_{i=1}^n \mathbf{x}_i^T \quad \mathbf{V}_i^{-1} \mathbf{x}_i \right) \left(\sum_{i=1}^n \mathbf{x}_i^T \mathbf{V}_i^{-1} \mathbf{x}_i \right)^{-1} \\ &= \left(\sum_{i=1}^n \mathbf{x}_i^T \mathbf{V}_i^{-1} \mathbf{x}_i \right)^{-1} = \frac{1}{n} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^T \mathbf{V}_i^{-1} \mathbf{x}_i \right)^{-1}\end{aligned}$$

- This is not the ‘full’ sandwich – and note that without knowledge of \mathbf{V}_i the cancellation wouldn’t work
- We do need to know σ_b^2 and σ_Y^2 to calculate this variance – but with consistent $\widehat{\sigma}_b^2$ and $\widehat{\sigma}_Y^2$, by results from Chapter 2, the usual ‘plug-in’ $\widehat{\text{Var}}[\hat{\beta}]$ gives valid large- n inference **with no further assumptions** – i.e. this is semi-parametric
- The cancellation occurs with any \mathbf{V}_i , and inference works if \mathbf{V}_i ’s free parameters can be estimated consistently

Mixed models: inference on β

Some hopefully-familiar estimates that will work;

$$\begin{aligned}\hat{\sigma}_b^2 &= \frac{1}{n} \sum_{i=1}^n \frac{1}{n_i(n_i - 1)} \sum_{j \neq j'} (Y_{ij} - \mathbf{X}_{ij}^T \hat{\boldsymbol{\beta}})(Y_{ij'} - \mathbf{X}_{ij'}^T \hat{\boldsymbol{\beta}}) \\ \hat{\sigma}_Y^2 &= \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} (Y_{ij} - \mathbf{X}_{ij}^T \hat{\boldsymbol{\beta}})^2 \right) - \hat{\sigma}_b^2\end{aligned}$$

- Up to $n - p$ correction factors, these are equivalent to the scale and correlation estimates used in linear regression GEE
- But the simpler variance estimate ($\hat{\mathbf{A}}^{-1}/n$ not $\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}/n$, see previous slide) – that uses the homoskedasticity assumptions – will work better in small samples
- The asymptotics ‘kick in’ faster, so e.g. coverage of 95% intervals can be expected to work better
- The `sandwich=FALSE` argument in `geem()` provides this, also the ‘model-based’ approaches in `geepack`’s `geese()`, but it has no standard name – so call it QL, say what you did, & why

Mixed models: London!

Doing this for the London schools example, using `geem()`:

```
geem1 <- geem(Y~LRT,id=school,data=schools,corstr="exchangeable",sandwich=FALSE)
outm <- data.frame(
  est  = geem1$beta,
  SE   = sqrt(diag(geem1$naiv.var)),
  loCI = geem1$beta + qnorm(0.025)*sqrt(diag(geem1$naiv.var)),
  hiCI = geem1$beta + qnorm(0.975)*sqrt(diag(geem1$naiv.var)),
  Z    = geem1$beta/sqrt(diag(geem1$naiv.var)),
  p    = pchisq( geem1$beta^2/diag(geem1$naiv.var), df=1, lower=FALSE ))
row.names(outm) <- geem1$coefnames
signif(outm,2)
      est      SE    loCI  hiCI      Z      p
(Intercept) 0.015 0.0330 -0.049  0.08  0.47  6.4e-01
LRT          0.047 0.0013  0.045  0.05 37.00 7.3e-306
# Noting that geem's phi = sigb^2 + sigy^2, alpha = sigb^2/phi
sqrt( geem1$alpha*geem1$phi )           # sigb
[1] 0.1954928 # Std Dev of school effects & given X
sqrt( geem1$phi- geem1$alpha*geem1$phi ) # sigy
[1] 0.7982854 # Std Dev of student scores, within-school (much bigger!)
```

- The `summary()` method for `geem` objects does not work with `sandwich=FALSE`
- A reminder: this is not typical use of `geem()`

Mixed models: London!

Computing issues aside, what does the output mean?

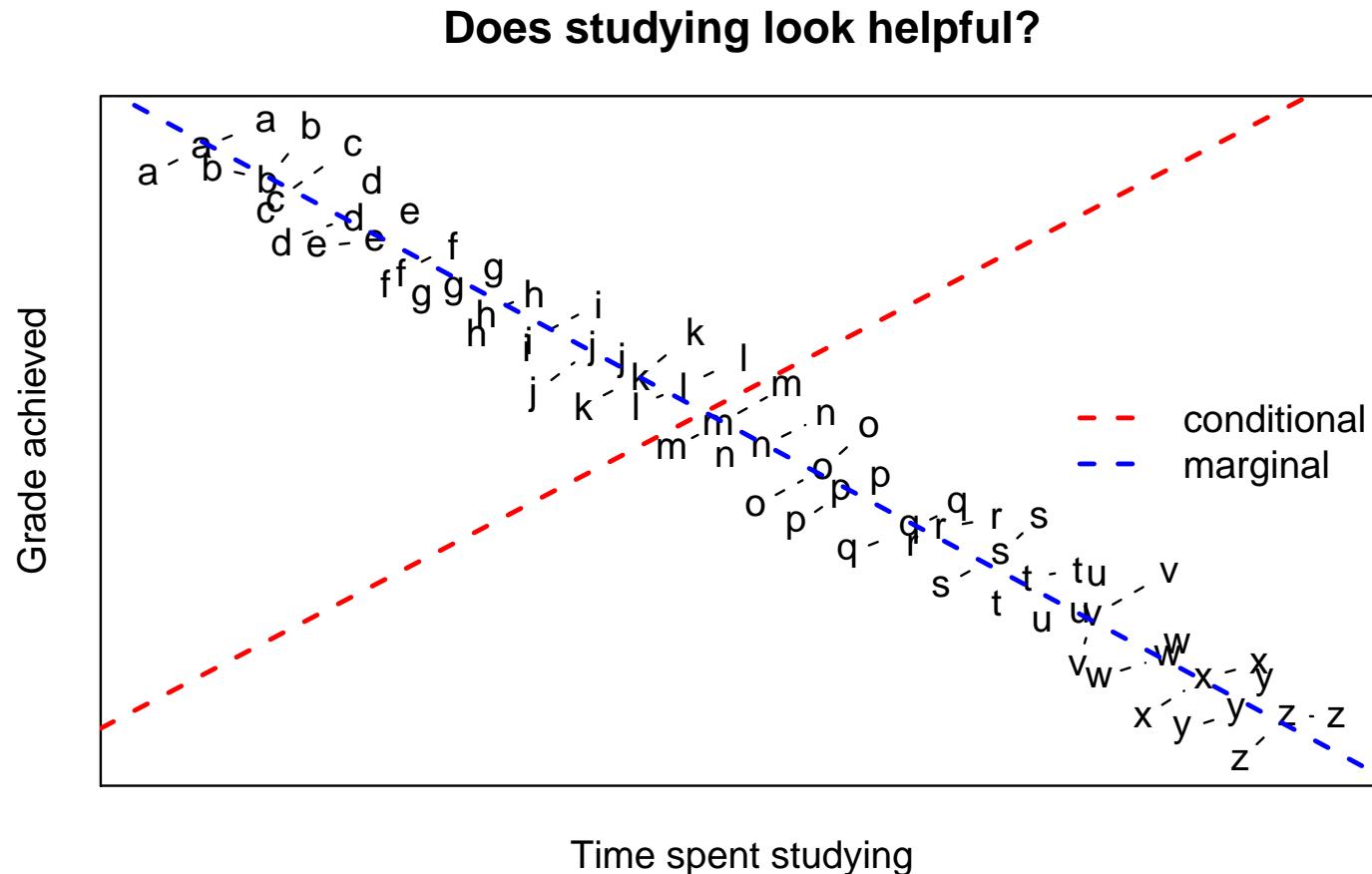
- $\hat{\beta}_0 = 0.015(-0.049, 0.080)$
 - **Marginal** Mean Y in subjects with $X_{ij} = 0$
 - **Conditional** Mean Y in subjects with $X_{ij} = 0$ in an average school ($b_i = 0$)
- $\hat{\beta}_1 = 0.047(0.045, 0.050)$
 - **Marginal** Difference in Mean Y per 1-unit difference in X , comparing subjects who may or may not be in the same cluster
 - **Conditional** Difference in Mean Y per 1-unit difference in X , comparing subjects who are in the same school (i.e. keeping b_i fixed)

As in Chapter 2, the marginal statements don't rely on the assumption about the random effects; we just need the mean model and between-cluster independence.

Q. Do the conditional statements need these assumptions?

Mixed models: marginal? conditional?

Q. Do the conditional statements need these assumptions?



One explanation: the b_i have $\mathbb{E}[b_i|X_i] \neq 0$ – i.e. b_i is strongly negatively correlated with X_i , violating the i.i.d. assumption

Mixed models: marginal? conditional?

Q. Do the conditional statements need these assumptions?

Property A1 (learning about conditional parameters from a marginal model) is **extremely** specialized;

- The ‘simple’ version (see next slides) only holds for (correctly-specified) linear links and i.i.d. random effects. Even then, care must be taken to specify \mathbf{V}_i correctly, in order to get valid standard errors, using non-sandwich QL methods
- If you can assume i.i.d. random effects with linear link, GEE **does** describe conditional parameters. (Elsewhere this is not true.) GEE sandwich SE estimates are robust, but may not be as efficient as non-sandwich QL methods
- A more complex version of this result holds for log-linear links – see later in the course
- Beyond this, there is **no** simple way to recover conditional β by fitting marginal models, without extra assumptions

Mixed models: marginal? conditional?

Back to the schools example: perhaps random intercepts don't capture everything unique to a school, in terms of average test score given X_{ij} . If the effect of X_{ij} also changes...

$$\begin{aligned}\mathbb{E}[Y_{ij}|X_{ij}, \mathbf{b}_i] &= \beta_0 + b_{i0} + \beta_1 X_{ij} + b_{i1} X_{ij} \\ &= (\beta_0 + b_{i0}) + (\beta_1 + b_{i1}) X_{ij}, \\ \text{Var}[Y_{ij}|X_{ij}, \mathbf{b}] &= \sigma_Y^2, \\ \mathbb{E}[\mathbf{b}_i|\mathbf{X}_i] &= \mathbf{0}_2, \\ \text{Var}[\mathbf{b}_i|\mathbf{X}_i] &= \mathbf{G}_{2 \times 2}.\end{aligned}$$

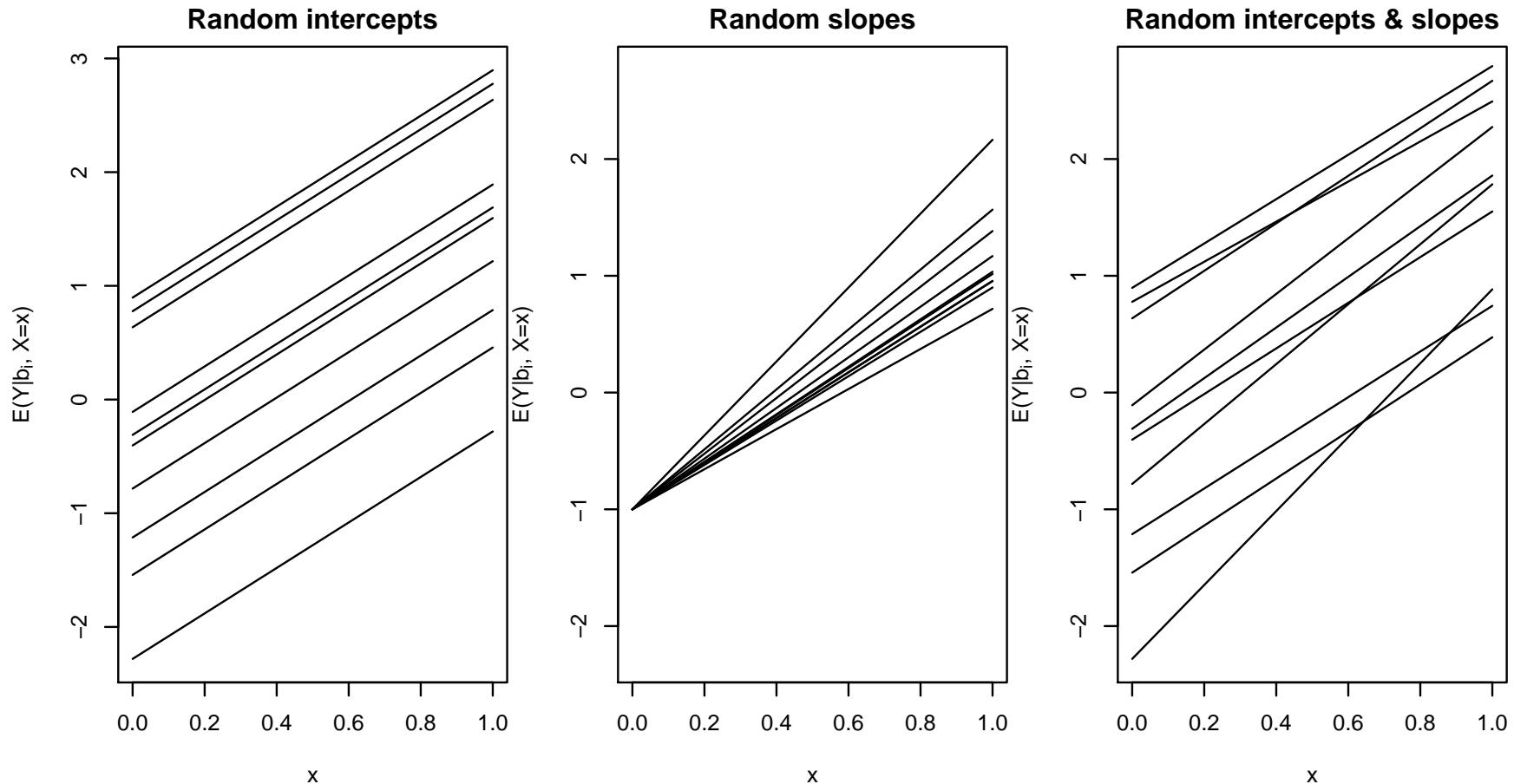
Marginally, using the usual notation;

$$\mathbb{E}[Y_{ij}|X_{ij}] = \beta_0 + \beta_1 X_{ij}, \quad \text{Var}[\mathbf{Y}_i|\mathbf{X}_i] = \sigma_Y^2 I_2 + \mathbf{X}_i \mathbf{G} \mathbf{X}_i^T.$$

We do get marginal β =conditional β , but no off-the-shelf GEE software permits the stated variance structure. Other software (coming soon!) does perform the sandwich free inference, but restricts how elements of \mathbf{G} and σ_Y^2 are estimated.

Mixed models: marginal? conditional?

Exactly which terms do you want to be random?



Random slopes alone are seldom used in practice, but are not totally implausible if baseline mean is well-controlled.

Mixed models: marginal? conditional?

Generalizing the idea of random slopes, and relaxing the i.i.d assumption on $Y_{ij}|X_{ij}, \mathbf{b}$;

$$\begin{aligned}\mathbb{E}[\mathbf{Y}_i|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i] &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i \\ \text{Var}[\mathbf{Y}_i|\mathbf{X}_i, \mathbf{b}] &= \phi\mathbf{R}_i \\ \mathbb{E}[\mathbf{b}_i|\mathbf{X}_i] &= \mathbf{0} \\ \text{Var}[\mathbf{b}_i|\mathbf{X}_i] &= \mathbf{G}_i,\end{aligned}$$

where covariates in \mathbf{Z} may or may not overlap with those in \mathbf{X} .
The induced moments are

$$\mathbb{E}[\mathbf{Y}_i|\mathbf{X}_i, \mathbf{Z}_i] = \mathbf{X}_i\boldsymbol{\beta}, \quad \text{Var}[\mathbf{Y}_i|\mathbf{X}_i] = \phi\mathbf{R}_i + \mathbf{Z}_i\mathbf{G}_i\mathbf{Z}_i^T.$$

- For example, \mathbf{R}_i might describe relatedness of subjects in family i – which has an impact even after accounting for covariates (\mathbf{X}) and family-specific responses to covariates (\mathbf{Z})
- If \mathbf{X} and \mathbf{Z} do share covariates, notice the induced heteroskedasticity in the marginal moments, wrt these covariates
- Weighted least-squares $\hat{\boldsymbol{\beta}}$ with any consistent estimates of unknowns in $\phi, \mathbf{R}, \mathbf{G}$ again provide valid inference for $\boldsymbol{\beta}$

Linear Mixed Models: Normality

Normality is death

Theodor Adorno
German Philosopher, Sociologist, and
Musicologist



*Nobody realizes that some people expend
tremendous energy merely to be normal*

Albert Camus
French Philosopher, Anarchist, Journalist
and Nobel Laureate (Literature)

Linear Mixed Models: Normality

Keeping the same mean and covariances, we now move to fully *parametric* inference. With homoskedastic Normal errors (ϵ_{ij}) and Normal random effects this defines *Linear Mixed Models* (LMMs). But why these assumptions?

- Because we believe that the unmeasured influences on each Y_{ij} and b_i represent the averaging of equally many (roughly) equally-variable and independent quantities, i.e. the CLT applies, in addition to the mean model
- Because MLEs and related methods work well* under moderate departures from these assumptions
- Because they give tractable likelihoods, making theory easier
- Because everyone else does it, and most software is coded that way

* A little bit of non-constant variance never hurt anyone

Ray Carroll



LMMs: using MLEs

The general definition of LMMs – also known as *Normal-Normal* models;

$$\begin{aligned} b_i | \mathbf{X}_i, \mathbf{Z}_i &\stackrel{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{G}(\alpha)) \\ \mathbf{Y}_i | b_i, \mathbf{X}_i, \mathbf{Z}_i &\stackrel{\text{indep}}{\sim} N(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i b_i, \phi \mathbf{R}_i(\alpha)), \end{aligned}$$

where the vector of ‘covariance parameters’ α determine between-cluster variation (\mathbf{G}) and within-cluster variation (\mathbf{R}_i).

Following the earlier results, it follows that

$$\begin{aligned} \mathbb{E}[\mathbf{Y}_i | \mathbf{X}_i, \mathbf{Z}_i] &= \mathbf{X}_i \boldsymbol{\beta} \\ \text{Cov}[\mathbf{Y}_i | \mathbf{X}_i, \mathbf{Z}_i] &= \mathbf{Z}_i \mathbf{G}(\alpha) \mathbf{Z}_i^T + \phi \mathbf{R}_i(\alpha), \end{aligned}$$

and for convenience we define $\Sigma_i = \text{Cov}[\mathbf{Y}_i | \mathbf{X}_i, \mathbf{Z}_i]$.

Through marginalization properties of the multivariate Normal distribution (see Stat 512/513) it follows that $\mathbf{Y}_i | \mathbf{X}_i, \mathbf{Z}_i$ **is also Normally distributed** – a very helpful result.

LMMs: using MLEs

First, it means that the likelihood of random variables we have observed (i.e. the \mathbf{Y}_i , not the \mathbf{b}_i) is

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) = - \sum_{i=1}^n \log |\Sigma_i(\boldsymbol{\alpha}, \phi)| - \sum_{i=1}^n (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta})^T \Sigma_i(\boldsymbol{\alpha}, \phi)^{-1} (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta})$$

...plus an ignorable constant.

For fixed $\boldsymbol{\alpha}, \phi$, this likelihood can be maximized in a familiar way. Setting the score function (i.e. $\frac{\partial l}{\partial \boldsymbol{\beta}}$) to zero, we obtain

$$\hat{\boldsymbol{\beta}} | \boldsymbol{\alpha}, \phi = \left(\sum_{i=1}^n \mathbf{X}_i^T \Sigma_i(\boldsymbol{\alpha}, \phi)^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^n \mathbf{X}_i^T \Sigma_i(\boldsymbol{\alpha}, \phi)^{-1} \mathbf{Y}_i \right)$$

...i.e. a weighted least-squares estimate, with weights for cluster i proportional to the inverse of the assumed-known within-cluster covariances Σ_i . Notice that $\hat{\boldsymbol{\beta}}$ is unbiased and also Normally-distributed.

LMMs: using MLEs

Second, because of the $\mathbf{A}^{-1}\mathbf{B}$ cancellation we saw on 3.12, the covariance of $\hat{\beta}$ is

$$\text{Var}[\hat{\beta}|\mathbf{X}, \mathbf{Z}] = \left(\sum_{i=1}^n \mathbf{X}_i^T \Sigma_i(\alpha, \phi)^{-1} \mathbf{X}_i \right)^{-1}.$$

Under the Normal-Normal assumptions and with known α, ϕ , the earlier large-sample results for inference are therefore **exact results** – i.e. ‘95% confidence intervals’ will actually cover the true β in **exactly** 95% of replicate studies.

For testing (Wald or Score or Likelihood Ratio) we similarly exploit

$$(\hat{\beta} - \beta)^T \text{Var}[\hat{\beta}|\mathbf{X}, \mathbf{Z}]^{-1} (\hat{\beta} - \beta) \sim \chi_p^2,$$

to get well-calibrated p -values assessing null hypotheses involving β , for any n . Optimality properties hold here too – e.g. UMPness. (For hypotheses on linear functions of β , replace β by $\mathbf{Q}\beta$, and degrees of freedom p by the rank of \mathbf{Q} .)

LMMs: using MLEs

Third, with unknown α, ϕ , the Fisher information is

$$\begin{aligned}\mathcal{I}_n(\beta, \alpha, \phi) &= -\mathbb{E} \left[\frac{\partial^2}{\partial \beta \partial \beta^T} l(\beta, \alpha, \phi) \right] \\ &= \sum_{i=1}^n \mathbf{X}_i \Sigma_i(\alpha, \phi)^{-1} \mathbf{X}_i^T \equiv n \mathbf{A}.\end{aligned}$$

So under these assumptions, the MLE $\hat{\beta}$ meets the Crámer-Rao lower bound and is the minimum variance unbiased estimate. Less formally, it will be hard for any well-behaved $\hat{\beta}$ to do better than the MLE at many values of β .

All 3 results **only** follow **if** we assume Normality. Why do so?

Everybody believes in the Normal law of errors: the experimenters, because they think it can be proved by mathematics; and the mathematicians, because they believe it has been established by experimental observation.

Gabriel Lippman, French Physicist
...writing to Poincaré (1854–1912)

LMMs: using MLEs

With unknown α, ϕ , having assumed the Normal-Normal model, it's still reasonable to use MLEs to estimate these as well as β , but pretty optimality results are not available. The additional terms in the score function are

$$\frac{\partial l}{\partial \alpha} = \sum_{i=1}^n (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta})^T \frac{\partial \Sigma_i^{-1}}{\partial \alpha} (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta}) - \text{tr} \left(\Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \alpha} \right).$$

- The derivative of a matrix with respect to a vector is a tensor
- To solve for the MLE, iterate between setting this expression to zero and calculating $\hat{\boldsymbol{\beta}}|\hat{\alpha}, \hat{\phi}$ from the weighted least squares formula given earlier. Keep going until convergence
- The $\Sigma_i(\hat{\alpha})$ must remain positive definite, i.e. ‘inside the teabag’. For the single- α choices we saw in GEE this is not a big problem. Keen people: for more general structures there are several parameterizations that maintain positive-definiteness; the Cholesky decomposition gives one popular parameterization

LMMs: dental growth!

The `nlme` package's `lme()` function implements maximum likelihood, for a wide range of LMMs. (See also `lmer()` in `lme4`). The `nlme` package also contains the `Orthodont` data;

```
library("nlme")
d4 <- data.frame(Orthodont) # the dental data
d4$id   <- d4$Subject
d4$male <- d4$Sex=="Male"    # relabeled, as before

#random intercepts
lme1 <- lme(fixed=distance~I(age-8)*male,
            random=~1|id,
            method="ML", data=d4)

#random intercepts and slopes (unstructured G)
lme2 <- lme(fixed=distance~I(age-8)*male,
            random=~I(age-8)|id,
            method="ML", data=d4)

#random intercepts and slopes (diagonal G)
lme2a <- lme(fixed=distance~I(age-8)*male,
              random=reStruct(~I(age-8)|id, pdClass="pdDiag"),
              method="ML", data=d4)
```

LMMs: dental growth!

Key bits of output; (random intercepts only)

```
> summary(lme1)
Linear mixed-effects model fit by maximum likelihood
Random effects:
Formula: ~1 | id
            (Intercept) Residual
StdDev:    1.740851 1.369159 # Estimates sigma_b, sigma_Y

Fixed effects: distance ~ I(age - 8) * male
                Value Std.Error DF  t-value p-value
(Intercept)      21.209091 0.6402989 79 33.12374 0.0000
I(age - 8)       0.479545 0.0940671 79  5.09791 0.0000
maleTRUE         1.406534 0.8317727 25  1.69101 0.1033
I(age - 8):maleTRUE 0.304830 0.1221968 79  2.49458 0.0147

Correlation:
              (Intr) I(g-8) mlTRUE
I(age - 8)   -0.441
maleTRUE      -0.770  0.339 # estimates Corr[\hat\beta]
I(age - 8):maleTRUE 0.339 -0.770 -0.441

> 1.741^2/(1.741^2 + 1.369^2)
[1] 0.6179269 #estimated intra-cluster correlation
```

LMMs: dental growth!

Interpretation of the output;

- Intercept: mean distance (in mm) for average children ($b_i = 0$) who happen to be female, at age 8
- The $I(\text{age} - 8)$ coefficient (0.48 mm/year) and 95% interval (0.30, 0.66) provides our estimate of the difference in expected distance, per additional year of age, in measurements **in the same child**, who happens to be female
- The $I(\text{age} - 8):\text{maleTRUE}$ coefficient is our estimate of the difference in ‘slope’ above, comparing boys to girls
- The `maleTRUE` coefficient is difficult to interpret as a within-person difference in mean (why?) but *can* be interpreted as the difference in the expected outcome, comparing 8-year old boys to girls assuming both are ‘average’ (i.e. $b_i = 0$)
- We estimate that, for observations in the same child and given age and sex, outcomes have correlation 0.62.

LMMs: dental growth!

Key bits of output; (random slopes and intercepts)

```
> summary(lme2)
Linear mixed-effects model fit by maximum likelihood
```

Random effects:

```
Formula: ~I(age - 8) | id
Structure: General positive-definite, Log-Cholesky parametrization
           StdDev   Corr
(Intercept) 1.7045122 (Intr)
I(age - 8)  0.1541351 -0.031
Residual    1.3100447 # was 1.369
```

Fixed effects: distance ~ I(age - 8) * male

	Value	Std.Error	DF	t-value	p-value
(Intercept)	21.209091	0.6226530	79	34.06246	0.0000
I(age - 8)	0.479545	0.1017049	79	4.71507	0.0000
maleTRUE	1.406534	0.8088500	25	1.73893	0.0943
I(age - 8):maleTRUE	0.304830	0.1321185	79	2.30724	0.0237

To dig out relevant parts of output, use `fixef()`, `vcov()`, `VarCorr()`, and the `lmeObject` help page.

LMMs: dental growth!

Key bits of output; (indep't random slopes and intercepts)

```
> summary(lme2a)
Linear mixed-effects model fit by maximum likelihood
```

Random effects:

```
Formula: ~I(age - 8) | id
Structure: Diagonal
    (Intercept) I(age - 8) Residual
StdDev:     1.694440  0.1504866 1.312446
```

Fixed effects: distance ~ I(age - 8) * male

	Value	Std.Error	DF	t-value	p-value
(Intercept)	21.209091	0.6203879	79	34.18682	0.0000
I(age - 8)	0.479545	0.1013345	79	4.73230	0.0000
maleTRUE	1.406534	0.8059075	25	1.74528	0.0932
I(age - 8):maleTRUE	0.304830	0.1316374	79	2.31568	0.0232

... note the *very* slight changes in Std.Error

LMMs: dental growth!

The fitted covariances of the random effects are not displayed very clearly in the `summary`. Check you know where each of these terms comes from;

- Random intercepts alone (`lme1`)

$$\text{Var}[b_i] = 1.74^2$$

- ‘Unstructured’ Random intercepts, slopes (`lme2`)

$$\text{Var}[\{b_{0i}, b_{1i}\}] = \begin{pmatrix} 1.70^2 & -0.031 \times 1.70 \times 0.15 \\ & 0.15^2 \end{pmatrix}$$

- Independent random intercepts, slopes (`lme2a`)

$$\text{Var}[\{b_{0i}, b_{1i}\}] = \begin{pmatrix} 1.69^2 & 0 \\ & 0.15^2 \end{pmatrix}$$

LMMs: dental growth!

With random intercepts only, the intraclass correlation tells us what proportion of the overall variability in all observations \mathbf{Y} can be explained by cluster-specific effects (and not noise).

Otherwise, there is no standard summary. However, it's usually sensible to interpret these variance terms by comparing them to the size of the $\hat{\beta}$;

- Mild-to-large between-subject variability in dental length: compare e.g. $\hat{\sigma}_b=1.74$ to the age effect(s), also the effect of being male
- For random slopes (`lme2` and `lme2a`) the between-subject variability in age effect seems mild: compare 0.15 to the main effect of age (0.48 mm/year) also the difference in main effect of age seen in males versus females

LMMs: lme() options

How these match up to our notation;

- **fixed**: fixed effects (\mathbf{X}_i) – via usual formula syntax
- **random**: random effects (\mathbf{Z}_i) – formula, or formula within a call to `reStruct()` that specifies `pdClass` (\mathbf{G})
- **correlation**: serial dependence (\mathbf{R}_i)
- **weights**: for observation-specific terms in log-likelihood

Among the options for **correlation**; (or write your own)

- **NULL**: independence (the default)
- **corAR1**: autoregressive process of order 1
- **corCompSymm**: compound symmetry structure
- **corSymm**: no additional structure

And for `pdClass` in `reStruct()`; (or again, write your own)

- **pdSymm**: no structure (the default)
- **pdDiag**: diagonal, i.e. independence
- **pdIdent**: multiple of an identity
- **pdCompSymm**: compound symmetry

LMMs: not quite like other MLEs

Basically, the elegant theory of MLEs (familiar from 570) applies to LMMs, just like any other parametric model. However, we will consider some obstacles LMMs present;

- Estimating scale parameters (e.g. σ_Y^2 or $\phi\mathbf{R}$) through MLEs is biased, and we get **slightly** better estimates (REML) by tweaking the likelihood-maximization
- Unlike models seen in 570, in LMMs data does not have to be *pathological* (i.e. weird and implausible) for estimates (e.g. $\hat{\mathbf{R}}$) to end up on the boundary of the parameter space. This affects the reference distribution for large-sample tests of some hypotheses
- How to do inference on the b_i ? These are random – not fixed unknowns – so e.g. likelihood theory doesn't apply
- Diagnostics get a little trickier; is it enough to check that $Y_i - \mathbf{X}_i\hat{\beta}$ follows the right distribution ($\phi\mathbf{R}_i, \mathbf{G}$) or should we look specifically for $b_i \sim N(0, \mathbf{G})$? How to do this when we didn't even observe the b_i ?

LMMs: REML estimates

In classical linear models, you should be aware (from e.g. 533) that the MLE's dispersion estimate is biased downwards;

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n} \chi_{n-p}^2$$

A similar property holds for LMMs, which we illustrate for one special case. For LMMs where every cluster has the same \mathbf{X}_i , and (common) Σ_i is unstructured, the MLE $\hat{\beta}$ is the usual OLS estimate of terms in the mean model, here written

$$\hat{\beta} = \frac{1}{n} (\mathbf{X}_i^T \mathbf{X}_i)^{-1} \mathbf{X}_i^T \sum_{i=1}^n \mathbf{Y}_i$$

and the (consistent) ML estimate of Σ is

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (\mathbf{Y}_i - \mathbf{X}_i \hat{\beta}) (\mathbf{Y}_i - \mathbf{X}_i \hat{\beta})^T,$$

... which is also a natural 'plug-in' estimate.

LMMs: REML estimates

The expectation of this estimate is

$$\begin{aligned}\mathbb{E}[\widehat{\Sigma}] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(\mathbf{Y}_i - \mathbf{X}_i \widehat{\boldsymbol{\beta}})(\mathbf{Y}_i - \mathbf{X}_i \widehat{\boldsymbol{\beta}})^T] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(\mathbf{Y}_i - \boldsymbol{\mu}_i + \boldsymbol{\mu}_i - \mathbf{X}_i \widehat{\boldsymbol{\beta}})(\mathbf{Y}_i - \boldsymbol{\mu}_i + \boldsymbol{\mu}_i - \mathbf{X}_i \widehat{\boldsymbol{\beta}})^T] \\ &= \frac{1}{n} \sum_{i=1}^n \boldsymbol{\Sigma} - \text{Cov}[\mathbf{Y}_i, \widehat{\boldsymbol{\beta}}^T] \mathbf{X}_i^T - \mathbf{X}_i \text{Cov}[\widehat{\boldsymbol{\beta}}, \mathbf{Y}_i^T] + \mathbf{X}_i \text{Var}[\widehat{\boldsymbol{\beta}}] \mathbf{X}_i^T \\ &= \boldsymbol{\Sigma} - \sum \mathbf{h}_i / n - \mathbf{h}_i \boldsymbol{\Sigma} / n + \mathbf{h}_i \boldsymbol{\Sigma} \mathbf{h}_i / n \\ &= \frac{n-1}{n} \boldsymbol{\Sigma} + \frac{1}{n} (\mathbf{I} - \mathbf{h}_i) \boldsymbol{\Sigma} (\mathbf{I} - \mathbf{h}_i),\end{aligned}$$

where \mathbf{h}_i is the hat matrix $\mathbf{X}_i(\mathbf{X}_i^T \mathbf{X}_i)^{-1} \mathbf{X}_i^T$, identical across all clusters.

We see that $\widehat{\boldsymbol{\Sigma}}$ is biased ‘too small’; for any vector \mathbf{u} in the null space of \mathbf{h}_i , we get $\mathbb{E}[\mathbf{u}^T \widehat{\boldsymbol{\Sigma}} \mathbf{u}] = \frac{n-1}{n} \mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u}$. This is a space of dimension $n_i - p$, so the problem is worse with larger p .

LMMs: REML estimates

In general, ML estimates of dispersion are biased small;

- We are not accounting for the fact that the elements of $\hat{\beta}$ are estimated, not known; the dispersion of points around fitted mean responses tends to be smaller than dispersion around true mean responses
- Too-small variances lead to ‘pointier’ likelihoods, with higher maxima
- When each n_i is large (compared to p) the problem is less severe – but the design matrices \mathbf{x}_i also play a role
- Increasing n reduces the bias problem – provided we keep the dimension of β and parameters in Σ_i fixed (recall Neyman-Scott) and we don’t shrink the n_i .

Keen people: We *can* get consistency when p grows with n , but only when it grows very slowly, e.g. $p \approx \log(n)$. See the 580s.

LMMs: REML estimates

Restricted (or Residual) Maximum Likelihood (REML) estimation provides less-biased estimators of the dispersion, in LMMs.

We use ‘stacked’ notation, where the model states that (marginally)

$$\mathbf{Y} \sim N(\mathbf{x}\boldsymbol{\beta}, \mathbf{V})$$

for $(\sum_{i=1}^n n_i) \times p$ ‘stacked’ design matrix \mathbf{x} , corresponding stacked vector \mathbf{Y} , and where \mathbf{V} is block diagonal with entries Σ_i . Writing the ‘hat’ matrix

$$\mathbf{H} = \mathbf{x}(\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T,$$

i.e. projection onto the space spanned by columns of \mathbf{x} , we can define orthonormal B of dimension $(\sum_{i=1}^n n_i) \times (\sum_{i=1}^n n_i) - p$ such that

$$\begin{aligned} B^T B &= I \\ BB^T &= I - \mathbf{H}. \end{aligned}$$

LMMs: REML estimates

In this stacked notation, the MLE for β is

$$\hat{\beta} = (\mathbf{x}^T \mathbf{V}^{-1} \mathbf{x})^{-1} \mathbf{x}^T \mathbf{V}^{-1} \mathbf{Y}$$

which as we know is distributed as

$$\hat{\beta} \sim N\left(\beta, (\mathbf{x}^T \mathbf{V}^{-1} \mathbf{x})^{-1}\right).$$

To describe the rest of the data, consider the OLS residuals

$$\mathbf{e}_{OLS} = (I - \mathbf{H})\mathbf{Y} = BB^T\mathbf{Y}$$

'Half-way' to these residuals is $B^T\mathbf{Y}$ – and this multiple of \mathbf{Y} usefully inherits $(I - \mathbf{H})$'s orthogonality to \mathbf{x} – note that

$$\begin{aligned} B^T \mathbf{x} &= B^T BB^T \mathbf{x} \\ &= B^T (I - \mathbf{H}) \mathbf{x} \\ &= B^T \mathbf{0} = \mathbf{0}. \end{aligned}$$

LMMs: REML estimates

Exploiting this property, we find that

$$\begin{aligned}\mathbb{E}[B^T \mathbf{Y}] &= B^T \mathbf{x} \boldsymbol{\beta} \\&= 0 \\ \text{Var}[B^T \mathbf{Y}] &= B^T \mathbf{V} B \\ \text{Cov}[B^T \mathbf{Y}, \hat{\boldsymbol{\beta}}] &= \text{Cov}[B^T \mathbf{Y}, (\mathbf{x}^T \mathbf{V}^{-1} \mathbf{x})^{-1} \mathbf{x}^T \mathbf{V}^{-1} \mathbf{Y}] \\&= B^T \mathbf{V} \mathbf{V}^{-1} \mathbf{x} (\mathbf{x}^T \mathbf{V}^{-1} \mathbf{x})^{-1} \\&= B^T \mathbf{x} (\mathbf{x}^T \mathbf{V}^{-1} \mathbf{x})^{-1} \\&= 0\end{aligned}$$

This means;

- $\hat{\boldsymbol{\beta}}$ and $B^T \mathbf{Y}$ partition \mathbf{Y} into two parts, of length p and $(\sum_{i=1}^n n_i) - p$
- Each part's covariance is of full rank, and the two parts are independent
- The $B^T \mathbf{Y}$ terms have mean zero and a variance that depends only on Σ

LMMs: REML estimates

This suggests estimating Σ by finding the MLE of the *restricted* or *residual* likelihood corresponding to just the $B^T \mathbf{Y}$ terms, using this $\hat{\Sigma}$ to maximize with respect to $\hat{\beta}$, and iterating until convergence. This approach is known as REML.

These definitions give enough information to do this, but further simplification can be done. Ignoring a multiple (a Jacobian) that only depends on \mathbf{x} , the likelihood of $B^T \mathbf{Y}$ can be written

$$\frac{(2\pi)^{-\frac{1}{2}\sum_{i=1}^n n_i} |\mathbf{V}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{Y}-\mathbf{x}\beta)^T \mathbf{V}^{-1} (\mathbf{Y}-\mathbf{x}\beta)}}{(2\pi)^{-\frac{1}{2}((\sum_{i=1}^n n_i)-p)} \left|(\mathbf{x}^T \mathbf{V}^{-1} \mathbf{x})^{-1}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\hat{\beta}-\beta)^T \mathbf{x}^T \mathbf{V}^{-1} \mathbf{x}(\hat{\beta}-\beta)}}.$$

Ignoring terms free of \mathbf{V} , and exploiting

$$(\mathbf{Y}-\mathbf{x}\beta)^T \mathbf{V}^{-1} (\mathbf{Y}-\mathbf{x}\beta) = (\mathbf{Y}-\mathbf{x}\hat{\beta})^T \mathbf{V}^{-1} (\mathbf{Y}-\mathbf{x}\hat{\beta}) + (\hat{\beta}-\beta)^T \mathbf{x}^T \mathbf{V}^{-1} \mathbf{x}(\hat{\beta}-\beta),$$

the relevant log-likelihood is then

$$-\frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} \log |\mathbf{x}^T \mathbf{V}^{-1} \mathbf{x}| - \frac{1}{2} (\mathbf{Y} - \mathbf{x}\hat{\beta})^T \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{x}\hat{\beta}).$$

LMMs: REML estimates

Compare this to what we maximize for the MLE for \mathbf{V} terms;

$$\begin{aligned}\text{REML: } & -\frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} \log |\mathbf{x}^T \mathbf{V}^{-1} \mathbf{x}| - \frac{1}{2} (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\beta}})^T \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\beta}}) \\ \text{MLE: } & -\frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\beta}})^T \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\beta}})\end{aligned}$$

- REML adds a term penalizing small values of \mathbf{V} – the log-determinant of a $p \times p$ matrix
- This term will usually be tiny compared to $\log |\mathbf{V}|$ since \mathbf{V} has $\sum_{i=1}^n n_i$ rows and columns
- But having p large relative to n is exactly when we worry about overfitting and too-small $\hat{\mathbf{V}}$ – recall Neymann-Scott
- So expect to see little difference – but where there **is** a difference, REML is preferred

If our rather *ad hoc* use of $B^T \mathbf{Y}$ doesn't appeal, a conditional likelihood interpretation exists (Smyth and Verbyla 1996) and a Bayesian interpretation (Harville 1974) and in some circumstances one via ANOVA (Verbeke and Molenberghs 2000, pg 46) ... among others.

LMMs: using REML in lme()

REML is the default in `lme()` – but it's still good practice to mention you used REML. Here, we use the example model from slide 3.29;

```
library("nlme")
d4 <- data.frame(Orthodont) # the dental data, again
d4$id   <- d4$Subject
d4$male <- d4$Sex=="Male"   # usual relabeling
rem11 <- lme(fixed=distance~I(age-8)*male,
             random=~1|id, data=d4)      # no method="ML" this time
> rem11
Linear mixed-effects model fit by REML # default
```

In this example, results are very similar to use of ML;

	Int	I(age-8)	male	I(age-8):male	σ_Y
ML	21.2 (0.63)	0.48 (0.092)	1.41 (0.82)	0.31 (0.120)	1.37
REML	21.2 (0.65)	0.48 (0.094)	1.41 (0.84)	0.31 (0.121)	1.38

REML is also the default in e.g. PROC MIXED (SAS) and most other widely-used software.

LMMs: using REML in `lme()`

- REML is viewed primarily as a method for getting better standard error estimates. If variance \mathbf{V} does not affect $\hat{\beta}$ (e.g. independence, exchangeable under balanced designs) then $\hat{\beta}_{MLE} = \hat{\beta}_{REML}$
- REML estimates of variance terms are biased – but less biased than the MLE, typically
- Using REML, Wald tests of β are fine – see `lme()` output for examples. But do say explicitly you are using “Wald tests”
- Why? As REML doesn’t use a full likelihood, LR tests using it may not be valid. (But for tests of variance parameters, LR tests are valid, see Jon’s Book, §8.5.3)
- This is easy to forget, because the `anova()` method for `lme()` output doesn’t warn you. But as `?anova.lme` does state, “likelihood comparisons are not meaningful for objects fit using REML with different fixed effects”

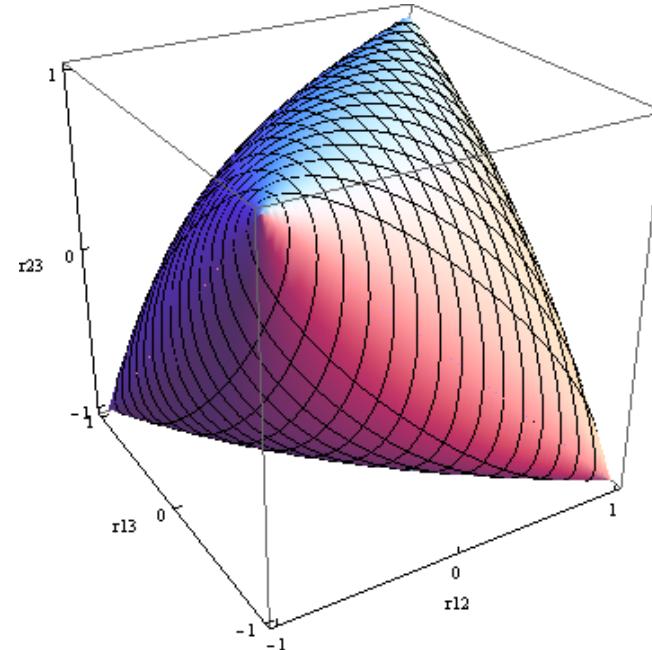
LMMs: boundary-valued estimates

It's easy to overlook that LMMs impose non-trivial constraints on the possible Σ_i , the covariance within a cluster. For example, with just random intercepts (e.g. slide 3.8) then the correlation any pair of observations in a given cluster is

$$\text{Corr}[Y_{ij}, Y_{ij'}|\mathbf{X}_i] = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_Y^2} \equiv \alpha.$$

Q1. Where in the teabag (right) does this correlation matrix have to lie?

Q2. Do data (i.e. sample correlations) have the same constraints?



LMMs: boundary-valued estimates

To see the effect on $\hat{\alpha} = \hat{\sigma}_b^2 / (\hat{\sigma}_b^2 + \hat{\sigma}_Y^2)$, we generate data from

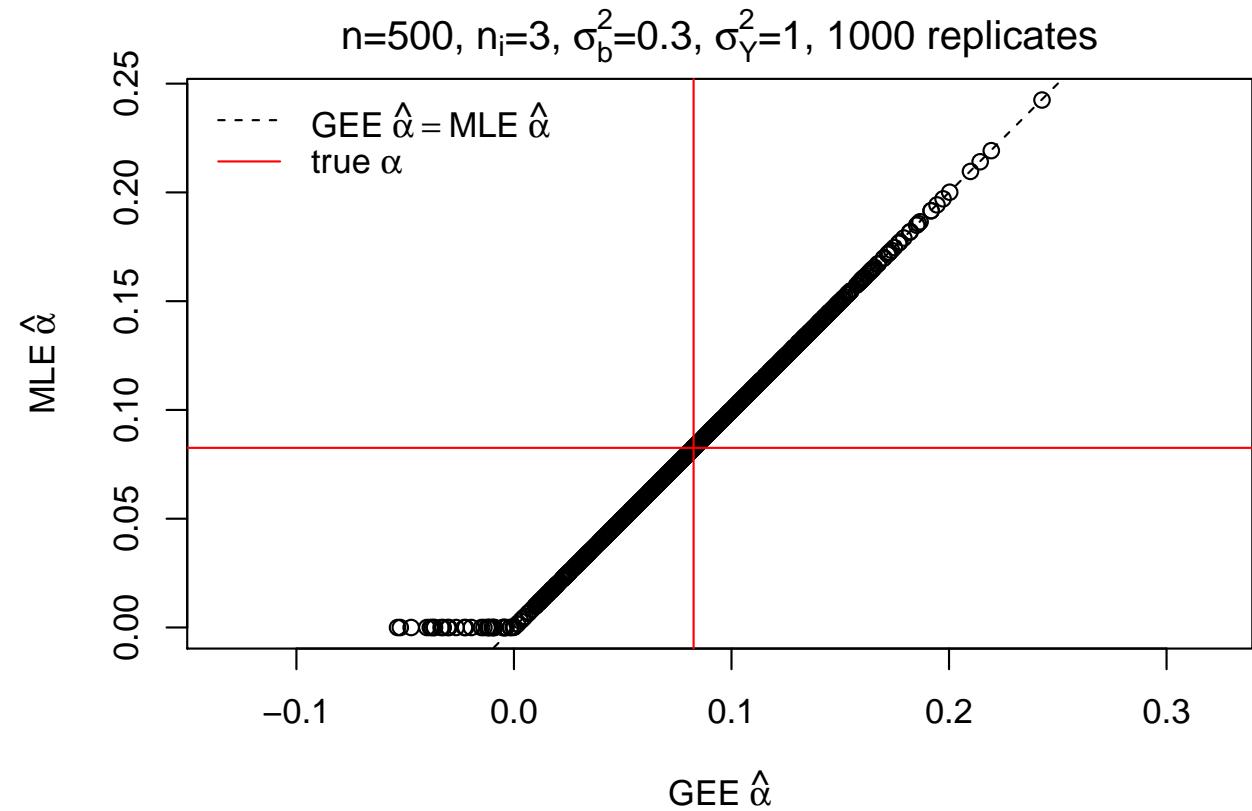
$$b_i \sim N(0, \sigma_b^2), \quad Y_{ij}|b_i \sim N(b_i, \sigma_Y^2),$$

and regress on just an intercept in two ways;

MLE: `lm(y~1, random=~1|id, ...)`

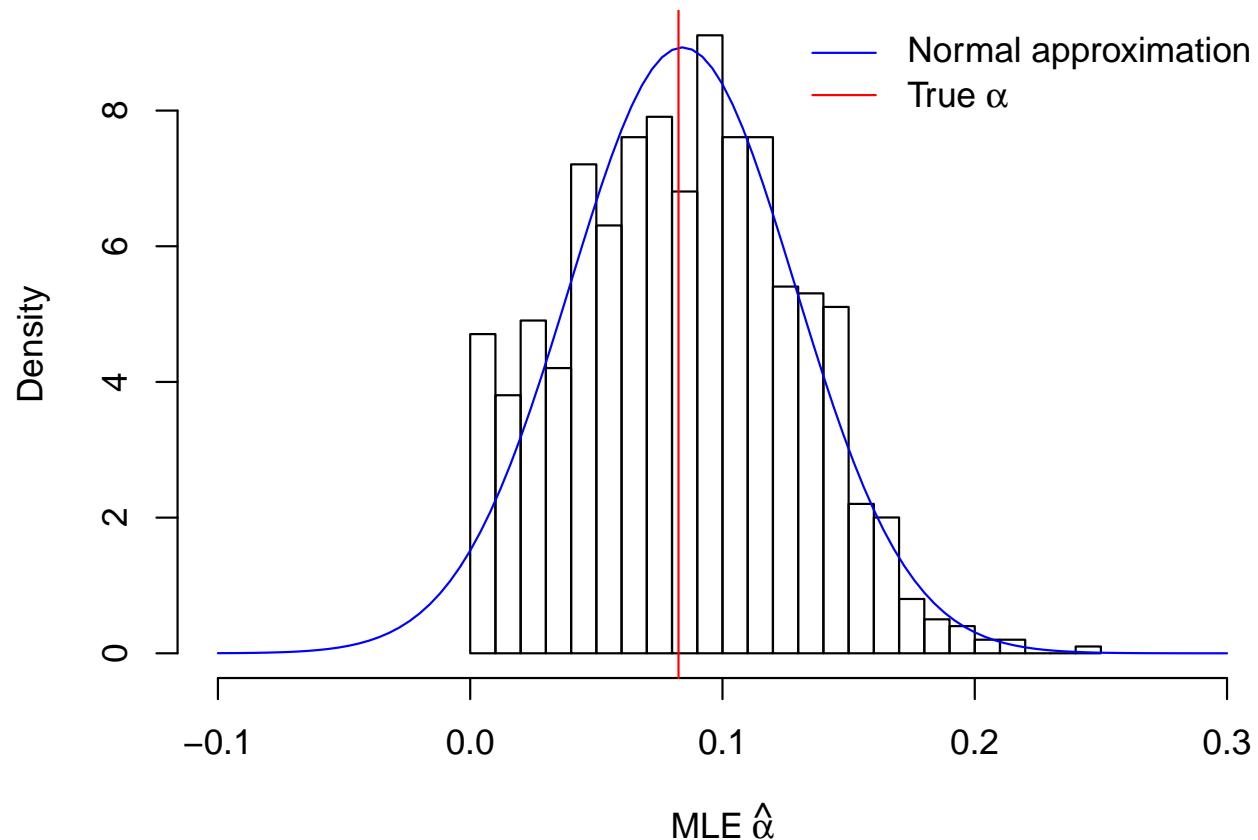
GEE: `gee(y~1, id=id, family="exchangeable", ...)`

For almost all datasets the methods agree, but when data suggest $\alpha < 0$ then the MLE goes to the boundary;



LMMs: boundary-valued estimates

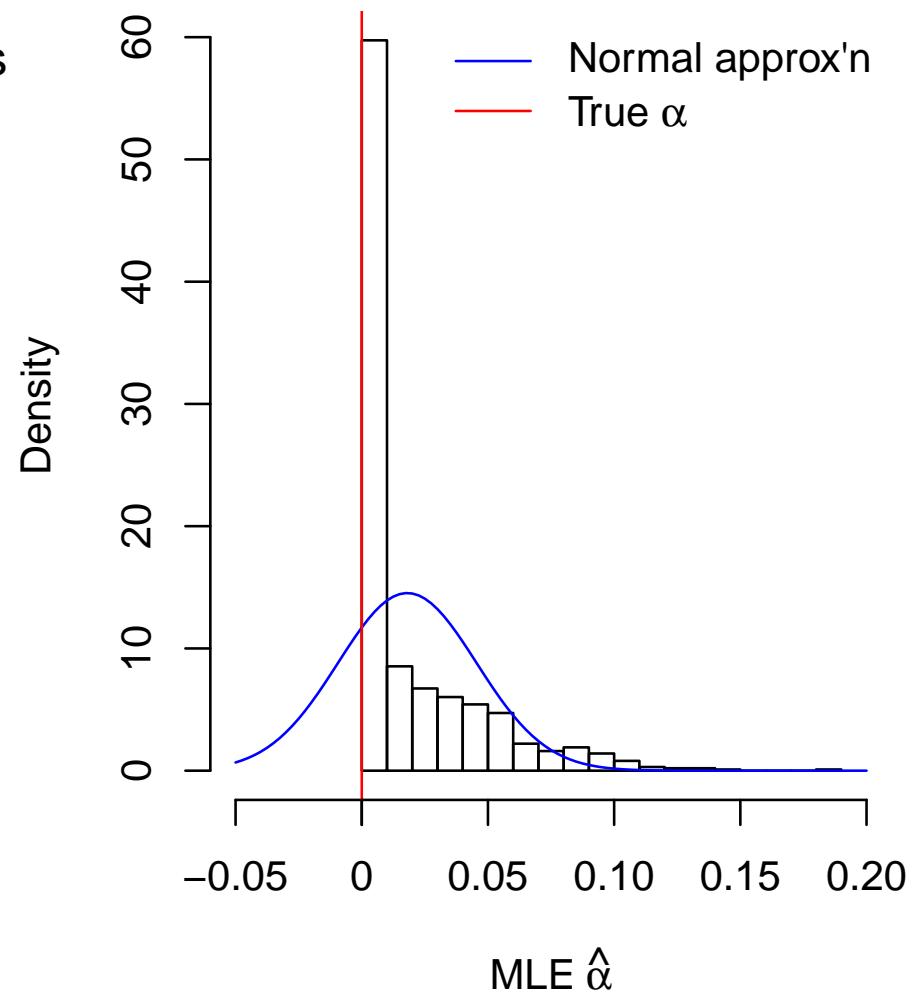
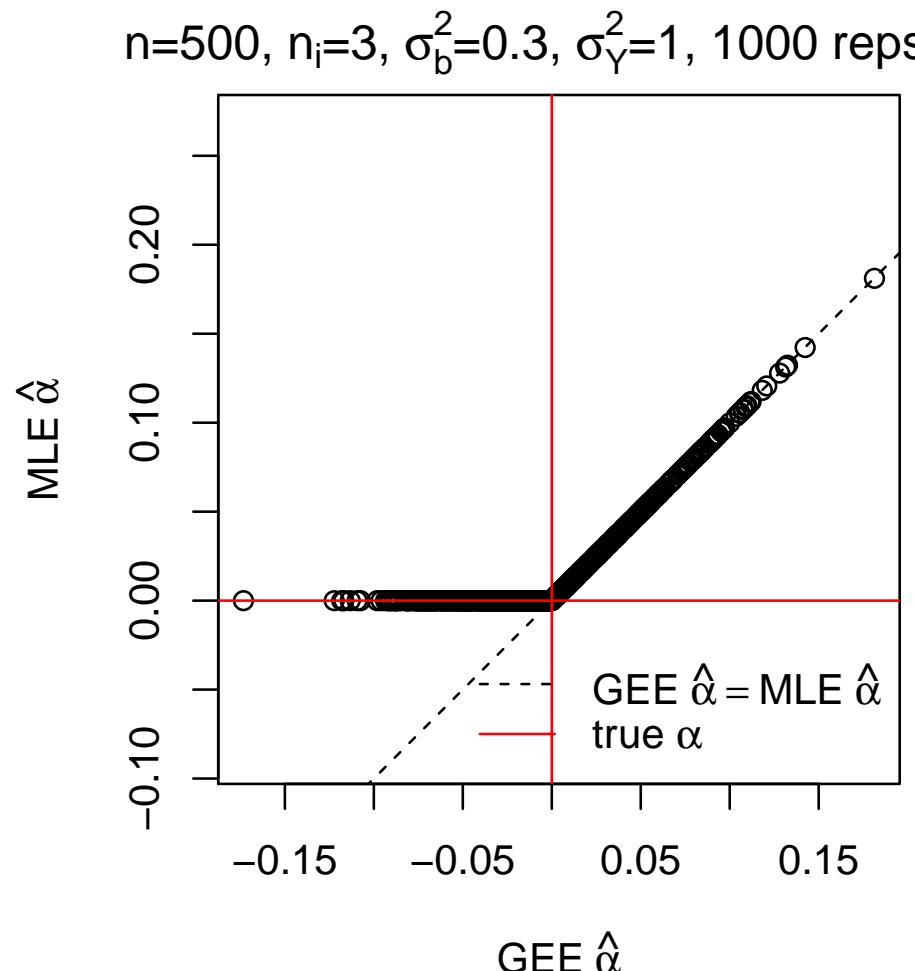
Looking at just the MLEs for $\alpha = \sigma_b^2 / (\sigma_b^2 + \sigma_Y^2)$;



For $\sigma_b^2 > 0$, i.e. $\alpha > 0$, confidence intervals built around Normality require large n to work well – to make the ‘spike’ at zero negligible.

LMMs: boundary-valued estimates

Now the same pictures but with $\sigma_b^2 = 0$, i.e. no clustering – but the analysis doesn't know this;



LMMs: boundary-valued estimates

Suppose we were testing $H_0 : \alpha = 0$, i.e. a null hypothesis about the variance components. Unlike tests on mean parameters (β), **under the null** and **even with large n** , asymptotic Normality of $\hat{\alpha}$ does **not** hold. (This is true for MLE and REML)

- The same property affects LR test statistics, which do not follow χ^2_1 under the null, even with n large
- The appropriate reference distribution (here) for the LR test is actually a 50:50 mix of χ^2_0 and χ^2_1
- Some software (Stata) can spot this, in some cases, but R does not. Instead, it just uses χ^2_1 as the reference distribution – a conservative approximation
- Just fitting the model can be a numerical challenge, when $\hat{\sigma}_b^2$ is on the boundary; Newton-Raphson algorithms may need ‘tweaking’ (e.g. step-halving). `lme()` and later `lmer()` are good, but nothing is foolproof, or can be foolproof. Try different start values in `lmer()`.

LMMs: boundary-valued estimates

With multiple variance components, things get even messier;
consider having random slopes and intercepts;

$$Y_{ij}|X_{ij}, b_{0i}, b_{1i} \sim N(\beta_0 + b_{0i} + \beta_1 X_{ij} + b_{1i} X_{ij}, \sigma_Y^2)$$
$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{b0}^2 & r\sigma_{b0}\sigma_{b1} \\ r\sigma_{b0}\sigma_{b1} & \sigma_{b1}^2 \end{pmatrix}\right)$$

... and the MLE fits $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_Y, \hat{\sigma}_{b0}, \hat{r}, \hat{\sigma}_{b1}$.

- **Q.** If testing $\sigma_{b1} = 0$, i.e. that everyone has the same slope, what other parameter would be irrelevant *under the null hypothesis*?
- Should *really* compare the likelihood ratio statistic (from our data) to a mixture* of $\chi_0^2, \chi_1^2, \chi_2^2$ – that we'd see, repeating the experiment under the null

*a hard-to-calculate, hard-to-explain, unattractive mixture

LMMs: boundary-valued estimates

From Stata's documentation; (very good on this topic)

Because of these complications, appropriate and sufficiently general computation methods for the more-than-one-parameter case have yet to be developed. Theory has demonstrated that, whatever the distribution of the LR test statistic, its tail probabilities are bound above by those of the chi-squared distribution with degrees of freedom equal to the full number of restricted parameters.

- It's standard to compute p -values come from this conservative approach
- If the p -value is overwhelmingly tiny, conservatism doesn't matter. If the p -value is very large, it's also unlikely to matter
- But otherwise, there can be considerable differences
- If you really need this test, non-trivial math awaits

LMMs: London! – boundary MLEs

We return to the London schools example, in which the religious denomination (if any) of the schools is recorded;

Denomination denom	Ch. of England CofE	Roman Catholic RC	None
# schools	3	9	26
# students	398	120	1460

- The students per school differs somewhat by `denom` – and plausibly by other characteristics do too
- ...so the extent to which unmeasured factors affect outcomes Y may differ by `denom`. This suggests fitting a different σ_b for each value of $D=\text{denom}$
- In the usual notation;

$$\begin{aligned} b_i | D_i &\stackrel{i.i.d.}{\sim} N(\mathbf{0}, \sigma_{bD_i}^2) \\ Y_{ij} | b_i, \mathbf{X}_{ij} &\stackrel{\text{indep}}{\sim} N(\mathbf{X}_{ij}^T \boldsymbol{\beta} + b_i, \sigma_Y^2). \end{aligned}$$

LMMs: London! – boundary MLEs

Fitting this with `lme()`, the formula syntax is used – and by default, it includes an intercept;

```
> lme1 <- lme(Y~LRT+ male, random=reStruct(~denom|school, pdClass="pdDiag"),
               data=schools, method="ML")
> getVarCov(lme1)
Random effects variance covariance matrix
      (Intercept) denomNone denomRC
(Intercept)    0.03278  0.000000   0.00
denomNone      0.00000  0.006829   0.00
denomRC        0.00000  0.000000   0.17
```

- Matrix Z_i is $n_i \times 3$ (see slide 3.23). Its 3 columns are 1_{n_i} , and two indicators for $D_i = \text{None}$, $D_i = \text{RC}$. CofE is the reference group
- The fitted $\hat{\sigma}_{bD}^2$ are 0.033 (CofE), 0.033+0.007 (None), and 0.033+0.17 (RC)
- $\hat{\sigma}_Y^2 = 0.62$, so the estimated ICC is notably larger in RC schools; $\hat{\alpha} = 0.22$ versus almost zero for None, CofE
- Note that the fitted σ_{bD}^2 **must** be larger in the two non-reference groups – they are baseline **plus** a positive term

LMMs: London! – boundary MLEs

Perhaps an easier way to parameterize the problem;

```
> lme2 <- lme(Y~LRT + male, random=reStruct(~denom -1|school, pdClass="pdDiag"),
+                 data=schools, method="ML")
> getVarCov(lme2)
Random effects variance covariance matrix
      denomCofE  denomNone  denomRC
denomCofE  0.03278  0.000000  0.00000
denomNone  0.00000  0.039607  0.00000
denomRC   0.00000  0.000000  0.20276
```

- Now each column of Z_i is an indicator for each level of `denom` – there is no ‘reference’
- The fitted values are identical; there is no restriction on the ordering of the `denom`-specific σ_{bD}^2 ; each is free to take any non-negative value
- Having a rationale for a specific pattern of heteroskedasticity is unusual, so this is typically a good approach

LMMs: London! – boundary MLEs

But either through believing that RC schools **must** have smallest σ_{bD}^2 – or just through careless coding, perhaps you did this;

```
> schools$denom3 <- relevel(schools$denom, "RC")
> lme3 <- lme(Y~LRT + male, random=reStruct(~denom3|school, pdClass="pdDiag"),
+                 data=schools, method="ML")
> getVarCov(lme3)
Random effects variance covariance matrix
      (Intercept) denom3CofE denom3None
(Intercept)   0.04821 0.0000e+00 0.0000e+00
denom3CofE    0.00000 1.0215e-07 0.0000e+00
denom3None    0.00000 0.0000e+00 8.4776e-09
```

- We get two parameters on the boundary of the fitted model
 - the $\hat{\sigma}_{bD}^2$ terms for CofE and None are equivalent to zero, numerically
- Note that the asymptotic behaviour of the $\hat{\sigma}_{bD}^2$ will be sharply different, depending on whether the reference σ_{bD}^2 is bigger/smaller than values for other D
- So **be careful** fitting random slopes. Is the induced pattern of heteroskedasticity sensible? (i.e. $\Sigma_i = \phi \mathbf{R}_i + \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i^T$)

LMMs: London! – boundary MLEs

For completeness; R code to implement LR tests of H_0 : that within-school covariance does **not** differ by denomination;

```
lme0 <- lme(Y~LRT+ male, random=reStruct(~1|school, pdClass="pdDiag"),
              data=schools, method="ML")
lme2 <- lme(Y~LRT+ male, random=reStruct(~denom -1|school, pdClass="pdDiag"),
              data=schools, method="ML")
anova(lme0, lme2)
      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
lme0     1 5 4742.922 4770.871 -2366.461
lme2     2 7 4744.007 4783.136 -2365.004 1 vs 2 2.915013  0.2328
```

- This test uses χ^2_2 as the reference – which is conservative
- Regardless of boundary issues, REML LR tests are only valid if the two models have the same fixed effects

```
schools$denom3 <- relevel(schools$denom, "RC")
lme3 <- lme(fixed=Y~LRT + male, random=reStruct(~denom3|school, pdClass="pdDiag"),
            data=schools, method="ML")
anova(lme0, lme3) # using the highly-constrained parameterization
      Model df      AIC      BIC    logLik   Test      L.Ratio p-value
lme0     1 5 4742.922 4770.871 -2366.461
lme3     2 7 4746.922 4786.051 -2366.461 1 vs 2 1.970366e-06      1 # eek!
```

LMMs: predicting the b_i



Oh! Drumossie, thy bleak moor shall... be stained with the best blood of the Highlands. Heads will be lopped off by the score, and no mercy shall be shown or quarter given on either side.

Prediction by the **Brahan Seer**, in the 16th century
Interpreted as predicting the Battle of Culloden (**1745**)

NB the Seer did not predict his own execution

LMMs: predicting the b_i

Defining random effects is fairly straightforward; they represent e.g. a cluster-specific intercept, that we would know accurately given enough data from that cluster, e.g. enough students from one school.

Frequentist inference on random effects is trickier;

- Typically n_i is small/modest for any cluster, so we can't rely on asymptotic definitions/approximations
- If we re-ran the study*, we'd get different clusters and hence estimate different b_i . Notions of consistency, bias and variance don't automatically apply

Nevertheless, it's natural to want to know about the b_i in specific clusters, e.g. how different is school i from average? Or how might student j in school i perform with covariates \mathbf{X}_{ij} ?

* ...following our classical motivation of the b_i as random draws from a population of cluster effects. Use of random effects models in other situations will be discussed later – see e.g. Hodges & Clayton (2010)

LMMs: predicting the b_i

Of course, the data provide **some** information about the b_i . To formalize what we can say about b_i given observed \mathbf{Y} , first write the LMM as

$$\begin{aligned} b_i | \mathbf{X}_i = \mathbf{x}_i, \mathbf{Z}_i = \mathbf{z}_i &\stackrel{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{G}) \\ \epsilon_i | \mathbf{X}_i = \mathbf{x}_i, \mathbf{Z}_i = \mathbf{z}_i &\stackrel{indep}{\sim} N(\mathbf{0}, \phi \mathbf{R}_i) \\ \mathbf{Y}_i &= \mathbf{x}_i \boldsymbol{\beta} + \mathbf{z}_i b_i + \epsilon_i, \end{aligned}$$

where the \mathbf{G} and \mathbf{R}_i may depend on some parameter(s) α . This implies that the joint distribution of \mathbf{Y}_i and b_i is

$$\begin{pmatrix} \mathbf{Y}_i \\ b_i \end{pmatrix} \middle| \mathbf{X}_i = \mathbf{x}_i, \mathbf{Z}_i = \mathbf{z}_i \stackrel{indep}{\sim} N \left(\begin{bmatrix} \mathbf{x}_i \boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \Sigma_i & \mathbf{z}_i \mathbf{G} \\ \mathbf{G} \mathbf{z}_i^T & \mathbf{G} \end{bmatrix} \right),$$

where $\Sigma_i = \mathbf{z}_i \mathbf{G} \mathbf{z}_i^T + \phi \mathbf{R}_i$ as before. Multivariate Normality implies*

$$b_i | \mathbf{Y}_i, \mathbf{X}_i = \mathbf{x}_i, \mathbf{Z}_i = \mathbf{z}_i \stackrel{ind}{\sim} N \left(\mathbf{G} \mathbf{z}_i^T \Sigma_i^{-1} (\mathbf{Y}_i - \mathbf{x}_i \boldsymbol{\beta}), (\phi^{-1} \mathbf{z}_i^T \mathbf{R}_i^{-1} \mathbf{z}_i + \mathbf{G}^{-1})^{-1} \right).$$

* ... left as an exercise for keen people

LMMs: predicting the b_i

On its own, $\mathbb{E}[b_i | \mathbf{Y}_i]$ is known as the *Best Linear Unbiased Predictor* (BLUP) of b_i – because among linear functions of \mathbf{Y}_i that are unbiased for b_i , it has minimum variance.

As the b_i are random, it is formally called a ‘predictor’ not an ‘estimate’ – see Robinson (1991).

We can’t evaluate the BLUP. But a natural ‘plug-in’ version of it is

$$\tilde{b}_i = \hat{\mathbf{G}}\mathbf{Z}_i^T \hat{\Sigma}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \hat{\beta}),$$

where all the estimated terms use $\hat{\beta}, \hat{\alpha}$ from MLE or REML. Confusingly, this \tilde{b}_i also gets called the BLUP, or BLUP estimate.

LMMs: predicting the b_i

What type of estimate/predictor is \tilde{b}_i ?

- It's biased! (as an estimate of the b_i we actually sampled).
But noting that we can write it as

$$\tilde{b}_i = \hat{G}\mathbf{Z}_i^T(\mathbf{Z}_i\hat{\mathbf{G}}\mathbf{Z}_i^T + \phi\hat{\mathbf{R}}_i)^{-1}(\mathbf{Z}_i b_i + \mathbf{X}_i(\beta - \hat{\beta}) + \epsilon_i),$$

if $\phi\mathbf{R}_i$ is a small component of the inverse term, approximation cancellation occurs and this bias may be small

- It's not an MLE – Neyman-Scott told us that MLEs for b_i are in general a Bad Idea
- It's not fully Bayesian – though it behaves like it; note that using \tilde{b}_i we obtain fitted values

$$\begin{aligned}\hat{\mathbf{Y}}_i &= \mathbf{X}_i\hat{\beta} + \mathbf{Z}_i\tilde{b}_i \\ &= \mathbf{X}_i\hat{\beta} + \mathbf{Z}_i\hat{\mathbf{G}}\mathbf{Z}_i^T\hat{\Sigma}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\hat{\beta}) \\ &= (I - \mathbf{Z}_i\hat{\mathbf{G}}\mathbf{Z}_i^T\hat{\Sigma}_i^{-1})\mathbf{X}_i\hat{\beta} + \mathbf{Z}_i\hat{\mathbf{G}}\mathbf{Z}_i^T\hat{\Sigma}_i^{-1}\mathbf{Y}_i,\end{aligned}$$

i.e. a combination of the estimated population average $\mathbb{E}[\mathbf{Y}_i|\mathbf{X}_i]$ and the observed \mathbf{Y}_i

LMMs: predicting the b_i

The most common way to think about \tilde{b}_i is as an *empirical Bayes* estimate of b_i :

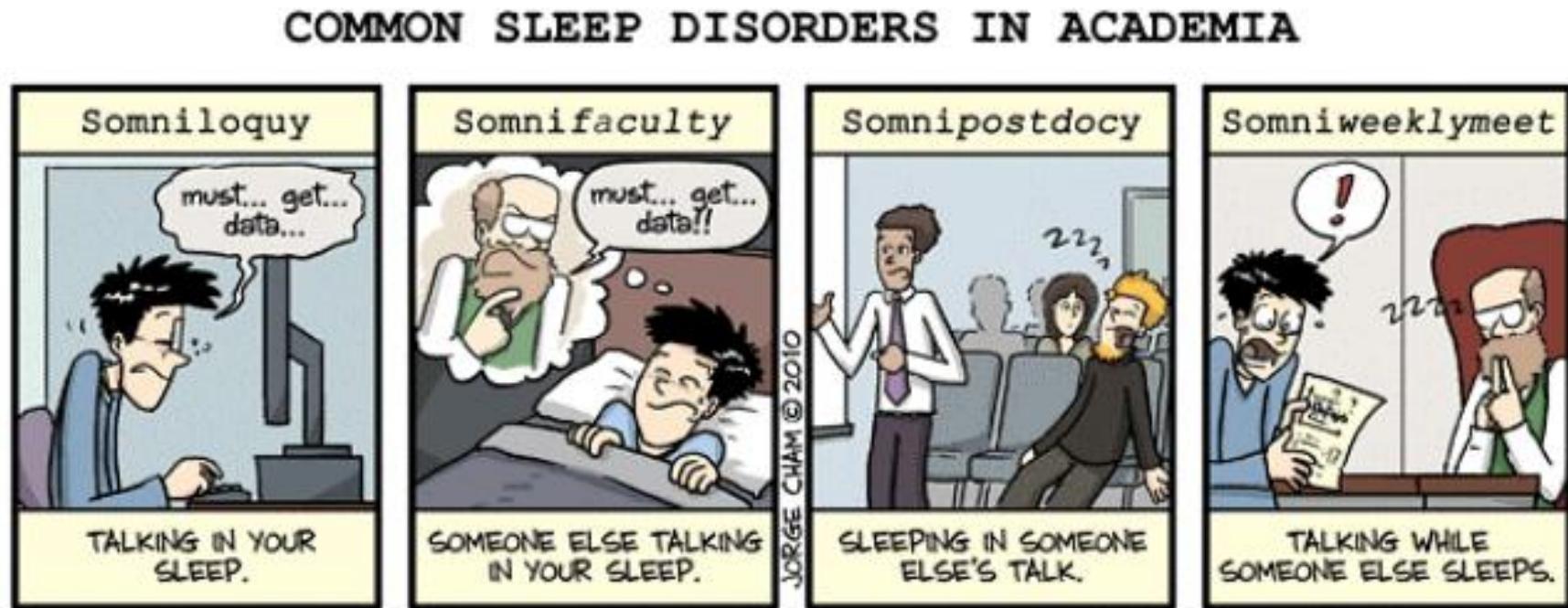
- \tilde{b}_i is the posterior mean for b_i , in a Bayesian calculation that ‘plugs-in’ $\beta = \hat{\beta}$, $\phi = \hat{\phi}$ and the MLEs for any terms in \mathbf{G} , i.e. generates a prior based on the data
- Could also describe \tilde{b}_i as a ‘plug-in’ estimate of the conditional **mode** of $b_i|\mathbf{Y}_i$ – this generalizes better, beyond LMMs
- Asymptotically in the number of clusters, each \tilde{b}_i is a good approximation to the mean/mode of Your posterior beliefs about that b_i , using any* proper prior for the parameters
- `nlme` author Doug Bates calls this a “statistical no-man’s land” – and Song’s Chap 9 simply punts on interpretation. Think of \tilde{b}_i (and corresponding intervals) as an approximation to ‘Your beliefs about the unknown cluster-specific effects b_i ’, having seen the data

Actually-Bayesian estimates will be given later.

* ... subject to regularity conditions, in particular the b_i being i.i.d.

LMMs: sleep!

An example with which you may empathize...

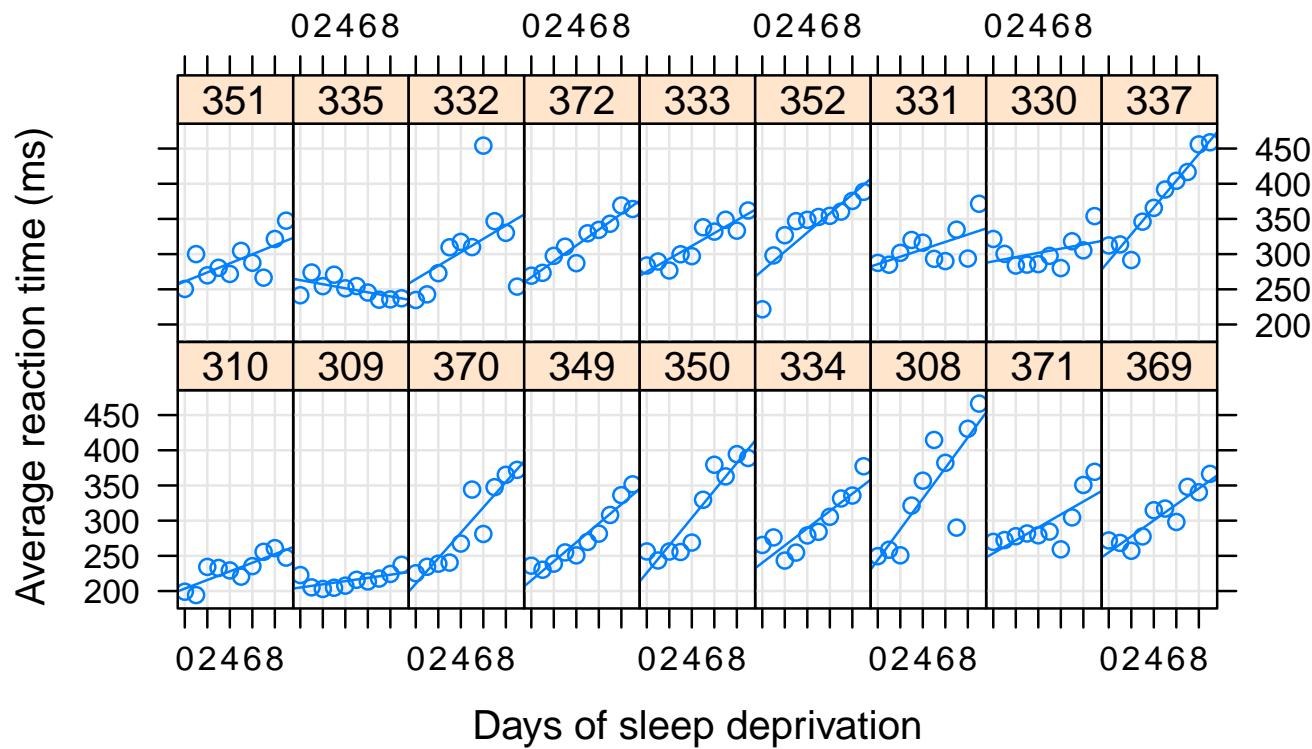


A professor is one who talks in someone else's sleep.

W. H. Auden (1907–1973)
Anglo-American poet, writer, and professor

LMMs: sleep!

The data comes from a small study on sleep deprivation; 18 truck drivers were deprived 3hrs sleep each night, for 10 nights, and their reaction times were measured;



Lines indicate subject-specific linear regressions

LMMs: sleep!

It seems clear there are between-subject differences in reaction time, and also (perhaps?) between-subject differences in response to sleep deprivation.

We consider a model which ignores the clustering (primarily for its point estimate) and several different random effects models – the Days variable is centered throughout;

```
library("lme4")                      # contains the (balanced) sleepstudy data
sleepstudy$DaysC <- with(sleepstudy, Days - mean(Days)) # centering

library("nlme")
lm0c   <- lm( Reaction ~ DaysC, sleepstudy) # ignore clustering
lme1   <- lme(Reaction ~ DaysC, random=~1|Subject, data=sleepstudy)
lme2a   <- lme(Reaction ~ DaysC, random=reStruct(~DaysC|Subject, pdClass="pdDiag")
               data=sleepstudy) # diagonal G
lme2b   <- lme(Reaction ~ DaysC, random=~DaysC|Subject,
               data=sleepstudy) # unrestricted G
```

The default for vector b_i fits an unstructured covariance \mathbf{G} , using Cholesky decompositions to keep the correlation within the teabag.

LMMs: sleep!

Some helpful utilities for `lme` output:

- `fitted()`: Gives the \hat{Y}_{ij} fitted values from slide 3.63
- `fitted(,level=0)`: Gives the $\mathbf{X}_{ij}^T \hat{\beta}$, fitted population means
- `ranef()`: Gives the \tilde{b}_i
- `predict()`: Gives $x^T \hat{\beta} + z^T \tilde{b}_i$ for specified covariates/cluster IDs, supplied in a `newdata` data frame. For example, to pull out the subject-specific intercepts $\hat{\beta}_0 + \tilde{b}_{i0}$:

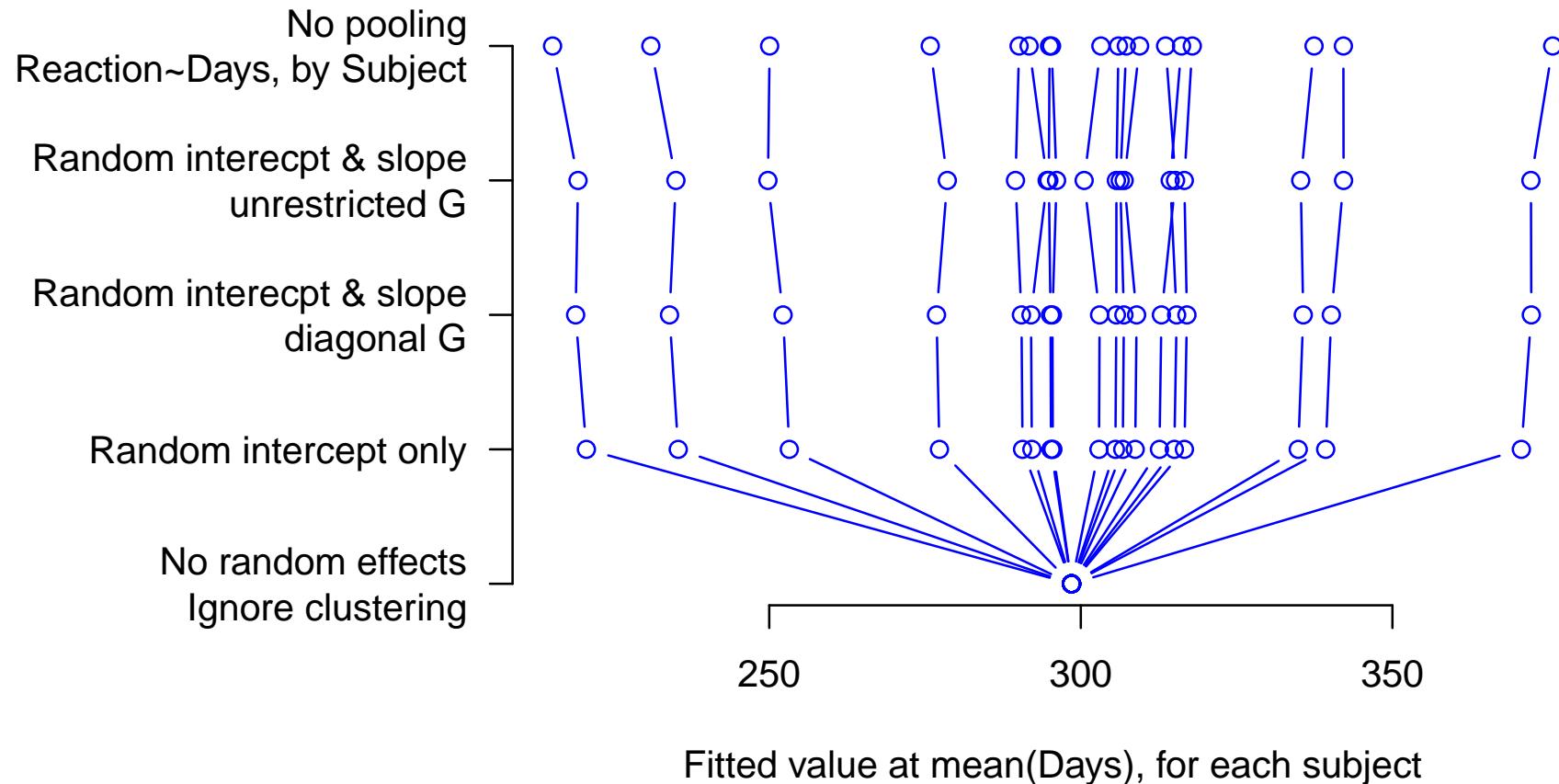
```
newdata0 <- data.frame(DaysC=rep(0,18), Subject=unique(sleepstudy$Subject) )  
predict(lme2a, newdata=newdata0)
```

- `coef()`: Gives coefficients $\hat{\beta}_k + \tilde{b}_{ki}$ – so `coef(lm2a)` is a simpler way to do the example above
- `intervals()`: Confidence intervals for the fixed effects (β) and covariance terms ($\mathbf{G}, \phi, \mathbf{R}$) – the latter use transformations from parameterizations that are unconstrained

For a full list, see `methods(class="lme")`. No prediction intervals around the \tilde{b}_i are available from `lme` – but `lme4`'s `lmer()` can provide these. (See later examples)

LMMs: sleep!

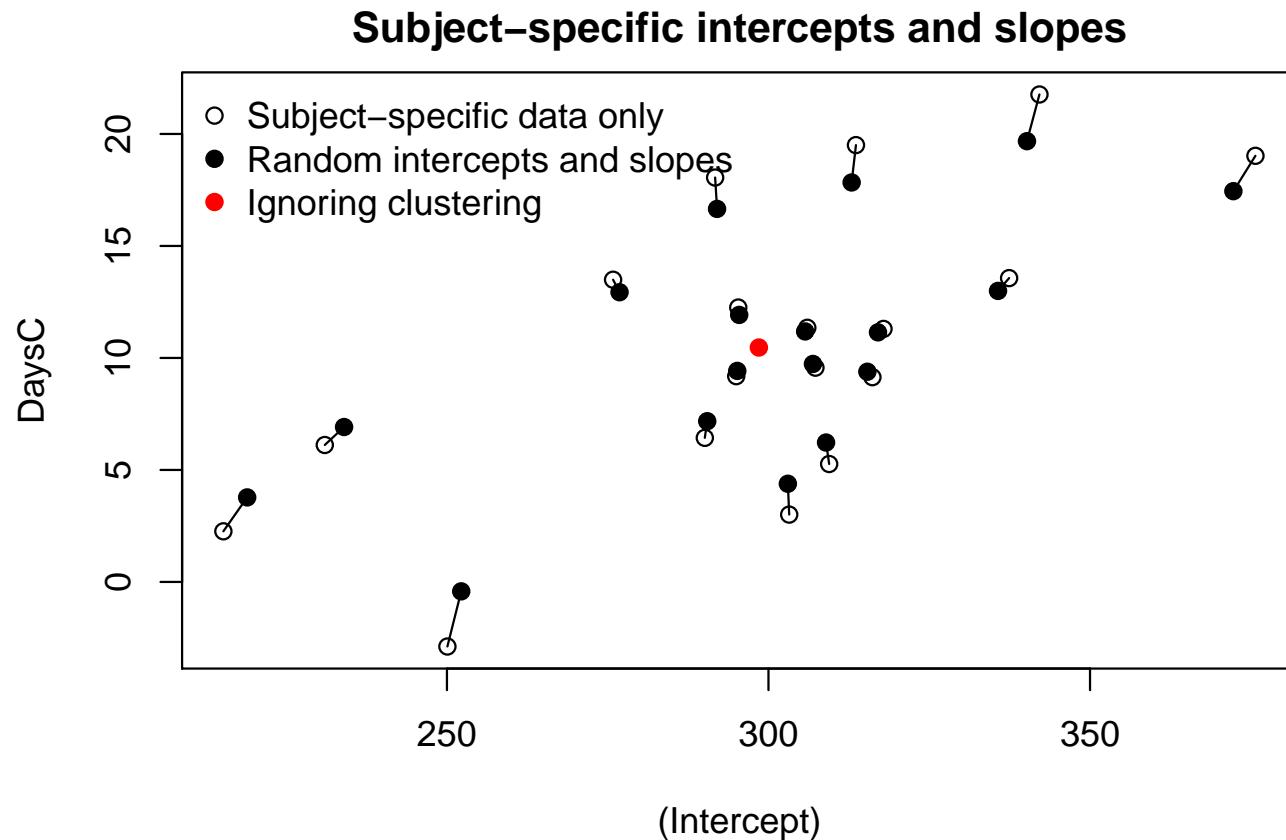
Compared to subject-specific analyses, the fitted values \hat{Y}_{ij} are ‘shrunk’ together;



Broadly, more shrinkage occurs with more homogeneity assumed across the clusters – but there is no single ‘right’ choice

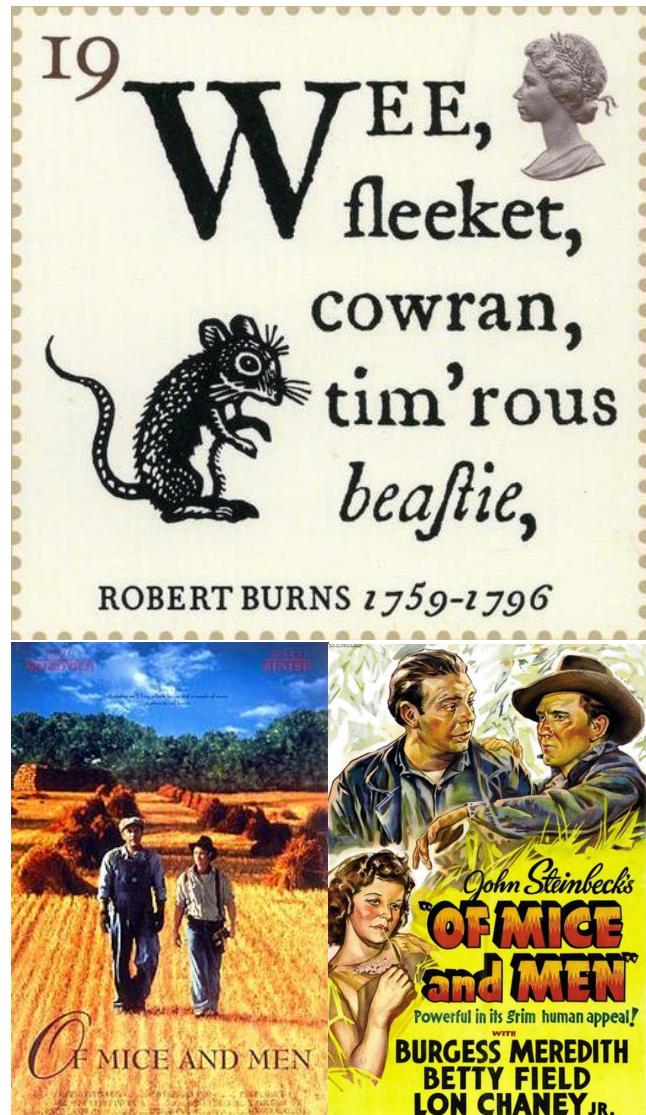
LMMs: sleep!

Predicted intercepts ($\hat{\beta}_0 + \tilde{b}_{i0}$) and slopes ($\hat{\beta}_1 + \tilde{b}_{i1}$) are also shrunk more with more assumed homogeneity;



Both here and for \hat{Y} (and in general for LMMs) extreme observations get shrunk more than near-average values.

LMMs: diagnostics



*...O, what a panic's in thy breastie!
Thou need na start awa sae hasty
Wi bickering brattle!*

*I wad be laith to rin an' chase thee,
Wi' murdering pattle.*

*But Mousie, thou art no thy lane,
In proving foresight may be vain:
The best-laid schemes o' mice an' men
Gang aft agley,
An' lea'e us nought but grief an' pain,
For promis'd joy!*

from 'To a Mouse'
by Robert Burns

LMMs: diagnostics

The best-laid general statement of LMMs;

$$\begin{aligned} b_i | \mathbf{X}_i, \mathbf{Z}_i &\stackrel{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{G}(\boldsymbol{\alpha})) \\ \mathbf{Y}_i | \mathbf{b}_i, \mathbf{X}_i, \mathbf{Z}_i &\stackrel{indep}{\sim} N(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i, \phi \mathbf{R}_i(\boldsymbol{\alpha})), \end{aligned}$$

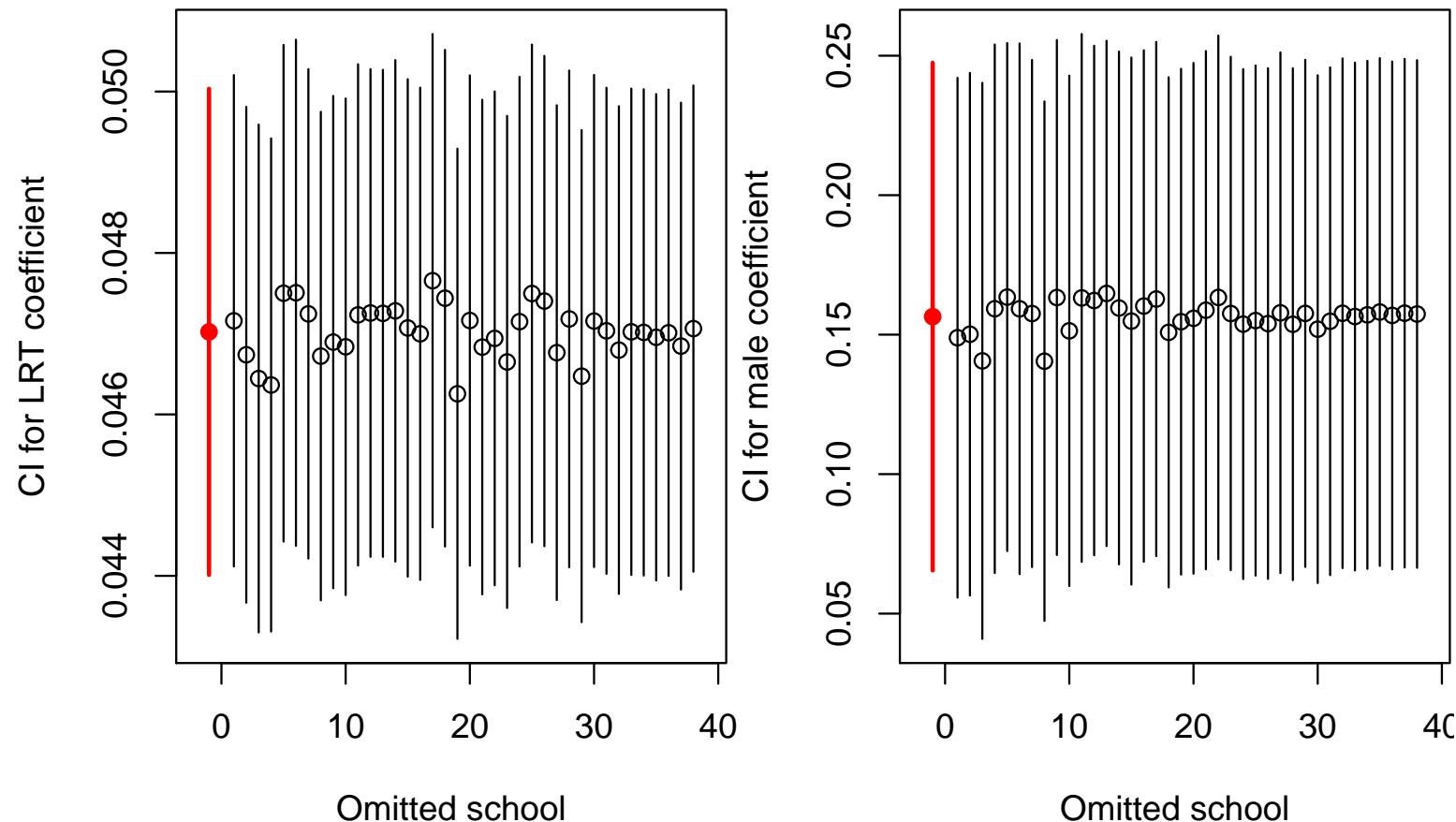
In approximate order of how much to worry about them, we can check;

- Large n : are results sensitive to any specific cluster?
- Mean model:
 - Is the marginal mean $\mathbb{E}[Y_{ij} | \mathbf{X}_{ij}] = \mathbf{X}_{ij}^T \boldsymbol{\beta}$?
 - Is the conditional mean $\mathbb{E}[Y_{ij} | \mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i] = \mathbf{X}_{ij}^T \boldsymbol{\beta} + \mathbf{Z}_{ij}^T \mathbf{b}_i$?
- Variance model:
 - Is the marginal variance $\text{Var}[\mathbf{Y}_i | \mathbf{X}_i, \mathbf{Z}_i] = \boldsymbol{\Sigma}_i = \phi \mathbf{R}_i + \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i^T$?
 - Is the conditional variance $\text{Var}[\mathbf{Y}_i | \mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i] = \phi \mathbf{R}_i$?
- Normality: are ϵ_i and b_i i.i.d Normal?

But see 2.83 for (usually much more important) first steps to do **before** any of what follows – and don't expect to have much power, unless the Tay Bridge really has fallen down.

LMMs: diagnostics

Leave-one-out analyses can address sensitivity to individual clusters – here with $\hat{\beta}$ from `lme(Y~LRT+male, ~1|School)`



Can also use n_i on x -axis, or $\#\text{males}_i$, etc. Leave-one-out for observations is similar but slower – and note n drives asymptotics.

LMMs: diagnostics

To check mean models, we use residuals, broadly as for GEE. But in LMMs just as there can be more than one ‘fitted value’ (see 3.68) we can define various residuals;

- *Stage zero/marginal* (population-level) residuals are the differences between outcomes Y_{ij} and fitted population mean for those covariates;

$$e_{ij} = Y_{ij} - \mathbf{X}_{ij}^T \hat{\boldsymbol{\beta}}$$

- *Stage one* (subject-level) residuals are the differences between outcomes Y_{ij} and their *estimated/predicted* cluster-specific mean;

$$\tilde{\epsilon}_{ij} = Y_{ij} - \mathbf{X}_{ij}^T \hat{\boldsymbol{\beta}} - \mathbf{Z}_{ij}^T \tilde{\mathbf{b}}_i$$

- Predicted random effects $\tilde{\mathbf{b}}_i$ are sometimes called *stage two* residuals – they describe how different cluster i is from the mean cluster, i.e. different from zero

LMMs: diagnostics

Following 2.90–2.95, the *response residuals* e_{ij} and \tilde{e}_{ij} could be plotted against fitted values and/or other functions of covariates to assess the mean models.

To avoid paying excessive attention to noisier residuals (i.e. those from observations with larger $\text{Var}[Y_{ij}]$) we could instead use *Pearson residuals*

$$e_{ij}/\sqrt{(\hat{\phi}\hat{\mathbf{R}}_i + \mathbf{z}_i\hat{\mathbf{G}}\mathbf{z}_i)_{jj}}, \quad \tilde{e}_{ij}/\sqrt{(\hat{\phi}\hat{\mathbf{R}}_i)_{jj}}.$$

The square of these (or their absolute value) could be used to check the variance models, assuming the mean model(s) were acceptable – again, just like we did for GEE.

These are not ‘wrong’, but can be misleading. Because of within-cluster correlation means, the individual residuals are not independent – so the value of a smoothed line may be driven by fewer observations than the scatterplot suggests.

LMMs: diagnostics

To counteract this, we can instead use *normalized residuals* – which divide each cluster-specific vector of residuals by a matrix square root of the corresponding variance. The matrix square root we use exploits *Cholesky factorization*;

- Any positive definite \mathbf{V} can be written as the cross-product of an upper-triangular matrix \mathbf{L} ; $\mathbf{V} = \mathbf{L}^T \mathbf{L}$
- This result is very useful in e.g. numerical linear algebra – note that R has a `chol()` function
- In particular, \mathbf{L} can be inverted quickly

The normalized residuals are the elements of

$$e_i^* = \mathbf{L}_e^{-1} e_i, \quad \tilde{\epsilon}_i^* = \mathbf{L}_\epsilon^{-1} \epsilon_i$$

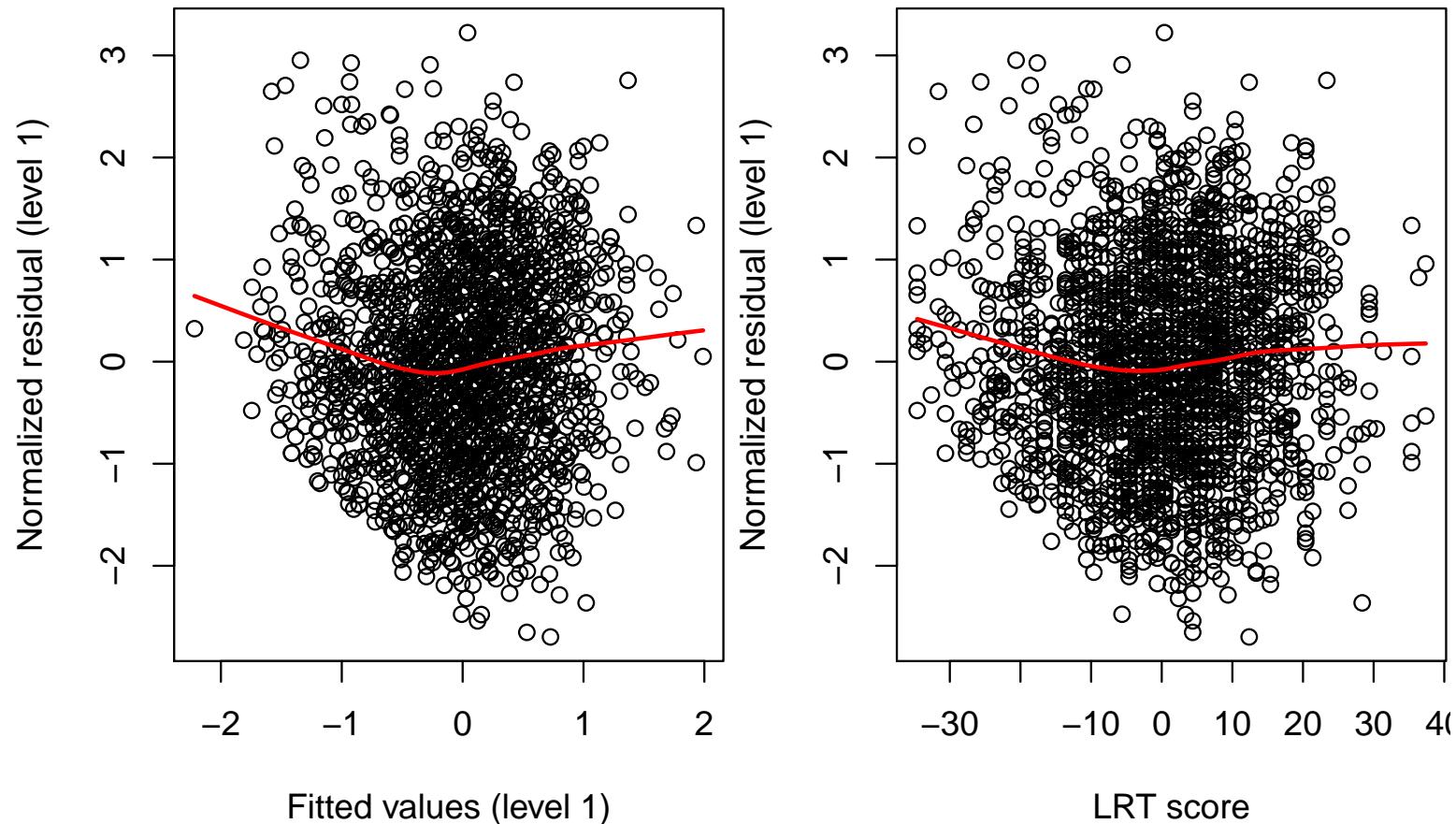
where \mathbf{L} is the Cholesky factorization of the relevant estimated covariance matrix.

Happily, the `residuals()` method for `lme` objects handles all this;

- `level` = 0 or 1 for stage 0 or 1 residuals
- `type` = `response`, `pearson`, or `normalized`

LMMs: diagnostics

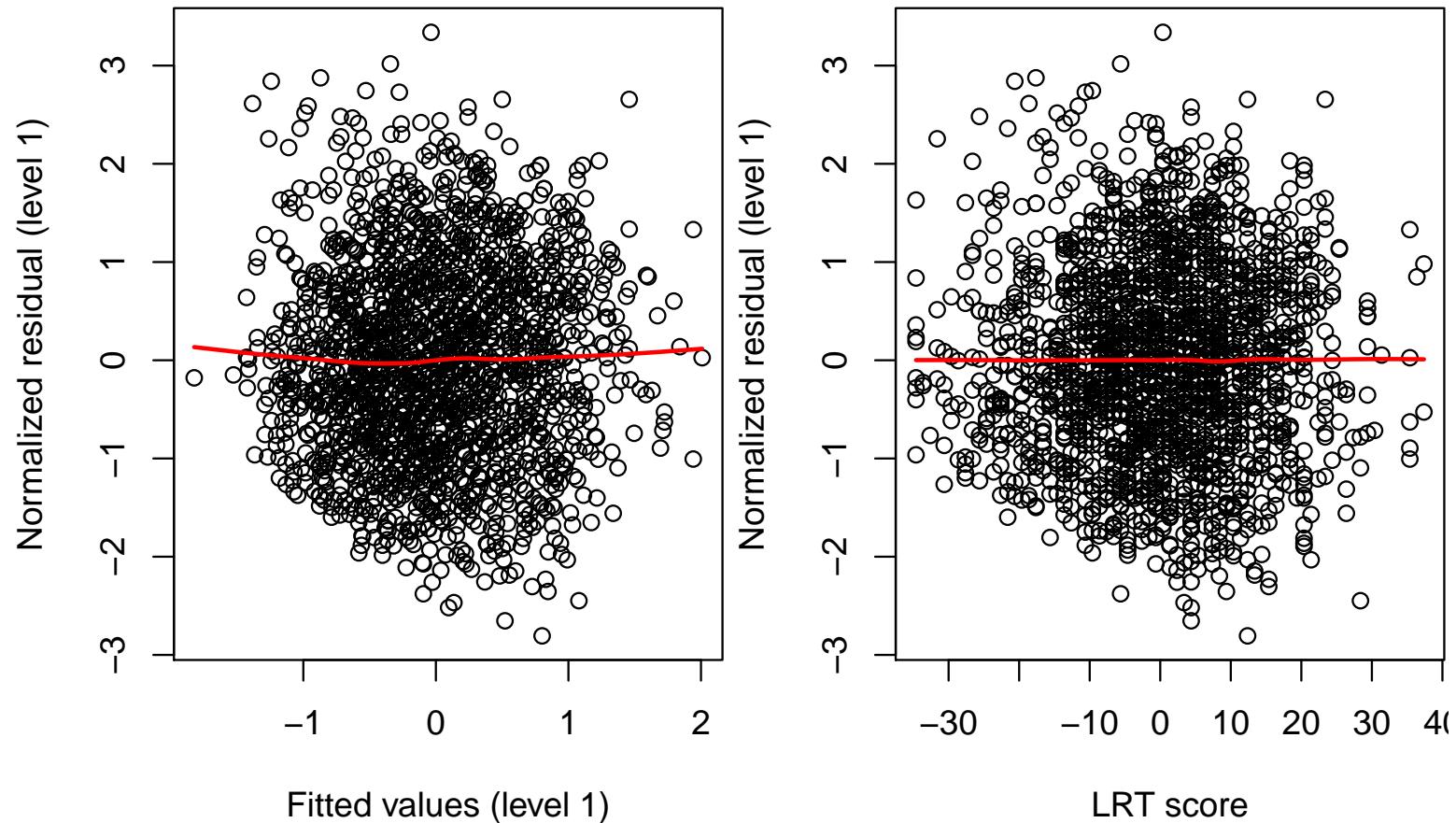
Back to the schools example; plotting and smoothing normalized level 1 residuals from `lme(Y~LRT+male, ~1|School)`



This suggests that $\mathbb{E}[Y]$, given sex & school, may not be linear in LRT – important if e.g. you think LRT confounds sex's effect.

LMMs: diagnostics

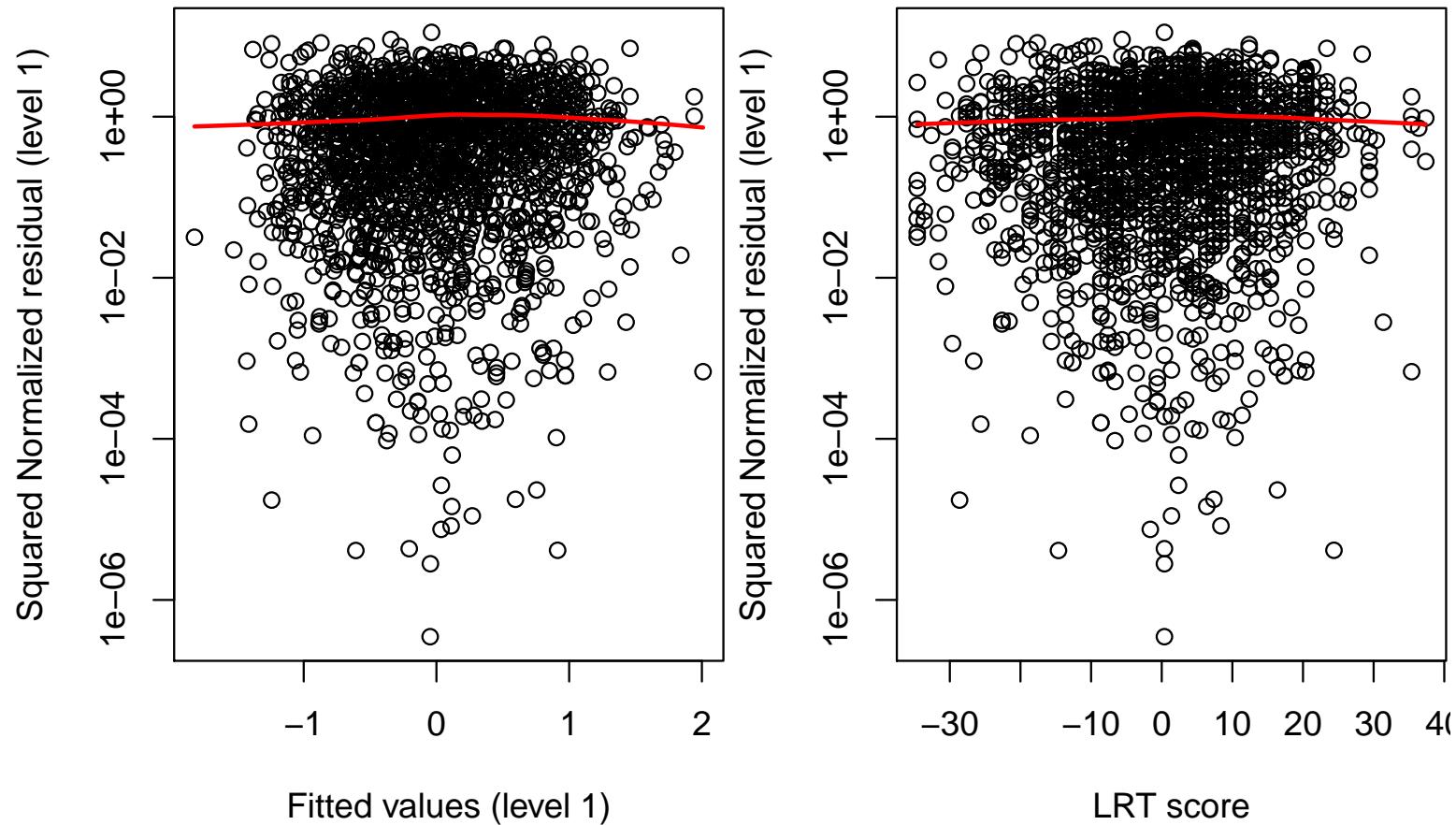
Using a cubic spline instead to represent LRT , with knots at the quintiles of LRT – i.e. `lme(Y~ns(LRT,5)+male, ~1|School)`



With this much data, fitting a flexible model like this is a good place to start the analysis.

LMMs: diagnostics

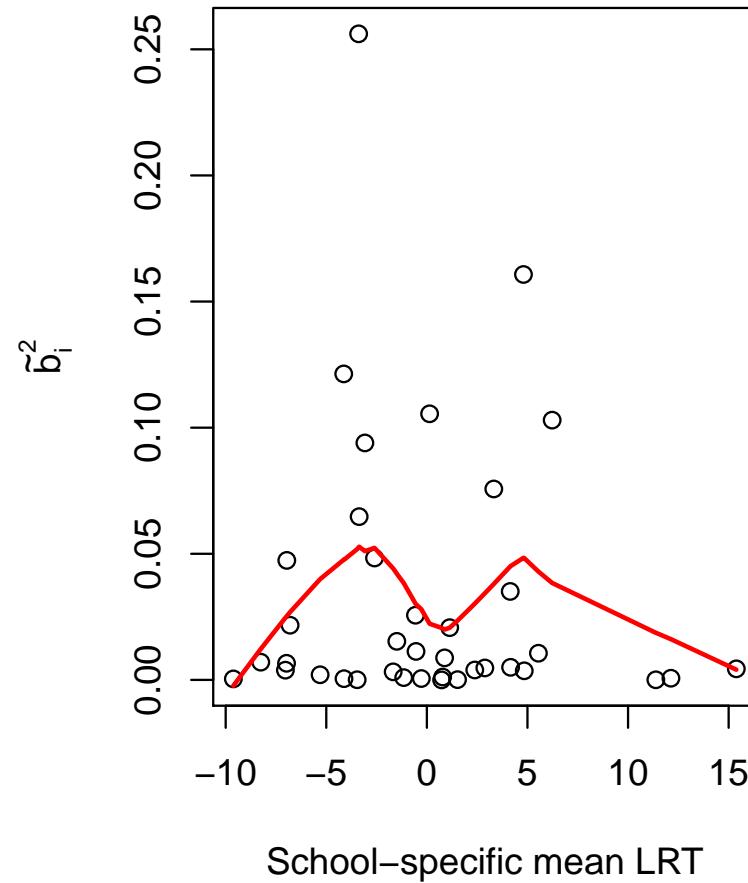
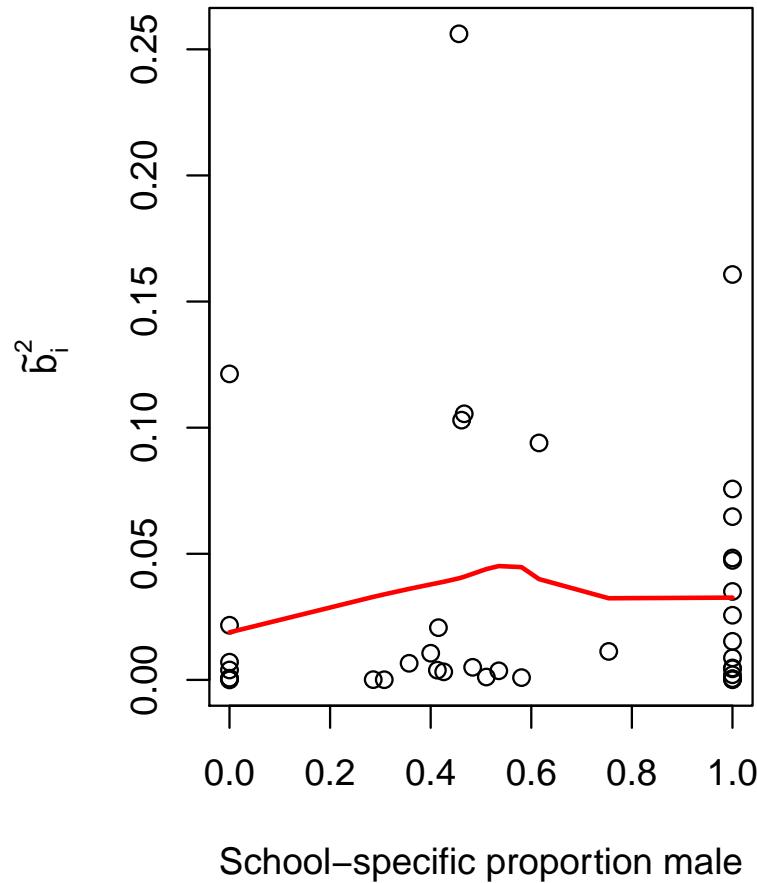
With the spline approach, checking the homoskedastic errors assumption using the level one normalized residuals;



Unlike GEE, validity of inference really does depend on this homoskedasticity assumption.

LMMs: diagnostics

To assess the homoskedasticity of the random effects, i.e. $\text{Var}[\tilde{b}_i | \mathbf{X}_i, \mathbf{Z}_i] = \mathbf{G}$, plot the squared \tilde{b}_i (or a function of them) against covariates (or a function of them);



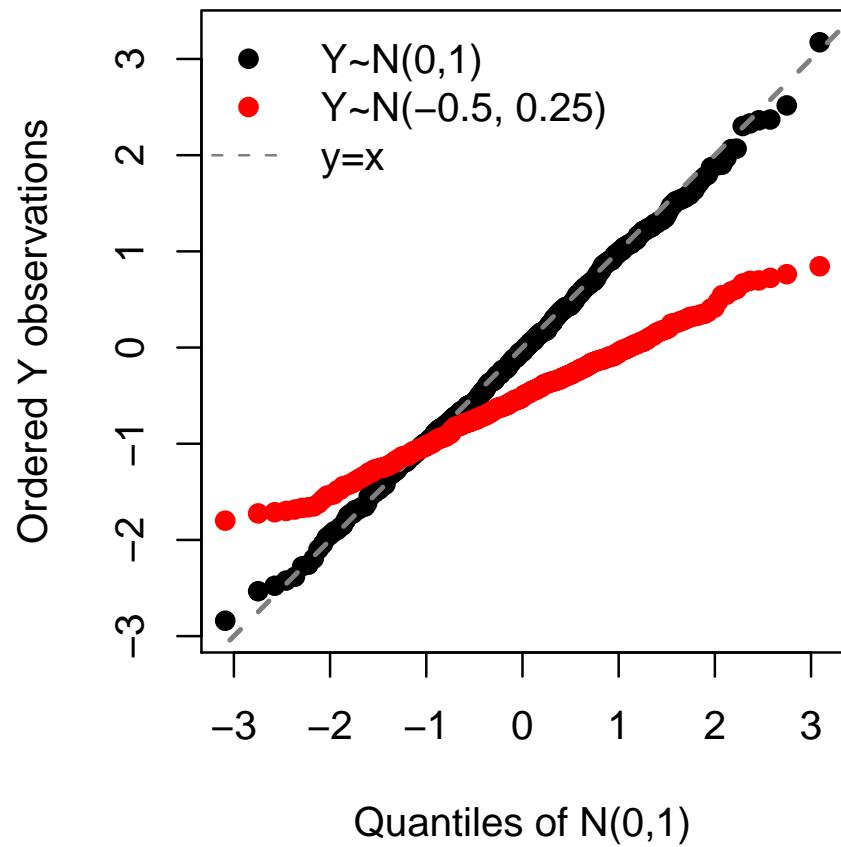
School 17 (see slide 3.6–3.7) has a large negative \tilde{b}_i .

LMMs: diagnostics

Finally, checking Normality* of the distribution of b_i or ϵ_i . Surprisingly histograms are **not** the best tool for this – they provide little visual information on the ‘tails’.

A *QQ plot* (‘quantile-quantile’) is better – plot the **ordered** sample values against the assumed distribution’s quantiles. ($n=500$, right)

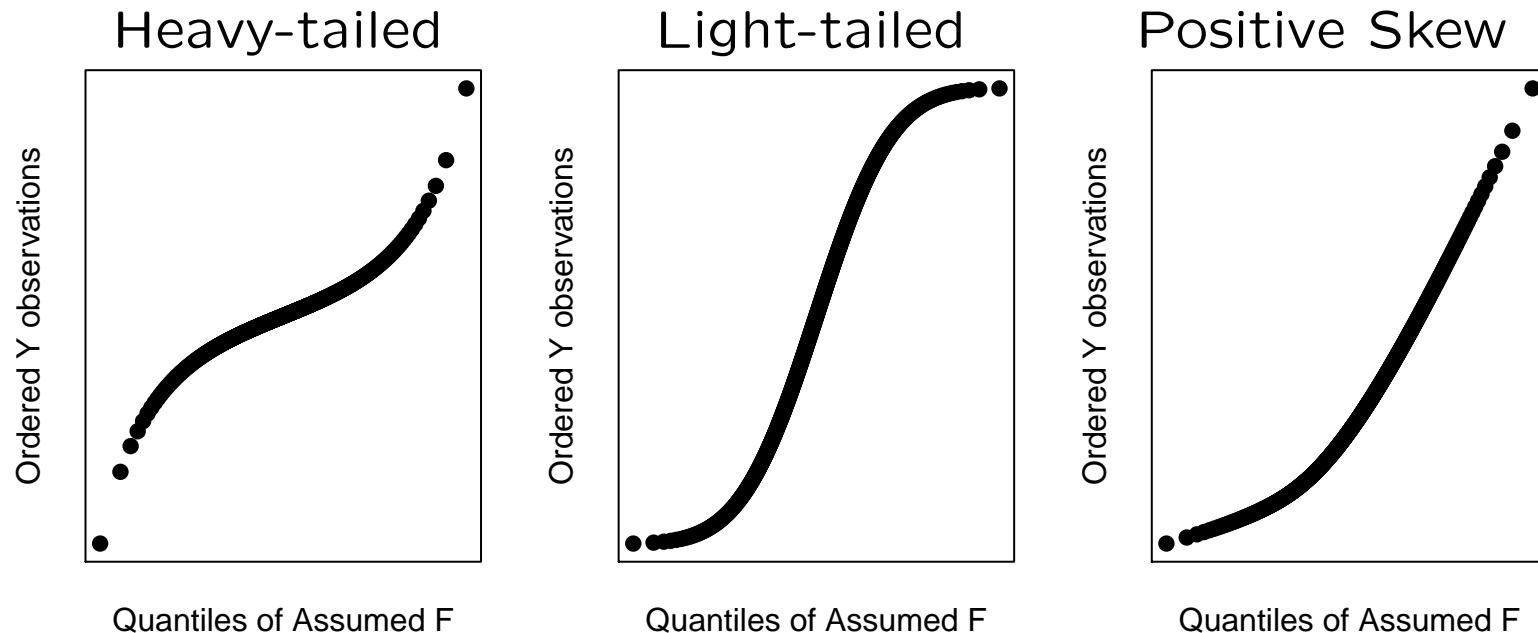
If the assumed distribution is correct, points lie on $y = x$, roughly. If it is correct up to location/scale, points lie on **some** straight line.



* ... which, again, is usually the least important LMM assumption

LMMs: diagnostics

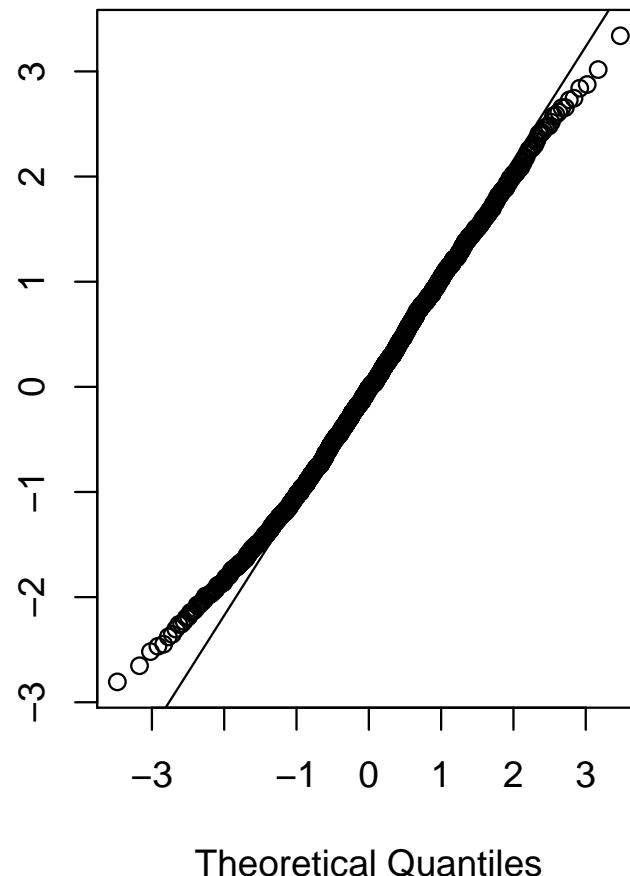
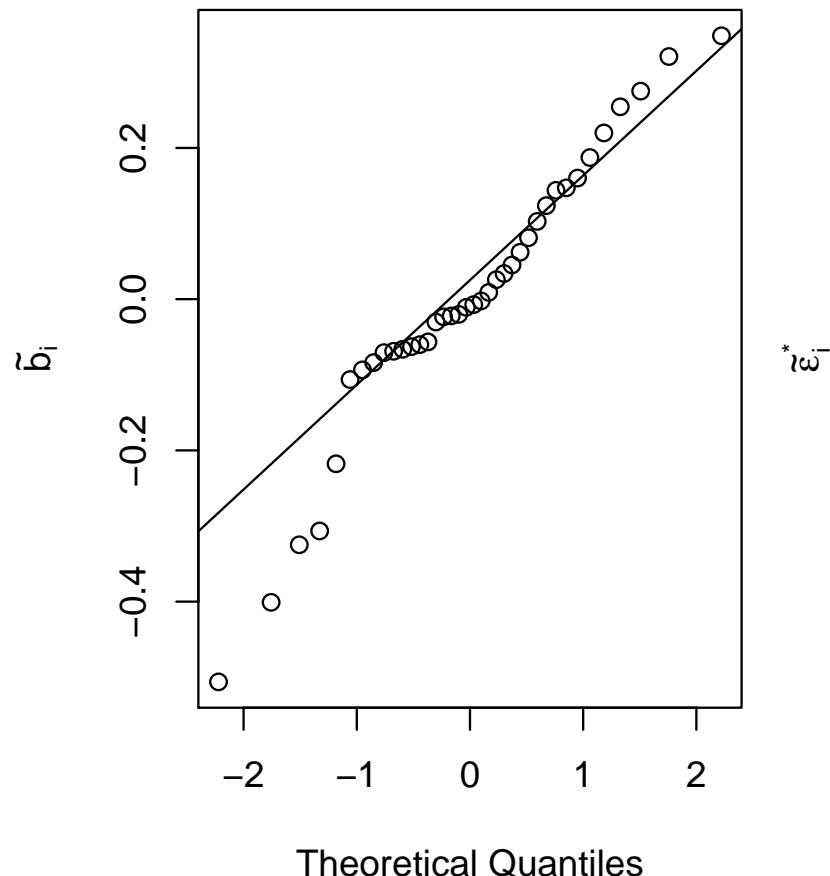
Violations of assumptions follow characteristic shapes;



- With n points, the x-axis values are `qnorm(ppoints(n))` – or use e.g. `qt()`,`qchisq()`,`qbinom()` to compare to other distributions
- Use order statistics (see STAT 512/513), transformed, to get 95% prediction intervals for each x -axis value – the tails of the plot are noisier than the middle

LMMs: diagnostics

For the London schools LMM, made using `qqnorm()`. The ‘guide’ line – from `qqline()` – goes through the 25% and 75%-iles.



Problems with mean & variance models may also show up here, as heavy tails, so don't use these plots in isolation.

LMMs: diagnostics

QQ plots for LMMs can easily deceive. Consider data from the following model, based on Verbeke and Lesaffre (1996);

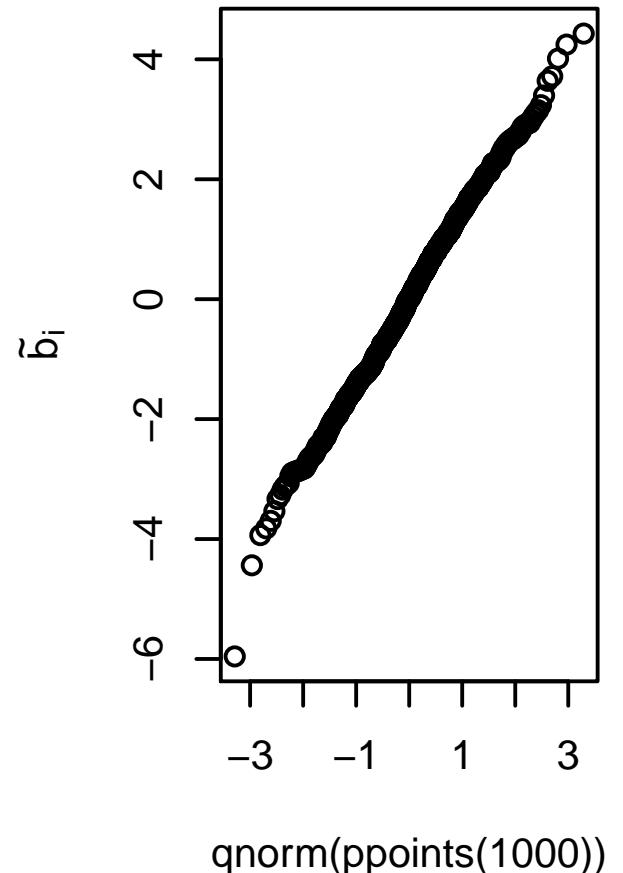
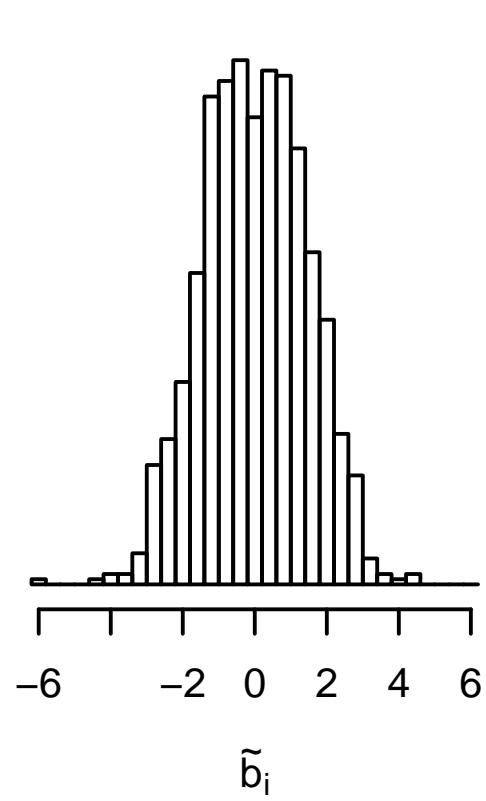
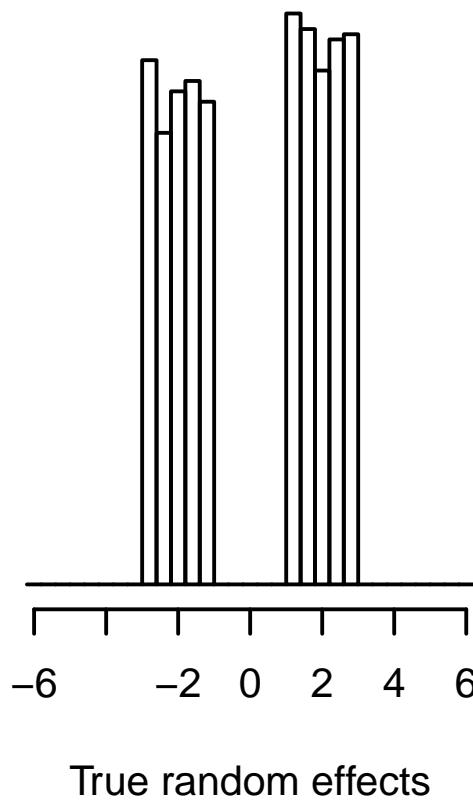
$$\begin{aligned} a_i &\stackrel{i.i.d.}{\sim} \frac{1}{2}U(-3, -1) + \frac{1}{2}U(1, 3) \\ Y_{ij}|a_i &\stackrel{\text{indep}}{\sim} N(a_i, \sigma_Y^2) \end{aligned}$$

for $1 \leq j \leq n_i \equiv 5$ and $1 \leq i \leq n = 1000$, with $\sigma_Y = 6$ – i.e. many small clusters, where the within-cluster noise is comparable to the between-cluster variability

- We will fit a standard LMM to this data, regressing Y on just an intercept, with assumed-Normal random effect intercepts
- As the first and second moments have been specified correctly, this LMM will provide valid inference on mean model parameters, with large n – which we have
- Estimates of fixed variance terms (e.g. σ_Y , ICC) will also be consistent
- However, the BLUP \tilde{b}_i do rely on Normality, and the data provide little information to the contrary...

LMMs: diagnostics

Really little information: some scary pictures;



LMMs: diagnostics

- Assumptions of Normal random effects cannot be reliably tested – at least not from fitted random effects \tilde{b}_i
- Normality of the b_i is not to blame – the \tilde{b}_i will ‘shrink’ to look like whatever we assume
- BLUP is not to blame – any method that assumes a parametric form would have the same difficulty; the way that \tilde{b}_i ‘borrows strength’ from the other clusters’ data is determined by this choice
- The problem is less bad if the data is very informative about each b_i ... but then you’re less likely to be using mixed models in the first place!
- There is no good objective way to infer* the ‘correct’ shrinkage factor. Expect to have to justify your choice of **independently** of the observed data (or show robustness to it) – particularly if you make inference on the b_i

* This is not true for prediction, where shrinkage is also used

LMMs: diagnostics summary

Final notes on LMM diagnostics;

- The usual concerns apply: diagnostics cannot be guaranteed to find model-misspecifications that matter for your inference
- Compared to GEE & 570, there are more assumptions, and more ways to check them. The specificity of what each diagnostic tells you can be poor
- Large-enough n , mean model and homoskedasticity assumptions are usually the priorities
- Normality... not so much
- But when interest lies in the b_i do expect aspects of the output to be sensitive to assumed Normality

Also recall slide 2.83; the diagnostics here are (just) for checking that your LMM, chosen based on context, is not grossly wrong.

LMMs: summary

Final notes on LMMs;

- Very widely used for clustered outcomes – along with GEE it is a standard tool
- Close connections with GEE linear regression with exchangeable working correlation matrix – but assumption of i.i.d. random effects from a population makes conditional interpretation of parameters possible
- Being fully-parametric, MLEs can be used for population parameters – though REML is the default
- Boundary-valued variances can be expected
- Predicting the random effects may be useful on its own, and also sets up diagnostics... but relies heavily on assuming Normality

GLMMs

*Fhairshon had a son,
Who married Noah's daughter,
And nearly spoil'd ta flood
By trinking up ta water—*

*Which he would have done,
I at least believe it,
Had ta mixture peen
Only half Glenlivet.*

From *The Massacre of the Macpherson*, by W.E. Aytoun
Scottish lawyer, humorist & poet (1813-1865)



*Love makes the world go round? Not at all.
Whisky makes it go round twice as fast.*
Compton Mackenzie (1883–1972)
Writer & raconteur, author of *Whisky Galore!*



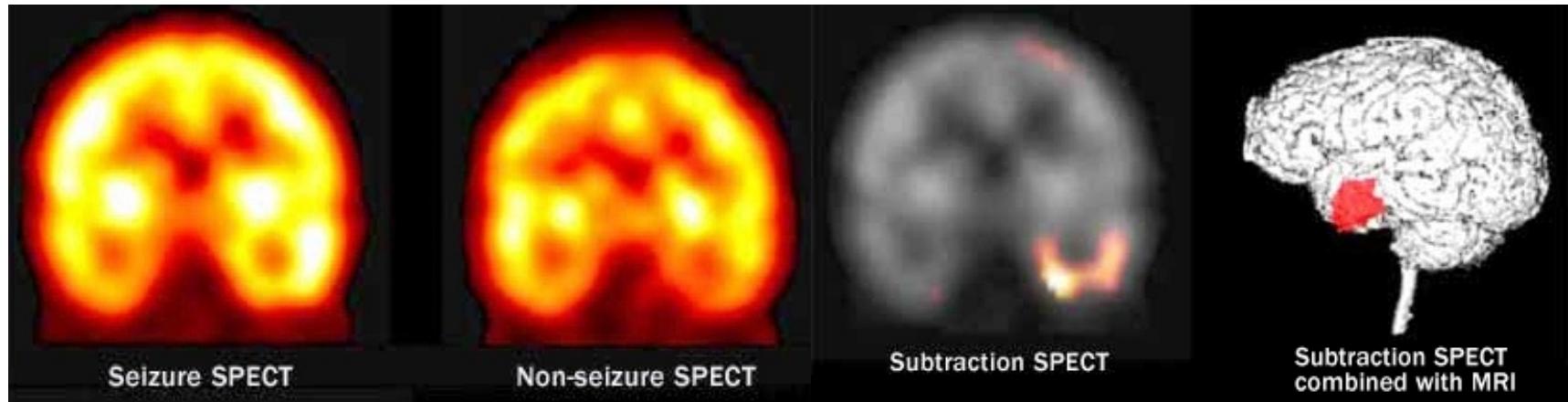
GLMMs

We're nearing the end of our guided tour of frequentist analysis of mixed models, and so provide a taste of the really fancy stuff; *Generalized Linear Mixed Models* (GLMMs).

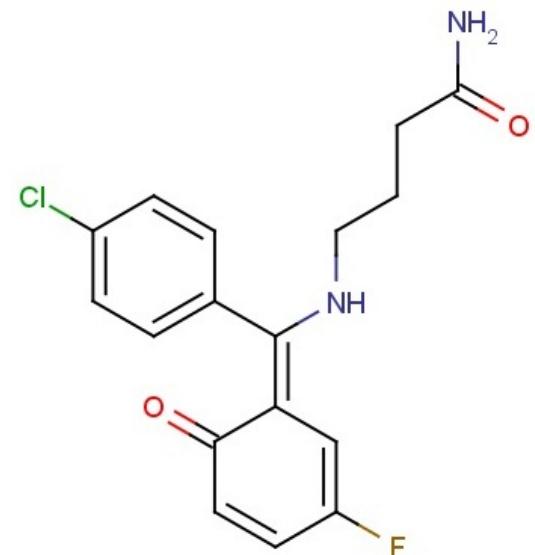
Some tasting notes to relish first;

- Computation – specifically numerical integration – becomes more challenging, and limits what models can be used fit
- The `lme4` package is among the best-available tools, however
- In GLMMs, connections with GEE (marginal/conditional) exist in only very special situations – so explaining the differences is a common task
- Predicting random effects and/or diagnostics, as we just saw for LMMs

GLMMs: progabide!

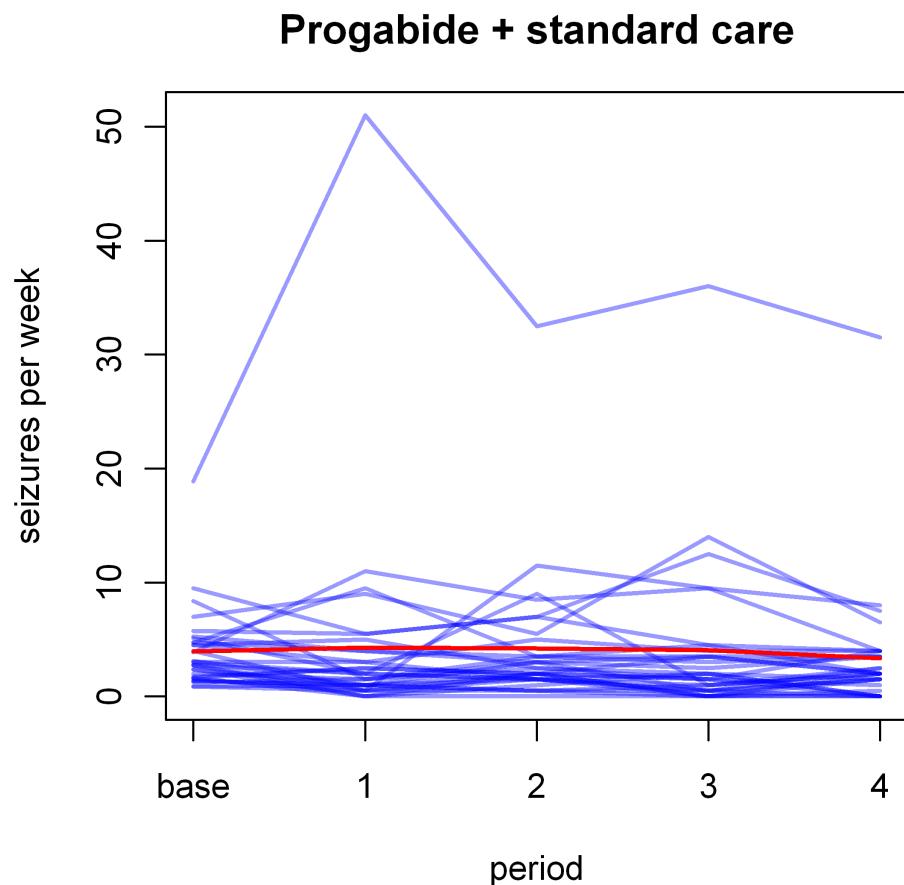
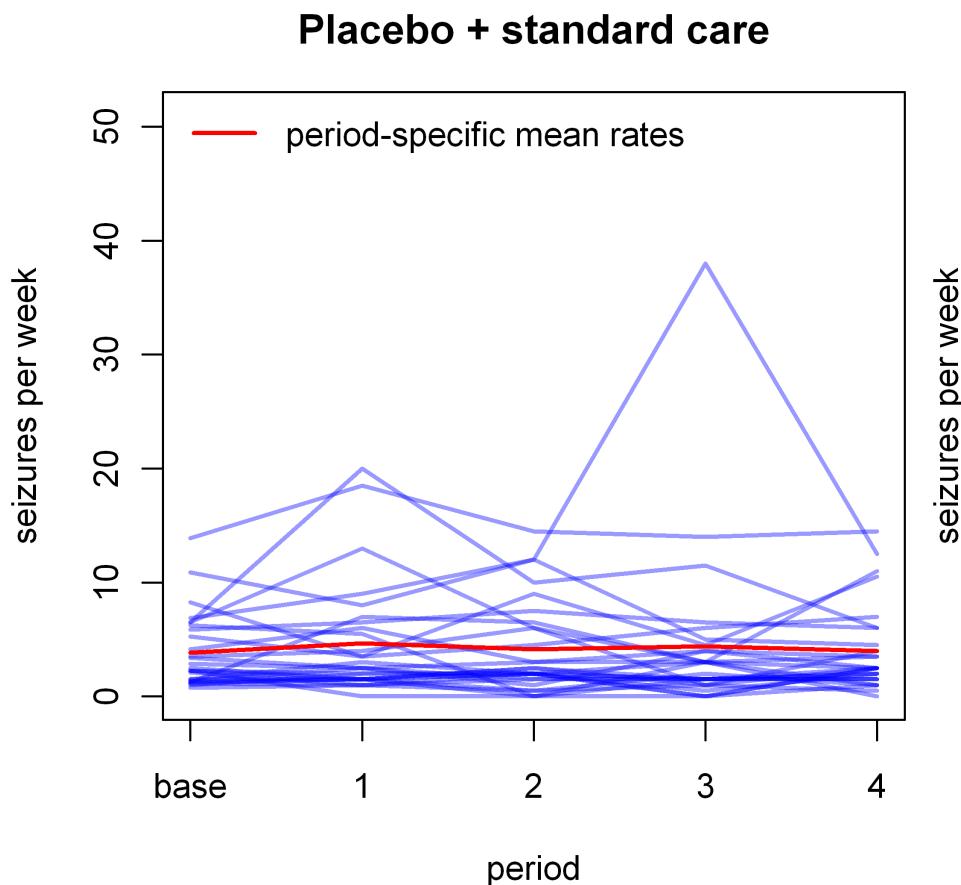


The seizure data (available on the class site, in long form) contains data on 59 subjects in a randomized trial of progabide, versus placebo, plus standard care. The numbers of epileptic seizures are recorded in an initial 8-week baseline period, followed by four 2-week periods.



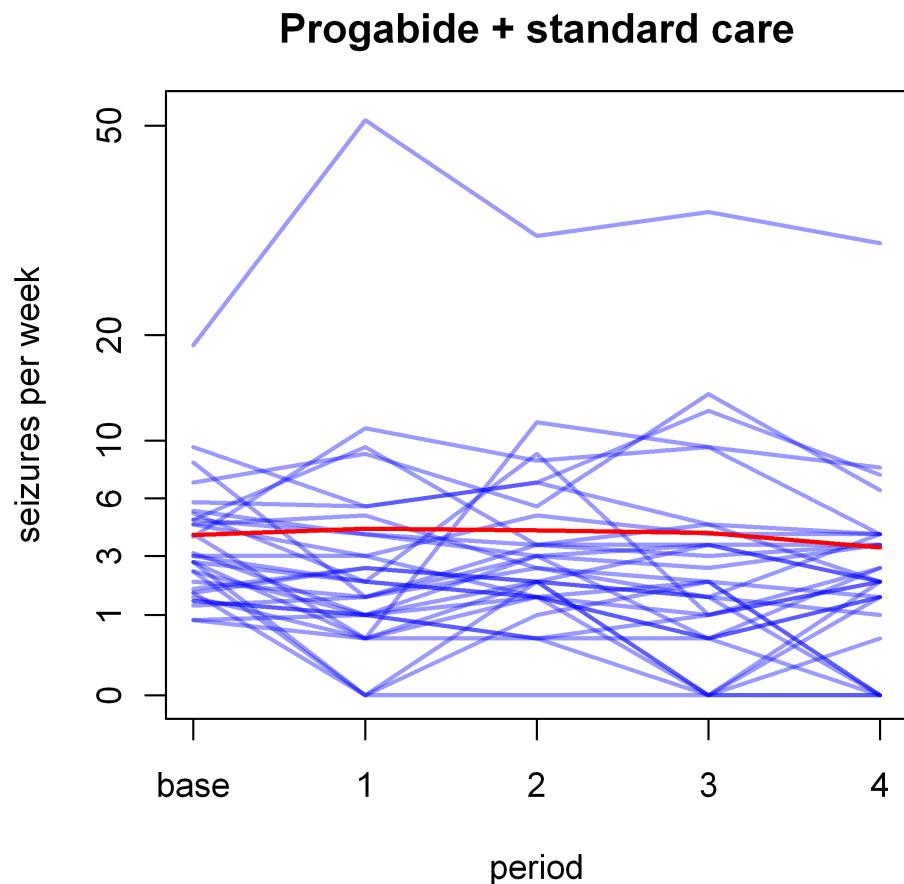
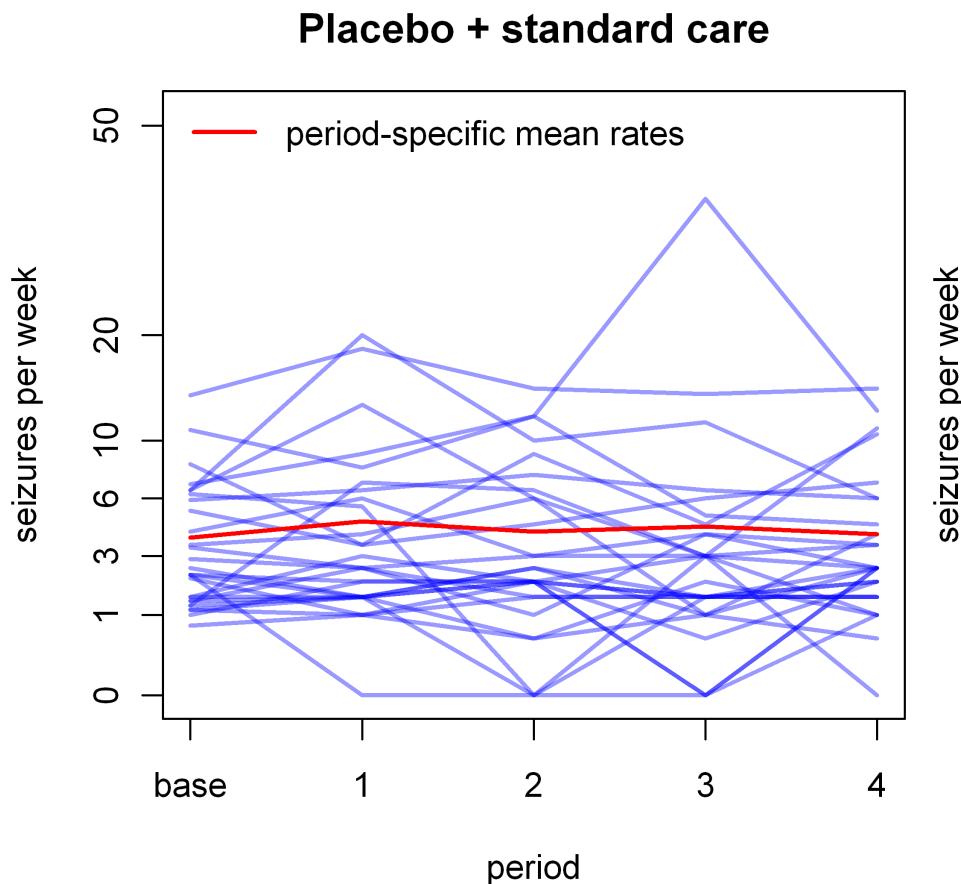
GLMMs: progabide!

The data; (one line = one person = one cluster)



GLMMs: progabide!

The data; y-axis is square-root transformed



GLMMs: probabide!

Generalizing what we saw for LMMs, our model is

$$\begin{aligned} b_i | \mathbf{X}_{ij} &\stackrel{i.i.d.}{\sim} N(0, \sigma^2) \\ Y_{ij} | b_i, \mathbf{X}_{ij} &\stackrel{\text{indept}}{\sim} \text{Poisson}(\mu_{ij}) \\ \mu_{ij} &= \exp(\log(t_{ij}) + b_i + \beta_0 + \beta_1 \text{base}_{ij} + \beta_2 \text{trt}_{ij} + \beta_3 \text{base}_{ij} \text{trt}_{ij}) \\ &\equiv \exp(O_{ij} + b_i + \mathbf{X}_{ij}^T \boldsymbol{\beta}) \end{aligned}$$

where

- $1 \leq i \leq n = 59$ and $1 \leq j \leq n_i \equiv 5$
- ‘base’ is an indicator for baseline period observations
- ‘trt’ is an indicator for probabide treatment
- $t_{ij} = 8, 2$ for baseline, non-baseline respectively
- Offset (O_{ij}) is just a covariate with coefficient fixed at 1

With the observations following a GLM conditional on b_i , this is a *generalized linear mixed model* (GLMM) – with fixed effects including an offset, and random intercepts.

GLMMs: probabide!

As we saw for LMMs, maximizing the log-likelihood of the observed values $\mathbf{Y}|\mathbf{X}$ – marginalized over the unobserved \mathbf{b}_i – will give MLEs for $\boldsymbol{\beta}$ and σ .

- For Normal-Normal models, need to do Fisher-scoring steps involving matrix expressions – but nothing worse
- In the probabide example, the log-likelihood is

$$\sum_{i=1}^n \log \left(\int \prod_{j=1}^{n_i} \frac{e^{-e^{O_i+b_i+\mathbf{X}_{ij}^T\boldsymbol{\beta}}} e^{Y_{ij}(O_i+b_i+\mathbf{X}_{ij}^T\boldsymbol{\beta})}}{Y_{ij}!} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-b_i^2/2\sigma^2} db_i \right)$$

... which does not simplify – note the $e^{-e^{b_i+...}}$ terms

- *Numerical integration* is required for each of the n integrals – followed by *numerical optimization* with respect to parameters $\boldsymbol{\beta}$ and σ

GLMMs: probabide! – coding by hand

Numerical computing is a huge research area – here are some tools, and where to find others;

- Numerical integration ([Task View](#))
 - `integrate()`
 - `adaptIntegrate()` in the `cubature` package
 - The `R2Cuba` interface to the C library `Cuba`
- Numerical optimization ([Task View](#))
 - `nlm()` implements a Newton-style algorithm for minimization, using numerical first & second derivatives
 - `optim()` is similar, but with different options
 - `nlminb()` minimizes subject to ‘box’ constraints (such as $\sigma^2 > 0$ or $-1 < \alpha < 1$)

Several optimization routines use (numerical) first and second derivatives, and so without further work will output the approximate Hessian at the MLE, i.e. the observed Fisher information matrix, used in standard error estimation.

GLMMs: probabide! – coding by hand

We will use `adaptIntegrate()`, which uses *adaptive Gaussian quadrature*.

You saw non-adaptive Gaussian quadrature in 570; the integral of any function can be approximated by the following sum;

$$\int f(b) \, db \approx \sum_{r=1}^{n_w} f(b_r) w_r.$$

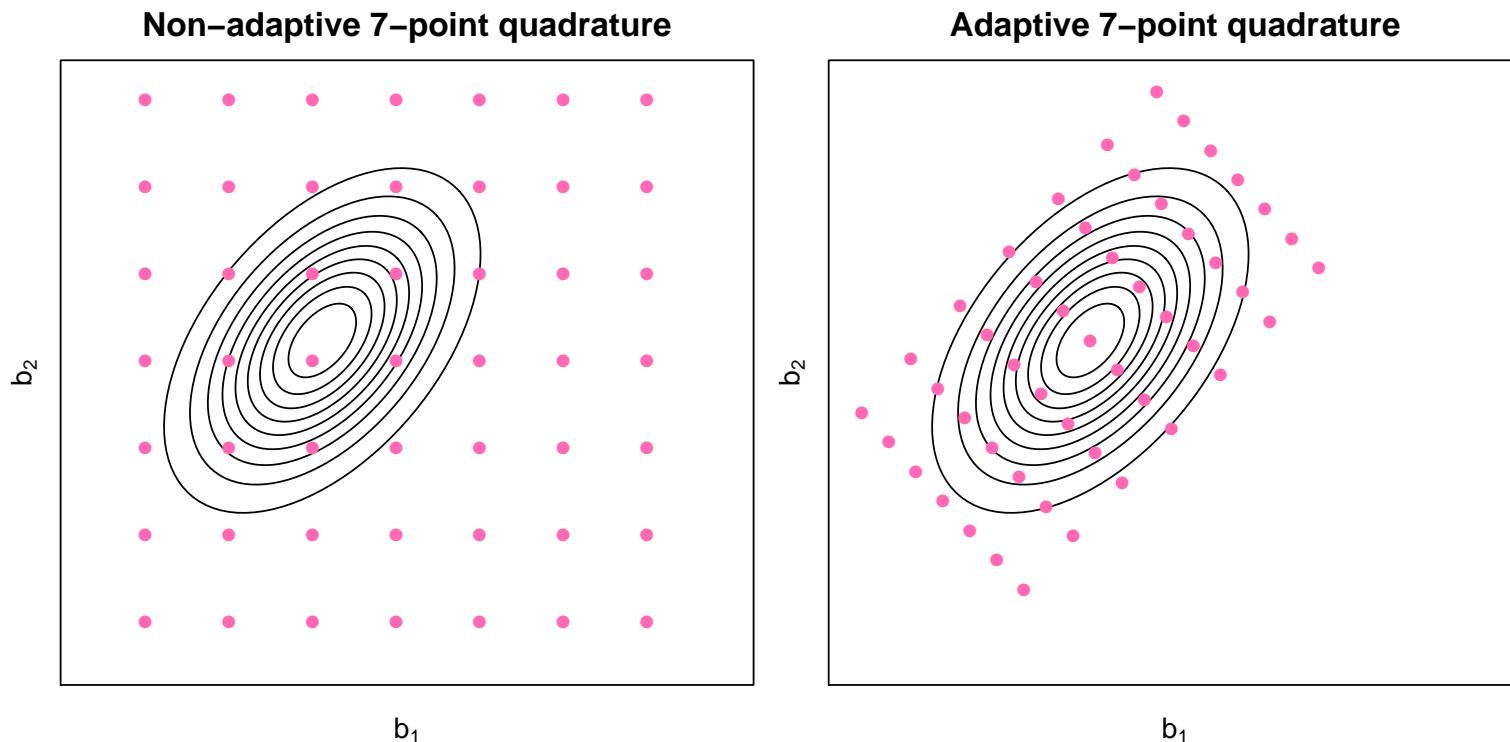
In Gauss-Hermite quadrature, we write $f(b) = g(b)e^{-b^2}$, and set the design points b_r to be the zeroes of Hermite polynomials $H_{n_w}(b)$ of degree n_w , and the weights w_r to be

$$w_r = \frac{w^{n_w-1} n_w! \sqrt{\pi}}{n_w^2 [H_{n_w-1}(b_r)]^2}.$$

A simple but surprisingly useful special case uses just one design point ($b_1 = \text{argmax}_b(f)$) – this is *Laplace approximation*.

GLMMs: probabide! – coding by hand

Adaptive quadrature is smarter; it picks the design points to reflect the location and scale of the integrand – here in 2D;



With either a ‘nice’ integrand, or many design points, this method will work well, for up to (say) five dimensions. Details are in Rabe-Hesketh *et al* (2005).

GLMMs: progabide! – coding by hand

We first fit it ‘by hand’; the data is in data frame `seiz.l`

```
library("cubature")
# First, a function to do all the cluster-specific integrals;
iliik.all <- function(beta, sigma){
  by(seiz.l, seiz.l$id, function(cluster){
    Xi <- with(cluster, cbind(
      log(t), rep(1, dim(cluster)[1]), base, trt, base*trt ))
    fixed <-  Xi %*% c(1, beta) # offset has coefficient==1
    integrand <- function(b){
      prod(dpois(cluster$y, exp(fixed+b) )) * dnorm(b, 0, sigma)
    }
    adaptIntegrate( integrand, -10, 10)$integral      # from cubature
  })}      # also consider integrating from e.g. -5*sigma to 5*sigma

# A function giving the full log-likelihood
llik <- function(beta, sigma){
  sum(log(unlist(iliik.all(beta, sigma))))}

# ... which we maximize
nlm1 <- nlm( f= function(p){ -llik(p[1:4], exp(p[5])) }, # note use of log-sigma
            p=c(1,0,0,0,-0.5), hessian=TRUE )
```

GLMMs: probabide! – coding by hand

This takes about 1m20s to fit, on my laptop. `nlm()` takes only 17 steps, but each involves 59 numeric integrals.

What are the point estimates, and estimate standard errors?

```
> nlm1$est  
[1] 1.1445132 -0.1118348 -0.1261635 0.1047226 -0.2495669  
> sqrt(diag(solve(nlm1$hess)))  
[1] 0.15205772 0.04687664 0.21003451 0.06503020 0.09607870
```

... and several start values (`p`) agree. Other computing notes;

- Use of the non-adaptive `integrate()` is not sufficiently accurate, here. The integrals are of functions which are almost zero everywhere; this is a challenge, for any numerical method. `integrate()` also requires vectorized integrands – and gives you no warning about this
- `optim()`'s default Hessian-calculator can give non-positive-definite output – `eigen()` may help you spot this
- Expect computation time to scale with n – and grow (badly!) with dimension of the random effects

GLMMs: progabide! – coding by hand

In practice, the output might be reported as;

- $\exp(\hat{\beta}_0) = 3.1$ (2.3, 4.2): rate of seizures per week, in non-baseline periods, on placebo (in an average cluster)
- $\exp(\hat{\beta}_1) = 0.88$ (0.58, 1.30): within-subject rate ratio comparing baseline to non-baseline periods, on placebo
- $\exp(\hat{\beta}_2) = 0.89$ (0.82, 0.98): within-subject rate ratio due to drug (versus placebo), in non-baseline periods
- $\exp(\hat{\beta}_3) = 1.10$ (0.98, 1.30): within-subject ratio of rate ratio comparing baseline to non-baseline periods, on treatment versus on placebo
- $\hat{\sigma} = 0.78$ (0.65, 0.94): standard deviation of random effects

To aid interpretation of $\hat{\sigma}$, note that a difference in $\log(\mu_{ij})$ of 0.78 corresponds to a RR of 2.2 (1.9, 2.6).

GLMMs: using glmer()

As you'd expect, tools are available that make the coding quicker – though the same basic methods are used, 'under the hood'. The `lme4` package is R's main resource for fitting GLMMs – it generalizes and extends what was in `nlme`. The workhorse of `lme4` is `glmer()` – though `lmer()` is also available for the special case of LMMs, where `family="gaussian"`.

For the progabide example; (compare with output on 3.100)

```
> library("lme4")
> glmer1 <- glmer(y~offset(log(t))+base*trt+ (1 | id) ,data=seiz.l,family=poisson)
> glmer1
Generalized linear mixed model fit by the Laplace approximation
Random effects:
Groups Name        Variance Std.Dev.
id      (Intercept) 0.60639  0.77871 # 0.77871^2=0.60639
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.14457   0.15189   7.535 4.87e-14 ***
baseTRUE    -0.11184   0.04688  -2.385   0.0171 *
trt         -0.12616   0.20986  -0.601   0.5477
baseTRUE:trt 0.10471   0.06504   1.610   0.1074
```

GLMMs: using `glmer()`

Notes on using `glmer()`:

- The `formula` argument's syntax specifies random effects – e.g $y \sim x + (1|id)$ for random intercepts. This differs from `lme`'s separate `fixed` and `random` formulae
- For random slopes use e.g. $y \sim x + (1+x|id)$ – with unconstrained $\text{Var}[b_i] = G$
- MLEs are provided – there is no REML option
- Random effects variances going to zero can cause numerical difficulties. Do check the full output, not just $\hat{\beta}$
- The `lme4` syntax makes it straightforward to have *crossed* random effects – for example random effects for students and the subjects they studied. Just specify $y \sim x + (1|id) + (1|subject)$ – and be prepared to wait. (In `lme()`, specify crossed random effects by setting `random` as a `list`)
- `glmer()` uses `nlminb()` for optimization. It is very convenient, but not *that* much cleverer than coding ‘by hand’

GLMMs: using glmer()

glmer() defaults to Laplace approximation – it is fast, and usually not terribly inaccurate. But the nAGQ option lets us use more quadrature points;

```
> glmer100 <- glmer(y~offset(log(t))+base*trt+ (1 | id),data=seiz.l,family=poisson,  
nAGQ=100)  
> glmer100  
Generalized linear mixed model fit by the adaptive Gaussian Hermite approximation  
Random effects:  
 Groups Name        Variance Std.Dev.  
 id      (Intercept) 0.60638  0.7787  
Fixed effects:  
             Estimate Std. Error z value Pr(>|z|)  
(Intercept)  1.14457   0.15189   7.535 4.87e-14 ***  
baseTRUE     -0.11184   0.04688  -2.385   0.0171 *  
trt         -0.12616   0.20986  -0.601   0.5477  
baseTRUE:trt  0.10471   0.06504   1.610   0.1074
```

This takes 7s. If results differ, use a range of values; also consider different start values. Also beware rounding errors using very large nAGQ, e.g. 1000.

GLMMs: using glmer()

For bivariate random effects (takes 45s);

```
> glmer2 <- glmer(y~offset(log(t))+base*trt+ (1 + base |id),
                     data=seiz.l, family=poisson, nAGQ=20)
> glmer2
Generalized linear mixed model fit by the adaptive Gaussian Hermite approximation
Random effects:
 Groups Name        Variance Std.Dev. Corr
 id     (Intercept) 0.83131  0.91176
         baseTRUE     0.23853  0.48839 -0.661 # of random effects
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.070246  0.177571  6.027 1.67e-09 ***
baseTRUE     0.001521  0.108974  0.014  0.9889
trt        -0.255646  0.246085 -1.039  0.2989
baseTRUE:trt 0.306492  0.151657  2.021  0.0433 *
Correlation of Fixed Effects: # estimates correlation of beta-hat's
              (Intr) bsTRUE trt
baseTRUE     -0.634
trt        -0.722  0.457
bsTRUE:trt   0.455 -0.719 -0.639
```

A trick; for independent random intercepts and slopes just use $y \sim x + (1|id) + (0+x|id)$ – no intercept in second term.

GLMMs: using glmer()

Likelihood ratio tests are available through an `anova()` method;

```
> anova(glmer1, glmer2)
Data: seiz.1
Models:
glmer1: y ~ offset(log(t)) + base * trt + (1 | id)
glmer2: y ~ offset(log(t)) + base * trt + (1 + base | id)
      Df     AIC     BIC   logLik   Chisq Chi Df Pr(>Chisq)
glmer1  5  970.66  989.10 -480.33
glmer2  7  785.12  810.92 -385.56  189.55      2 < 2.2e-16 ***
```

As with LMMs, this reference distribution is conservative.

See `methods(class="merMod")` and `help("merMod-class")` for more utility functions, including;

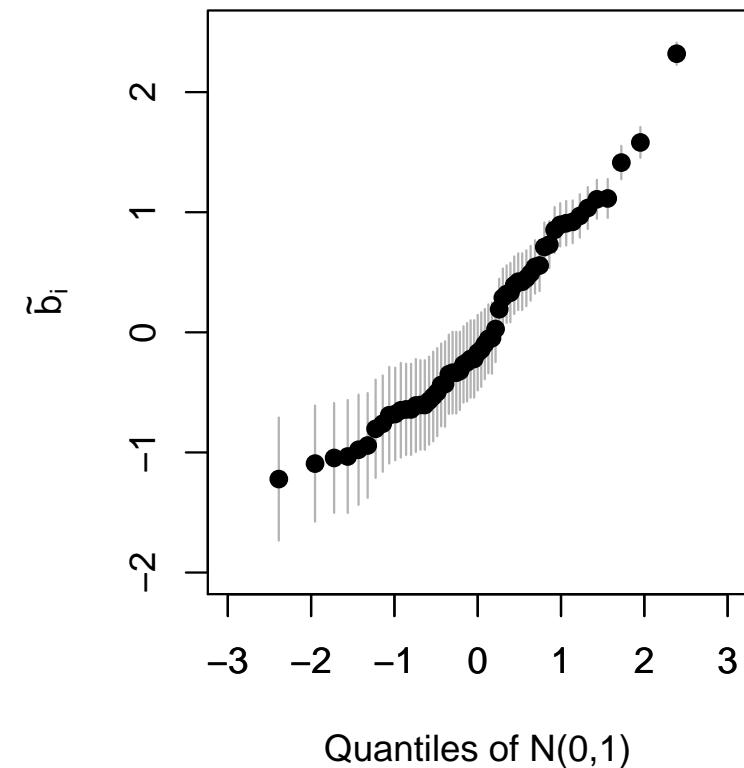
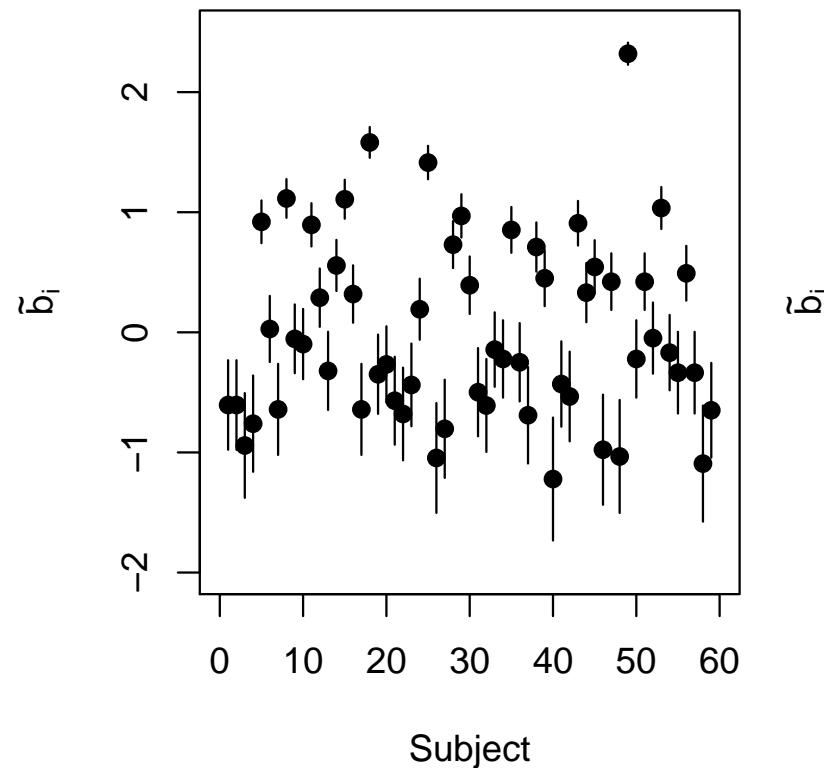
- `fixef()`: for the fixed effects estimates
- `vcov()`: for the corresponding variance estimates
- `VarCorr()`: estimated variance of the b_i
- `ranef()`: predictions of the b_i – plug-in values of the conditional **mode** of the $b_i|\mathbf{Y}, \mathbf{X}, \mathbf{Z}$ (recall BLUPs from 3.62)

Note: `lme4` uses S4 methods; `as.data.frame()` can convert to S3.

GLMMs: using `glmer()`

In `lme4`, `ranef()` gives the (approximate) conditional variance of the $b_i | \mathbf{Y}, \mathbf{X}, \mathbf{Z}$. The easiest way to get at these is with the `arm` package's `se.ranef()` function – e.g. `se.ranef(glmer1)`

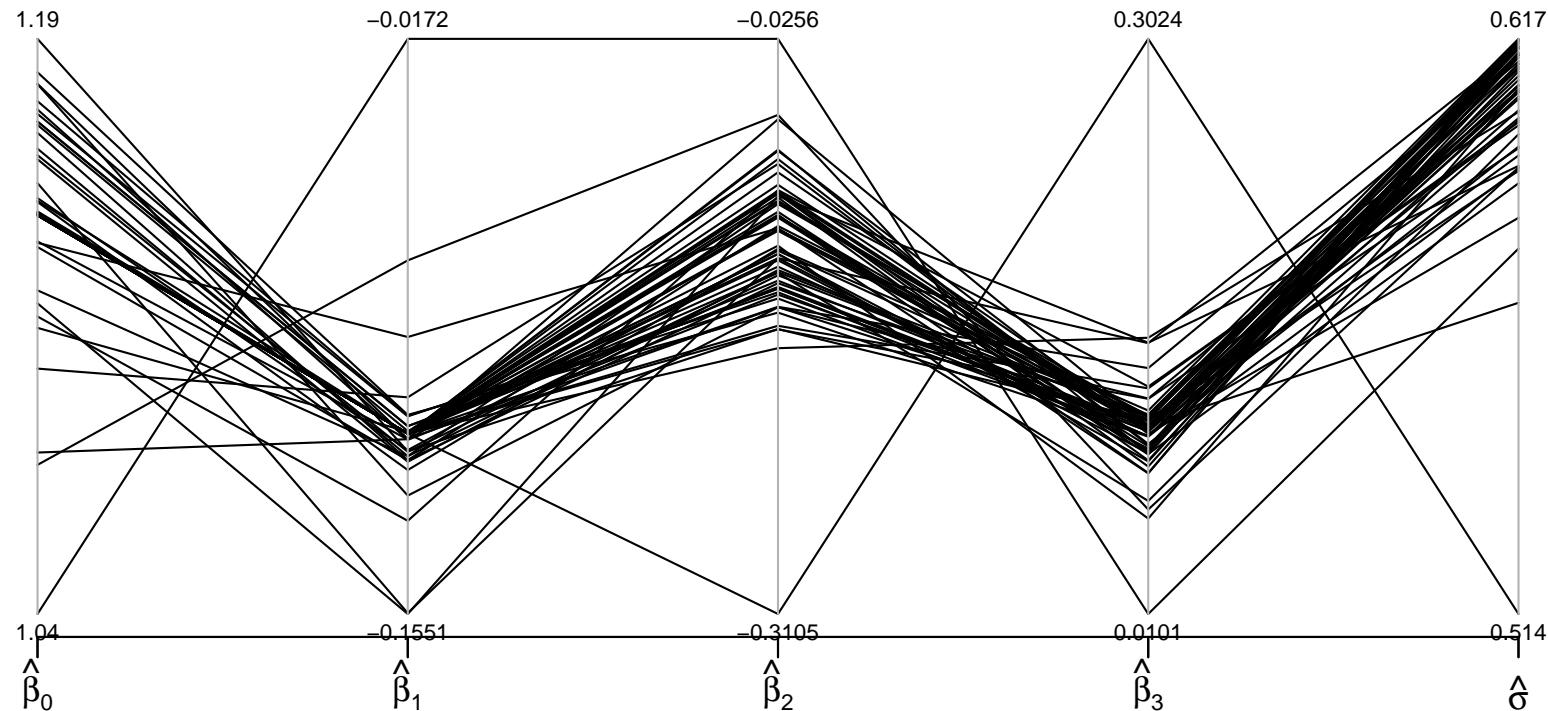
Using these, we can obtain 95% *prediction intervals*, specifying values of b_i for which the observed data would be surprising;



GLMMs: using `glmer()` – diagnostics

The QQ plot can be used as a diagnostic for the i.i.d. Normality of the b_i – though the constant variance assumption is typically much more important than the Normality.

Other familiar diagnostics also work for GLMMs. Below, we plot leave-one-out estimates of $\hat{\beta}$ and σ – which line corresponds to leaving out #49?



GLMMs: using `glmer()` – diagnostics

Also following what we saw for LMMs, fitted values (including predicted $\tilde{\mathbf{b}}_i$) and corresponding residuals can be used to perhaps diagnose problems with the mean-model assumptions.

With link function $g^{-1}()$ we define;

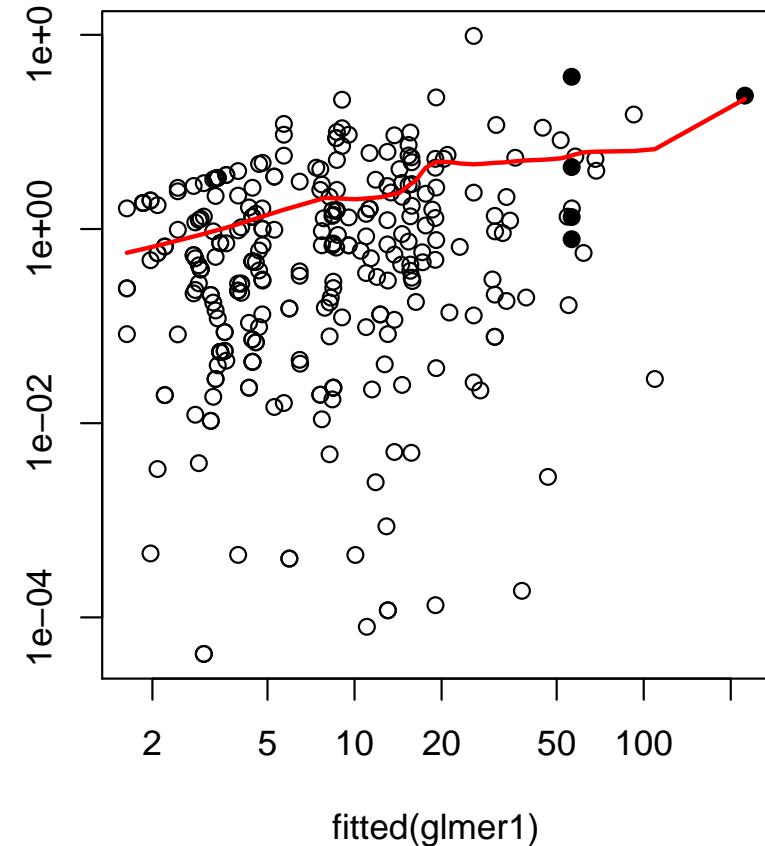
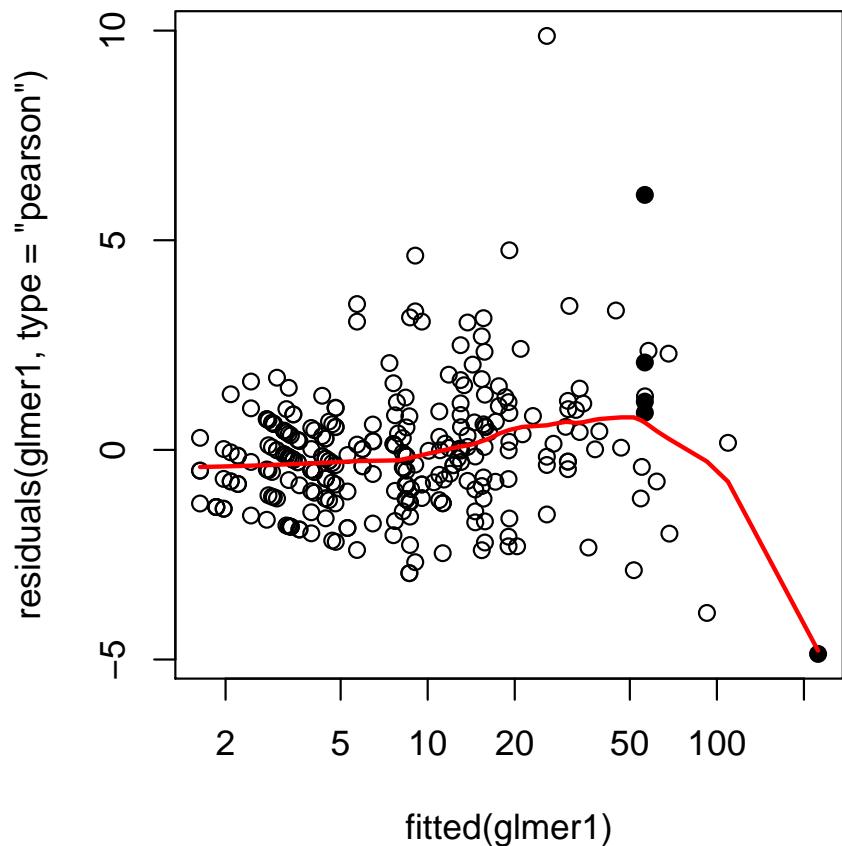
$$\begin{aligned}\text{Fitted values: (linear predictors)} \quad & \mathbf{X}_{ij}^T \hat{\boldsymbol{\beta}} + \mathbf{Z}_{ij}^T \tilde{\mathbf{b}}_i \\ \text{(responses)} \quad & \hat{Y}_{ij} = g(\mathbf{X}_{ij}^T \hat{\boldsymbol{\beta}} + \mathbf{Z}_{ij}^T \tilde{\mathbf{b}}_i) \\ \text{Residuals: (response, level 1)} \quad & \tilde{\epsilon}_{ij} = Y_{ij} - \hat{Y}_{ij} \\ \text{(Pearson, level 1)} \quad & \tilde{\epsilon}_{ij}^\dagger = \frac{Y_{ij} - \hat{Y}_{ij}}{\sqrt{\widehat{\text{Var}}[Y_{ij} | \mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i]}}\end{aligned}$$

- In R, use `fitted()` to get the fitted responses – or instead `predict(, type="link")` to get the linear predictors
- The `residuals()` method allows `type="response"` and `"pearson"`. (And also `"deviance"` – see 570 for definition)

Normalizing with Cholesky factorization is not available, nor is averaging over the \mathbf{b}_i to produce ‘stage 0’ residuals.

GLMMs: using `glmer()` – diagnostics

Plotting Pearson residuals (squared) against fitted values, with the usual smoother, highlighting #49;



The mean-variance relationship shows signs of violation – but it's hard to unpick this from a mis-specified mean.

Mixed models: other MLE tools (*)

Quadrature-based methods are most common, and for up to 5-dimensional random effects are (cautiously) recommended.

The EM algorithm[†] **used** to be standard for LMMs; we view the random effects $\{b_i\}$ and errors $\{\epsilon_i\}$ as missing data. But convergence (particularly finding the exact maximum) can be slow; using Newton-Raphson methods like `lme()` is now default.

Penalized Quasi-Likelihood (PQL, Breslow and Clayton 1993, Breslow 2003) is an iterative process for approximately fitting GLMMs; it fits a series of LMMs (using REML) using **working** dependent variables and weights that change at each iteration – thus generalizing Fisher scoring. For small clusters and with binary outcomes its variance estimates can behave poorly. See `glmmPQL()` in MASS; also note PQL routines **used** to be available in `lme4` – and now are not.

[†] see 513/the 580s for a full overview

Mixed models: induced marginal models



The numerical delicacy of fitting GLMMs (and other hierarchical models) is one non-trivial why non-experts should be extremely cautious using them.

- Compare this with GEE and/or QL, which require assumption on mean and variance – at most
- Do the results of 3.15–3.20 extend, giving (often-desirable) conditional inference from less-parametric assumptions?

Mixed models: induced marginal models

Consider the mean in **log-linear** GLMMs;

$$\begin{aligned} \mathbf{b}_i | \mathbf{X}_{ij}, \mathbf{Z}_{ij} &\stackrel{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{G}) \\ \mathbb{E}[Y_{ij} | \mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i] &= e^{\mathbf{Z}_{ij}^T \mathbf{b}_i + \mathbf{X}_{ij}^T \boldsymbol{\beta}} \\ \mathbb{E}[Y_{ij} | \mathbf{X}_{ij}, \mathbf{Z}_{ij}] &= \mathbb{E}\left[e^{\mathbf{Z}_{ij}^T \mathbf{b}_i}\right] e^{\mathbf{X}_{ij}^T \boldsymbol{\beta}} = e^{\frac{1}{2}\mathbf{Z}_{ij}^T \mathbf{G} \mathbf{Z}_{ij} + \mathbf{X}_{ij}^T \boldsymbol{\beta}} \end{aligned}$$

The RHS is a form that we *might* fit, in GEE with `link="log"`

- With just a random intercept, the intercept returned by log-linear GEE of Y on x is \neq fixed-effect intercept β_0 , but the other coefficients **are** consistent for corresponding β_k
- Otherwise it gets complicated; the marginal mean model is linear in \mathbf{X}_i but has quadratic terms in \mathbf{Z} – perhaps many of them
- We **are** exploiting the log-linear mean model, and (in general) Normality of $\mathbf{b}_i | \mathbf{X}, \mathbf{Z}$. If \mathbf{X} and \mathbf{Z} share no covariates, Normality is not required

Early GLMM authors got confused on this (Grömping 1996)

Mixed models: induced marginal models

This comparison, for the progabide example:

	GLMM		GEE*		
	Est	Est SE	Est	Est SE_r	Est SE_m
β_0 : (Intercept)	1.14	0.15	1.46	0.19	1.15
β_1 : base	-0.11	0.05	-0.11	0.12	0.69
β_2 : trt	-0.13	0.21	-0.08	0.36	1.62
β_3 : base:trt	0.10	0.07	0.10	0.21	0.97
σ : (RE std dev)	0.779				
α : (Exch corr'n)			0.771		

- Assuming log-linear $\mathbb{E}[\mathbf{Y}|\mathbf{X}, \mathbf{Z}, \mathbf{b}]$ and i.i.d Normal $b|\mathbf{X}, \mathbf{Z}$, $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ are just different estimators of the **same** parameters
- β_0^{GLMM} and β_0^{GEE} are different parameters
- GEE_r provide valid inference on $\beta_1, \beta_2, \beta_3$ with no further assumptions
- GEE_m requires that $\text{Var}[\mathbf{Y}_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}] = \phi \mathbb{E}[\mathbf{Y}_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}]$ and that $\text{Corr}[\mathbf{Y}_i|\mathbf{X}_i, \mathbf{Z}_i]$ really is compound symmetric, with same α in every cluster

* with Poisson family, same \mathbf{X} , exchangeable \mathbf{R} , robust and model-based SEs

Mixed models: induced marginal models

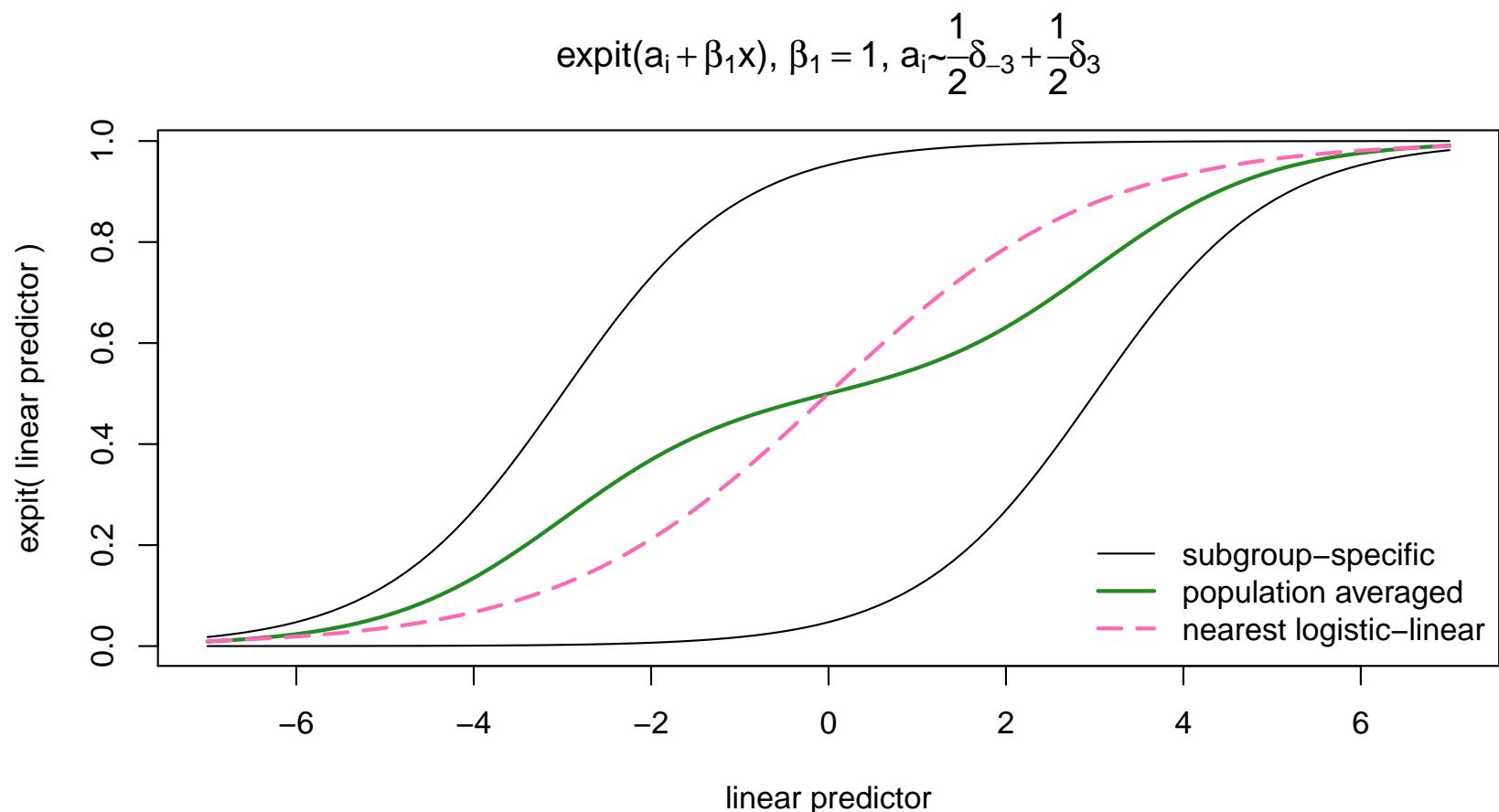
Combined with the earlier LMM discussion, we've seen that 'slope' parameters β from GLMMs with linear and log-linear link functions can be *collapsible* – that is, the parameter has the same value, regardless if we average over (sub)populations, or not. (Recall you saw this term in 570)

Unfortunately, these situations are the exception and not the rule. Essentially everything else is *non-collapsible*.

- In particular, odds ratios (defined using logistic link functions) are non-collapsible
- Under non-collapsibility, for inference on conditional β expect to have to make some assumptions (e.g. fit GLMMs, but see also Zeger and Heagerty 2000)
- Just calling GEE estimates 'biased' or 'inconsistent' for conditional β is **wrong** and **unhelpful**, if you don't say which parameter is of interest. See the discussion of Crouchley and Davies (1999) for 'free & frank exchange' of ideas

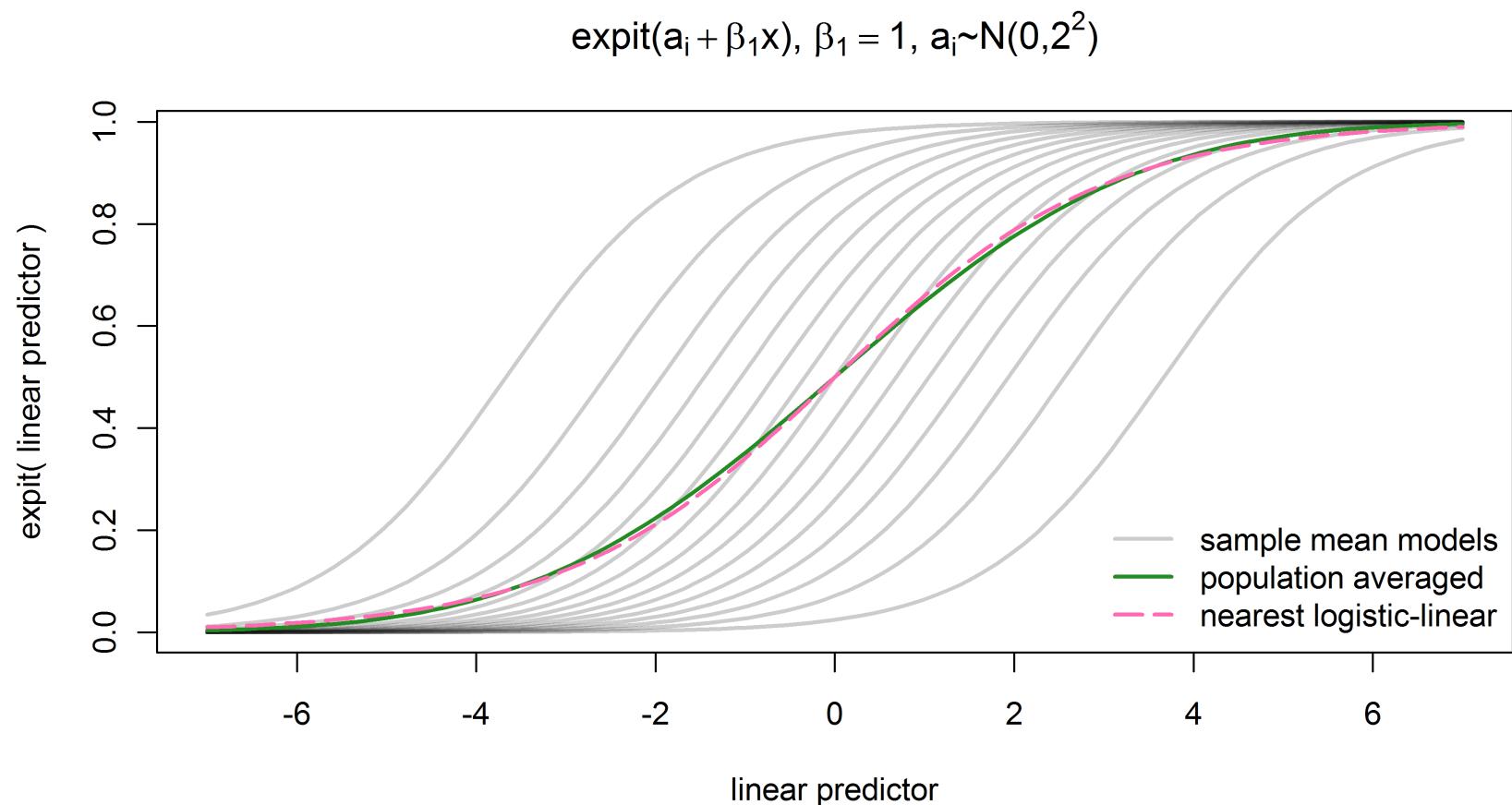
Mixed models: induced marginal models

Non-collapsibility of odds ratios, over two clusters;



Mixed models: induced marginal models

With $N(0, \sigma^2)$ random intercepts; $\beta^{\text{marg}} \approx 0.65$ but $\beta^{\text{cond}} = 1$



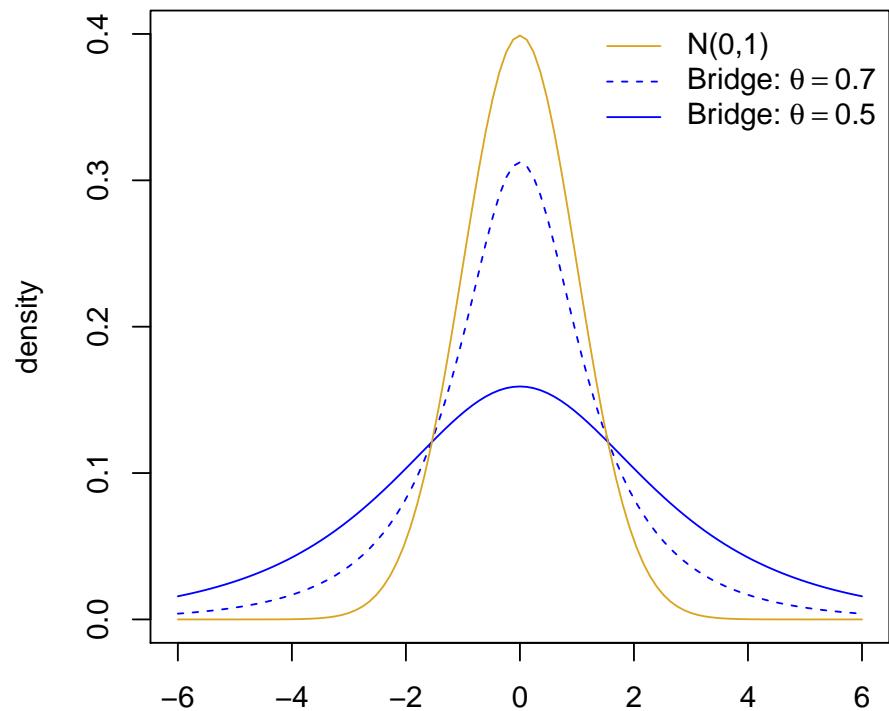
Mixed models: induced marginal models

For most random effects distributions, no closed form exists for the difference between β^{marg} and β^{cond} . But Wang and Louis (2003) reverse-engineered a ‘Bridge’ distribution that makes the relationship straightforward;

For i.i.d random intercepts a_i in logistic regression, they assume the density;

$$h(a) = \frac{1}{2\pi} \frac{\sin(\theta\pi)}{\cosh(\theta a) + \cos(\theta\pi)},$$

for $0 \leq \theta \leq 1$.



For any θ , the Bridge distribution is heavier-tailed than Normal, but has finite moments. For θ fairly near 1 it is a realistic choice

Mixed models: induced marginal models

The pretty, pretty property it was reverse-engineered to have;

$$\begin{aligned}\mathbb{E}[Y_{ij} | \mathbf{X}_{ij} = \mathbf{x}, a_i] &= \text{expit}(a_i + \mathbf{x}^T \boldsymbol{\beta}) \\ \Rightarrow \mathbb{E}[Y_{ij} | \mathbf{X}_{ij} = \mathbf{x}] &= \text{expit}(\beta_0^* + \theta \mathbf{x}^T \boldsymbol{\beta}), \text{ for some } \beta_0^*\end{aligned}$$

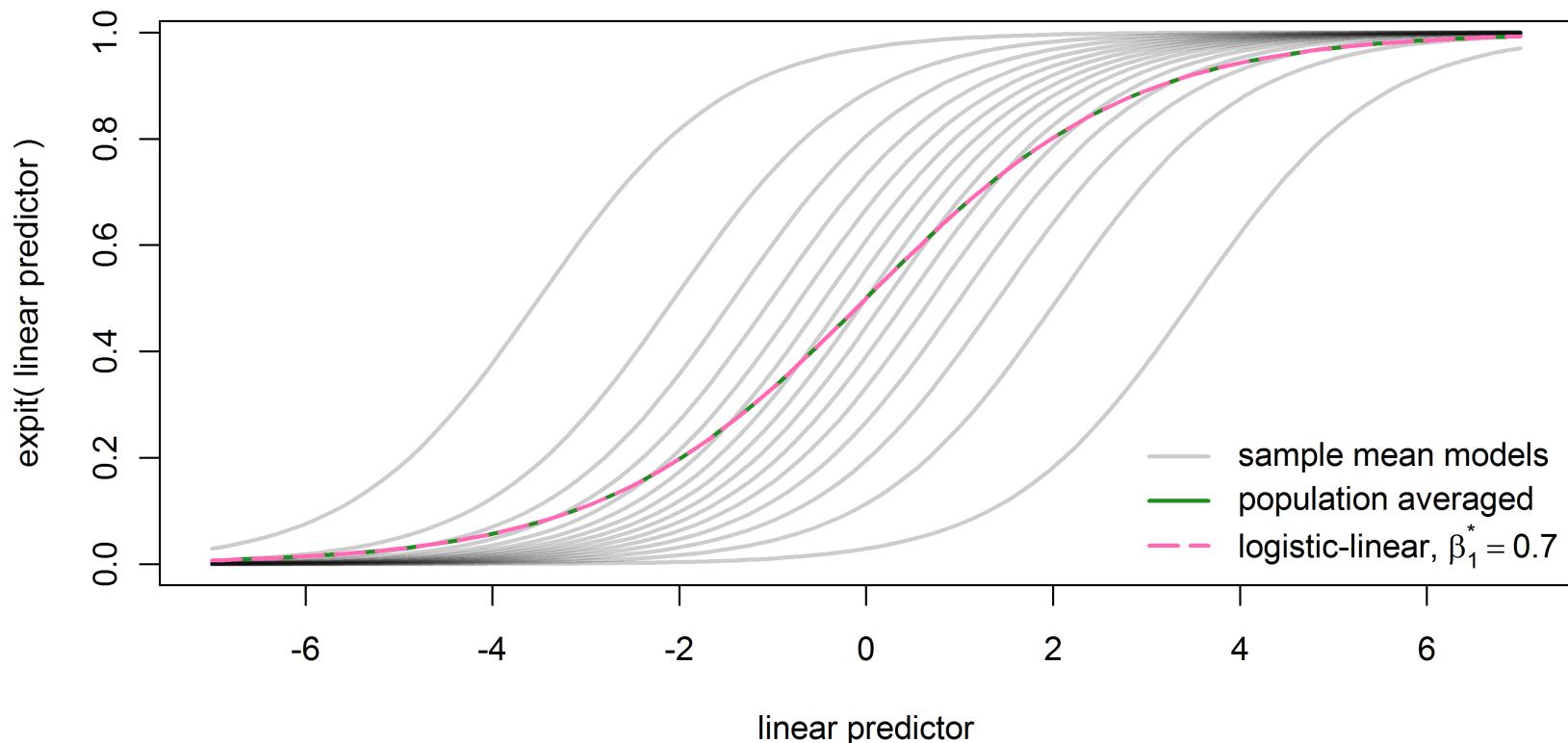
Furthermore, $\text{Corr}[Y_{ij}, Y_{ij'}] = 1 - \theta$. So if the $a_i \sim \text{Bridge}$, GEE with exchangeable \mathbf{R}_i where slope coefficients are rescaled by $(1 - \hat{\alpha})^{-1}$ estimates conditional parameter $\boldsymbol{\beta}$.

- Where $n_i \equiv 1$, no amount of data can ever distinguish $\{\theta, \boldsymbol{\beta}\}$ from $\{\lambda\theta, \lambda^{-1}\boldsymbol{\beta}\}$. So estimation of $\boldsymbol{\beta}$ is **hopeless** for $n_i \equiv 1$, without knowing the exact random-effects distribution
- With small clusters, we may learn very little about θ , meaning that estimating $\boldsymbol{\beta}^{\text{conditional}}$ remains difficult, without more information. (Also, likelihoods may not be unimodal)
- For $b_i \sim N(0, \sigma^2)$, the ‘correction’ factor depends on \mathbf{X} but is $\approx 1/\sqrt{1 + \frac{3 \times 16^2}{15^2 \pi^2} \sigma^2}$ (Zeger et al 1988) ... = 0.648 if $\sigma = 2$

Mixed models: induced marginal models

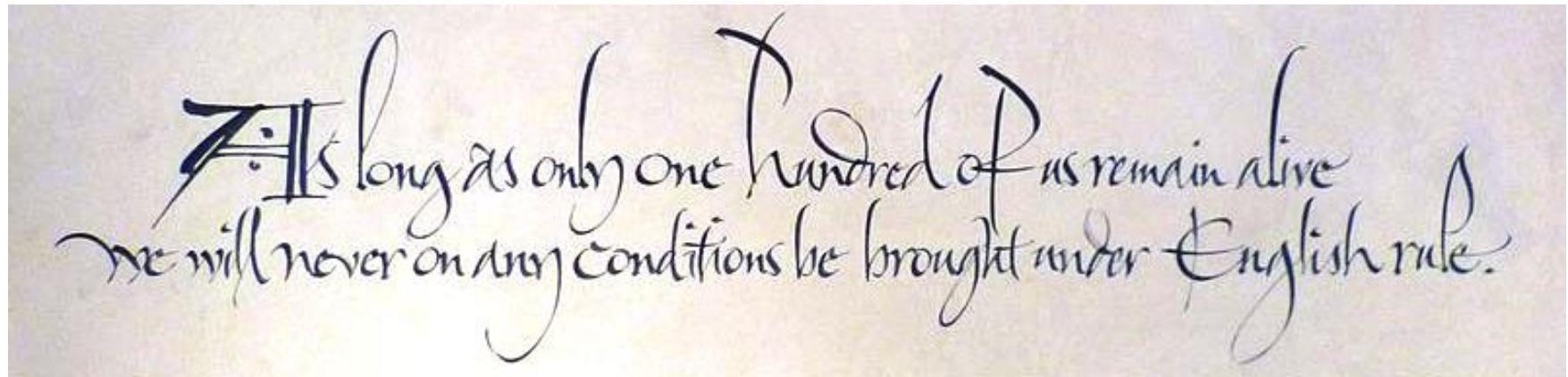
Cluster-specific means, with Bridge random intercepts;

$$\text{expit}(a_i + \beta_1 x), \beta_1 = 1, a_i \sim \text{Bridge}(\theta=0.7)$$



Mixed models: conditional likelihoods

*Quia quamdiu Centum ex nobis viui remanserint, nuncquam
Anglorum dominio aliquatenus volumus subiugari.*



From the Declaration of Arbroath (1320, right), a follow-up to the Battle of Bannockburn (1314) ... but followed by the Act of Union (1707)



Mixed models: conditional likelihoods

Like Anglo-Scottish relationships, GLMMs present long-standing difficulties. As well as trying to work round them by insisting on particular forms of link function (collapsibility) and/or random effects distributions (Bridge), another – rather drastic – approach is to use a form of likelihood that does not rely on any specification of the random effects **at all**.

- This is what *conditional likelihood* methods achieve – given certain canonical-link GLMs for $Y_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i$, it's possible to obtain inference without specifying **any** distributional assumptions for the \mathbf{b}_i . (They can even be non-random)
- This idea was introduced briefly in 570, for matched pair designs

As we'll see, the considerable robustness of this approach comes at a price – it can't be applied to every problem. It is also not available in standard software, so expect to code by hand.

Mixed models: conditional likelihoods

As well as the usual independence between clusters, the key assumptions here are that

$$\mathbb{E}[Y_{ij}|\mathbf{b}_i, \mathbf{X}_{ij} = \mathbf{x}, \mathbf{Z}_{ij} = \mathbf{z}] = g(\mathbf{x}^T\boldsymbol{\beta} + \mathbf{z}^T\mathbf{b}_i) \equiv \mu_{ij}$$

and that, through an assumed canonical-link one-parameter GLM, the likelihood contribution for each cluster can be written

$$L_i \propto \prod_{j=1}^{n_i} \exp(Y_{ij}\mu_{ij} - \psi(\mu_{ij})),$$

where $\psi'(\mu) = g(\mu)$, and constants free of $\boldsymbol{\beta}$ and \mathbf{b}_i are omitted. The log-likelihood for each cluster is

$$l_i = \mathbf{Y}_i^T \mathbf{x}_i \boldsymbol{\beta} + \mathbf{Y}_i^T \mathbf{z}_i \mathbf{b}_i - \sum_{j=1}^{n_i} \psi(\mu_{ij}).$$

- Examples include Poisson family-log link, and binomial family-logistic link
- By factorization, $\sum_i \mathbf{Y}_i^T \mathbf{x}_i$ is sufficient for $\boldsymbol{\beta}$, and each $\mathbf{Y}_i^T \mathbf{z}_i$ is sufficient for \mathbf{b}_i

Mixed models: conditional likelihoods

Sufficiency means that the $\mathbf{Y}_i^T \mathbf{z}_i$ tell us everything the data can about the \mathbf{b}_i – so if we consider datasets where these are fixed (i.e. we *condition* on them), then any dependence of the remaining randomness on the \mathbf{b}_i will have gone.

More formally, and conditioning on $\mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}$ throughout;

$$\begin{aligned}\mathbb{P}[\mathbf{Y}_i | \mathbf{b}_i = \mathbf{b}] &= \mathbb{P}[\mathbf{Y}_i | \mathbf{b}_i = \mathbf{b}, \mathbf{Y}_i^T \mathbf{z}_i] \times \mathbb{P}[\mathbf{Y}_i^T \mathbf{z}_i | \mathbf{b}_i = \mathbf{b}] \\ &= \frac{\exp(\mathbf{Y}_i^T \mathbf{x}\beta + \mathbf{Y}_i^T \mathbf{z}\mathbf{b})}{\sum_{\mathbf{Y}' : \mathbf{Y}'^T \mathbf{z} = \mathbf{Y}_i^T \mathbf{z}} \exp(\mathbf{Y}'^T \mathbf{x}\beta + \mathbf{Y}'^T \mathbf{z}\mathbf{b})} \times \mathbb{P}[\mathbf{Y}_i^T \mathbf{z} | \mathbf{b}_i = \mathbf{b}] \\ &= \frac{\exp(\mathbf{Y}_i^T \mathbf{x}\beta)}{\sum_{\mathbf{Y}' : \mathbf{Y}'^T \mathbf{z} = \mathbf{Y}_i^T \mathbf{z}} \exp(\mathbf{Y}'^T \mathbf{x}\beta)} \times \mathbb{P}[\mathbf{Y}_i^T \mathbf{z} | \mathbf{b}_i = \mathbf{b}]\end{aligned}$$

where the blue elements cancel, making the first term completely free of \mathbf{b}_i . The *conditional likelihood* is the product of this term, over all clusters, viewed as a function of β . (For continuous \mathbf{Y}_i , replace summations by integration)

Mixed models: conditional likelihoods

Conditional likelihood methods just treat the product of the conditional terms (denoted $L_C = \prod_{i=1}^n L_{Ci}$) as a regular likelihood;

- The cMLE, $\hat{\beta} = \operatorname{argmax}_{\beta} L_C(\beta)$ is a natural point estimate (given the assumptions, and viewing conditioning as reasonable)
- The inverse of the conditional observed Fisher information

$$\mathcal{I}^{*-1} = - \left(\frac{\partial^2}{\partial \beta \partial \beta^T} \log(L_C(\beta)) \Big|_{\beta=\hat{\beta}} \right)^{-1}$$

provides an estimate of $\operatorname{Var}[\hat{\beta}_{cMLE}]$, that can be used to give standard errors and intervals as usual

- This approach can be justified through standard likelihood arguments. For example, intervals using $\hat{\beta}, \mathcal{I}^*$ must give asymptotic 95% coverage when sampling with $\{\mathbf{Y}_i^T \mathbf{z}_i, b_i\}$ fixed, and so gives 95% coverage overall.

Mixed models: cMLE for a 2×2 table

To show what's going on, we illustrate L_C for a single 2×2 table – i.e. n_i conditionally independent binary outcomes, with binary covariate X . The data is

	$Y_{ij} = 0$	$Y_{ij} = 1$	Total
$X_{ij} = 0$	6	9	$r_{i0} = 15$
$X_{ij} = 1$	2	10	$r_{i1} = 12$
Total	$C_{i0} = 8$	$C_{i1} = 19$	$n_i = 27$

– and as usual, the X_{ij} are viewed as fixed. Under a logistic mean model, with random intercepts, the probabilities for each cell are

	$Y_{ij} = 0$	$Y_{ij} = 1$	Total
$X_{ij} = 0$	$\frac{1}{1+\exp(\beta_0+b_i)}$	$\frac{\exp(\beta_0+b_i)}{1+\exp(\beta_0+b_i)}$	1
$X_{ij} = 1$	$\frac{1}{1+\exp(\beta_0+b_i+\beta_1)}$	$\frac{\exp(\beta_0+b_i+\beta_1)}{1+\exp(\beta_0+b_i+\beta_1)}$	1

Mixed models: cMLE for a 2×2 table

The sufficiency argument tells us to condition on a column total, C_{i1} . Denoting T_i as $\#\{(X_{ij}, Y_{ij}) = (1, 1)\}$ this gives conditional likelihood

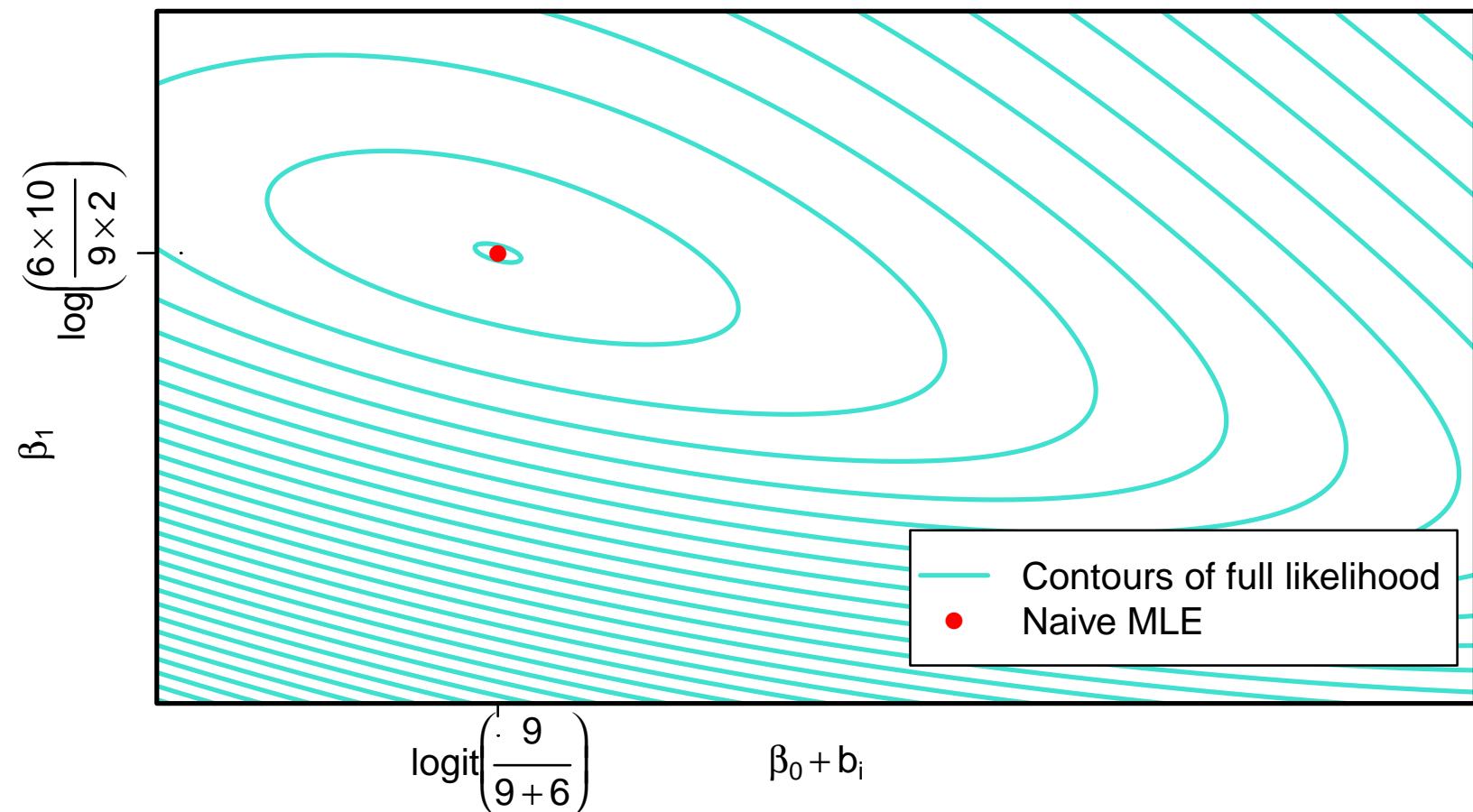
$$\begin{aligned} L_{Ci}(\beta_1) &= \frac{\binom{r_{i0}}{C_{i1}-T_i} \frac{\exp(\beta_0+b_i)^{C_{i1}-T_i}}{(1+\exp(\beta_0+b_i))^{r_{i0}}} \binom{r_{i1}}{T_i} \frac{\exp(\beta_0+b_i+\beta_1)^{T_i}}{(1+\exp(\beta_0+b_i+\beta_1))^{r_{i1}}}}{\sum_t \binom{r_{i0}}{C_{i1}-t} \frac{\exp(\beta_0+b_i)^{C_{i1}-t}}{(1+\exp(\beta_0+b_i))^{r_{i0}}} \binom{r_{i1}}{t} \frac{\exp(\beta_0+b_i+\beta_1)^t}{(1+\exp(\beta_0+b_i+\beta_1))^{r_{i1}}}} \\ &= \frac{\binom{r_{i1}}{T_i} \binom{r_{i0}}{C_{i1}-T_i} \exp(T_i\beta_1)}{\sum_t \binom{r_{i1}}{t} \binom{r_{i0}}{C_{i1}-t} \exp(t\beta_1)}, \end{aligned}$$

where t indexes all tables with the same row and column totals as observed – ‘impossible’ binomial coefficients are zero

For large values of n_i , evaluating the summands for all ‘possible’ t is a *major* hassle – an entire statistical industry has grown up around enumerating them efficiently. (The same calculations are used in Fisher’s exact test)

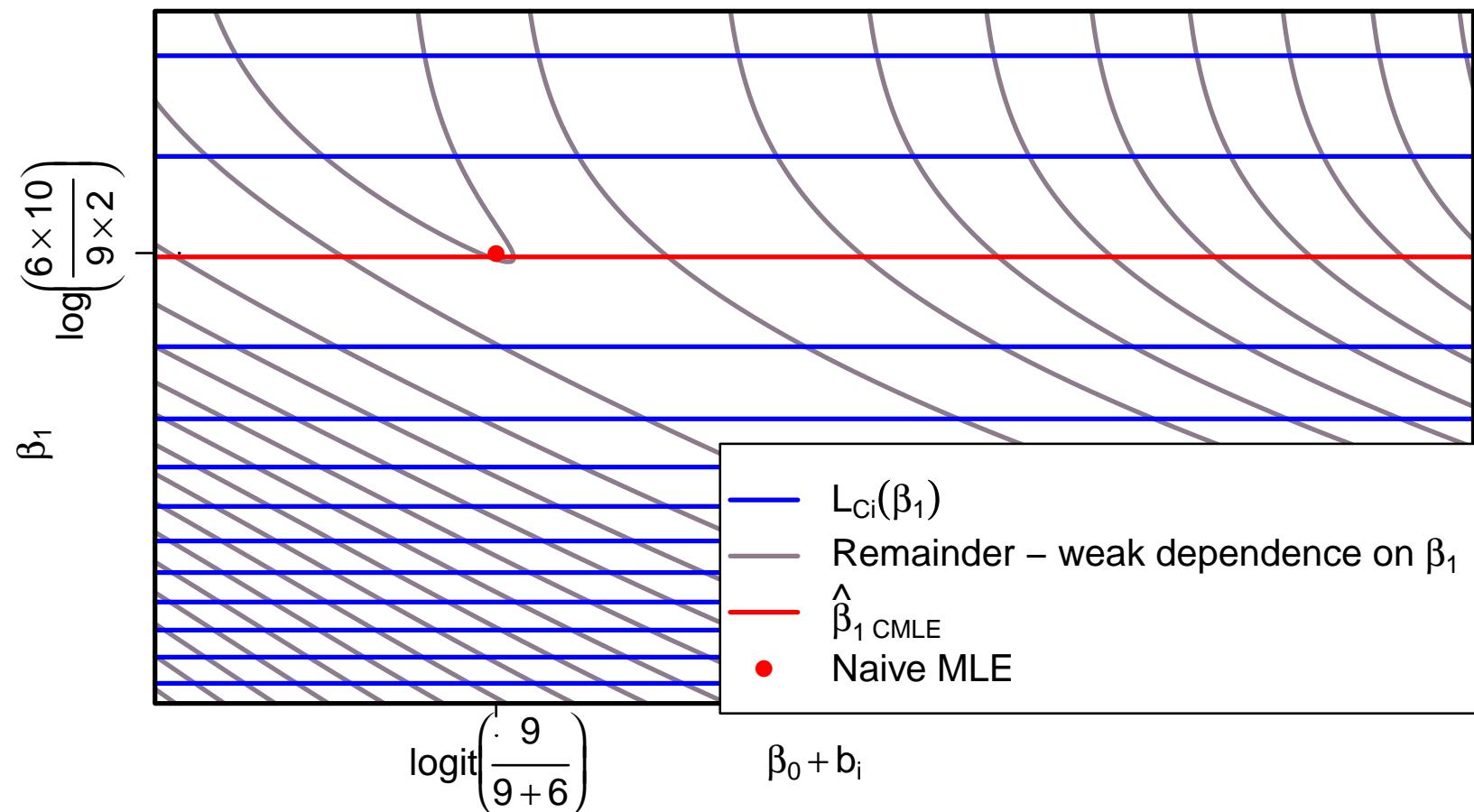
Mixed models: cMLE for a 2×2 table

For our one table, here's the full likelihood;



Mixed models: cMLE for a 2×2 table

... and how it factors into L_{Ci} and a remainder term;



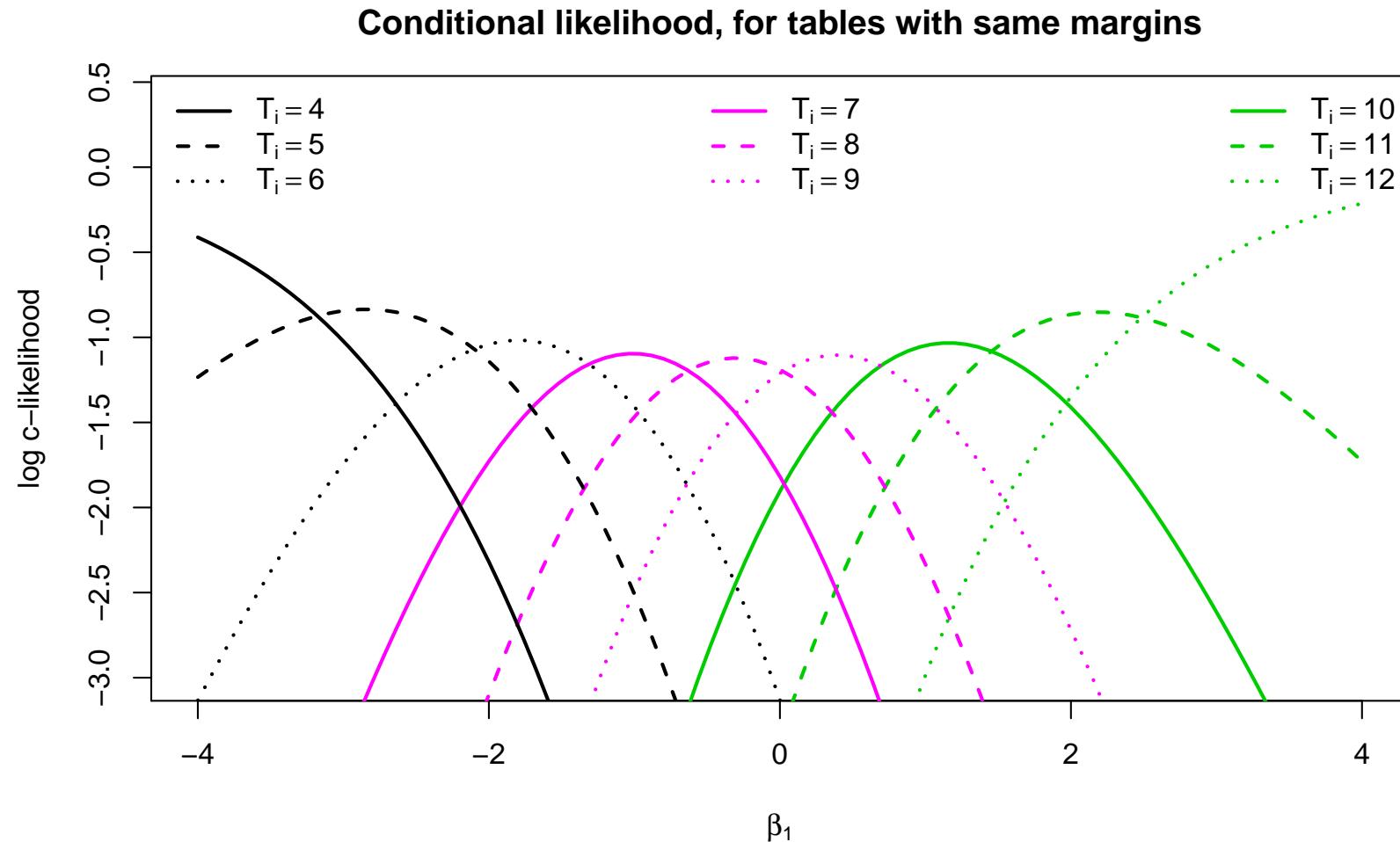
Mixed models: cMLE for a 2×2 table

- The estimates of β_1 are similar here – this is a large cluster. With many small n_i this need not happen – see the HW
- Both β_0 **and** the b_i are removed – because the intercept is in both \mathbf{X} and \mathbf{Z} – and no inference on them is possible
- As the images show, the marginal totals $\{C_{0i}, C_{1i}\}$ *do* contain information about β_1 – but perhaps not much. It cannot be accessed without making assumptions that ‘unlock’ it
- With many small clusters, and general \mathbf{X} , this method is called *conditional logistic regression* – see `clogit()` in the `survival` package

In general – not just in 2×2 tables – conditioning does not ‘buy’ us any precision; we arguably **lose** a little information about β_1 by conditioning. But compared to other estimates, the variance of the cMLE is typically easier to estimate – enabling better-calibrated inference, with small(ish) n .

Mixed models: cMLE for a 2×2 table

Plotting L_{C_i} for all tables with $\{n_i, r_{i0}, C_{i0}\} = \{27, 15, 8\}$ shows this;



With $\mathbf{Y}_i^T \mathbf{z}_i$ fixed, $\mathcal{I}^* \approx \text{Var}[\hat{\beta}]^{-1}$ is quite stable across different T_i .

Mixed models: cMLE – progabide!

We return to the progabide example, where;

$$\begin{aligned} b_i | \mathbf{X}_{ij} &\stackrel{i.i.d.}{\sim} N(0, \sigma^2) \\ Y_{ij} | b_i, \mathbf{X}_{ij} &\stackrel{\text{indept}}{\sim} \text{Poisson}(\mu_{ij}) \\ \mu_{ij} &= \exp(\log(t_j) + b_i + \beta_0 + \beta_1 \text{base}_j + \beta_2 \text{trt}_i + \beta_3 \text{base}_j \text{trt}_i) \end{aligned}$$

The cluster-specific means are

j	1	2, 3, 4, 5
$\text{trt} = 0$	$e^{\log(8) + b_i + \beta_0 + \beta_1}$	$e^{\log(2) + b_i + \beta_0}$
$\text{trt} = 1$	$e^{\log(8) + b_i + \beta_0 + \beta_1 + \beta_2 + \beta_3}$	$e^{\log(2) + b_i + \beta_0 + \beta_2}$

- Note that $Y_{i+} = \sum_{j=2}^5 Y_{ij} \sim Po(4\mu_{i2})$ is Poisson with mean $4\mu_{i2}$
- By previous theory, $\mathbf{Y}_i^T \mathbf{z}_i = T_i = \sum_{j=1}^5 Y_{ij}$ is sufficient for b_i
- Also note that $T_i \sim Po(8e^{\beta_0 + b_i}(1 + e^{\beta_1}))$ for $\text{trt}=0$, and $Po(8e^{\beta_0 + b_i + \beta_2}(e^{\beta_1} + e^{\beta_3}))$ for $\text{trt}=1$

Mixed models: cMLE – progabide!

So after conditioning, each cluster with $\text{trt}=0$ contributes

$$\begin{aligned} L_{Ci} &= \frac{\frac{e^{-8e^{\beta_0+b_i+\beta_1}}(8e^{\beta_0+b_i+\beta_1})^{Y_{i1}}}{Y_{i1}!} \frac{e^{-8e^{\beta_0+b_i}}(8e^{\beta_0+b_i})^{Y_{i+}}}{Y_{i+}!}}{e^{-8e^{\beta_0+b_i}(1+e^{\beta_1})}(8e^{\beta_0+b_i}(1+e^{\beta_1}))^{T_i}/T_i!} \\ &= \binom{T_i}{Y_{i1}} \left(\frac{e^{\beta_1}}{1+e^{\beta_1}}\right)^{Y_{i1}} \left(\frac{1}{1+e^{\beta_1}}\right)^{T_i-Y_{i1}} \end{aligned}$$

and similarly with $\text{trt}=1$ we get

$$L_{Ci} = \binom{T_i}{Y_{i1}} \left(\frac{e^{\beta_1+\beta_3}}{1+e^{\beta_1+\beta_3}}\right)^{Y_{i1}} \left(\frac{1}{1+e^{\beta_1+\beta_3}}\right)^{T_i-Y_{i1}}.$$

Together, these should be recognizable as the likelihood for unclustered logistic regression - where β_1 is the intercept, and β_3 the log odds ratio for treatment.

Mixed models: cMLE – progabide!

Doing this logistic regression in R:

```
> Ti <- aggregate( subset(seiz.l, t>1)$y, by=list(seiz.l$id), sum )[,2]
> Yi1 <- subset(seiz.l, t==8)$y
> trtbas <- subset(seiz.l, t==8)$trt # treatment at time j=1
> summary( glm( cbind(Yi1, Ti-Yi1) ~ trtbas , family=binomial) )
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.11184    0.04688 -2.386   0.017 *
trtbas       0.10473    0.06503  1.610   0.107

> summary(glmer1)
Formula: y ~ offset(log(t)) + base * trt + (1 | id)
Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.14457    0.15192  7.534 4.92e-14 ***
baseTRUE     -0.11183    0.04671 -2.394   0.0167 *
trt          -0.12615    0.20984 -0.601   0.5477
baseTRUE:trt  0.10472    0.06480  1.616   0.1061
```

- The effects of baseline and its interaction with treatment are almost identical to the GLMM's MLE
- No treatment effect (!) as this is constant over each cluster

Mixed models: cMLE – progabide!

Final notes on conditional likelihood methods;

- While extremely robust – and as efficient as they can be without further assumptions – their use is very limited. Tidy conditioning arguments that excise all b_i are only available in canonical link GLMs – this is the Pitman-Koopman-Darmois theorem
- The big penalty is that the intercept (and other terms collinear with it, within-cluster) are eliminated
- This includes the b_i , so if you want to compare different b_i – in any way – use of conditional likelihood would be “disastrous” (McCulloch, Searle and Neuhaus, pg 202)
- Use of conditional likelihood also provides some robustness against informative cluster sizes (see Neuhaus and McCulloch, 2011)

Mixed models: Bayes



While not Scottish, an inspirational and uplifting anthem...

Mixed models: Bayes

...anthem of the International Society for Bayesian Analysis;

Bayesians!
– *won't you listen to me,
I said, Bayesians!*
– *find out what you can be,
So just come on!
– to the ISBA,
It will boost your career today!*



L-R: Carlin, George, Stern, Berger

*It's fun to be in the I.S.B.A!
It's fun to be in the I.S.B.A!
You can work on your tan,
You can swim in the sea,
You can hang out with Arnie Z!*



by Jon Wakefield (1966–)
self-styled Bayesian God

Mixed models: Bayes

In essence, Bayesian analysis for mixed models just adds a prior (and the Bayesian description of uncertainty through the language of probability) onto the likelihood structures we saw earlier. So what's special about being Bayesian*, here?

- Bayes offers a natural way to say ‘these clusters are indistinguishable’, that motivates using *random* effects models even for cluster effects that were *fixed* by design
- Default Bayesian calculations can fit (i.e. give the posterior for) essentially **any** model, given time, and permit flexible inference
- No extra work/thought to estimate random effects

Of course, you may **also** have a substantive prior. Much less attractive is that the same wrong-model concerns apply here, just as with mixed model MLEs – that ‘You’ believe the model/prior may not convince Anyone Else.

* ... apart from the songs, t-shirts, hanging out with Arnie Z etc

Bayes: exchangeability of data

In 570 you saw that it's difficult to describe **lack** of information, using the (Bayesian) language of probability – use of ‘flat’ or ‘flattish’ priors may do this in practice, but they really don’t express the idea directly.

It’s much easier to express **inability to distinguish** different variables. Formally, we say an (ordered) vector of outcomes

$$\{Y_1, Y_2, Y_3, \dots, Y_n\}$$

is *exchangeable* iff it has the same distribution as any permutation of this vector. For example, the distribution of these;

$$\{Y_1, Y_2, Y_3, Y_4, Y_5\}$$

$$\{Y_2, Y_3, Y_4, Y_5, Y_1\}$$

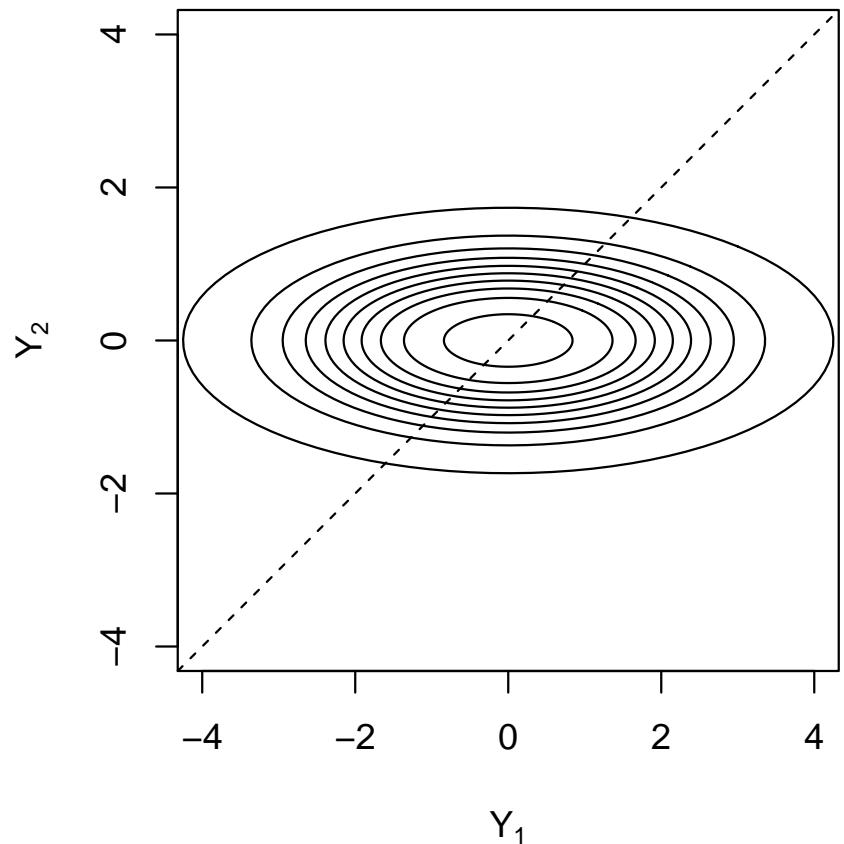
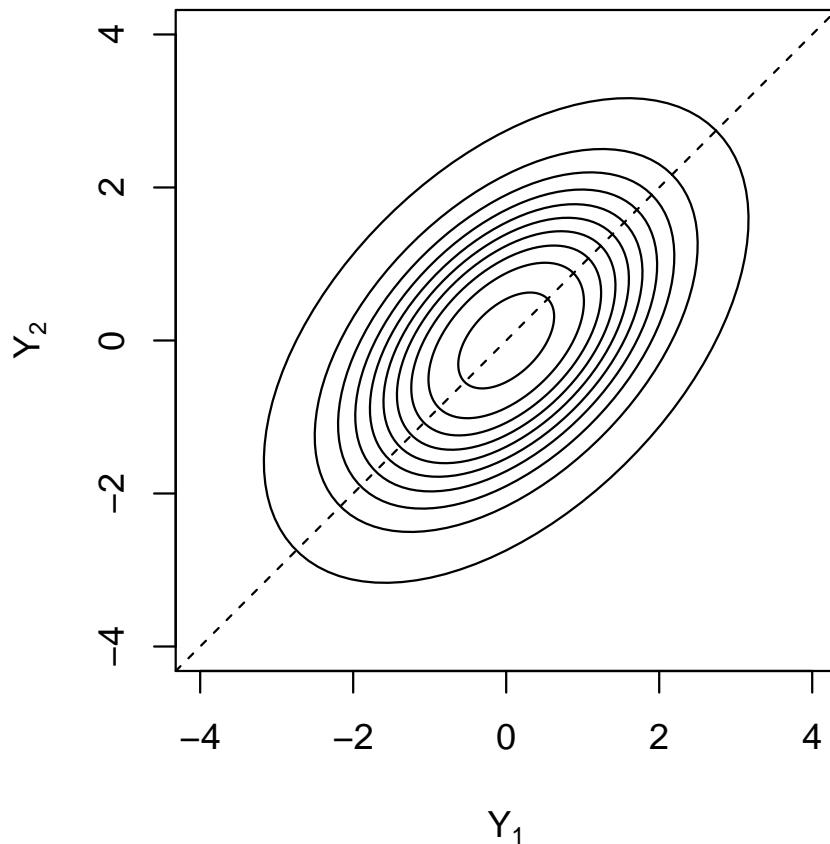
⋮

$$\{Y_{\pi(1)}, Y_{\pi(2)}, Y_{\pi(3)}, Y_{\pi(4)}, Y_{\pi(5)}\}$$

would be the same, for any permutation $\pi(\cdot)$.

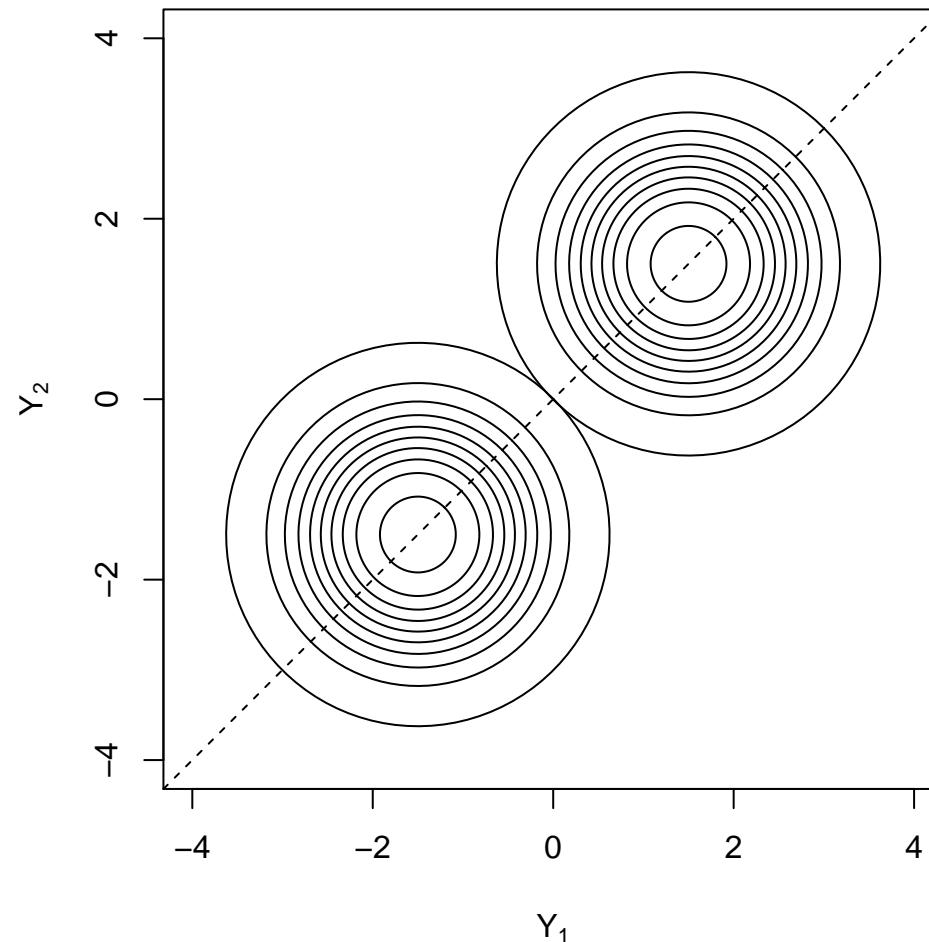
Bayes: exchangeability of data

The notation may look intimidating, but this is just a symmetry property; e.g. for density $f(y_1, y_2)$, is $f(y_1, y_2) = f(y_2, y_1)$?



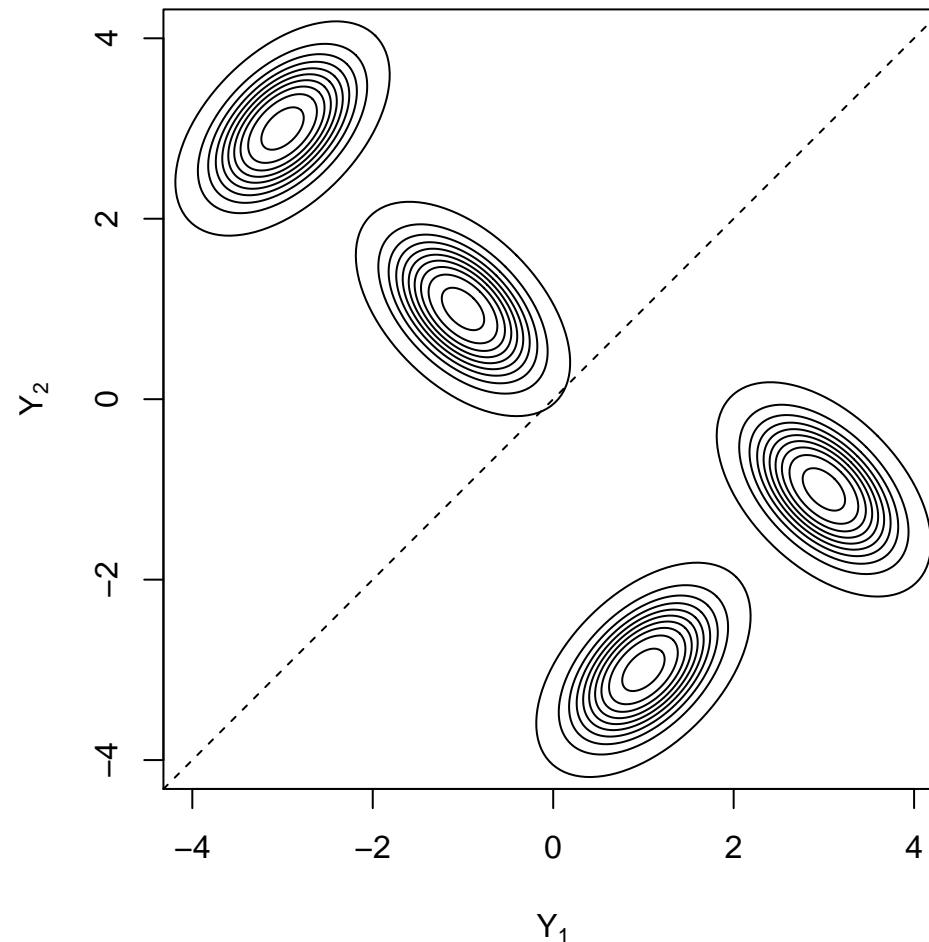
Bayes: exchangeability of data

Identically distributed? Exchangeable? I.i.d?



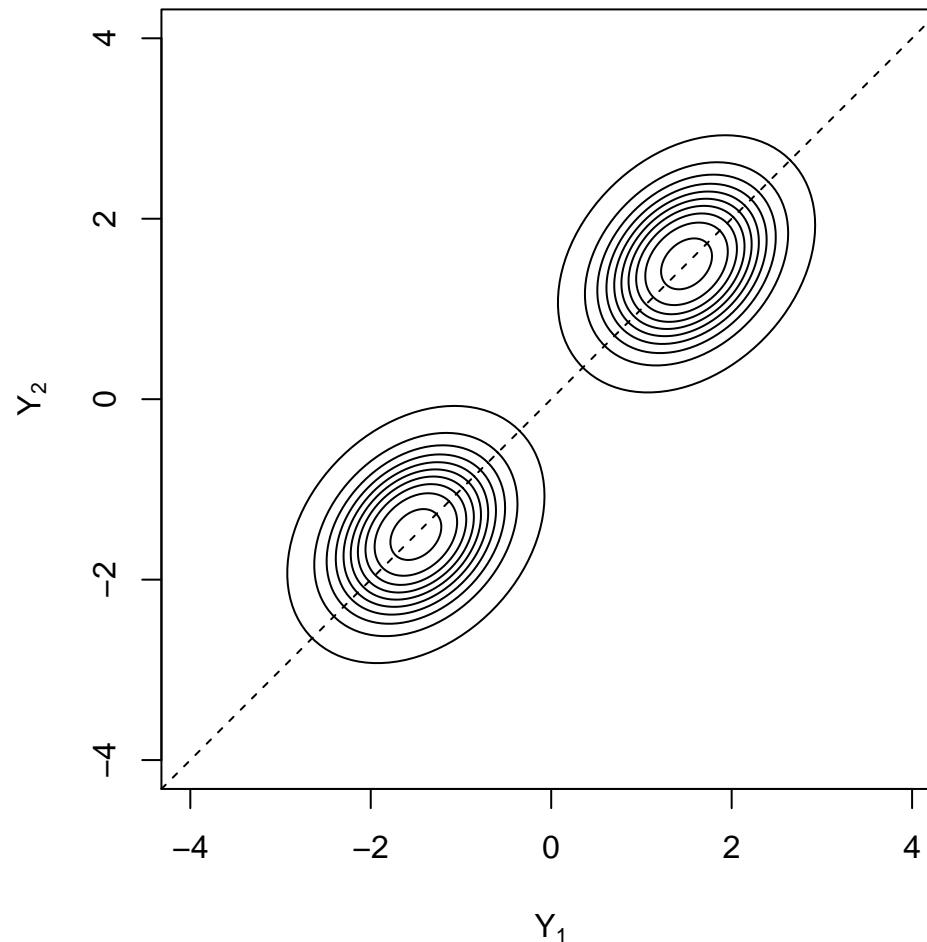
Bayes: exchangeability of data

Identically distributed? Exchangeable? I.i.d?



Bayes: exchangeability of data

Identically distributed? Exchangeable? I.i.d?

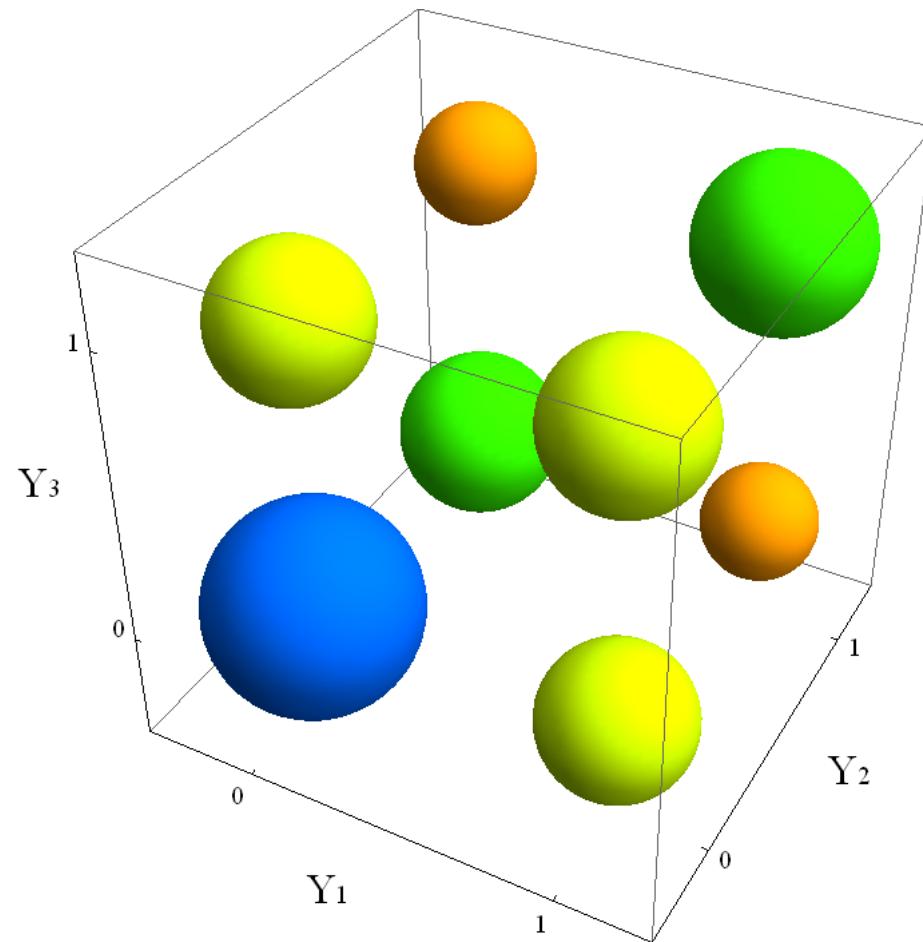


Bayes: exchangeability of data

An example with binary $\{Y_1, Y_2, Y_3\}$; (colors indicate probabilities)

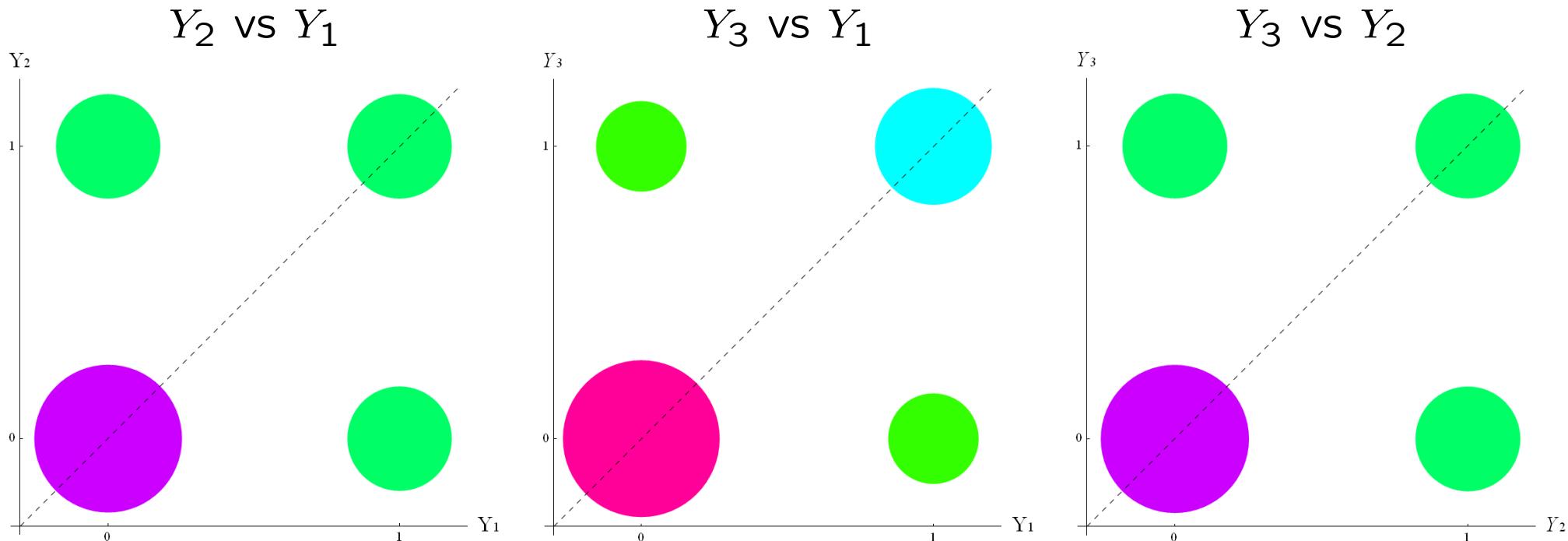
Are Y_1, Y_2, Y_3 identically distributed? Exchangeable? I.i.d?

Y_1	Y_2	Y_3	$\mathbb{P}[\mathbf{Y}]$
0	0	0	6/20
0	0	1	2/20
0	1	0	3/20
0	1	1	1/20
1	0	0	2/20
1	0	1	2/20
1	1	0	1/20
1	1	1	3/20



Bayes: exchangeability of data

Full exchangeability does not hold, but any two variables are exchangeable; (colors indicate probabilities again)



The variables $\{Y_1, Y_2, Y_3\}$ are 2-exchangeable; the concept can be generalized to n -exchangeability.

Bayes: exchangeability of data

A more realistic example; say you had (just) the rates of measles incidence in 15 counties in the US, in just one year – see 2.50. It would be unrealistic to claim these 15 outcomes are i.i.d., but **so long as their labels don't matter** we **can** claim they are exchangeable.

If measles epidemics are short-lived, and we aren't taking (effective!) steps to stop them, then outcomes over 7 years within a county might **also** be exchangeable, i.e.;

$$\begin{aligned}\mathbb{P}[Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \dots, Y_{i7} = y_{i7}] \\ = \mathbb{P}[Y_{i1} = y_{i\pi(1)}, Y_{i2} = y_{i\pi(2)}, \dots, Y_{i7} = y_{i\pi(7)}], \text{ for all } \pi(\cdot).\end{aligned}$$

However, auto-regressive properties could violate this, e.g. in post-epidemic years we assign many more resources to disease prevention. But the vector-valued county-specific outcomes $\{\mathbf{Y}_i\}$ may still be exchangeable, here.

Bayes: exchangeability of data

For outcomes $\{Y_1, Y_2, \dots, Y_n\}$ it should be reasonably clear that

i.i.d|random effects \Rightarrow exchangeable \Rightarrow identically distributed.

For the LH statement, recall models like;

$$\begin{aligned} b_i &\stackrel{i.i.d.}{\sim} N(0, \sigma^2) \\ \mu_{ij} &= \exp(\beta_0 + b_i) = \mu_i \\ Y_{ij}|b_i &\stackrel{indep}{\sim} \text{Pois}(\mu_i) \end{aligned}$$

Within cluster i , the $\{Y_{ij}\}$ are exchangeable; shuffling the outcomes within that cluster would not matter. Similarly, the set of vectors $\{\mathbf{Y}_i\}$ is also exchangeable; shuffling the clusters would not matter.

From the earlier graphics, we know

identically distributed $\not\Rightarrow$ exchangeable.

But what's the link between conditional i.i.d and exchangeability?

Bayes: exchangeability of data

A remarkably strong result holds;

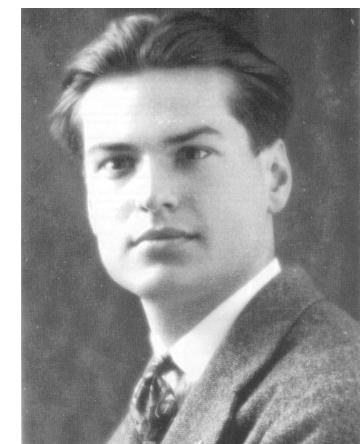
$$\text{exchangeable} \Leftrightarrow \text{i.i.d|random effects}$$

If $\{Y_1, Y_2, \dots\}$ is an exchangeable sequence of real-valued random variables, then the density* of any sample can be written

$$f(Y_1, Y_2, \dots, Y_n) = \int \prod_{i=1}^n f(Y_i|\boldsymbol{\theta}) dH(\boldsymbol{\theta}).$$

This result is *de Finetti's Theorem*.

For exchangeable data, de Finetti tells us that the only (logical) models treat the data as i.i.d. given some $\boldsymbol{\theta}$, and integrate over $H(\boldsymbol{\theta})$. ‘Being Bayesian’ just means selecting an *a priori* sensible H .



Bruno de Finetti
in 1928
3.148

*or measure, for non-continuous variables

Bayes: exchangeability of data

With exchangeable observations within exchangeable clusters, de Finetti tells us that we must be able to write the density as

$$f(\{\mathbf{Y}_i\}) = \int \prod_{i=1}^n \left(\int \prod_{j=1}^{n_i} f(Y_{ij}|\boldsymbol{\theta}_i) dH_w(\boldsymbol{\theta}_i) \right) dH_b(H_w),$$

where H_w and H_b respectively govern the distribution of random effects **within** and **between** clusters.

If, say, between-cluster heterogeneity is determined by a single random intercept, we are back to (familiar) random effects models – with a prior on the parameters in H , and β .

De Finetti's result, and many others like it, say that some form of mixed model is the only way to obtain exchangeable data; this is a strong motivation for use of mixed models.

Bayes: exchangeability of data

However...

- θ may be high-dimensional – and your data may not say much about H , as we've seen before
- De Finetti tells you **nothing** about what mixed model to use – it just says one exists, or the data wouldn't be exchangeable
- Claiming the data are exchangeable when they're not looks contrived, or lazy, or worse. For example, the weights of “butterflies in Brazil, ball-bearings in Birmingham and brussel sprouts in Belgium”* would only be considered exchangeable variables by someone **very** badly-informed
- Expect to have to justify exactly how data from all the clusters informs you (known as *borrowing strength*) about parameters specific to any particular cluster. This is not easy, and hand-waving about CLT ideas may not be enough in practice

*example due to Barnard, quoted by Harding (1972)

Bayes: exchangeability of data

Operating without covariates is somewhat stylized. In the measles data (say for one year's data), we actually know the rates of inoculation X_i and population size $n.\text{children}_i$ for each county. Given this information, exchangeability of outcomes is not plausible; the expectation of Y_i *should* reflect the 'offset' term, and (maybe) the inoculation rate.

Hence, in the absence of other knowledge, and assuming that the distribution of Y_i depends on \mathbf{X}_i only through a linear combination $\mathbf{X}_i\beta$, we might model the Y_{ij} as;

$$\begin{aligned} b_i &\stackrel{i.i.d.}{\sim} N(0, \sigma^2) \\ \mu_{ij} &= \exp(\log(n.\text{children}_{ij}) + b_i + \beta_0 + \beta_1 X_i) \\ Y_{ij}|b_i, X_{ij} &\stackrel{\text{indep}}{\sim} \text{Pois}(\mu_{ij}) \end{aligned}$$

Under this model the $\{\mathbf{Y}_i\}$ are *conditionally exchangeable*, a.k.a. *partially exchangeable*. We are assuming that, **if** the counties had all had the same inoculation rates and population sizes, their observed counts would be exchangeable.

Bayes: exchangeability in priors

We have discussed De Finetti's theorem with regard to random data. But it is just a probability result, so must also apply in Bayesian inference, where (as per slide 1.30) we use the probability calculus to describe beliefs.

De Finetti for Bayesian work;

If 'You' believe the labels on parameters $\theta_1, \theta_2, \dots, \theta_K$ don't matter, then 'Your' beliefs are equivalent to those where the θ_k are i.i.d. from some prior $\pi(\cdot)$

- Parameters describing cluster-specific effects in otherwise-indistinguishable clusters should be i.i.d. in the prior – saying they're all identical is a special case
- ... de Finetti doesn't say from what distribution the θ_k are i.i.d – but it does say one exists
- This applies *whether or not* you sampled the clusters at random

Bayes: exchangeability in priors

A Bayesian GLMM, for the measles data – with centered rate and exchangeable county effects (computations to follow)

$$\begin{aligned} 1/\sigma_b^2 &\sim \Gamma(0.05, 0.0164) \\ \beta_0, \beta_1 &\stackrel{i.i.d.}{\sim} N(0, 100) \\ b_i &\stackrel{i.i.d.}{\sim} N(0, \sigma_b^2) \\ Y_{ij}|b_i, X_{ij} &\stackrel{\text{indep}}{\sim} \text{Pois}(\mu_{ij}) \\ \mu_{ij} &= \exp(\beta_0 + \log(\text{n.children}_{ij}) + b_i + \beta_1(X_i - 69.2)) \end{aligned}$$

Estimates, compared to the analogous GLMM MLE, and GEE;

	Bayes	GLMM MLE	GEE
β_0	-8.65 (-9.3, -7.9)	-8.65 (-9.3, -8.0)	-7.94 (-8.4, -7.5)
β_1	-0.130 (-0.23, -0.04)	-0.136 (-0.21, -0.07)	-0.11 (-0.14, -0.08)
σ_b	1.36 (0.89, 2.30)	1.20	
α			-0.025

Bayesian estimates are posterior median, with central 95% credible intervals; others are default 95% confidence intervals

Bayes: exchangeability in priors

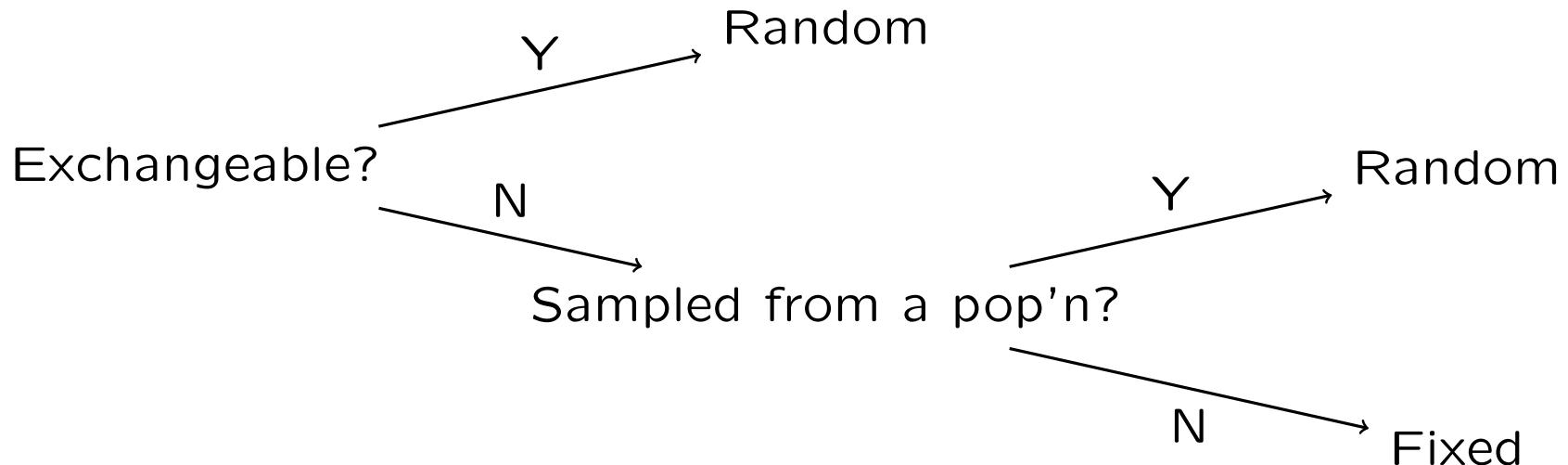
Two distinct settings you might use the Bayesian analyses – or an approximation of it;

		Random clusters	Fixed clusters
$1/\sigma_b^2$	$\sim \Gamma(0.05, 0.0164)$	Prior	Prior
β_0, β_1	$i.i.d. \sim N(0, 100)$		
b_i	$i.i.d. \sim N(0, \sigma_b^2)$	Model	Model
$Y_{ij} b_i, X_{ij}$	$indep \sim Pois(\mu_{ij})$		
μ_{ij}	$= \exp(O_{ij} + b_i + \mathbf{X}_{ij}^T \boldsymbol{\beta})$		

- In this example, it's unlikely that we sampled counties at random – so the RH column applies
- Looking at e.g. microbiota abundance in the gut of n randomly-sampled adults, the LH column might apply
- But **exactly** the same Bayesian calculation is used – and with large n the likelihood and Bayes inference for $\boldsymbol{\beta}$ and σ_b^2 will agree, by Bernstein-Von Mises

Bayes: exchangeability in priors

How we choose between fixed and random has changed, over the years. This tree is adapted from McCulloch et al's 2008 book;



Particularly with random effects models, you must also address what to estimate; the fixed effects, the variance of the random effects, or the random effects themselves – or some combination?

- See recent papers by Hodges et al describing the choice
- Using frequentist random effects inference on truly-fixed clusters *has* been suggested – see Greenland (2000)

Bayes: advanced exchangeability (*)

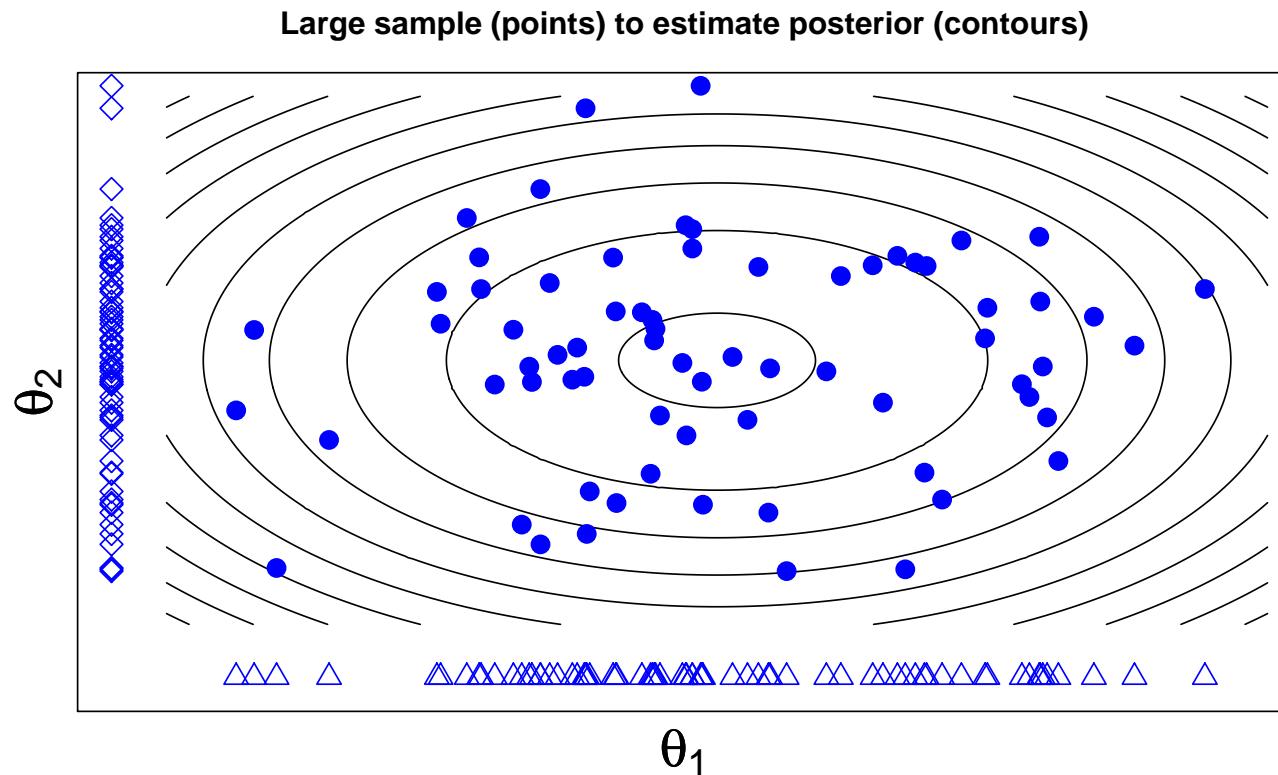
Exchangeability is arguably the most fundamental statistical property;

- Exchangeability is sufficient to prove many Central Limit Theorems; we don't need full independence
- Remarkably, exchangeability plus sufficiency of the sample mean and variance *implies* the data are *i.i.d* $N(\mu, \sigma^2)$ with some prior on μ, σ
- Pairing exchangeability with sufficiency of other statistics motivates other distributions; see Diaconis & Freedman work from the 1980s
- Poirier (2010) says this approach reflects a *healthy attitude toward parametric likelihoods*: *they are not intended to be “true” properties of reality*. Exchangeability \equiv our ignorance, and sufficiency \equiv our preference for low-dimensional summaries. Neither property may reflect reality (i.e. the true F) but both are relevant for inference.

See the Bernardo review paper, on the class site.

Bayes: computation

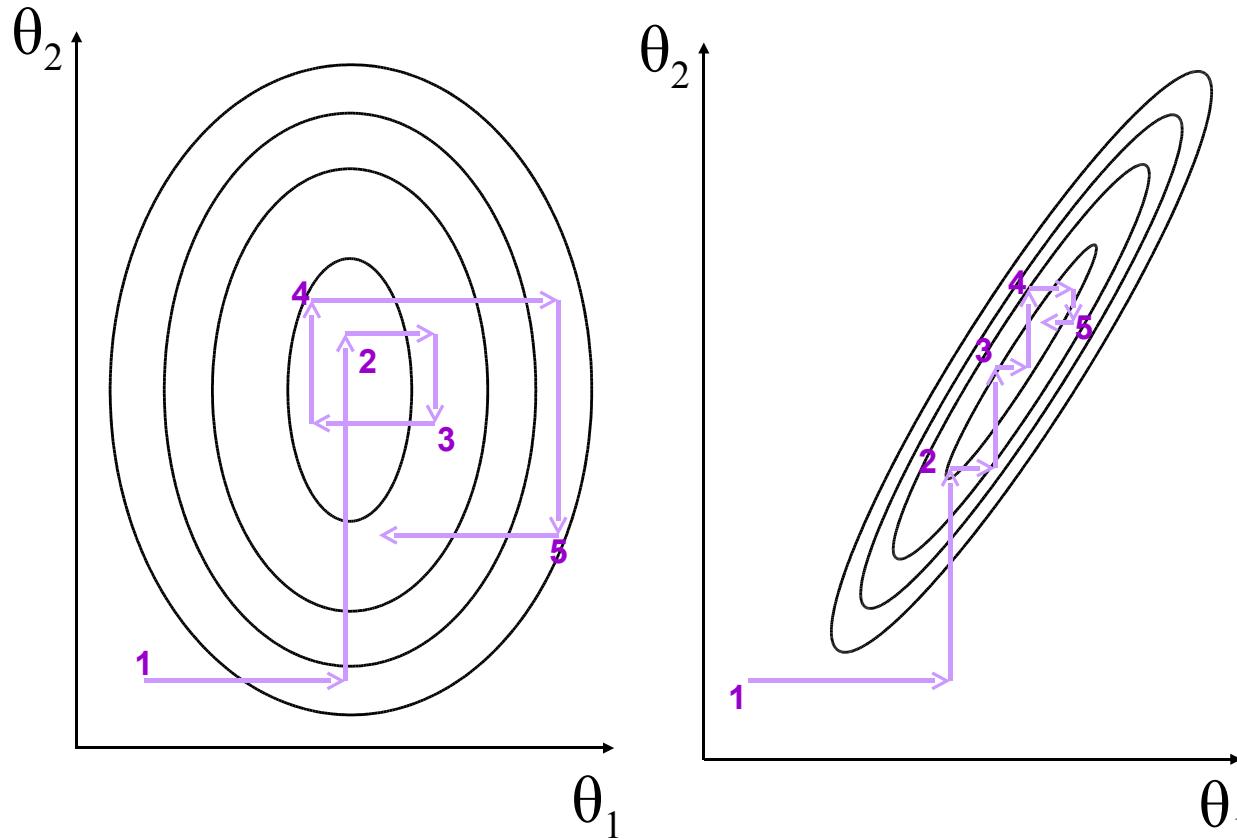
Recall (from 570) the Monte Carlo method's big picture;



We want a large sample from some distribution – i.e. the posterior. It **does not matter** if we get there by taking i.i.d. samples (see e.g. rejection sampling) or via some form of dependent sampling.

Bayes: computation

Here's one way; (the *Gibbs Sampler* – in two examples)



Gibbs updates parameters ‘one at a time’, using $\pi(\theta_1|\theta_2)$, then $\pi(\theta_2|\theta_1)$. The samples in the sequence $\boldsymbol{\theta}^{(s)}$ (a *Markov Chain*) **are** dependent, but the posterior **is** covered appropriately.

Bayes: computation

Algebra for the same thing; as you know, the posterior is

$$\pi(\theta_1, \theta_2 | \mathbf{Y}) \propto f(\mathbf{Y} | \theta_1, \theta_2) \times \pi(\theta_1, \theta_2),$$

and is usually intractable. But it is equivalent to

$$\pi(\theta_1, \theta_2 | \mathbf{Y}) = p(\theta_2 | \mathbf{Y}) \times p(\theta_1 | \theta_2, \mathbf{Y}),$$

and conditional $p(\theta_1 | \theta_2, \mathbf{Y})$ may be more readily-available.

Gibbs uses *just* the conditionals, iterating between these steps:

$$\begin{aligned}\theta_1^{(s)} &\sim p(\theta_1 | \theta_2^{(s-1)}, \mathbf{Y}) \\ \theta_2^{(s)} &\sim p(\theta_2 | \theta_1^{(s)}, \mathbf{Y})\end{aligned}$$

to produce the sequence

$$(\theta_1^{(0)}, \theta_2^{(0)}), (\theta_1^{(1)}, \theta_2^{(1)}), \dots, (\theta_1^{(s)}, \theta_2^{(s)}), \dots$$

- If the run is long enough ($s \rightarrow \infty$), this sequence *is* a sample from $\pi(\theta_1, \theta_2 | \mathbf{Y})$, no matter where you started
- For more parameters, update each θ_k in turn, then start again
- Gibbs is a (clever) special case of Metropolis-Hastings

Bayes: Gibbs sampling – rats!

Before mixed models, consider the ‘rats’ example from HW1;

$$\begin{aligned} Y_i | X_i = x_i &\stackrel{\text{indep}}{\sim} \text{Exp}\left(e^{-\beta_0 - \beta_1 x_i}\right) \\ \Leftrightarrow L(\boldsymbol{\beta}) &= \exp\left(\sum_i -\beta_0 - \beta_1 x_i - Y_i e^{-\beta_0 - \beta_1 x_i}\right) \\ &= \exp\left(-n\beta_0 - \beta_1 \sum_i x_i - e^{-\beta_0} \sum_i Y_i e^{-\beta_1 x_i}\right) \end{aligned}$$

with independent $U(-5, 5)$ priors on β_0 and β_1 .

To do Gibbs Sampling, we have to sample **somehow** from the posterior distributions of β_0 given β_1 , and β_1 given β_0 . How?

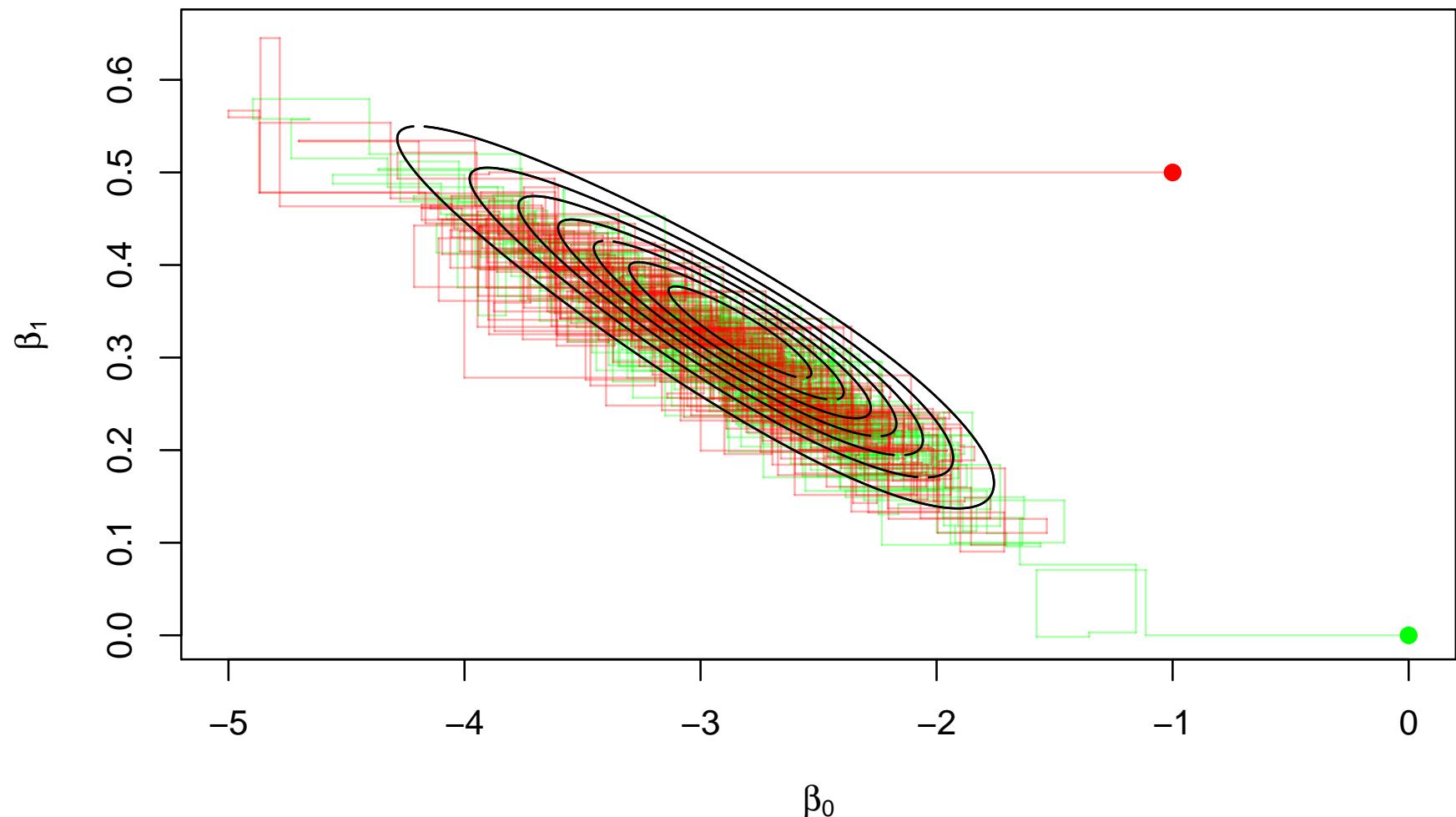
- Exploit conjugacy when possible - for example, $\beta_0 | \beta_1$ here has a (truncated) log-Gamma distribution. This is fast, but requires cleverness
- Otherwise employ brute force, e.g. some form of rejection sampling. Taking this approach here...

Bayes: Gibbs sampling – rats!

```
x <- c(6.1,4.2, 0.5,8.8,1.5,9.2,8.5,8.7,6.7,6.5,6.3,6.7, 0.2,8.7,7.5)
y <- c(0.8,3.5,12.4,1.1,8.9,2.4,0.1,0.4,3.5,8.3,2.6,1.5,16.6,0.1,1.3)
X <- cbind(rep(1, 15), x)
lik <- function(beta){ prod(dexp( y, rate=exp(X %*% beta))) }
gen.beta1 <- function(beta0){
  glm1 <- glm(y ~ -1 + offset(beta0*rep(-1,n)) + I(-x), family=Gamma(link=log))
  M1 <- lik(c(beta0, glm1$coef)) # upper limit on prior*likelihood
  repeat({ beta1.star <- runif(1, -5, 5) # reject'n sampling to get beta1|beta0
    accept.p <- lik(c(beta0, beta1.star))/M1
    if(rbinom(1, 1, accept.p)==1) break()
  })
  beta1.star}
gen.beta0 <- function(beta1){
  glm0 <- glm(y ~ -1 + rep(-1,n) + offset(-beta1*x), family=Gamma(link=log))
  M0 <- lik(c(glm0$coef, beta1))
  repeat({ beta0.star <- runif(1, -5, 5) # reject'n sampling to get beta0|beta1
    accept.p <- lik(c(beta0.star, beta1))/M0
    if(rbinom(1, 1, accept.p)==1) break()
  })
  beta0.star}
set.seed(4)
beta.sample <- matrix(NA, 1001, 2)
beta.sample[1,] <- c(0,0) # starting values
for(i in 1:1000){           # do 1000 Gibbs updates
  beta.sample[i+1,1] <- gen.beta0(beta.sample[ i,2])
  beta.sample[i+1,2] <- gen.beta1(beta.sample[i+1,1]) }
```

Bayes: Gibbs sampling – rats!

Initial chain values, for two different starting values;



Contours indicate prior \times likelihood, proportional to posterior.

Bayes: Gibbs sampling – dyestuff!

The same basic principles apply using Gibbs (or Metropolis-Hastings) to mixed models – the random effects are treated just like any other unknown. However, with more unknowns there are more full conditionals to figure out and sample from.

To illustrate this, we do Gibbs sampling for a covariate-free LMM that is balanced (all n_i equal). The model/prior are;

$$\begin{aligned}\tau_b = 1/\sigma_b^2 &\sim \Gamma(\text{shape} = \alpha_b, \text{rate} = \theta_b) \\ \tau_Y = 1/\sigma_Y^2 &\sim \Gamma(\text{shape} = \alpha_Y, \text{rate} = \theta_Y) \\ \beta_0 &\sim U(-\infty, \infty), \text{ i.e. flat - improper} \\ b_i &\stackrel{i.i.d.}{\sim} N(0, \sigma_b^2) \\ Y_{ij}|b_i &\stackrel{\text{indept}}{\sim} N(\beta_0 + b_i, \sigma_Y^2)\end{aligned}$$

The ‘inverse-Gamma’ priors on σ_b^2, σ_Y^2 lead to conjugacy.

Bayes: Gibbs sampling – dyestuff!

The full posterior (for all unknowns) is proportional to

$$\frac{e^{-\theta_b \tau_b} \tau_b^{\alpha_b - 1} \theta_b^{\alpha_b} e^{-\theta_Y \tau_Y} \tau_Y^{\alpha_Y - 1} \theta_Y^{\alpha_Y}}{\Gamma(\alpha_b) \Gamma(\alpha_Y)} \frac{e^{-\frac{\tau_b}{2} \mathbf{b}^T \mathbf{b}}}{(2\pi)^{n/2} \tau_b^{-n/2}} \prod_{i=1}^n \frac{e^{-\frac{\tau_Y}{2} (\mathbf{Y}_i - (\beta_0 + b_i) \mathbf{1})^T (\mathbf{Y}_i - (\beta_0 + b_i) \mathbf{1})}}{(2\pi)^{n_i/2} \tau_Y^{-n_i/2}}$$

Identifying terms in each unknown, we get all the conditional distributions – for each unknown given all the rest;

Conditional	Kernel	Distribution
$\tau_b \beta_0, \tau_Y, \mathbf{b}, \mathbf{Y}$	$\propto e^{-(\theta_b + \frac{1}{2} \mathbf{b}^T \mathbf{b}) \tau_b} \tau_b^{\alpha_b + \frac{n}{2} - 1}$	$\Gamma(\alpha_b + \frac{n}{2}, \theta_b + \frac{1}{2} \mathbf{b}^T \mathbf{b})$
$\tau_Y \beta_0, \tau_b, \mathbf{b}, \mathbf{Y}$	$\propto e^{-(\theta_Y + \frac{1}{2} S_{++}^2) \tau_Y} \tau_Y^{\alpha_Y + \frac{nn_i}{2} - 1}$	$\Gamma(\alpha_Y + \frac{nn_i}{2}, \theta_Y + \frac{1}{2} S_{++}^2)$
$\beta_0 \tau_b, \tau_Y, \mathbf{b}, \mathbf{Y}$	$\propto e^{-\frac{\tau_Y}{2} (nn_i \beta_0^2 - 2\beta_0 \sum_i n_i (\bar{Y}_i - b_i))}$	$N\left(\frac{\sum_i (\bar{Y}_i - b_i)}{n}, \frac{1}{nn_i \tau_Y}\right)$
$b_i \tau_b, \tau_Y, \beta_0, \mathbf{b}_{[-i]}, \mathbf{Y}$	$\propto e^{-\frac{1}{2} ((\tau_b + \tau_Y n_i) b_i^2 - 2b_i \tau_Y n_i (\bar{Y}_i - \beta_0))}$	$N\left(\frac{\tau_Y n_i}{\tau_b + \tau_Y n_i} (\bar{Y}_i - \beta_0), \frac{1}{\tau_b + \tau_Y n_i}\right)$

where $S_{++}^2 = \sum_{ij} (Y_{ij} - \beta_0 - b_i)^2$ and $\bar{Y}_i = n_i^{-1} \sum_{j=1}^{n_i} Y_{ij}$, and the Gamma distributions are again given in terms of shape and rate.

Bayes: Gibbs sampling – dyestuff!

We implement this ‘Bayesian random-effects ANOVA’ for the small dyestuff data seen in HW7 – using inverse-Gamma(0.01, 0.01) priors for the variance parameters. First, set up the prior, various constants, and a starting value;

```
dyestuff <- read.table("dyestuff.txt", header=TRUE)
# prior and constants
alphab <- 0.01
thetab <- 0.01
alphaY <- 0.01
thetaY <- 0.01
n      <- length(unique(dyestuff$batch)) #i.e. 6
ni     <- 5
Yibar <- aggregate(dyestuff$y, list(dyestuff$batch), mean)[,2]

#initial values - these only have to be vaguely sane
beta0k <- mean(dyestuff$y)
taubk  <- 1/var(dyestuff$y)
tauYk  <- 2/var(dyestuff$y)
bk     <- Yibar
```

Bayes: Gibbs sampling – dyestuff!

Then do the Gibbs Sampling, for iterations $k = 1, 2, \dots, K$:

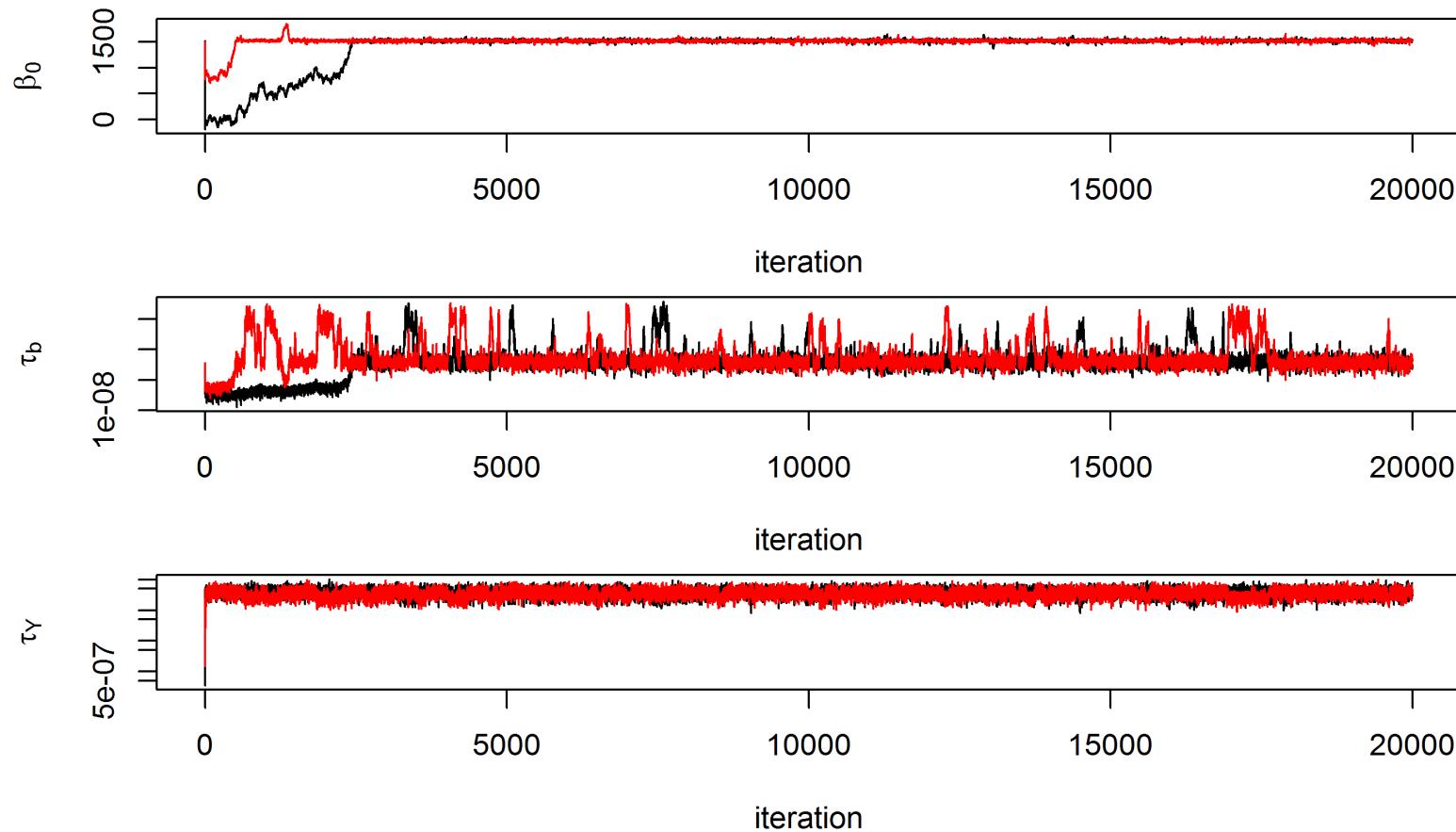
```
bigK <- 50000
posterior <- as.data.frame( matrix(NA, bigB, 9) )
names(posterior) <- c("beta0","taub","tauY",paste("b",1:6, sep=""))
set.seed(4)
posterior[1,] <- c(beta0k, taubk, tauYk, bk)

for(k in 2:bigK){
  taubk <- rgamma(1, shape=alphab+n/2, rate=thetab+sum(bk^2)/2)
  Spp <- sum( (dyestuff$y-beta0k-bk[dyestuff$batch])^2 )
  tauYk <- rgamma(1, shape=alphaY + n*ni/2, rate=thetaY+Spp/2)
  beta0k <- rnorm(1, mean(Yibar-bk), sqrt(1/(n*ni*tauYk)) )
  bk <- rnorm(n,tauYk*ni*(Yibar-beta0k)/(taubk+tauYk*ni),sqrt(1/(taubk+tauYk*ni)))
  posterior[k,] <- c(beta0k, taubk, tauYk, bk) }
```

- Set up the output **first** and then fill it in! (Or R's memory management will badly slow this, at large k)
- Everything is conjugate here, but any form of sampling could be used
- When coding, be careful with ‘intermediate’ calculations, e.g. Spp here. Updates must use most recent version of **every** other unknown

Bayes: Gibbs sampling – dyestuff!

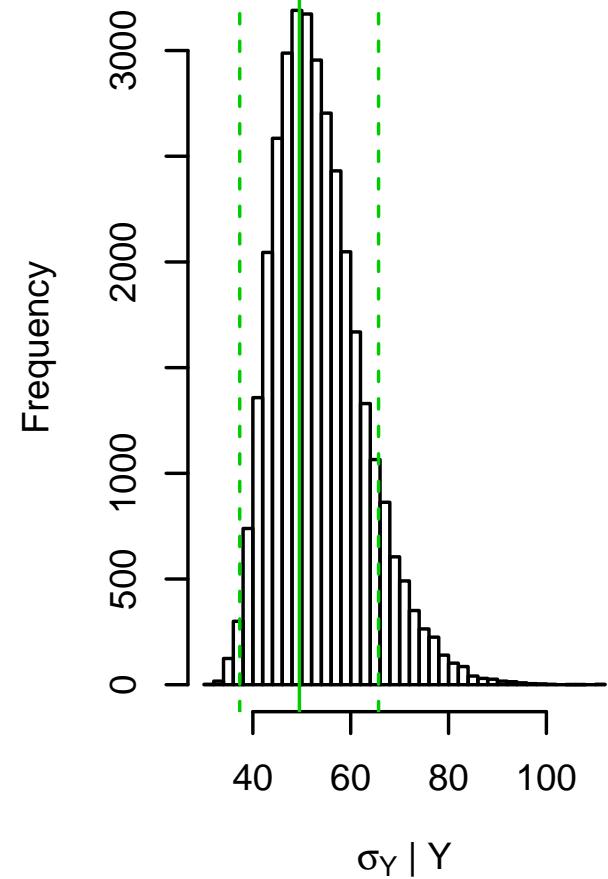
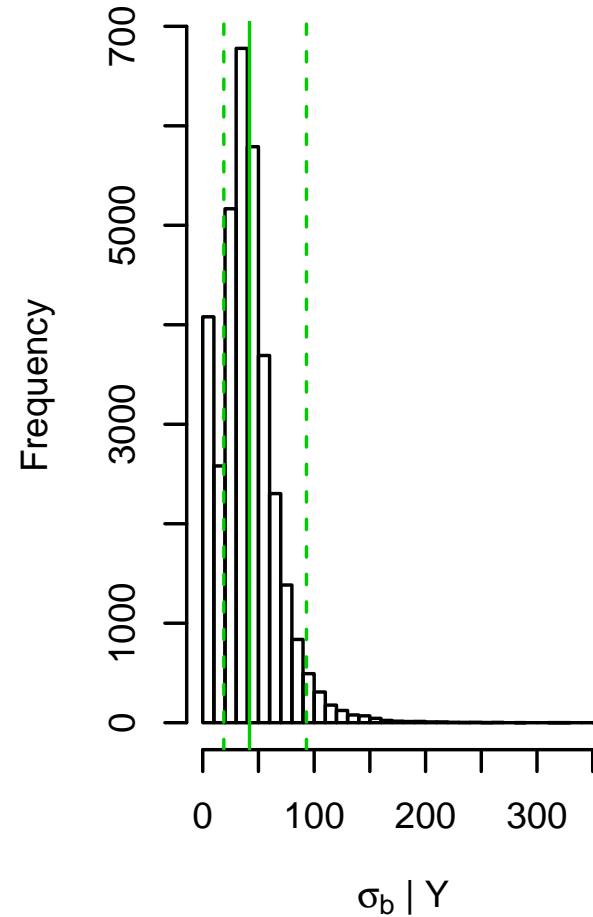
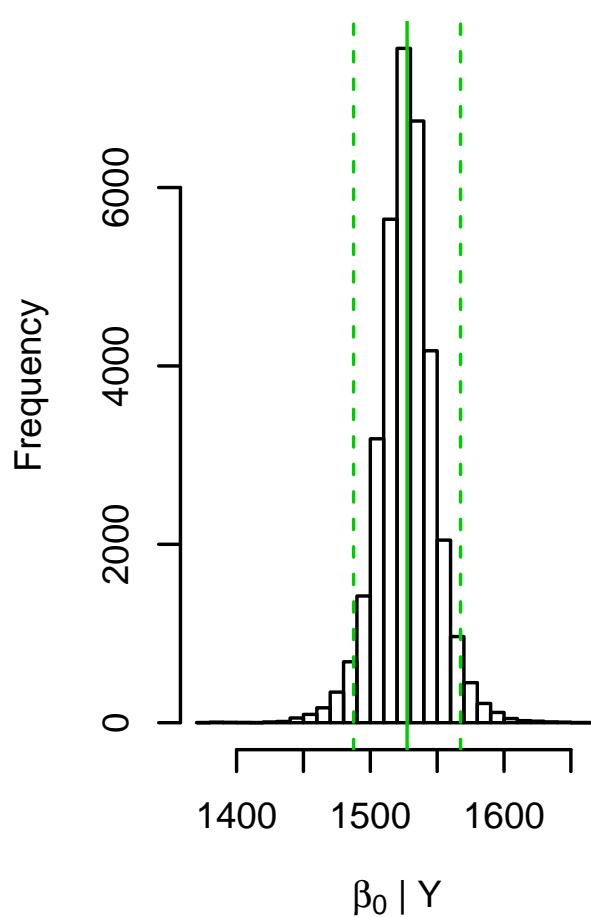
β_0, τ_b, τ_Y from two example chains with different initial values;



- ‘Converging’ means the chains’ initial values no longer matter
- Discard first 3000 ‘burn-in’ iterations (assuming b look okay)
- Output as PNGs, not PDFs!

Bayes: Gibbs sampling – dyestuff!

Posteriors for the fixed effects;



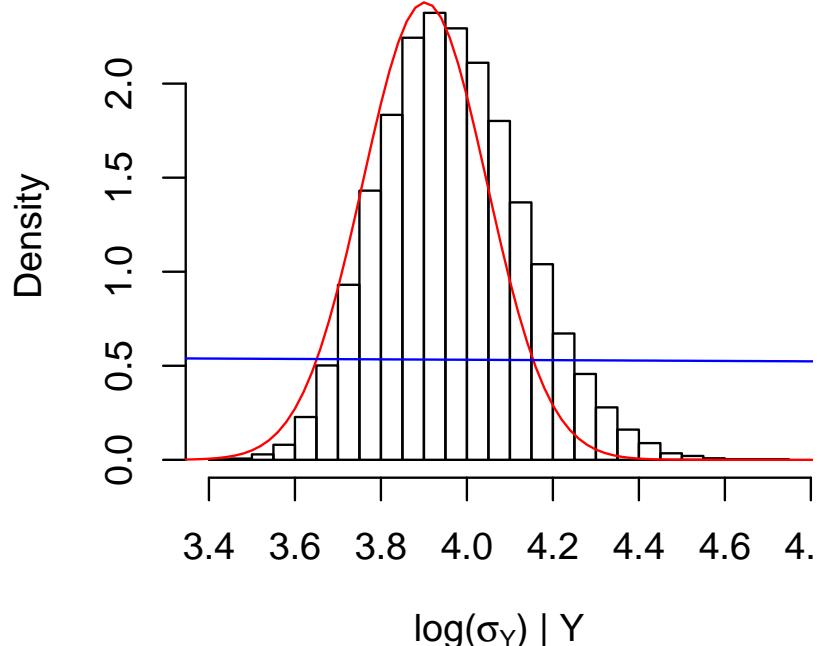
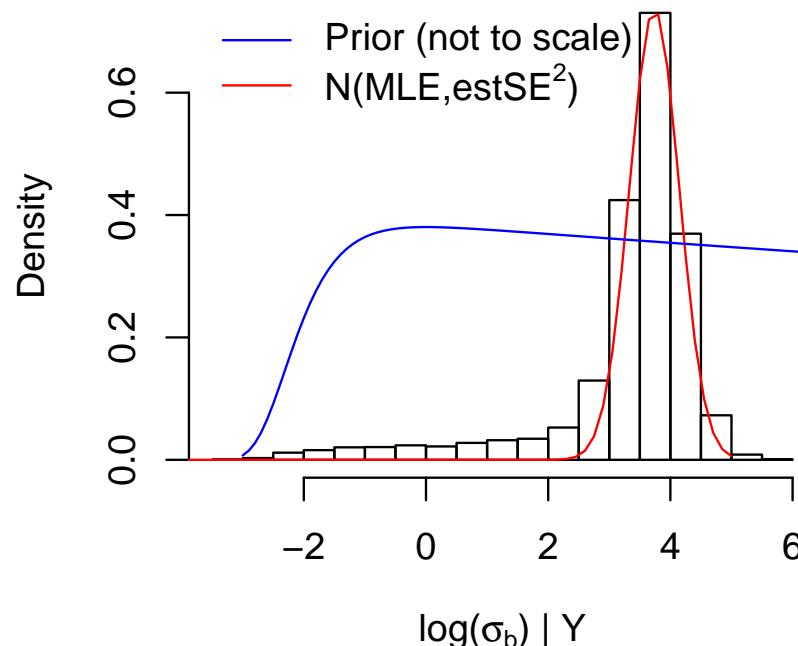
Vertical lines indicate MLE estimates and likelihood-based CIs.

Bayes: Gibbs sampling – dyestuff!

Bayesian median and 2.5%, 97.5%ile intervals, with MLE and approximate 95% CIs, and the posterior support of those CIs.

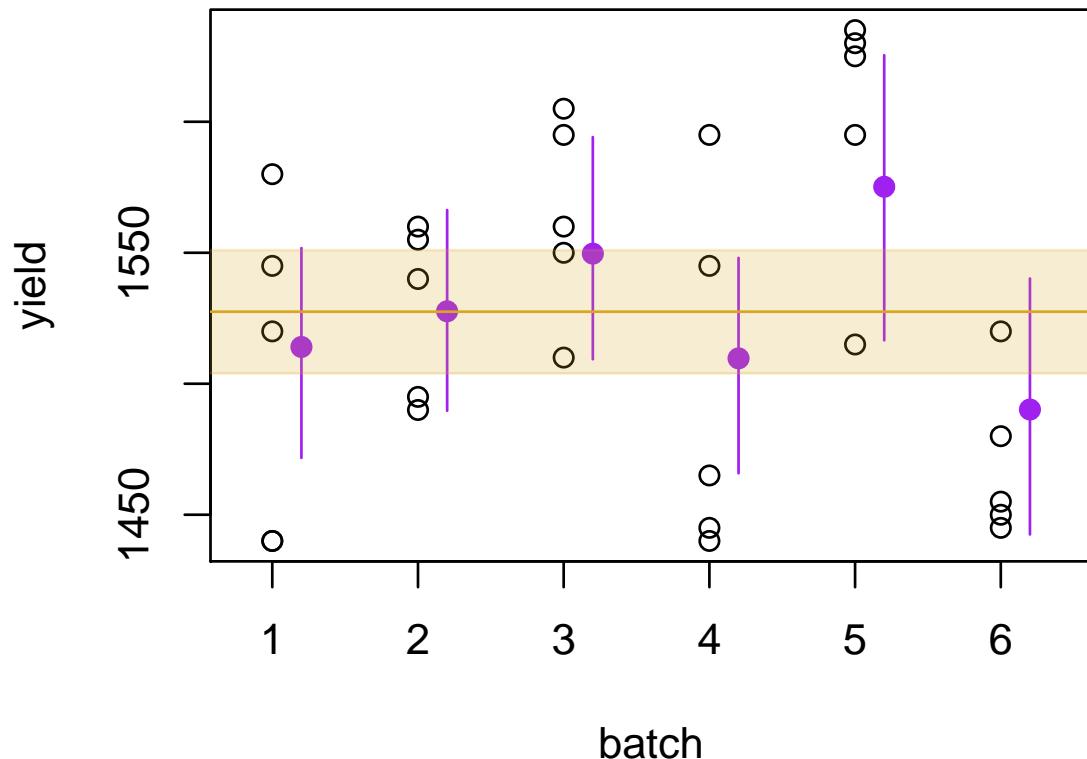
	β_0	σ_b	σ_Y
Bayes	1527 (1483,1571)	37.6 (0.4, 102)	52.3 (39.3, 74.8)
MLE	1527 (1488,1568)	42.0 (18.9, 93)	42.0 (18.9, 93.1)
Support	0.934	0.780	0.891

Q. What's going on for the variance components?



Bayes: Gibbs sampling – dyestuff!

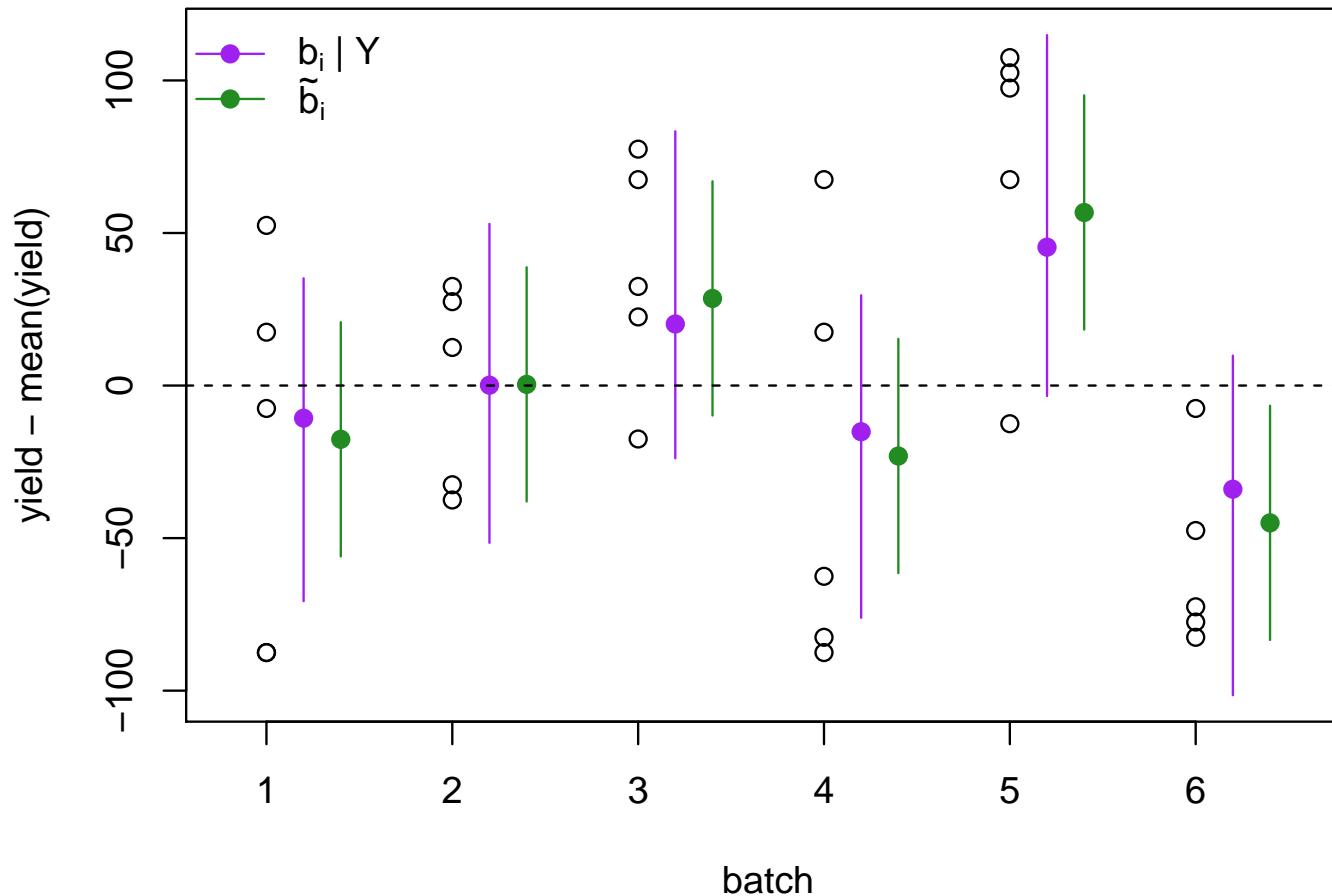
Here are the data, with credible intervals for fitted batch-specific means under the model, and MLEs \pm CI assuming $\sigma_b = 0$, i.e. truly no clustering;



- The data rules out $\sigma_Y = 0$, but not $\sigma_b = 0$
- Hence, likelihood for $\log \sigma_b$ has a heavy left tail, making the prior matter there
- Expect priors for τ_Y, τ_b to be **somewhat** informative; using $\Gamma(0.001, 0.001)$ was recommended in the past – and now isn't. Often, having **some** substantive info will really help

Bayes: Gibbs sampling – dyestuff!

Comparing Bayes and Empirical Bayes for the b_i :



The difference is partly the prior – but the empirical Bayes \tilde{b}_i uses the data twice, so can be expected to overstate precision to some extent.

Bayes: Gibbs sampling – dyestuff!

Using MCMC (and Gibbs) makes many other jobs straightforward. For example, it's easy to get;

- Posterior summaries of other functions, e.g. the batch-specific means $\beta_0 + b_i$, ICC $\sigma_b^2 / (\sigma_b^2 + \sigma_Y^2)$, etc
- Posterior summaries of non-smooth functions, e.g. the rank of each b_i , or whether $\max|b_i - b_{i'}|$ is below some tolerable threshold
- Posterior support for regions of the parameter space – e.g. $\beta > 0$, closely related to one-sided p -values (Casella & Berger 1987)
- Bayesian significance testing, comparing ‘signal-to-noise’ $\mathbb{E}[\theta|\mathbf{Y}]^2/\text{Var}[\theta|\mathbf{Y}]$ to a threshold – closely related to two-sided p -values (Rice 2010)
- Posterior predictive sampling: at each step k , generate a new b_i or Y_{ij}^* , to see the distribution of cluster effects or observations You should expect to see in a new sample, given what You know

Bayes: underflow

When coding samplers it's common that, in some regions of the parameter space Θ , evaluating the likelihood and/or prior will involve very small and/or large numbers, e.g. $\exp(-10)$, 10^{99} .

- Your computer may internally round the very small numbers to zero (underflow) and/or having trouble adding numbers of very different magnitudes
- This issue goes away as computers get better
- Almost always, any trouble occurs in parts of Θ where there is very little support – i.e. parts you can ignore, in practice
- MCMC samplers can get ‘lost’ out in the tails; numeric errors dominate the (weak) information about where to move next
- Sampling starting values from very diffuse priors (e.g. $N(0, 10^{10})$) can lead to these problems – do try to start somewhere ballpark plausible

Bayes: off-the-shelf MCMC

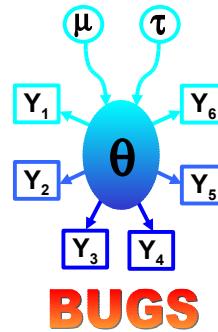
Using either Gibbs or Metropolis-Hastings, doing the algebra and/or coding to sample from the appropriate proposal distribution(s) is often non-trivial – for a complex model and/or non-smooth prior, making it work can be a large chunk of a thesis or dissertation.

But in many settings this work is also mundane; particularly for Gibbs, once we specify the model and priors, the process of getting samples from the posterior **can** be done with no original thought – we just need to compute the conditional likelihoods, and be able to sample from the corresponding density.

If you want MCMC but coding is the bottleneck – consider;

- **WinBUGS** (next)
- ... or JAGS, OpenBUGS, NIMBLE and now Stan
- For specific (G)LMMs, the **MCMCglmm** package runs chains directly, **brms** runs Stan. We'll also see INLA – as in 570

Bayes: WinBUGS



Started in 1989, the **B**ayesian analysis **U**sing **G**ibbs **S**ampling (BUGS) project has developed software where users specify only model and prior – everything else is internal. WinBUGS is the most comprehensive version.

- The model/prior syntax is very similar to R
- ... with some annoying wrinkles – variance/precision, also column major ordering in matrices
- Can be ‘called’ from R – see e.g. R2WinBUGS, but you still need to code the model

Child cancers 'not caused by Sellafield'

Before we try it on GLMMs, a tiny GLM example ($n = 1, Y = 4$);



$$Y|\theta \sim \text{Pois}(E \exp(\theta))$$

$$\theta \sim N(0, 1.797^2)$$

$$E = 0.25$$

Bayes: WinBUGS

One (sane) way to code this in the BUGS language;

```
model{  
    Y~dpois(lambda)          ...Poisson distribution, like R  
    lambda <- E*exp(theta)   ...syntax follows R  
    E <- 0.25                ...constants could go in data  
    theta~dnorm(m,tau)       ...prior for  $\theta$   
    m <- 0  
    tau <- 1/v               tau = precision NOT variance!  
    v <- 1.797*1.797  
}  
                                ...finish the model  
  
#data  
list(Y=4)                      Easiest way to input data  
#inits  
list(theta=0)                   Same list format; or use gen.inits
```

Bayes: WinBUGS

Notes on all this; (not a substitute for reading the manual!)

- This should look familiar, from the models we have been writing out. In particular ‘ \sim ’ is used to denote distributions of data *and* parameters
- All ‘nodes’ appear **once** on the LHS; hard work is done on RHS
- No formulae allowed when specifying distributions
- Data nodes *must* have distributions. Non-data nodes *must* have priors – it’s easy to forget these
- Write out regressions ‘by hand’; $\text{beta0} + \text{beta1} * \text{x1} + \dots$
- This language can’t do everything; BUGS does not allow e.g.

$\text{Y} \leftarrow \text{U} + \text{V}$

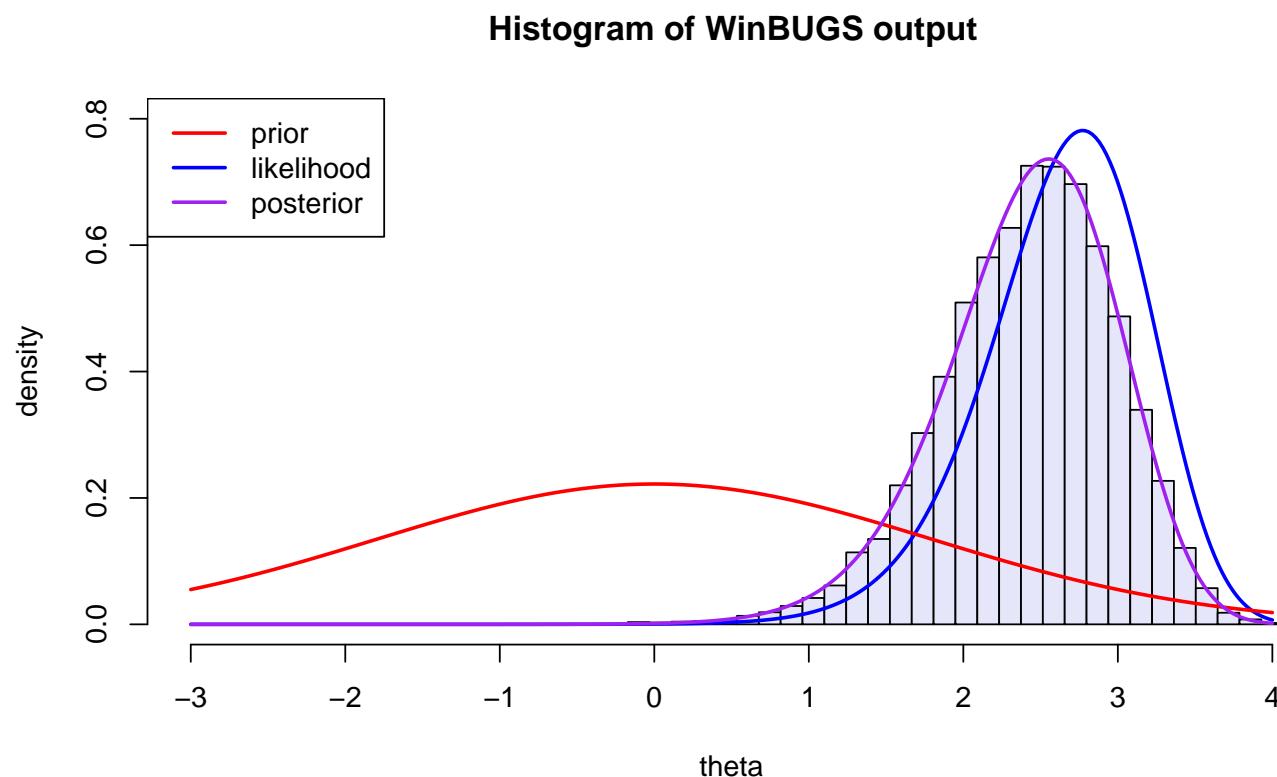
$\text{U} \sim \text{dnorm}(\text{meanu}, \tau_{\text{uu}}); \text{V} \sim \text{dt}(\text{meanv}, \tau_{\text{uv}}, k)$

#data

list(Y=...)

Bayes: WinBUGS

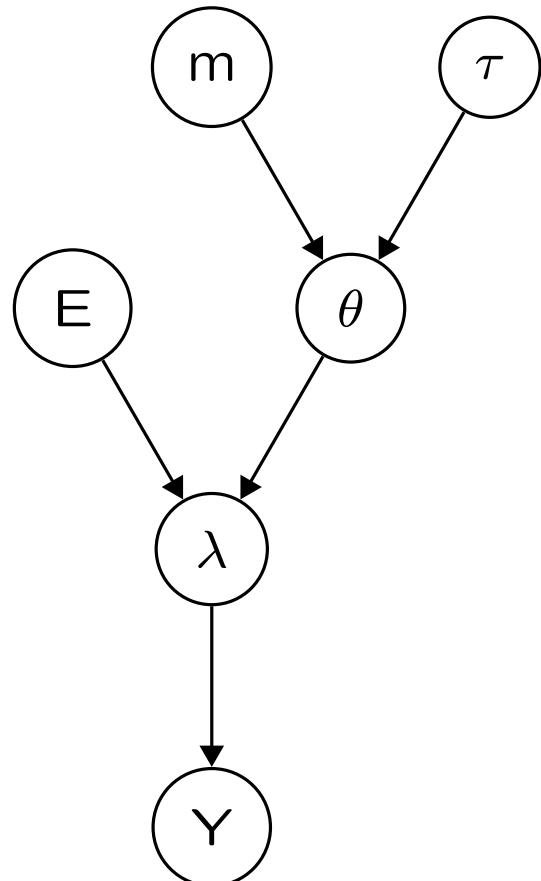
From 10,000 iterations, how do we do? (Note ‘MC error’ estimates Monte Carlo error in the posterior mean)



node	mean	sd	MC error	2.5%	median	97.5%
theta	2.422	0.5608	0.005246	1.229	2.466	3.388

Bayes: WinBUGS

Under the hood, here's how WinBUGS 'thinks':



- It's a DAG; arrows represent stochastic relationships (not causality)
- Some texts use square nodes for observed variables (Y , here)
- To do a Gibbs update, we need to know/work out the distribution of a node conditional on **only** its parents, children, and its children's other parents*.

* This set is a node's 'Markov blanket'. The idea saves a lot of effort, and is particularly useful when fitting random effects models.

Bayes: WinBUGS

- As well as the Markov blanket idea, WinBUGS uses what it knows about conjugacy to substitute closed form integrals in the calculations, where it can. (e.g. using inverse-gamma priors on Normal variances)
- Otherwise, it chooses from a hierarchy of sampling methods; Metropolis-Hastings is the last resort
- Because of this, and the complexity of turning a model into a sampling scheme, don't expect too much help from the error messages
- Even when the MCMC is working correctly, it is possible you may be fitting a ridiculous, unhelpful model. Just like R, WinBUGS' authors assume you take responsibility for that

Also note Gibbs-style sampling is not ideal for every job. But it *is* a good method to consider when fitting GLMMs.

Bayes: WinBUGS

WinBUGS model file for a GLMM – the `measles` example from 3.153, with the data in matrix format; (see 3.184)

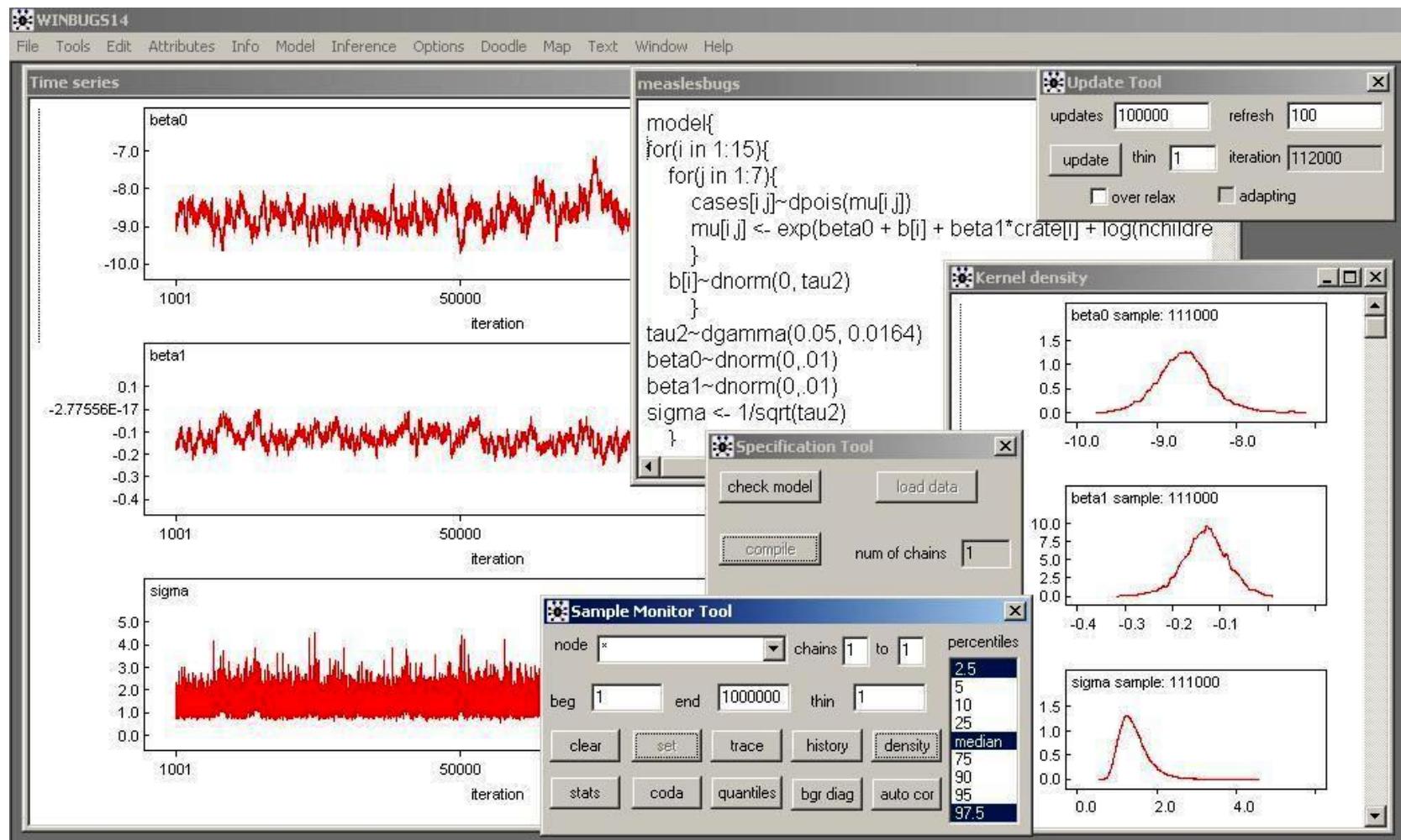
```
model{
  for(i in 1:15){
    for(j in 1:7){
      cases[i,j]~dpois(mu[i,j])
      mu[i,j] <- exp(beta0 + b[i] + beta1*crate[i] + log(nchildren[i,j]))
    }
    b[i]~dnorm(0, tau) # precision, NOT variance!
  }
  tau~dgamma(0.05, 0.0164) # check the manual for parameterization
  beta0~dnorm(0,.01) # precision NOT variance
  beta1~dnorm(0,.01) # precision NOT variance
  sigma <- 1/sqrt(tau)
}

# some starting values (or use gen.inits, but beware diffuse priors)
list( b=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0), beta0=0.7, beta1=0, tau=1)
```

This is not the only way to code it - you could collapse the loops, for example. The ordering of statements doesn't matter.

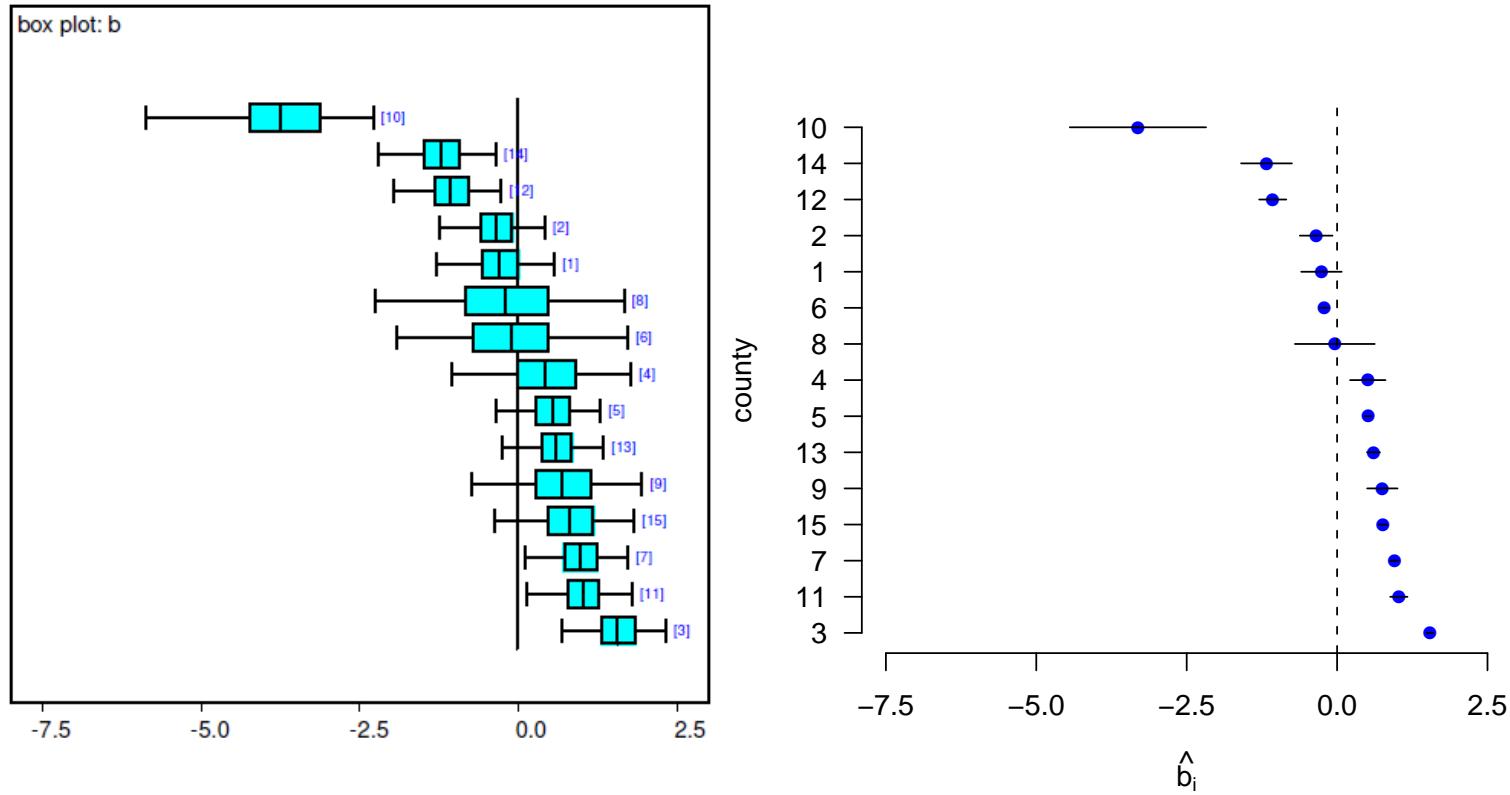
Bayes: WinBUGS

Screenshot of pointy-clicky WinBUGS doing the measles analysis – stats reported earlier



Bayes: WinBUGS

WinBUGS' summary of $b|Y$ and analogous `ranef()` output;



Again, the 'full' (and exact) Bayes version gives wider intervals.
For better MCMC graphics, dump your chains with the coda button and use R.

Bayes: WinBUGS

Pointy-clicky WinBUGS is great for setting up your model and checking it gives sane output. But once you've done this and want to obtain reproducible output (and pretty graphs) from a long chain.

First, in R, set up the data; (compatible with your model)

```
my.inits <- list( # set up initial values, for 2 chains;
  list( b=rep(0,15),      beta0=0.7, beta1=0, tau=1),
  list( b=seq(0,1,l=15), beta0=-2, beta1=0.5, tau=0.5)
)

measles.long <- list( # data in list format
  cases      = matrix( measles$cases, 15, 7, byrow=TRUE ),
  crate      = measles$rate[seq(1,99,7)] - 69.2, # centering
  nchildren  = matrix( measles$children, 15, 7, byrow=TRUE )
)
```

Next, save (just) the model code from 3.181 in a text file – say `measlesbugscode.txt`.

Bayes: WinBUGS

Finally, have R run WinBUGS, dump its results in files, and read their contents in your R session;

```
library("R2WinBUGS") # run WinBUGS from an R session
bugs1 <- bugs( data=measles.long, inits=my.inits,
                parameters.to.save=c("beta0","beta1","tau"),
                model.file="measlesbugscode.txt",
                bugs.seed=4,
                n.chains=2, n.iter=200000,
                n.burnin=6000, n.thin=1,
                working.dir="C:/Users/kenrice/Desktop/571/notes/figs",
                bugs.dir="C:/Program Files/winbugs14/WinBUGS14",
                codaPkg=TRUE)

library("mcmc")
chain1 <- read.coda(output.file=bugs1[1],
                      index.file="C:/Users/kenrice/Desktop/571/notes/figs/codaIndex.txt")
chain2 <- read.coda(output.file=bugs1[2],
                      index.file="C:/Users/kenrice/Desktop/571/notes/figs/codaIndex.txt")
chains <- mcmc.list(chain1, chain2)
```

The main `bugs()` arguments should be familiar from earlier MCMC work. Use `plot()`, `print()`, `summary()` etc on `chains`.

Bayes: WinBUGS – MCMC diagnostics

The R2WinBUGS package eliminates all the pointing and clicking, and the coda package has other useful utilities;

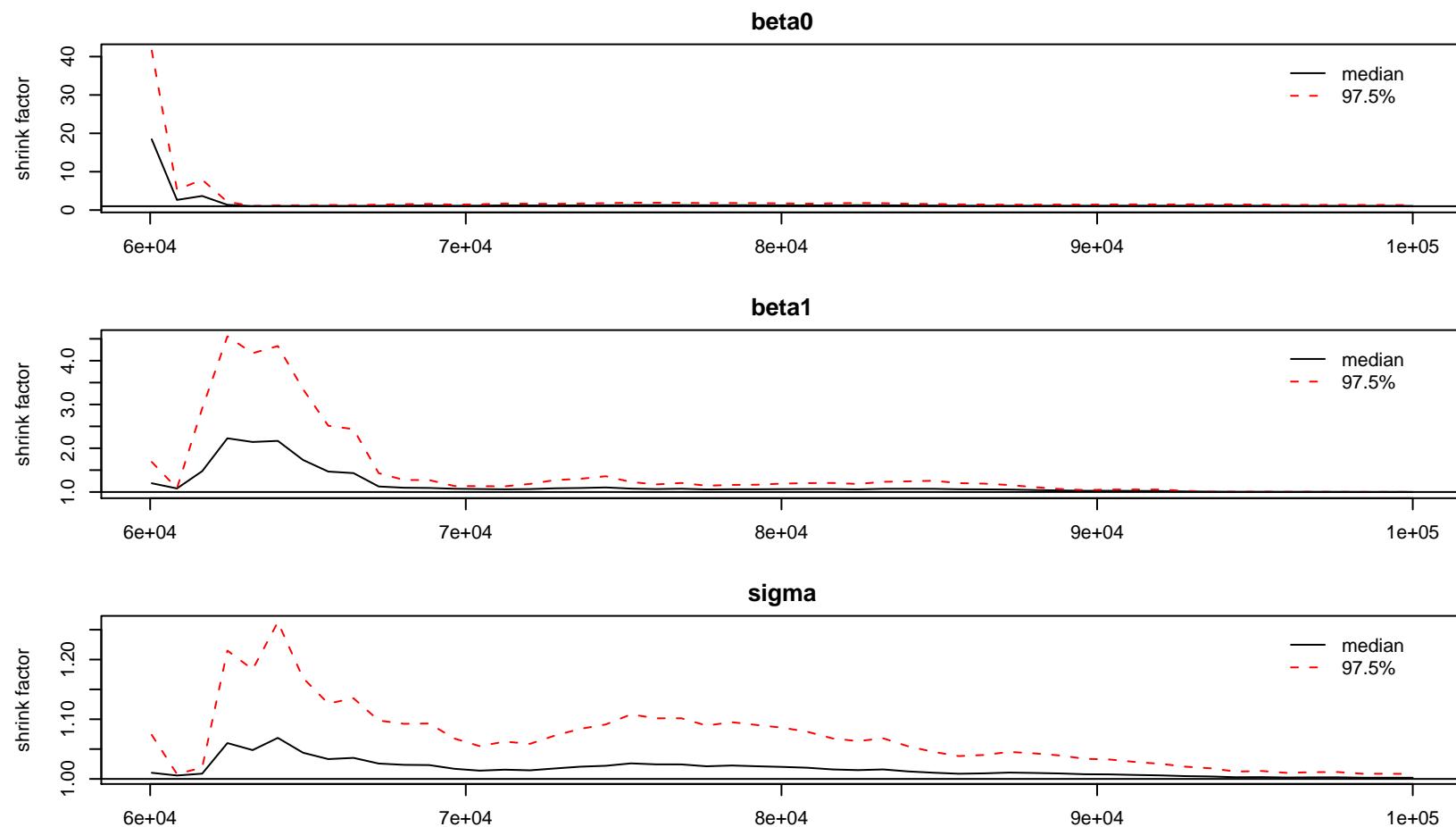
- To check convergence of the chain for some θ , the *Gelman-Rubin diagnostic* compares within-chain variance (W) to between-chain variance (B), using tools from LMMs*. For a converged chain we should get $R = W/B \approx 1$.
- Similar ideas provide the *effective sample size*, i.e. **roughly** how many i.i.d posterior samples the chain represents

```
> gelman.diag( chains )
Potential scale reduction factors: # i.e. R
      Point est. Upper C.I.          # see also gelman.plot()
beta0        1.00      1.01
beta1        1.04      1.15
tau          1.00      1.01
> effectiveSize( chains )
    beta0      beta1       tau
246.8357  310.3243 8058.6007 # suggests 200,000 iterations a *bit* small
```

*... nothing to do with our fitting a mixed model to the data!

Bayes: WinBUGS – MCMC diagnostics

After convergence, $R \rightarrow 1$ as the chain grows;



Bayes: WinBUGS – MCMC diagnostics

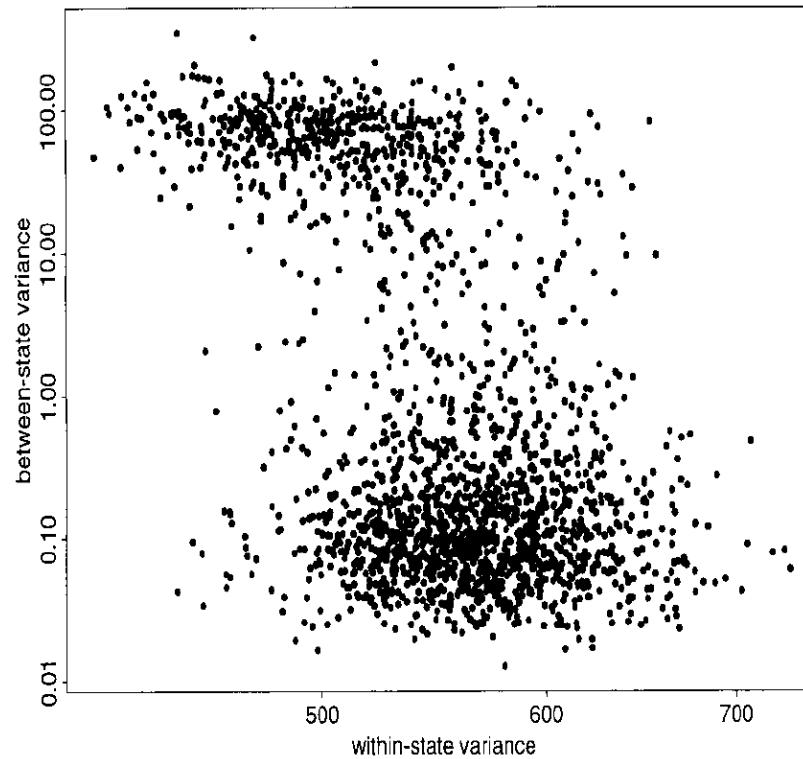
Other notes on MCMC diagnostics; (see Stat 516/7/8 for more)

- Running multiple chains helps you learn quickly about the posterior's *qualities*
- Running a single **very** long chain helps ensure your multiple chains didn't miss anything, It also precisely *quantifies* the posterior
- Formally, it is impossible to rule out all local modes – but MLEs face the same difficulty, and we use them too
- See Gelman and Shirley (2010) on the class site, and Charlie Geyer's skepticism, also the `coda` package documentation

Like model diagnostics, MCMC diagnostics are not perfect – but they may prevent you from making a gross error, and this makes them better than nothing. Report what you did, and why.

Bayes: WinBUGS – MCMC diagnostics

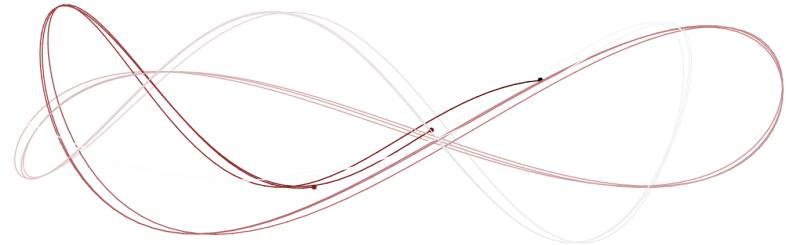
A fairly-famous gross error; Jim Hodges' (JRSSB, 1998) hand-coded mixed-model analysis of exchangeable data from 50 states;



Using only 750 (!) iterations, JH found only *one* mode. A longer chain (JW, discussant) finds the other mode – JH now uses it as a teaching example. Modern MCMC uses $\gg 750$ iterations.

Bayes: Stan

Stan is similar to BUGS, WinBUGS, JAGS etc – but new & improved;



- Coded in C++, for faster updating, it runs the *No U-Turn Sampler* – cleverer than WinBUGS' routines
- The `rstan` package lets you run chains from R, just like we did with `R2WinBUGS`
- Some modeling limitations – no discrete parameters – but becoming popular; works well with some models where WinBUGS would struggle
- Basically the same modeling language as WinBUGS – but Stan allows R-style vectorization
- Requires declarations (like C++) – unlike WinBUGS, or R – so models require a bit more typing...

Bayes: Stan

```
data {  
    int<lower=0> N; // number of observations  
    int<lower=0> n; // number of counties  
    int<lower=0, upper=n> county[N];  
    int<lower=0> cases[N];  
    vector[N] crate;  
    vector[N] offset; }  
parameters {  
    real beta0;  
    real beta1;  
    real<lower=0> tau;  
    real b[n]; }  
transformed parameters {  
    real sigb;  
    sigb <- 1/sqrt(tau); }  
model {  
    vector[N] logmu;  
    beta0~normal(0,10);  
    beta1~normal(0,10);  
    tau~gamma(0.05, 0.0164);  
    b~normal(0,sigb);  
    for (i in 1:N)  
        logmu[i] <- offset[i] + beta0 + b[county[i]]+ beta1*crate[i];  
    cases ~ poisson_log(logmu); }
```

Bayes: Stan

With this stored in `measlesstancode.stan`, to run the analysis;

```
measles.list <- list( cases=measles$cases, offset=log( measles$children ),  
  crate=measles$rate - 69.2, county=measles$county, N=nrow(measles), n=15)  
stan1 <- stan(file = "measlesstancode.stan", data = measles.list,  
  iter = 10000, chains = 1)  
# Elapsed Time: 78.475 seconds (Total)  
print(stan1)                                # edited output -  
  50%    2.75%    97.5% n_eff # compare with 3.153 & 3.186  
beta0 -8.63    -9.43    -7.84    811  
beta1 -0.14    -0.22    -0.04   1174  
sigb   1.35     0.89    2.24   1503
```

- Up to Monte Carlo error, the same as we got from WinBUGS
- ≈ 1000 independent samples from chain length 10,000 is **much** better than WinBUGS' chain of 200,000
- Compilation can take time, but once done Stan will re-use the sampling code
- Diagnostics work as before; `print()` gives Gelman-Rubin R , for graphical diagnostics use `traceplot()`

Bayes: INLA

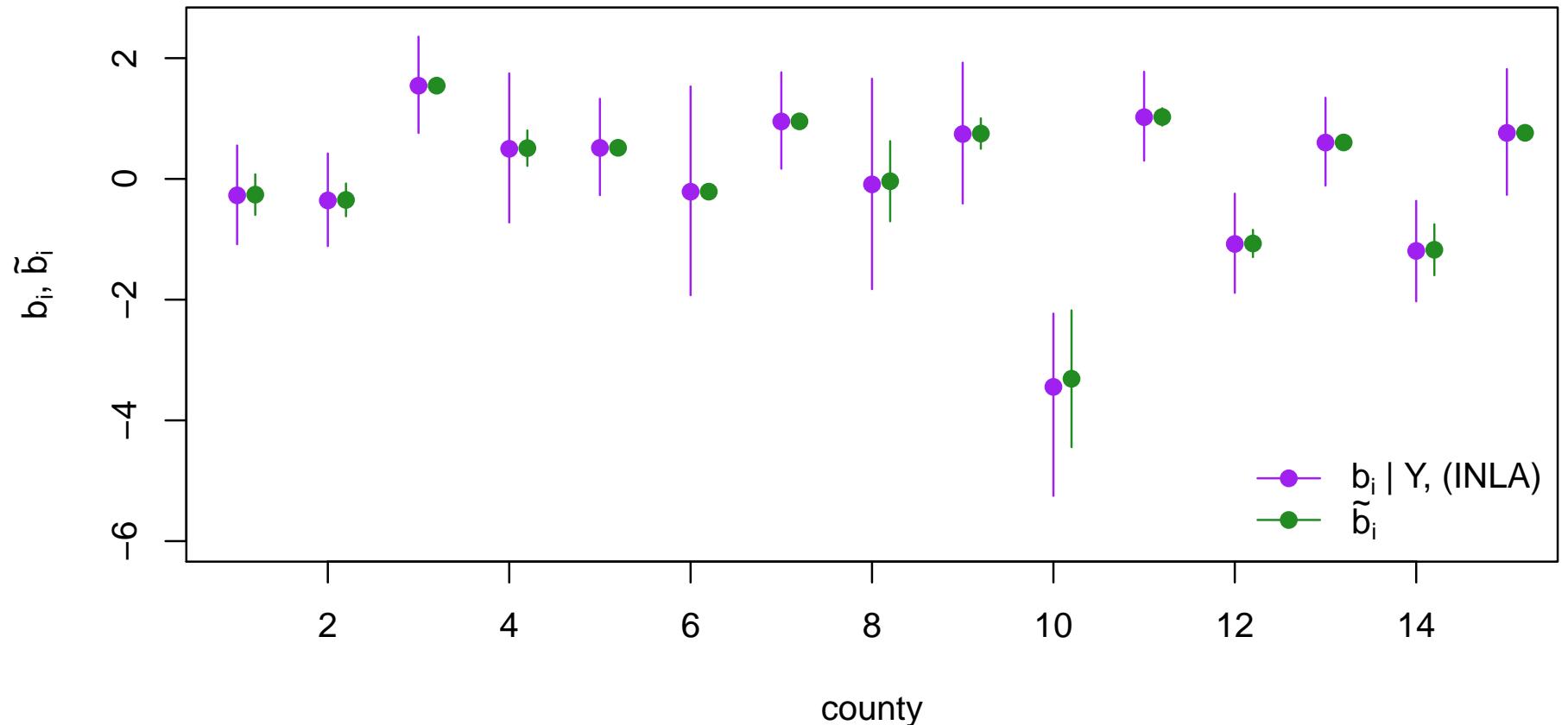
You met INLA in 570, in some detail. Its application to (G)LMMs is conceptually no different in 571 – but the mixed model and prior must be specified. For the measles example;

```
library("INLA")
hyperprior <- list(theta=list(prior="loggamma",param=c(0.05,0.0164)))
fixedprior <- list( mean.intercept=0, prec.intercept=.01,
                     mean=0,                  prec=.01 )
formula1 <- cases ~ crate + offset(log(children)) +
               f(county,model="iid",hyper=hyperprior)
inla1 <- inla(formula1,family="poisson",data=measles, control.fixed= fixedprior)
# fixed effects
round(inla1$summary.fixed[,c(4,3,5)], 2)
      0.5quant 0.025quant 0.975quant
(Intercept) -8.65        -9.39       -7.95
crate        -0.14        -0.22       -0.05
# and for "sigb"
round(1/sqrt(inla1$summary.hyperpar[c(4,5,3)]),2)
      1.37        0.89        2.24
```

- Priors for fixed β and distribution of random b_i enter in two parts – in `f()` and `control.fixed` (see also `control.family`)
- Compare with Stan – no MC steps here so very reliable
- Also see `inla1$summary.random` for $b|Y$, the random effects

Bayes: INLA

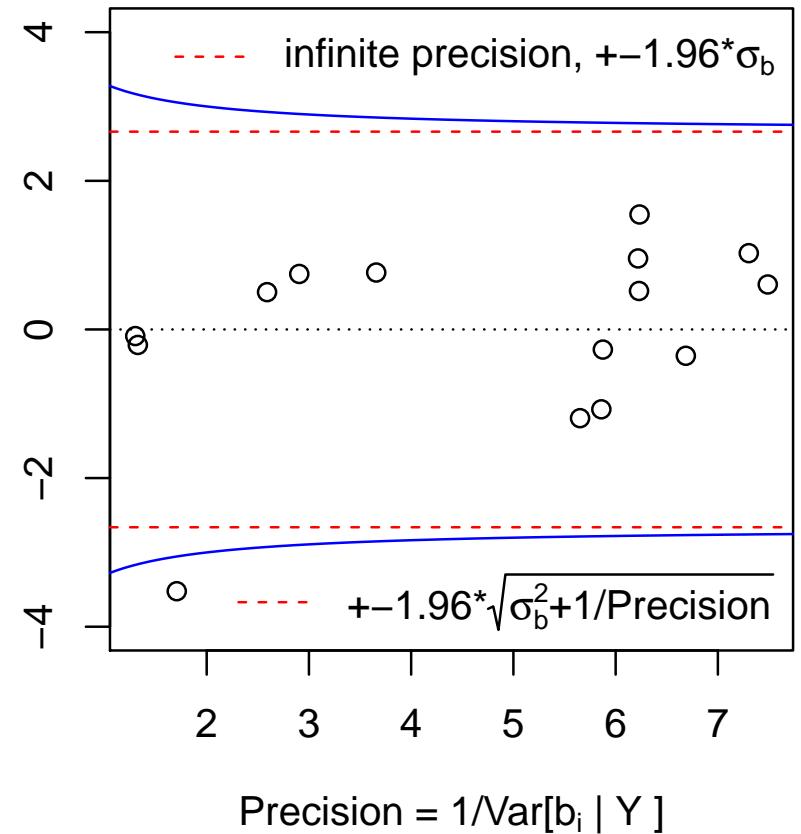
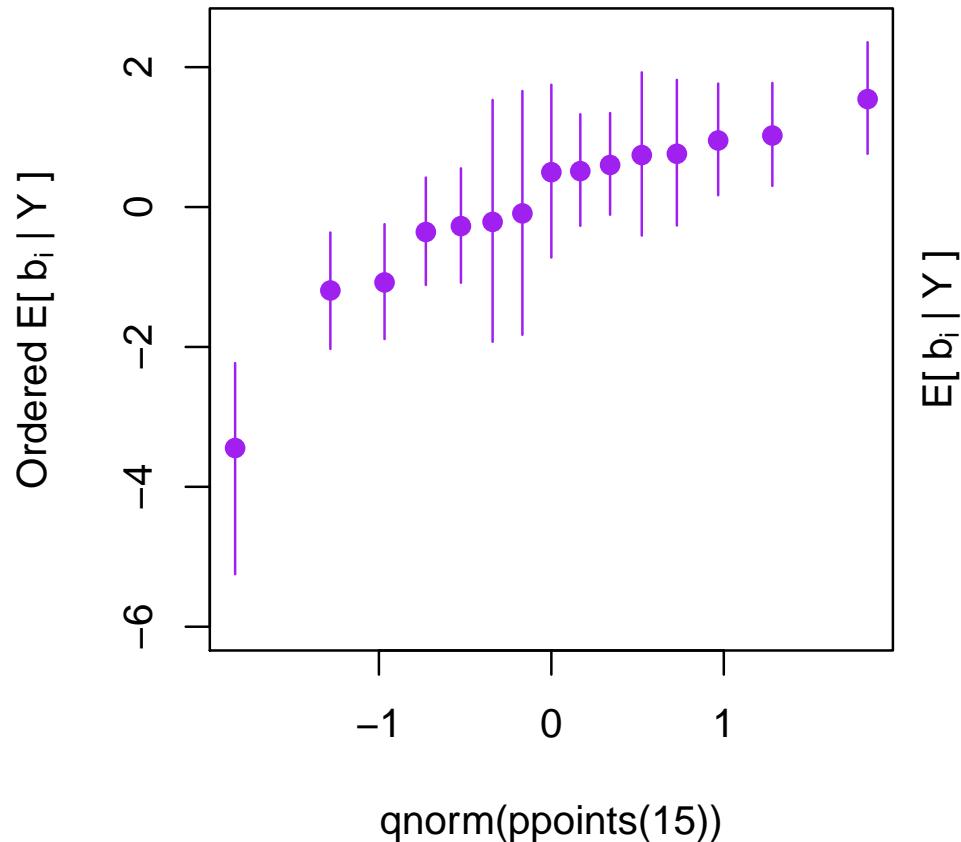
Inference for the random effects, versus `glmer()`:



- As with dyestuff example, E-Bayes looks over-precise
- Harder to compare against the data, here

Bayes: model diagnostics

An alternative to QQ plots, if you have uncertainties you believe;



The RH is a *funnel plot* – with no ordering, it's more obvious that more extreme b_i are (just?) noisier

Bayes: model diagnostics

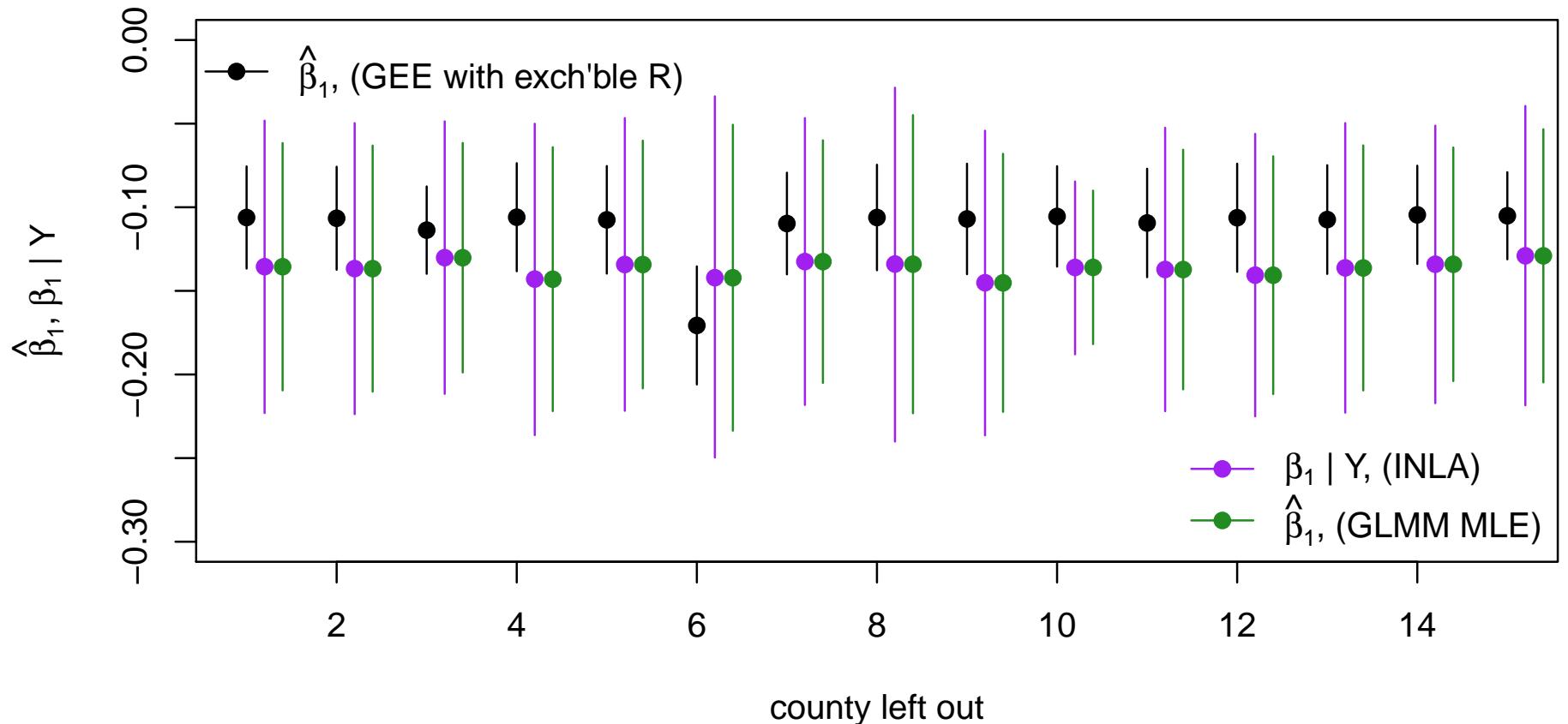
Diagnostics under Bayesian approaches are surprisingly controversial; one view is that Bayes is (just) a way to describe Your uncertainty based on Your prior – including Your model – and the data, so there is no need to check the model.

- There are many other views. One approach is *subjunctive Bayes*; with **this** prior and the data, **this** is the posterior one would have. With **that** prior... etc
- *Prior-data conflict* is a major concern; a Bayesian *vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule?* (Stephen Senn)
- Conflict from different parts of the data is also a concern

With INLA's speed, we can address the last point in a familiar way;

Bayes: model diagnostics

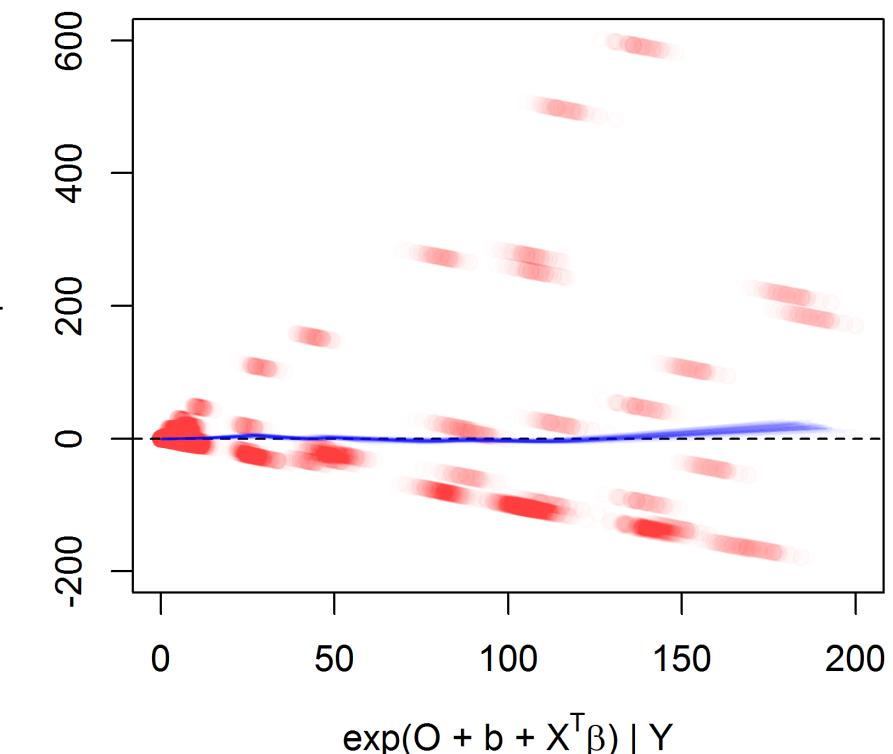
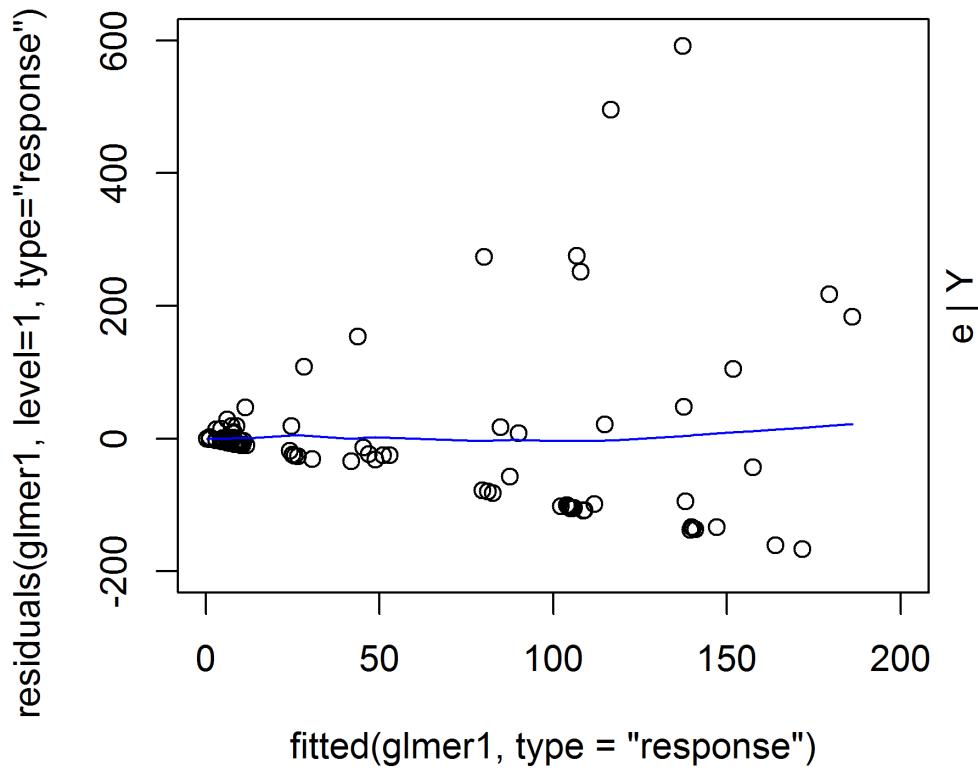
Here, leaving out each cluster in turn (see 2.87)



- with MCMC, checking 15 chains would be required. NB with a correct model, GEE estimates the same parameter as GLMM, here (see 3.113) so some cause for concern

Bayes: model diagnostics

Using the posterior for unknown e_{ij} , where $e_{ij} = Y_{ij} - e^{O_{ij} + b_i + X_{ij}^T \beta}$;



The Bayes version uses Stan output (100 well-spaced posterior samples) with a smoother at each iteration. **Q.** Based on either plot, what can you tell about mean=variance assumption?

Bayes: model diagnostics

Notes on diagnostics for Bayesian GLMM fits;

- This is an open research area – expect to have to justify anything you do, and code it yourself
- Getting some idea of the uncertainty around the plotted points and the smoother can help a lot, particularly if they are very noisy. (For the smoother, one could use the bootstrap, although it won't work well everywhere)
- While the posterior describes Your uncertainty, but it **also** captures how very strange You should consider the data You observed
- Some authors (e.g. Gelman Carlin *et al*, chapter 6) suggest giving tail-area statements of how extreme some function of the observed data is, relative to data generated from the posterior predictive distribution. But these are in general hard to calibrate – see papers on the course site

Mixed models: missing data



*At e'en at the gloamin, nae swankies are roamin
'Bout stacks wi the lassies at bogle tae play
But ilk ane sits dreary, lamentin her deary –
The Flooers of the Forest are a' wede away*

From *The Flowers of the Forest*
by Jean Elliot (1727–1805), Scottish poet

Mixed models: missing data

As the quarter is largely *a'* wede away, we reprise missing data (2.130–2.168) only briefly. Using previous notation, recall;

- The best solution is not to have any missing data... but this often doesn't happen, even in well-run studies
- If we condition on observed \mathbf{X} , only missing Y are a problem
- The pattern of missingness determines whether complete case analysis is valid – e.g. for GEE, generally need MCAR, where $\mathbb{P}[M|\mathbf{Y}, \mathbf{X}] = \mathbb{P}[M|\mathbf{X}]$
- Under MAR, multiple imputation (MI) is an attractive way to exploit the data we do have

Two important questions we can address briefly;

- Q1.** When is complete-case analysis valid, fitting mixed models?
- Q2.** Can we do MI without it being an ugly hack?

Mixed models: is complete-case okay?

Under MCAR, complete-case mixed model analysis is valid, just like GEE. To see this;

- For MLEs, note we are solving EEs and follow exactly the same steps as on 2.135
- For Bayes, note the likelihood contribution is

$$f(Y_{ij}|\mathbf{X}_{ij})^{1-M_{ij}} \mathbb{P}[M_{ij}|\mathbf{X}_{ij}] = \begin{cases} f(Y_{ij}|\mathbf{X}_{ij}) \mathbb{P}[M_{ij} = 0|\mathbf{X}_{ij}], & M_{ij} = 0 \\ \mathbb{P}[M_{ij} = 1|\mathbf{X}_{ij}], & M_{ij} = 1 \end{cases},$$

so **assuming the missingness and analysis models share no parameters**, the complete-case analysis gets the posterior correct up to a constant multiple – which is all that's required for inference on the analysis model

- MCAR is often implausible in practice (see 2.137)

Note there's nothing special here about mixed models, these statements also hold for e.g. 570's independent outcomes.

Mixed models: is complete-case okay?

A stronger result, that's important in practice;

Under Missing At Random (MAR), i.e. where

$$f(\mathbf{M}|\mathbf{Y}, \mathbf{X}) = f(\mathbf{M}|\mathbf{Y}^o, \mathbf{Y}^m, \mathbf{X}) = f(\mathbf{M}|\mathbf{Y}^o, \mathbf{X}),$$

model-based complete case analysis* is valid

Why does this hold? Again using the informal Ch2 notation;

$$\begin{aligned} f(\mathbf{Y}, \mathbf{M}|\mathbf{X}) &= f(\mathbf{M}|\mathbf{Y}, \mathbf{X})f(\mathbf{Y}|\mathbf{X}) \\ &= f(\mathbf{M}|\mathbf{Y}^o, \mathbf{X})f(\mathbf{Y}|\mathbf{X}) \\ f(\mathbf{Y}^o, \mathbf{Y}^m, \mathbf{M}|\mathbf{X}) &= f(\mathbf{M}|\mathbf{Y}^o, \mathbf{X})f(\mathbf{Y}^o, \mathbf{Y}^m|\mathbf{X}). \end{aligned}$$

Only one \mathbf{Y}^m appears on each side, so integrating it out;

$$f(\mathbf{Y}^o, \mathbf{M}|\mathbf{X}) = f(\mathbf{M}|\mathbf{Y}^o, \mathbf{X})f(\mathbf{Y}^o|\mathbf{X}) \propto f(\mathbf{Y}^o|\mathbf{X}),$$

again, **if** analysis and missingness models share no parameters.

* ...with assumed-correct model and large-enough n for any asymptotics. If using model-based analysis, prior must permit likelihood domination

Mixed models: is complete-case okay?

While MAR is still a questionable assumption, modeling's robustness to MAR missingness is a reason many users give, when asked to justify complete-case analysis over;

- Complete-case GEE, which needs MCAR
- Weighted or imputed GEE, which need MAR and/or very strong assumptions, and non-trivial work to fit them

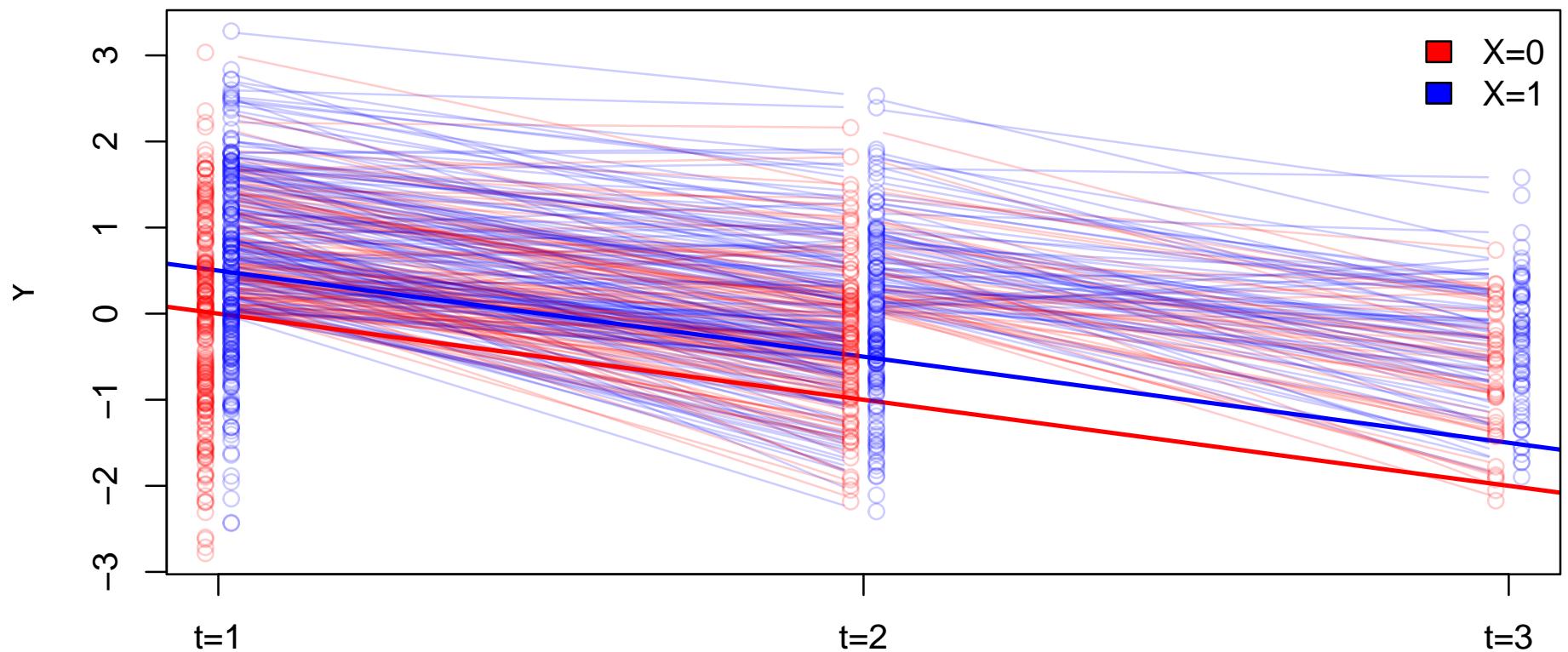
To illustrate this, a challenging example from HW6;

$$\begin{aligned} b_i &\stackrel{i.i.d.}{\sim} N(0, \sigma_b^2) \\ X_i &\stackrel{i.i.d.}{\sim} Bern(0.5) \\ Y_{it}|b_i, X_i = x &\stackrel{indep}{\sim} N(\beta_0 + b_i + \beta_1 t + \beta_2 x, \sigma_Y^2), \text{ for } t = 1, 2, 3, \end{aligned}$$

where if any $Y_{it} < 0$ then all subsequent $Y_{it'}$ are missing.

Mixed models: is complete-case okay?

Typical data; (Here with $n = 500$, and $\sigma_b = 1, \sigma_Y = 0.5, \beta_0 = 1, \beta_1 = -1, \beta_2 = 0.5$, and superimposing true means for $b_i = 0$)



- High rates of missingness, e.g. 50%, for $X = 0$ at $t = 2$
- Highly informative cluster size

Mixed models: is complete-case okay?

For a single sample with $n = 10,000$; (estimates & std errs)

Truth		LMM MLE	GEE	
			Indept R	Exch'ble R
$\hat{\beta}_0$	1	0.98 (0.02)	0.47 (0.02)	—
$\hat{\beta}_1$	-1	-0.98 (0.01)	-0.41 (0.01)	—
$\hat{\beta}_2$	0.5	0.48 (0.02)	0.36 (0.02)	—
$\frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}_Y^2}, \hat{\alpha}$	0.8	0.799	0	1.06 (!!?)

- Complete-case GEE with independence working R is a disaster, for conditionally-specified β
- With exchangeable R there are boundary problems – the $\hat{\alpha}, \hat{\phi}$ used don't account (well) for clusters with $n_i = 1$ – this is not common in practice. NB not all GEE code crashes out, even though it should!
- Complete-case LMM MLE gets the right answer **despite** not using knowledge of missingness model, beyond that it is MAR

Mixed models: is complete-case okay?

You do need to know (or assume) MAR for complete-case analysis to be valid. For an example, consider modifying the previous example so dropout occurs **immediately** if $Y < 0$; no negative Y are ever observed, and complete case-analysis cannot fit the correct mean models.

- This is an example of MNAR: M depends on the missing Y
- Sensitivity analyses of different **assumed** MNAR mechanisms is then the best you can do

Why? Fundamentally, because...

MAR, versus MNAR, is not a testable assumption

To see this intuitively, note that to do such a test you would need to compare the distribution of the missing outcomes to those that have been observed – which is impossible.

Mixed models: is complete-case okay?

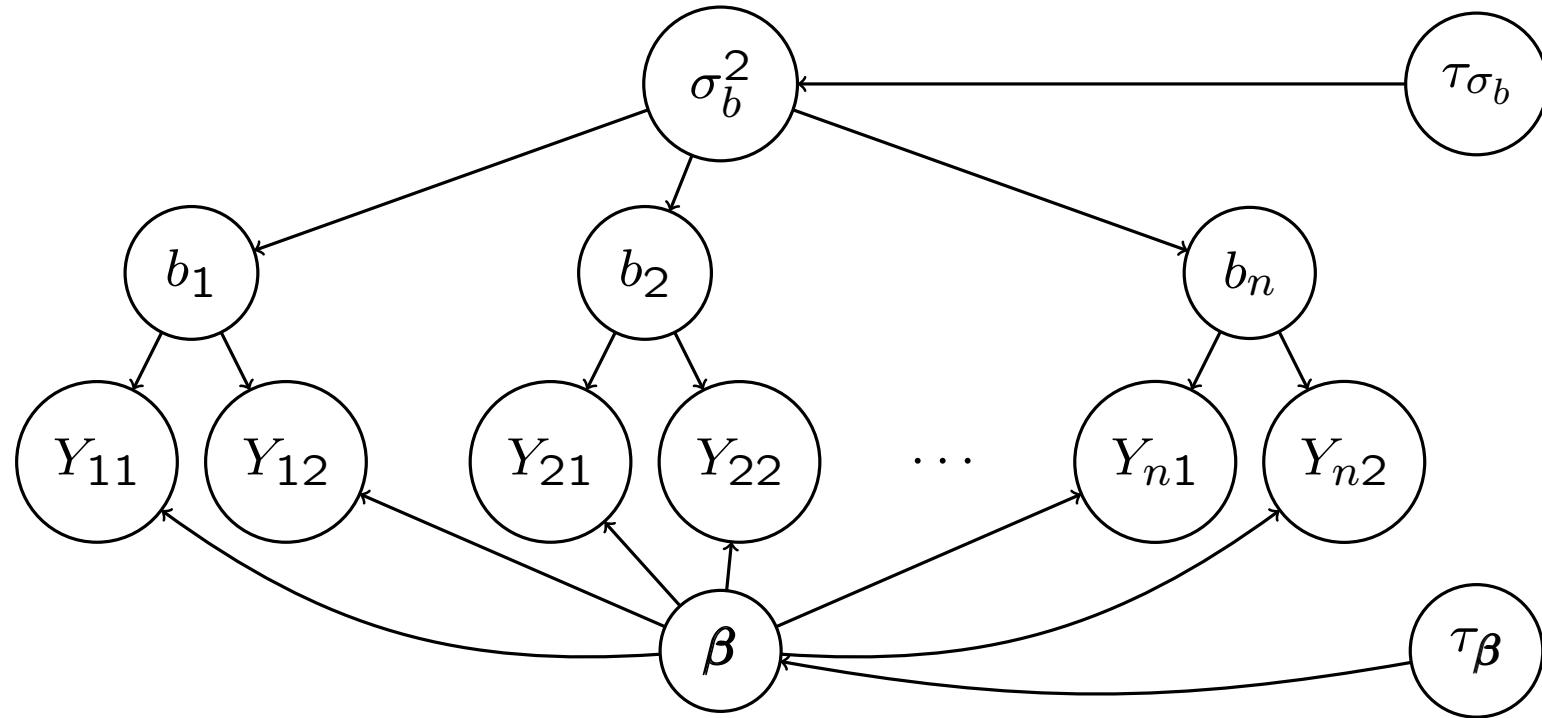
Put another way – and far more formally;

- Every MNAR model has a MAR counterpart with equally good fit (2008), i.e. the two are indistinguishable, and complete-case is only valid under MAR
- This hasn't stopped past researchers from trying to test MAR! For example, Zhou et al (1999) proposed a likelihood ratio test of H_0 : MAR – but unless one has strong *a priori* beliefs in MNAR, small p -values from the test are **not** good evidence against MAR
- Indexes suggesting sensitivity to MAR assumptions **are** available (Troxel et al 2004) but may not be easily calibrated

It's worth re-emphasizing that **context really matters** here. With no idea about what patterns of missingness are likely, it's very difficult to make useful conclusions.

Mixed models: imputation for free

The DAG for a mixed model;



- To update what we know about an unknown node, sample from the conditional: the unknown node given **all** the others
- If missing=unknown, the DAG (and corresponding models/priors) tells us how to impute/update

Mixed models: imputation for free

Missing=unknown is another way to say MAR – and hence justify using observed Y and covariates to impute missing Y .

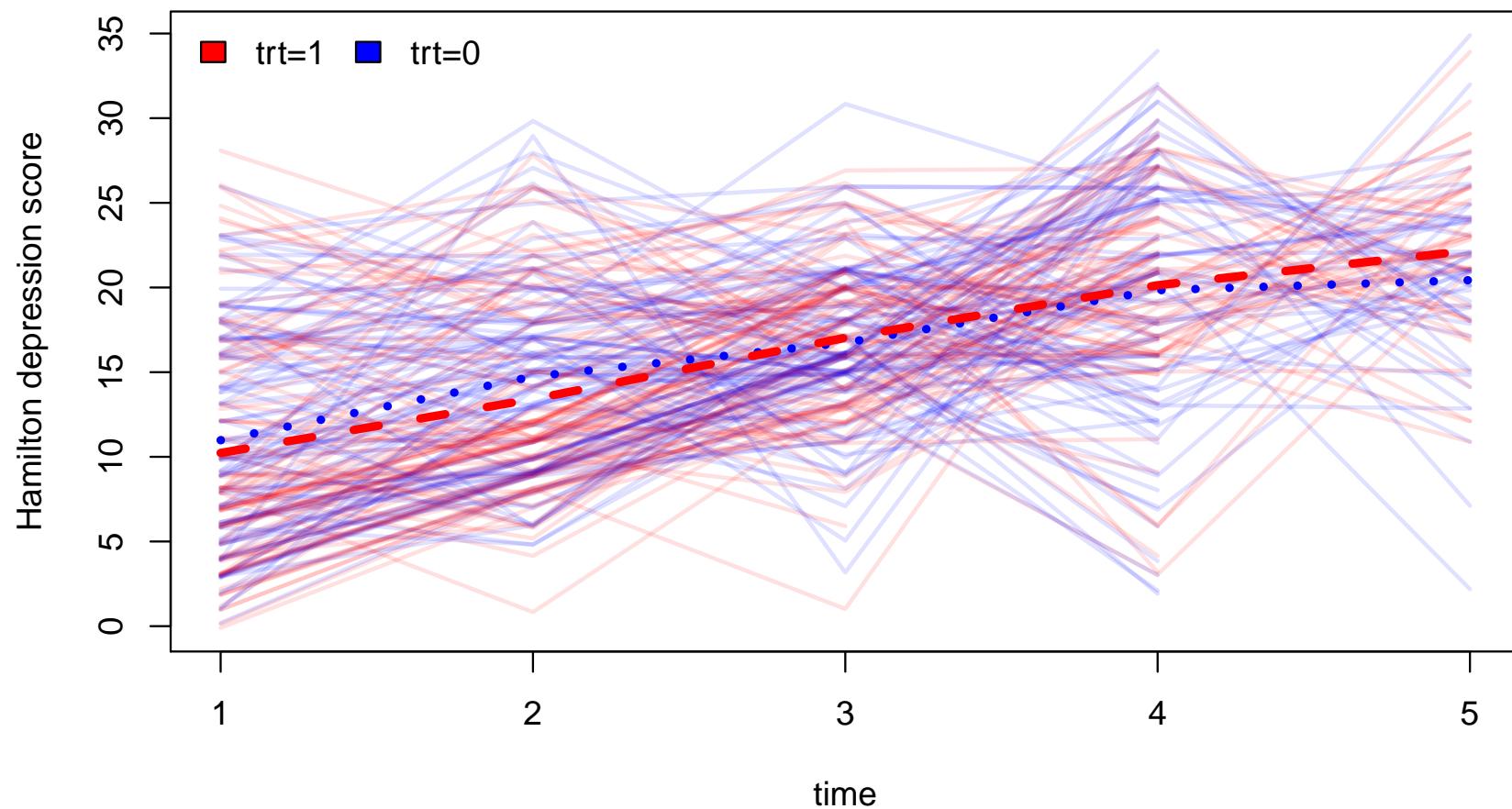
- MAR assumes that the observed outcomes tell us all we need to know to usefully impute the missing outcomes
- MAR makes standard multiple imputation ‘work’ (see Chapter 2)
- Really, the fully-Bayesian imputation is **what MI approximates** – recall that to be ‘proper’ the MI parameters had to be generated from a posterior distribution

This **very useful** approach is not specific to mixed models: **any** Bayesian analysis can use the same idea, assuming MAR.

Also useful: when it sees NAs, WinBUGS does the imputation updates **automatically**. Stan can do them but needs a careful model specification. INLA can not impute in this way.

Mixed models: imputation for free

An example; $n = 200$ subjects from a randomized trial of antidepressants, with time-specific means.



At $t = 4$, 7%/17% of outcomes are missing in control/treatment.
At $t = 5$ this rises to 54%/56%.

Mixed models: imputation for free

An LMM for this data, with the constraint of no treatment/control difference in mean Y at $t = 1$:

$$\begin{aligned} b_{i0} &\stackrel{i.i.d.}{\sim} N(0, \sigma_{b0}^2) \\ b_{i1} &\stackrel{i.i.d.}{\sim} N(0, \sigma_{b1}^2) \\ \mu_{it} | \mathbf{b}_i, \text{trt}_i &= (\beta_0 + b_{i0}) + (\beta_1 + b_{i1} + \beta_2 \text{trt}_i)(t - 1) \\ Y_{it} | \mathbf{b}_i, \text{trt}_i &\stackrel{\text{indep}}{\sim} N(\mu_{it}, \sigma_Y^2) \end{aligned}$$

- Treatment effect is difference in slope
- Fitting random intercepts and slopes – or viewing them as exchangeable
- Depression scores are integers – but mean and variance assumptions are far more important than Normality

Assuming the model is correct, complete-case analysis is valid.
But does imputation buy back any precision?

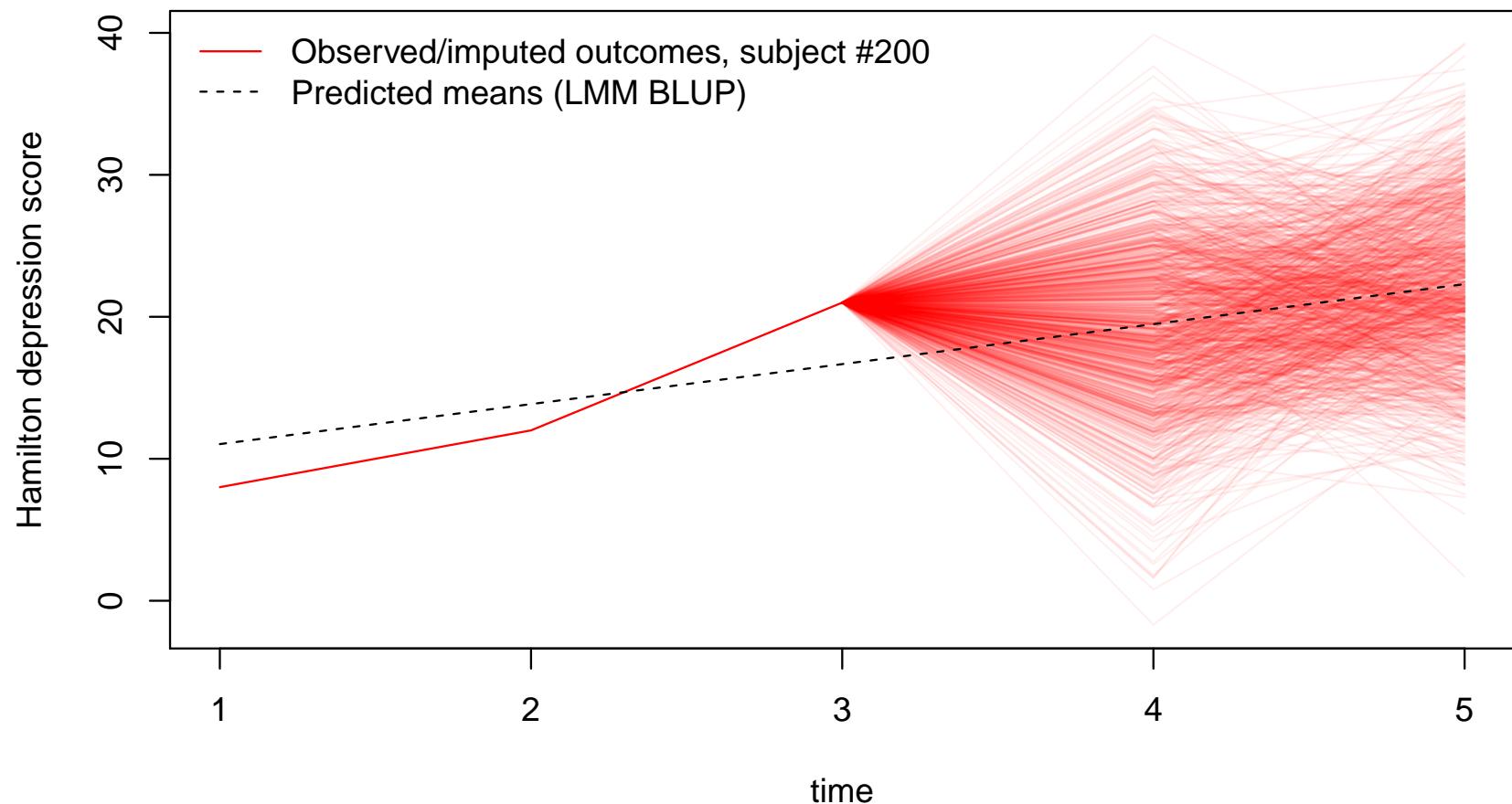
Mixed models: imputation for free

WinBUGS code for the model, with the data in long format – note no special commands needed for NAs;

```
model{
  for(i in 1:bigN){ # i indexes observations
    y[i] ~ dnorm(mu[i], tauY)
    mu[i] <- int[i] + slope[i]*(time[i]-1)
    int[i] <- beta0 + b0[id[i]]
    slope[i] <- beta1 + beta2*trt[i] + b1[id[i]]
  }
  for(j in 1:n){ # j indexes people
    b0[j] ~ dnorm(0, taub0)
    b1[j] ~ dnorm(0, taub1)
  }
  taub0 <- 1/pow(sigb0,2)
  taub1 <- 1/pow(sigb1,2)
  sigb0 ~ dunif(0,5) # U(0,5) priors on Rand Eff SDs
  sigb1 ~ dunif(0,5)
  tauY ~ dgamma(0.01, 0.01) # prior on 1/Var(Y|b)
  beta0 ~ dnorm(0,0.001)      # priors on Fixed Effs
  beta1 ~ dnorm(0,0.001)
  beta2 ~ dnorm(0,0.001)
}
```

Mixed models: imputation for free

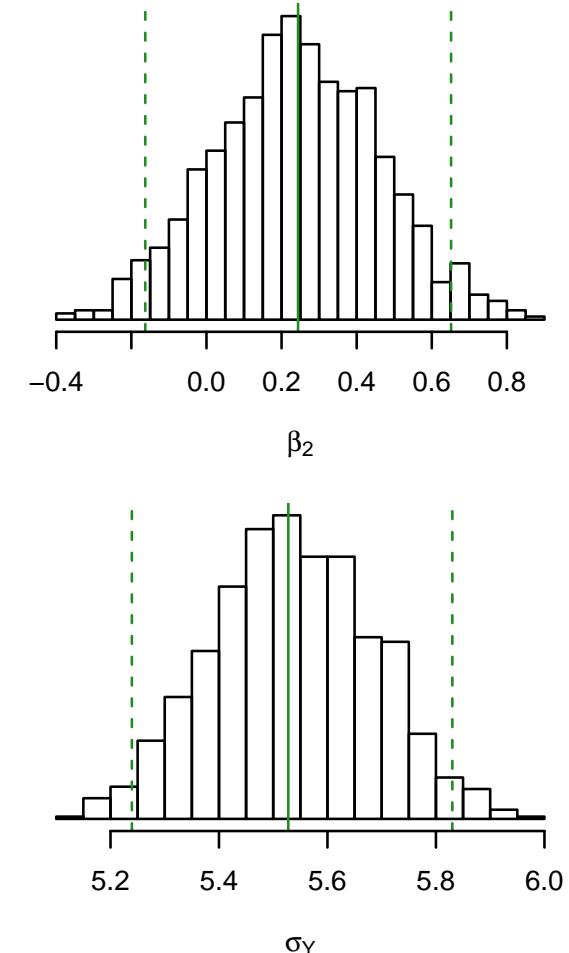
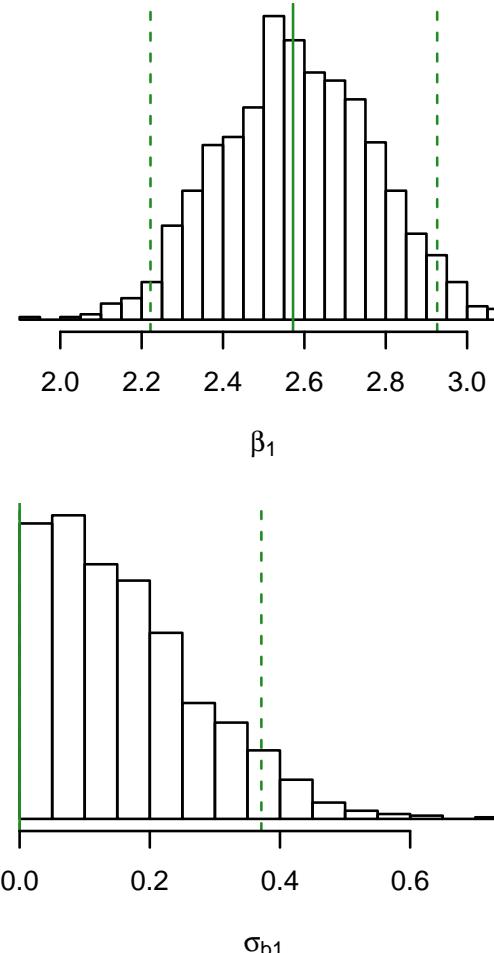
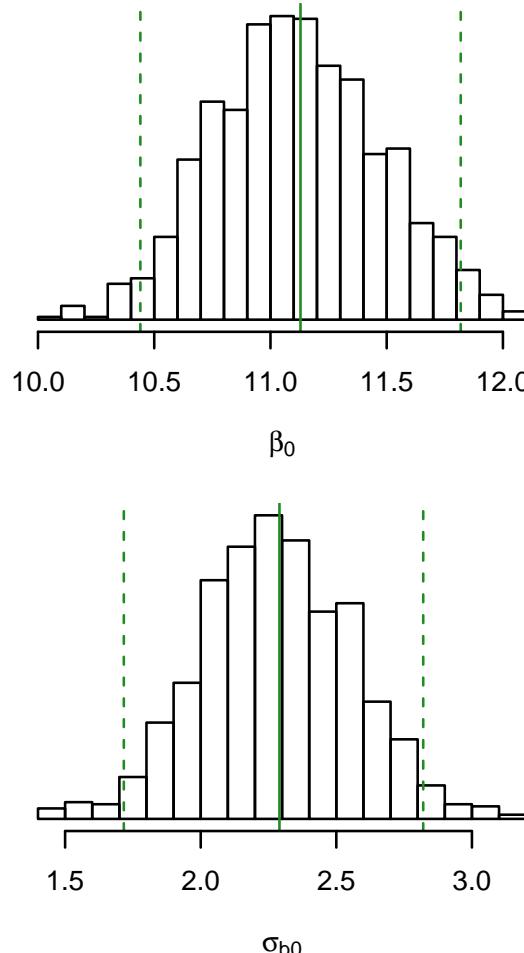
Illustrating the imputation for one subject;



The imputations ‘shrink’ toward the cluster-specific mean outcome, which borrows strength from all clusters.

Mixed models: imputation for free

WinBUGS output, with large- n LMM 95% CIs (complete-case);



Note the MLE for $\sigma_{b1} = 0$, on the boundary.

Mixed models: imputation for free

Comparing inference, with complete-case LMM;

Parameter:	β_0		β_1		β_2	
	Est	CI	Est	CI	Est	CI
Bayes	11.11	10.45,11.82	2.58	2.22,2.94	0.24	-0.19, 0.67
LMM-cc	11.13	10.44,11.82	2.57	2.22,2.93	0.24	-0.16,0.65

Parameter:	σ_{b0}	σ_{b1}	σ_Y
Bayes	2.28	1.73,2.84	0.14
LMM-cc	2.29	1.72,2.82	0

Bayesian estimates are medians and 2.5%, 97.5%-iles.

- The differences are trivial, here – can think of imputation as providing (somewhat) better weighting
- σ_{b1} behavior is due to boundary MLE, not missing data. Exercise for keen people: implement Bayesian inference for complete-case data only, and check it matches full-data Bayes

Mixed models: imputation for free

Final notes on Bayesian imputation under MAR;

- If MAR is reasonable, this is a good first approach
- Sensitivity analyses would state/fit a MNAR model for the missingness indicators, and check for differences. This too can be done in WinBUGS/Stan, but code for the missingness model is needed
- Under MAR, the pattern of missingness (beyond being MAR) is **ignorable** – so the WinBUGS imputation will work, but so will other congenial approaches
- Numerically, expect to have to give sensible starting values, to avoid underflow issues (see 3.173)
- For discrete unknowns (e.g. b_i or Y_{ij}), Gibbs can get ‘stuck’ if e.g. all unknown $b_i = 0$; sampling from the conditionals for other unknowns never moves in finite time. More generally, be wary of multimodality in complex modeling

See the course site for Gelman & Hill’s chapter on missing data, and other resources. Also see Biost 531 – this is a large topic

Mixed models: summary

Main points from this chapter:

- Scientific interest often lies in conditional parameters: to learn about these we need to rely on assumptions
- Connections with GEE, and marginal parameters, exist in a few useful special cases, but not in general
- Numerical fitting of mixed-models uses sophisticated tools
- Bayesian inference is a natural approach, both for fitting and interpreting models
- Model-based methods require only MAR, not MCAR. Imputation can still be used, and under MAR is particularly straightforward using MCMC

CHAPTER 4: EPILOGUE



*Should auld acquaintance be forgot,
and never brought to mind?
Should auld acquaintance be forgot,
and auld lang syne?*

*For auld lang syne, my jo,
for auld lang syne,
we'll tak a cup o' kindness yet,
for auld lang syne.*

CHAPTER 4: EPILOGUE

Some auld acquaintances **you** should try not to forget;

- Motivation for correlated data – efficiency for marginal comparisons, within-cluster comparisons, that require clustered data.
- Robust methods; GEE – mostly marginal inference. Gives robustness to everything except the mean model (diagnostics) possible loss of efficiency (more diagnostics) and some computational issues. Generally requires MCAR
- Model-based methods; (G)LMM – mostly conditional inferences. Modeling assumptions must be \sim right, and may be difficult/impossible to check (still more diagnostics). Computation can be slow & challenging. Requires MAR
- Some connections between these; e.g. LMMs with homoskedastic random effects, β is *both* marginal and conditional. But typically expect to justify which you use

CHAPTER 4: EPILOGUE

'Bigger picture' topics;

- Robustness is desirable, but typically comes at the price of **some** efficiency and/or worse small-sample behavior – compared to correctly-specified mixed models
- Diagnostics (model and MCMC) can help spot egregious errors, but will not reliably detect more subtle errors
- Bayes and non-Bayes methods have **very** close connections, provided you make sure both answer the same question. More generally, a single method can have multiple interpretations, and knowing them helps understand that method
- You are expected to be able to code up the methods we have discussed – and also to use off-the-shelf code wisely
- **Context really matters.** For inference, expect to rely heavily on context when choosing what to do

Extensions beyond 571 (*)

As noted in Chapter 0, 570/571's data-based examples are 'toy', i.e. most of them just illustrate methods. Real-life data analyses are messier, require more time and use much more contextual knowledge. So for more experience applying 570/571's material;

- Stat/Biost 579: data analysis and report-writing. Example-centered discussions of what analyses to do and why, and how to present them. Varies with instructors' style/area of interest
- Biost 590/Stat 599: consulting – with some lecture material but mostly hands-on collaboration with local scientists
- Stat 528: similar to both of the above, but aimed at Stat MS students

In my experience, correlated data is a key area where consulting clients need help; dealing with missing data is another key area.

Extensions beyond 571 (*)

The independence of clusters is *not* always reasonable in practice;

Covariation in the socioeconomic determinants of self rated health and happiness: a multivariate multilevel analysis of individuals and communities in the USA

S V Subramanian, Daniel Kim, Ichiro Kawachi

J Epidemiol Community Health 2005;59:664–669. doi: 10.1136/jech.2004.025742

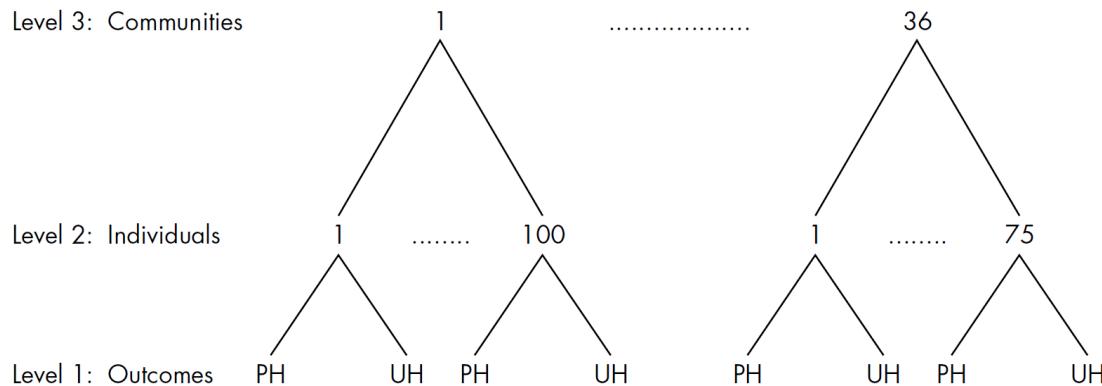


Figure 1 Multivariate multilevel structure of responses (PH, poor health; UH, unhappy) at level 1 nested within individuals at level 2 nested within communities at level 3

One level of clustering doesn't allow some dependence within community, and *more* dependence within individual. Mixed models extend to *hierarchical* models – see Gelman and Hill.

Extensions beyond 571 (*)

In spatial work, every outcome is typically correlated with every other outcome; can't split the data into multiple independent clusters.

Right: from *Spatial analysis of the distribution of Lyme disease in Wisconsin*, Kitron & Kazmierczak, Amer J of Epidemiology, 1997.

See BIOST/EPI 555 for specialized methods.

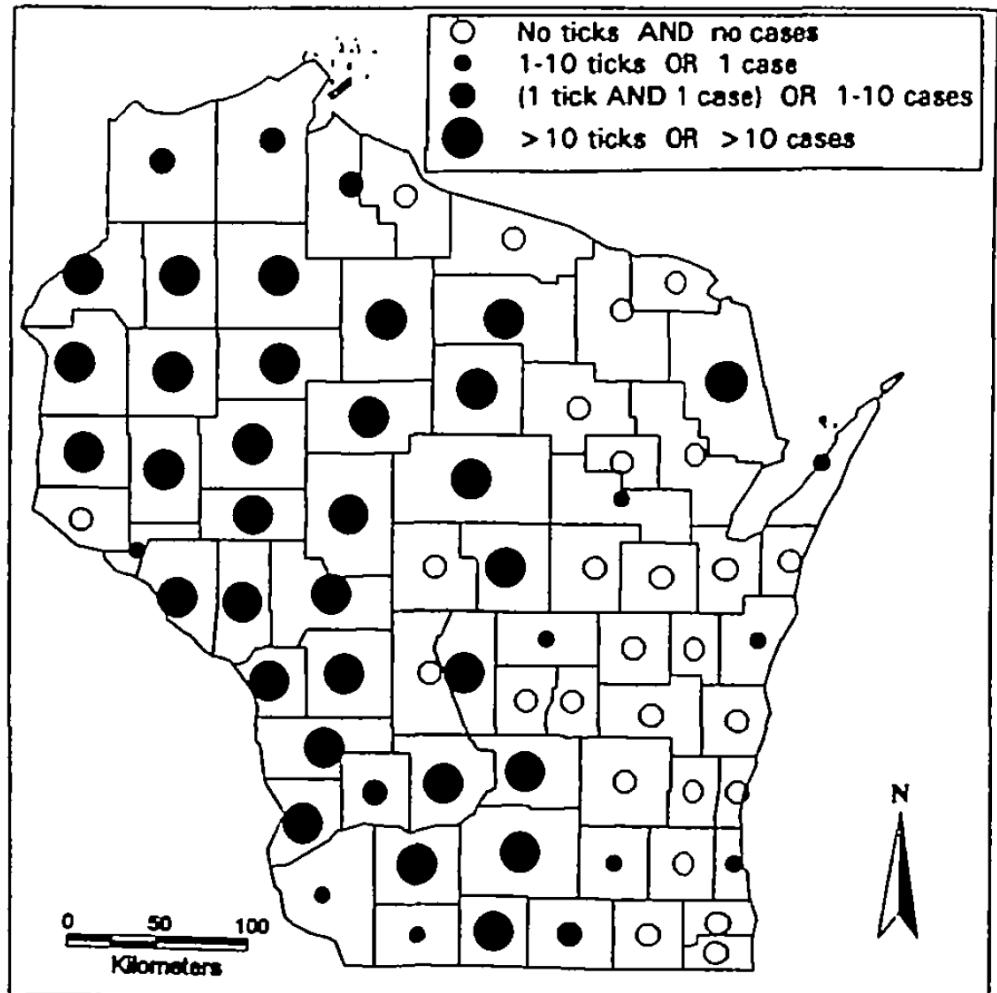


FIGURE 4. Lyme disease endemicity in Wisconsin counties, 1970–1995, as determined by county of exposure for human cases and known distribution of *Ixodes scapularis*.

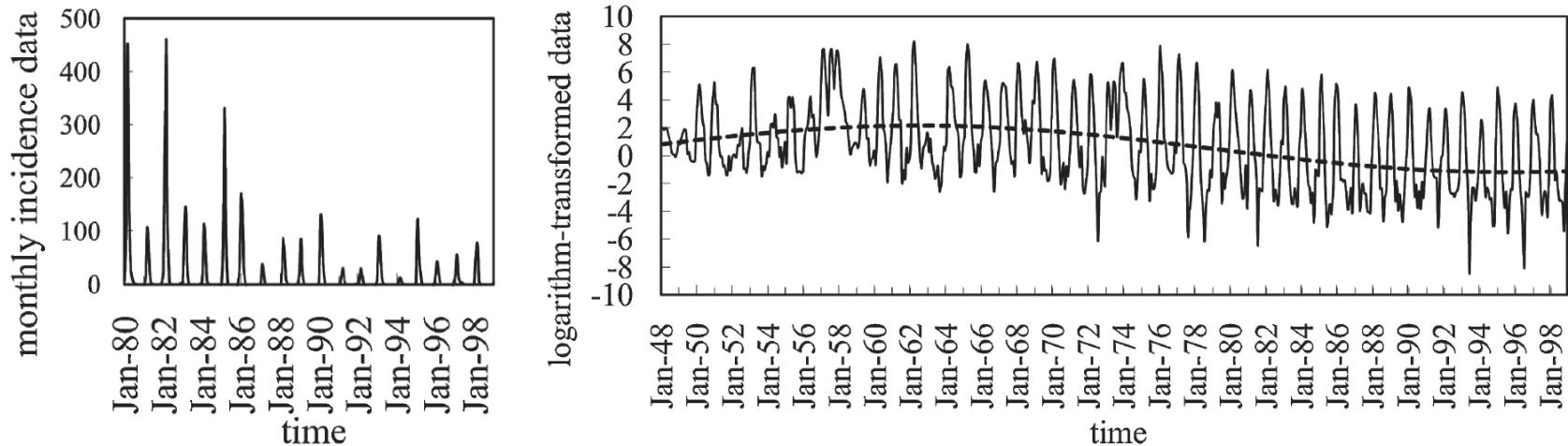
Extensions beyond 571 (*)

In analysis of time series, **everything** is correlated with everything else – no contributions from independent clusters.

Time Series Analysis of Incidence Data of Influenza in Japan

Ayako Sumi¹, Ken-ichi Kamo², Norio Ohtomo³, Keiji Mise⁴, and Nobumichi Kobayashi¹

J Epidemiol 2011;21(1):21-29
[doi:10.2188/jea.JE20090162](https://doi.org/10.2188/jea.JE20090162)



Methods tend to be model-based: see AMATH 582, FISH 506, STAT 519/520, and much work in econometrics.

Look ahead to 572 (*)

In 571, much of the methodology is new (e.g. INLA, lme4) – the product of recent methods research.

In Stat/Biost 572, as training for doing dissertation methods research, students develop skills in reading, understanding and critiquing contributions to this literature of statistical methods.

In the course, each student reads a chosen methods paper, and presents an in-depth critique of it*, both as a talk** and as a written report.

* This means reproducing the proofs, simulations, and if possible the examples – and sometimes correcting errors

** Actually 3 talks, all on the chosen paper

Look ahead to 572 (*)

Q. Where do I look for the papers?

- Ideally, in methods literature in an area you are already interested in
- See, for example, the Other Resources page on 571's class site
- A list of suggestions from faculty will be available – note this is a good way to start assessing potential PhD advisors, if you haven't done so already
- Papers must be approved by the 572 instructor – so don't delay looking for potential papers

Also note:

- Choosing a paper read in an earlier year is discouraged
- Stat students: you should read different papers (and topics) for each sequence
- Some *limited* faculty input may be available

Cheerio!



Aw things haes an end, an a pudden haes twa