Family-based genetic association modelling in a multistage sample

Thomas Lumley
@tslumley
t.lumley@auckland.ac.nz

(with Xudong Huang, Alastair Scott)

Biometrics Regional Conference, Hobart, 2015

The usual story

Once upon a time,

- there was a fearsome Problem
- along came a brave Method
- armed with Powerful maximal inequalities or Vast simulations
- so the Problem was vanquished

and we all live happily until the next episode

Today's story

In a galaxy far, far away

- researchers landed on a fascinating Problem
- and found Alien researchers also occupying it
- leading to some Skirmishes
- until everyone realised the Truth was more Complicated

and the Episode ends in an unsatisfactory but realistic truce.

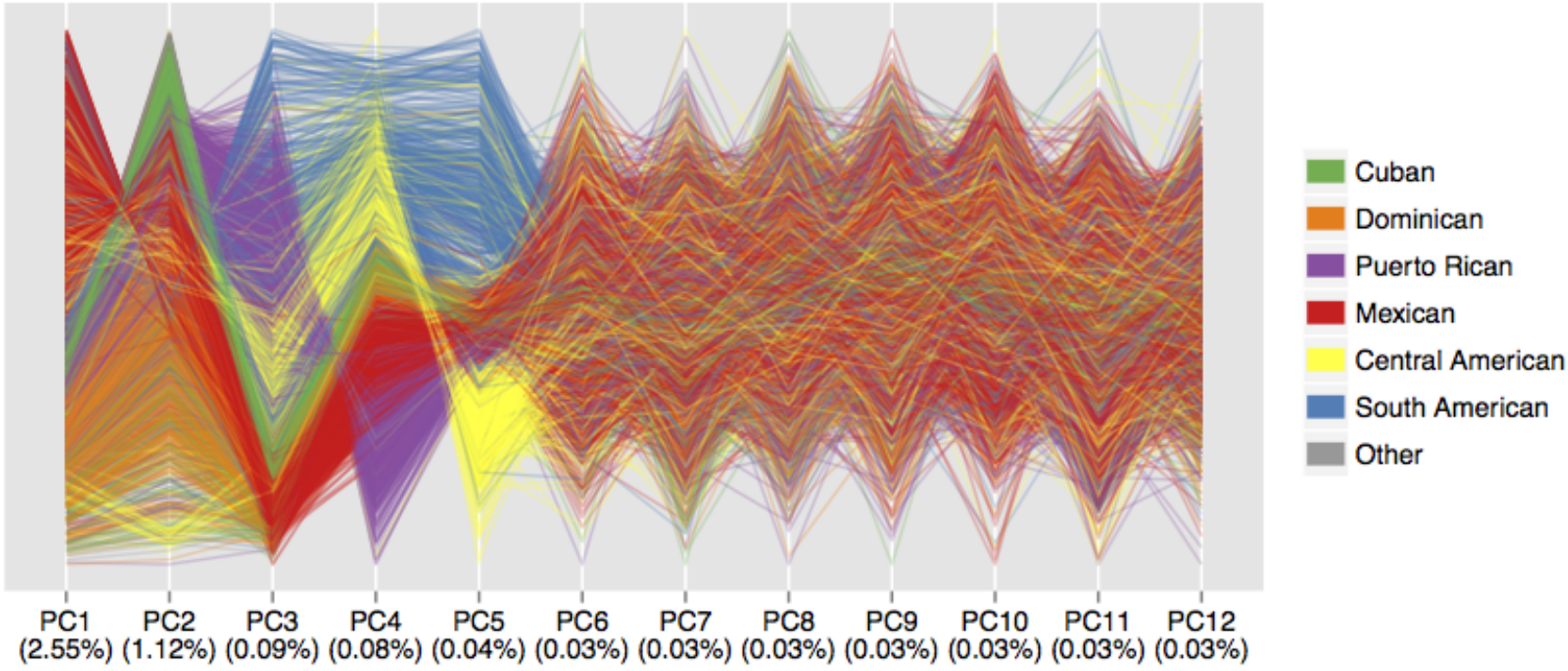
Basic Problem

- We have a complex (e.g. multistage, unequal-probability) sample from a population
- We would like to fit a mixed model to the **population** data distributions
- The sampling design is **not** part of our target of inference: just a nuisance
- We have sampling probabilities, but not necessarily the variables they are based on.

Genetic association

- A multistage probability sample of US, by census block and household: 12,000 people
- Genetic analyses need mixed model with relatedness and ancestry: 1 million SNPs
- Sampling is not part of genetic question: prefer not to have to model it

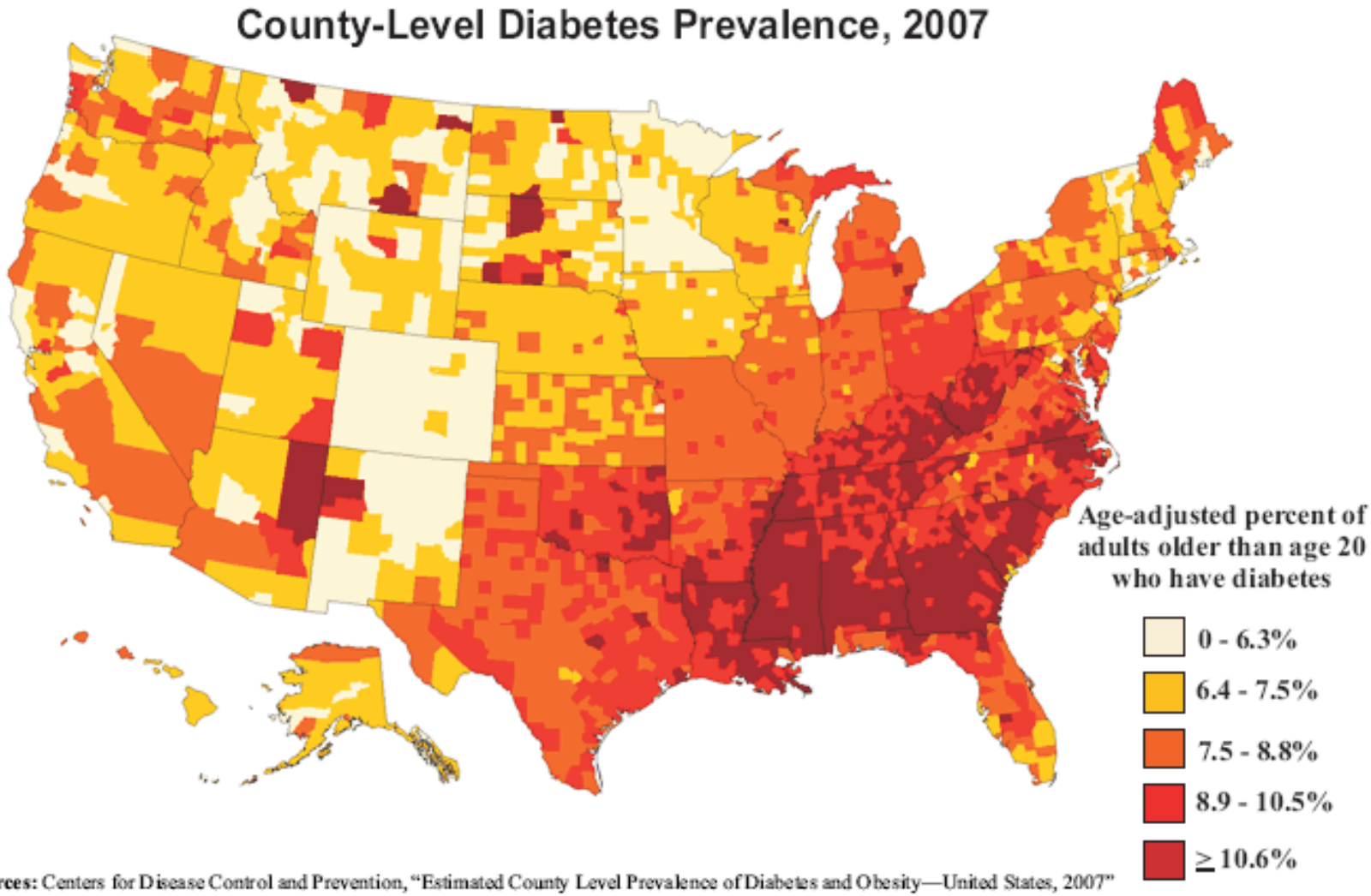
Ancestry (parallel coordinate plot, PCA)



Related problems

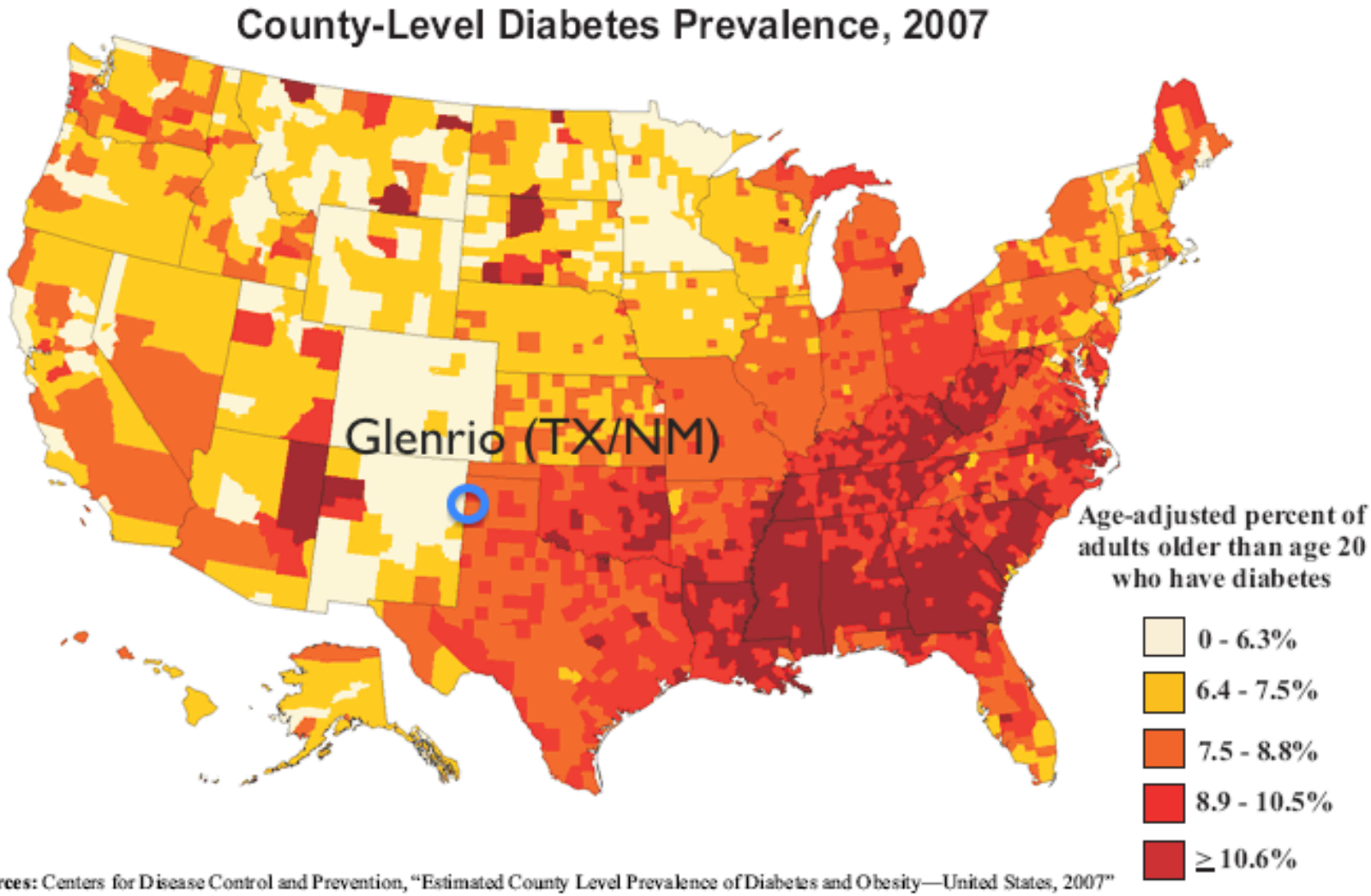
- Spatial smoothing models: need to model correlation between adjacent units
- Data sampled using one set of administrative boundaries; model uses a different set (e.g. primary school/secondary school)

Diabetes stops at the state border?



Sources: Centers for Disease Control and Prevention, "Estimated County Level Prevalence of Diabetes and Obesity—United States, 2007" *Morbidity and Mortality Weekly Report* 58 No. 45 (Nov. 20, 2009):1259-1263.

Diabetes stops at the state border?



Sources: Centers for Disease Control and Prevention, "Estimated County Level Prevalence of Diabetes and Obesity—United States, 2007" *Morbidity and Mortality Weekly Report* 58 No. 45 (Nov. 20, 2009):1259-1263.

Model(s)

Regression model:

$$Y_i = X_i\beta + Z_i b + \epsilon_i$$

with $b \sim N(0, V)$ and $\epsilon \sim N(0, \sigma^2)$

Sampling model:
$R_i \in \{0, 1\}$ is sampling indicator, with

$$E[R_i] = \pi_i, \quad E[R_i R_j] = \pi_{ij},$$

known at least for all units (pairs, triples) in the sample.

Population size N, sample size n.

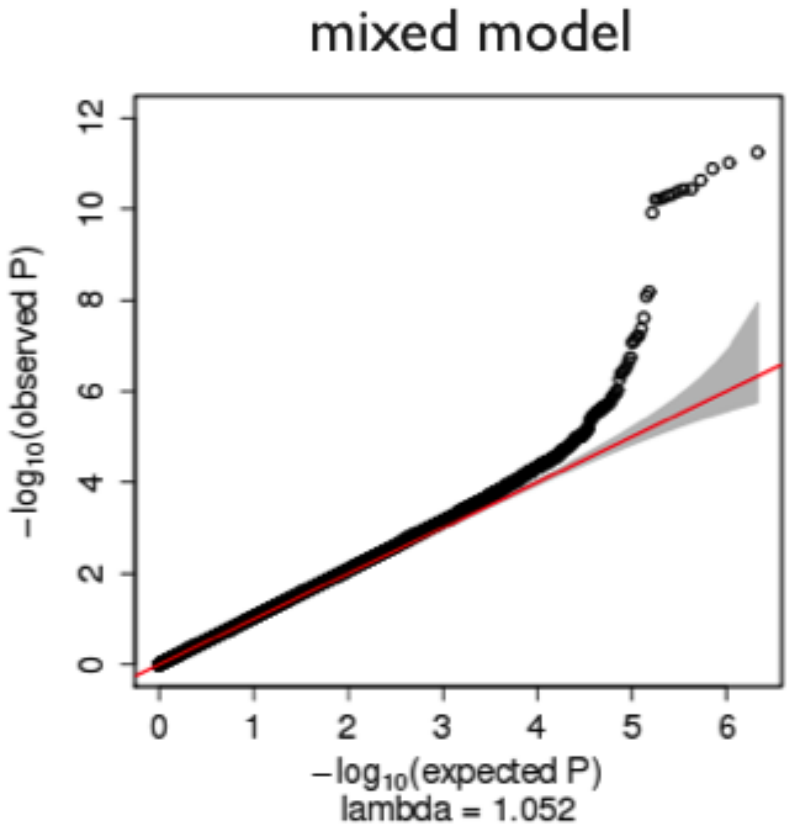
The Conflict

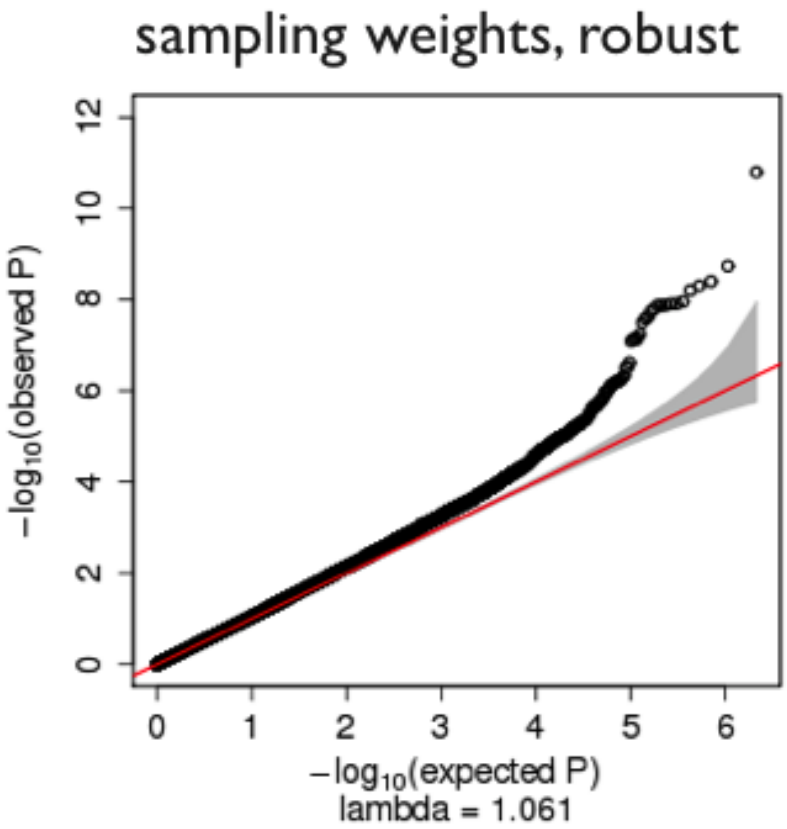
- Geneticists: fit the mixed model and ignore the sampling. What harm can it do?
- Samplers: fit the sampling weights and ignore the correlation (`,robust cluster()`). What harm can it do?

Samplers are right in theory; geneticists are right in practice: information loss matters.

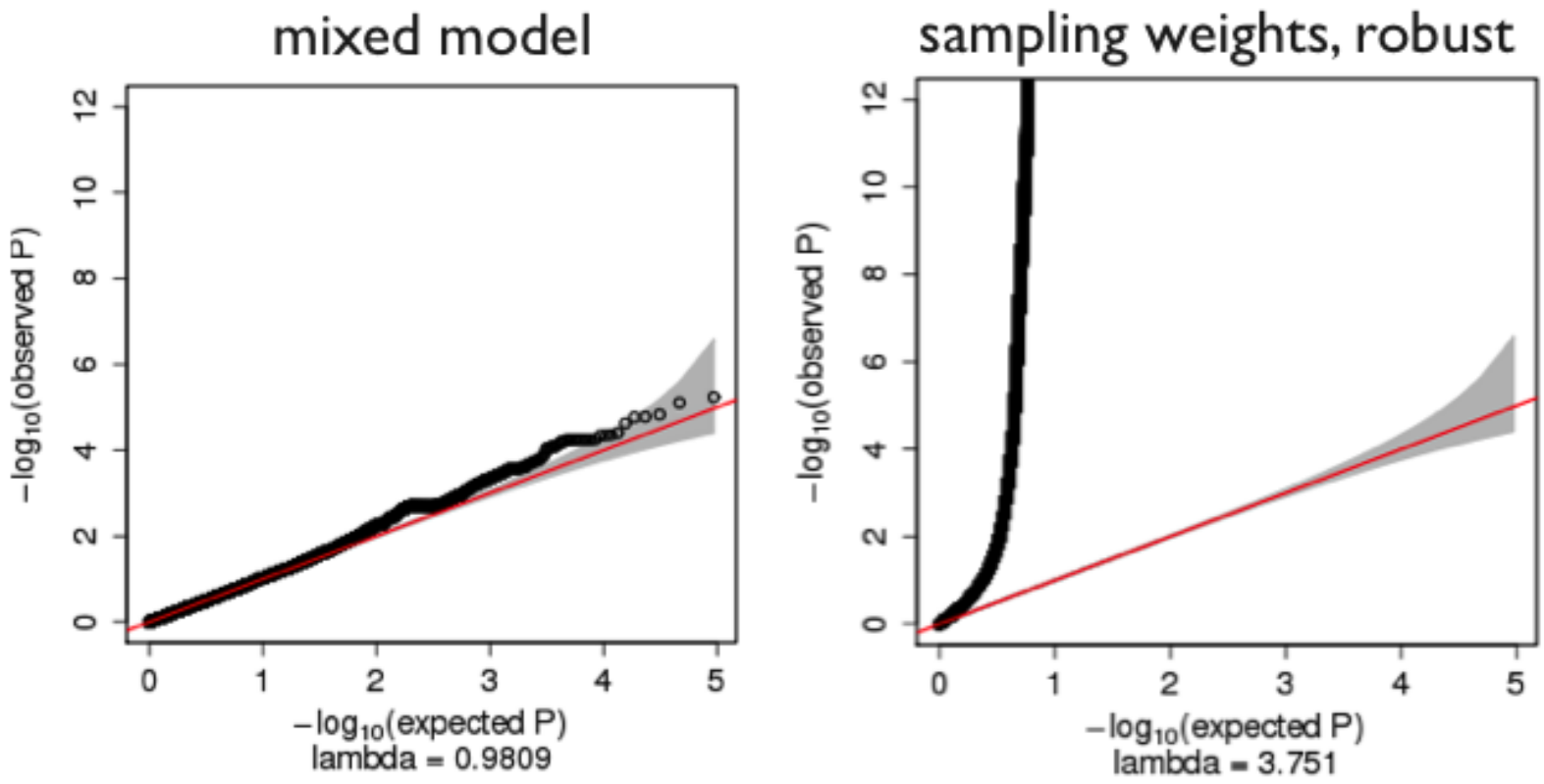
Can we compromise?

log(p-value) distribution: common ($\gtrsim 0.1\%$) variants





log(p-value) distribution: rare ($\lesssim 0.1\%$) variants



Reweighting

Population total

$$\sum_{i=1}^{N} x_i$$

Estimated by

$$\sum_{i=1}^{N} \frac{R_i}{\pi_i} x_i$$

Reweighting

Population total

$$\beta = \arg\min \sum_{i=1}^{N} (y_i - x_i\beta)^2$$

Estimated by

$$\hat{\beta} = \arg\min \sum_{i=1}^{N} \frac{R_i}{\pi_i} (y_i - x_i\beta)^2$$

Reweighting

Population total

$$\sum_{i=1}^{N} x_i(y_i - x_i\beta) = 0$$

Estimated by

$$\sum_{i=1}^{N} \frac{R_i}{\pi_i} x_i(y_i - x_i\hat{\beta}) = 0$$

Reweighting?

Mixed model loglikelihood

$$\ell(\beta, \theta) = -\frac{1}{2} \log |V(\theta)| - \frac{1}{2}(y - x\beta)^T V^{-1}(\theta)(y - x\beta)$$

Sequential likelihood?

The case in the literature:

- random effects are iid at each stage of the model
- simple random sampling (clusters within strata) at each stage of the design
- model and design structure are the same

Can use the sequential independence to reweight based on sampling probabilities at each stage (with suitable rescaling for 'degrees of freedom')

[Doesn't handle the SoL genetic data]

Composite likelihood?

Likelihood for a pair of observations is still Gaussian, with appropriate submatrix of $V(\theta)$ as variance.

Unweighted pairwise composite loglikelihood (Lindsey; Heagerty & Lele)

$$\ell_C(\beta, \theta) = \sum_{i,j} \ell_{ij}(\beta, \theta)$$

with

$$\ell_{ij} = -\frac{1}{2} \log \begin{vmatrix} \sigma_i^2 & \sigma_{ij} \\ \sigma_{ij} & \sigma_j^2 \end{vmatrix} - \frac{1}{2} \begin{pmatrix} y_i - x_i\beta \\ y_j - x_j\beta \end{pmatrix}^T \begin{pmatrix} \sigma_i^2 & \sigma_{ij} \\ \sigma_{ij} & \sigma_j^2 \end{pmatrix}^{-1} \begin{pmatrix} y_i - x_i\beta \\ y_j - x_j\beta \end{pmatrix}$$

Computation

There are $\binom{n}{2}$ pairs: 72 million for genetic example

We need to fit a million models.

Analytic simplifications:

- Use only some of the pairs, based on (i, j) correlation
- Fit adjustment model fully, then just use one-step update for each genetic variant

Still a lot of work.

Write loglikelihood and derivatives for each pair as vectorised code, using explicit formula for determinant, inverse

If n isn't too large, run all together, otherwise run for each i.

Parallelizes very well, either directly or by outsourcing to database

Efficiency

Folklore says composite likelihood is pretty efficient

Intuition says two moments should be pretty good for Normal

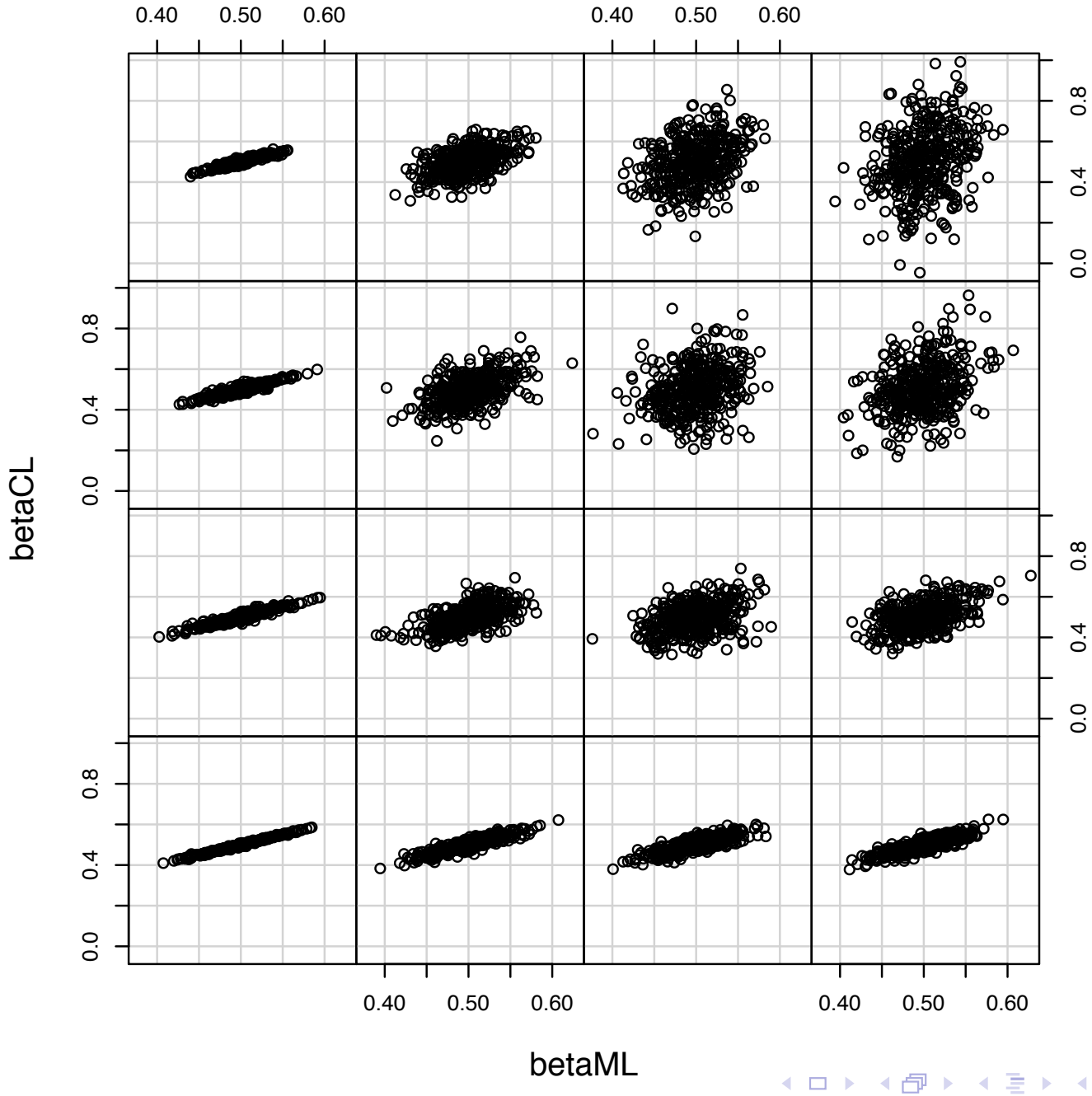
For β, how does it compare to GLS plugging in variance component estimates?

Example:

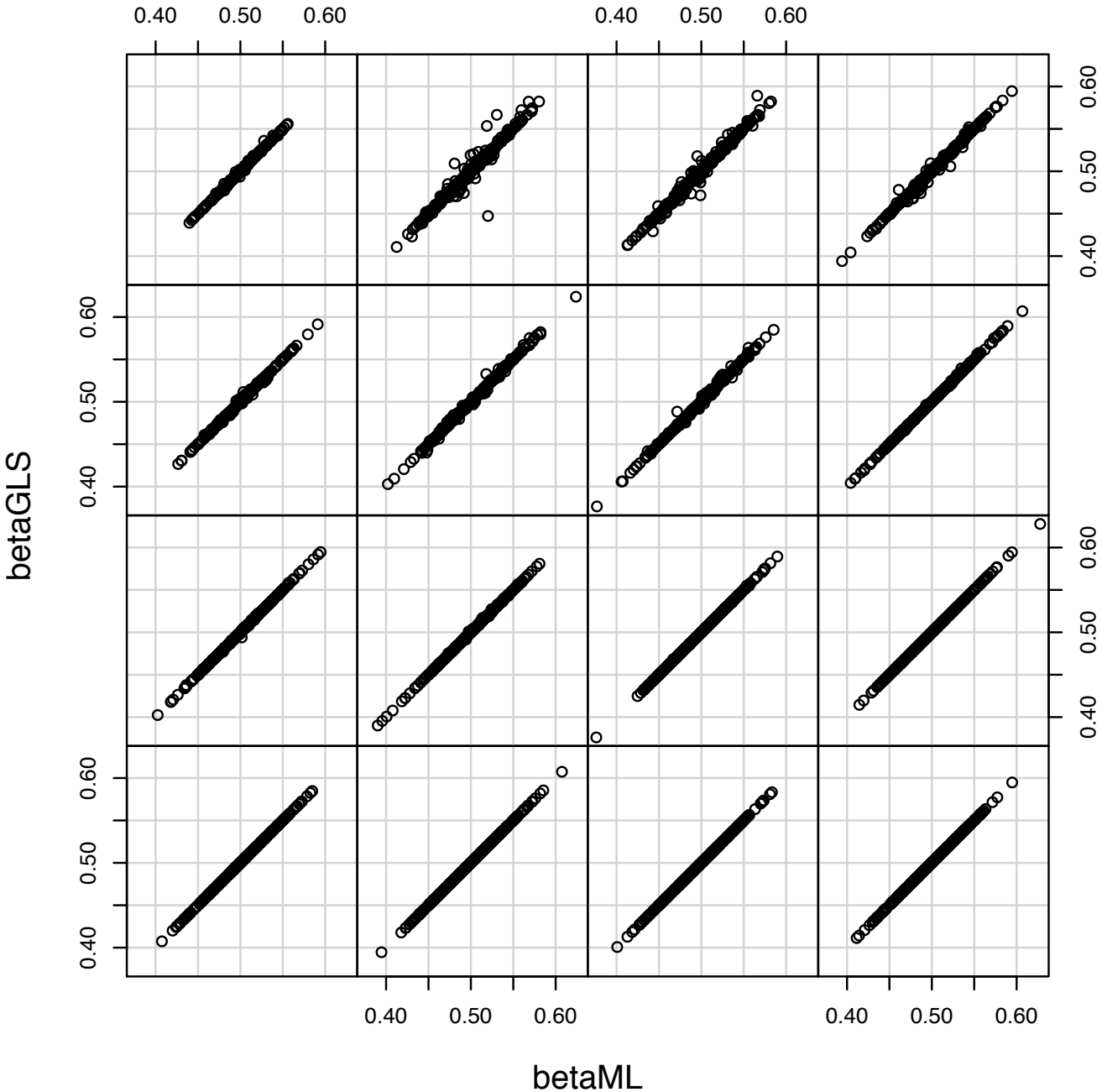
$$Y_i = a_i + \beta X_{ij} + \epsilon_{ij}$$

where $X_{ij} = b_i + \eta_j$

Efficiency (random intercept)



Efficiency (random intercept)



Sampling

Weighted pairwise log likelihood

$$\hat{\ell}_C(\beta, \theta) = \sum_{i,j} \frac{R_i R_j}{\pi_{ij}} \ell_{ij}(\beta, \theta)$$

In literature only for special case where model and sampling structure the same; generalises easily.

Efficiency

Assume the variance components known

- problem reduces to Generalised Least Squares

$$Y_i = x_i\beta + \epsilon_i$$

with $cov[Y]^{-1} = \Omega$ known (or Ω^{-1} known and *sparse*)

- Census parameter: $\beta^* = (X^T\Omega X)^{-1}(X^T\Omega Y)$
- Pairwise estimating equation

$$\sum_{i=1}^{N} x_i\omega_{ii}(y_i - x_i\beta) + \sum_{i\neq j}^{N} x_i\omega_{ij}(y_j - x_j\beta) = 0.$$

Sampling

Consider simple random sampling without replacement.
Design-based inference **should** be close to model based (?)

- $\pi_{ij} = n(n-1)/N(N-1) \approx \pi_i^2$

$$\sum_{R_i=1} x_i \omega_{ii} (y_i - x_i \beta) + \sum_{i \neq j,\, R_{ij}=1} \frac{N-1}{n-1} x_i \omega_{ij} (y_j - x_j \beta) = 0$$

Hugely more weight on $i \neq j$ terms (?!)

Resulting weight matrix (ω_{ij}/π_{ij}) not even positive definite.

Why not OLS?

Advantage of generalised least squares over ordinary least squares was efficiency, but weighting ruins it.

Why isn't

$$\sum_{i=1}^{N} x_i \omega_{ii} (y_i - x_i \beta) = 0.$$

weighted to

$$\sum_{i=1}^{N} \frac{R}{\pi_i} x_i \omega_{ii} (y_i - x_i \beta) = 0$$

design-consistent?

Contextual confounding

Suppose we are interested in $Y_i|X_i$ (**your** genes), not $Y_i|\mathbf{X}$ (everyone's genes) and β is the relevant coefficient

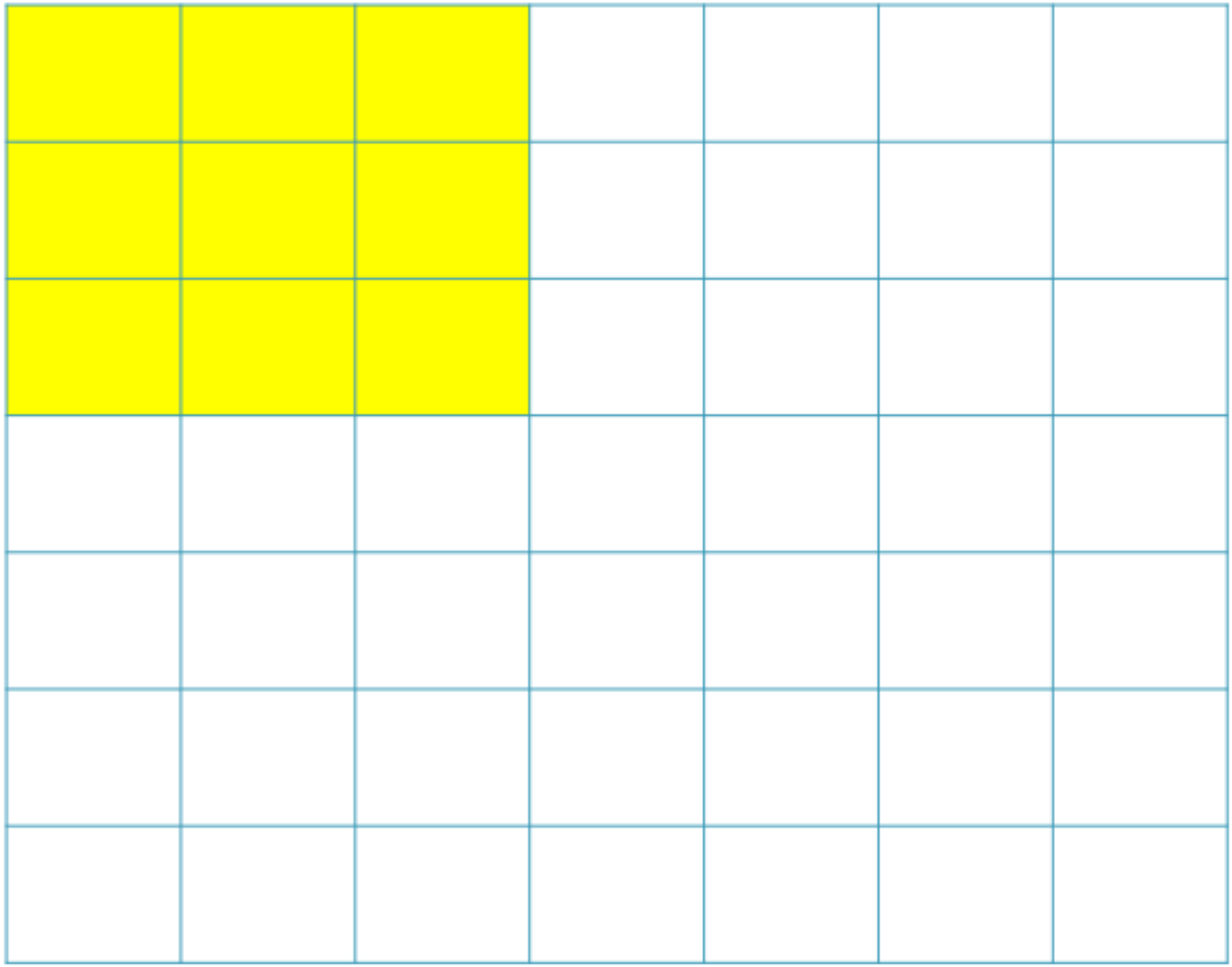
If $(Y_j - x_j\beta)$ is correlated with x_i,

$$E[x_i(Y_j - x_j\beta)] \neq 0.$$

A GLS estimator of β is typically biased, but the OLS estimator isn't.

[Pepe & Anderson 1994; Pan, W., Connett, J.E. and Louis, T.A. (2000)]

Why $i \neq j$ upweighted



Rethink design-consistency?

Truly design-based estimator sacrifices efficiency to reproduce bias of census estimator in the presence of contextual confounding

Probably not a good tradeoff.

- We can change model to get $\omega_{ij} = 0$ where $\pi_{ij} \ll \pi_i, \pi_j$
- Can we assume "no contextual confounding" and get consistent estimator without upweighting of $i \neq j$ terms?

What's random?

Need to be more detailed about model: what π_{ij} can depend on.

- If $R \perp Y \mid X$ or $R \perp X \mid Y$, can reweight just using π_i.
- If $R \perp \{X, Y\} \mid Z$, can use just $E[\pi_{ij}|Z]$
- If no contextual confounding, can use $\pi_{ij}/E[\pi_{ij}|x_i]$
- If also correctly-specified cross-sectional mean, can use $\pi_{ij}/E[\pi_{ij}|x_i, x_j]$

Given a set of estimating functions, it's just GMM or Gauss–Markov Theorem.

Similar to Lai & Small (2005, JRSSB) on time-dependent covariates in longitudinal data.

Summary

- Design-based inference for mixed models is hard
- Composite likelihood gives a practical approach
- But weighting is more confusing than it looks
- Doing better than OLS requires assumptions
- The assumptions need to be case-specific.

Questions?