Random Effects Old and New

- Many things that are now called "random effects" would not have been recognized as REs 50 years ago.
- The distinction between old- & new-style REs has consequences.

Outline

- Old style: Definition, example.
- Some background: Three senses of "probability"
- New style: They implement smoothing/shrinkage, and they're part of the model's mean, not its error variance.
- Consequences: Old & new REs require distinct ways to
  - Do inference and prediction
  - Do simulations for evaluating statistical methods
  - Interpret analytical artifacts

Definition: Old-style random effects

Scheffé (1959, p. 238):

- The *levels* of the RE are *draws* from a *population*,
- The draws are not of interest in themselves but only as samples from the larger population.

Example of old-style random effects

New objective methods to count and measure nerve fibers in skin and mucosa (Kennedy Lab).

A recent study (Panoutsopoulou et al 2013) had:

- 25 "normal" (non-diabetic) subjects
- Skin sampled by biopsy and blister (method)
- Samples taken on the calf and on the foot (locations).

This design has three old-style random effects:

- Subject main effect
- Method-by-subject interaction
- Location-by-subject interaction

It also has a residual (error term) = method-by-location-by-subject interaction

This design has three old-style random effects:

- Subject main effect
- Method-by-subject interaction
- Location-by-subject interaction

These random effects describe how:

- Average nerve density varies between subjects
- Blister minus biopsy varies between subjects
- Foot minus calf varies between subjects

These *are* old-style random effects

- The levels of the REs (subjects) are a sample (tho' not formal).
- The levels are not interesting in themselves but only as representatives of non-diabetic adults.
- The object was to measure, in that population, differences between methods and locations.
- These random effects are part of error variation: They capture (nuisance) correlation within subject.

A new measurement on a new person would involve a new draw of all the variance components.

A new measurement on one of these 25 people would involve a new draw on only the error term.

Some background: Three senses of "probability"

1. Draws from a random mechanism, either one we create and control, or one we imagine is out in the world. (Frequentist notion.)

2. A person's uncertainty about an unknown quantity. (Subjective Bayesian notion.)

3. A descriptive device.

Example of #3:

The heights of US-born 52-year-old males employed by U of M

can be described as looking like n iid draws from $N(\mu, \sigma^2)$.

Though this doesn't imply anyone's height is a draw from a random mechanism or is even uncertain.

Definition (sort of): New-style random effects

A new-style RE differs from old-style REs in at least one of these ways:

1. The levels of the effect are not draws from a population because there is no population. The mathematical form of a random effect is used for convenience only.

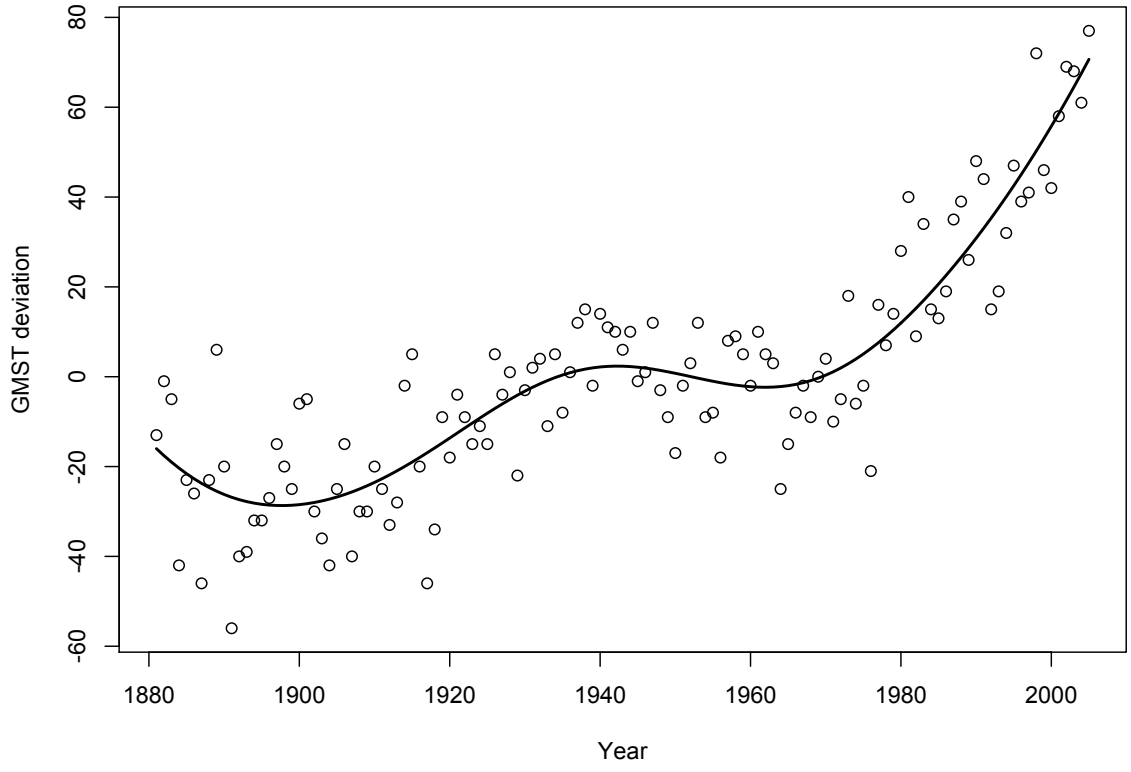
2. The levels of the effect come from a meaningful population but they are the whole population and these particular levels are of interest.

3. A sample has been drawn, but the samples are all from a single level of (draw from) the model's random effect, and that level is of interest.

(1) The levels of the effect are not draws from a pop'n; there is no pop'n. The mathematical form of a random effect is used for convenience only.

Example: Mixed linear model representation of penalized splines

Object: Draw a smooth curve through the data.



A penalized spline is just a linear model with constrained coefficients.

Minimizing

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \alpha\mathbf{u}'\mathbf{u}$$

is formally identical to estimating $(\boldsymbol{\beta}, \mathbf{u})$ in the mixed linear model

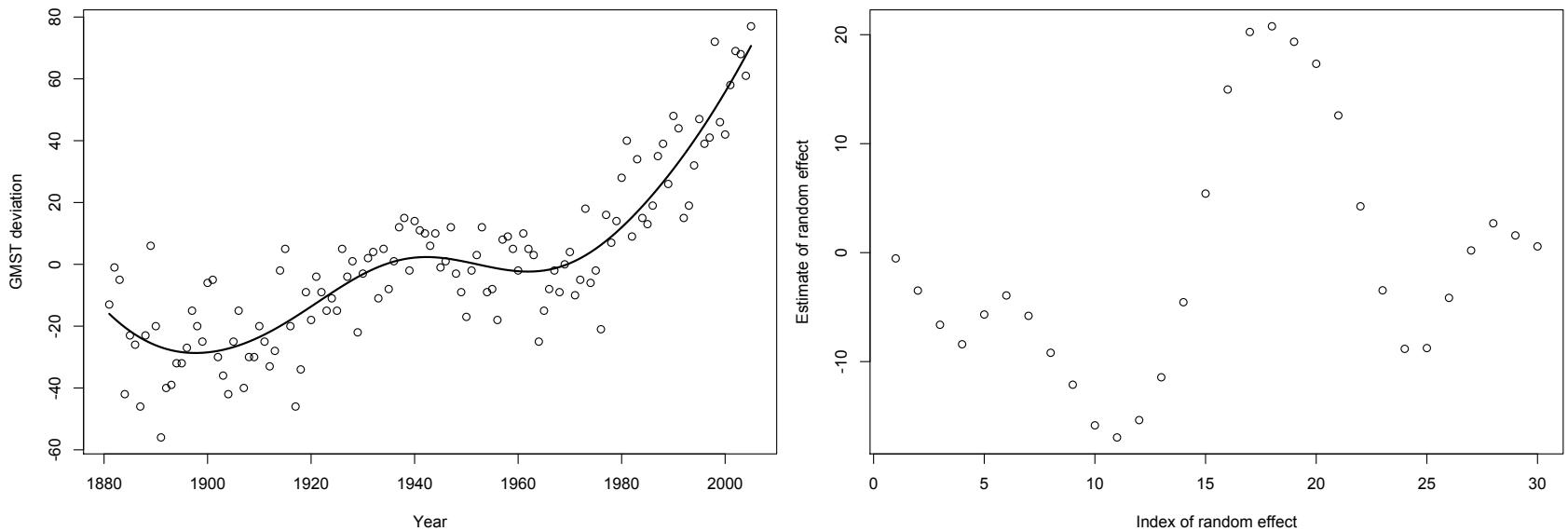
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \quad \epsilon_i \sim N(0, \sigma^2), \quad u_j \sim N(0, \tau^2)$$

when σ^2 and τ^2 are taken as given; $\alpha = \sigma^2/\tau^2$

RWC (p. 138): "[T]he mixed model formulation of penalized splines [is] a convenient fiction . . . [It] is a reasonable (though not compelling) Bayesian prior for a smooth curve, and [max RL] estimates of the smoothing parameter . . . generally behave well".

The analysis has the *form* of an RE analysis,
but $\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}$ is not a draw from a random mechanism.

And the fitted u_j (right panel below) don't look like iid $N(0, \tau^2)$:



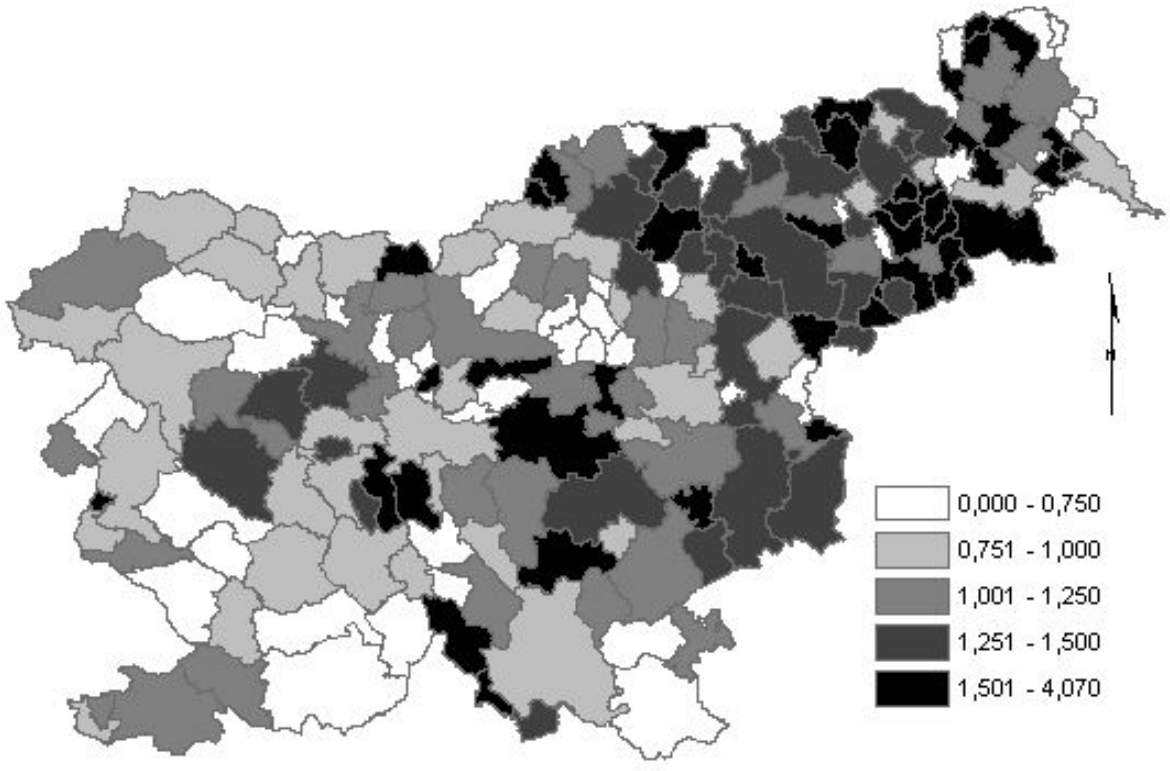
- Like β, $\mathbf{u}$ is just part of the model's mean structure; $\mathbf{u}$'s distribution just constrains the estimates of the u_j.
- *We choose* this model to serve a particular purpose.
- It is senseless to imagine that this model generated the data, or that draws could be made from it.
- The random-effect form merely provides a flexible family of smooth mean functions and some discipline on the fit.

. . . After adopting this convenient fiction, RWC behave like conscientious statisticians, checking for heteroskedastic errors, non-linearity, etc.

(2) The levels of the effect come from a meaningful pop'n but they're the whole pop'n and are of interest.

Example: Stomach cancer in Slovenia, 1995-2001

Standardized incidence ratio, by municipality



Disease maps are commonly smoothed using models with REs, e.g., this oft-used model by Besag, York, and Mollié (1991).

O_i = stomach cancers in municipality i $\sim$ Poisson with mean

$$\log \mathrm{E}(O_i) = \log(E_i) + \beta SEc_i + S_i + H_i$$

- E_i = expected # of cancers.
- SEc_i = SES, centered
- $\mathbf{H} = (H_1, ..., H_{194})' \sim$ iid $N(0, \tau^2)$ (heterogeneity)
- $\mathbf{S} = (S_1, ..., S_{194})' \sim$ ICAR (spatial clustering), with
  - $p(\mathbf{S}|\sigma^2) \propto \exp(-\mathbf{S}'\mathbf{Q}\mathbf{S}/2\sigma^2)$; $\mathbf{Q}$ captures spatial neighbor pairs.

Unlike the spline

- There is a meaningful population, BUT
- the municipalities are the whole population and they're of interest.

Like the spline:

- Perhaps some random process produced the O_i, but
- $\mathbf{S} + \mathbf{H}$ was not produced by a draw from CAR + iid, and
- even though new counts could be made for 2002-2008, they wouldn't be an iid draw from the same mechanism.

It is hard to see how $\mathbf{S} + \mathbf{H}$ can usefully be understood as a draw from a random mechanism.

The intuition motivating a spatial model – near municipalities are more similar than far – is descriptive, not mechanical.

Less problematic: **S**'s CAR model is descriptive (3rd sense of probability).

We could say the 192 Slovenian S_i, if observed, would look like a draw from a CAR model.

Like the heights of 52-year old men, this doesn't mean **S** was drawn from a random mechanism; it's just a way to describe a group of constants.

Some would say it's natural to think of S's CAR distribution as a statement of subjective probability (2nd sense of probability).

If we view **S**'s distribution as

- description, we can use that description in a statistical method;
- subjective probability, we can use it in a Bayesian computation.

Either way:

- The random effect is a device *we choose* for a particular purpose.
- It is senseless to imagine that this model *generated* the data.

(3) A sample has been drawn, but the samples are all from one level of (draw from) the model's random effect, and that level is of interest

In some region, we want to know the fraction of iron at a given depth.

We label this fraction $\mu + W(s)$, at location s.

μ and $W(s)$ are fixed but unknown.

We observe $y(s_i)$ at location s_i and model it as

$$y(s_i) = \mu + W(s_i) + \text{error}(s_i), \text{ with iid error}(s_i).$$

$W(s)$ is commonly modeled as a Gaussian process:

$$(W(s_1), ..., W(s_n)) \sim \text{multivariate normal.}$$

This is unlike the GMST and Slovenia examples:

- The s_i are a sample of possible locations.
- These s_i are not so interesting; we want to know about the region, which we can call a population.
- We could draw new s_i or make new measurements at the original s_i.

So far this sounds like an old-style random effect, but it's not.

A draw from this GP model is a function on the 2-D region of interest.

As with the spline, there's no population: The RE's "population" is the hypothetical infinite population of draws from the GP.

As in the stomach-cancer example:

- Exactly 1 draw has been made from this RE; no more can or will be.
- The whole point of gathering data is to learn about this one draw. If we measure y_i at new s_i, we just learn more about this one draw.
- We may describe $W(\mathbf{s})$ by saying $\{W(\mathbf{s}_i)\}$ looks like a draw from $N(\mathbf{0}, \Sigma)$; or that this distribution describes our uncertainty about $W(\mathbf{s})$, and we can deploy either of these in a statistical method.

In these respects, $\{W(s_i)\}$ is identical to $\mathbf{S} + \mathbf{H}$ in the Slovenian stomach-cancer example.

Comments on new-style random effects, 1

New-style random effects can all be understood as formal devices that implement smoothing or shrinkage.

This is obvious for penalized splines.

This is less clear for the other examples, perhaps because we habitually think of spatial models in terms of covariance matrices.

Comments on new-style random effects, 2

New-style random effects are part of the model's mean structure, not part of its variance structure.

The distribution of a new-style random effect does not embody or represent the mechanism that produced the data.

$\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}, \mathbf{S}$, or $W(s)$ is just a group of fixed, unknown constants.

The random effect's distribution is something we choose.

In a given situation, some choices are better than others.

But "better" means they give better estimates of the unknowns, not that they better represent the mechanism that produced the data.

Comments on new-style random effects, 3

Maybe the true $\mathbf{X}\beta + \mathbf{Zu}, \mathbf{S}$, or $W(s)$ arose as a draw from some random mechanism.

But it makes no sense to imagine further draws, and the value of that single draw is of interest.

As *hypothesized mechanisms* of the process that produced the data, these models are silly.

- If we make this mistake, we are mistaking the shovel for the process that produced the soil.

It is more accurate to think of these models as simply descriptive, in a superficial manner, and useful for producing estimates.

Comments on new-style random effects, 4

Re the Slovenian example, I've heard

- "$S_i + H_i$ is only in the model as an error term, like the error term in linear regression."

Well, what "error" does $S_i + H_i$ capture?

- Local variation in the mean of the data-generating process that's not captured by SEc.
- That is, local bias or lack of fit in the model's fixed effects.
- Berkson distinguished this kind of error from "classical" error, e.g., the error term in linear regression.

New-style REs are Berksonian, part of the model's mean, not its error.

Comments on new-style random effects, 5

Some spatial analyses do involve old-style random effects.

Example:

- Ozone in the Boston area, daily data for m years.
- Days may be an old-style random effect.

But not necessarily ...

- Suppose we're interested in a specific week.
- We may describe that week's spatial ozone gradient using an RE, but it's part of the model's mean, not its error variance.
- There is a meaningful sense in which this fixed but unknown feature of Boston was drawn from a probability distribution, but that sense is not relevant to our question.

Practical consequences: Inference & Prediction

"Inference" = Analyses focused on the present dataset and on mechanisms that supposedly generated it.

"Prediction" = Analyses focused on data related to the present dataset but as yet unobserved or unknown.

Practical consequences: Inference

Bayesian:

- Bayesian calculations are the same, old-style or new-style.
- But the distinction is relevant to prior specification.

Non-Bayesian: Old-style REs are deeply embedded in the terminology.

- "BLUP": "Unbiased" refers to the expectation over random effects as well as error terms.
- A penalized spline fit is a BLUP; it flattens peaks and fills valleys. These are biases.
- So the term "BLUP" really shouldn't be used for new-style REs.

But this confusion has more serious consequences ...

Example: Confidence interval for a penalized spline fit

Representing the spline with the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \quad \epsilon_i \sim N(0, \sigma^2), \quad u_j \sim N(0, \tau^2)$$

If we take $\mathbf{u}$ as a new-style RE

- A CI should use $\text{var}(\hat{f}(x)|\mathbf{u})$.
- This CI's coverage is too low for x where $\hat{f}(x)$ is biased, because it is centered at $E[\hat{f}(x)|\mathbf{u}]$, not $f(x)$.
- The obvious fix is to subtract the bias from $\hat{f}(x)$.

But if we take $\mathbf{u}$ as an old-style RE

- The solution is to widen the CI at each x.
- But the interval is still centered in the wrong place.
- Coverage is still too low in areas of high curvature and too high in areas of low curvature.

Practical consequences: Prediction

For old-style REs (e.g., nerve-density example), two cases:

(1) Predicting biopsy and blister nerve density measurements from a new subject's calf and foot.

- Each new measurement's variance is the sum of all three REs (subjects, method-by-subject, location-by-subject), and residual.

(2) Predicting a new biopsy from the foot of an already-sampled subject.

- Each new measurement's variance is just residual error variance.
- Predictive SE also accounts for uncertainty about the values of the three REs for that subject.

Nobody disputes this. However ...

Prediction for New-style REs

All are like predicting a new measurement on an already-sampled subject:

- No new sampling of the RE is possible.
- You must condition on the existing "draw" of the RE.
- The new measurement's variance is error variance only.

In the mineral-exploration example, it is possible

- to measure again at an observed s_i or
- to measure at a new s_0.

In either case, the RE $W(s)$ has already been drawn and is just unknown.

Otherwise, making a prediction would involve re-drawing the process that produced the ore seam – obviously ridiculous.

Practical consequences: Interpreting analytical artifacts

Spatial confounding as in the Slovenia dataset:

Interpretation depends on whether the RE is old-or new-style

Practical Consequences: Simulation experiments

Principle:

- A simulation experiment is intended to answer specific questions.
- The experiment's design must enable it to answer those questions.

For old-style REs the questions all amount to: Measure behavior of estimates, tests, intervals for FEs and variance components.

We'd answer these questions (using the nerve-density example) by simulating a subject's 4 measurements this way:

- Make a draw of each of the three REs, and
- Make four draws from residual error.
- Each fake observation is a sum of true FEs, drawn REs, and error.

This mimics the way the real data were generated: by sampling subjects.

Simulation experiments for new-style random effects

- Draws never need to be made from a new-style RE.
- It's usually incorrect and self-defeating to draw from a new-style RE.

Questions that might be asked about methods using new-style REs:

- What is the fit's bias or MSE at specific predictor values?
- What is the coverage of a particular type of confidence interval?

To answer such questions, it is *wrong* to draw from a new-style RE.

- Argument from first principles: New-style REs are only convenient fictions; taking them literally is a conceptual error.
- Pragmatic argument: Simulating from new-style REs does not produce data with relevant features and is thus self-defeating.

Argument from first principles

A new-style RE is just a convenient fiction; taking it literally is a conceptual error.

In evaluating a penalized spline procedure (e.g., a basis), the question is how well it captures turns, peaks, valleys.

Therefore, data should be simulated by adding residual error to specific true f(x) having turns, peaks, valleys.

Simulating data by drawing

$$f(x) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}, \quad u_j \sim N(0, \tau^2)$$

- has a different true $f(x)$ for each draw and
- leaves out precisely the most relevant features.

Pragmatic argument

Draws from models with new-style REs are not consistent with our intuition, which arises from *fitting* such models.

Applied to a spline with a truncated quadratic basis:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \quad \epsilon_i \sim N(0, \sigma_e^2), \quad u_j \sim N(0, \sigma_s^2)$$

$\mathbf{u}$ contains changes in the quadratic coefficient at the knots.

Fitting this p-spline, larger $\sigma_s^2 \Rightarrow$ wigglier, rougher fit.

Draws from $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, u_j \sim N(0, \sigma_s^2)$ don't behave like this.

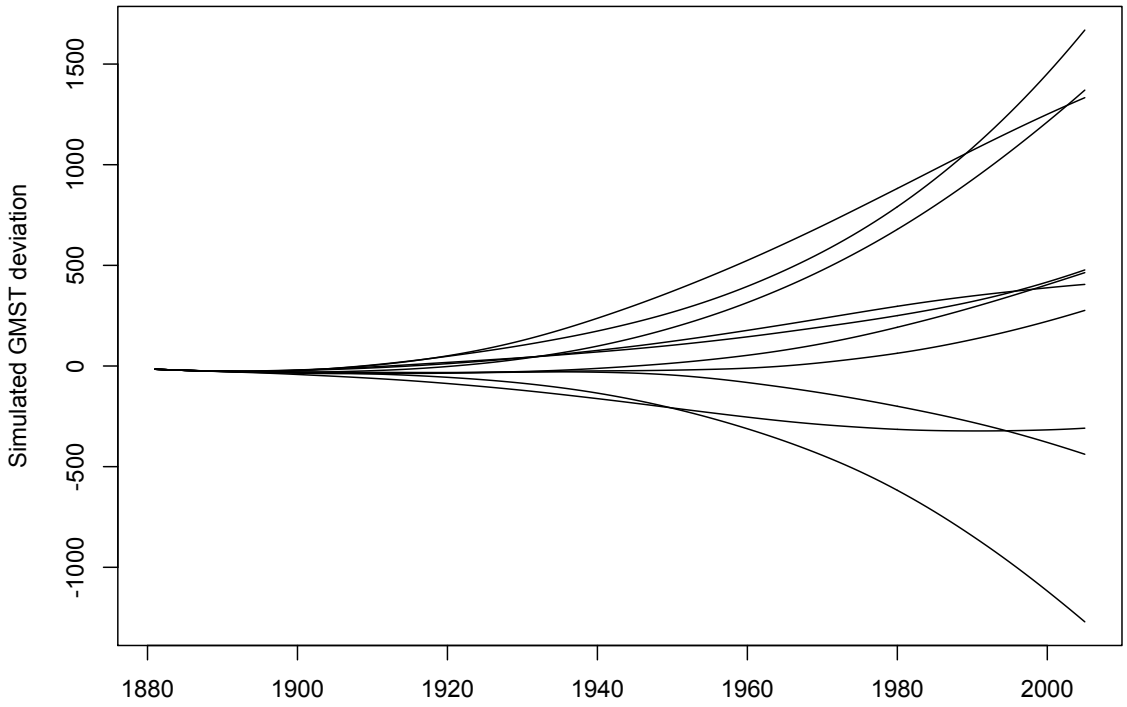
Proof of pragmatic argument

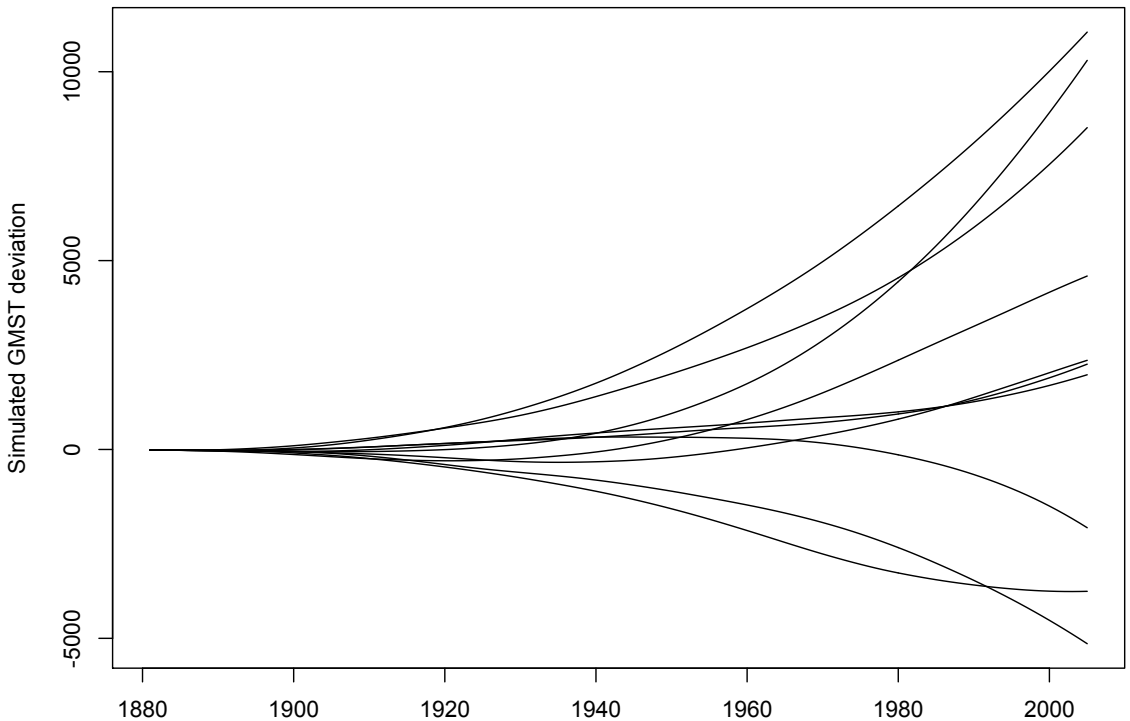
Use the spline fit to the GMST data: truncated-quadratic basis, 30 knots.

Draw 10 curves from $\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}$ using the estimated $\boldsymbol{\beta}$ and $\mathbf{u}$'s estimated variance $\hat{\tau}^2 = 947$.

Draw 10 more curves, with only one change: $\tau^2 = 94{,}700$.

The results are on the next page. Can you tell which is which?


Simulated GMST deviation
1500
1000
500
0
-500
-1000
1880
1900
1920
1940
1960
1980
2000


Simulated GMST deviation
10000
5000
0
-5000
1880
1900
1920
1940
1960
1980
2000

The draws with bigger variance only have a larger vertical scale.

These curves lack interesting features; to produce an interesting feature, you need a spectacularly improbable $\mathbf{u}$.

Instead, to answer any kind of real question, you must specify interesting true $f(x)$ and simulate datasets *by adding residual error only*.

Sometimes it seems harmless to draw a true curve from a new-style RE and then repeatedly add residual errors.

But this gives only bland curves lacking features you'd want in a simulation experiment.

OK, so what about animal breeding?

(About which I know hardly anything . . .)

Dataset: It's a sample of cows (but you'd include all cows if you could).

Object: Identify superior animals for breeding.

- The cows are a sample (of existing and potential cows).
- The variance components are meaningful, not just a fiction.
- You could and indeed do make new draws (of potential cows).

BUT . . .

The whole point is to identify superior animals $\Rightarrow$ you're interested in the levels of the "cow" random effect.

So, how do you simulate to test methods for animal breeding?