If you believe in things
that you don't understand
then you suffer

Stevie Wonder, "Superstition"

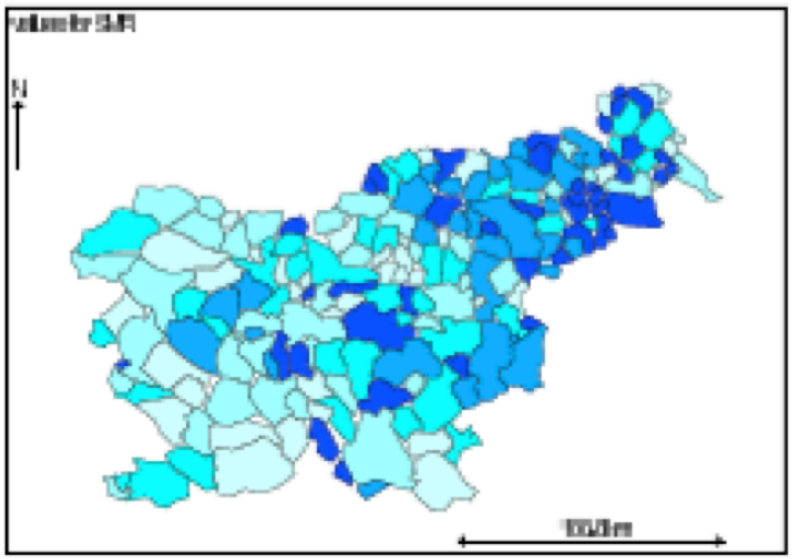
For Slovenian municipality i, i = 1, . . . , 192, we have

SIR_i = (observed stomach cancers 1995-2001 / expected)$_i$ = O_i/E_i

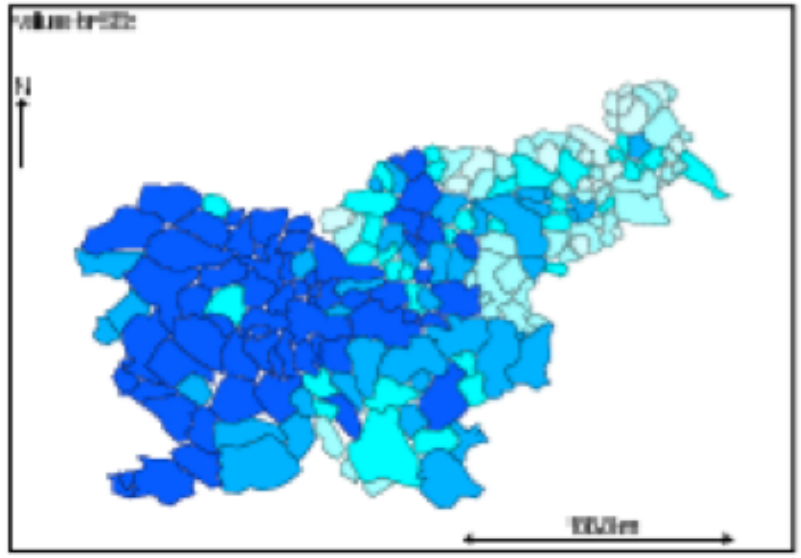
SEc_i = Socioeconomic score (centered and scaled)$_i$

Dark is high. There's a clear negative association.

SIR



SEc



First do a non-spatial analysis: $O_i \sim \text{Poisson}(\mu_i)$, where

$\log \mu_i = logE_i + \alpha + \beta SEc_i$

No surprise: $\beta | O_i \sim$ median -0.14, 95% interval -0.17 to -0.10.

Now do a spatial analysis:

$\log \mu_i = logE_i + \alpha + \beta SEc_i + S_i + H_i$

$\mathbf{S} = (S_1, \ldots, S_{194})' \sim$ improper CAR, precision parameter τ_s,
$\mathbf{H} = (H_1, \ldots, H_{194})' \sim$ iid Normal mean 0, precision τ_s.

SURPRISE!! $\beta | O_i \sim$ median -0.02, 95% interval -0.10 to +0.06.

DIC improved 1153.0 $\rightarrow$ 1081.5; pD: 2.0 $\rightarrow$ 62.3.

The obvious association has gone away. What happened?

Two premises underlying much of the course

(1) Writing down a model and using it to analyze a dataset = specifying a function from the data to inferential or predictive summaries.

- It is essential to understand that function from data to summaries.
- When I fit *this* model to *this* dataset, why do I get *this* result, and how much would the result change if I made *this* change to the data or model?

(2) We must distinguish between the model *we choose* to analyze a given dataset and the process that *we imagine* produced the data.

In particular, *our* choice to use a model with random effects does not imply that those random effects correspond to any random mechanism out there in the world.

Mixed linear models in the standard form

Mixed linear models are commonly written in the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \text{ where}$$

- The observation $\mathbf{y}$ is an n-vector;
- $\mathbf{X}$, the fixed-effects design matrix, is $n \times p$ and known;
- $\boldsymbol{\beta}$, the fixed effects, is $p \times 1$ and unknown;
- $\mathbf{Z}$, the random-effects design matrix, is $n \times q$ and known;
- $\mathbf{u}$, the random effects, is $q \times 1$ and unknown;
- $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G}(\boldsymbol{\phi}_G))$, where $\boldsymbol{\phi}_G$ is unknown;
- $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{R}(\boldsymbol{\phi}_R))$, where $\boldsymbol{\phi}_R$ is unknown; and
- The unknowns in $\mathbf{G}$ and $\mathbf{R}$ are $\boldsymbol{\phi} = (\boldsymbol{\phi}_G, \boldsymbol{\phi}_R)$.

The novelty here is the random effects

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \text{ where}$$

$$\mathbf{u} \sim N(\mathbf{0}, \mathbf{G}(\phi_G)) \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{R}(\phi_R))$$

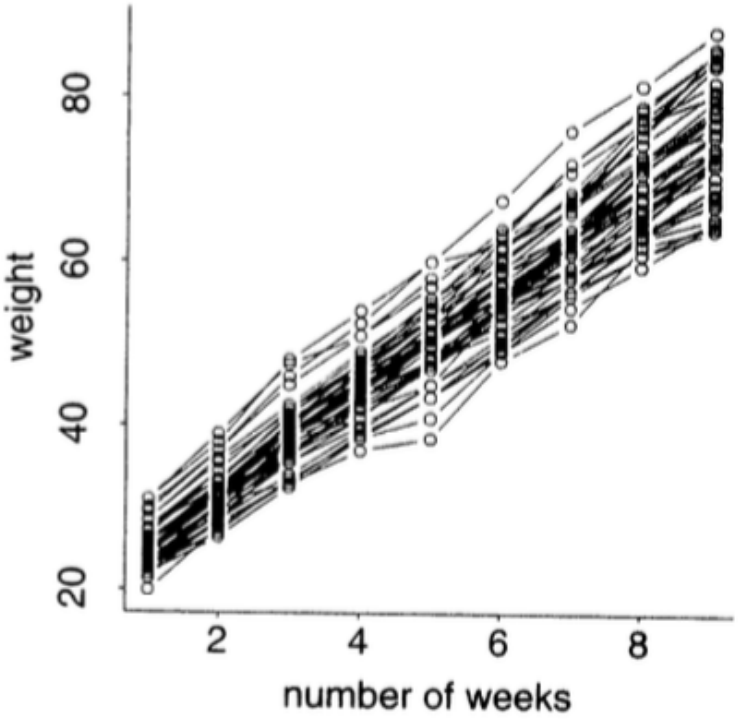
Original meaning (old-style random effects): A random effect's levels are *draws from a population*, and *the draws are not of interest in themselves* but only as samples from the larger population, which *is* of interest.

The random effects $\mathbf{Zu}$ are a way to model sources of variation that affect several observations the same way, e.g., all observations in a cluster.

Current meaning (new-style random effects): Also, a random effect $\mathbf{Zu}$ is a kind of model that is flexible because it has many parameters but that avoids overfitting because $\mathbf{u}$ is constrained by means of its covariance $\mathbf{G}$.

Example: Pig weights (RWC Sec. 4.2), Longitudinal data

48 pigs (i = 1, . . . , 48), weighed 1x/week for 9 weeks (j = 1, . . . , 9)



Dumb model: $\text{weight}_{ij} = \beta_0 + \beta_1 \text{ week}_j + \epsilon_{ij}, \quad \epsilon_{ij}$ iid $N(0, \sigma^2)$

This model assigns all variation *between* pigs to ϵ_{ij}, when pigs evidently vary in both intercept and slope.

Pig weights: A less dumb model

$$\begin{aligned} \text{weight}_{ij} &= \beta_{0i} + \beta_1 \text{ week}_j + \epsilon_{ij}, \quad \epsilon_{ij} \text{ iid } N(0, \sigma_e^2) \\ &= \beta_0 + u_i + \beta_1 \text{ week}_j + \epsilon_{ij} \end{aligned}$$

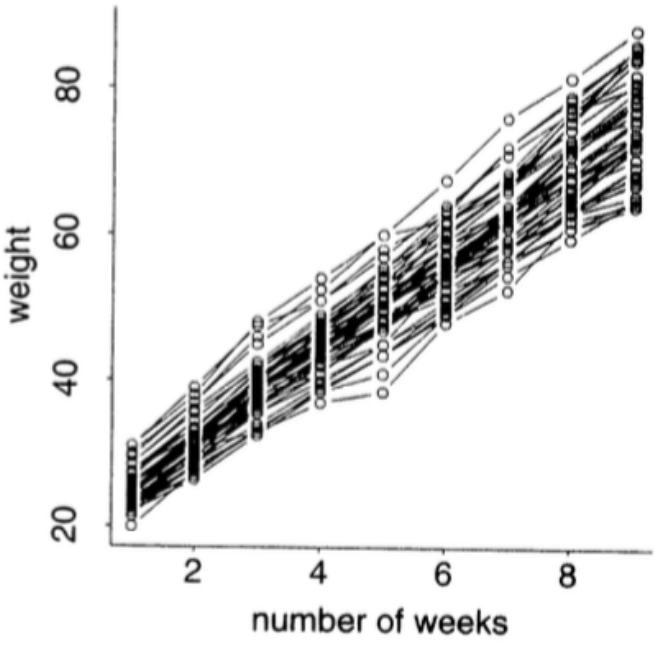
where β_0 and β_1 are fixed and u_i iid $N(0, \sigma_u^2)$

This model

- shifts pig i's line up or down as $u_i > 0$ or $u_i < 0$.
- partitions variation into
  - variation between pigs, u_i, and
  - variation within pigs, ϵ_{ij}.

Note that $\text{cov}(\text{weight}_{ij}, \text{weight}_{ij'}) = \text{cov}(u_i, u_i) = \sigma_u^2$.

Pig weights: Possibly not dumb model



$$\begin{aligned} \text{weight}_{ij} &= \beta_0 + u_{i0} + (\beta_1 + u_{i1})\ \text{week}_j + \epsilon_{ij}, \\ \begin{pmatrix} u_{i0} \\ u_{i1} \end{pmatrix} &\sim N\left(\mathbf{0}, \begin{bmatrix} \sigma^2_{00} & \sigma^2_{01} \\ \sigma^2_{01} & \sigma^2_{11} \end{bmatrix}\right) \end{aligned}$$

This is the so-called "random regressions" model.

Random regressions model in the standard form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \boldsymbol{\epsilon}, \text{ where}$$

$$\mathbf{X} = \left[\begin{array}{cc} 1 & week_1 \\ \vdots & \vdots \\ 1 & week_9 \\ \hline \vdots & \vdots \\ \hline 1 & week_1 \\ \vdots & \vdots \\ 1 & week_9 \end{array}\right], \quad \mathbf{Z} = \left[\begin{array}{cc|cc|c|cc} 1 & week_1 & & & & & \\ \vdots & \vdots & \multicolumn{2}{c|}{\mathbf{0}_{9\times 2}} & \cdots & \multicolumn{2}{c}{\mathbf{0}_{9\times 2}} \\ 1 & week_9 & & & & & \\ \hline & & 1 & week_1 & & & \\ \multicolumn{2}{c|}{\mathbf{0}_{9\times 2}} & \vdots & \vdots & \cdots & \multicolumn{2}{c}{\mathbf{0}_{9\times 2}} \\ & & 1 & week_9 & & & \\ \hline \vdots & & \vdots & & \ddots & \multicolumn{2}{c}{\vdots} \\ \hline & & & & & 1 & week_1 \\ \multicolumn{2}{c|}{\mathbf{0}_{9\times 2}} & \multicolumn{2}{c|}{\mathbf{0}_{9\times 2}} & \cdots & \vdots & \vdots \\ & & & & & 1 & week_9 \end{array}\right],$$

Random regressions model in the standard form (2)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \text{ where}$$

$$\mathbf{y} = \begin{bmatrix} weight_{1,1} \\ \vdots \\ weight_{1,9} \\ \hline weight_{2,1} \\ \vdots \\ weight_{2,9} \\ \hline \vdots \\ weight_{48,9} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} u_{10} \\ u_{11} \\ \hline u_{20} \\ u_{21} \\ \hline \vdots \\ \hline u_{48,0} \\ u_{48,1} \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_{1,1} \\ \vdots \\ \epsilon_{1,9} \\ \hline \epsilon_{2,1} \\ \vdots \\ \epsilon_{2,9} \\ \hline \vdots \\ \epsilon_{48,9} \end{bmatrix}$$

Conventional fits of this model commonly give an estimate of ± 1 for the correlation between the random intercept and slope. Nobody knows why.

Example: Molecular structure of a virus

Peterson et al. (2001) hypothesized a molecular description of the outer shell (prohead) of the bacteriophage virus ϕ29.

Goal: Break the prohead or phage into constituent molecules and weigh them; the weights (and other information) test the hypotheses about the components of the prohead (gp7, gp8, gp8.5, etc.)

There were four steps in measuring each component:

- Select a *parent*: 2 prohead parents, 2 phage parents.
- Prepare a *batch* of the parent.
- On a *gel date*, create electrophoresis gels, separate the molecules on the gels, cut out the relevant piece of the gel.
- Burn gels in an *oxidizer run*; each gel gives a weight for each molecule.

Here's a sample of the design for the gp8 molecule.

Parent	Batch	Gel Date	Oxidizer Run	gp8 Weight
1	1	1	1	244
1	1	2	1	267
1	1	2	1	259
1	1	2	1	286
1	3	1	1	218
1	3	2	1	249
1	3	2	1	266
1	3	2	1	259
1	7	4	3	293
⋮	⋮	⋮	⋮	⋮
1	7	4	3	297
1	7	5	4	315
⋮	⋮	⋮	⋮	⋮
1	7	5	4	283
1	7	7	4	311
⋮	⋮	⋮	⋮	⋮
1	7	7	4	334
2	2	1	1	272
2	2	2	1	223

To analyze these data, I treated each of the four steps as an old-style random effect.

Model: For y_i the i^{th} measurement of the number of gp8 molecules

$$
\begin{aligned}
y_i &= \mu + \text{parent}_{j(i)} + \text{batch}_{k(i)} + \text{gel}_{l(i)} + \text{run}_{m(i)} + \epsilon_i \\
\text{parent}_{j(i)} &\overset{iid}{\sim} N(0, \sigma_p^2), \quad j = 1, \ldots, 4 \\
\text{batch}_{k(i)} &\overset{iid}{\sim} N(0, \sigma_b^2), \quad k = 1, \ldots, 9 \\
&\quad \text{batches are nested within parents} \\
\text{gel}_{l(i)} &\overset{iid}{\sim} N(0, \sigma_g^2), \quad l = 1, \ldots, 11 \\
\text{run}_{m(i)} &\overset{iid}{\sim} N(0, \sigma_r^2), \quad m = 1, \ldots, 7 \\
\epsilon_i &\overset{iid}{\sim} N(0, \sigma_e^2), \quad i = 1, \ldots, 98
\end{aligned}
$$

Deterministic functions $j(i)$, $k(i)$, $l(i)$, and $m(i)$ map i to parent, batch, gel, and run indices.

Molecular-structure model in the standard form

$$\ln \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \qquad \mathbf{X} = \mathbf{1}_{98}, \qquad \boldsymbol{\beta} = \mu,$$

$$\mathbf{Z}_{98\times 31} = \begin{array}{cccc} \overbrace{1000}^{\text{parent (4 cols)}} & \overbrace{10\ldots0}^{\text{batch (9 cols)}} & \overbrace{10\ldots0}^{\text{gel (11 cols)}} & \overbrace{10\ldots0}^{\text{run (7 cols)}} \\ 1000 & 10\ldots0 & 01\ldots0 & 10\ldots0 \\ \vdots & \vdots & \vdots & \vdots \\ 0001 & 00\ldots1 & 00\ldots1 & 00\ldots1 \end{array},$$

$$\mathbf{u}_{31\times 1} = [\text{parent}_1, \ldots, \text{parent}_4, \text{batch}_1, \ldots, \text{batch}_9, \text{gel}_1, \ldots, \text{gel}_{11}, \text{run}_1, \ldots, \text{run}_7]',$$

$$\mathbf{G} = \begin{bmatrix} \sigma_p^2\mathbf{I}_4 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_b^2\mathbf{I}_4 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_g^2\mathbf{I}_4 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \sigma_o^2\mathbf{I}_4 \end{bmatrix}$$

Example: Rating vocal fold images

ENT docs evaluate speech/larynx problems by taking videos of the inside of the larynx during speech and having trained raters assess them.

Standard method (2008): strobe lighting with period slightly longer than the vocal folds' period.

Kendall (2009) tested a new method using high-speed video (HSV), giving a direct view of the folds' vibration.

Each of 50 subjects was measured using both *image forms* (strobe/HSV).

Each of the 100 images was assessed by ≥ 1 raters, CR, KK, KUC.

Each rating consisted of 5 quantities on continuous scales.

Interest: compare image forms, compare raters, measure their interaction.

Subject ID	Imaging Method	Rater	% Open Phase
1	strobe	CR	56
1	strobe	KK	70
1	strobe	KK	70
1	HSV	KUC	70
1	HSV	KK	70
1	HSV	KK	60
2	strobe	KUC	50
2	HSV	CR	54
3	strobe	KUC	60
3	strobe	KUC	70
3	HSV	KK	56
4	strobe	CR	65
4	HSV	KK	56
5	strobe	KK	50
5	HSV	KUC	55
5	HSV	KUC	67
6	strobe	KUC	50
6	strobe	KUC	50
6	HSV	KUC	50

Vocal folds example: the design

This is a repeated-measures design

- The subject effect is "study number".
- There are 3 fixed effect factors:
  - image form, varying entirely within subject;
  - rater, varying both within and between subjects;
  - image form-by-rater interaction.

The random effects are:

- study number;
- study number by image form;
- study number by rater;
- residual (3-way interaction)

First rows of the FE design matrix $\mathbf{X}$

Intercept	Image form	Rater		Interaction	
1	0	1	0	0	0
1	0	0	0	0	0
1	0	0	0	0	0
1	1	0	1	0	1
1	1	0	0	0	0
1	1	0	0	0	0
1	0	0	1	0	0
1	1	1	0	1	0
1	0	0	1	0	0
1	0	0	1	0	0
1	1	0	0	0	0
1	0	1	0	0	0
1	1	0	0	0	0
1	0	0	0	0	0
1	1	0	1	0	1
1	1	0	1	0	1
1	0	0	1	0	0
			⋮		

Random effects: Construct $\mathbf{u}$ first, then $\mathbf{Z}$

Study number: $u_1, \ldots, u_{50}$, one per subject.

Study number by image form: $u_{1H}, u_{1S}, \ldots, u_{50H}, u_{50S}$, one per subject/form.

Study number by rater: Complicated! One per unique combination of a rater and a study number:

$$
\begin{array}{lll}
u_{10,CR}, & u_{10,KK}, & u_{10,KUC} \\
u_{60,CR}, & & u_{60,KUC} \\
 & u_{66,KK}, & u_{66,KUC} \\
u_{79,CR}, & u_{79,KK}, & \\
 & \vdots & \\
u_{931,CR}, & & \\
u_{967,CR} & &
\end{array}
$$

Non-zero columns of the RE design matrix **Z**

Sub ID	Form	Rater	Sub	Sub-by-Method	Sub-by-Rater
1	strobe	CR	100	100000	1000000
1	strobe	KK	100	100000	0100000
1	strobe	KK	100	100000	0100000
1	HSV	KUC	100	010000	0010000
1	HSV	KK	100	010000	0100000
1	HSV	KK	100	010000	0100000
2	strobe	KUC	010	001000	0001000
2	HSV	CR	010	000100	0000100
3	strobe	KUC	001	000010	0000010
3	strobe	KUC	001	000010	0000010
3	HSV	KK	001	000001	0000001
		⋮		⋮	

JMP's RL maximizing algorithm converged for 4 of 5 outcomes. The designs are identical, so the difference arises from **y**. Nobody knows how.

A key point of these examples — and of this course — is that we now have **tremendous ability to fit models** in this form but **little understanding of how the data determine the fits**.

This course and book are my stab at developing that understanding.

The next section's purpose is to emphasize three points:

- The theory and methods of mixed linear models are strongly connected to the theory and methods of linear models, though the differences are important;
- The restricted likelihood is the posterior distribution from a particular Bayesian analysis; and
- Conventional and Bayesian analyses are incomplete or problematic in many respects.

Doing statistics with MLMs: Conventional analysis

The conventional analysis usually proceeds in three steps:

- Estimate $\boldsymbol{\phi} = (\boldsymbol{\phi}_G, \boldsymbol{\phi}_R)$, the unknowns in **G** and **R**.
- Treating $\hat{\boldsymbol{\phi}}$ as if it's true, estimate $\boldsymbol{\beta}$ and **u**.
- Treating $\hat{\boldsymbol{\phi}}$ as if it's true, compute tests and confidence intervals for $\boldsymbol{\beta}$ and **u**.

We will start with estimating $\boldsymbol{\beta}$ and **u** (mean structure), move to estimating $\boldsymbol{\phi}$ (variance structure), and then on to tests, etc.

The obvious problem — treating $\hat{\boldsymbol{\phi}}$ as if it's true — is well known.

Mean structure estimates (β and $\mathbf{u}$)

This area has a long history and a lot of jargon has accrued:

- fixed effects are *estimated*;
- random effects are *predicted*;
- fixed + random effects are (estimated) best linear unbiased predictors, (E)BLUPs

I avoid this as much as I can and instead use the more recent ...

Unified approach, based on likelihood ideas

- Assume normal errors and random effects (the same approach is used for non-normal models).
- For estimates of β and $\mathbf{u}$, use likelihood maximizing values *given* the variance components.

We have: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \quad \mathbf{u} \sim N_q(0, \mathbf{G}(\boldsymbol{\phi}_G)), \quad \boldsymbol{\epsilon} \sim N_n(0, \mathbf{R}(\boldsymbol{\phi}_R))$

In the conventional view, we can write the joint density of $\mathbf{y}$ and $\mathbf{u}$ as

$$f(\mathbf{y}, \mathbf{u}|\boldsymbol{\beta}, \boldsymbol{\phi}) = f(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \boldsymbol{\phi}_R) f(\mathbf{u}|\boldsymbol{\phi}_G),$$

Taking the log of both sides gives $\log f(\mathbf{y}, \mathbf{u}|\boldsymbol{\beta}, \boldsymbol{\phi}) =$

$$\begin{aligned} K \quad &- \quad \frac{1}{2}\log|\mathbf{R}(\boldsymbol{\phi}_R)| - \frac{1}{2}\log|\mathbf{G}(\boldsymbol{\phi}_G)| \\ &- \quad \frac{1}{2}\left\{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'\mathbf{R}(\boldsymbol{\phi}_R)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \mathbf{u}'\mathbf{G}^{-1}(\boldsymbol{\phi}_G)\mathbf{u}\right\}. \end{aligned}$$

Treat $\mathbf{G}$ and $\mathbf{R}$ as known, set $\mathbf{C} = [\mathbf{X}|\mathbf{Z}]$, estimate $\boldsymbol{\beta}$ and $\mathbf{u}$ by minimizing

$$\underbrace{\left[\mathbf{y} - \mathbf{C}\begin{pmatrix}\boldsymbol{\beta}\\ \mathbf{u}\end{pmatrix}\right]' \mathbf{R}^{-1}\left[\mathbf{y} - \mathbf{C}\begin{pmatrix}\boldsymbol{\beta}\\ \mathbf{u}\end{pmatrix}\right]}_{\text{“likelihood”}} + \underbrace{[\boldsymbol{\beta}|\mathbf{u}]\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{-1}\end{pmatrix}\begin{bmatrix}\boldsymbol{\beta}\\ \mathbf{u}\end{bmatrix}}_{\text{“penalty”}}$$

It's easy to show that for $\mathbf{C} = [\mathbf{X}|\mathbf{Z}]$, this gives point estimates:

$$\begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix}_{\phi} = \left[\mathbf{C}'\mathbf{R}^{-1}\mathbf{C} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{pmatrix}\right]^{-1} \mathbf{C}'\mathbf{R}^{-1}\mathbf{y}$$

The tildes and subscript mean these estimates depend on ϕ.

- Omit the penalty term and this is just the GLS estimate.
- The penalty term is an extra piece of information about $\mathbf{u}$; this is how it affects the estimates of both $\mathbf{u}$ and $\boldsymbol{\beta}$.

These estimates give fitted values:

$$\tilde{\mathbf{y}}_{\phi} = \mathbf{C}\begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix}_{\phi} = \mathbf{C}\left[\mathbf{C}'\mathbf{R}^{-1}\mathbf{C} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{pmatrix}\right]^{-1} \mathbf{C}'\mathbf{R}^{-1}\mathbf{y}.$$

Insert estimates for $\mathbf{G}$ and $\mathbf{R}$ to give EBLUPs.

Estimates of ϕ_G and ϕ_R, the unknowns in $\mathbf{G}$ and $\mathbf{R}$

Obvious [?] approach: Write down the likelihood and compute MLEs.

But this has problems.

(1) [Dense, obscure] It's not clear exactly what the likelihood is.

(2) [Pragmatic] Avoid problem (1) by getting rid of $\mathbf{u}$, writing

$$\begin{aligned} \mathrm{E}(\mathbf{y}|\boldsymbol{\beta}) &= \mathbf{X}\boldsymbol{\beta} \\ \mathrm{cov}(\mathbf{y}|\mathbf{G},\mathbf{R}) &= \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R} \equiv \mathbf{V} \\ \text{So } \mathbf{y} &\sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}), \text{ for } \mathbf{V} \equiv \mathbf{V}(\mathbf{G},\mathbf{R}). \end{aligned}$$

But maximizing this likelihood gives biased estimates.

Example: If $X_1, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$, the ML estimator $\hat{\sigma}^2_{ML} = \frac{1}{n}\sum(X_i - \bar{X})^2$ is biased.

MLEs are biased because they don't account for estimating fixed effects.

Solution: The restricted (or residual) likelihood

Let $\mathbf{Q}_X = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ be the projection onto $\mathcal{R}(\mathbf{X})^{\perp}$, the orthogonal complement of the column space of the fixed-effect design matrix $\mathbf{X}$.

First try: The residuals $\mathbf{Q}_X\mathbf{y} = \mathbf{Q}_X\mathbf{Z}\mathbf{u} + \mathbf{Q}_X\boldsymbol{\epsilon}$ are n-variate normal with mean $\mathbf{0}$ and covariance $\mathbf{Q}_X\mathbf{Z}\mathbf{G}\mathbf{Z}'\mathbf{Q}_X + \mathbf{Q}_X\mathbf{R}\mathbf{Q}_X$.

The resulting likelihood is messy: $\mathbf{Q}_X\mathbf{y}$ has a singular covariance matrix.

Instead, do the same thing but with a non-singular covariance matrix.

Solution: The restricted (or residual) likelihood

Second try: $\mathbf{Q}_X$ has spectral decomposition $\mathbf{K}_0\mathbf{D}\mathbf{K}_0'$,

where $\mathbf{D} = \text{diag}(1, \ldots, 1, \mathbf{0}_p')$, for $p = \text{rank}(\mathbf{X})$;

$\mathbf{K}_0$ is an orthogonal matrix with first $n - p$ columns $\mathbf{K}$.

Then $\mathbf{Q}_X = \mathbf{K}\mathbf{K}'$; $\mathbf{K}$ is $n \times (n - p)$, its columns are a basis for $\mathcal{R}(\mathbf{X})^{\perp}$.

Project $\mathbf{y}$ onto the column space of $\mathbf{K}$ = the residual space of $\mathbf{X}$:

$\mathbf{K}'\mathbf{y} = \mathbf{K}'\mathbf{Z}\mathbf{u} + \mathbf{K}'\boldsymbol{\epsilon}$ is $(n - p)$-variate normal with mean 0 and covariance $\mathbf{K}'\mathbf{Z}\mathbf{G}\mathbf{Z}'\mathbf{K} + \mathbf{K}'\mathbf{R}\mathbf{K}$, which is non-singular.

The likelihood arising from $\mathbf{K}'\mathbf{y}$ is the restricted (residual) likelihood.

Pre-multiplying the previous equation by any non-singular $\mathbf{B}$ with $|\mathbf{B}| = 1$ gives the same restricted likelihood.

The RL also arises as a Bayesian marginal posterior

The previous construction has intuitive content: Attribute to $\boldsymbol{\beta}$ the part of $\mathbf{y}$ that lies in $\mathcal{R}(\mathbf{X})$ and use only the residual to estimate $\mathbf{G}$ and $\mathbf{R}$.

Unfortunately, when derived this way, the RL is obscure.

It can be shown that you get the same RL as the Bayesian marginal posterior of $\boldsymbol{\phi}$ with flat priors on everything. I'll derive that posterior.

The RL as a marginal posterior

Use the likelihood from a few slides ago

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}), \text{ for } \mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$$

with prior $\pi(\boldsymbol{\beta}, \phi_G, \phi_R) \propto 1$. (DON'T USE THIS in a Bayesian analysis.)

Then the posterior is

$$\pi(\boldsymbol{\beta}, \phi_G, \phi_R) \propto |\mathbf{V}|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right).$$

To integrate out $\boldsymbol{\beta}$: expand the quadratic form and complete the square:

$$\begin{aligned} & (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ = \; & \mathbf{y}'\mathbf{V}^{-1}\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\ = \; & \mathbf{y}'\mathbf{V}^{-1}\mathbf{y} - \tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}} + (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}), \end{aligned}$$

for $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, the GLS estimate given $\mathbf{V}$.

Integrating out $\boldsymbol{\beta}$ is just the integral of a multivariate normal density, so

$$\int L(\boldsymbol{\beta}, \mathbf{V}) d\boldsymbol{\beta} = K|\mathbf{V}|^{-\frac{1}{2}}|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left[\mathbf{y}'\mathbf{V}^{-1}\mathbf{y} - \tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}}\right]\right)$$

$$\text{for } \tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

Take the log and expand $\tilde{\boldsymbol{\beta}}$ to get the log restricted likelihood

$$\begin{aligned} \log \text{RL}(\boldsymbol{\phi}|\mathbf{y}) = K &- 0.5\left(\log|\mathbf{V}| + \log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|\right) \\ &- 0.5\ \mathbf{y}'\left[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\right]\mathbf{y}, \end{aligned}$$

where $\mathbf{V} = \mathbf{Z}\mathbf{G}(\phi_G)\mathbf{Z}' + \mathbf{R}(\phi_R)$ is a function of $\boldsymbol{\phi} = (\phi_G, \phi_R)$.

This is more explicit than the earlier expression, though not closed-form.

Standard errors for $\hat{\boldsymbol{\beta}}$ in standard software

For the model $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\phi}))$, $\mathbf{V}(\boldsymbol{\phi}) = \mathbf{Z}\mathbf{G}(\boldsymbol{\phi}_G)\mathbf{Z}' + \mathbf{R}(\boldsymbol{\phi}_R)$:

Set $\boldsymbol{\phi} = \hat{\boldsymbol{\phi}}$, giving $\hat{\mathbf{V}}$.

$\hat{\boldsymbol{\beta}}$ is the GLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$,

which has the familiar covariance form $\text{cov}(\hat{\boldsymbol{\beta}}) \approx (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$,

giving $SE(\hat{\beta}_i) \approx \text{cov}(\hat{\boldsymbol{\beta}})_{ii}^{0.5}$.

Non-Bayesian alternative: Kenward-Roger approximation (in SAS):

$$\begin{aligned}
\text{cov}(\hat{\boldsymbol{\beta}}) &\approx (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1} \\
&+ \text{adjust for bias of } (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1} \text{as an estimate of } (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \\
&+ \text{adjust for bias of } (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \text{ (Kackar-Harville)}
\end{aligned}$$

Two covariance matrices for $(\boldsymbol{\beta}, \mathbf{u})'$

Unconditional covariance:

$$\text{cov}\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} - \mathbf{u} \end{pmatrix} = \left[\mathbf{C}'\mathbf{R}(\hat{\phi}_R)^{-1}\mathbf{C} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}(\hat{\phi}_G)^{-1} \end{pmatrix}\right]^{-1},$$

This is wrt the distributions of $\boldsymbol{\epsilon}$ and $\mathbf{u}$ and gives the same SEs as above.

Conditional on $\mathbf{u}$.

$$\begin{aligned}\text{cov}\left(\begin{array}{c|c} \hat{\boldsymbol{\beta}} & \\ \hat{\mathbf{u}} & \mathbf{u} \end{array}\right) &= \left[\mathbf{C}'\mathbf{R}(\hat{\phi}_R)^{-1}\mathbf{C} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}(\hat{\phi}_G)^{-1} \end{pmatrix}\right]^{-1} \mathbf{C}'\mathbf{R}^{-1}\mathbf{C} \\ &\quad \times \left[\mathbf{C}'\mathbf{R}(\hat{\phi}_R)^{-1}\mathbf{C} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}(\hat{\phi}_G)^{-1} \end{pmatrix}\right]^{-1}.\end{aligned}$$

This is wrt the distribution of $\boldsymbol{\epsilon}$. Later we'll see how this can be useful.

Tests for fixed effects in standard software

Tests and intervals for linear combinations $l'\beta$ use

$$z = \frac{l'\hat{\beta}}{(l'\hat{\text{cov}}(\hat{\beta})l)^{0.5}} \approx N(0,1). \tag{1}$$

RWC (p. 104):

> *[T]heoretical justification of [(1)] for general mixed models is somewhat elusive owing to the dependence in* **y** *imposed by the random effects. Theoretical back-up for [(1)] exists in certain special cases, such as those arising in analysis of variance and longitudinal data analysis. . . . For some mixed models, including many used in . . . this book, justification of [(1)] remains an open problem.*

Use them at your own risk. The LRT has the same problem.

RWC suggest a parametric bootstrap instead (e.g., p. 144).

Tests and intervals for random effects

I'll say little about these because (a) they have only asymptotic rationales, which are incomplete, and (b) mostly they perform badly.

- Restricted LRT: Compare **G**, **R** values for the same **X**.
- Test for $\sigma_s^2 = 0$: Asymptotic distribution of LRT is a mixture of χ^2s.
- The XICs: AIC, BIC (aka Schwarz criterion), etc.
- Satterthwaite approximate confidence interval for a variance (SAS):

$$\frac{\nu\hat{\sigma}_s^2}{\chi^2_{\nu,1-\alpha/2}} \leq \sigma_s^2 \leq \frac{\nu\hat{\sigma}_s^2}{\chi^2_{\nu,\alpha/2}}$$

The denominators are quantiles of χ^2_ν for $\nu = 2\left[\hat{\sigma}_s^2/(\text{Asymp SE } \hat{\sigma}_s^2)\right]^2$.