Statistical Modelling

Antony Overstall
University of Southampton

©2018

(Chapters 1–2 closely based on original notes by
Anthony Davison, Jon Forster & Dave Woods)

2. Beyond the Generalised Linear Model

Overview

1.   Generalised linear models

2.   Overdispersion

3.   Correlation

4.   Random effects models

Generalised Linear Models

GLM recap

$y_1, \ldots, y_n$ are observations of response variables $Y_1, \ldots, Y_n$ assumed to be independently generated by a distribution of the same exponential family form, with means $\mu_i \equiv \mathrm{E}(Y_i)$ linked to explanatory variables $X_1, X_2, \ldots, X_p$ through

$$g(\mu_i) = \eta_i \equiv \beta_0 + \sum_{r=1}^{p} \beta_r x_{ir} \equiv x_i^{\mathrm{T}} \beta$$

GLMs have proved remarkably effective at modelling real world variation in a wide range of application areas.

GLM failure

However, situations frequently arise where GLMs do not adequately describe observed data.

This can be due to a number of reasons including:

☐ The mean model cannot be appropriately specified as there is dependence on an unobserved (or unobservable) explanatory variable.

☐ There is excess variability between experimental units beyond that implied by the mean/variance relationship of the chosen response distribution. e.g. under a Poisson $\text{var}(Y_i) = E(Y_i) = \mu_i$

☐ The assumption of independence is not appropriate.

☐ Complex multivariate structure in the data requires a more flexible model class

Overdispersion

Example 1: toxoplasmosis

The table below gives data on the relationship between rainfall (x) and the proportions of people with toxoplasmosis (y/m) for 34 cities in El Salvador.

City	y	x	City	y	x	City	y	x
1	5/18	1620	12	3/5	1800	23	3/10	1973
2	15/30	1650	13	8/10	1800	24	1/6	1976
3	0/1	1650	14	0/1	1830	25	1/5	2000
4	2/4	1735	15	53/75	1834	26	0/1	2000
5	2/2	1750	16	7/16	1871	27	7/24	2050
6	2/8	1750	17	24/51	1890	28	46/82	2063
7	2/12	1756	18	3/10	1900	29	7/19	2077
8	6/11	1770	19	23/43	1918	30	9/13	2100
9	33/54	1770	20	3/6	1920	31	4/22	2200
10	8/13	1780	21	0/1	1920	32	4/9	2240
11	41/77	1796	22	3/10	1936	33	8/11	2250
						34	23/37	2292
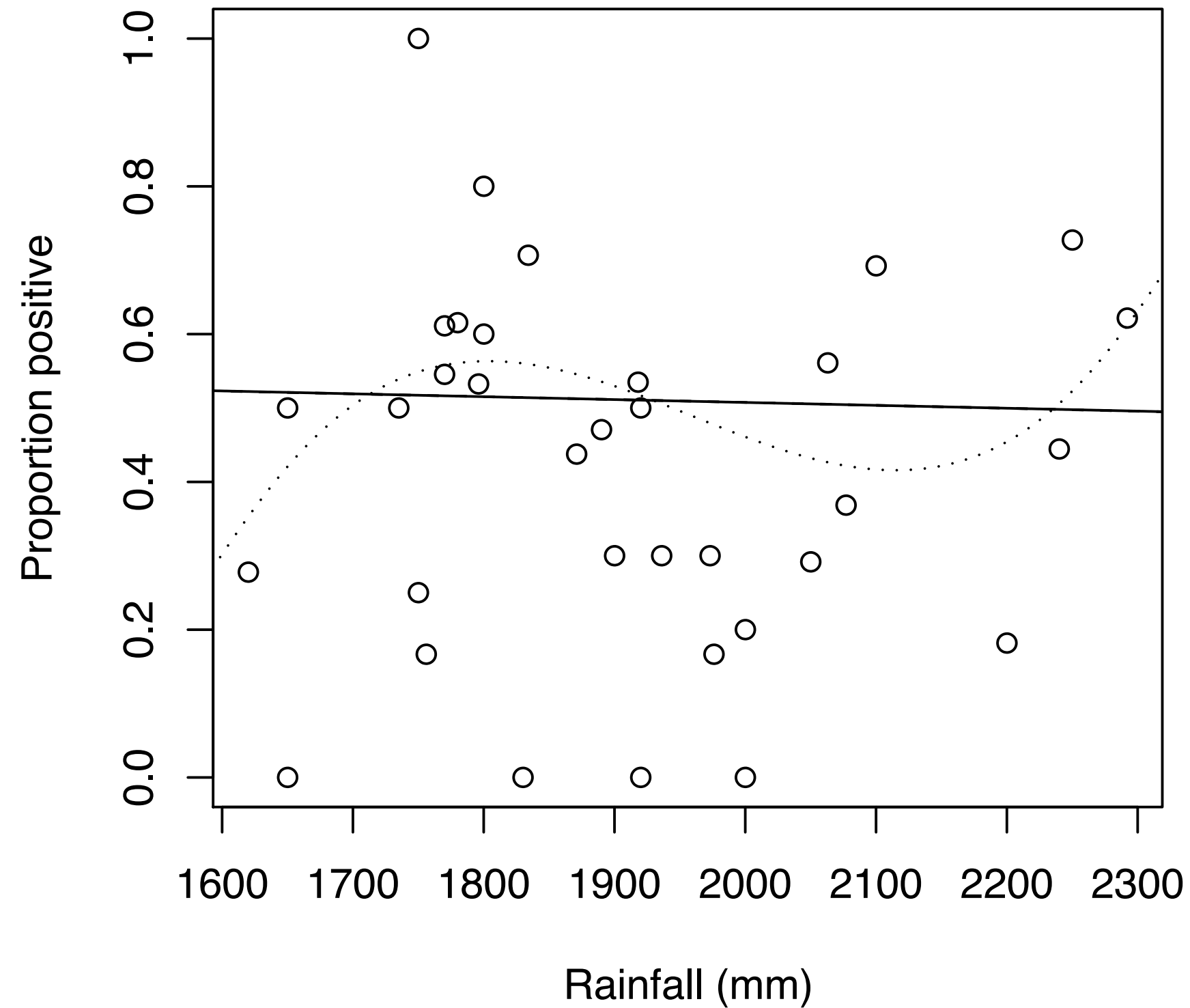
Example

Toxoplasmosis data and fitted models

Example

$$Y_i = \frac{Z_i}{m_i} = \frac{\# \text{ people with toxo}}{\# \text{ people tested}}$$

$$m_i Y_i = Z_i \sim \text{binomial}(m_i, \mu_i)$$
$$\text{logit}(\mu_i) = x_i^T \beta$$

Fitting various binomial logistic regression models relating toxoplasmosis incidence to rainfall:

Comparison of Cubic to Constant

$$\Delta D = 74.21 - 62.62$$
$$= 11.59 \sim X^2_3$$
$$p\text{-value} = 0.009$$

Model	df	deviance	AIC
Constant	33	74.21	166.90
Linear	32	74.09	168.78
Quadratic	31	74.09	170.78
Cubic	30	62.62	161.33

So evidence in favour of the cubic over other models, but a poor fit ($X^2 = 58.21$ on 30df).

This is an example of **overdispersion** where residual variability is greater than would be predicted by the specified mean/variance relationship

$$\text{var}(Y) = \frac{\mu(1-\mu)}{m}.$$

Example

QQ-plot of residuals from cubic model
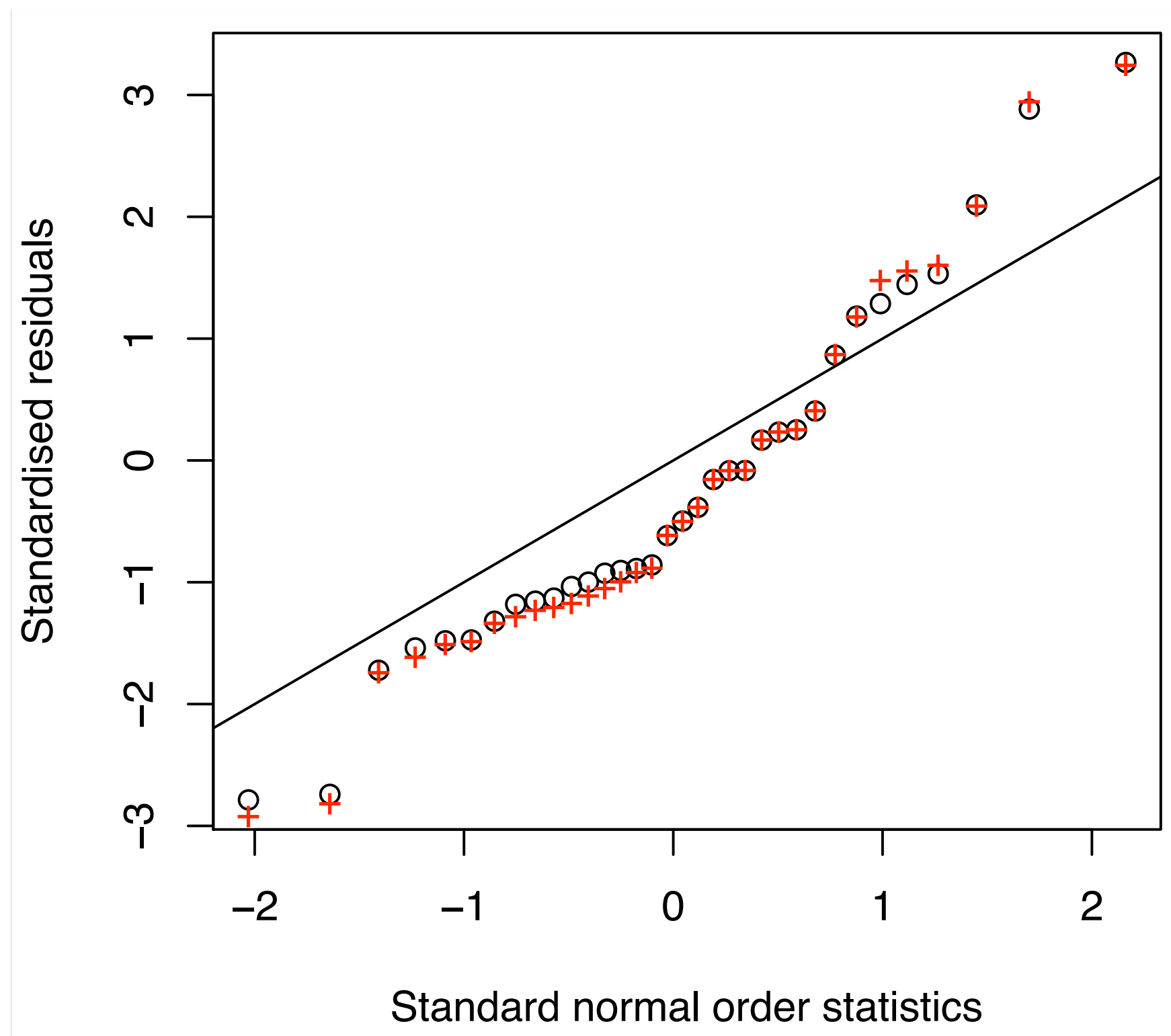
O = Pearson

+ = deviance

Toxoplasmosis residual plot

Quasi-likelihood

A quasi-likelihood approach to accounting for overdispersion models the mean and variance, but stops short of a full probability model for Y.

For a model specified by the mean relationship $g(\mu_i) = \eta_i = x_i^{\mathrm{T}}\beta$, and variance $\mathrm{var}(Y_i) = \sigma^2 V(\mu_i)/m_i$, the quasi-likelihood equations are

$$\sum_{i=1}^{n} x_i \frac{y_i - \mu_i}{\sigma^2 V(\mu_i) g'(\mu_i)/m_i} = 0$$

← Proportional to Score function from GLM.

If $V(\mu_i)/m_i$ represents $\mathrm{var}(Y_i)$ for a standard distribution from the exponential family, then these equations can be solved for β using standard GLM software.

Provided the mean and variance functions are correctly specified, asymptotic normality for $\widehat{\beta}$ still holds.
The dispersion parameter σ^2 can be estimated using

$$\widehat{\sigma}^2 \equiv \frac{1}{n-p-1} \sum_{i=1}^{n} \frac{m_i(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}$$

Quasi-likelihood for toxoplasmosis data

Assuming the same mean model as before, but $\mathrm{var}(Y_i) = \sigma^2 \mu_i (1 - \mu_i)/m_i$, we obtain $\widehat{\sigma}^2 = 1.94$ with $\widehat{\beta}$ (and corresponded fitted mean curves) as before.

Comparing cubic with constant model, one now obtains

$$F = \frac{(74.21 - 62.62)/3}{1.94} = 1.99$$

$$= Reg \frac{MS}{Res\ MS}$$

$$\sim F_{3,30}$$

which provides much less compelling evidence in favour of an effect of rainfall on toxoplasmosis incidence.

$$p\text{-value} = 0.14$$

Reasons

To construct a full probability model in the presence of overdispersion, it is necessary to consider **why** overdispersion might be present.

Possible reasons include:

☐ There may be an important explanatory variable, other than rainfall, which we haven't observed.

☐ Or there may be many other features of the cities, possibly unobservable, all having a small individual effect on incidence, but a larger effect in combination. Such effects may be individually undetectable – sometimes described as *natural excess variability between units*.

Reasons: unobserved heterogeneity

When part of the linear predictor is 'missing' from the model,

$$\eta_i^{\text{true}} = \eta_i^{\text{model}} + \eta_i^{\text{diff}}$$

"$x_i^\top \beta$

] All models are wrong

We can compensate for this, in modelling, by assuming that the missing $\eta_i^{\text{diff}} \sim F$ in the population. Hence, given η_i^{model}

$$\mu_i \equiv g^{-1}(\eta_i^{\text{model}} + \eta_i^{\text{diff}}) \sim G$$

$\eta_i = g(\mu_i)$

where G is the distribution induced by F. Then

$$\mathrm{E}(Y_i) = \mathrm{E}_G[\mathrm{E}(Y_i \mid \mu_i)] = \mathrm{E}_G(\mu_i)$$

$$\mathrm{var}(Y_i) = \mathrm{E}_G(V(\mu_i)/m_i) + \mathrm{var}_G(\mu_i)$$

$= E_G\left(\mathrm{var}\left(Y_i|\mu_i\right)\right) + \mathrm{var}_G\left(E\left(Y_i|\mu_i\right)\right)$

$\frac{V(\mu_i)}{m_i}$ μ_i

Direct models

One approach is to model the Y_i directly, by specifying an appropriate form for G.

For example, for the toxoplasmosis data, we might specify a **beta-binomial** model, where $m_i Y_i \sim \text{binomial}(m_i, \mu_i)$

$$\mu_i \sim \text{Beta}(k\mu_i^*, k[1 - \mu_i^*]) \quad \text{logit}(\mu_i^*) = x_i^\tau \beta$$

leading to

$$\text{E}(Y_i) = \mu_i^*, \qquad \text{var}(Y_i) = \frac{\mu_i^*(1 - \mu_i^*)}{m_i}\left(1 + \frac{m_i - 1}{k + 1}\right)$$

with $(m_i - 1)/(k + 1)$ representing an overdispersion factor.

k controls overdispersion

Estimate k

$$\mu_i \sim Beta\left(\overset{G}{\underset{\text{``}}{k\mu_i^*}}, \overset{\beta}{k(1-\mu_i^*)}\right)$$

where G denotes α and β respectively.

$$E_G(\mu_i) = \mu_i^*$$

$$var_G(\mu_i) = \frac{\mu_i^*(1-\mu_i^*)}{k+1}$$

$$E_G\left(\mu_i(1-\mu_i)\right) = \frac{k\mu_i^*(1-\mu_i^*)}{k+1}$$

- $E(Y_i) = E_G(\mu_i) = \mu_i^*$

- $var(Y_i) = E_G\left(\frac{V(\mu_i)}{m_i}\right) + var_G(\mu_i)$

$$= E_G\left(\frac{\mu_i(1-\mu_i)}{m_i}\right) + \frac{\mu_i^*(1-\mu_i^*)}{k+1}$$

$$= \frac{k\mu_i^*(1-\mu_i^*)}{m_i(k+1)} + \frac{\mu_i^*(1-\mu_i^*)}{k+1} = \frac{\mu_i^*(1-\mu_i^*)}{m_i}\left(1 + \frac{m_i-1}{k+1}\right)$$

Direct models: fitting

Models which explicitly account for overdispersion can, in principle, be fitted using your preferred approach, e.g. the beta-binomial model has likelihood

$$f(y \mid \mu^*, k) \propto \prod_{i=1}^{n} \frac{\Gamma(k\mu_i^* + m_i y_i)\Gamma\{k(1 - \mu_i^*) + m_i(1 - y_i)\}\Gamma(k)}{\Gamma(k\mu_i^*)\Gamma\{k(1 - \mu_i^*)\}\Gamma(k + m_i)}.$$

Similarly the corresponding model for count data specifies a gamma distribution for the Poisson mean, leading to a *negative binomial* marginal distribution for Y_i.

However, these models have limited flexibility and can be difficult to fit, so an alternative approach is usually preferred.

A random effects model for overdispersion

A more flexible, and extensible approach models the excess variability by including an extra term in the linear predictor

$$\eta_i = x_i^{\mathrm{T}} \beta + u_i \qquad (1)$$

where the u_i can be thought of as representing the 'extra' variability between units, and are called **random effects**.

The model is completed by specifying a distribution F for u_i in the population − almost always, we use

$$u_i \sim N(0, \sigma^2)$$

for some unknown σ^2.

We set $\mathrm{E}(u_i) = 0$, as an unknown mean for u_i would be unidentifiable in the presence of the intercept parameter β_0.

Random effects: likelihood

The parameters of this random effects model are usually considered to be (β, σ^2) and therefore the likelihood is given by

$$
\begin{aligned}
& \quad\quad\quad\quad\quad\quad\quad\quad = \int f(y, u \mid \beta, \sigma^2)\, du \\
f(y \mid \beta, \sigma^2) \;=&\; \int f(y \mid \beta, u, \sigma^2) f(u \mid \beta, \sigma^2)\,\mathrm{d}u \quad\begin{array}{l} y \mid \beta, u \text{ is indep} \\ \quad \text{of } \sigma^2 \\ u \mid \sigma^2 \text{ is indep} \\ \quad \text{of } \beta \end{array} \\
=&\; \int f(y \mid \beta, u) f(u \mid \sigma^2)\,\mathrm{d}u \\
&\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad \begin{array}{l} y_1, \dots, y_n \mid \beta, u \text{ are} \\ \quad \text{indep} \end{array} \\
=&\; \int \prod_{i=1}^{n} f(y_i \mid \beta, u_i) f(u_i \mid \sigma^2)\,\mathrm{d}u_i \quad \begin{array}{l} u_1, \dots u_n \mid \sigma^2 \\ \text{are indep} \end{array} \;(2)
\end{aligned}
$$

where $f(y_i \mid \beta, u_i)$ arises from our chosen exponential family, with linear predictor (1) and $f(u_i \mid \sigma^2)$ is a univariate normal p.d.f.

Often no further simplification of (2) is possible, so computation needs careful consideration – we will come back to this later.

Dependence

Toxoplasmosis example revisited

$Y_{ij} = \begin{cases} 1 & \text{if thoxo} \\ 0 & \text{otherwise} \end{cases}$

j^{th} person in i^{th} city

We can think of the toxoplasmosis proportions Y_i in each city (i) as arising from the sum of binary variables, Y_{ij}, representing the toxoplasmosis status of individuals (j), so $m_i Y_i = \sum_{j=1}^{m_i} Y_{ij}$. Then

$$\text{var}(Y_i) = \frac{1}{m_i^2} \sum_{j=1}^{m_i} \text{var}(Y_{ij}) + \frac{1}{m_i^2} \sum_{j \neq k} \text{cov}(Y_{ij}, Y_{ik})$$

$$= \frac{\mu_i(1-\mu_i)}{m_i} + \frac{1}{m_i^2} \sum_{j \neq k} \text{cov}(Y_{ij}, Y_{ik})$$

So any positive correlation between individuals induces overdispersion in the counts.

Dependence: reasons

There may be a number of plausible reasons why the responses corresponding to units within a given **cluster** are dependent (in the toxoplasmosis example, cluster = city)

One compelling reason is the unobserved heterogeneity discussed previously.

In the 'correct' model (corresponding to η_i^{true}), the toxoplasmosis status of individuals, Y_{ij}, are independent, so

$$Y_{ij} \perp\!\!\!\perp Y_{ik} \mid \eta_i^{\text{true}} \quad \Leftrightarrow \quad Y_{ij} \perp\!\!\!\perp Y_{ik} \mid \eta_i^{\text{model}}, \eta_i^{\text{diff}}$$

However, in the absence of knowledge of η_i^{diff}

$$Y_{ij} \not\!\perp\!\!\!\perp Y_{ik} \mid \eta_i^{\text{model}}$$

Hence conditional (given η_i^{diff}) independence between units in a common cluster i becomes marginal dependence, when marginalised over the population distribution F of unobserved η_i^{diff}.

Random effects and dependence

The correspondence between positive intra-cluster correlation and unobserved heterogeneity suggests that intra-cluster dependence might be modelled using random effects, For example, for the individual-level toxoplasmosis data

$$Y_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(\mu_{ij}), \quad \log \frac{\mu_{ij}}{1 - \mu_{ij}} = x_{ij}^{\mathrm{T}}\beta + u_i, \quad u_i \sim N(0, \sigma^2)$$

which implies

$$Y_{ij} \not\perp\!\!\!\perp Y_{ik} \mid \beta, \sigma^2$$

Intra-cluster dependence arises in many applications, and random effects provide an effective way of modelling it.

Marginal models

Random effects modelling is not the only way of accounting for intra-cluster dependence.

A **marginal model** models $\mu_{ij} \equiv \mathrm{E}(Y_{ij})$ as a function of explanatory variables, through $g(\mu_{ij}) = x_{ij}^{\mathrm{T}}\beta$, and also specifies a variance relationship $\mathrm{var}(Y_{ij}) = \sigma^2 V(\mu_{ij})/m_{ij}$ and a model for $\mathrm{corr}(Y_{ij}, Y_{ik})$, as a function of μ and possibly additional parameters.

It is important to note that the parameters β in a marginal model have a different interpretation from those in a random effects model, because for the latter

$$\mathrm{E}(Y_{ij}) = \mathrm{E}(g^{-1}[x_{ij}^{\mathrm{T}}\beta + u_i]) \neq g^{-1}(x_{ij}^{\mathrm{T}}\beta) \quad (\text{unless } g \text{ is linear}).$$

☐ A random effects model describes the mean response at the subject level ('subject specific')

☐ A marginal model describes the mean response across the population ('population averaged')

GEEs

As with the quasi-likelihood approach above, marginal models do not generally provide a full probability model for Y. Nevertheless, β can be estimated using **generalised estimating equations (GEEs)**.

The GEE for estimating β in a marginal model is of the form

$$\sum_i \left(\frac{\partial \mu_i}{\partial \beta}\right)^{\mathrm{T}} \mathrm{var}(Y_i)^{-1}(Y_i - \mu_i) = 0$$

Similar to score function for GLM

$\frac{x_i}{g'(\mu_i)}$

where $Y_i = (Y_{ij})$ and $\mu_i = (\mu_{ij})$

Consistent covariance estimates are available for GEE estimators.

Furthermore, the approach is generally robust to mis-specification of the correlation structure.

For the rest of this module, we focus on fully specified probability models.

Clustered data

Examples where data are collected in clusters include:

☐ Studies in biometry where **repeated measures** are made on experimental units. Such studies can effectively mitigate the effect of between-unit variability on important inferences.

☐ Agricultural field trials, or similar studies, for example in engineering, where experimental units are arranged within **blocks**

☐ Sample surveys where collecting data within clusters or **small areas** can save costs

Of course, other forms of dependence exist, for example spatial or serial dependence induced by arrangement in space or time of units of observation.

Example 2: Rat growth

The table below is extracted from a data set giving the weekly weights of 30 young rats.

|			Week		
Rat	1	2	3	4	5
1	151	199	246	283	320
2	145	199	249	293	354
3	147	214	263	312	328
4	155	200	237	272	297
5	135	188	230	280	323
6	159	210	252	298	331
7	141	189	231	275	305
8	159	201	248	297	338
…	…	…	…	…	…
30	153	200	244	286	324

Rat is cluster

Example

One line
per rat

Rat growth data

A simple model

Letting Y represent weight, and X represent week, we can fit the simple linear regression

$$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$$

with resulting estimates $\widehat{\beta}_0 = 156.1$ (2.25) and $\widehat{\beta}_1 = 43.3$ (0.92) Residuals show clear evidence of an unexplained difference between rats
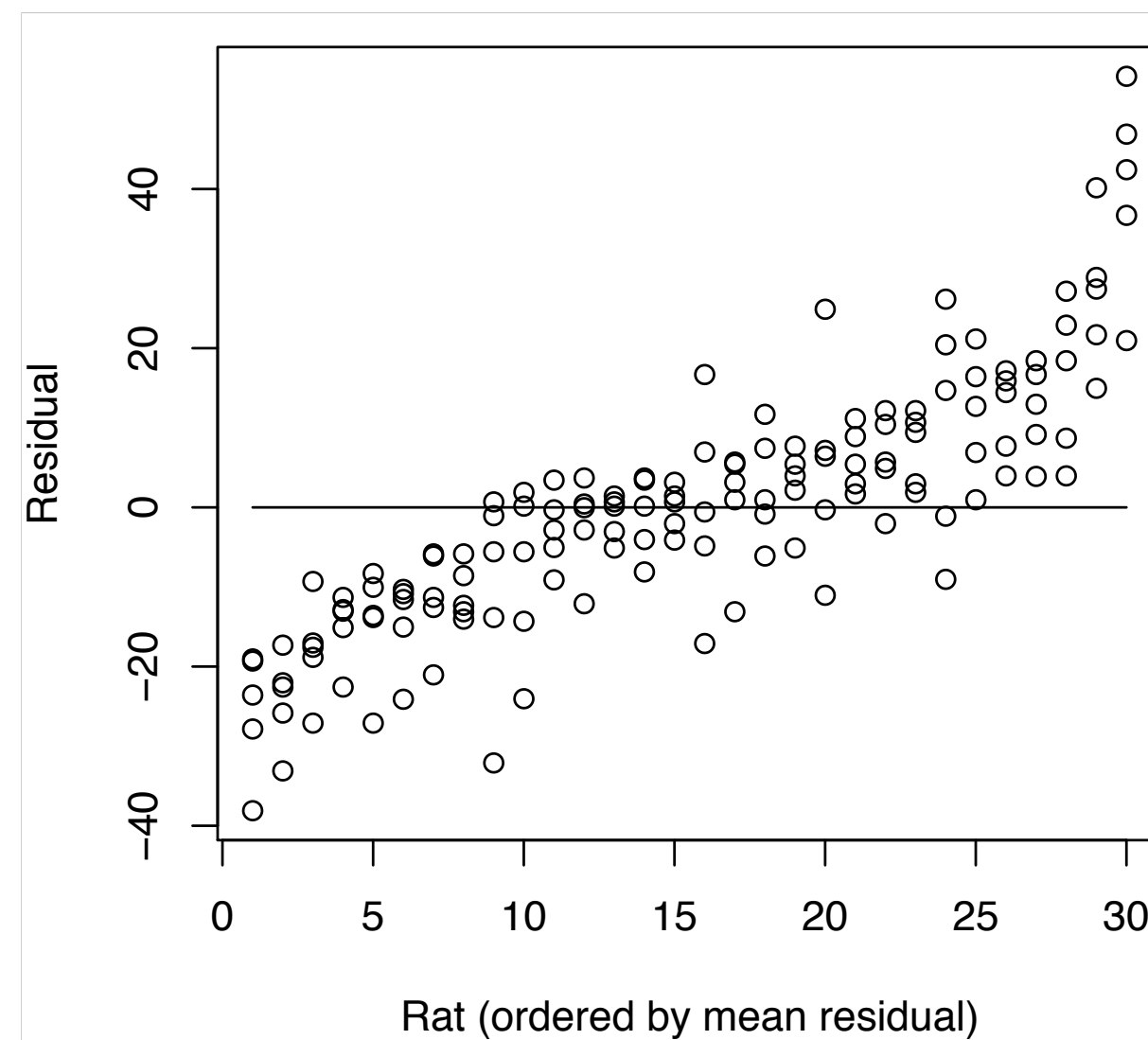
Model elaboration

Naively adding a (fixed) effect for animal gives

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_i + \epsilon_{ij}.$$

Residuals show evidence of a further unexplained difference between rats in terms of dependence on x.
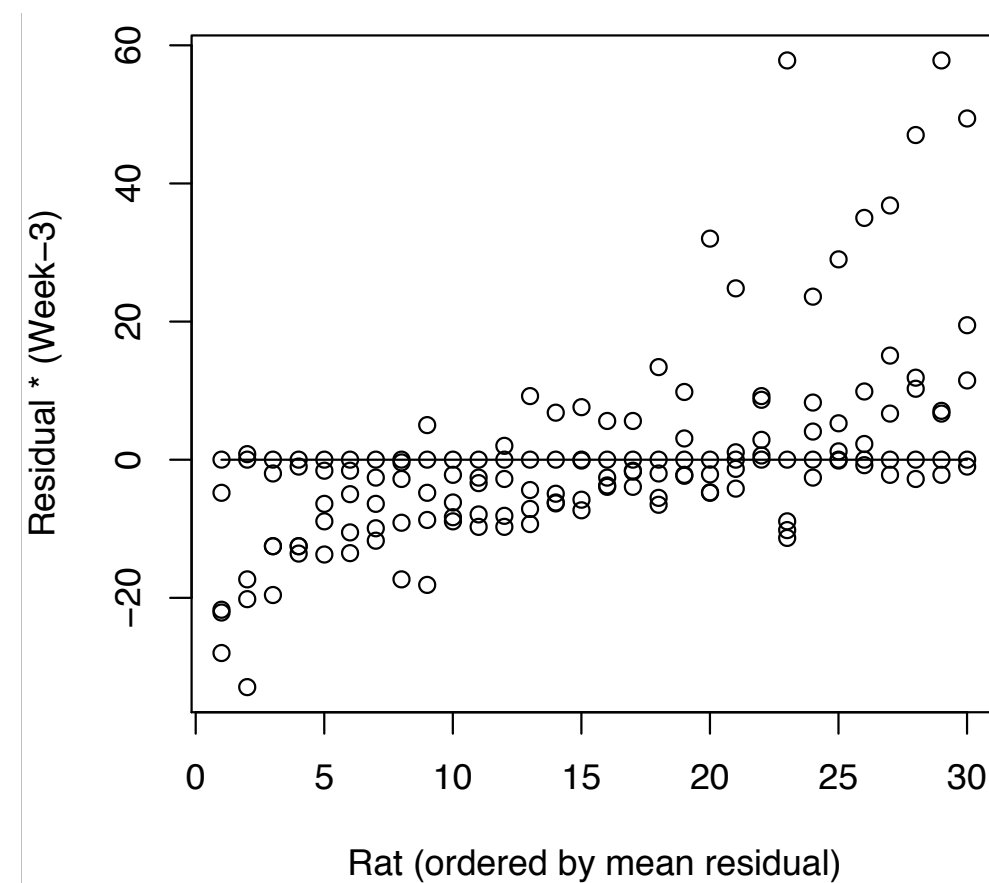
More complex cluster dependence required.

Random Effects and Mixed Models

Linear mixed models

A linear mixed model (LMM) for observations $y = (y_1, \ldots, y_n)$ has the general form

(handwritten: fixed effects, random effects)

$$Y \sim N(\mu, \Sigma), \quad \mu = X\beta + Zb, \quad b \sim N(0, \Sigma_b), \tag{3}$$

(handwritten annotations: $n \times p$, $p \times 1$, $n \times Q$, $Q \times 1$, $Q \times Q$)

where X and Z are matrices containing values of explanatory variables. Usually, $\Sigma = \sigma^2 I_n$.

A typical example for clustered data might be

$$Y_{ij} \overset{\text{ind}}{\sim} N(\mu_{ij}, \sigma^2), \quad \mu_{ij} = x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}}b_i, \quad b_i \overset{\text{ind}}{\sim} N(0, \Sigma_b^*), \tag{4}$$

where x_{ij} contain the explanatory data for cluster i, observation j and (normally) z_{ij} contains that sub-vector of x_{ij} which is allowed to exhibit extra between cluster variation in its relationship with Y.

In the simplest (random intercept) case, $z_{ij} = (1)$, as in equation (1).

k clusters n_i units in i^{th} cluster

$$X = \begin{bmatrix} x_{11}^T \\ \vdots \\ x_{1n_1}^T \\ x_{21}^T \\ \vdots \\ \vdots \\ x_{kn_k}^T \end{bmatrix} \begin{array}{l} \Big\} \text{Cluster 1} \\[20pt] \Big\} \text{Cluster 2} \\[30pt] \Big\} \text{Cluster k} \end{array}$$

$$Z = \begin{bmatrix} z_{11}^T & 0 & & 0 \\ \vdots & \vdots & & \vdots \\ z_{1n_1}^T & 0 & & 0 \\ 0 & z_{2\cdot}^T & & \\ \vdots & \vdots & & \\ & & & z_{k1}^T \\ & & & \vdots \\ 0 & 0 & & z_{kn_k}^T \end{bmatrix}$$

$b_1 \quad b_2 \quad \cdots \quad b_k$

$$\Big\} \text{Cluster 1}$$
$$\Big\} \text{Cluster k}$$

$$b^T = \left(b_1^T, \dots b_k^T \right) \qquad n = \sum_{i=1}^{k} n_i$$

$$\Sigma_b = \begin{pmatrix} \Sigma_b^* & 0 \\ & \ddots & \\ 0 & & \Sigma_b^* \end{pmatrix}$$

LMM example

A plausible LMM for k clusters with $n_1, \ldots, n_k$ observations per cluster, and a single explanatory variable x (e.g. the rat growth data) is

$$y_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})x_{ij} + \epsilon_{ij}, \quad (b_{0i}, b_{1i})^{\mathrm{T}} \overset{\text{ind}}{\sim} N(0, \Sigma_b^*).$$

This fits into the general LMM framework (3) with $\Sigma = \sigma^2 I_n$ and

$$X = \begin{pmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{kn_k} \end{pmatrix}, \quad Z = \begin{pmatrix} Z_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & Z_k \end{pmatrix}, \quad Z_i = \begin{pmatrix} 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{in_i} \end{pmatrix},$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix}, \quad b_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix}, \quad \Sigma_b = \begin{pmatrix} \Sigma_b^* & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Sigma_b^* \end{pmatrix}$$

where Σ_b^* is an unspecified 2×2 positive definite matrix.

Variance components

The term **mixed model** refers to the fact that the linear predictor $X\beta + Zb$ contains both fixed effects β and random effects b. Under an LMM, we can write the marginal distribution of Y directly as

$$Y \sim N(X\beta, \Sigma + Z\Sigma_b Z^{\mathrm{T}}) \qquad (5)$$

where X and Z are matrices containing values of explanatory variables. Hence $\mathrm{var}(Y)$ is comprised of two **variance components**.

Other ways of describing LMMs for clustered data, such as (4) (and their generalised linear model counterparts) are as **hierarchical** models or **multilevel** models. This reflects the two-stage structure of the model, a conditional model for $Y_{ij} \mid b_i$, followed by a marginal model for the random effects b_i.

Sometimes the hierarchy can have further levels, corresponding to clusters nested within clusters, for example, patients within wards within hospitals, or pupils within classes within schools.

$$f(y \mid \beta, \Sigma, \Sigma_b) = \int f(y \mid \beta, b, \Sigma) \, f(b \mid \Sigma_b) \, db$$

$$\underbrace{N(X\beta + Zb, \Sigma)}_{} \qquad \underbrace{N(0, \Sigma_b)}_{}$$

Discussion: Why random effects?

It would be perfectly possible to take a model such as (4) and ignore the final component, leading to fixed cluster effects (as we did for the rat growth data).

The main issue with such an approach is that inferences, particularly predictive inferences can then only be made about those clusters present in the observed data.
Random effects models, on the other hand, allow inferences to be extended to a wider population (at the expense of a further modelling assumption).

↗ toxoplasmosis data

It also can be the case, as in (1) with only one observation per 'cluster', that fixed effects are not identifiable, whereas random effects can still be estimated. Similarly, some treatment variables must be applied at the cluster level, so fixed treatment and cluster effects are aliased.

↖ eg. in the rat data, gender is fixed

Finally, random effects allow 'borrowing strength' across clusters by shrinking fixed effects towards a common mean.

Discussion: A Bayesian perspective

A Bayesian LMM supplements (3) with prior distributions for β, Σ and Σ_b.

In one sense the distinction between fixed and random effects is much less significant, as in the full Bayesian probability specification, both β and b, as unknowns have probability distributions, $f(\beta)$ and $f(b) = \int f(b \mid \Sigma_b) f(\Sigma_b) \mathrm{d}\Sigma_b$

Indeed, prior distributions for 'fixed' effects are sometimes constructed in a hierarchical fashion, for convenience (for example, heavy-tailed priors are often constructed this way).

e.g. a t-distribution for β $\qquad$ $\beta \mid \tau^2 \sim N(0, \tau^2)$ $\qquad$ $\tau^2 \sim IG$

The main difference is the possibility that random effects for which we have no relevant data (for example cluster effects for unobserved clusters) might need to be predicted.

LMM fitting

The likelihood for $(\beta, \Sigma, \Sigma_b)$ is available directly from (5) as

$$f(y \mid \beta, \Sigma, \Sigma_b) \propto |V|^{-1/2} \exp\left(-\tfrac{1}{2}(y - X\beta)^{\mathrm{T}} V^{-1}(y - X\beta)\right) \quad (6)$$

where $V = \Sigma + Z\Sigma_b Z^{\mathrm{T}}$. This likelihood can be maximised directly (usually numerically).

However, mles for variance parameters of LMMs can have large downward bias (particularly in cluster models with a small number of observed clusters).
Hence estimation by **REML** – *REstricted* (or *REsidual*) Maximum Likelihood is usually preferred.

REML proceeds by estimating the variance parameters (Σ, Σ_b) using a *marginal likelihood* based on the residuals from a (generalised) least squares fit of the model $\mathrm{E}(Y) = X\beta$.

REML

degenerate because $I_n - H$ is not of full rank

In effect, REML maximizes the likelihood of any linearly independent sub-vector of $(I_n - H)y$ where $H = X(X^T X)^{-1} X^T$ is the usual hat matrix. As

$$(I_n - H)y \sim N(0, (I_n - H)V(I_n - H))$$

this likelihood will be free of β. It can be written in terms of the full likelihood (6) as

$r = B(I_n - H)y \quad \leftarrow$ *$n-p$ linearly independent residuals*

$$f(r \mid \Sigma, \Sigma_b) \propto f(y \mid \widehat{\beta}, \Sigma, \Sigma_b)|X^T V X|^{1/2} \qquad (7)$$

distribution of r is independent of β

where

$$\widehat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y \qquad (8)$$

is the usual generalised least squares estimator given known V. Having first obtained $(\widehat{\Sigma}, \widehat{\Sigma}_b)$ by maximising (7), $\widehat{\beta}$ is obtained by plugging the resulting $\widehat{V}$ into (8).

Note that REML maximised likelihoods cannot be used to compare different fixed effects specifications, due to the dependence of 'data' r in $f(r \mid \Sigma, \Sigma_b)$ on X.

Estimating random effects

A natural predictor $\tilde{b}$ of the random effect vector b is obtained by minimising the mean squared prediction error $\mathrm{E}[(\tilde{b} - b)^{\mathrm{T}}(\tilde{b} - b)]$ where the expectation is over both b and y.

This is achieved by

$$\tilde{b} = \mathrm{E}(b \mid y) = (Z^{\mathrm{T}}\Sigma^{-1}Z + \Sigma_b^{-1})^{-1}Z^{\mathrm{T}}\Sigma^{-1}(y - X\beta) \tag{9}$$

giving the **Best Linear Unbiased Predictor** (BLUP) for b, with corresponding variance

$$\mathrm{var}(b \mid y) = (Z^{\mathrm{T}}\Sigma^{-1}Z + \Sigma_b^{-1})^{-1}$$

The estimates are obtained by plugging in $(\widehat{\beta}, \widehat{\Sigma}, \widehat{\Sigma}_b)$, and are **shrunk** towards 0, in comparison with equivalent fixed effects estimators.

Any component, b_k of b with no relevant data (for example a cluster effect for an as yet unobserved cluster) corresponds to a null column of Z, and then $\tilde{b}_k = 0$ and $\mathrm{var}(b_k \mid y) = [\Sigma_b]_{kk}$, which may be estimated if, as is usual, b_k shares a variance with other random effects.

Bayesian estimation: the Gibbs sampler

Bayesian estimation in LMMs (and their generalised linear model counterparts) generally proceeds using **Markov Chain Monte Carlo (MCMC)** methods, in particular approaches based on the **Gibbs sampler**. Such methods have proved very effective.

MCMC computation provides posterior summaries, by **generating** a **dependent** sample from the posterior distribution of interest. Then, any posterior expectation can be estimated by the corresponding Monte Carlo sample mean, densities can be estimated from samples etc.

MCMC will be covered in detail in APTS: Computer Intensive Statistics. Here we simply describe the (most basic) Gibbs sampler.

To generate from $f(y_1, \ldots, y_n)$, (where the component y_is are allowed to be multivarate) the Gibbs sampler starts from an arbitrary value of y and updates components (sequentially or otherwise) by generating from the conditional distributions $f(y_i \mid y_{\setminus i})$ where $y_{\setminus i}$ are all the variables other than y_i, set at their currently generated values.

Hence, to apply the Gibbs sampler, we require conditional distributions which are available for sampling.

Bayesian estimation for LMMs

For the LMM

$$Y \sim N(\mu, \Sigma), \quad \mu = X\beta + Zb, \quad b \sim N(0, \Sigma_b)$$

with corresponding prior densities $f(\beta)$, $f(\Sigma)$, $f(\Sigma_b)$, we obtain the *conditional* posterior distributions

$$
\begin{aligned}
f(\beta \mid y, \text{rest}) &\propto \phi(y - Zb; X\beta, V) f(\beta) \\
f(b \mid y, \text{rest}) &\propto \phi(y - X\beta; Zb, V) \phi(b; 0, \Sigma_b) \\
f(\Sigma \mid y, \text{rest}) &\propto \phi(y - X\beta - Zb; 0, V) f(\Sigma) \\
f(\Sigma_b \mid y, \text{rest}) &\propto \phi(b; 0, \Sigma_b) f(\Sigma_b)
\end{aligned}
$$

(handwritten annotations:) N then conditional distribution of β will be N; N; IW (inverse Wishart) ↳ generalisation of IG to pos-def matrices; IW

where $\phi(y; \mu, \Sigma)$ is a $N(\mu, \Sigma)$ p.d.f. evaluated at y.

We can exploit **conditional conjugacy** in the choices of $f(\beta)$, $f(\Sigma)$, $f(\Sigma_b)$ making the conditionals above of known form and hence straightforward to sample from. The conditional independence $(\beta, \Sigma) \perp\!\!\!\perp \Sigma_b \mid b$ is also helpful.

See Practical 2 for further details.

Example: Rat growth revisited

Here, we consider the model

$$y_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})x_{ij} + \epsilon_{ij}, \quad (b_{0i}, b_{1i})^{\mathrm{T}} \overset{\mathrm{ind}}{\sim} N(0, \Sigma_b)$$

where $\epsilon_{ij} \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ and Σ_b is an unspecified covariance matrix. This model allows for random (cluster specific) slope and intercept.

Estimates obtained by REML (ML in brackets) are

Parameter	Estimate	Standard error
β_0	156.05	2.16 (2.13)
β_1	43.27	0.73 (0.72)
$\Sigma_{00}^{1/2} = s.d.(b_0)$	10.93 (10.71)	
$\Sigma_{11}^{1/2} = s.d.(b_1)$	3.53 (3.46)	
$Corr(b_0, b_1)$	0.18 (0.19)	
σ	5.82 (5.82)	

As expected ML variances are smaller, but not by much.

Example: Fixed v. random effect estimates

The shrinkage of random effect estimates towards a common mean is clearly illustrated.
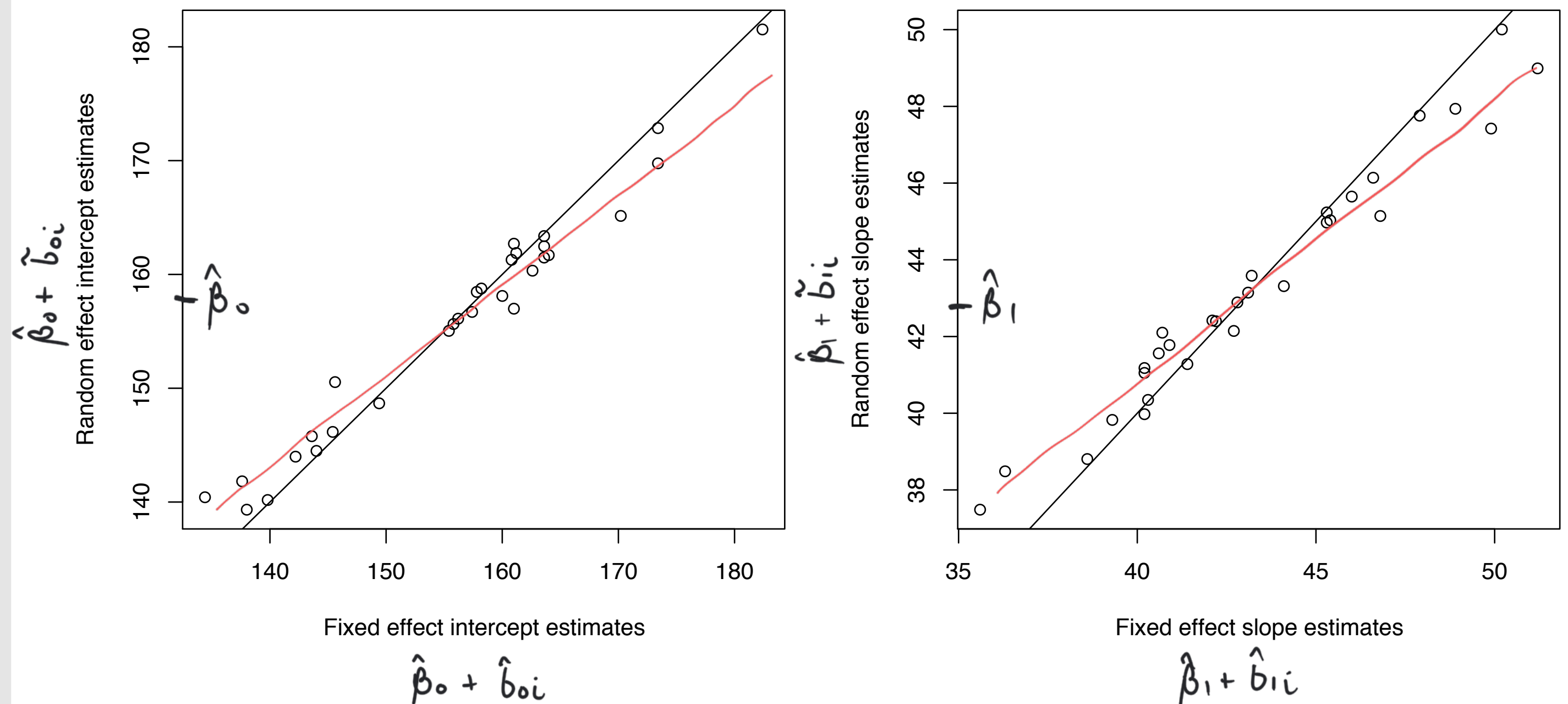
Random effects estimates 'borrow strength' across clusters, due to the Σ_b^{-1} term in (9). Extent of this is determined by cluster similarity. This is usually considered to be a desirable behaviour.

Random effect shrinkage

The following simple example illustrates (from a Bayesian perspective) why and how random effects are shrunk to a common value. Suppose that $y_1, \ldots, y_n$ satisfy

$$y_j \mid \theta_j \overset{\text{ind}}{\sim} N(\theta_j, v_j), \quad \theta_1, \ldots, \theta_n \mid \mu \overset{\text{iid}}{\sim} N(\mu, \sigma^2), \quad \mu \sim N(\mu_0, \tau^2),$$

where $v_1, \ldots, v_n$, σ^2, μ_0 and τ^2 are assumed known here. Then, the usual posterior calculations give us

$$\text{E}(\mu \mid y) = \frac{\mu_0/\tau^2 + \sum y_j/(\sigma^2 + v_j)}{1/\tau^2 + \sum 1/(\sigma^2 + v_j)}, \; \text{var}(\mu \mid y) = \frac{1}{1/\tau^2 + \sum 1/(\sigma^2 + v_j)},$$

and

$$\text{E}(\theta_j \mid y) = (1 - w)\text{E}(\mu \mid y) + w y_j,$$

where

$$w = \frac{\sigma^2}{\sigma^2 + v_j}.$$

If θ_j was a fixed effect

$E(\theta_j|y) = y_j$

Example: Diagnostics

Normal Q-Q plots of intercept (panel 1) and slope (panel 2) random effects and residuals v. week (panel 3)
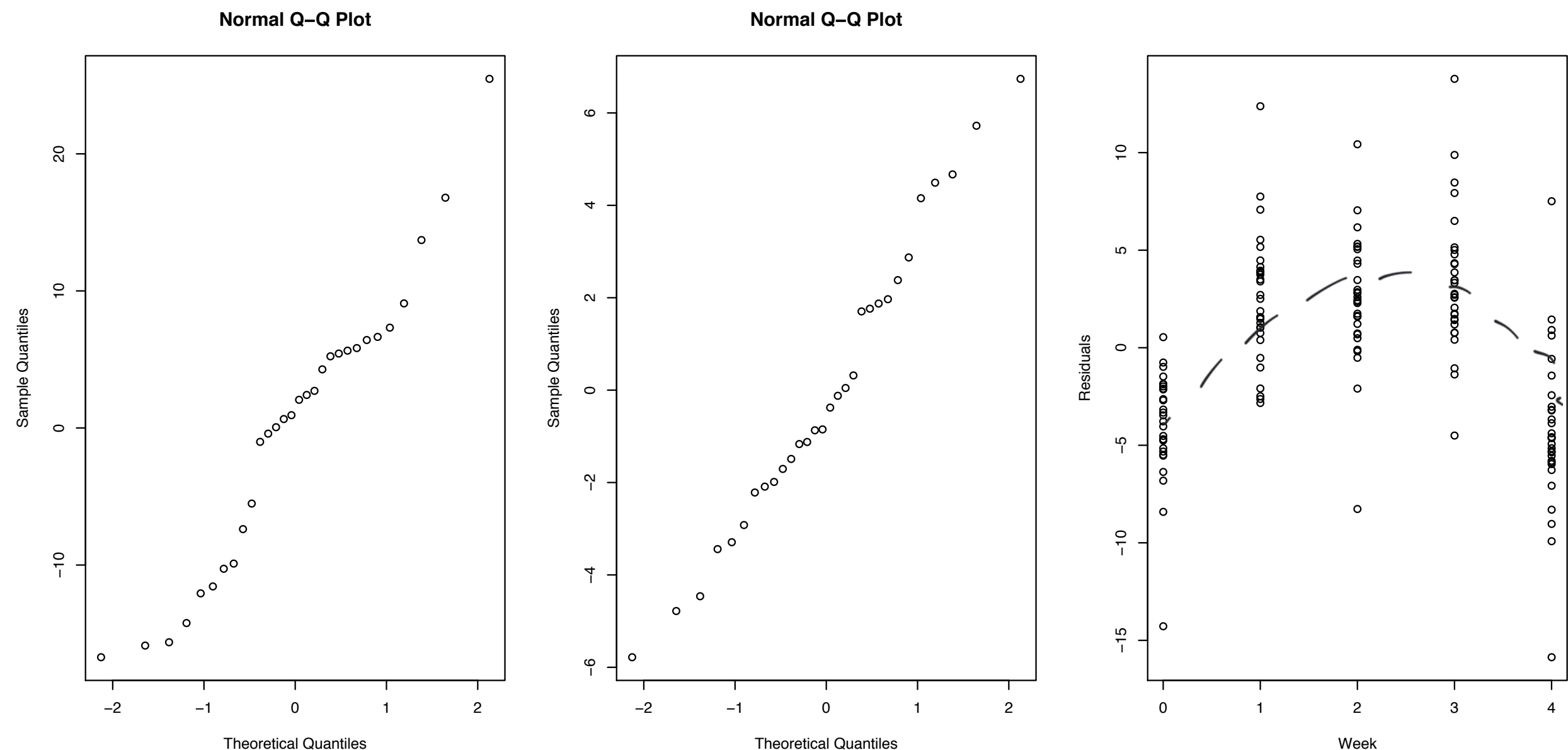
Evidence of a common quadratic effect, confirmed by AIC (1036 v. 1099) and BIC (1054 v. 1114) based on full ML fits. AIC would also include a cluster quadratic effect (BIC equivocal).

not REML

Generalised linear mixed models

Generalised linear mixed models (GLMMs) generalise LMMs to non-normal data, in the obvious way:

$$
Y_i \overset{\text{ind}}{\sim} F(\cdot \mid \mu_i, \sigma^2), \quad g(\mu) \equiv \begin{pmatrix} g(\mu_1) \\ \vdots \\ g(\mu_n) \end{pmatrix} = X\beta + Zb, \quad b \sim N(0, \Sigma_b)
$$

$$(10)$$

where $F(\cdot \mid \mu_i, \sigma^2)$ is an exponential family distribution with $\mathrm{E}(Y) = \mu$ and $\mathrm{var}(Y) = \sigma^2 V(\mu)/m$ for known m. Commonly (e.g. Binomial, Poisson) $\sigma^2 = 1$, and we shall assume this from here on.

It is not necessary that the distribution for the random effects b is normal, but this usually fits. It is possible (but beyond the scope of this module) to relax this.

GLMM example

A plausible GLMM for binary data in k clusters with $n_1, \ldots, n_k$ observations per cluster, and a single explanatory variable x (e.g. the toxoplasmosis data at individual level) is

$$Y_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(\mu_i), \quad \log \frac{\mu_i}{1 - \mu_i} = \beta_0 + b_{0i} + \beta_1 x_{ij}, \quad b_{0i} \overset{\text{ind}}{\sim} N(0, \sigma_b^2)$$

$$(11)$$

[note: no random slope here]. This fits into the general GLMM framework (10) with

$$X = \begin{pmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{kn_k} \end{pmatrix}, \quad Z = \begin{pmatrix} Z_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & Z_k \end{pmatrix}, \quad Z_i = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix},$$

$$\beta = (\beta_0, \beta_1)^{\mathrm{T}}, \quad b = (b_{01}, \ldots, b_{0k})^{\mathrm{T}}, \quad \Sigma_b = \sigma_b^2 I_k$$

[or equivalent binomial representation for city data, with clusters of size 1.]

GLMM likelihood

The marginal distribution for the observed Y in a GLMM does not usually have a convenient closed-form representation.

Bernoulli in toxo example

$$
\begin{aligned}
f(y \mid \beta, \Sigma_b) &= \int f(y \mid \beta, b, \Sigma_b) f(b \mid \beta, \Sigma_b) \mathrm{d}b \\
&\qquad\qquad\qquad\qquad\qquad\qquad \longrightarrow N(0, \Sigma_b) \\
&= \int f(y \mid \beta, b) f(b \mid \Sigma_b) \mathrm{d}b \\
&= \int \prod_{i=1}^{n} f\left(y_i \mid g^{-1}([X\beta + Zb]_i)\right) f(b \mid \Sigma_b) \mathrm{d}b.
\end{aligned} \tag{12}
$$

For **nested** random effects structures, some simplification is possible. For example, for (11)

$$
f(y \mid \beta, \sigma_b^2) \propto \prod_{i=1}^{n} \int \frac{\exp\left(\sum_j y_{ij}(\beta_0 + b_{0i} + \beta_1 x_{ij})\right)}{\left\{1 + \exp\left(\sum_j y_{ij}(\beta_0 + b_{0i} + \beta_1 x_{ij})\right)\right\}^{n_k}} \phi(b_{0i}; 0, \sigma_b^2) \mathrm{d}b_{0i}
$$

a product of one-dimensional integrals.

GLMM fitting: quadrature

$$\int f(b)\, \phi(b)\, db \approx \sum_{q=1}^{Q} w_q f(b_q)$$

Fitting a GLMM by likelihood methods requires some method for approximating the integrals involved.

The most reliable when the integrals are of low dimension is to use Gaussian quadrature (see APTS: Statistical computing). For example, for a one-dimensional cluster-level random intercept b_i we might use

$$\int \prod_j f\left(y_{ij} \mid g^{-1}(x_i^{\mathrm{T}}\beta + b_i)\right) \phi(b_i \mid 0, \sigma_b^2) \mathrm{d}b_i$$

$$\approx \sum_{q=1}^{Q} w_q \prod_j f\left(y_{ij} \mid g^{-1}(x_i^{\mathrm{T}}\beta + b_{iq})\right)$$

for suitably chosen weights $(w_q, q = 1, \ldots, Q)$ and quadrature points $(b_{iq}, q = 1, \ldots, Q)$

Effective quadrature approaches use information about the mode and dispersion of the integrand (can be done adaptively).

For multi-dimensional b_i, quadrature rules can be applied recursively, but performance (in fixed time) diminishes rapidly with dimension.

GLMM fitting: Penalised quasi-likelihood

An alternative approach to fitting a GLMM uses penalised quasi-likelihood (PQL).

The most straightforward way of thinking about PQL is to consider the adjusted dependent variable v constructed when computing mles for a GLM using Fisher scoring

$$v_i = (y_i - \mu_i)g'(\mu_i) + \eta_i$$

Now, for a GLMM,

$$\mathrm{E}(v \mid b) = \eta = X\beta + Zb$$

and

$$\mathrm{var}(v \mid b) = W^{-1} = \mathrm{diag}\left(\mathrm{var}(y_i)g'(\mu_i)^2\right),$$

where W is the weight matrix used in Fisher scoring.

1) Given current η update v and W

2) Given v and W calculate weighted least squares estimate of β

GLMM fitting: PQL continued

Hence, approximating the conditional distribution of v by a normal distribution, we have

$$v \sim N(X\beta + Zb, W^{-1}), \quad b \sim N(0, \Sigma_b) \tag{13}$$

where v and W also depend on β and b.

PQL proceeds by iteratively estimating β, b and Σ_b for the linear mixed model (13) for v, updating v and W at each stage, based on the current estimates of β and b.

An alternative justification for PQL is as using a Laplace-type approximation to the integral in the GLMM likelihood.

A full Laplace approximation (expanding the complete log-integrand, and evaluating the Hessian matrix at the mode) is an alternative, equivalent to one-point Gaussian quadrature.

Laplace approximation

$$f(y \mid \beta, \Sigma_b) = \int f(y \mid \beta, b) \, f(b \mid \Sigma_b) \, db$$

$$= \int h(b \mid y, \beta, \Sigma_b) \, db$$

Take 2nd order Taylor series expansion of $\log h(b \mid y, \beta, \Sigma_b)$

about $\quad \bar{b} = \arg\max_{b \in \mathbb{R}^q} h(b \mid y, \beta, \Sigma_b)$

$$\log h(b \mid y, \beta, \Sigma_b) \approx \log h(\bar{b} \mid y, \beta, \Sigma_b) \;+\; \underbrace{\frac{\partial \log h(\bar{b} \mid y, \Sigma_b, \beta)}{\partial b}}_{\;\approx 0}(b - \bar{b})$$

$$+ \frac{1}{2}(b - \bar{b})^T \underbrace{\frac{\partial^2 \log h(\bar{b} \mid y, \Sigma_b, \beta)}{\partial b \, \partial b^T}}_{= -H^{-1}} (b - \bar{b})$$

$$f(y \mid \beta, \Sigma_b) \approx h(\bar{b} \mid y, \beta, \Sigma_b) \int \exp\left(-\frac{1}{2}(b - \bar{b})^T H^{-1}(b - \bar{b})\right) db$$

$$= (2\pi)^{q/2} |H|^{1/2} \, h(\bar{b} \mid y, \beta, \Sigma_b)$$

Using PQL, estimates of random effects b come 'for free'. With Gaussian quadrature, some extra effort is required to compute $\mathrm{E}(b \mid y)$ – further quadrature is an obvious possibility.

There are drawbacks with PQL, and the best advice is to use it with caution.

☐ It can fail badly when the normal approximation that justifies it is invalid (for example for binary observations)

☐ As it does not use a full likelihood, model comparison should not be performed using PQL maximised 'likelihoods'

Likelihood inference for GLMMs remains an area of active research and vigorous debate. Recent approaches include HGLMs (hierarchical GLMs) where inference is based on the h-likelihood $f(y \mid \beta, b) f(b \mid \Sigma)$.

Bayesian estimation for GLMMs

Bayesian estimation in GLMMs, as in LMMs, is generally based on the Gibbs sampler. For the GLMM

$$Y_i \overset{\text{ind}}{\sim} F(\cdot \mid \mu), \quad g(\mu) = X\beta + Zb, \quad b \sim N(0, \Sigma_b)$$

with corresponding prior densities $f(\beta)$ and $f(\Sigma_b)$, we obtain the *conditional* posterior distributions

$$f(\beta \mid y, \text{rest}) \quad \propto \quad f(\beta) \prod_i f(y_i \mid g^{-1}(X\beta + Zb))$$

$$f(b \mid y, \text{rest}) \quad \propto \quad \phi(b; 0, \Sigma_b) \prod_i f(y_i \mid g^{-1}(X\beta + Zb))$$

$$f(\Sigma_b \mid y, \text{rest}) \quad \propto \quad \phi(b; 0, \Sigma_b) f(\Sigma_b)$$

For a conditionally conjugate choice of $f(\Sigma_b)$, $f(\Sigma_b \mid y, \text{rest})$ is straightforward to sample from. The conditionals for β and b are not generally available for direct sampling, but there are a number of ways of modifying the basic approach to account for this.

Toxoplasmosis revisited

Estimates and standard errors obtained by ML (quadrature), Laplace and PQL for the individual-level model

$$Y_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(\mu_i), \quad \log \frac{\mu_i}{1 - \mu_i} = \beta_0 + b_{0i} + \beta_1 x_{ij}, \quad b_{0i} \overset{\text{ind}}{\sim} N(0, \sigma_b^2)$$

Parameter	Estimate (s.e.)		
	ML	Laplace	PQL
β_0	-0.1384 (1.452)	-0.1343 (1.440)	-0.115 (1.445)
β_1 $(\times 10^6)$	7.215 (752)	5.930 (745.7)	0.57 (749.2)
σ_b	0.5209	0.5132	0.4946
AIC	65.75	65.96	—

Toxoplasmosis continued

Estimates and standard errors obtained by ML (quadrature), Laplace and PQL for the extended model

$$\log \frac{\mu_i}{1 - \mu_i} = \beta_0 + b_{0i} + \beta_1 x_{ij} + \beta_1 x_{ij}^2 + \beta_1 x_{ij}^3.$$

Parameter	Estimate (s.e.)		
	ML	Laplace	PQL
β_0	-335.5 (137.3)	-335.1 (136.3)	-330.8 (143.4)
β_1	0.5238 (0.2128)	0.5231 (0.2112)	0.5166 (0.222)
β_2 ($\times 10^4$)	-2.710 (1.094)	-2.706 (1.086)	-3 (1.1)
β_3 ($\times 10^8$)	4.643 (1.866)	4.636 (1.852)	0 (0)
σ_b	0.4232	0.4171	0.4315
AIC	63.84	63.97	—

So for this example, a good agreement between the different computational methods.