APTS Statistical Modelling

Helen Ogden
based on notes by Anthony Davison, Jon Forster, Dave Woods and Antony Overstall

April 2022

1 Model selection

1.1 Introduction

> All models are wrong, but some models are useful
>
> – *George Box (1919–2013)*

Statisticians construct models to simplify reality, to gain understanding, to compare scientific, economic, or other theories, and to predict future events or data. We rarely believe in our models, but regard them as temporary constructs, which should be subject to improvement. Often we have several models and must decide which, if any, is preferable.

Criteria for model selection include:

- Substantive knowledge, from previous studies, theoretical arguments, dimensional or other general considerations.
- Sensitivity to failure of assumptions: we prefer models that provide valid inference even if some of their assumptions are invalid.
- Quality of fit of models to data: we could use informal measures such as residuals, graphical assessment, or more formal or goodness-of-fit tests.
- For reasons of economy we seek 'simple' models.

There may be a very large number of plausible models for us to compare. For instance, in a linear regression with p covariates, there are 2^p possible combinations of covariates: for each covariate, we need to decide whether or not to include that variable in the model. If $p = 20$ we have over a million possible models to consider, and the problem becomes even more complex if we allow for transformations and interactions in the model.

To focus and simplify discussion we will consider model selection among parametric models, but the ideas generalise to semi-parametric and non-parametric settings.

Example 1.1. A logistic regression model for binary responses assumes that $Y_i \sim \text{Bernoulli}(\pi_i)$, with a linear model for log odds of 'success'

$$\log\left\{\frac{P(Y_i = 1)}{P(Y_i = 0)}\right\} = \log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^T\beta.$$

The log-likelihood for β based on independent responses with covariate vectors $x_1, \ldots, x_n$ is

$$\ell(\beta) = \sum_{j=1}^{n} y_j x_j^T\beta - \sum_{j=1}^{n} \log\left\{1 + \exp(x_j^T\beta)\right\}$$

A good fit gives large fitted loglikelihood $\widehat{\ell} = \ell(\widehat{\beta})$ where $\widehat{\beta}$ is the MLE under the model.

The `SMPracticals` package contains a dataset called `nodal`, which relates to the the nodal involvement (`r`) of 53 patients with prostate cancer, with five binary covariates `aged`, `stage`, `grade`, `xray` and `acid`. Considering only of models without any interaction between the 5 binary covariates, there are still $2^5 = 32$ possible logistic regression models for this data. We can rank these models according to fitted loglikelihood $\widehat{\ell}$. Figure 1 summarises this as a plot of the number of parameters against the fitted loglikelihood for each of the 32 models under consideration.

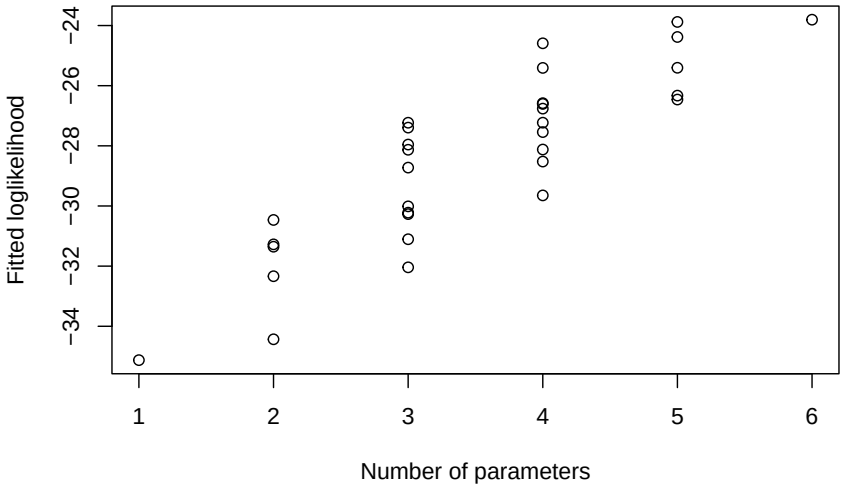

Figure 1: Fitted loglikelihood for 32 possible logistic regression models for the `nodal` data

Adding terms always increases the loglikelihood $\hat{\ell}$, so taking the model with highest $\hat{\ell}$ would give the full model. We need to a different way to compare models, which should trade off quality of fit (measured by $\hat{\ell}$) and model complexity (number of parameters).

1.2 Criteria for model selection

1.2.1 Likelihood inference under the wrong model

Suppose the (unknown) **true model** is $g(y)$, that is, $Y_1, \ldots, Y_n \sim g$. Suppose we have a **candidate model** $f(y;\theta)$, under which we assume $Y_1, \ldots, Y_n \sim f(y;\theta)$, which we wish to compare against other candidate models. For each candidate model, we will first find maximum likelihood estimate $\hat{\theta}$ of the model parameters, then use some criteria based on the fitted loglikelihood $\hat{\ell} = \ell(\hat{\theta})$ to compare candidate models.

We do not assume that our candidate models are correct: there may be no value of θ such that $f(.;\theta) = g(.)$. Before we can decide on an appropriate criterion for choosing between models, we first need to understand the asymptotic behaviour of $\hat{\theta}$ and $\ell(\hat{\theta})$ without the usual assumption that the model is correctly specified.

The log likelihood $\ell(\theta)$ will be maximised at $\hat{\theta}$, and

$$\bar{\ell}(\hat{\theta}) = n^{-1}\ell(\hat{\theta}) \to \int \log f(y;\theta_g) g(y)\, dy, \quad \text{almost surely as } n \to \infty,$$

where θ_g minimises the Kullback–Leibler divergence

$$KL(f_\theta, g) = \int \log\left\{\frac{g(y)}{f(y;\theta)}\right\} g(y)\, dy.$$

Theorem 1.1. *Suppose the true model is g, that is, $Y_1, \ldots, Y_n \sim g$, but we assume that $Y_1, \ldots, Y_n \sim f(y;\theta)$. Then under mild regularity conditions, the maximum likelihood estimator $\hat{\theta}$ has asymptotic distribution*

$$\hat{\theta} \;\sim\; N_p\left\{\theta_g, I(\theta_g)^{-1} K(\theta_g) I(\theta_g)^{-1}\right\}, \tag{1}$$

where

$$K(\theta) = n \int \frac{\partial \log f(y;\theta)}{\partial \theta} \frac{\partial \log f(y;\theta)}{\partial \theta^T} g(y)\, dy,$$
$$I(\theta) = -n \int \frac{\partial^2 \log f(y;\theta)}{\partial \theta \partial \theta^T} g(y)\, dy.$$

The likelihood ratio statistic has asymptotic distribution

$$W(\theta_g) = 2\left\{\ell(\hat{\theta}) - \ell(\theta_g)\right\} \sim \sum_{r=1}^{p} \lambda_r V_r,$$

where $V_1, \ldots, V_p \sim \chi^2_1$*, and the* λ_r *are eigenvalues of* $K(\theta_g)^{1/2} I(\theta_g)^{-1} K(\theta_g)^{1/2}$*. Thus* $E\{W(\theta_g)\} = \mathrm{tr}\{I(\theta_g)^{-1} K(\theta_g)\}$*.*

Under the correct model, θ_g is the 'true' value of θ, $K(\theta) = I(\theta)$, $\lambda_1 = \cdots = \lambda_p = 1$, and we recover the usual results.

In practice $g(y)$ is of course unknown, and then $K(\theta_g)$ and $I(\theta_g)$ may be estimated by

$$\hat{K} = \sum_{j=1}^{n} \frac{\partial \log f(y_j;\hat{\theta})}{\partial \theta} \frac{\partial \log f(y_j;\hat{\theta})}{\partial \theta^T}, \quad \hat{J} = -\sum_{j=1}^{n} \frac{\partial^2 \log f(y_j;\hat{\theta})}{\partial \theta \partial \theta^T};$$

the latter is just the observed information matrix. We may then construct confidence intervals for θ_g using (1) with variance matrix $\hat{J}^{-1}\hat{K}\hat{J}^{-1}$.

1.2.2 Information criteria

Using the fitted likelihood $\bar{\ell}(\hat{\theta})$ to choose between models leads to overfitting, because we use the data twice: first to estimate θ, then again to evaluate the model fit. If we had another independent sample $Y_1^+, \ldots, Y_n^+ \sim g$ and computed

$$\bar{\ell}^+(\hat{\theta}) = n^{-1} \sum_{j=1}^{n} \log f(Y_j^+;\hat{\theta}),$$

then we would not have this problem, suggesting that we choose the candidate model that maximises

$$\Delta = E_g\left[E_g^+\left\{\bar{\ell}^+(\hat{\theta})\right\}\right],$$

where the inner expectation is over the distribution of the Y_j^+, and the outer expectation is over the distribution of $\hat{\theta}$.

Since $g(.)$ is unknown, we cannot compute Δ directly. We will show that $\bar{\ell}(\hat{\theta})$ is a biased estimator of Δ, but by adding an appropriate penalty term we can obtain an approximately unbiased estimator of Δ, which we can use for model comparison.

We write

$$E_g\{\bar{\ell}(\hat{\theta})\} = \underbrace{E_g\{\bar{\ell}(\hat{\theta}) - \bar{\ell}(\theta_g)\}}_{a} + \underbrace{E_g\{\bar{\ell}(\theta_g)\} - \Delta}_{b} + \Delta$$

We will find expressions for a and b, which will give us the bias in using $\bar{\ell}(\hat{\theta})$ to estimate Δ, and allow us to correct for this bias.

We have

$$a = E_g\{\bar{\ell}(\hat{\theta}) - \bar{\ell}(\theta_g)\} = \frac{1}{2n} E_g\{W(\theta_g)\} \approx \frac{1}{2n}\,\mathrm{tr}\{I(\theta_g)^{-1} K(\theta_g)\}.$$

Results on inference under the wrong model may be used to show that

$$b = E_g\{\bar{\ell}(\theta_g)\} - \Delta \approx \frac{1}{2n}\,\mathrm{tr}\{I(\theta_g)^{-1} K(\theta_g)\},$$

where the second term is a penalty that depends on the model dimension. We will not prove this here.

Putting this together, we have

$$E_g\{\bar{\ell}(\hat{\theta})\} = \Delta + a + b = \Delta + \frac{1}{n}\,\mathrm{tr}\{I(\theta_g)^{-1} K(\theta_g)\},$$

so remove the bias in using $\bar{\ell}(\hat{\theta})$ to estimate Δ, we aim to maximise

$$\bar{\ell}(\hat{\theta}) - \frac{1}{n}\,\mathrm{tr}(\hat{J}^{-1}\hat{K}).$$

Equaivalently, we can maximise

$$\hat{\ell} - \mathrm{tr}(\hat{J}^{-1}\hat{K}),$$

or equivalently **minimise**

$$2\{\mathrm{tr}(\hat{J}^{-1}\hat{K}) - \hat{\ell}\},$$

the Network Information Criterion (NIC).

Let $p = \dim(\theta)$ be the number of parameters for a model, and $\hat{\ell}$ the corresponding maximised log likelihood. There are many other information criteria with a variety of penalty terms:

- $2(p - \hat{\ell})$ (AIC—Akaike Information Criterion)
- $2(\frac{1}{2}p \log n - \hat{\ell})$ (BIC—Bayes Information Criterion)
- $\mathrm{AIC_c}$, $\mathrm{AIC_u}$, DIC, EIC, FIC, GIC, SIC, TIC, ...
- Mallows $C_p = RSS/s^2 + 2p - n$ commonly used in regression problems, where RSS is residual sum of squares for candidate model, and s^2 is an estimate of the error variance σ^2.

Example 1.2. AIC and BIC can both be used to choose between the 2^5 models previously fitted to the nodal involvement data. In this case, both prefer the same model, which includes three of the five covariates: `acid`, `stage` and `xray` (so has four free parameters).

Figure 2 shows the AIC and BIC for each model, against the number of free parameters. BIC increases more rapidly than AIC after the minimum, as it penalises more strongly against complex models.

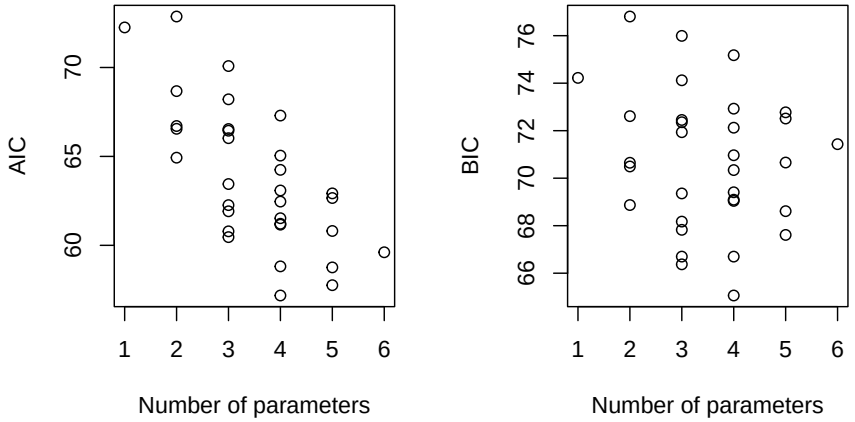

Figure 2: AIC and BIC for 32 possible logistic regression models for the nodal data

1.2.3 Theoretical properties of information criteria

We may suppose that the true underlying model is of infinite dimension, and that by choosing among our candidate models we hope to get as close as possible to this ideal model, using the data available. If so, we need some measure of distance between a candidate and the true model, and we aim to minimise this distance. A model selection procedure that selects the candidate closest to the truth for large n is called **asymptotically efficient**.

An alternative is to suppose that the true model is among the candidate models. If so, then a model selection procedure that selects the true model with probability tending to one as $n \to \infty$ is called **consistent**.

We seek to find the correct model by minimising $\mathrm{IC} = c(n,p) - 2\hat{\ell}$, where the penalty $c(n,p)$ depends on sample size n and model dimension p

- Crucial aspect is behaviour of differences of IC.
- We obtain IC for the true model, and $\mathrm{IC_+}$ for a model with one more parameter.

Then

$$\begin{aligned} P(\mathrm{IC_+} < \mathrm{IC}) &= P\left\{c(n,p+1) - 2\hat{\ell}_+ < c(n,p) - 2\hat{\ell}\right\} \\ &= P\left\{2(\hat{\ell}_+ - \hat{\ell}) > c(n,p+1) - c(n,p)\right\}. \end{aligned}$$

and in large samples

$$\begin{aligned}&\text{for AIC, } c(n,p+1)-c(n,p)=2\\&\text{for NIC, } c(n,p+1)-c(n,p)\approx 2\\&\text{for BIC, } c(n,p+1)-c(n,p)=\log n\end{aligned}$$

In a regular case $2(\hat{\ell}_+ - \hat{\ell}) \sim \chi^2_1$, so as $n \to \infty$,

$$P(\text{IC}_+ < \text{IC}) \to \begin{cases} 0.16, & \text{AIC, NIC,} \\ 0, & \text{BIC.} \end{cases}$$

Thus AIC and NIC have non-zero probability of over-fitting, even in very large samples, but BIC does not.

1.3 Variable selection for linear models

Consider a normal linear model

$$Y_{n\times 1} = X^\dagger_{n\times q}\beta_{q\times 1} + \epsilon_{n\times 1}, \quad \epsilon \sim N_n(0, \sigma^2 I_n),$$

with design matrix $X^\dagger$ with columns x_r, for $r \in \mathcal{X} = \{1, \dots, q\}$. We choose a model corresponding to a subset $\mathcal{S} \subseteq \mathcal{X}$ of columns of $X^\dagger$, of dimension $p = |S|$.

Terminology:

- the **true** model corresponds to the subset $\mathcal{T} = \{r : \beta_r \neq 0\}$, and $|\mathcal{T}| = p_0 < q$;
- a **correct** model contains $\mathcal{T}$ but has other columns also, corresponding subset $\mathcal{S}$ satisfies $\mathcal{T} \subset \mathcal{S} \subset \mathcal{X}$ and $\mathcal{T} \neq \mathcal{S}$;
- a **wrong** model has subset $\mathcal{S}$ lacking some x_r for which $\beta_r \neq 0$, and so $\mathcal{T} \not\subset \mathcal{S}$.

We aim to identify $\mathcal{T}$. If we choose a wrong model, we will have bias, whereas if we choose a correct model, we may increase the variance. We seek to choose a model which balances the bias and variance.

To identify $\mathcal{T}$, we fit a candidate model $Y = X\beta + \epsilon$, where columns of X are a subset $\mathcal{S}$ of those of $X^\dagger$. The fitted values are

$$X\hat{\beta} = X\{(X^TX)^{-1}X^TY\} = HY = H(\mu+\epsilon) = H\mu + H\epsilon,$$

where $H = X(X^TX)^{-1}X^T$ is the **hat matrix** and $H\mu = \mu$ if the model is correct

Following the reasoning for AIC, suppose we also have independent dataset Y_+ from the true model, so $Y_+ = \mu + \epsilon_+$. Apart from constants, previous measure of prediction error is

$$\Delta(X) = n^{-1}EE_+\left\{(Y_+ - X\hat{\beta})^T(Y_+ - X\hat{\beta})\right\},$$

with expectations over both Y_+ and Y.

Theorem 1.2. *We have*

$$\begin{aligned}\Delta(X) &= n^{-1}\mu^T(I-H)\mu + (1+p/n)\sigma^2 \\ &= \begin{cases} n^{-1}\mu^T(I-H)\mu + (1+p/n)\sigma^2 & \textit{if model is wrong,} \\ (1+p_0/n)\sigma^2 & \textit{if model is true,} \\ (1+p/n)\sigma^2 & \textit{if model is correct.} \end{cases}\end{aligned}$$

The **bias** term $n^{-1}\mu^T(I-H)\mu > 0$ unless the model is correct, and is reduced by including useful terms. The **variance** term $(1+p/n)\sigma^2$ is increased by including useless terms. Ideally we would choose covariates X to minimise $\Delta(X)$, but this is impossible, as it depends on unknowns μ, σ. We will have to estimate $\Delta(X)$.

Proof. Consider data $y = \mu + \epsilon$ to which we fit the linear model $y = X\beta + \epsilon$, obtaining fitted values

$$X\hat{\beta} = Hy = H(\mu + \epsilon)$$

where the second term is zero if μ lies in the space spanned by the columns of X, and otherwise is not.

We have a new data set $y_+ = \mu + \epsilon_+$, and we will compute the average error in predicting y_+ using $X\hat{\beta}$, which is

$$\Delta = n^{-1}E\left\{(y_+ - X\hat{\beta})^T(y_+ - X\hat{\beta})\right\}.$$

Now

$$y_+ - X\hat{\beta} = \mu + \epsilon_+ - (H\mu + H\epsilon) = (I - H)\mu + \epsilon_+ - H\epsilon.$$

Therefore

$$(y_+ - X\hat{\beta})^T(y_+ - X\hat{\beta}) = \mu^T(I - H)\mu + \epsilon^T H\epsilon + \epsilon_+^T\epsilon_+ + A$$

where $E(A) = 0$, which gives the result. □

Example 1.3. We consider an example with $n = 20$, $p_0 = 6$, and $\sigma^2 = 1$. In this example, the true model is a a degree five polynomial. Figure 3 shows $\log(\Delta(X))$ for models of increasing polynomial degree, from a quadratic model ($p = 3$) to a degree 14 polynomial ($p = 15$).

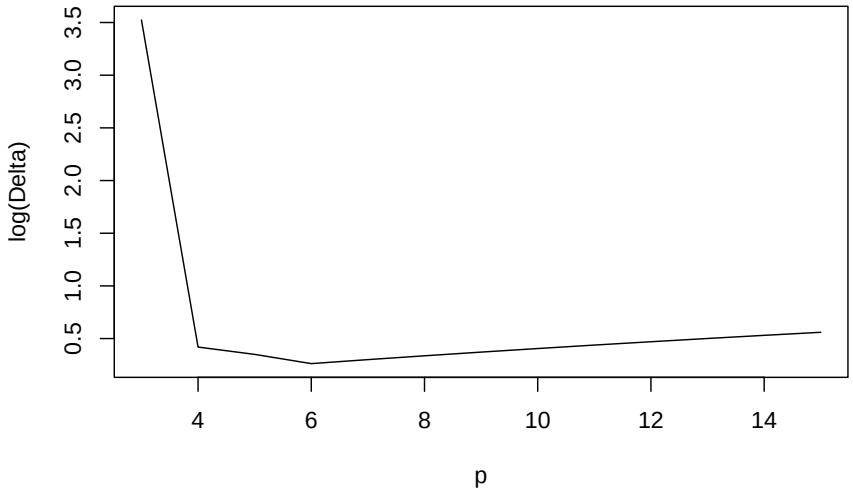

Figure 3: $\log(\Delta(X))$ for models with varying polynomial degree

The minimum of $\Delta(X)$ is at $p = p_0 = 6$: There is a sharp decrease in bias as useful covariates are added, and a slow increase with variance as the number of variables p increases.

If n is large, can split the data into two parts (X', y') and (X^*, y^*), say, and use one part to estimate the model, and the other to compute the prediction error; then choose the model that minimises

$$\hat{\Delta} = \frac{1}{n'}(y' - X'\hat{\beta}^*)^T(y' - X'\hat{\beta}^*) = \frac{1}{n'}\sum_{j=1}^{n'}(y'_j - x'_j\hat{\beta}^*)^2.$$

Usually the dataset is too small for this, so we often use **leave-one-out cross-validation**, which is the sum of squares

$$n\hat{\Delta}_{\mathrm{CV}} = \mathrm{CV} = \sum_{j=1}^{n}(y_j - x_j^T\hat{\beta}_{-j})^2,$$

where $\hat{\beta}_{-j}$ is estimate computed without (x_j, y_j). This seems to require n fits of model, but in fact

$$\mathrm{CV} = \sum_{j=1}^{n}\frac{(y_j - x_j^T\hat{\beta})^2}{(1 - h_{jj})^2},$$

where $h_{11}, \ldots, h_{nn}$ are diagonal elements of H, and so can be obtained from one fit.

A simpler (and often more stable) version uses **generalised cross-validation**, which is the sum of squares

$$\mathrm{GCV} = \sum_{j=1}^{n} \frac{(y_j - x_j^T \hat{\beta})^2}{\{1 - \mathrm{tr}(H)/n\}^2}.$$

Theorem 1.3. *We have*

$$E(GCV) = \mu^T (I-H)\mu/(1-p/n)^2 + n\sigma^2/(1-p/n) \approx n\Delta(X).$$

Proof. We need the expectation of $(y - X\hat{\beta})^T (y - X\hat{\beta})$, where $y - X\hat{\beta} = (I-H)y = (I-H)(\mu + \epsilon)$, and squaring up and noting that $E(\epsilon) = 0$ gives

$$E\left\{(y - X\hat{\beta})^T (y - X\hat{\beta})\right\} = \mu^T (I-H)\mu + E\left\{\epsilon^T (I-H)\epsilon\right\} = \mu^T (I-H)\mu + (n-p)\sigma^2.$$

Now note that $\mathrm{tr}(H) = p$ and divide by $(1-p/n)^2$ to give (almost) the required result, for which we need also $(1-p/n)^{-1} \approx 1 + p/n$, for $p \ll n$. □

We can minimise either GCV or CV. Many variants of cross-validation exist. Typically we find that model chosen based on CV is somewhat unstable, and that GCV or k-fold cross-validation works better. A standard strategy is to split data into 10 roughly equal parts, predict for each part based on the other nine-tenths of the data, then find the model that minimises this estimate of prediction error.

1.4 A Bayesian perspective on model selection

In a parametric model, data y is assumed to be realisation of $Y \sim f(y;\theta)$, where $\theta \in \Omega_\theta$.

Separate from data, we have prior information about parameter θ summarised in a prior density $\pi(\theta)$. The model for the data is $f(y \mid \theta) \equiv f(y;\theta)$ The posterior density for θ is given by Bayes' theorem:

$$\pi(\theta \mid y) = \frac{\pi(\theta) f(y \mid \theta)}{\int \pi(\theta) f(y \mid \theta)\, d\theta}.$$

Here $\pi(\theta \mid y)$ contains all information about θ, conditional on observed data y. If $\theta = (\psi, \lambda)$, then inference for ψ is based on **marginal posterior density**

$$\pi(\psi \mid y) = \int \pi(\theta \mid y)\, d\lambda.$$

Suppose we have M alternative models for the data, with respective parameters $\theta_1 \in \Omega_{\theta_1}, \ldots, \theta_m \in \Omega_{\theta_m}$. Typically the dimensions of Ω_{θ_m} are different.

We enlarge the parameter space to give an **encompassing model** with parameter

$$\theta = (m, \theta_m) \in \Omega = \bigcup_{m=1}^{M} \{m\} \times \Omega_{\theta_m}.$$

Thus we need priors $\pi_m(\theta_m \mid m)$ for the parameters of each model, plus a prior $\pi(m)$ giving pre-data probabilities for each of the models. Overall, we have

$$\pi(m, \theta_m) = \pi(\theta_m \mid m)\pi(m) = \pi_m(\theta_m)\pi_m,$$

say.

Inference about model choice is based on marginal posterior density

$$\pi(m \mid y) = \frac{\int f(y \mid \theta_m)\pi_m(\theta_m)\pi_m\, d\theta_m}{\sum_{m'=1}^{M} \int f(y \mid \theta_{m'})\pi_{m'}(\theta_{m'})\pi_{m'}\, d\theta_{m'}} = \frac{\pi_m f(y \mid m)}{\sum_{m'=1}^{M} \pi_{m'} f(y \mid m')}.$$

We can write

$$\pi(m, \theta_m \mid y) = \pi(\theta_m \mid y, m)\pi(m \mid y),$$

so Bayesian updating corresponds to

$$\pi(\theta_m \mid m)\pi(m) \mapsto \pi(\theta_m \mid y, m)\pi(m \mid y)$$

and for each model $m = 1, \ldots, M$ we need

- the posterior probability $\pi(m \mid y)$, which involves the marginal likelihood $f(y \mid m) = \int f(y \mid \theta_m, m)\pi(\theta_m \mid m)\, d\theta_m$; and
- the posterior density $f(\theta_m \mid y, m)$.

If there are just two models, can write

$$\frac{\pi(1 \mid y)}{\pi(2 \mid y)} = \frac{\pi_1}{\pi_2}\frac{f(y \mid 1)}{f(y \mid 2)},$$

so the posterior odds on model 1 equal the prior odds on model 1 multiplied by the **Bayes factor** $B_{12} = f(y \mid 1)/f(y \mid 2)$.

Suppose the prior for each θ_m is $N(0, \sigma^2 I_{d_m})$, where $d_m = \dim(\theta_m)$. Then, dropping the m subscript for clarity,

$$\begin{aligned} f(y \mid m) &= \sigma^{-d/2}(2\pi)^{-d/2} \int f(y \mid m, \theta) \prod_r \exp\left\{-\theta_r^2/(2\sigma^2)\right\} d\theta_r \\ &\approx \sigma^{-d/2}(2\pi)^{-d/2} \int f(y \mid m, \theta) \prod_r d\theta_r, \end{aligned}$$

for a highly diffuse prior distribution (large σ^2).

The Bayes factor for comparing the models is approximately

$$\frac{f(y \mid 1)}{f(y \mid 2)} \approx \sigma^{(d_2 - d_1)/2} g(y),$$

where $g(y)$ depends on the two likelihoods but is independent of σ^2. Hence, *whatever the data tell us about the relative merits of the two models*, the Bayes factor in favour of the simpler model can be made arbitrarily large by increasing σ.

This illustrates **Lindley's paradox**, and implies that we must be careful when specifying prior dispersion parameters to compare models.

If a quantity Z has the same interpretation for all models, it may be necessary to allow for model uncertainty. In prediction, each model may be just a vehicle that provides a future value, not of interest *per se*.

The predictive distribution for Z may be written

$$f(z \mid y) = \sum_{m=1}^{M} f(z \mid y, m) P(m \mid y)$$

where

$$P(m \mid y) = \frac{f(y \mid m)P(m)}{\sum_{m'=1}^{M} f(y \mid m')P(m')}.$$

2 Beyond Generalised Linear Models

2.1 Generalised Linear Models

$y_1, \ldots, y_n$ 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 $\mu_i \equiv E(Y_i)$ linked to explanatory variables $X_1, X_2, \ldots, X_p$ through

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

GLMs have proved remarkably effective at modelling real world variation in a wide range of application areas. 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.
- The assumption of independence is not appropriate.
- Complex multivariate structure in the data requires a more flexible model class

2.2 Overdispersion

2.2.1 An example of overdispersion

Example 2.1. The dataset `toxo` in `SMPracticals` provides data on the number of people testing positive for toxoplasmosis (`r`) out of the number of people tested (`m`) in 34 cities in El Salvador, along with the annual rainfall in mm (`rain`) in those cities.

We can fit various logistic regression models for relating toxoplasmosis incidence to rainfall. If we consider logistic models with a polynomial dependence on rainfall, AIC and stepwise selection methods both prefer a cubic model. For simplicity here, we compare the cubic model and a constant model, in which there is no dependence on rainfall.

```
mod_const <- glm(r/m ~ 1, data = toxo, weights = m,
                 family = "binomial")
mod_cubic <- glm(r/m ~ poly(rain, 3), data = toxo, weights = m,
                 family = "binomial")
```

Figure 4 shows the fitted proportions testing positive under the two models.

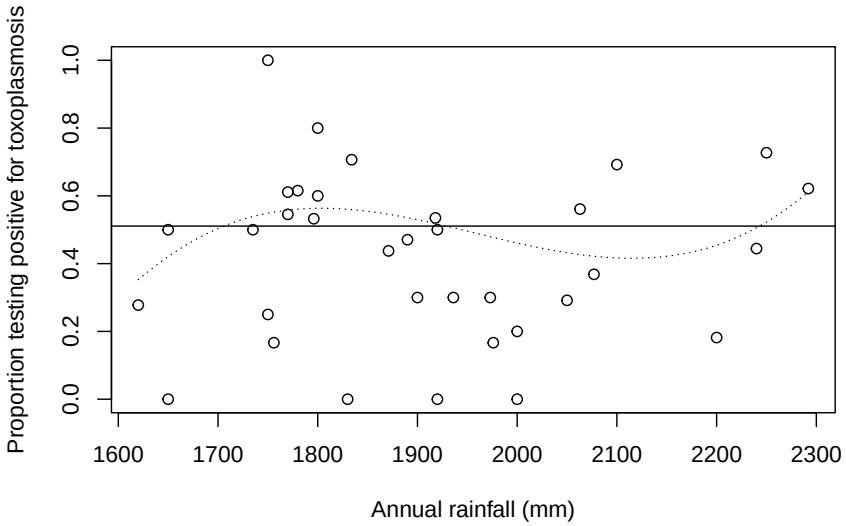

Figure 4: Propotion of people testing positive for toxoplasmosis against rainfall, with fitted proportions under constant (solid line) and cubic (dotted line) logistic regression models

We can conduct a hypothesis test to compare the models:

```
anova(mod_const, mod_cubic, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: r/m ~ 1
## Model 2: r/m ~ poly(rain, 3)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        33     74.212
## 2        30     62.635  3   11.577 0.008981 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is apparently evidence to reject the null model (that there is no effect of rain on the probability of testing positive for toxoplasmosis) in favour of the cubic model.

However, we find that the residual deviance for the cubic model (62.63) is much larger than the residual degrees of freedom (30). This is an indicator of **overdispersion**, where the residual variability is greater than would be predicted by the specified mean/variance relationship

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

2.2.2 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 $g(\mu_i) = \eta_i = x_i^T\beta$, and variance $\text{var}(Y_i) = \sigma^2 V(\mu_i)/m_i$, the quasi-likelihood equations are

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

If $V(\mu_i)/m_i$ represents $\text{var}(Y_i)$ 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 $\hat{\beta}$ still holds.

The dispersion parameter σ^2 can be estimated using

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

Example 2.2. Fitting the same models as before, but with $\text{var}(Y_i) = \sigma^2\mu_i(1-\mu_i)/m_i$, we get

```
mod_const_quasi <- glm(r/m ~ 1, data = toxo, weights = m,
                       family = "quasibinomial")
mod_cubic_quasi <- glm(r/m ~ poly(rain, 3), data = toxo, weights = m,
                       family = "quasibinomial")
```

We find the estimates of the β coefficients are the same as before, but now we estimate σ^2 as 1.94 under the cubic model.

Comparing the cubic with the constant model, we now obtain

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

```
anova(mod_const_quasi, mod_cubic_quasi, test = "F")
```

```
## Analysis of Deviance Table
##
## Model 1: r/m ~ 1
## Model 2: r/m ~ poly(rain, 3)
##   Resid. Df Resid. Dev Df Deviance      F Pr(>F)
## 1        33     74.212
## 2        30     62.635  3   11.577 1.9888 0.1369
```

After accounting for overdispersion, there is much less compelling evidence in favour of an effect of rainfall on toxoplasmosis incidence.

2.2.3 Models for overdispersion

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*.

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

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

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}

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

where G is the distribution induced by F. Then

$$E(Y_i) = E_G[E(Y_i \mid \mu_i)] = E_G(\mu_i)$$
$$\text{var}(Y_i) = E_G(V(\mu_i)/m_i) + \text{var}_G(\mu_i)$$

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

$$\mu_i \sim \text{Beta}(k\mu_i^*, k[1-\mu_i^*])$$

leading to

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

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

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

$$\eta_i = x_i^T \beta + u_i \tag{2}$$

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

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

$$\begin{aligned} f(y \mid \beta, \sigma^2) &= \int f(y \mid \beta, u, \sigma^2) f(u \mid \beta, \sigma^2) du \\ &= \int f(y \mid \beta, u) f(u \mid \sigma^2) du \\ &= \int \prod_{i=1}^{n} f(y_i \mid \beta, u_i) f(u_i \mid \sigma^2) du_i \end{aligned} \tag{3}$$

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

Often no further simplification of (3) is possible, so computation needs careful consideration – we will come back to this later.

2.3 Dependence

2.3.1 Toxoplasmosis example revisited

Example 2.3. 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

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

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

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}.

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

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

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.

2.3.2 Marginal models and generalised estimating equations

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

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

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

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').

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

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

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.

2.3.3 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 2.4. The `rat.growth` data in `SMPracticals` gives the weekly weights (`y`) of 30 young rats. Figure 5 shows the weight against week separately for each rat.

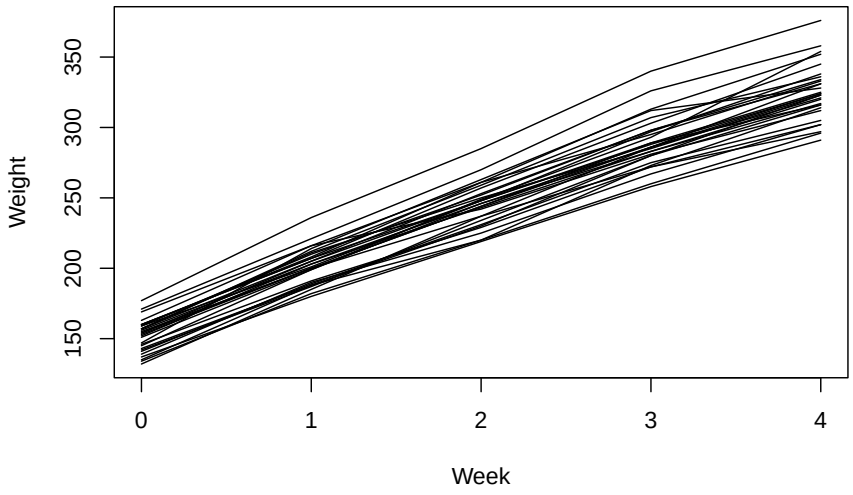

Figure 5: Individual rat weight by week, for the rat growth data

Writing y_{ij} for the jth observation of the weight of rat i, and x_{ij} for the week in which this record was made, we can fit the simple linear regression

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

with resulting estimates $\hat{\beta}_0 = 156.1$ (2.25) and $\hat{\beta}_1 = 43.3$ (0.92). Figure 6 shows the residuals from this model, separately for each rat, showing clear evidence of an unexplained difference between rats.

2.4 Linear mixed models

2.4.1 Model statement

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

$$Y \sim N(\mu, \Sigma), \quad \mu = X\beta + Zb, \quad b \sim N(0, \Sigma_b), \tag{4}$$

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

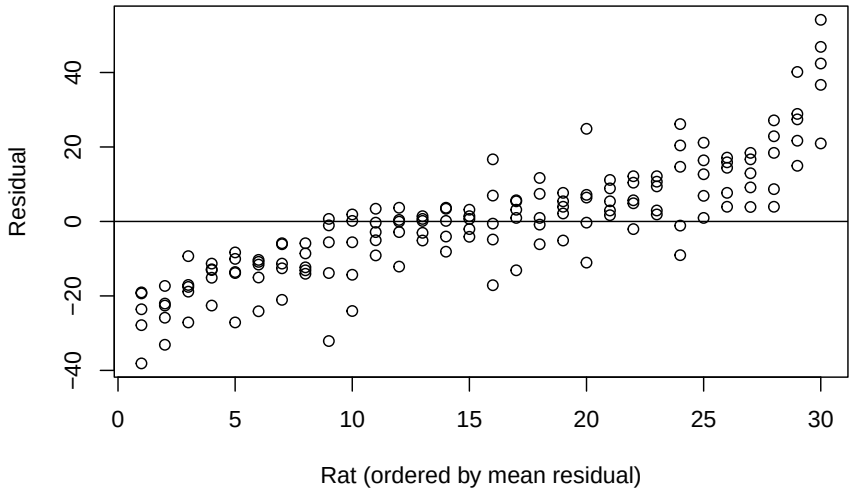

Figure 6: Residuals from a simple linear regression for each rat in the rat.growth data

A typical example for clustered data might be

$$Y_{ij} \sim N(\mu_{ij}, \sigma^2), \quad \mu_{ij} = x_{ij}^T\beta + z_{ij}^T b_i, \quad b_i \sim N(0, \Sigma_b^*), \tag{5}$$

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 (2).

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

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

This fits into the general LMM framework (4) 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.

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

$$Y \sim N(X\beta, \Sigma + Z\Sigma_b Z^T) \tag{6}$$

where X and Z are matrices containing values of explanatory variables.

Hence $\text{var}(Y)$ is comprised of two **variance components**. Other ways of describing LMMs for clustered data, such as (5) (and their generalised linear model counterparts) are known 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.

2.4.2 Discussion: why random effects?

Instead of including random effects for clusters, e.g.

$$y_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})x_{ij} + \epsilon_{ij},$$

we could use separate fixed effects for each cluster, e.g.

$$y_{ij} = \beta_{0i} + \beta_{1i}x_{ij} + \epsilon_{ij}.$$

However, inferences can then only be made about those clusters present in the observed data. Random effects models allow inferences to be extended to a wider population. It also can be the case, as in the original toxoplasmosis example with only one observation per 'cluster', that fixed effects are not identifiable, whereas random effects can still be estimated. Random effects also allow 'borrowing strength' across clusters by shrinking fixed effects towards a common mean.

2.4.3 LMM fitting

The likelihood for $(\beta, \Sigma, \Sigma_b)$ is available directly from (6) as

$$f(y \mid \beta, \Sigma, \Sigma_b) \propto |V|^{-1/2} \exp\left(-\frac{1}{2}(y - X\beta)^T V^{-1}(y - X\beta)\right) \tag{7}$$

where $V = \Sigma + Z\Sigma_b Z^T$. 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 $E(Y) = X\beta$.

In effect, REML maximizes the likelihood of any linearly independent sub-vector of $(I_n - H)y$ where $H = X(X^TX)^{-1}X^T$ 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 (7) as

$$f(r \mid \Sigma, \Sigma_b) \propto f(y \mid \hat{\beta}, \Sigma, \Sigma_b)|X^TVX|^{1/2} \tag{8}$$

where

$$\hat{\beta} = (X^TV^{-1}X)^{-1}X^TV^{-1}y \tag{9}$$

is the usual generalised least squares estimator given known V.

Having first obtained $(\hat{\Sigma}, \hat{\Sigma}_b)$ by maximising (8), $\hat{\beta}$ is obtained by plugging the resulting $\hat{V}$ into (9).

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.

2.4.4 Estimating random effects

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

This is achieved by

$$\tilde{b} = E(b \mid y) = (Z^T\Sigma^{-1}Z + \Sigma_b^{-1})^{-1}Z^T\Sigma^{-1}(y - X\beta) \tag{10}$$

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

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

The estimates are obtained by plugging in $(\hat{\beta}, \hat{\Sigma}, \hat{\Sigma}_b)$, 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 $\text{var}(b_k \mid y) = [\Sigma_b]_{kk}$, which may be estimated if, as is usual, b_k shares a variance with other random effects.

2.4.5 Rat growth revisited

Example 2.5. Here, we consider the model

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

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

We may fit the model in R by using the `lme4` package:

```
library(lme4)
rat_rs <- lmer(y ~ week + (week | rat), data = rat.growth)
rat_rs
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ week + (week | rat)
##    Data: rat.growth
## REML criterion at convergence: 1084.58
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  rat      (Intercept) 10.933
##           week         3.535   0.18
##  Residual              5.817
## Number of obs: 150, groups:  rat, 30
## Fixed Effects:
## (Intercept)         week
##      156.05        43.27
```

We could also consider the simpler random intercept model

$$y_{ij} = \beta_0 + b_{0i} + \beta_1 x_{ij} + \epsilon_{ij}, \quad b_{0i} \sim N(0, \sigma_b^2).$$

```
rat_ri <- lmer(y ~ week + (1 | rat), data = rat.growth)
rat_ri
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ week + (1 | rat)
##    Data: rat.growth
## REML criterion at convergence: 1127.169
## Random effects:
##  Groups   Name        Std.Dev.
##  rat      (Intercept) 13.851
##  Residual              8.018
## Number of obs: 150, groups:  rat, 30
## Fixed Effects:
## (Intercept)         week
##      156.05        43.27
```

We might compare the models with AIC or BIC, but in order to do so we need to refit the models with maximum likelihood rather than REML.

```
rat_rs_ML <- lmer(y ~ week + (week | rat), data = rat.growth, REML = FALSE)
rat_ri_ML <- lmer(y ~ week + (1 | rat), data = rat.growth, REML = FALSE)
c(AIC(rat_rs_ML), AIC(rat_ri_ML))
```

```
## [1] 1101.124 1139.204
```

```
c(BIC(rat_rs_ML), BIC(rat_ri_ML))
```

```
## [1] 1119.188 1151.246
```

By either measure, we prefer the random slopes model.

An alternative fixed effects model would be to fit a model with separate intercepts and slopes for each rat

$$y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \epsilon_{ij}.$$

Figure 7 shows parameter estimates from the random effects model against those from the fixed effects model, demonstrating shrinkage of the random effect estimates towards a common mean. Random effects estimates 'borrow strength' across clusters, due to the Σ_b^{-1} term in (10). The extent of this is determined by cluster similarity.

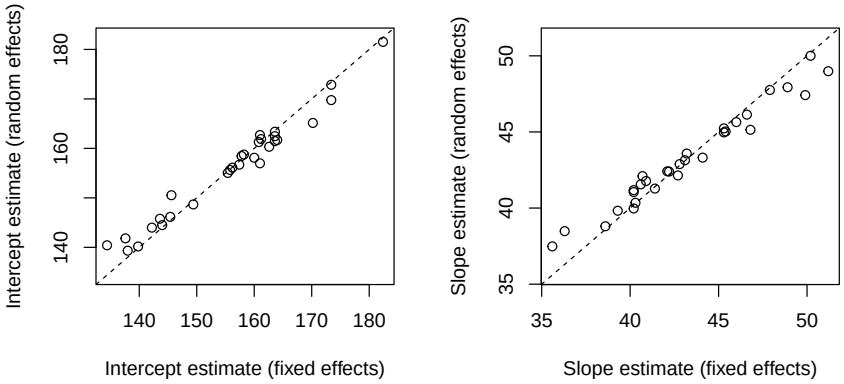

Figure 7: Parameter estimates from the random effects model against those from the fixed effects model for the rat growth data

2.4.6 Bayesian inference: the Gibbs sampler

Bayesian inference for 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 multivariate) the Gibbs sampler starts from an arbitrary value of y and updates components (sequentially or otherwise) by generating from the conditional distributions $f(y_i \mid y_{\,i})$ where $y_{\,i}$ 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.

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

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

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.

2.5 Generalised linear mixed models

2.5.1 Model setup

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

$$Y_i \sim F(\cdot \mid \mu_i, \sigma^2), \quad g(\mu) \equiv \begin{pmatrix} g(\mu_1) \\ \vdots \\ g(\mu_n) \end{pmatrix} = X\beta + Zb, \;\; b \sim N(0, \Sigma_b) \tag{11}$$

where $F(\cdot \mid \mu_i, \sigma^2)$ is an exponential family distribution with $E(Y) = \mu$ 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.

Example 2.6. 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

$$Y_{ij} \sim \text{Bernoulli}(\mu_{ij}), \quad \log \frac{\mu_{ij}}{1-\mu_{ij}} = \beta_0 + b_{0i} + \beta_1 x_{ij}, \quad b_{0i} \sim N(0, \sigma_b^2) \tag{12}$$

Note there is no random slope here. This fits into the general GLMM framework (11) with

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

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

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

2.5.2 GLMM likelihood

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

$$\begin{aligned} f(y \mid \beta, \Sigma_b) &= \int f(y \mid \beta, b, \Sigma_b) f(b \mid \beta, \Sigma_b) db \\ &= \int f(y \mid \beta, b) f(b \mid \Sigma_b) db \\ &= \int \prod_{i=1}^{n} f\left(y_i \mid g^{-1}([X\beta + Zb]_i)\right) f(b \mid \Sigma_b) db. \end{aligned} \tag{13}$$

For *nested* random effects structures, some simplification is possible. For example, for (12)

$$f(y \mid \beta, \sigma_b^2) = \prod_{i=1}^{k} \int \prod_j f\left(y_{ij} \mid g^{-1}(x_i^T\beta + b_i)\right) \phi(b_i \mid 0, \sigma_b^2) db_i,$$

a product of one-dimensional integrals.

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

$$\int \prod_j f\left(y_{ij} \mid g^{-1}(x_i^T\beta + b_i)\right) \phi(b_i \mid 0, \sigma_b^2) db_i \approx \sum_{q=1}^{Q} w_q \prod_j f\left(y_{ij} \mid g^{-1}(x_i^T\beta + b_{iq})\right)$$

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.

An alternative approach is to use a Laplace approximation to the likelihood. Writing

$$h(b) = \prod_{i=1}^{n} f\left(y_i \mid g^{-1}([X\beta + Zb]_i)\right) f(b \mid \Sigma_b)$$

for the integrand of the likelihood, a (first-order) Laplace approximation approximates $h(.)$ as an unnormalised multivariate normal density function

$$\tilde{h}(b) = c\, \phi_k(b; \hat{b}, V),$$

where

- $\hat{b}$ is found by maximizing $\log h(.)$ over b.
- the variance matrix V is chosen so that the curvature of $\log h(.)$ and $\log \tilde{h}(.)$ agree at $\hat{b}$.
- c is chosen so that $\tilde{h}(\hat{b}) = h(\hat{b})$.

The first-order Laplace approximation is equivalent to adaptive Gaussian quadrature with a single quadrature point. Quadrature provides accurate approximations to the likelihood. For some model structures, particularly those with crossed rather than nested random effects, the likelihood integral may be high-dimensional, and it may not be possible to use quadrature. In such cases, a Laplace approximation is often sufficiently accurate for most purposes, but this is not guaranteed.

Another alternative is to use Penalized Quasi Likelihood (PQL) for inference, which is very fast but often inaccurate. PQL can fail badly in some cases, particularly with binary observations, and its use is not recommended. Likelihood inference for GLMMs remains an area of active research and vigorous debate.

Example 2.7. For the individual-level model

$$Y_{ij} \sim \text{Bernoulli}(\mu_i), \quad \log \frac{\mu_i}{1-\mu_i} = \beta_0 + b_{0i} + \beta_1 x_i, \quad b_{0i} \sim N(0, \sigma_b^2), \tag{14}$$

the estimates and standard errors obtained by ML (quadrature), Laplace and PQL are shown in Table 1. For the extended model

$$\log \frac{\mu_i}{1-\mu_i} = \beta_0 + b_{0i} + \beta_1 x_{ij} + \beta_1 x_{ij}^2 + \beta_1 x_{ij}^3, \tag{15}$$

the estimates and standard errors are shown in Table 2. For this example, there is a good agreement between the different computational methods.

Table 1: Estimates (with standard errors in brackets) obtained by various approximations to the likelihood for model (14).

Parameter	ML	Laplace	PQL
β_0	-0.1384 (1.452)	-0.1343 (1.440)	-0.115 (1.445)
$\beta_1(\times 10^6)$	7.215 (752)	5.930 (745.7)	0.57 (749.2)
σ_b	0.5209	0.5132	0.4946
AIC	65.75	65.96	—

Table 2: Estimates (with standard errors in brackets) obtained by various approximations to the likelihood for model (15).

Parameter	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)
$\beta_2(\times 10^4)$	-2.710 (1.094)	-2.706 (1.086)	-3 (1.1)
$\beta_3(\times 10^8)$	4.643 (1.866)	4.636 (1.852)	0 (0)
σ_b	0.4232	0.4171	0.4315
AIC	63.84	63.97	—

2.5.3 Bayesian inference for GLMMs

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

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

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

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

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.

3 Nonlinear models

3.1 Basic nonlinear models

So far we have only considered models where the link function of the mean response is equal to the linear predictor, i.e. in the most general case of the generalised linear mixed model (GLMM)

$$\mu_{ij} = E(y_{ij}), \qquad g(\mu_{ij}) = \eta_{ij} = x_{ij}^T \beta + z_{ij}^T b_i,$$

and where the response distribution for y is from the exponential family of distributions. The key point is that the linear predictor is a linear function of the parameters. Linear models, generalised linear models (GLMs) and linear mixed models (LMMs) are all special cases of the GLMM.

These "linear" models form the basis of most applied statistical analyses. Usually, there is no scientific reason to believe these linear models are true for a given application.

We begin by considering nonlinear extensions of the normal linear model

$$y_i = x_i^T \beta + \epsilon_i, \tag{16}$$

where $\epsilon_i \sim N(0, \sigma^2)$, independently, where β are the p regression parameters. Instead of the mean response being the linear predictor $x_i^T \beta$, we could allow it to be a nonlinear function of parameters, i.e.

$$y_i = \eta(x_i, \beta) + \epsilon_i, \tag{17}$$

where $\epsilon_i \sim N(0, \sigma^2)$, independently, where β are the p nonlinear parameters. The model specified by (17) has the linear model (16) as a special case when $\eta(x, \beta) = x^T \beta$.

Nonlinear parameters can be of two different types:

- **Physical parameters** have meaning within the science underlying the model, $\eta(x, \beta)$. Estimating the value of physical parameters contributes to scientific understanding.
- **Tuning parameters** do not have physical meaning. Their presence is often as a simplification of a more complex underlying system. Their estimation is to make the model fit best to reality.

How might the function $\eta(x, \beta)$ be specified?

- **Mechanistically** – prior scientific knowledge is incorporated into building a mathematical model for the mean response. This can often be complex and $\eta(x, \beta)$ may not be available in closed form.
- **Phenomenologically (empirically)** – a function $\eta(x, \beta)$ may be posited that appears to capture the non-linear nature of the mean response.

Example 3.1. The response, y, is the uptake of calcium (in nmoles per mg) at time x (in minutes) by $n = 27$ cells in "hot" suspension. Figure 8 shows calcium uptake against time.

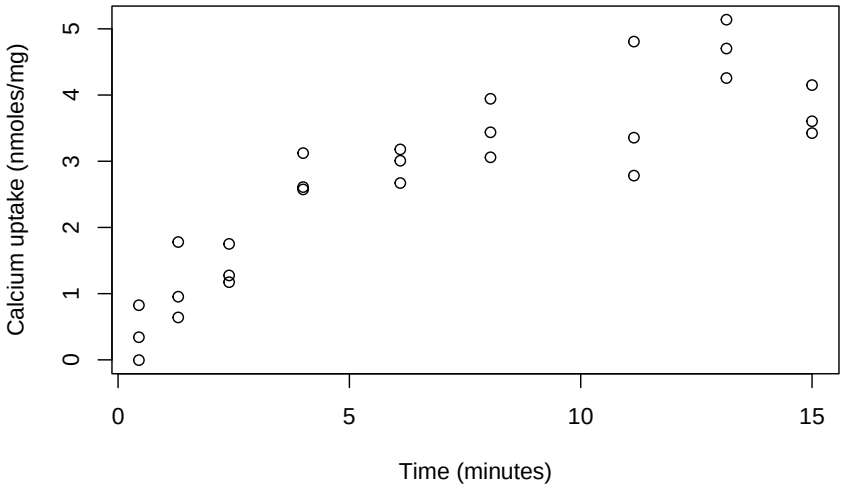

Figure 8: Calcium uptake against time

We see that calcium uptake "grows" with time. There is a large class of phenomenological models for growth curves. Consider the non-linear model with

$$\eta(x, \beta) = \beta_0 \left(1 - \exp\left(-x/\beta_1\right)\right). \tag{18}$$

This is derived by assuming that the rate of growth is proportional to the calcium remaining, i.e.

$$\frac{d\eta}{dx} = (\beta_0 - \eta)/\beta_1.$$

The solution (with initial condition $\eta(0, \beta) = 0$) to this differential equation is (18). Here β_0 is the final size of the population, and β_1 (inversely) controls the growth rate.

Figure 9 shows fitted lines for the three different models added to the plot of calcium uptake against time.

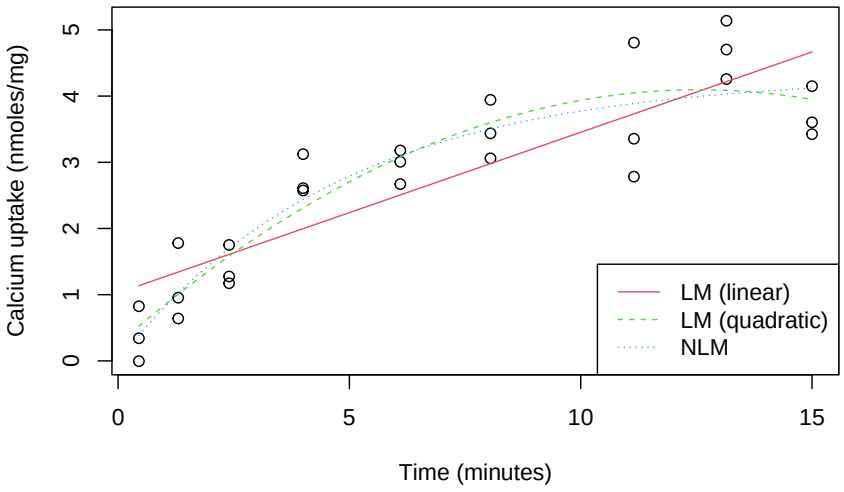

Figure 9: Calcium uptake against time, with expected uptake from three models overlaid

A comparison of the goodness-of-fit for the three models is shown in Table 3. The goodness-of-fit for the quadratic and nonlinear models is identical (to 2 decimal places). Since the nonlinear model is simpler (fewer parameters), it is the preferred model.

Table 3: A comparison of the goodness-of-fit of three models for the calcium data.

Model	Parameters (p)	$l(\hat{\beta})$	AIC
Linear model (slope)	2	-28.70	63.40
Linear model (quadratic)	3	-20.95	49.91
Non-linear model	2	-20.95	47.91

3.2 Extending the nonlinear model

3.2.1 Introduction

Nonlinear models can be extended to non-normal responses and clustered responses in the same way as linear models. Here, we consider clustered responses and briefly discuss the nonlinear mixed model.

Example 3.2. Theophylline is an anti-asthmatic drug. An experiment was performed on $n = 12$ individuals to investigate the way in which the drug leaves the body. The study of drug concentrations inside organisms is called *pharmacokinetics*. An oral dose, D_i, was given to the ith individual at time $t = 0$, for $i = 1, \ldots, n$. The concentration of theophylline in the blood was then measured at 11 time points in the next 25 hours. Let y_{ij} be the theophylline concentration (mg/L) for individual i at time t_{ij}. Figure 10 shows the concentration of theophylline against time for each of the individuals. There is a sharp increase in concentration followed by a steady decrease.

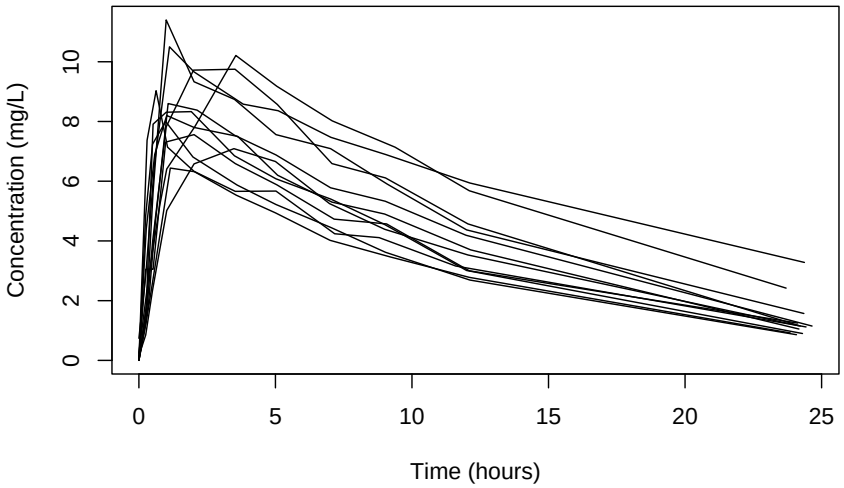

Figure 10: Concentration of theophylline against time for each of the individuals in the study

Compartmental models are a common class of model used in pharmacokinetics studies. If the initial dosage is D, then a two-compartment open pharmacokinetic model is

$$\eta(\beta, D, t) = \frac{D\beta_1\beta_2}{\beta_3(\beta_2 - \beta_1)} \left(\exp\left(-\beta_1 t\right) - \exp\left(-\beta_2 t\right)\right),$$

where the (positive) nonlinear parameters are

- β_1, the elimination rate which controls the rate at which the drug leaves the organism;
- β_2, the absorption rate which controls the rate at which the drug enters the blood;
- β_3, the clearance which controls the volume of blood for which a drug is completely removed per time unit.

Initially ignore the dependence induced from repeated measurements on individuals and assume the following basic nonlinear model

$$y_{ij} = \eta(\beta, D_i, t_{ij}) + \epsilon_{ij},$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$. Figure 11 gives a plot of the residuals for each subject in the study under this model, showing evidence of an unexplained difference between individuals.

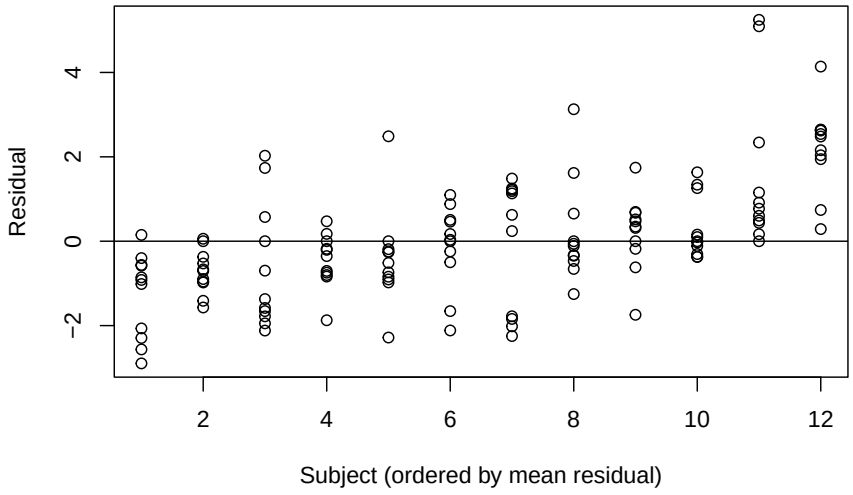

Figure 11: Residuals for each individual in the theopylline study, assuming a basic nonlinear model which ignores dependence

3.2.2 Nonlinear mixed effects models

A nonlinear mixed model is

$$y_{ij} = \eta(\beta + b_i, x_{ij}) + \epsilon_{ij},$$

where

$$\epsilon_{ij} \sim N(0, \sigma^2), \qquad b_i \sim N(0, \Sigma_b),$$

and Σ_b is a $q \times q$ covariance matrix.

This model specifies that $\beta_i = \beta + b_i$ are the nonlinear parameters for the ith cluster, i.e. the cluster-specific nonlinear parameters. In the case of the Theophylline example, each individual would have unique elimination rate, absorption rate and clearance. It follows that $\beta_i \sim N(\beta, \Sigma_b)$. The mean, β, of the cluster-specific nonlinear parameters across all individuals are the population nonlinear parameters.

We might like to specify the model in a way such that only a subset of the nonlinear parameters can be different for each individual, and the remainder fixed for all individuals.

Suppose $q \leq p$ nonlinear parameters are can be different for each individual, then a more general way of writing the nonlinear mixed model is

$$y_{ij} = \eta(\beta + Ab_i, x) + \epsilon_{ij},$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ and $b_i \sim N(0, \Sigma_b)$. Here Σ_b is a $q \times q$ covariance matrix and A is a $p \times q$ binary matrix. A allows the specification of the fixed and varying nonlinear parameters.

The linear mixed model is a special case of the nonlinear mixed model where

$$\eta(\beta, x) = x^T \beta.$$

Then

$$\eta(\beta + Ab, x) = x^T (\beta + Ab) = x^T \beta + x^T Ab,$$

so $z = A^T x$. For a random intercept model, where $q = 1$, $A = (1, 0, \ldots, 0)$.

Example 3.3. We fit the nonlinear mixed model, allowing all of the nonlinear parameters to vary across individuals, i.e. $A = I_3$.

Estimates:

$$\begin{array}{rclrcl} \hat{\beta}_1 & = & 0.0864 & \hat{\Sigma}_{b11} & = & 0.0166 \\ \hat{\beta}_2 & = & 1.6067 & \hat{\Sigma}_{b22} & = & 0.9349 \\ \hat{\beta}_3 & = & 0.0399 & \hat{\Sigma}_{b33} & = & 0.0491 \end{array}$$

We have AIC = 372.6.

The estimated value of Σ_{b11} is "small" so we fit the nonlinear mixed model, allowing absorption rate and clearance to vary across individuals, i.e.

$$A = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Estimates:

$$\begin{array}{lllllll} \hat{\beta}_1 & = & 0.0859 & & & & \\ \hat{\beta}_2 & = & 1.6032 & \hat{\Sigma}_{b22} & = & 0.6147 \\ \hat{\beta}_3 & = & 0.0397 & \hat{\Sigma}_{b33} & = & 0.0284 \end{array}$$

We have AIC = 368.6. No further model simplifications reduce the AIC.

3.2.3 Extensions to non-normal responses

Nonlinear models can be extended to non-normal responses in the same way as linear models. The most general model is the generalised nonlinear mixed model (GNLMM), which assumes y_{ij} is from exponential family,

$$E(y_{ij}) = \mu_{ij}, \qquad g(\mu_{ij}) = \eta(\beta + Ab_i, x_{ij}).$$

This model has the following special cases:

linear model	nonlinear model
linear mixed model	nonlinear mixed model
generalised linear model	generalised nonlinear model
generalised linear mixed model	

There are various technical and practical issues related to fitting nonlinear models (some are common to GLMs and GLMMs). For instance:

- we need to use some approximation of likelihood function (since random effects are integrated out),
- sometimes optimisation routines to find estimates do not converge to a global maximum of the likelihood,
- evaluating $\eta(\beta, x)$ is sometimes computationally expensive.

These are all areas of current research.