APTS Statistical Modelling

Chapter 1: Model Selection

Helen Ogden

April 2021

Why do we model?

> *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

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.

Comparing 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: logistic regression

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 $\hat{\ell} = \ell(\hat{\beta})$ where $\hat{\beta}$ is the MLE under the model.

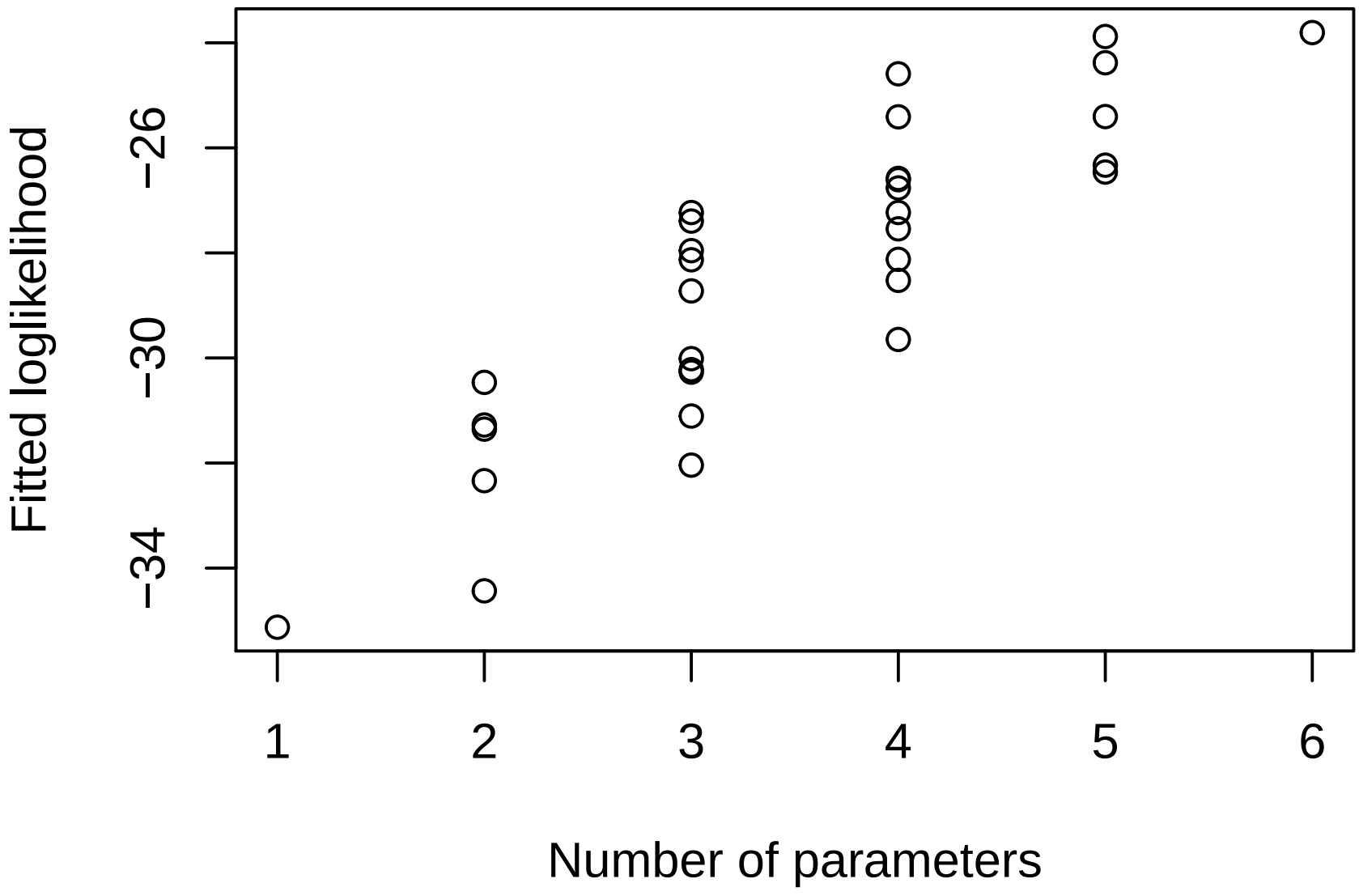
Comparing models for the nodal involvement data

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 to fitted loglikelihood $\hat{\ell}$, and summarise this by plotting the number of parameters against the loglikelihood for each of the 32 models under consideration.

Comparing models by fitted loglikelihood



Comparing models by fitted loglikelihood

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).

Kullback–Leibler discrepancy

Given (unknown) **true model** $g(y)$, and **candidate model** $f(y;\theta)$, the Kullback–Leibler discrepancy is

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

Jensen's inequality implies that

$$\int \log g(y) g(y)\,dy \geq \int \log f(y;\theta) g(y)\,dy, \tag{1}$$

i.e. $KL(f_\theta, g) \geq 0$.

Kullback–Leibler discrepancy for model choice

If θ_g is the value of θ that maximizes the expected log likelihood on the right of (1), then it seems natural to choose the candidate model that maximises

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

which should be an estimate of $\int \log f(y; \theta_g) g(y)\, dy$.

However as $\bar{\ell}(\hat{\theta}) \geq \bar{\ell}(\theta_g)$, by definition of $\hat{\theta}$, this estimate is biased upwards.

We need to correct for the bias, but in order to do so, we first need to understand the properties of likelihood estimators when the assumed model f is not the true model g.

Likelihood inference under the wrong model

Suppose the true model is g, that is, $Y_1, \ldots, Y_n \sim g$, but we assume that $Y_1, \ldots, Y_n \sim f(y; \theta)$. 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 minimizes the Kullback–Leibler discrepancy

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

Asymptotic distribution of the MLE

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

$$\hat{\theta} \;\sim\; N_p\left\{\theta_g, I(\theta_g)^{-1} K(\theta_g) I(\theta_g)^{-1}\right\}, \tag{2}$$

where

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

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

Asymptotic distribution of the MLE

If $g(y) = f(y;\theta)$, so that the supposed density is correct, then θ_g is the true θ, then $K(\theta_g) = I(\theta)$, and (2) reduces to the usual approximation.

In practice $g(y)$ is 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 (2) with variance matrix $\hat{J}^{-1}\hat{K}\hat{J}^{-1}$.

Asymptotic distribution of the likelihood ratio statistic

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 $E\{W(\theta_g)\} = \text{tr}\{I(\theta_g)^{-1} K(\theta_g)\}$.

Bias in $\bar{\ell}(\hat{\theta})$

We have

$$\begin{aligned} E_g\left\{\bar{\ell}(\hat{\theta})\right\} &= E_g\left\{\bar{\ell}(\theta_g)\right\} + E\left\{\bar{\ell}(\hat{\theta}) - \bar{\ell}(\theta_g)\right\} \\ &= E_g\left\{\bar{\ell}(\theta_g)\right\} + \frac{1}{2n} E\left\{W(\theta_g)\right\}, \\ &\approx E_g\left\{\bar{\ell}(\theta_g)\right\} + \frac{1}{2n}\,\mathrm{tr}\{I(\theta_g)^{-1}K(\theta_g)\}, \end{aligned}$$

where E_g denotes expectation over the data distribution g. The bias is positive because I and K are positive definite matrices.

Emulating a training set

We need to fix two problems with using $\bar{\ell}(\hat{\theta})$ to choose the best candidate model:

- upward bias, as $\bar{\ell}(\hat{\theta}) \geq \bar{\ell}(\theta_g)$ because $\hat{\theta}$ is based on $Y_1, \ldots, Y_n$;
- no penalisation if the dimension of θ increases.

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 both problems disappear, 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 $\hat{\theta}$.

Emulating a training set

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

$$\Delta \approx \int \log f(y;\theta_g) g(y)\, dy - \frac{1}{2n} \operatorname{tr}\{I(\theta_g)^{-1} K(\theta_g)\},$$

where the second term is a penalty that depends on the model dimension.

Recall we have

$$E_g\left\{\bar{\ell}(\hat{\theta})\right\} \approx \int \log f(y;\theta_g) g(y)\, dy + \frac{1}{2n} \operatorname{tr}\{I(\theta_g)^{-1} K(\theta_g)\}.$$

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

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

Network Information Criterion

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

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

or equivalently 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).

Other information criteria

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:

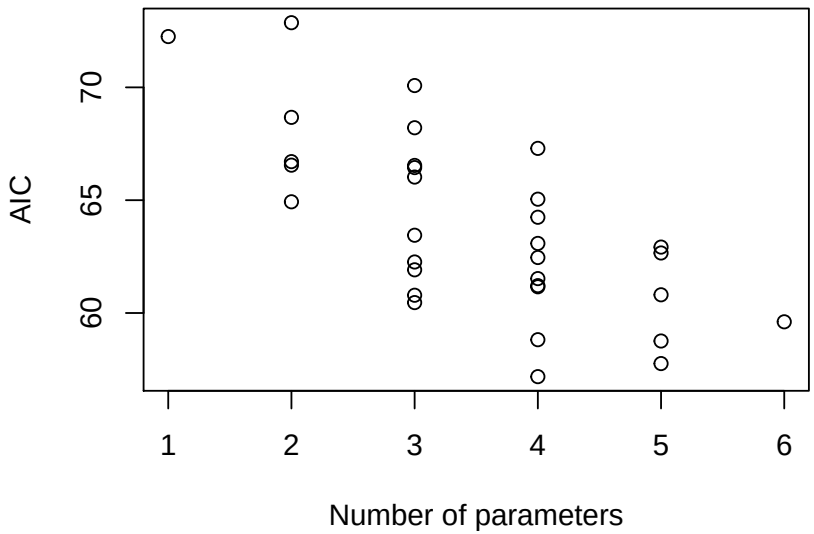
- $2(p - \hat{\ell})$ (AIC—Akaike Information Criterion)
- $2(\frac{1}{2}p \log n - \hat{\ell})$ (BIC—Bayes Information Criterion)
- AIC_c, AIC_u, DIC, EIC, FIC, GIC, SIC, TIC, ...
- Mallows $C_p = RSS/s^2 + 2p - n$ commonly used in regression problems, where RSS is residual sum of squares for candidate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

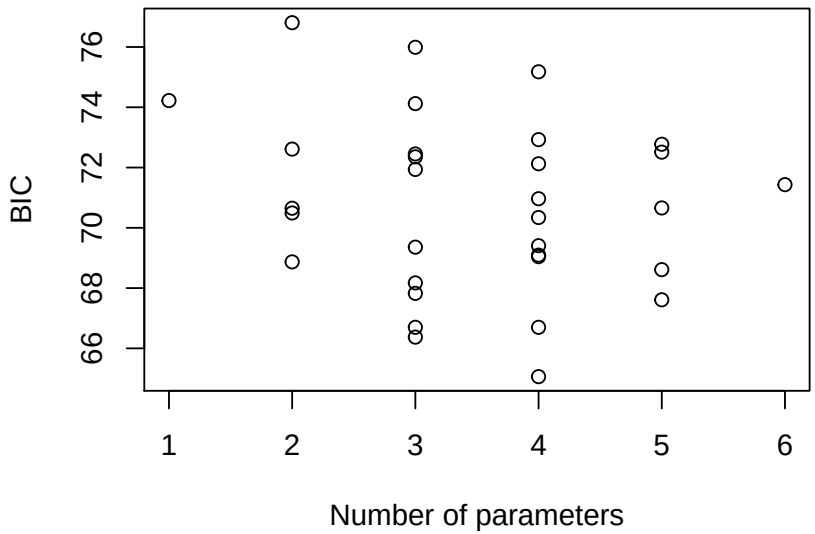
Example: return to nodal involvement

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).

Example: return to nodal involvement

We can plot out 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.





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**.

Consistency of information criteria

We seek to find the correct model by minimising 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 IC_+ for a model with one more parameter.

Consistency of information criteria

Then

$$P(\text{IC}_+ < \text{IC}) = P\left\{c(n,p+1) - 2\hat{\ell}_+ < c(n,p) - 2\hat{\ell}\right\}$$
$$= P\left\{2(\hat{\ell}_+ - \hat{\ell}) > c(n,p+1) - c(n,p)\right\}.$$

and in large samples

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

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.

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

- 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.

Prediction error

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

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

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 of Theorem 1.2

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.

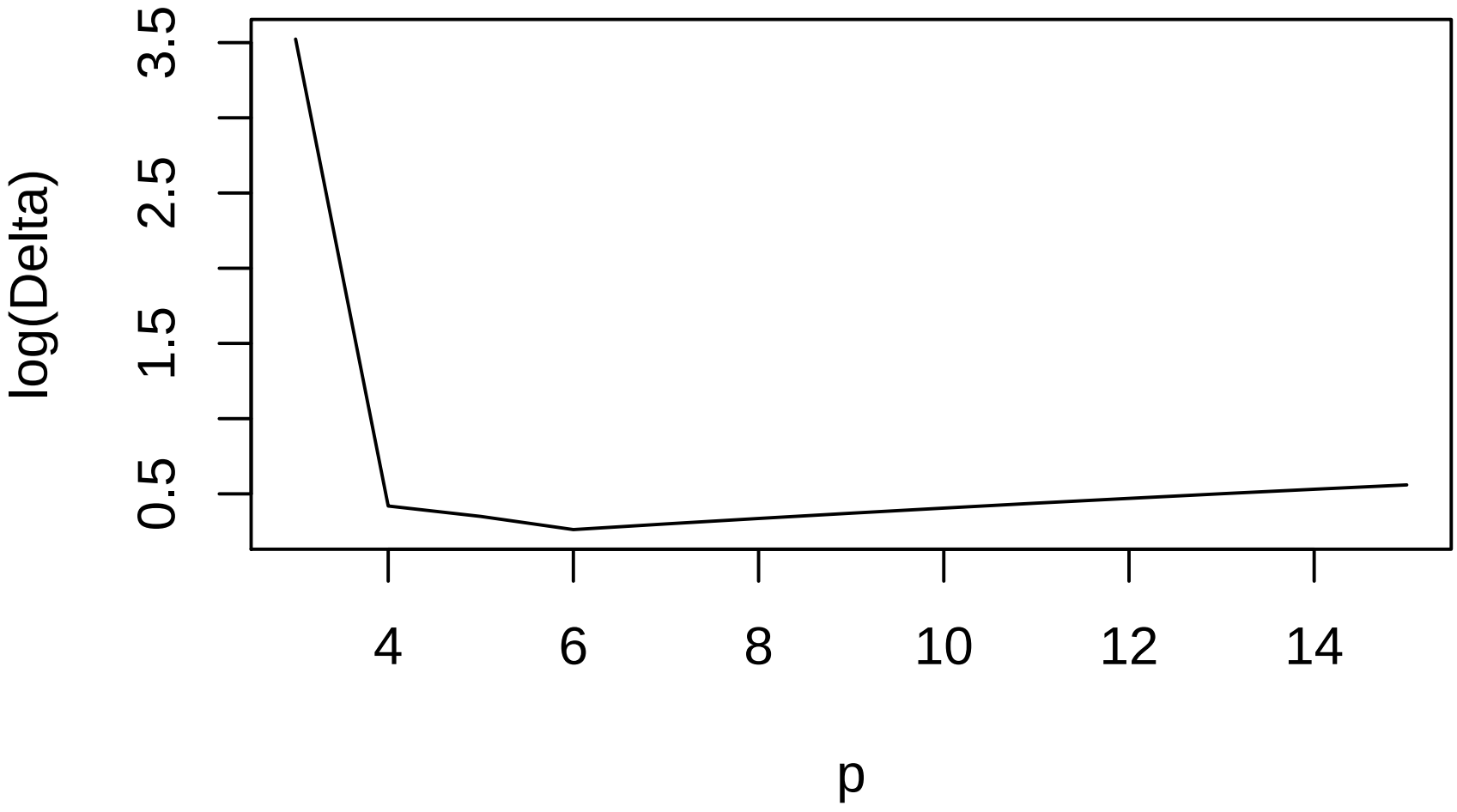
Example

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.

We plot $\log(\Delta(X))$ against p for models of increasing polynomial degree, from a quadratic model ($p = 3$) to a degree 14 polynomial ($p = 15$).

$\log(\Delta(X))$ against p



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.

Cross-validation

If n is large, can split the data into two parts (X', y') and (X^*, y^*), 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

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

where $\hat{\beta}_{-j}$ is estimate computed without (x_j, y_j).

Simplified computation of CV

This seems to require n fits of model, but in fact

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

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

Generalised cross-validation

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

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

Theorem 1.3

We have

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

Proof of Theorem 1.3

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

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

Now note that $\text{tr}(H) = p$ 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 $p \ll n$.

Variants of cross-validation

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.

Bayesian inference

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**

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

The encompassing model

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.

A Bayesian perspective on model comparison

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 $f(\theta_m \mid y, m)$.

The Bayes factor

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** $B_{12} = f(y \mid 1)/f(y \mid 2)$.

Sensitivity of the marginal likelihood

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,

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

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

Sensitivity of the marginal likelihood

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.

Model averaging

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

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

APTS Statistical Modelling

Chapter 2: Beyond Generalised Linear Models

Helen Ogden

April 2021

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.

When do we need to go beyond GLMs?

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

An example of overdispersion

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.

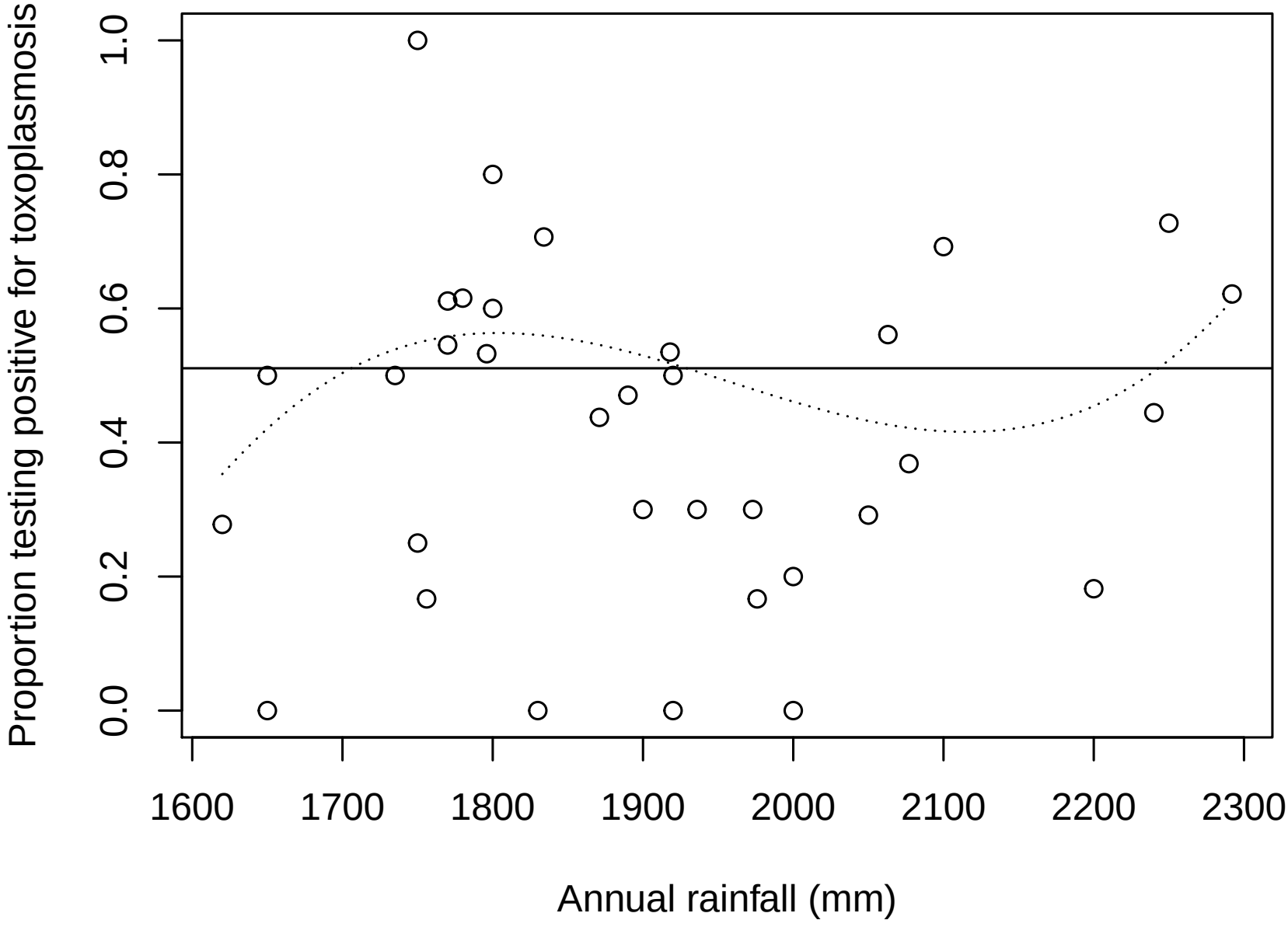
We can fit various logistic regression models for relating toxoplasmosis incidence to rainfall.

An example of overdispersion

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")
```

Comparing constant and cubic model



Comparing constant and cubic model

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.

Evidence of overdispersion

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

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

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 $\text{var}(Y_i) = \boldsymbol{\sigma^2} V(\mu_i)/m_i$, 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

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

Quasilikelihood for the toxoplasmosis data

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 constant and 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.

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*.

Adding missing explanatory variables to the model

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.

Modelling the Y_i directly: a beta-binomial model

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.

The beta-binomial model has likelihood

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

Modelling the Y_i directly

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

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

$$\eta_i = x_i^T \beta + u_i \tag{1}$$

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

Likelihood for random effects model

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

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

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.

Toxoplasmosis example revisited

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.

Reasons for clustering

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}.

Modelling clustering with random effects

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.

Marginal models

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').

Generalised estimating equations

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.

Clustered data

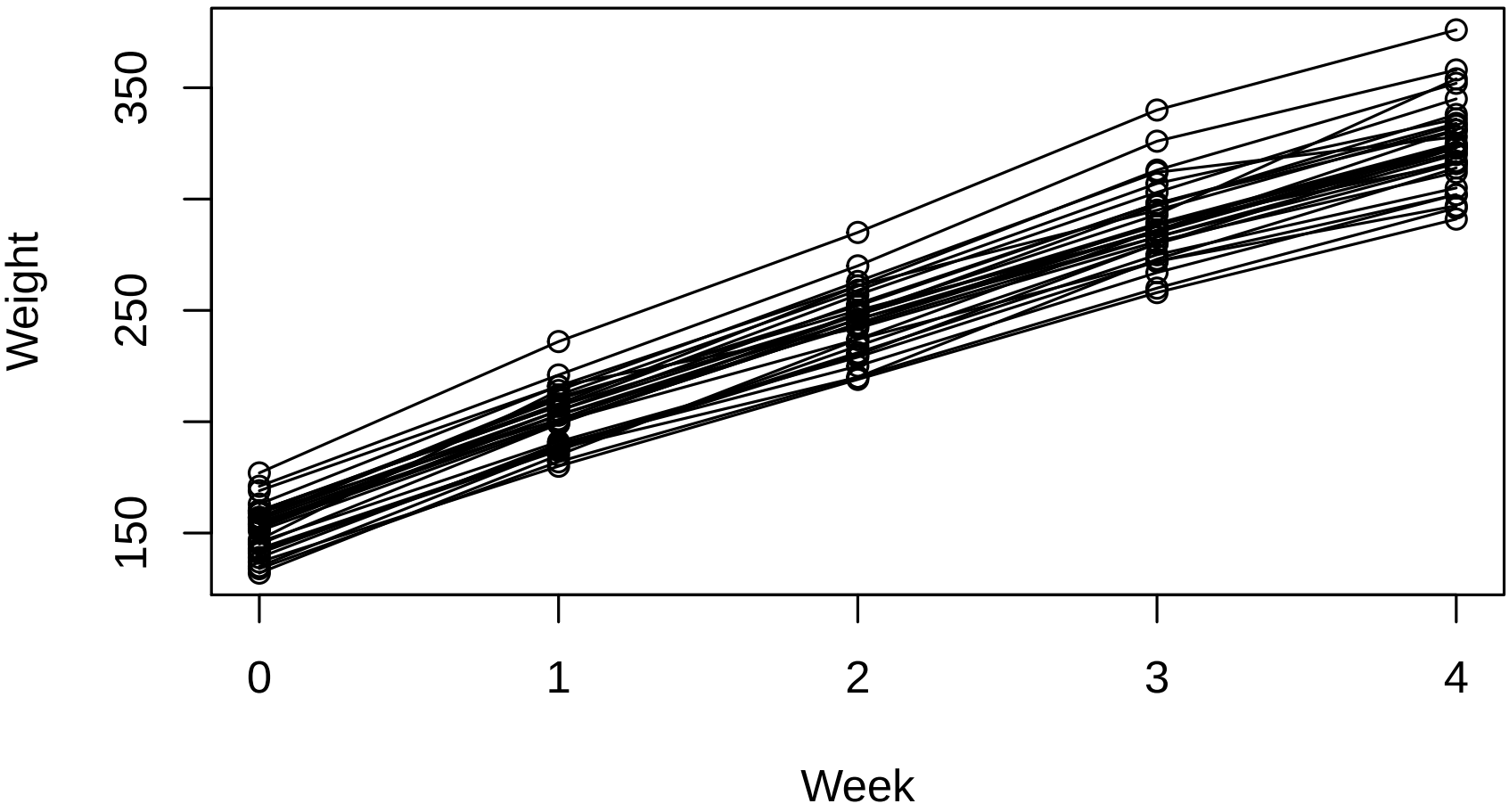
Examples where data are collected in clusters include:

- Studies in biometry where **repeated measures** are made on experimental units.
- Agricultural field trials where experimental units are arranged within **blocks**.
- Sample surveys where collecting data within clusters or **small areas** can save costs.

Other forms of dependence also exist, for example spatial or serial dependence induced by arrangement in space or time of units of observation.

Rat growth example

The `rat.growth` data in `SMPracticals` gives the weekly weights (`y`) of 30 young rats. We can plot the weight against week separately for each rat:



Simple linear regression

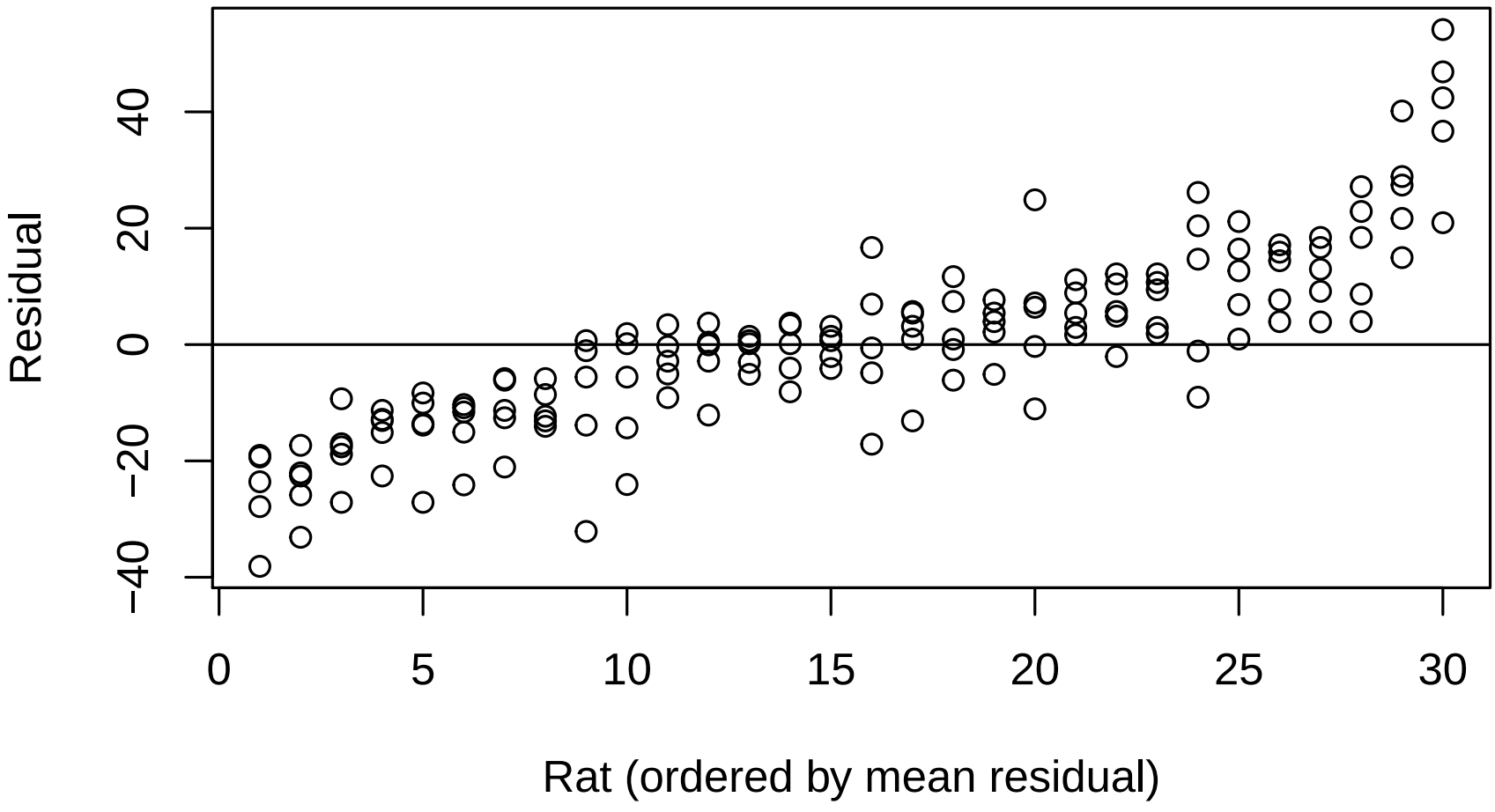
Writing y_{ij} for the jth observation of the weight of rat i, and x_{ij} for the week in which this record was made, we can fit the simple linear regression

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

with resulting estimates $\hat{\beta}_0 = 156.1$ (2.25) and $\hat{\beta}_1 = 43.3$ (0.92).

Residual plots for simple linear regression

Residuals show clear evidence of an unexplained difference between rats:



Linear mixed models

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

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

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

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

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, $z_{ij} = (1)$.

A random slopes model

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 (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.

Mixed models

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

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

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

Hence var(Y) is comprised of two **variance components**.

Hierarchical models

Other ways of describing LMMs for clustered data (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.

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.

LMM fitting

Recall that

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

so the likelihood for $(\beta, \Sigma, \Sigma_b)$

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

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).

REML

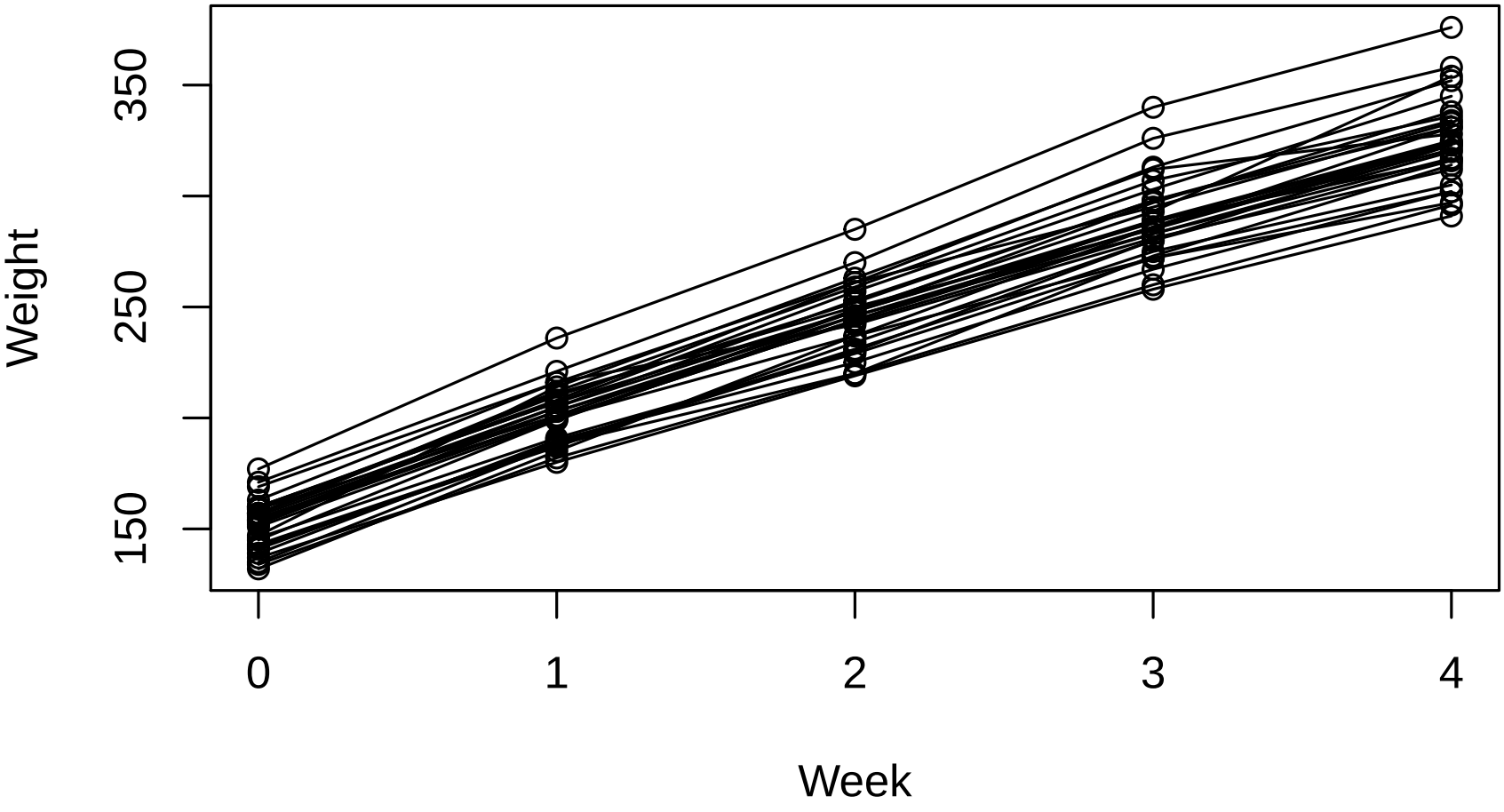
Estimation by **REML** – *REstricted* (or *REsidual*) Maximum Likelihood is usually preferred.

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 $E(Y) = X\beta$.

However, REML maximised likelihoods cannot be used to compare different fixed effects specifications, because these residuals (used as "data" in REML) depend on X.

Return to the rat growth example

Recall the rat growth data:



A random slopes model for the rat growth data

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.

Fitting the model in R

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
```

A random intercept model

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
```

Comparing models with information criteria

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.

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

$$\tilde{b} = E(b \mid y) = (Z^T \Sigma^{-1} Z + \Sigma_b^{-1})^{-1} Z^T \Sigma^{-1} (y - X\beta) \qquad (6)$$

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.

Return to rat growth example

Our preferred model `rat_rs` for the rat growth data was the random slopes model

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

We can find estimates of the rat-specific intercepts and slopes with

```
coef(rat_rs)$rat
```

```
##    (Intercept)     week
## 1     155.0375 42.39649
## 2     150.5353 48.98945
## 3     161.2795 45.64680
## 4     156.9845 37.48403
## 5     141.8160 45.14045
## 6     162.4677 43.58348
## 7     146.1565 41.27861
## 8     158.4606 45.02854
## 9     181.5240 50.00399
## 10    139.3235 39.97359
## 11    162.6955 47.93547
## 12    143.9827 42.14362
## 13    156.1070 43.14287
## 14    172.8526 47.75630
## 15    165.1396 38.48420
```

Alternative fixed effects model

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

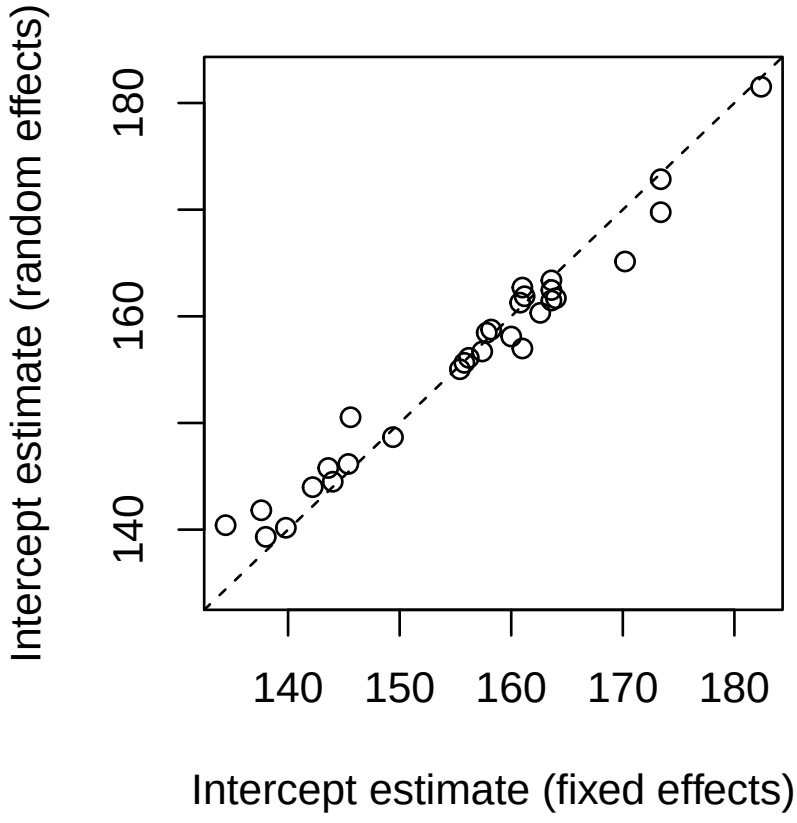
$$y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \epsilon_{ij}.$$

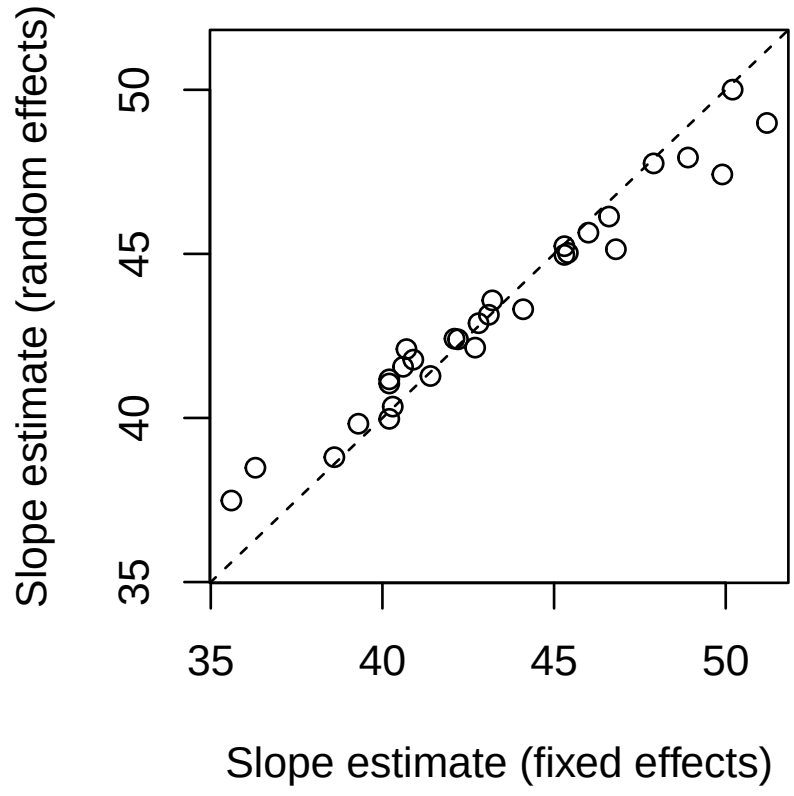
```
rat_fs <- lm(y ~ rat * week, data = rat.growth)
rat_fs
```

```
##
## Call:
## lm(formula = y ~ rat * week, data = rat.growth)
##
## Coefficients:
## (Intercept)         rat2         rat3         rat4         rat5         rat6
##       155.4         -9.8          5.4          5.6        -17.8          8.2
##        rat7         rat8         rat9        rat10        rat11        rat12
##       -10.0          2.4         27.0        -17.4          5.6        -13.2
##       rat13        rat14        rat15        rat16        rat17        rat18
##         0.8         18.0         14.8          8.2        -11.8          4.6
##       rat19        rat20        rat21        rat22        rat23        rat24
##         8.2          2.0          2.8        -11.4         -6.0          8.6
##       rat25        rat26        rat27        rat28        rat29        rat30
##       -21.0          5.8         18.0          7.2        -15.6          0.4
##        week    rat2:week    rat3:week    rat4:week    rat5:week    rat6:week
##        42.2          9.0          3.8         -6.6          4.6          1.0
```

Shrinkage towards a common mean

Comparing the estimates from the random slopes and this fixed effects model:





Random effects estimates 'borrow strength' across clusters. The extent of this is determined by cluster similarity.

Generalised linear mixed models

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

$$Y_i \sim F(\cdot \mid \mu_i, \sigma^2), \quad g(\mu) \equiv \begin{pmatrix} g(\mu_1) \\ \vdots \\ g(\mu_n) \end{pmatrix} = X\beta + Zb, \quad b \sim N(0, \Sigma_b) \tag{7}$$

where $F(\cdot \mid \mu_i, \sigma^2)$ is an exponential family distribution with $E(Y) = \mu$ and $\text{var}(Y) = \sigma^2 V(\mu)/m$ 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.

A random interecept GLMM for binary data

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

$$Y_{ij} \sim \text{Bernoulli}(\mu_{ij}), \quad \log \frac{\mu_{ij}}{1-\mu_{ij}} = \beta_0 + b_{0i} + \beta_1 x_{ij}, \quad b_{0i} \sim N(0, \sigma_b^2) \tag{8}$$

Note there is no random slope here. This fits into the general GLMM framework (7) 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.

GLMM likelihood

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

$$\begin{aligned} f(y \mid \beta, \Sigma_b) &= \int f(y \mid \beta, b, \Sigma_b) f(b \mid \beta, \Sigma_b) db \\ &= \int f(y \mid \beta, b) f(b \mid \Sigma_b) db \\ &= \int \prod_{i=1}^{n} f\left(y_i \mid g^{-1}([X\beta + Zb]_i)\right) f(b \mid \Sigma_b) db. \quad (9) \end{aligned}$$

For *nested* random effects structures, some simplification is possible. For example, for (8)

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

a product of one-dimensional integrals.

Approximating the likelihood

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. 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.

Laplace approximation to the likelihood

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 $\tilde{h}(\hat{b}) = h(\hat{b})$.

Laplace approximation to the likelihood

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.

Penalized Quasi Likelihood

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.

Toxoplasmosis example revisited

For the individual-level model

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

the estimates and standard errors obtained by ML (quadrature), Laplace and PQL are:

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 example revisited

For the extended model

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

the estimates and standard errors are:

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, there is a good agreement between the different computational methods.

APTS Statistical Modelling

Chapter 3: Nonlinear models

Helen Ogden

April 2021

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.

Extensions of the normal linear model

We begin by considering nonlinear extensions of the normal linear model

$$y_i = x_i^T \beta + \epsilon_i, \tag{1}$$

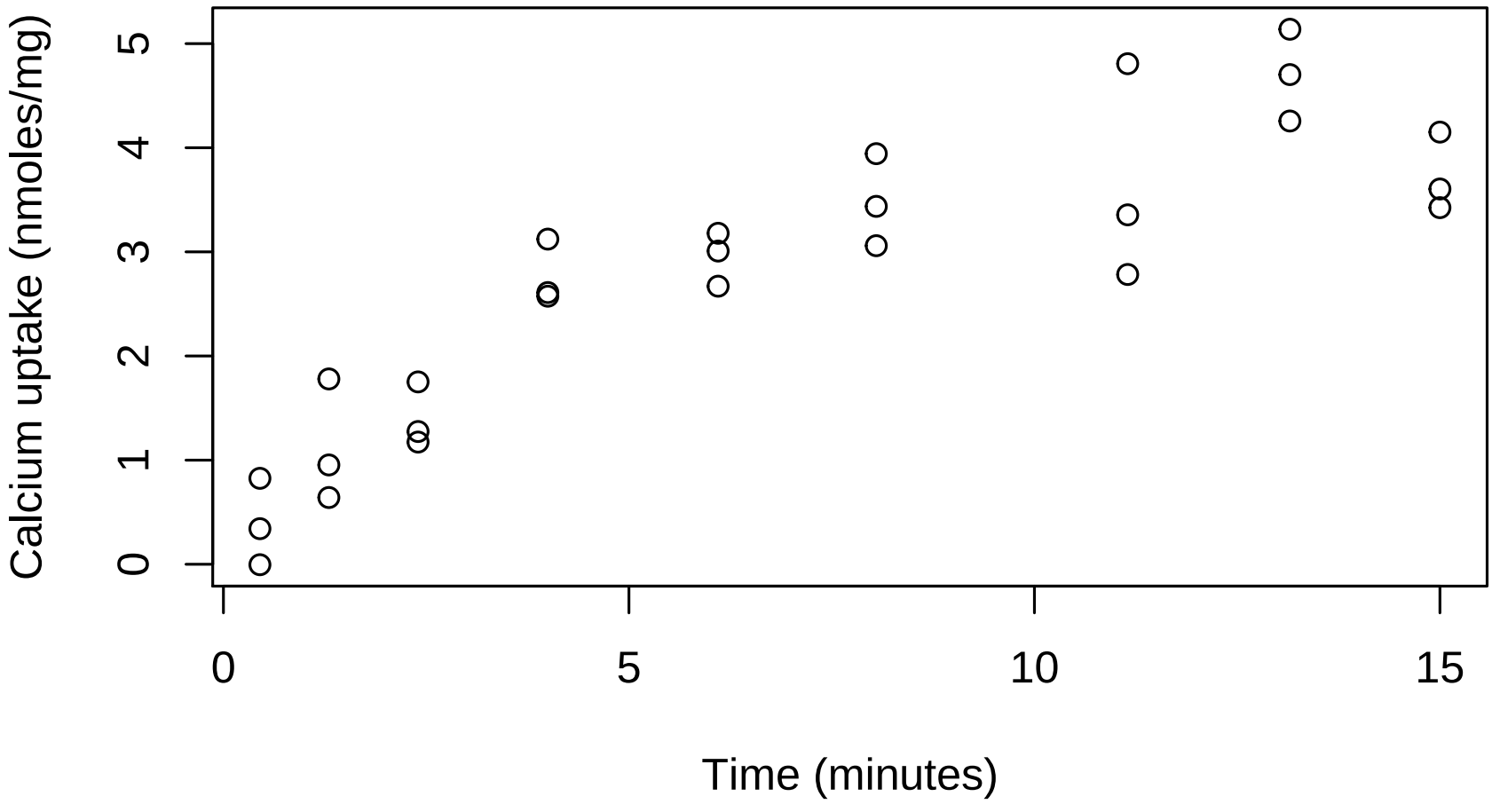
where $\epsilon_i \sim N(0, \sigma^2)$, independently, where β are the p regression parameters. Instead of the mean response being the linear predictor $x_i^T \beta$, we could allow it to be a nonlinear function of parameters, i.e.

$$y_i = \eta(x_i, \beta) + \epsilon_i, \tag{2}$$

where $\epsilon_i \sim N(0, \sigma^2)$, independently, where β are the p nonlinear parameters. The model specified by (2) has the linear model (1) as a special case when $\eta(x, \beta) = x^T \beta$.

Example: `calcium` data

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



Nonlinear parameters

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 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.

Return to the `calcium` data

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

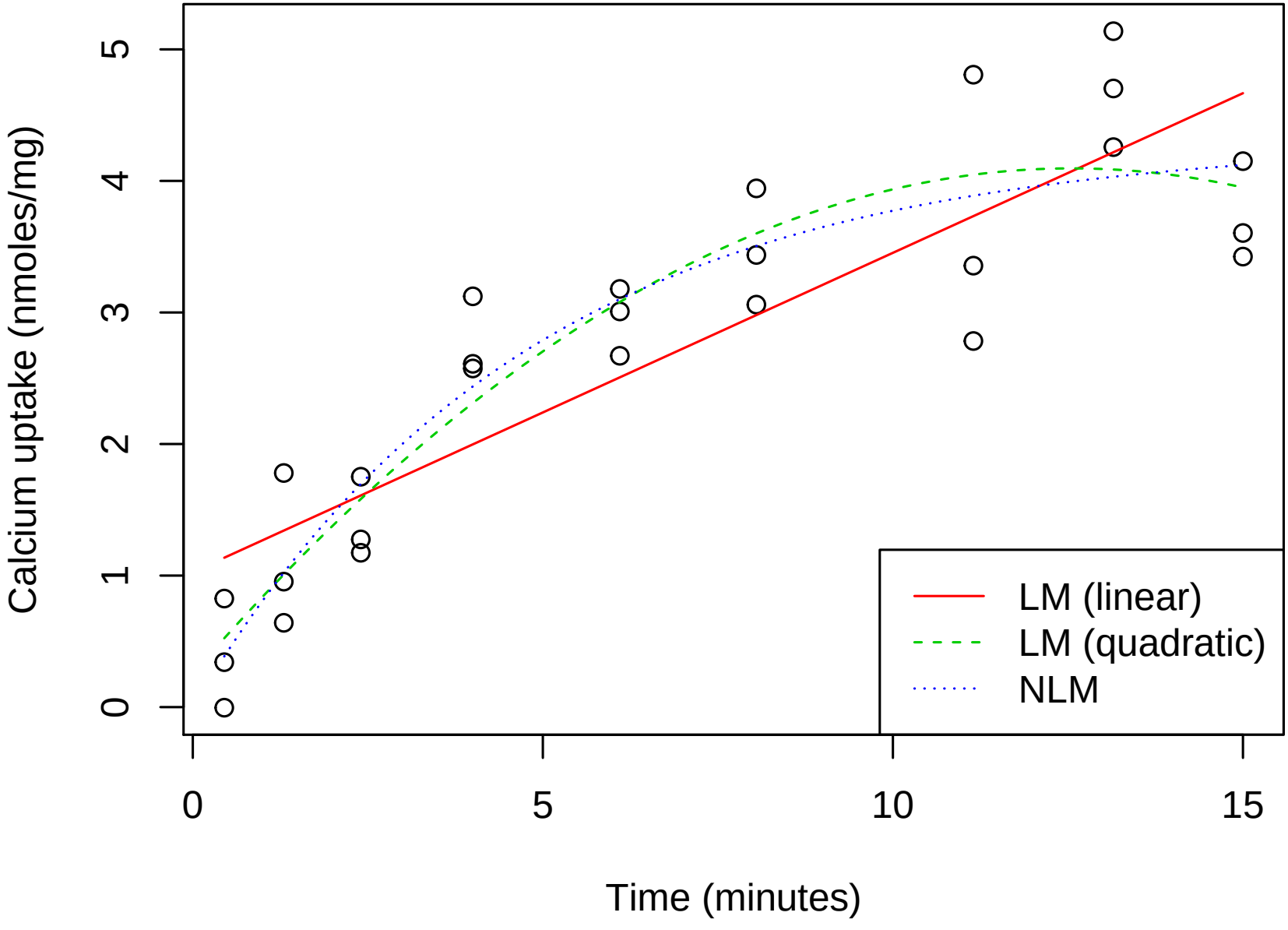
$$\eta(x, \beta) = \beta_0 \left(1 - \exp\left(-x/\beta_1\right)\right). \tag{3}$$

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 (3). Here β_0 is the final size of the population, and β_1 (inversely) controls the growth rate.

Models for the `calcium` data



Models for the `calcium` data

A comparison of the goodness-of-fit for the three models:

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

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.

Extending the nonlinear model

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.

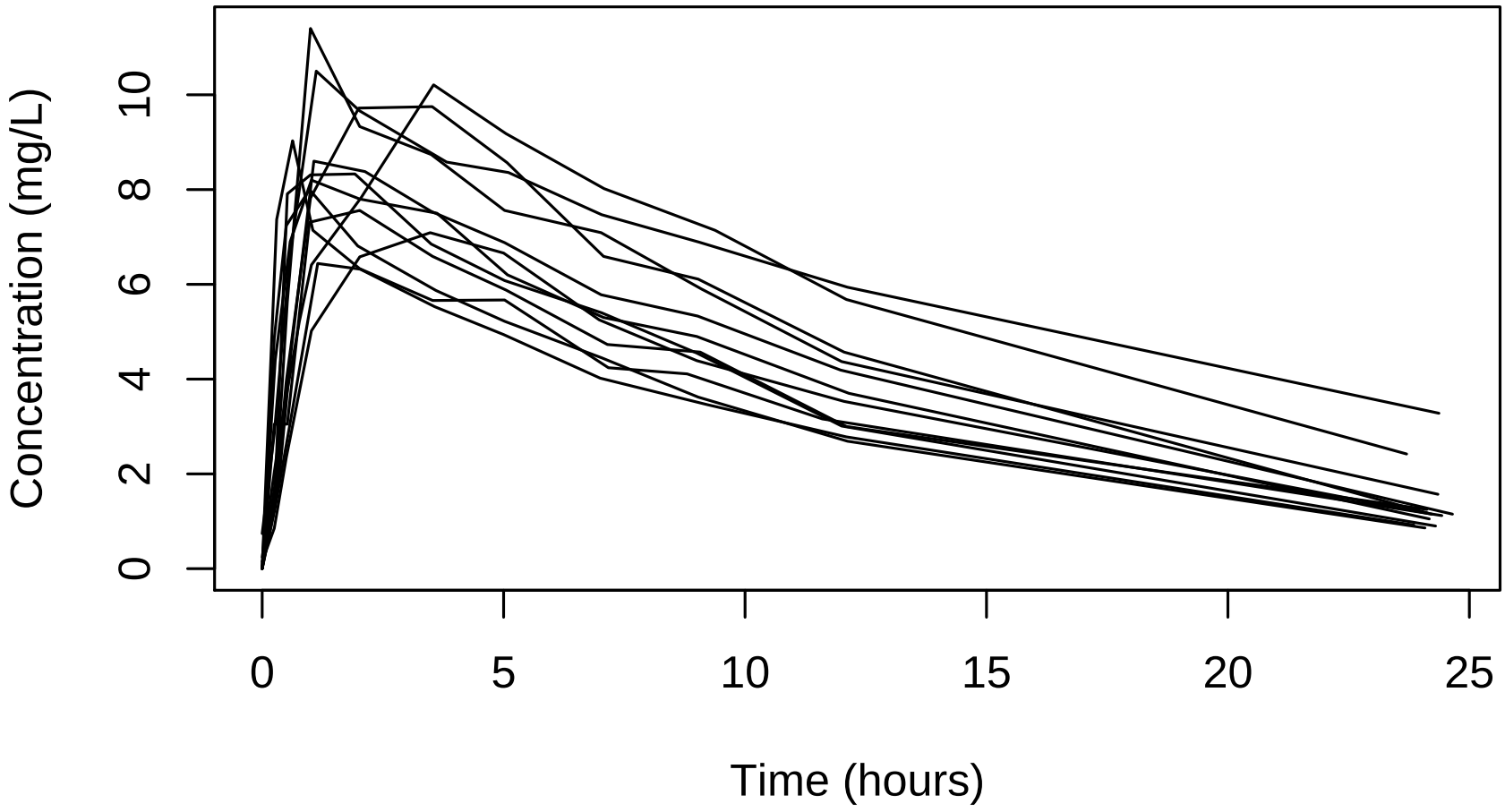
Example: Theophylline data

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}.

Example: Theophylline data



For each individual, there is a sharp increase in concentration followed by a steady decrease.

Compartmental models

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.

Compartmental models ignoring dependence

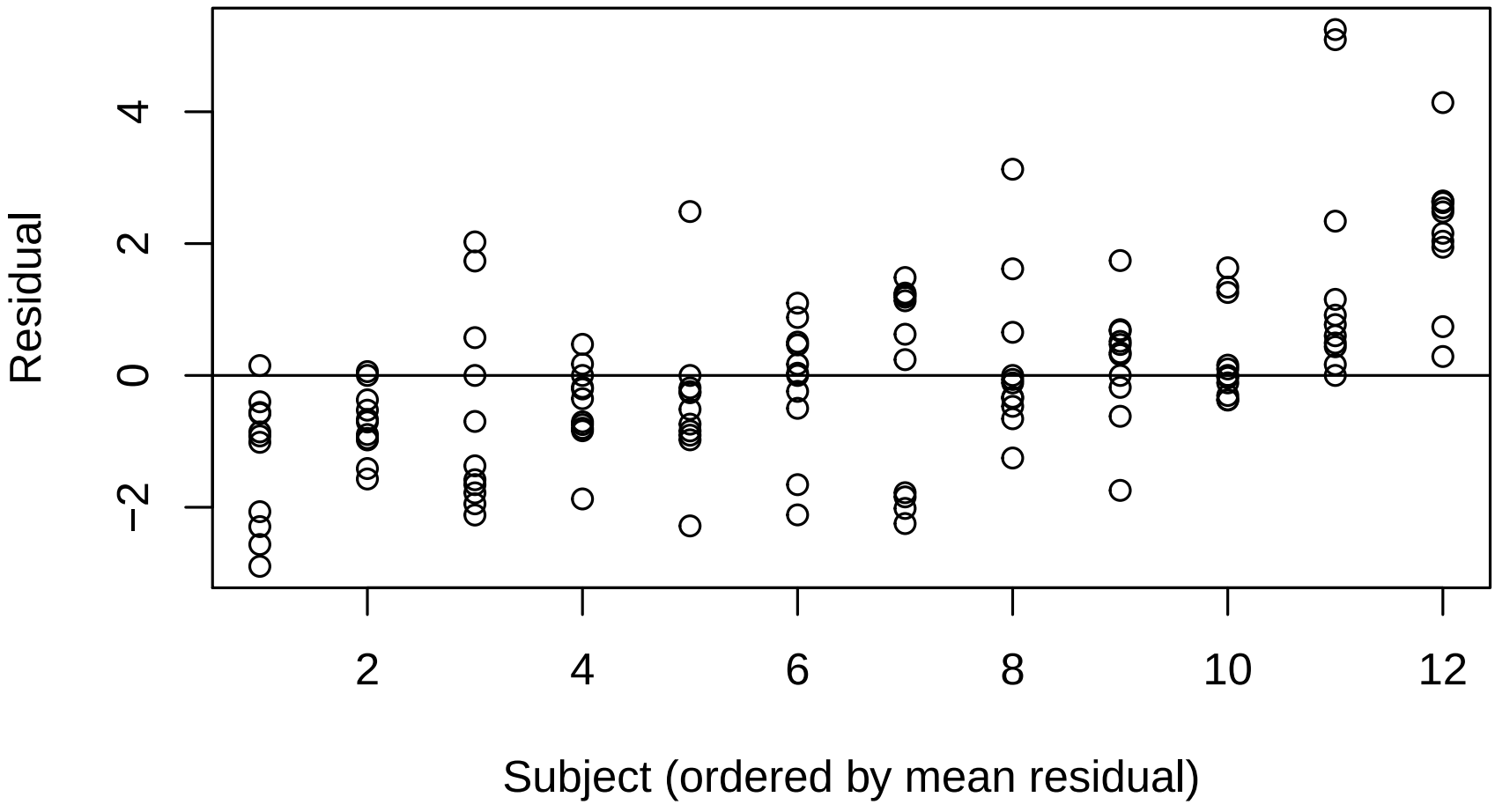
Initially ignore the dependence induced from repeated measurements on individuals and assume the following basic nonlinear model

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

where $\epsilon_{ij} \sim N(0, \sigma^2)$.

Compartmental models for the theophylline data

Residuals show evidence of an unexplained difference between individuals.



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.

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.

Extending the nonlinear mixed effects model

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.

Special case: the linear mixed model

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 $q = 1$, $A = (1, 0, \ldots, 0)$.

Return to the theophylline example

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 $= 372.6$.

Return to the theophylline example

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.

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	

Issues with fitting nonlinear models

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.