Statistical Modelling

Antony Overstall

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

Statistical Modelling

1. Model Selection
2. Beyond the Generalised Linear Model
3. Design of Experiments

1. Model Selection

Overview

1. Basic ideas
2. Linear model
3. Bayesian inference

Basic Ideas

Why model?

George E. P. Box (1919–2013):

> All models are wrong, but some models are useful.

- ☐ Some reasons we construct models:
  - to simplify reality (efficient representation);
  - to gain understanding;
  - to compare scientific, economic, ... theories;
  - to predict future events/data;
  - to control a process.
- ☐ We (statisticians!) rarely believe in our models, but regard them as temporary constructs subject to improvement.
- ☐ Often we have several and must decide which is preferable, if any.

Criteria for model selection

- ☐ Substantive knowledge, from prior studies, theoretical arguments, dimensional or other general considerations (often qualitative)
- ☐ Sensitivity to failure of assumptions (prefer models that are robustly valid)
- ☐ Quality of fit—residuals, graphical assessment (informal), or goodness-of-fit tests (formal)
- ☐ Prior knowledge in Bayesian sense (quantitative)
- ☐ Generalisability of conclusions and/or predictions: same/similar models give good fit for many different datasets
- ☐ ... but often we have just one dataset ...

Motivation

Even after applying these criteria (but also before!) we may compare many models:

- ☐ linear regression with p covariates, there are 2^p possible combinations of covariates (each in/out), before allowing for transformations, etc.— if $p = 20$ then we have a problem;
- ☐ choice of bandwidth $h > 0$ in smoothing problems
- ☐ the number of different clusterings of n individuals is a Bell number (starting from $n = 1$): 1, 2, 5, 15, 52, 203, 877, 4140, 21147, 115975, ...
- ☐ we may want to assess which among 5×10^5 SNPs on the genome may influence reaction to a new drug;
- ☐ ...

For reasons of economy we seek 'simple' models.

Albert Einstein (1879–1955)

'Everything should be made as simple as possible, **but no simpler**.'

William of Occam (?1288–?1348)

Occam's razor: **Entia non sunt multiplicanda sine necessitate**: **entities should not be multiplied beyond necessity**.

Setting

- To focus and simplify discussion we will consider parametric models, but the ideas generalise to semi-parametric and non-parametric settings
- We shall take generalised linear models (GLMs) as example of moderately complex parametric models:
  - Normal linear model has three key aspects:
    - ▷ *structure for covariates*: *linear predictor* $\eta = x^{\mathrm{T}}\beta$;
    - ▷ *response distribution*: $y \sim N(\mu, \sigma^2)$; and
    - ▷ relation $\eta = \mu$ between $\mu = \mathrm{E}(y)$ and η.
  - GLM extends last two to
    - ▷ y has density
    $$f(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y; \phi)\right\},$$
    where θ depends on η; *dispersion parameter* ϕ is often known; and
    - ▷ $\eta = g(\mu)$, where g is monotone *link function*.

Logistic regression

- Commonest choice of link function for binary reponses:
$$\Pr(Y = 1) = \pi = \frac{\exp(x^{\mathrm{T}}\beta)}{1 + \exp(x^{\mathrm{T}}\beta)}, \quad \Pr(Y = 0) = \frac{1}{1 + \exp(x^{\mathrm{T}}\beta)},$$
giving linear model for log odds of 'success',
$$\log\left\{\frac{\Pr(Y = 1)}{\Pr(Y = 0)}\right\} = \log\left(\frac{\pi}{1 - \pi}\right) = x^{\mathrm{T}}\beta.$$
- Log likelihood for β based on independent responses $y_1, \ldots, y_n$ with covariate vectors $x_1, \ldots, x_n$ is
$$\ell(\beta) = \sum_{j=1}^{n} y_j x_j^{\mathrm{T}}\beta - \sum_{j=1}^{n} \log\left\{1 + \exp(x_j^{\mathrm{T}}\beta)\right\}$$
- Good fit gives small deviance $D = 2\left\{\ell(\tilde{\beta}) - \ell(\widehat{\beta})\right\}$, where $\widehat{\beta}$ is model fit MLE and $\tilde{\beta}$ is unrestricted MLE.

Nodal involvement data

Table 1: Data on nodal involvement: 53 patients with prostate cancer have nodal involvement (r), with five binary covariates age etc.

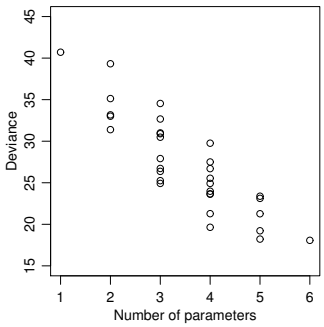
m	r	age	stage	grade	xray	acid
6	5	0	1	1	1	1
6	1	0	0	0	0	1
4	0	1	1	1	0	0
4	2	1	1	0	0	1
4	0	0	0	0	0	0
3	2	0	1	1	0	1
3	1	1	1	0	0	0
3	0	1	0	0	0	1
3	0	1	0	0	0	0
2	0	1	0	0	1	0
2	1	0	1	0	0	1
2	1	0	0	1	0	0
1	1	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	
1	1	0	0	1	0	1
1	0	0	0	0	1	1
1	0	0	0	0	1	0

Nodal involvement deviances

Deviances D for 32 logistic regression models for nodal involvement data. $+$ denotes a term included in the model.

age	st	gr	xr	ac	df	D	age	st	gr	xr	ac	df	D
					52	40.71	+	+	+			49	29.76
+					51	39.32	+	+		+		49	23.67
	+				51	33.01	+	+			+	49	25.54
		+			51	35.13	+		+	+		49	27.50
			+		51	31.39	+		+		+	49	26.70
				+	51	33.17	+			+	+	49	24.92
+	+				50	30.90		+	+	+		49	23.98
+		+			50	34.54		+	+		+	49	23.62
+			+		50	30.48		+		+	+	49	19.64
+				+	50	32.67			+	+	+	49	21.28
	+	+			50	31.00	+	+	+	+		48	23.12
	+		+		50	24.92	+	+	+		+	48	23.38
	+			+	50	26.37	+	+		+	+	48	19.22
		+	+		50	27.91	+		+	+	+	48	21.27
		+		+	50	26.72		+	+	+	+	48	18.22
			+	+	50	25.25	+	+	+	+	+	47	18.07

Nodal involvement



- Adding terms
  - always increases the log likelihood $\widehat{\ell}$ and so reduces D,
  - increases the number of parameters,

  so taking the model with highest $\widehat{\ell}$ (lowest D) would give the full model
- We need to trade off quality of fit (measured by D) and model complexity (number of parameters)

Log likelihood

- Given (unknown) **true model** $g(y)$, and **candidate model** $f(y;\theta)$, Jensen's inequality implies that

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

  with equality if and only if $f(y;\theta) \equiv g(y)$.
- If θ_g is the value of θ that maximizes the expected log likelihood on the right of (1), then it is natural to choose the candidate model that maximises

$$\overline{\ell}(\widehat{\theta}) = n^{-1} \sum_{j=1}^{n} \log f(y;\widehat{\theta}),$$

  which should be an estimate of $\int \log f(y;\theta) g(y)\, dy$. However as $\overline{\ell}(\widehat{\theta}) \geq \overline{\ell}(\theta_g)$, by definition of $\widehat{\theta}$, this estimate is biased upwards.
- We need to correct for the bias, but in order to do so, need to understand the properties of likelihood estimators when the assumed model f is not the true model g.

Wrong model

Suppose the true model is g, that is, $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} g$, but we assume that $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y;\theta)$. The log likelihood $\ell(\theta)$ will be maximised at $\widehat{\theta}$, and

$$\overline{\ell}(\widehat{\theta}) = n^{-1}\ell(\widehat{\theta}) \xrightarrow{\text{a.s.}} \int \log f(y;\theta_g) g(y)\, dy, \quad n \to \infty,$$

where θ_g minimizes the Kullback–Leibler discrepancy

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

θ_g gives the density $f(y;\theta_g)$ closest to g in this sense, and $\widehat{\theta}$ is determined by the finite-sample version of $\partial KL(f_\theta, g)/\partial\theta$, i.e.

$$0 = n^{-1}\sum_{j=1}^{n} \frac{\partial \log f(y_j;\widehat{\theta})}{\partial\theta}.$$

Wrong model II

Theorem 1 *Suppose the true model is g, that is, $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} g$, but we assume that $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y;\theta)$. Then under mild regularity conditions the maximum likelihood estimator $\widehat{\theta}$ satisfies*

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

where f_{θ_g} is the density minimising the Kullback–Leibler discrepancy between f_θ and g, I is the Fisher information for f, and K is the variance of the score statistic. The likelihood ratio statistic

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

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

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.

Note: 'Proof' of Theorem 1

Expansion of the equation defining $\widehat{\theta}$ about θ_g yields

$$\widehat{\theta} \doteq \theta_g + \left\{ -n^{-1} \sum_{j=1}^{n} \frac{\partial^2 \log f(y_j;\theta_g)}{\partial\theta\partial\theta^{\mathrm{T}}} \right\}^{-1} \left\{ n^{-1} \sum_{j=1}^{n} \frac{\partial \log f(y_j;\theta_g)}{\partial\theta} \right\}$$

and a modification of the usual derivation gives

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

where the *information sandwich* variance matrix depends on

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

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

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 of course unknown, and then $K(\theta_g)$ and $I(\theta_g)$ may be estimated by

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

the latter is just the observed information matrix. We may then construct confidence intervals for θ_g using (2) with variance matrix $\widehat{J}^{-1}\widehat{K}\widehat{J}^{-1}$.
Similar expansions lead to the result for the likelihood ratio statistic.

Out-of-sample prediction

- We need to fix two problems with using $\overline{\ell}(\widehat{\theta})$ to choose the best candidate model:
  - upward bias, as $\overline{\ell}(\widehat{\theta}) \geq \overline{\ell}(\theta_g)$ because $\widehat{\theta}$ is based on $Y_1, \ldots, Y_n$;
  - no penalisation if the dimension of θ increases.
- If we had another independent sample $Y_1^+, \ldots, Y_n^+ \overset{\text{iid}}{\sim} g$ and computed

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

then both problems disappear, suggesting that we choose the candidate model that maximises

$$\mathrm{E}_g\left[\mathrm{E}_g^+\left\{\overline{\ell}^+(\widehat{\theta})\right\}\right],$$

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

Information criteria

- Previous results on wrong model give

$$\mathrm{E}_g\left[\mathrm{E}_g^+\left\{\overline{\ell}^+(\widehat{\theta})\right\}\right] \doteq \int \log f(y;\theta_g) g(y)\, dy - \frac{1}{2n}\mathrm{tr}\{I(\theta_g)^{-1}K(\theta_g)\},$$

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

- We want to estimate this based on $Y_1, \ldots, Y_n$ only, and get

$$\mathrm{E}_g\left\{\overline{\ell}(\widehat{\theta})\right\} \doteq \int \log f(y;\theta_g) g(y)\, dy + \frac{1}{2n}\mathrm{tr}\{I(\theta_g)^{-1}K(\theta_g)\},$$

- To remove the bias, we aim to maximise

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

  where

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

  the latter is just the observed information matrix.

Note: Bias of log likelihood

To compute the bias in $\overline{\ell}(\widehat{\theta})$, we write

$$\begin{aligned}
\mathrm{E}_g\left\{\overline{\ell}(\widehat{\theta})\right\} &= \mathrm{E}_g\left\{\overline{\ell}(\theta_g)\right\} + \mathrm{E}\left\{\overline{\ell}(\widehat{\theta}) - \overline{\ell}(\theta_g)\right\} \\
&= \mathrm{E}_g\left\{\overline{\ell}(\theta_g)\right\} + \frac{1}{2n}\mathrm{E}\left\{W(\theta_g)\right\}, \\
&\doteq \mathrm{E}_g\left\{\overline{\ell}(\theta_g)\right\} + \frac{1}{2n}\mathrm{tr}\{I(\theta_g)^{-1}K(\theta_g)\},
\end{aligned}$$

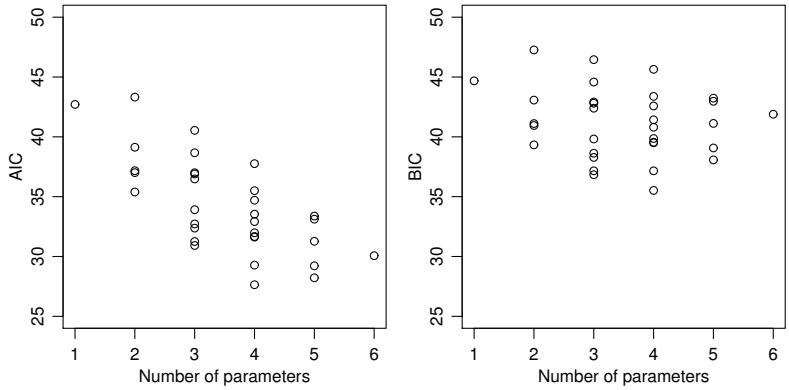
where E_g denotes expectation over the data distribution g. The bias is positive because I and K are positive definite matrices.

Information criteria

- Let $p = \dim(\theta)$ be the number of parameters for a model, and $\widehat{\ell}$ the corresponding maximised log likelihood.
- For historical reasons we choose models that **minimise** similar criteria
  - $2(p - \widehat{\ell})$ (AIC—Akaike Information Criterion)
  - $2\{\mathrm{tr}(\widehat{J}^{-1}\widehat{K}) - \widehat{\ell}\}$ (NIC—Network Information Criterion)
  - $2(\frac{1}{2}p\log n - \widehat{\ell})$ (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.

Nodal involvement data

AIC and BIC for 2^5 models for binary logistic regression model fitted to the nodal involvement data. Both criteria pick out the same model, with the three covariates `st`, `xr`, and `ac`, which has deviance $D = 19.64$. Note the sharper increase of BIC after the minimum.



Theoretical aspects

- ☐ 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**.

Properties of AIC, NIC, BIC

- ☐ We seek to find the correct model by minimising $\mathrm{IC} = c(n,p) - 2\widehat{\ell}$, 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. Then

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

  and in large samples

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

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

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

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

Linear Model

Variable selection

- ☐ Consider normal linear model

$$Y_{n\times 1} = X^{\dagger}_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}, \quad \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n),$$

  where **design matrix** $X^{\dagger}$ has full rank $p < n$ and columns x_r, for $r \in \mathcal{X} = \{1, \ldots, p\}$. Subsets $\mathcal{S}$ of $\mathcal{X}$ correspond to subsets of columns.
- ☐ Terminology
  - the **true** model corresponds to subset $\mathcal{T} = \{r : \beta_r \neq 0\}$, and $|\mathcal{T}| = q < p$;
  - 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 $\mathcal{T} \not\subset \mathcal{S}$.
- ☐ Aim to identify $\mathcal{T}$.
- ☐ If we choose a wrong model, have bias; if we choose a correct model, increase variance—seek to balance these.

Stepwise methods

- **Forward selection**: starting from model with constant only,
  1. add each remaining term separately to the current model;
  2. if none of these terms is significant, stop; otherwise
  3. update the current model to include the most significant new term; go to 1
- **Backward elimination**: starting from model with all terms,
  1. if all terms are significant, stop; otherwise
  2. update current model by dropping the term with the smallest F statistic; go to 1
- **Stepwise**: starting from an arbitary model,
  1. consider 3 options—add a term, delete a term, swap a term in the model for one not in the model;
  2. if model unchanged, stop; otherwise go to 1

Nuclear power station data

```
> nuclear
     cost  date t1 t2  cap pr ne ct bw cum.n pt
1  460.05 68.58 14 46  687  0  1  0  0    14  0
2  452.99 67.33 10 73 1065  0  0  1  0     1  0
3  443.22 67.33 10 85 1065  1  0  1  0     1  0
4  652.32 68.00 11 67 1065  0  1  1  0    12  0
5  642.23 68.00 11 78 1065  1  1  1  0    12  0
6  345.39 67.92 13 51  514  0  1  1  0     3  0
7  272.37 68.17 12 50  822  0  0  0  0     5  0
8  317.21 68.42 14 59  457  0  0  0  0     1  0
9  457.12 68.42 15 55  822  1  0  0  0     5  0
10 690.19 68.33 12 71  792  0  1  1  1     2  0
...
32 270.71 67.83  7 80  886  1  0  0  1    11  1
```

Nuclear power station data

	Full model		Backward		Forward	
	Est (SE)	t	Est (SE)	t	Est (SE)	t
Constant	-14.24 (4.229)	-3.37	-13.26 (3.140)	-4.22	-7.627 (2.875)	-2.66
date	0.209 (0.065)	3.21	0.212 (0.043)	4.91	0.136 (0.040)	3.38
log(T1)	0.092 (0.244)	0.38				
log(T2)	0.290 (0.273)	1.05				
log(cap)	0.694 (0.136)	5.10	0.723 (0.119)	6.09	0.671 (0.141)	4.75
PR	-0.092 (0.077)	-1.20				
NE	0.258 (0.077)	3.35	0.249 (0.074)	3.36		
CT	0.120 (0.066)	1.82	0.140 (0.060)	2.32		
BW	0.033 (0.101)	0.33				
log(N)	-0.080 (0.046)	-1.74	-0.088 (0.042)	-2.11		
PT	-0.224 (0.123)	-1.83	-0.226 (0.114)	-1.99	-0.490 (0.103)	-4.77
s (df)	0.164 (21)		0.159 (25)		0.195 (28)	

Backward selection chooses a model with seven covariates also chosen by minimising AIC.

Stepwise Methods: Comments

- ☐ Systematic search minimising AIC or similar over all possible models is preferable—not always feasible.
- ☐ Stepwise methods can fit models to purely random data—main problem is no objective function.
- ☐ Sometimes used by replacing F significance points by (arbitrary!) numbers, e.g. $F = 4$
- ☐ Can be improved by comparing AIC for different models at each step—uses AIC as objective function, but no systematic search.

Prediction error

- ☐ To identify $\mathcal{T}$, we fit candidate model

$$Y = X\beta + \varepsilon,$$

  where columns of X are a subset $\mathcal{S}$ of those of $X^\dagger$.
- ☐ Fitted value is

$$X\widehat{\beta} = X\{(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y\} = HY = H(\mu + \varepsilon) = H\mu + H\varepsilon,$$

  where $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is the **hat matrix** and $H\mu = \mu$ if the model is correct.
- ☐ Following reasoning for AIC, suppose we also have independent dataset Y_+ from the true model, so $Y_+ = \mu + \varepsilon_+$
- ☐ Apart from constants, previous measure of prediction error is

$$\Delta(X) = n^{-1}\mathrm{E}\ \mathrm{E}_+\left\{(Y_+ - X\widehat{\beta})^{\mathrm{T}}(Y_+ - X\widehat{\beta})\right\},$$

  with expectations over both Y_+ and Y.

Prediction error II

- Can show that

$$\Delta(X) = \begin{cases} n^{-1}\mu^{\mathrm{T}}(I-H)\mu + (1+p/n)\sigma^2, & \text{wrong model,} \\ (1+q/n)\sigma^2, & \text{true model,} \\ (1+p/n)\sigma^2, & \text{correct model;} \end{cases} \quad (3)$$

  recall that $q < p$.
- **Bias**: $n^{-1}\mu^{\mathrm{T}}(I-H)\mu > 0$ unless model is correct, and is reduced by including useful terms
- **Variance**: $(1+p/n)\sigma^2$ increased by including useless terms
- Ideal would be to choose covariates X to minimise $\Delta(X)$: impossible—depends on unknowns μ, σ.
- Must estimate $\Delta(X)$

Note: Proof of (3)

Consider data $y = \mu + \varepsilon$ to which we fit the linear model $y = X\beta + \varepsilon$, obtaining fitted value

$$X\widehat{\beta} = Hy = H(\mu + \varepsilon)$$

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 $y_+ = \mu + \varepsilon_+$, and we will compute the average error in predicting y_+ using $X\widehat{\beta}$, which is

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

Now

$$y_+ - X\widehat{\beta} = \mu + \varepsilon_+ - (H\mu + H\varepsilon) = (I-H)\mu + \varepsilon_+ - H\varepsilon.$$

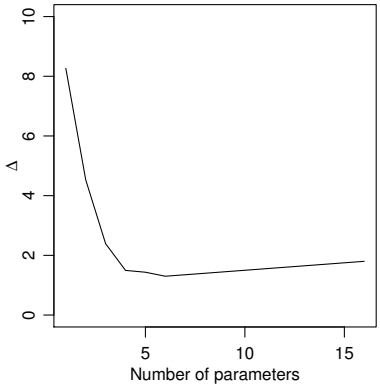
Therefore

$$(y_+ - X\widehat{\beta})^{\mathrm{T}}(y_+ - X\widehat{\beta}) = \mu^{\mathrm{T}}(I-H)\mu + \varepsilon^{\mathrm{T}}H\varepsilon + \varepsilon_+^{\mathrm{T}}\varepsilon_+ + A$$

where $\mathrm{E}(A) = 0$; this gives that

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

Example



$\Delta(X)$ as a function of the number of included variables p for data with $n = 20$, $q = 6$, $\sigma^2 = 1$. The minimum is at $p = q = 6$:

- ☐ there is a sharp decrease in bias as useful covariates are added;
- ☐ there is a slow increase with variance as the number of variables p increases.

Cross-validation

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

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

- ☐ Usually dataset is too small for this; use **leave-one-out cross-validation** sum of squares

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

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

- ☐ Seems to require n fits of model, but in fact

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

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

Cross-validation II

- Simpler (more stable?) version uses **generalised cross-validation** sum of squares

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

- Can show that

$$\mathrm{E}(\mathrm{GCV}) = \mu^{\mathrm{T}}(I-H)\mu/(1-p/n)^2 + n\sigma^2/(1-p/n) \approx n\Delta(X) \tag{4}$$

  so try and minimise GCV or CV.

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

Note: Derivation of (4)

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

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

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 $p \ll n$.

Other selection criteria

- Corrected version of AIC for models with normal responses:

$$\mathrm{AIC_c} \equiv n \log \widehat{\sigma}^2 + n\frac{1 + p/n}{1 - (p+2)/n},$$

  where $\widehat{\sigma}^2 = \mathrm{RSS}/n$. Related (unbiased) $\mathrm{AIC_u}$ replaces $\widehat{\sigma}^2$ by $S^2 = \mathrm{RSS}/(n-p)$.

- Mallows suggested

$$C_p = \frac{SS_p}{s^2} + 2p - n,$$

  where SS_p is RSS for fitted model and s^2 estimates σ^2.

- Comments:
  - AIC tends to choose models that are too complicated; $\mathrm{AIC_c}$ cures this somewhat
  - BIC chooses true model with probability $\to 1$ as $n \to \infty$, if the true model is fitted.

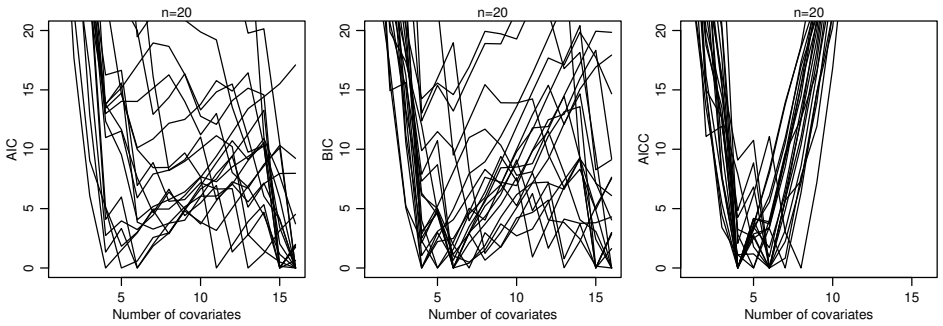
Simulation experiment

Number of times models were selected using various model selection criteria in 50 repetitions using simulated normal data for each of 20 design matrices. The true model has $p = 3$.

n		Number of covariates						
		1	2	3	4	5	6	7
10	C_p		131	504	91	63	83	128
	BIC		72	373	97	83	109	266
	AIC		52	329	97	91	125	306
	$\mathrm{AIC_c}$	15	398	565	18	4		
20	C_p		4	673	121	88	61	53
	BIC		6	781	104	52	30	27
	AIC		2	577	144	104	76	97
	$\mathrm{AIC_c}$		8	859	94	30	8	1
40	C_p			712	107	73	66	42
	BIC			904	56	20	15	5
	AIC			673	114	90	69	54
	$\mathrm{AIC_c}$			786	105	52	41	16

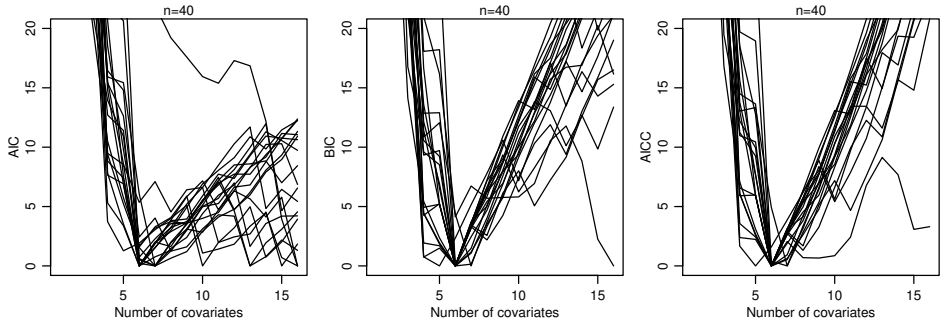
Simulation experiment

Twenty replicate traces of AIC, BIC, and $\mathrm{AIC_c}$, for data simulated with $n = 20$, $p = 1, \ldots, 16$, and $q = 6$.



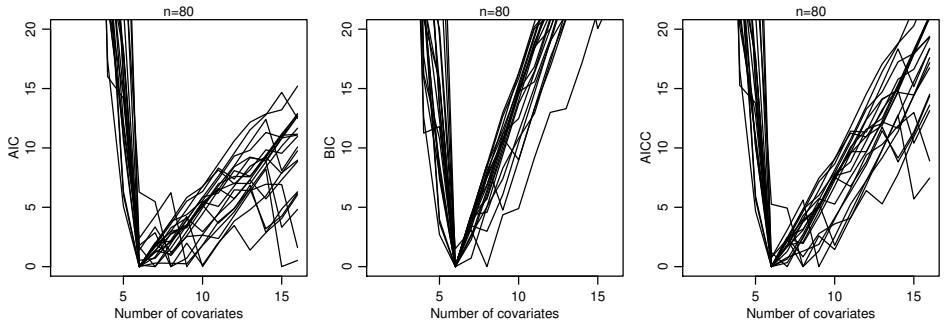
Simulation experiment

Twenty replicate traces of AIC, BIC, and AIC_c, for data simulated with $n = 40$, $p = 1, \ldots, 16$, and $q = 6$.



Simulation experiment

Twenty replicate traces of AIC, BIC, and AIC_c, for data simulated with $n = 80$, $p = 1, \ldots, 16$, and $q = 6$.



As n increases, note how

- ☐ AIC and AIC_c still allow some over-fitting, but BIC does not, and
- ☐ AIC_c approaches AIC.

Thomas Bayes (1702–1761)

Bayes (1763/4) *Essay towards solving a problem in the doctrine of chances*. Philosophical Transactions of the Royal Society of London.

Bayesian inference

Parametric model for data y assumed to be realisation of $Y \sim f(y;\theta)$, where $\theta \in \Omega_\theta$.

Frequentist viewpoint (cartoon version):

- ☐ there is a true value of θ that generated the data;
- ☐ this 'true' value of θ is to be treated as an unknown constant;
- ☐ probability statements concern randomness in hypothetical replications of the data (possibly conditioned on an ancillary statistic).

Bayesian viewpoint (cartoon version):

- ☐ all ignorance may be expressed in terms of probability statements;
- ☐ a joint probability distribution for data and all unknowns can be constructed;
- ☐ Bayes' theorem should be used to convert prior beliefs $\pi(\theta)$ about unknown θ into posterior beliefs $\pi(\theta \mid y)$, conditioned on data;
- ☐ probability statements concern randomness of unknowns, conditioned on all known quantities.

Mechanics

- Separate from data, we have prior information about parameter θ summarised in density $\pi(\theta)$
- Data model $f(y \mid \theta) \equiv f(y; \theta)$
- Posterior density given by Bayes' theorem:

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

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

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

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 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 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

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

  say.
- Inference about model choice is based on marginal posterior density

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

Inference

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

Sensitivity of the marginal likelihood

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

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

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.

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*;
  - physical parameters (means, variances, etc.) may be suitable for averaging, but care is needed.
- The predictive distribution for Z may be written

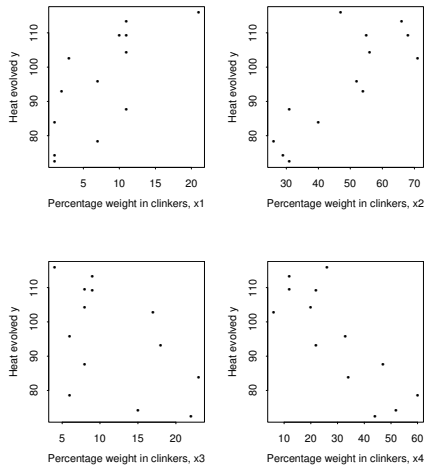
$$f(z \mid y) = \sum_{m=1}^{M} f(z \mid y, m) \Pr(m \mid y)$$

where

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

Example: Cement data

Percentage weights in clinkers of 4 four constitutents of cement $(x_1, \ldots, x_4)$ and heat evolved y in calories, in $n = 13$ samples.



Example: Cement data

```
> cement
   x1 x2 x3 x4     y
1   7 26  6 60  78.5
2   1 29 15 52  74.3
3  11 56  8 20 104.3
4  11 31  8 47  87.6
5   7 52  6 33  95.9
6  11 55  9 22 109.2
7   3 71 17  6 102.7
8   1 31 22 44  72.5
9   2 54 18 22  93.1
10 21 47  4 26 115.9
11  1 40 23 34  83.8
12 11 66  9 12 113.3
13 10 68  8 12 109.4
```

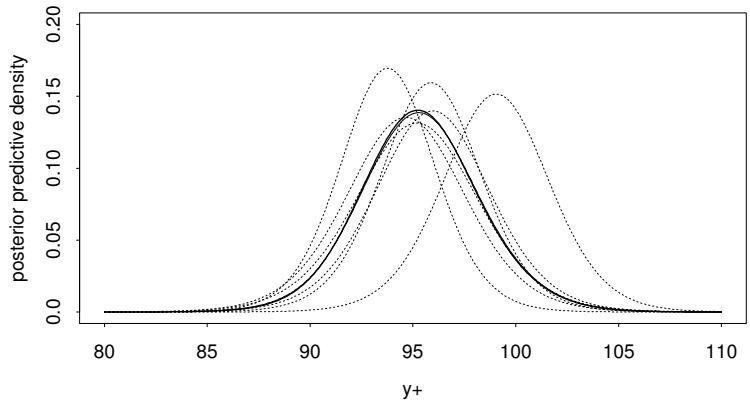

Example: Cement data

Bayesian model choice and prediction using model averaging for the cement data ($n = 13, p = 4$). For each of the 16 possible subsets of covariates, the table shows the log Bayes factor in favour of that subset compared to the model with no covariates and gives the posterior probability of each model. The values of the posterior mean and scale parameters a and b are also shown for the six most plausible models; $(y_+ - a)/b$ has a posterior t density. For comparison, the residual sums of squares are also given.

Model	RSS	$2 \log B_{10}$	$\Pr(M \mid y)$	a	b
– – – –	2715.8	0.0	0.0000		
1 – – –	1265.7	7.1	0.0000		
– 2 – –	906.3	12.2	0.0000		
– – 3 –	1939.4	0.6	0.0000		
– – – 4	883.9	12.6	0.0000		
1 2 – –	57.9	45.7	0.2027	93.77	2.31
1 – 3 –	1227.1	4.0	0.0000		
1 – – 4	74.8	42.8	0.0480	99.05	2.58
– 2 3 –	415.4	19.3	0.0000		
– 2 – 4	868.9	11.0	0.0000		
– – 3 4	175.7	31.3	0.0002		
1 2 3 –	48.11	43.6	0.0716	95.96	2.80
1 2 – 4	47.97	47.2	0.4344	95.88	2.45
1 – 3 4	50.84	44.2	0.0986	94.66	2.89
– 2 3 4	73.81	33.2	0.0004		
1 2 3 4	47.86	45.0	0.1441	95.20	2.97

Example: Cement data

Posterior predictive densities for cement data. Predictive densities for a future observation y_+ with covariate values x_+ based on individual models are given as dotted curves. The heavy curve is the average density from all 16 models.



DIC

- How to compare complex models (e.g. hierarchical models, mixed models, Bayesian settings), in which the 'number of parameters' may:
  - outnumber the number of observations?
  - be unclear because of the regularisation provided by a prior density?
- Suppose model has 'Bayesian deviance'

$$D(\theta) = -2 \log f(y \mid \theta) + 2 \log f(y)$$

for some normalising function $f(y)$, and suppose that samples from the posterior density of θ are available and give $\overline{\theta} = \mathrm{E}(\theta \mid y)$.

- One possibility is the **deviance information criterion (DIC)**

$$D(\overline{\theta}) + 2p_D,$$

where the number of associated parameters is

$$p_D = \overline{D(\theta)} - D(\overline{\theta}).$$

- This involves only (MCMC) samples from the posterior, no analytical computations, and reproduces AIC for some classes of models.

2. Beyond the Generalised Linear Model

slide 52

Overview

1. Generalised linear models
2. Overdispersion
3. Correlation
4. Random effects models

Generalised Linear Models

slide 54

GLM recap

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

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

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

GLM failure

However, situations frequently arise where GLMs do not adequately describe observed data.
This can be due to a number of reasons including:

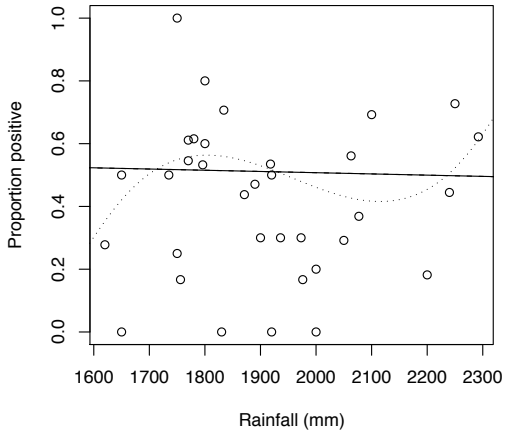
- ☐ The mean model cannot be appropriately specified as there is dependence on an unobserved (or unobservable) explanatory variable.
- ☐ There is excess variability between experimental units beyond that implied by the mean/variance relationship of the chosen response distribution.
- ☐ The assumption of independence is not appropriate.
- ☐ Complex multivariate structure in the data requires a more flexible model class

Example 1: toxoplasmosis

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

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

Example



Toxoplasmosis data and fitted models

Example

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

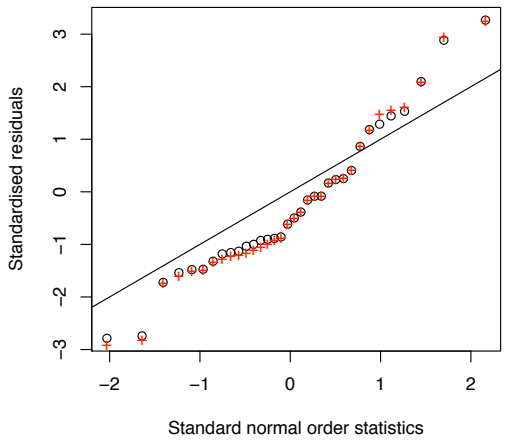
Model	df	deviance
Constant	33	74.21
Linear	32	74.09
Quadratic	31	74.09
Cubic	30	62.62

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

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

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

Example



Toxoplasmosis residual plot

Quasi-likelihood

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

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

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

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

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

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

Quasi-likelihood for toxoplasmosis data

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

Comparing cubic with constant model, one now obtains

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

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

Reasons

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

Possible reasons include:

- ☐ There may be an important explanatory variable, other than rainfall, which we haven't observed.
- ☐ Or there may be many other features of the cities, possibly unobservable, all having a small individual effect on incidence, but a larger effect in combination. Such effects may be individually undetectable – sometimes described as *natural excess variability between units*.

Reasons: unobserved heterogeneity

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

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

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

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

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

Direct models

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

For example, for the toxoplasmosis data, we might specify a **beta-binomial** model, where

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

leading to

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

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

Direct models: fitting

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

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

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

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

A random effects model for overdispersion

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

$$\eta_i = x_i^{\mathrm{T}}\beta + u_i \tag{5}$$

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

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

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

for some unknown σ^2.
We set $\mathrm{E}(u_i) = 0$, as an unknown mean for u_i would be unidentifiable in the presence of the intercept parameter β_0.

Random effects: likelihood

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

$$\begin{aligned} f(y \mid \beta, \sigma^2) &= \int f(y \mid \beta, u, \sigma^2) f(u \mid \beta, \sigma^2) \mathrm{d}u \\ &= \int f(y \mid \beta, u) f(u \mid \sigma^2) \mathrm{d}u \\ &= \int \prod_{i=1}^{n} f(y_i \mid \beta, u_i) f(u_i \mid \sigma^2) \mathrm{d}u_i \end{aligned} \tag{6}$$

where $f(y_i \mid \beta, u_i)$ arises from our chosen exponential family, with linear predictor (5) and $f(u_i \mid \sigma^2)$ is a univariate normal p.d.f.

Often no further simplification of (6) 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

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

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

Dependence: reasons

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

One compelling reason is the unobserved heterogeneity discussed previously.
In the 'correct' model (corresponding to η_i^{true}), the toxoplasmosis status of individuals, Y_{ij}, are independent, so

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

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

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

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

Random effects and dependence

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

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

which implies

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

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

Marginal models

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

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

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

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

- ☐ A random effects model describes the mean response at the subject level ('subject specific')
- ☐ A marginal model describes the mean response across the population ('population averaged')

GEEs

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

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

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

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

Consistent covariance estimates are available for GEE estimators.

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

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

Clustered data

Examples where data are collected in clusters include:

- ☐ Studies in biometry where **repeated measures** are made on experimental units. Such studies can effectively mitigate the effect of between-unit variability on important inferences.
- ☐ Agricultural field trials, or similar studies, for example in engineering, where experimental units are arranged within **blocks**
- ☐ Sample surveys where collecting data within clusters or **small areas** can save costs

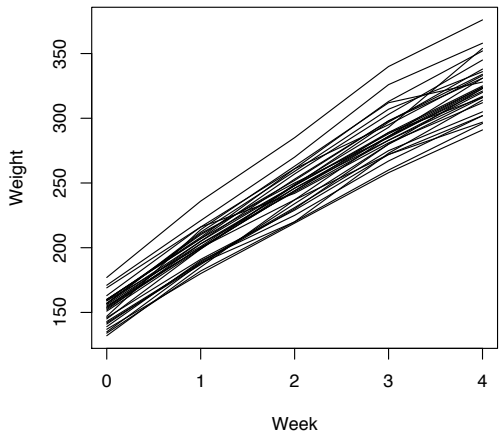
Of course, other forms of dependence exist, for example spatial or serial dependence induced by arrangement in space or time of units of observation.

Example 2: Rat growth

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

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

Example



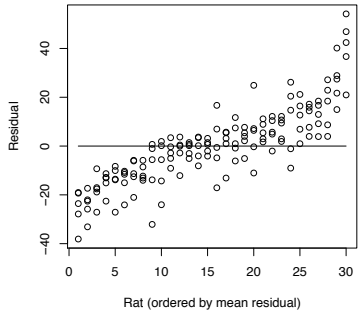
Rat growth data

A simple model

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

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

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

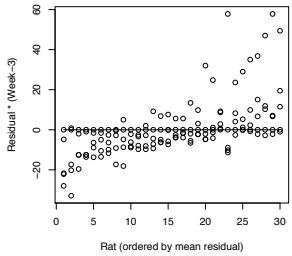


Model elaboration

Naively adding a (fixed) effect for animal gives

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

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



More complex cluster dependence required.

Linear mixed models

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

$$Y \sim N(\mu, \Sigma), \quad \mu = X\beta + Zb, \quad b \sim N(0, \Sigma_b), \tag{7}$$

where X and Z are matrices containing values of explanatory variables. Usually, $\Sigma = \sigma^2 I_n$.
A typical example for clustered data might be

$$Y_{ij} \overset{\text{ind}}{\sim} N(\mu_{ij}, \sigma^2), \quad \mu_{ij} = x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}} b_i, \quad b_i \overset{\text{ind}}{\sim} N(0, \Sigma_b^*), \tag{8}$$

where x_{ij} contain the explanatory data for cluster i, observation j and (normally) z_{ij} contains that sub-vector of x_{ij} which is allowed to exhibit extra between cluster variation in its relationship with Y. In the simplest (random intercept) case, $z_{ij} = (1)$, as in equation (5).

LMM example

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

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

This fits into the general LMM framework (7) with $\Sigma = \sigma^2 I_n$ and

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

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

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

Variance components

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

$$Y \sim N(X\beta, \Sigma + Z\Sigma_b Z^{\mathrm{T}}) \tag{9}$$

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

Other ways of describing LMMs for clustered data, such as (8) (and their generalised linear model counterparts) are as **hierarchical** models or **multilevel** models. This reflects the two-stage structure of the model, a conditional model for $Y_{ij} \mid b_i$, followed by a marginal model for the random effects b_i.

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

Discussion: Why random effects?

It would be perfectly possible to take a model such as (8) and ignore the final component, leading to fixed cluster effects (as we did for the rat growth data).

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

It also can be the case, as in (5) with only one observation per 'cluster', that fixed effects are not identifiable, whereas random effects can still be estimated. Similarly, some treatment variables must be applied at the cluster level, so fixed treatment and cluster effects are aliased.

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

Discussion: A Bayesian perspective

A Bayesian LMM supplements (7) with prior distributions for β, Σ and Σ_b.

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

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

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

LMM fitting

The likelihood for $(\beta, \Sigma, \Sigma_b)$ is available directly from (9) as

$$f(y \mid \beta, \Sigma, \Sigma_b) \propto |V|^{-1/2} \exp\left(-\tfrac{1}{2}(y - X\beta)^{\mathrm{T}} V^{-1}(y - X\beta)\right) \tag{10}$$

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

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

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

REML

In effect, REML maximizes the likelihood of any linearly independent sub-vector of $(I_n - H)y$ where $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ 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 (10) as

$$f(r \mid \Sigma, \Sigma_b) \propto f(y \mid \widehat{\beta}, \Sigma, \Sigma_b)|X^{\mathrm{T}}VX|^{1/2} \tag{11}$$

where

$$\widehat{\beta} = (X^{\mathrm{T}}V^{-1}X)^{-1}X^{\mathrm{T}}V^{-1}y \tag{12}$$

is the usual generalised least squares estimator given known V.
Having first obtained $(\widehat{\Sigma}, \widehat{\Sigma}_b)$ by maximising (11), $\widehat{\beta}$ is obtained by plugging the resulting $\widehat{V}$ into (12).
Note that REML maximised likelihoods cannot be used to compare different fixed effects specifications, due to the dependence of 'data' r in $f(r \mid \Sigma, \Sigma_b)$ on X.

Estimating random effects

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

$$\tilde{b} = \mathrm{E}(b \mid y) = (Z^{\mathrm{T}}\Sigma^{-1}Z + \Sigma_b^{-1})^{-1}Z^{\mathrm{T}}\Sigma^{-1}(y - X\beta) \tag{13}$$

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

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

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

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

Bayesian estimation: the Gibbs sampler

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

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

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

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

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

Bayesian estimation for LMMs

For the LMM

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

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

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

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

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

See Practical 3 for further details.

Example: Rat growth revisited

Here, we consider the model

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

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

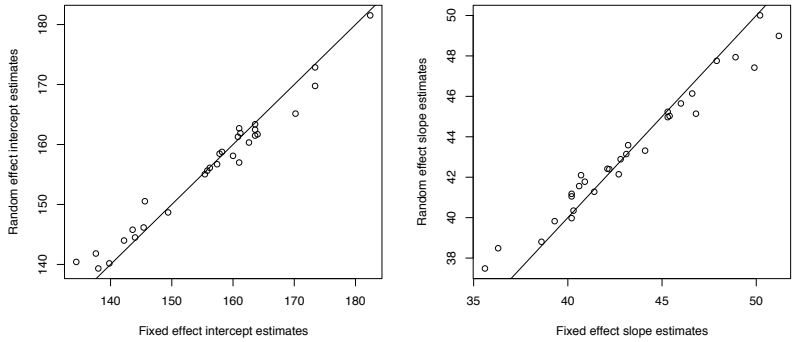
Estimates obtained by REML (ML in brackets) are

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

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

Example: Fixed v. random effect estimates

The shrinkage of random effect estimates towards a common mean is clearly illustrated.



Random effects estimates 'borrow strength' across clusters, due to the Σ_b^{-1} term in (13). Extent of this is determined by cluster similarity. This is usually considered to be a desirable behaviour.

Random effect shrinkage

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

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

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

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

and

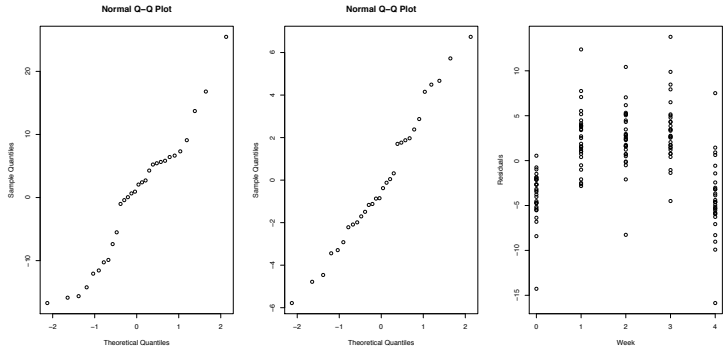
$$\mathrm{E}(\theta_j \mid y) = (1 - w)\mathrm{E}(\mu \mid y) + w y_j,$$

where

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

Example: Diagnostics

Normal Q-Q plots of intercept (panel 1) and slope (panel 2) random effects and residuals v. week (panel 3)



Evidence of a common quadratic effect, confirmed by AIC (1036 v. 1099) and BIC (1054 v. 1114) based on full ML fits. AIC would also include a cluster quadratic effect (BIC equivocal).

Generalised linear mixed models

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

$$Y_i \overset{\text{ind}}{\sim} F(\cdot \mid \mu_i, \sigma^2), \quad g(\mu) \equiv \begin{pmatrix} g(\mu_1) \\ \vdots \\ g(\mu_n) \end{pmatrix} = X\beta + Zb, \;\; b \sim N(0, \Sigma_b) \tag{14}$$

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

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

GLMM example

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

$$Y_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(\mu_i), \quad \log \frac{\mu_i}{1-\mu_i} = \beta_0 + b_{0i} + \beta_1 x_{ij}, \quad b_{0i} \overset{\text{ind}}{\sim} N(0, \sigma_b^2) \tag{15}$$

[note: no random slope here]. This fits into the general GLMM framework (14) with

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

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

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

GLMM likelihood

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

$$\begin{aligned} f(y \mid \beta, \Sigma_b) &= \int f(y \mid \beta, b, \Sigma_b) f(b \mid \beta, \Sigma_b) \mathrm{d}b \\ &= \int f(y \mid \beta, b) f(b \mid \Sigma_b) \mathrm{d}b \\ &= \int \prod_{i=1}^{n} f\left(y_i \mid g^{-1}([X\beta + Zb]_i)\right) f(b \mid \Sigma_b) \mathrm{d}b. \end{aligned} \tag{16}$$

For **nested** random effects structures, some simplification is possible. For example, for (15)

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

a product of one-dimensional integrals.

GLMM fitting: quadrature

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

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

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

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

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

GLMM fitting: Penalised quasi-likelihood

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

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

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

Now, for a GLMM,

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

and

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

where W is the weight matrix used in Fisher scoring.

GLMM fitting: PQL continued

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

$$v \sim N(X\beta + Zb, W^{-1}), \quad b \sim N(0, \Sigma_b) \tag{17}$$

where v and W also depend on β and b.

PQL proceeds by iteratively estimating β, b and Σ_b for the linear mixed model (17) for v, updating v and W at each stage, based on the current estimates of β and b.

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

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

GLMM fitting: discussion

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

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

- ☐ It can fail badly when the normal approximation that justifies it is invalid (for example for binary observations)
- ☐ As it does not use a full likelihood, model comparison should not be performed using PQL maximised 'likelihoods'

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

Bayesian estimation for GLMMs

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

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

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

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

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.

Toxoplasmosis revisited

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

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

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

Toxoplasmosis continued

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

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

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

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

3. Nonlinear Models

slide 106

Overview

1. Basic nonlinear models
2. Extending the nonlinear model
3. Computationally expensive nonlinear models
4. Model discrepancy

Basic nonlinear models

slide 108

Linear 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)

$$\begin{aligned} \mu_{ij} &= \mathrm{E}(y_{ij}) \\ g(\mu_{ij}) &= \eta_{ij} = x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}}b_i, \end{aligned}$$

  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.
- The GLMM has the following special cases
  - linear models;
  - generalised linear models (GLMs);
  - linear mixed models (LMMs).
- 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. However, they might be "useful".

Nonlinear models

- Begin by assuming that y has a normal distribution and the link function, g, is the identity link and $z_{ij} = 0$, i.e.

$$y_i = x_i^T\beta + \epsilon_i, \tag{18}$$

  where $\epsilon_i \sim N(0, \sigma^2)$, independently, where β are the p regression parameters.
- Consider extending this model so that instead of the mean response being the linear predictor $x_i^T\beta$, it is a nonlinear function of parameters, i.e.

$$y_i = \eta(x_i, \beta) + \epsilon_i, \tag{19}$$

  where $\epsilon_i \sim N(0, \sigma^2)$, independently, where β are the p nonlinear parameters.
- Obviously, the model specified by (19) has the linear model (18) as a special case when $\eta(x, \beta) = x^{\mathrm{T}}\beta$.
- *Note that, sometimes the term nonlinear model is used to describe any model which is not a linear model (18), which would include GLMs and GLMMs.*

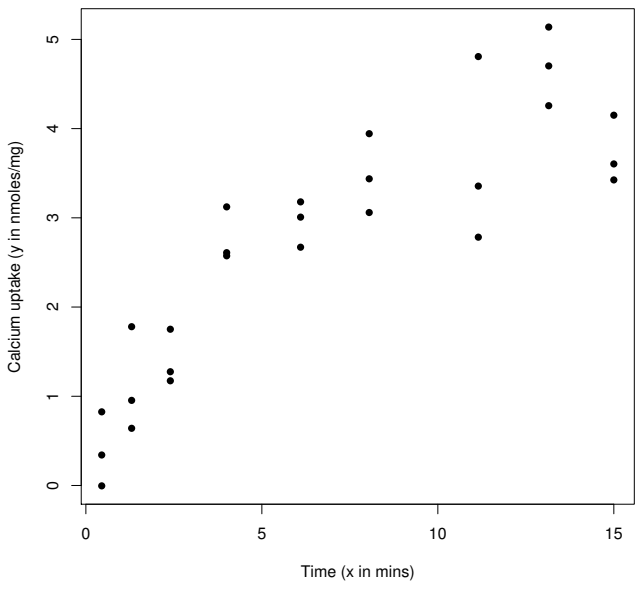
Example - Calcium Data

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

```
> calcium
       x        y
1   0.45  0.34170
2   1.30  1.77967
3   2.40  1.75136
4   4.00  3.12273
5   6.10  3.17881
6   8.05  3.05959
7  11.15  4.80735
8  13.15  5.13825
9  15.00  3.60407
10  0.45 -0.00438
11  1.30  0.95384
12  2.40  1.27497
13  4.00  2.60958
```


Example - Calcium Data

Plot of calcium uptake against time.



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 presence is often as a simplification of a more complex underlying system. Their estimation is to make the model fit best to reality.

In most cases, in a linear model, the regression parameters are tuning parameters.

Advantages and disadvantages

Advantages

- Can incorporate prior scientific knowledge through the function $\eta(x, \beta)$.
- Can fit simpler models, i.e. less parameters, to adequately describe observed data than by using linear models.
- Can provide extrapolated predictions (typically discouraged for linear models).
- Can directly contribute to scientific understanding through the estimation of physical parameters.

Disadvantages

- Need to specify the function $\eta(x, \beta)$.
- There are computational problems associated with these models.
- *All models are wrong*.

Specifying $\eta(x, \beta)$

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 - Calcium Data

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

$$\eta(x, \beta) = \beta_0 \left(1 - \exp\left(-x/\beta_1\right)\right). \tag{20}$$

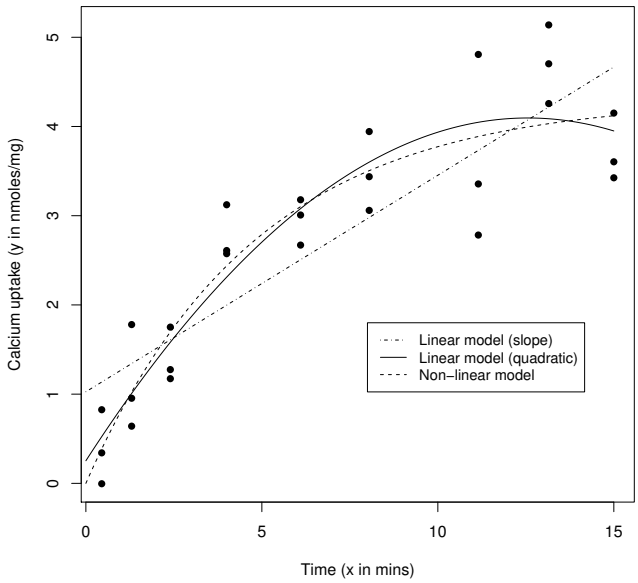
- This is derived by assuming that the rate of growth is proportional to the calcium remaining, i.e.

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

- The solution to this differential equation is (20).
- Interpretation of parameters:
  - β_0 - final size of population;
  - β_1 - (inversely) controls growth rate.

Example - Calcium Data

- Plot of calcium uptake against time.
- Includes fitted lines for three different models



Example - Calcium Data

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

Model	Parameters (p)	$l(\widehat{\theta})$	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 (less parameters), it is the preferred model.

Extending the nonlinear model

Introduction

- Nonlinear models can be extended to
  1. non-normal responses;
  2. clustered responses;

  in the same way as linear models.
- Here, we consider clustered responses and briefly discuss the nonlinear mixed model.

Example - Theophylline

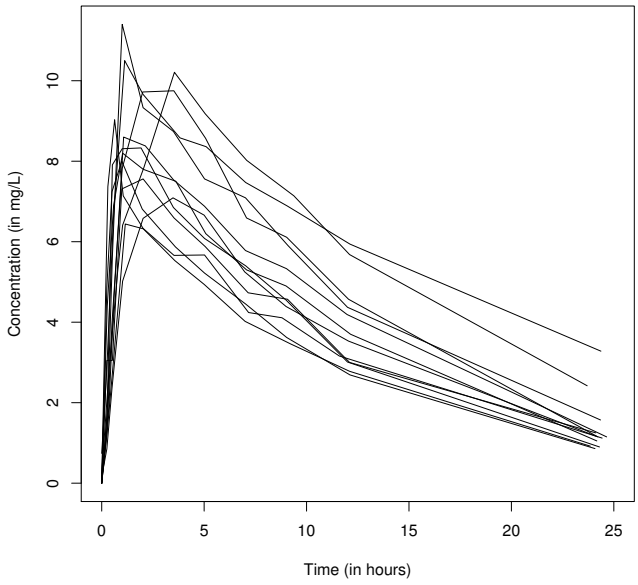
- Theophylline is an anti-asthmatic drug.
- An experiment was performed on $n = 12$ individuals to investigate the way in which the drug leaves the body.
- The study of drug concentrations inside organisms is called *pharmacokinetics*.
- An oral dose, D_i, was given to the ith individual at time $t = 0$, 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

```
> Theoph
Grouped Data: conc ~ Time | Subject
    Subject   Wt Dose  Time  conc
1         1 79.6 4.02  0.00  0.74
2         1 79.6 4.02  0.25  2.84
3         1 79.6 4.02  0.57  6.57
4         1 79.6 4.02  1.12 10.50
5         1 79.6 4.02  2.02  9.66
6         1 79.6 4.02  3.82  8.58
7         1 79.6 4.02  5.10  8.36
8         1 79.6 4.02  7.03  7.47
9         1 79.6 4.02  9.05  6.89
10        1 79.6 4.02 12.12  5.94
11        1 79.6 4.02 24.37  3.28
12        2 72.4 4.40  0.00  0.00
13        2 72.4 4.40  0.27  1.72
```


Example - Theophylline

Plot of concentration of theophylline against time for each of the individuals.



There is a sharp increase in concentration followed by a steady decrease.

Example - Theophylline

- 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

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

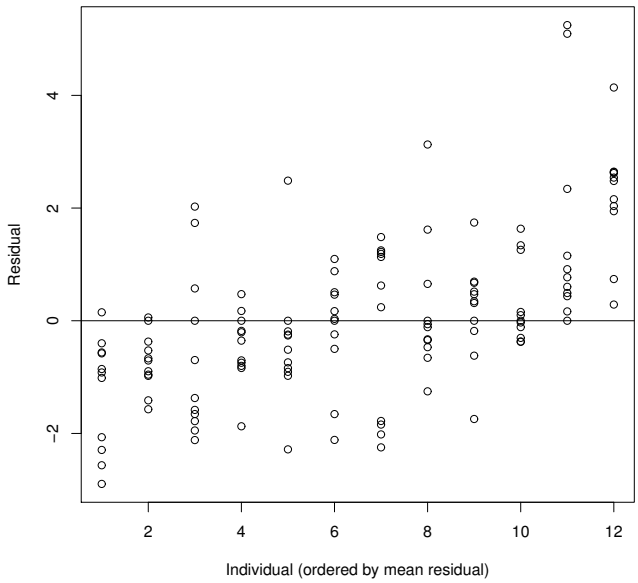
  where the (positive) nonlinear parameters are
  - β_1 is the elimination rate and controls the rate at which the drug leaves the organism;
  - β_2 is the absorption rate and controls the rate at which the drug enters the blood;
  - β_3 is the clearance and controls the volume of blood for which a drug is completely removed per time unit.

Example - Theophylline

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

$$\begin{aligned} \epsilon_{ij} &\sim N(0, \sigma^2), \\ b_i &\sim N(0, \Sigma_b), \end{aligned}$$

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.
- Obviously, $\beta_i \sim N(\beta, \Sigma_b)$. The mean, β, of the cluster-specific nonlinear parameters across all individuals are the population nonlinear parameters.

Non-linear mixed effects models

- 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

$$\begin{aligned} \epsilon_{ij} &\sim N(0, \sigma^2) \\ b_i &\sim N(0, \Sigma_b), \end{aligned}$$

where 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 of linear mixed models

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

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

- Then

$$\begin{aligned} \eta(\beta + Abx) &= x^{\mathrm{T}}(\beta + Ab) \\ &= x^{\mathrm{T}}\beta + x^{\mathrm{T}}Ab, \end{aligned}$$

so $z = A^{\mathrm{T}}x$.

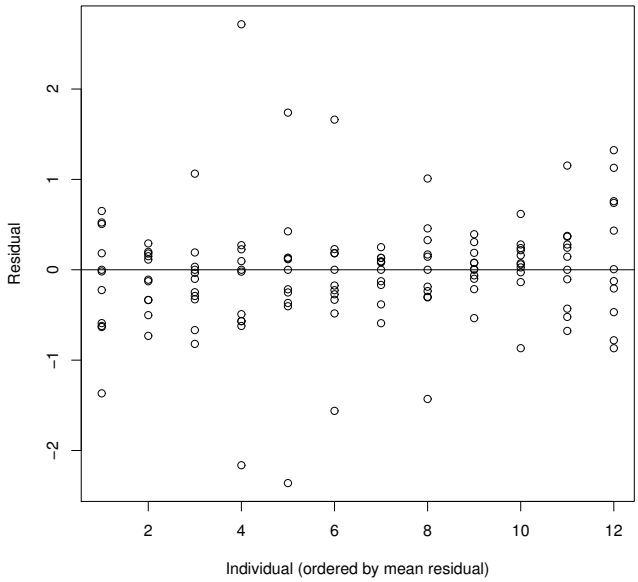
- For a random intercept model, where $q = 1$, $A = (1, 0, \ldots, 0)$.

Example - Theophyline

- Returning to the Theophyline example, we fit the nonlinear mixed model, allowing all of the nonlinear parameters to vary across individuals, i.e. $A = I_3$.
- Estimates:

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

- $\mathrm{AIC} = 372.6$



Example - Theophyline

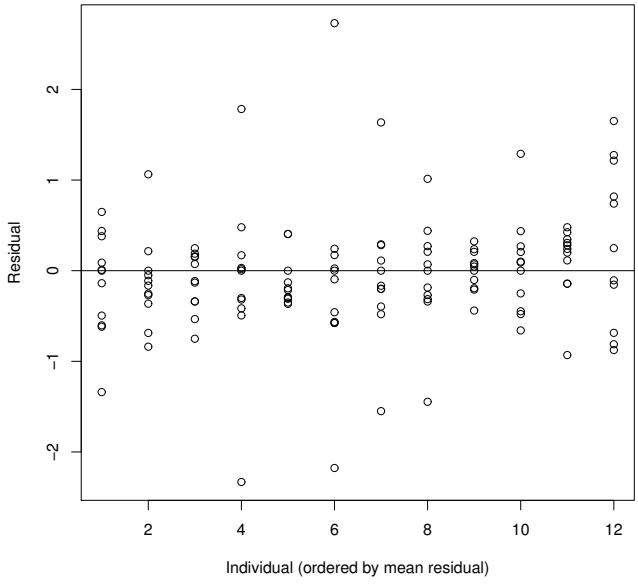
- The estimated value of Σ_{b11} is "small"so we fit the nonlinear mixed model, allowing absorption rate and clearance to vary across individuals, i.e.

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

- Estimates:

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

- $\mathrm{AIC} = 368.6$
- No further model simplifications reduce the AIC.



Extensions to nonnormal responses

- Nonlinear models can be extended to nonnormal responses in the same way as linear models.
- The most general model is the generalised nonlinear mixed model (GNLMM).
- y_{ij} is from exponential family.
- $\mathrm{E}(y_{ij}) = \mu_{ij}$.
- $g(\mu_{ij}) = \eta(\beta + Ab_i, x_{ij})$.
- 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

There are various technical and practical issues related to fitting nonlinear models (some of these are common to GLMs and GLMMs).

- ☐ Approximation of likelihood function (random effects are integrated out)
- ☐ Convergence of optimisation routines to find estimates
- ☐ Existence of estimates
- ☐ Reliability of asymptotic inference
- ☐ Computational expense of evaluating $\eta(\beta, x)$.
- ☐ *All models are wrong*.

These are all areas of current research.

Computationally expensive nonlinear models slide 133

Computationally expensive nonlinear models

- ☐ It typically requires many evaluations of $\eta(\beta, x)$ to fit a nonlinear model, either to find the estimate of β or to generate an MCMC sample.
- ☐ What happens if the non-linear model $\eta(\beta, x)$ is computationally expensive?
- ☐ For example, $\eta(\beta, x)$ could be the numerical solution to a system of differential equations where the exact solution is not available in closed form.
- ☐ The numerical solution to $\eta(\beta, x)$, implemented in computer code, is computationally expensive to evaluate.
- ☐ This can render model fitting to be infeasible.

Computer experiments and emulators

- ☐ One approach is to develop an approximation, $\widehat{\eta}(\beta, x)$, to the non-linear model $\eta(\beta, x)$.
- ☐ Evaluation of $\widehat{\eta}(\beta, x)$ replaces evaluation of $\eta(\beta, x)$ in all model fitting procedures.
- ☐ The approximation is typically called an *emulator* or *surrogate*.
- ☐ How is such an emulator constructed?
- ☐ Answer: via a computer experiment. This topic is briefly discussed here and will be covered in much more detail on the APTS Week 4 module: **Design of Experiments and Studies**

Computer experiments and emulators

- Let $z = (\beta, x)$ be the inputs to the nonlinear model such that $\eta(\beta, x) = \eta(z)$. Let d be the dimension of z.
- The nonlinear model is evaluated at a "small" number, m, of inputs

$$\zeta = \left\{z^1, \ldots, z^m\right\},$$

  where $z^i = (\beta^i, x^i)$, for $i = 1, \ldots, m$.
- Finally, let $\eta^i = \eta(z^i)$, for $i = 1, \ldots, m$ and $\eta = \left(\eta^1, \ldots, \eta^m\right)$.

Gaussian Process Emulators

- The most commonly-used emulator is a Gaussian Process (GP) emulator.
- Here any finite collection of evaluations of $\eta(z)$ is assumed to have a multivariate normal distribution.
- Suppose $\eta_0 = \eta(z^0) = \eta(\beta^0, x^0)$ is the value of the nonlinear model we wish to predict.
- Assumption:

$$\begin{pmatrix} \eta \\ \eta^0 \end{pmatrix} \sim N\left(\begin{pmatrix} \theta \\ \vdots \\ \theta \end{pmatrix}, \tau^2 \begin{pmatrix} C & c^{\mathrm{T}} \\ c & 1 \end{pmatrix}\right),$$

  i.e. a multivariate normal with (marginally) common mean θ and variance τ^2.
- Note that

C_{ij} – correlation between $\eta(z^i)$ and $\eta(z^j)$

c_i – correlation between $\eta(z^i)$ and $\eta(z^0)$

Gaussian Process Emulators

- By the properties of the multivariate normal

$$\eta^0 | \eta \sim N\left(\theta + c^{\mathrm{T}} C^{-1}(\eta - \theta 1_m), \tau^2\left(1 - c^{\mathrm{T}} C^{-1} c\right)\right),$$

  where 1_m is a vector of m ones.
- Structure is imposed on the elements of C and c as follows

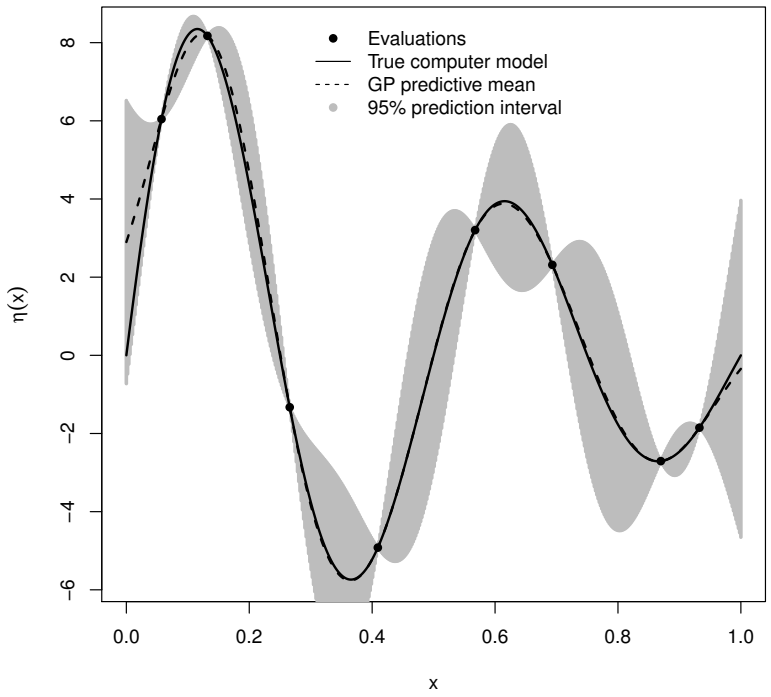
$$\begin{aligned} C_{ij} &= \kappa(z^i, z^j; \rho) \\ c_{ij} &= \kappa(z^i, z^0; \rho) \end{aligned}$$

  where $\kappa(\cdot, \cdot; \rho)$ is a correlation function depending on ρ.
- A commonly-used correlation function is the squared exponential:

$$\kappa(z^i, z^j; \rho) = \exp\left(-\sum_{k=1}^{d} \rho_i \left(z_k^i - z_k^j\right)^2\right).$$

- θ, τ^2 and ρ can be estimated via maximum likelihood or a Bayesian approach taken.

Example



Model discrepancy

Model discrepancy

- Under the basic nonlinear model we are assuming

$$y = \eta(\beta, x) + \epsilon,$$

  i.e. the observed responses are given by the nonlinear model plus some random error.
- However, all models are wrong.
- In Chapter 1, we accounted for this by considering more complex models.
- If $\eta(\beta, x)$ is a mechanistic model, there is not really scope to make it more complex.

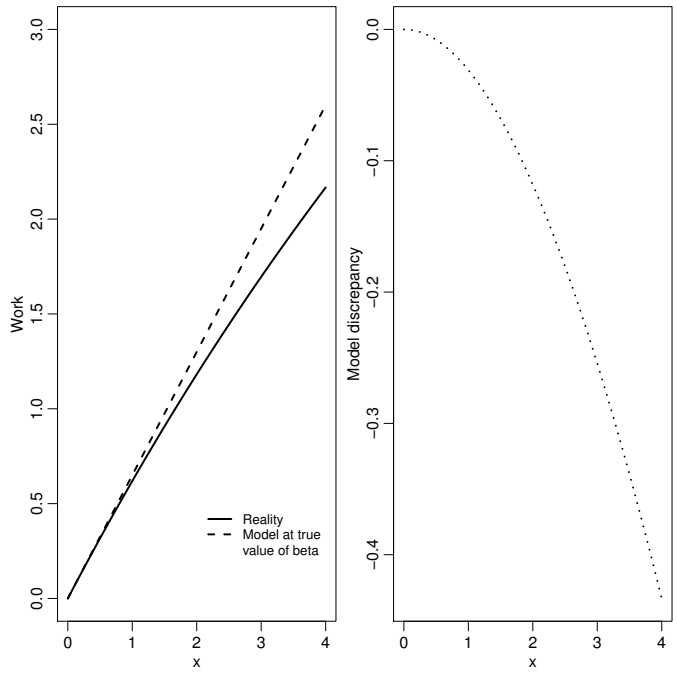
Model discrepancy

- Let $\mu(x)$ be true system depending on x.
- We observe
$$y = \mu(x) + \epsilon.$$
- $\eta(\beta, x)$ is our model and is our best guess at $\mu(x)$ where β are the true value of the parameters.
- Assume
$$\mu(x) = \eta(\beta, x) + \delta(x),$$
where $\delta(x)$ is the difference between reality and our model, i.e. the model discrepancy.
- Therefore
$$y = \eta(\beta, x) + \delta(x) + \epsilon.$$
- The model discrepancy is an unknown function.
- Taking a Bayesian approach, a prior is placed on this function. In particular, the Kennedy & O'Hagan (2001) framework places a Gaussian process prior on this function.
- Therefore, we are explicitly modelling the model discrepancy.

Example - Illustrative

- This example is adapted from Brynjarsdottir & O'Hagan (2014).
- Suppose reality is such that
$$\mu(x) = \frac{\beta x}{1+x/20},$$
where $\beta = 0.65$ is the true value of nonlinear parameter.
- Our model is such that
$$\eta(\beta, x) = \beta x.$$
- The model discrepancy is then
$$\delta(x) = \frac{x}{20+x}.$$

Example - Illustrative

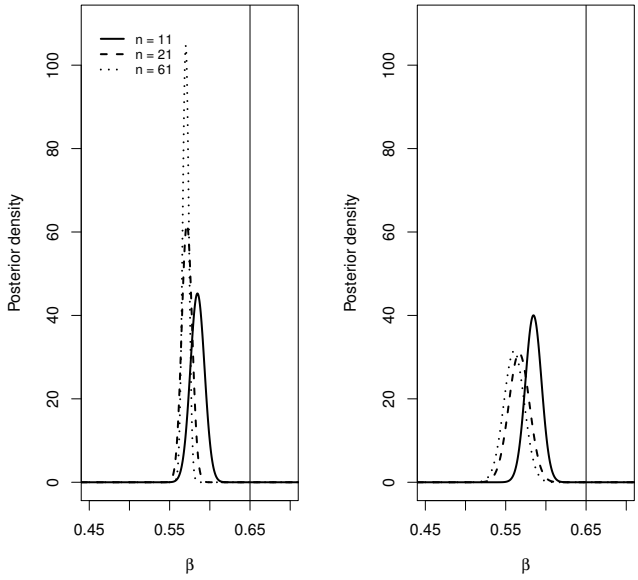


Example - Illustrative

- ☐ Consider an experiment to achieve the three aims of
  1. Interpolation prediction, i.e. predict value of y for $x = 2$;
  2. Extrapolation prediction, i.e. predict value of y for $x = 6$;
  3. Estimate value of β.
- ☐ We observe the response y at n values of $x \in [0, 4]$.
- ☐ We take two different approaches:
  1. Ignore mode discrepancy - just fit basic nonlinear model;
  2. Use Kennedy & O'Hagan framework with nonlinear model with model discrepancy.

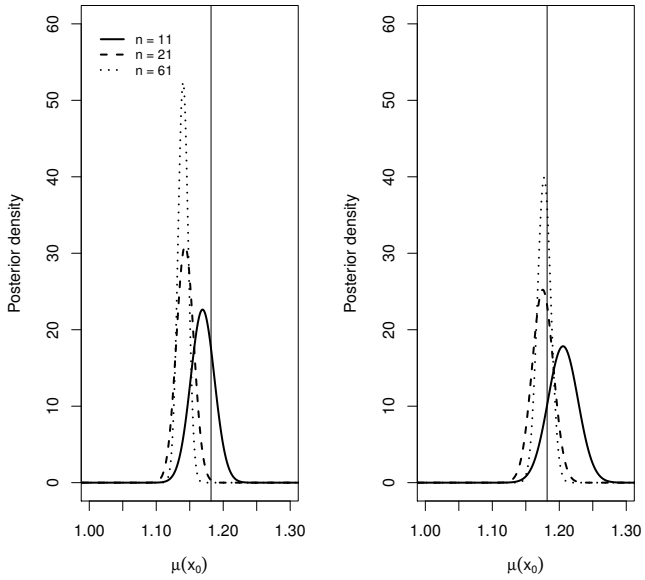
Estimation

Posterior density of β under basic nonlinear model (left) and nonlinear model with model discrepancy (right).



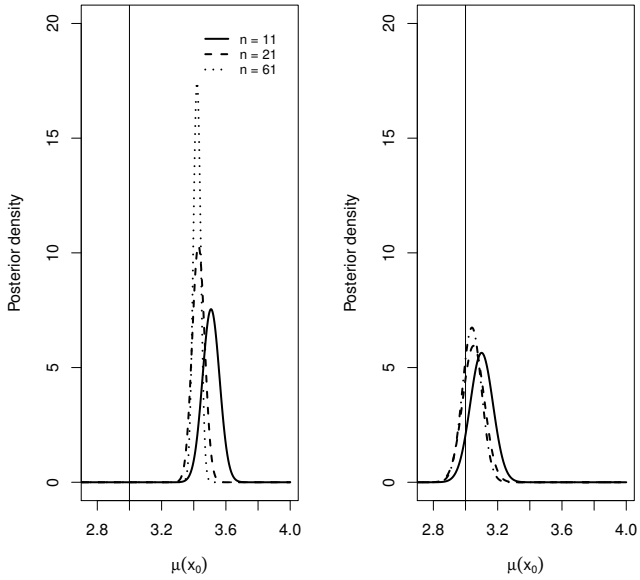
Example - Interpolation

Posterior density of $\mu(x_0)$ (with $x_0 = 2$) under basic nonlinear model (left) and nonlinear model with model discrepancy (right).



Example - Extrapolation

Posterior density of $\mu(x_0)$ (with $x_0 = 6$) under basic nonlinear model (left) and nonlinear model with model discrepancy (right).



Model discrepancy - discussion

- ☐ Modelling the model discrepancy with a Gaussian process alleviated the problem with interpolation prediction but not for extrapolation prediction or parameter estimation.
- ☐ Brynjarsdottir & O'Hagan (2014) considered using a constrained Gaussian process to incorporate prior information on the model discrepancy (e.g. value at $x = 0$ and monotoncity) and this eased the problem for parameter estimation but not for extrapolation prediction.
- ☐ How to account for model discrepancy remains an open research problem.