Statistical Modelling

Antony Overstall

(Chapters 1–3 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

slide 1

Overview

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

Basic Ideas

slide 3

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 ...

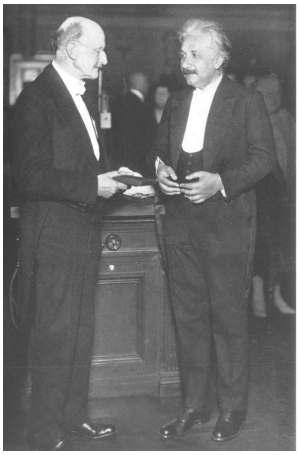
Motivation

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.

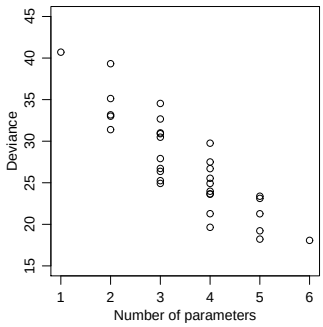
m	r	age	stage	grade	xray	acid
6	5	0	1	1	1	1
6	1	0	0	0	0	1
4	0	1	1	1	0	0
4	2	1	1	0	0	1
4	0	0	0	0	0	0
3	2	0	1	1	0	1
3	1	1	1	0	0	0
3	0	1	0	0	0	1
3	0	1	0	0	0	0
2	0	1	0	0	1	0
2	1	0	1	0	0	1
2	1	0	0	1	0	0
1	1	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	
1	1	0	0	1	0	1
1	0	0	0	0	1	1
1	0	0	0	0	1	0

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

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

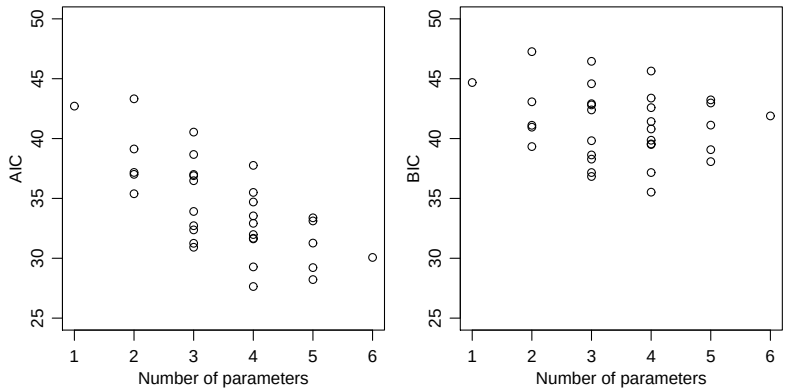
where E_g denotes expectation over the data distribution g. The bias is positive because I and K are positive definite matrices.

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 slide 23

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

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

  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

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

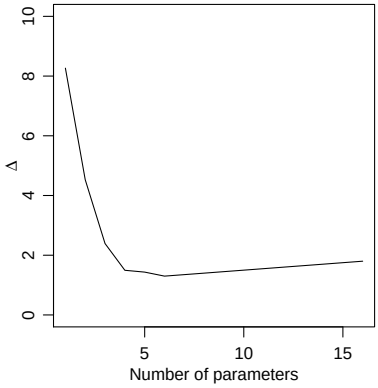
Therefore

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

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.

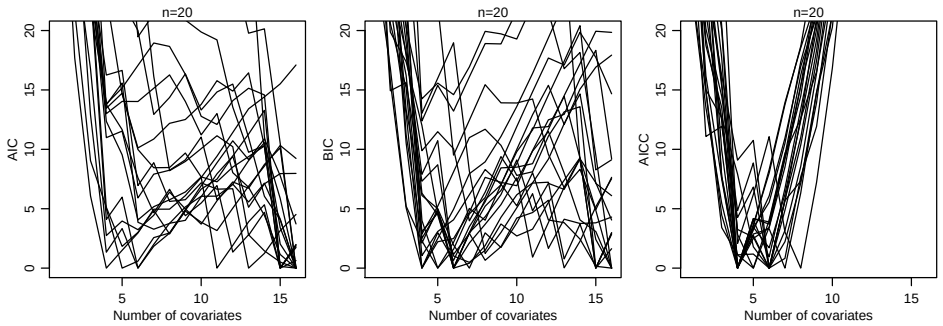
Simulation experiment

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

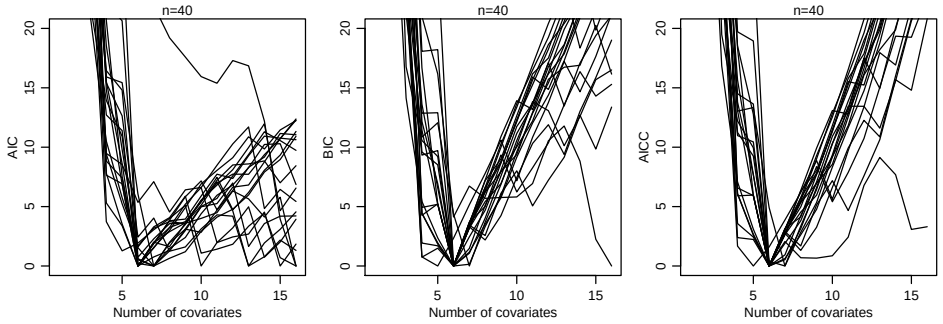
Simulation experiment

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



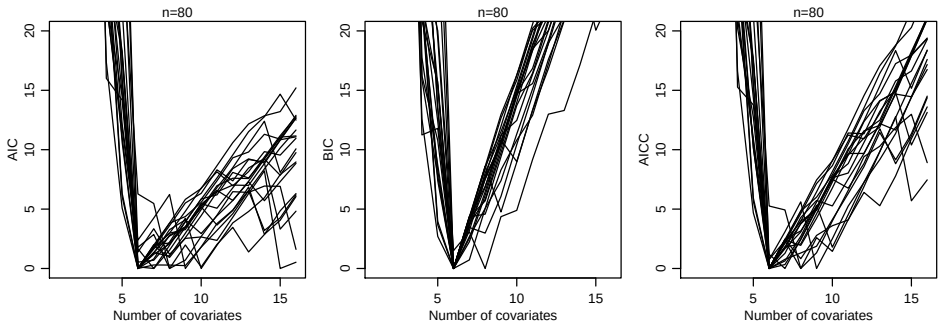
Simulation experiment

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



Simulation experiment

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



As n increases, note how

- ☐ AIC and $\mathrm{AIC_c}$ still allow some over-fitting, but BIC does not, and
- ☐ $\mathrm{AIC_c}$ 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

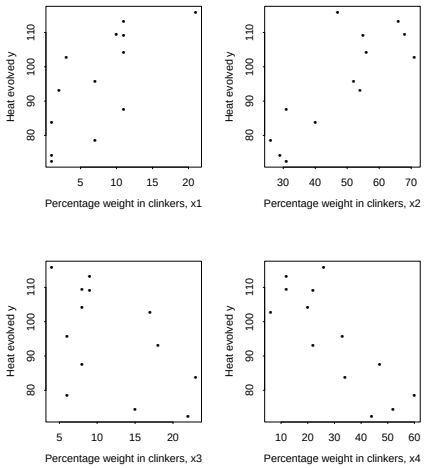
$$f(z \mid y) = \sum_{m=1}^{M} f(z \mid y, m) \Pr(m \mid y)$$

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
```

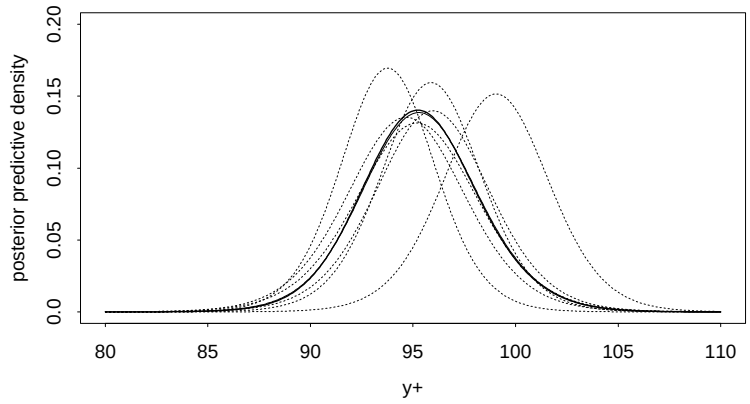

Example: Cement data

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:

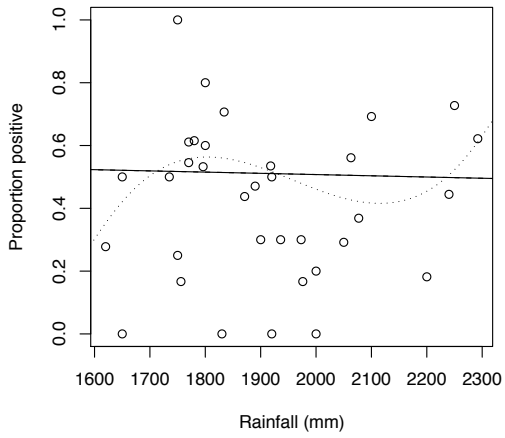
- ☐ The mean model cannot be appropriately specified as there is dependence on an unobserved (or unobservable) explanatory variable.
- ☐ There is excess variability between experimental units beyond that implied by the mean/variance relationship of the chosen response distribution.
- ☐ The assumption 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:

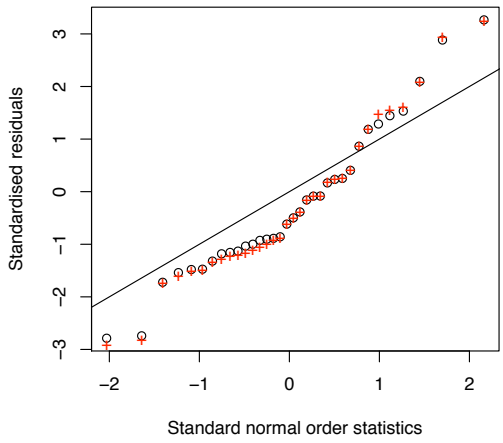
Model	df	deviance
Constant	33	74.21
Linear	32	74.09
Quadratic	31	74.09
Cubic	30	62.62

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

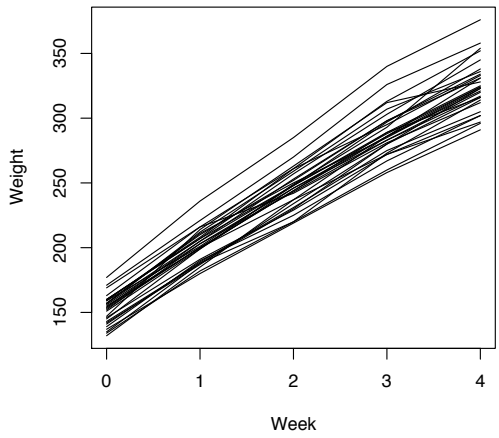
Of course, other forms of dependence exist, for example spatial or serial dependence induced by arrangement in space or time of units of observation.

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



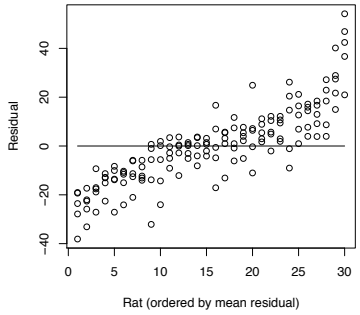
Rat growth data

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

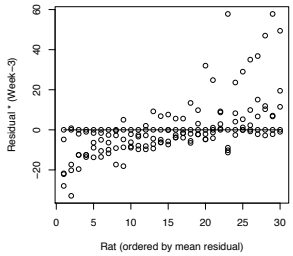


Model elaboration

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.

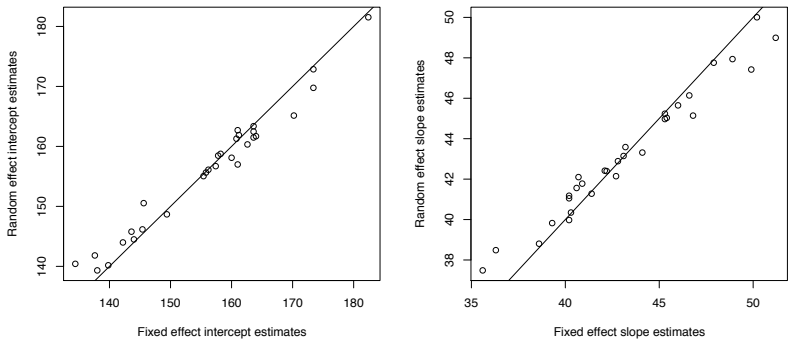
Estimates obtained by REML (ML in brackets) are

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

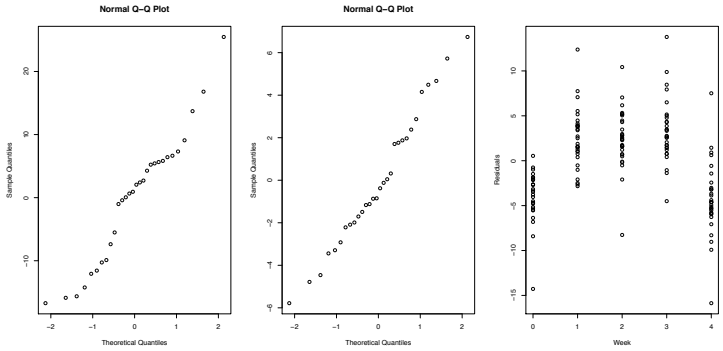
$$\mathrm{E}(\theta_j \mid y) = (1 - w)\mathrm{E}(\mu \mid y) + w y_j,$$

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. Design of Experiments

Overview

1. Introduction and principles of experimentation
2. Factorial designs
3. Regular fractional factorial designs
4. D-optimality and non-regular designs
5. Approximate designs

Introduction and principles of experimentation

Modes of data collection

- ☐ Observational studies
- ☐ Sample surveys
- ☐ Designed experiments

Definition: An experiment is a procedure whereby controllable factors, or features, of a system or process are deliberately varied in order to understand the impact of these changes on one or more measurable responses.

- ☐ Agriculture
- ☐ Industry
- ☐ Laboratory and in silico

Ronald A. Fisher (1890 – 1962)

Role of experimentation

Why do we experiment?

- ☐ Key to the scientific method (hypothesis – **experiment** – observe – infer – conclude)
- ☐ Potential to establish **causality**...
- ☐ ... and to understand and improve complex systems depending on many factors
- ☐ Comparison of treatments, factor screening, prediction, optimisation, ...

Design of experiments: a statistical approach to the arrangement of the operational details of the experiment (e.g. sample size, specific experimental conditions investigated, ...) so that the quality of the answers to be derived from the data is as high as possible

Definitions

- **Treatment** – entities of scientific interest to be studied in the experiment
  e.g. varieties of crop, doses of a drug, combinations of temperature and pressure
- **Unit** – smallest subdivision of the experimental material such that two units may receive different treatments
  e.g. plots of land, subjects in a clinical trial, samples of reagent
- **Run** – application of a treatment to a unit

A unit-treatment statistical model

$$y_{ij} = \tau_i + \varepsilon_{ij}$$

$i = 1, \ldots, t$; $j = 1, \ldots, n_t$

- y_{ij} – measured response arising from the jth unit to which treatment i has been applied
- τ_i – treatment effect: expected response from application of the ith treatment
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ – random deviation from the expected response

The aims of the experiment will often be achieved by estimating comparisons between the treatment effects, $\tau_k - \tau_l$

Experimental precision and accuracy are largely obtained through **control** and **comparison**

Example

Fabrication of integrated circuits (Wu & Hamada, 2009, p.155)

- An initial step in fabricating integrated circuits is the growth of a epitaxial layer on polished silicon wafers via chemical deposition
- Unit
  - A set of six wafers (mounted in a rotating cylinder)
- Treatment
  - A combination of settings of the factors:
    - A: rotation method (x_1)
    - B: nozzle position (x_2)
    - C: deposition temperature (x_3)
    - D: deposition time (x_4)

Principles of experimentation

- Replication
  - The application of each treatment to multiple experimental units
    - Provides an estimate of experimental error against which to judge treatment differences
    - Reduces the variance of the estimators of treatment differences

Principles of experimentation

- ☐ Randomisation
  - Randomise allocation of units to treatments, order in which the treatments are applied, ...
    - ▷ Protects against lurking (uncontrolled) variables and subjectively in allocation of treatments to units
- ☐ Blocking
  - Account for systematic differences between batches of experimental units by arranging them in homogeneous blocks
    - ▷ If the same treatment is applied to all units, within-block variation in the response would be much less than between-block
    - ▷ Compare treatments within the same block and hence eliminate block effects

Factorial designs

Example revisited

Fabrication of integrated circuits (Wu & Hamada, 2009, p.155)

- ☐ An initial step in fabricating integrated circuits is the growth of a epitaxial layer on polished silicon wafers via chemical deposition
- ☐ Unit
  - A set of six wafers (mounted in a rotating cylinder)
- ☐ Treatment
  - A combination of settings of the factors:
    - ▷ A: rotation method (x_1)
    - ▷ B: nozzle position (x_2)
    - ▷ C: deposition temperature (x_3)
    - ▷ D: deposition time (x_4)
- ☐ **Assume each factor has two-levels, coded -1 and +1**

Treatments and a regression model

- Each factor has two levels, $x_k = \begin{cases} -1 \\ +1 \end{cases}$, $k = 1, 2, 3, 4$
- A treatment is then defined as a combination of four values of -1, $+1$
  - E.g. $x_1 = -1$, $x_2 = -1$, $x_3 = +1$, $x_4 = -1$
  - Specifies the settings of the process
- Assume each treatment effect is determined by a regression model in the four factors, e.g.

$$\tau = \beta_0 + \sum_{i=1}^{4} \beta_i x_i + \sum_{i=1}^{4} \sum_{j>i}^{4} \beta_{ij} x_i x_j + \sum_{i=1}^{4} \sum_{j>i}^{4} \sum_{k>j}^{4} \beta_{ijk} x_i x_j x_k + \beta_{1234} x_1 x_2 x_3 x_4$$

(Two-level) Factorial design

Run	x_1	x_2	x_3	x_4	$\overline{y}$
1	-1	-1	-1	-1	13.59
2	-1	-1	-1	+1	14.59
3	-1	-1	+1	-1	14.05
4	-1	-1	+1	+1	14.24
5	-1	+1	-1	-1	13.94
6	-1	+1	-1	+1	14.65
7	-1	+1	+1	-1	14.14
8	-1	+1	+1	+1	14.40
9	+1	-1	-1	-1	13.72
10	+1	-1	-1	+1	14.67
11	+1	-1	+1	-1	13.90
12	+1	-1	+1	+1	13.84
13	+1	+1	-1	-1	13.88
14	+1	+1	-1	+1	14.56
15	+1	+1	+1	-1	14.11
16	+1	+1	+1	+1	14.30

- Treatments in standard order
- $\overline{y}$ - average response from the six wafers

Regression model

Regression model and least squares

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

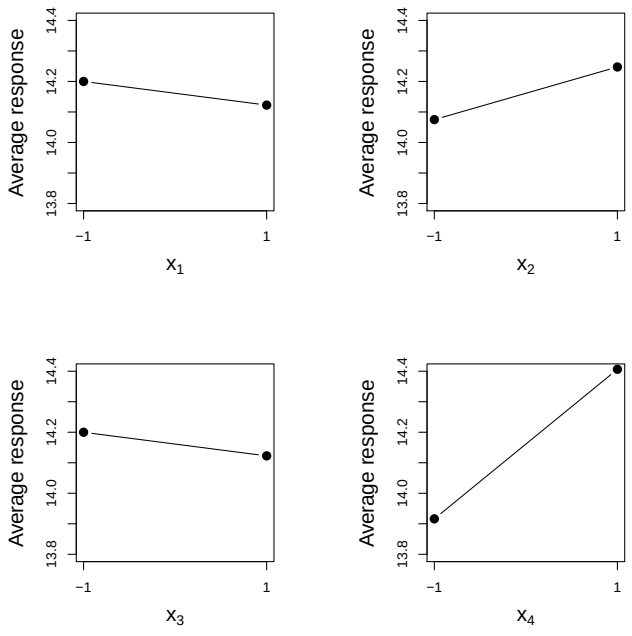
$$\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$$

- ☐ $n = 16$, $p = 16$
- ☐ Model matrix X contains intercept, linear and cross-product terms (up to 4th order)
- ☐ Information matrix $X^{\mathrm{T}}X = nI$
  - $\widehat{\beta} = \frac{1}{n}X^{\mathrm{T}}Y$
  - Regression coefficients are estimated by independent contrasts in the data

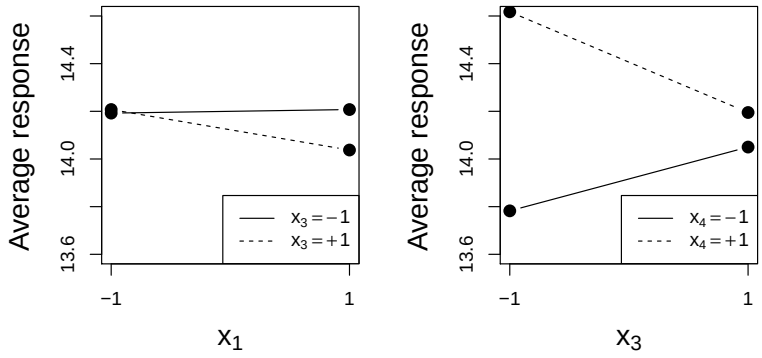
Main effects and interactions

- ☐ Main effect of x_i = [Average response when $x_i = 1$] − [Average response when $x_i = -1$]
- ☐ Interaction between x_i and x_j = [Average response when $x_i x_j = 1$] − [Average response when $x_i x_j = -1$]
- ☐ Main effect of $x_i = 2\beta_i$
- ☐ Interaction between x_i and $x_j = 2\beta_{ij}$
- ☐ Higher order interactions defined similarly

Main effects

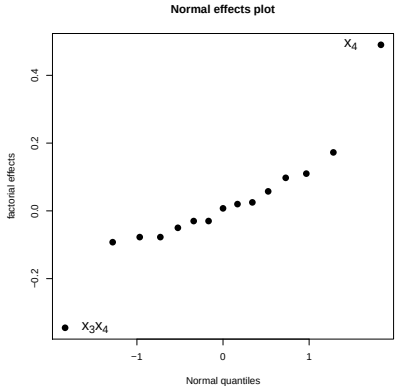


Interactions



Orthogonality

- $X^{\mathrm{T}}X = nI \Rightarrow \widehat{\beta}$ are independently normally distributed with equal variance
- Hence, can treat the identification of important effects (e.g. non-zero β) as an outlier identification problem



- Plot ordered factorial effects against quantiles from a standard normal
- Outlying effects are identified as important

Replication

- An unreplicated factorial design provides no model-independent estimate of σ^2 (Gilmour & Trinca, 2012, JRSSC)
  - Any unsaturated model does provide an estimate, but it may be biased by ignored (significant) model terms
  - This is one reason why graphical (or associated) analysis methods are popular
- Replication increases the power of the design
  - Common to replicate a centre point
    - Provides a portmanteau test of curvature
    - Allows unbiased estimation of σ^2

Principles of factorial experimentation

- Effect sparsity
  - The number of important effects in a factorial experiment is small relative to the total number of effects investigated (cf Box & Meyer, 1986, Technometrics)
- Effect hierarchy
  - Lower-order effects are more likely to be important than higher-order effects
  - Effects of the same order are equally likely to be important
- Effect heredity
  - Interactions where at least one parent main effect is important are more likely to be important themselves

Regular fractional factorial designs slide 127

Example

Production of bacteriocin (Morris, 2011, p.231)

- Bacteriocin is a natural food preservative grown from bacteria
- Unit
  - A single bio-reaction
- Treatment
  - A combination of settings of the factors:
    - A: amount of glucose (x_1)
    - B: initial inoculum size (x_2)
    - C: level of aeration (x_3)
    - D: temperature (x_4)
    - E: amount of sodium (x_5)
- Assume each factor has two-levels, coded -1 and $+1$

Choosing subsets of treatments

- Factorial designs can require a large number of runs for only a moderate number of factors ($2^5 = 32$)
- Resource constraints (e.g. cost) may mean not all 2^m combinations can be run
- Lots of degrees of freedom are devoted to estimating higher-order interactions
  - e.g. in a 2^5 experiment, 16 d.f. are used to estimate 3 factor and higher-order interactions
  - Principles of effect hierarchy and sparsity suggest this is wasteful
- Need to trade-off **what you want to estimate** against the **number of runs** you can afford

Example

Run	x_1	x_2	x_3	x_4	x_5
1	-1	-1	-1	+1	+1
2	-1	-1	+1	-1	-1
3	-1	+1	-1	+1	-1
4	-1	+1	+1	-1	+1
5	+1	-1	-1	-1	+1
6	+1	-1	+1	+1	-1
7	+1	+1	-1	-1	-1
8	+1	+1	+1	+1	+1

- 8=32/4 $= 2^5/2^2 = 2^{5-2}$
- Need a principled way of choosing one-quarter of the runs from the factorial design that leads to clarity in the analysis

Example

	x_1	x_2	x_3		x_4	x_5	
Run	A	B	C	AB	AC	BC	ABC
1	-1	-1	-1	+1	+1	+1	-1
2	-1	-1	+1	+1	-1	-1	+1
3	-1	+1	-1	-1	+1	-1	+1
4	-1	+1	+1	-1	-1	+1	-1
5	+1	-1	-1	-1	-1	+1	+1
6	+1	-1	+1	-1	+1	-1	-1
7	+1	+1	-1	+1	-1	-1	-1
8	+1	+1	+1	+1	+1	+1	+1

- Seven orthogonal columns in 8 runs
- Assign factors to columns
- $x_4 = x_1x_3$, $x_5 = x_2x_3 \Rightarrow x_4x_5 = x_1x_2$

The defining relation and alias scheme

- If two columns are equal, their product must be the constant column (the identity)
- This gives us the defining relation ...
  - $I = x_1x_3x_4 = x_2x_3x_5 = x_1x_2x_4x_5$
- ... from which we can obtain the aliasing scheme
  - $x_1 = x_3x_4 = x_1x_2x_3x_5 = x_2x_4x_5$
  - $x_2 = x_1x_2x_3x_4 = x_3x_5 = x_1x_4x_5$
  - $x_3 = x_1x_4 = x_2x_5 = x_1x_2x_3x_4x_5$
  - $x_4 = x_1x_3 = x_2x_3x_4x_5 = x_1x_2x_5$
  - $x_5 = x_1x_3x_4x_5 = x_2x_3 = x_1x_2x_4$
  - $x_1x_2 = x_2x_3x_4 = x_1x_3x_5 = x_4x_5$
  - $x_1x_5 = x_3x_4x_5 = x_1x_2x_3 = x_2x_4$

The alias matrix

- If more than one effect in each alias string is non-zero, the least squares estimators will be biased
  - Assume model $Y = X_1\beta_1 + \varepsilon$
  - True model $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$

$$\begin{aligned} E(\widehat{\beta}_1) &= (X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}E(Y) \\ &= (X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}(X_1\beta + X_2\beta_2) \\ &= \beta + (X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}X_2\beta_2 \\ &= \beta + A\beta_2 \end{aligned}$$

- A is the alias matrix
  - If the columns of X_1 and X_2 are not orthogonal, $\widehat{\beta}_1$ is biased

The alias matrix

- For the 2^{5-2} example:
  - X_1 - 8×6 matrix with columns for the intercept and 5 linear terms ("main effects")
  - X_2 - 8×10 matrix with columns for the 10 product terms ("two factor interactions")

$$A = \begin{array}{c} \begin{array}{cccccccccc} 12 & 13 & 14 & 15 & 23 & 24 & 25 & 34 & 35 & 45 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

- Regular designs have entries in A only equal to 0, ± 1

The role of fractional factorial designs in a sequential strategy

- Typically, in a first experiment, fractional factorial designs are used in screening
  - Investigate which of many factors have a substantive effect on the response
  - Main effects and interactions
  - Centre points to check for curvature
- At second and later stages, augment the design
  - To resolve ambiguities due to the aliasing of factorial effects ("break the alias strings")
  - To allow estimation of curvature and prediction from a more complex model

D-optimality and non-regular designs slide 136

Introduction

- Regular fractional factorial designs have to have N equal to a power of the number of levels
  - E.g. 2^{5-2}, $3^{3-1} \times 2$
  - This inflexibility in run sizes can be a problem in practical experiments
- Non-regular designs can have any number of runs (usually $N \geq p$)
  - Often the clarity provided by the regular designs is lost
    - ▷ No defining relation or straightforward aliasing scheme
    - ▷ Partial aliasing and fractional entries in A
- One approach is finding non-regular designs is via a design optimality criterion

D-optimality

- Notation: Let ζ denote a design; that is, the treatments to be applied in the experiment and their replications
- Assuming the model $Y = X\beta + \varepsilon$, a **D-optimal design** $\zeta^\star$ maximises

$$\phi(\zeta) = \det\left(X^{\mathrm{T}}X\right)$$

- That is, a D-optimal design maximises the determinant of the Fisher information matrix
  - Equivalent to minimising the volume of the joint confidence ellipsoid for β
- Also useful to define a Bayesian version, with R a prior precision matrix

$$\phi_B(\zeta) = \det\left(X^{\mathrm{T}}X + R\right)$$

Comments

- D-optimal designs are model dependent
  - If the model (i.e. the columns of X) changes, the optimal design may change
  - Model robust design is an active area of research
- D-optimality promotes orthogonality in the X matrix
  - If there are sufficient runs, the D-optimal design will be orthogonal
  - For particular models and choices of N, regular fractional factorial designs are D-optimal
- There are many other optimality criteria, tailored to other experimental goals
  - Prediction, model discrimination, space-filling, ...

Example: Plackett-Burman design

- $m = 11$ factors in $N = 12$ runs, first-order (main effects) model
- A particular D-optimal design is the following orthogonal array

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
2	-1	-1	-1	-1	-1	1	1	1	1	1	1
3	-1	-1	1	1	1	-1	-1	-1	1	1	1
4	-1	1	-1	1	1	-1	1	1	-1	-1	1
5	-1	1	1	-1	1	1	-1	1	-1	1	-1
6	-1	1	1	1	-1	1	1	-1	1	-1	-1
7	1	-1	1	1	-1	-1	1	1	-1	1	-1
8	1	-1	1	-1	1	1	1	-1	-1	-1	1
9	1	-1	-1	1	1	1	-1	1	1	-1	-1
10	1	1	1	-1	-1	-1	-1	1	1	-1	1
11	1	1	-1	1	-1	1	-1	-1	-1	1	1
12	1	1	-1	-1	1	-1	1	-1	1	1	-1

Example: Plackett-Burman design

- This 12-run PB design is probably the most studied non-regular design
- Orthogonal columns
- Complex aliasing between main effects and two factor interactions
  - The alias matrix A has entries 0, $\pm\frac{1}{3}$, e.g.

$$\begin{aligned}E(\widehat{\beta}_1) =&\beta_1 + \tfrac{1}{3}(-\beta_{23} - \beta_{24} - \beta_{25} + \beta_{26} - \beta_{27} - \beta_{28} + \beta_{29} + \beta_{2(10)} - \beta_{2(11)}\\ &+ \beta_{34} - \beta_{35} - \beta_{36} + \beta_{37} - \beta_{38} + \beta_{39} - \beta_{3(10)} - \beta_{3(11)}\\ &+ \beta_{45} + \beta_{46} - \beta_{47} - \beta_{48} - \beta_{49} - \beta_{4(10)} - \beta_{4(11)}\\ &- \beta_{56} - \beta_{57} - \beta_{58} - \beta_{59} + \beta_{5(10)} + \beta_{5(11)}\\ &- \beta_{67} + \beta_{68} - \beta_{69} - \beta_{6(10)} - \beta_{6(11)}\\ &+ \beta_{78} - \beta_{79} + \beta_{7(10)} - \beta_{7(11)}\\ &- \beta_{89} - \beta_{8(10)} + \beta_{8(11)}\\ &- \beta_{9(10)} + \beta_{9(11)}\\ &- \beta_{(10)(11)}) .\end{aligned}$$

Example: supersaturated design

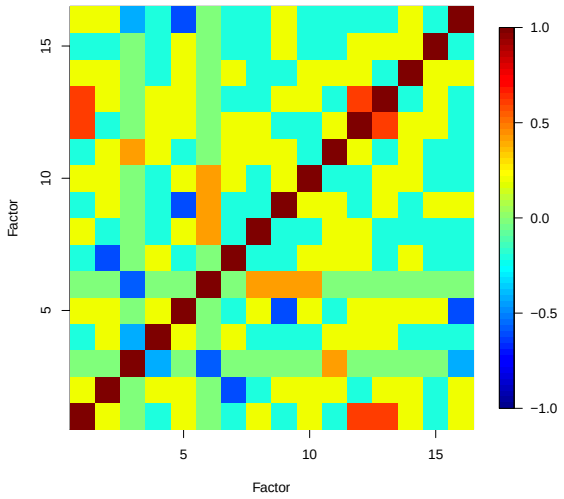
- Screening designs with fewer runs than factors
  - Can't use ordinary least squares as X does not have full column rank
  - Bayesian D-optimality with $R = (0|\tau I)$, i.e. a first column of zeros and then a diagonal matrix
- Supersaturated experiment used by GlaxoSmithKline in the development of a new oncology drug
  - 16 factors: e.g. Temperature, solvent amount, reaction time
  - $N = 10$ runs
  - Bayesian D-optimal design with $\tau = 0.2$

Example: supersaturated design

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	-1	1	1	1	1	-1	-1	1	-1	-1	1	-1	-1	-1	-1	-1
2	1	1	-1	1	1	-1	-1	-1	-1	-1	-1	1	1	1	1	1
3	-1	-1	-1	-1	-1	1	-1	1	1	-1	-1	-1	-1	-1	1	1
4	1	1	1	-1	1	1	-1	1	1	1	1	1	1	1	1	-1
5	-1	-1	1	-1	1	-1	1	-1	-1	1	-1	-1	-1	1	1	-1
6	-1	-1	1	1	-1	-1	1	-1	1	-1	1	1	1	-1	1	-1
7	1	-1	-1	1	1	1	1	1	-1	1	-1	1	1	-1	-1	-1
8	-1	1	-1	1	-1	1	1	-1	1	1	1	-1	-1	1	-1	1
9	1	1	1	-1	-1	-1	-1	-1	1	1	-1	-1	1	-1	-1	1
10	1	-1	1	-1	-1	-1	1	1	-1	-1	1	1	-1	1	-1	1

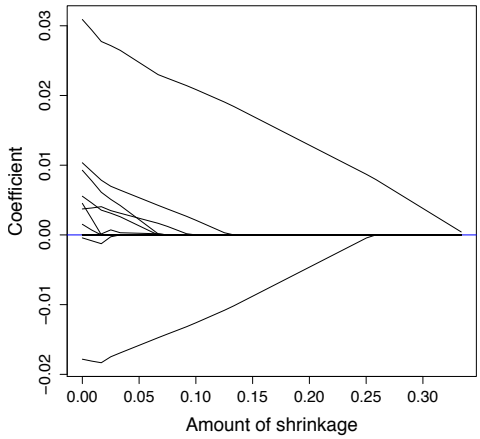
Example: supersaturated design

- ☐ Partial aliasing between main effects
- ☐ Heatmap of column correlations



Example: supersaturated design

- ☐ Analysis via regularised (penalised) methods (Dantzig selector, Candes & Tao, 2007, Annals of Statistics)
- ☐ Shrink small coefficients to zero



A generalised definition of a design

Exact design

$$\zeta = \left\{ \begin{array}{cccc} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_s \\ r_1 & r_2 & \cdots & r_s \end{array} \right\}, \qquad \sum_{i=1}^{s} r_i = n,$$

$0 < r_i \leq n$ integer.

Approximate design

$$\zeta = \left\{ \begin{array}{cccc} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_s \\ \omega_1 & \omega_2 & \cdots & \omega_s \end{array} \right\}, \qquad \sum_{i=1}^{s} \omega_i = 1,$$

$0 < \omega_i \leq 1.$

Vector $\boldsymbol{x}_i$ is a **support** (or distinct) point

Converts a discrete problem (choice of r_i) into a continuous problem (choice of ω_i)

Information matrix

We can define the Fisher information for an approximate design ζ

$$M_\beta(\zeta) = -\sum_{i=1}^{s} \omega_i \frac{\partial^2 \log f(y_i;\beta)}{\partial\beta\partial\beta^{\mathrm{T}}}$$

$$\left(= \sum_{i=1}^{s} \omega_i \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} \quad \text{for the linear model} \right)$$

□ $\boldsymbol{x}_i^{\mathrm{T}}$ is the ith row of X

Now, an approximate D-optimal maximises $\phi(\zeta) = \det\left(M_\beta(\zeta)\right)$

Example: logistic regression

One-variable logistic regression, $Y(x) \sim Bin\{m, \pi(x)\}$

$$\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \eta(x) = \beta_0 + \beta_1 x$$

$$M_\beta(\zeta) = \sum_{i=1}^{s} \omega_i \pi(x_i)\{1-\pi(x_i)\}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$$

$$= \sum_{i=1}^{s} \omega_i \frac{e^\eta}{(1+e^\eta)^2} \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$$

Example: logistic regression

D-optimal design **in terms of** η

$$\zeta^\star = \left\{ \begin{array}{cc} \eta = -1.5434 & \eta = 1.5434 \\ 0.5 & 0.5 \end{array} \right\}$$

- Clearly, the choice of x that will achieve these values of η depends on β
  - $\beta_0 = 0$, $\beta_1 = 1$, $x = \pm 1.5434$
- **Locally** D-optimal, i.e. optimal design depends on the values of the model parameters
  - This is common for nonlinear models

General equivalence theorem

Approximate designs allow powerful theory, including the general equivalence theorem

The key result of this theorem is a **necessary and sufficient condition for D-optimality**; namely, that a D-optimal design $\zeta^\star$ must satisfy

$$\psi(\boldsymbol{x}, \zeta^\star) = \pi(\boldsymbol{x})\{1 - \pi(\boldsymbol{x})\}\boldsymbol{x}^{\mathrm{T}} M_\beta^{-1}(\zeta^\star)\boldsymbol{x} \leq p \quad \text{for all } \boldsymbol{x}$$

- p is the number of parameters in β
- $\psi(\boldsymbol{x}, \zeta^\star)$ is the derivative of objective function $\phi(\zeta)$ in the direction of $\boldsymbol{x}$, i.e. it tells us how quickly the function is increasing/decreasing

Example: logistic regression

Check on necessary and sufficient conditions

$$\begin{aligned} \psi(x, \zeta) =& \pi(\boldsymbol{x})\{1 - \pi(\boldsymbol{x})\}\left\{ \sum_{i=1}^{s} \omega_i \tfrac{e^\eta}{(1+e^\eta)^2} x_i^2 \right. \\ & \left. - 2x \sum_{i=1}^{s} \omega_i \tfrac{e^\eta}{(1+e^\eta)^2} x_i + x^2 \sum_{i=1}^{s} \omega_i \tfrac{e^\eta}{(1+e^\eta)^2} \right\} \end{aligned}$$

Example: logistic regression