Statistical Modelling

Antony Overstall
University of Southampton

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

Statistical Modelling

1. Model Selection
2. Beyond the Generalised Linear Model
3. Non-linear models

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 . . .

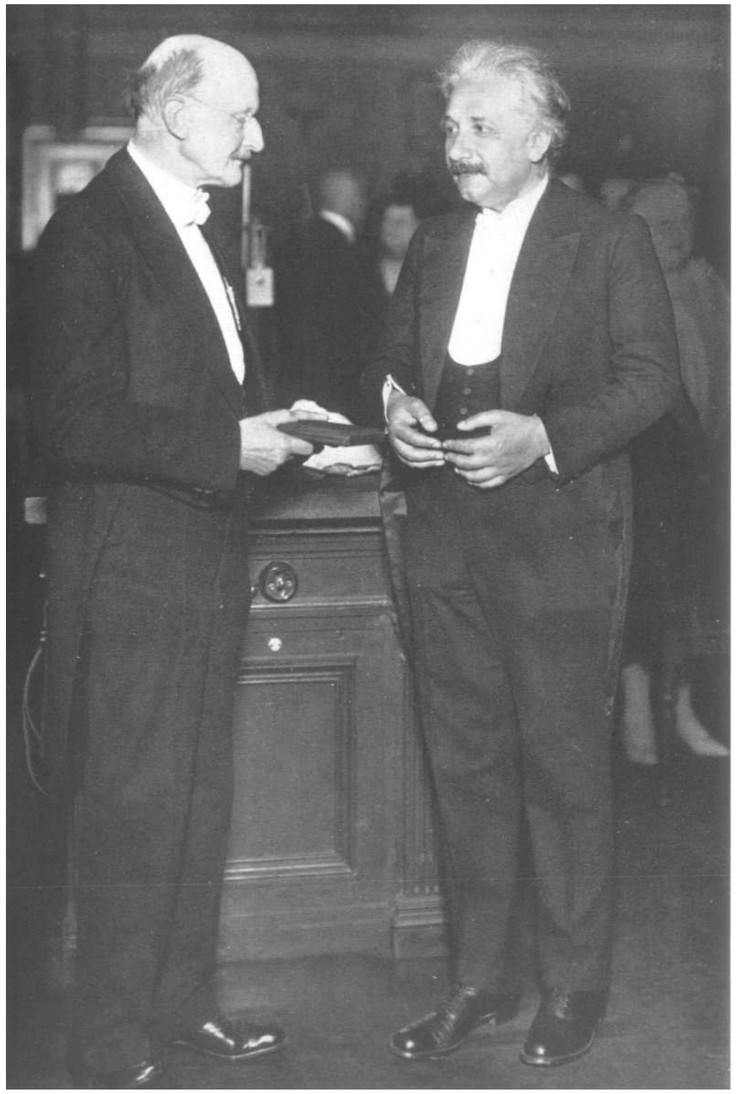
Motivation

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)

Given two models that fit the data equally well – choose the simpler model.

What happens if one model fits a better but is much more complex?

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

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

$\ell(\tilde{\beta})$ is the maximised log-likelihood under the saturated model, ie one parameter per response

- 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

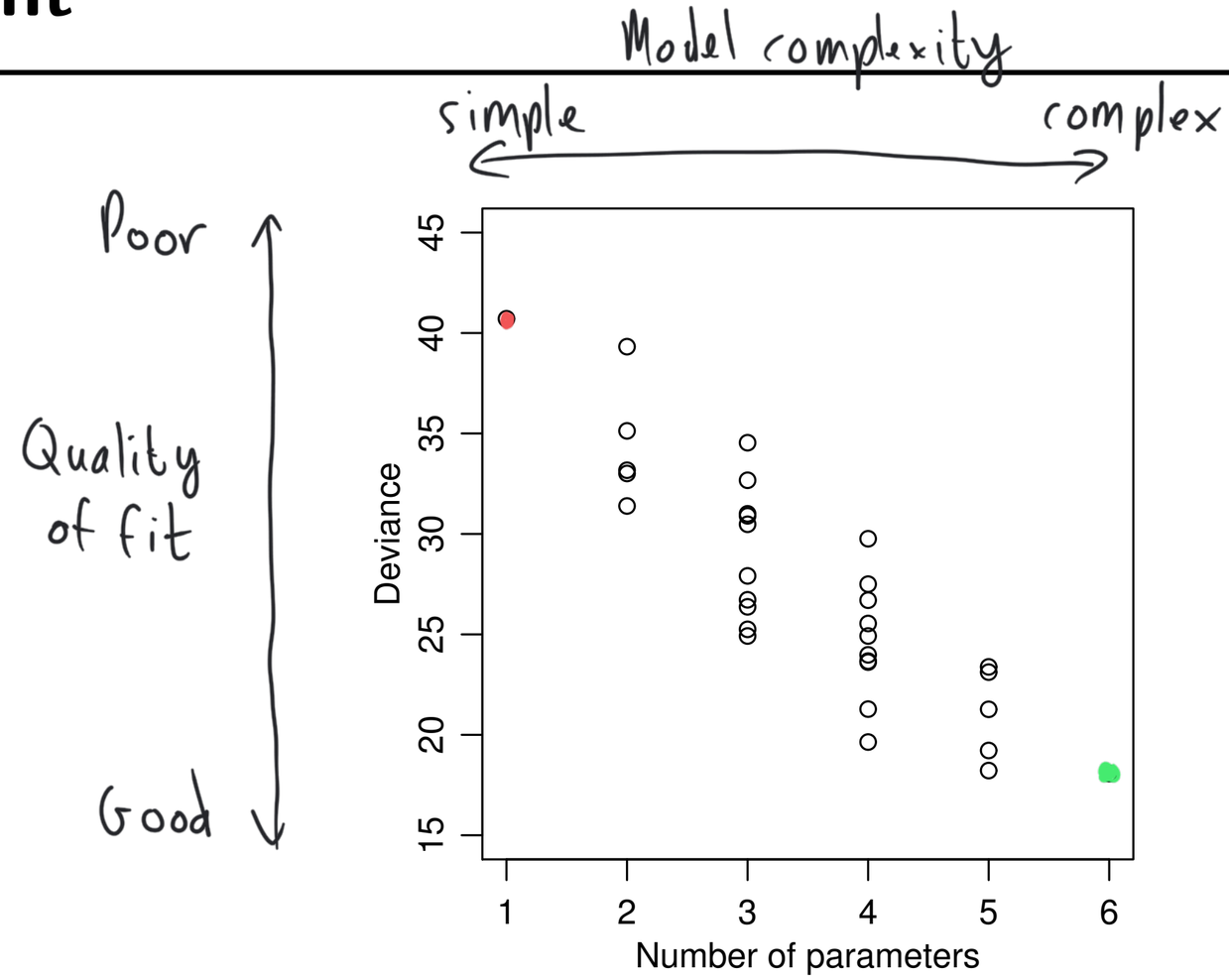
number of parameters

$df = n - p$

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

- Candidate models $f_1(y;\theta_1), \ldots, f_K(y;\theta_K)$

- 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

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

  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.

Slide 15

Jensen's inequality: If φ is convex then

$$\varphi(E(X)) \leq E(\varphi(X))$$

Since $\varphi(x) = -\log x$ is convex

$$E(\log X) \leq \log E(X)$$

$$\int \log \frac{f(y;\theta)}{g(y)}\, g(y)\, dy = E_g\left[\log \frac{f(Y;\theta)}{g(Y)}\right]$$

Negative

Kullback-Leibler discrepancy between g and $f(y;\theta)$

$$\leq \log E_g\left[\frac{f(Y;\theta)}{g(Y)}\right]$$

$$= \log \int \frac{f(y;\theta)}{g(y)}\, g(y)\, dy$$

$$= \log \int f(y;\theta)\, dy = 0$$

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.

Slide 17

$\hat{\theta}$ is defined by

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

Take 1st order Taylor series expansion of RHS about θ_g

$$0 \approx n^{-1} \sum_{j=1}^{n} \frac{\partial \log f(y_j; \theta_g)}{\partial \theta} + n^{-1} \sum_{j=1}^{n} \frac{\partial^2 \log f(y_j; \theta_g)}{\partial \theta \partial \theta^T} (\hat{\theta} - \theta_g)$$

Rearranging

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

$\approx -I(\theta_g)$ $\sim N(0, K(\theta_g))$

$$\Rightarrow \quad \hat{\theta} \sim N\left(\theta_g, I(\theta_g)^{-1} K(\theta_g) I(\theta_g)^{-1}\right)$$

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

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

$\mathrm{tr}(\hat{J}^{-1}\hat{K}) \approx p$

  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.

Slide 19

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

Take a 2nd order Taylor series expansion of $\bar{\ell}^{+}(\hat{\theta})$ about θ_g

$$\bar{\ell}^{+}(\hat{\theta}) \approx \bar{\ell}^{+}(\theta_g) + \frac{\partial \bar{\ell}^{+}(\theta_g)}{\partial \theta}(\hat{\theta} - \theta_g) + \frac{1}{2}(\hat{\theta} - \theta_g)^{T} \frac{\partial^2 \bar{\ell}^{+}(\theta_g)}{\partial \theta \partial \theta^{T}} (\hat{\theta} - \theta_g)$$

$$E_g^{+}\left(\bar{\ell}^{+}(\hat{\theta})\right) \approx \int \log f(y;\theta_g)\, g(y)\, dy + E_g^{+}\left[\frac{\partial \bar{\ell}^{+}(\theta_g)}{\partial \theta}\right](\hat{\theta} - \theta_g)$$

$$+ \frac{1}{2}(\hat{\theta} - \theta_g)^{T}\left(-\tfrac{1}{n} I(\theta_g)\right)(\hat{\theta} - \theta_g)$$

$E(\underline{x}) = 0$
$\mathrm{var}(\underline{x}) = \Sigma$
$E[\underline{x}^{T} A \underline{x}] = \mathrm{tr}(A\Sigma)$

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

Information criteria

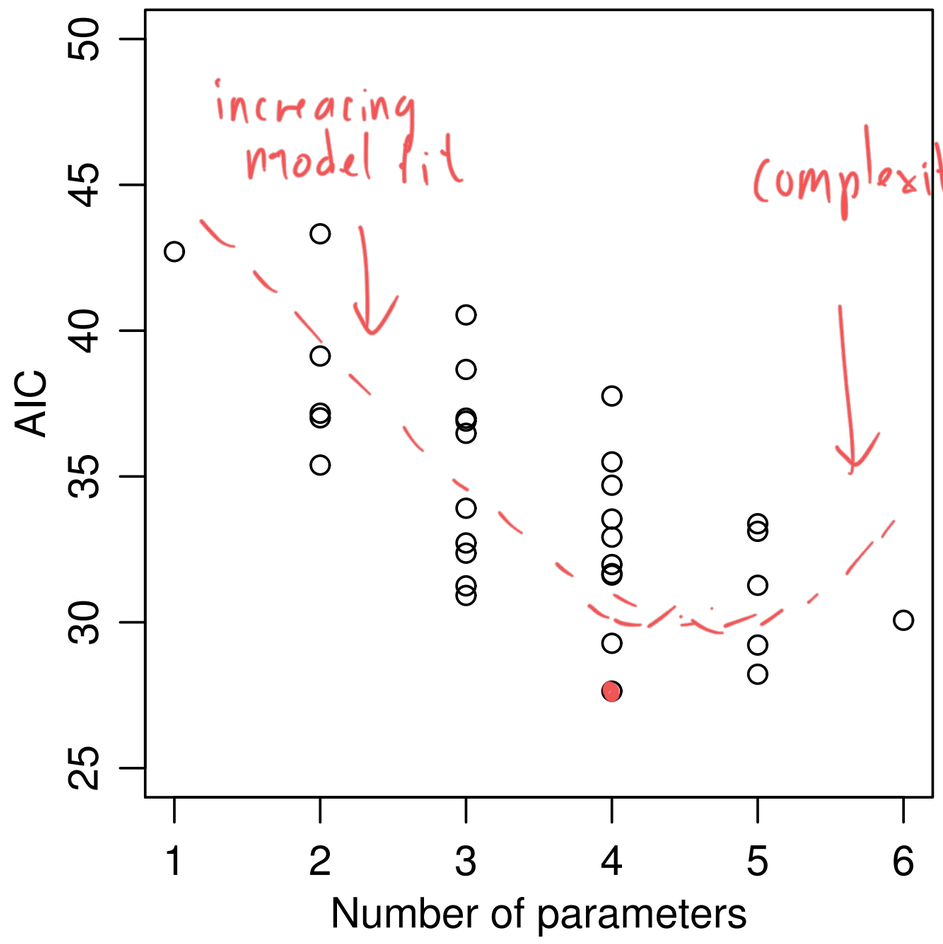
max $\bar{l}(\hat{\theta}) - p_n$

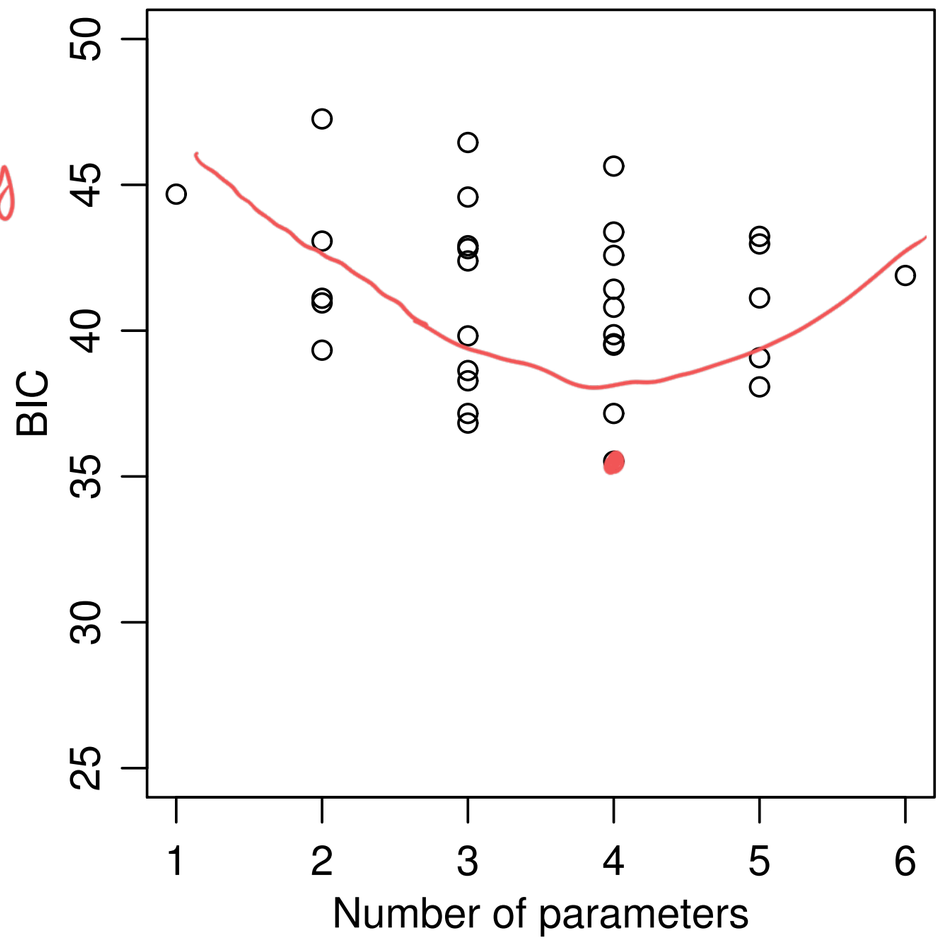
min $(p - \ell(\hat{\theta}))$

- 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 $X^\dagger$.

$X = (X^\dagger, E)$

$E(X\hat{\beta}) = H\mu$
$\text{var}(X\hat{\beta}) = \sigma^2 H$

- 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.

↳ idempotent + symmetric $HH^T = HH = H$

- 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.

SLIDE 30

Suppose z is an rv with $E(z) = m$

$\text{var}(z) = V$

Then $E(z^T \Lambda z) = tr(\Lambda V) + m^T \Lambda m$

First

$$E_+\left[(Y_+ - X\hat{\beta})^T (Y_+ - X\hat{\beta})\right]$$

$$= \underbrace{tr(\sigma^2 I_n)}_{=\sigma^2 n} + (\mu - X\hat{\beta})^T(\mu - X\hat{\beta})$$

$$\Delta(X) = n^{-1} E\left[\sigma^2 n + (\mu - X\hat{\beta})^T(\mu - X\hat{\beta})\right]$$

$$= \sigma^2 + \frac{1}{n}\left(tr(\sigma^2 H) + \mu^T \underbrace{(I-H)^T(I-H)}_{I-H}\mu\right)$$

$$= \sigma^2 + \left(\frac{\sigma^2 p}{n} + n^{-1}\mu^T(I-H)\mu\right) = \sigma^2\left(1 + \frac{p}{n}\right) + \frac{\mu^T(I_n - H)\mu}{n}$$

$$\Lambda = I_n$$

$$m = E(Y_+ - X\hat{\beta}) = \mu - X\hat{\beta}$$

$$V = \text{var}(Y_+ - X\hat{\beta}) = \sigma^2 I_n$$

$$E(\mu - X\hat{\beta}) = \mu - H\mu = (I-H)\mu$$

$$\text{var}(\mu - X\hat{\beta}) = \sigma^2 H$$

If model is correct then $H\mu = \mu$ and

$$\Delta(x) = \sigma^2\left(1 + \frac{p}{n}\right)$$

If model is true then $H\mu = \mu$ and $p = q$

$$\Delta(x) = \sigma^2\left(1 + \frac{q}{n}\right)$$

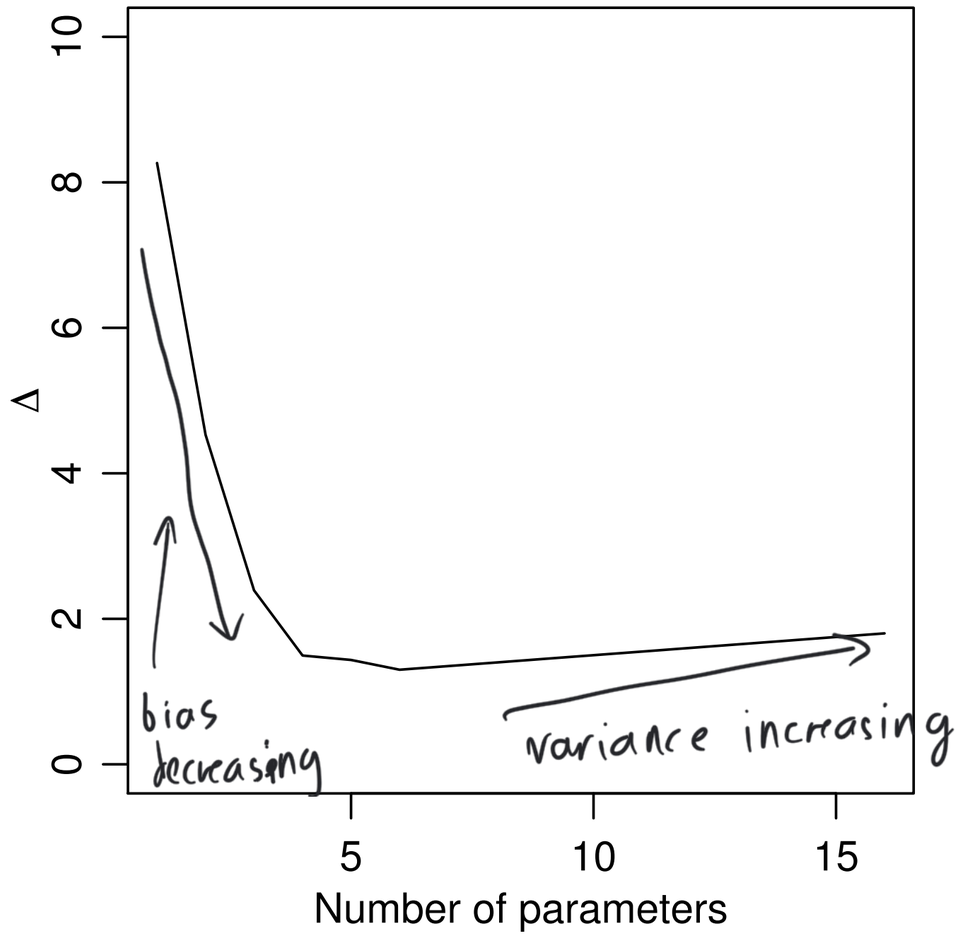
Prediction error II

- 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 $\Delta(X)$

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

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

$\hat{\beta}^*$ is computed (X^*, y^*)

- 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

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

Approximate $h_{jj} = \frac{\mathrm{tr}(H)}{n}$

- Can show that

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

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.

Other selection criteria

$$AIC = n\log\hat{\sigma}^2 + 2p + n$$

- Corrected version of AIC for models with normal responses:

$$\begin{aligned} \mathrm{AIC_c} &\equiv n\log\widehat{\sigma}^2 + n\frac{1+p/n}{1-(p+2)/n}, \\ &= AIC + k(n,p) \end{aligned}$$

$k(n,p) \to 0$ as $n \to \infty$

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

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

Can be shown to be an approximation to AIC

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.

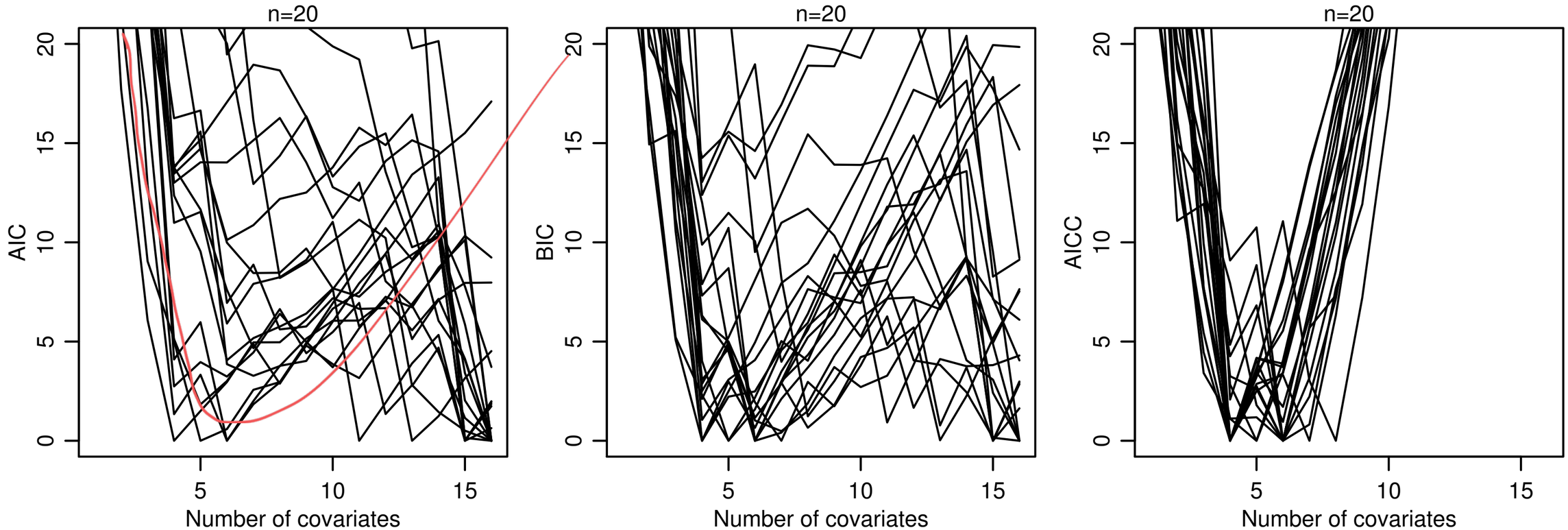
Simulation experiment

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	

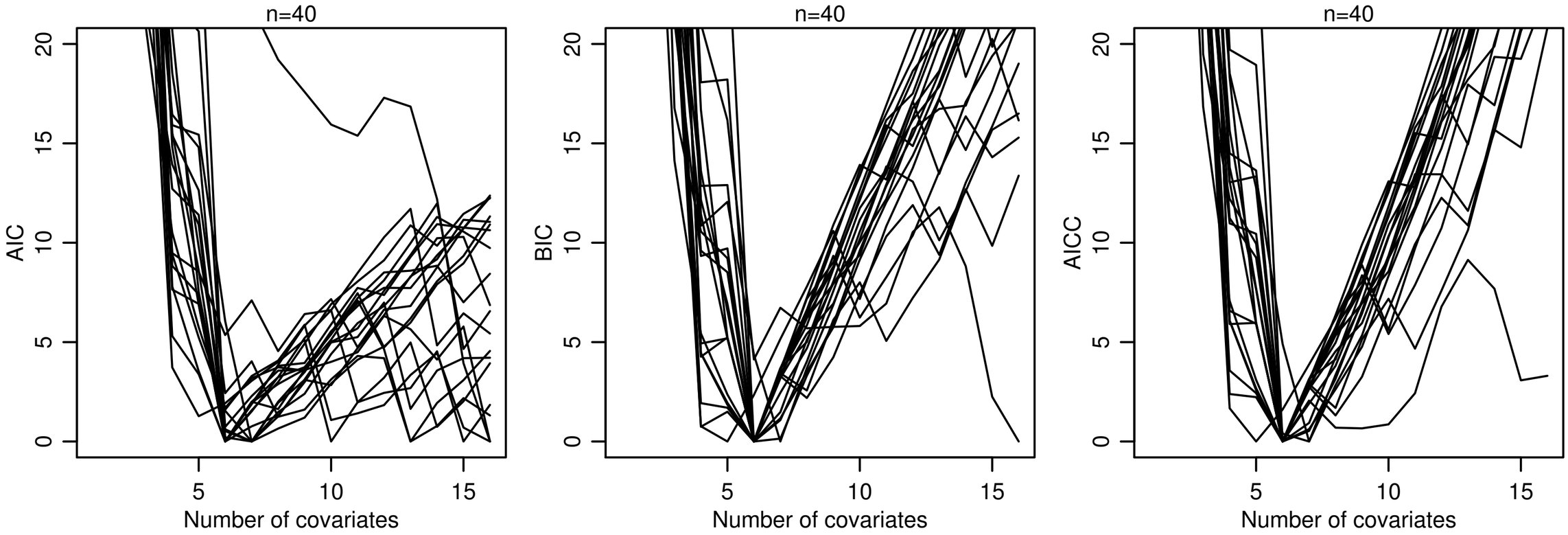
Simulation experiment

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



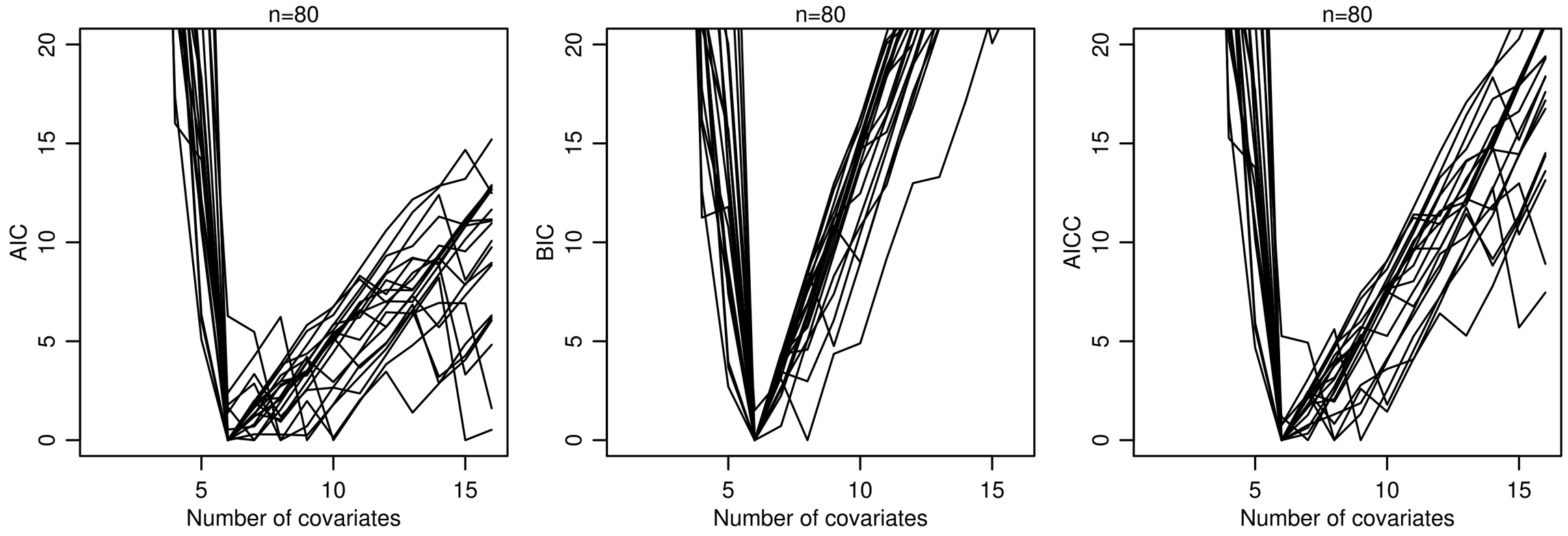
Simulation experiment

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.

Bayesian Inference

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 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

$$\pi(m, \theta_m) = \pi(\theta_m \mid m)\pi(m) = \pi_m(\theta_m)\pi_m, \qquad \sum_{m=1}^{M} \pi_m = 1$$

say. $= \int \pi(m, \theta_m \mid y)\, d\theta_m = \int \frac{f(y \mid \theta_m, m)\, \pi_m(\theta_m)\pi_m}{\sum \int f(y \mid \theta_m, m)\, \pi_m(\theta_m)\pi_m\, d\theta_m}\, d\theta_m$

- Inference about model choice is based on marginal posterior density

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

$\int f(y; \theta_m)\pi_m(\theta_m)d\theta_m$ = marginal likelihood / evidence

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,

$$\begin{aligned} f(y \mid m) &= \sigma^{-d/2}(2\pi)^{-d/2} \int f(y \mid m, \theta) \prod_{r=1}^{d} \exp\left\{-\theta_r^2/(2\sigma^2)\right\} \mathrm{d}\theta_r \\ &\approx \sigma^{-d/2}(2\pi)^{-d/2} \int f(y \mid m, \theta) \prod_{r=1}^{d} \mathrm{d}\theta_r, \end{aligned}$$

for a highly diffuse prior distribution (large σ^2). The Bayes factor for comparing the models is approximately

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

Model 1 is simpler than Model 2 ie. $d_1 < d_2$

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

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

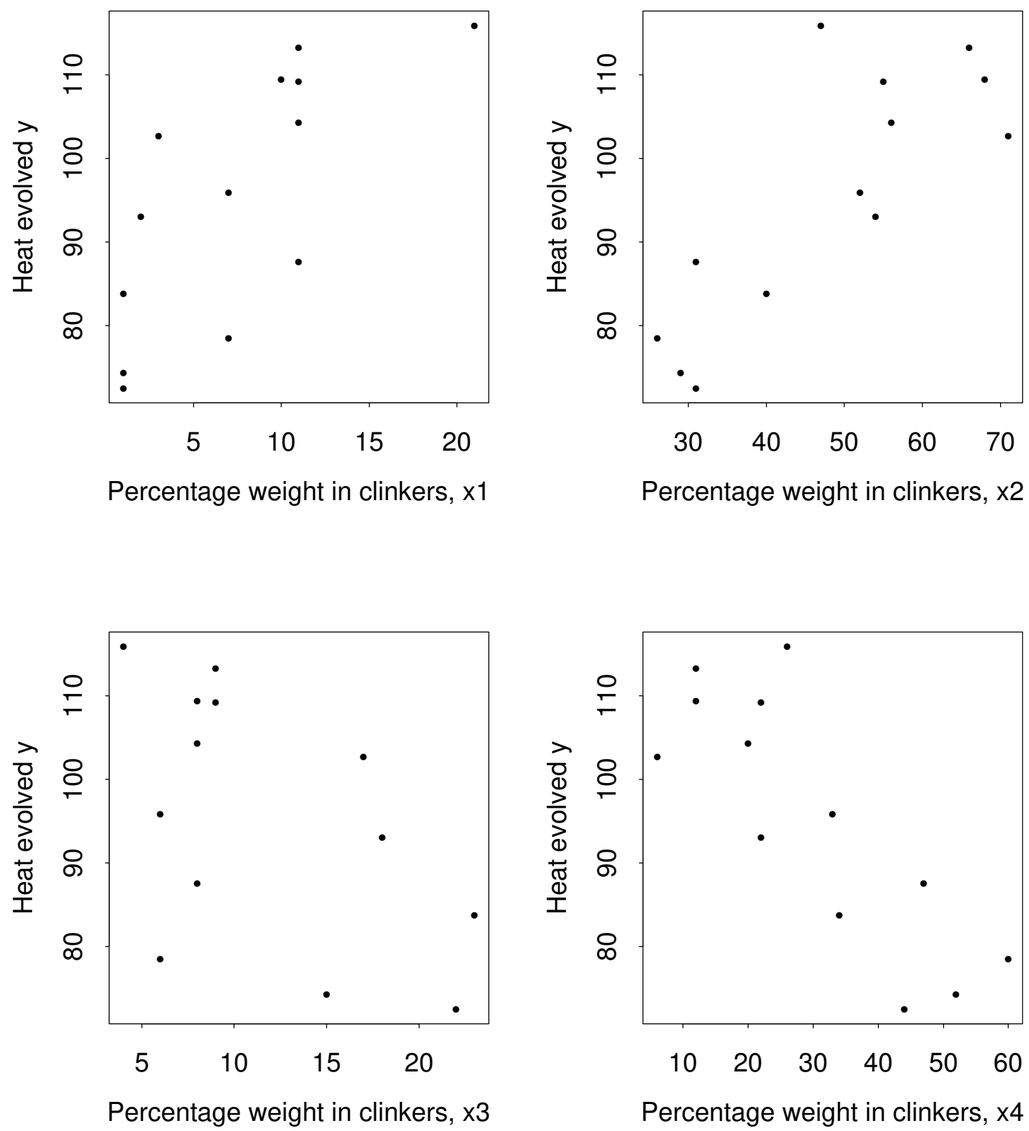
In prediction $f(z \mid m)$

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
```

Fit a linear model.

i.e.

$$y = X\beta + \varepsilon$$

X can have 1 to 5 columns

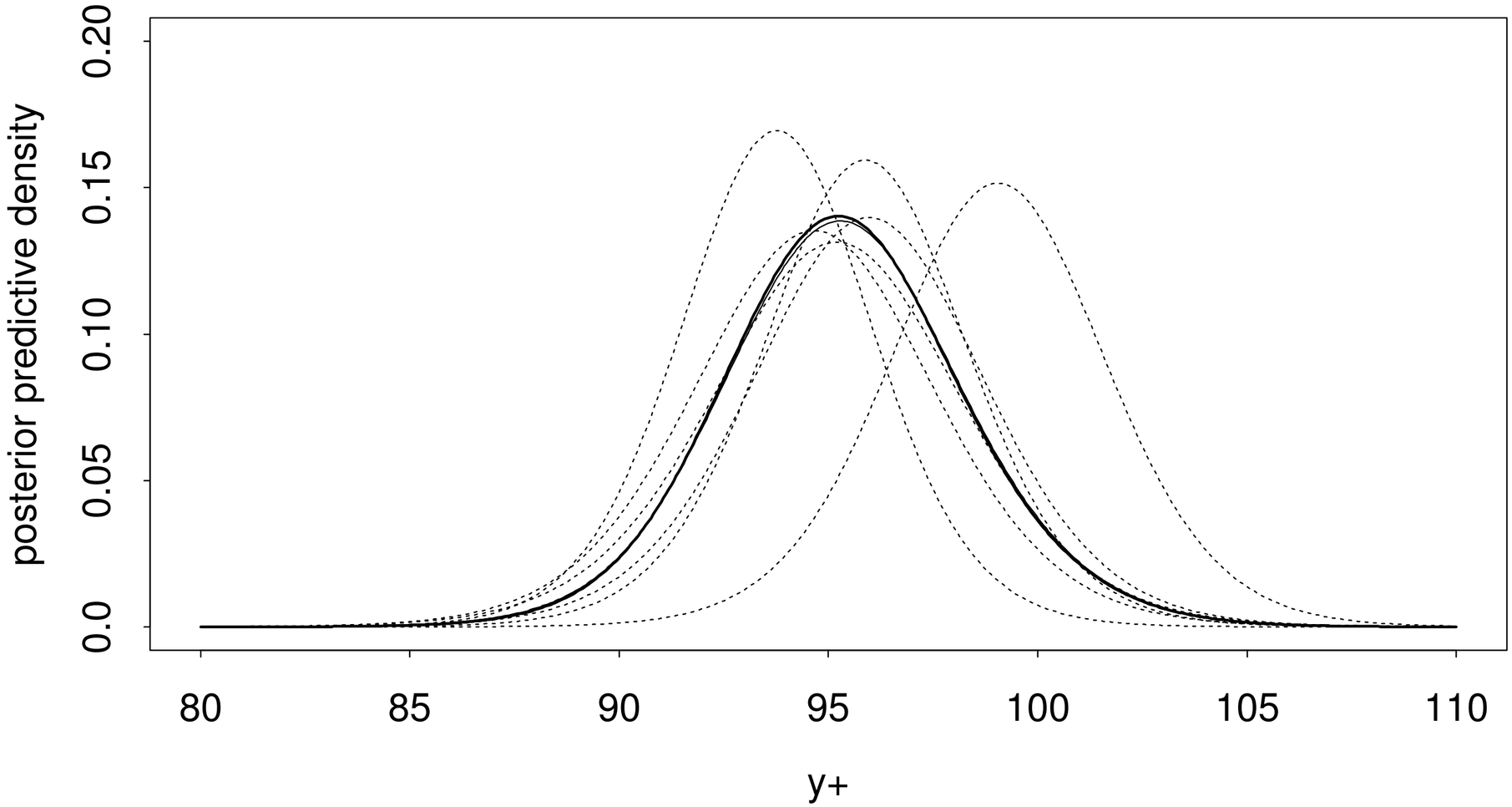
Example: Cement data

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 $\theta = \mathrm{E}(\theta \mid y)$.

- 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.