Selecting Amongst
Large Classes of Models

Brian D. Ripley

Professor of Applied Statistics
University of Oxford

`ripley@stats.ox.ac.uk`
`http://stats.ox.ac.uk/∼ripley`

Manifesto

Statisticians and other users of statistical methods have been choosing models for a long time, but the current availability of large amounts of data and of computational resources means that model choice is now being done on a scale which was not dreamt of 25 years ago.

Unfortunately, the practical issues are probably less widely appreciated than they used to be, as statistical software and the advent of AIC, BIC and all that has made it so much easier for the end user to trawl through literally thousands of models (and in some cases many more).

Traditional distinctions between ‘parametric’ and ‘non-parametric’ models are often moot, when people now (attempt to) fit neural networks with half a million parameters.

Where do the models come from?

- Sometimes a set of models is provided based on subject-matter theory. In my experience good theory is very rare. Sometimes called *mechanistic* models. One example is the Black–Scholes theory of option pricing.
- Most often some simple restrictions are placed on the behaviour we expect to find, for example linear models, $AR(p)$ processes, factorial models with limited interactions. Sometimes called *empirical* models.

  Note that these can be very broad classes if transformations of variables (on both sides) are allowed.
- We now have model classes that can approximate any reasonable model, for example neural networks. And we may have subsets within these such as (generalized) additive models. Nowadays we may have the data and the computational resources to fit such models.

Why do we want to choose a model?

It took me a long while to realize how profound a question that was.

Explanation *vs* Prediction

This causes a lot of confusion. For *explanation*, Occam's razor applies and we want

> an explanation that is as simple as possible,
> but no simpler
>
> attrib Einstein

and we do have a concept of a 'true' model, or at least a model that is a good working approximation to the truth, for

> all models are false, but some are useful
>
> G.E.P. Box, 1976

Explanation is like doing scientific research.

On the other hand, *prediction* is like doing engineering development. All that matters is that it works. And if the aim is prediction, model choice should be based on the quality of the predictions.

Workers in pattern recognition have long recognised this, and used *validation sets* to choose between models, and *test sets* to assess the quality of the predictions from the chosen model.

One of my favourite teaching examples is

> Ein-Dor, P. & Feldmesser, J. (1987) Attributes of the performance of central processing units: a relative performance prediction model. *Communications of the ACM* **30**, 308–317.

which despite its title selects a subset of transformed variables. The paper is a wonderful example of how **not** to do that, too.

More on CPUs' performance

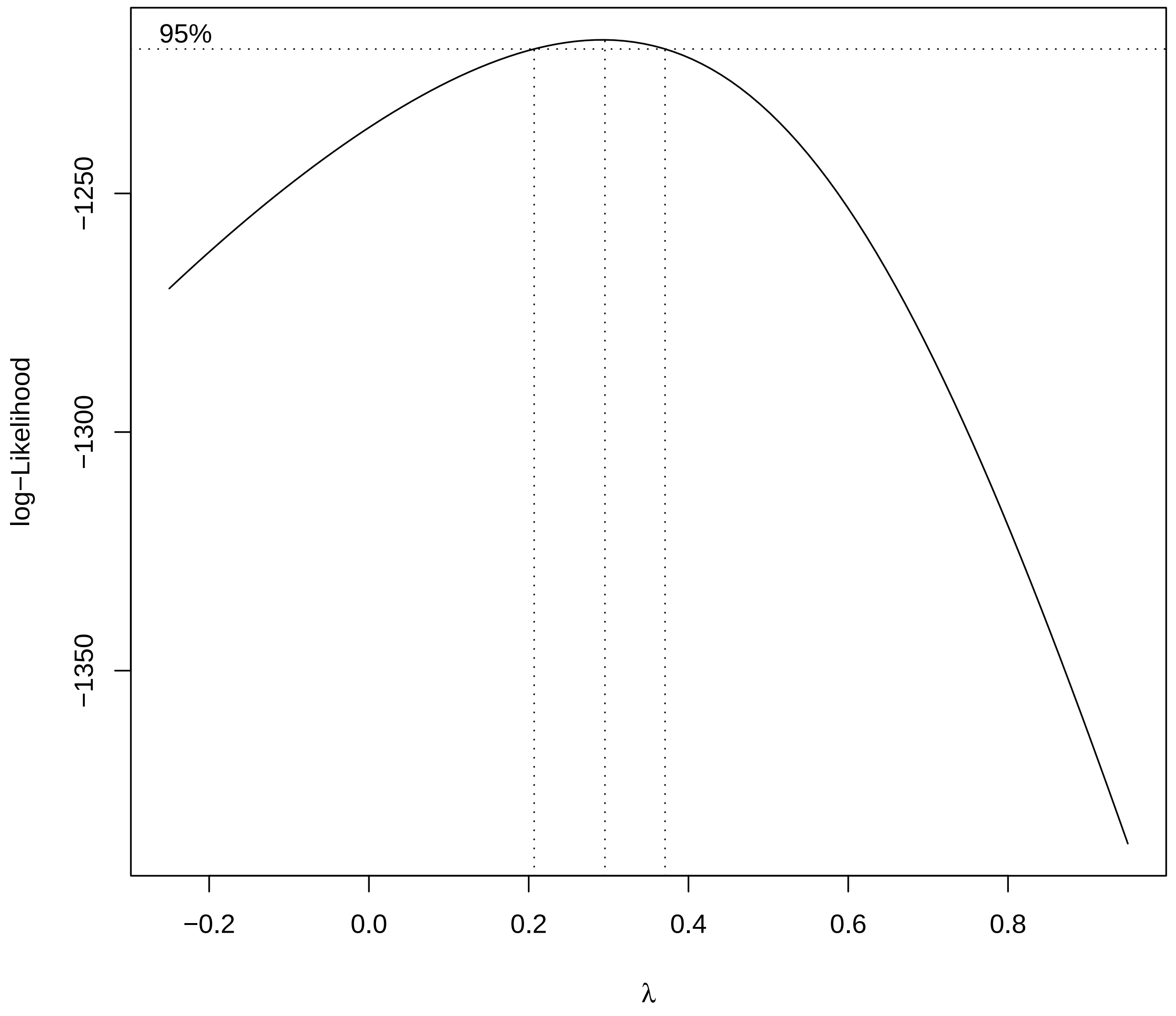
There were six machine characteristics:

- the cycle time (nanoseconds),
- the cache size (Kb),
- the minimum and maximum possible main memory size (Kb)
- the minimum and maximum possible number of channels.

How much memory and how many channels the actual machine tested had is unspecified.

The original paper gave a linear regression for the **square root** of performance, but log scale looks more intuitive. We have a technology to test that, from Box & Cox (1964).

Box–Cox transformations



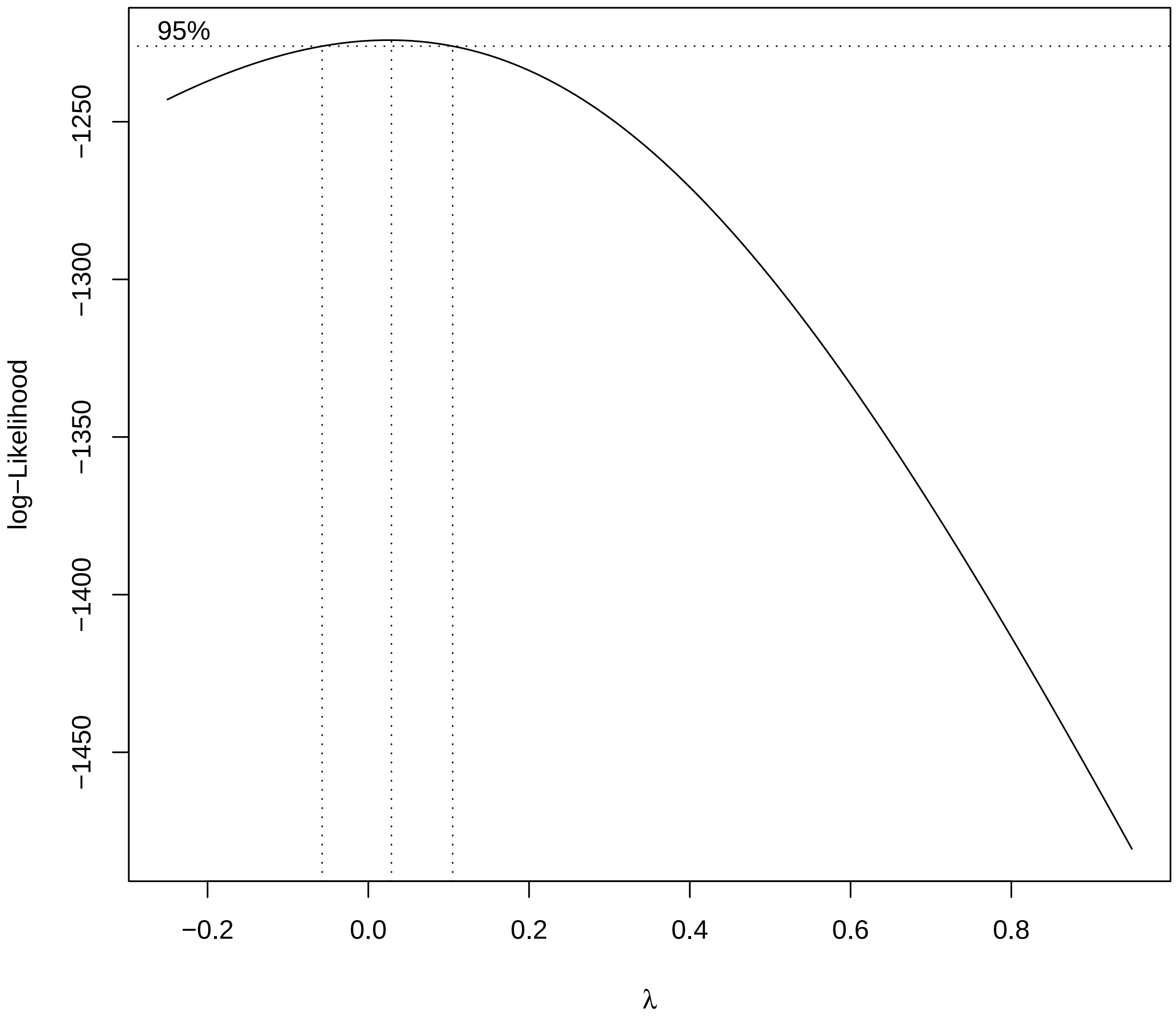
That is not what we were expecting!

Caveat: what did we transform?

We only transformed the response: it is natural to transform the regressors as well, so we need to choose several transformations simultaneously. We have technology to do that, even with non-parametric smooth functions (ACE, AVAS, . . .) but it is not very reliable.

Old-fashioned methods work: we discretized the continuous regressors into four groups and used these as categorical predictors.

Box–Cox transformations revisited



which is rather satisfying.

Why select a model at all?

It does seem a widespread misconception that model choice is about

choosing the best model

For *explanation* we ought to be alert to be possibility of there being several (roughly) equally good explanatory models.

I learnt that from David Cox after having already done a lot of informal model choice in applied problems in which I would have benefited from offering several alternative solutions.

Simplicity helps both with communicating the concepts embodied in the model and in what psychologists call *generalization*, the ability to 'work' in scenarios very different from those in which the model was studied. So there is a premium on few models.

For *prediction* I find a good analogy is that of choosing between expert opinions: if you have access to a large panel of experts, how would you use their opinions?

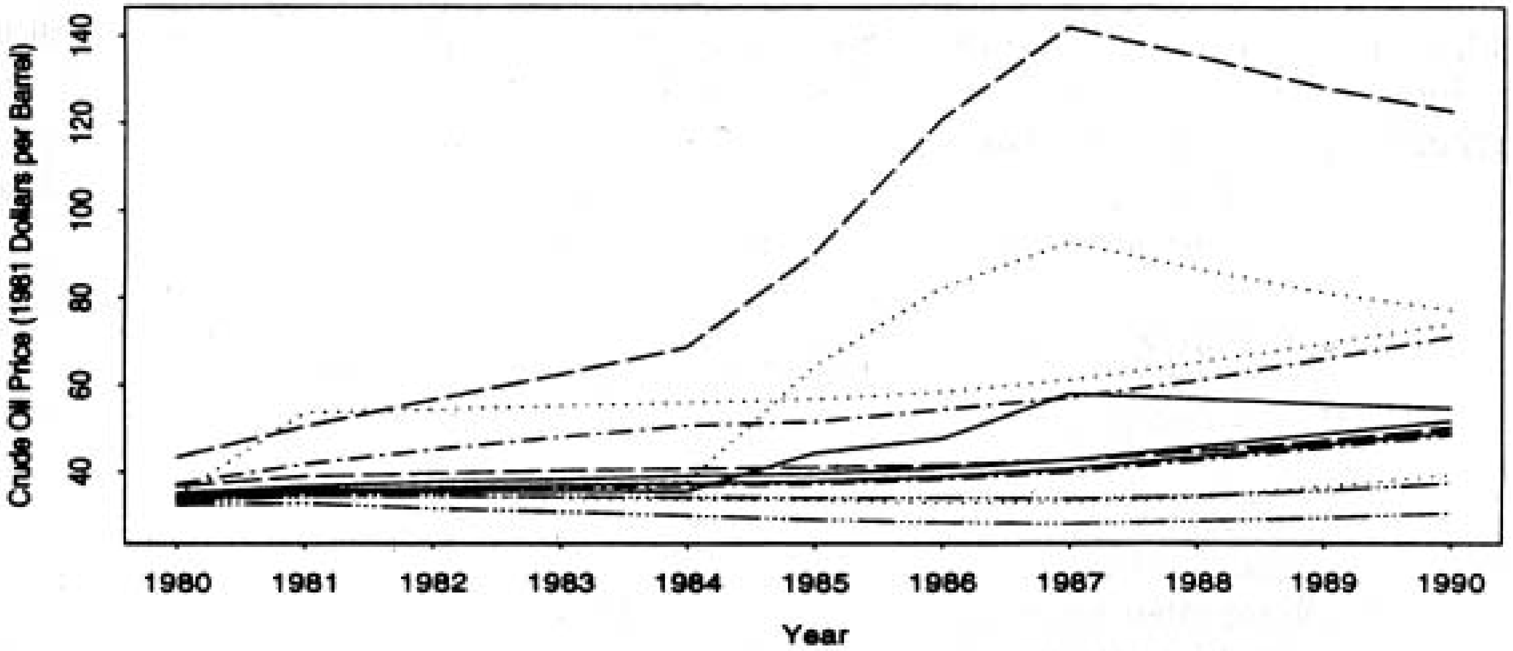
People do tend to pick one expert (‘guru’) and listen to him/her, but it would seem better to seek a consensus view, which translates to *model averaging* rather than model choice. Our analogy is with experts, which implies some prior selection of people with a track record: one related statistical idea is the *Occam’s window* (Madigan & Raftery, 1994) which keeps only models with a reasonable record.

Because the model may be used in scenarios very different from those in which it was tested, generalization is still important, and *other things being equal* a mechanistic model or a simple empirical model has more chance of reflecting the data-generation mechanism and so of generalizing. But other things rarely *are* equal.

All the models/experts may be wrong

Note that taking a consensus view only helps sometimes with generalization.

Draper (1995) has a graph of predictions of oil prices for 1981–90 made in 1980. The analysts were all confident, differed considerably from each other, and were all way off — the oil price was $13 in 1986!



Computational cost

A major reason to choose a model appears still to be computational cost, a viewpoint of Geisser (1993). Even if we can fit large families of models, we may have time to consider the predictions only from a few.

A much-quoted example is a NIST study on reading hand-written ZIP codes, which have to be read in about 1/2 second each to be useful in a sorting machine.

An historical perspective — Model choice in 1977

That's when I started to learn about this.

- The set of models one could consider was severely limited by computational constraints, although packages such as GLIM 3.77 were becoming available.
- Stepwise selection was the main formal tool, using hypothesis tests between a pair of nested models, e.g. F tests for regressions.

  Few people did enough tests to worry much about multiple comparisons issues.
- Residual plots were used, but they were crude plots and limited to small datasets.

There was very little attempt to deal with choosing between models that were genuinely different explanations: Cox's (1961) 'tests of separate families of hypotheses' existed but was little known and less used.
But the world was changing . . . .

Cross-validation

A much misunderstood topic!

Leave-one-out CV

The idea is that given a dataset of N points, we use our model-building procedure on each subset of size $N-1$, and predict the point we left out. Then the set of predictions can be summarized by some measure of prediction accuracy. Idea goes back at least as far as Mosteller & Wallace (1963), and Allen's (1971, 4) PRESS (prediction sum-of-squares) used this to choose a set of variables in linear regression.

Stone (1974) / Geisser (1975) pointed out we could apply this to many aspects of model choice, including parameter estimation.

NB: This is *not* jackknifing *a la* Quenouille and Tukey.

Having to do model-building N times can be prohibitive unless there are computational shortcuts.

V-fold cross-validation

Divide the data into V sets, and amalgamate $V-1$ of them, build a model and predict the result for the remaining set. Do this V times leaving a different set out each time.

How big should V be? We want the model-building problem to be realistic, so want to leave out a small proportion. We don't want too much work. So usually V is 3–10.

One early advocate of this was the CART book (Breiman, Friedman, Olshen & Stone, 1984) and program.

Does it work?

Leave-one-out CV does not work well in general. It makes too small changes to the fit.

10-fold CV often works well, but sometimes the result is very sensitive to the partitioning used. We can ‘average’ over several random partitions.

Often better for comparisons than for absolute values of performance.

How prediction accuracy is measured can be critical.

AIC, BIC and all that

Akaike (1973, 1974) introduced a criterion for model adequacy, first for time-series models and then more generally. He relates how his secretary suggested he call it ‘An Information Criterion’, AIC.

This has a very appealing simplicity:

$$AIC = -2\log(\text{maximized likelihood}) + 2p$$

where p is the number of estimated parameters. Choose the model with the smallest AIC (and perhaps retain all models within 2 of the minimum).

Despite that, quite a few people have managed to get it wrong!

This is similar to Mallows’ C_p criterion for regression,

$$C_p = \text{RSS}/\sigma^2 + 2p - N$$

and is the same if σ^2 is known. Both are of the form

$$\text{measure of fit} + \text{complexity penalty}$$

Schwarz's (1978) criterion, often called BIC or SBC, replaces 2 by $log\, n$ for a suitable definition of n, the size of the dataset. In the original regression context this is just the number of cases.

BIC was anticipated by work of Harold Jeffreys in the 1930's.

Derivation of AIC

Suppose we have a dataset of size N, and we fit a model to it by maximum likelihood, and measure the fit by the *deviance* D (constant minus twice maximized log-likelihood). Suppose we have m (finite) nested models.

Hypothetically, suppose we have another dataset of the same size, and we compute the deviance D^* for that dataset *at the MLE for the first dataset*. We would expect that D^* would be bigger than D, on average. In between would be the value D' if we had evaluated the deviance at the true parameter values. Some Taylor-series expansions show that

$$E\,D^* - E\,D' \approx p, \qquad E\,D' - E\,D \approx p$$

and hence $AIC = D + 2p$ is (to this order) an unbiased estimator of $E\,D^*$. And that is a reasonable measure of performance, the Kullback-Leibler divergence between the true model and the plug-in model (at the MLE).

These expectations are over the dataset under the assumed model.

Crucial assumptions

1. The model is true! Suppose we use this to select the order of an $AR(p)$ model. If the data really came from an $AR(p_0)$ model, all models with $p \geq p_0$ are true, but those with $p < p_0$ are not even approximately true.

   This assumption can be relaxed. Takeuchi (1976) did so, and his result has been rediscovered by Stone (1977) and many times since. p gets replaced by a much more complicated formula.

2. The models are nested – AIC is widely used when they are not.

3. Fitting is by maximum likelihood. Nowadays many models are fitted by penalized methods or Bayesian averaging .... That can be worked through too, in NIC or Moody's p_{eff}.

4. The Taylor-series approximations are adequate. People have tried various refinements, notably AICC (or AIC$_c$) given by

   $$AICC = D + 2p\Big(\frac{N}{N-p+1}\Big)$$

   Also, the MLEs need to be in the interior of the parameter space, even when a simpler or alternative model is true. (Not likely to be true for variance components for example.)

5. AIC is a reasonably good estimator of $E\,D^*$, or at least that differences between models in AIC are reasonably good estimators of differences in $E\,D^*$.

   This seems the Achilles' heel of AIC.
   $AIC = O_p(N)$ but the variability as an estimate is $O_p(\sqrt{N})$. This reduces to $O_p(1)$ for differences between models *provided they are nested*.

AIC has been criticised in asymptotic studies and simulation studies for tending to over-fit, that is choose a model at least as large as the true model. That is a virtue, not a deficiency: this is a prediction-based criterion, not an explanation-based one.

AIC is asymptotically equivalent to leave-one-out CV for iid samples and using deviance as the loss function (Stone, 1977), and in fact even when the model is not true NIC is equivalent (Ripley, 1996).

Bayesian approaches

Note the plural — I think Bayesians are rarely Bayesian in their model choices. Assume M (finite) models, exactly one of which is true.

In the Bayesian formulation, models are compared via $P\{M \mid \mathcal{T}\}$, the posterior probability assigned to model M.

$$P\{M \mid \mathcal{T}\} \propto p(\mathcal{T} \mid M) p_M,$$

$$p(\mathcal{T} \mid M) = \int p(\mathcal{T} \mid M, \theta) p(\theta) \, \mathrm{d}\theta$$

so the ratio in comparing models M_1 and M_2 is proportional to $p(\mathcal{T} \mid M_2)/p(\mathcal{T} \mid M_1)$, known as the *Bayes factor*.

However, a formal Bayesian approach then averages predictions from models, weighting by $P\{M \mid \mathcal{T}\}$, unless a very peculiar loss function is in use. And this has been used for a long time, despite recent attempts to claim the credit for ‘Bayesian Model Averaging’.

Suppose we just use the Bayes factor as a guide. The difficulty is in evaluating $p(\mathcal{T} \mid M)$. Asymptotics are not useful for Bayesian methods, as the prior on θ is often very important in providing smoothing, yet asymptotically negligible.

We can expand out the log posterior density via Laplace approximation and drop various terms, eventually reaching

$$\log p(\mathcal{T} \mid M) \approx L(\widehat{\theta}; \mathcal{T}) - \tfrac{1}{2} \log |H|.$$

where H is the Hessian of the log-likelihood and we needed to assume that the prior is very diffuse.

For an iid random sample of size n from the assumed model, the penalty might be roughly proportional to $-(\frac{1}{2} \log n)\, p$ provided the parameters are identifiable. This is Schwarz's BIC up to a factor of two. As with AIC, the model with minimal BIC is chosen.

Crucial assumptions

1. The data were derived as an iid sample. (What about e.g. random effects models?) (Originally for linear models only.)
2. Choosing a single model is relevant in the Bayesian approach.
3. The model is true.
4. The prior can be neglected. We may not obtain much information about parameters which are rarely effective, even in very large samples.
5. The simple asymptotics are adequate and that the rate of data collection on each parameter would be the same. We should be interested in comparing different models for the same N, and in many problems p will be comparable with N.

Note that as this is trying to choose an explanation, we would expect it to neither overfit nor underfit, and there is some theoretical support for that.

There are other (semi-)Bayesian approaches, including DIC.

Model averaging

For prediction purposes (and that applies to almost all Bayesians) we should average the predictions over models. We **do not choose** a single model.

What do we average?

The probability predictions made by the models.

For linear regression this amounts to averaging the coefficients over the models (being zero where a regressor is excluded), and this becomes a form of shrinkage.
[Other forms of shrinkage like ridge regression may be as good at very much lower computational cost.]

Note that we may not want to average over all models. We may want to choose a subset for computational reasons, or for plausibility.

How do we choose the weights?

- In the Bayesian theory this is clear, via the Bayes factors. In practice this is discredited. Even if we can compute them accurately (and via MCMC we may have a chance), we assume that one and exactly one model is true. In practice Bayes factors can depend on aspects of model inadequacy which are of no interest.
- Via cross-validation (goes back to Stone, 1974).
- Via bootstrapping (LeBlanc & Tibshirani, 1993).
- As an extended estimation problem, with the weights depending on the sample via a model (e.g. a multiple logistic); so-called *stacked generalization* and *mixtures of experts*.

Bagging, boosting, random forests

Model averaging ideas have been much explored in the field of classification trees.

In *bagging* models are fitted from bootstrap resamples of the data, and weighted equally.

In *boosting* each additional model is chosen to (attempt to) repair the inadequacies of the current averaged model by resampling biased towards the mistakes.

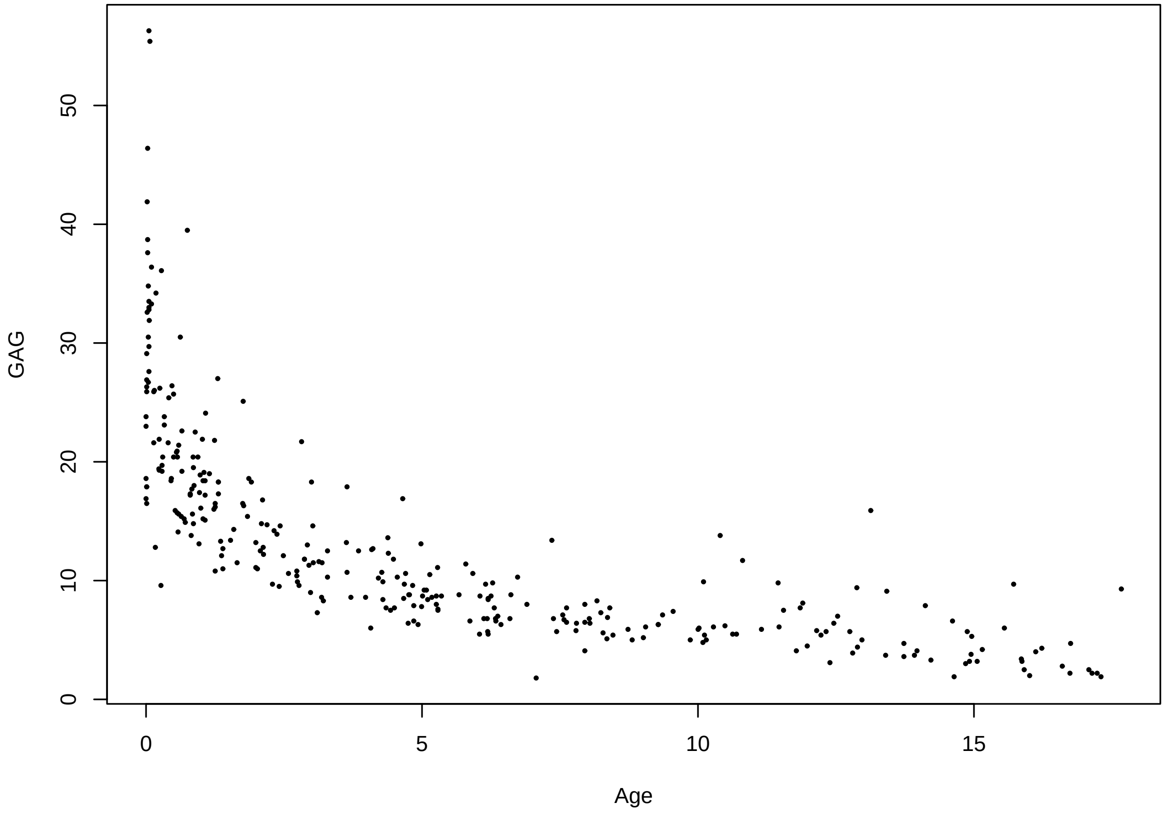
In *random forests* the tree-construction algorithm randomly restricts itself at the choice of each split.

(Practical) model selection in 2004

- The concept of a model ought to be much, much larger than in 1977. Even a decade ago, people attempted to fit neural networks with half a million free parameters.
- Many models are not fitted by maximum likelihood, to very large datasets.
- Model classes can often overlap in quite extensive ways.

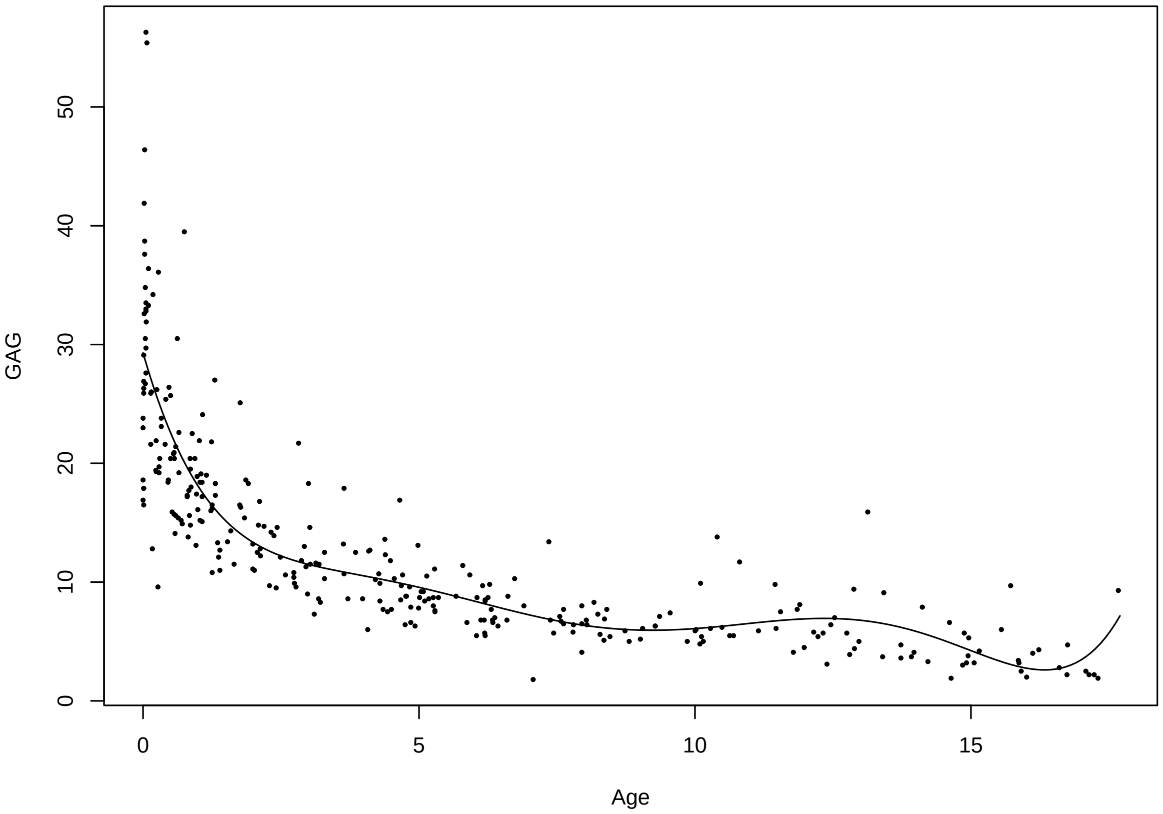
Calibrating GAG in urine

Susan Prosser measured the concentration of the chemical GAG in the urine of 314 children aged 0—18 years. Her aim was to establish ‘normal’ levels at different ages.



Clearly we want to fit a smooth curve. What? Polynomial? Exponential?

Choosing the degree of a polynomial by F-tests gives degree 6.



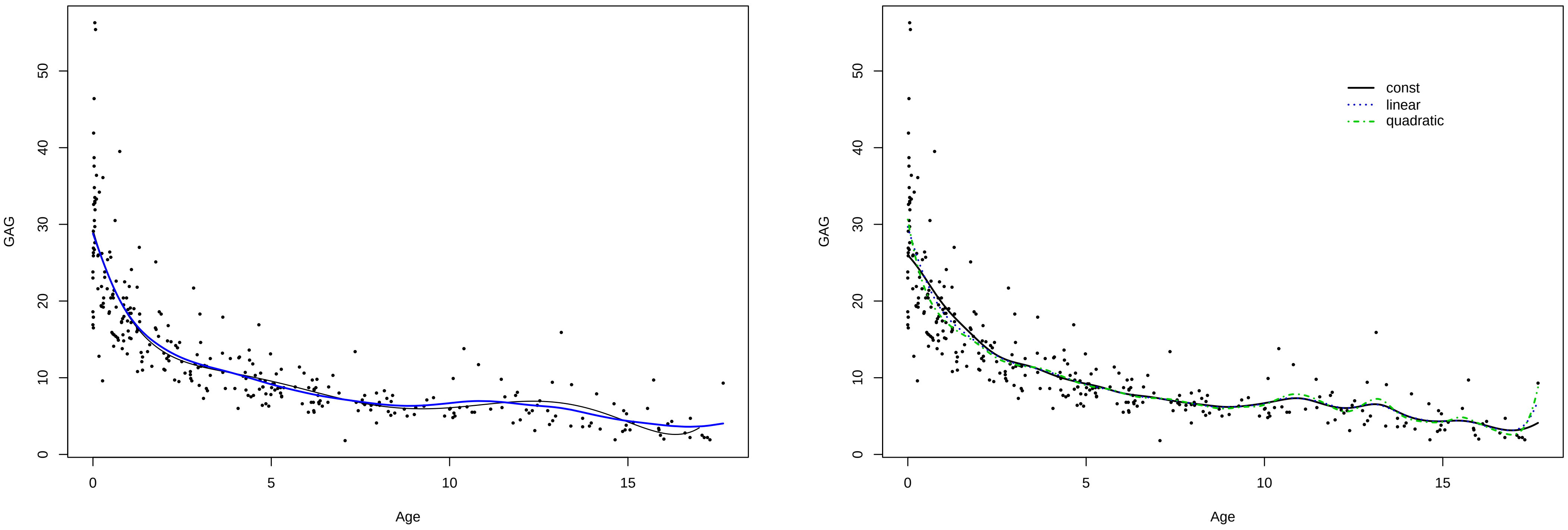
Is this good enough?

Smoothing splines would be the numerical analyst's way to fit a smooth curve to such a scatterplot. The issue is 'how smooth' and in this example it has been chosen automatically by GCV.

```
> plot(GAGurine, pch=20)
> lines(smooth.spline(Age, GAG), lwd = 3, col="blue")
```

An alternative would be *local* polynomials, using a kernel to define 'local' and choosing the bandwidth automatically.

```
> plot(GAGurine, pch=20)
> (h <- dpill(Age, GAG))
> lines(locpoly(Age, GAG, degree = 0, bandwidth = h))
> lines(locpoly(Age, GAG, degree = 1, bandwidth = h), lty = 3)
> lines(locpoly(Age, GAG, degree = 2, bandwidth = h), lty = 4)
```


GAG
0
10
20
30
40
50
0
5
10
15
Age
GAG
0
10
20
30
40
50
0
5
10
15
Age
const
linear
quadratic

(Practical) model selection in 2004

. . .

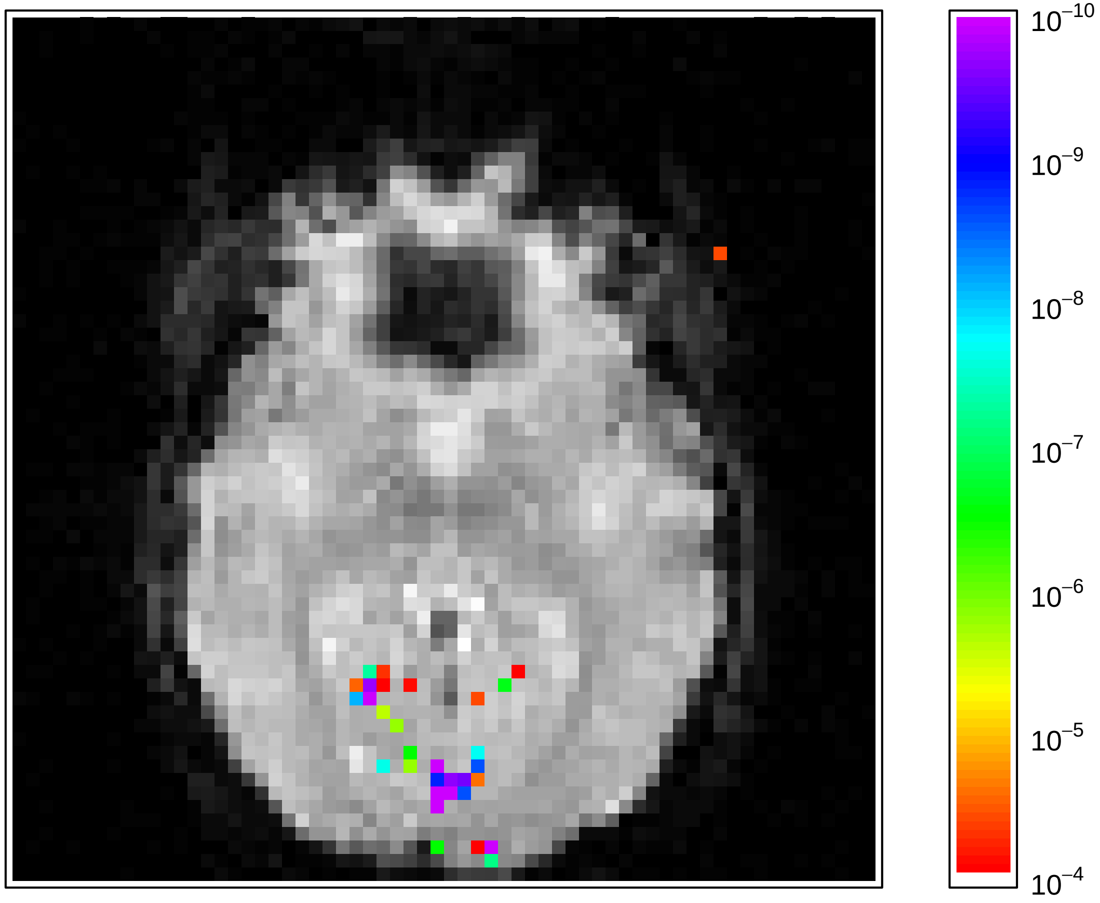
- There are lots of formal ‘figures of adequacy’ for a model. Some have proved quite useful, but
  - Their variability as estimators can be worrying large.
  - Computation, e.g. of ‘effective number of degrees of freedom’, can be difficult.
  - Their implicit measure of performance can be overly sensitive to certain aspects of the model which are not relevant to our problem.

  The assumptions of the theories need to be checked, as the criteria are used way outside their known spheres of validity (and in some cases where they are clearly not valid).

- Nowadays people do tens of thousands of significance tests, or more.

Plotting multiple p values

p-value image of a single fMRI brain slice thresholded to show p-values below 10^{-4} and overlaid onto an image of the slice. Colours indicate differential responses within each cluster. An area of activation is shown in the visual cortex.



- Formal training/validation/test sets, or the cross-validatory equivalents, are a very general and safe approach.
- ‘Regression diagnostics’ are often based on approximations to overfitting or case deletion. Now we can (and some of us do) fit extended models with smooth terms or use fitting algorithms that downweight groups of points. (I rarely use least squares these days.) It is still all too easy to select a complex model just to account for a tiny proportion of aberrant observations.
- Alternative explanations with roughly equal support are commonplace. Model averaging seems a good solution. Selecting several models, studying their predictions and taking a consensus is also a good idea, *when time permits* and when *non-quantitative information is available*.

Epilogue

My memory (which I hope is reliable enough) is that I first encountered ‘Nelder’ as an commentator in an ornithology journal, playing Sherlock Holmes over the suspiciously large number of rare birds reported from near Hastings at around the turn of the 20th century.

My friend and co-author Bill Venables (an avid birdwatcher) tells me John is celebrating his 80th birthday by birdwatching in Australia, including visiting Kakadu National Park in NT (highly recommended from our 2003 visit).

So here is a little practice, with an Australian bias.