Model Choice

Lecture 2

Brian Ripley

http://www.stats.ox.ac.uk/∼ripley/ModelChoice/

Bayesian approaches

Note the plural — I think Bayesians are rarely Bayesian in their model selection, and Geisser's quote showed that model choice is perhaps not a strict Bayesian concept.

Assume M (finite) models, exactly one of which is true, and let $\mathcal{T}$ indicate the data in the training set.

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*.

We assume (often implicitly) that models have equal prior probabilities.

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 very 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 a Laplace approximation and drop various terms, eventually reaching

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

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

BIC aka SBC

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 minus two. As with AIC, the model with minimal BIC is chosen. Thus

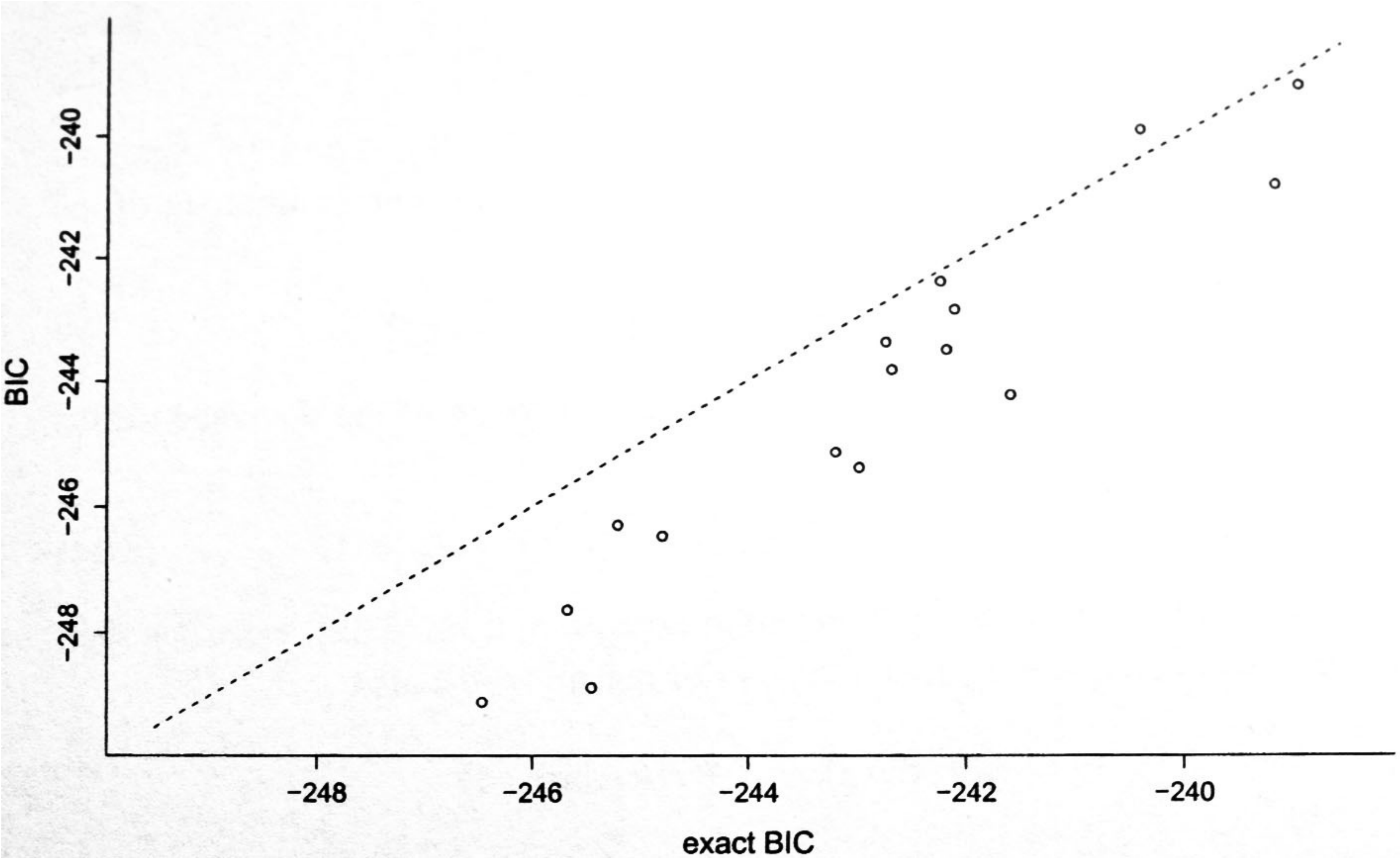
$$BIC = -2\log p(\mathfrak{T} \mid M) + \text{const} \approx -2L(\widehat{\theta}; \mathfrak{T}) + (\log n)\,p$$

There are a number of variants, depending on e.g. which terms are dropped in the approximations and what estimator is used for $\widehat{\theta}$ (MLE, MAP, …).

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. There is a true model.
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 systematically overfit nor underfit.



BIC *vs* the exact formula, for variable selection in a 4-variable logistic regression on 189 subjects.

Hannan–Quinn and asymptotics

The criterion of Hannan & Quinn (1979) is

$$HQ = 2L(\widehat{\theta}; \mathcal{T}) - (2c \log \log n)p$$

where $c > 1$ is a constant. It was originally derived for determining the (true) order p_0 of an $AR(p)$ process.

Like BIC it is (under regularity conditions) *strongly consistent*: that is if there is a single true p the strategy of using BIC or HQ will asymptotically choose p with probability tending to one. This is also true of AIC.

If there is not a true p we can ask for the model which is *least false*, that the $\widehat{P}$ which is closest to the true P in the sense of Kullback-Leibler divergence.

The problem is that for nested models such as $AR(p)$, if one model is true then so is any larger model. In that case, BIC and HQ are still strongly consistent, but AIC will tend to choose increasingly complex models and will select an order $\widehat{p} \geq p_0$ with probability tending to one.

For prediction, the relevant asymptotics concern the efficiency of the predictions, for example the mean square error or error rate of the predictions.

For variable selection in a regression, AIC, PRESS and C_p are asymptotically efficient, and BIC and HQ are not. Similarly for $AR(p)$, AIC and AICC are asymptotically efficient and BIC and HQ are not.

Best of both worlds?

AIC has desirable asymptotic properties for prediction and BIC for explanation. Is it possible to have both?

People have tried, but Yang (*Biometrika* 2005) showed that the answer is essentially no (and still no if model averaging is allowed).

Example – Blood groups

Landsteiner's classification of human blood groups as [O A B AB] clearly suggests two 'antigenes' are involved (O = 'ohne'). There were two theories as to what Mendelian genetics are involved: one had three alleles at a single locus, the other two alleles at each of two loci, each of which give two-parameter models for the frequencies of occurrences of the blood types.

Bernstein (1924) collected data on 502 Japanese living in Korea, and found $N_A = 212, N_B = 103, N_{AB} = 39, N_O = 148$. This gives $-2L(\widehat{\theta}; \mathcal{T})$ (at the MLEs) of 1254.2 and 1293.9.

The models are not nested, and they have the same number of free parameters. So although we cannot use a testing framework (except perhaps that of Cox (1961)), both AIC and BIC can be used and strongly support the single-locus model.

Deviance Information Criterion (DIC)

From Spiegelhalter *et al.* (2002).

For a Bayesian setting where prior information is not negligible, and the model is assumed to be a good approximation but not necessarily true.

Consider the *deviance* to be

$$D(\theta) = \text{deviance}(\theta) = \text{const}(\mathcal{T}) - 2L(\theta; \mathcal{T})$$

In GLMs we use $D(\widehat{\theta})$ as the (scaled) (residual) deviance.

Define

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

The first overline means averaging θ over $p(\theta \mid \mathcal{T})$, and the second means our estimate of the 'least false' parameter value, usually the posterior mean of θ (but perhaps the median or mode of the posterior distribution). Then define

$$DIC = D(\bar{\theta}) + 2\,p_D$$

Clearly DIC is AIC-like, but

- Like NIC it allows for non-ML fitting, in particular it can incorporate the regularization effect of the prior that should reduce the effective number of parameters.
- It is not necessary (but is usual) that $p_D \geq 0$.
- DIC is explicitly meant to apply to non-nested non-IID problems.
- DIC is intended to be approximated via MCMC samples from the posterior density of θ given $\mathcal{T}$. On the other hand, DIC needs an explicit formula for the likelihood (up to a model-independent normalizing constant).

In practice p_D is often close to p, and indeed asymptotically is equivalent to the modified p used by Takeuchi and in NIC.

Focussed Information Criterion (FIC)

AIC is aimed at prediction in general, but supposed we are interested in a specific prediction. Claeskens & Hjort (2003–) considered how this should affect the strategy for model selection. The criterion they derive is rather complicated, but called *FIC*.

Suppose we have a (multiple) linear regression and we are interested in prediction at a specific point in the covariate space, or we are interested in a single regression coefficient. Then it is reasonable that the model we choose for prediction might depend on which point or which coefficient.

Note though that that applies to the whole model selection strategy, not just to a formal criterion—the panoply of models will very likely depend on the question.

Example of FIC

A classic example is the dataset `menarche` in MASS on age of menarche of 3918 girls in Warsaw. This can be fitted by a logistic regression, but there is some evidence that non-linear terms in age are needed, especially for small ages.

Thus FIC selects a linear model for predictions near the center of the age distribution, a quadratic model for the first quartile and a quartic model for prediction of the 1% quantile (at 10.76 years).

However, there are some strange features of this dataset: 27% of the data are observations that all 1049 17-year-old girls had reached menarche. It is unclear that using all the data (or an unweighted MLE) is appropriate for the questions asked.

Shrinkage

The reason we might want to use variable selection for prediction is not that we think some of the variables are irrelevant (if we did, why were they measured in the first place?), but that they are not relevant enough.

The issue is the age-old one of bias *vs* variance. Adding a weakly relevant variable reduces the bias of the predictions, but it also increases the uncertainty of the predictions.

Consider a multiple linear regression. For small values of the true regression coefficient β_i, the mean square error of the (interesting) predictions will decrease if $\widehat{\beta}_i$ is forced to be zero, whereas for larger values it will increase. Criteria such as FIC are estimating if the effect of dropping the variable is beneficial.

Can we avoid this? Yes, if we don't force $\widehat{\beta}_i$ to zero, but we *shrink* it towards zero. This results in biased but less variable predictions.

Ridge regression

For this slide only suppose we have an orthogonal (or even orthonormal) regression.

Suppose we shrink all our regression coefficients (or all but the intercept) by a factor λ for $0 < \lambda \leq 1$. You can show that this always reduces the MSE for small λ. So some (unknown) amount of shrinkage is always worthwhile. James & Stein showed that for $p > 2$ there was a formula for shrinkage that would always reduce MSE, but unfortunately it might correspond to $\lambda < 0$. Sclove's variant is

$$\tilde{\beta}_i = \left(1 - \frac{c}{F}\right)^+ \widehat{\beta}_i$$

where $c > 0$ and F is the F-statistic for all the coefficients being zero.

(`http://www.stat.ucla.edu/~cocteau/stat120b/lectures/lecture4.pdf`).

In *ridge regression* we minimize the sum of squares plus λ times the sum of squares of the coefficients (usually omitting the intercept, and suitably scaling the variables).

Lots of related ideas, including really setting small $\tilde{\beta}_i = 0$ and shrinking larger ones (e.g. ‘LASSO’).

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? Basic decision theory tells us:

> *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. [Box quote!] In practice Bayes factors can depend heavily 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.

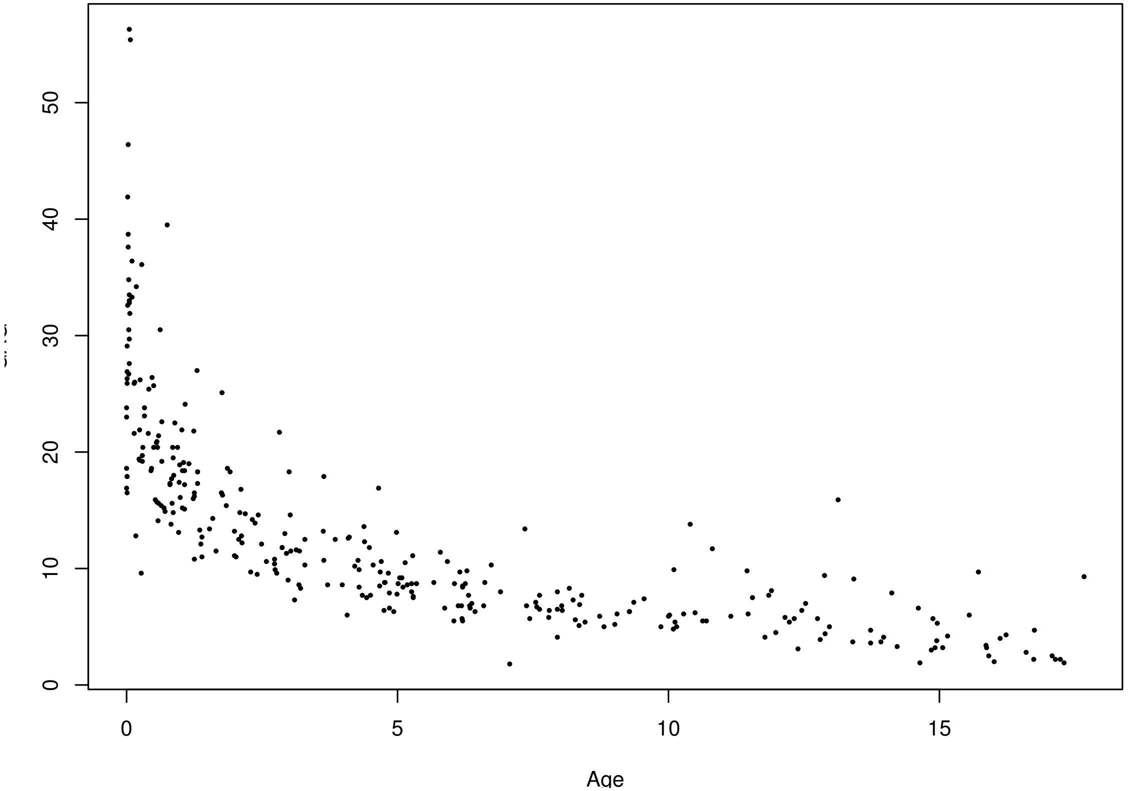
In *random forests* the tree-construction algorithm randomly restricts itself at the choice of each split.

(Practical) model selection in 2010

- The concept of a model ought to be much, much larger than in 1977. Even in the 1990s 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.

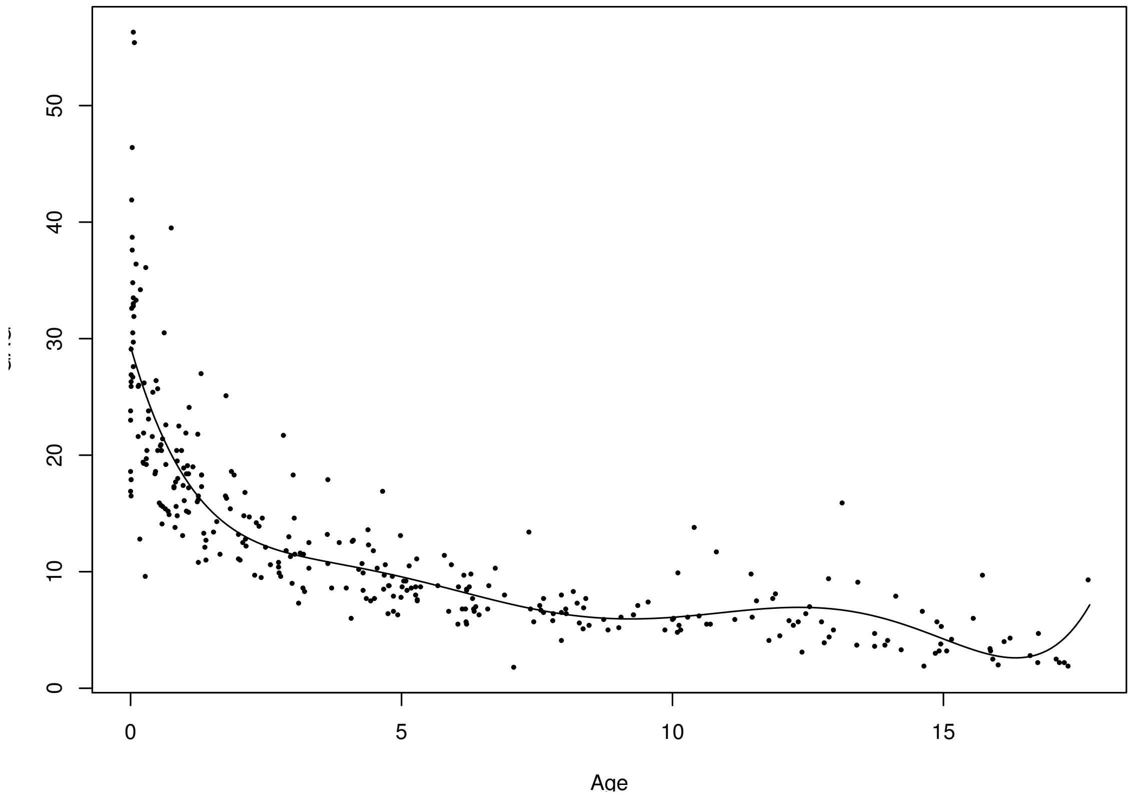
Calibrating GAG in urine

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?

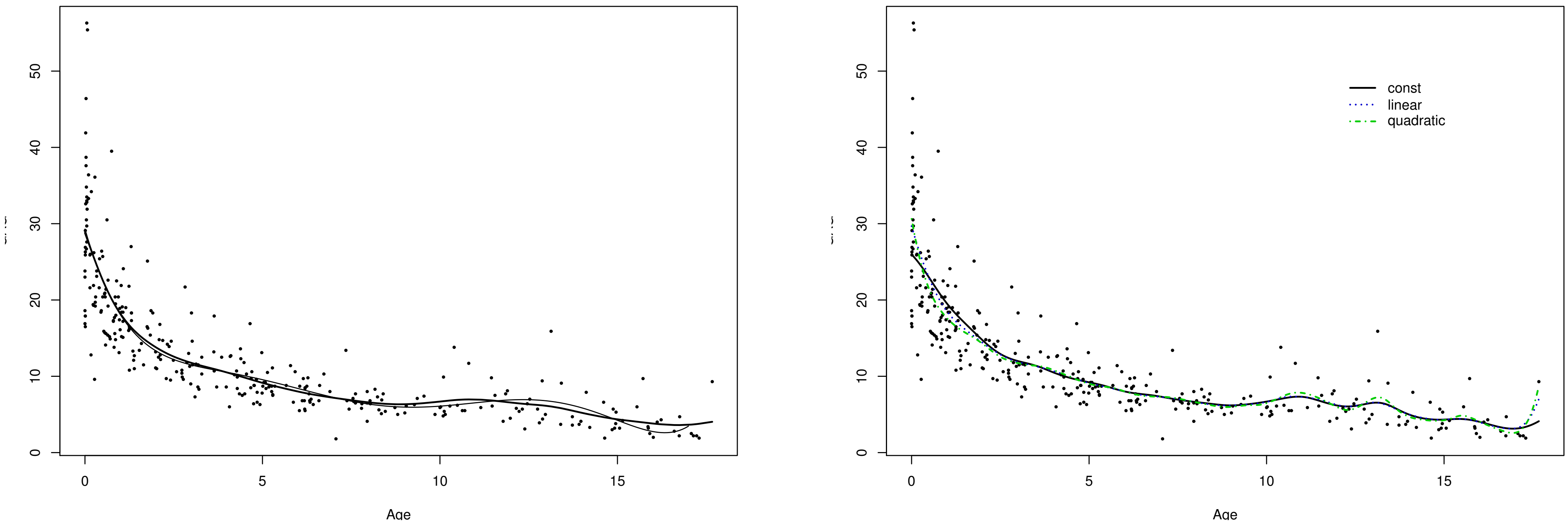
Smoothing splines would be the numerical analyst's way to fit a smooth curve to such a scatterplot. The issue is 'how smooth' and in this example it has been chosen automatically by GCV.

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

Neural networks are another global non-linear model that are usually fitted by penalized least squares.

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)
```


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

(Practical) model selection in 2010

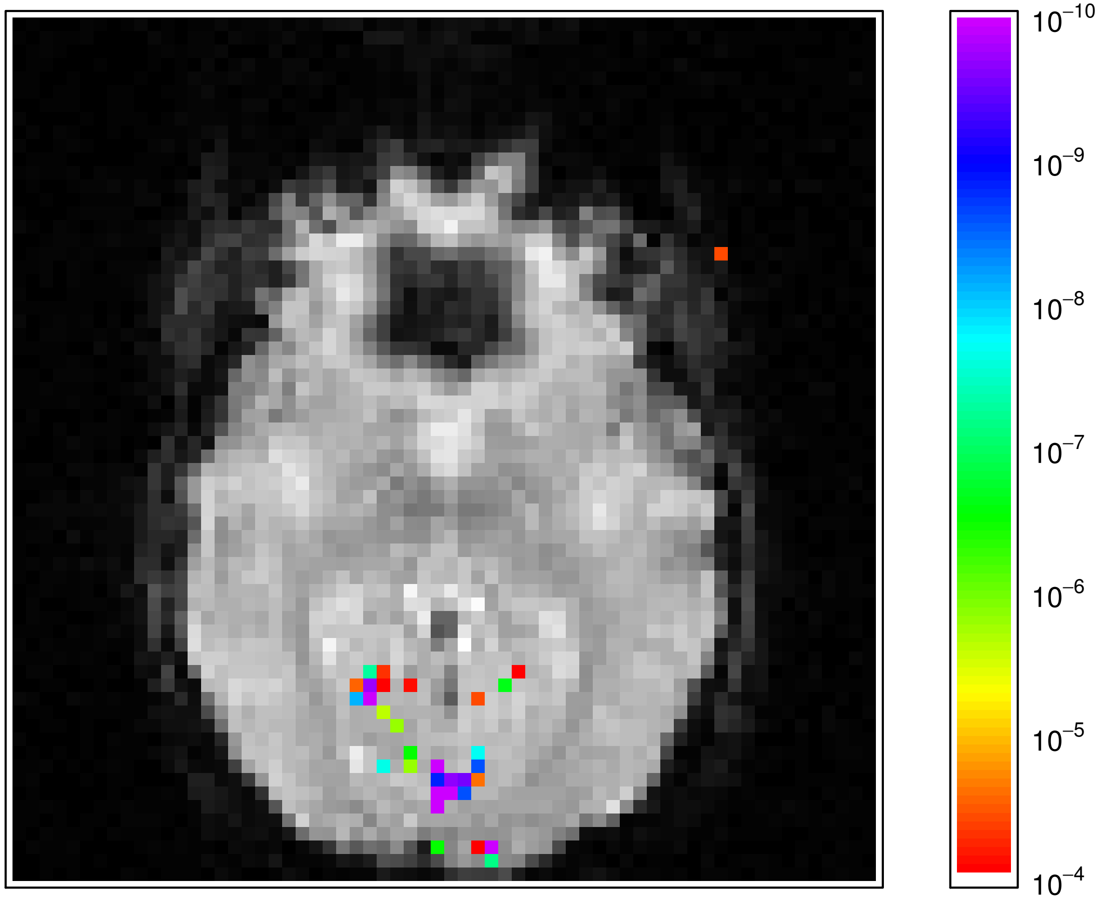
. . .

- 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*.