Appropriate Model Selection

1 Scientific Issues

1.1 Start with good scientific question(s) and hypotheses
- 1.1.1 good questions in science
- 1.1.2 avoid null hypothesis vs. alternative hypothesis dichotomy
- 1.1.3 multiple working hypotheses, all plausible

1.2 Limit the model set
- 1.2.1 avoid spurious results from considering too many models
- 1.2.2 too many models –> all w_i values small.

1.3 Do not fret about lack of "true" model
- 1.3.1 info. theoretic approach does not assume "true" model among alternatives considered
- 1.3.2 information theoretic approach does not assume full reality can be parameterized
- 1.3.3 finding best model ≠ finding "truth"
- 1.3.4 goal: find best-fitted model in set considered

1.4 Do not combine Information-theoretic methods with statistical hypothesis tests
- 1.4.1 no need to state α, *P*-values, etc.
- 1.4.2 instead, evaluate relative support for each model
- 1.4.3 extended analysis: multimodel inference based on model averaging, etc

2 Methodological Issues

2.1 Model each hypothesis well
- – seek advice of statistician or modeler
- – consider nonlinear models

2.2 Determine model selection uncertainty
- – beware results of analysis lacking uncertainty estimates (e.g., regression software output)
- – predictions may differ among models;
- –> need estimate of model selection precision (uncertainty)

2.3 With overdispersed (count) data, use $QAIC_c$
- – see Burnham and Anderson (2002) p.67-70.

2.4 Do not explain data post-hoc
- 2.4.1 Clearly distinguish between a priori hypotheses and post-hoc data exploration
- 2.4.2 Conduct analysis based on a priori thinking, followed by some post hoc consideration; Never do the reverse.

2.5 Statistical significance vs. quantitative evidence
- 2.5.1 statistical significance ≠ biological importance!
- 2.5.2 *P*-values do not provide weight of evidence for ideas

2.6 Assess goodness of fit using global model

2.7 Provide all relevant information, for each model considered:
- 2.7.1 Maximized log-likelihood
- 2.7.2 Number of estimated parameters (K)
- 2.7.3 Information criterion used (e.g., AIC_c)
- 2.7.4 Criterion differences (Δ_i)
- 2.7.5 Akaike weights (w_i)

3 Avoiding Mistakes

3.1 Incorrect number of parameters, *K*

3.2 Using AIC when AIC_c warranted

3.3 Comparing AIC across different data sets

3.4 Comparing AIC among different response variables (*y*)

3.5 Failure to converge on numerical parameter estimates
– e.g., failure to find global maximum of log-likelihood function

4 Multimodel Inference: Model Averaging

4.1 If one model clearly best (e.g., $w_i \geq 0.90$), draw inferences from that model.

4.2 If no single model clearly superior, better to use weighted estimate for prediction.

4.3 Use Akaike weights (w_i) to generate weighted estimate.

4.4 Model averaging:

$$\hat{\hat{\theta}} = \sum_{i=1}^{R} w_i \hat{\theta}_i$$

$\hat{\hat{\theta}}$ = model averaged estimate of θ
$\hat{\theta}$ = predicted value of θ
w_i = Akaike weight of model i
R = number of models considered

5 Model Selection Uncertainty

5.1 Crudely analogous to standard error of the mean

5.2 Result of model selection repeated with different, independent data set?
– different model selected? different parameter estimates?

5.3 Unconditional variance estimate for parameter θ:

$$\text{var}(\hat{\hat{\theta}}) = \left[\sum w_i \sqrt{\text{var}(\hat{\theta}_i \mid g_i) + (\hat{\theta}_i - \hat{\hat{\theta}})^2} \right]^2$$

5.4 Use for model averaged estimator $\hat{\hat{\theta}}$, or maximum likelihood estimate $\hat{\theta}$ of selected model.

6 Confidence Set for Best Model

6.1 Analogous to confidence interval for parameter estimate.
6.2 Three approaches to determine confidence set
6.2.1 Treat Akaike weights as posterior probabilities.
Sum Akaike weights from largest to smallest, until sum $\geq 1-\alpha$.
(e.g., 0.95 for 95% confidence set) Confidence set is all models included in sum.

6.2.2 Treat Δ_i as random variable with sampling distribution.
General guidelines, for independent observations, large sample sizes, nested models:

Δ_i	Empirical Support
0–2	strong
4–7	weak
>10	essentially none

6.2.3 Relative model likelihood, with threshold value for confidence set.

Select cutoff point (C) for ratio of likelihoods for models in confidence set:

$$\frac{\mathcal{L}(g_i \mid x)}{\mathcal{L}(g_{\min} \mid x)} > 1/C, \qquad \text{note: } \frac{\mathcal{L}(g_i \mid x)}{\mathcal{L}(g_{\min} \mid x)} \equiv \exp\left(-\frac{1}{2}\Delta_i\right)$$

where x are the data, $\mathcal{L}(g_i \mid x)$ is likelihood of model i, and $g_{\min}$ is the best model.

For example, if $C = 5$, then confidence set would include all models with $\Delta_i < 3.2$.

If $C = 20$, then confidence set would include all models with $\Delta_i < 6$.

7 Relative Importance of Variables

7.1 Analogous to confidence interval for parameter estimate.

7.2 Multiple regression: common (but poor) practice to select final model (e.g., stepwise), then conclude included variables are important, unselected variables not important.

7.3 Problem: ignores model selection uncertainty

7.4 Better method: sum Akaike weights (w_i) across all models

7.5 Relative importance of variable j = sum w_i over all models containing j; $w_+(j)$

– larger $w_+(j) \Rightarrow$ variable j more important, relative to other variables

7.6 Example: linear regression with up to 3 independent variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

8 possible models (1: variable included; 0: excluded), with Akaike weights, w_i

x_1	x_2	x_3	w_i
0	0	0	0.00
1	0	0	0.10
0	1	0	0.01
0	0	1	0.05
1	1	0	0.04
1	0	1	0.50
0	1	1	0.15
1	1	1	0.15

Best model: $w_i = 0.5$, i.e., $P\{\text{best model}\} = 1/2$

Summed weights for x_1, $w_+(1) = 0.79$

x_2 not in best model, but nonzero importance: $w_+(2) = 0.35$

Summed weights for x_3, $w_+(3) = 0.85$; much greater than best model itself

Conc: variables ordered by importance: x_3, x_1, x_2 with importance weights 0.85, 0.79, 0.35

Further Reading:

Anderson, DR, KP Burnham. 2002. Avoiding pitfalls when using information-theoretic methods. *J. Wildl. Manage.* 66:912-918.

Buckland, ST, KP Burnham, NH Augustin. 1997. Model selection: an integral part of inference. *Biometrics* 53:603-618.

Burnham, KP, Anderson, DR. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. Springer-Verlag, New York.