Model selection

Outline for today

- The problem of model selection
- Choose among models by a criterion rather than significance testing
- Criteria: Mallow's C_p and AIC
- Search strategies: All subsets; stepAIC
- Several models may fit about equally well
- The science part: formulate a set of candidate models

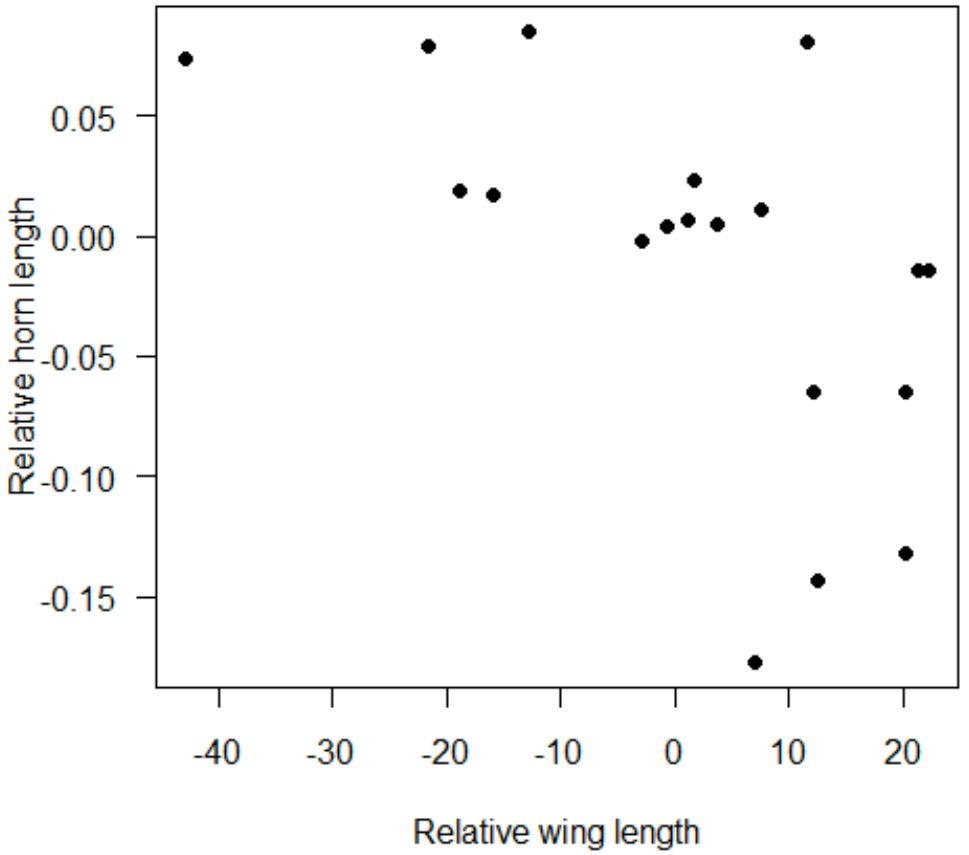
The problem of model selection (example problems)

- When fitting a polynomial, how to decide the complexity: straight line regression, quadratic, cubic, … when to stop.
- How to decide which variables to retain and which to discard when building a multiple regression model.
- When can interaction terms be removed from an ANOVA model.
- How best to decide among candidate models representing different biological processes.

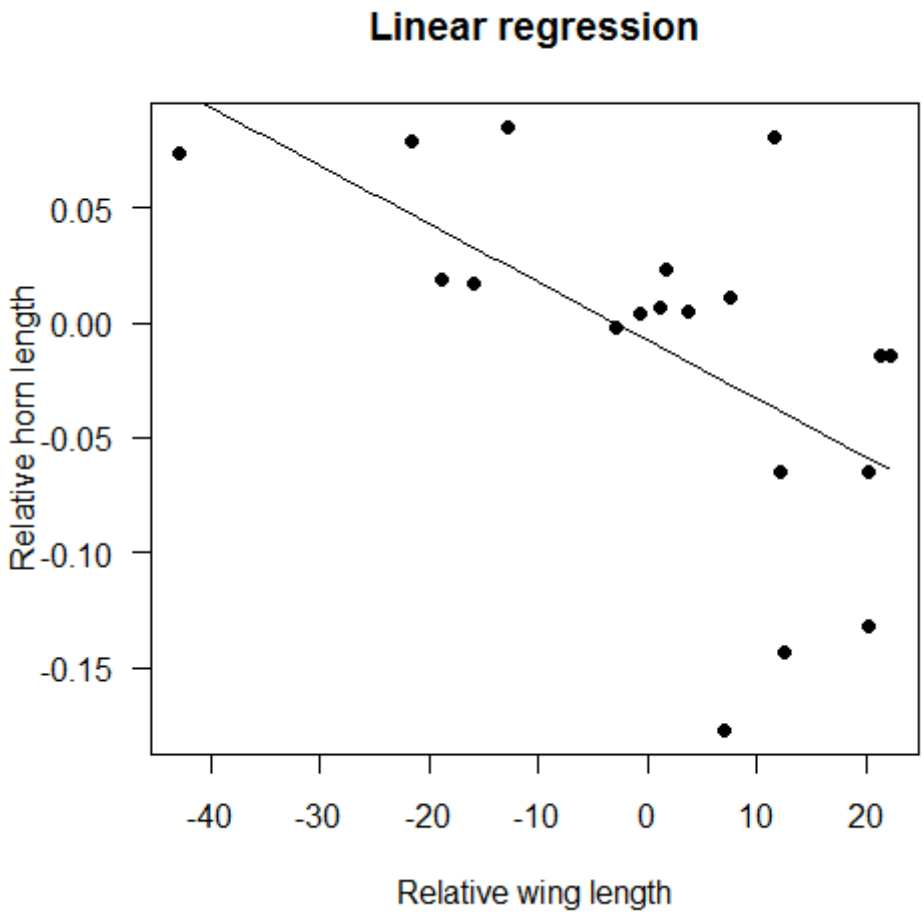
Example 1: Fit a polynomial regression model

Data: Trade-off between the sizes of wings and horns in 19 females of the beetle *Onthophagus sagittarius*. Both variables are size corrected.

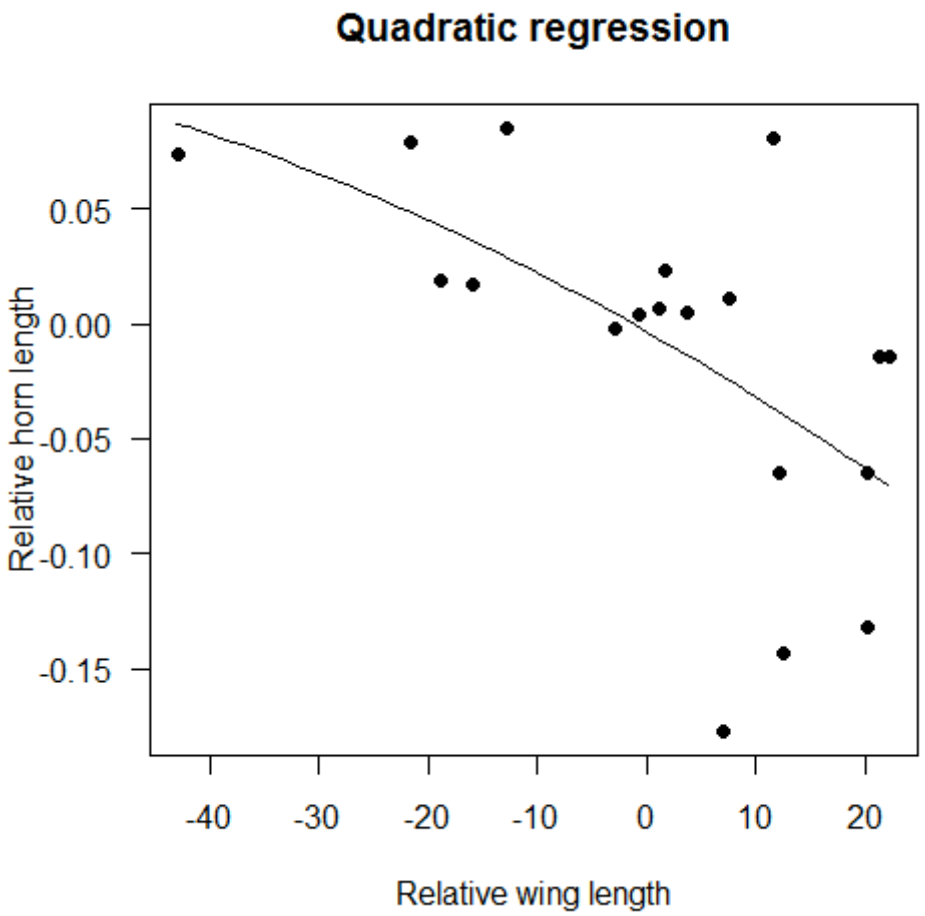
Emlen, D. J. 2001. Costs and the diversification of exaggerated animal structures. Science 291: 1534-1536.



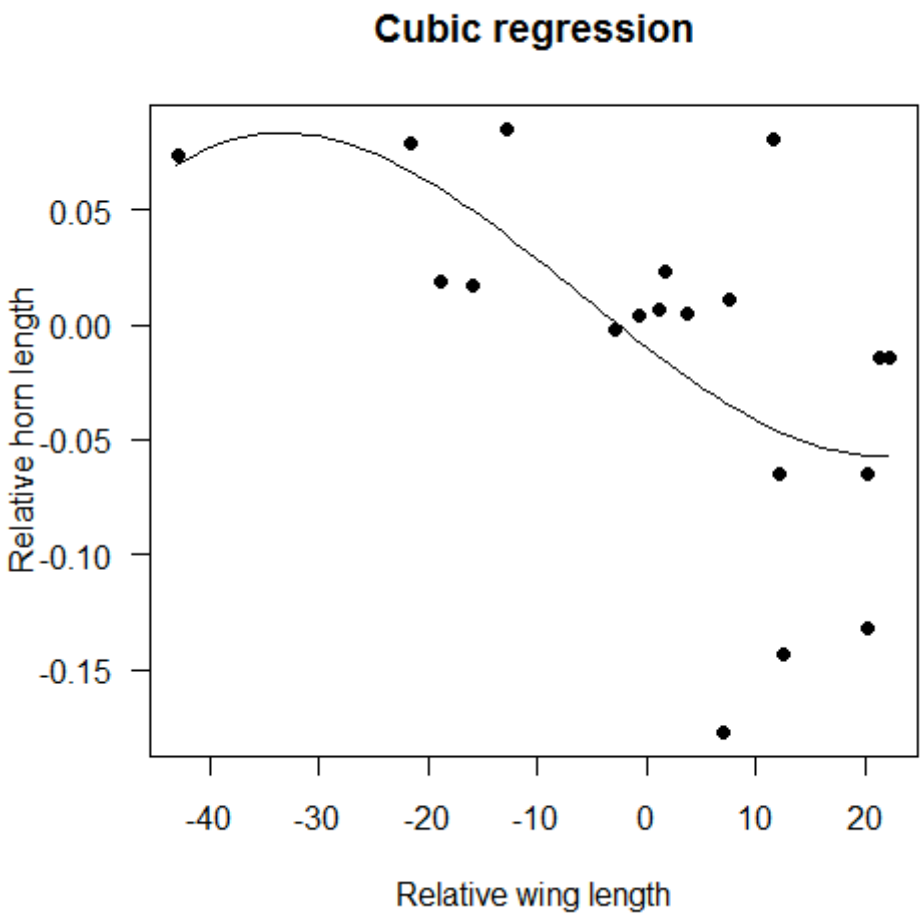
Start with a linear regression



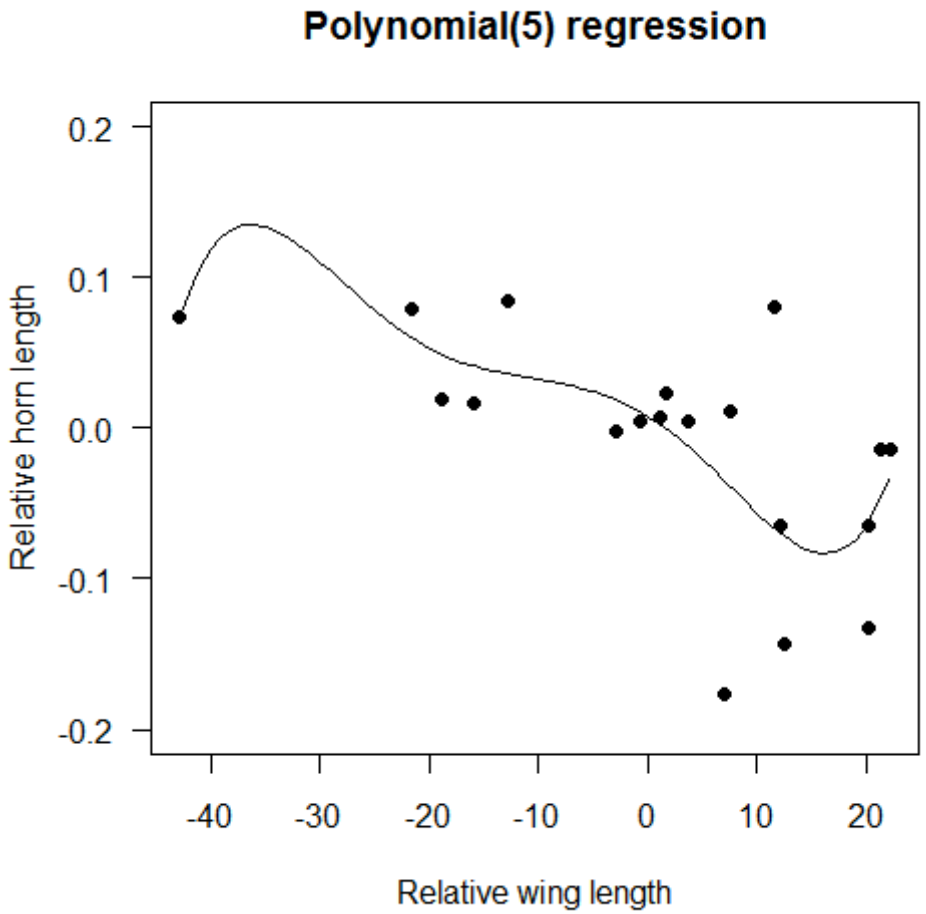
Why not a quadratic regression instead (polynomial degree 2)



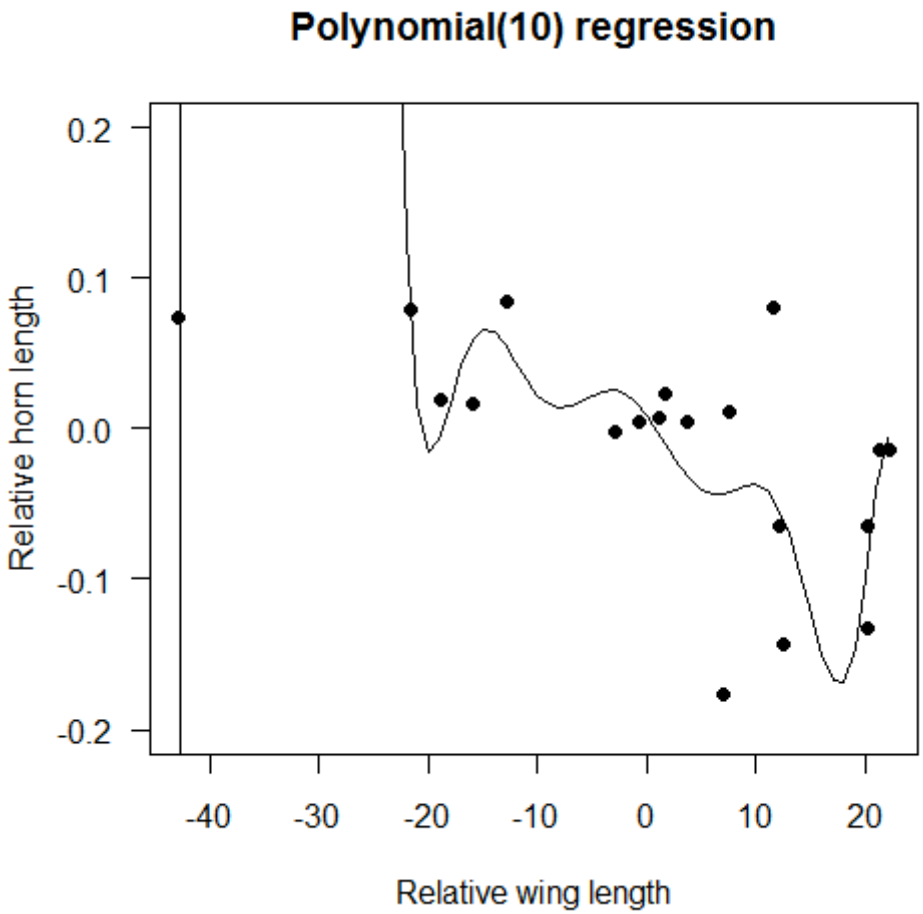
How about a cubic polynomial regression (degree 3)



Better still, a polynomial degree 5

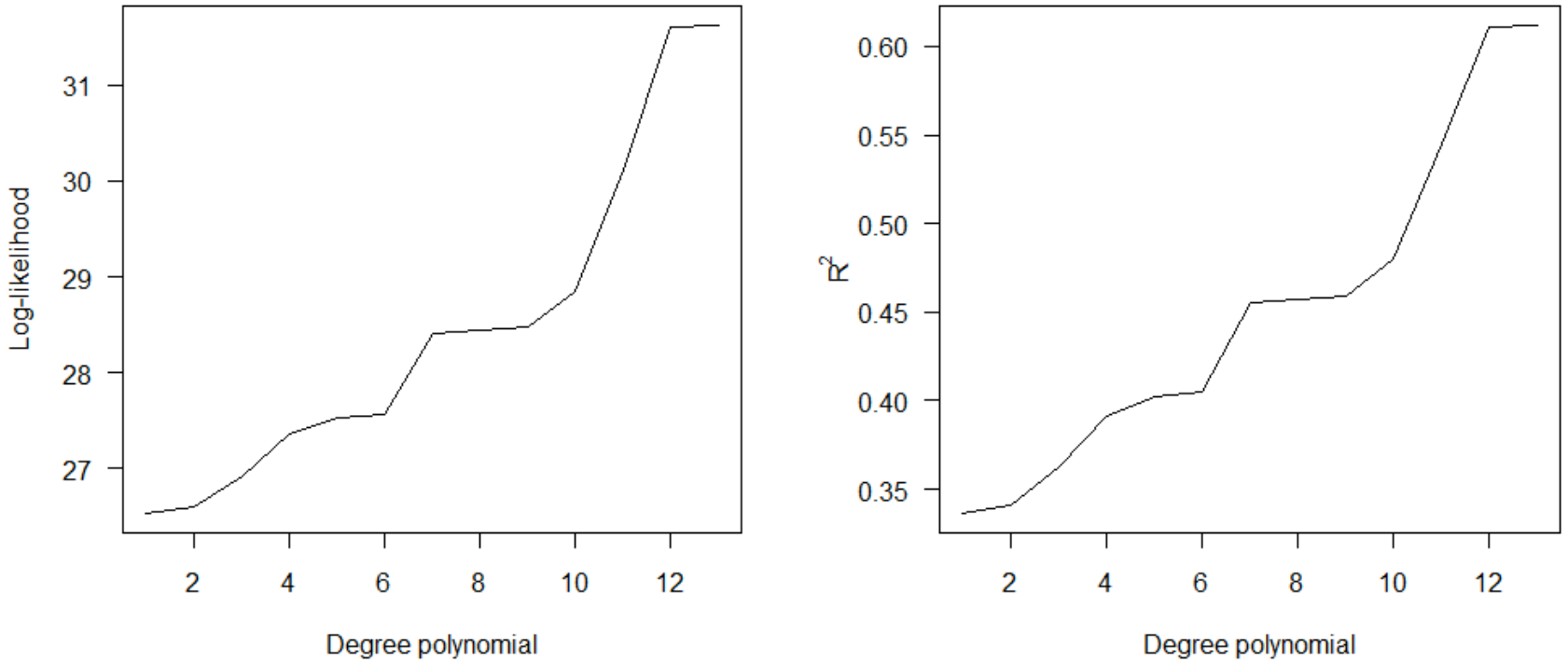


A polynomial, degree 10

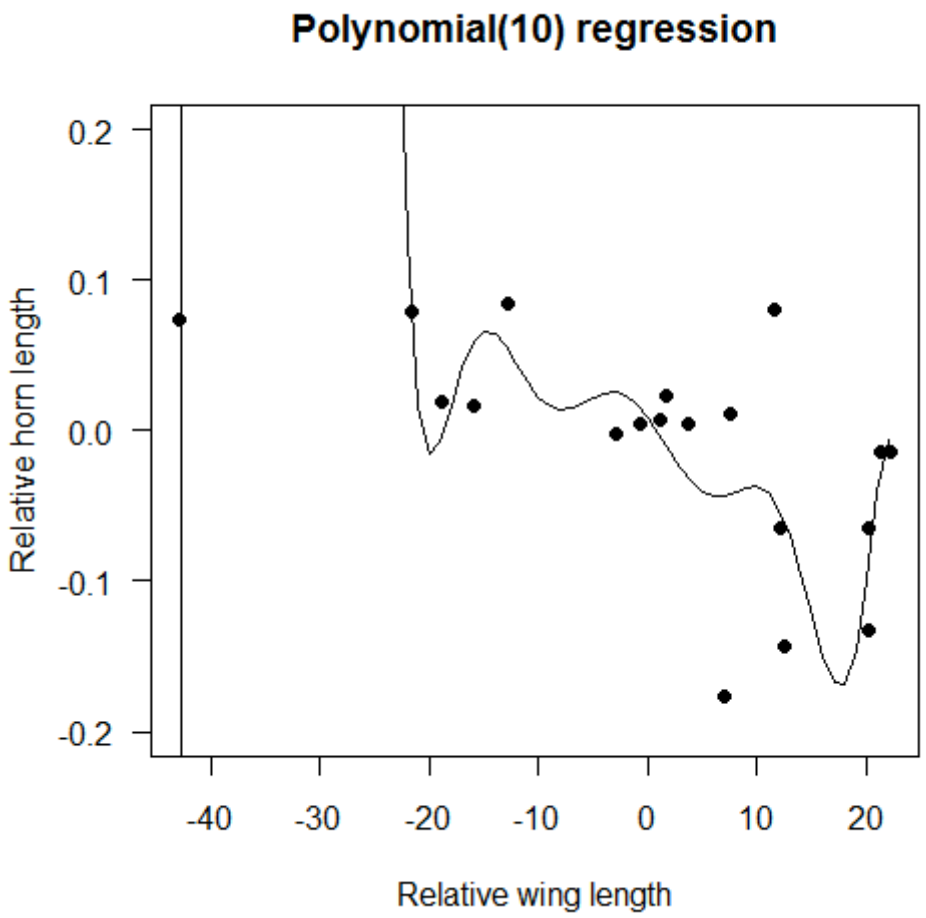


R^2 and log-likelihood increase with number of parameters in model

Isn't this good? Isn't this what we want, the best fit possible to data?



What is wrong with a complicated graph that best fits the data

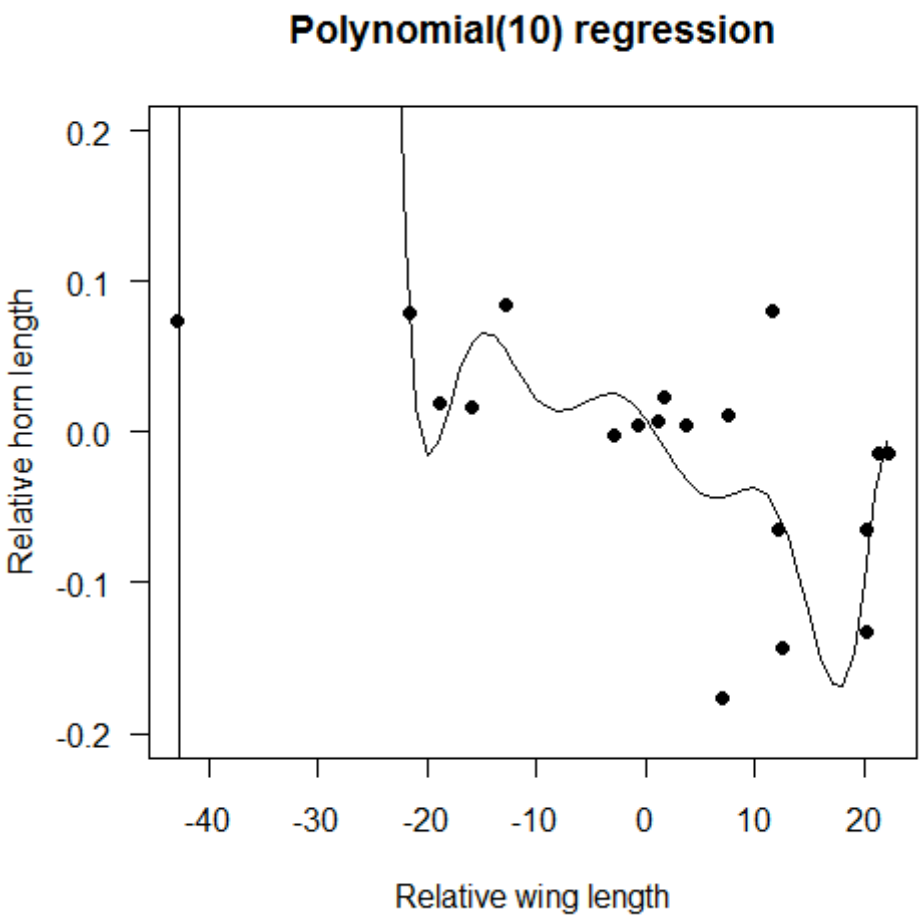


Does it violate a principle of parsimony

Fit no more parameters than is necessary. If two or more models fit the data almost equally well, prefer the simpler model.

"models should be pared down until they are minimal adequate"

-- Crawley 2007, p325



Should we simplify by stepwise elimination of terms

An approach in common use (e.g., recommended by Crawley).

Approach involves a cycle of deleting model terms that are not statistically significant and then refitting. Continue until only statistically significant terms remain.

But does this approach actually yield the “best” model. Or is it a recipe for committing a sequence of Type 2 errors.

Each instance in which a variable is dropped from the model involves “accepting” a null hypothesis.

“Cross-validation score” is one way to estimate prediction error:

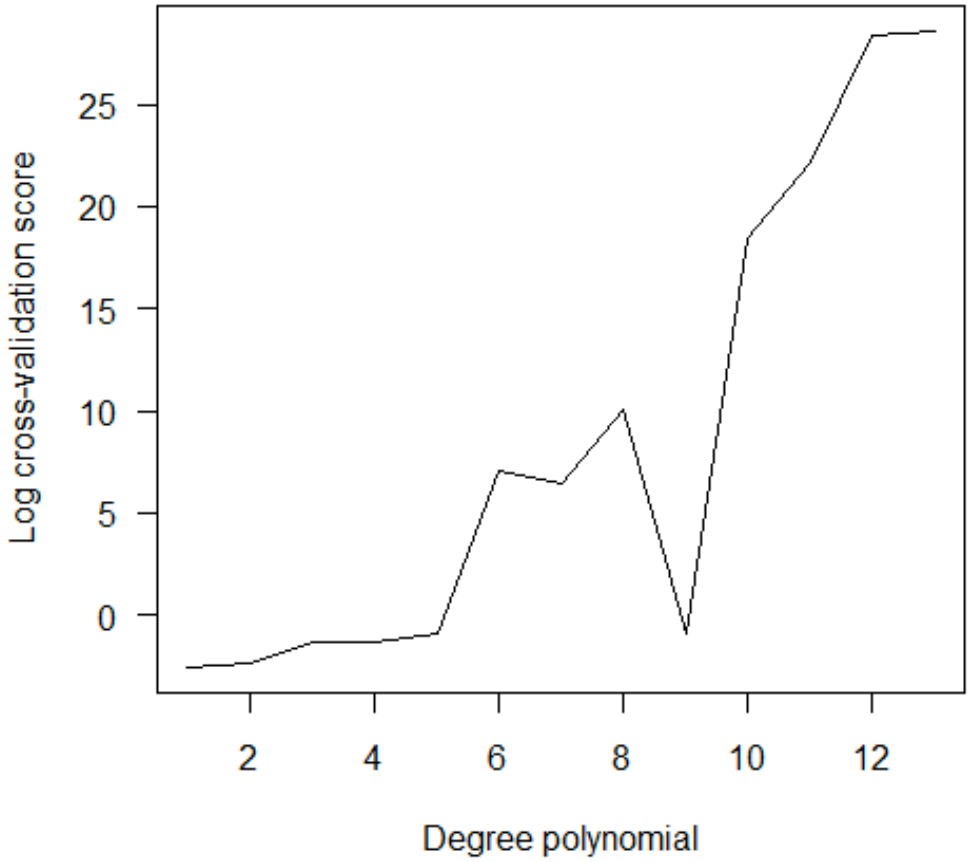
$$\text{cv.score} = \sum e_{(i)}^2$$

where

$$e_{(i)}^2 = y_i - \hat{y}_{(i)}$$

$\hat{y}_{(i)}$ is the predicted value for y_i when the model is fitted to the data leaving out y_i.

In this example, the score worsens with increasing numbers of parameters in the model. Here, linear is best!



Time for a pre-break shocker

“All models are wrong but some are useful.”

George Box

[break]

Let's reconsider our objectives

- Want a model that predicts well
- Want one that approximates the true relationship between the variables
- Would like to be able to evaluate a wider array of models than those in which one or more "reduced" models is necessarily a subset of the other, "full" model.

Note: Reduced vs. full models are referred to as "nested models", whereas models in which one is not the subset of the other are called "non-nested" models. Not to be confused with nested experimental designs or sampling designs.

How to accomplish these goals

To answer this, we need

- A **criterion** to compare models:
  - Mallow's C_p
  - AIC (Akaike's Information Criterion)
  - BIC (Bayesian Information Criterion)

- A **strategy** for searching the candidate models

Mallow's C_p is frequently used in multiple regression

Criterion: Mallow's C_p.

$$C_p = \frac{\mathrm{SS}_{\mathrm{error}}}{\hat{\sigma}^2} - n + 2p$$

$\mathrm{SS}_{\mathrm{error}}$ is the error sum of squares for the model with *p* predictors
$\hat{\sigma}^2$ is the estimated error mean square of the true model (e.g., all predictors).
n is the sample size.
p is the number of predictors (including intercept).

C_p estimates the mean square prediction error. It is related to AIC.

The *2p* behaves like a penalty for including too many predictors (explanatory variables). This feature is shared with all other model selection criteria.

Mallow's C_p is frequently used in multiple regression
Here we show its use in "all subsets regression"

Strategy: Test all possible models and select the one with smallest C_p

Implemented in the `leaps` package in R, It uses an efficient algorithm to choose among a potentially huge number of models.

So we are not dealing with data from an experiment here, where we can make intelligent choices based on the experimental design. Typically we are modeling observational data.

By investigating all possible subsets of variables, we are admitting that the only intelligent decision we've made is the choice of variables to try. No other scientific insight was used to decide an *a priori* set of models.

Example 2a: Ant species richness

Data: Effects of latitude, elevation, and habitat on ant species richness

Gotelli, N.J. & Ellison, A.M. (2002b). Biogeography at a regional scale: determinants of ant species density in bogs and forests of New England. Ecology, 83, 1604–1609.

```
ants
   site nspecies habitat latitude elevation
1   TPB        6  forest    41.97       389
2   HBC       16  forest    42.00         8
3   CKB       18  forest    42.03       152
4   SKP       17  forest    42.05         1
...
23  TPB        5     bog    41.97       389
24  HBC        6     bog    42.00         8
25  CKB       14     bog    42.03       152
26  SKP        7     bog    42.05         1
...
```

(Bog and forest sites were technically paired by latitude and elevation, but residuals were uncorrelated, so we'll follow authors' lead in treating data as independent for the purposes of this exercise)

Example 2a: Ant species richness

Full regression model:

```
z<- lm( log(nspecies) ~ habitat * latitude * elevation)
```

“Leaps” requires that all variables be numeric (we can disguise habitat as a numeric variables by scoring: 0=bog, 1=forest)

All subsets of Habitat, Latitude, Elevation and their 2- and 3-way interactions were tested.

Not all the models are necessarily sensible (what sort of intellect would deliberately fit a model with a 3-way interaction and no main effects). The approach is most sensibly applied to large sets of numeric variables, neglecting interactions)

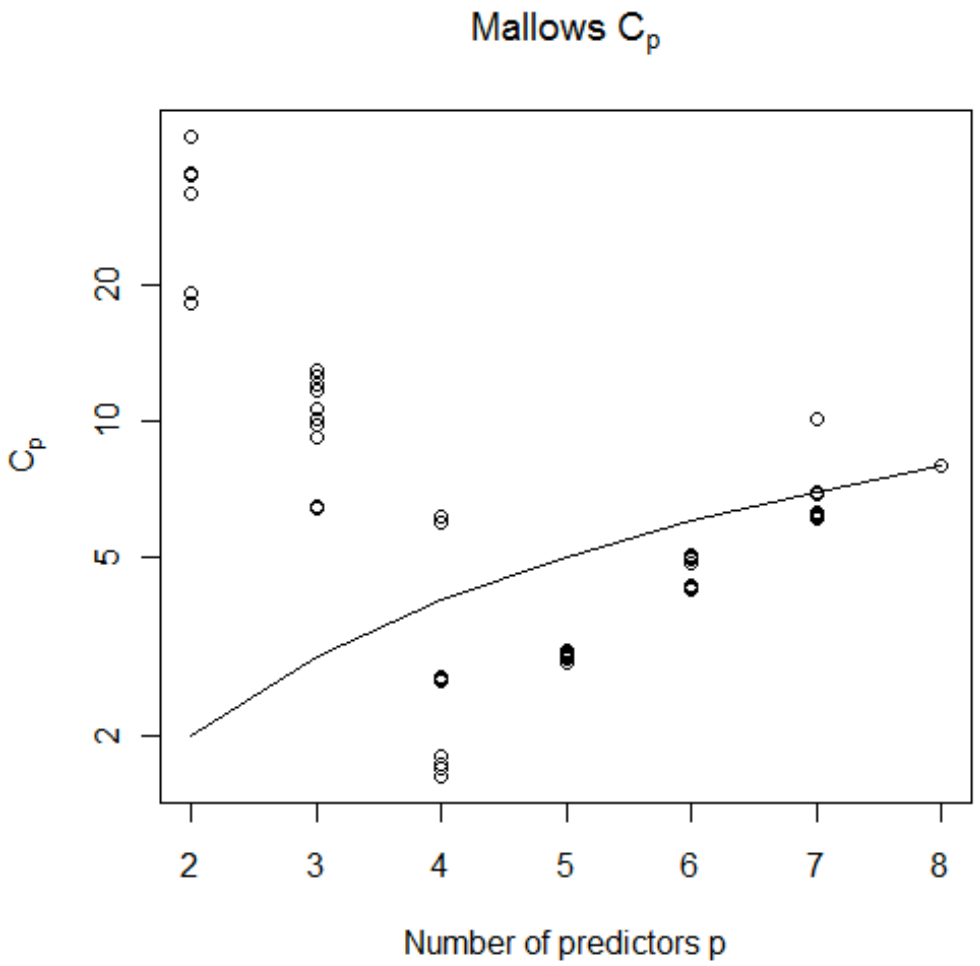
Example 2a: Ant species richness

`leaps` saves the top 10 models for each value of *p*.

The line in the figure indicates
$C_p = p$
(vertical axis is in log units)

The best model has
4 predictors (3 variables plus intercept)

But other models fit the data nearly as well, i.e., all those for which $C_p < p$

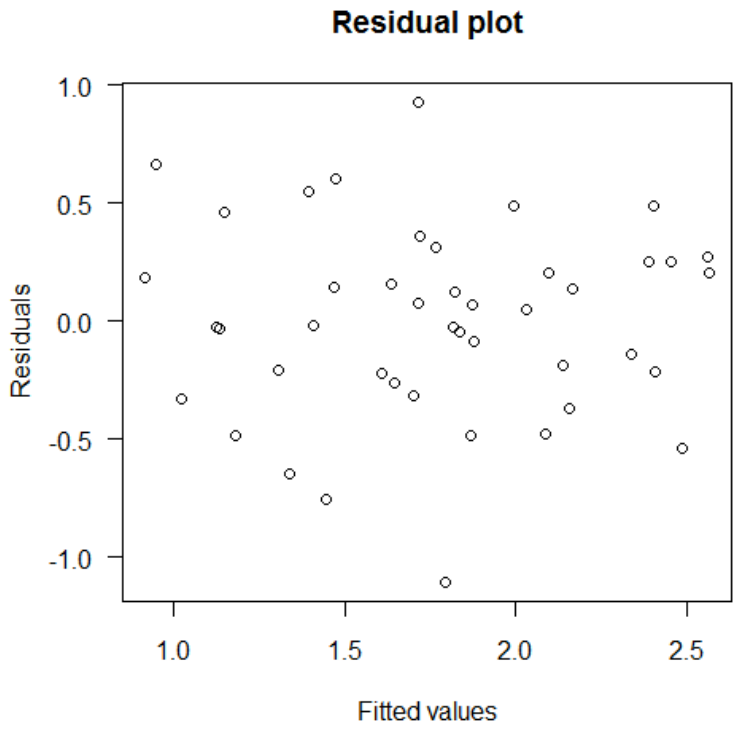


Example 2a: Ant species richness

Best model (smallest C_p):

```
z<- lm( log(nspecies) ~ habitat + latitude + elevation)
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.3180285  2.6101963   3.953 0.000306 ***
habitat      0.6898845  0.1269432   5.435 2.94e-06 ***
latitude    -0.2007838  0.0609920  -3.292 0.002085 **
elevation   -0.0010856  0.0004049  -2.681 0.010610 *
```



Example 2a: Conclusions

If regression is purely for prediction, all of the models with $C_p < p$ predict about equally well. In which case there's no reason to get carried away about exactly which model is the "best".

Interpretation is more complex if regression is used for explanation. If numerous models are nearly equally good at fitting the data, it is difficult to claim to have found the predictors that "explain" the response.

But since, like correlation, "regression is not causation", it is not possible to find the true causes of variation in the explanatory variable without experimentation anyway.

AIC (Akaike's Information Criterion)

Criterion: minimize AIC.

$$\text{AIC} = -2 \ln L(\text{fitted model} \mid \text{data}) + 2k$$

k is the number of parameters estimated in the model (including intercept and σ^2)

First part of AIC is the log-likelihood of the model given the data.

Second part is 2*k*, which acts like a penalty for the number of variables in the model (this is an interpretation, not its raison d'etre).

Just as with the log-likelihood, what matters is not AIC itself but the <u>difference</u> in AIC between models.

Criterion: minimize AIC.

$$\mathrm{AIC} = -2 \ln L(\text{fitted model} \mid \text{data}) + 2k$$

AIC is an estimate of the expected distance ("information lost") between the fitted model and the "true" model.

There are two reasons why a model fitted to data might depart from the truth.

1. The model is biased (contains too few parameters, underestimates the complexity of reality)
2. There is not enough data to yield good estimates of many parameters.

AIC yields a balance between these two sources of information loss.

AIC (Akaike's Information Criterion)

Search strategy: One method is the stepwise procedure for selection of variables implemented in the `stepAIC` command in the MASS library.

Can use for categorical and numerical variables.

`stepAIC` obeys "marginality restrictions". Not all terms are on equal footing. For example

- x^2 is not fitted unless x is also present in the model
- the interaction a:b is not fitted unless both a and b are also present
- a:b:c not fitted unless all two-way interactions of a, b, c, are present

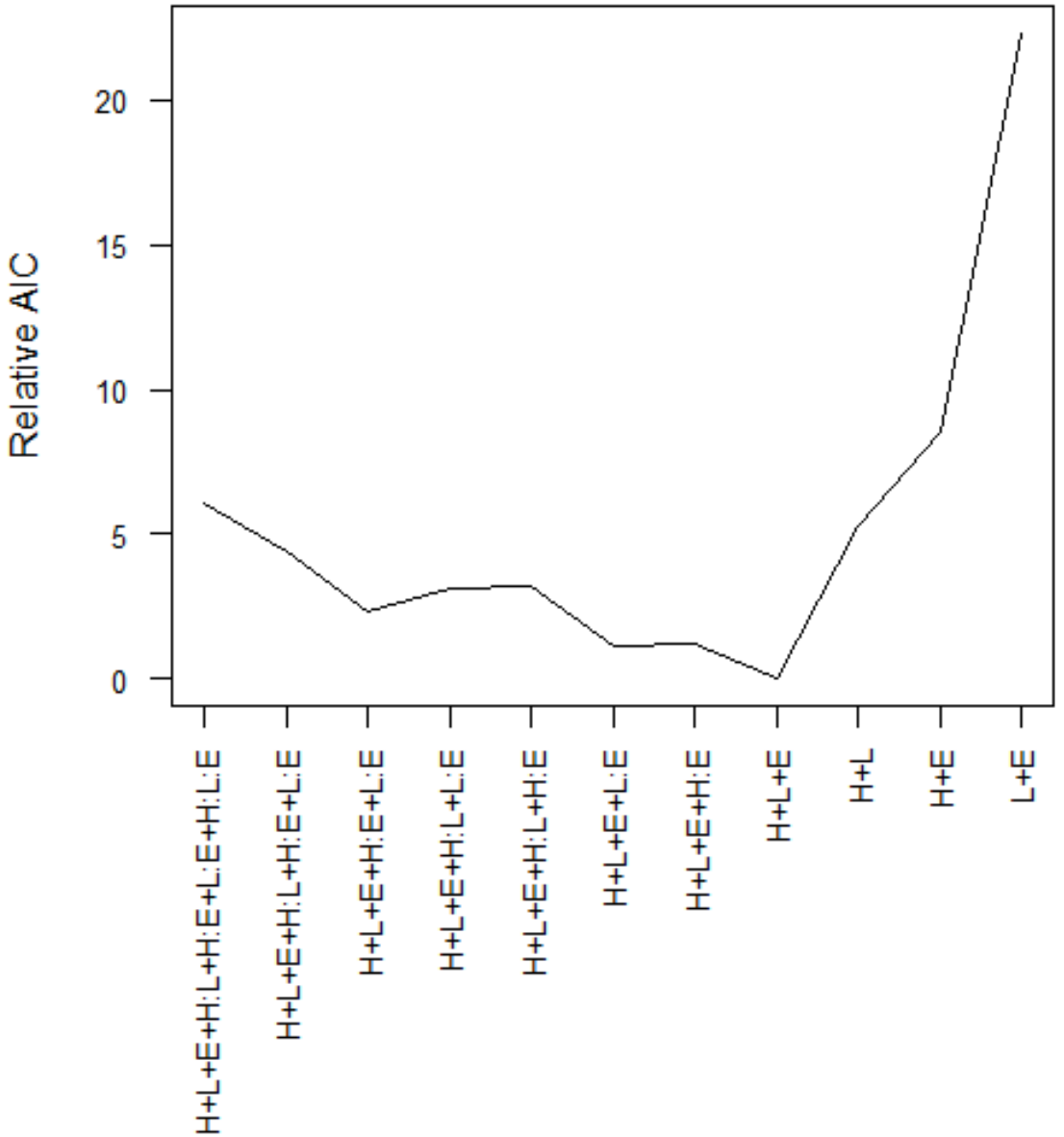
The search algorithm is therefore intelligent and economical

Example 2b: Ant species richness

Same data as that analyzed earlier with `leaps`.

AIC difference (Δ) is plotted (relative to the minimum AIC)

"Best" model is again the model with the three additive terms Habitat, Latitude, and Elevation

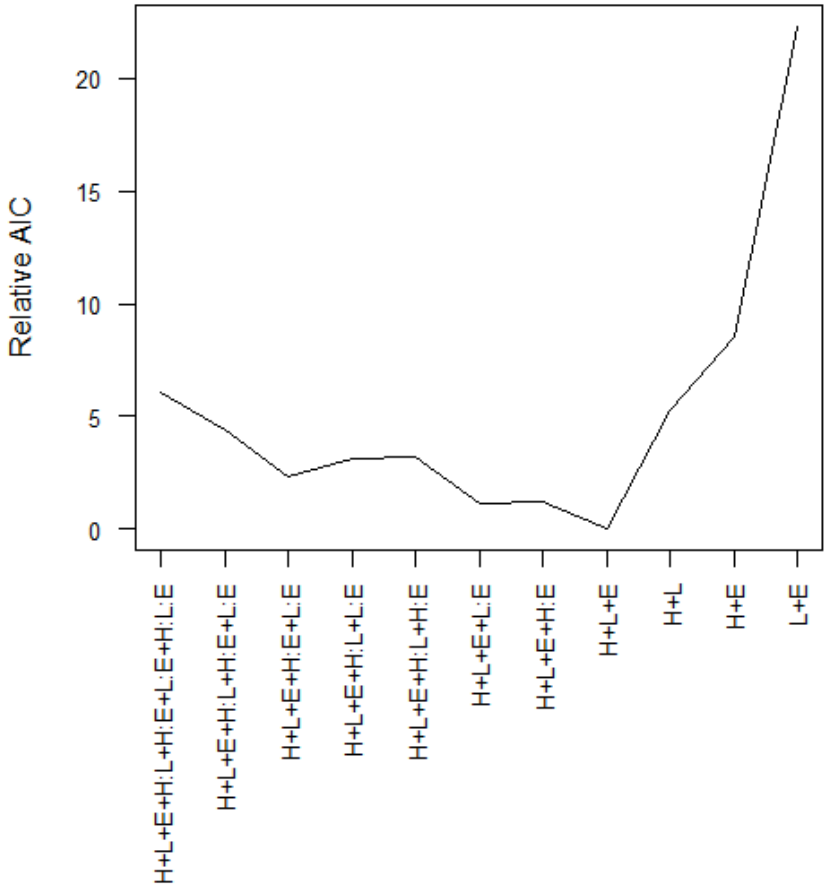


Differs from classical statistical approaches

AIC ranks the models from best to worst. Several models may be about equally good. Differences between them are not tested statistically. No null model. No *P*-value. No model is formally “rejected”.

AIC difference (Δ) is the difference between AIC for a model and that for the “best” model

<u>AIC difference (Δ) support</u>
0 – 2 Substantial support
4 – 7 Considerably less
> 10 Essentially none



Model uncertainty

AIC difference (Δ) support
0 – 2 Substantial support
4 – 7 Considerably less
> 10 Essentially none

The reason for model uncertainty is sampling error. Keep in mind that the data being used to select the "best" model is sampled from a population, and would be different if we returned to that same population for another sample.

This is why we retain multiple models in our set having reasonable levels of support. Think of it as a "confidence set" of models, analogous to a confidence interval for a parameter estimate. The null-hypothesis-significant-testing approach provides no equivalent procedure.

Going further: Multimodel Inference

Avoids the need to base inference solely conditional upon the single "best" model.

Multimodel Inference allows inferences to be made about a parameter based on a set of models that are ranked and weighted according to level of support from the data.

"Model averaging" is an example: a model-average estimate takes a weighted estimate of the parameter estimates from each model deemed to have sufficient support.

The best source for further information is

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. 2nd. New York, Springer

Last example: Selecting among candidate models

The information-theoretic approach shows it true advantage when comparing alternative conceptual or mathematical models to data

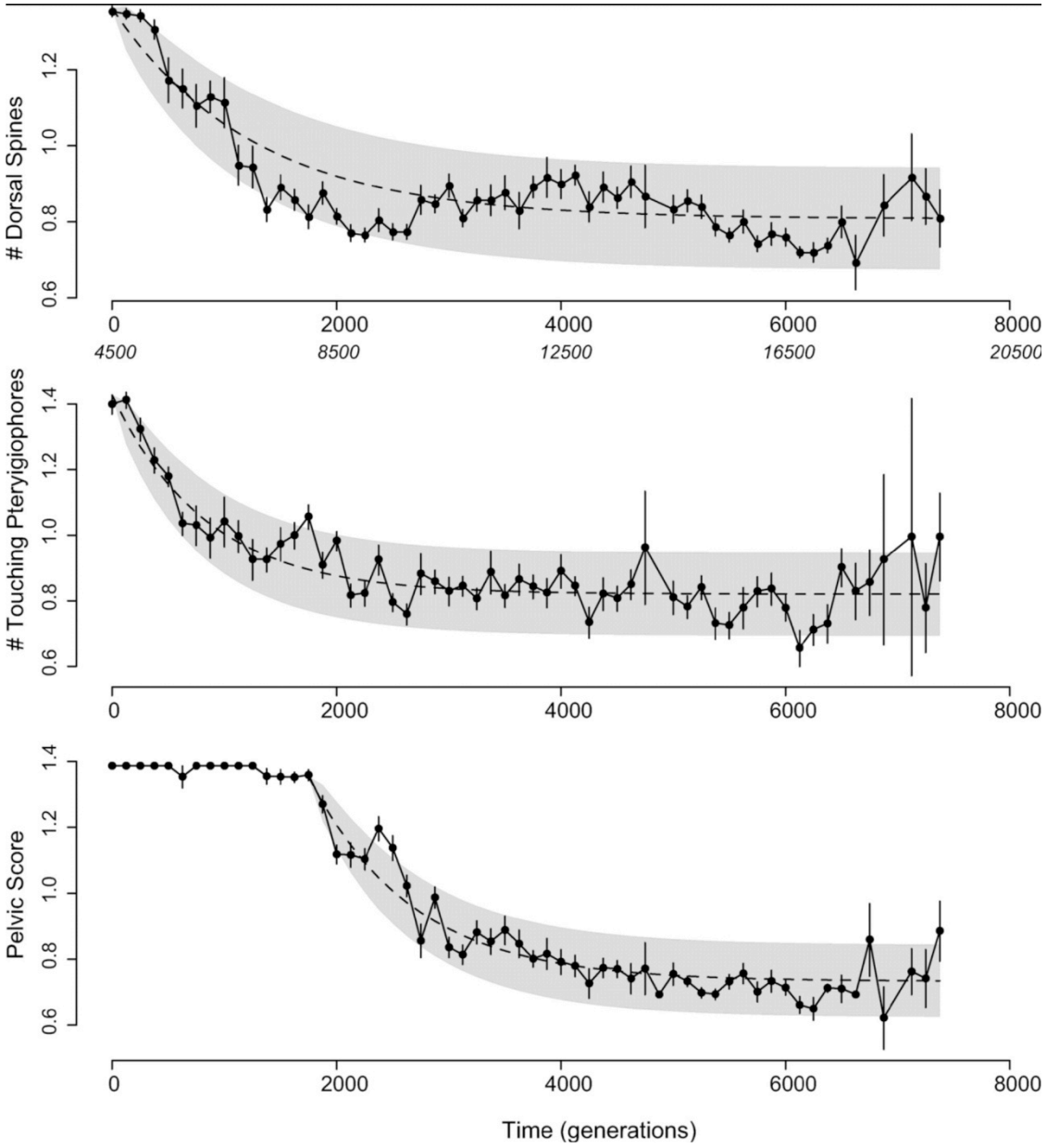
This is where data dredging ends and science begins:
Formulating a set of candidate models

No model is considered the “null” model. Rather, all models are evaluated on the same footing.

Example 3: Adaptive evolution in the fossil record

Data: Armor measurements of 5000 fossil *Gasterosteus doryssus* (threespine stickleback) from an open pit diatomite mine in Nevada. Time=0 corresponds to the first appearance of a highly-armored form in the fossil record.

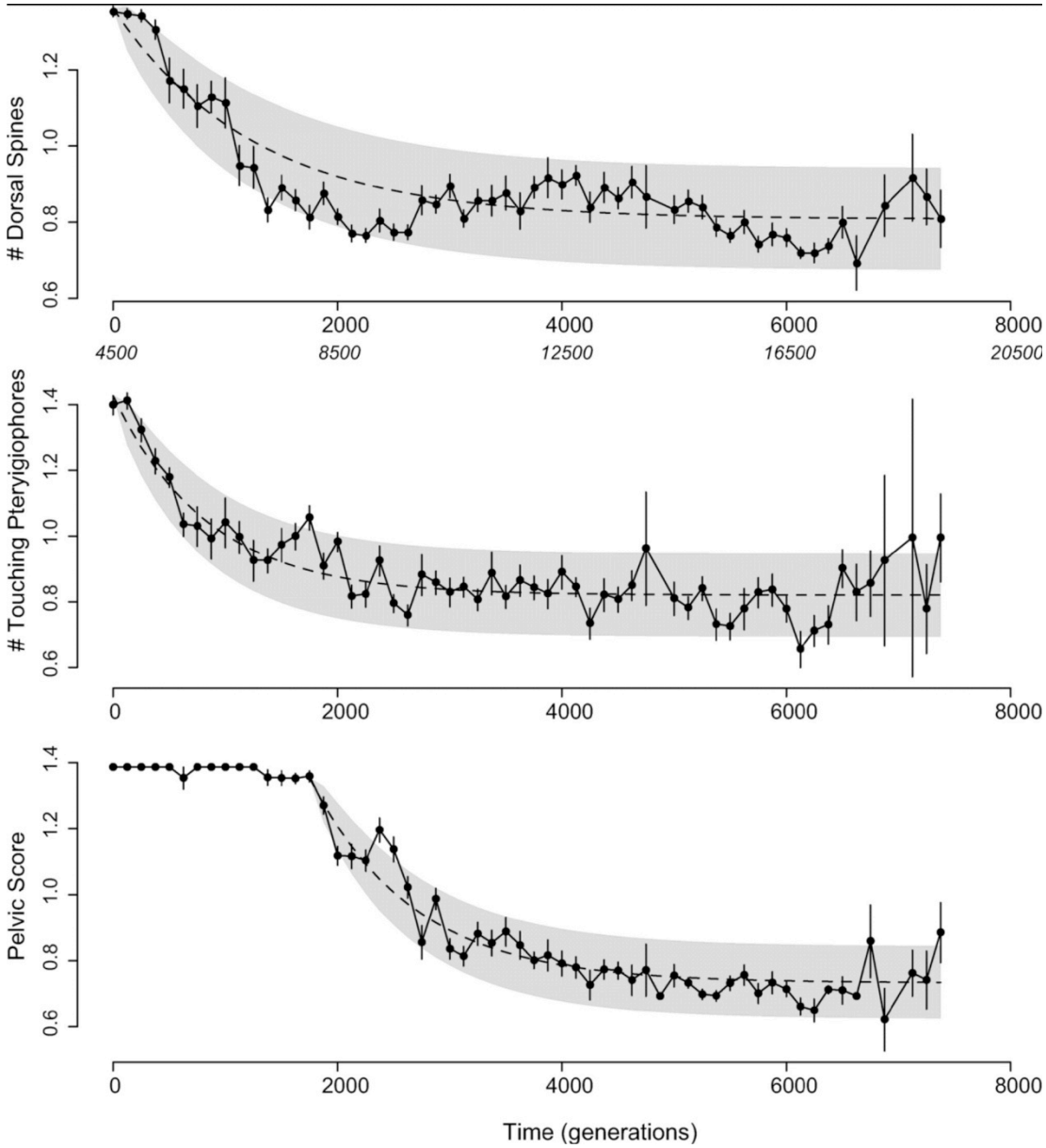
G. Hunt, M. A. Bell & M. P Travis 2008, *Evolution* 62: 700–710.



Example 3: Adaptive evolution in the fossil record

A previous analysis was not able to reject a null hypothesis of random drift in the trait means.

1 generation = 2 years



Example 3: Adaptive evolution in the fossil record

Hunt et al used AIC to compare the fits of two evolutionary models fitted to the data.

1. Neutral random walk (like Brownian motion)
Two parameters need to be estimated from the data: 1) initial trait mean; 2) variance of the random step size each generation.

2. Adaptive peak shift (Orstein–Uhlenbeck process)
Four parameters to be estimated: 1) initial trait mean; 2) variance of the random step size each generation; 3) phenotypic position of the optimum; 4) strength of the "pull" toward the optimum.

Example 3: Adaptive evolution in the fossil record

Results: AIC difference (Δ) of neutral model is large (no support)

Trait	Model	logL	K	AIC_C	Akaike weight	LRT
No. of dorsal spines	Neutral	86.48	2	−168.73	0.002	
	Adaptive	94.94	4	−181.11	**0.998**	16.92, $P = 0.0003$
Pterygiophores	Neutral	65.91	2	−127.59	0.001	
	Adaptive	74.80	4	−140.84	**0.999**	17.78, $P = 0.0002$
Pelvic score	Neutral	58.38	2	−112.46	0.001	
	Adaptive	68.33	4	−127.65	**0.999**	19.89, $P = 0.00005$

The adaptive model beats neutral drift for all three traits.

Akaike weight is the weight of evidence in favor of a model being the actual best model for the situation at hand, assuming that one of the models in the set really is the best. A 95% confidence set of models is obtained by ranking the models and summing the weights until that sum is ≥ 0.95.

Example 3: Adaptive evolution in the fossil record

Trait	Model	logL	K	AIC_C	Akaike weight	LRT
No. of dorsal spines	Neutral	86.48	2	−168.73	0.002	
	Adaptive	94.94	4	−181.11	**0.998**	16.92, $P = 0.0003$
Pterygiophores	Neutral	65.91	2	−127.59	0.001	
	Adaptive	74.80	4	−140.84	**0.999**	17.78, $P = 0.0002$
Pelvic score	Neutral	58.38	2	−112.46	0.001	
	Adaptive	68.33	4	−127.65	**0.999**	19.89, $P = 0.00005$

Backsliding from the model selection approach, the authors showed that the adaptive model rejects neutrality in a likelihood ratio test (here the models are *not* on equal footing – one of them, the simpler, is set as the null hypothesis).

This suggests that specifying the alternative model more precisely is already more effective than conventional null-hypothesis testing, where the alternative hypothesis is merely “everything but the null hypothesis”

Conclusion: Model Selection

There are better options for model selection that than stepwise elimination using null hypothesis testing.

These approaches work best when thoughtful science is used to specify the candidate models under consideration (rather than evaluating all possible models).

Working with a set of models that fit the data about the same rather than with the one single best model will take some getting used to.

If you want more certainty about which variables are the ones that really matter, then you will need to do an experiment.

Discussion paper for next week:

Cohen. J. 1994. The earth is round ($p < 0.05$). Am. Psych. 49: 997-1003.

Download from "**assignments**" tab on course web site.

Presenters: Susannah & Ian

Moderators: Kim & Allison