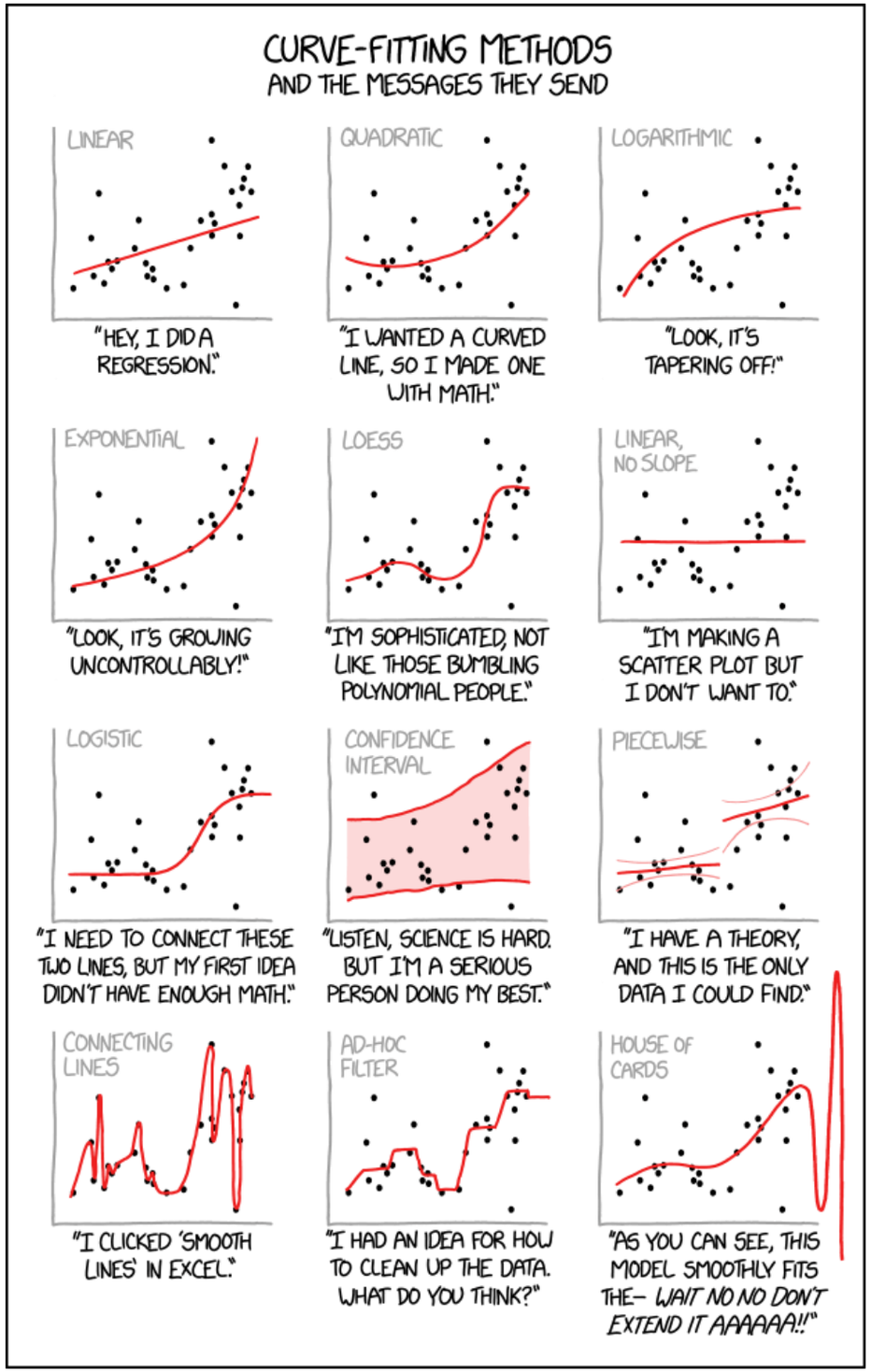
xkcd.com/2048



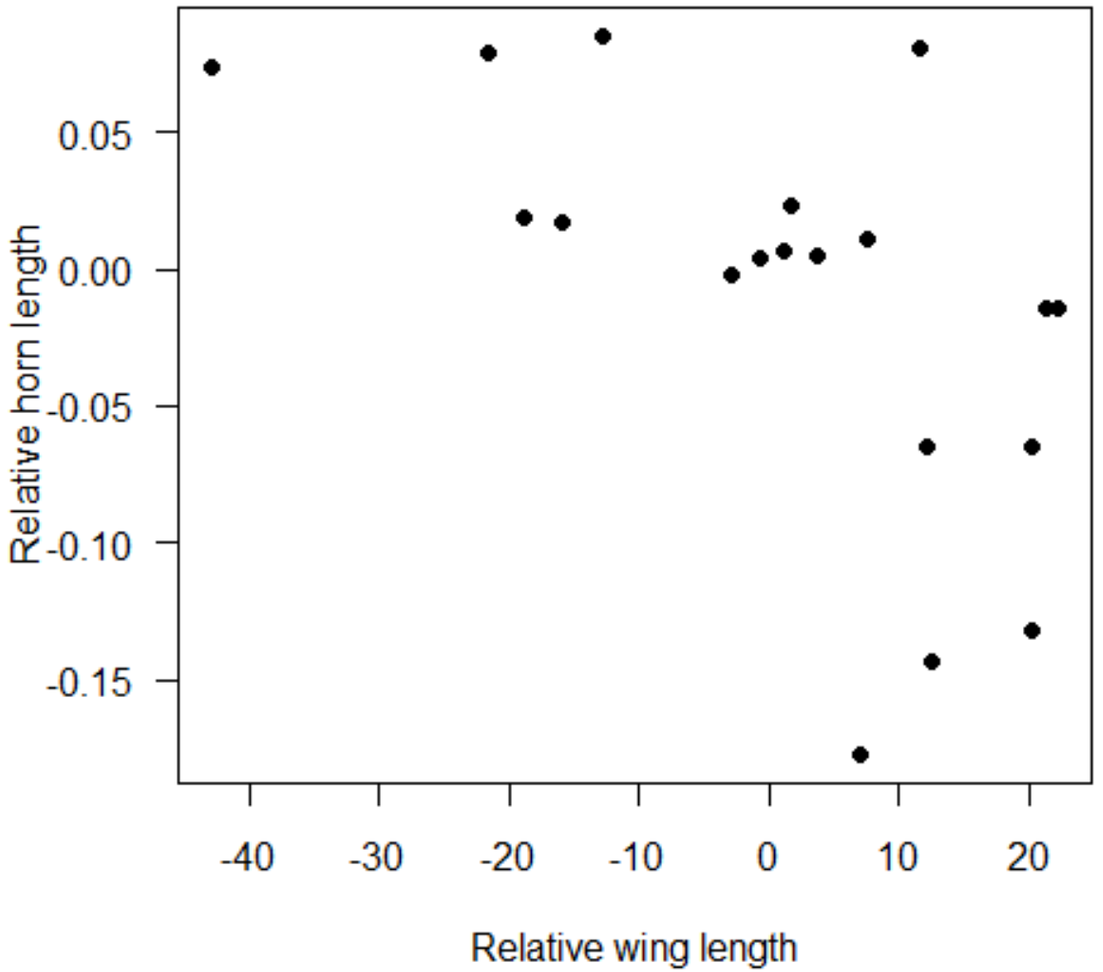
Outline for today

- Example 1: polynomial regression – which degree is best?
- The problem of model selection
- Choose among models using an explicit criterion
- Goals of model selection
- AIC criterion
- Search strategies: dredge(); stepAIC()
- Example 2: Predicting ant species richness
- Several models may fit about equally well
- The science part: formulate a set of candidate models
- Example 3: Adaptive evolution in the fossil record

Example 1: Fit a polynomial regression model – which degree is best?

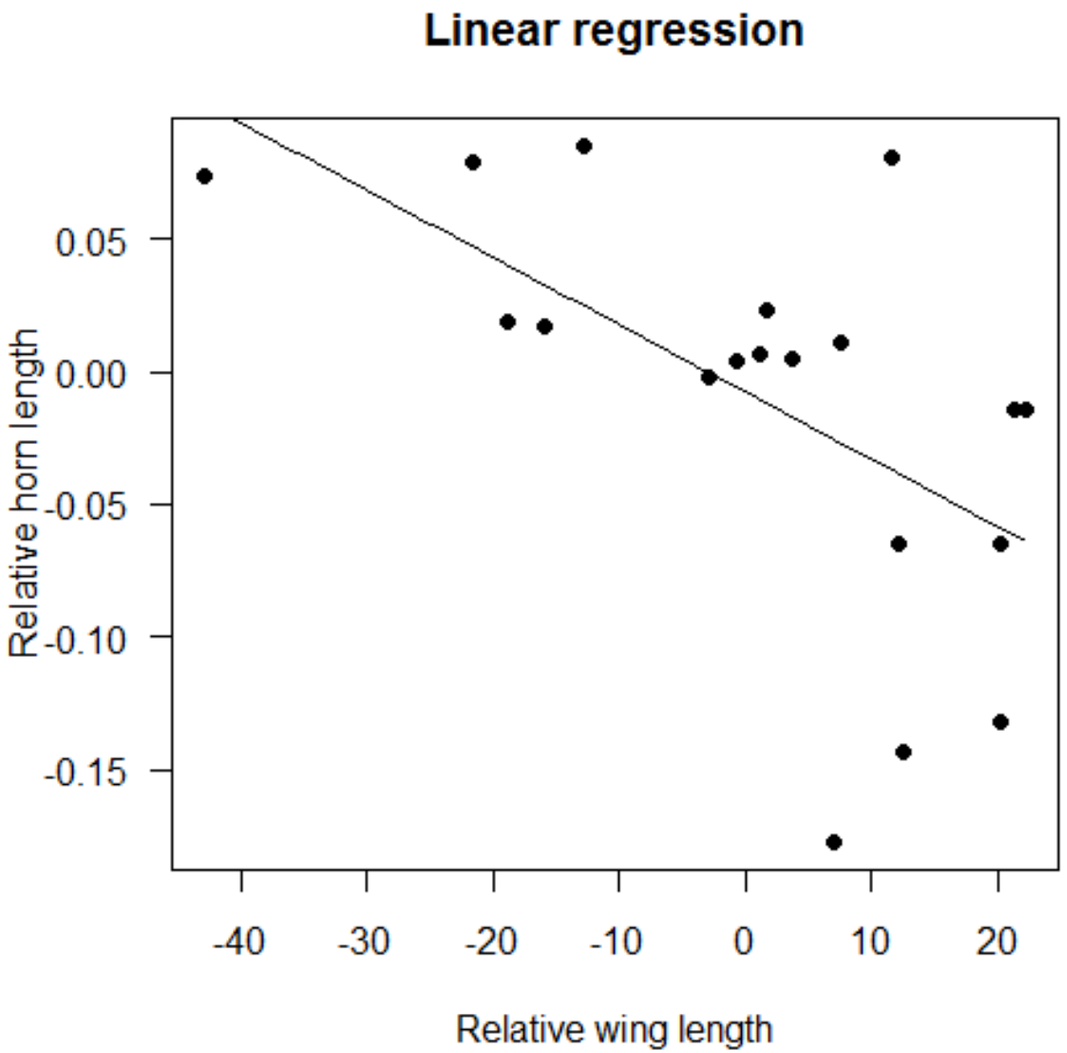
Data: Trade-off between the sizes of wings and horns in 19 females of the beetle *Onthophagus sagittarius*. Both variables are size corrected.

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



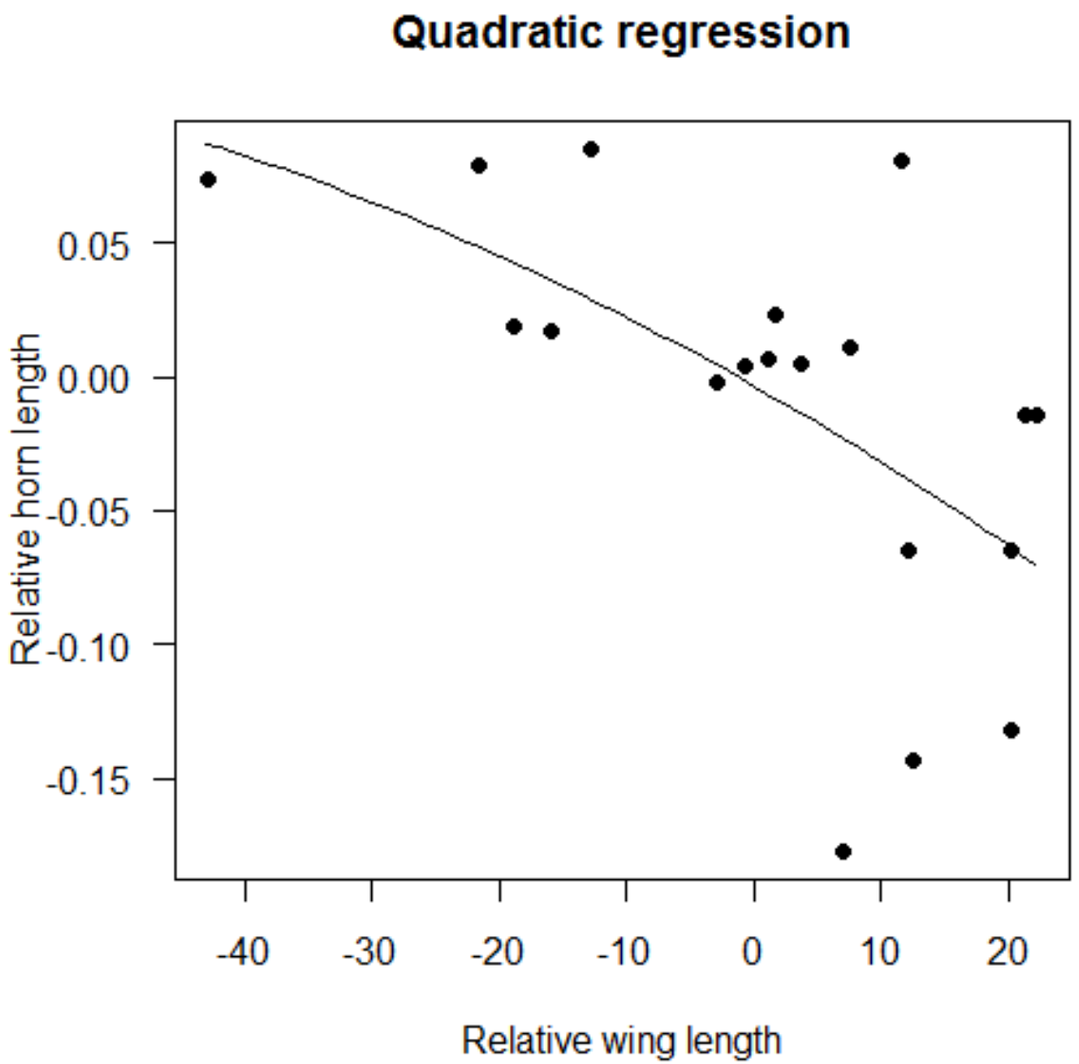
Example 1: Fit a polynomial regression model – which degree is best?

Start with a linear regression



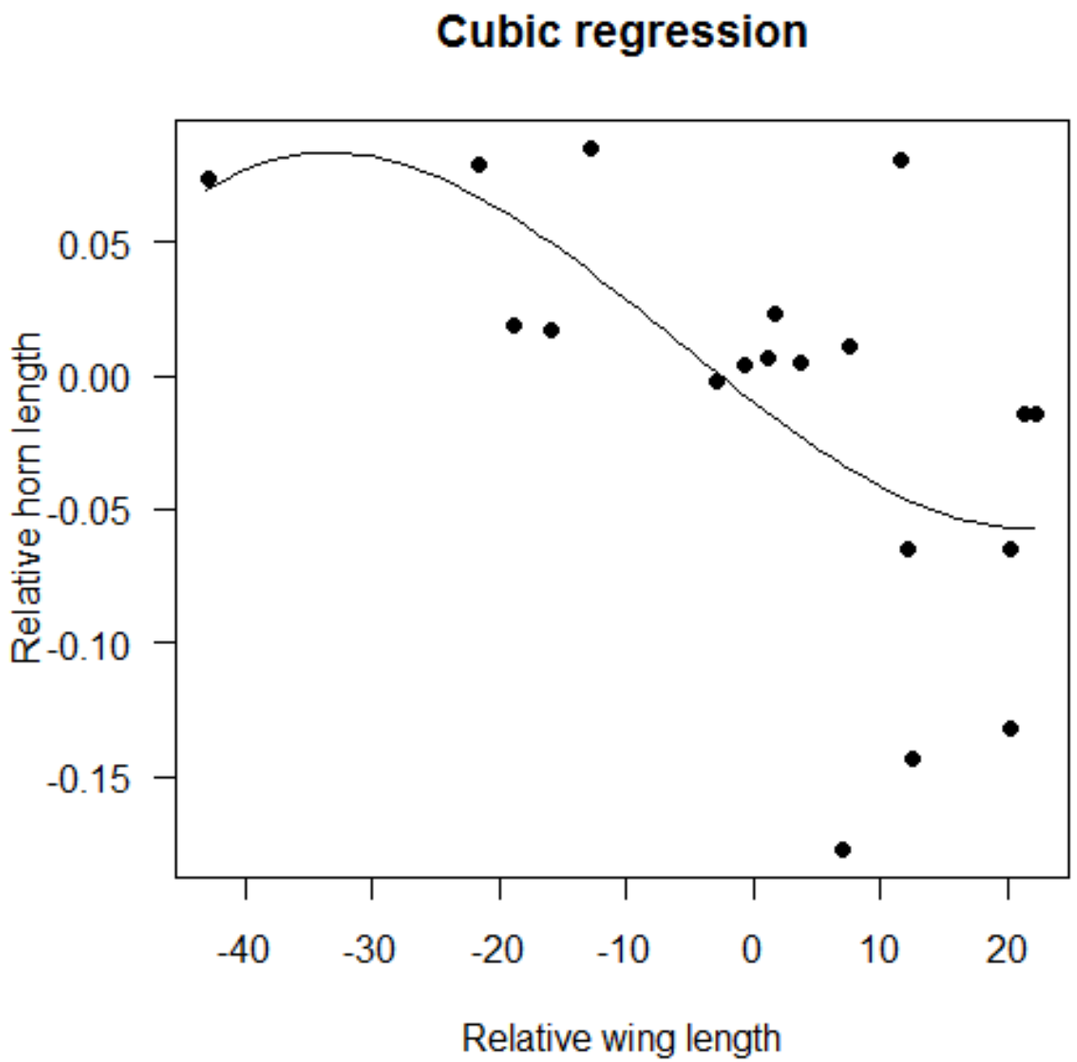
Example 1: Fit a polynomial regression model – which degree is best?

Why not a quadratic regression instead (polynomial degree 2)



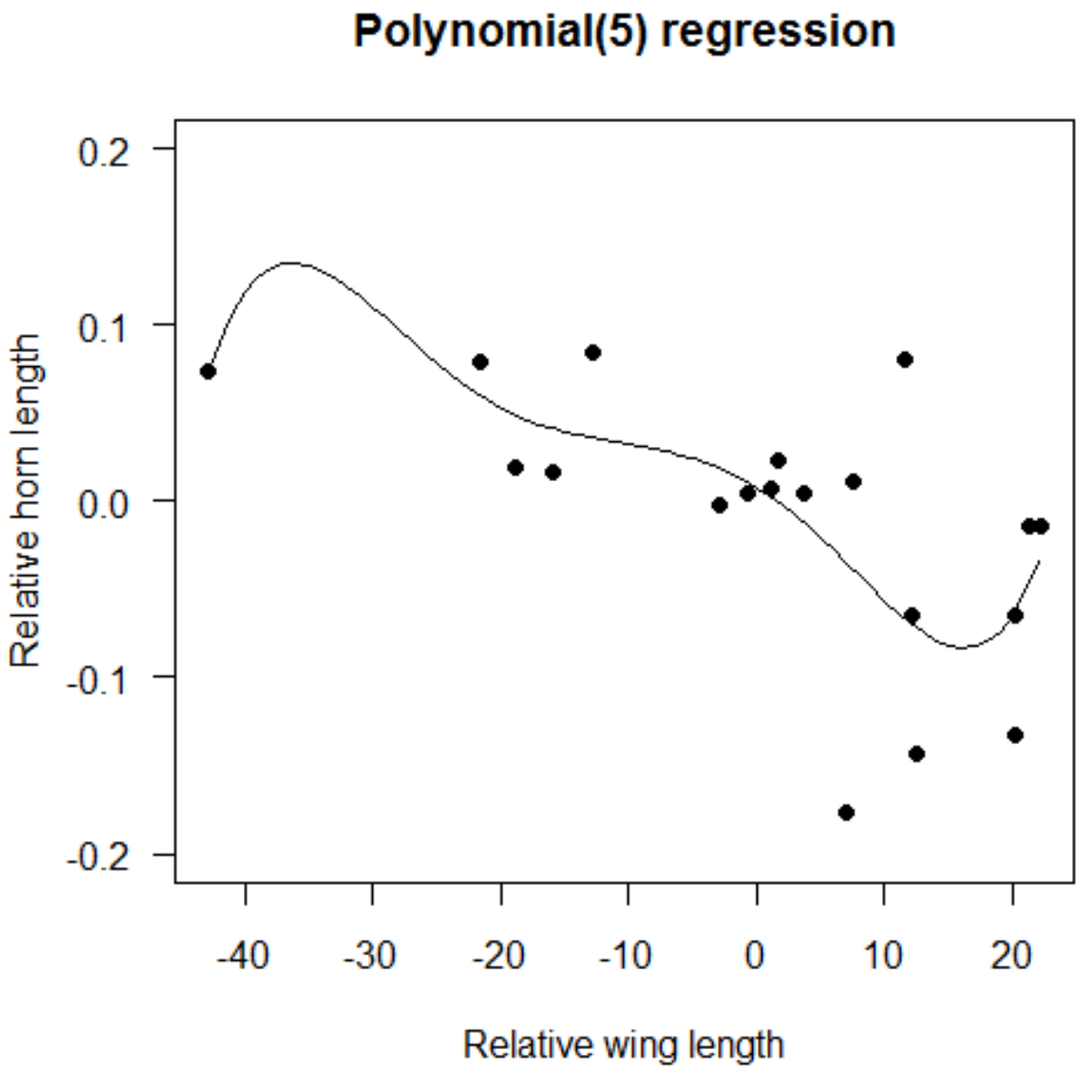
Example 1: Fit a polynomial regression model – which degree is best?

How about a cubic polynomial regression (degree 3)



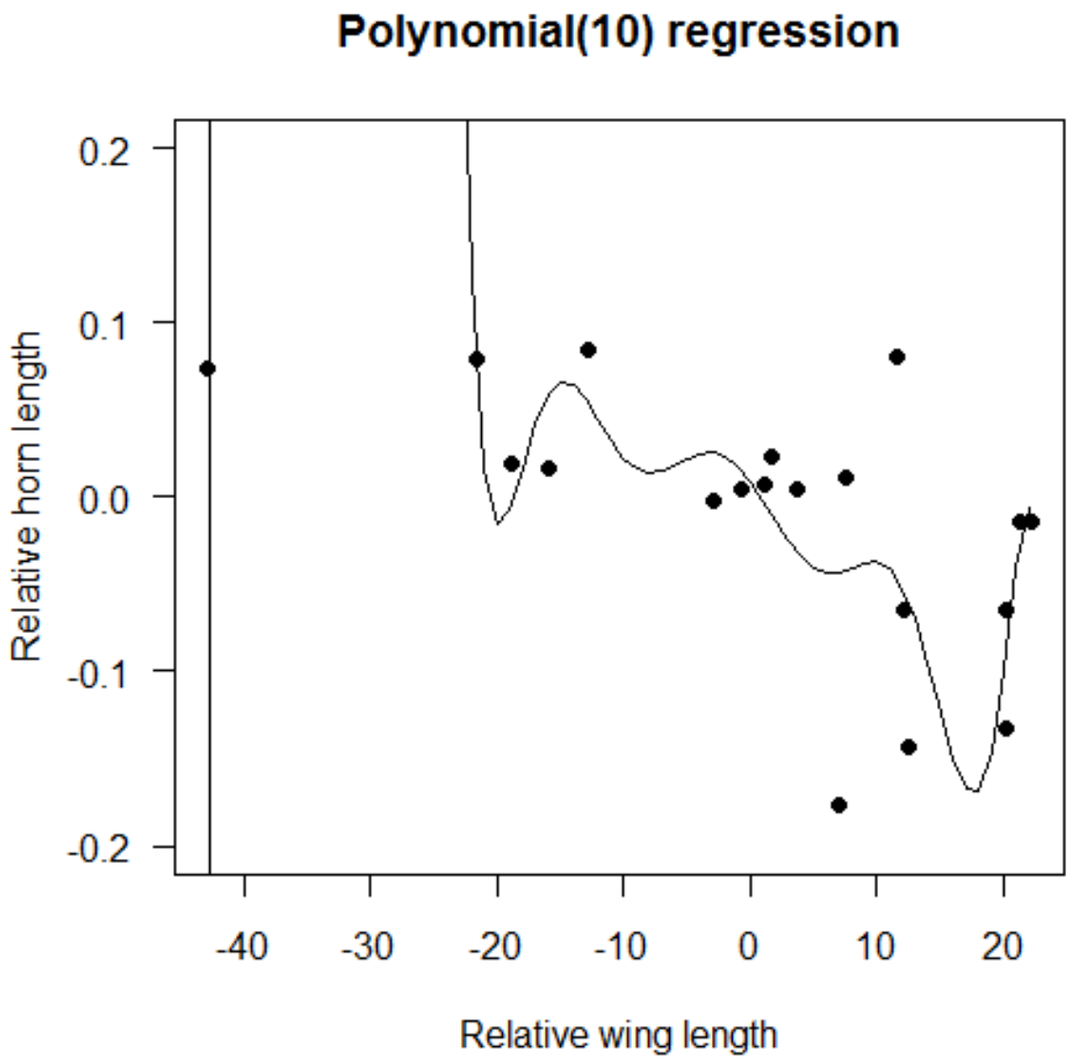
Example 1: Fit a polynomial regression model – which degree is best?

Better still, a polynomial degree 5



Example 1: Fit a polynomial regression model – which degree is best?

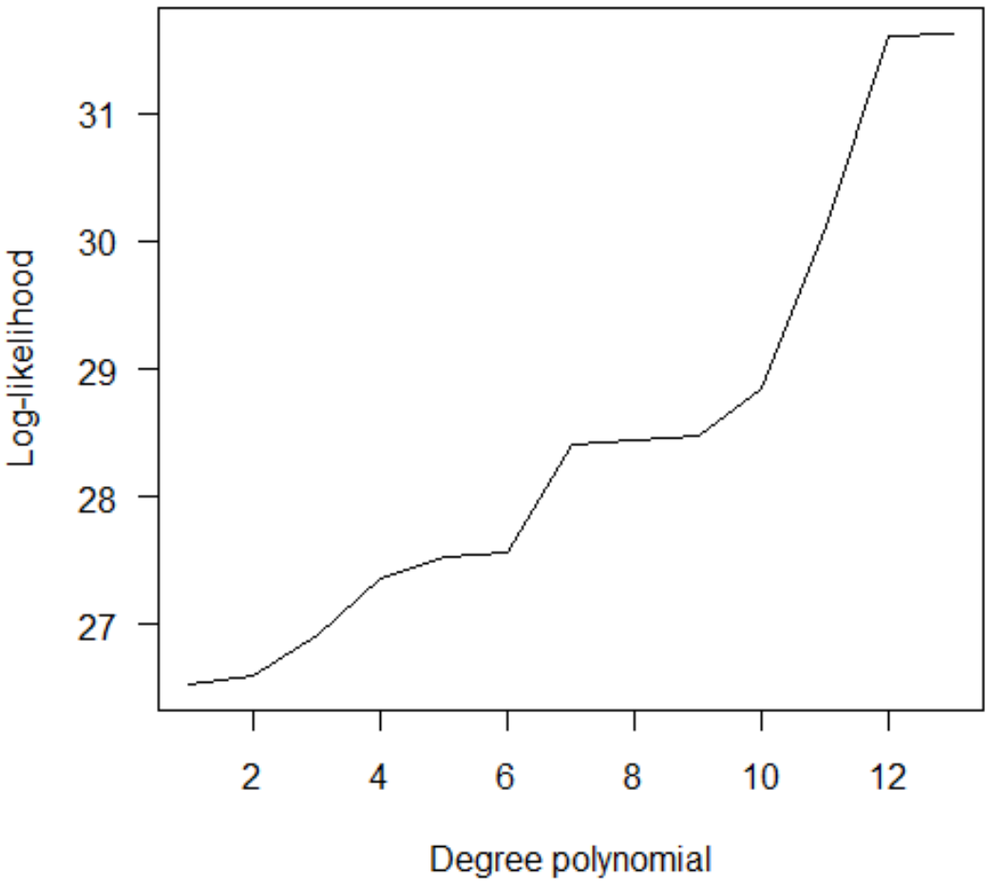
A polynomial, degree 10

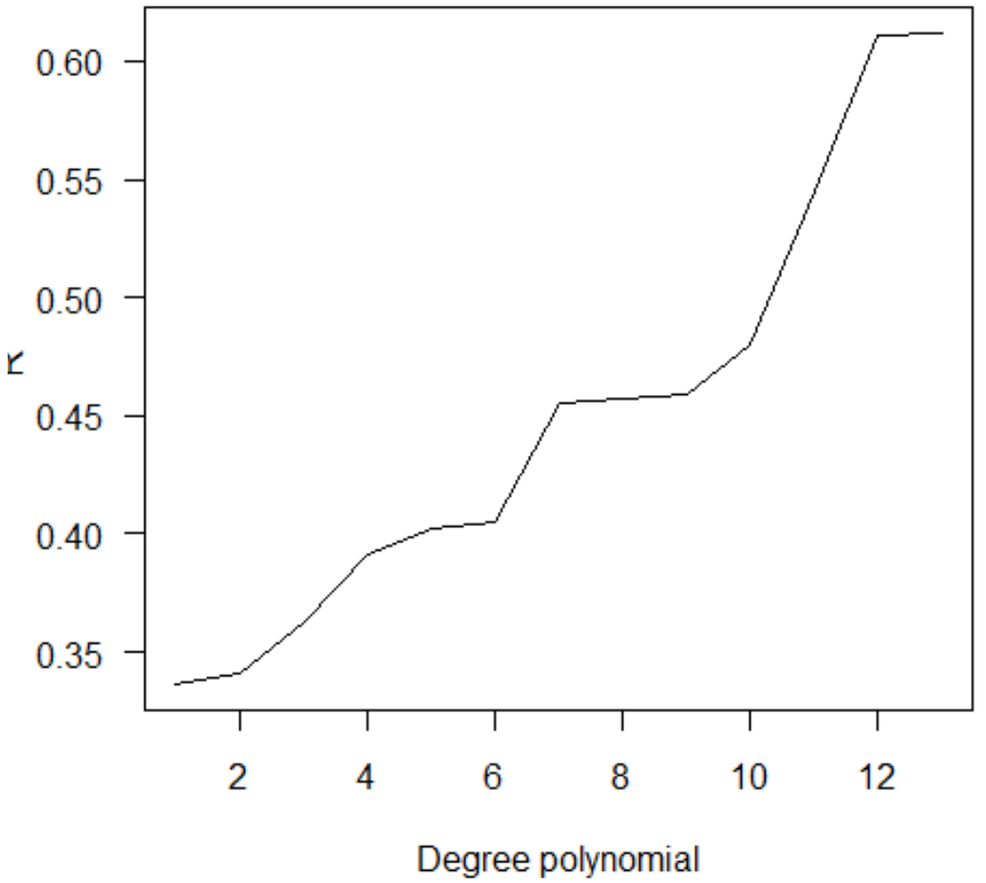


The problem of model selection

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

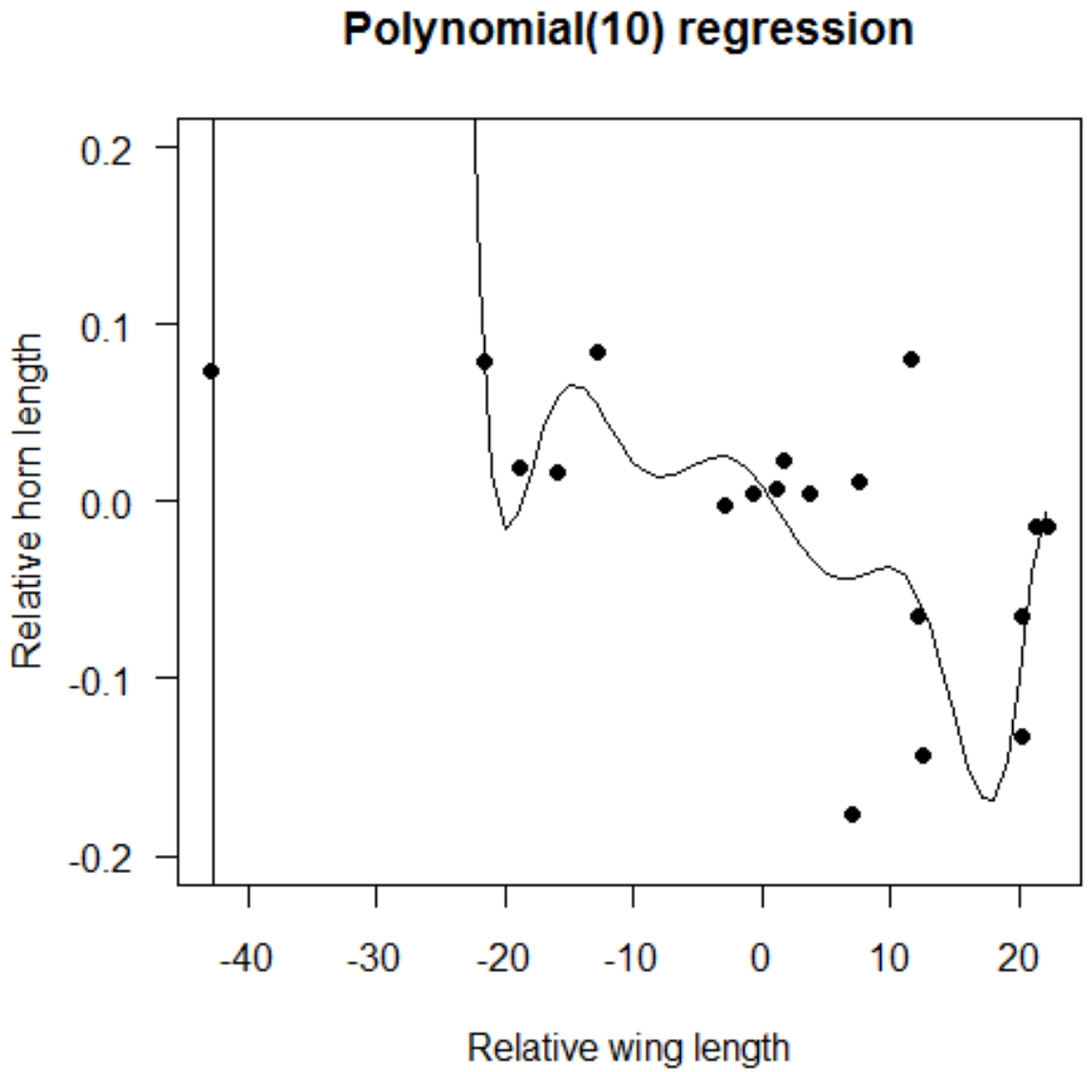
Isn't this good? Isn't this what we want – the best fit possible to data?





The problem of model selection

What is wrong with this picture?



The problem of model selection

Does it violate some principle?

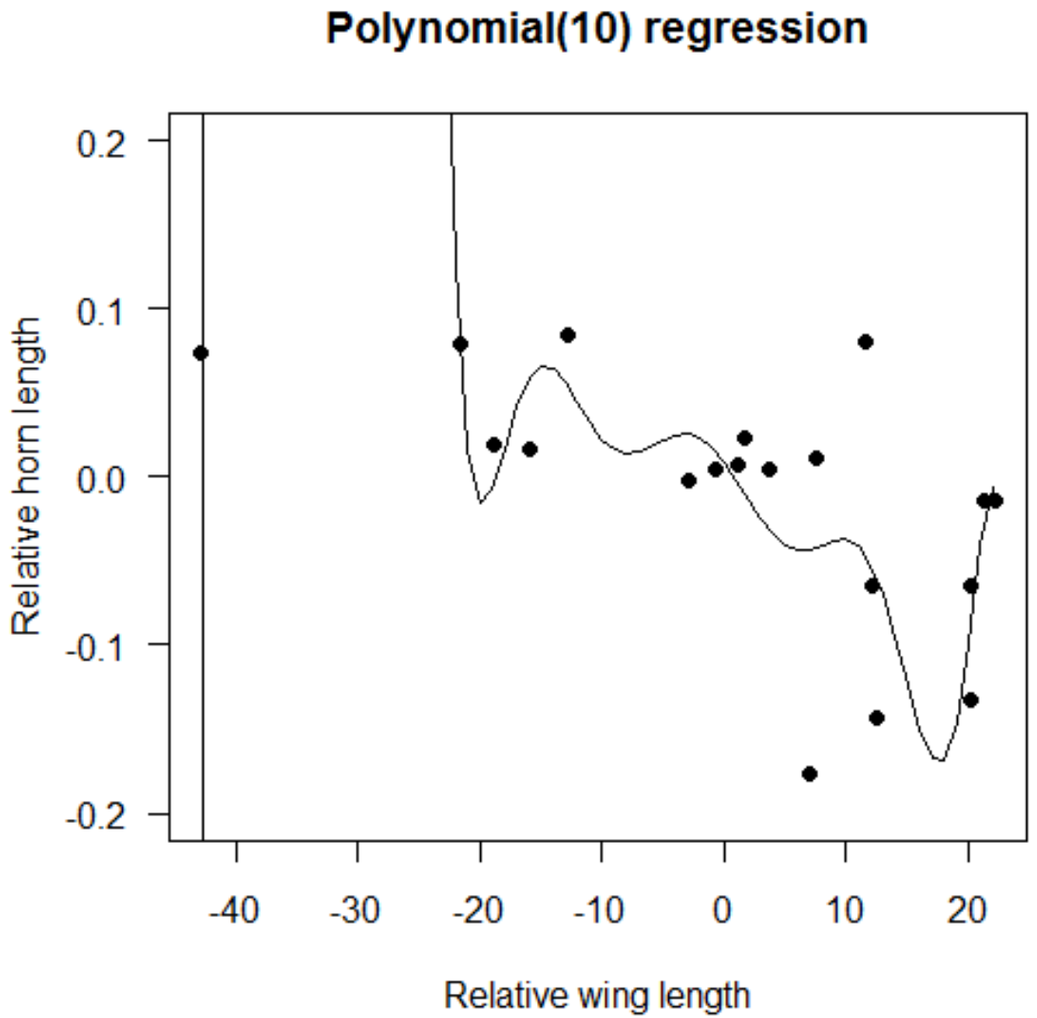
Parsimony principle: Fit no more parameters than is necessary. If two or more models fit the data almost equally well, prefer the simpler model.

But, why?

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

-- Crawley 2007, p325

But how is "minimal adequate" decided? What criterion is used?



The problem of model selection

Stepwise multiple regression, using stepwise elimination of terms, is a common practice.

This approach involves fitting a multiple regression with many variables, followed by a cycle of deleting model terms that are not statistically significant and then refitting. Continue until only statistically significant terms remain.

The procedure ends us up with a single, final model, the “minimum adequate model.”

But is this a good idea? Does it yield the best model?

Typically, we are only modeling **observational** data. We are not dealing with data from an experiment, where we can make intelligent choices about which model to fit based on the experimental design.

Does stepwise elimination of terms actually yield the "best" model?

1. What criterion are we actually using to decide which model is "best"?

2. Each step in which a variable is dropped from the model involves "accepting" a null hypothesis. What happens if we drop a false null hypothesis? How can a sequence of Type 2 errors lead us to the "best" model?

3. How repeatable is the outcome of stepwise regression? With a different sample, would stepwise elimination bring us to the same model again?

4. Might models with different subsets of variables fit the data nearly as well as the current "best"?

Instead: choose among models using an explicit criterion

One reasonable criterion: choose the model that best predicts a new observation.

"Cross-validation score" is one way to measure prediction error:

$$\text{CVscore} = \sum e_{(i)}^2$$

Where the prediction error $e_{(i)}^2 = \left(y_i - \hat{y}_{(i)}\right)^2$

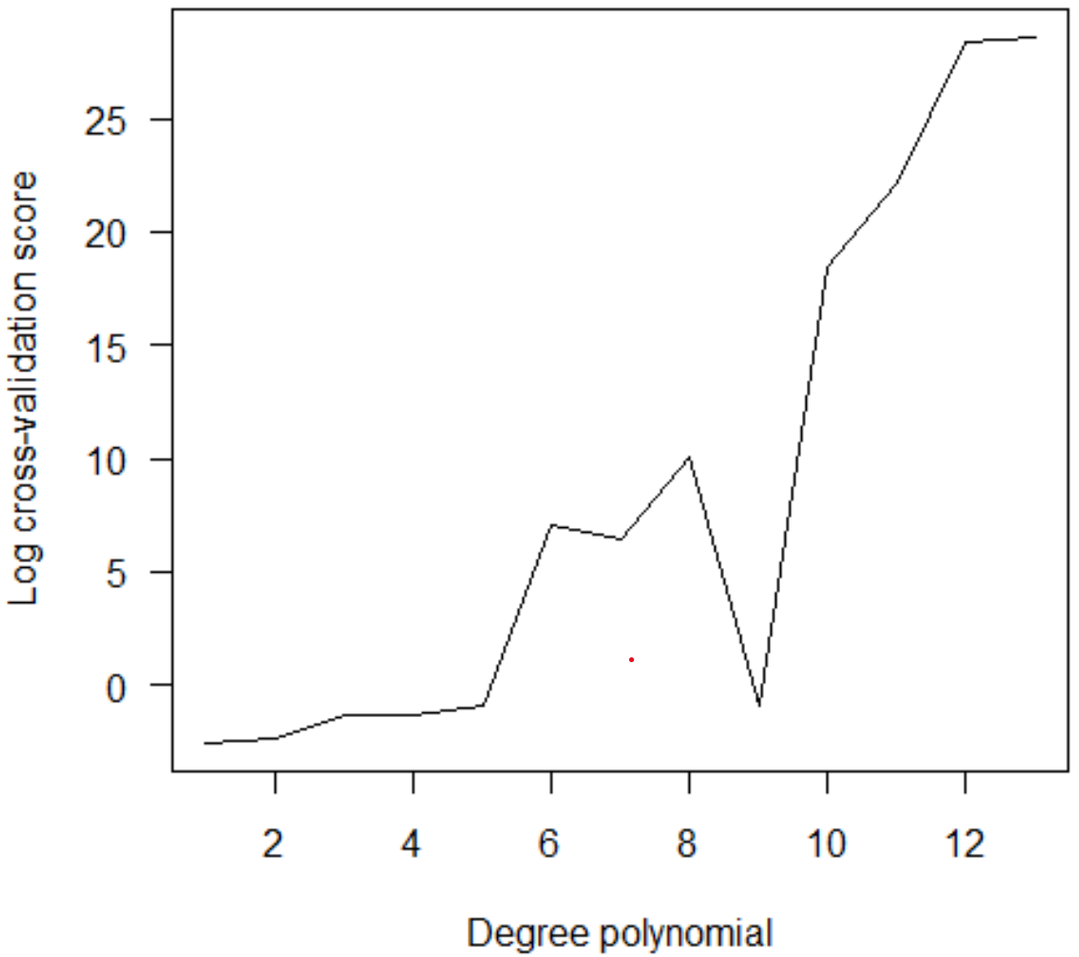
y_i is a measurement of the response variable.

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

A larger *CVscore* corresponds to a worse prediction (more prediction error).

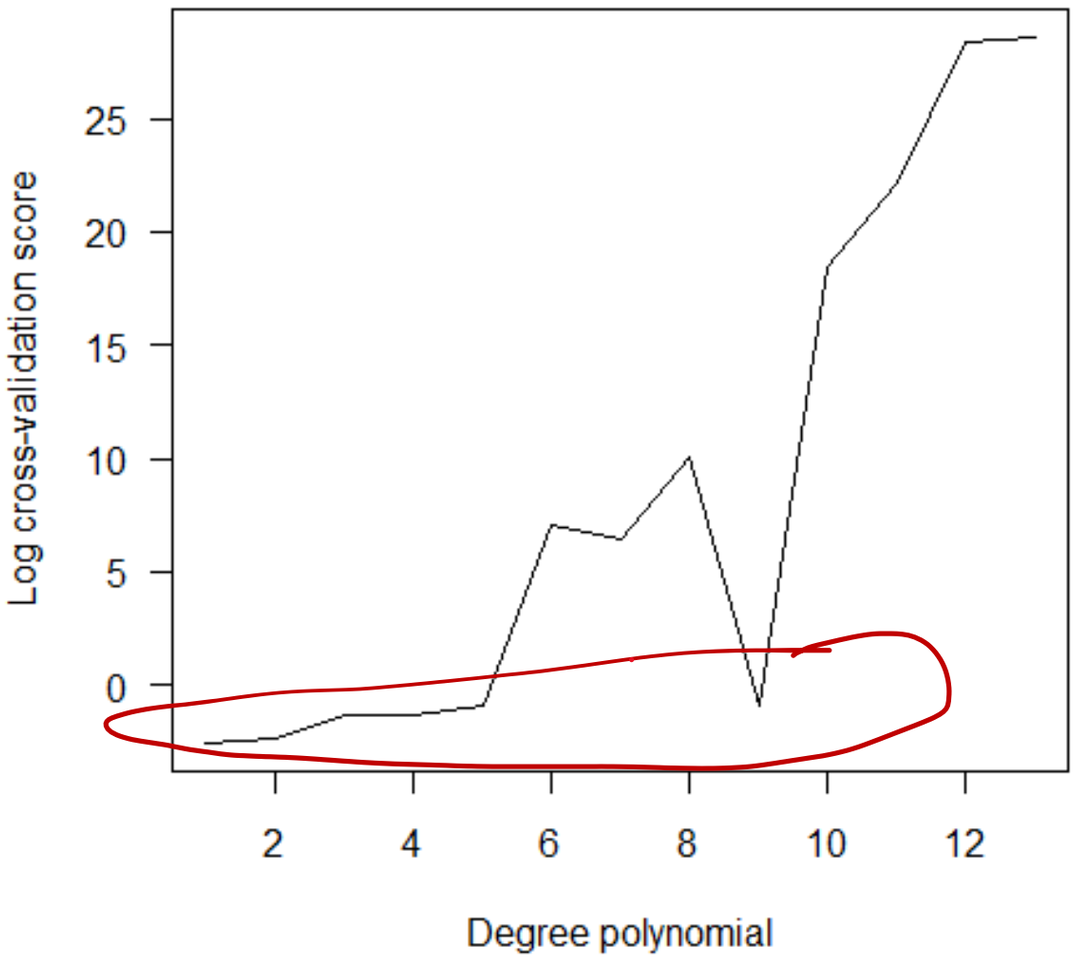
Choose among models using an explicit criterion

In our beetle example, the *CVscore* increases (prediction error worsens) with increasing numbers of parameters in the model. Here, the simple linear regression was “best”.



Choose among models using an explicit criterion

But some *other polynomials do nearly equally well*!



What determines prediction errors?

Prediction errors result from both bias and sampling variance (sampling error) in model parameter estimates. The effects of bias and sampling variance are inversely related (the *bias-variance tradeoff*).

The coefficients of the simplest model are likely to be biased – the true relationship is likely to be more complex. But the coefficients of a simple model have relatively low sampling error (low sample variance) compared to a more complex model. They are likely to be close to their parameter values.

The coefficients of most complex models have lower bias (their long-run means over repeated sampling are close to their true values), but the coefficients of complex models have high sampling error (high sample variance). In any sample they are likely to be far from their parameter values.

Prediction error is typically minimized somewhere in-between.

What determines prediction errors?

The simplest models have low sampling variance but high bias because of missing terms (the truth is more complex).

The most complex models have low bias but high sampling variance – they poorly estimate too many parameters ("overfitting").

Training error: how well a model fits the data used to fit the model.
Test error: how well a model fits a new sample of data.

Hastie et al. (2009)

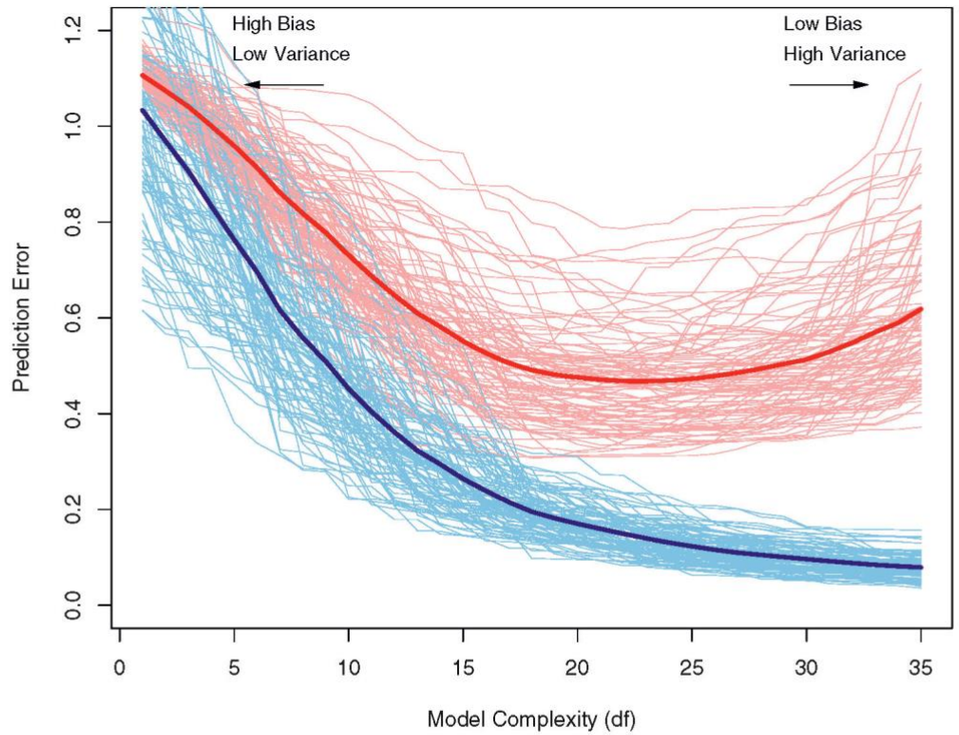


FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\text{err}}$*, while the light red curves show the conditional test error* $\text{Err}_{\mathcal{T}}$ *for* 100 *training sets of size* 50 *each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\text{E}[\overline{\text{err}}]$.

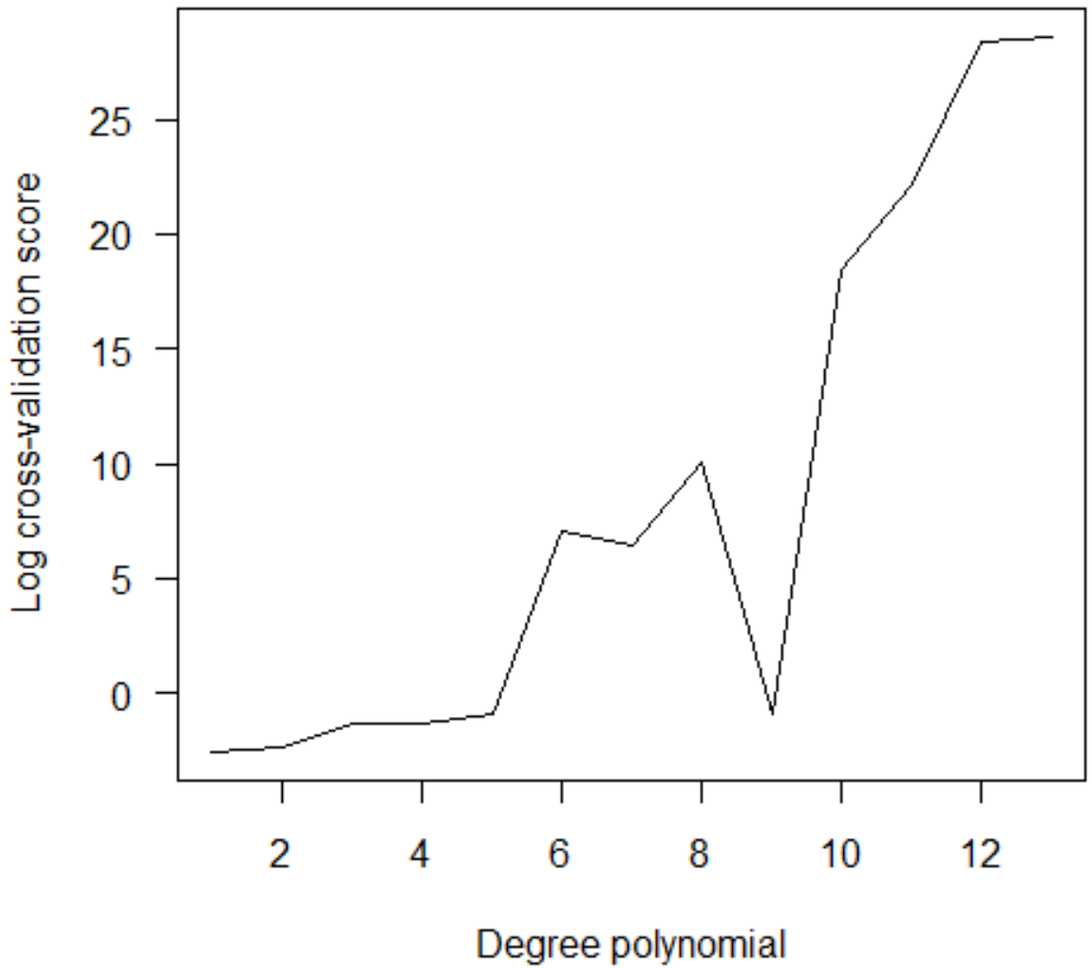
The problem of model selection

Another problem with my polynomial regression analysis is that I'm data dredging.

I didn't have any hypotheses to help guide my search. This can lead to non-reproducible results.

For example my 9th degree polynomial was surprisingly good at prediction. But is there any good, *a priori* reason to include it among the set of candidate models to evaluate?

Where's the science in all this?



Goals of model selection

Some reasonable objectives:

- A model that predicts well (note: not the same as the true model).
- A model that approximates the true relationship between the variables.
- Plus, a set of models that fit the data nearly as well as the “best” model.
- Ability to compare non-nested* models, not just restricted to “full” versus “reduced” models having a subset of terms in the full model.

*Full and reduced are referred to as “nested models”, because the one contains a subset of the terms occurring in the other. Models in which the terms contained in one are not a subset of the terms in the other are called “non-nested” models. (Don’t confuse with nested experimental designs or nested sampling designs.)

Goals of model selection

To accomplish these goals, we need a model selection approach that includes:

- A **criterion** to compare models:
  - *CVscore*
  - AIC (Akaike's Information Criterion)
  - BIC (Bayesian Information Criterion)
- A **strategy** for searching the candidate models

Again, we are typically modeling **observational** data, not data from an experiment where we can make intelligent choices about which model to fit based on the experimental design.

AIC (Akaike's Information Criterion)

Criterion: minimize AIC.

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

k is the number of parameters estimated in the model (including intercept and the variance of the residuals, s^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 – the price paid for including *k* variables in the model (this is an interpretation, not why the 2*k* is part of the formula).

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

AIC (Akaike's Information Criterion)

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

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. Bias: The fitted model may contain too few parameters, underestimating the complexity of reality.
2. Variance: High sampling error (low precision) of model parameter estimates. There are not enough data to yield good estimates of the many parameters of a complex model.

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

Example 2: Ant species richness

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

	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
...					

n = 44 sites

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

Example 2: Ant species richness

`dredge()` in `MuMIn` package in R. Provide model with all desired terms:

```
zfull <- lm(log(nspecies) ~ habitat * latitude * elevation)
zdredge <- dredge(zfull, evaluate = TRUE, rank = "AIC")
```

Model selection table (variable names abbreviated; "+" refers to categorical term)

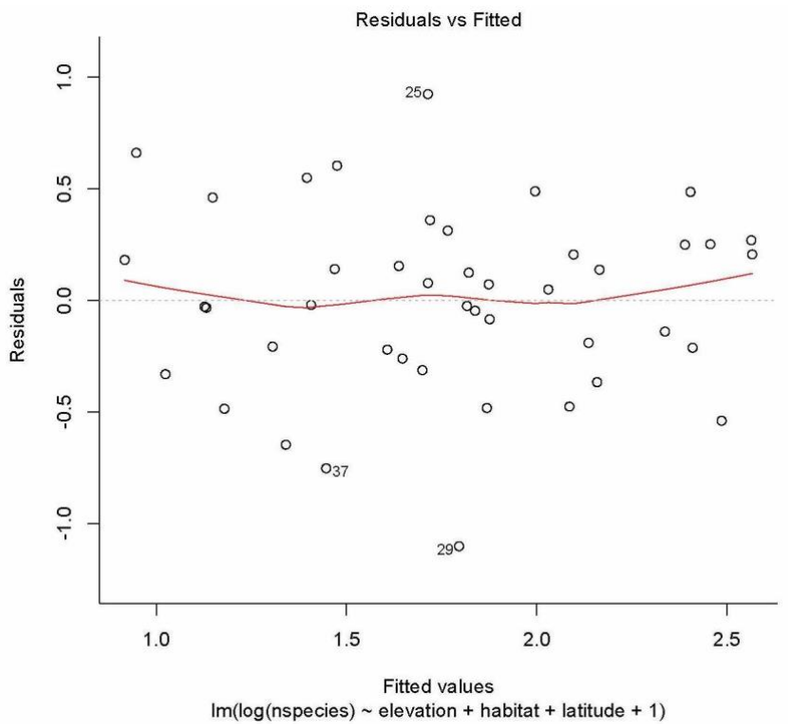
```
# ("df" is k, the number of parameters: all coefficients plus 1 more for σ² of residuals)
 (Int)        elv hbt     ltt elv:hbt   elv:ltt hbt:ltt elv:hbt:ltt df  logLik  AIC delta weight
10.320 -0.0010860   + -0.2008                                       5 -22.273 54.5  0.00  0.288
13.810 -0.0166000   + -0.2826         0.0003621                     6 -21.846 55.7  1.14  0.162
10.240 -0.0007565   + -0.2008       +                               6 -21.895 55.8  1.24  0.155
 9.794 -0.0010860   + -0.1886                       +               6 -22.251 56.5  1.95  0.108
13.730 -0.0162700   + -0.2826       + 0.0003621                     7 -21.460 56.9  2.37  0.088
13.290 -0.0166000   + -0.2704         0.0003621       +             7 -21.823 57.6  3.10  0.061
10.100 -0.0007605   + -0.1974       +                 +             7 -21.893 57.8  3.24  0.057
13.590 -0.0162700   + -0.2792       + 0.0003621       +             8 -21.458 58.9  4.37  0.032
11.320              + -0.2301                                       4 -25.909 59.8  5.27  0.021
15.680 -0.0255800   + -0.3283       + 0.0005794       +           + 9 -21.299 60.6  6.05  0.014
10.800              + -0.2179                         +             5 -25.890 61.8  7.23  0.008
 1.736 -0.0013240   +                                               4 -27.548 63.1  8.55  0.004
 1.659 -0.0009951   +               +                               5 -27.250 64.5  9.95  0.002
 1.428              +                                               3 -31.875 69.7 15.20  0.000
10.660 -0.0010860     -0.2008                                       4 -34.438 76.9 22.33  0.000
...
```

Example 2: Ant species richness

"Best" model (smallest AIC) is the model with the three additive terms Habitat, Latitude, and Elevation.

```
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 2: Ant species richness

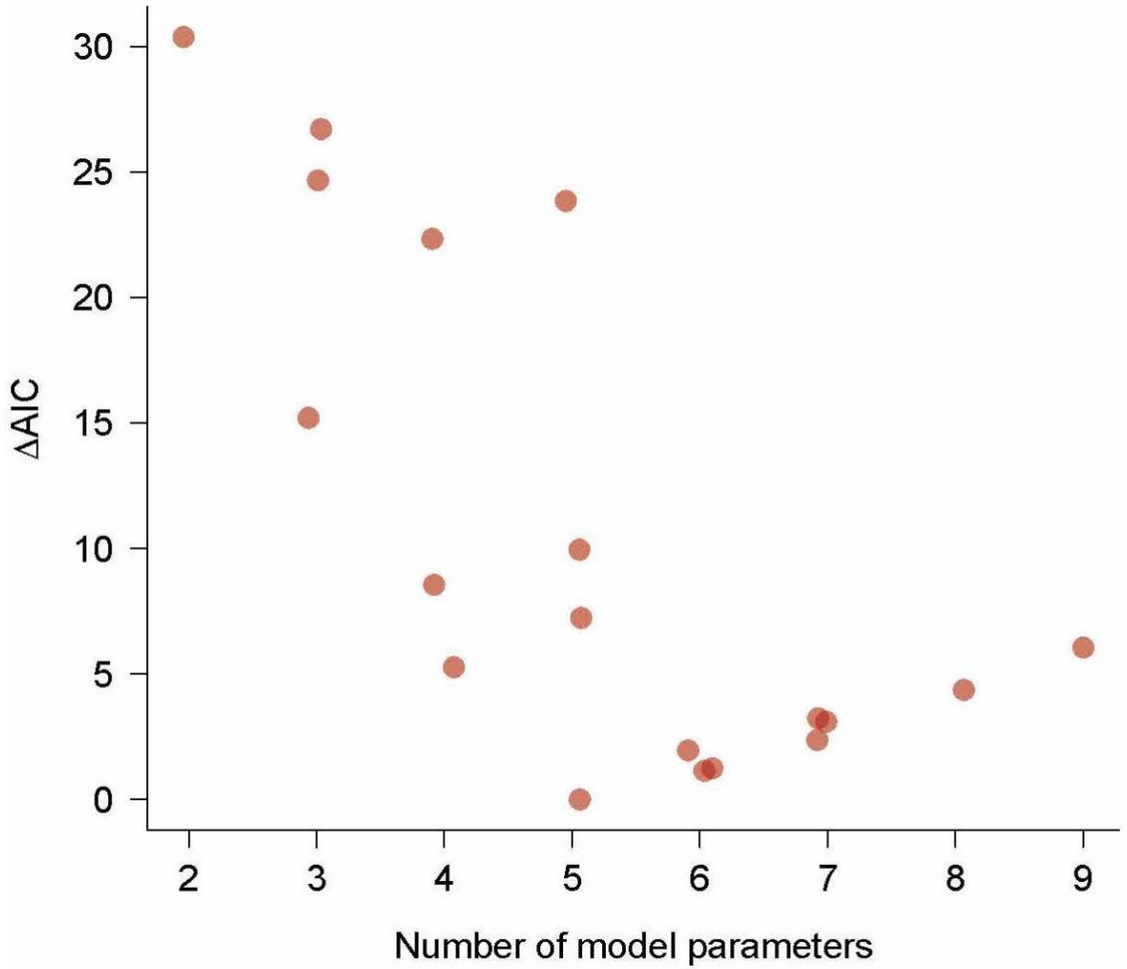
Other models are nearly as good as the best.

Each dot refers to a model.
AIC difference (Δ) is the difference between a model's AIC score and that of the "best" model.

The best model has 5 parameters (includes 1 for variance of residual)
But a few other models fit the data nearly as well.

<u>AIC difference (Δ) support</u>

0 – 2 Substantial support
4 – 7 Considerably less support
> 10 Essentially no support

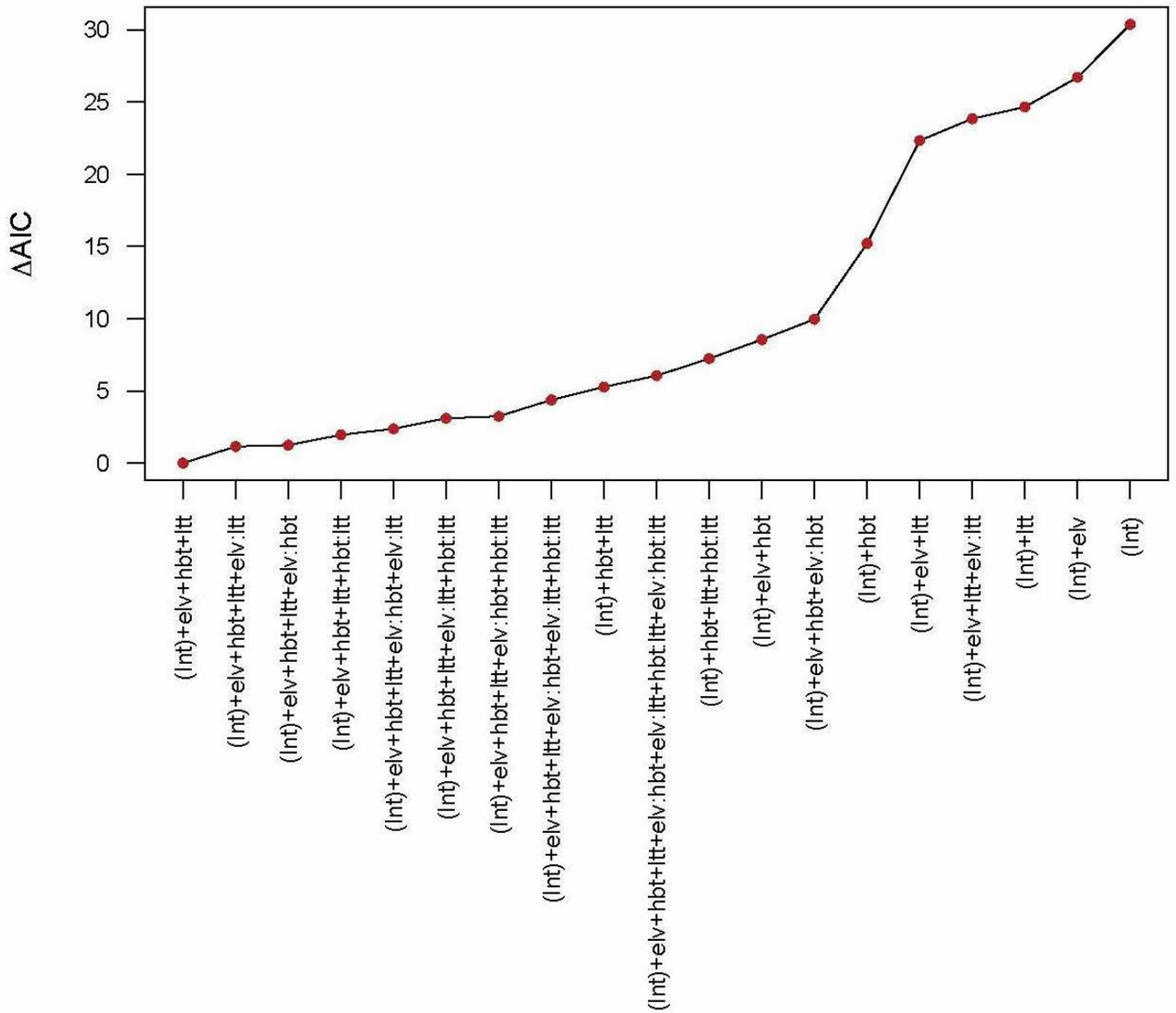


Example 2: Ant species richness

Models ranked by ΔAIC

AIC difference (Δ) support
0 – 2 Substantial support
4 – 7 Considerably less support
> 10 Essentially no support

Think of a cutoff based on AIC score can be used to generate a "95% confidence set of models", analogous to a 95% confidence interval for a parameter.



Example 2: Ant species richness

Another way to form a “95% confidence set of models”, analogous to a 95% confidence interval for a parameter, is to use cumulative model weights.

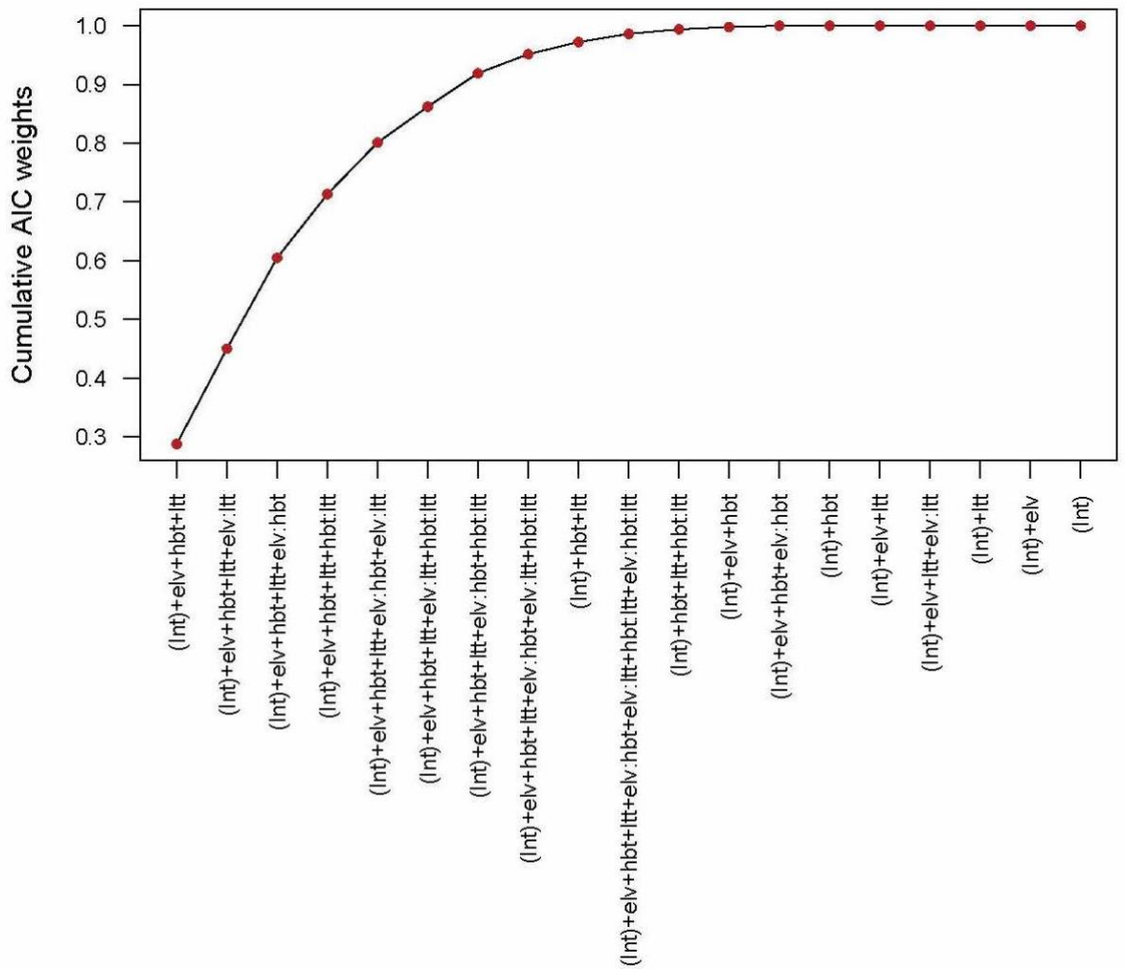
AIC weights measure support that a given model is the “best” model, assuming that the “best” model is one of the set of models being compared.

```
subset(zdredge, cumsum(zdredge$weight) <= .95)
```

(Int)	elv	hbt	ltt	elv:hbt	elv:ltt	hbt:ltt	df	logLik	AIC	delta	weight
10.320	-0.0010860		+ -0.2008				5	-22.273	54.5	0.00	0.313
13.810	-0.0166000		+ -0.2826		0.0003621		6	-21.846	55.7	1.14	0.177
10.240	-0.0007565		+ -0.2008	+			6	-21.895	55.8	1.24	0.168
9.794	-0.0010860		+ -0.1886			+	6	-22.251	56.5	1.95	0.118
13.730	-0.0162700		+ -0.2826		+ 0.0003621		7	-21.460	56.9	2.37	0.096
13.290	-0.0166000		+ -0.2704		0.0003621	+	7	-21.823	57.6	3.10	0.067
10.100	-0.0007605		+ -0.1974	+		+	7	-21.893	57.8	3.24	0.062

Example 2: Ant species richness

Another way to form a “95% confidence set of models”, analogous to a 95% confidence interval for a parameter, is to use cumulative model weights.



Example 2: Conclusions

If regression is purely for prediction, all of the models with relatively small ΔAIC predict about equally well. This means there's no reason to get too excited over a single "best" model. Present the confidence set of models, the same way you would a confidence interval for a parameter.

The interpretation is more complex if regression is used for explanation, to identify the causes of variation in *y*. Numerous models might be nearly equally good at fitting the data, in which case the causes of variation in *y* are more elusive.

Keep in mind that, like correlation, "regression is not causation". It is not possible to find the true causes of variation in the explanatory variable without an experiment.

AIC (Akaike's Information Criterion)

Search strategies:
One method is a stepwise procedure for selection of variables implemented by `stepAIC()` in the `MASS` library in R. Another is `dredge()` in the `MuMIn` package, which search all subsets while obeying restrictions.

Both methods obey restrictions. Not all terms are on equal footing. E.g.,

- Squared term 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

However, keep in mind that often we are data dredging. The only intelligent decision we've made is the choice of variables to include in our dredge. No other scientific insight was used to decide an *a priori* set of models.

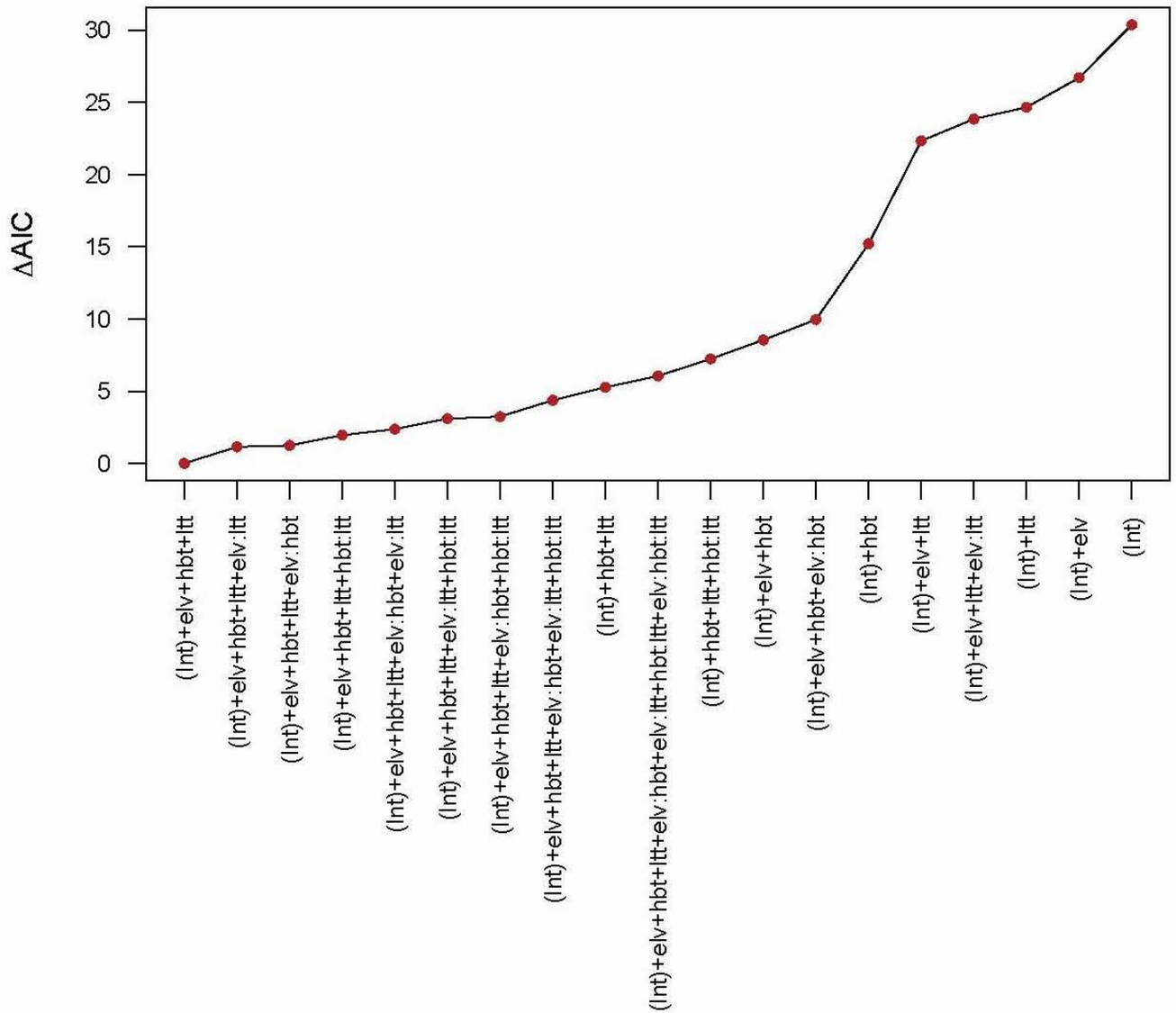
How AIC differs from classical statistical approaches

No null hypothesis.

No hypothesis testing.

No *P*-value.

No model is formally “rejected”.

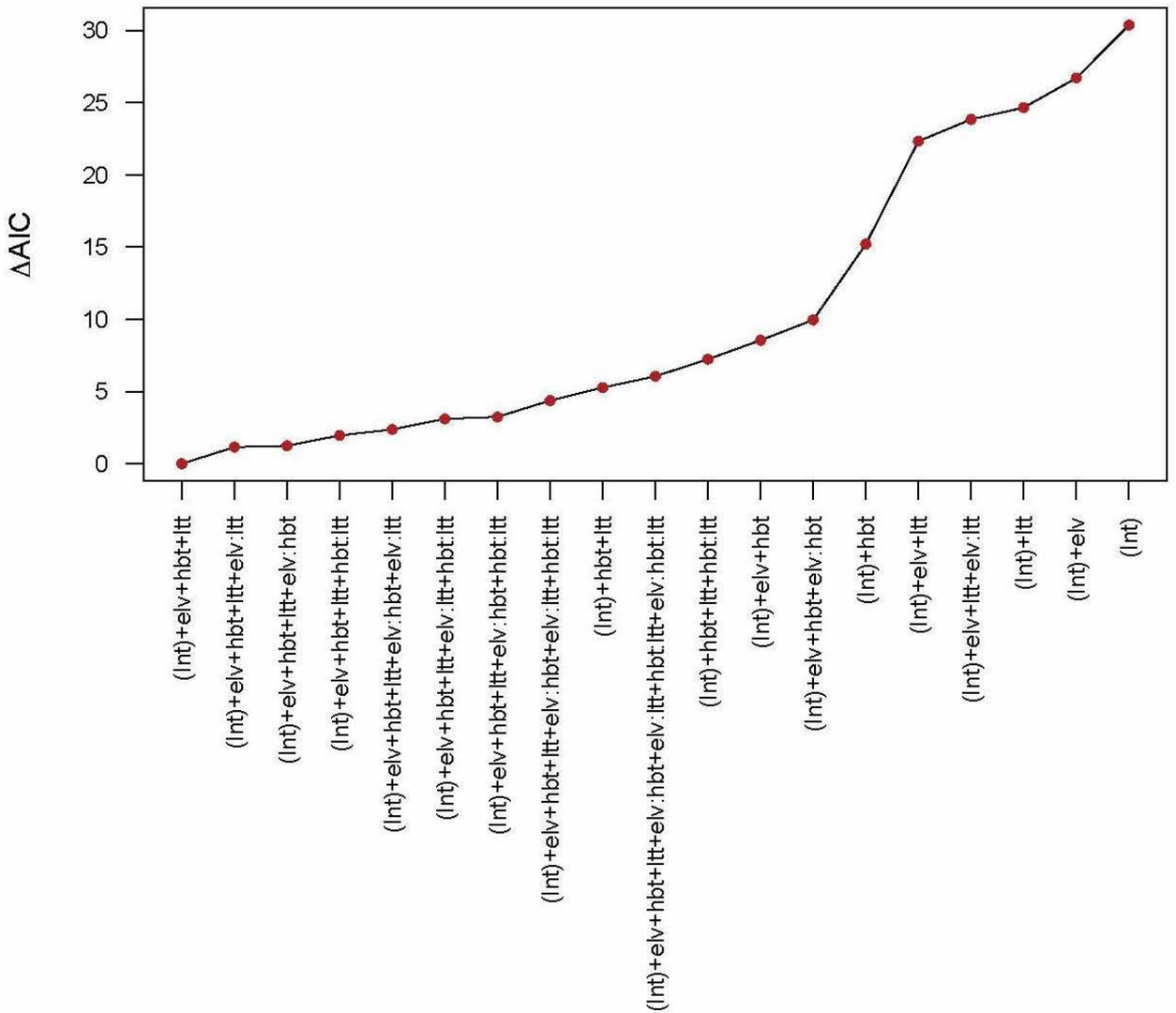


How AIC differs from classical statistical approaches

Several models may be about equally good.

Your “best” model isn’t necessarily the true model.

This is because AIC balances the bias-variance trade-off. It does a good job to minimize information loss, on average. A larger data set might yield a more complex model.



How AIC differs from classical statistical approaches

Model uncertainty

AIC difference (Δ) support
0 – 2 Substantial support
4 – 7 Considerably less support
> 10 Essentially no support

The reason for model uncertainty is sampling error. Remember always 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.

Think of all the models that have some support as constituting a “confidence set” of models, analogous to a confidence interval when estimating a parameter.

Going further: Multimodel Inference

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

"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.

Implemented in `MuMIn` package in R.

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

Avoid data-dredging by formulating a set of candidate models

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

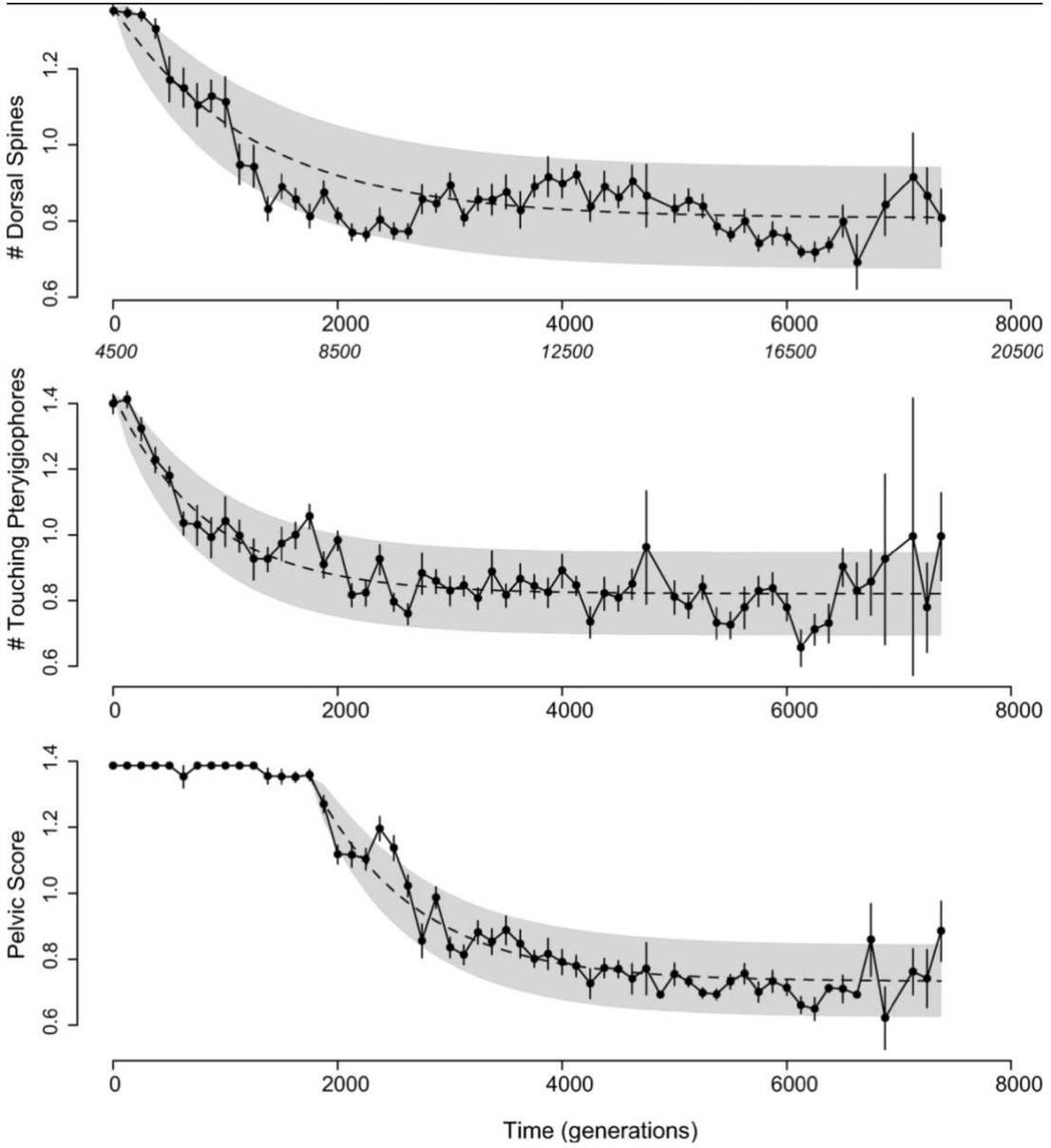
This is where data dredging ends and science begins.

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.

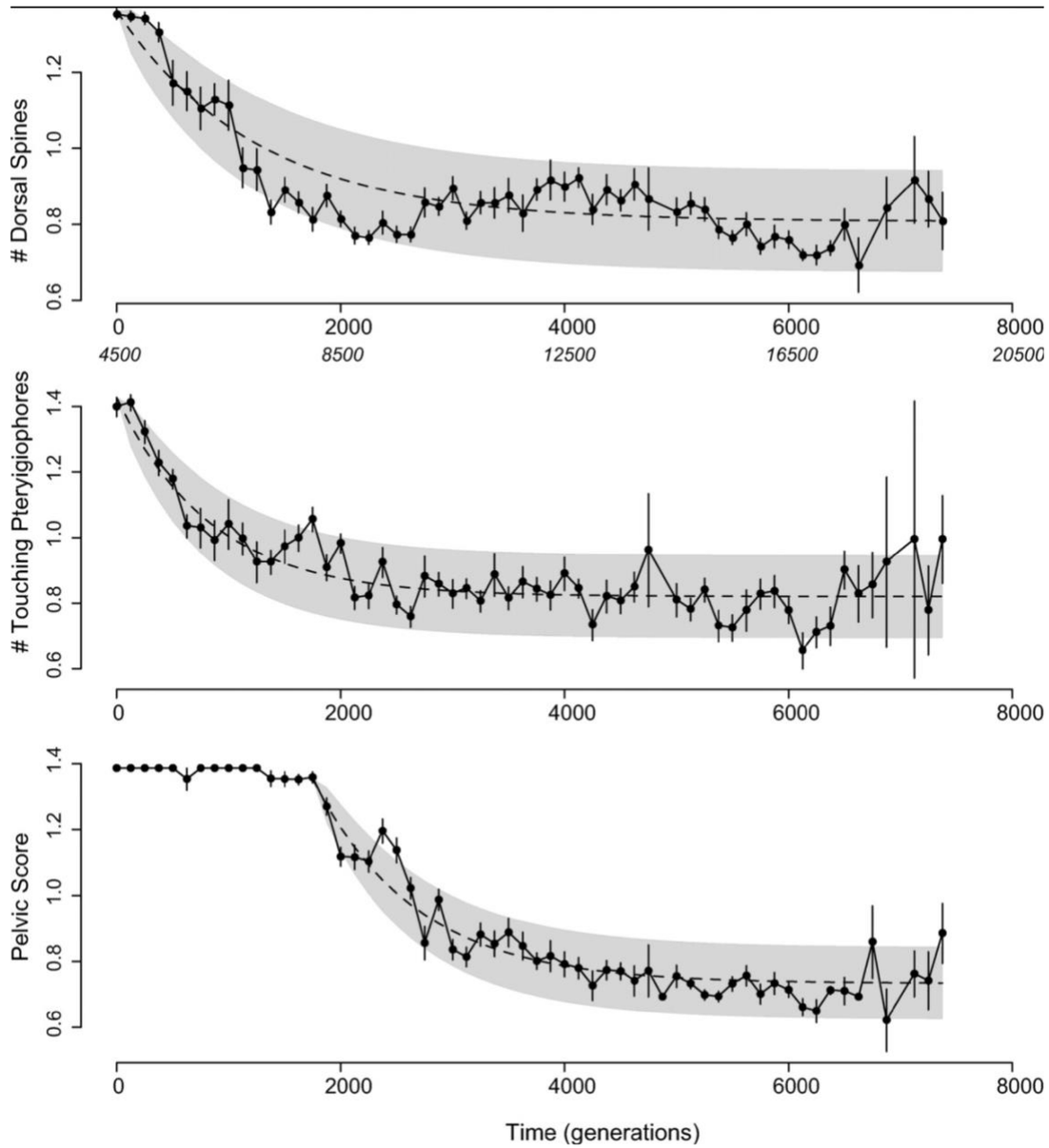
G. Hunt, M. A. Bell & M. P Travis 2008, *Evolution* 62: 700–710.



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 the AIC criterion 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 a single “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
No. of dorsal spines	Neutral	86.48	2	−168.73	0.002
	Adaptive	94.94	4	−181.11	**0.998**
Pterygiophores	Neutral	65.91	2	−127.59	0.001
	Adaptive	74.80	4	−140.84	**0.999**
Pelvic score	Neutral	58.38	2	−112.46	0.001
	Adaptive	68.33	4	−127.65	**0.999**

The adaptive model beats neutral drift for all three traits.

Akaike weight is the weight of evidence in favor of a model being the best model among the set being considered, and 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 the cumulative sum reaches 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, $P = 0.00005$

Stepping back 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 even under the conventional hypothesis testing framework, specifying 2 specific candidate models is already superior to an approach in which the alternative hypothesis is merely “everything but the null hypothesis.”

Conclusions

Stepwise elimination of terms and null hypothesis significance testing is not the ideal approach for model selection. Information-theoretic approaches have explicit criteria and defensible properties.

This approach compares models but needs no *P*-values because there is no null model.

These IT approaches work best when thoughtful science is used to specify the candidate models under consideration before testing (minimizing data dredging).

Working with a set of models that fit the data about equally well, rather than with the one single best model, recognizes that there is model uncertainty.

If you want more certainty about which variables cause variation in the response variable, then you will need to do an experiment.

Digression: Exploring your data can be good

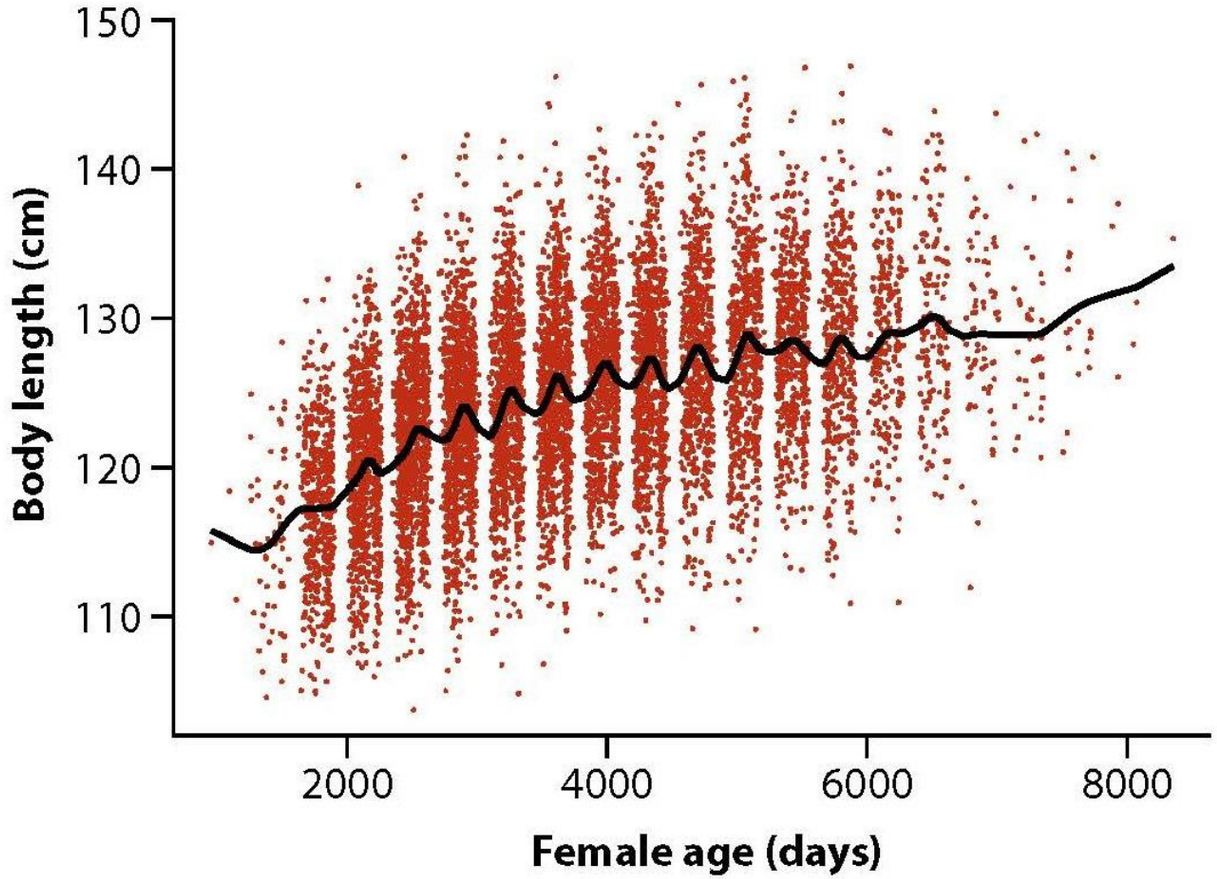


Figure 17.8-3 Measurements of body length as a function of age for female fur seals, with the "spline" fit in black.

Discussion paper for next week:

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

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

Presenters:

Moderators: