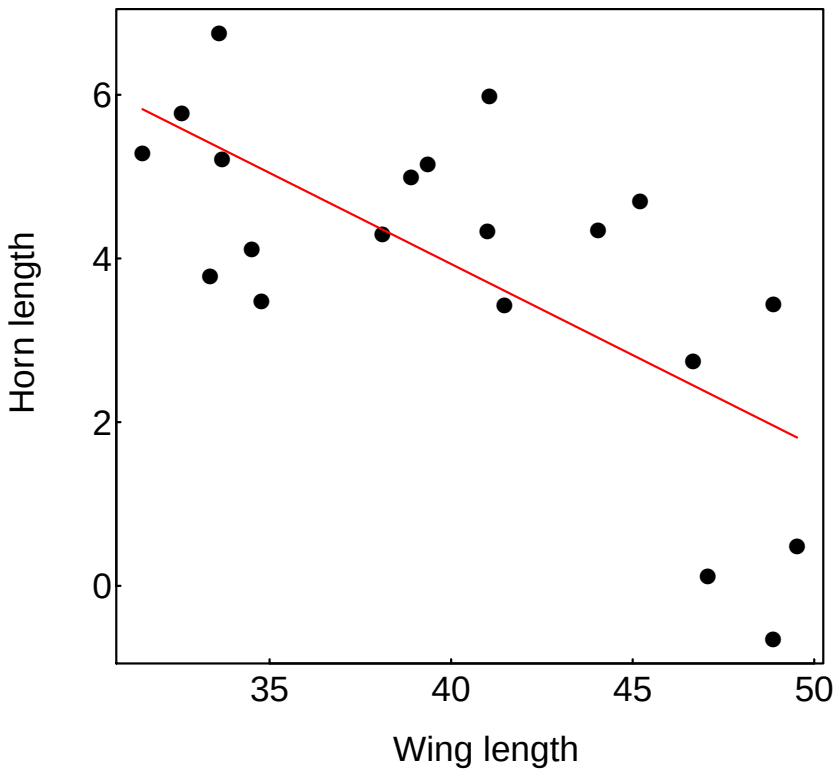
BISC-869, Model selection

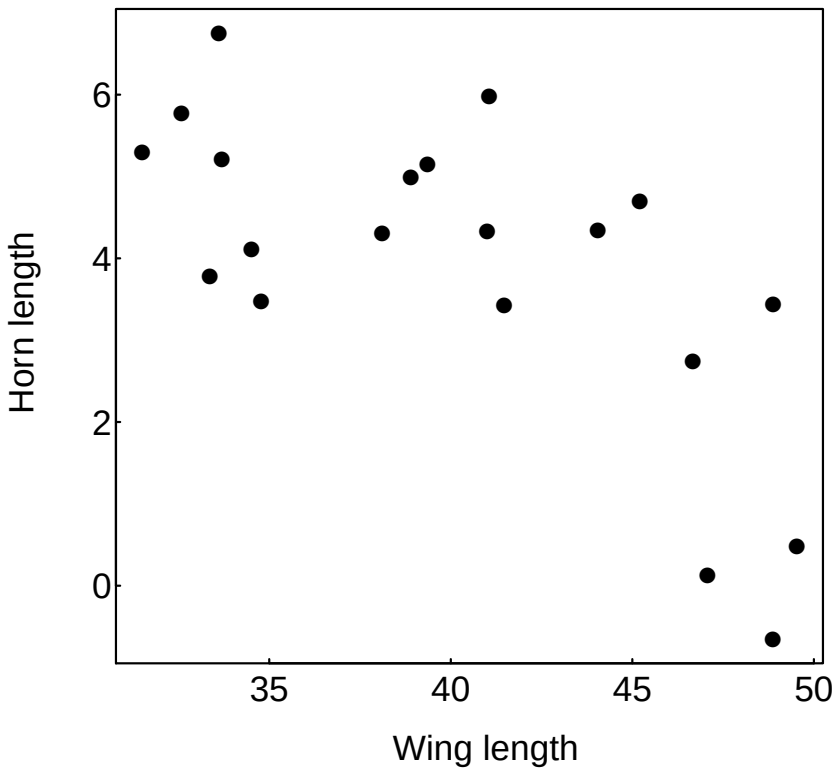
March 15, 2022

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

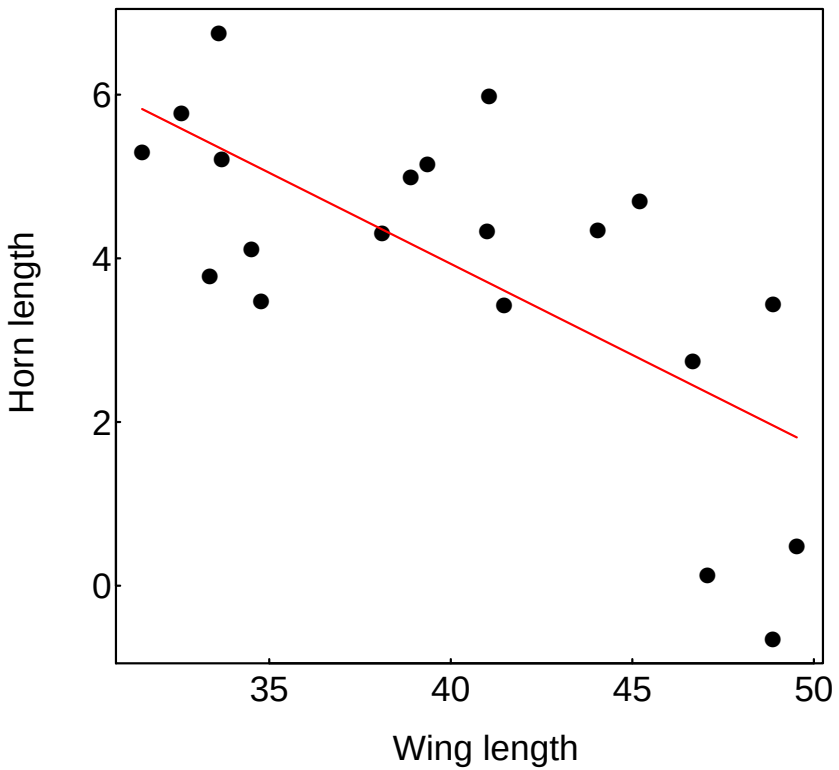
Emlen, D. J. 2001. Costs and the diversification of exaggerated animal structures. *Science* 291: 1534–1536.



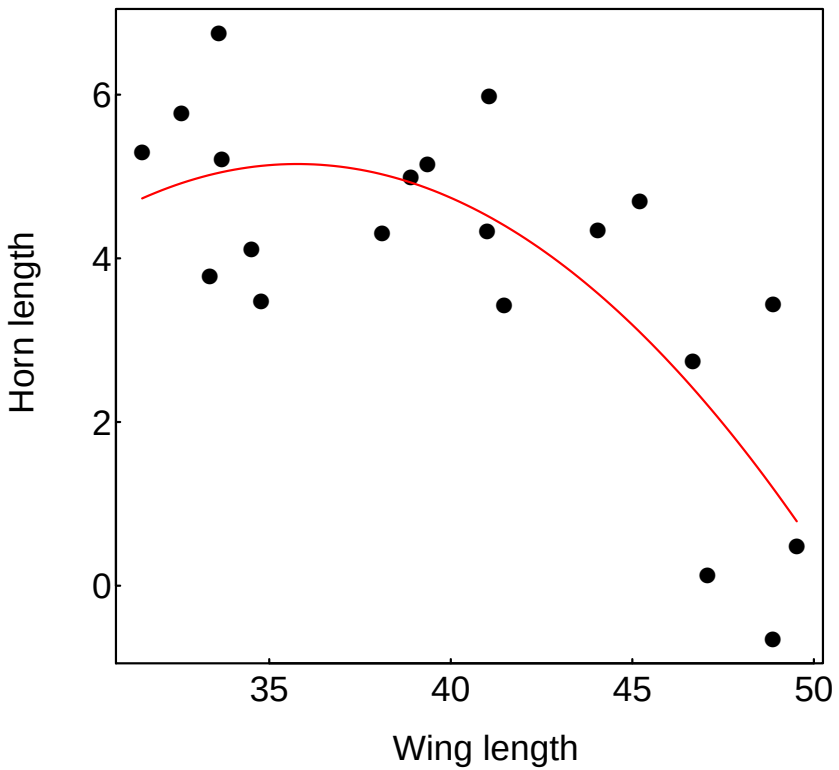
Start with a linear regression:



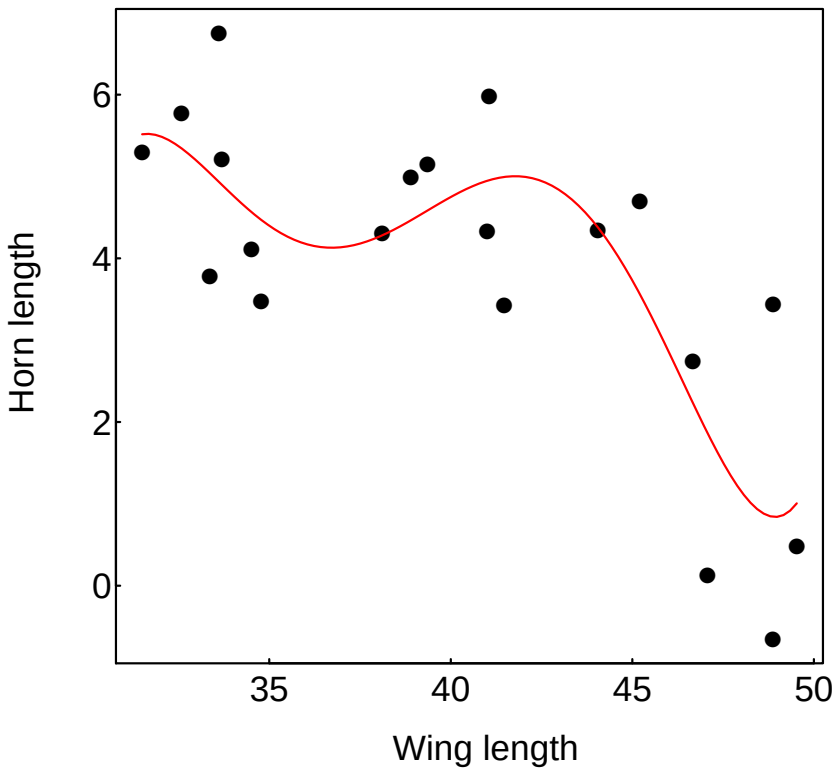
Start with a linear regression:



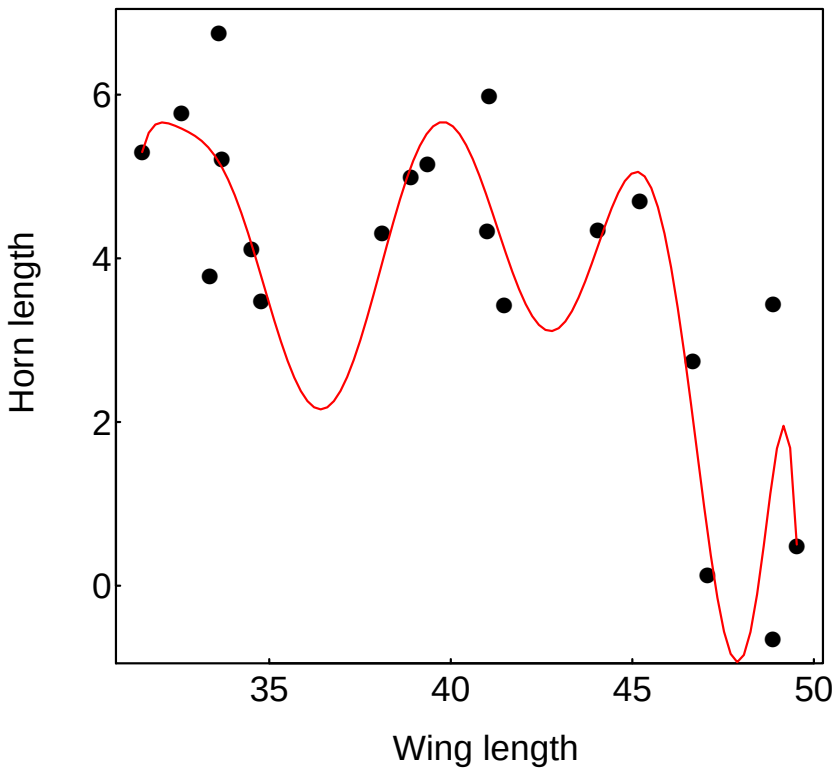
Now lets try a quadratic regression (polynomial, degree 2):



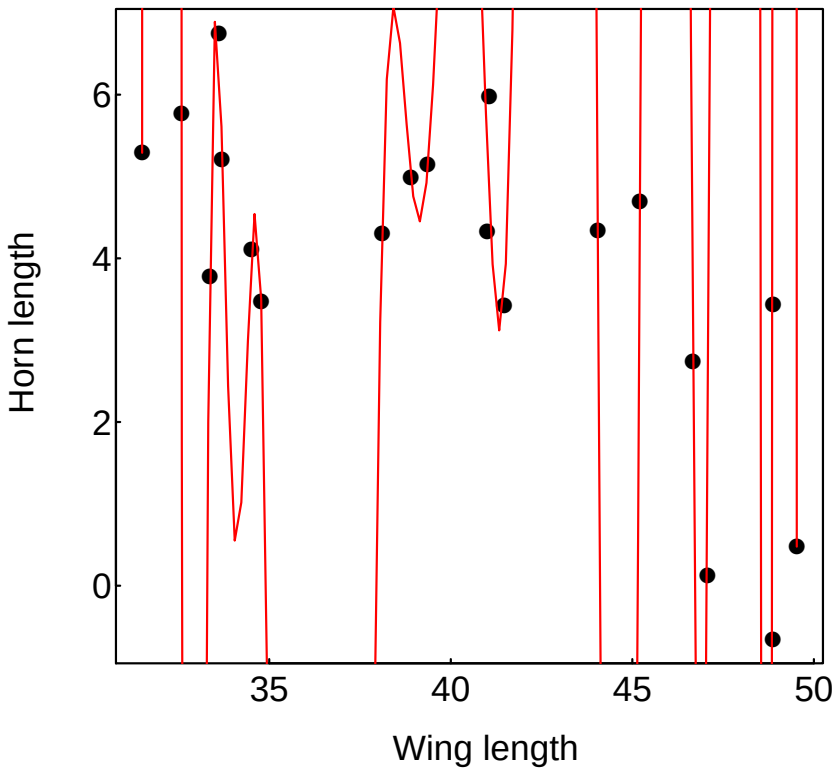
$\cdots$ polynomial, degree 5:



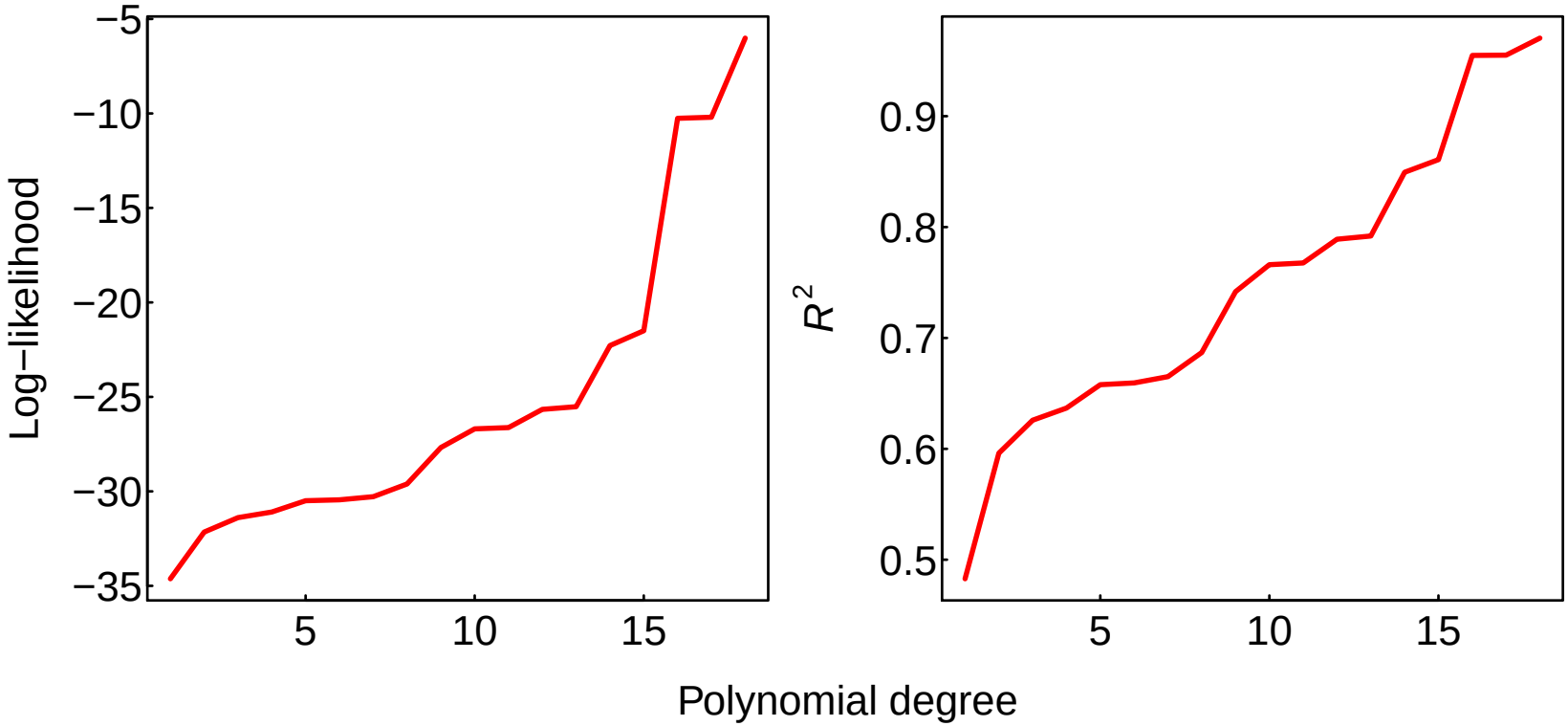
$\cdots$ polynomial, degree 10:



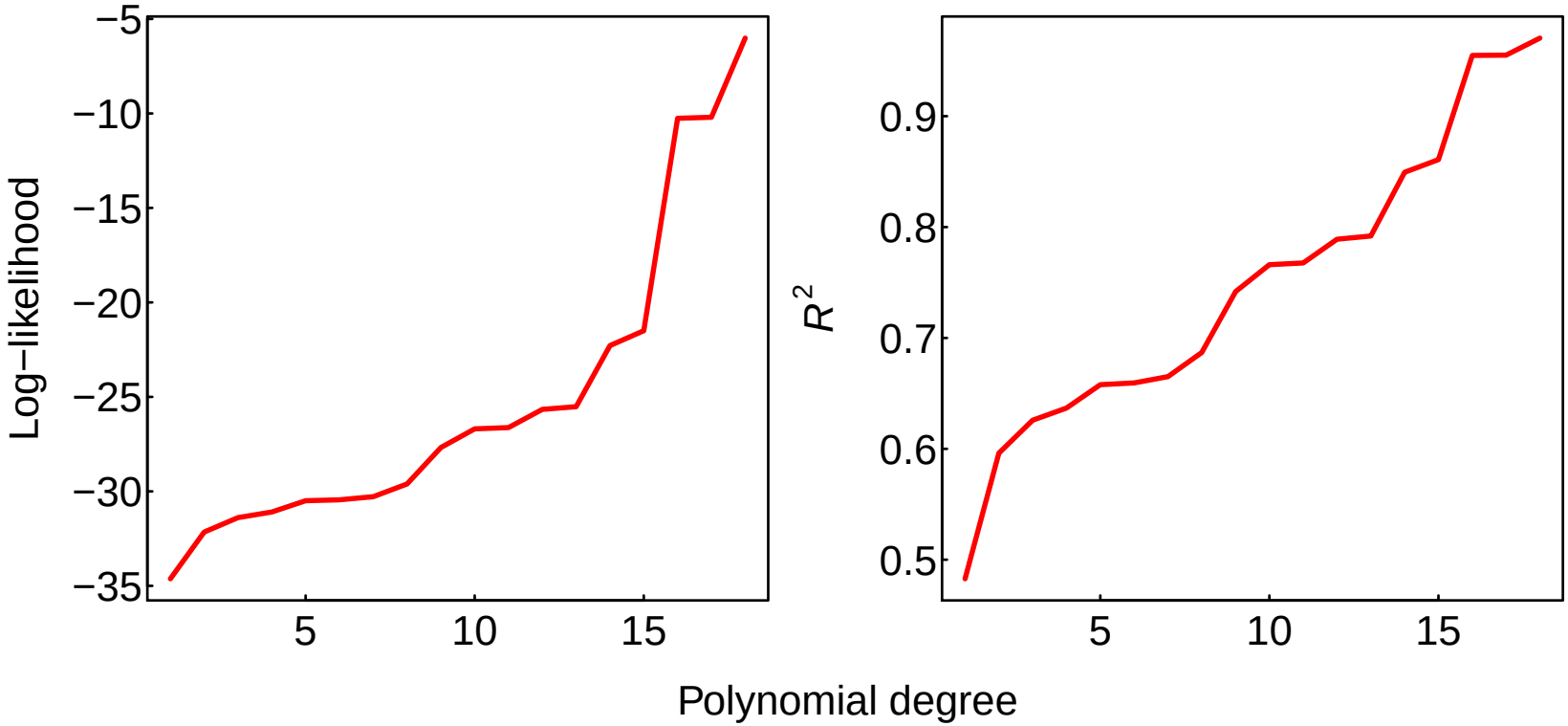
$\cdots$ polynomial, degree 18:



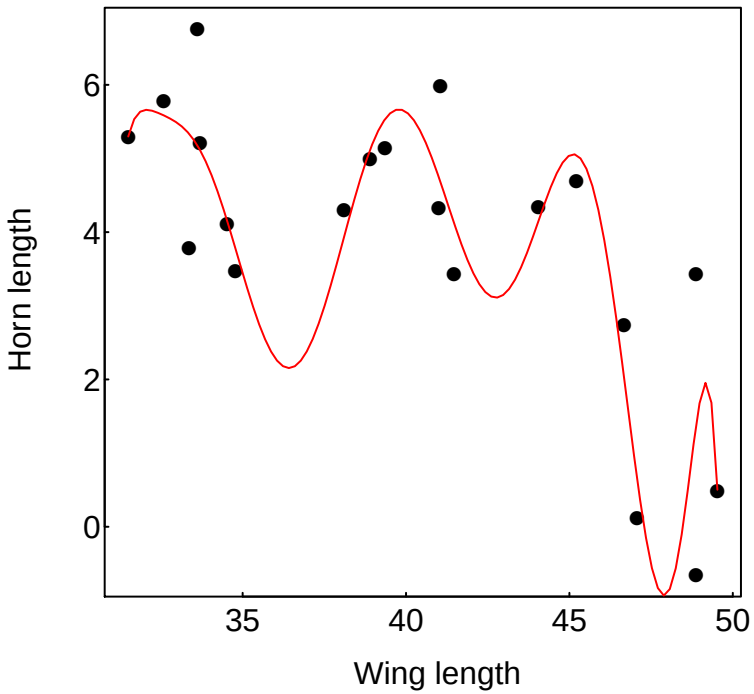
R^2 and log-likelihood increase with number of parameters in model.



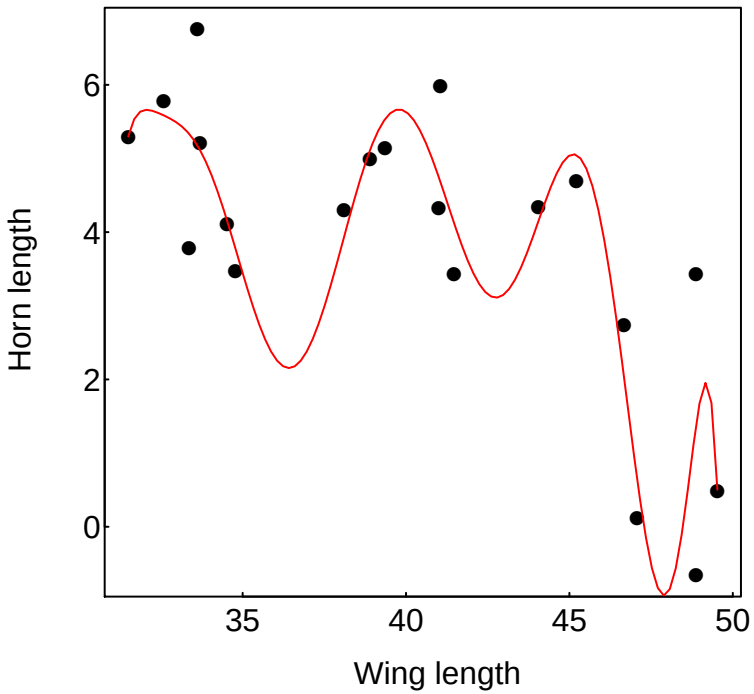
R^2 and log-likelihood increase with number of parameters in model.



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

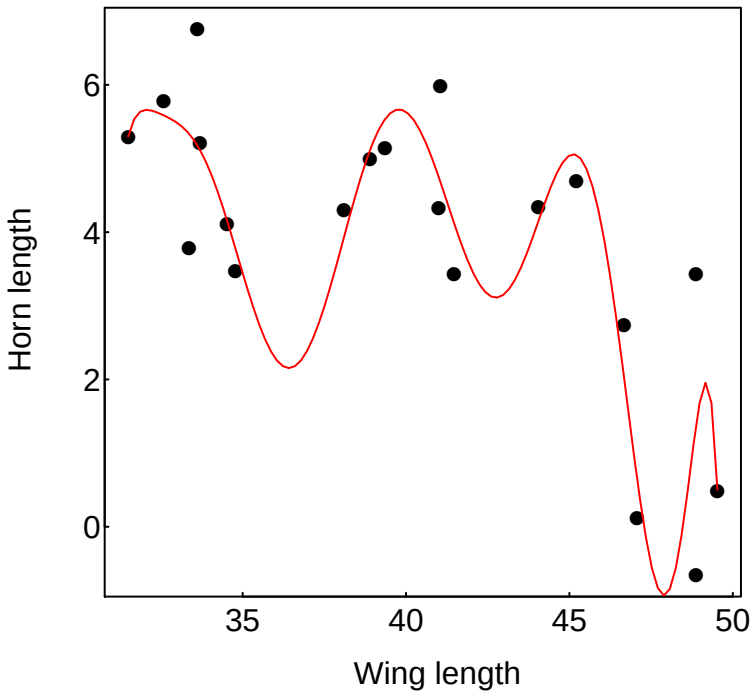


What is wrong with this picture?



What is wrong with this picture?

Does it violate some principle?

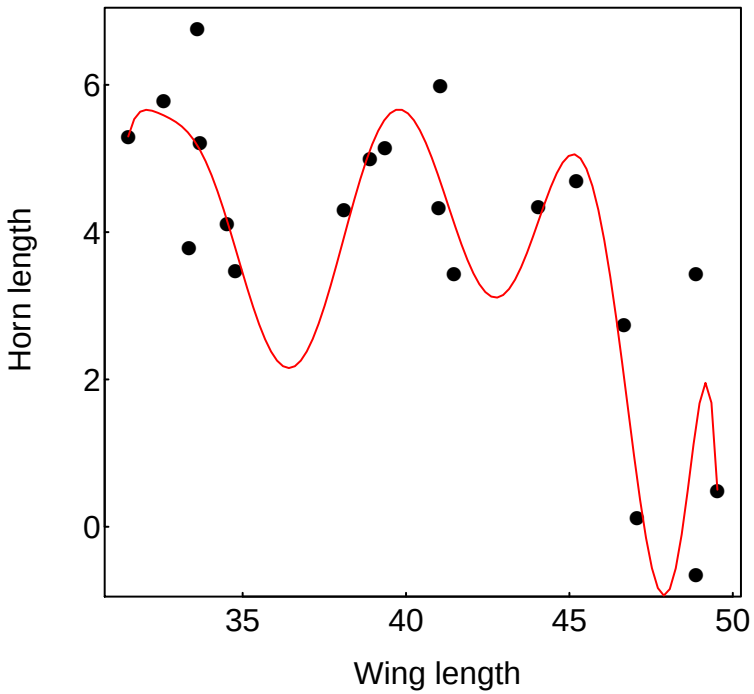


What is wrong with this picture?

Does it violate some principle?

Parsimony principle: Fit no more parameters than are 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



What is wrong with this picture?

Does it violate some principle?

Parsimony principle: Fit no more parameters than are 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

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

Stepwise elimination of terms is a common practice.

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.

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 up with a single, final model, the "minimum adequate model."

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 up with a single, final model, the "minimum adequate model."

Does stepwise elimination of terms actually yield the "best" model? What criterion are we actually using to decide which model is "best"?

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 up with a single, final model, the "minimum adequate model."

Does stepwise elimination of terms actually yield the "best" model? What criterion are we actually using to decide which model is "best"?

Each time we drop a variable from the model, we are "accepting" a null hypothesis. What happens if we accept a false null hypothesis? Does a sequence of Type II errors inevitably bring us to the wrong answer?

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 up with a single, final model, the "minimum adequate model."

Does stepwise elimination of terms actually yield the "best" model? What criterion are we actually using to decide which model is "best"?

Each time we drop a variable from the model, we are "accepting" a null hypothesis. What happens if we accept a false null hypothesis? Does a sequence of Type II errors inevitably bring us to the wrong answer?

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

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 up with a single, final model, the "minimum adequate model."

Does stepwise elimination of terms actually yield the "best" model? What criterion are we actually using to decide which model is "best"?

Each time we drop a variable from the model, we are "accepting" a null hypothesis. What happens if we accept a false null hypothesis? Does a sequence of Type II errors inevitably bring us to the wrong answer?

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

Might models with different subsets of variables fit the data nearly as well?

Choose among models using an explicit criterion

A reasonable criterion: choose the model that predicts best.

A reasonable criterion: choose the model that predicts best.

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

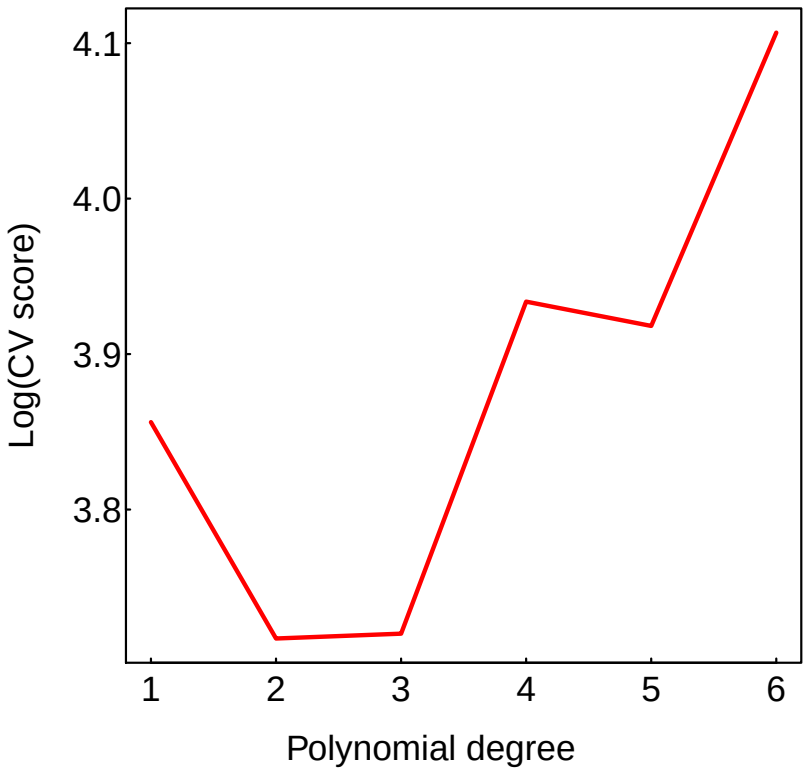
$$\text{CVscore} = \sum e_{(i)}^2$$

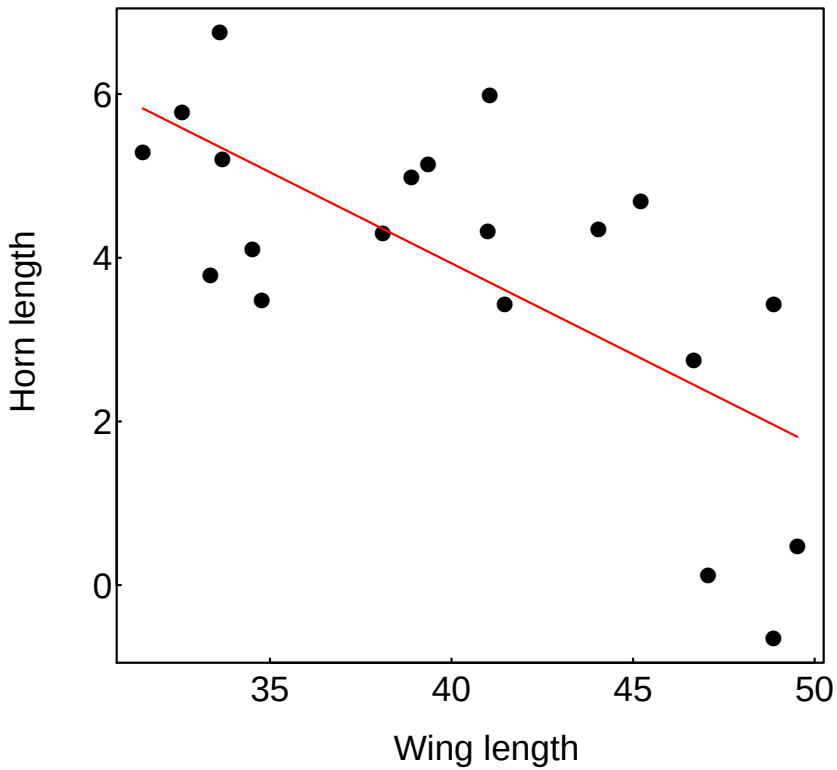
where:

- $e_{(i)}^2 = (y_i - y_{(i)})^2$.
- y_i are the observations for the response variable.
- $y_{(i)}$ is the predicted value for y_i when the model is fitted to the data leaving out y_i.

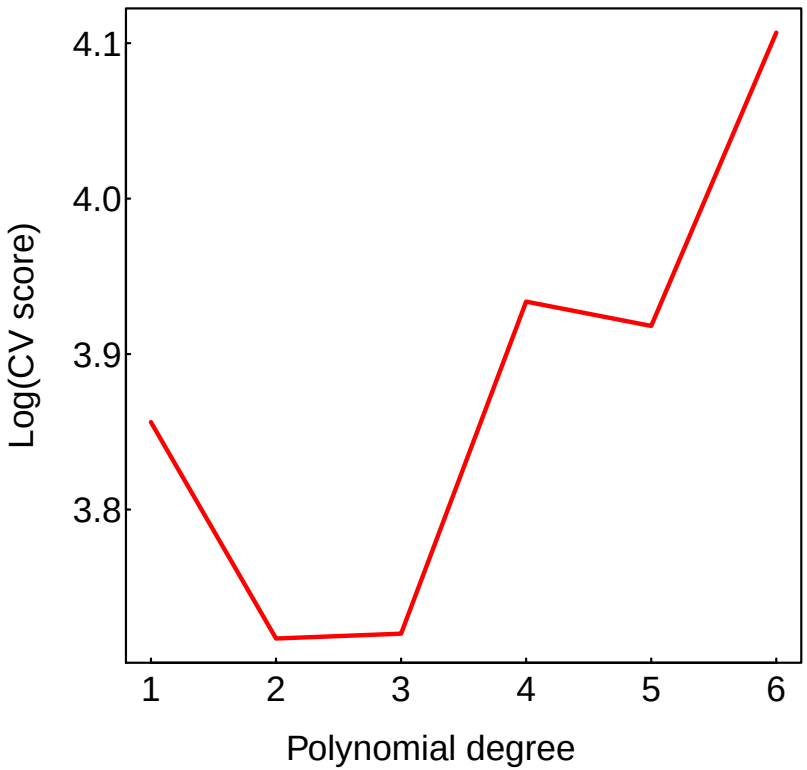
Larger CVscore corresponds to worse prediction (more prediction error).

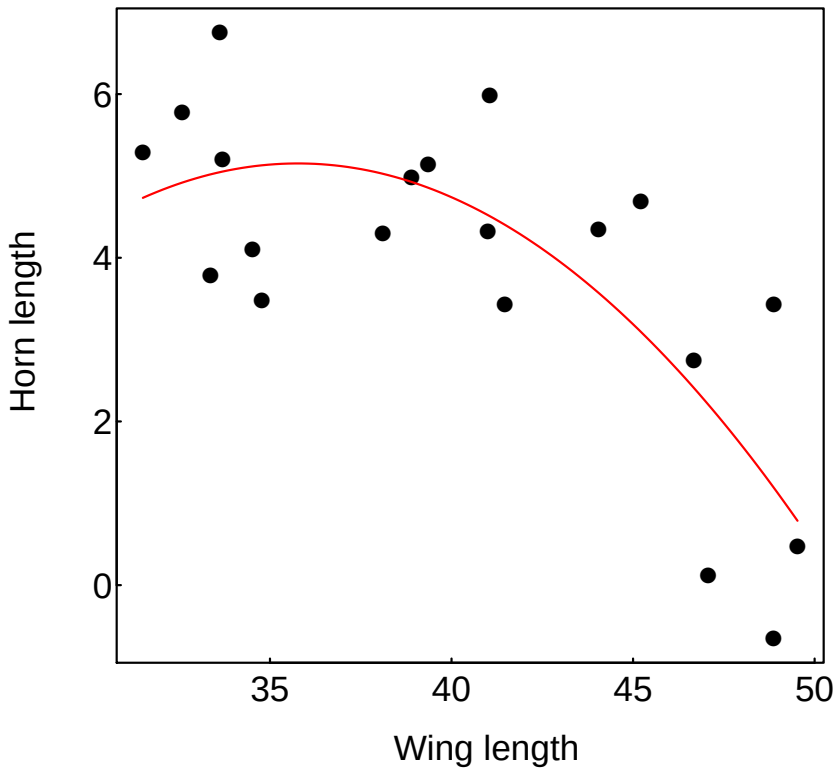
In our beetle example, the CVscore increases (prediction error worsens) with increasing numbers of parameters in the model. Here, the quadratic linear regression was "best". But cubic does nearly equally well.



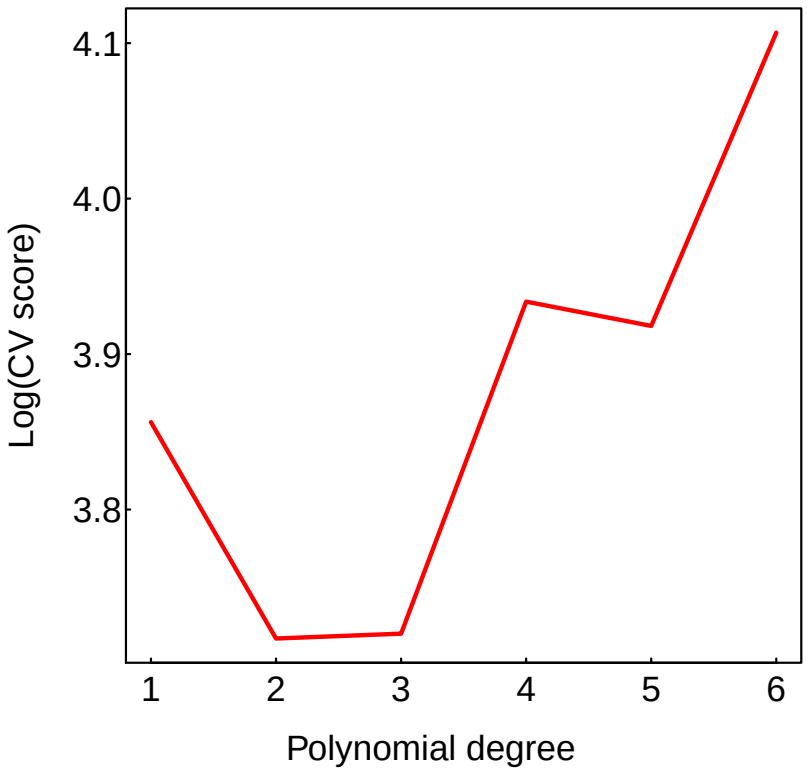


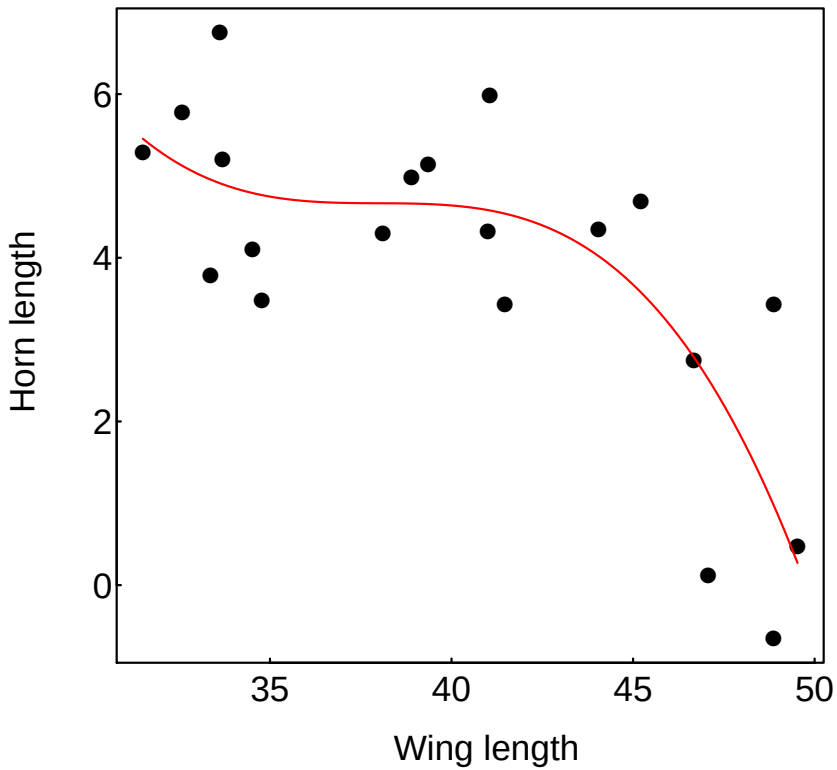
In our beetle example, the CVscore increases (prediction error worsens) with increasing numbers of parameters in the model. Here, the quadratic linear regression was "best". But cubic does nearly equally well.



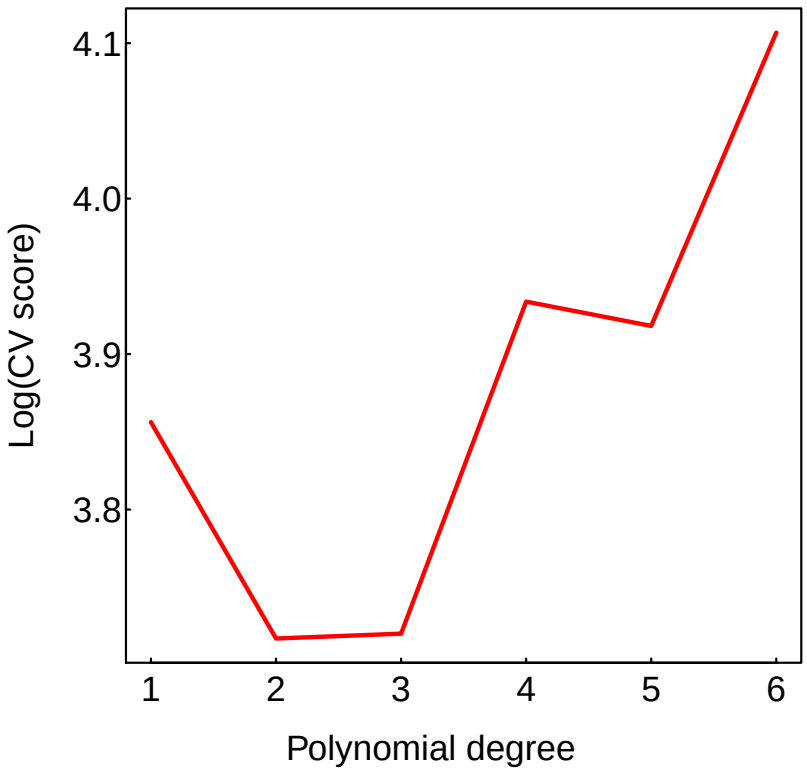


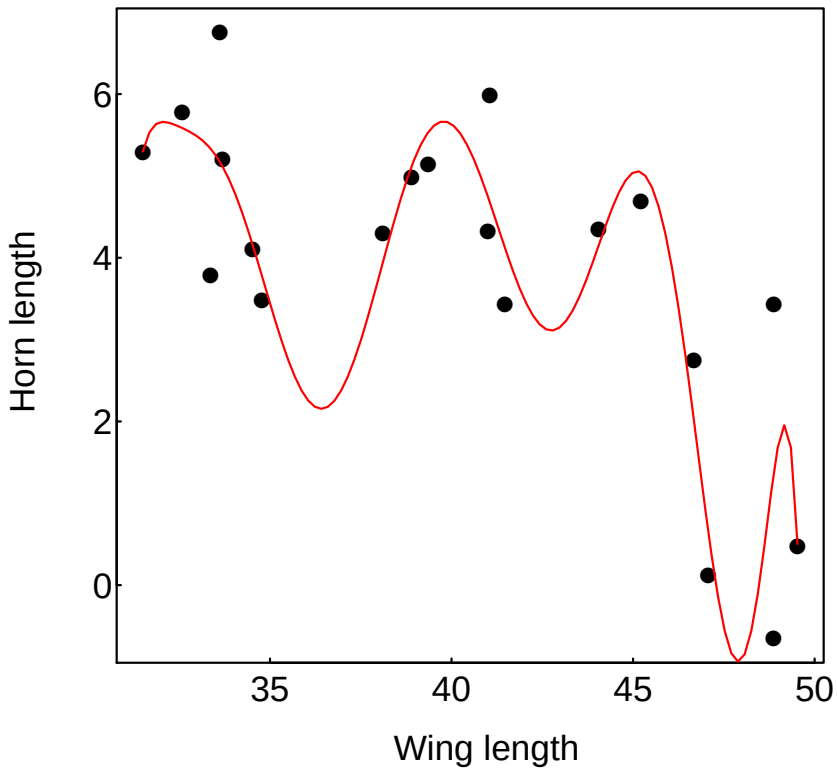
In our beetle example, the CVscore increases (prediction error worsens) with increasing numbers of parameters in the model. Here, the quadratic linear regression was "best". But cubic does nearly equally well.





In our beetle example, the CVscore increases (prediction error worsens) with increasing numbers of parameters in the model. Here, the quadratic linear regression was "best". But cubic does nearly equally well.





Prediction worsens as models become complex because of bias-variance tradeoff.

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

Prediction worsens as models become complex because of bias-variance tradeoff.

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:** There is not enough data to yield good estimates of many parameters, leading to high sampling error (low precision).

Prediction worsens as models become complex because of bias-variance tradeoff.

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:** There is not enough data to yield good estimates of many parameters, leading to high sampling error (low precision).

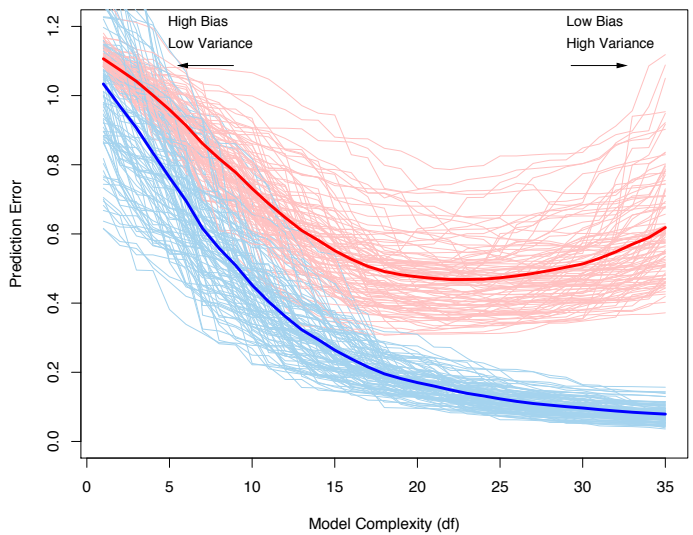


FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\text{err}}$*, while the light red curves show the conditional test error* $\text{Err}_{\mathcal{T}}$ *for* 100 *training sets of size* 50 *each,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

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.

Prediction worsens as models become complex because of bias-variance tradeoff.

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:** There is not enough data to yield good estimates of many parameters, leading to high sampling error (low precision).

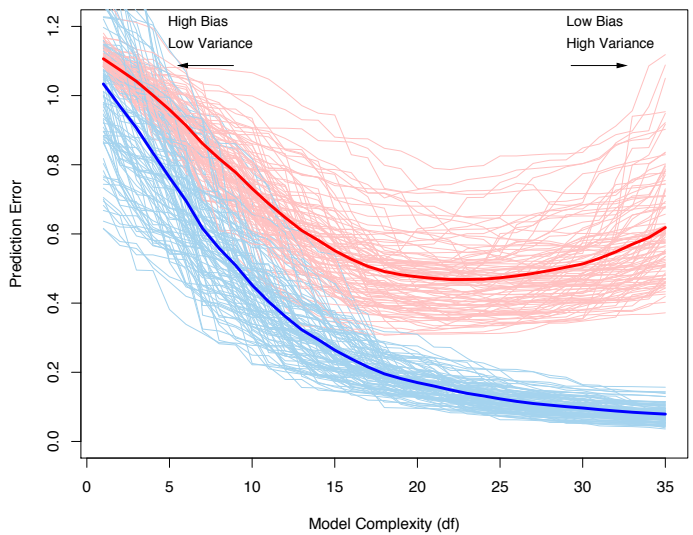


FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\text{err}}$*, while the light red curves show the conditional test error* $\text{Err}_{\mathcal{T}}$ *for* 100 *training sets of size* 50 *each,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

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.

The simplest models have low variance but high bias resulting from missing terms.

Prediction worsens as models become complex because of bias-variance tradeoff.

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:** There is not enough data to yield good estimates of many parameters, leading to high sampling error (low precision).

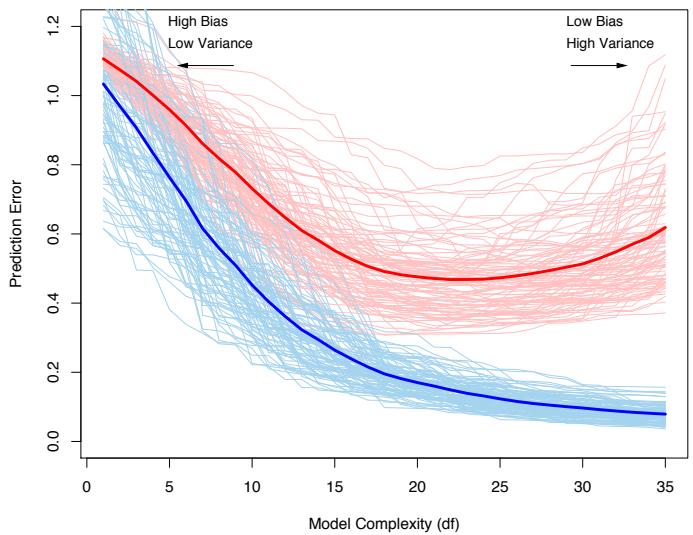


FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\text{err}}$*, while the light red curves show the conditional test error* $\text{Err}_{\mathcal{T}}$ *for* 100 *training sets of size* 50 *each,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

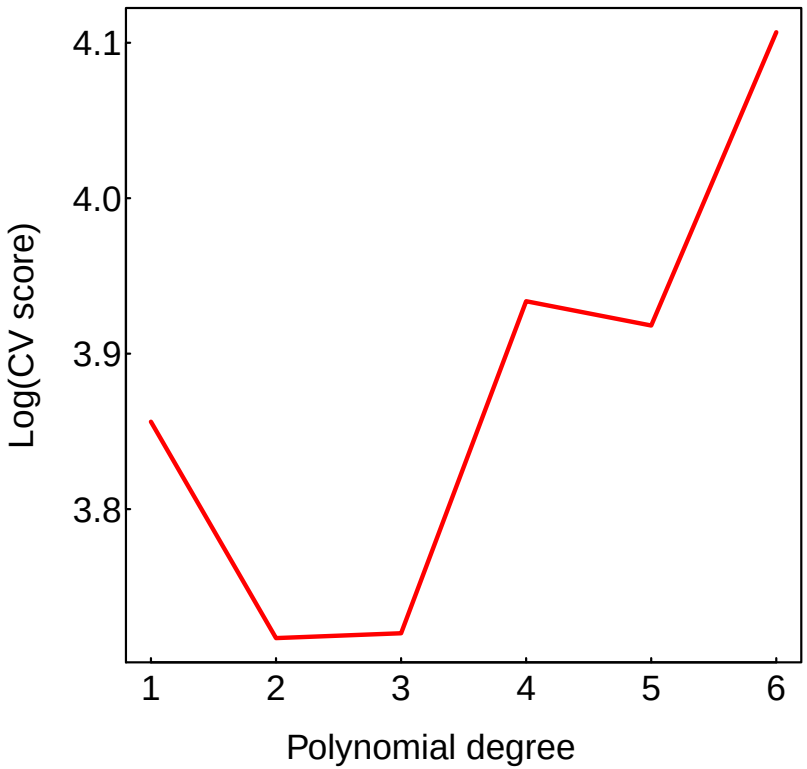
Training error: how well a model fits the data used to fit the model.

Test error: how well a model fits a new sample of data.

The simplest models have low variance but high bias resulting from missing terms.

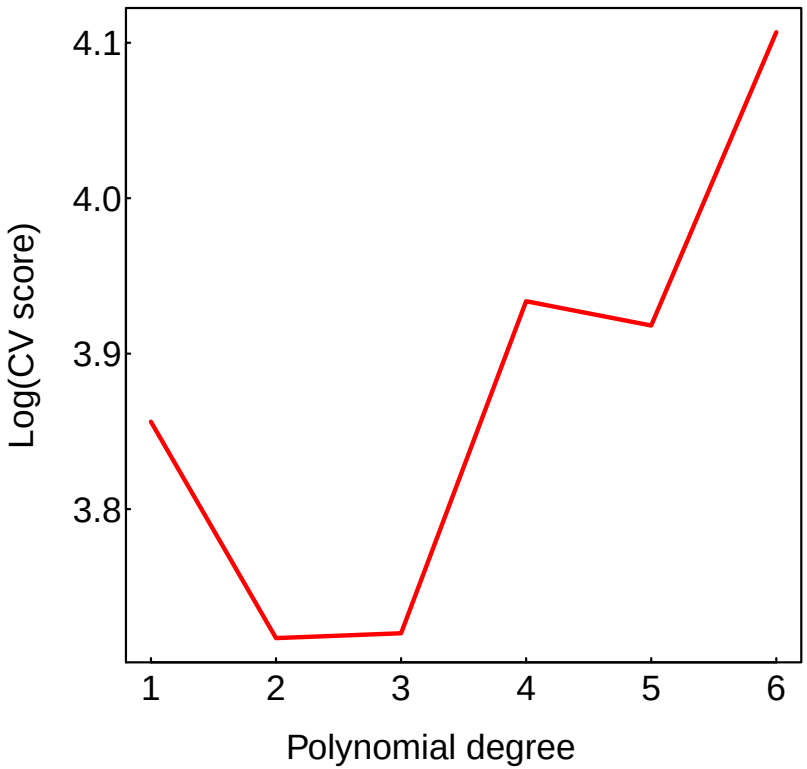
The most complex models have low bias but high variance resulting from estimating too many parameters ("overfitting") with limited data.

What else is worrying about our polynomial regression analysis?



We're data dredging. We didn't have any hypotheses to help guide our search. This too can lead to non-reproducible results.

What else is worrying about our polynomial regression analysis?



We're data dredging. We didn't have any hypotheses to help guide our search. This too can lead to non-reproducible results.

E.g., the 10^{th} degree polynomial is surprisingly good at prediction. But is there any good, *a priori*, reason to include it among the set of candidate models to evaluate?

Some reasonable objectives:

- A model that predicts well.

Some reasonable objectives:

- A model that predicts well.
- A model that approximates the true relationship between the variables.

Some reasonable objectives:

- A model that predicts well.
- A model that approximates the true relationship between the variables.
- Information on which models fit the data nearly as well as the "best" model.

Some reasonable objectives:

- A model that predicts well.
- A model that approximates the true relationship between the variables.
- Information on which models fit the data nearly as well as the "best" model.
- To compare non-nested* models, rather than just compare a "full" model to "reduced" models having a subset of its terms.

*Reduced vs. full models 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.

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

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

- A **criterion** to compare models:

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

- A **criterion** to compare models:
  - Mallow's C_p

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

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

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

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

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

- 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, proposed in 1973, is frequently used in multiple regression.

Mallow's C_p, proposed in 1973, is frequently used in multiple regression.

$$C_p = \frac{SS_{\text{error}}}{\hat{\sigma}^2} - n + 2p$$

where:

- SS_{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 (explanatory variables) in the model (including the intercept)

Mallow's C_p, proposed in 1973, is frequently used in multiple regression.

$$C_p = \frac{SS_{\text{error}}}{\hat{\sigma}^2} - n + 2p$$

where:

- SS_{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 (explanatory variables) in the model (including the intercept)

C_p estimates the mean square prediction error.

Mallow's C_p, proposed in 1973, is frequently used in multiple regression.

$$C_p = \frac{SS_{\text{error}}}{\hat{\sigma}^2} - n + 2p$$

where:

- SS_{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 (explanatory variables) in the model (including the intercept)

C_p estimates the mean square prediction error.

It is equivalent to AIC in the case of multiple regression with independent normal errors.

Mallow's C_p, proposed in 1973, is frequently used in multiple regression.

$$C_p = \frac{SS_{\text{error}}}{\hat{\sigma}^2} - n + 2p$$

where:

- SS_{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 (explanatory variables) in the model (including the intercept)

C_p estimates the mean square prediction error.

It is equivalent to AIC in the case of multiple regression with independent normal errors.

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

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

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

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

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

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

Models for which $C_p < p$ are all considered about as good.

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

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

Models for which $C_p < p$ are all considered about as good.

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

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

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

Models for which $C_p < p$ are all considered about as good.

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

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: 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.

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.

```
head(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
5   CB        9  forest    42.05       210
6   RP       15  forest    42.17        78
```

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.

```
head(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
5   CB        9  forest    42.05       210
6   RP       15  forest    42.17        78
```

```
tail(ants)
```

```
   site nspecies habitat latitude elevation
39  PEA        3     bog    44.29       468
40  CHI        2     bog    44.33       362
41  MOL        3     bog    44.50       236
42  COL        2     bog    44.55        30
43  MOO        5     bog    44.76       353
44  CAR        5     bog    44.95       133
```

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

Regression model with all possible terms:

Regression model with all possible terms:

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

Regression model with all possible terms:

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

This evaluates all subsets of `habitat`, `latitude`, `elevation` and their 2- and 3-way interactions.

Regression model with all possible terms:

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

This evaluates all subsets of `habitat`, `latitude`, `elevation` and their 2- and 3-way interactions.

`leaps` requires that all variables be numeric (I disguised habitat as a numeric variable by scoring: 0=bog, 1=forest).

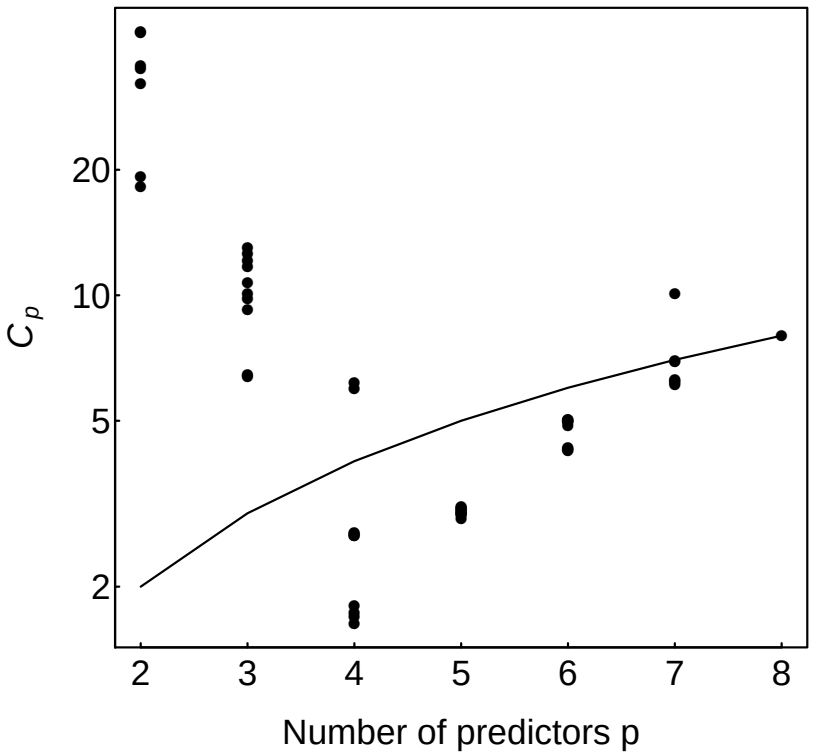
Regression model with all possible terms:

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

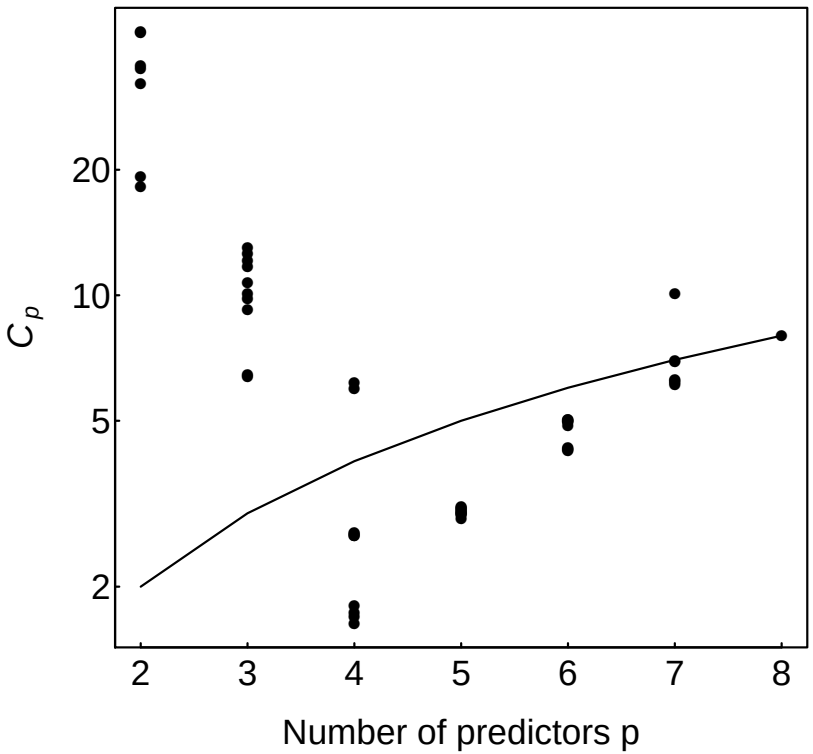
This evaluates all subsets of `habitat`, `latitude`, `elevation` and their 2- and 3-way interactions.

`leaps` requires that all variables be numeric (I disguised habitat as a numeric variable by scoring: 0=bog, 1=forest).

Not all the evaluated models are necessarily sensible (dubious to fit a model with a 3-way interaction and no main effects).

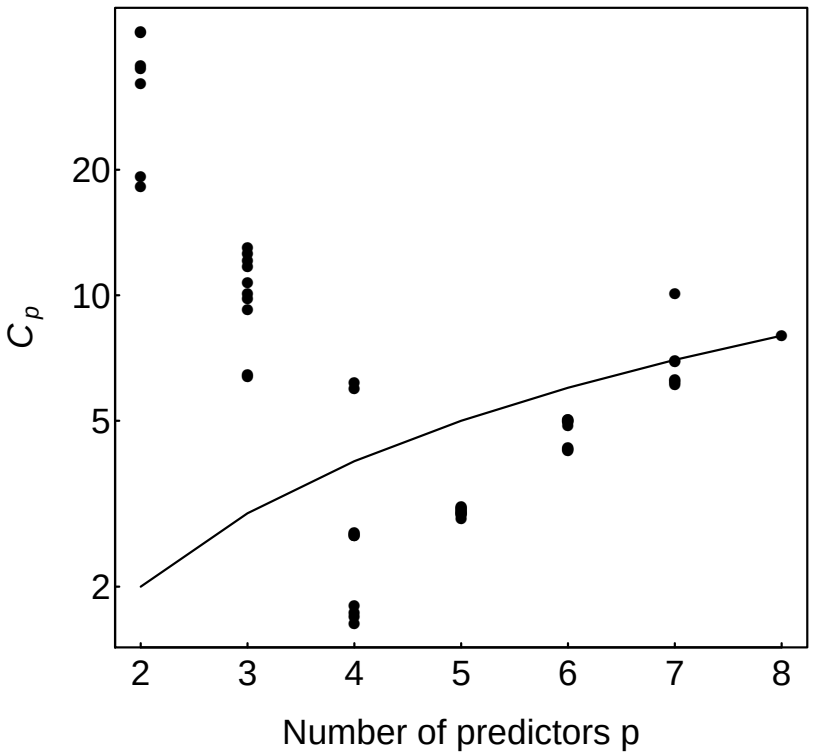


By default, `leaps` saves the top 10 models for each value of p.



By default, `leaps` saves the top 10 models for each value of p.

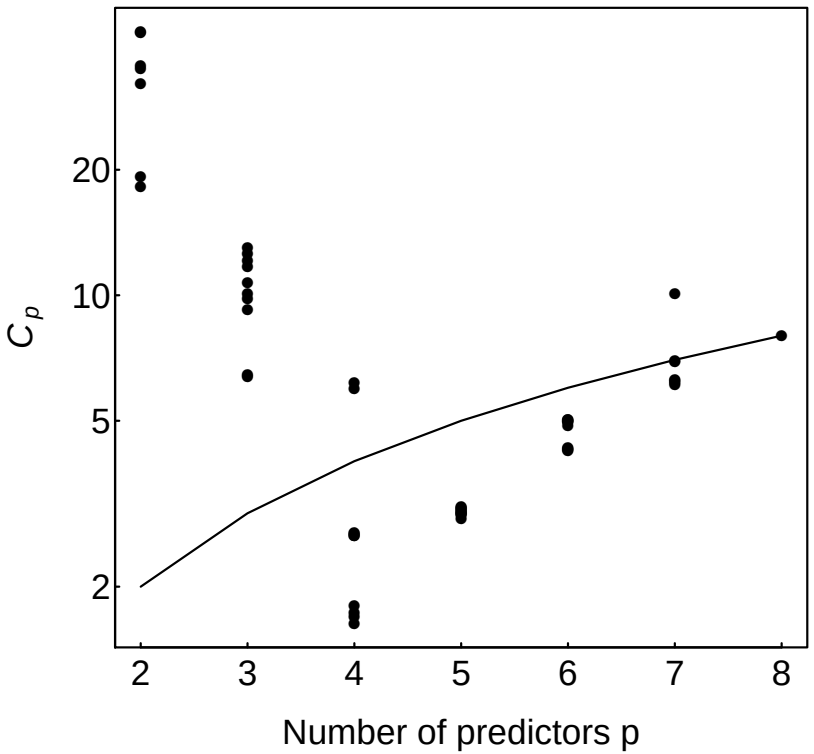
The line in the figure indicates $C_p = p$ (vertical axis is in log units).



By default, `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).



By default, `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 $C_p < p$.

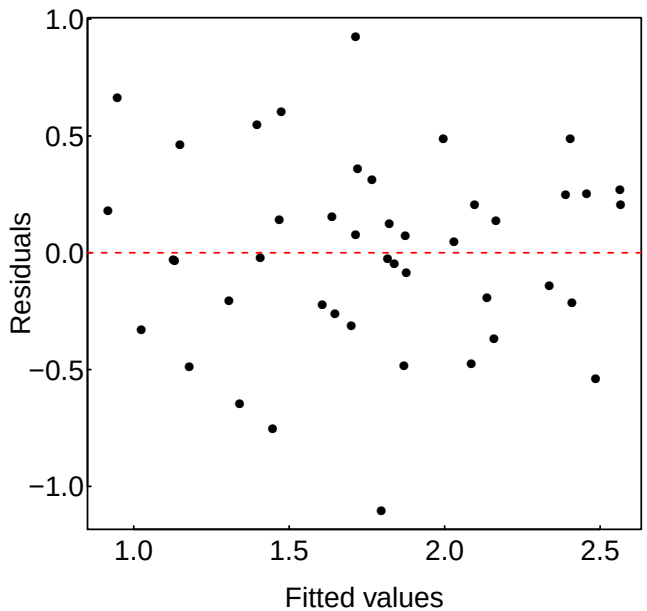
Best model (smallest C_p):

Note: I have shortened variable names

Best model (smallest C_p): *Note: I have shortened variable names*

```
out <- lm(log(nspecies)~hab + lat + ele)
Residuals:
     Min       1Q   Median       3Q      Max
-1.10279 -0.23082  0.01417  0.25020  0.92499

Coefficients:
             Estimate Std. Error t value   Pr(>|t|)
(Intercept) 10.3180285  2.6101963   3.953   0.000306 ***
hab          0.6898845  0.1269432   5.435 0.00000294 ***
lat         -0.2007838  0.0609920  -3.292   0.002085 **
ele         -0.0010856  0.0004049  -2.681   0.010610 *
```



A total of 34 models had $C_p < p$

hab	lat	ele	hab.lat	hab.ele	lat.ele	hab.lat.ele
TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE
TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE
FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE
TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE
TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE
TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE
FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE
TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE
TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE
FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE
FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE
TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE
TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE
FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE
FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE
TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE
TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE
FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE
FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE
TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE
TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE
TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE
FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE
FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE
TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE

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 with excitement over your single "best" model.

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 with excitement over your single "best" model.

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 "best explain" the response.

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 with excitement over your single "best" model.

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 "best explain" the response.

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 experimentation.

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

Criterion: minimize AIC.

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

Criterion: minimize AIC.

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

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

Criterion: minimize AIC.

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.

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

Criterion: minimize AIC.

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 $2k$, which acts like a penalty - the price paid for including k variables in the model (this is an interpretation, not why the $2k$ is part of the formula).

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

Criterion: minimize AIC.

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 $2k$, which acts like a penalty - the price paid for including k variables in the model (this is an interpretation, not why the $2k$ is part of the formula).

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

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

Criterion: minimize AIC.

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 $2k$, which acts like a penalty - the price paid for including k variables in the model (this is an interpretation, not why the $2k$ is part of the formula).

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

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

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

Criterion: minimize AIC.

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 $2k$, which acts like a penalty - the price paid for including k variables in the model (this is an interpretation, not why the $2k$ is part of the formula).

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

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

AIC yields a balance between bias and variance, the two sources of information loss.

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

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

Can use this for categorical and numerical variables.

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

Can use this for categorical and numerical variables.

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

- 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

The search algorithm is therefore intelligent and economical.

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

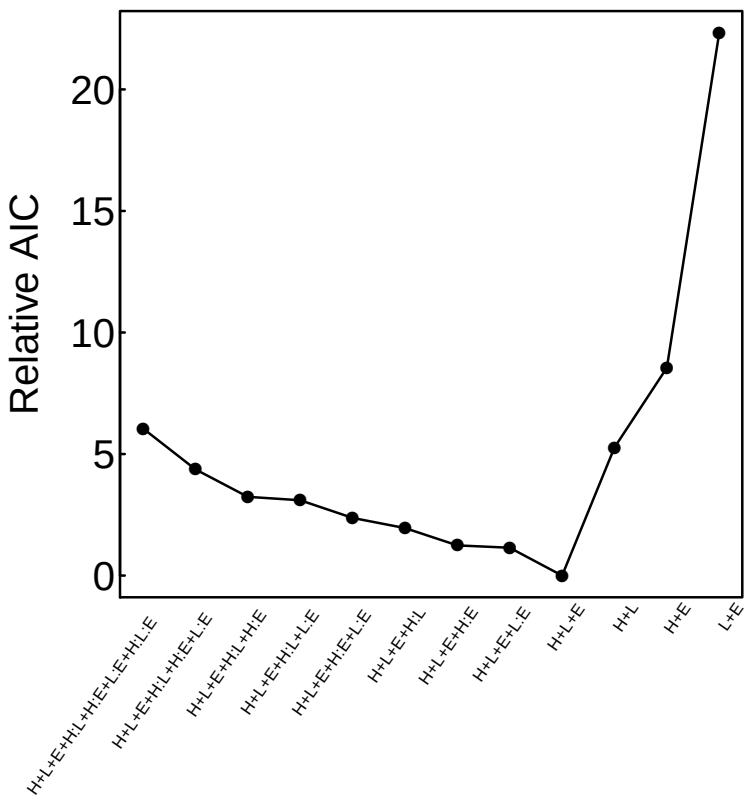
Can use this for categorical and numerical variables.

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

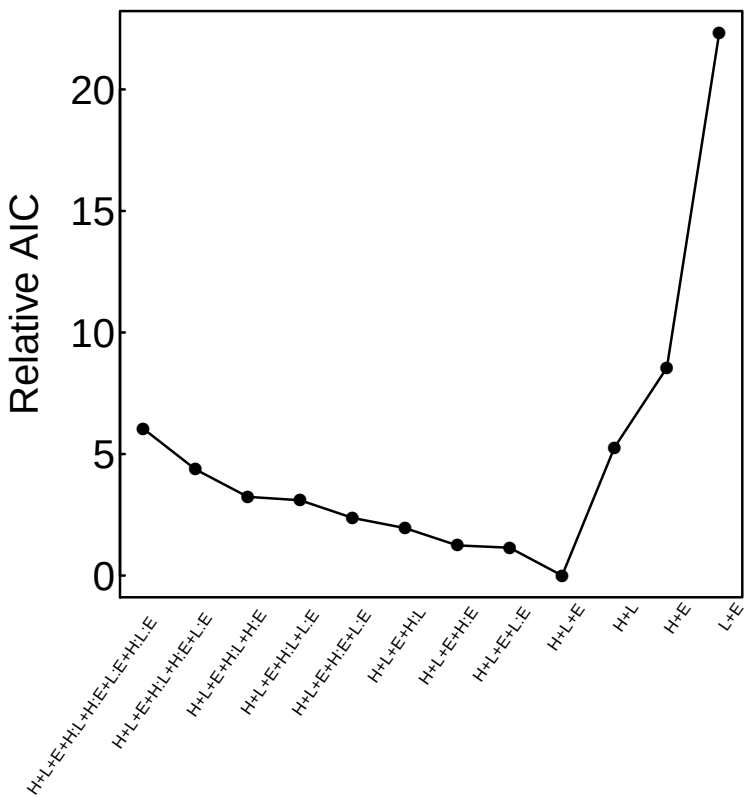
- squared term x^2 is not fitted unless x is also present in the model
- the interaction a:b is not fitted unless both a and b are also present
- a:b:c not fitted unless all two-way interactions of a, b, c, are present

The search algorithm is therefore intelligent and economical.

However, we are still data dredging.

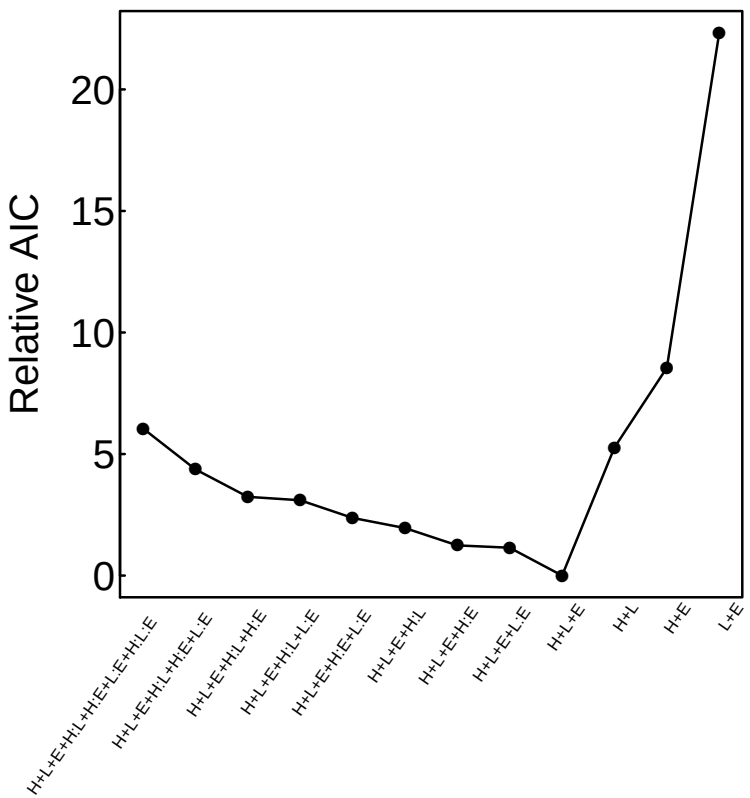


Same data as that analyzed earlier.



Same data as that analyzed earlier.

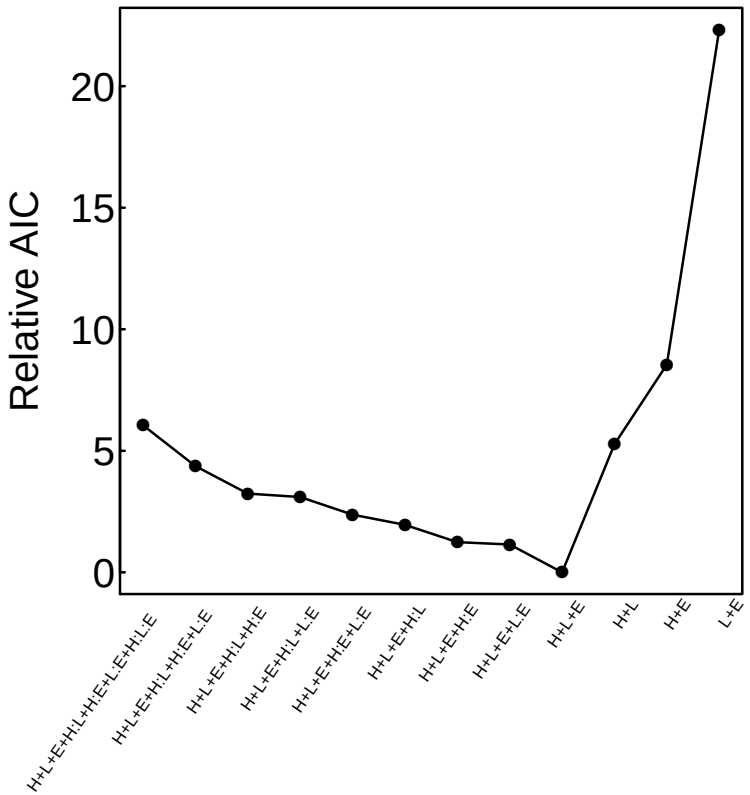
AIC difference (Δ) is the difference between a model's AIC score and that of the "best" model.



Same data as that analyzed earlier.

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

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



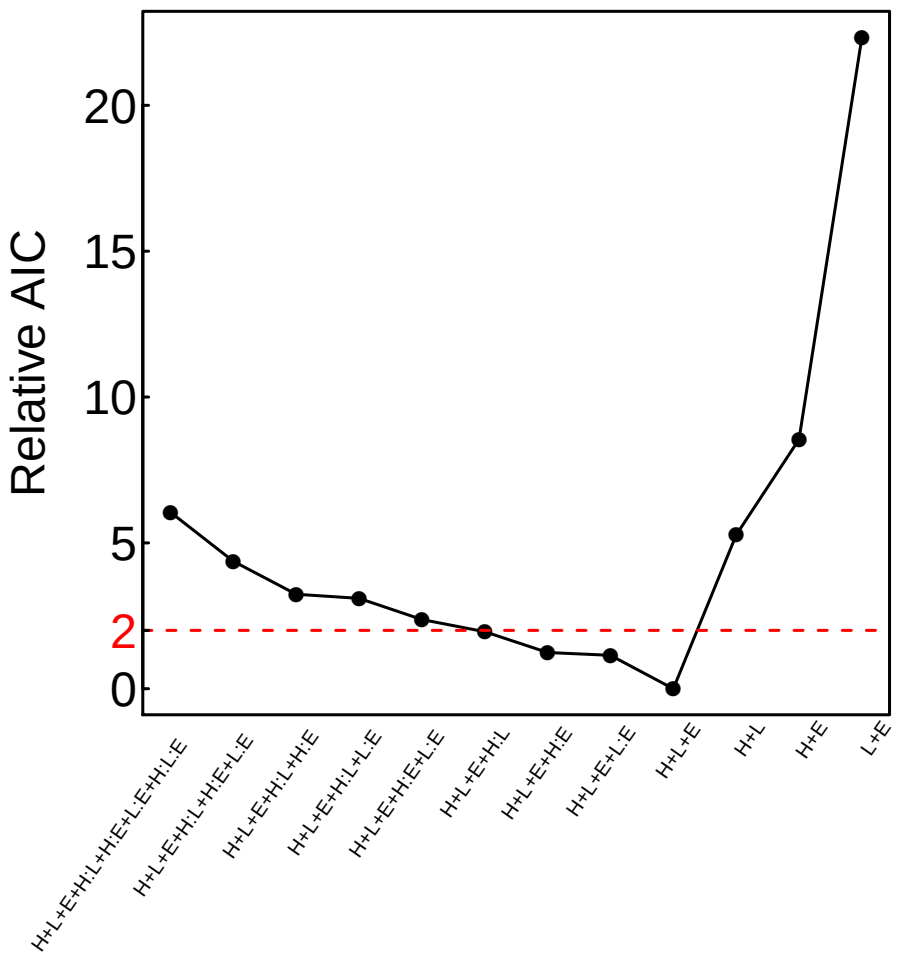
- No hypothesis testing.
- No null model.
- No *P*-value.
- No model is formally 'rejected'.

Several models may be about equally good.

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

Several models may be about equally good.

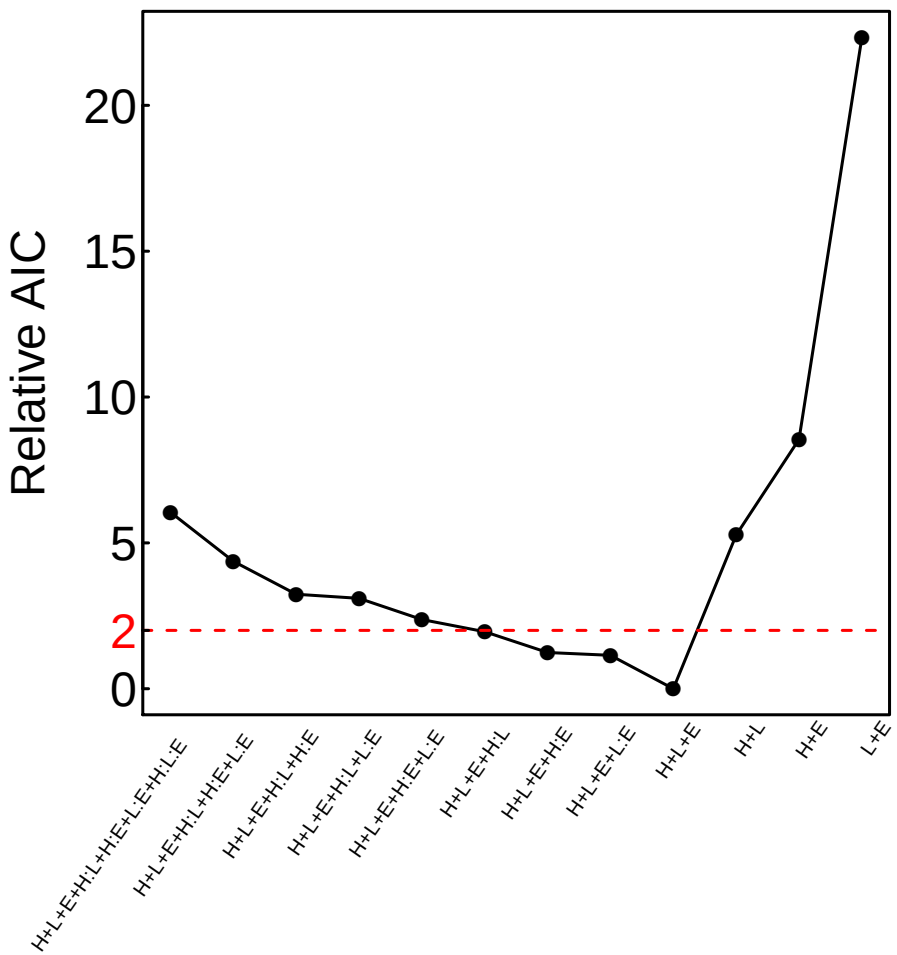
Δ AIC	Support
0–2	Substantial support
4–7	Considerably less support
> 10	Essentially no support



Several models may be about equally good.

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

Your "best" model isn't necessarily the true model.

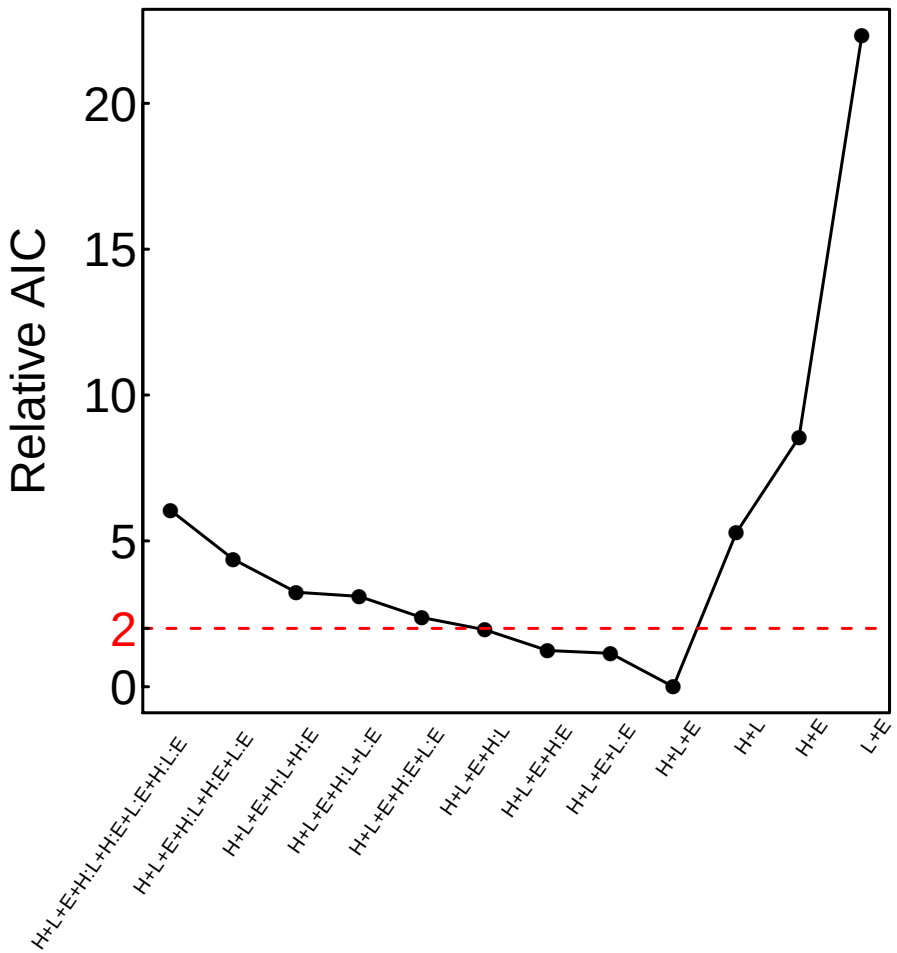


Several models may be about equally good.

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

Your "best" model isn't necessarily the true model.

Remember: AIC balances the bias-variance trade-off. It does a good job to minimize information loss, on average.



Using drop1 to find the 'best' model

```
out1 <- lm(log(nspecies)~H*L*E, data=ants)
drop1(out1, test='F')
```

```
Model:
log(nspecies) ~ H * L * E
       Df Sum of Sq    RSS     AIC F value Pr(>F)
<none>              6.7832 -66.268
H:L:E   1  0.049136 6.8324 -67.951  0.2608 0.6127
```

No support for three-way interaction ($\Delta AIC < 2$).

Using drop1 to find the 'best' model

```
out1 <- lm(log(nspecies)~H*L*E, data=ants)
drop1(out1, test='F')
```

```
Model:
log(nspecies) ~ H * L * E
       Df Sum of Sq    RSS     AIC F value Pr(>F)
<none>              6.7832 -66.268
H:L:E   1  0.049136 6.8324 -67.951  0.2608 0.6127
```

No support for three-way interaction ($\Delta AIC < 2$).

```
out2 <- update(out1, .  -H:L:E)
drop1(out2, test='F')
```

```
Model:
log(nspecies) ~ H + L + E + H:L + H:E + L:E
       Df Sum of Sq    RSS     AIC F value Pr(>F)
<none>              6.8324 -67.951
H:L     1  0.000553 6.8329 -69.947  0.0030 0.9567
H:E     1  0.114279 6.9467 -69.221  0.6189 0.4365
L:E     1  0.136456 6.9688 -69.081  0.7390 0.3955
```

$\Delta AIC < 2$ for each two-way interaction.

Using drop1 to find the 'best' model

```
out1 <- lm(log(nspecies)~H*L*E, data=ants)
drop1(out1, test='F')
```

```
Model:
log(nspecies) ~ H * L * E
       Df Sum of Sq    RSS     AIC F value Pr(>F)
<none>              6.7832 -66.268
H:L:E   1  0.049136 6.8324 -67.951  0.2608 0.6127
```

No support for three-way interaction ($\Delta AIC < 2$).

```
out2 <- update(out1, ~  -H:L:E)
drop1(out2, test='F')
```

```
Model:
log(nspecies) ~ H + L + E + H:L + H:E + L:E
       Df Sum of Sq    RSS     AIC F value Pr(>F)
<none>              6.8324 -67.951
H:L     1  0.000553 6.8329 -69.947  0.0030 0.9567
H:E     1  0.114279 6.9467 -69.221  0.6189 0.4365
L:E     1  0.136456 6.9688 -69.081  0.7390 0.3955
```

$\Delta AIC < 2$ for each two-way interaction.

```
out3 <- update(out2, ~  -H:L -H:E -L:E)
drop1(out3, test='F')
```

```
Model:
log(nspecies) ~ H + L + E
       Df Sum of Sq     RSS     AIC F value      Pr(>F)
<none>              7.0904 -72.320
H       1    5.2353 12.3258 -49.990 29.5348 0.000002939 ***
L       1    1.9210  9.0114 -63.771 10.8371    0.002085 **
E       1    1.2742  8.3646 -67.048  7.1881    0.010610 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Keep all main effects (dropping any main effect increases AIC by > 2).

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.

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.

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.

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.

A good 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

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

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

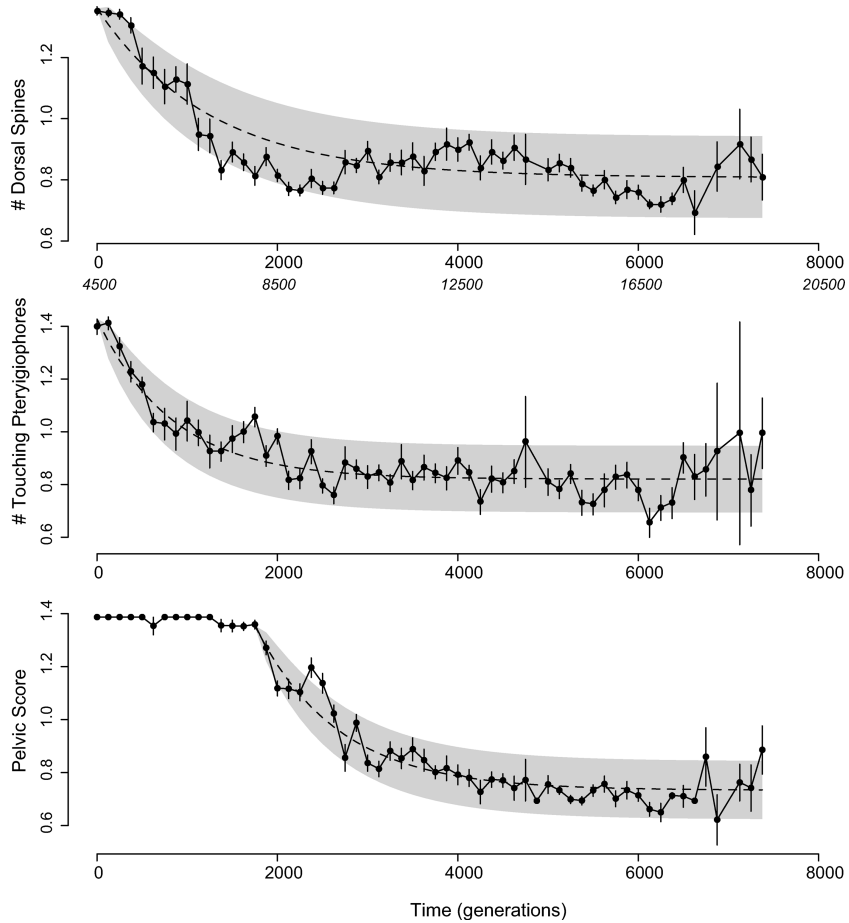
This is where data dredging ends and science begins.

The information-theoretic approach shows its 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.

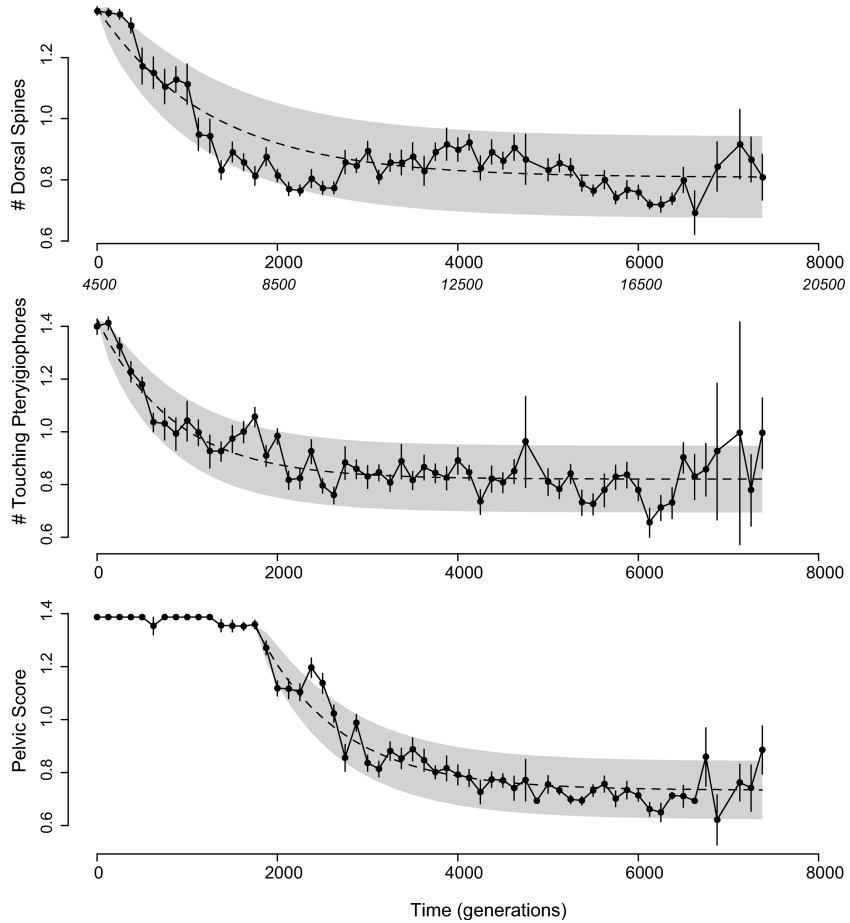
Example: Adaptive evolution in the fossil record



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

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.

Example: Adaptive evolution in the fossil record

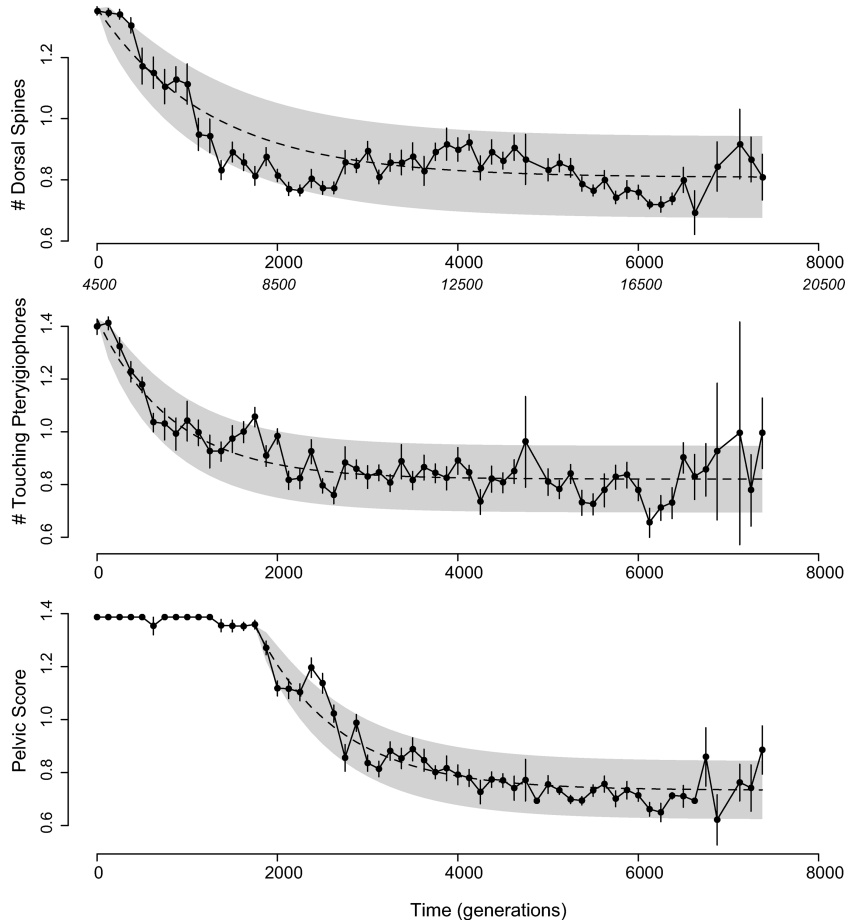


G. Hunt, M. A. Bell & M. P Travis 2008, *Evolution* 62: 700–710.

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.

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

Example: Adaptive evolution in the fossil record



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

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.

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

1 generation = 2 years.

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

1. **Neutral random walk** (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.

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, $P = 0.00005$

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, $P = 0.00005$

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

The adaptive model beats neutral drift for all three traits.

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, $P = 0.00005$

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

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 that sum is 0.95.

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, $P = 0.00005$

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

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 that sum is 0.95.

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).

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, $P = 0.00005$

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

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 that sum is 0.95.

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."

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

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

Using this approach involves giving up on P-values.

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

Using this approach involves giving up on P-values.

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

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

Using this approach involves giving up on P-values.

These information theoretic 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.

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

Using this approach involves giving up on P-values.

These information theoretic 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.