Model Selection and Model Averaging

Michael Noonan

February 7, 2021

Biol 520C: Statistical modelling for biological data

Table of contents

Housekeeping

Housekeeping

- Rubric and time slots for the first talk are posted on Canvas.

Housekeeping

- Rubric and time slots for the first talk are posted on Canvas.
- If you're still struggling with R have a look at Prof. Schluter's R tips page for a similar course run on the Vancouver campus. https://www.zoology.ubc.ca/ schluter/R/

The ΔAIC Grey Zone

Recap

Recap

Last lecture we covered two approaches for model selection:

Recap

Last lecture we covered two approaches for model selection:

1. **Likelihood-ratio tests**

Recap

Last lecture we covered two approaches for model selection:

1. **Likelihood-ratio tests**
   LRTs are quick easy

Recap

Last lecture we covered two approaches for model selection:

1. **Likelihood-ratio tests**
   LRTs are quick easy, well grounded theoretically

Recap

Last lecture we covered two approaches for model selection:

1. **Likelihood-ratio tests**
   LRTs are quick easy, well grounded theoretically, result in clear outcomes

Recap

Last lecture we covered two approaches for model selection:

1. **Likelihood-ratio tests**
   LRTs are quick easy, well grounded theoretically, result in clear outcomes, but become unwieldy for complex models.

Last lecture we covered two approaches for model selection:

1. **Likelihood-ratio tests**
   LRTs are quick easy, well grounded theoretically, result in clear outcomes, but become unwieldy for complex models.

   If some of the models aren't nested in one another, there is no way to compare them using LRTs.

Last lecture we covered two approaches for model selection:

1. **Likelihood-ratio tests**
   LRTs are quick easy, well grounded theoretically, result in clear outcomes, but become unwieldy for complex models.

   If some of the models aren't nested in one another, there is no way to compare them using LRTs.

2. **Information criteria**
   IC don't require nested models

Recap

Last lecture we covered two approaches for model selection:

1. **Likelihood-ratio tests**
   LRTs are quick easy, well grounded theoretically, result in clear outcomes, but become unwieldy for complex models.

   If some of the models aren't nested in one another, there is no way to compare them using LRTs.

2. **Information criteria**
   IC don't require nested models, well grounded theoretically

Recap

Last lecture we covered two approaches for model selection:

1. **Likelihood-ratio tests**
   LRTs are quick easy, well grounded theoretically, result in clear outcomes, but become unwieldy for complex models.

   If some of the models aren't nested in one another, there is no way to compare them using LRTs.

2. **Information criteria**
   IC don't require nested models, well grounded theoretically, easy to apply in practice.

Recap

Last lecture we covered two approaches for model selection:

1. **Likelihood-ratio tests**
   LRTs are quick easy, well grounded theoretically, result in clear outcomes, but become unwieldy for complex models.

   If some of the models aren't nested in one another, there is no way to compare them using LRTs.

2. **Information criteria**
   IC don't require nested models, well grounded theoretically, easy to apply in practice.

   Just because a model is the best fit out of a pool of candidates doesn't mean it's any good.

Recap

Last lecture we covered two approaches for model selection:

1. **Likelihood-ratio tests**
   LRTs are quick easy, well grounded theoretically, result in clear outcomes, but become unwieldy for complex models.

   If some of the models aren't nested in one another, there is no way to compare them using LRTs.

2. **Information criteria**
   IC don't require nested models, well grounded theoretically, easy to apply in practice.

   Just because a model is the best fit out of a pool of candidates doesn't mean it's any good.

   Sometimes IC approaches result in a clear winner, other times...

Clicker question: i

Clicker question: i

```
FIT <- gls(num_sp ~ latitude + elevation, data = data, method = "ML")

dredge(FIT)

Global model call: gls(model = num_sp ~ latitude + elevation, data = data, method = "ML")
---
Model selection table
  (Intrc)     elvtn   lattd df   logLik AICc delta weight
4  11.120 -0.001373 -0.2018  4   -4.059 18.5  0.00  0.922
2   2.489 -0.001613          3   -8.672 24.7  6.21  0.041
3  12.390           -0.2388  3   -8.857 25.0  6.58  0.034
1   2.114                    2  -13.224 31.1 12.61  0.002
Models ranked by AICc(x)
```

Which parameters would you include in your model?

A — latitude and elevation.

B — Only elevation.

C — Only latitude.

D — Only the intercept.

Clicker question: i


```
FIT <- gls(num_sp ~ latitude + elevation, data = data, method = "ML")

dredge(FIT)

Global model call: gls(model = num_sp ~ latitude + elevation, data = data, method = "ML")
---
Model selection table
  (Intrc)     elvtn    lattd df   logLik AICc delta weight
4  11.120 -0.001373 -0.2018  4   -4.059 18.5  0.00  0.922
2   2.489 -0.001613          3   -8.672 24.7  6.21  0.041
3  12.390           -0.2388  3   -8.857 25.0  6.58  0.034
1   2.114                    2 -13.224 31.1 12.61  0.002
Models ranked by AICc(x)
```

Which parameters would you include in your model?

A — latitude and elevation.

B — Only elevation.

C — Only latitude.

D — Only the intercept.

Clicker question: ii

Clicker question: ii

Model selection table

	(Int)	β_1	β_2	β_3	β_4	β_5	β_6	df	logLik	AICc	delta	weight
33	-103.000						0.05326	4	-10370.46	20748.9	0.00	0.202
49	-103.000					+	0.05326	4	-10370.46	20748.9	0.00	0.202
34	-104.200	0.25270					0.05385	5	-10369.64	20749.3	0.36	0.168
50	-104.200	0.25270				+	0.05385	5	-10369.64	20749.3	0.36	0.168
1	4.208							3	-10372.99	20752.0	3.07	0.043
17	4.208					+		3	-10372.99	20752.0	3.07	0.043
2	4.212	-0.05585						4	-10372.23	20752.5	3.55	0.034
18	4.212	-0.05585				+		4	-10372.23	20752.5	3.55	0.034
37	-51.480			-0.028750			0.05640	5	-10372.58	20755.2	6.24	0.009

Which parameters would you include in your model?

A — β_1, β_6, and β_5

B — Only β_6

C — β_6 and β_5

D — A, B, and C are good options.

Clicker question: ii

Model selection table

	(Int)	β_1	β_2	β_3	β_4	β_5	β_6	df	logLik	AICc	delta	weight
33	-103.000						0.05326	4	-10370.46	20748.9	0.00	0.202
49	-103.000					+	0.05326	4	-10370.46	20748.9	0.00	0.202
34	-104.200	0.25270					0.05385	5	-10369.64	20749.3	0.36	0.168
50	-104.200	0.25270				+	0.05385	5	-10369.64	20749.3	0.36	0.168
1	4.208							3	-10372.99	20752.0	3.07	0.043
17	4.208					+		3	-10372.99	20752.0	3.07	0.043
2	4.212	-0.05585						4	-10372.23	20752.5	3.55	0.034
18	4.212	-0.05585				+		4	-10372.23	20752.5	3.55	0.034
37	-51.480			-0.028750			0.05640	5	-10372.58	20755.2	6.24	0.009

Which parameters would you include in your model?

A — β_1, β_6, and β_5

B — Only β_6

C — β_6 and β_5

D — A, B, and C are good options.

No clear winner?

No clear winner?

Sometimes ΔAICc based model selection leads to a clear 'winner'

No clear winner?

Sometimes ΔAICc based model selection leads to a clear 'winner', other times the differences between the top models are miniscule.

No clear winner?

Sometimes ΔAICc based model selection leads to a clear 'winner', other times the differences between the top models are miniscule.

As a modeler, you need to use this information to make some decision. But what do you do?

No clear winner?

Sometimes ΔAICc based model selection leads to a clear ‘winner’, other times the differences between the top models are miniscule.

As a modeler, you need to use this information to make some decision. But what do you do?

As good scientists, we turn to the literature for an answer:

No clear winner?

Sometimes ΔAICc based model selection leads to a clear 'winner', other times the differences between the top models are miniscule.

As a modeler, you need to use this information to make some decision. But what do you do?

As good scientists, we turn to the literature for an answer:

Burnham & Anderson (2002)

1. ΔAIC 0–2: Substantial supp.
2. ΔAIC 4–7: Considerably less supp.
3. ΔAIC >10: Essentially no supp.

No clear winner?

Sometimes ΔAICc based model selection leads to a clear 'winner', other times the differences between the top models are miniscule.

As a modeler, you need to use this information to make some decision. But what do you do?

As good scientists, we turn to the literature for an answer:

Burnham & Anderson (2002)

1. ΔAIC 0–2: Substantial supp.
2. ΔAIC 4–7: Considerably less supp.
3. ΔAIC $>$10: Essentially no supp.

Burnham *et al.* (2011)

1. ΔAIC 0–7: Plausible
2. ΔAIC 7–14: Equivocal
3. ΔAIC $>$14: Implausible

No clear winner?

Sometimes ΔAICc based model selection leads to a clear 'winner', other times the differences between the top models are miniscule.

As a modeler, you need to use this information to make some decision. But what do you do?

As good scientists, we turn to the literature for an answer:

Burnham & Anderson (2002)

1. ΔAIC 0–2: Substantial supp.
2. ΔAIC 4–7: Considerably less supp.
3. ΔAIC >10: Essentially no supp.

Burnham *et al.* (2011)

1. ΔAIC 0–7: Plausible
2. ΔAIC 7–14: Equivocal
3. ΔAIC >14: Implausible

Not very helpful...

ΔAIC of 2

ΔAIC of 2

Because there are no p-values or clear cutoffs tied to ΔAIC values, the literature is confused on what we should do.

ΔAIC of 2

Because there are no *p*-values or clear cutoffs tied to ΔAIC values, the literature is confused on what we should do.

One recurring approach is to use ΔAIC <2 as a cut-off

ΔAIC of 2

Because there are no *p*-values or clear cutoffs tied to ΔAIC values, the literature is confused on what we should do.

One recurring approach is to use ΔAIC <2 as a cut-off, which comes from the recommendations of Burnham & Anderson (2002).

ΔAIC of 2

Because there are no *p*-values or clear cutoffs tied to ΔAIC values, the literature is confused on what we should do.

One recurring approach is to use ΔAIC <2 as a cut-off, which comes from the recommendations of Burnham & Anderson (2002).

Lower AIC values are better (all else being equal), but what's so special about ΔAIC <2?

The ΔAIC $= 2$ Threshold

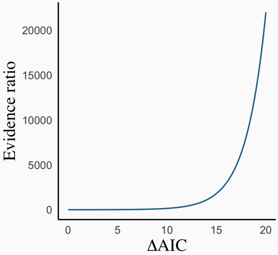
ΔAIC of 2 and the evidence ratio

ΔAIC of 2 and the evidence ratio

The ΔAIC values for any given model are linked to the evidence ratio for the best model as $e^{-(\frac{\Delta \mathrm{AIC}}{2})}$.

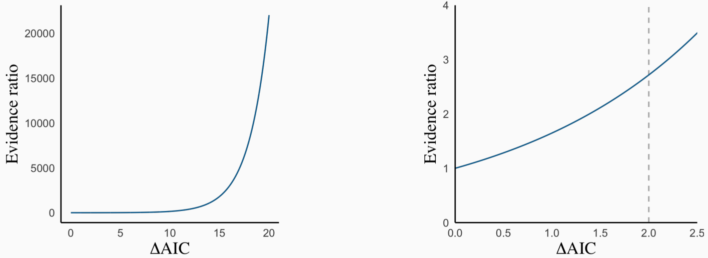
ΔAIC of 2 and the evidence ratio

The ΔAIC values for any given model are linked to the evidence ratio for the best model as $e^{-\left(\frac{\Delta \mathrm{AIC}}{2}\right)}$.



ΔAIC of 2 and the evidence ratio

The ΔAIC values for any given model are linked to the evidence ratio for the best model as $e^{-\left(\frac{\Delta\mathrm{AIC}}{2}\right)}$.



With ΔAIC $\lesssim$ 1.38 the evidence $<$ 2, with ΔAIC $\lesssim$ 2, evidence $<$ 3.

So in the ΔAIC $<$ 2 regime models are only $\sim$twice as likely as the AIC 'best' model.

ΔAIC of 2 and AIC

ΔAIC of 2 and AIC

The evidence ratio gives us a feel for why this range of values might be important, but it's still a bit murky.

ΔAIC of 2 and AIC

The evidence ratio gives us a feel for why this range of values might be important, but it's still a bit murky.

The equation for AIC is:

$$\mathrm{AIC} = -2\ln(\mathcal{L}) + 2K$$

ΔAIC of 2 and AIC

The evidence ratio gives us a feel for why this range of values might be important, but it's still a bit murky.

The equation for AIC is:

$$\mathrm{AIC} = -2\ln(\mathcal{L}) + 2K$$

The equation for ΔAIC is:

$$\Delta\mathrm{AIC} = -2\log\mathcal{L}(\theta_1) + 2\log\mathcal{L}(\theta_2) - 2(K_2 - K_1)$$

ΔAIC of 2 and AIC

The evidence ratio gives us a feel for why this range of values might be important, but it's still a bit murky.

The equation for AIC is:

$$\mathrm{AIC} = -2\ln(\mathcal{L}) + 2K$$

The equation for ΔAIC is:

$$\Delta\mathrm{AIC} = -2\log\mathcal{L}(\theta_1) + 2\log\mathcal{L}(\theta_2) - 2(K_2 - K_1)$$

What happens if two models have the $\sim$ the same likelihood and only differ by one parameter?

ΔAIC of 2 and AIC cont.

ΔAIC of 2 and AIC cont.

If $\mathcal{L}(\theta_1) \approx \mathcal{L}(\theta_2)$ and $K_1 = K_2 + 1$

ΔAIC of 2 and AIC cont.

If $\mathcal{L}(\theta_1) \approx \mathcal{L}(\theta_2)$ and $K_1 = K_2 + 1$

$$\Delta \mathrm{AIC} = -2 \log \mathcal{L}(\theta_1) + 2 \log \mathcal{L}(\theta_2) - 2(K_2 - K_1)$$

ΔAIC of 2 and AIC cont.

If $\mathcal{L}(\theta_1) \approx \mathcal{L}(\theta_2)$ and $K_1 = K_2 + 1$

$$\Delta\mathrm{AIC} = -2\log\mathcal{L}(\theta_1) + 2\log\mathcal{L}(\theta_2) - 2(K_2 - K_1)$$

$$\Delta\mathrm{AIC} = -2\log\mathcal{L}(\theta_1) + 2\log\mathcal{L}(\theta_1) - 2(K_2 - (K_2 + 1))$$

ΔAIC of 2 and AIC cont.

If $\mathcal{L}(\theta_1) \approx \mathcal{L}(\theta_2)$ and $K_1 = K_2 + 1$

$$\Delta\mathrm{AIC} = -2\log\mathcal{L}(\theta_1) + 2\log\mathcal{L}(\theta_2) - 2(K_2 - K_1)$$

$$\Delta\mathrm{AIC} = -2\log\mathcal{L}(\theta_1) + 2\log\mathcal{L}(\theta_1) - 2(K_2 - (K_2 + 1))$$

$$\Delta\mathrm{AIC} = \cancel{-2\log\mathcal{L}(\theta_1) + 2\log\mathcal{L}(\theta_1)}$$

ΔAIC of 2 and AIC cont.

If $\mathcal{L}(\theta_1) \approx \mathcal{L}(\theta_2)$ and $K_1 = K_2 + 1$

$$\Delta\mathrm{AIC} = -2\log\mathcal{L}(\theta_1) + 2\log\mathcal{L}(\theta_2) - 2(K_2 - K_1)$$

$$\Delta\mathrm{AIC} = -2\log\mathcal{L}(\theta_1) + 2\log\mathcal{L}(\theta_1) - 2(K_2 - (K_2 + 1))$$

$$\Delta\mathrm{AIC} = \cancel{-2\log\mathcal{L}(\theta_1) + 2\log\mathcal{L}(\theta_1)} - 2(\cancel{K_2} - \cancel{K_2} - 1)$$

ΔAIC of 2 and AIC cont.

If $\mathcal{L}(\theta_1) \approx \mathcal{L}(\theta_2)$ and $K_1 = K_2 + 1$

$$\Delta \text{AIC} = -2 \log \mathcal{L}(\theta_1) + 2 \log \mathcal{L}(\theta_2) - 2(K_2 - K_1)$$

$$\Delta \text{AIC} = -2 \log \mathcal{L}(\theta_1) + 2 \log \mathcal{L}(\theta_1) - 2(K_2 - (K_2 + 1))$$

$$\Delta \text{AIC} = \cancel{-2 \log \mathcal{L}(\theta_1) + 2 \log \mathcal{L}(\theta_1)} - 2(\cancel{K_2} - \cancel{K_2} - 1)$$

$$\Delta \text{AIC} = -2(-1)$$

ΔAIC of 2 and AIC cont.

If $\mathcal{L}(\theta_1) \approx \mathcal{L}(\theta_2)$ and $K_1 = K_2 + 1$

$$\Delta\mathrm{AIC} = -2\log\mathcal{L}(\theta_1) + 2\log\mathcal{L}(\theta_2) - 2(K_2 - K_1)$$

$$\Delta\mathrm{AIC} = -2\log\mathcal{L}(\theta_1) + 2\log\mathcal{L}(\theta_1) - 2(K_2 - (K_2 + 1))$$

$$\Delta\mathrm{AIC} = \cancel{-2\log\mathcal{L}(\theta_1) + 2\log\mathcal{L}(\theta_1)} - 2(\cancel{K_2} - \cancel{K_2} - 1)$$

$$\Delta\mathrm{AIC} = -2(-1) = 2$$

ΔAIC of 2 and AIC cont.

If $\mathcal{L}(\theta_1) \approx \mathcal{L}(\theta_2)$ and $K_1 = K_2 + 1$

$$\Delta\mathrm{AIC} = -2\log\mathcal{L}(\theta_1) + 2\log\mathcal{L}(\theta_2) - 2(K_2 - K_1)$$

$$\Delta\mathrm{AIC} = -2\log\mathcal{L}(\theta_1) + 2\log\mathcal{L}(\theta_1) - 2(K_2 - (K_2 + 1))$$

$$\Delta\mathrm{AIC} = \cancel{-2\log\mathcal{L}(\theta_1) + 2\log\mathcal{L}(\theta_1)} - 2(\cancel{K_2} - \cancel{K_2} - 1)$$

$$\Delta\mathrm{AIC} = -2(-1) = 2$$

Because the penalty term is 2K, models with the $\sim$ the same likelihood that only differ by one parameter will, by definition, have a $\Delta\mathrm{AIC}$ of 2.

ΔAIC of 2 and AIC[2]

So this sheds some more light on the ΔAIC of 2 threshold

ΔAIC of 2 and AIC[2]

So this sheds some more light on the $\Delta\mathrm{AIC}$ of 2 threshold and explains why in practice you'll find models that differ from the 'best fit' model by only one parameter within the $\Delta\mathrm{AIC}$ of 2 threshold.

So this sheds some more light on the $\Delta\mathrm{AIC}$ of 2 threshold and explains why in practice you'll find models that differ from the 'best fit' model by only one parameter within the $\Delta\mathrm{AIC}$ of 2 threshold.

Var 1	Var 2	Var 3	Var 4	Var 5	AIC	ΔAIC	Weight
	-0.539 (± 0.244)		**-0.602** (± 0.190)		1789.73	0.00	0.26
	-0.674 (± 0.336)		**-0.609** (± 0.192)	0.173 (± 0.295)	1791.38	1.65	0.12
	-0.544 (± 0.245)	0.003 (± 0.008)	**-0.566** (± 0.214)		1791.60	1.86	0.10
-0.090 (± 0.333)	**-0.541** (± 0.244)		**-0.574** (± 0.217)		1791.66	1.93	0.10
			-0.641 (± 0.201)		1792.23	2.50	0.08
-0.070 (± 0.335)	**-0.670** (± 0.336)		**-0.586** (± 0.220)	0.167 (± 0.296)	1793.34	3.61	0.04
			-0.622 (± 0.198)	-0.212 (± 0.222)	1793.34	3.61	0.04
	-0.662 (± 0.344)	0.001 (± 0.008)	**-0.591** (± -0.591)	0.155 (± 0.316)	1793.35	3.62	0.04

Source: Mark Brewer

Second best model is $\mathcal{M}_1 + \mathrm{Var}_5$, third best model is $\mathcal{M}_1 + \mathrm{Var}_3$, fourth best is $\mathcal{M}_1 + \mathrm{Var}_1$, all within the $\Delta\mathrm{AIC}$ of 2 threshold.

ΔAIC of 2 and p-values

ΔAIC of 2 and *p*-values

The likelihood ratio test allows for estimating *p*-values for the support between a pair of models.

ΔAIC of 2 and p-values

The likelihood ratio test allows for estimating p-values for the support between a pair of models.

$$P = \Pr(\chi^2_k > \lambda)$$

ΔAIC of 2 and *p*-values

The likelihood ratio test allows for estimating *p*-values for the support between a pair of models.

$$P = \Pr(\chi^2_k > \lambda)$$

$$\Delta \mathrm{AIC} = \lambda - 2(K_2 - K_1)$$

ΔAIC of 2 and *p*-values

The likelihood ratio test allows for estimating *p*-values for the support between a pair of models.

$$P = \Pr(\chi^2_k > \lambda)$$

$$\Delta \mathrm{AIC} = \lambda - 2(K_2 - K_1)$$

$$P = \Pr(\chi^2_k > \Delta \mathrm{AIC} + 2(K_2 - K_1))$$

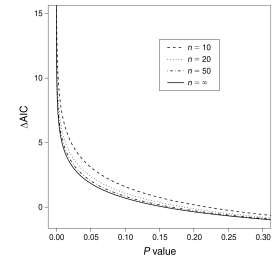
ΔAIC of 2 and *p*-values

The likelihood ratio test allows for estimating *p*-values for the support between a pair of models.

$$P = \Pr(\chi^2_k > \lambda)$$

$$\Delta\mathrm{AIC} = \lambda - 2(K_2 - K_1)$$

$$P = \Pr(\chi^2_k > \Delta\mathrm{AIC} + 2(K_2 - K_1))$$



(Murtaugh, 2014)

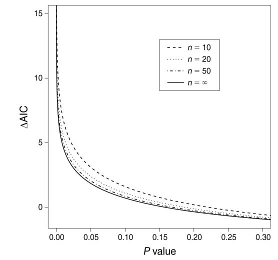
ΔAIC of 2 and *p*-values

The likelihood ratio test allows for estimating *p*-values for the support between a pair of models.

$$P = \Pr(\chi^2_k > \lambda)$$

$$\Delta \mathrm{AIC} = \lambda - 2(K_2 - K_1)$$

$$P = \Pr(\chi^2_k > \Delta \mathrm{AIC} + 2(K_2 - K_1))$$



(Murtaugh, 2014)

There's a one-to-one relationship between ΔAIC and *p*-values e.g., a ΔAIC of 2 when models differ by 1 parameter corresponds to a p-value of $\sim$0.047

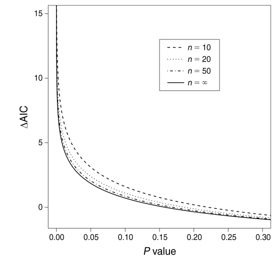
ΔAIC of 2 and p-values

The likelihood ratio test allows for estimating p-values for the support between a pair of models.

$$P = \Pr(\chi^2_k > \lambda)$$

$$\Delta\mathrm{AIC} = \lambda - 2(K_2 - K_1)$$

$$P = \Pr(\chi^2_k > \Delta\mathrm{AIC} + 2(K_2 - K_1))$$



(Murtaugh, 2014)

There's a one-to-one relationship between ΔAIC and p-values e.g., a ΔAIC of 2 when models differ by 1 parameter corresponds to a p-value of $\sim$0.047 (but a p-value of $\sim$0.015 when they differ by 10 parameters).

ΔAIC of 2 and p-values

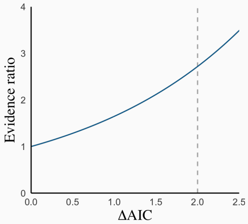
ΔAIC of 2 and p-values

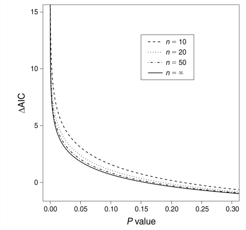
Putting the pieces together:

ΔAIC of 2 and *p*-values

Putting the pieces together:

ΔAIC $= 2$ corresponds to an evidence ratio of $\sim$2.7





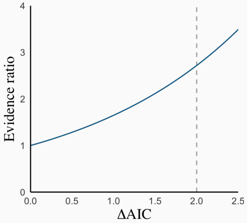
(Murtaugh, 2014)

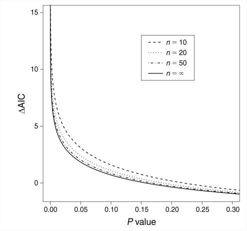
ΔAIC of 2 and p-values

Putting the pieces together:

$\Delta\text{AIC} = 2$ corresponds to an evidence ratio of $\sim$2.7

When two models have identical likelihoods, they can not have $\Delta\text{AIC} < 2$ (caps our capacity to distinguish models).





(Murtaugh, 2014)

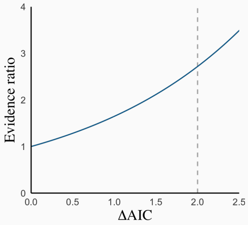
ΔAIC of 2 and *p*-values

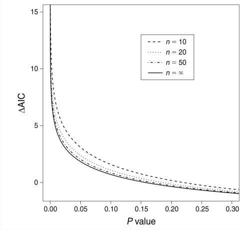
Putting the pieces together:

ΔAIC = 2 corresponds to an evidence ratio of ∼2.7

When two models have identical likelihoods, they can not have ΔAIC <2 (caps our capacity to distinguish models).

A ΔAIC of 2 with a differences of 1 parameter corresponds to a *p*-value of ∼0.047, meaning the complexity is a significant improvement.





(Murtaugh, 2014)

ΔAIC of 2

ΔAIC of 2

Now we we have a better idea of why ΔAIC <2 is pervasive in the literature.

ΔAIC of 2

Now we we have a better idea of why ΔAIC <2 is pervasive in the literature.

But are we any closer to knowing what to do when we have a number of top contenders (i.e., ΔAIC <2)?

AIC Overfitting

AIC Overfitting

Before we decide what to do when multiple models are reasonable contenders we need to explore how AIC model selection performs in practice.

AIC Overfitting

Before we decide what to do when multiple models are reasonable contenders we need to explore how AIC model selection performs in practice.

It's well known in the statistics literature that AIC has a tendency to overfit

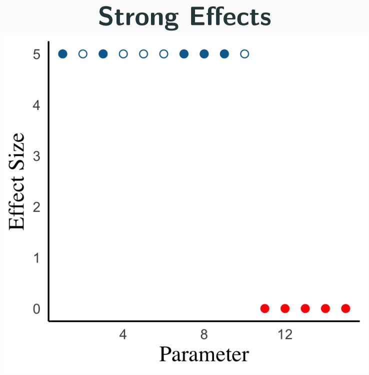
AIC Overfitting

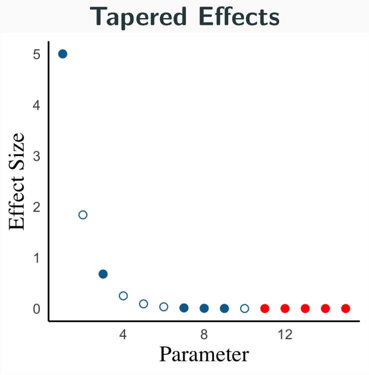
Before we decide what to do when multiple models are reasonable contenders we need to explore how AIC model selection performs in practice.

It's well known in the statistics literature that AIC has a tendency to overfit, but what does that look like?

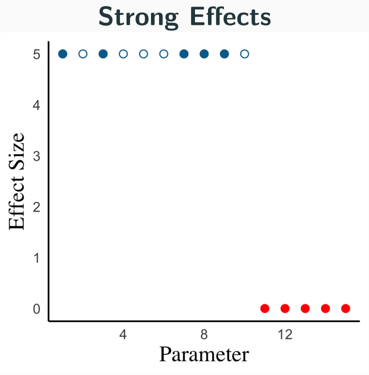
Simulation Study

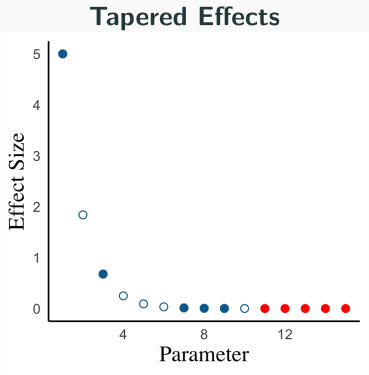
Simulation Study



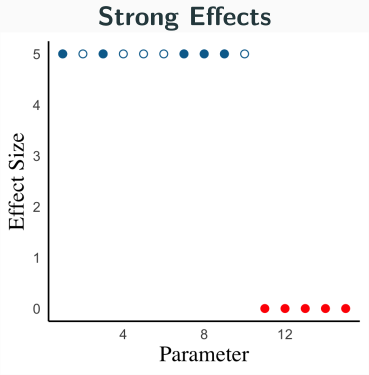


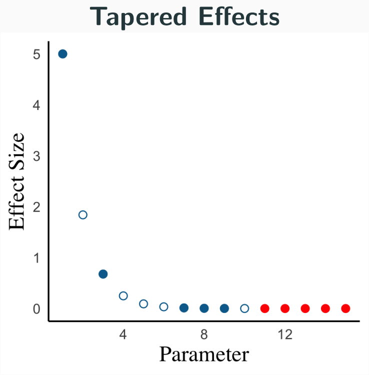
Simulation Study





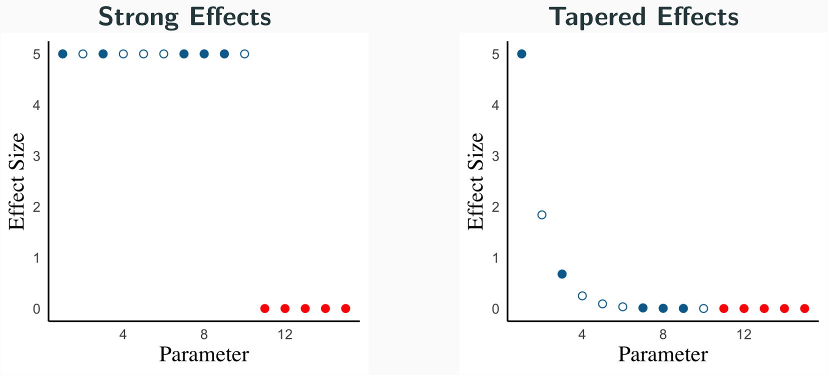
Randomly used 5 of them





Randomly used 5 of them, and added 5 other ‘noise’ parameters.

Simulation Study

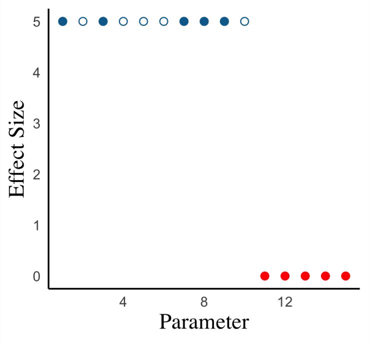


Randomly used 5 of them, and added 5 other ‘noise’ parameters.

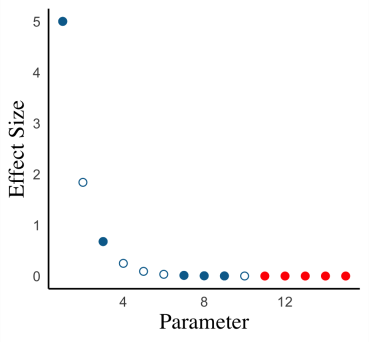
Fit models

Simulation Study

Strong Effects



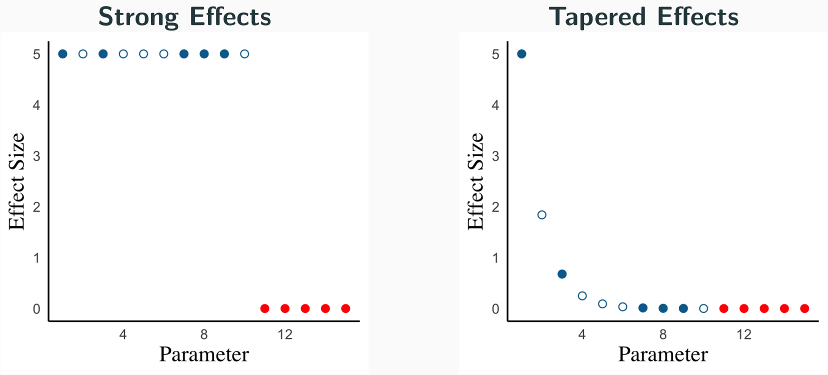
Tapered Effects



Randomly used 5 of them, and added 5 other ‘noise’ parameters.

Fit models, selected the best by AIC and AICc

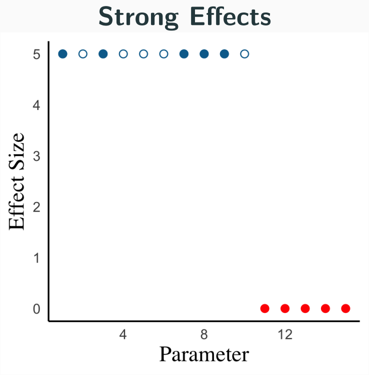
Simulation Study

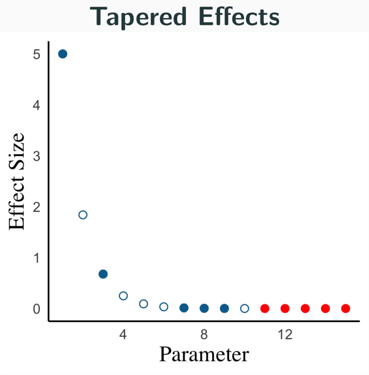


Randomly used 5 of them, and added 5 other 'noise' parameters.

Fit models, selected the best by AIC and AICc, compared which parameters were selected based on the true system

Simulation Study



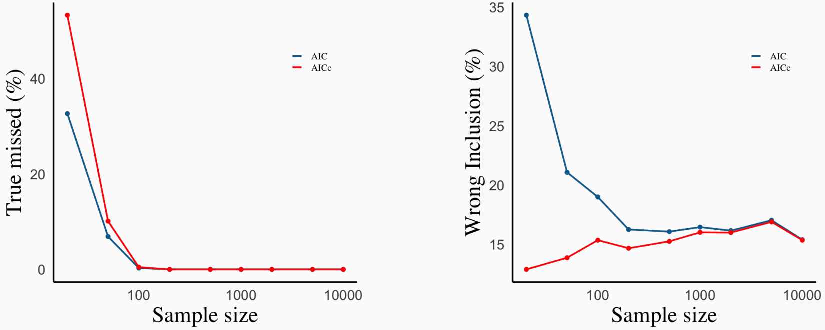


Randomly used 5 of them, and added 5 other 'noise' parameters.

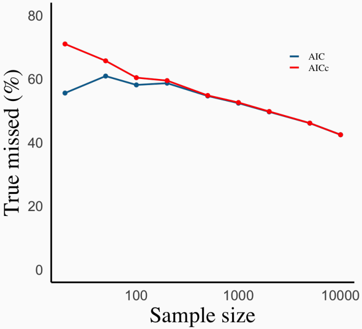
Fit models, selected the best by AIC and AICc, compared which parameters were selected based on the true system, repeated this 1000s of times

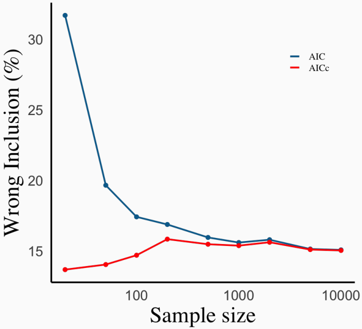
Strong Effects Results

Strong Effects Results



Tapered Effects Results

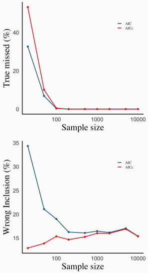




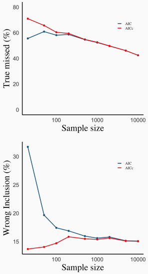
AIC/AICc performance

THE UNIVERSITY OF BRITISH COLUMBIA
Okanagan Campus

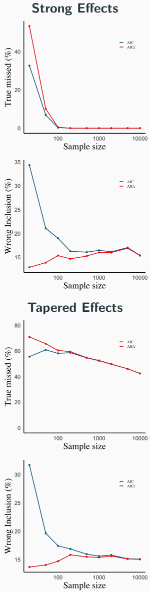
Strong Effects



Tapered Effects



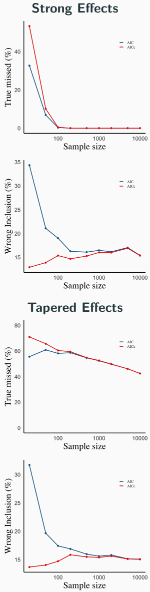
For systems with strong effect sizes, both AIC and AICc identified true parameters well for high *n*.



AIC/AICc performance

For systems with strong effect sizes, both AIC and AICc identified true parameters well for high *n*.

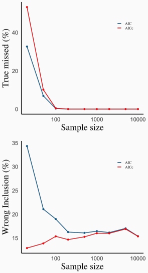
For systems with tapered effect sizes, neither AIC nor AICc identified all of the true parameters well



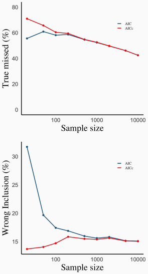
For systems with strong effect sizes, both AIC and AICc identified true parameters well for high n.

For systems with tapered effect sizes, neither AIC nor AICc identified all of the true parameters well even with n was extremely high.

Strong Effects



Tapered Effects

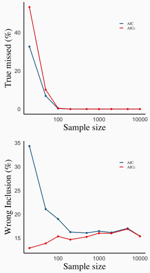


For systems with strong effect sizes, both AIC and AICc identified true parameters well for high n.

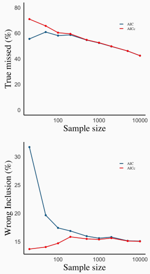
For systems with tapered effect sizes, neither AIC nor AICc identified all of the true parameters well even with n was extremely high.

When n was small, AIC missed fewer true parameters than AICc, but at the cost of more false positives.

Strong Effects



Tapered Effects



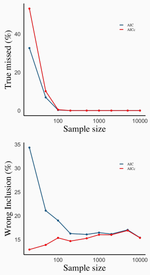
For systems with strong effect sizes, both AIC and AICc identified true parameters well for high *n*.

For systems with tapered effect sizes, neither AIC nor AICc identified all of the true parameters well even with *n* was extremely high.

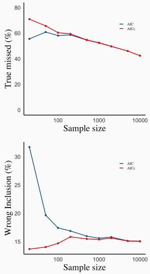
When *n* was small, AIC missed fewer true parameters than AICc, but at the cost of more false positives.

For both systems AIC and AICc consistently identified noise parameters as being important.

Strong Effects



Tapered Effects



AIC/AICc performance

AIC/AICc performance

In practice, AIC/AICc tend to overfit (pick too many parameters)

AIC/AICc performance

In practice, AIC/AICc tend to overfit (pick too many parameters), but when effect sizes are small they can also miss key parameters.

AIC/AICc performance

In practice, AIC/AICc tend to overfit (pick too many parameters), but when effect sizes are small they can also miss key parameters.

When ΔAIC/AICc values are small our ability to distinguish between models gets murky.

AIC/AICc performance

In practice, AIC/AICc tend to overfit (pick too many parameters), but when effect sizes are small they can also miss key parameters.

When ΔAIC/AICc values are small our ability to distinguish between models gets murky.

This leaves us with a number of different options:

AIC/AICc performance

In practice, AIC/AICc tend to overfit (pick too many parameters), but when effect sizes are small they can also miss key parameters.

When ΔAIC/AICc values are small our ability to distinguish between models gets murky.

This leaves us with a number of different options:

1. Be cautious and pick the most parsimonious model.

AIC/AICc performance

In practice, AIC/AICc tend to overfit (pick too many parameters), but when effect sizes are small they can also miss key parameters.

When ΔAIC/AICc values are small our ability to distinguish between models gets murky.

This leaves us with a number of different options:

1. Be cautious and pick the most parsimonious model.

2. Be liberal and include all contending parameters in the final model.

AIC/AICc performance

In practice, AIC/AICc tend to overfit (pick too many parameters), but when effect sizes are small they can also miss key parameters.

When ΔAIC/AICc values are small our ability to distinguish between models gets murky.

This leaves us with a number of different options:

1. Be cautious and pick the most parsimonious model.
2. Be liberal and include all contending parameters in the final model.
3. Conduct likelihood-ratio tests on the top models.

AIC/AICc performance

In practice, AIC/AICc tend to overfit (pick too many parameters), but when effect sizes are small they can also miss key parameters.

When ΔAIC/AICc values are small our ability to distinguish between models gets murky.

This leaves us with a number of different options:

1. Be cautious and pick the most parsimonious model.
2. Be liberal and include all contending parameters in the final model.
3. Conduct likelihood-ratio tests on the top models.
4. Perform model averaging.

Model Averaging

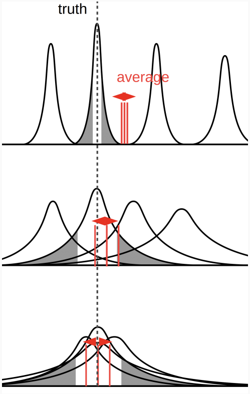
Model averaging

Model averaging

Model averaging refers to the practice of using several models at once for making predictions or inferring parameters.

Model averaging refers to the practice of using several models at once for making predictions or inferring parameters.

The Idea: If a model is misspecified, the parameters estimates may be too high/low (e.g., noise parameters soaking up effects or some effects getting inflated to make up for missing parameters).



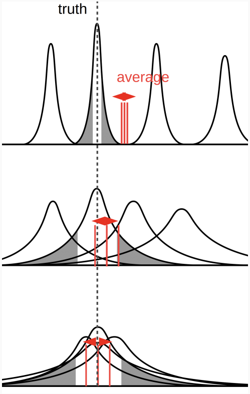
(Dormann *et al.*, 2018)

Model averaging

Model averaging refers to the practice of using several models at once for making predictions or inferring parameters.

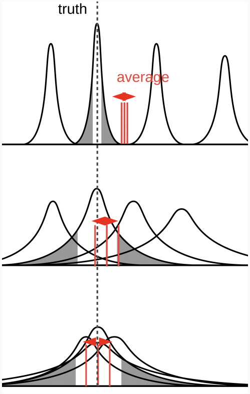
The Idea: If a model is misspecified, the parameters estimates may be too high/low (e.g., noise parameters soaking up effects or some effects getting inflated to make up for missing parameters).

Averaging parameter values from different models, with biases in either way, should cancel out and reduce bias in the average.



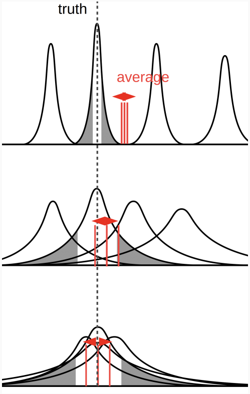
(Dormann *et al.*, 2018)

Model averaging cont.



(Dormann *et al.*, 2018)

Variance between models is also important!
(Dormann *et al.*, 2018)

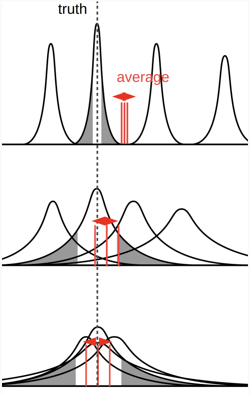


(Dormann *et al.*, 2018)

Model averaging cont.

Variance between models is also important! (Dormann *et al.*, 2018)

If all models make identical, and wrong predictions, this would cancel any benefit of averaging.

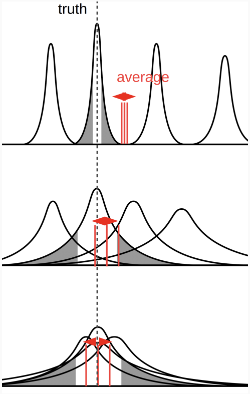


(Dormann *et al.*, 2018)

Variance between models is also important!
(Dormann *et al.*, 2018)

If all models make identical, and wrong predictions, this would cancel any benefit of averaging.

If, however, model predictions are incorrect but vary evenly around the truth, there are substantial benefits to averaging.



(Dormann *et al.*, 2018)

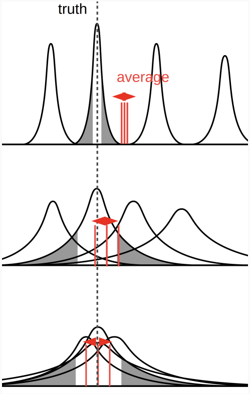
Model averaging cont.

Variance between models is also important! (Dormann *et al.*, 2018)

If all models make identical, and wrong predictions, this would cancel any benefit of averaging.

If, however, model predictions are incorrect but vary evenly around the truth, there are substantial benefits to averaging.

In other words, you only then I get the full benefits of model averaging when models have very different parameter estimates.



(Dormann *et al.*, 2018)

Challenge: If we throw all models we can possible think of into an averaging procedure, but only a few are actually reasonable, then the junk can ruin the averaged model.

Model weighting

Challenge: If we throw all models we can possible think of into an averaging procedure, but only a few are actually reasonable, then the junk can ruin the averaged model.

One solution you see in many papers is to only average plausible models (e.g., only models within Δ AICc of 2).

Model weighting

Challenge: If we throw all models we can possible think of into an averaging procedure, but only a few are actually reasonable, then the junk can ruin the averaged model.

One solution you see in many papers is to only average plausible models (e.g., only models within Δ AICc of 2).

Another solution is to somehow weight all of the different models based on their 'plausibility'.

Model weighting

Challenge: If we throw all models we can possible think of into an averaging procedure, but only a few are actually reasonable, then the junk can ruin the averaged model.

One solution you see in many papers is to only average plausible models (e.g., only models within Δ AICc of 2).

Another solution is to somehow weight all of the different models based on their 'plausibility'.

A third solution is to take a hybrid approach of averaging a reduced pool of unequally weighted models.

Model weighting

Challenge: If we throw all models we can possible think of into an averaging procedure, but only a few are actually reasonable, then the junk can ruin the averaged model.

One solution you see in many papers is to only average plausible models (e.g., only models within Δ AICc of 2).

Another solution is to somehow weight all of the different models based on their ‘plausibility’.

A third solution is to take a hybrid approach of averaging a reduced pool of unequally weighted models.

The question is how do we assign model weights?

Model weights

Model weights

In ecology, model averaging is dominated by the IC framework popularised by Burnham & Anderson (2002).

Model weights

In ecology, model averaging is dominated by the IC framework popularised by Burnham & Anderson (2002).

We saw earlier that model likelihoods provide a formal measure of evidence for each of the models in the set:

$$\ell_i = e^{\left(-\frac{\Delta \mathrm{AIC}_i}{2}\right)}$$

Model weights

In ecology, model averaging is dominated by the IC framework popularised by Burnham & Anderson (2002).

We saw earlier that model likelihoods provide a formal measure of evidence for each of the models in the set:

$$\ell_i = e^{(-\frac{\Delta \mathrm{AIC}_i}{2})}$$

The 'probability' of each model, $\mathcal{M}_i$, given the data and the N possible models, can be computed as a measure of strength of evidence (Burnham *et al.*, 2011):

$$w_i = \Pr(\mathcal{M}_i|\mathrm{data}) = \frac{\ell_i}{\sum_{i=1}^{N} \ell_i}$$

Model weights

In ecology, model averaging is dominated by the IC framework popularised by Burnham & Anderson (2002).

We saw earlier that model likelihoods provide a formal measure of evidence for each of the models in the set:

$$\ell_i = e^{\left(-\frac{\Delta \mathrm{AIC}_i}{2}\right)}$$

The 'probability' of each model, $\mathcal{M}_i$, given the data and the N possible models, can be computed as a measure of strength of evidence (Burnham *et al.*, 2011):

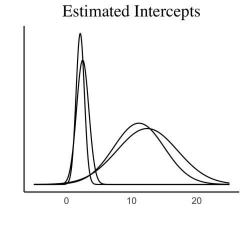
$$w_i = \Pr(\mathcal{M}_i|\mathrm{data}) = \frac{\ell_i}{\sum_{i=1}^{N} \ell_i} \qquad \text{where} \qquad \sum_{i=1}^{N} \ell_i = 1$$

Model averaging in R

Model averaging in R


```
library(nlme)
library(MuMIn)

data <- read.csv("Ant_Richness.csv")

FIT <- gls(num_sp ~ latitude + elevation, data = data, method = "ML", na.action = na.fail)

FITS <- dredge(FIT)

FITS

Global model call: gls(model = num_sp ~ latitude + elevation, data = data, method = "ML",
    na.action = na.fail)
```

```
---
Model selection table
  (Intrc)     elvtn   lattd df  logLik AICc delta weight
4  11.120 -0.001373 -0.2018  4  -4.059 18.5  0.00  0.922
2   2.489 -0.001613          3  -8.672 24.7  6.21  0.041
3  12.390           -0.2388  3  -8.857 25.0  6.58  0.034
1   2.114                    2 -13.224 31.1 12.61  0.002
Models ranked by AICc(x)
```

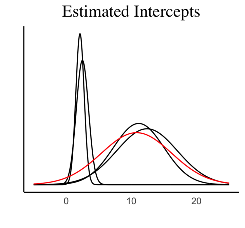


Model averaging in R cont.


```
AVG.FIT <- model.avg(FITS)
```

Model averaging in R cont.


```
AVG.FIT <- model.avg(FITS)

summary(AVG.FIT)

Component models:
       df  logLik   AICc  delta  weight
12      4   -4.06  18.47   0.00    0.92
1       3   -8.67  24.68   6.21    0.04
2       3   -8.86  25.05   6.58    0.03
(Null)  2  -13.22  31.08  12.61    0.00

Term codes:
elevation   latitude
        1          2

Model-averaged coefficients:
(full average)
              Estimate Std. Error Adjusted SE z value Pr(>|z|)
(Intercept) 10.7869966  3.2388986   3.3934619   3.179  0.00148 **
elevation   -0.0013333  0.0004967   0.0005211   2.558  0.01051 *
latitude    -0.1943997  0.0758035   0.0794094   2.448  0.01436 *

(conditional average)
              Estimate Std. Error Adjusted SE z value Pr(>|z|)
(Intercept) 10.7869966  3.2388986   3.3934619   3.179  0.00148 **
elevation   -0.0013832  0.0004323   0.0004612   2.999  0.00271 **
latitude    -0.2031600  0.0650028   0.0693561   2.929  0.00340 **
```



Benefits of model averaging

IC based model averaging via `MuMIn` is one of the many ways you can cary out model averaging

Benefits of model averaging

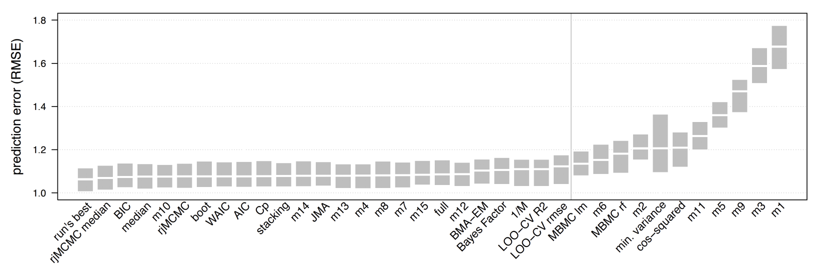
IC based model averaging via `MuMIn` is one of the many ways you can cary out model averaging, but is any one method better than any other?

Benefits of model averaging

IC based model averaging via `MuMIn` is one of the many ways you can cary out model averaging, but is any one method better than any other? Also is model averaging better than just picking a model?

Benefits of model averaging

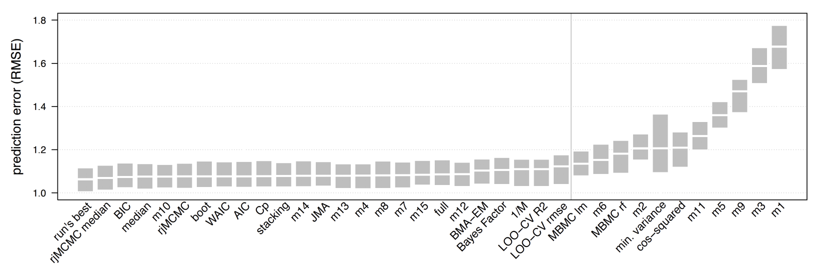
IC based model averaging via `MuMIn` is one of the many ways you can cary out model averaging, but is any one method better than any other? Also is model averaging better than just picking a model?



(Dormann *et al.*, 2018)

Benefits of model averaging

IC based model averaging via `MuMIn` is one of the many ways you can cary out model averaging, but is any one method better than any other? Also is model averaging better than just picking a model?



(Dormann *et al.*, 2018)

"We found little in our results to justify the dominance of AIC-based model averaging. And model-averaging did not necessarily outperform single models."

Model averaging take-home

Model averaging take-home

Model averaging has no super-powers. Like most other statistical methods, model averaging has benefits and costs, and you must weight them to decide which approach is best for your problem.

Model averaging take-home

Model averaging has no super-powers. Like most other statistical methods, model averaging has benefits and costs, and you must weight them to decide which approach is best for your problem.

Benefits include a possible reduction of predictive error and improved parameter estimates.

Model averaging take-home

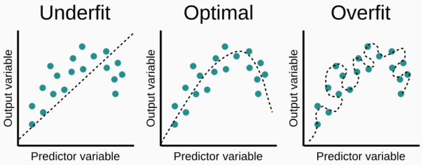
Model averaging has no super-powers. Like most other statistical methods, model averaging has benefits and costs, and you must weight them to decide which approach is best for your problem.

Benefits include a possible reduction of predictive error and improved parameter estimates.

Costs include extra work/computation time, the fact that it does not always work, and that confidence intervals and p-values are difficult to provide.

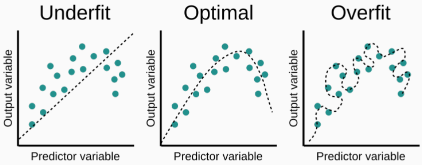
Model Selection and Averaging Recap

Our goal when building models is to identify the fit that optimally balances over- and under-fitting.



Model selection and averaging

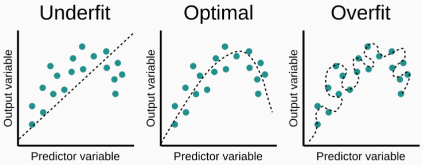
Our goal when building models is to identify the fit that optimally balances over- and under-fitting.



In practice, there is no perfect solution for doing this and how you proceed is part science part art.

Model selection and averaging

Our goal when building models is to identify the fit that optimally balances over- and under-fitting.



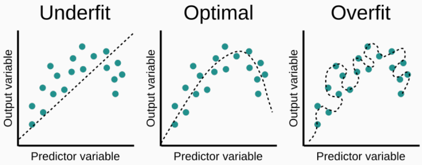
In practice, there is no perfect solution for doing this and how you proceed is part science part art.

Know your data

Model selection and averaging

THE UNIVERSITY OF BRITISH COLUMBIA
Okanagan Campus

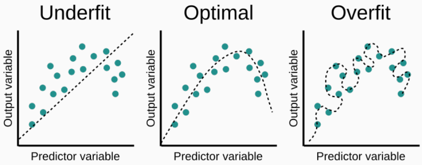
Our goal when building models is to identify the fit that optimally balances over- and under-fitting.



In practice, there is no perfect solution for doing this and how you proceed is part science part art.

Know your data, keep your research question in mind

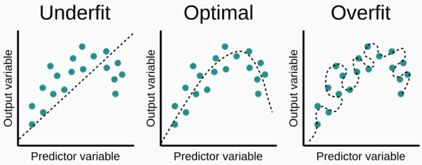
Our goal when building models is to identify the fit that optimally balances over- and under-fitting.



In practice, there is no perfect solution for doing this and how you proceed is part science part art.

Know your data, keep your research question in mind, proceed cautiously

Our goal when building models is to identify the fit that optimally balances over- and under-fitting.

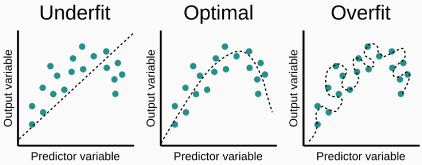


In practice, there is no perfect solution for doing this and how you proceed is part science part art.

Know your data, keep your research question in mind, proceed cautiously, check model assumptions

Model selection and averaging

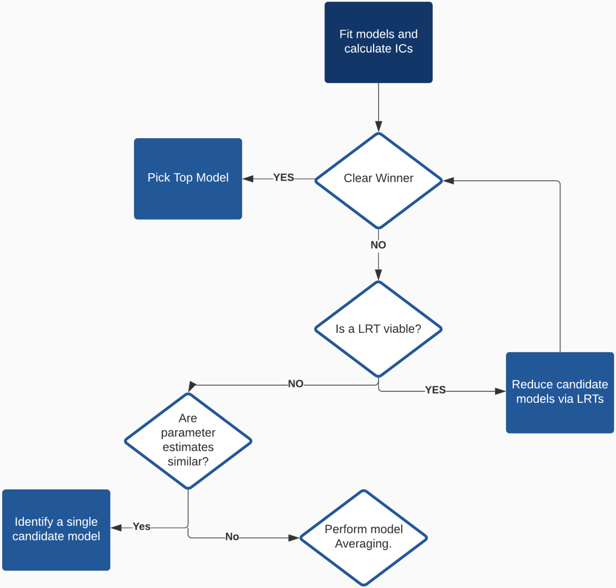
Our goal when building models is to identify the fit that optimally balances over- and under-fitting.



In practice, there is no perfect solution for doing this and how you proceed is part science part art.

Know your data, keep your research question in mind, proceed cautiously, check model assumptions, and always check model quality/performance.

Pragmatic workflow



References

Burnham, K.P. & Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer New York, New York, NY.

Burnham, K.P., Anderson, D.R. & Huyvaert, K.P. (2011). AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, 65, 23–35.

Dormann, C.F., Calabrese, J.M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C.M., Ciuti, S., Elith, J., Gerstner, K. *et al.* (2018). Model averaging in ecology: A review of bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs*, 88, 485–504.

Murtaugh, P.A. (2014). In defense of p values. *Ecology*, 95, 611–617.