Likelihood-based model selection

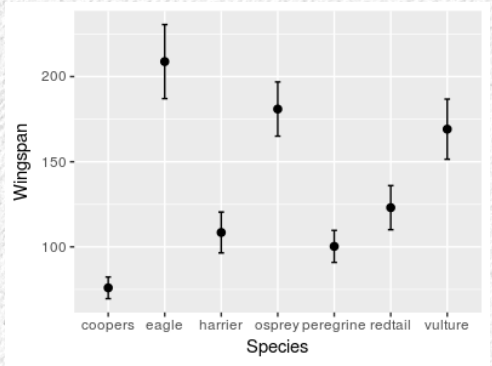
- Models as hypotheses (again, for real this time)
- Approximating models
- Support for competing hypotheses
- Information-theoretic criteria (AIC)
- Model uncertainty

Models are hypotheses

- In the sense that...
  - Hypothesize the response depends on a predictor by including the predictor in the model
    - Hypothesize the response is independent of a predictor by leaving it out of the model
  - Hypothesize observations are different on average by splitting them into different categories
    - Hypothesize that levels of a categorical variable are not different on average by combining them
  - Hypothesizes that the response to one predictor depends on the level of another by including an interaction
    - Hypothesize that the response to one predictor is independent of the response to the other by excluding an interaction

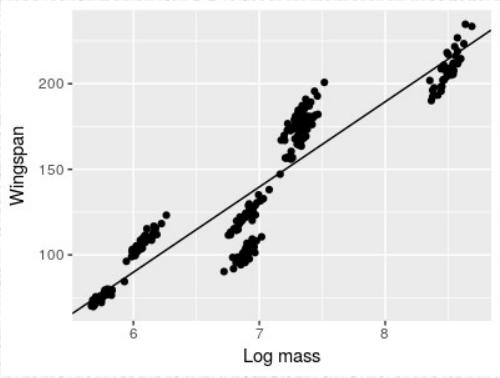
Wingspan differs between species

WS ~ species



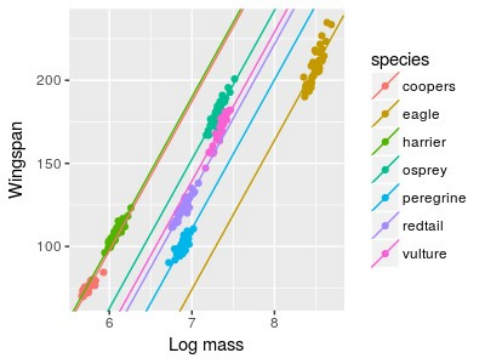
Wingspan depends on log mass

WS ~ log.mass



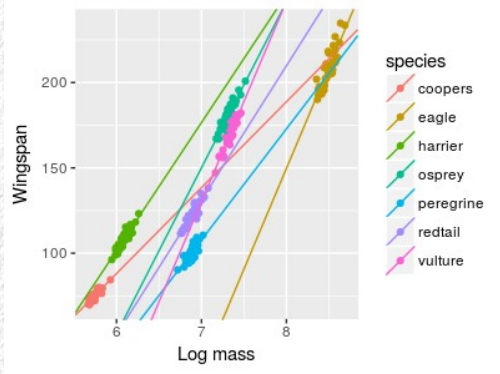
Wingspan differs between species, and depends on log mass

WS ~ species + log.mass



Relationship between wingspan and log mass depends on species

WS ~ species * log.mass



The problem to be solved...

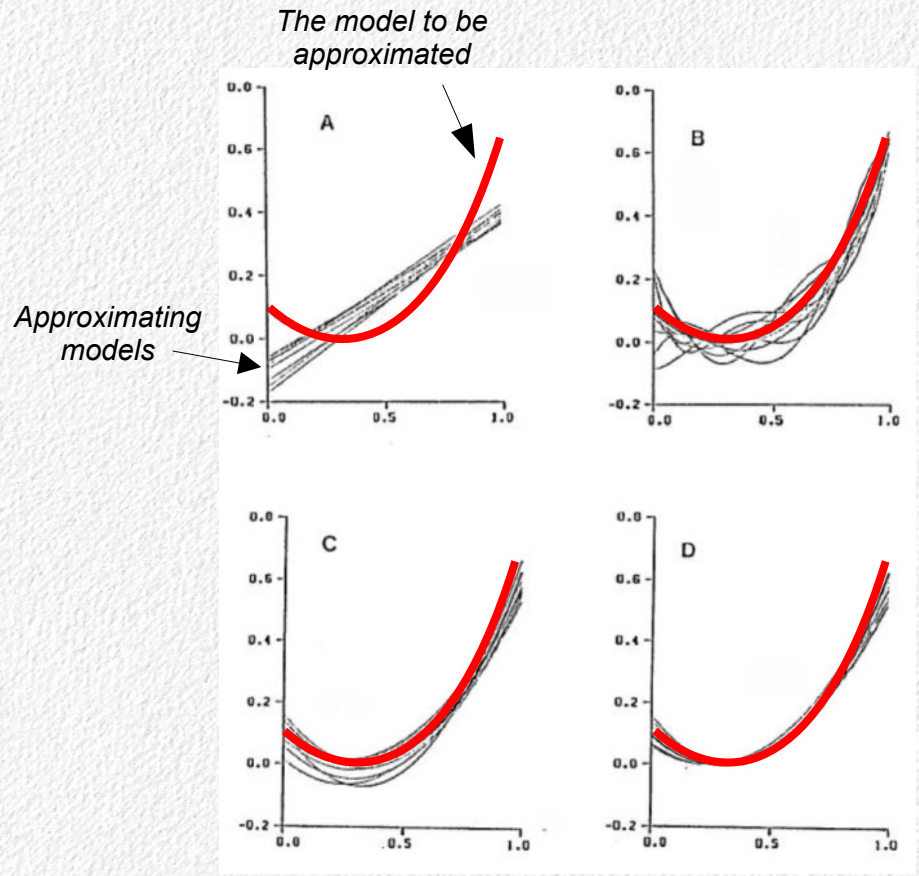
- According to statistician George E.P. Box “all models are wrong, but some are useful”
- Wrong in that:
  - They are abstractions, simplifications
  - Do not contain all of the information in the data
- Useful if:
  - They help us derive knowledge from data
- How do we judge which models are the least wrong and most useful? We find the model best supported by the data

The distance between two models

- K-L distance (or K-L information) = measure of distance between any two models
  - Developed by Solomon Kullback and Richard Leibler
- Assume one of the models, f(x), is the True model (that is, we assume f(x) exists but we don't know its properties)
- Second model is g(x|Φ) – used to approximate the True model, with parameter Φ
- K-L distance measures how much information about the true model, f(x), is lost when it's approximated by g(x|Φ)
  - Smaller values = less loss of information, which is better

$$I(f,g)=\int f(x)\ln\left(\frac{f(x)}{g(x|\phi)}\right)dx$$

Different approximations to a model



Artificial example – the red quadratic curve is the "true model"

Which looks like the best approximation?

If K-L distance is a good measure of distance between models, it should distinguish these four cases

Should also have "good properties", such as transitivity, which means that:

If B is better supported than A,

and

C is better supported than B,

then

C should be better supported than A

We don’t know the True Model...problem?

- The true model, f(x), is part of the equation, but we don’t know what it is
- Solution: compare different models to one another
  - f(x) is the one that generated the data → f(x) will have the highest likelihood given the data
  - The closer a model is to f(x) the better an approximation it is, and the better supported it will be by the data
  - Therefore, *of the models being considered*, the one with the best support from the data is the one that is closest to f(x)

Using likelihoods to approximate K-L distance

- We can approximate K-L distance using “Akaike’s Information Criterion”:

$$AIC = -2\ln\left(\mathscr{L}\left(model \middle| the\ data\right)\right) + 2K$$

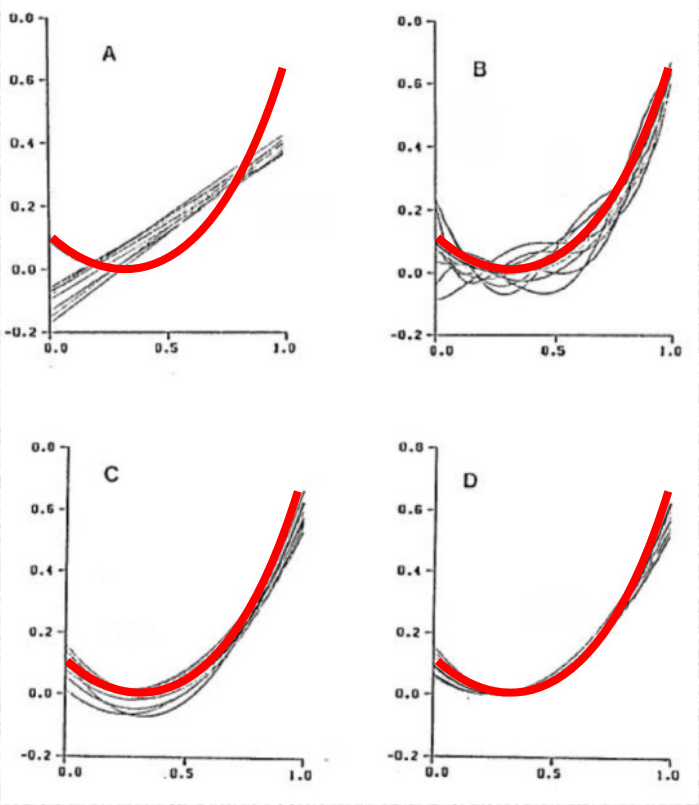
- AIC equals K-L distance up to an (unknown) additive constant (AIC = KL + C)
- Balances fit (likelihood) and complexity (# parameters = K)
  - Higher the likelihood the smaller -2ln(L(model|data)) is
  - Greater the number of parameters the larger 2K is
- Comparing AIC’s for two models, Model 1 and Model 2
  - $AIC_1 - AIC_2 = KL_1 + C - (KL_2 + C) = KL_1 - KL_2$
  - Therefore, the difference between AIC's is also the differences between K-L distances
  - We can know AIC_1-AIC_2 = KL_1 - KL_2 even if we don’t know f(x), or C

AIC values

- As expected, best model is D

  D: AIC = 215.01
  C: AIC = 216.80
  B: AIC = 221.07
  A: AIC = 228.58

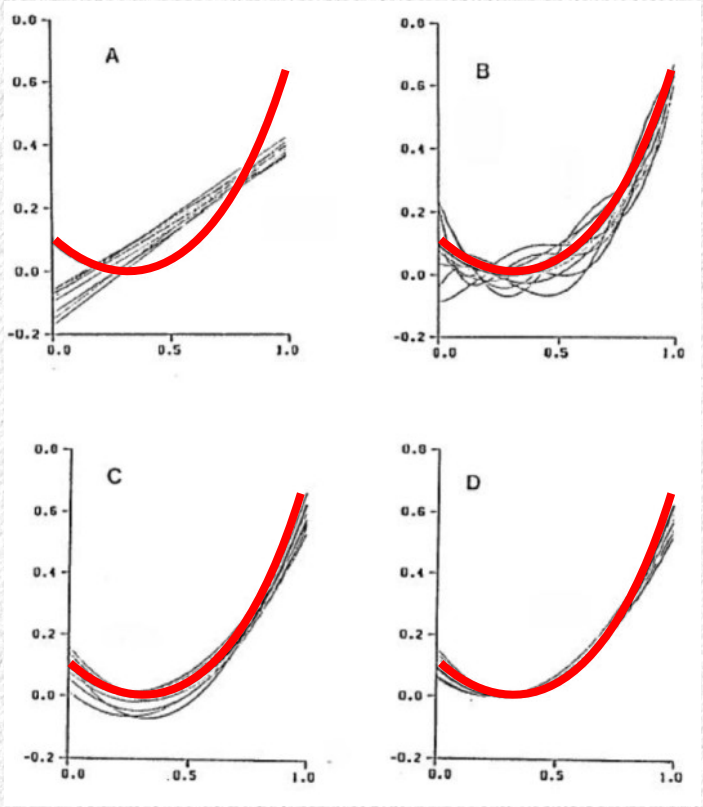
- The magnitudes are unimportant, only the difference between them matters



ΔAIC

- Interpretation of results focuses on differences in AIC between models, symbolized as ΔAIC
  - Identify the model with the lowest AIC
  - Subtract smallest AIC from all the model AIC's → ΔAIC
- ΔAIC's indicate differences in support for models in the data
  - ΔAIC = 0 is best supported
  - ΔAIC less than 2 indicates fairly equivalent support
  - ΔAIC between 4 and 7 indicate substantial differences in support
  - ΔAIC greater than 10 indicates essentially no support for a model relative to the best supported

Example



Best model is D, but C is also well supported

ΔAIC's:

D = 0.00
C = 1.79
B = 6.06
A = 13.57

Refinements to AIC

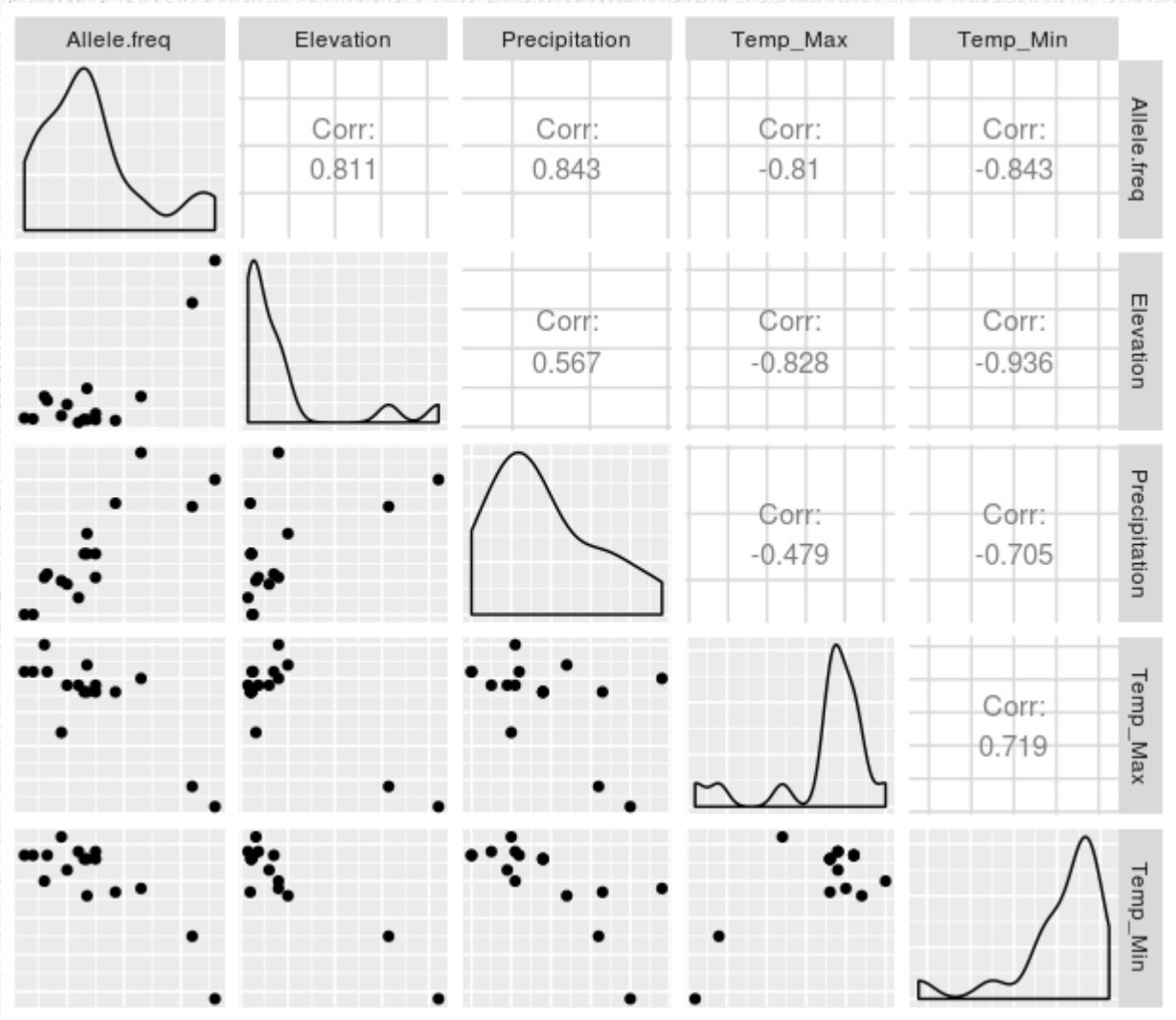
- For sample sizes per parameter (n/K) < 40 use AIC_c

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1}$$

- Penalty for complexity is larger when n is small
- Why? Adding more parameters is a bigger problem with less data, and should incur a greater penalty

Example: butterfly data

- Frequency of alleles for a gene are thought to be changing due to differences in environmental conditions
- Temperature and precipitation are leading candidates, but they also are correlated
- Is one explanation better supported than the other? Or are they indistinguishable from one another?



The basic patterns

Several high correlations

Elevation is correlated with precipitation, max, and min temperature

Precipitation, max and min are inter-correlated with one another

Which variables best explain allele frequency?

Do we need both temperature measurements?

Do we need both min temp and max temp?

Allele frequency was modeled with max temp, min temp, and both together

Max and min is Allele.freq ~ Max + Min
Max is Allele.freq ~ Max
Min is Allele.freq ~ Min

Model	***K***	***AIC***	***dAIC***
Max, Min	3	73.18	0.00
Min	2	76.80	3.62
Max	2	79.59	6.41

Should we drop a temperature?

What environmental characteristics best explain variation in this gene?

- Use maximum temp, minimum temp, precipitation and elevation to predict gene frequency
- Biologically sensible hypotheses, not necessarily all possible
  - Include each alone
  - A model with max, min, precip
  - A model with max*min*precip
  - Models with two-way interactions between max*precip, min*precip

The full set of models compared

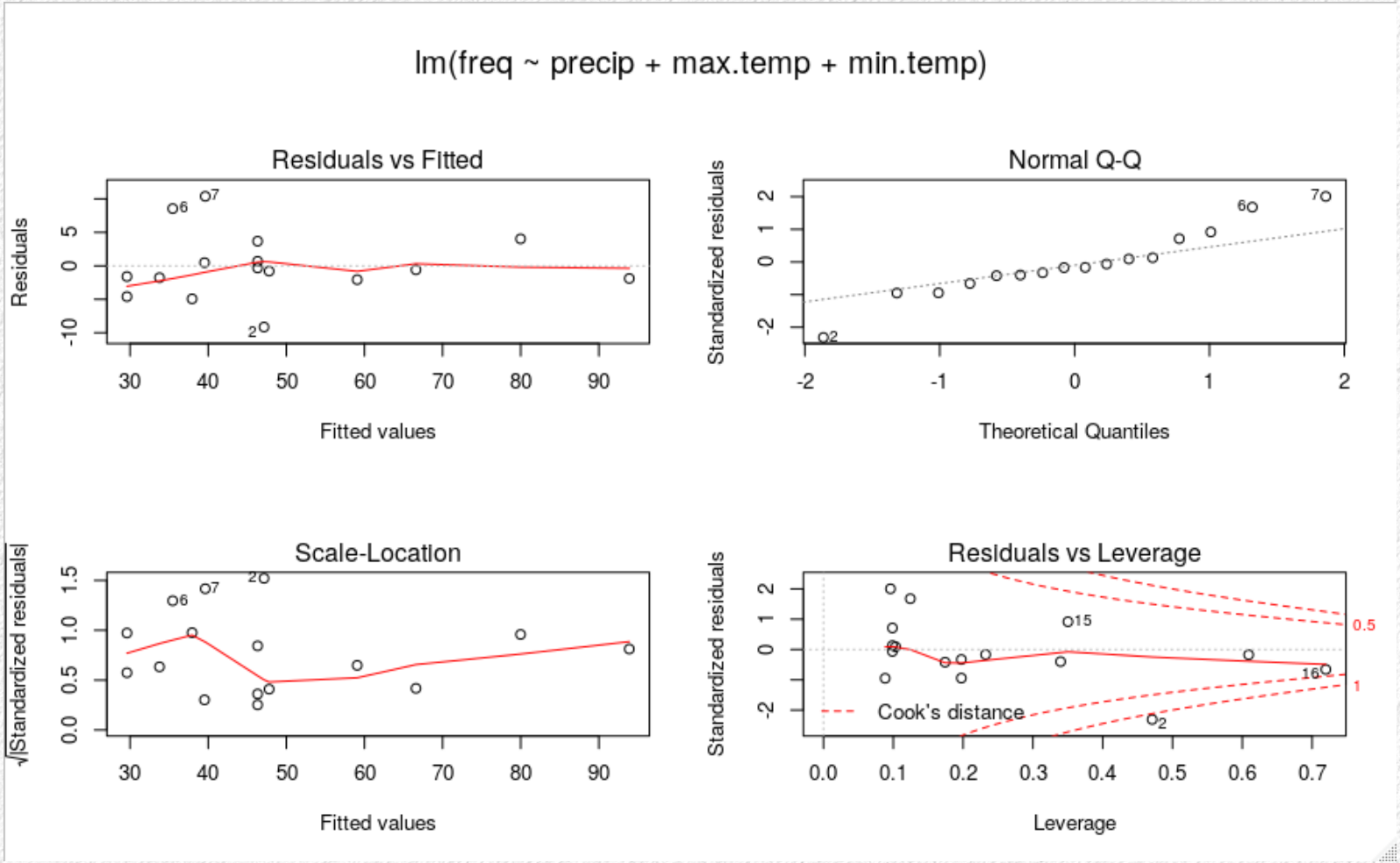
Model	*R^2*	*K*	*AIC*	*dAIC*	*AICc*	*dAICc*
Max, Min, Precip	0.932	4	57.58	0.00	61.22	0.00
Max, Min x Precip	0.935	5	58.97	1.39	64.97	3.75
Max x Precip, Min	0.933	5	59.31	1.73	65.31	4.09
Max x Precip, Min x Precip	0.938	6	60.06	2.48	69.40	8.18
Max x Min	0.856	4	71.76	14.17	75.39	14.17
Max, Min	0.796	3	73.18	15.60	75.18	13.96
Precip	0.711	2	76.77	19.19	77.69	16.47
Min	0.71	2	76.80	19.22	77.72	16.51
Max x Min x Precip	0.957	8	58.40	0.81	78.97	17.75
Elevation	0.657	2	79.51	21.93	80.44	19.22
Max	0.655	2	79.59	22.01	80.52	19.30

Can have different predictors, but all must use the same response (allele frequency)
Is the best supported model the one with the highest R^2?
Compare AIC to AICc – sample size important?

Model fit, assumptions

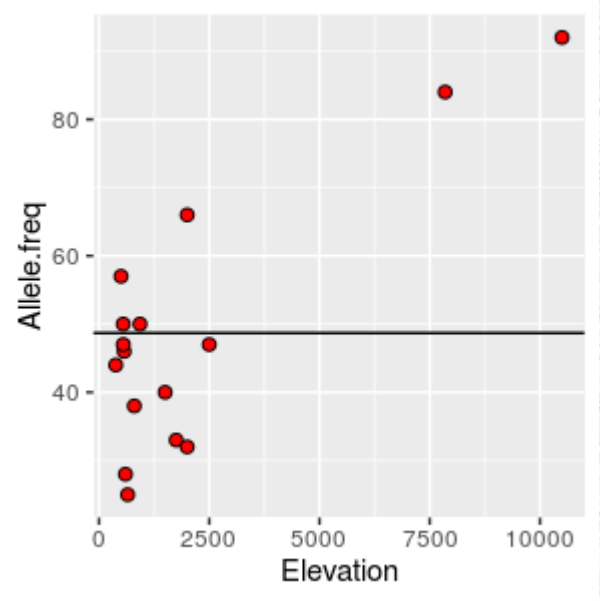
- We must still pay attention to meeting assumptions, measures of explained variation
  - Likelihoods are based on a specified distribution of residuals (we're assuming normal)
  - Our analysis only tells us which model is best relative to the others under consideration – could be the best is still terrible
- Need to check assumptions on well-supported models
- Need to check that best supported models are any good at all

No obvious problems in the model fit



What if the best supported model sucks?

- Lack of a null hypothesis means we aren't considering the “no effect” possibility
- We can fix this with an **intercept only** model
  - Allele frequency is modeled as independent of all the predictors
  - The intercept is the mean of the response variable (mean allele.freq)
- If none of the models is better supported than the intercept only model, then none of them should be interpreted
- Note that the null hypothesis isn’t special with this method – just another hypothesis



With intercept only added...

Model	*R^2*	*K*	*AIC*	*dAIC*	*AICc*	*dAICc*
Max, Min, Precip	0.932	4	57.58	0.00	61.22	0.00
Max, Min x Precip	0.935	5	58.97	1.39	64.97	3.75
Max x Precip, Min	0.933	5	59.31	1.73	65.31	4.09
Max x Precip, Min x Precip	0.938	6	60.06	2.48	69.40	8.18
Max x Min	0.856	4	71.76	14.17	75.39	14.17
Max, Min	0.796	3	73.18	15.60	75.18	13.96
Precip	0.711	2	76.77	19.19	77.69	16.47
Min	0.71	2	76.80	19.22	77.72	16.51
Max x Min x Precip	0.957	8	58.40	0.81	78.97	17.75
Elevation	0.657	2	79.51	21.93	80.44	19.22
Max	0.655	2	79.59	22.01	80.52	19.30
Intercept only	0	1	94.64	37.06	94.93	33.71

So, all the models in the set are better than the null, but Max + Min + Precip is the best of the bunch

Model uncertainty

- There may not be a single best-supported hypothesis in the set under consideration
- We can increase the chances of a clear winner by:
  - Designing informative experiments
    - Measure variables for which predictions of competing hypotheses are different
    - Maximize the amount of independent variation between predictors to minimize confounding (that is, use good experimental design)
  - Increasing sample size
- But, we need a way to evaluate our confidence in our best model

Measuring model uncertainty

- We can calculate “Akaike weights” that help us deal with uncertainty about degree of support for competing models
- Measure the probability that a model would be selected as best if the experiment were repeated
  - Vary between 0 and 1
  - Sum to 1 across the set of models being compared
- Ideally, the best-supported hypothesis will have a weight near 1, and the rest will have weights near 0

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum \exp\left(-\frac{1}{2}\Delta\right)}$$

Akaike weights

Model	*K*	*AICc*	*dAICc*	*w*
Max, Min, Precip	4	61.22	0.00	0.7680
Max, Min x Precip	5	64.97	3.75	0.1177
Max x Precip, Min	5	65.31	4.09	0.0994
Max x Precip, Min x Precip	6	69.40	8.18	0.0129
Max x Min	4	75.39	14.17	0.0006
Max, Min	3	75.18	13.96	0.0007
Precip	2	77.69	16.47	0.0002
Min	2	77.72	16.51	0.0002
Max x Min x Precip	8	78.97	17.75	0.0001
Elevation	2	80.44	19.22	0.0001
Max	2	80.52	19.30	0.0000
Intercept only	1	94.93	33.71	0.0000

The best supported model is expected to be the best model 76.8% of the time, if the study was repeated

The second and third best supported models would be selected 10-12% of the time

Do p-values give the same impression about model support?

Model	*dAICc*	*w*	*p*
Max, Min, Precip	0.00	0.7680	0.0000002762
Max, Min x Precip	3.75	0.1177	0.0000018545
Max x Precip, Min	4.09	0.0994	0.0000020802
Max x Precip, Min x Precip	8.18	0.0129	0.0000096551
Max x Min	14.17	0.0006	0.0000536565
Max, Min	13.96	0.0007	0.0000322064
Precip	16.47	0.0002	0.0000406060
Min	16.51	0.0002	0.0000411862
Max x Min x Precip	17.75	0.0001	0.0000752443
Elevation	19.22	0.0001	0.0001392917
Max	19.30	0.0000	0.0001445422
Intercept only	33.71	0.0000	

If only one of these was presented with a p-value, would we doubt it was well supported?

What if multiple models are well supported?

- Say so! Interpret all well-supported models, discuss the similarities and differences
  - Separating the well-supported from the poorly supported models is worthwhile, even if more than one are well supported
- The importance of individual variables can be measured by how often they occur across multiple models
- “Model averaging” - average the coefficients across all retained models to obtain estimates of effects

Interpreting the predictors

- The Method of Support is based on assessing support for **models**
- But, we understand the results in terms of **variables**
- When there is a single, clearly best-supported model we interpret the variables in the usual way
  - Slopes/standardized coefficients
  - Partial effect sizes
- When there isn't a clearly best-supported model, we can sum the weights of models variables appear in to get a measure of importance for predictors

Evidence of the importance of Min, Max, and Precip

Model	***dAICc***	***w***		Variable	Sum of w's
Max, Min, Precip	0.00	0.7680		Min	0.99970
Max, Min x Precip	3.75	0.1177		Max	0.99955
Max x Precip, Min	4.09	0.0994		Precip	0.99834
Max x Precip, Min x Precip	8.18	0.0129			
Max x Min	14.17	0.0006			
Max, Min	13.96	0.0007			
Precip	16.47	0.0002			
Min	16.51	0.0002			
Max x Min x Precip	17.75	0.0001			
Elevation	19.22	0.0001			
Max	19.30	0.0000			
Intercept only	33.71	0.0000			

Post-hoc procedures

- If you apply the method to an ANOVA model, you would know which model is best supported, but would not know which groups are different
- Could give up and resort to Tukey tests
- Or, a model selection-based approach would be to:
  - Estimates of means and confidence intervals
  - Compare models with merged factor levels

Example: merged factor levels

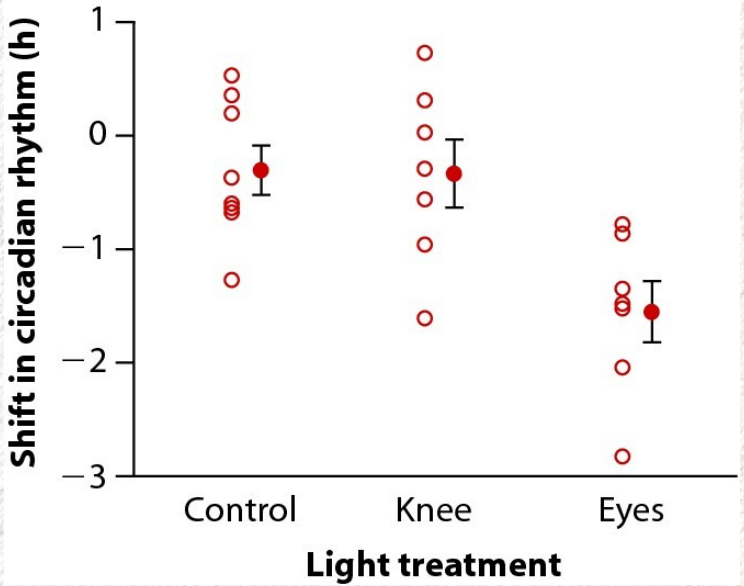
- Approach as: what are the **homogeneous subsets**?
- The model with groupings that are best supported by the data is interpreted
- No need to test all possible – can order them by smallest to largest mean, and compare possible ordered groupings
- Example of sleep cycle data

Circadian cycles

- When you travel, you get jet lag – sleepy at the wrong time
- Exposure to the light/dark cycle at your destination eventually restores your normal sleep cycle
- One study found that shining light on the back of the knee could help shift circadian rhythms
- Controversial result (no reason for it to be true)
- Re-tested, comparing the amount of shift in circadian rhythm for:
  - Untreated controls
  - Light shined on the backs of knees
  - Light shined in eyes

The data

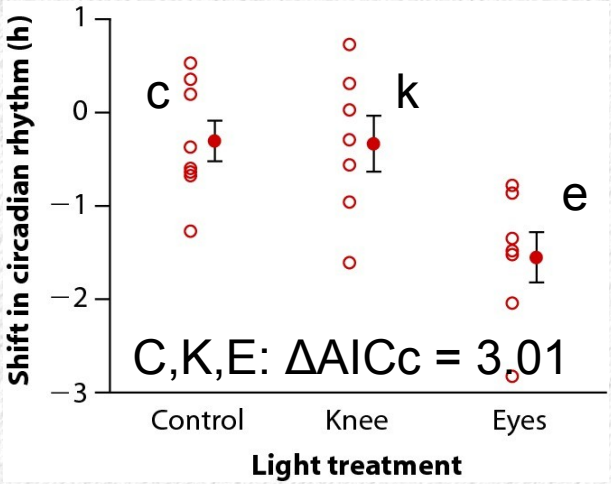
- Use the following groupings for light treatments:
  - Control, Knee, Eye (unchanged treatment column)
  - Control, Knee&Eye
  - Control&Knee, Eye
  - Control&Eye, Knee
  - Control&Knee&Eye (intercept only)
- Run one model for each of these groupings
- Interpret the best supported model

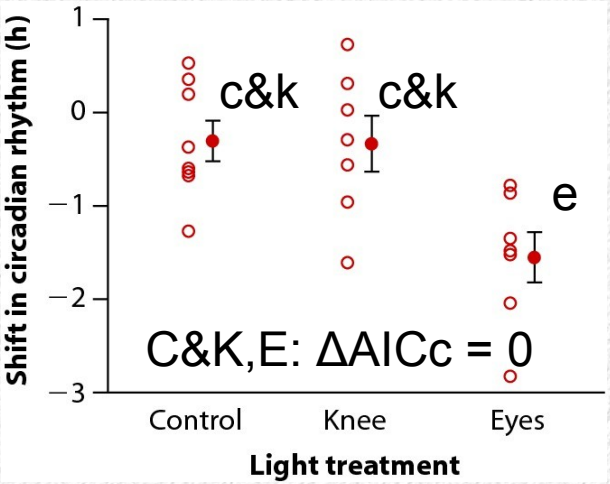


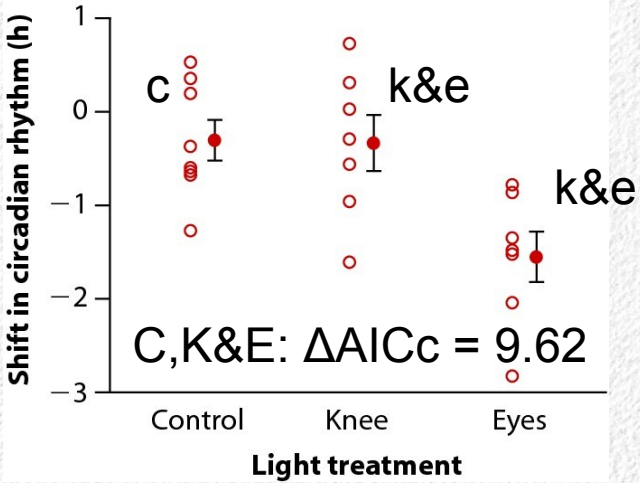
Treatment	c.ke	ck.e	ce.k	Shift
Control	Control	ControlOrKnee	ControlOrEye	0.53
Control	Control	ControlOrKnee	ControlOrEye	0.36
Control	Control	ControlOrKnee	ControlOrEye	0.2
Control	Control	ControlOrKnee	ControlOrEye	-0.37
Control	Control	ControlOrKnee	ControlOrEye	-0.6
Control	Control	ControlOrKnee	ControlOrEye	-0.64
Control	Control	ControlOrKnee	ControlOrEye	-0.68
Control	Control	ControlOrKnee	ControlOrEye	-1.27
Knee	KneeOrEye	ControlOrKnee	Knee	0.73
Knee	KneeOrEye	ControlOrKnee	Knee	0.31
Knee	KneeOrEye	ControlOrKnee	Knee	0.03
Knee	KneeOrEye	ControlOrKnee	Knee	-0.29
Knee	KneeOrEye	ControlOrKnee	Knee	-0.56
Knee	KneeOrEye	ControlOrKnee	Knee	-0.96
Knee	KneeOrEye	ControlOrKnee	Knee	-1.61
Eyes	KneeOrEye	Eyes	ControlOrEye	-0.78
Eyes	KneeOrEye	Eyes	ControlOrEye	-0.86
Eyes	KneeOrEye	Eyes	ControlOrEye	-1.35
Eyes	KneeOrEye	Eyes	ControlOrEye	-1.48
Eyes	KneeOrEye	Eyes	ControlOrEye	-1.52
Eyes	KneeOrEye	Eyes	ControlOrEye	-2.04
Eyes	KneeOrEye	Eyes	ControlOrEye	-2.83

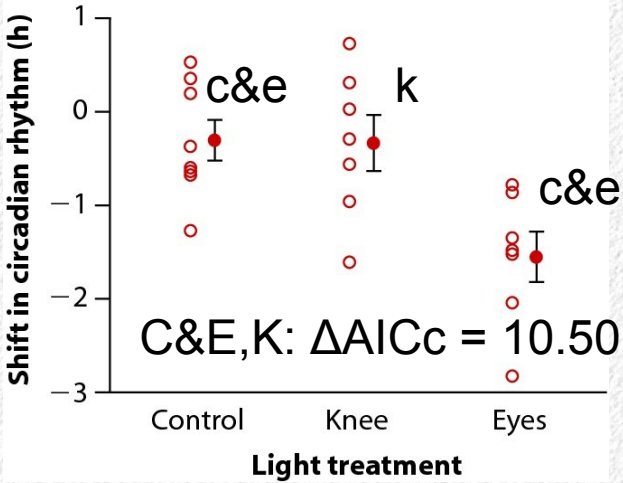
Re-coding the data

- Predictor columns represent different possible models
- Treatment is model c.k.e (all three different)
- cke (no differences) is an “intercept only” model = no groups, just likelihood of grand mean given the data (no column needed)









Intercept only: ΔAICc = 9.82

What's the model?

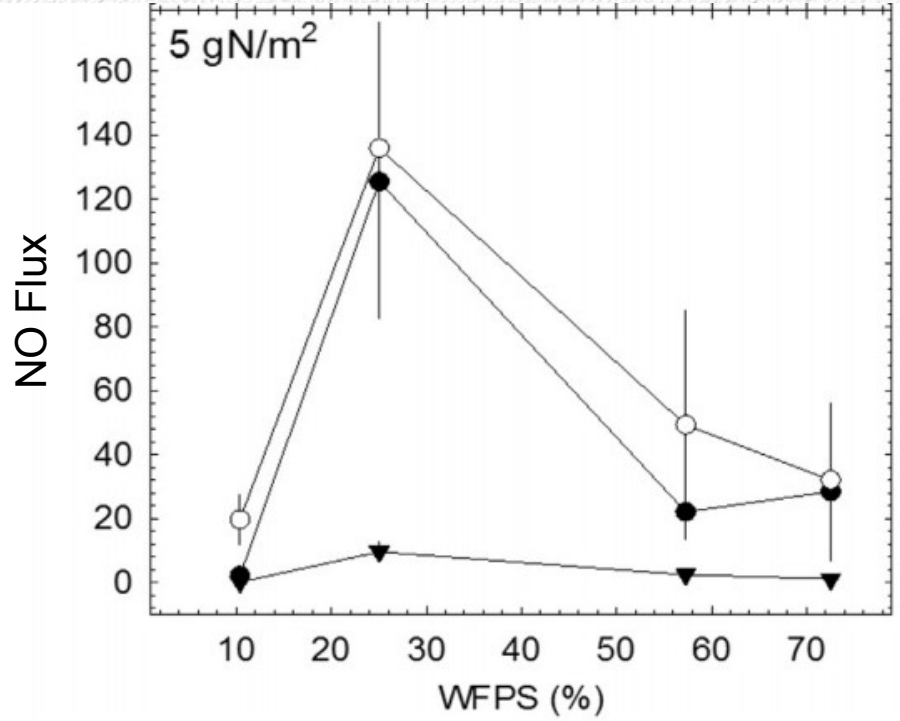


Fig. 3. The mean (±SE, n = 5) NO flux as a function of water-filled pore space (WFPS) and temperature in soil exposed to 0 g N/m^2 (control; top panel), 2 g N/m^2 (middle-panel), and 5 g N/m^2 (top-panel). The plotting symbols indicate temperature treatments; ice-water bath (inverted-triangles), room temperature (black-circles), and hot-water bath (white-circles).