Model Selection

Likelihood Ratio Tests and Information Criteria

Advanced Biostatistics

Dean Adams

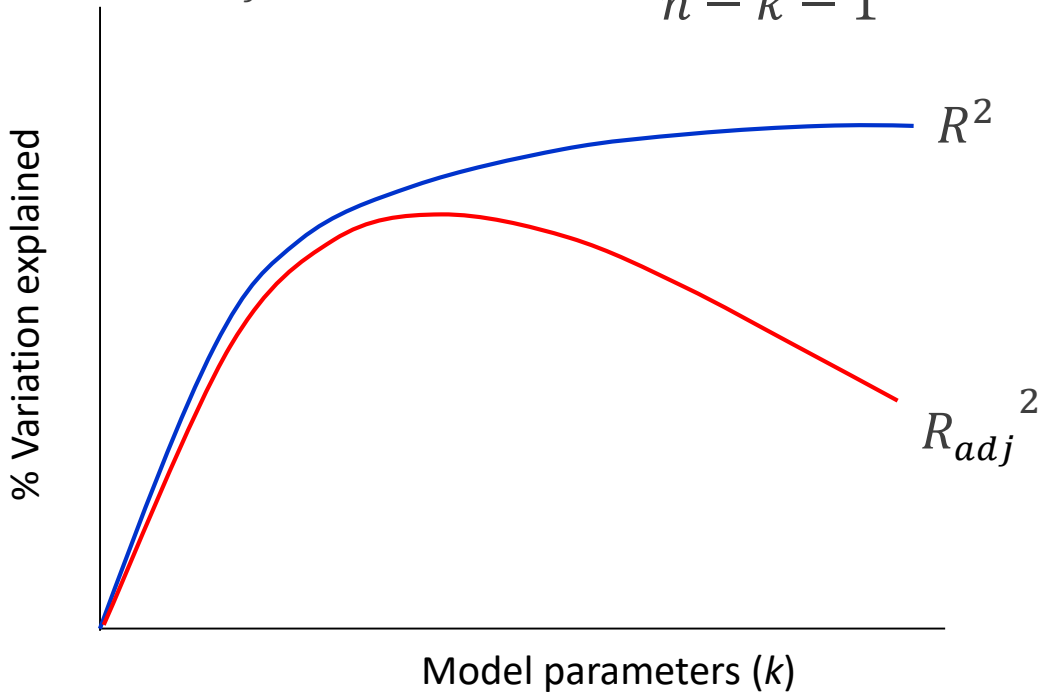
Lecture 11

EEOB 590C

Often, (especially in observational studies) it is unknown what 'explanatory' variables are important sources of variation.

We can always 'dump' any or all explanatory variables into a model, but there are some problems with this.

$$R_{adj}{}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$



Increasing model parameters (*k*) results in:

1) Reduction of explained information relative to the model "size"

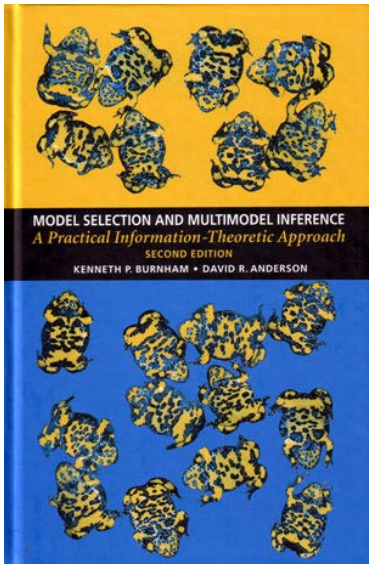
2) Potential loss of statistical power (i.e., eventually model *df* > error *df*)

The goal is to **evaluate two or more models with different sets of parameters** to determine if one is 'better' (i.e., explains more variation in the response given certain criteria).

→ Likelihood Ratio Tests: statistical comparison of nested models

→ Information Theoretic approaches: ranking of models based on parsimony

→ Cross-validation procedures: measure robustness of a specific model



Burnham and Anderson 2002. Model selection and multimodel inference: A Practical Information Theoretic Approach. Springer.

Likelihood ratio tests statistically compare one model (e.g., full model) with an alternative model that is nested within the full model (e.g., reduced model).

Model likelihood: the likelihood of set of model parameters (**β**), given the data (**Y**), is equal to the probability of the observed outcomes given the parameters.

$$L\left(\hat{\boldsymbol{\beta}} \middle| \mathbf{Y}\right)$$

For normally distributed error (uni- or multi-variate), the maximum likelihood estimator (MLE) is:

$$L(\hat{\boldsymbol{\beta}}|\mathbf{Y}) = \frac{1}{(\sqrt{2\pi})^{np}|\boldsymbol{\Sigma}|^{n/2}} e^{-\left[\sum(\mathbf{Y}_i-\bar{\mathbf{Y}})^t\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\bar{\mathbf{Y}})\right]/2-n\left[\sum(\bar{\mathbf{Y}}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{Y}}-\boldsymbol{\mu})\right]/2}$$

Multivariate normal probability density function (with $\boldsymbol{\Sigma}$ = error covariance matrix), which is maximized when $\bar{\mathbf{Y}} = \boldsymbol{\mu}$ (i.e., when exp = 0)

For $\bar{\mathbf{Y}} = \boldsymbol{\mu}$ $\log(L(\hat{\boldsymbol{\beta}})) = -\frac{1}{2}n\log|\boldsymbol{\Sigma}| - \frac{n}{2}\log(2\pi) - \frac{n}{2}$

This is a constant; if two nested models are compared, then the difference in equal constants is 0

For comparing nested models, the log-likelihood of a linear model is thus:

$$\log(L(\hat{\boldsymbol{\beta}})) = -\frac{1}{2}n\log|\boldsymbol{\Sigma}|$$

$$\log\left(L(\hat{\boldsymbol{\beta}}_{full})\right) - \log\left(L(\hat{\boldsymbol{\beta}}_{reduced})\right)$$

$$= -\frac{1}{2}n\left(\log|\boldsymbol{\Sigma}_{full}| - \log|\boldsymbol{\Sigma}_{reduced}|\right)$$

$$\log\left(\frac{L(\hat{\boldsymbol{\beta}}_{full})}{L(\hat{\boldsymbol{\beta}}_{reduced})}\right) = -\frac{1}{2}n\log\left(\frac{|\boldsymbol{\Sigma}_{full}|}{|\boldsymbol{\Sigma}_{reduced}|}\right)$$

This is a likelihood **ratio**!

$$LRT = -2\log\left(\frac{L(\hat{\boldsymbol{\beta}}_{full})}{L(\hat{\boldsymbol{\beta}}_{reduced})}\right) = n\log\left(\frac{|\boldsymbol{\Sigma}_{full}|}{|\boldsymbol{\Sigma}_{reduced}|}\right)$$

LRT ~ X^2, df = $(\Delta k)p$
where p: number of response variables
k_i: number of parameters of model i

$$LRT = -2\log\left(\frac{L(\hat{\boldsymbol{\beta}}_{full})}{L(\hat{\boldsymbol{\beta}}_{reduced})}\right) = n\log\left(\frac{|\boldsymbol{\Sigma}_{full}|}{|\boldsymbol{\Sigma}_{reduced}|}\right)$$

$\boldsymbol{\Sigma}$: error covariance matrix

-Likelihood ratio tests involve statistically comparing one model with an alternative model that is nested within the full model.

-Although not explicitly stated, we have performed many likelihood ratio tests thus far, e.g., MANOVA with Wilks' Λ

-Let **T** be the *p x p* SSCP matrix for the data, partitioned into **H** (SSCP predicted) and **E** (SSCP residual error)

For nested models,

$\mathbf{T} = \mathbf{H}_{Full} + \mathbf{E}_{Full} = \mathbf{H}_{Reduced} + \mathbf{E}_{Reduced}$

A difference in models is thus

$\Delta\mathbf{H} = \mathbf{H}_{Full} - \mathbf{H}_{Reduced} = \mathbf{E}_{Reduced} - \mathbf{E}_{Full}$

$$\Lambda = \frac{|\mathbf{E}_{Full}|}{|\mathbf{E}_{Full} + \Delta\mathbf{H}|}$$ ← Typical definition

$$\Lambda = \frac{|\mathbf{E}_{Full}|}{|\mathbf{E}_{Full} + (\mathbf{E}_{Reduced} - \mathbf{E}_{Full})|}$$

$$\Lambda = \frac{|\mathbf{E}_{Full}|}{|\mathbf{E}_{Reduced}|}$$ ← Alternative definition involves a ratio of model error determinants

➢ A log-transformation of Wilks Λ, scaled by sample size, is an *LRT*. Wilks Λ is a modified *LRT*!

$$LRT = -2\log\left(\frac{L(\widehat{\boldsymbol{\beta}}_{full})}{L(\widehat{\boldsymbol{\beta}}_{reduced})}\right) = n\log\left(\frac{|\boldsymbol{\Sigma}_{full}|}{|\boldsymbol{\Sigma}_{reduced}|}\right)$$

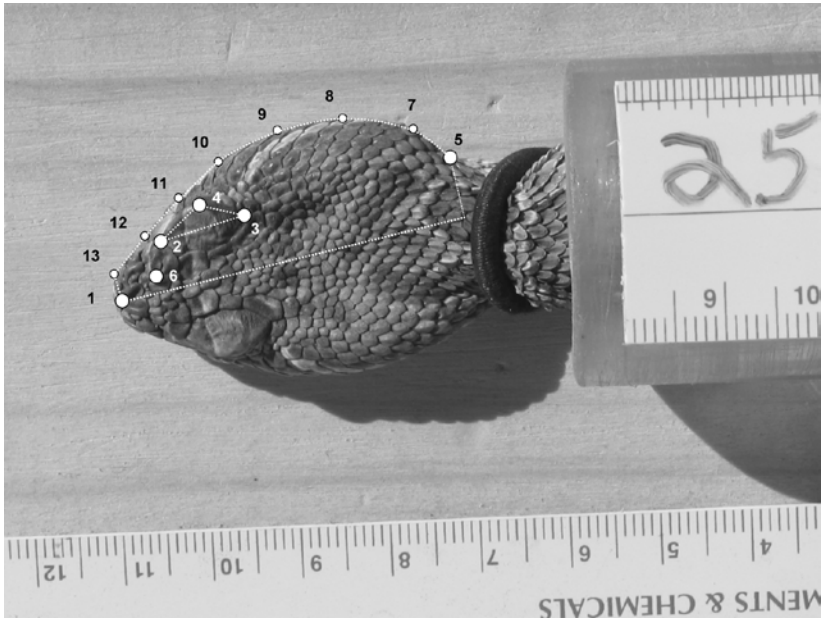
$\boldsymbol{\Sigma}$: error covariance matrix

Likelihood ratio tests involve statistically comparing one model with an alternative model that is nested within the full model.

The null hypothesis of an *LRT* is that the additional parameters in the full model offer no improvement over the simpler set of parameters in the reduced model. A significant *LRT* (rejection of the null hypothesis) implies that significantly more variation of the dependent variable was described by the additional parameters of the full model. Therefore, the full model is considered a significant improvement over the reduced model.

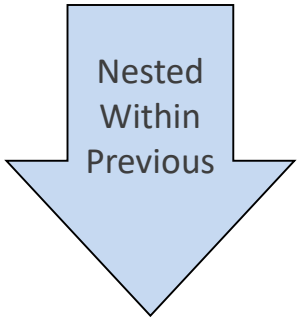
Head shape of (live!) prairie rattlesnakes

Compared several biological hypotheses explaining morphological variation



Smith and Collyer, 2008. In *The biology of rattlesnakes.*

Models considered:



FULL: Shape = Sex + Region + Sex × Reg. + SVL

Sex-reduced: Shape = Region + SVL

Region-reduced: Shape = SVL

SVL-reduced: Shape = Region

*LRT*s:

Full vs. Sex-reduced → Test of sexual dimorphism

Sex-reduced vs. Region-reduced → Test of regional variation

Sex-Reduced vs SVL-reduced → Test of shape/size variation

Smith and Collyer, 2008. In *The biology of rattlesnakes.*

Results

	Full Model for LRT	
Reduced by:	Sex + Region +Sex × Region + SVL	Region + SVL
SVL+Region	Pillai = 0.459, P = 0.375	
Region		Pillai = 0.554, P = 0.00039
SVL		Pillai = 0.346, P = 0.00042

Conclusions:

Sex is not an important source of head shape variation but Region and SVL are certainly important

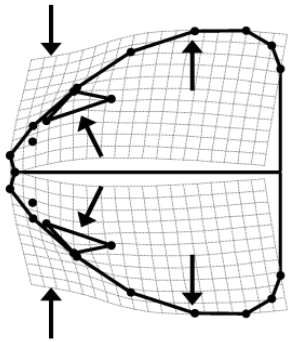
When a MANCOVA was performed, sexual dimorphism was significant, but in this case, the Sex × Region term remains in the reduced model

LRTs demonstrate that adding parameters for Sex and Sex × Region does not offer a significant improvement over a simpler model!

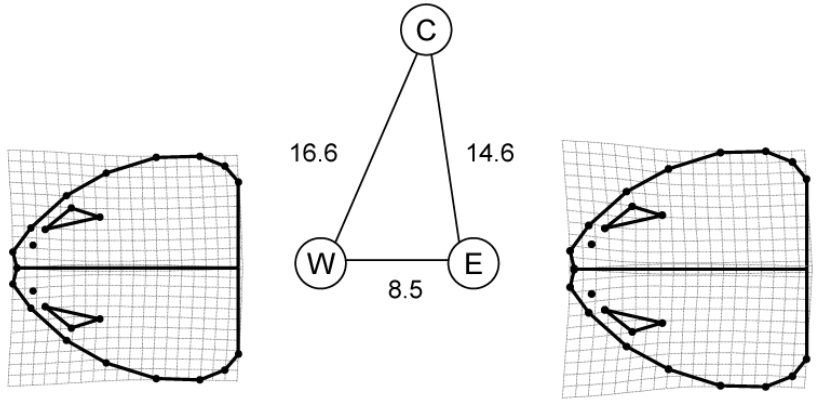
Smith and Collyer, 2008. In *The biology of rattlesnakes.*

2-d Plane containing the three regional head shapes (East, West, Central) in the p = 15 multivariate shape space

Hibernacula = sandy mammal burrows (large holes)



Hibernacula = rocky buttes (small, narrow holes)



Hence, regional variation shows an important biological signal. If sexual dimorphism was tested in the classical sense (e.g., MANOVA), one might conclude it was meaningful. However, it is less meaningful than locally adapted head shape divergence.

Smith and Collyer, 2008. In *The biology of rattlesnakes.*

Advantages

Provides a statistical assessment of model difference, based on importance of parameters that differ between models – ELEGANT!

Can be performed on any level of hierarchical structure (e.g., a model with parameters A, B, C, D, E, F can be compared to one with parameters A and E, one with just F, or one with B, C, and D in the same manner).

Disadvantages

As long as *n:kp* ratio is large, significant differences between models are likely, even for small effects.

Models must be nested comparisons?

$$\log\left(\frac{L(\hat{\boldsymbol{\beta}}_{full})}{L(\hat{\boldsymbol{\beta}}_{reduced})}\right) = -\frac{1}{2}n\log\left(\frac{\left|\boldsymbol{\Sigma}_{full}\right|}{\left|\boldsymbol{\Sigma}_{reduced}\right|}\right)$$

Really? What about this equation implies that models must be nested?

B is nested into A: the log-ratio of likelihoods is always positive
- A is the null and Q evaluates the improved fit of B vs A

Non-nested models: not clear which is the "null"
the log-ratio is not always positive

$$Q = 2\log\left(\frac{L(\hat{\boldsymbol{\beta}}_A)}{L(\hat{\boldsymbol{\beta}}_B)}\right)$$

Procedure:

1. Calculate observed (Q_{obs})
2. Simulate data under model A, and generate distribution of Q_{rand}
3. Compare Q_{obs} to Q_{rand}
4. Simulate data under model B and repeat

Possible outcomes:

- Q significant when A null, but not when B null: (B is preferred)
- Q significant when B null, but not when A null: (A is preferred)
- Q significant when both are null (neither model preferred)
- Q not significant when both are null (no discrimination between models)

See Lewis et al. 2011. *A unified approach to model selection using the likelihood ratio test*. Methods in Ecology and Evolution 2: 155-162.

Approach that ranks candidate models based on score: typically AIC (Akaike's information criterion)*

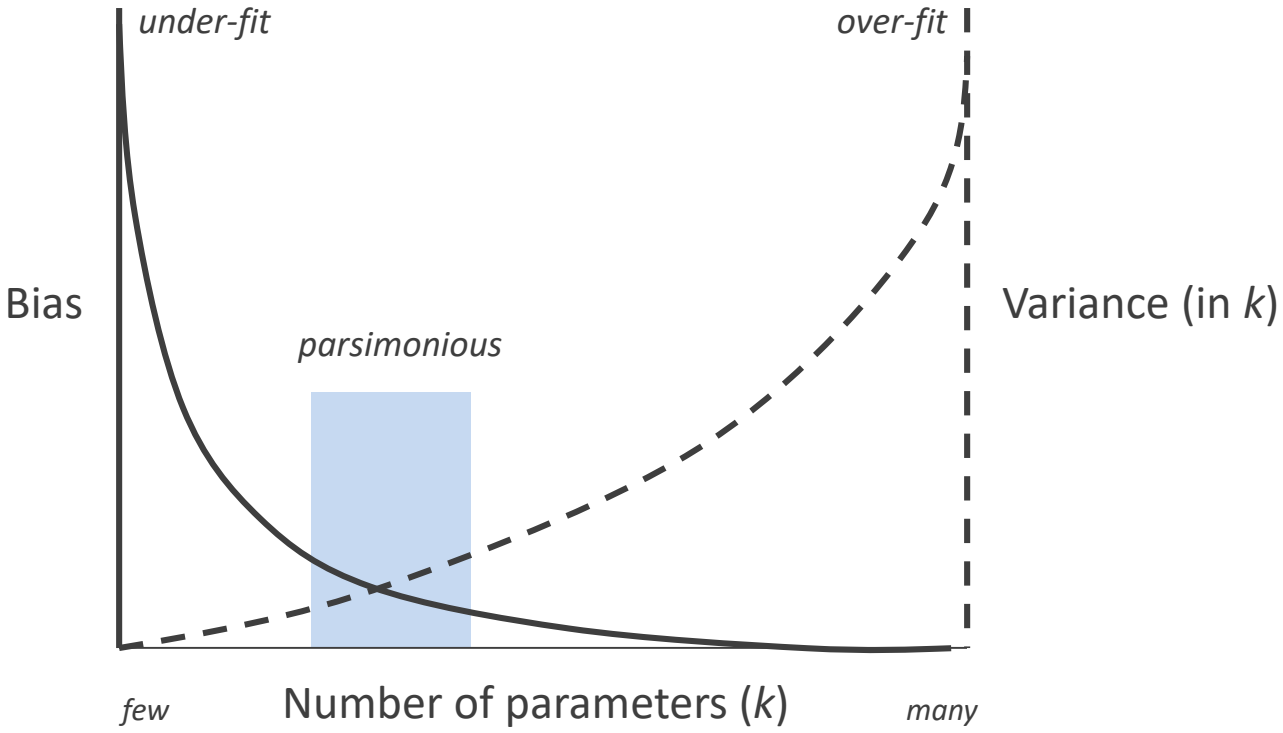
Model selected that explains the most information with the fewest parameters (a 'parsimony' criterion)

IT scores "reward" an increase in likelihood, but "penalize" the increase in number of parameters

Not based on significance testing!

Select the 'best' model found when difference between models is $\Delta AIC > 4$

Note: why $\Delta AIC > 4$? Because for 1df, $\chi^2 = 3.814 \sim 4$ (from frequentist stats!)



This illustration is hypothetical: to measure sampling bias and variance of parameters, true population parameters would need to be known.

The most commonly used criterion to score and compare candidate models

Developed by Hirotugu Akaike (1973, 1974)

The most general form is (Bedrick and Tsai 1994):

$$\mathrm{AIC} = -2\log\left(L\left(\widehat{\boldsymbol{\beta}}\right)\right) + 2[pk + 0.5p(p+1)]$$

For univariate data, this is simplified to*: $\mathrm{AIC} = -2\log\left(L\left(\widehat{\boldsymbol{\beta}}\right)\right) + 2(k+1)$

* But careful with erroneously applying the univariate version to multivariate data!!!

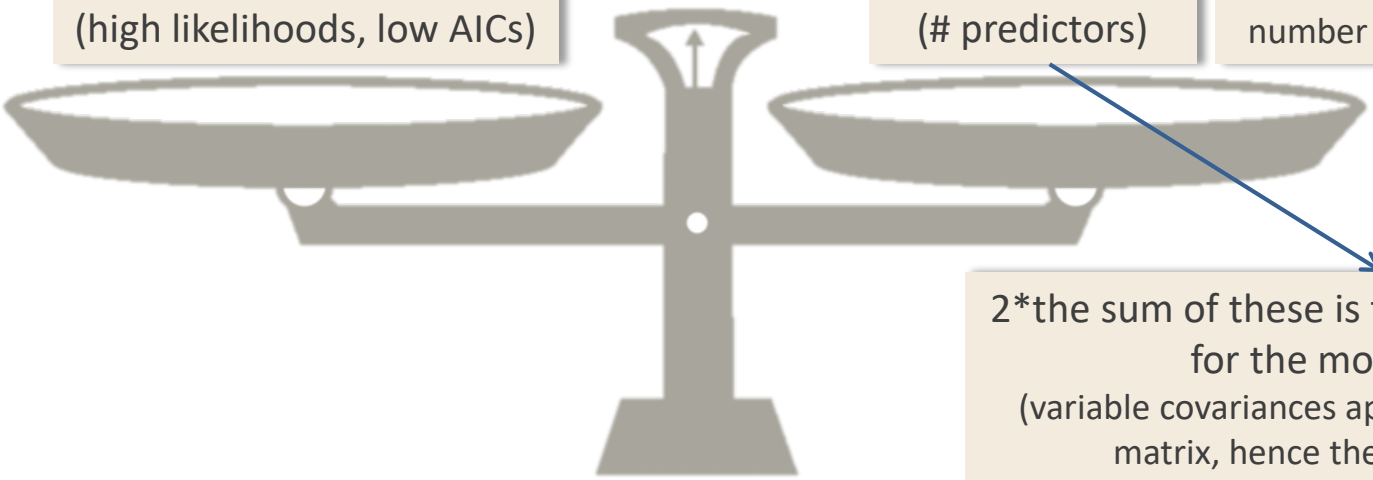
For model comparison, a model is considered better than another when ΔAIC > 4
(because with df = 1, $\chi^2 = 3.814 \approx 4$)

$$\mathrm{AIC} = -2\log\left(L\left(\widehat{\boldsymbol{\beta}}\right)\right) + 2[pk + 0.5p(p+1)]$$

Model likelihood
(high likelihoods, low AICs)

Dimensions of β
(# predictors)

Dimensions of the Σ matrix
(increasing the number of response variables geometrically increases the number of covariances among them)



2*the sum of these is the "***parameter penalty***" for the model likelihood
(variable covariances appear twice in a covariance matrix, hence the multiplication by 2)

Less error means a higher probability that the observed values are produced by model parameters (higher likelihood).
Decreasing error decreases AIC.
Adding model parameters without lowering error increases AIC.
Lower AIC is better.

Important detail: the parameter penalty is arbitrary

AICc (second order AIC)

$$\mathrm{AICc} = -2\log\left(L\left(\hat{\boldsymbol{\beta}}\right)\right) + 2K * \left(\frac{n}{n-K-1}\right)$$

where *K* is the bracketed portion of the parameter penalty of AIC (this is severely crippling for multivariate data, but is considered a "sample size correction" when *n* is low for univariate data).

QAIC (quasi-likelihood AIC)

$$Q\mathrm{AIC} = \left[-2\log\left(L\left(\hat{\boldsymbol{\beta}}\right)\right)/\hat{c}\right] + 2K$$

where *c* is a dispersion index equal to χ^2/df (i.e., corrects for over-dispersion)

Bayesian IC

$$\mathrm{BIC} = -2\log\left(L\left(\hat{\boldsymbol{\beta}}\right)\right) + 2K\log(n)$$

… and many others (see Burnham and Anderson 2002)

Model	Sex	Region	Sex*Reg	SVL	Sex*SVL	Reg* SVL	Sex*Reg* SVL	AIC	Δ AIC*
Full	X	X	X	X	X	X	X	-13285.574	90.46583
MANCOVA	X	X	X	X				-13341.302	34.73799
I	X			X	X			-13338.149	37.89086
II	X			X				-13355.83	20.20989
III	X							-13336.489	39.55036
IV		X		X		X		-13358.317	17.72253
V		X		X				-13376.04	0
VI		X						-13360.637	15.40285
VII				X				-13365.501	10.5392

- $\Delta AIC_i = AIC_i - min(AIC)$. The highlighted models are those with low AIC values and warrant further inspection. Burnham and Anderson (2002) offer a scale ($\Delta AICi > 2$) for judging $\Delta AICi$ values, though $\Delta AIC > 4$ commonly used.
- In this example, *LRT*s were performed on the 4 highlighted models and model V was significantly better than the two simpler models and the one more complex model.

Smith and Collyer, 2008. In *The biology of rattlesnakes.*

In the previous example, models were nested and LRT and IC gave the same results.

Burnham and Anderson (1998) believe that ITMS should be performed in place of LRT.

> They provide a 'thought' experiment, where models become more complex, and indicate no change in AIC accompanied with a $k = 8$ change in model parameters, which eventually results in a significant LRT (i.e., the additional 8 parameters caused a significant difference between full and reduced models)

Their logic, however, is flawed because AIC contains model likelihood; thus no change in AIC would require a linear -1:1 change between model likelihood and the parameter penalty described by additional parameters – this is impossible!

With nested models, AIC offers no advantage over LRT for the following reason (using multivariate normal definition of model likelihood):

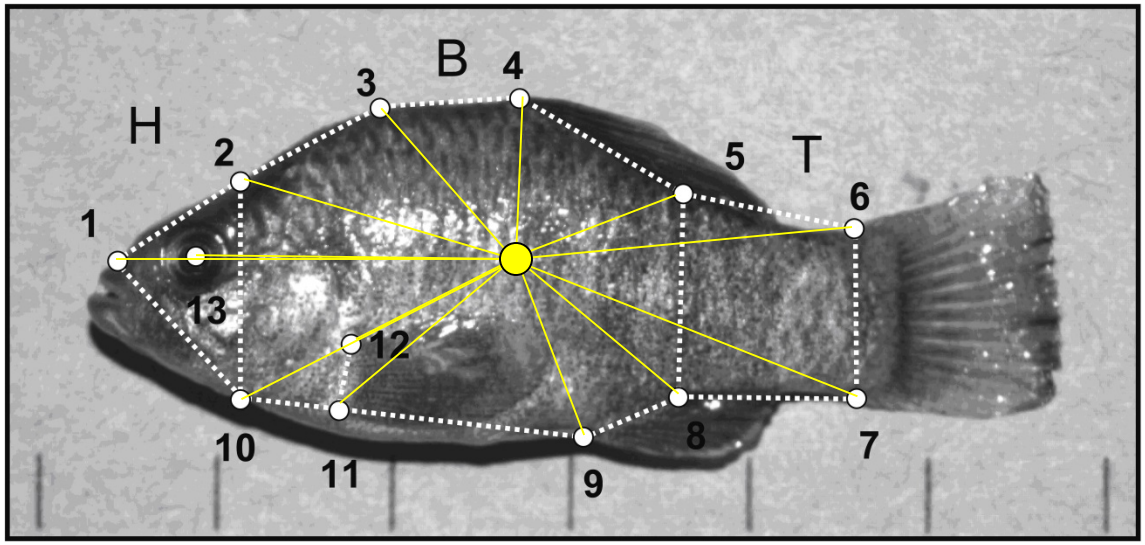
$$\Delta\text{AIC} = \left[n\left[\log\left(\frac{|\boldsymbol{\Sigma}_{full}|}{n^p}\right) + p\right] + 2\left[p(k_{full}) + 0.5p(p+1)\right]\right] - \left[n\left[\log\left(\frac{|\boldsymbol{\Sigma}_{red}|}{n^p}\right) + p\right] + 2\left[p(k_{red}) + 0.5p(p+1)\right]\right]$$

$$= n\left[\log\left(\frac{|\boldsymbol{\Sigma}_{full}|}{n^p}\right) - \log\left(\frac{|\boldsymbol{\Sigma}_{red}|}{n^p}\right)\right] + 2p\left(k_{full} - k_{red}\right)$$

$$= n\log\left[\frac{|\boldsymbol{\Sigma}_{full}|}{|\boldsymbol{\Sigma}_{red}|}\right] + 2p\Delta k \quad \Rightarrow \quad \boxed{\Delta\text{AIC} = \text{LRT} + 2\text{p}\Delta k}$$

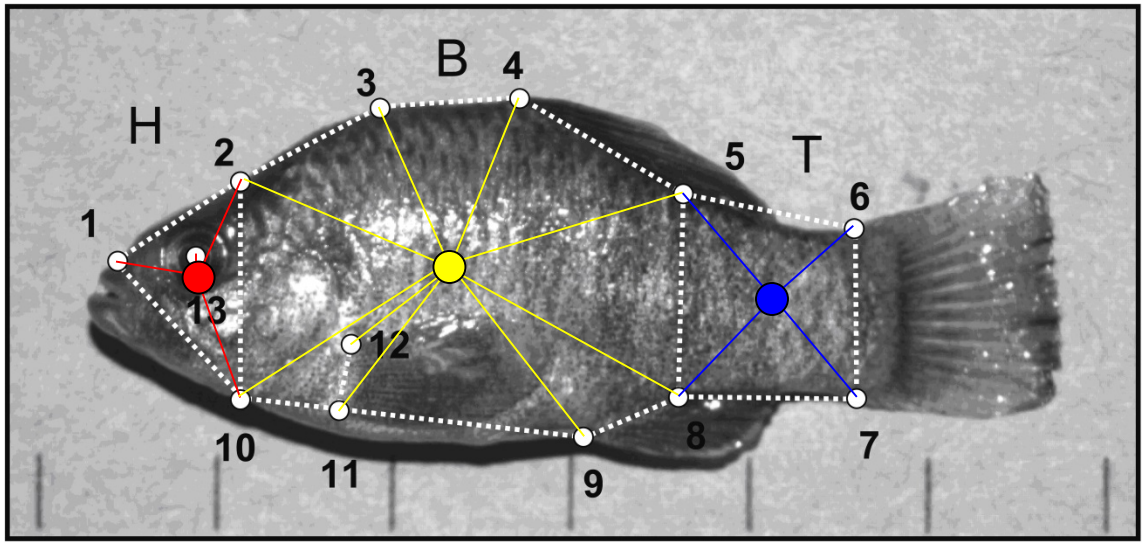
Thus, the difference in AIC values for two models is the LRT plus a parameter penalty!

Non-nested models for pupfish body shape variation



Size = Centroid size (CS), the square root of summed squared distances from landmarks to centroid

CS Total



CS separately for head (H), body (B), and tail (T)

Models of body shape variation:

1) Population differences only
2) Population differences and shape allometry (using overall body size)
3) Population differences and regional shape allometries (using separate measures of H, B, and T size)

* Models 2 and 3 are not nested

Models		Wilks Λ	df_{Model}	df_{Error}	*P*	Δ AIC
1	Population	0.0263389	154	1,177.166	<0.0001	2,412.292
2	Pop. + CS_{Total}	0.0053559	176	1,321.522	<0.0001	2,220.942
3	**Pop. + CS_H + CS_B + CS_T**	**0.0000001**	**220**	**1,586.936**	**<0.0001**	**0**

Conclusion: A model with regional allometries explains much more shape variation *despite the additional parameters needed* for this model!

In this example, a clear difference among models was observed. This is not always the case. Things get tricky when there is *model uncertainty*.

A single best model cannot always be identified: sometimes ΔAIC are 'close'

Model (defined by parameter sets)	ΔAIC
A	3
A + B	0
C	2

One way of looking at the support for different models are Akaike weights:

$$w_i = \frac{\exp\left(-0.5(\Delta\mathrm{AIC}_i)\right)}{\sum_{i=1}^{c}\exp\left(-0.5(\Delta\mathrm{AIC}_i)\right)}$$

Such that $w_1 + w_2 + \ldots + w_C = 1$

Model weight is akin to the probability of model *i* given the set of *C* candidate models

Thus

Model	ΔAIC	w_i
A	3	0.122
A + B	0	0.547
C	2	0.331

Model Averaging:

$\boldsymbol{\beta}^* = 0.122\,\boldsymbol{\beta}_A + 0.547\,\boldsymbol{\beta}_{A+B} + 0.331\,\boldsymbol{\beta}_C$

Note that parameters in A have more weight because they appear in two different models

Model Averaging:

$\boldsymbol{\beta}^* = 0.122\ \boldsymbol{\beta}_A + 0.547\ \boldsymbol{\beta}_{A+B} + 0.331\ \boldsymbol{\beta}_C$

Note that parameters in A have more weight because they appear in two different models

This is VERY bad practice!

The choice of models to include is arbitrary

The result are non-existent coefficients

The likelihood (and AIC) of the average model might be worse than those of candidate models

Model averaging can be *dangerous!*

Consider this example:

Models		F	df_{Model}	df_{Error}	*P*	Δ AIC
1	Region + SVL	5.537	3	103	0.0014	1413.114
2	Region + SVL + Region *SVL	3.826	5	101	0.00324	1414.552

From typical recommendation (ΔAIC < 2), models should be averaged

However, LRT of 1 vs. 2 is not significant (F = 1.228, df = 2, p = 0.2984), so the interaction does not add anything to explaining variance

Also, $AIC_{aver.mod}$ = 1415.718, **worse** than AICs of other models!

Model averaging *NOT* appropriate here (and probably not elsewhere)

Model averaging assumes weights are appropriate, but:

- If a random variable (i.e., an X with no explanatory power) is retained in all models, it receives a weight of 1.0, despite explaining no variation in Y!

-For multivariate, distribution of weights becomes essentially binary

Δk	p = 1	2	3	4	5	6	7	8
0	2	6	12	20	30	42	56	72
1	4	10	18	28	40	54	70	88
2	6	14	24	36	50	66	84	104
3	8	18	30	44	60	78	98	120
4	10	22	36	52	70	90	112	136
5	12	26	42	60	80	102	126	152
6	14	30	48	68	90	114	140	168
7	16	34	54	76	100	126	154	184
8	18	38	60	84	110	138	168	200

Parameter penalty with no change in model likelihood

Akaike's weights are practically binary with $p > 4$

Weight distributions

Δk	p = 1	2	3	4	5	6	7	8
0	0.632199	0.864665	0.950213	0.981684	0.993262	0.997521	0.999088	0.999665
1	0.111476	0.006686	0.000123	0.000001	0.000000	0.000000	0.000000	0.000000
2	0.023951	0.000456	0.000003	0.000000	0.000000	0.000000	0.000000	0.000000
3	0.006950	0.000052	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	0.002014	0.000006	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
5	0.000578	0.000001	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
6	0.000173	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
7	0.000054	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
8	0.000017	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

While intuitive, AIC weights and model averaging provides little insight

Advantages

Can consider non-nested models

Penalizes model likelihood by the number of parameters needed in the model

Limitations

Choice of parameter penalty is arbitrary. Although there are sound statistical reasons to choose one type of information criterion (IC) over another, there are no statistical distributions for ICs, and ICs do not incorporate biological reasoning.

Indices are univariate even if data are multivariate. ICs only measure magnitude of deviation of models, but not direction in the parameter space.

Disadvantages

Not really a statistical test, and proponents of ITMS have argued strongly that statistical tests are useless. However, the statistical theory of LRT is well-founded and does not necessarily differ from IC for nested models.

Model averaging – proceed with extreme caution ***IF AT ALL***

The goal of model selection is to select the best model from a set of candidate models. But what if all the models are garbage?

Cross-validation evaluates the robustness of a particular model

It is related to model selection in that the model that is most robust (from a set of candidates) may be considered the best.

Procedure:

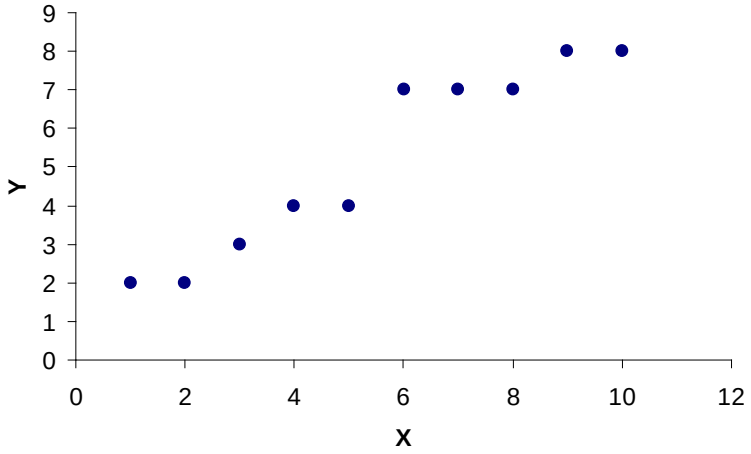
break the data in two subsets

estimate parameters using one subset

use those parameters to predict values in the second subset

measure error

A robust model is one where error is minimized.



Two models: $\hat{y}_i = \bar{y}$

$\hat{y}_i = b_0 + b_1 x_i$

Step 1: Randomly choose 5 values and estimate coefficients for both models

Step 2: Use these coefficients to estimate the value for the other 5 (for both models)

Step 3: Calculate the residuals for the remaining 5 values (for both models)

Step 4: Describe the cross-validation criterion as the sum of squared residuals (SS)

Repeat 1000 times

Results

	Mean (SS)	Var (SS)
Mean model	23.73	44.80
Regression model	1.59	0.95

Regression model obviously more robust

Model selection is a good idea when it is not clear what potential sources of variation exist (i.e., maybe not needed for an experimental study where treatments are controlled).

Model selection might put you in the right ballpark, but it does not make inferences for you.

Never should one abandon biological reasoning, and trust the outcome of naïve model selection (i.e., if you know that a certain parameter needs to be in the model – perhaps organism size – then neither choose a candidate model that lacks this parameter, nor trust an index ranking that excludes it as important).

Model selection is a tool in a large toolbox of statistical and analytical procedures. It is a tool that should probably be used early, followed by more rigorous methods. It SHOULD NOT be the focus of the data analysis.

Although this lecture dealt with regression/ANOVA type linear models, model selection is quite prevalent in other disciplines.

A major use of model selection is in phylogenetics. Model selection can be used to help infer the "best" phylogenetic relationship of taxa:

LRT can be used to compare simpler trees to full trees.

IC can be used to compare trees with different models of evolutionary rates.

CV can be used, removing taxa and estimating trees at each iteration, to evaluate tree-robustness.

IC is also quite popular for Mark/Recapture and population growth studies (see Burnham and Anderson 2002).

IC or LRT with a randomization test might be useful for comparisons of the same model but different least-squares methods of coefficient estimation (e.g., ordinary least-squares versus phylogenetic least squares)