AUTOMATED MODEL SELECTION

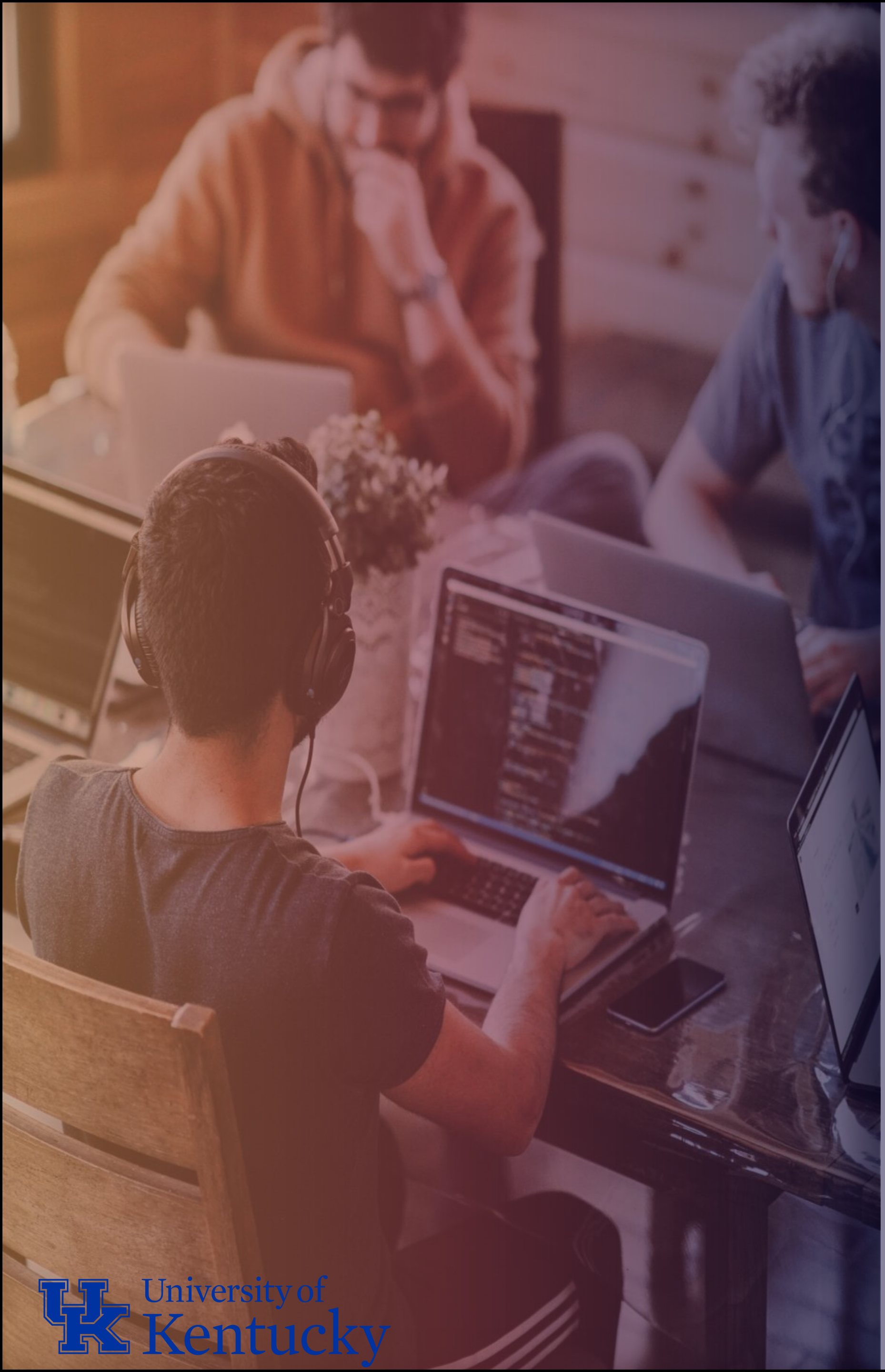
Animal Science

V. L. Daley
National Animal Nutrition Program
Modeling Commitee
Postdoctoral Research Associate

OUTLINE

- Multimodel inference
  - Introduction
  - Framework
    - ✓ Potential candidate variables
    - ✓ Global mixed model
    - ✓ Set of candidate models
    - ✓ Model selection
    - ✓ Evaluation

University of Kentucky



Multimodel Inference

Animal Science

✓Automated **“Model selection”**

- Automated model selection is a procedure to select the best model from a set of candidate models.

✓ **“Multimodel Inference”**

- Information-theoretic approaches
- Formal inference to be based on more than one model

(Burnham and Anderson 1992, 2001, 2002, 2004)

Brief History of Multimodel Inference

1922

R.A. Fisher

1. Model specification, 2. Estimation of parameters,
3. Estimation of precision

1948

Shannon

Mathematical theory of communication.

1951

Kullback–Leibler (K-L information)

Distance between "full reality" and a "model“ The best model loses the least information relative to other models in the set.

1973 1974 1985 1994

Hirotugu Akaike

- Model selection criterion based on K-L information
- AIC is an estimate of the K-L information.
- A set of a priori candidate models, the AIC is computed for each model
- Akaike’s approach allowed model selection

1998 2002…

Burnham and Anderson

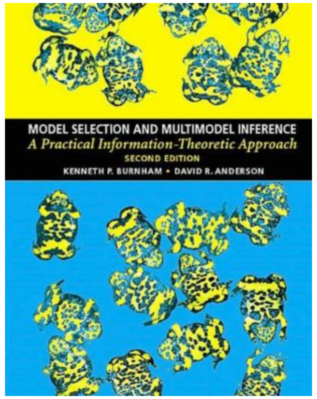
Model Selection, biological science, candidate model, approximate model

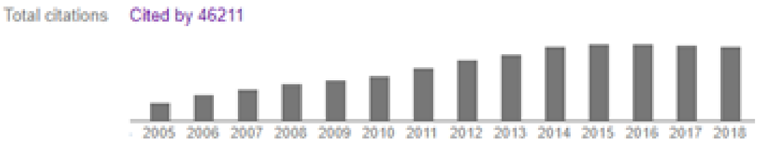
2013…?

Dairy & Animal Science

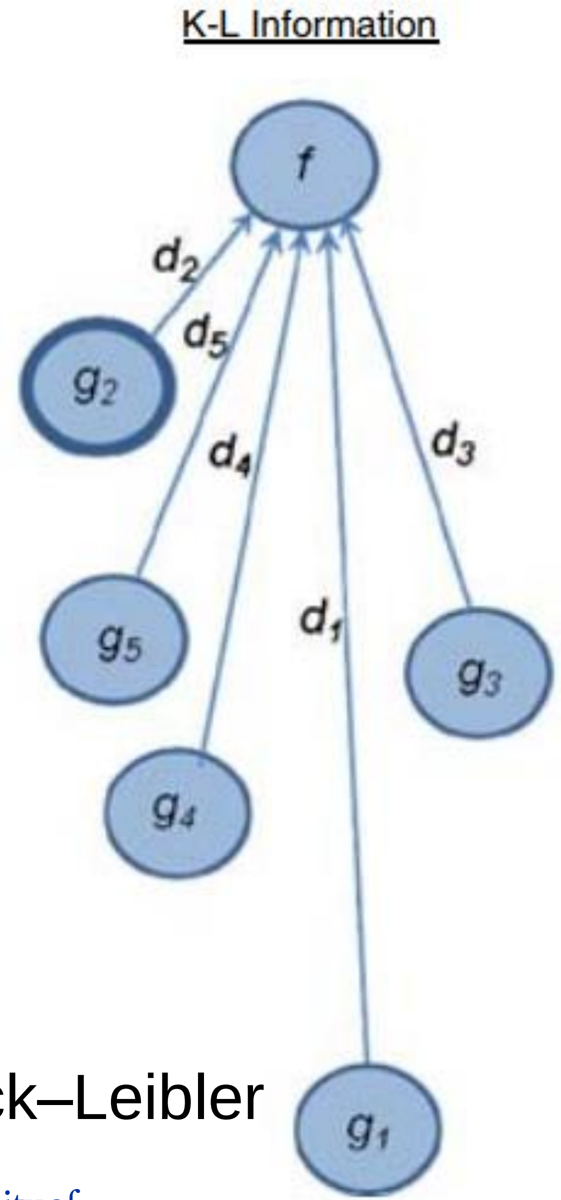
Model Selection and Multimodel Inference

Authors
Kenneth P Burnham
David R Anderson

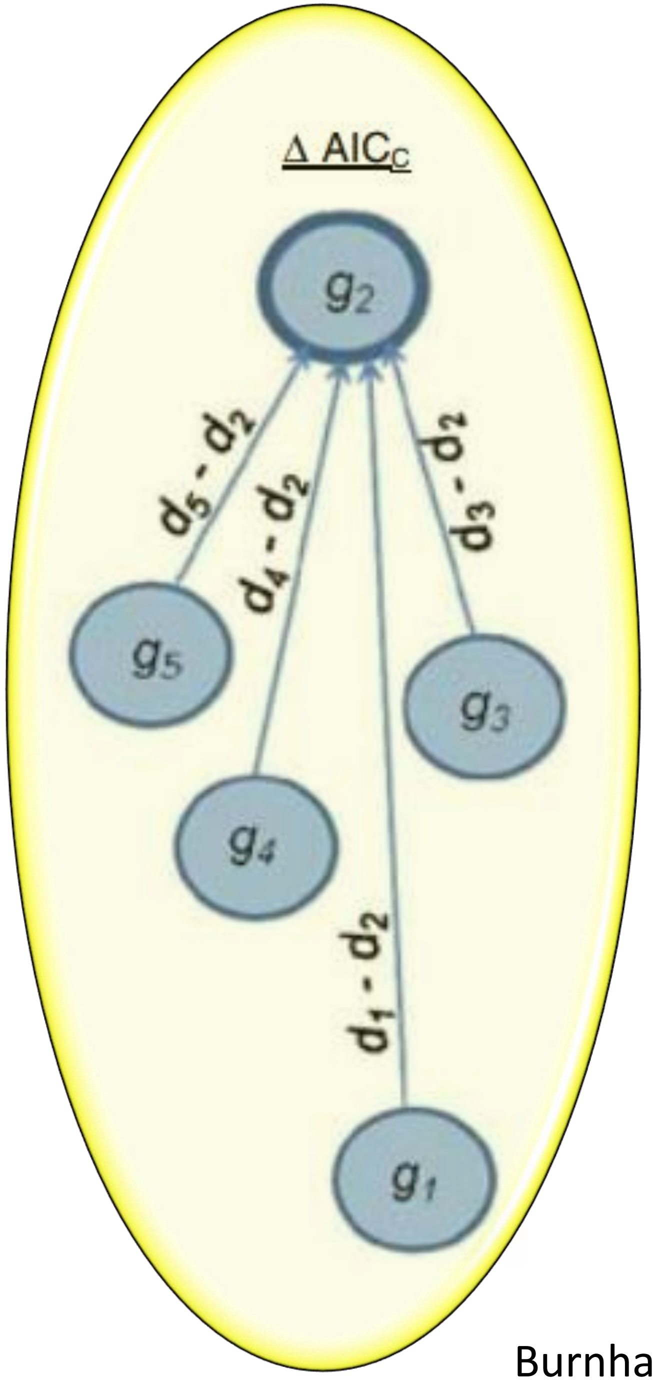




Brief History of Multimodel Inference



Kullback–Leibler



Δ values = the estimated distance of the various models to the best model (model g2).

Burnham et al. Behav Ecol Sociobiol (2011) 65:23–35

AIC versus AICc

- A second order bias correction for AIC
- Sample sizes are small

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

$$\mathrm{AICc} = \mathrm{AIC} + (2K(K+1))/(n-K-1)$$

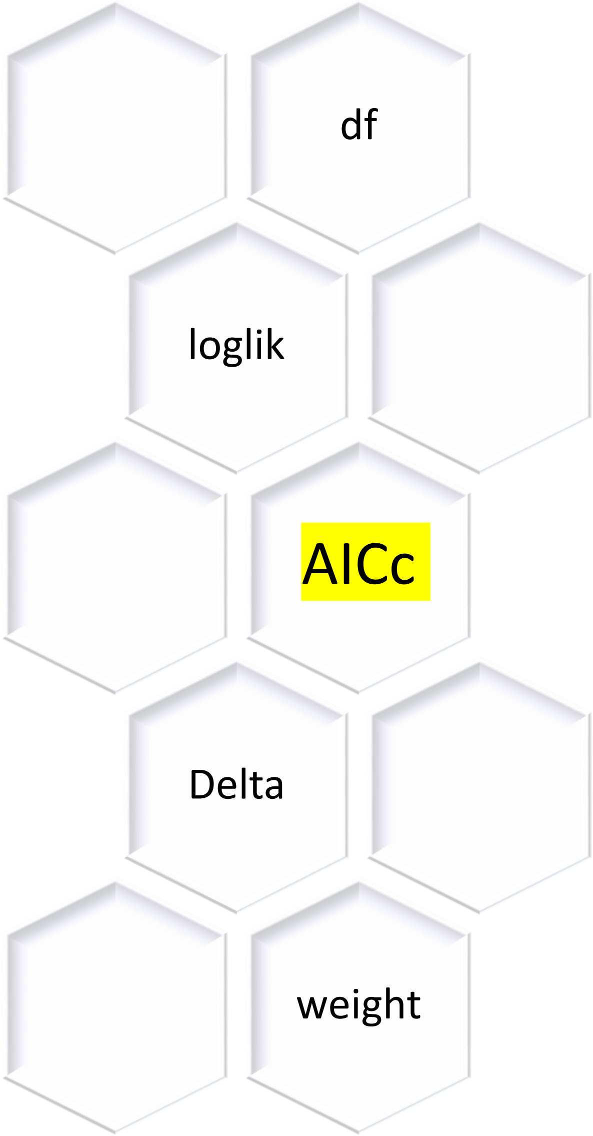
- As sample size (n) increases, AICc converges to AIC.

$\mathcal{L}$ = Likelihood function

K = number of parameters in the model

n = sample size

Small sample sizes (n/K < ≈ 40)



Sugiura (1978) and Hurvich and Tsai (1989)

Datasets in Multimodel Inference

1. Dataset of variables

- ✓ Representative
- ✓ Objective of the study
- ✓ Outliers
- ✓ Biological evaluation

2. Dataset of models

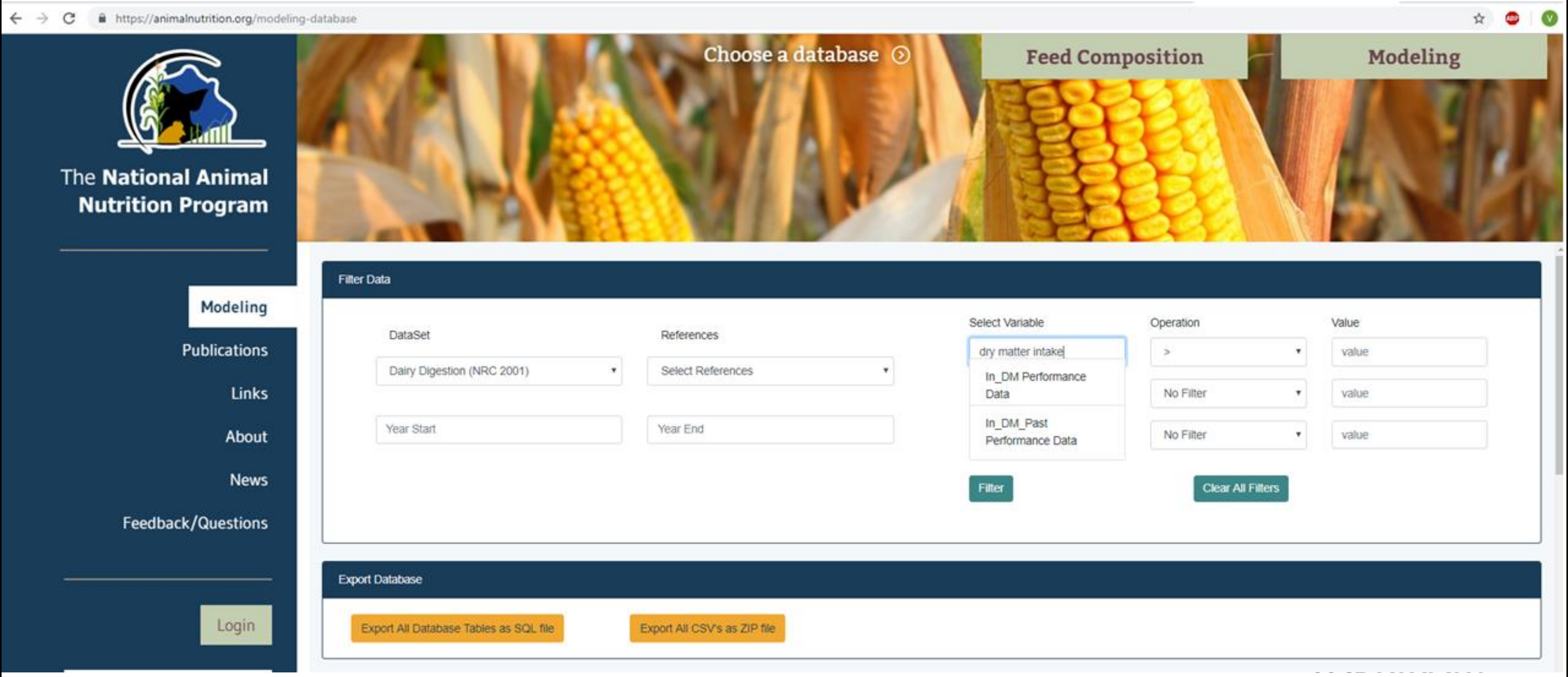
- ✓ Assumed there is a best model (well estimated).
- ✓ Dataset for "Model selection"
- ✓ Selection based on information criterion.
- ✓ Framework and methodology.
- ✓ Inference based on the full set of models.
- ✓ Mathematical and philosophical background.

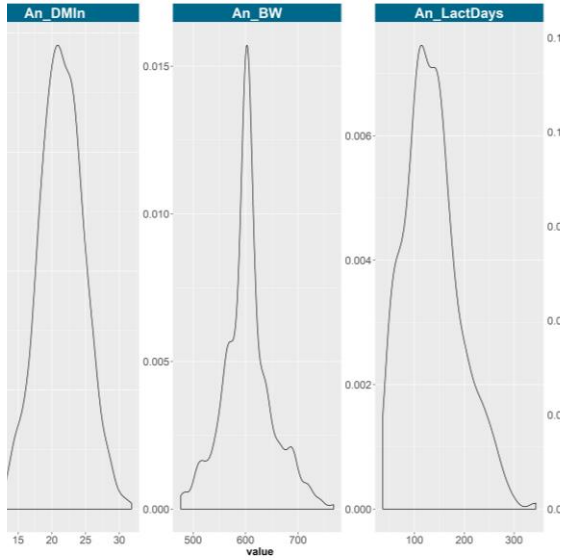
University of Kentucky

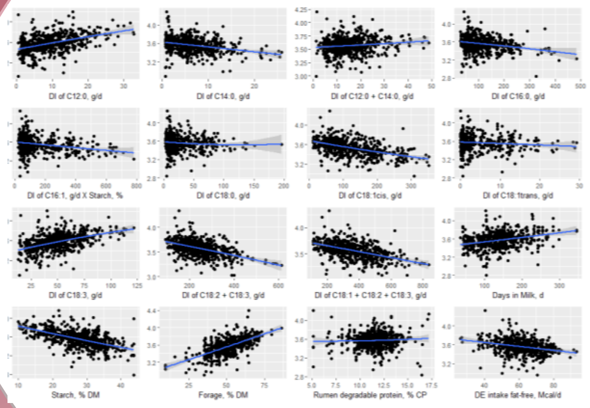
FABRIKAM

Search for Database

https://animalnutrition.org/

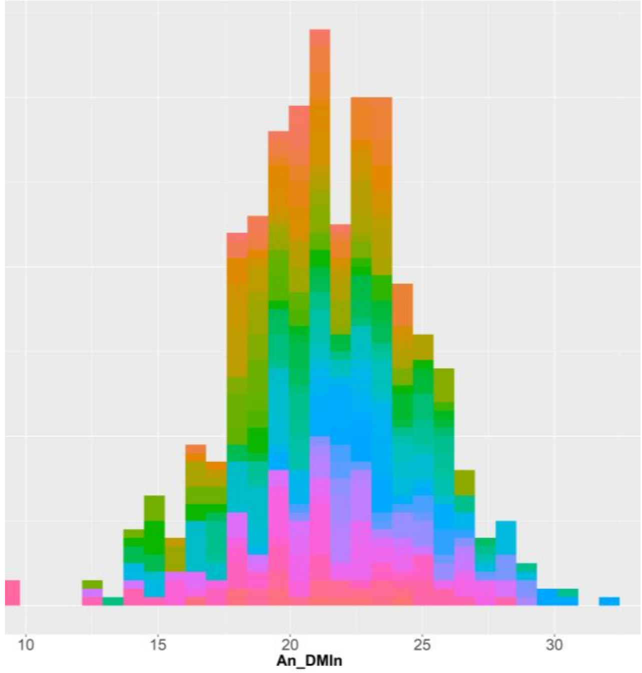






EXAMPLE OF DATA VISUALIZATION

Biological coherence and outliers



University of Kentucky

Example of a Descriptive Statistics Table

	vars	n	mean	sd	median	min	max	range
PubID*	1	645	-	-	-	1	156	155
TID	2	645	-	-	-	-	-	-
An_DMIn	3	645	21.41	3.25	21.32	12.6	30.8	18.2
An_BW	4	645	603.97	46.95	602.7	476	768	292
An_LactDays	5	645	135.79	55.58	129	42	344	302
Obs_MilkProd	6	645	32.47	6.8	32.64	16.8	53.8	37
Obs_MilkFatp	7	645	3.56	0.42	3.57	2.26	4.78	2.52
Obs_MilkPrtp	8	645	3.09	0.21	3.09	2.57	3.9	1.33
Dt_Forage	9	645	50.81	10.29	50	9.61	86.23	76.62
Dt_NDF	10	645	31.95	5.53	31.45	21.41	52.26	30.85
Dt_ForNDF	11	645	23.39	5.59	22.89	5.04	48.32	43.28

MODEL DEVELOPMENT

Automated Model Selection

National Animal Nutrition Program
Leveraging Resources, Linking Researchers

Multimodel Inference Framework

Potential candidate variables → Global Model → Set of candidate models → Model selection → Evaluation

1. Potential candidate variables

- ✓ Potential variables that might or might not appear in the best model
- ✓ Objective of the study
- ✓ Prior knowledge from scientific literature
- ✓ Biologically relevant variables
- ✓ Large or small
- ✓ Power of each variable
- ✓ Association among variables

2. Global Model

- Overparameterization
- Interaction
- Large number of variables
- Fixed and random effects
- Weight

```
lmer(An_DMIn ~ x1 + x2 + x3 + x4 + x5 + x6 +  (1|PubID),
data=d, weights = sqrt(N_study), REML=FALSE)
```

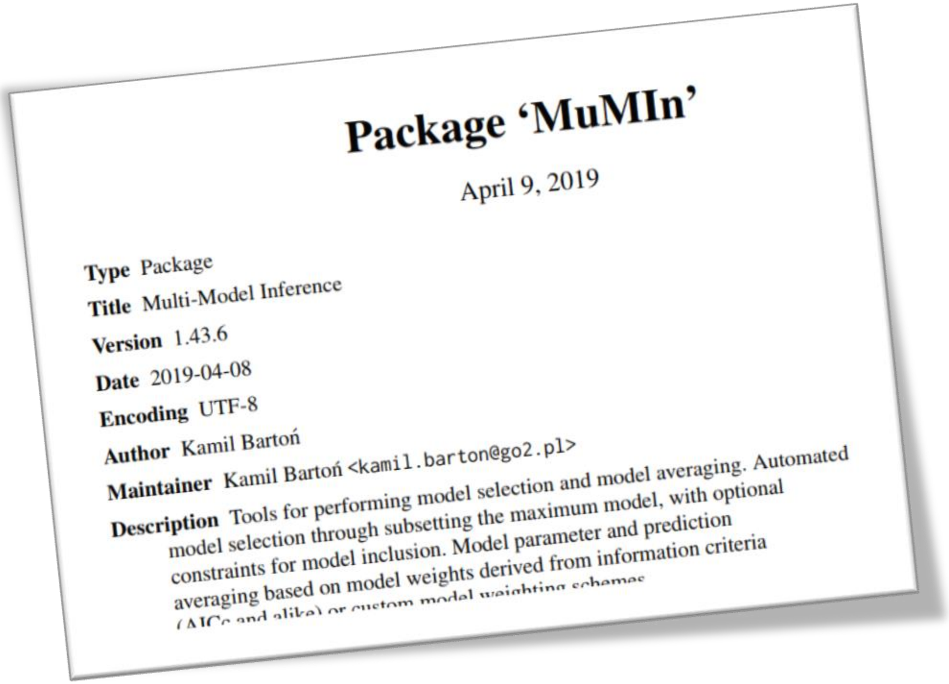
3. Generating a set of models

- **Dredge function**

dredge(**global.model**, beta = c("none", "sd", "partial.sd"), evaluate = TRUE, **rank = "AICc"**, fixed = NULL, m.lim = NULL, m.min, m.max, subset, trace = FALSE, varying, extra, ct.args = NULL, **...**)

- **Pdredge: Parallel Computation**

pdredge(**global.model**, cluster = NA, beta = c("none", "sd", "partial.sd"), evaluate = TRUE, rank = "AICc", fixed = NULL, m.lim = NULL, m.min, m.max, subset, trace = FALSE, varying, extra, ct.args = NULL, check = FALSE, **...**)



Package 'MuMIn'

April 9, 2019

Type Package

Title Multi-Model Inference

Version 1.43.6

Date 2019-04-08

Encoding UTF-8

Author Kamil Bartoń

Maintainer Kamil Bartoń <kamil.barton@go2.pl>

Description Tools for performing model selection and model averaging. Automated model selection through subsetting the maximum model, with optional constraints for model inclusion. Model parameter and prediction averaging based on model weights derived from information criteria

4. Set of Candidate Models

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

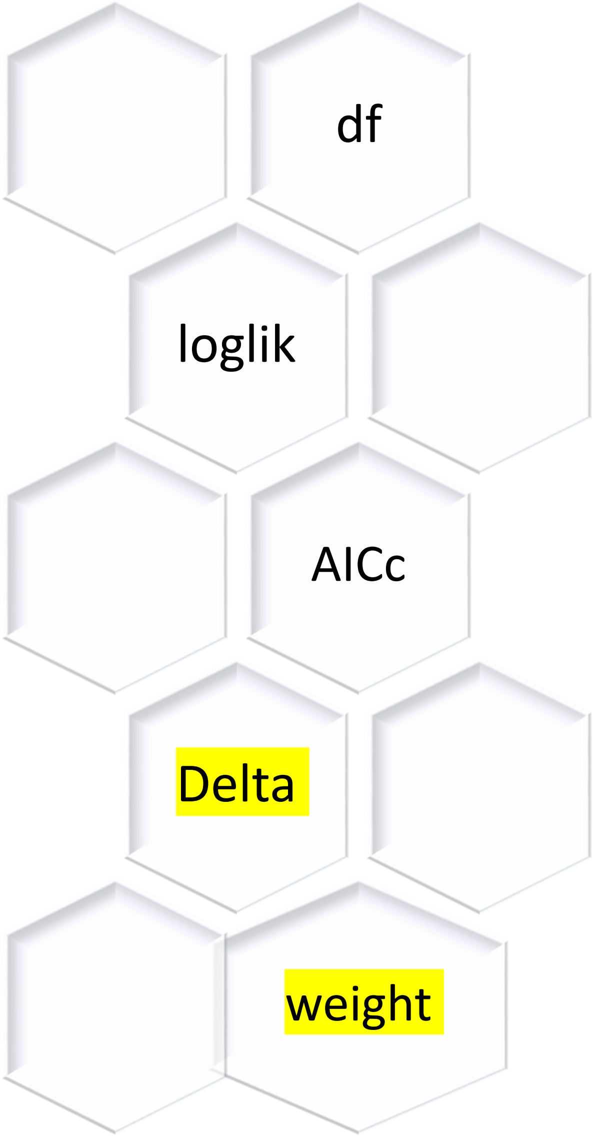
Go to file/function Addins

Exercise #3 R script 1.R allmods

Filter

	(Intercept)	An_BW	An_LactDays	Dt_FA	Dt_Forage	Dt_St	Obs_MilkProd	df	logLik	AICc	delta	weight
47	1.253	Model 1 4	0.013060979	-0.145082657	-1.454809e-02	NA	0.3108421	8	-1000.711	2017.648	0.000000	4.513738e-01
39	0.629	Model 2 3	0.011976634	-0.147986272	NA	NA	0.3166825	7	-1002.578	2019.333	1.684523	1.944226e-01
63	1.203	Model 3 7	0.013050949	-0.144719929	-1.427533e-02	0.001387305	0.3108330	9	-1000.704	2019.691	2.042979	1.625208e-01
55	0.422	Model 4 2	0.012032693	-0.145708250	NA	0.007622286	0.3160340	8	-1002.347	2020.921	3.273039	8.786280e-02

Model selection



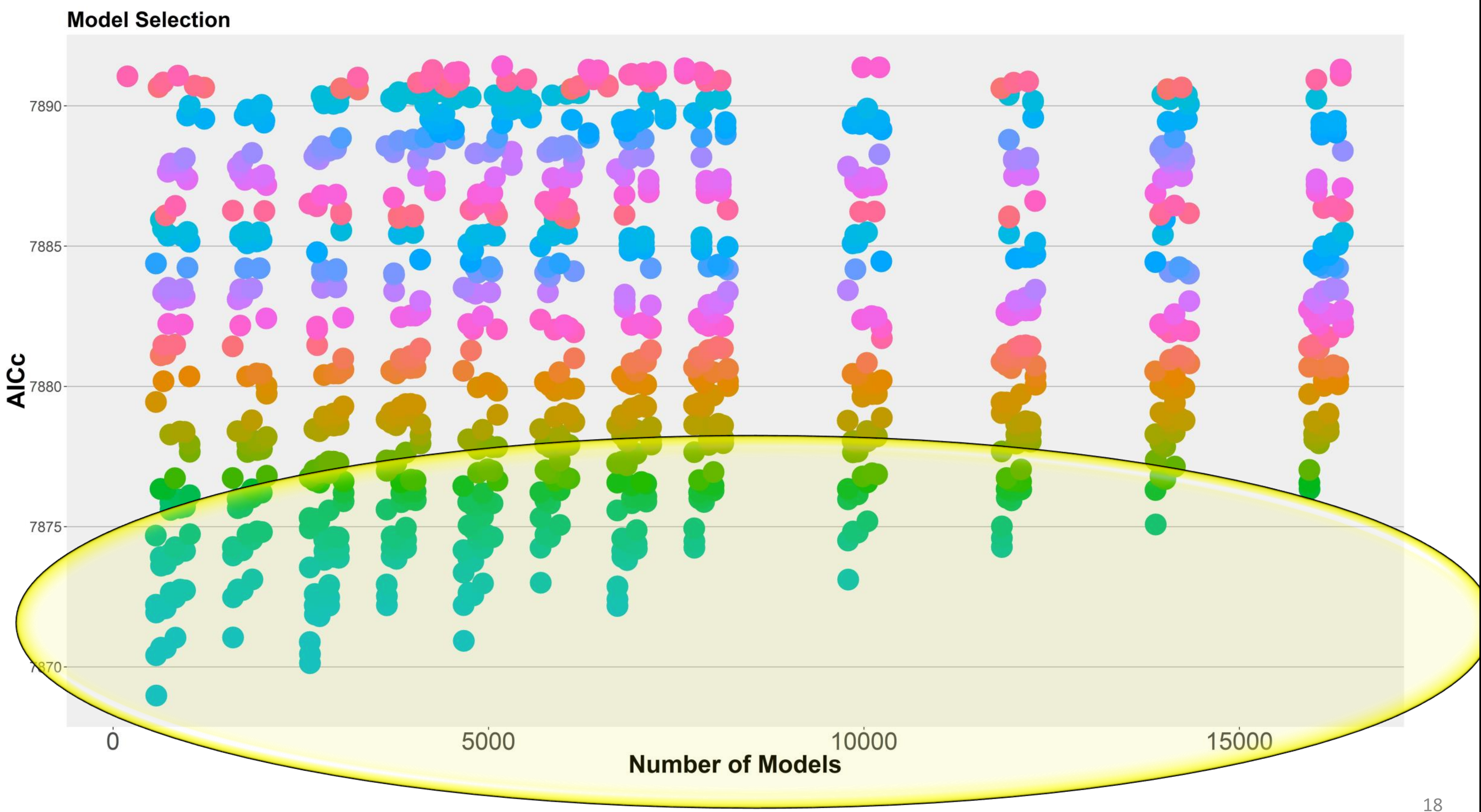
Delta

- AIC differences, relative to the smallest AIC value in the set of models.
- AICi − AICmin
- These values are estimates of the expected K-L information (or distance) between the selected (best) model and the ith model.

Weight

The relative likelihood of the model, given the data. These are normalized to sum to 1, are denoted by wi, and interpreted as probabilities.

Burnham and Anderson (2002)


Model Selection
AICc
7890
7885
7880
7875
7870
0
5000
10000
15000
Number of Models

MODEL SELECTION

Collect All Models

Best Models

AICc

Anova

Test derivated models

Biologic coherence and repeatability

Variance inflation factors (**VIF**)

MODEL EVALUATION

- ✓ Variance inflation factors
- ✓ Concordance correlation coefficient
- ✓ Root mean square error
- ✓ Cross Evaluation

University of Kentucky

MODEL EVALUATION

Biological coherence

Testing on the training data

Repeated cross-evaluation

Multimodel Inference – Key Points

All models are wrong, but some are useful.

Sample size. Large versus small datasets.

Balance between under- and overfitted models.

There is a “best model,” but not a true model!

Model selection:

Priori thinking and biological sense.

It is expected that the results tend to support one or more hypotheses.

TAKEAWAYS

PREVIOUS KNOWLEDGE IS REQUIRED

USEFUL FOR LARGE DATASETS

SELECT THE BEST MODELS BASED ON BOTH BIOLOGICAL SENSE AND INFERENCE ADOPTED

VERIFY THE CONSISTENCY OF ESTIMATED PARAMETERS ACROSS CANDIDATE SET OF MODELS

CROSS EVALUATION

University of Kentucky

THANK YOU

National Animal Nutrition Program
Leveraging Resources, Linking Researchers

VT College of Agriculture and Life Sciences
Virginia Tech

USDA
United States Department of Agriculture
National Institute of Food and Agriculture

Veridiana Daley

veridi7@vt.du

https://www.linkedin.com/in/Veridiana L. Daley

HANDS-ON LESSONS

Practice

Development of empirical models to predict the dry matter intake of dairy cows

1- CLEAN THE DATASET AND PLOT THE VARIABLES.

2- DEVELOP A SET OF CANDIDATE MODELS.

4- SELECT THE BEST 4 MODELS BASED ON THE AIC_C

5- IF TIME ALLOWS, EVALUATE THE BEST 4 MODELS

National Animal Nutrition Program
Leveraging Resources, Linking Researchers

University of Kentucky

THANK YOU

National Animal Nutrition Program
Leveraging Resources, Linking Researchers

VT™

Veridiana Daley

veridi7@vt.du

USDA
United States Department of Agriculture
National Institute of Food and Agriculture

https://www.linkedin.com/in/Veridiana L. Daley