Model Based Inference in the Life Sciences: A Primer on Evidence

David R. Anderson

Model Based Inference in the Life Sciences: A Primer on Evidence

Springer

David R. Anderson
Colorado Cooperative Fish & Wildlife Research Unit
Department of Fish, Wildlife and Conservation Biology
Colorado State University
Fort Collins, CO 80523
USA
anderson@cnr.colostate.edu

ISBN: 978-0-387-74073-7 e-ISBN: 978-0-387-74075-1

Library of Congress Control Number: 2007934534

The photograph on the cover is of Auguste Rodin's bronze sculpture, The Thinker. The statue has become an icon representing intellectual activity and hence, reflects a focus of this textbook. Deep thought is required to hypothesize an array of plausible science hypotheses. Methods are now available to provide a strength of evidence for members of this array and these are simple to compute and understand. It is this hard thinking that is so vital to empirical science. Photo was courtesy of Dick Johnsen.

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

To my daughters, Tamara E. and Adrienne M. Anderson

Preface

This book is about the "information-theoretic" approaches to rigorous inference based on Kullback–Leibler information. My objective in writing this book is to provide an introduction to making rigorous statistical inferences from data and models about hypotheses in the life sciences. The goal of this primer is to explain the information-theoretic approaches with a focus on application. I stress science philosophy as much as statistical theory and I wade into some ideas from information theory because it is so interesting. The book is about hypothesizing science alternatives and providing quantitative evidence for these.

In 1973 Hirotugu Akaike made a world class discovery when he found a linkage between K–L information and statistical theory through the maximized log-likelihood function. Statistical theory developed since the mid-1970s allows science hypotheses, represented by mathematical models, to be ranked from best to worst. In addition, the discrete probability of each hypothesis i, given the data, can be easily computed. These can be viewed as Bayesian posterior model probabilities and are quite important in making inferences about the science hypotheses of interest. The likelihood of each hypothesis, given the data, and evidence ratios between hypotheses i and j are also available, and easy to interpret. All of these new features are simple to compute and understand and go far beyond traditional methods. While many of the examples are biological, I hope students and scientists in other fields (e.g., social sciences, medicine, economics, and many other disciplines) can learn from this primer. Several examples are ecological as that has been my interest; however, the specific examples used are far less important than the science context and trying to understand new approaches; I could not include an example from all of the many subdisciplines.

Tosio Kitagawa (1986) noted that the information-theoretic methods are

> "*… a challenge to conventional statistics as well as a proposal for a new approach to statistical analysis. The reader may find some aspects of the approach controversial insofar as they imply a criticism of conventional mathematical statistics, such as the use of statistical tests, individual sampling distribution theory, and statistical tables*."

I find that some people are still struggling with these new approaches 20 years later. Perhaps this reticence is healthy for science as new ideas must be carefully evaluated and scrutinized; however, we must not let "progress ride on a hearse" either.

I have tried to write this as a science textbook; in a sense it is a companion to the books I have written on this subject with Ken Burnham in 1998 and 2002. Those books contain statistical theory, derivations, proofs, some comparisons with other approaches, and were written at a more advanced level. The present primer tries to be well above a "cookbook" but well below a highly technical treatment; this is a book largely for people new to these effective approaches to empirical science. The book provides a consistent strategy (the concepts of evidence and evolving hypothesis sets) for rapid learning and a way of thinking about science and discovery; a road map of sorts. I provide several examples and many insights on modeling; however, I must say clearly that this is not a primer on modeling.

In the first 4 chapters I cover some material to motivate thinking about plausible science hypotheses (the most important issue), data, information, K–L information, and various measures of evidence and support for members of a set of science hypotheses and their corresponding models. Several examples continue through the chapters as new developments are introduced. At this point, the basics of model based inference under the "information-theoretic" approach will have been laid out. But then, like many good novels – there is a twist. Instead of trying to identify the best science hypothesis (and its model) from the set of hypotheses, I refocus on making formal inference based on all the models – "multimodel inference." In many cases it is desirable to make predictions from all the hypotheses in an *a priori* set – one facet of multimodel inference. These procedures allow model averaging and unconditional measures of precision. Those people thinking this jump will surely be difficult will be pleasantly surprised. The main approaches to multimodel inference under this approach can be understood in 1–2 h of lecture and discussion – they are relatively simple but effective. I hope readers will conceptualize their empirical work in science as multimodel inference. This mental image will help focus on the importance of deriving a set of good, plausible science hypotheses (the hard thinking), gathering quality data, and using modern methods to provide quantitative evidence for each of the science hypotheses of interest.

I want to be quick to say that there are other valid approaches to making inferences from empirical data and I make no effort to deny these. There are general theories related to cross validation, nonparametric statistics, bootstrapping, and

Bayesian approaches to mention only a few. In addition, there are a number of new theories for model selection for linear models; I have omitted reference to these special cases but admit that, with further development, they may someday have wider application. Of the four general theories I noted, only the Bayesian approaches have the breadth and depth of those based on information theory. All have their strengths and I encourage some understanding of these approaches. I will make passing reference to some of these alternatives. I am pro-Bayesian and am interested in areas of commonality between the information-theoretic methods and Bayesian methods. Frequentists and Bayesians have waged a long and protracted philosophical war; I do not want to see the information-theoretic approaches join the conflict.

I consider the various null hypothesis testing approaches to be only of historical interest at this stage (2007), except perhaps in the analysis of data from strict experiments where the design stipulates a single model (i.e., design based inference). In general I think scientists serious about their work must move beyond testing sterile null hypotheses to modern methods and the substantial advantages they provide. I offer several comparisons.

This primer is written to be useful for seniors in excellent undergraduate science programs at top universities. Perhaps more realistically, the book is aimed at graduate students, post-doctoral fellows, as well as established scientists in academia, government agencies, and various science institutes. A basic statistical background is essential to easily understand the material in this book: sampling theory, simple experimental designs, measures of variability and covariability (e.g., sampling variances and covariances, standard errors, coefficients of variation, various approaches to confidence intervals, and sampling correlations), "regression" (e.g., β_i as partial regression coefficients, R^2, residual variance σ^2, residual sum of squares RSS), and goodness-of-fit concepts.

Ideally, the reader would have had some introduction to Fisher's likelihood approaches (e.g., maximum likelihood estimates, profile likelihood intervals). It is hard to understand why there is so much emphasis on least squares approaches even in graduate courses for nonstatistics majors as this narrow approach comes at the expense of the much more general and useful likelihood methods. In addition, likelihood is foundational to the Bayesian approaches. Readers with the required background will find the quantitative issues easy; it is the deeper conceptual issues that will challenge nearly everyone (e.g., model selection bias). This is the fun and rewarding part of science – thinking hard. Readers lacking exposure to null hypothesis testing will find the material here easier to understand than their counterparts. Still, readers should expect to have to reread some material and contemplate the examples given to chase a full understanding of the material.

A *Remarks* section is found near the end of most chapters and some people will find these unordered comments interesting; however, I suggest this material might best be saved for a second reading. This material includes historical comments, technical notes, and other tangential issues that I thought might

interest many readers. In a sense, the *Remarks* are a grouping of what would otherwise be "footnotes," which I often find interesting, but sometimes distracting from the main points. Most chapters end with some nontraditional exercises. Comments on answers to some of the exercises can be found at www.springer.com/978-0-387-74073-7 Each chapter includes a photo and short biography of people who have made major contributions to this literature. I think it is important to recognize and learn from people who came before us and made substantial contributions to science.

This is not an introductory text as I assume a basic knowledge of statistics, the ability to conceptualize science hypotheses (H_i), represent these by mathematical models (g_i), obtain estimates of model parameters (θ), their sampling covariance matrix (Σ), goodness-of-fit tests, and residual analysis. Given this backdrop, new and deeper questions can be asked and answers quantified effectively and simply. I believe this material is fun and exciting if you are a scientist who is serious about advanced work on problems where there are substantial stochasticities and complexities. The material is very broad, but I say less about models for multivariate responses and random effects (as opposed to so-called fixed effects) models.

Many people in the life sciences leave graduate programs with little or no exposure to quantitative thinking and methods and this is an increasingly serious issue, limiting both their contributions to science and their career growth. Many PhD-level people lack any working knowledge of calculus, statistics, matrix algebra, computer programming, numerical methods, and modeling. Some people think that is why they are in biology – "because then I don't have to learn that quantitative stuff." I can certainly understand the sentiment; however, there are ample reasons to reconsider, even later in life. Quantification becomes essential in real world problems as a science matures in a given discipline.

In a sense, undergraduate students are taught a small fraction of material that is *already known* in their field and associated disciplines. People need this background information. Graduate work is quite different (or should be), as students are taught effective philosophies and methods to help them learn how to understand things *new* to their field of science. First one wants to know the current "edge" of knowledge on some issue. Second, one wants to push that edge further as new things are learned from the science process. These are things that cannot be found in a book or on the Internet; the discovery of *new* things – this is what science is all about. Good undergraduate programs try to blur the line between these extremes and this is healthy for science. Graduate programs try to help students shift gears into considering methodologies and philosophies for rapid learning of new things; things that no one has discovered (yet). These are often the harder, more complex issues as our predecessors have solved the easier problems. The information-theoretic approaches represent an effective science strategy and allow one to shift into 6th or even 7th gear and that is what makes learning this material both important and fun.

I wanted to write a short book and try to make some main points that I think are important to people coming to these subjects for the first time. This is a

book about doing empirical science. I do not expect everyone to agree with every philosophical or analytical aspect. Few of the ideas are originally mine as I have taken from thoughts and results from many others as I try to synthesize the more fundamental issues for the reader. This synthesis comes from about 40 years of experience, studying the work of others and trying to form a coherent philosophy about an effective way to do empirical science. I hope people will take what they find useful and be willing to forge ahead in areas where they have found better approaches. I intend to remain interested in this broad subject and will always enjoy hearing comments from colleagues, many of whom I have not yet met.

I want to fully acknowledge my closest friend and colleague over the last 34 years, Ken Burnham. I (reluctantly) wrote this text alone as I am trusting Ken to complete his book on experimental design. Ken has had an powerful influence on my thinking about science philosophy, statistics, information theory, and model based inference. Several other people helped in various ways and I am proud to acknowledge Peter Caley and Jim Hone for their help with my use of their ferret data and Lianne Ball and Paul Doherty for their help with the Palm Springs ground squirrel example. I benefited greatly from extensive review comments offered by Peter Beerli, Barry Grand, Benedikt Schmidt, and Bill Thompson. I also want to thank Bill Gould, Paul Lukacs, Dave Otis, and Eric Stolen for their advice and encouragement. The photo of Thomas Chamberlin was provided by the Edgar Fahs Smith collection at the University of Pennsylvania. John Kimmel at Springer was both patient and encouraging as he helped me through the writing and publishing process.

Fort Collins, CO

David R. Anderson
June, 2007

Contents

About the Author

David R. Anderson received B.S. and M.S. degrees in wildlife biology from Colorado State University and a Ph.D. in theoretical ecology from the University of Maryland. He spent 9 years as a research biologist at the Patuxent Wildlife Research Center in Maryland and 9 years as Leader of the Utah Cooperative Wildlife Research Unit and professor in the Wildlife Science Department at Utah State University. He was a Senior Scientist with the Biological Resources Division within the U.S. Geological Survey, Leader of the Colorado Cooperative Fisheries and Wildlife Research Unit, and a professor in the Department of Fishery and Wildlife Biology at Colorado State University until his retirement in 2003. He has been at Colorado State University since 1984 where he holds an emeritus professorship. He is president of the Applied Information Company in Fort Collins, Colorado. He has published 18 books and monographs, 107 papers in peer reviewed national and international journals, and 45 book chapters, government scientific reports, conference proceedings, and transactions. He is the recipient of numerous professional awards for scientific and academic contributions, including the 2004 Aldo Leopold Memorial Award and Medal.

Glossary

Terms

Akaike weight	The probability that model i is the actual (fitted) K–L best model in the set
Asymptotic	A result or procedure where sample size goes to infinity as a limit
Bias	(Of an estimator) Bias = $\mathrm{E}(\hat{\theta}) - \theta$.
Deductive inference	Reasoning from the general to the particular. Central in logic
Deviance	A fundamental term in likelihood theory. In this book we can usually get by with deviance = -2 log $\mathcal{L}$, that is, negative 2 times the value of the log-likelihood at its maximum point
Effect size	A general term to reflect a measure of the magnitude of some parameter. In a simple experiment, the effect size is often the difference in treatment vs. control means, $\mu_c - \mu_t$. In regression and other models, the effect size is just the (regression) parameter, β_j. In some survival studies, the effect size is defined as the ratio of treatment and control survival probabilities, φ_t / φ_c
Entropy	A measure of disorder or randomness. A highly technical issue as there is more than one form. A short introduction is given in the *Remarks* section in Chap. 3
Estimate	The computed value of an estimator, given a particular set of sample data (e.g., $\hat{\theta} = 9.8$)
Estimator	A function of the sample data that is used to estimate some parameter. A simple example is $\hat{p} = y / n$ for a binomial proportion. An estimator is a random variable and denoted by a "hat" (e.g., $\hat{p}$ or $\hat{\theta}$). Some estimators do

	not have a simple "closed form" and rely on numerical methods to compute their numerical value
Evidence ratio	A ratio of the model probabilities for models i and j in the set, $E_{i,j}$. Used as a quantitative measure of the strength of evidence for any two hypotheses i and j
Global model	Usually the most highly dimensioned model in the set; used primarily for goodness of fit assessment. At least some models in the set are often nested within the global model
iid	Abbreviation for "independent and identically distributed"
Inductive inference	Reasoning from a sample to the population from which the sample was drawn. Central to statistical inference and fundamental to empirical science
Likelihood	A relative value useful in comparing entities. Not a probability as likelihoods do not sum or integrate to 1. Likelihoods are 0 or positive. For example, one can compute the likelihood of various values of p, given the data (n and y) and the binomial model. A single likelihood value is not useful; at least one more value is needed as likelihood values are relative (comparative) to some reference value. Appendix A
Log-likelihood	The natural logarithm of the likelihood function and fundamental in both statistical and information theory
Mean squared error	A measure of performance or accuracy, often in prediction, and defined as the sum of squared bias + variance
Model probability	The discrete probability of model i being the actual best model in terms of K–L information
Negentropy	The negative of entropy, also equal to K–L information
Nested models	A model that is a special case of another model is said to be "nested." A linear model $E(Y) = \beta_0 + \beta_1(x)$ is nested within the quadratic model $E(Y) = \beta_0 + \beta_1(x) + \beta_2(x^2)$
Occam's razor	Taken from thirteenth-century English monk is the well worn idea of the importance of simplicity. The "razor" is the concept "shave away all that is unnecessary"
Parsimony	Classically, this concept is a bias versus variance trade-off. It implies a balancing between the evils of over-fitting and under-fitting. This term should not mean just a "smaller" model as it is sometimes used. Instead, parsimony refers to some trade-off between too few and too many parameters, given a particular sample size. Closely related to Occam's razor

Precision	A property of an estimator related to the amount of variation among estimates from repeated samples. Precision is measured by the sampling variance, standard error, coefficient of variation, and various types of confidence interval. Precision and information are closely related
Predictive mean squared error	Conceptually the expected value of the variance + squared bias. Practically, this can be estimated as $\mathbf{E}[(\hat{Y}_i) - \mathbf{E}(Y_i)]^2$, where $\hat{Y}_i$ is the predicted value from the *i*th sample
Pretending variable	Slang for the case where a model containing an unrelated variable enters the model set with a Δ value of about 2 and is judged as being a "good" model; however, the deviance was not changed. Thus the variable is "pretending" to be important by being in a "good" model, but since the fit was not improved, the variable must be recognized as unimportant. Further evidence of this can be gleaned from examination of the confidence interval for the associated parameter estimate
Probability	Many people consider probabilities to be only long-term frequencies; others (e.g., Bayesians) have the expanded view that probabilities can convey a quantification of belief. In either case, they are nonnegative quantities and sum or integrate to 1 and range between 0 and 1, inclusive

Symbols

AIC	Akaike's Information Criterion, $=-2\log(\mathcal{L}(\theta\|x)) + 2K$ or just $-2\log(\mathcal{L}) + 2K$ in shorthand notation
AIC_{min}	The estimate of expected K–L information for the best model in the set, given the data. For example, given the model set $(g_1, g_2,\ldots, g_R)$ and the data x, if the information criterion is minimized for model g_6, then min = 6, signifying that AIC_6 is the minimum over AIC_1, AIC_2,…, AIC_R. The minimum AIC is a random variable over samples. This notation, indicating the index number in $\{1, 2,\ldots, R\}$ that minimizes expected K–L information, also applies to AICc, QAICc, and TIC
AICc	A second-order AIC, useful when sample size is small in relation to the number of model parameters to be estimated (K). $AICc = -2\log(\mathcal{L}(\theta\|x) + 2K + 2(K(K+1))/(n-K-1)$
β_j	Standard notation for a (partial) regression coefficient ("slopes" relating to the *j*th predictor variable)

BIC	Bayesian Information Criterion (also termed SIC in some literature for Schwarz's information criterion)
$\text{cov}(\hat{\theta}_i, \hat{\theta}_j)$	The sampling covariance of two estimators $\hat{\theta}_i$ and $\hat{\theta}_j$, respectively. This is a measure of codependence and reflects the fact that both estimates, *i* and *j*, come from the same data set and, therefore, might be related (dependent)
c	A simple variance inflation factor used in quasi-likelihood methods where there is overdispersion of count data (e.g., extrabinomial variation). $c \equiv 1$ under independence
Δ_i	AIC differences, relative to the smallest AIC value in the model set. The best model has $\Delta_i \equiv 0$. Formally, $\Delta_i = \text{AIC}_i - \text{AIC}_{\text{min}}$). These values are estimates of the expected K–L information (or distance) between the best (selected) model and the *i*th model. These difference apply to AIC, AICc, QAICc, and TIC
e_i	The *i*th residual in regression analysis, $y_i - \hat{y}_i$
$E(\hat{\theta})$	An operator meaning to take the statistical expectation of the estimator $\hat{\theta}$. Roughly an average of the parameter estimates taken over an infinite number of realizations from the stochastic process that generated the data for a fixed sample size (Appendix B)
$E_{i,j}$	The evidence ratio; the relative likelihood of hypothesis *i* vs. hypothesis *j* or, equivalently, model *i* versus model *j*. A formal measure of the strength of evidence of any two science hypotheses *i* and *j* in the candidate set
$f(x)$	Used to denote hypothetical "truth" or "full reality," the process that produces multivariate data, x. This conceptual or hypothetical "probability distribution" is considered to be infinite dimensional (i.e., an infinite number of "entities," not necessarily what we term "parameters")
$g_i(x)$	Used to denote the model representing science hypotheses *i*. These models are a function of the data (x), thus the notation $g_i(x)$. The set of R candidate models is represented simply as $g_1, g_2, \ldots, g_R$
GOF	Goodness-of-fit test or statistic
H_i	The *i*th science hypothesis
H_o	The "null" hypothesis, the hypothesis tested in null hypothesis testing
H_a	The "alternative" hypothesis associated with null hypothesis testing

K	The number of estimable parameters in an approximating model. Some parameters are confounded with another in some models and are then not "identifiable." In such cases, the parameter count (K) should add 1 parameter for the confounded pair (not 2)
K–L	Kullback–Leibler information (or distance, discrepancy, number)
LS	Least squares method of estimation ("regression")
$\mathcal{L}(\theta \mid x)$	Likelihood function of the model parameters, given the data x
$\mathcal{L}(\theta \mid x, g_i)$	Extended notation to denote the fact that the likelihood function always assumes the data and the specific model g_i are *given*
log(•)	The natural logarithm ($\log_e$). All logarithms in this book are natural (Naperian) logarithms
$\log(\mathcal{L})$	Shorthand notation for the log-likelihood function
$\log(\mathcal{L}(\theta \mid x, g_i))$	Extended notation to denote the fact that the log-likelihood function always assumes the data and the specific model are given
logit(θ)	The logit transform: $\text{logit}(\theta) = \log(\theta/(1-\theta))$, where $0 < \theta < 1$
g_i	Shorthand notation for the candidate models considered. See $g_i(x)$
ML	Maximum Likelihood method of estimation (Appendix A)
MLE	Maximum Likelihood Estimate (or estimator)
n	Sample size. However, some problems do not have a simple sample size as the effective sample size varies by parameter
QAICc	A version of AICc for overdispersed count data where quasi-likelihood adjustments are required, hence $\hat{c}$ is used
θ	Used to denote a generic parameter vector (such as a set of conditional survival probabilities, S_i or a set of regression coefficients, β_i)
$\hat{\theta}$	An estimator of the generic parameter vector θ. Usually these are MLEs. The "hat" denotes an estimate or estimator, rather than the parameter value
$\rho_{x,y}$	The population correlation coefficient between variables x and y
R	The number of candidate hypotheses or models in the set
RSS	The residual sum of squares in least squares methods. Often referred to as the error sum of squares or sum of

	squares due to error (SSE). The RSS is $\Sigma(e_i)^2$ for $i = 1, 2,\ldots, n$
σ^2	The residual variance in "regression." Here I will use the MLE of this quantity; σ^2 = RSS / n and not the more usual "unbiased" LS estimator (i.e., RSS/(n–K)
se or se($\hat{\theta}$)	Standard error and standard error of the estimator $\hat{\theta}$. Used as a measure of precision (or repeatability)
TIC	Takeuchi information criterion
Tr	The matrix trace operator; the sum of the diagonal elements of a square matrix
var($\hat{\theta}$)	The sampling variance of the estimator $\hat{\theta}$. The square root of this quantity is the standard error. Both are measures of precision
w_i	Akaike weights. Used with any of the information criteria that are estimates of Kullback–Leibler information (e.g., AIC, AICc, QAICc, TIC). Estimates of the probability of model i being the K–L best model, given the data and the model set. These are analogous to Bayesian posterior model probabilities
$W_+(j)$	The sum of Akaike weights over all models that contain the explanatory variable j
X or X matrix	The data or matrix of data
$\propto$	A symbol meaning "proportional to"
$\equiv$	A symbol meaning "defined as"
$\approx$	A symbol meaning "approximately equal to"
$\mid$	A symbol meaning that entities to the right are "given" or "conditional upon" or known, as in $\mathcal{L}(\theta \mid x)$
<<	A symbol meaning "much less than"

Definitions of other statistical terms are given by Everitt (1998).