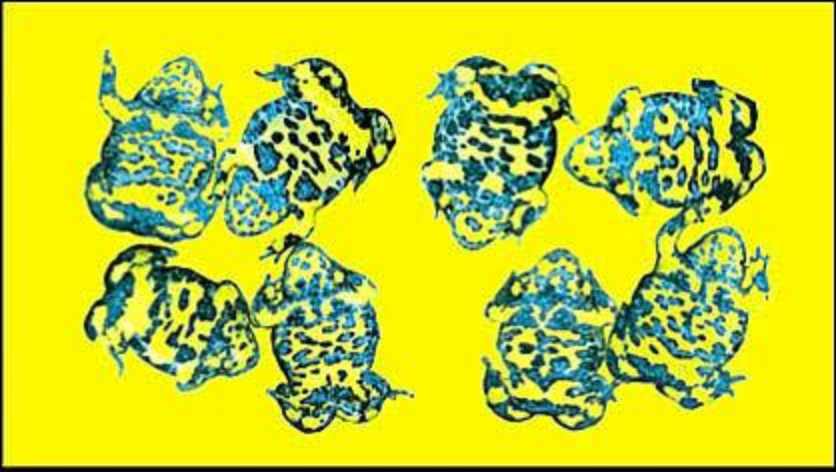
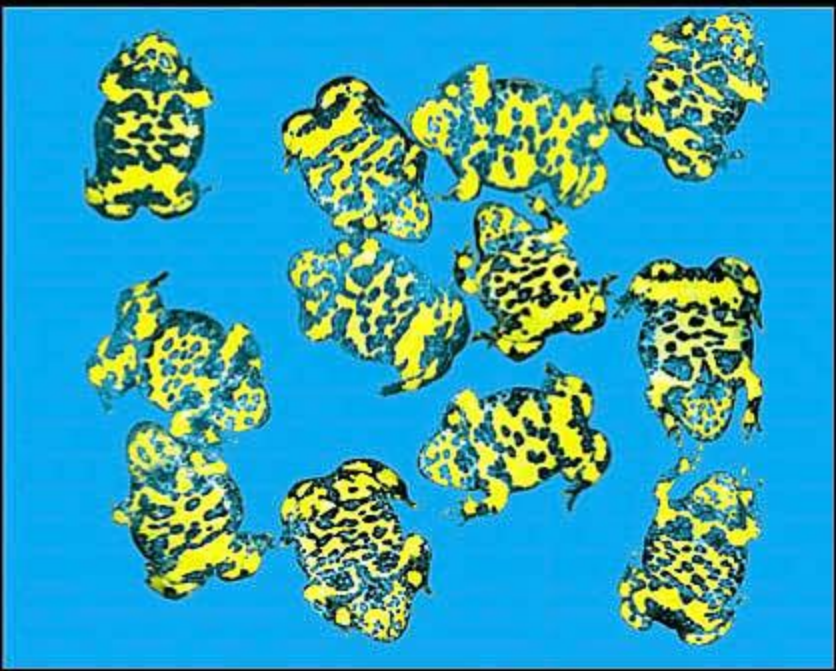


MODEL SELECTION AND MULTIMODEL INFERENCE

A Practical Information-Theoretic Approach

SECOND EDITION

KENNETH P. BURNHAM • DAVID R. ANDERSON



Springer
New York
Berlin
Heidelberg
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Model Selection and Multimodel Inference

A Practical Information-Theoreti

Second Edition

With 31 Illustrations

Springer

and Wildlife Research Unit
Colorado State University
Fort Collins, CO 80523-1484
USA

Cover Illustration: The cover was assembled from photos of the yellow-be
gata) taken by Jonas Barandum as part of his Ph.D. program at the Univers
have individually identifiable patterns on their abdomen from a few weeks
that remain unchanged until death. Two pairs are duplicates—but which tw
Cover photographs by Dr. Jonas Barandum, St. Gallen, Switzerland. Cover

Library of Congress Cataloging-in-Publication Data
Burnham, Kenneth P.
Model selection and multimodel inference : a practical information-th
/ Kenneth P. Burnham, David R. Anderson.—2nd ed.
p. cm.
Rev. ed. of: Model selection and inference. © 1998.
Includes bibliographical references (p.).
ISBN 0-387-95364-7 (alk. paper)
1. Biology—Mathematical models. 2. Mathematical statistics. I. B
Model selection and inference. II. Title.
QH323.5 B87 2002
570′.1′51—dc21

ISBN 0-387-95364-7 Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2 SPIN 10853081

www.springer-ny.com

Springer-Verlag New York Berlin Heidelberg
A member of BertelsmannSpringer Science+Business Media GmbH

To my mother and father, Lucille R. (deceased) and J
(deceased), and my son and daughter, Shawn P. and

To my parents, Charles R. (deceased) and Leta M. A
Dalene F. Anderson; and my daughters, Tam
Adrienne M. Anderson

This page intentionally left blank

Preface

We wrote this book to introduce graduate students and
various scientific disciplines to the use of information-the
the analysis of empirical data. These methods allow the
of a "best" model and a ranking and weighting of the r
a pre-defined set. Traditional statistical inference can th
selected best model. However, we now emphasize that in
approaches allow formal inference to be based on more th
timodel inference). Such procedures lead to more robust
cases, and we advocate these approaches throughout the

The second edition was prepared with three goals in r
tried to improve the presentation of the material. Boxes r
tial expressions and points. Some reorganization has been
flow of concepts, and a new chapter has been added. Ch
been streamlined in view of the detailed theory provided
ond, concepts related to making formal inferences from r
(multimodel inference) have been emphasized throughou
ticularly in Chapters 4, 5, and 6. Third, new technical mat
to Chapters 5 and 6. Well over 100 new references to the
are given. These changes result primarily from our exper
several seminars, workshops, and graduate courses on ma
tion. In addition, we have done substantially more thinkin
reading the literature since writing the first edition, and
led to further insights.

Information theory includes the celebrated Kullback–L
tween two models (actually, probability distributions), a

maximized log-likelihood. His measure, now called *Akai*
terion (AIC), provided a new paradigm for model selecti
empirical data. His approach, with a fundamental link to
is relatively simple and easy to use in practice, but littl
classes and far less understood in the applied sciences tha

We do not accept the notion that there is a simple "true r
ical sciences. Instead, we view modeling as an exercise i
of the explainable information in the empirical data, in th
being a sample from some well-defined population or proc
views modeling as a fabric in the tapestry of science. Se
proximating model represents the inference from the da
"effects" (represented by parameters) can be supported b
on Akaike's information criterion (and various extension
parsimonious model as a basis for statistical inference. M
on information theory represents a quite different appro
sciences, and the resulting selected model may differ subs
selection based on some form of statistical null hypothesi

We recommend the information-theoretic approach for
from observational studies. In this broad class of studies, w
ious hypothesis-testing approaches have no theoretical ju
often perform poorly. For classic experiments (control–
domization and replication) we generally support the tra
(e.g., analysis of variance); there is a very large literature or
However, for complex experiments we suggest considerati
tory models, hence on estimation of the size and precis
effects and on parsimony, with far less emphasis on "tes
ses, leading to the arbitrary classification "significant" vers
Instead, a strength of evidence approach is advocated.

We do not claim that the information-theoretic methods
best for a particular situation. They do represent a unified
an extension of likelihood theory, an important applica
theory, and they are objective and practical to employ acros
empirical problems. Inference from multiple models, or the
"best" model, by methods based on the Kullback–Leibler
certainly better than other methods commonly in use now (
testing of various sorts, the use of R^2, or merely the use
model). In particular, subjective data dredging leads to ov
the attendant problems in inference, and is to be strongly
in more confirmatory studies.

Parameter estimation has been viewed as an optimiza
least eight decades (e.g., maximize the log-likelihood or m
sum of squared deviations). Akaike viewed his AIC and
"... a natural extension of the classical maximum likelih

beyond merely the computation and interpretation of AIC
nious model for inference from empirical data; it refocuse
on a variety of considerations and modeling prior to the ac
Model selection, under the information-theoretic approac
tempts to identify the (likely) best model, orders the r
worst, and produces a weight of evidence that each mode
an inference.

Several methods are given that allow model selection un
porated into estimates of precision (i.e., multimodel infer
is to present and illustrate a consistent methodology that
lation, model selection, estimation of model parameters a
in a unified manner, under a compelling common framev
explain other information criteria (e.g., AIC_c, $QAIC_c$, a
several examples to illustrate various technical issues, i
parisons with BIC, a type of dimension consistent criter
provide many references to the technical literature for th
further on these topics.

This is an applied book written primarily for biologi
using models for making inferences from empirical data
science book; we say relatively little about decision makir
management science. Research biologists working either
laboratory will find simple methods that are likely to be u
gations. Researchers in other life sciences, econometrics,
and medicine might also find the material useful but wi
examples that have been taken largely from ecological stu
vertebrates, as these are our interests. Applied statistician
information-theoretic methods presented here quite useful
native to the null hypothesis testing approach that has bec
uninformative. We hope material such as this will find its
where applied data analysis and associated science philoso
book might be useful as a text for a course for students w
rience and education in statistics and applied data analysi
audience includes honors or graduate students in the bio
statistical sciences. Those interested in the empirical scien
terial useful because it offers an effective alternative to (
yet often both complex and uninformative, null hypothesi
and (2) the far less taught, but potentially very useful, Ba

Readers should ideally have some maturity in the quan
experience in data analysis. Several courses in contempor
and methods as well as some philosophy of science would
ful in understanding the material. Some exposure to likeli
essential, but those with experience only in least squares
will gain some useful insights. Biologists working in a

conceptual issues may preclude beginners. Chapters 1–4 a
all readers because they provide the essential material, in
multimodel inference. Chapters 5 and 6 present more d
some new research results. Few readers will be able to
presented here after just one reading of the material; som
ditional consideration will often be necessary to understan
Underlying theory is presented in Chapter 7, and this mat
and more mathematical. A high-level summary of the ma
is provided in Chapter 8.

We intend to remain active in this subject area after thi
been published, and we invite comments from colleague
learn more and understand differing points of view. We ho
not appear too dogmatic or idealized. We have tried to syn
we believe are important and incorporate these as recomm
in several of the chapters. This book is an effort to exp
multimodel inference in some depth. We realize that there a
and that some people may still wish to test null hypothe
building models of empirical data, and that others may
attitude toward data dredging than we advocate here. We
other model selection methods, such as cross-validation
of Bayesian methods. Indeed, we just learned (March, 20
derived as a Bayesian result and have added a note on this is
the final page proofs (see Section 6.4.5). However, in the
science, we are compelled by the a priori approach of buildi
to represent research hypotheses, the use of information
a basis for selecting a best approximating model; model
multimodel inference methods, when truth is surely very
likelihood theory for deriving parameter estimators; and
selection uncertainty into statistical inferences. In particu
moving beyond mere selection of a single best model by
methods of multimodel inference.

Several people have helped us as we prepared the two e
In particular, we acknowledge C. Chatfield, C. Hurvich,
J. Rotella, R. Shibata, and K. Wilson for comments on
original manuscript. We are grateful to three anonymous
ments that allowed us to improve the first edition. D. Oti
served as the reviewers for the second edition and offere
that were helpful; we greatly appreciate their excellent su
cussions with S. Buckland, R. Davis, R. Shibata, and
useful. S. Beck, K. Bestgen, D. Beyers, L. Ellison, A. Fra
B. Lubow, C. McCarty, M. Miller, and T. Shenk provide
sights as part of a graduate course on model selection me
from the authors. C. Flather allowed us to use his data

by M. Zablan under the supervision of G. White. C. Sou
use his kangaroo data from Wallaby Creek. P. Lukacs con
analysis and some of the Monte Carlo studies of the body
J. Kullback allowed us to use a photo of his father, and H
R. Shibata, and K. Takeuchi kindly sent us photos and b
that appear in the book. Chelsea Publishing Company all
photo of L. Boltzmann from the book *Wissenschaftliche
Ludwig Boltzmann*, and the International Biometric Soc
use of a photo of R. Fisher (from *Biometrics* 1964, taken
ton). J. Barandun provided the toad photos for the cover,
the cover design, and B. Schmidt helped in coordination.
S. Kane, B. Klein, A. Lyman, and T. Sundlov helped obta
J. Kimmel and L. Farkas helped in countless ways as we pr
of this book.

We are happy to acknowledge the long-term cooperators
operative Fish and Wildlife Research Unit: the Colorado
Colorado State University, the Biological Resources Divis
logical Survey, and the Wildlife Management Institute. G
faculty within the Department of Fisheries and Wildlife
State University provided a forum for our interests in th
ical data. We extend our appreciation to several federal
Department of the Interior, particularly the U.S. Geologi
support of our long-term research interests.

Fort Collins, Colorado

This page intentionally left blank

Contents

About the Authors

Drs. Kenneth P. Burnham and David R. Anderson have wo
for the past 28 years and have jointly published 9 books
graphs and 66 journal papers on a variety of scientific is
are both in the Colorado Cooperative Fish and Wildlife R
orado State University, where they conduct research, tea
and mentor graduate students.

Ken Burnham has a B.S. in biology and M.S. and Ph.
tics. For 29 years post-Ph.D. he has worked as a statis
developing statistical theory in several areas of life scien
ogy and wildlife, most often in collaboration with sub
Ken has worked (and lived) in Oregon, Alaska, Marylan
Research Center), North Carolina (U.S. Department of
Carolina State University, Statistics Department), and
USGS-BRD). He is the recipient of numerous professio
Distinguished Achievement Medal from the American Sta
and Distinguished Statistical Ecologist Award from INT
Congress of Ecology). Ken is a Fellow of the American Sta

David Anderson received B.S. and M.S. degrees in wi
Ph.D. in theoretical ecology. He is currently a Senior Sc
logical Resources Division within the U.S. Geological Su
in the Department of Fishery and Wildlife Biology. He
Patuxent Wildlife Research Center in Maryland and 9 y
Utah Cooperative Wildlife Research Unit and professor in
Department at Utah State University. He has been at Colo
since 1984. He is the recipient of numerous professional
and academic contributions, including the Meritorious Se
U.S. Department of the Interior.

This page intentionally left blank

Glossary

Notation and abbreviations generally used are given bel
for specific examples can be found in those sections.

AIC	Akaike's information criterion.
AIC_{min}	The estimate of relative, expected K-L inf model in the set, given the data. For examp $g_1, g_2, \ldots, g_R$ and the data x, if the inf minimized for model g_6, then $min = 6$, is the minimum over $\text{AIC}_1, \ldots, \text{AIC}_R$. T a random variable over samples. This no index number in $\{1, 2, \ldots, R\}$ that mini information, also applies to AIC_c, QAIC
AIC_{best}	In any set of models, one will be the model, hence the actual best AIC model. $E_f(\text{AIC})$ is minimized is denoted by the *min* is a random variable (like $\hat{\theta}$), *best* i value can be determined using Monte "best" model is the same model over (of which we have only a single sample applies to AIC_c, QAIC_c, and TIC.
AIC_c	A second-order AIC, necessary for small
Akaike weights	The relative likelihood of the model, give normalized to sum to 1, are denoted by u probabilities.

	the data. Such a *best* model can be found methods and represents a statistical expec consider the set $\mathrm{E}(\mathrm{AIC}_i)$, where $i = 1, 2$ model where $\mathrm{E}(\mathrm{AIC}_i)$ is minimized is AIC, AIC_c, QAIC_c, or TIC could be used
Bias	(of an estimator) Bias $= \mathrm{E}(\hat{\theta}) - \theta$.
BIC	Bayesian information criterion (Akaike 1978), also termed SIC in some literature
c	A simple variance inflation factor used methods where there is overdispersion extra binomial variation).
Δ_i	AIC differences, relative to the smallest A R models. Hence, AIC values are rescaled constant such that the model with the m has $\Delta_i = 0$. Formally, $\Delta_i = \mathrm{AIC}_i - \mathrm{A}$ are estimates of the expected K-L infor between the selected (best) model and t differences apply to AIC, AIC_c, QAIC_c,
Δ_p	A "pivotal" value, analogous to $(\theta$ $\mathrm{AIC}_{best} - \mathrm{AIC}_{min}$.
df	Degrees of freedom as associated with hy df is the difference between the number null and alternative hypotheses in stand tests.
$\mathrm{E}(\hat{\theta})$	The statistical expectation of the estimat
Estimate	The computed value of an estimator, giv sample data (e.g., $\hat{\theta} = 9.8$).
Estimator	A function of the sample data that is us parameter. An estimator is a random var by a "hat" (e.g., $\hat{\theta}$).
Evidence ratio	The relative likelihood of model i ver $\mathcal{L}(g_i \mid data)/\mathcal{L}(g_j \mid data)$, which is identica
$f(x)$	Used to denote "truth" or "full reality produces multivariate data x. This co distribution is often considered to be a infinite-dimensional space.
$g_i(x)$	Used to denote the set of candidate mod sized to provide an adequate approximatio of empirical data. The expression $g_i(x \mid$ necessary to clarify that the function inv

	R candidate models is represented as $\hat{g}_i$. Also, $\hat{g}_i = g_i(x\|\hat{\theta})$.
Global model	A highly parameterized model containi associated parameters thought to be impo an a priori consideration of the problem is a global model, all other models in the of this global model.
K	The number of estimable parameters i model.
K-L	Kullback–Leibler distance (or discre number).
LRT	Likelihood ratio test.
LS	Least squares method of estimation.
$\mathcal{L}(\theta \mid x, g)$	Likelihood function of the model param x and the model g.
$\mathcal{L}(g_i\|x)$	The discrete likelihood of model g_i, give
$\log(\cdot)$	The natural logarithm ($\log_e$).
$\text{logit}(\theta)$	The logit transform: $\text{logit}(\theta) = \log($ $0 < \theta < 1$.
g_i	Shorthand notation for the candidate mo
min	An index to denote the fitted model tha formation criterion, given the data. The model selected, based on minimizing th rion, given the data. AIC, AIC_c, QAIC_c, in this context.
ML	Maximum likelihood method of estimati
MLE	Maximum likelihood estimate (or estima
n	Sample size. In some applications there m relevant sample size (e.g., in random effe
Parsimony	The concept that a model should be as concerning the included variables, model ber of parameters. Parsimony is a desire model used for inference, and it is usual able tradeoff between squared bias and va estimators. Parsimony lies between the over-fitting.
Precision	A property of an estimator related to the among estimates from repeated samples.

	quasi-likelihood adjustments are required
π_i	Model selection probabilities (or relative
	from Monte Carlo studies or the bootstra
R	The number of candidate models in the s
	One of these models is the estimated be
	sense of a specific model $g(x\|\hat{\theta})$, where
	eters have been estimated) for the data
	model (possibly the same model) is th
	model (g_{best}) to use as a basis for inferen
τ_i	Prior probability of model i. Also used
	redundancy (Section 4.6).
θ	Used to denote a generic parameter vec
	conditional survival probabilities S_i).
$\hat{\theta}$	An estimator of the generic parameter θ.
θ_0	The optimal parameter value in a given m
	sample size, but ignoring estimation issu
	This is the value that minimizes K-L inf
	model structure.
TIC	Takeuchi's information criterion.
w_i	Akaike weights. Used with any of the info
	are estimates of expected Kullback–Leibl
	AIC_c, QAIC, TIC). The w_i sum to 1 and
	as the probability that model i is the actu
	model for the sampling situation conside
$w_+(j)$	Sum of Akaike weights over all mode
	explanatory variable j. These sums are
	selection problems where one wants a
	importance of the explanatory variable
	estimates that are robust to model selecti
χ^2	A test statistic distributed as chi-square
	grees of freedom df. Used here prima
	goodness-of-fit test of the global model
	data.
$\approx$	Approximately equal to.
$\sim$	Distributed as.

Introduction

1.1 Objectives of the Book

This book is about making valid inferences from scientifi
ingful analysis depends on a model of the information in t
objective is to provide scientists, including statisticians,
giving practical advice for the analysis of empirical data u
theoretic paradigm. We first assume that an exciting scienti
carefully posed and relevant data have been collected, fo
perimental design or probabilistic sampling program. Alt
and models to represent them, should be carefully cons
stage of the investigation. Often, little can be salvaged i
been seriously flawed or if the question was poorly posed
alize, of course, that these issues are never as ideal as one w
proper attention must be placed on the collection of data (C
Anderson 2001). We stress inferences concerning the st
of biological systems, relevant parameters, valid measu
formal prediction.

There are many studies where we seek an understanding
pecially causal ones. There are many studies to understan
are important because of the parameters in them and rela
between and among variables. These parameters have re
pretations, even when they relate to quantities that are no
(e.g., survival probabilities, animal density in an area, ge
interaction terms). Science would be very limited without

ters. Observables and prediction are often critical, but sel
these issues.

The first objective of this book is to outline a consis
sues surrounding the analysis of empirical data. Inducti
statistical inference about a defined population or process
sample or experimental data set. "Data analysis" leadin
is the integrated process of careful a priori model formu
tion, parameter estimation, and measurement of precision (
component due to model selection uncertainty). We do no
selection should be treated as an activity that precedes
model selection is a critical and integral aspect of scienti
leads to valid inference.

A philosophy of thoughtful, science-based, a priori mo
Often, one first develops a global model (or set of mode
several other plausible candidate (sub)models postulated t
proximations to information in the data at hand. This form
models. Science and biology play a lead role in this a prior
careful consideration of the problem. A simple example of
alternative scientific hypotheses might be helpful at this e
the importance of an interaction between age (a) and w
a particular animal population. A model including such a
have the main effects plus the interaction; $a + w + a *$
$a + w$ lacks the interaction term. Information-theoretic m
lines of quantitative evidence concerning the importance
interaction.

The modeling and careful thinking about the problem
that have often received relatively little attention in statistic
for nonmajors), partly because such classes rarely conside
or philosophy of data analysis. A proper a priori model-bu
to avoid "data dredging," which leads to overfitted models
covery" of effects that are actually spurious (Anderson 2
has often been a rush to "get to the data analysis" and begin
the data and compute various estimates of interest or con
tests. We realize that these other philosophies may have th
in more exploratory investigations.

The second objective is to explain and illustrate method
at the interface of information theory and mathematical st
of an estimated "best approximating model" from the a pr
models. In particular, we review and explain the use of A
criterion (AIC) in the selection of a model (or small set
statistical inference. AIC provides a simple, effective, a
for the selection of an estimated "best approximating mo
sis and inference. Model selection includes "variable sele

quite different from the usual methods based on null hypot
selection based on information theory is not the only reason
is what we are focusing on here because of its philosophica
advantages.

The practical use of information criteria, such as Aka
lection is relatively recent (the major exception being in
where AIC has been used routinely for the past two decad
information theory and mathematical statistics started wit
book. Akaike considered AIC to be an extension of R. A
theory. These are all complex issues, and the literature is o
and scattered widely throughout books and research journ
to bring this relatively new material into a readable text fo
ily) the biological and statistical sciences. We provide a
many of which are biological, to illustrate various aspec
application.

In contrast, hypothesis testing as a means of selecting
much longer exposure in science. Many seem to feel mo
the hypothesis testing paradigm in model selection, and
the results of a test as *the* standard by which other ap
judged (we believe that they are wrong to do so). Bayesia
selection and inference have been the focus of much recen
the technical level of this material often makes these app
to many in the biological sciences. A variety of cross-valic
based methods have been proposed for model selection,
like very reasonable approaches. The computational dem
Bayesian and cross-validation methods for model selectio
(often 1–3 orders of magnitude higher than information-th
especially if there are more than a dozen or so high-di
models.

The theory presented here allows estimates of "model se
inference problems that arise in using the same data fo
tion and the associated parameter estimation and inferenc
uncertainty is ignored, precision is often overestimated,
interval coverage is below the nominal level, and predic
rate than expected. Another problem is the inclusion of
or factors, with no assessment of the reliability of their s
eral methods for dealing with model- and variable-selec
suggested and examples provided. Incorporating model s
into estimators of precision is an active area of research,
additional approaches developed in the coming years.

The third objective is to present a number of approach
inference from more than one model in the set. That is,
inferences from only the model estimated to be the best, r

research area for Bayesians for the past several years (
Model averaging can be easily done under an information
Model averaging has several practical and theoretical adv
in prediction or in cases where a parameter of interest oc
els. Confidence sets on models is another useful approac
models in the set represent a logical ordering (e.g., a set
ing chronic treatment effects over 1, 2, . . . , t time periods)
importance of explantory variables in a general regression
assessed by summing certain quantities across models. MM
useful in certain conflict resolution issues (Anderson et a

Current practice often would judge a variable as impor
based on whether that variable was in or out of the selected
regression, based on hypothesis testing). Such procedures
dichotomy (see Breiman 2001) and are not in the spirit of a
MMI allows us to discard simplistic dichotomies and foc
ranking models and variables as to their relative value an

Modeling is an art as well as a science and is direct
good approximating model of the information in empiri
for statistical inference from those data. In particular, th
eters estimated from data should be substantially less th
or inference is likely to remain somewhat preliminary (e.
mentions a regression problem with 757 variables and a sa
absurd to think that valid inference is likely to come from
data). In cases where there are relatively few data per est
small-sample version of AIC is available (termed AIC_c)
routinely rather than AIC. There are cases where quasi-lik
appropriate when count data are overdispersed; this theo
criteria such as QAIC and $QAIC_c$, and these extensions
following material.

Simple models with only 1-2 parameters are not the
book; rather, we focus on models of more complex syst
mation has been firmly considered to be an optimizatio
decades, and AIC formulates the problem of model selec
tion problem across a set of candidate models. Minimiz
operation with results that are easy to interpret. Models c
and scaled, allowing full consideration of other good m
the estimated "best approximating model." Evidence ra
strength of evidence for alternative hypotheses. Competin
AIC values close to the minimum, are also useful in the
selection uncertainty. Inference should often be based o
model, unless the data clearly support only a single model
some approaches are provided to allow inference from
models, including model averaging.

McQuarrie and Tsai (1998) present both theoretical and
model selection in regression and time series analysis,
results of large-scale Monte Carlo simulation studies.

1.2 Background Material

Data and stochastic models of data are used in the empiri
inferences concerning both processes and parameters of in
1981, Lunneborg 1994, and Shenk and Franklin 2001 f
ciples). Statistical scientists have worked with researche
sciences for many years to improve methods and underst
processes. This book provides practical, omnibus method
ference from models that are good approximations to biol
data. We focus on statistical evidence and try to avoid an
such as "significant or not significant." A broad definition
here. A single, simple data set might be the subject of anal
data collected from several field sites or laboratories are t
comprehensive analysis. The data might commonly be
tioned by age, sex, species, treatment group, or within sev
geographic areas. In linear and nonlinear regression model
explanatory variables. There are often factors (variables) w
and large effects in these information-rich data sets (the
effect sizes). Parameters in the model represent the effects
focus on modeling philosophy, model selection, estimati
eters, and valid measures of precision under the relative
information-theoretic methods. Valid inference rests upon
addition to the critical considerations relating to problem
design, and protocol for data collection.

1.2.1 Inference from Data, Given a Model

R. A. Fisher (1922) discussed three aspects of the gener
inference: (1) model specification, (2) estimation of mo
(3) estimation of precision. Here, we prefer to partition
into two components: formulation of a set of candidate m
of a model (or small number of models) to be used in ma
much of the twentieth century, methods have been availab
efficiently estimate model parameters and their precision
covariance matrix). Fisher's *likelihood theory* has been t
approach to these issues, but it *assumes* that the model
(and correct, i.e., a true model) and that only the paramet

of a log-linear model for the analysis of count data displa
table. The parameters in these models can be estimated us
hood (ML) methods. That is, if one assumes or somehow
model, methods exist that are objective and asymptotically
ing model parameters and the sampling covariance struc
that model. A more challenging example might be to assu
propriately modeled by a 3-parameter gamma distributio
use the method of maximum likelihood to estimate thes
and the model-based 3×3 sampling covariance matrix. C
model, and if the sample size is "large," then maximum
estimators of parameters that are consistent (i.e., asympto
variance tending to zero), fully efficient (i.e., minimum v
sistent estimators), and normally distributed. With sma
assuming an appropriate model, ML estimators often hav
where bias $\equiv \mathrm{E}(\hat{\theta}) - \theta$. Such bias is usually a trivial c
often substantially less than the $\mathrm{se}(\hat{\theta})$, and bias-adjuste
ten be found if this is deemed necessary. The sampling
estimators are often skewed with small samples, but pro
vals or log-based intervals or bootstrap procedures can
asymmetric confidence intervals with good coverage pro
**the maximum likelihood method provides an objecti
for estimation of model parameters and the sampling**
given an appropriate model.

1.2.2 Likelihood and Least Squares Theory

Biologists have typically been exposed to least squares
classes in applied statistics. LS methods for linear models
to compute, and therefore they enjoyed an early history c
burg 1985). In contrast, Fisher's likelihood methods of
numerical methods and were thus not popular prior to th
ability of personal computers and the development of easy
theory has many similarities with likelihood theory, and it
mators of the structural parameters (but not σ^2) for linear a
when the residuals are assumed to be independent and no
is now easy to allow alternative error structures (i.e., nonn
as Poisson, gamma or log-normal) for regression and othe
either a likelihood or quasi-likelihood framework (e.g., M
1989, Heyde 1997), but more difficult in an LS framewo

The concepts underlying both estimation methods are
understand (Silvey 1975). Consider the simple linear reg
sponse variable (y) is modeled as a linear function of an

variance σ^2. Under LS the estimates of β_0 and β_1 are
$\sum(\epsilon_i)^2$— hence the name *least squares*. The parameter
minimize the average squared error terms (ϵ_i) and define
is the "best fit." Hundreds of statistics books cover the th
for least squares estimation in linear and nonlinear model
the ϵ_i are assumed to be independent, normally distribute

Likelihood methods are much more general, far less tau
tics courses, and slightly more difficult to understand at
much of this book relies on an understanding of likeliho
brief introduction is given here. While likelihood theory i
lying both frequentist and Bayesian statistics, there are no
of applied books solely on this important subject (good ex
Cullagh and Nelder 1989, Edwards 1992, Azzalini 1996
Severini 2000).

The theory underlying likelihood begins with a probabi
parameters (θ). Specifically, model g describes the proba
the data, given the model parameters and a specific mod
$g(x|\theta, \mathit{model})$. A simple example is the binomial probab
θ is the probability of a "success"; let this be the paramet
could be the observation of $y = 15$ successes out of $n = 4$
Then, the discrete probability of getting 15 successes out
parameter ($p \equiv 0.4$) and the binomial model, is

$$g(y, n|p, \mathit{binomial}) = \binom{n}{y} p^y (1-p)^{n-y},$$

$$g(15, 40|p = 0.4, \mathit{binomial}) = \frac{40!}{15!25!}(0.4)^{15}(1-$$

The key point is that for this calculation, the model (her
and its parameters (here $p = 0.4$) are known in advance
In very simple problems such as this, an excellent model
be considered given (such is rarely the case in the real
not sure what model might be used). Then one observes t
$n = 40$) and can compute the probability of the data, giv
parameters.

In much of science, neither the model parameters nor
However, data can be collected in a way that allows th
estimated if a good model can be found or assumed. The
is the basis for such parameter estimation and is a functi
p, given the data and the binomial model:

$$\mathcal{L}(p|y, n, \mathit{binomial}) = \binom{n}{y} p^y (1-p)^n$$

$$\mathcal{L}(p|15, 40, \text{binomial}) = \frac{40!}{15!25!}(p)^{15}(1-p)^{25}$$

Clearly, the likelihood is a function of (only) the unkno
this example); everything else is known or assumed. Th
and the associated likelihood function differ only in terms
given. In the probability model, the parameters, the model
are known, and interest lies in the probability of observi
(the data, y given n in this simple example). In the like
data are given (observed) and the model is assumed (but
lies in estimating the unknown parameters; thus, the like
of only the parameters. The probability model of the dat
function of the parameters are closely related; they mer
of the data and the parameters, given a model. The bino
does not contain the unknown parameter p and is often
contain any information about the unknown parameters
to compute if $n > 50$).

The notation for the likelihood function is very helpful
consider the general expression $\mathcal{L}(\theta|data, model)$. If we f
vention of letting x represent the empirical data and g a
model, then $\mathcal{L}(\theta|x, g)$ is read as "the likelihood of a partic
of the unknown parameter θ (θ is usually a vector), giv
particular model g."

A well-known example will help illustrate the concep
n pennies and observing y "heads." Assuming that the f
and that each penny has an equal probability of a head,
is an obvious model choice in this simple setting. The li
$\mathcal{L}(p|y, n,$ binomial), where p is the (unknown) probabil
given the data (y and n) and the binomial model, one ca
hood that p is 0.15 or 0.73 or any other value between 0 a
(a relative, not absolute, value) is a function of the unk
Given this formalism, one might compute the likelihoo
the unknown parameter p and pick the most likely one as
p, *given* the data and the model. It seems compelling to
that is "most likely." This is Fisher's concept of *maximu
tion*; he published this when he was 22 years old as a third
at Cambridge University! He reasoned that the best esti
parameter (given data and a model) was that which was
the name *maximum likelihood*, ML. The ML estimate
mial model happens to have a closed-form expression
$\hat{p} = y/n = 7/11 = 0.6363$. That is, the numerical v
maximizes the likelihood function. In most real-world ca
form estimator either does not exist or cannot be found
difficulty.

hypothesis tests, the basis for profile likelihood intervals,
quantities (such as model selection criteria). More general
includes the broad concept of *support* (Edwards 1992)
the essential basis for Bayesian approaches to statistica
likelihood is the backbone of statistical theory, whereas
viewed as a limited special case and, while very useful
applications, is not foundational in modern statistics.

For many purposes the natural logarithm of the likeliho
tial; written as $\log(\mathcal{L}(\theta|data, model))$, or $\log(\mathcal{L}(\theta|x, mod$
is clear, just $\log(\mathcal{L}(\theta))$ or even just $\log(\mathcal{L})$. Often, one s
$\log(\mathcal{L}(\theta|x))$, without it being clear that a particular mode
vanced feature of $\log(\mathcal{L})$ is that it, by itself, is a type of *inf*
θ and the model (Edwards 1992:22–23). The log-likeliho
model where 11 pennies are flipped and 7 heads are obse

$$\log(\mathcal{L}(p|y, n, binomial)) = \log\binom{n}{y} + y \cdot \log(p) + (n -$$

$$= \log\binom{11}{7} + 7 \cdot \log(p) + (1$$

$$= 5.79909 + 7 \cdot \log(p) + (4)$$

A property of logarithms for values between 0 and 1 is
negative quadrant; thus, values of discrete log-likelihood fu
(unless some additive constants have been omitted). Figu
of the likelihood (a) and log-likelihood (b) functions wh
flipped, 7 heads were observed, and the binomial mode
value of $p = 0.636$ maximizes both the likelihood an
function; this value is denoted by $\hat{p}$ and is the maxim
mate (MLE). Relatively little information is contained in
size ($n = 11$) and this is reflected in the broad shape o
sample size been 5 times larger, with $n = 55$ and 35
likelihood and log-likelihood functions would be more
and d). In fact, the sampling variance is derived from th
likelihood function around its maximum point. In the usu
vector, a variance–covariance matrix can be estimated bas
tives of the log-likelihood function. These procedures w
here.

The value of the log-likelihood function at its maxim
important quantity, and it is this point that defines the
estimate. In the example with 11 flips and 7 heads, the val
log-likelihood is -1.411 (Figure 1.1b). This result is con

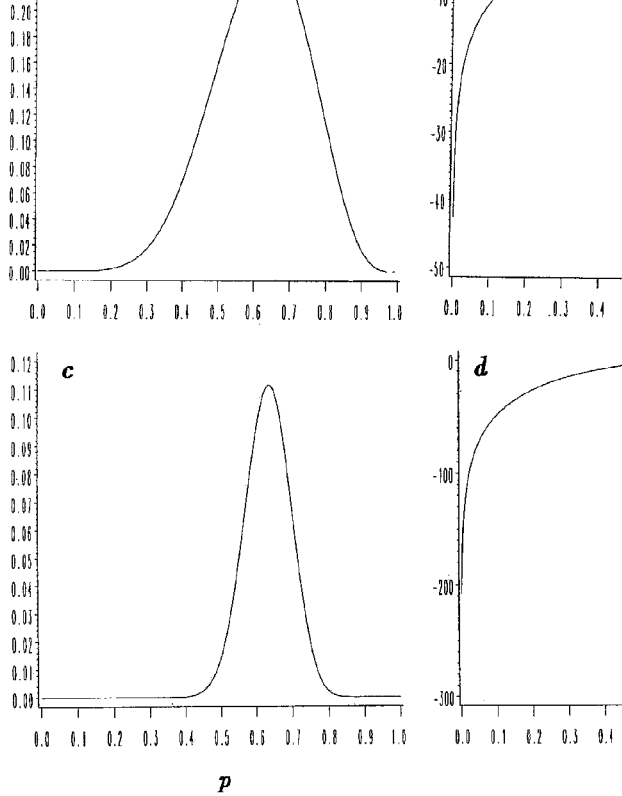


FIGURE 1.1. Plots of the binomial likelihood (a) and log-likelih
$n = 11$ penny flips and the observation that $y = 7$ of these were
plots of the binomial likelihood (c) and log-likelihood (d) functio
times larger; $n = 55$ penny flips and the observation that $y = 35$ of
the differing scales on the Y axis.

log-likelihood function

$$\log(\mathcal{L}(p|y, n, \textit{binomial})) = \log\binom{n}{y} + y \cdot \log(p) + (n$$

and substituting the MLE ($\hat{p} = 0.6363$) and the data (y a

$$-1.411 = 5.79909 + 7 \cdot \log(0.6363) + (4) \cdot \log($$

Thus, when one sees reference to a maximized $\log(\mathcal{L}(\theta))$ th
a numerical value (e.g., -1.411).

Many do not realize that the common procedure for setti
interval (i.e., $\hat{\theta} \pm 1.96 \cdot \widehat{\text{se}}(\hat{\theta})$) is merely an approximatio

achieved confidence interval coverage can be much less th
say, 95%). For example, if the binomial parameter is near 0
of the estimator $\hat{\theta}$ will be nonnormal (asymmetric) unles
very large. In general, rather than use the simple approx
a 95% interval using the log-likelihood function; this pr
is called a *profile likelihood interval*. This is not a simple
approximation has seen heavy use in applied data analysis
the full theory for profile likelihood intervals here, but will
the binomial case where $n = 11$, $y = 7$, $\hat{p} = 0.6363$, and
likelihood value is -1.411. Here, we start with 3.84, whic
the chi-squared distribution with 1 degree of freedom. On
1.92, and this value is subtracted from the maximum point
function: $-1.411 - 1.92 = -3.331$. Now, numerically,
values of p that are associated with the values of the log-li
-3.331. These 2 values are the endpoints of an exact 95% li
interval. In this example, the 95% likelihood interval is (0

Biologists familiar with LS but lacking insight into likeli
benefit from an example. Consider a multiple linear regre
dependent variable y is hypothesized to be a function of r
tor) variables x_j $(j = 1, 2, \ldots, r)$. Here the residuals ϵ_i c
are assumed to be independent, normally distributed with
σ^2, and the model structure is expressed as

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r + \epsilon_i, \qquad i$$

Hence

$$E(y_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r, \qquad i$$

and $E(y_i)$ is a linear function of $r + 1$ parameters. The co

$$\epsilon_i = y_i - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r) = y$$

have the joint probability distribution $g(\underline{\epsilon}|\underline{\theta})$, where $\underline{\theta}$ is a
parameters ($\beta_0, \beta_1, \ldots, \beta_r$, and σ). Here, corresponding
has the model

$$g(\epsilon_i|\underline{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left[\frac{\epsilon_i}{\sigma}\right]^2}.$$

The likelihood is simply the product of these over the n
preted as a function of the unknown parameters, given
model structure, and the normality assumption:

$$\mathcal{L}(\underline{\theta}|\underline{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left[\frac{\epsilon_i}{\sigma}\right]^2} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^-$$

of $\underline{\beta}$ is identical to the usual LS regression estimators (t
tor of σ^2 differs slightly). This formalism shows, *given*
between the data, the model, and the parameters to be ob
using either LS or ML.

In all fitted linear models the residual sum of squares (

$$\text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2,$$

where

$$\begin{aligned}\hat{\epsilon}_i &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_r \\ &= y_i - \hat{\text{E}}(y_i)\end{aligned}$$

The ML estimator is $\hat{\sigma}^2 = \text{RSS}/n$, while the estimator un
LS case is $\hat{\sigma}^2 = \text{RSS}/(n-(r+1))$. This shows that ML
of σ^2 differ by a factor of $n/(n-(r+1))$; often a trivial d
sample size is small. The maximized likelihood is

$$\mathcal{L}(\hat{\underline{\theta}}|x) = \left[\frac{1}{\sqrt{2\pi}\hat{\sigma}}\right]^n e^{-\frac{1}{2}n},$$

or

$$\log(\mathcal{L}(\hat{\underline{\theta}})) = -\frac{1}{2}n\log(\hat{\sigma}^2) - \frac{n}{2}\log(2\pi) -$$

The additive constants can often be discarded from the log
they are constants that do not influence likelihood-based
all standard linear models, we can take

$$\log(\mathcal{L}(\hat{\underline{\theta}})) \approx -\frac{1}{2}n\log(\hat{\sigma})^2.$$

This result is important in model selection theory becaus
mapping from LS analysis results (e.g., the RSS or the M
maximized value of the log-likelihood function for con
linear models with normal residuals. Note that the log-like
to an arbitrary additive constant in this usual case. If the
linear and nonlinear models or if the residual distributions
gamma, and log-normal), then all the terms in the log
retained, without omitting any constants. Most uses of th
relative to its maximum, or to other likelihoods at thei
curvature of the log-likelihood function at the maximum.

The number of parameters $K = r+2$ in these linear mod
intercept (say, β_0), the r regression coefficients ($\beta_1, \ldots,$
variance (σ^2). Often, one (erroneously) considers only th
eters being estimated as the intercept and the slope paran

Sir Ronald Aylmer Fisher was born in 1890 in East Finchley, Londo
Australia, in 1962. This photo was taken when he was approximatel
was one of the foremost scientists of his time, making incredible cor
and applied statistics and genetics. Details of his life and many scie
are found in Box (1978). He published 7 books (one of these ha
printed in 7 languages) and nearly 300 journal papers. Most releva
book is Fisher's likelihood theory and parameter estimation using h
likelihood.

however, in the context of model selection, the number
include σ^2 and thus $K = r + 2$. If the method of LS i
rameter estimators, one must use the regression-based e
$(n - (r + 1))/n = (n - K + 1)/n$ to obtain the ML est
estimation, we minimize $\text{RSS} = n\hat{\sigma}^2$, which for all para
itself is equivalent to maximizing $-\frac{1}{2} \cdot n \log(\hat{\sigma}^2)$.

There is a close relationship between LS and ML me
nonlinear models, where the ϵ_i are assumed to be norm
example, the LS estimates of the structural model parame
equivalent to the MLEs. Likelihood (and related Bayesian)
extensions to the many other classes of models and, with
of computing equipment, likelihood methods are finding in
statisticians and researchers in other scientific disciplines (
1995 for background).

1.2.3 The Critical Issue: "What Is the Best Mod

While hundreds of books and countless journal papers de
model parameters and their associated precision, relativel

In fact, Fisher believed at one time that model specifica
field of mathematical statistics, and this attitude prevailed
community until at least the early 1970s. "*What is the*
is *the* critical question in making valid inference from d
sciences.

The likelihood function $\mathcal{L}(\theta|x, \mathit{model})$ makes it clear tha
θ, data and the model are taken as *given*. Before one can co
that $\theta = 5.3$, one must have data and a particular stati
an investigator will have empirical data for analysis, it
model is known or given. Rather, a number of alternative r
somehow considered as well as the specific explanatory
in modeling a response variable. This issue includes th
problem in multiple regression analysis. If one has data
ML theory can be used to estimate the unknown paran
quantities useful in making statistical inferences. Howe
the best to use for making inferences? What is the basis f
"best"?

Model selection relates to fitted models: given the data
model, then the MLEs of the model parameters have be
Inference relates to theoretical models. It is necessary to

(1) models as structure only (θ value irrelevant),
(2) models as structure, plus specific θ_o (this is the theore
(3) models as structure, plus MLE $\hat{\theta}$, fitted to data,
(4) models as structure by fitting, downplaying θ.

If a poor or inappropriate model (3, above) is used, t
on the data and this model will often be poor. Thus, it is
select (i.e., infer) an appropriate model (1, above) for the
data set; however, this is not the same as trying to find the
selection methods with a deep level of theoretical suppo
particularly, methods that are easy to use and widely applic
of "applicability" means that the methods have good oper
for realistic sample sizes. As Potscher (1991) noted, asym
of little value unless they hold for realized sample sizes.

A simple example will motivate some of the concept
(1992 and 1996) studied patterns of avian species-accum
forested landscapes in the eastern United States using
Breeding Bird Survey (Bystrak 1981). He derived an a pri
models from two sources: (1) the literature on species are
the power or exponential models were suggested) and (2)
the literature for functions that increased monotonically to
1.1). Which model should be used for the analysis of th
Clearly, none of these 9 models are likely to be the "t

response variable (y) is the number of accumulated species, and t
(x) is the accumulated number of samples. Nine models and their nu
shown to motivate the question, "Which fitted model should be use
from these data?"

Model structure	Number o parameters (
$E(y) = ax^b$	3
$E(y) = a + b\log(x)$	3
$E(y) = a(x/(b+x))$	3
$E(y) = a(1 - e^{-bx})$	3
$E(y) = a - bc^x$	4
$E(y) = (a + bx)/(1 + cx)$	4
$E(y) = a(1 - e^{-bx})^c$	4
$E(y) = a\left(1 - [1 + (x/c)^d]^{-b}\right)$	5
$E(y) = a[1 - e^{-(b(x-c))^d}]$	5

[a] There are $K-1$ structural parameters and one residual variance parameter, σ
$E(\epsilon) = 0$, $V(\epsilon) = \sigma^2$.

the index data from the Breeding Bird Survey over the yea
Flather wanted an approximating model that fit the data we
in making inferences about bird communities on the scale
In this first example, the number of parameters in the cand
only from 3 to 5. Which approximating model is "best" fo
from these data is answered philosophically by the prin
(Section 1.4) and operationally by several information-
Chapter 2. Methods for estimating model selection unce
rating this into inferences are given in Chapter 2 and illus
and 5.

Note, in each case, that the response variable y is be
than mixing models of y with $\log(y)$, or other transformat
variable (Table 1.1). These models are in the sense of 1 ab
is given but the parameter values are unspecified. Given a
can be used to obtain $\hat{\theta}$ in the sense of 3 above. In some of t
the model parameters are derived from theory, without th
specific empirical data. Such cases seem to be the except
sciences, where model parameters must usually be estim
using least squares or likelihood theory.

1.2.4 Science Inputs: Formulation of the Set of C

Model specification or formulation, in its widest sense, i
difficult than estimating the model parameters and their p

jective art, that is why scientists must be trained, educa
in their discipline. The published literature and experien
sciences can be used to help formulate a set of a priori ca
most original, innovative part of scientific work is the p
proper question. Good approximating models, each repr
hypothesis, in conjunction with a good set of relevant data
into the underlying biological process and structure.

Lehmann (1990) asks, "where do models come from,"
logical examples (also see Ludwig 1989, Walters 1996, Li
arise from questions about biology and the manner in w
tems function. Relevant theoretical and practical questio
variety of sources (see Box et al. 1978, O'Connor and Spot
ally, these questions come from the scientific literature, re
experiments, personal experience, or contemporary deba
tific community. More practical questions stem from re
controversies, biomonitoring programs, quasi-experimen
hearings.

Chatfield (1995b) suggests that there is a need for m
(than is usually evident) and a *better balance* between the
question), analysis theory, and data. This suggestion ha
literature for decades. One must conclude that it has not be
in applied science or statistics courses. **Our science cultur
do enough to expect and enforce critical thinking.** Too
is focused on the analysis theory and data analysis, wi
about the reason for the study in the first place (see Hayne
examples).

Tukey (1980) argues for the need for deep thinking and e
analysis, and that the results of these activities lead to good
and confirmatory data analysis. In the exploratory phas
importance of a flexible attitude and plotting of the data. H
the computation of test statistics, P-values, and so forth
data analysis. Tukey concludes that to implement the co
properly we need to do a lot of exploratory work.

The philosophy and theory presented here must rest on w
and careful planning and execution of field or laboratory p
books exist giving information on these important issues (
Cook and Campbell 1979, Mead 1988, Hairston 1989, D
1991, Eberhardt and Thomas 1991, Manly 1992, Skalsk
Thompson 1992, Scheiner and Gurevitch 1993, Cox and R
and Zimmermann 2000). Chatfield (1991) reviews statisti
that these might be avoided. Research workers are urged t
to these critical issues. Methods given here should not b
poorly designed work. In the following material we will

Development of the a priori set of candidate models
a global model: a model that has many parameters, inc
relevant effects, and reflects causal mechanisms thought
science of the situation. The global model should also
sign and attributes of the system studied. Specification
should not be based on a probing examination of the dat
some early point, one should investigate the fit of the glob
(e.g., examine residuals and measures of fit such as R^2,
χ^2 goodness-of-fit tests) and proceed with analysis only
the global model provides an acceptable fit to the data.
parameters can then be derived as special cases of the gl
of reduced models represents plausible alternatives based
hypothesized about the process under study. Generally, alt
involve differing numbers of parameters; the number of p
differ by at least an order of magnitude across the set o
Chatfield (1995b) writes concerning the importance of
siderations such as accepted theory, expert background k
information in addition to known constraints on both th
and the variables in the models. All these factors should b
the makeup of the set of candidate models, prior to actua

The more parameters used, the better the fit of the mod
achieved. Large and extensive data sets are likely to suppo
and this should be considered in the development of the set
**If a particular model (parametrization) does not mak
this is reason to exclude it from the set of candidate m
in the case where causation is of interest**. In developing
models, one must recognize a certain balance between k
and focused on plausible hypotheses, while making it b
against omitting a very good a priori model. While this
considered, we advise the inclusion of all models that seem
justification, prior to data analysis. While one must wor
to both underfitting and overfitting, it seems that mode
damaging than underfitting (Shibata 1989). We recomme
considerable amount of careful, a priori thinking in arriving
models (see Peirce 1955, Burnham and Anderson 1992, C

Freedman (1983) noted that when there are many,
variables ($x_1, x_2, \ldots, x_{50}$) used to predict a response va
selection methods will provide regression equations w
"significant" F values, and many "significant" regress
shown by large t values, *even if the explanatory variab
of y*. This undesirable situation occurs most frequently
variables is of the same order as the number of observa
known as Freedman's paradox, was illustrated by Freed

methods. Miller (1990) notes that estimated regression co
away from zero in such cases; this is a type of model selec
resolution of this paradox is in the a priori modeling con
the number of candidate models small, achieving a large
to the number of parameters to be estimated, and basing
than one model.

It is not uncommon to see biologists collect data on 5
variables in the blind hope that some analysis method a
will "find the variables that are significant" and sort out the
(Olden and Jackson 2000). This shotgun strategy will lik
spurious correlations (Anderson et al. 2001b), and it is p
use of many of the traditional multivariate analysis met
components, stepwise discriminant function analysis, c
methods, and factor analysis) found in the biological lit
that mostly spurious results will be found using this unthin
see Flack and Chang 1987 and Miller 1990), and we enc
to give very serious consideration to a well-founded set
and predictor variables (as a reduced set of possible predi
minimizing the inclusion of spurious variables and relat
are not alone in collecting a small amount of data on a v
variables. A. J. Miller (personal communication) indica
data sets in other fields with as many as 1,500 variables v
cases is less than 40 (a purely statistical search for meanin
such data is doomed to failure).

After a carefully defined set of candidate models has be
left with the evidence contained in the data; the task of the
this evidence from analyzing the data. Questions such a
supported by the data?" can be answered objectively. This
allows a clear place for experience (i.e., prior knowled
results of past studies, the biological literature, and curren
the modeling process formally. Then, one turns to the
important" within a sense of parsimony. In some cases, c
of the number and nature of the predictor variables to be
will suffice in defining the candidate models. This proc
initial set of, say, 15–40 predictor variables and a cons
smaller set to use in the set of candidate models. Using A
methods one can only hope to select the best model fr
models are not in the set of candidates, they cannot be d
selection (i.e., data analysis) algorithms.

We lament the practice of generating models (i.e., "mo
in the total absence of real data, and yet "inferences" are m
structure, and functioning of the real world based on stu
We do not object to the often challenging and stimulating

of 28 candidate models for predator–prey relationships t
their "credibility" and "parsimony." However, as is often t
empirical data available on a variety of taxa to pursue the
ous manner (also see Turchin and Batzli (2001), who sug
a system of 2–3 differential equations, for vegetation–h
interactions). Such exercises help us sort out ideas that i
their logical consequences are explored. Modeling exer
our logical and quantitative abilities. Modeling exercises
into how the world *might* function, and hence modeling
alternative hypotheses to be explored with real data. Our
the confusing of presumed insights from such models w
the real world (see Peters 1991, Weiner 1995). An infe
to some aspect of the real world is justified only after
shown to adequately fit relevant empirical data (this will
when the model in its totality has been fit to and tested a
Gause (1934) had similar beliefs when he stated, "Mathem
independent of experiments are of but small importance .

The underlying philosophy of analysis is important h
conservative approach to the overall issue of *strategy* in th
the biological sciences with an emphasis on a priori consid
to be considered. *Careful, a priori consideration of alte
often require a major change in emphasis among many p*
an unfamiliar concept to both biologists and statisticians, w
a tendency to use either a traditional model or a model with
software, making its use easy (Lunneborg 1994). This a
contrast to strategies advocated by others who view model
as a highly iterative and interactive exercise. Such a strate
deliberate data dredging and should be reserved for early
of initial investigation. Such an exploratory avenue is no
book.

Here, we advocate the deliberate exercise of carefully
say, 4–20 alternative models as potential approximation
level information in the data available and the scient
addressed (Lytle 2002 provides an advanced example). S
lems might have as many as 70–100 or more models tha
consider. The number of candidate models is often larger
We find that people tend to include many models that a
than the data could reasonably support (e.g., models wit
parameters). There need to be some well-supported guid
to help analysts better define the models to be considere
els, developed without first deeply examining the data, c
candidate models." The science of the issue enters the ar
priori set of candidate models.

for data analysis are the "true model" that generates the bi
serve (see, for example, Bancroft and Han 1977). We beli
reality) in the biological sciences has essentially infinite d
full reality cannot be revealed with only finite samples of d
those data. It is generally a mistake to believe that there is a
in the biological sciences and that during data analysis th
covered and its parameters estimated. Instead, biological s
with many small effects, interactions, individual heteroge
and environmental covariates (most being unknown to us
to identify a model that provides a good *approximation* t
The words "true model" represent an oxymoron, except i
Carlo studies, whereby a model is used to generate "data" u
numbers (we will use the term "generating model" for s
studies). The concept of a "true model" in biology seem
may even be a source of confusion about the nature of ap
(e.g., see material on BIC and related criteria in Chapter

A model is a simplification or approximation of reality
reflect all of reality. Taub (1993) suggests that unproductiv
true models can be avoided by simply recognizing that a r
definition. Box (1976) noted that "all models are wrong, b
While a model can never be "truth," a model might be rank
to useful, to somewhat useful to, finally, essentially usel
methods try to rank models in the candidate set relative to
any of the models is actually "good" depends primarily
data and the science and a priori thinking that went into the
(reality) is elusive (see deLeeuw 1988). Proper modeling a
what inferences the data support, not what full reality m
1982:14–15, Lindley 1986). Models, used cautiously, tell
supported by the (finite) data available." Increased sampl
allows us to chase full reality, but never quite catch it.

The concept of truth and the false concept of a true
surprisingly important. Often, in the literature, one see
model or simply *the* model as if to be vague as to the exac
Bayesians seem to say little about the subject, even as t
of the prior probabilities on models. Consider the simple
size (n) at time t,

$$n_{t+1} = n_t \cdot s_t,$$

where s is the survival probability during the interval fro
a correct model in the sense that it is algebraically and d
rect; however, it is not an exact representation or model
is not explanatory; it is definitional (it is a tautology, be

be improved if average population survival probability
able from a beta distribution; still, this is far from a mo
truth, even in this very simple setting. Individual variati
be caused by biotic and abiotic variables in the environ
exact model of full reality would have, at the very least,
individual as a nonlinear function of a large number of env
and their interaction terms. Even in this simple case, it is
cannot expect any mathematical model to represent full
true models in the biological sciences. We will take a s
models g_i, without pretending that one represents full rea
"true."

In using some model selection methods it is assumed tha
models contains the "true model" that generated the data
this assumption, unless we use a data set generated by M
as a tutorial example (e.g., Section 3.4), and then we wil
condition clear. In the analysis of real data, it seems unv
that the "true model" is included in the set of candidate mo
true model exists at all. Even if a "true model" did exist an
using some method, it would not be good as a fitted model
(i.e., understanding or prediction) about some biological
numerous parameters would have to be estimated from th
precision of these estimated parameters would be quite lo

Often the investigator wants to simplify some represe
order to achieve an understanding of the dominant aspects
study. If we were given a nonlinear formula with 200 p
could make correct predictions, but it would be difficul
main dynamics of the system without some further simpl
Thus, one should tolerate some inexactness (an inflated er
a simpler and more useful understanding of the phenome

In particular, we believe that there are tapering effect size
systems; that is, there are often several large, important
many smaller effects, and, finally, followed by a myriad o
These effects may be sequentially unveiled as sample size
dominant, effects might be relatively easy to identify and
fairly poor analysis methods, while the second-order eff
treatment effect or an interaction term) might be more di
still smaller effects can be detected only with very larg
Kareiva 1994 and related papers), while the smallest effects
being detected, even with very large samples. Rare events t
may be very important but quite difficult to study. Approxi
be related to the amount of data and information available
appropriately support only simple models with few para
comprehensive data sets will support, if necessary, more

dimensioned model with relatively few parameters might
full truth or reality. Biologists should not believe that a
exists that generates the data observed, although some
might be of relatively low dimension and could be well a
fairly simple model. The issue of a range of tapering effe
in epidemiology, where Michael Thun notes, "... you
from a big thing. What's very hard to do is to tell a littl
at all" (Taubes 1995). *Full reality will always remain elu*
sciences.

At a more advanced conceptual level, these is a concep
about the population (or process or system) under stud
and the goal is to express this information in a more com
form using a "model." Conceptually, this is a change in co
to using a different "alphabet." The data have only a fin
information. The *goal* of model selection is to achieve
translation so that no information is lost; in fact, we cann
The data can be ideally partitioned into *information* and
of the data is not information. However, noise could cont
we cannot decode. Conceptually, the role of a good mod
so as to separate information from noise.

Our main emphasis in modeling empirical data is to un
ical structure, process, or system. Sometimes prediction
here, however, one would hopefully have an understand
of the system as a basis for making trustworthy predicti
developing a set of candidate models prior to intensive
ing one that is "best," and estimating the parameters of
precision (using maximum likelihood or least squares me
strategy is a basis for valid inferences, and there are sev
methods to allow additional inferences and insights. In pa
to allow formal inference from more than one model, and
advantages (Hoeting et al. 1999). Statistical science is n
of mathematics, but rather it is concerned with the develo
theory of information using what is known or postulate
of the matter. In our investigations into these issues we
by how much uncertainty there is in selecting a good ap
the variability in terms of what model is selected or c
independent data sets, for example, is often large.

1.2.6 An Ideal Approximating Model

We consider some properties of an ideal model for valid i
ysis of data. It is important that the best model is selected
that were defined prior to data analysis and based on the

model would be appropriately simple, based on concepts
thermore, precise, unbiased estimators of parameters wou
accurate estimators of precision. The best model would id
confidence interval coverage close to the nominal level (
confidence intervals of minimum width. Achieved confid
age is a convenient index to whether parameter estimat
precision are adequate. Finally, one would like as good a
the structure of the system as the information permits. T
adjusted R^2 can be computed and σ^2 estimated as a mea
plained or residual variation, respectively. Ideally, the pa
model would have biological interpretations. If predictio
having the above issues in place might warrant some ten
predictions. There are many cases where two or more m
tied for "best," and this should be fully recognized in furth
ence, especially when they produce different predictions.
might be 4–10 models that have at least some support, an
scrutiny in reaching conclusions from the data, based on i
than a single model.

1.3 Model Fundamentals and Notation

This section provides a conceptualization of some importa
as they are used in this book. Some of these classes are part
model selection. A general notation is introduced that is i
to readers.

1.3.1 Truth or Full Reality f

While there are no models that exactly represent full realit
full truth can be denoted as f. The concept of f is abstra
which we want to make inferences, based on data and ap
We use the notation $f(x)$ to denote that integration is ov
we do not want to convey the notion that f is a function
arise from full reality and can be used to make formal in
truth, if data collection has been carefully planned and
experimental design has been achieved.

1.3.2 Approximating Models $g_i(x|\theta)$

We use the notation $g_i(x|\theta)$ or often, if the context is cl
ith approximating model. We use θ to represent gener

trials (p). The models g_i are discrete or continuous prob
and our focus will be on their associated likelihoods, $\mathcal{L}$
log-likelihoods $\log(\mathcal{L}(\theta|data,\ model))$. Notation for the
sometimes be shortened to $\log(\mathcal{L}(\theta|x,\ g))$ or even $\log(\mathcal{L}$
R models will have been defined prior to data analysis. T
only the form of the model, leaving the unknown parame

A simple example will aid in the understanding of th
a study of mortality (μ_c) as a function of concentration
ical compound. The size (s) of the animal (binary as sm
group covariate (z, such as gender) are also recorded, beca
esized to be important in better understanding the con
function. Investigators might consider mortality probabili
time interval to be a logistic function of concentration,
$c = 0, 1, 2, 4, 8,$ and 16. The full structure of the logist
variables are included in the model can be written as,

$$\mu_c = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 c + \beta_2 s + \beta_3 z}$$

Use of the logistic link function allows the expression to l
model structure,

$$\text{logit}(\mu_c) = \log_e\left(\frac{\mu_c}{1-\mu_c}\right) = \beta_0 + \beta_1 c + \beta_2$$

Here the data (y) are binary for mortality (dead or alive),
and gender (male and female), while concentration is recor
The response variable $y = 1$ if the animal died and 0 if it liv
concentration. Then,

$$\text{Prob}\{y = 1|c, s, z\} = \mu_c$$

for n individuals at concentration c, size s, and gender z.
is proportional to

$$\mathcal{L}(\mu_c|data,\ model) = \prod_{i=1}^{n} (\mu_c(i))^{y_i}\,(1 - \mu_c($$

Thus, a set of approximating structural models might be c
science of the issue. The stochastic part of the model is assu
The models are alternatives, defined prior to data analysi
in the strength of evidence for each of the alternative hyp
by models. Five ($R = 5$) structural models will be used f

$$g_1(x): \quad \text{logit}(\mu_c) = \beta_0 + \beta_1 c + \beta_2 s + \beta$$
$$g_2(x): \quad \text{logit}(\mu_c) = \beta_0 + \beta_1 c + \beta_2 s,$$
$$g_3(x): \quad \text{logit}(\mu_c) = \beta_0 + \beta_1 c \qquad\qquad + \beta$$

These models specify the structural form (including how covariates enter), but not the parameter values (the β_i); eac are independent Bernoulli random variables. The first mod model. The second model represents the hypothesis that (z) is unimportant, while the third model is like the first, is hypothesized to be unimportant. The fifth model impl constant and not a function of concentration. Often, eno the compound that model g_5 is not worth exploration. O or complementary log-log, or probit function could have the hypothesized relationships in this example, rather tha

1.3.3 The Kullback–Leibler Best Model $g_i(x|\theta_0)$

For given full reality (f), data (x), sample size (n), and is a best model in the sense of Kullback-Leibler inform Chapter 2). That is, given the possible data, the form of possible parameter values, K-L information can be comp in the set and the model best approximating full reality d

The parameters that produce this conceptually best sing $g(x|\theta)$, are denoted by θ_0, Of course, this model is gene but can be estimated; such estimation involves computi parameters in each model ($\hat{\theta}$) and then *estimating* K-L in for model selection and inference. The MLEs converge and the concept of bias is with respect to θ_0, rather than parameters" associated with full reality f.

1.3.4 Estimated Models $g_i(x|\hat{\theta})$

Estimated models have specific parameter values from M based on the given data and model. If another, replicate da and based on the same sample size, the parameter estimate what; the amount of difference expected is related to m (e.g., standard errors and confidence intervals). It is impor the model form $g_i(x|\theta)$ from specific estimates of this m and the process of parameter estimation, $g_i(x|\hat{\theta})$.

In the models of mortality as a function of concentratio (above), there are associated likelihoods and log-likelihood can be used to obtain the MLEs $\hat{\beta}_0$ and $\hat{\beta}_1$ for model g likelihood function is

$$\mathcal{L}(\beta_0, \beta_1|data,\ model) = \prod_{i=1}^{n} (\mu_c(i))^{y_i} (1 - \mu_c$$

$$1 + \exp\{-(\beta_0 + \beta_1 c)\}$$

Thus, the only parameters in the likelihood are β_0 and β_1 one can obtain the MLEs. The value of the maximized lo estimated variance–covariance matrix can also be comput we have only the model form $g(x|\theta)$ we have an infinite where all such models have the same form but different all of these models there is a unique K-L best model. Co how to find this model, given f.

1.3.5 Generating Models

Monte Carlo simulation is a very useful and general app and applied statistics (Manly 1991). These procedures rec specified as the basis for generating Monte Carlo data. Suc reality, and thus we call it a *generating model*. It is "trut of computerized truth. One should not confuse a generat based on Monte Carlo data with full reality f.

1.3.6 Global Model

Ideally, the global model has in it all the factors or var important. Other models are often special cases of this g is not always a global model. If sample size is small, it m fit the global model. Goodness-of-fit tests and estimates parameter for count data should be based (only) on the glo cept of overdispersion is relatively model-independent; h must be used to compute or model any overdispersion thou data. Thus, the most highly parametrized model will ser for assessing overall fit and estimating a parameter associa sion. In the models of mortality (above), model g_1 woul model.

The advantage of this approach is that if the global adequately, then a selected model that is more parsimon data (this is an empirical result, not a theorem). Parsimon should not lead to a model that does not fit the data (this pr for the selection methods we advocate here). Thus, goodn and the estimation of overdispersion parameters should be global model (this could also be computed for the selecte

In summary, we will use the word "model" to mean di fully, the context will be clear. Certainly it is important t between f and g. The general structural form is denote specifying the numerical value of the parameter θ (e.g., m

by the space over which θ varies. Frequently, we will refe
MLEs (the most likely, given the data and the model) have
cases we will mean the best model, $g(x|\theta_0)$, which is one
K-L best relative to f).

1.3.7 Overview of Stochastic Models in the Biol

Models are useful in the biological sciences for underst
of systems, estimating parameters of interest and their a
covariance matrix, predicting outcomes and responses, a
hypotheses. Such models might be used for "relational" o
poses or might be used for prediction. In the following ma
the main types of models used in the biological sciences. A
meant to be exhaustive, it will allow the reader an impress
of models of empirical data that we will treat under an in
framework.

Simple linear and multiple linear regression models (Sel
Smith 1981, Brown 1993) have seen heavy use in the biol
the past four decades. These models commonly employ
parameters, and the statistical theory is fully developed (
squares or likelihood theory). Similarly, analysis of varia
models have been widely used, and the theory underlyi
closely related to regression models and is fully developed
of general linear models). Theory and software for this w
are readily available.

Nonlinear regression models (Gallant 1987, Seber and
et al. 1995) have also seen abundant use in the biologic
regression is a common example). Here, the underlying
lihood based, and some classes of nonlinear models requ
software. In general, nonlinear estimation is a more adva
somewhat less well understood by many practicing resea

Other types of models used in the biological sciences
linear (McCullagh and Nelder 1989, Morgan 1992, 20
additive (Hastie and Tibshirani 1990) models (these can b
regression models). These modeling techniques have seen
past decade. Multivariate modeling approaches such as m
and regression, canonical correlation, factor analysis, pr
analysis, and discriminate function analysis have had a che
biological and social sciences, but still see substantial use
and McCulloch 1990). Log-linear and logistic models
become widely used for count data. Time series models (B
1987, 1991) are used in many biological disciplines. Va
organism's growth (Brisbin et al. 1987, Gochfeld 1987) hav

Compartmental models are a type of state transition m
continuous response and are usually based on systems of
differential equations (Brown and Rothery 1993, Matis an
are discrete state transition models using the theory of Ma
1971); these have found use in a wide variety of fields inc
ical models of disease transmission. More advanced meth
wide application include the class of models called "ran
and deLeeuw 1998).

Models to predict population viability (Boyce 1992),
type of Leslie matrix, are much used in conservation bi
alternative model forms given serious evaluation. A com
that these models are rarely based on empirical data; the fo
its parameter values are often merely only "very rough g
by the lack of empirical data (White 2000).

Biologists in several disciplines employ differential equ
research (see Pascual and Kareiva 1996 for a reanalysis of
data and Roughgarden 1979 for examples in population ge
ary ecology). Many important applications involve explo
(Myers et al. 1995). Computer software exists to allow
be estimated using least squares or maximum likelihood
and Splus). These are powerful tools in the analysis of em
beg the issue of "what model to use."

Open and closed capture–recapture (Lebreton et al. 19
ery (Brownie et al. 1985) models represent a class of mod
multinomial distributions (see issues 5 and 6 of volum
of Applied Statistics, 1995). Distance sampling theory (B
2001) relies on models of the detection function and ofte
metric models. Parameters in these models are nearly alw
maximum likelihood.

Spatial models (Cressie 1991 and Renshaw 1991) are
the biological sciences, allowing the biologist to take adva
sets (e.g., geographic information systems). Stein and
shown how Kriging (perhaps the most widely used spati
expressed as a least squares problem, and the developme
Monte Carlo methods such as the Gibbs sampler (Robe
Chen et al. 2000) allow other forms of spatial models
squares or maximum likelihood (Augustin et al. 1996). Fu
for methods widely used on biological data has been c
and Brooks (1990). Geographic information systems pote
numbers of covariates for biological models, so that mode
particularly important.

Spatiotemporal models are potentially invaluable to t
most researchers model changes over space or time, and

spatiotemporal modeling, model selection will be an imp
such a framework. A step towards this general framework
land and Elston (1993), who modeled changes in the sp
wildlife.

There are many other examples where modeling of dat
tal role in the biological sciences. Henceforth, we will ex
that cannot be put into a likelihood or quasi-likelihood
framework and models that do not explicitly relate to emp
squares formulations are merely special cases that have
hood formulation in usual practice. There are general in
approaches for models well outside the likelihood framew
less 1994, Ishiguo et al. 1997, Hurvich and Simonoff 1
and b). There are now model selection methods for nonpa
splines, kernel methods, martingales, and generalized e
Thus, methods exist for nearly all classes of models we m
the theoretical or applied biological sciences.

1.4 Inference and the Principle of Parsim

1.4.1 Avoid Overfitting to Achieve a Good Mode

Consider two analysts studying a small set of biological d
linear regression model. The first exclaims that a particula
excellent fit to the data. The second notices that 22 para
the regression and states, "Yes, but you have used enough
elephant!" This seeming conflict between increasing mod
numbers of parameters to be estimated from the data led W
the question, "How many parameters *does* it take to fit an e
that about 30 parameters would do reasonably well (Fig
had he fit 36 parameters to his data, he could have
fit.

Wel's finding is both insightful and humorous, but it de
pretation for our purposes here. His "standard" is itself only
even lacks ears, a prominent elephantine feature; hardly t
would have been a large, digitized, high-resolution photog
too, would have been only a model (and not truth). Perh
should have been used as truth, but this begs the questio
should we use?" This simple example will encourage thi
ality, "true models," and approximating models and moti
parsimony in the following section. **William of Occam su
teenth century that one "shave away all that is unnec
often referred to as *Occam's razor*. Occam's razor has

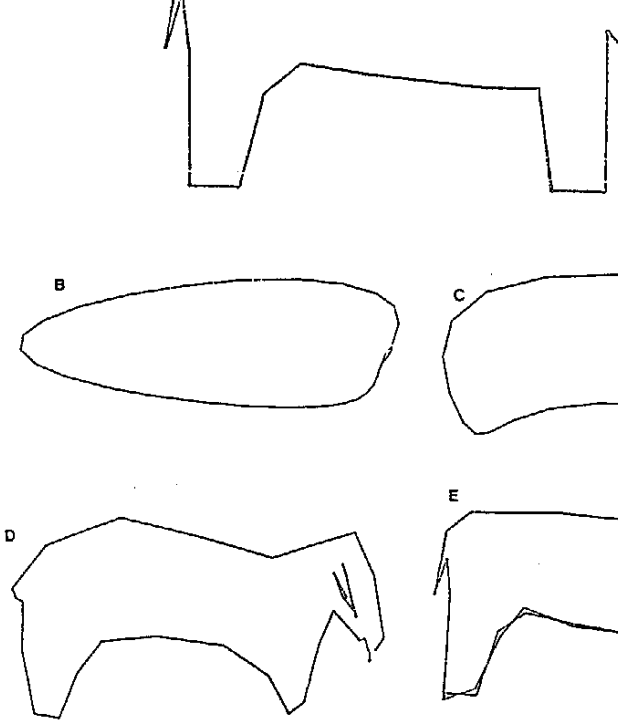


FIGURE 1.2. "How many parameters does does it take to fit an el
by Wel (1975). He started with an idealized drawing (A) defined
least squares Fourier sine series fits of the form $x(t) = \alpha_0 + \sum \alpha_i$
$\beta_0 + \sum \beta_i \sin(it\pi/36)$ for $i = 1, \ldots, N$. He examined fits for $K =$
in B–E) and stopped with the fit of a 30 term model. He concluded
"may not satisfy the third-grade art teacher, but would carry most
preliminary design."

in both science and technology, and it is embodied in t
simony. Albert Einstein is supposed to have said, "Everyt
as simple as possible, but no simpler."

Success in the analysis of real data and the resulting infe
importantly on the choice of a best approximating model.
biological sciences should be based on a parsimonious m
accurate approximation to the structural information in t
should not be viewed as searching for the "true model." M
selection are essentially concerned with the "art of appr
1974).

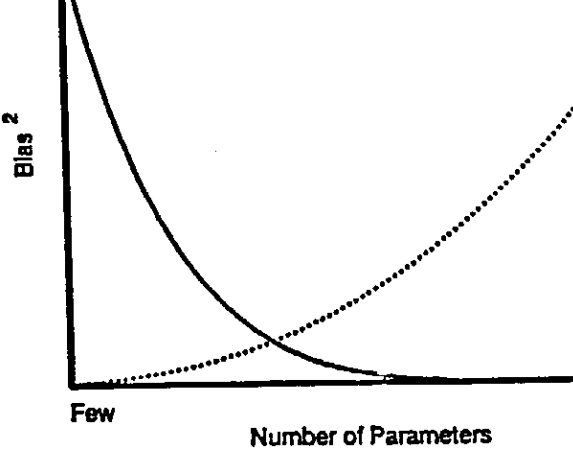


FIGURE 1.3. The *principle of parsimony*: the conceptual tradeof
(solid line) and variance vs. the number of estimable parameters in th
selection methods implicitly employ some notion of this tradeoff.
model need not occur exactly where the two curves intersect. Fu
attainable with finite samples and usually lies well to the right of the
approximating model lies (the tradeoff region). Bias decreases an
increases as the number of parameters in a model increases.

1.4.2 The Principle of Parsimony

If the fit is improved by a model with more parameters, th
stop? Box and Jenkins (1970:17) suggested that the *pri*
should lead to a model with "... the smallest possible nu
for adequate representation of the data." Statisticians v
parsimony as a bias versus variance tradeoff. In general
variance increases as the dimension of the model (K) increa
ten, we may use the number of parameters in a model as a n
of structure inferred from the data. The fit of any model
increasing the number of parameters (e.g., the elephant-fi
ever, a tradeoff with the increasing variance must be cons
model for inference. Parsimonious models achieve a pro
bias and variance. All model selection methods are base
the principle of parsimony (Breiman 1992, Zhang 1994).

In understanding the utility of an approximate model f
is convenient to consider two undesirable possibilities: u
fitted models. Here, we must avoid judging a selected mo
supposed "true model," as occurs when data are simulated
very simple, model using Monte Carlo methods. In this ca

10 parameters). This interpretation is often of little valu
ignores the principle of parsimony and its implications an
conception that such a simple true model exists in biolog
believe that truth is essentially infinite-dimensional, then o
defined in terms of the number of parameters in the fitted r
this use of the terms "underfitted" and "overfitted" that su
of a low-dimensional "true model" as a "standard."

Instead, we reserve the terms underfitted and overfitte
to a "best approximating model" (Section 1.2.6). Here, a
would ignore some important replicable (i.e., conceptuall
other samples) structure in the data and thus fail to ident
actually supported by the data. In this case, bias in the p
is often substantial, and the sampling variance is underest
resulting in poor confidence interval coverage. Underfi
miss important treatment effects in experimental settings
as judged against a best approximating model, are ofte
parameter estimators, but have estimated (and actual) san
are needlessly large (the precision of the estimators is p
could have been accomplished with a more parsimoniou
treatment effects tend to be identified, and spurious va
with overfitted models. Shibata (1989) argues that unde
more serious issue in data analysis and inference than ov
assessment breaks down in many exploratory studies whe
be only 35–80 and there are 20–80 explanatory variables
may expect substantial overfitting and many effects that a
(Freedman 1983, Anderson et al. 2001b).

The concept of parsimony and a bias versus variance
portant. Thus we will provide some additional insights (al
Forster and Sober 1994, and Jaffe and Spirer 1987). The
tion and analysis is to make inferences from the sample th
the population. The inferences relate to the *information* a
system under study as inferred from the models considere
estimated in each model. A paramount consideration is th
good precision, of any inference reached. When we ima
samples, there will be some recognizable features commo
samples. Such features are the sort of inference about wh
strong inferences (from our single sample). Other featu
say, 60% of the samples yet still reflect something real ab
process under study, and we would hope to make weaker in
these. Yet additional features appear in only a few samp
be best included in the error term (σ^2) in modeling. If o
inference about these features quite unique to just the sin
as if they applied to all (or most all) samples (hence to t

replicable over samples is underfitting. The data are not
rather we approximate the structural information in the
ble over such samples (see Chatfield 1996, Collopy et al.
that structure with a model form and parameter estimate
"sampling variation" that must also be estimated (inferre

True replication is very advantageous, but this tends
in the case of strict experiments where replication and
foundation. Such experimental replication allows a valid
variation (σ^2). An understanding of these issues makes on
when observational studies seem possible and practical, a
seem less feasible.

A best approximating model is achieved by properly
of underfitting and overfitting. Stone and Brooks (199
"... straddling pitfalls of underfitting and overfitting."
is achieved when bias and variance are controlled to achie
val coverage at approximately the nominal level and where
minimum. Proper model selection rejects a model that is fa
tempts to identify a model in which the error of approximat
to random fluctuations are well balanced (Shibata 1983,
selection methods are "parsimonious" (e.g., BIC, Schwa
realistic situations, to select models that are too simple (i.
bias is large, precision is overestimated, and achieved con
erage is well below the nominal level. Such instances are
inference. One has only a highly precise, quite biased res

Sakamoto et al. (1986) simulated data to illustrate the c
and the errors of underfitting and overfitting models (Figu
(each with $n = 21$) were generated from the simple mode

$$y = e^{(x-0.3)^2} - 1 + \epsilon,$$

where x varied from 0 to 1 in equally spaced steps of 0.05,
Thus, in this case, they considered the generating model
rameters: 0.3, -1, and 0.01. They considered the set of ca
the approximating models) to be simple polynomials of c
table below.

Order	K	Approximating Model
0	2	$\mathrm{E}(y) = \beta_0$
1	3	$\mathrm{E}(y) = \beta_0 + \beta_1(x)$
2	4	$\mathrm{E}(y) = \beta_0 + \beta_1(x) + \beta_2(x^2)$
3	5	$\mathrm{E}(y) = \beta_0 + \beta_1(x) + \beta_2(x^2) + \beta_3(x^3)$
4	6	$\mathrm{E}(y) = \beta_0 + \beta_1(x) + \beta_2(x^2) + \beta_3(x^3) + \beta_4$
5	7	$\mathrm{E}(y) = \beta_0 + \beta_1(x) + \beta_2(x^2) + \beta_3(x^3) + \beta_4$

Thus, each of these 6 models was fit to each of the 10 sin

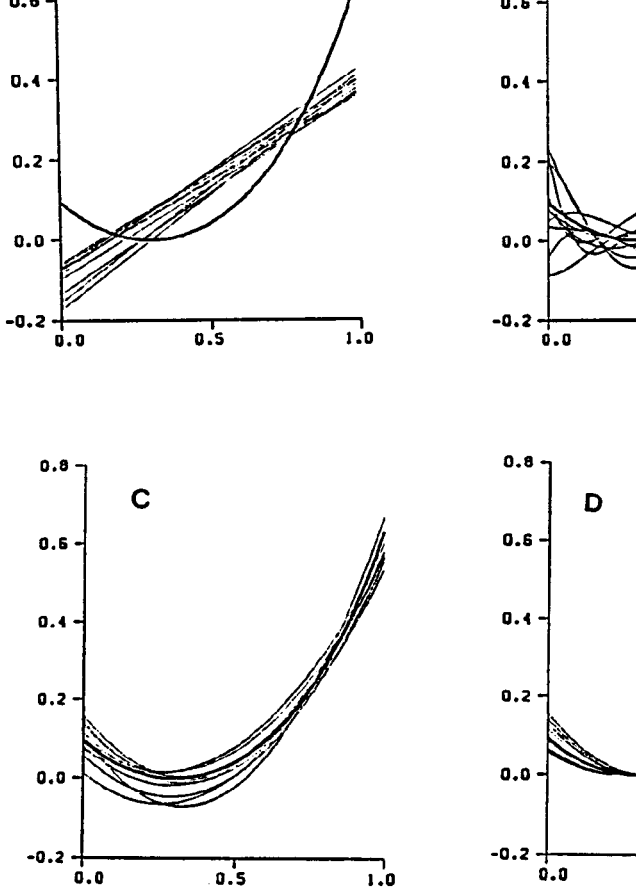


FIGURE 1.4. Ten Monte Carlo repetitions of data sets ($n = 21$) ge
$y = e^{(x-0.3)^2} - 1 + \epsilon; 0 \leq x \leq 1, \epsilon \sim N(0, .01)$ (from Sakamoto et al
order polynomial (A) clearly misidentifies the basic nonlinear stru
and unsatisfactory. A 5th-order polynomial (B) has too many paran
large variance, and will have poor predictive qualities because it
Neither A nor B is properly parsimonious, nor do they represent a bes
A 2nd-order polynomial seems quite good as an approximating m
that the function is nonnegative and has its minimum at $x = 0.3$,
model that enforces these conditions is improved further (D). In n
one lacks the benefit of simple plots and 10 independent data sets,
A–D. See Section 3.7 for a full analysis of these data.

1.4A). Here bias is obvious, the nonlinear structure of the
poorly approximated, and confidence interval coverage a
the model will be quite poor. Of course, there is *some* mode
5 models because they are only simple polynomial approxi
is illustrated in Figure 1.4B, where a 5th-order polynom
as an approximating model. Here, there is little evidence
quantity), precision is obviously poor, and it is difficult to
structure of the model. Prediction will be quite impreci
and it has features that do not occur in the generating m
one extrapolates beyond the range of the data (always a r
underfitting and overfitting are undesirable in judging ap
for data analysis.

If a second-order polynomial ($K = 4$) is used as the ap
the fits seem quite reasonable (Figure 1.4C), and one mi
ference from this model. Finally, *if* it were known a pri
of the situation that the function was nonnegative and had
at $x = 0.3$, then an improved quadratic approximating m
information very effectively (Figure 1.4D). The form of t

$$\mathrm{E}(y) = \beta_0(x + \beta_1)^2$$

with $K = 3$ (i.e., β_0, β_1, and σ^2), whereas the second-o
4 parameters. This example illustrates that valid statistic
partially dependent on the analysis process; the science c
play an important role through modeling. This particular
visual image of underfitting and overfitting in a simple ca
ating model and various approximating models can be ea
dimensions. Parsimony issues with real data in the biolog
always defy such a simple graphical approach because tru
rarely has 10 independent data sets on exactly the same p
high dimensions are problematic to produce and interpret
generating model contained no tapering effects. However
models do have tapering effects. Therefore, objective an
are needed that do not rely on simple graphics and can cop
complexities and high dimensionality.

1.4.3 Model Selection Methods

Model selection has most often been viewed, and hence ta
null hypothesis testing. Sequential testing has most often b
stepup (forward) or stepdown (backward) methods. Stepw
for variables to be added or deleted at each step. These te
remain popular in many computer software packages in
operating characteristics. Testing schemes are based on

tests are to be made (see Westfall and Young 1993), a
independent. Tests between models that are not nested
model is nested if it is a special case of another model; f
degree polynomial is nested within a fourth-degree pol
hypothesis testing is a very poor basis for model selectio
Sclove 1994b). McQuarrie and Tsai (1998) do not even tre
for a short appendix on stepwise regression—the final thre

Cross-validation has been suggested and well studied
selection (Mosteller and Tukey 1968, Stone 1974, 1977;
the data are divided into two partitions. The first partitio
fitting; and the second is used for model validation (so
partition has only one observation). Then a new partition
whole process is repeated hundreds or thousands of time
then chosen, such as minimum squared prediction error,
selection. There are several variations on this theme, and i
ology (Craven and Wahba 1979, Burman 1989, Shao 199
Hjorth 1994). These methods are quite computer intensiv
practical if more than about 15–20 models must be eva
size is large. Still, cross-validation offers an interesting a
selection.

Some analysts favor using a very general model in all
fitted model). We believe that this is generally poor pra
Others have a "favorite" model that they believe is gooc
nearly all situations. For example, some researchers alway
model (Buckland et al. 1993) with 2 parameters ($K = 2$)
model to the detection function in line transect sampling.
what reasonable for situations where a simple model su
to 3), but will be poor practice in more challenging mode
$10 \leq K \leq 30$ or more is required. These *ad hoc* rules ig
parsimony and data-based model selection, in which the
model to be used for inference.

If goodness-of-fit tests can be computed for all alterna
some are not nested within others, then one could use the m
parameters that "fits" (i.e., $P > 0.05$ or 0.10). However
fits can often be achieved by using models with more a
(e.g., the elephant-fitting problem), and this can make th
α very critical. A large α-level leads to overfitted model
problems. In addition, other problems may be encounte
underdispersion and low power if one must pool small ex
that the test statistic is chi-square distributed. Perhaps, mo
is no theory to suggest that this approach will lead to s
good inferential properties (i.e., an adequate bias vs. varia
achieved confidence interval coverage and width).

where R^2 is the usual coefficient of multiple determinatio
1981:91–92). Under this method, one selects the model ir
statistic is largest. McQuarrie and Tsai (1998) found this
poor (also see Rencher and Pun (1980). While adjusted
scriptive statistic, it is not useful in model selection. M
(Mallows 1973, 1995) is also used in LS regression with
a constant variance and in this special case provides a r
date models that is the same as the rankings under AIC (t
C_p vs. AIC, will differ, see Atilgan 1996). The selection
adjusted R^2 statistic and Mallows's C_p are related for
(see Seber 1977:362–369). Hurvich and Tsai (1989) and
(1998) provide some comparisons of AIC_c vs. several c
regression problems.

Bayesian researchers have taken somewhat different
sumptions, and have proposed several alternative methods
Methods such as CAIC, BIC (SIC), WIC, and HQ are ment
as well as full Bayesian model selection (see especially
These other Bayesian approaches to model selection and
current state of the art in statistics but may seem very di
and implement and are very computer intensive (e.g., Lau
and Carlin and Chib 1995). Draper (1995) provides a re
advanced methods (also see Potscher 1991). Spiegelhalt
developed a deviance information criterion (DIC) from a
that is analogous to AIC. This seems to represent a blendi
Bayesian thinking, resulting in an AIC-like criterion.

The general approach that we advocate here is one deriv
1974, 1977, 1978a and b, and 1981a and b), based on inf
it is discussed at length in this book. Akaike's informatio
has led to a number of alternative methods having desirab
selection of best approximating models in practice (e.g.,
and TIC—Chapters 2 and 7). Our general advocacy con
associated criteria is somewhat stronger than that of L
(1986) but similar in that they also recommend objective
some well-defined criterion with a strong, fundamental b

1.5 Data Dredging, Overanalysis of Data and Spurious Effects

The process of analyzing data with few or no a priori q
tively and iteratively searching the data for patterns and "s
called by the derogatory term "data dredging." Other ter

large data sets. Often the problem arises when data on l
been taken with little or no a priori motive or without b
science. No specific objectives or alternatives were in pla
sis; thus the data are submitted for analysis in the hope th
a plethora of null hypothesis test results will provide in
is significant." A model is fit, and variables not in that
create a new model, letting the data and intermediate resu
ther models and variables to be investigated. Patterns se
of the analysis are "chased" as new variables, cross pro
variables are added to the model and alternative transfor
new models are clearly based on the intermediate results f
analyses. The final model is the result of effective dredgi
everything remaining is "significant." Under this view, Ho
(1989:169) comment that "Model fitting is an iterative p
obtain the final model on the first pass through the data." I
that such a final model is probably overfitted and unstabl
considerably if other sample data were available on the
actual predictive performance (i.e., on new data) often we
be expected from the statistics provided by the terminal
field 1996, Wang 1993). The inferential properties of a pr
data analysis are very different. For example, (traditionall
of precision can be made from the model following data
1998).

1.5.1 Overanalysis of Data

If data dredging is done, the resulting model is very muc
fitted) to the data in a post hoc fashion, and the estimates o
to be overestimated. Such tailoring overdescribes the data
validity of inferences made about the information in the
tion of interest. Many naive applications of classical mult
merely "fishing trips" hoping to find "significant" linear
the many variables subjected to analysis (Rexstad et al. 1
Reid 2000).

Computer routines (e.g., SAS INSIGHT) and associated
dredging both easy and "effective." Some statistical liter
so-called *iterative process of model building* (e.g., Hend
1981). One looks for patterns in the residuals, employs va
ing variables in their decreasing order of "importance," a
models. Stepwise regression and discriminant functions, f
to search for "significant" variables; such methods are es
if many variables (Freedman's paradox) are available for
data are available on over 100 variables, and the sample si

White (2000:1097) notes, "It is widely acknowledg
searchers that data snooping [dredging] is a dangerous pr
but in fact it is endemic." Examples of data dredging in
tion of crossplots or a correlation matrix of the explanat
the response variable. These data-dependent activities c
linear or nonlinear relationships and interactions *in the s*
lead the investigator to consider additional models. These
avoided, because they probably lead to overfitted models v
eter estimates and inclusion of unimportant variables as re
(Anderson et al. 2001b). The sample may be well fit, but
valid inference from the sample to the population. This typ
exploratory data analysis has a place in the earliest stages
ological relationship but should probably remain unpublis
cases are not the subject of this book, and we can only
results of such procedures be treated as possible hypothe
Longford and Nelder 1999). New data should be collec
hypotheses effectively and then submitted for a compreh
priori strategy of analysis such as we advocate here.

Two types of data dredging might be distinguished. The
above; a highly interactive, data dependent, iterative pos
second is also common and also leads to likely overfittin
effects that are actually spurious. In this type, the invest
a priori information; thus "all possible models" are cons
(e.g., SAS PROC REG allows this as an option). Note t
models" approach usually does not include interaction te
various transformations such as $(x_1)^2$ or $1/x_3$ or $\log(x_2)$. In
problems, the number of candidate models in this approa
(e.g., 20 variables $>$ a million models, 30 variables $>$ a
least this second type is not explicitly data dependent, bu
dependent and leads to the same "sins." Also, it is usually
rather than taking the results of one set of analyses and inp
into the consideration of new models. Still, in some ap
software often can systematically search all such models r
and thus the strategy of trying all possible models (or at le
ber of models) continues, unfortunately, to be popular. W
situations could be substantially improved if the research
cus on the science of the situation before proceeding with s
approach.

Standard inferential tests and estimates of precision (e.g
tors of the sampling covariance matrix, given a model) are
model results from the first type of data dredging. Resul
misleading, and there is no valid basis to claim "significan
ally there is no way to estimate precision because of the s

a-levels or P-values. However, if there were 1,000 mod
would be 0.00005, instead of the usual 0.05! Problems
are often linked with the problems with hypothesis tes
Anderson et al. 2000). This approach is hardly satisfactor
ignored the issue and merely pretended that data dredging
that the usual inferential methods somehow still apply. **J**
referees rarely seem to show concern for the validity
clusions where substantial data dredging has occurre
methodology based on data dredging has been allowe
in an unthinking manner.

We certainly encourage people to understand their data a
the scientific questions of interest. We advocate some exa
prior to the formal analysis to detect obvious outliers and
determine a preliminary truncation point or the need for gro
of distance sampling data). One might examine the residu
chosen global model to determine likely error distributi
models (e.g., normal, lognormal, Poisson). However, if
is noticed while examining the residuals and this leads
variable, then we might suggest caution concerning data d
can be a fine line between a largely a priori approach and
dredging.

Thus, this book will address primarily cases where t
priori knowledge concerning the issue at hand and where
of good candidate models can be specified in advance of
Of course, there is some latitude where some (few) additio
investigated as the analysis proceeds; however, results fro
should be kept clearly separate from the purely a priori
that objective science is best served using a priori cons
limited peeking at plots of the data, parameter estimates f
els, correlation matrices, or test statistics as the analysis
condone data dredging in confirmatory analyses, but allow
in more preliminary explorations. If some limited data dre
careful analysis based on prior considerations, then we b
types of results should be carefully explained in resulting
1980). For this philosophy to succeed, there should be
consideration of alternative candidate models than has
past.

1.5.2 Some Trends

At the present time, nearly every analysis is done usi
biologists and researchers in other disciplines are increasi
methods for more generalized analyses. Standard compute

should generally be avoided, except in (1) the early stages
or (2) *after* a more confirmatory analysis has been done
the investigator should fully admit to the process that
results and should treat them much more cautiously than
the initial, a priori, approach. When done carefully, we e
explore their data beyond the important a priori phase.
We recommend a substantial, deliberate effort to get the a
models in place and try to obtain more confirmatory re
the post hoc issues that often arise after one has seen the
results.
Data dredging activities form a continuum, ranging from f
to the grievous (mortal). There is often a fine line betwee
our advice is to stay well toward the a priori end of the c
achieve a more confirmatory result.
One can always do post hoc analyses after the a priori an
never go from post hoc to a priori. Why not keep one's c
regard?
Grievous data dredging is endemic in the applied literature
taught or implied in statistics courses without the needed
the attendant inferential problems.
Running all possible models is a thoughtless approach an
of finding effects that are, in fact, spurious if only a sing
for inference. If prediction is the objective, model avera
estimates of precision should include model selection u
this case, surely one can often rule out many models on

allow likelihood methods to be used where LS methods
the past. LS methods will see decreasing use, and likel
see increasing use as we proceed into the twenty-first
methods allow a much more general framework for addres
(e.g., a choice of link functions and error distributions
logistic regression models). Another advantage in a likelih
confidence intervals with good properties can be set usi
intervals. Edwards (1976), Berger and Wolpert (1984), Az
(1997), and Morgan (2000) provide additional insights into
while Box (1978) provides the historical setting relating
methods.

During the past twenty years, modern statistical scien
away from traditional formal methodologies based on
sis testing (Clayton et al. 1986, Jones and Matloff 1
Bozdogan 1994, Johnson 1995, Stewart-Oaten 1995, N
1999, Anderson et al. 2000). The historic emphasis on hy

sizes and associated confidence intervals (Graybill and Iy
Reid 2000).

Most researchers recognize that we do not conduct ex
reject null hypotheses or claim statistical significance; we
than this. We typically want to compare meaningful (i.e
tives, or seek information about effects and their size a
interested in causation. **There has been too much forma
confusion that leads people to think that statistics an
is mostly about testing uninteresting or trivial null h
science is much more than this. We must move bey
testing-based thinking because it is so uninformative.**

In particular, hypothesis testing for model selection is
1981a) and will surely diminish in the years ahead. Th
theory that supports the notion that hypothesis testing w
a basis for model selection. There are not even general f
guidelines) that rigorously define how the various P-valu
arrive at a final model. How does one interpret dozens of
with differing power, to arrive at a good model? Only α
this case and generally fail to result in a final parsimonio
inferential properties. The multiple testing issue is probl
that likelihood ratio tests exist only for nested models.
within a data set are not independent, making inferences
of testing is arbitrary, and differing test order will often l
models. Model selection is dependent on the arbitrary choi
depend on both n and K to be useful in model selection;
this is lacking. Testing theory is problematic when nuisan
in the models being considered. Finally, there is the fa
null is probably false on simple a priori grounds (e.g., H
no effect, so the parameter θ is constant across treatme
$\theta_1 = \theta_2 = \cdots = \theta_k$). Rejection of such null hypotheses do
effect or parameter should be included in the approximati
testing approach is both common and somewhat absurd. A
have been well known in the literature for many years; th
ignored in the practical analysis of empirical data. Neste
interesting summary of quotations regarding hypothesis t

Unfortunately, it has become common to compute estim
a hypothesis test has been conducted and found to be nons
hoc power is not valid (Goodman and Berlin 1994, Gerar
and Heisey 2001). While a priori power and sample siz
important in planning an experiment or observational stu
hoc power are not valid and should not be reported (And

Computational restrictions prevented biologists from e
models until the past two decades or so. Thus, people tend

models as an integral component of data analysis. Compu
mitted more computer-intensive methods such as the vari
and bootstrapping approaches and other resampling sch
Duval 1993, Efron and Tibshirani 1993), and such techn
increasing use in the future.

The size or dimension (K) of some biological models ca
this has tended to increase over the past two decades. Ope
and band recovery models commonly have 20–40 estima
single data set and might have well over 200 parameters
of several data sets (see Burnham et al. 1987, Preface, for a
these trends). Analysis methods for structural equations co
30 parameters (Bollen and Long 1993). These are applicat
model specification and selection is essential to answer
inferences do the data support about the population?"

1.6 Model Selection Bias

The literature on model selection methods has increased
past 15–25 years; much of this has been the result of Akaik
in the mid-1970s. However, relatively little appears in the l
the properties of the parameter estimators, given that a da
selection procedure has been used (see Rencher and Pur
Tsai 1990, Miller 1990, Goutis and Casella 1995, Ye 19
used to both select a parsimonious model and estimate th
and their precision (i.e., the conditional sampling covarian
selected model). These issues prompt a concern for both
and model selection uncertainty (Section 1.7).

Bias in estimates of model parameters often arises wh
tion has been done. Miller (1990) provides a technical
selection bias in the context of linear regression. He notes
stepwise analysis of meteorological data with large sample
date models. When selecting only about 5 variables from
he found t statistics as large as 6, suggesting that a par
very highly significant, and yet even the sign of the corre
coefficient could be incorrect. Miller warns that P-value
tion software are totally without foundation, and large
coefficients are often caused by data-based model selecti

Consider a linear model where there is a response va
planatory variables x_j, where $j = 1, \ldots, 4$. Order is n
example, so for convenience let x_1 be, in fact, very import
somewhat important, while x_4 is barely important. Given a

data sets of the same size, from the same stochastic proc
and x_2 would be included in the model in nearly all case
inference from a sample data set to the population would
selected that included predictors x_1 and x_2 (essentially al
estimators of the regression coefficients associated with
would have good statistical properties with respect to bia
standard theory tends to hold for the estimators $\hat{\beta}_1$ and $\hat{\beta}$

Variable x_3 is somewhat marginal in its importance; a
that $|\beta_3|/\text{se}(\beta_3) \approx 1$, and thus its importance is somewhat
might be included in the model in only 15–30% of the 1,0
sets where it is selected, it tends to have an estimated re
that is biased away from zero. Thus, an inference from
concerning the population tend to exaggerate the importan
An inference from a data set in one of the remaining 70–
would imply that x_3 was of no importance. Neither of these

Variable x_4 is barely important at all (a tapering effect
$|\beta_4|/\text{se}(\beta_4) \approx \frac{1}{4}$. This variable might be included in only
of the 1,000 data sets and, when it is selected, there will l
(away from 0) in the estimator of this regression paramet
particular sample where this variable is included in the mo
the variable x_4 was much more important than is actually
the investigator has no way to know that $\hat{\beta}_4$, when selec
upper 5–10% of its sampling distribution). Then, if one
t-test, where $t = \hat{\beta}_4/\,\widehat{\text{se}}(\hat{\beta}_4)$, the likely decision will often
x_4 is significant, and should be retained in the model. Th
comes from the fact that the numerator in the test is bia
denominator is biased low. The analyst has no way to kno
is probably spurious.

When predictor variables x_3 and x_4 are included in mo
estimator for a σ^2 is negatively biased and precision is exa
types of bias are called model selection bias and can of
(Miller 1990, Ye 1998). Ye (1998) warns, "... the iden
structure bears little cost [i.e., including variables x_1 and x_2
through white noise has a heavy cost [i.e., including varia
Of course, in the analysis of real data, the investigator typ
which (if any) variables are dominant versus those that are
importance. Model selection bias is related to the proble
notion of tapering effect sizes, and Freedman's (1983) pa

The problem of model selection bias is particularly seric
is available to guide the analysis. Many exploratory studi
even thousands of models, based on a large number of ex
very often the number of models exceeds the size of the
model has been (somehow) selected, the analyst is usual

knowingly extracted some of the residual variation as if i
structure. When sample size is large, true replication e
relatively few models, these problems may be relatively
ever, often one has only a small sample size, no true re
models and variables; then model selection bias is usua
2000).

If, for example, x_3 is uncorrelated with x_1, x_2, and x_4,
of $\hat{\beta}_3$ is symmetric around β_3 and bias, given that x_3 is se
$\beta_3 = 0$, then $\mathrm{E}(\hat{\beta}_3) = 0$). This is an interesting result, but p
in practice because predictor variables are almost always
the case where $\beta_3 = 0$, but x_3 is highly correlated with
the correlation between x_1 and x_3 is high (even 0.5) and
variable x_3 is selected, it is much more likely to be when $\hat{\beta}$
where x_3 is selected, $\hat{\beta}_3$ tends to be positive. In cases w
between x_1 and x_3 is negative, then $\hat{\beta}_3$ tends to be negative
biased low. By itself, x_3 would have some predictive value
its correlation with x_1, which is actually correlated with th

If sample size is small and there are many variables and
the negative bias in $\hat{\sigma}^2$ is often severe. If the predictor
intercorrelated and only one (say x_{11}) is actually correlate
variable, then the estimates of the regression coefficients
stantially biased away from 0 in the subset of models w
predictor variable is selected. Leamer (1978), Copas (198
Gilchrist (1984), Breiman (1992), Zhang (1992*a*), and Cha
give insights into problems that arise when the same data a
the model and to make inferences from that model.

1.7 Model Selection Uncertainty

Model selection uncertainty also arises when the data are
selection and parameter estimation (Hjorth 1994:15–23).
been selected from a reasonable set of candidate models, t
rameter estimators might be small for several of the more
but might be substantial for variables associated with tap
ever, there is uncertainty as to the best model to use. From
one must ask whether β_3 or β_4 should be in the model; th
is a component of variance in the estimators.

Denote the sampling variance of an estimator $\hat{\theta}$,
$\mathrm{var}(\hat{\theta}|model)$. More generally, the sampling variance of
components: (1) $\mathrm{var}(\hat{\theta}|model)$ and (2) a variance compon

given the data, and estimates a conditional sampling va
lected model. Then estimated precision will be too small
component for model selection uncertainty is missing. M
tainty is the component of variance that reflects that mo
estimates which model is best, based on the single data s
(in the fixed set of models considered) may be selected a
replicate data set arising from the same experiment.

Failure to allow for model selection uncertainty often
sampling variances and covariances that are too low, an
confidence interval coverage will be below the nominal val
for coping with model selection uncertainty are at the fo
research; better methods might be expected in the coming y
the continued increases in computing power. Model sele
problematic in making statistical inferences; if the goal is o
then perhaps selection uncertainty is a minor issue.

One must keep in mind that there is often considerable u
lection of a particular model as the "best" approximating m
data are conceptualized as random variables; their value
if another, independent sample were available. It is this "s
that results in uncertain statistical inference from the part
analyzed. While we would like to make inferences that
other (hypothetical) data sets, our ability to do so is stil
with procedures such as AIC, with its cross-validation pro
dependent and identically distributed sample data. Various
resampling methods will further improve our assessment
our inferences, but it remains important to understand th
lection is accompanied by a substantial amount of uncert
technique can effectively allow insights into model uncer
similar issues are the subject of Chapter 5.

Perhaps we cannot totally overcome problems in estim
lowing a data-dependent selection method such as AIC (e.g
Ye 1998). This limitation certainly warrants exploration b
tion uncertainty is a quite difficult area of statistical infe
must also consider the "cost" of *not* selecting a good pars
the analysis of a particular data set. That is, a model is jus
independent of the data and used to approximate the da
ference. This procedure simply ignores both the uncerta
model selection and the benefits of selection of a model t
This naive strategy certainly will incur substantial costs in
ferences because model selection uncertainty is ignored (
Alternatively, one might be tempted into an iterative, hig
egy of data analysis (unadulterated data dredging). Again,
costs in terms of reliable inference using this approach. In

1.8 Summary

Truth in the biological sciences and medicine is extreme
we cannot hope to find exact truth or full reality from
nite amount of data. Thus, inference about truth must
approximating model. Likelihood and least squares meth
ous inference theory if the model structure is "given." H
scientific problems, the model is *not* "given." Thus, the cr
is the best model to use." This is the model selection prob

The emphasis then shifts to the careful a priori definition
models. This is where the science of the problem enters
there should be a good rationale for including each par
set, as well as a careful justification for why other models
degree to which these steps can be implemented suggests
analysis, rather than a more exploratory analysis. Critica
scientific question and modeling alternatives, prior to look
been underemphasized in many statistics classes in the pa
tant issues, and one must be careful not to engage in dat
this weakens inferences that might be made. Informatio
provide a simple way to select a best approximating mode
set of models.

In general, the information-theoretic approach shoul
searching for a single best model as a basis for inference.
tion uncertainty is included in estimates of precision, thi
in many cases. Instead, multimodel inference should be
to making valid inference. Here, models are ranked and s
understanding of model uncertainty over the set. These me
derstand and compute. Specific methodologies for this mo
are the subject of this book.

We cannot overstate the importance of the scientifi
formulation of multiple working hypotheses, and the bu
of models to clearly and uniquely represent these hypo
to be presented in the following chapters are "easy" to u
and interpret; however, they rest on both good science and
to the issue. We try to emphasize a more confirmatory enc
sciences, rather than exploratory work that has become
often led to so little (Anderson et al. 2000).

Data analysis is taken to mean the entire integrate
ori model specification, model selection, and estimation
their precision. Scientific inference is based on this pr

The principle of parsimony is fundamental in the selec
based selection of a parsimonious model is challenging. T
rewards for proper model selection in terms of valid i
substantial dangers in either underfitting or overfitting. H
has selected a good approximating model, there are issue
bias and model selection uncertainty. Perhaps these canno
but their effects can be lessened. These issues will be addr
to follow.

Zhang (1994) notes that for the analyst who is less conce
optimality it is more important to have available method
flexible enough to be used in a variety of practical situatio
theoretic methods fall in this broad class and, when use
reliable inference.

Information and Likelihood T
Basis for Model Selection and

Full reality cannot be included in a model; thus we seek
proximate the effects or factors supported by the empirica
of an appropriate approximating model is critical to statis
many types of empirical data. This chapter introduces c
mation theory (see Guiasu 1977), which has been a disci
mid-1940s and covers a variety of theories and methods t
to many of the sciences (see Cover and Thomas 1991 for a
Figure 2.1 is produced from their book and shows th
lationship of information theory to several other fields
Kullback–Leibler "distance," or "information," between
back and Leibler 1951) is introduced, discussed, and lin
entropy in this chapter. Akaike (1973) found a simple r
the Kullback–Leibler distance and Fisher's maximized
tion (see deLeeuw 1992 for a brief review). This relationsh
effective, and very general methodology for selecting a
for the analysis of empirical data.

Akaike introduced his "*entropy maximization principle*"
in the mid-1970s (Akaike 1973, 1974, 1977) as a theoret
selection. He followed this pivotal discovery with several
beginning in the early 1980s (Akaike 1981a and b, 198
This chapter introduces AIC and related criteria such as
TIC. No mathematical derivations of these criteria are giv
are given in full detail in Chapter 7. We urge readers to
derivation (given in Chapter 7), for without it, the simple

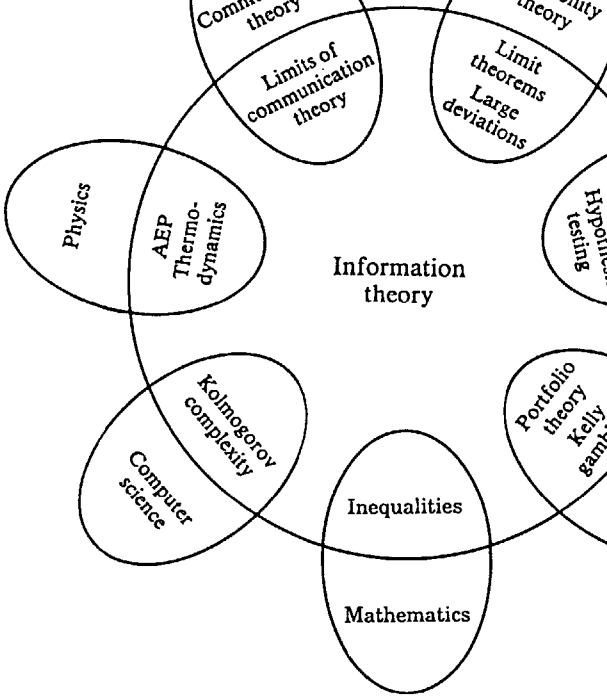


FIGURE 2.1. Information theory and its relationships to other disc
Thomas 1991). Information theory began in the mid-1940s, at the
context of this book, the most relevant components of informatio
information, entropy (from thermodynamics and communication
Leibler information.

underlying Kullback–Leibler information and the various
cannot be fully appreciated.

2.1 Kullback–Leibler Information or Dis Two Models

We begin without any issues of parameter estimation and
expressions for the models f and g, assuming that they ar
In initial sections of this chapter we will let both f and g b
distributions, since this will allow an understanding of
distance in a simple setting. However, we will soon switc

Kullback–Leibler Information

Kullback-Leibler information between models f an
continuous functions as the (usually multi-dimensional)

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

where log denotes the natural logarithm. The notation
"information lost when g is used to approximate f."

As a heuristic interpretation, **$I(f,g)$ is the distance fr**

We will use both interpretations throughout this book, s
ful. Of course, we seek an approximating model that loses
as possible; this is equivalent to minimizing $I(f, g)$, ov
is considered to be given (fixed), and only g varies ove
indexed by θ. Similarly, Cover and Thomas (1991) note t
is a measure of the inefficiency of assuming that the distri
true distribution is f.

Kullback–Leibler Information

The expression for the Kullback-Leibler information
case of discrete distributions such as the Poisson, binom
is

$$I(f,g) = \sum_{i=1}^{k} p_i \cdot \log\left(\frac{p_i}{\pi_i}\right).$$

Here, there are k possible outcomes of the underlying ra
true probability of the ith outcome is given by p_i, w
constitute the approximating probability distribution (i.e.
model). In the discrete case, we have $0 < p_i < 1$,
$\sum p_i = \sum \pi_i = 1$. Hence, here f and g correspond
respectively.

As in the continuous care the notation **$I(f,g)$** denotes t
when g is used to approximate f or the distance from

In the following material we will generally think of K-
continuous case and use the notation f and g for simplic

Well over a century ago measures were derived for asse
between two models or probability distributions. Most re
mann's (1877) concept of generalized entropy (see Sect
and thermodynamics (see Akaike 1985 for a brief revie
employed entropy in his famous treatise on communicatio
2001 for an exciting review of information theory, its pract
to evolution). Kullback and Leibler (1951) derived an infor

Ludwig Eduard Boltzmann, 1844–1906, one of the most famous made incredible contributions in theoretical physics. He received most of his work was done in Austria, but he spent some years i full professor of mathematical physics at the University of Graz, A His mathematical expression for entropy was of fundamental impo areas of science. The negative of Boltzmann's entropy is a measure o over half a century later by Kullback and Leibler. J. Bronowski wr "an irascible, extraordinary man, an early follower of Darwin, qua and everything that a human should be." Several books chronicle th of science, including Cohen and Thirring (1973) and Broda (1983) papers appear in Hasenöhrl (1909).

happened to be the negative of Boltzmann's entropy, no Kullback–Leibler (K-L) information or distance (but see K he preferred the term *discrimination information*). The back and Leibler's work was to provide a rigorous definit in relation to Fisher's "sufficient statistics." The K-L dis called the K-L discrepancy, divergence, information, and these terms as synonyms, but tend to use *distance* or *inform* to follow.

The Kullback–Leibler distance can be conceptualized tance" between two models, say f and g (Kullback 195 this is a measure of "discrepancy"; it is not a simple d measure from f to g is not the same as the measure directed, or oriented, distance (Figure 2.2). The K-L dis

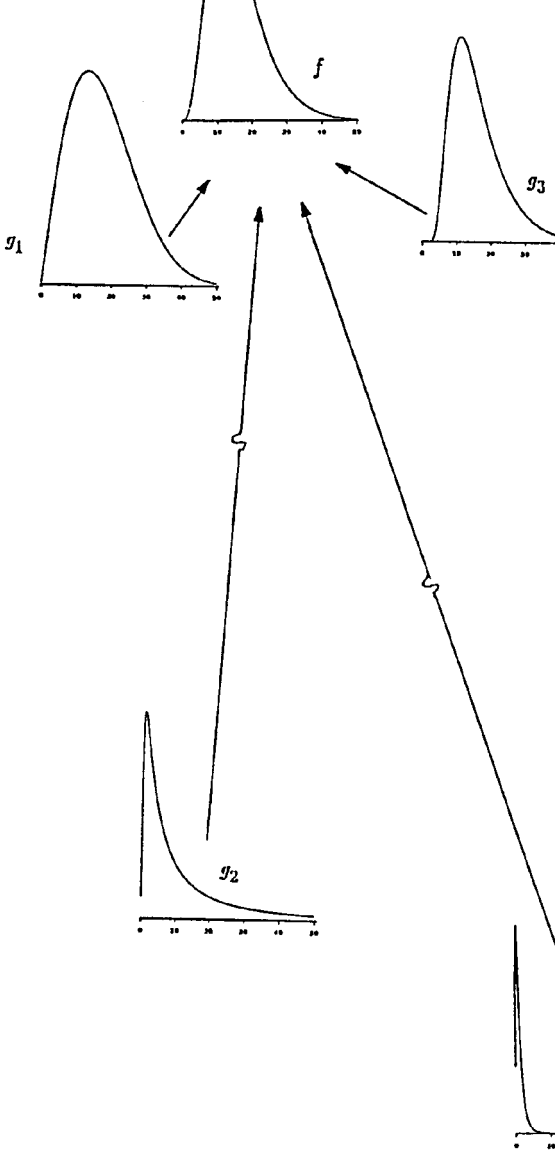


FIGURE 2.2. The Kullback–Leibler discrepancy $I(f, g_i)$ is a dir
various candidate models g_i to f. Knowing the K-L distances w
which of the 4 approximating models is *closest* to model f. Here,
the 4 approximating models are g_1 = Weibull (2, 20), g_2 = lognor
Gaussian (16, 64), and g_4 = F distribution (4, 10). In each case, th
known exactly (not estimated).

an extension of Shannon's concept of information (Ebrahimi and Soofi 1994) and is sometimes called a "relative entropy between models is a *fundamental quantity* in science and (see Akaike 1983) and is the logical basis for model sele with likelihood inference.

At a heuristic level, "information" is defined as -1 continuous probability density function or $-\log_e(p_i)$ fo Kullback–Leibler information is a type of "cross entropy ization. In either the continuous or discrete representation is an expected value (i.e., $\int f(x)(\cdot)dx$ for the continuou for the discrete case) of the logarithm of the ratio of the and g) or two discrete probabilities (p_i and π_i). In the can think of this as an average (with respect to f) of $\log_e($ crete case it is an average (with respect to the p_i) of the l (p_i/π_i). The foundations of these expressions are both de (see Boltzmann 1877, Kullback and Leibler 1951, or con information theory).

The K-L distance ($I(f, g)$) is always positive, except butions f and g are identical (i.e., $I(f, g) = 0$ if and c everywhere). More detail and extended notation will be in 7; here we will employ a simple notation and use it to generality in the sample data (x) and the multivariate fun

2.1.1 *Examples of Kullback–Leibler Distance*

An example will illustrate the K-L distances ($I(f, g_i)$). distribution with 2 parameters ($\alpha = 4$, $\beta = 4$). Then consi models g_i, each with 2 parameters (see below): Weibull Gaussian, and the F distribution. Details on these simple pr be found in Johnson and Kotz (1970). The particular paran the four g_i are not material here, except to stress that they not estimated. "Which of these parametrized distributions is answered by computing the K-L distance between ea 2.2). These are as follows:

	Approximating model	$I(f, g$
g_1	Weibull distribution ($\alpha = 2, \beta = 20$)	0.046
g_2	lognormal distribution ($\theta = 2, \sigma^2 = 2$)	0.672
g_3	inverse Gaussian ($\alpha = 16, \beta = 64$)	0.060
g_4	F distribution ($\alpha = 4, \beta = 10$)	5.745

Here, the Weibull distribution is closest to (loses the leas f, followed by the inverse Gaussian. The lognormal distrib

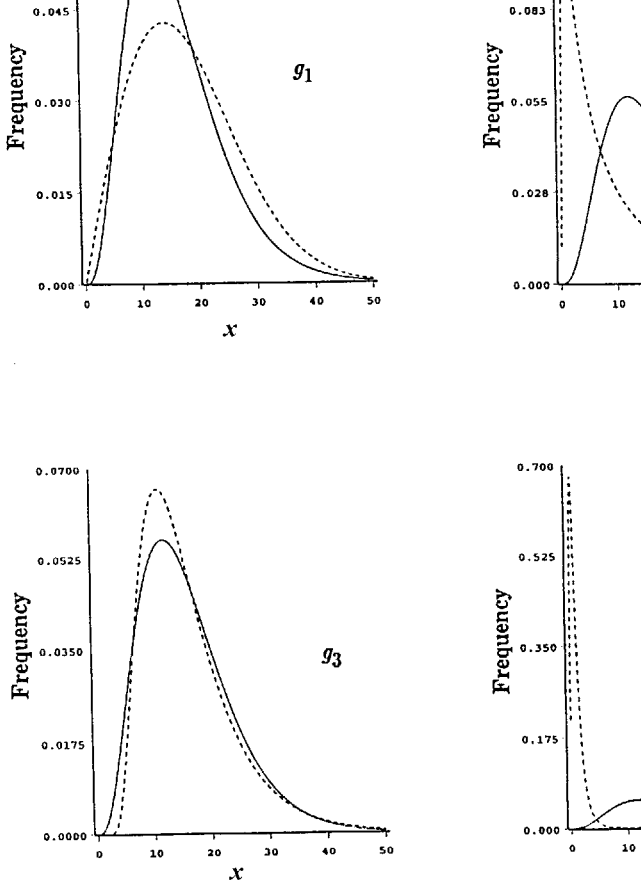


FIGURE 2.3. Plots of f (= gamma (4, 4), solid line) against each
models g_i (dashed lines) as a function of x. Here, g_1 = Weibull (
(2, 2), g_3 = inverse Gaussian (16, 64), and g_4 = F distribution (4, 1
cases can plots such as these be used to judge closeness between
same in all 4 graphs; it is merely scaled differently to allow the g_i
same graph.

while the F distribution is relatively far from the gamma
Figure 2.3).

Further utility of the K-L distance can be illustrated by
approximating models g_i might be closest to f when the
allowed to vary (i.e., what parameter values make each g
f?). Following a computer search of the parameter space
found that the *best* Weibull had parameters $\alpha = 2.120$ an

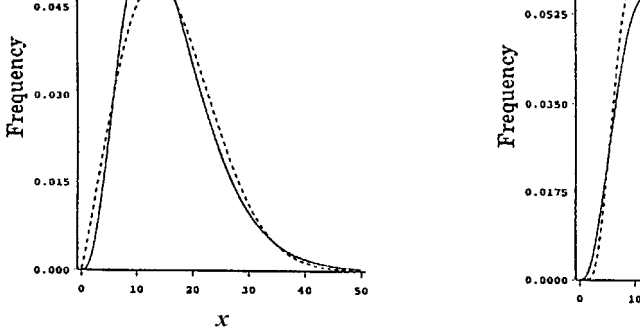


FIGURE 2.4. Plots of f (= gamma (4, 4)) against the best Weibu models. The Weibull model that was closest to f had parameters (2 distance = 0.02009, while the best lognormal had parameters (2 distance = 0.02195. Compare these optimally parametrized mode 2.3 (top).

K-L distance of 0.02009; this is somewhat closer than the c tion 0.04620 above. Using the same approach, the best lo parameters $\theta = 2.642$ and $\sigma^2 = 0.2838$ and a K-L distan the best inverse Gaussian model had parameters $\alpha = 16$ K-L distance of 0.03726, and the approximately best F rameters $\alpha \approx 300$, $\beta = 0.767$ and a K-L distance of appr K-L distance is not sensitive to α in this case, but is quite numerically). Thus, K-L distance indicates that the best W than is the best lognormal (Figure 2.4). Note that the forma distance requires knowing the true distribution f as well in the models g_i (i.e., parameter estimation has not yet be K-L distance cannot be computed for real-world problem

These values represent *directed* distances; in the firs $I(f, g_1) = 0.04620$, while $I(g_1, f) = 0.05552$ (in fa be interested in $I(g_1, f)$ since this is the information los approximate g!). The point here is that these are directed and $I(f, g_1) \neq I(g_1, f)$; *nor should they be equal, becau and model are not interchangeable*.

These are all univariate functions; thus one could mer same scale and visually compare each g_i to f; however, th will work only in the simplest cases. In addition, if two a butions are fairly close to f, it might be difficult to decid only visual inspection. Values of the K-L distance are n mean and variance of the distributions; rather, the distribut are the subject of comparison.

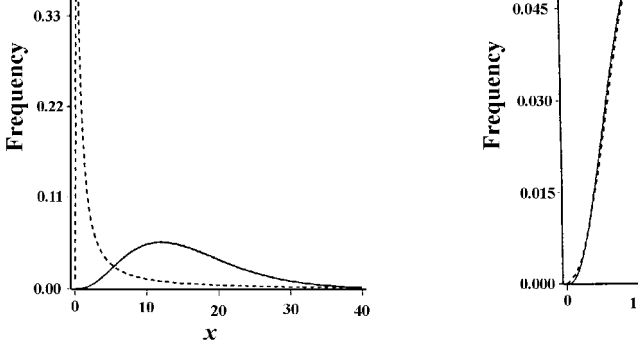


FIGURE 2.5. Plots of f (= gamma (4, 4)) against the best 2-param and the best 3-parameter (noncentral) F distribution. The best 2- poor approximation to f (K-L distance = 1.486), while the best 3 excellent approximation (parameters 1.322, 43.308, 18.856) with K- Approximating models with increasing numbers of parameters typic approximating models with fewer parameters.

The F distribution ($\alpha = 4$, $\beta = 10$) provided a relatively to the gamma distribution with ($\alpha = 4$, $\beta = 4$). Even the distribution remains a relatively poor approximation (K-L However, in general, adding more parameters will result in tion (e.g., the classic use of the Fourier series in the physic (1975) elephant-fitting problem). If we allow the addition (λ) in the F distribution (the noncentral F distribution), model ($\alpha = 1.322$, $\beta = 43.308$, and $\lambda = 18.856$) has a K 0.001097; this is better than any of the other 2-paramet (Figure 2.5). Closeness of approximation can always be more parameters to the candidate model. When we co parameters and the associated uncertainty, then the prin must be addressed (see Section 1.4), or overfitted models

In the remainder of the book we will want a more genera f, and we will use it to reflect truth or full reality. Here, rea a model; rather, it reflects the complex biological (and mea process that generated the observed data x. For this reason parametrize the complex function f, because it represent it might not even have parameters in a sense that would be modeling framework. In fact, thinking that truth is parame of (artificial) model-based conceptualization. Sometimes i f as full reality and let it have (conceptually) an infinite nu (see Section 1.2.4). This "crutch" of infinite-dimensionali concept of reality even though it is in some unattainable

rise to a set of data x. Finally, we will see how this concep
(f) collapses into a nonidentifiable constant in the contex

2.1.2 *Truth, f, Drops Out as a Constant*

The material above makes it obvious that both f and g (a
must be known to compute the K-L distance between thes
ever, if only relative distance is used, this requirement
$I(f, g)$ can be written equivalently as

$$I(f, g) = \int f(x) \log(f(x))dx - \int f(x) \log(g$$

Note that each of the two terms on the right of the ab
statistical expectation with respect to f (truth). Thus, the K
can be expressed as a difference between two statistical e

$$I(f, g) = \mathrm{E}_f \left[\log(f(x))\right] - \mathrm{E}_f[\log(g(x \mid$$

each with respect to the distribution f. This last expre
insights into the derivation of AIC.

The first expectation $\mathrm{E}_f[\log(f(x))]$ is a constant that
unknown true distribution, and it is clearly not known (i
f in actual data analysis). Therefore, treating this unkn
stant, a measure of *relative* directed distance is possib
Kapur and Kesavan 1992:155). Clearly, if one computed
tion $\mathrm{E}_f[\log(g(x \mid \theta))]$, one could estimate $I(f, g)$ up to a
$\mathrm{E}_f[\log(f(x))]$),

$$I(f, g) = C - \mathrm{E}_f[\log(g(x \mid \theta))],$$

or

$$I(f, g) - C = -\mathrm{E}_f[\log(g(x \mid \theta))].$$

The term $\big(I(f, g) - C\big)$ is a *relative* directed distance bet
$\mathrm{E}_f \left[\log(g(x \mid \theta))\right]$ becomes the quantity of interest for sel
For two models g_1 and g_2, if $I(f, g_1) < I(f, g_2)$, so g_1 is
$C < I(f, g_2) - C$, and hence $-E_f[\log(g_1(x|\theta))] <$
Moreover, $I(f, g_2) - I(f, g_1) \equiv -E_f[\log(g_2(x|\theta))] +$
we know how much better model g_1 is than model g_2.
we just do not know the absolute measure of how good
can identify the fact that model g_1 is better than g_2. No
estimation is involved here, but the concepts carry over to
timation occurs. From the preceding example, where f is
$\int f(x) \log(f(x))dx = 3.40970$, and this term is constan

The *relative* distances between the gamma $(4, 4)$ m
approximating models are shown below:

	Approximating model	Relative d $I(f, g_i)$
g_1	Weibull distribution ($\alpha = 2, \beta = 20$)	3.455
g_2	lognormal distribution ($\theta = 2, \sigma^2 = 2$)	4.082
g_3	inverse Gaussian ($\alpha = 16, \beta = 64$)	3.469
g_4	F distribution ($\alpha = 4, \beta = 10$)	9.155

Note that the ranking of "closeness" of the four can
is preserved, and the relative ranking of distance betw
unchanged, even though only relative distances are used.

Kullback-Leibler distance $I(f, g)$ is on a true ratio sc
true zero. In contrast, $-\int f(x)(\log(g(x|\theta)))dx \equiv -\mathrm{E}_f[$
interval scale and lacks a true zero. A difference of mag
same thing anywhere on the scale. Thus, $D = 10 = 12 -$
a difference of 10 means the same thing anywhere on the
$10 = V_1 - V_2$, regardless of the size of V_1 and V_2.

The calculation of the two components of K-L distan
fect based on a sample size of 1. If the sample size w
component would be 100 times larger, and the differen
components would also be 100 times larger. For exampl
$\int f(x)\log(f(x))dx = 3.40970 \times 100 = 340.970$ and E
Weibull) $= 3.45591 \times 100 = 345.591$. Thus, the differe
components of K-L distance would be 4.620; the *relativ*
when sample size is large. A large sample size magnifies
search hypotheses and the models used to represent them
size conveys a wide variety of advantages in making v

Typically, as in the example above, the analyst would
priori candidate models $g_i(x \mid \theta)$ and want to select the
a basis for data analysis and inference. Definition of "b
principle of parsimony and the related concept of a best ap
In data analysis, the parameters in the various candidate m
and must be estimated from the empirical data. This rep
distinction from the material above, since one usually ha
estimated parameters, denoted by $g_i(x \mid \hat{\theta})$. In this case,
of the relative directed distances between the unknown
data and the various candidate models $g_i(x \mid \hat{\theta})$. Then, kn
relative distance from each $g_i(x)$ to $f(x)$, we select the c
is *estimated* to be closest to truth for inference (Figure 2.
the model with the smallest estimated, *relative* distanc
select an approximating model that loses the least informa

irrelevant for comparison. (Similarly, it is interesting to no
likelihood function also involves an additive constant tha
models; this term is known, but generally ignored, since
to compute.) In practice, we can obtain only an *estimato*
distance from each approximating model $g_i(x \mid \hat{\theta})$ to f.

2.2 Akaike's Information Criterion: 1973

Akaike's (1973) seminal paper proposed the use of the l
formation or distance as a fundamental basis for model
K-L distance cannot be computed without full knowledg
ality) and the parameters (θ) in each of the candidate mo
found a rigorous way to estimate K-L information, bas
log-likelihood function at its maximum point.

Given a parametric structural model there is a unique
fact, minimizes K-L distance $I(f, g)$. This (unknown) mi
parameter depends on truth f, the model g through its stru
space, and the sample space (i.e., the structure and nature o
collected). In this sense there is a "true" value of θ under
let this value be θ_0. Then θ_0 is the absolute best value of θ
K-L information loss is minimized at θ_0. If one someho
g was, in fact, the K-L best model, then the MLE $\hat{\theta}$ wou
property of the model $g(x|\theta_0)$ as the minimizer of K-L, o
important feature involved in the derivation of AIC (Chap

In data analysis the model parameters must be estimated
substantial uncertainty in this estimation. Models based o
ters, hence on $\hat{\theta}$ not θ, represent a major distinction from th
parameters would be known. This distinction affects ho
distance as a basis for model selection. The difference b
θ_0 (we do not) and having the estimate $\hat{\theta}$ (we do) is quite

Selection Target

Akalke (1973, 1974, 1985, 1994) showed that the critical
applied K-L model selection criterion was to estimate

$$\mathbf{E}_y\mathbf{E}_x[\log(g(x|\hat{\theta}(y)))],$$

where x and y are independent random samples from th
and both statistical expectations are taken with respect
double expectation, both with respect to truth f, is the
selection approaches, based on K-L information.

over the set of R models considered.

It is tempting to just estimate $E_y E_x[\log(g(x|\hat{\theta}(y)))]$ $\log(\mathcal{L}(\hat{\theta})|data)$ for each model g_i. However, Akalke (19 maximized log-likelihood is biased upward as an esti selection target (above). He also found that under certa conditions are important, but quite technical) this bias is to K, the number of estimable parameters in the approxim an asymptotic result of fundamental importance.

The Key Result

Thus, an approximately unbiased estimator of

$$\mathbf{E_y E_x[\log(g(x|\hat{\theta}(y)))]}$$

for large samples and "good" models is

$$\mathbf{\log(\mathcal{L}(\hat{\theta}|data)) - K.}$$

This result is equivalent to

$$\mathbf{\log(\mathcal{L}(\hat{\theta}|data)) - K = constant - \hat{E}_{\hat{\theta}}[I(f}$$

where $\hat{g} = g(\cdot|\hat{\theta})$.

The bias-correction term ($K =$ the number of estimabl is a special case of a more general result derived by T described in the following section and in Chapter 7. **Ak relation between the relative expected K-L distance a log-likelihood has allowed major practical and theo model selection and the analysis of complex data se** Bozdogan 1987, and deLeeuw 1992).

Akaike's Information Criterion

Akaike (1973) then defined "*an information criterio* plying $\log(\mathcal{L}(\hat{\theta}|y)) - K$ by -2 ("taking historical reaso get

$$\mathbf{AIC = -2\log(\mathcal{L}(\hat{\theta}|y)) + 2K.}$$

This has become known as **"Akaike's information crit**

Thus, rather than having a simple measure of the di tween two models (i.e., the K-L distance), one has ins the expected, relative distance between the fitted mode true mechanism (perhaps of infinite dimension) that act observed data.

The expression $\log(\mathcal{L}(\hat{\theta}|y))$ is the numerical value of the maximum point (see Section 1.2.2). This maximum point

els). In some types of models there are some parameters t
estimable from the data, and these should not be counted in
can occur in the analysis of count data where a cell has no ob
a parameter that is identifiable becomes nonestimable for
timability can also arise due to inherent confounding (e.g.
and f_t in certain band recovery models of Brownie et al. 1
one computes AIC for each of the candidate models and se
the smallest value of AIC. It is this model that is estimate
the unknown reality that generated the data, from among t
considered. This seems a very natural, simple concept; sel
imating model that is estimated, on average, to be closes
Basing AIC on the expectation (over $\hat{\theta}$) of $\mathrm{E}_x[\log(g(x|\theta$
with a cross-validation property for independent and id
samples (see Stone 1977, Stoica et al. 1986, Tong 1994).
show that AIC asymptotically coincides with generalized
subset regression (also see review by Atilgan 1996).

**Of course, models not in the set remain out of co
useful in selecting the best model in the set; however, i
very poor, AIC will still select the one estimated to be
relatively best model might be poor in an absolute sense
must be made to ensure that the set of models is well f**

$I(f, g)$ can be made smaller by adding more known
rameters in the approximating model g. Thus, for a fixed
addition of parameters in a model g_i will allow it to be cl
when these parameters must be estimated (rather than bein
further uncertainty is added to the *estimation* of the relati
some point, the addition of still more estimated paramete
posite from desired effect (i.e., to reduce $\mathrm{E}_{\hat{\theta}}[I(f, \hat{g})]$ as de
the estimate of the relative K-L distance will increase be
estimated parameters that are not really needed to achieve
phenomenon can be seen by examination of the informa
minimized,

$$\mathbf{AIC} = -2\log(\mathcal{L}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})) + 2\boldsymbol{K},$$

where the first term on the right-hand side tends to de
rameters are added to the approximating model, while th
gets larger as more parameters are added to the approxi
is the tradeoff between bias and variance or the tradeoff b
and overfitting that is fundamental to the principle of par
1.4.2). Some investigators have considered K to be a m
ity," but this is unnecessary, though not irrational. We co

expected K-L information, thus parsimony arises as a b
proach. Further books and papers on the derivation of A
(1983, 1989), Linhart and Zucchini (1986), Bozdogan (1
(1991).

Usually, AIC is positive; however, it can be shifted by a
and some shifts can result in negative values of AIC. C
regression statistics (see Section 1.2.2) often results in neg
our work, we have seen minimum AIC values that range
numbers to as high as 340,000. **It is not the absolute siz
it is the relative values over the set of models considere
the differences between AIC values (Section 2.5), that**

The material to this point has been based on likelihoo
very general approach. In the special case of least squares
normally distributed errors, and apart from an arbitrary ad
can be expressed as a simple function of the residual sum

The Least Squares Case

If all the models in the set assume normally distribute
stant variance, then AIC can be easily computed from leas
statistics as

$$\mathbf{AIC} = n \log(\hat{\sigma}^2) + 2K,$$

where

$$\hat{\sigma}^2 = \frac{\sum \hat{\epsilon}_i^2}{n} \text{ (the MLE of } \sigma^2),$$

and $\hat{\epsilon}_i$ are the estimated residuals for a particular candidat
mistake with LS model fitting, when computing AIC, is to
σ^2 from the computer output, instead of computing the M
**Also, for LS model fitting, K is the total number of es
parameters, including the intercept and σ^2.**

Thus, AIC is easy to compute from the results of LS e
of linear models and is now included in the output of man
for regression analysis. However, the value of K is somet
correctly because either β_0 (the intercept) or σ^2 (or both) i
in determining K.

The fact that AIC is an estimate only of relative expe
almost unimportant. It is the fact that AIC is only an estim
distances from each model g_i to f that is less than idea
recognize that there is usually substantial uncertainty as t
a given data set. After all, these are stochastic biological p
relatively high levels of uncertainty.

the basic underlying structure of the model that fits the da
speak" is of interest to both biologists and statisticians
ing from empirical data. The data then help determine th
(order or dimension) of the approximating model used fo
determine what effects or factors are justified. In this se
given data set are conditional on sample size. We must
more data were available, then further effects could pro
supported. "Truth" is elusive; model selection tells us wha
support, not what full reality might be.

Akaike (1973) multiplied the bias-corrected log-likeli
torical reasons" (e.g., it is well known that -2 times t
ratio of two maximized likelihood values is asymptoticall
certain conditions and assumptions). The term -2 occu
contexts, so it was not unreasonable that Akaike perfor
eration to get his AIC. Two points frequently arise, and
here. First, the model associated with the minimum AIC
if the bias-corrected log-likelihood (i.e., $\log(\mathcal{L}) - K$) is n
-34, or -51.3, or any other negative number. Thus, the
changed by the multiplication of both terms by any negat
merely chose -2. Second, some investigators have not
link between K-L information and AIC and believed, th
2 in the second term in AIC was somehow "arbitrary"
bers should also be considered. This error has led to con
in the technical literature; clearly, K is the asymptotic bi
not arbitrary. Akaike chose to work with $-2\log(\mathcal{L})$, rath
the term $+2K$ is theoretically correct, for large sample s
terms (the log-likelihood and the bias correction) are mu
negative constant, the model where the criterion is mini
and there is nothing arbitrary.

It might be argued that we should have merely defined
model)); then AIC $= -2l + 2K$, making the criterion lool
may have advantages, we believe that the full notation
and helps in understanding exactly what is meant. The
breviations such as $\log(\mathcal{L}(\theta|x, g_i))$, makes it explicit tha
is a function of (only) the parameters (θ), while the data
say multinomial) must be given (i.e., known). These distin
important when we introduce the concept of a likelihood
data: $\mathcal{L}(g_i|data)$. Both concepts are fundamental and use
in this book and the notation serves an important purpose

If the approximating models in the candidate set are
then Takeuchi's information criterion (TIC) is an alternat
quite large. AIC is a special case of TIC, and as such, AI
approach to the estimation of relative expected K-L distan

At one point in Akaike's derivation of an estimator of K-L i
the assumption that the model set included f (full reality
subject of attention and criticism. Akaike maintained that
was asymptotically unbiased and free from any notion th
model or that such a true model was required to be in
models. This section will indicate that such claims were ju
another insight into the concept of parsimony. The key
important, little-known paper (in Japanese) by Takeuchi (
just 3 years after Akaike's initial breakthrough in 1973.

Takeuchi (1976) provides a very general derivation of a
rion, without taking expectations with respect to g. His cr
TIC (Takeuchi's information criterion) and was thought t
where the candidate models were not particularly close a
TIC has a more general bias-adjustment term to allow
justed to be an asymptotically unbiased estimate of rel
information,

$$\mathbf{TIC} = -2\log(\mathcal{L}) + 2 \cdot \mathbf{tr}(\boldsymbol{J}(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta})^{-1}$$

The $K \times K$ matrices $J(\theta)$ and $I(\theta)$ involve first and s
derivatives of the log-likelihood function, and "tr" denc
function. One might consider always using TIC and worr
quacy of the models in the set of candidates. This consid
issues that are problematic. First, one must *always* worry
the set of approximating models being considered; this
shortcut. Second, using the expanded bias adjustment te
estimation of the elements of the matrices $J(\theta)$ and $I($
in Chapter 7). Shibata (1999) notes that estimation error
ces can cause instability of the results of model selectio
where a candidate model has $K = 20$ parameters. The
and $I(\theta)$ are of dimension 20×20, and reliable estima
of each matrix will be difficult unless sample size is ver
that $\mathrm{tr}(J(\theta)I(\theta)^{-1})$ itself has a very simple parsimonious e
This is an interesting and important general result.

Thus, AIC is an approximation to TIC, where $\mathrm{tr}(J(\theta$
approximation is excellent when the approximating mod
comes poor when the approximating model is a poor. H
that are poor, the first term, $-2\log(\mathcal{L})$, dominates the crit
is poor and this term will tend to be relatively large, cor
better model. Thus, with the final approximation that $\mathrm{tr}(J($
can see that AIC is an asymptotically unbiased estimator
K-L information, derived without assuming that full reali
or that such a model is in the set of candidate models. V
portant contribution to the literature, it has rarely seen ap

be done, we expect $\mathrm{tr}(J(\hat{\theta})I(\hat{\theta})^{-1})$ to be very close to K.

2.4 Second-Order Information Criterion:

While Akaike derived an estimator of K-L information
poorly if there are too many parameters in relation to th
(Sugiura 1978, Sakamoto et al. 1986). Sugiura (1978) de
variant of AIC that he called c-AIC.

A Small Sample AIC

Hurvich and Tsai (1989) further studied this small-san
bias adjustment, which led to a criterion that is called A

$$\mathbf{AIC}_c = -2\log(\mathcal{L}(\hat{\theta})) + 2K\left(\frac{n}{n-K-1}\right.$$

where the penalty term is multiplied by the correction fa
This can be rewritten as

$$\mathbf{AIC}_c = -2\log(\mathcal{L}(\hat{\theta})) + 2K + \frac{2K(K+}{n-K-}$$

or, equivalently,

$$\mathbf{AIC}_c = \mathbf{AIC} + \frac{2K(K+1)}{n-K-1},$$

where n is sample size (also see Sugiura 1978).

**Unless the sample size is large with respect to the nu
parameters, use of $\mathbf{AIC}_c$ is recommended.**

AIC_c merely has an additional bias-correction term. If n
to K, then the second-order correction is negligible and
well. Findley (1985) noted that the study of bias correc
itself; the exact small-sample bias-correction term varie
(e.g., normal, exponential, Poisson). Bedrick and Tsai (19
refinement, but it is more difficult to compute (also see Hu
and 1995a and b, and Hurvich et al. 1990). While AIC
Gaussian assumptions for linear models (fixed effects), B
found this second-order approximation to the K-L dista
product multinomial models. **Generally, we advocate th
the ratio n/K is small (say < 40)**. In reaching a decisi
AIC vs. AIC_c, one must use the value of K for the highe
global) model in the set of candidates. If the ratio n/K
then AIC and AIC_c are similar and will strongly tend to se
One must use either AIC or AIC_c consistently in a given

2.5 Modification of Information Criterion Overdispersed Count Data

In general, if the random variable n represents a count discrete distribution (e.g., Poisson or binomial), it has a $\mu(\theta)$, and a known theoretical variance function, $\sigma^2(\theta)$ In a model of overdispersed data the expectation of n the variance model must be generalized, for example us factor, e.g., $\gamma(\theta)\sigma^2(\theta)$. The form of the factor $\gamma(\theta)$ can b by theoretical considerations and can be complex (see, e Nelder 1989). Overdispersion factors typically are smal above 1 to perhaps 3 or 4 if the model structure is correc is due to small violations of assumptions such as indepen homogeneity over individuals. Hence, a first approximat overdispersion is to use a simple constant c in place of γ generalized to more than one c for different partitions of

Count data have been known not to conform to simp tions based on binomial or multinomial distributions Fisher 1949, Armitage 1957, and Finney 1971). There a tistical models for count data (e.g., Poisson, binomial, multinomial). In these, the sampling variance is theoreti *assumption* (e.g., for the Poisson model, $\text{var}(n) = \text{E}(n$ model, $\text{var}(\hat{p}) = p(1-p)/n$. If the sampling variance e cal (model-based) variance, the situation is called "overdi here is on a lack of independence in the data leading t "extrabinomial variation." Eberhardt (1978) provides a c issues in the biological sciences. For example, Canada ge frequently mate for life, and the pair behaves almost as than as two independent "trials." The young of some spe with the parents for a period of time, which can also caus dence of individual responses. Further reasons for overdis systems include species whose members exist in schools of such populations can be expected to have positive corr viduals within the group; such dependence causes overdi type of overdispersion stems from parameter heterogeneit having unique parameters rather than the same paramet probability) applying to all individuals.

The estimators of model parameters often remain unbi of overdispersion, but the model-based theoretical varianc cision (McCullagh and Nelder 1989). To properly cope

for the analysis of overdispersed data (also see William
and Pregibon 1985, Moore 1987, and McCullagh and N
1999a). Hurvich and Tsai (1995b) provide information on
overdispersed data.

Cox and Snell (1989) discuss modeling of count data a
useful approximation is based on a single variance inflati
can be estimated from the goodness-of-fit chi-square statis
model and its degrees of freedom,

$$\hat{c} = \chi^2/\mathrm{df}.$$

The variance inflation factor should be estimated from th
and Snell (1989) assert that the simple approach of a cor
tion factor should often be adequate, as opposed to the
task of seeking a detailed model for the $\gamma(\theta)$. In a study
approaches on five data sets, Liang and McCullagh (1993)
overdispersion was clearly better than use of a single $\hat{c}$ in c
examined.

Given $\hat{c}$, empirical estimates of sampling variances (v
ances ($\mathrm{cov}_e(\hat{\theta}_i, \hat{\theta}_j)$) can be computed by multiplying t
theoretical (model-based) variances and covariances by
has long been used; see, e.g., Finney 1971). These empiric
ation (i.e., $\hat{c} \cdot \widehat{\mathrm{var}}_t(\hat{\theta}_i)$) must be treated as having the degr
to compute $\hat{c}$ for purposes of setting confidence limits (or
**The number of parameters (*K*) must include one for
the variance inflation factor, if used**. Generally, quasi-lik
(i.e., use of $\hat{c} > 1$) are made only if some distinct lack o
(for example, if the observed significance level $P \leq 0.$
goodness-of-fit degrees of freedom ≥ 10, as rough guide

We might expect $c > 1$ with real data but would not exp
4 if model structure is acceptable and only overdispersio
Eberhardt 1978). Substantially larger values of c (say, 6–1
partly by a model structure that is inadequate; that is, th
not account for an acceptable amount of variation in the da
methods of variance inflation are most appropriate only
structural adequacy of the model has been achieved. The
be computed only for the global model; one should not m
estimates of this variance inflation factor for each of th
in the set. The issue of the structural adequacy of the m
heart of good data analysis (i.e., the reliable identificati
versus residual variation in the data). Patterns in the good
(Pearson χ^2 or G-statistics) might be an indication of stru
the model. Of course, the biology of the organism in quest

When data are overdispersed and $\hat{c} > 1$, the proper li
(not just $\log(\mathcal{L})$). Principles of quasi-likelihood suggest s
to AIC and AIC_c; we denote these modifications by (Leb

$$\text{QAIC} = -\left[2\log(\mathcal{L}(\hat{\theta}))/\hat{c}\right] + 2K,$$

and

$$\text{QAIC}_c = -\left[2\log(\mathcal{L}(\hat{\theta}))/\hat{c}\right] + 2K + \frac{2K(K}{n - K}$$

$$= \text{QAIC} + \frac{2K(K+1)}{n-K-1}.$$

If an overdispersion factor is estimated, then one parame
K. Of course, when no overdispersion exists, then $c =$
for QAIC and QAIC_c reduce to AIC and AIC_c, respectiv
(1994) found that these criteria performed well in product
of capture–recapture data in the presence of differing leve

One must be careful when using some standard softw
SAS GENMOD), since they were developed some time
esis testing mode (i.e., adjusting χ^2 test statistics by $\hat{c}$ t
some cases, a separate estimate of c is made for each m
and covariances are multiplied by this model-specific esti
inflation factor. Some software packages compute an est
model, thus making the correct use of model selection c
one is careful. Instead, we recommend that the global mod
for the estimation of a single variance inflation factor c.

Overdispersed Count Data: A Revie

Try to ensure that the structural part of the data is w
global model.

If there is biological reason to suspect overdispersion,
sion parameter c can be estimated as χ^2/df, using the gl

If overdispersion is present, the log-likelihood of the
the data and the model, should be computed as

$$\frac{\log(\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{g}_i))}{\hat{\boldsymbol{c}}}.$$

The number of parameters K is now the number of para
account for the estimation of the overdispersion paramet

The estimated overdispersion parameter should gener
Otherwise, some structural lack of fit is probably enter
overdispersion. If $\hat{c} < 1$, just use $c = 1$.

and covariances should also be adjusted using $\hat{c}$ from the
there are few degrees of freedom left.

AIC for Overdispersed Count Data

Model selection should use either

$$\mathbf{QAIC} = -[2\log(\mathcal{L}(\hat{\theta}))/\hat{c}] + 2K,$$

or

$$\mathbf{QAIC}_c = -[2\log(\mathcal{L}(\hat{\theta}))/\hat{c}] + 2K + \frac{2K(K}{n-K}$$
$$= \mathbf{QAIC} + \frac{2K(K+1)}{n-K-1}$$

The variance–covariance matrix should be multiplie overdispersion parameter $\hat{c}$ (i.e., $\hat{c}(\mathrm{cov}(\hat{\theta}_i, \hat{\theta}_j))$.

Some commercial software computes AIC, while AIC and no general software package computes QAIC or Q cases, AIC, AIC_c, QAIC, and QAIC_c can be computed the material that is output from standard computer packag or least squares estimation). In general, we recommend information-theoretic criterion for count data, and we wil of the practical examples in Chapter 3. Of course, ofte parameter is near 1, negating the need for quasi-likeliho just as often the ratio n/K is large, negating the need fo correction term in AIC_c. AIC, AIC_c, and QAIC_c are all est K-L information. We often use the generic term "AIC" t criteria.

2.6 AIC Differences, Δ_i

AIC, AIC_c, QAIC_c, and TIC are all on a relative (or in strongly dependent on sample size. Simple differences estimates of $\mathrm{E}_{\hat{\theta}}[\hat{I}(f, g_i)] - \min \mathrm{E}_{\hat{\theta}}[\hat{I}(f, g_i)]$, where the ex estimated parameters and min is over the models.

The larger Δ_i is, the less plausible it is that the fitted K-L best model, given the data x. Some rough rules of thu are particularly useful for nested models:

Δ_i	Level of Empirical Support of Mode
0-2	Substantial
4-7	Considerably less
> 10	Essentially none.

AIC differences,

$$\Delta_i = \mathrm{AIC}_i - \mathrm{AIC}_{min},$$

over all candidate models in the set. We use the term "A
a generic sense here to mean AIC, AIC_c, QAIC_c, or TIC
estimate the relative expected K-L differences between f
Δ_i values are easy to interpret and allow a quick compar
candidate models and are also useful in computing Akail
2.9). The model estimated to be best has $\Delta_i \equiv \Delta_{min} \equiv 0$

Models with $\Delta_i > 10$ have either essentially no sup
omitted from further consideration, or at least those models
substantial explainable variation in the data. These guideli
is small (even as many as 100), but may break down in exp
there may be thousands of models. The guideline value
larger for nonnested models, and more research is needed
Linhart 1988). If observations are not independent, but
independent, then these simple guidelines cannot be expe
the log-likelihood is corrected for overdispersion in coun
c, then the guidelines above will be useful.

As an example, candidate models g_1, g_2, g_3, and g_4
3,400, 3,560, 3,380, and 3,415, respectively. Then one w
as the best single model as the basis for inference becaus
AIC value. Because these values are on a relative (interv
subtract, say, 3,380 (the minimum of the 4 values) from
have the following rescaled AIC values: 20, 180, 0, and
rescaling does not change the ranks of the models, nor the
in the AIC values. People are often surprised that Δ_i of
important, when the associated AIC values that led to the
order of 97,000 or 243,000.

AIC Differences

**It is not the absolute size of the AIC value, it is the r
particularly the AIC differences (Δ_i), that are impor**

An individual AIC value, by itself, is not interpretable
constant (interval scale). AIC is only comparative, relative
in the model set; thus such differences Δ_i are very impo

We can say with considerable confidence that in rea
several or more models and large sample size (say $n > 10$
model) a model having $\Delta_i = 20$, such as model g_4, w
approximating model for the data at hand.

K-L best model. Consider Δ_i values for 7 models as 0, 1
and 7.3. An important question is, how big a difference ma
asked in the sense of when a model is not to be considered
selected best model as plausibly the actual K-L best mode
used, for the sample size and data at hand. The question l
answer; it is like asking how far away from an MLE $\hat{\theta}$ ar
θ must be (assuming that the model is a good model) b
that an alternative θ is unlikely as "truth." This question c
with a confidence (or credibility) interval on θ based on
uncertainty. A conventionally accepted answer here is that
if it is further away than $\pm 2\,\widehat{\text{se}}(\hat{\theta})$ (there is a fundamental
a procedure). Relative scaling of alternative models can
using Akaike weights (Section 2.9) and evidence ratios (S

2.7 A Useful Analogy

In some ways, selection of a best approximating model
racing or other similar contests. The goal of such a race i
(fastest) car/driver combination, and the data represent r
race (e.g., the Indianapolis 500 in the USA, the 24 Heures
Only a relatively few car/driver combinations "qualify," ba
(e.g., 33 cars at Indianapolis)—this is like the set of car
only certain models "qualify," based on the science of the
be chaotic if all car/driver combinations with an interest
just as it makes little sense to include a very large numb
set of candidates (and risk Freedman's paradox). Cars th
not win, even though they might indeed have been the be
not failed to qualify. Similarly, models, either good or b
candidates remain out of consideration.

At the end of the race the results provide a ranking
car/driver combination, from first to last. Furthermore, if
of quality is available (e.g., elapsed time for each finis
"scaling" can be considered. Clearly, the primary inter
the race" or "which was the first"; this is like the model
AIC value. This answers the question, "Which is best in t
could differ for another (future) race or another data set,
unavailable to us.

Some (secondary) interest exists in the question, "W
place?" and in particular, was second place only thousan
hind the winner or 5 minutes behind? The race time result
these questions, as do the Δ_i values in model selection. In t

first place winner.), while in the second case, the inferc
a single car/driver combination as the clear best (with a
finish). The finishing times provide insights into the third
etc. In trying to understand the performance of car/drive
has considerable information from both the rankings and t
analogous to the AIC values (both the ranks and the Δ_i
2.9 and 2.10 will see how the Δ_i can be used to estimat
and these will provide additional insights. Note that the
winner is of little interest because of temperature differenc
and other variables; only the *relative* times for a given ra
terest. Similarly, the absolute values of AIC are also of li
they reflect sample size and some constants, among oth
of the maximized log-likelihood (i.e., $\log(\mathcal{L}(\hat{\theta}|x))$) varie
sample to sample. However, all comparisons of models ar
data, so this sample-to-sample variation is irrelevant. Co
log-likelihood values across data sets is like comparing
when some races are 500 miles whereas others are 400 o

The winner of the race is clearly the best for the particul
to make a broader inference concerning races for an enti
(i.e., ranks) from several races can be pooled or weighted.
inferences beyond a single observed data set can sometin
some type of model averaging using, for example, the
strap (details in Chapters 4 and 5) and the incorporation
uncertainty in estimators of precision.

The race result might not always select the best car/
because the fastest qualifying car/driver may have had ba
engine failure) and finished well back from the leader (if
model selection one has only one realization of the stoch
estimated relative distance as the basis for the selection of a
model (a winner). If the same race is held again with th
winner and order of finishers are likely to change some
new sample of data could be obtained, the model ranks
somewhat.

To carry the analogy a bit further, data dredging wou
watching a race as cars dropped out and others came to
continually shifts the bet and predicted winner, based on
lead at any point in time (i.e., an unfair advantage). In
prediction would surely be improved, but the rules of play
altered! Alternatively, the definition of winning might not
to the initiation of the race. Only after the race are the
based, in part, on who they think "ought" to win). Then, on
applicability of this specific prediction to other races. Ind
"new rules" when data dredging has been done. That is,

all the information and insights they might provide. How
leading to data dredging should be revealed, and results
discussed in this light.

Many realize that there is considerable variation in cars a
to race and track to track. Similarly, many are comfortab
there is often considerable sampling variation (uncertainty
estimate of a parameter from data set to data set. Simila
(races) could be taken, the estimated best model (car/dri
from sample to sample (or race to race). Both components
and model selection uncertainty should ideally be incorpo
of precision.

2.8 Likelihood of a Model, $\mathcal{L}(g_i|data)$

While the AIC differences Δ_i are useful in ranking the m
to quantify the plausibility of each model as being the act
This can be done by extending the concept of the likeliho
given both the data and model, i.e., $\mathcal{L}(\theta|x, g_i)$, to the conc
of the model given the data, hence $\mathcal{L}(g_i|x)$. Such quantiti
making inferences concerning the relative strength of evi
models in the set.

Likelihood of a Model, Given Data

The likelihood of model g_i, given the data, is simple t
model in the set:

$$\mathcal{L}(g_i|x) \propto \exp\left(-\frac{1}{2}\Delta_i\right),$$

where "$\propto$" means "is proportional to." Such likelihoods r
strength of evidence for each model.

Akaike (see, e.g., Akaike 1983b) advocates the above
relative likelihood of the model, given the MLEs of mod
on the same data. Such quantities can also be expressed a

$$C\mathcal{L}(\hat{\theta}|x, g_i)e^{-K},$$

where C is an arbitrary constant.

Model Probabilities

To better interpret the relative likelihood of a model, the set of R models, we normalize the $\mathcal{L}(g_i|x)$ to be a set weights," w_i, adding to 1:

$$w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^{R} \exp(-\frac{1}{2}\Delta_r)}.$$

The w_i depend on the entire set; therefore, if a model i during a post hoc analysis, the w_i must be recomputed f the newly defined set.

This idea of the likelihood of the model given the da model weights, has been suggested for many years by A 1978b, 1979, 1980, 1981b and 1983b; also see Bozdoga et al. 1991) and has been researched some by Buckland model weights seemed not to have a name, so we call th This name will herein apply also when we use AIC_c, QAI **A given w_i is considered as the weight of evidence in fav the actual K-L best model for the situation at hand *giv* models must be the K-L best model of that set of R m** that there are only R models and one of them must be best it is convenient to normalize the relative likelihoods to su

For the estimated K-L best model (let this be model g_{mi} for that model $\exp(-\frac{1}{2}\Delta_{min}) \equiv 1$. The odds for the i^{th} m the K-L best model are thus $\exp(-\frac{1}{2}\Delta_i)$ to 1, or just the " is convenient to reexpress such odds as the set of Akaike a Δ_i is, the smaller the w_i, and the less plausible is model K-L best model for f based on the design and sample siz weights provide an effective way to scale and interpret t weights also have other important uses and interpretations following chapters.

In general, likelihood provides a good measure of d evidence about parameter values, given a model and da 1997). We think that this concept extends to evidence model, given a set of models. That is, evidence for the represented by the likelihood of a model.

of these models might be the K-L best model for the c
compelled by a certain aspect of information theory itse
Jessop 1995). Let τ_i be the prior probability that model i is
Lacking any prior information, we set the τ_i all equal, and
In fact, doing so places all R of the models on an equal fo
as the K-L best model.

If there is prior information or belief, this opens the d
probabilities. Ignoring any model redundancy (this subjec
tion 4.6), τ_i is our prior state of information or belief tha
the data, provides the K-L best model for the design an
is a deceptively complex issue, as it relates both to idea
approximations to truth and to expected model fitting tra
sampling variances.

To us it seems impossible to have any real prior basi
differential assessment of the τ_i (other than on how th
structurally interrelated or partially redundant). Using th
principle of Jaynes (1957) we should take the τ_i to repres
tainty about all unknown aspects of the probability distrib
the τ_i. Thus we determine the τ_i that maximize the entropy
ject to constraints that express whatever information (in t
we have about the distribution. In the "no information" cas
we have is that $\sum \tau_i = 1$ (plus the essential $0 < \tau_i < 1$)
tropy (hence maximum uncertainty) prior is then $\tau_i \equiv 1/R$
field to delve into the aspects of information theory under
entropy principle. This principle is fundamentally tied t
entropy and to information theory and can be used to jus
Bayesian priors—when they exist. The interested reader
and Kesavan 1992, or the less technical Jessop 1995.]

Given any set of prior probabilities (the τ_i), generalized
given by

$$w_i = \frac{\mathcal{L}(g_i|\underline{x})\tau_i}{\sum_{r=1}^{R} \mathcal{L}(g_r|\underline{x})\tau_r}.$$

There may be occasions to use unequal prior probabilities,
above. However, in general, by Akaike weights we mean th
without the τ_i (this assumes $\tau_i = 1/R$).

The inclusion of prior probabilities (τ_i) in the w_i is
approach. The full Bayesian approach to model selection re
τ_i on the model and a prior probability distribution on the pa
g_i for each model. Then the derivation of posterior results
(usually achievable only by Markov chain Monte Carlo

and Raftery (1995), Hoeting and Ibrahim (1996), Raftery
Morgan (2000).

A brief comparison is given here of what we mean by
ities τ_i under this information-theoretic approach to mo
what seems to be meant by the prior probabilities of mo
approach. The Bayesian approach seems generally to ass
models, in the set of R models, is true. Hence, τ_i is ther
belief that model form g_i is the true model form (see, e
Under the information-theoretic approach we do not assu
the set of models, and $\tau_1, \ldots, \tau_R$ is a probability distribu
formation (or lack thereof) about which of the R models is
for the data. Information theory itself (Kapur and Kesavan
determination of the τ_i, generally as $\tau_i \equiv 1/R$. For data
that the issue cannot be which model structure is truth,
models considered is truth. Rather, the issue is, which m
data (i.e., when θ is *estimated*) is the best model for purp
the (finite) information in the data. Letting $\tau_i = \text{Prob}\{\text{be}$
g_i is the K-L best model}, then τ_i is about the "parameter
random variable g_{min}. Here, we use only $\tau_i = 1/R$.

2.10 Evidence Ratios

Using the hypothetical example in Section 2.6, the likelih
given the data, and the Akaike weights are given below:

Model	Δ_i	$\mathcal{L}(g_i\|x)$	Akaike weight u
1	0	1	0.431
2	1.2	0.54881	0.237
3	1.9	0.38674	0.167
4	3.5	0.17377	0.075
5	4.1	0.12873	0.056
6	5.8	0.05502	0.024
7	7.3	0.02599	0.010.

As weight of evidence for each model we can see that the
is not convincingly best; the evidence ratio for model
is only about 2 (i.e., $w_1/w_2 = 1.82$). This relatively w
best model suggests that we should expect to see a lo
selected best model from sample to sample if we cou
draw multiple independent samples; that is, the model s
is likely to be high. The evidence ratio for the best mod

$$\mathcal{L}(g_i|x)/\mathcal{L}(g_j|x)$$

or, equivalently, the ratio of Akaike weights w_i/w_j. Su
monly used, and we will term them **evidence ratios**. Su
the evidence about fitted models as to which is better in
sense.

In particular, there is often interest in the ratio w_1/w_j, v
estimated best model and j indexes the rest of the mode
ratios are not affected by any other model, hence do not
set of R models—just on models i and j. These evidence
to all other models besides i and j.

$0.431/0.024 = e^{(5.8/2)} = 18$, and we must conclude th
model 6 is the K-L best model; the evidence here is reaso
model 6.

There is a striking nonlinearity in the evidence ratios as
values. Consider the ratio $w_1/w_j(\equiv w_{min}/w_j)$,

$$\frac{w_1}{w_j} \equiv \frac{1}{e^{-1/2\Delta_j}} \equiv e^{1/2\Delta_j}$$

in the comparison of the evidence for the best model versu
Then, we have the following table:

Δ_j	Evidence ratio
2	2.7
4	7.4
8	54.6
10	148.4
15	1,808.0
20	22,026.5

This information helps to justify the rough rules of thum
the evidence for models being the best K-L model in the
provided some likelihood-based rules similar to these ov
Edwards (1992) and Royall (1997) for additional perspec
of evidence in a likelihood framework.

People may, at first, be frustrated that they do not have
point that provides a simple dichotomy to indicate wha
"significant" under the Neyman–Pearson null hypothes
where a decision is to be reached). Even knowing that st
is not particularly related to biological significance, an
arbitrary, some investigators seem to feel comfortable b

The approach we advocate is one of quantitative evidenc
interpret the quantitative evidence.

Consider a football game where the final score is 10 to
B, respectively. Here, one does not ask whether the win
A was "significant." Rather, one can see that the game
the score (the evidence). Further scrutiny of the evidenc
examining the total yards gained, the cumulative time
ball, the number of penalties, etc., for each team. Base
the evidence, one can reach a determination concerning
of the two teams. Furthermore, in this case, most ration
roughly the same determination, based on the evidence. S
had been 40 to 3 (the evidence), it would be clear that te
hapless opponent. Even in this case there is no concept of
much less any test of the null hypothesis based on the
the teams were of equal ability. Again, most rational pe
agree that team A was the better team on the day of the
evidence (40 vs. 3). Based on the evidence, people migh
an inference to other games between these two teams.
intermediate cases (10 vs. 16) where the evidence is not
the final touchdown occurred in overtime, in which case
interpret the evidence (10 to 16) differently. Again, a re
statistics might provide insights, but we should admit that
lead to a clear determination, accepted by all. One encoun
numerical evidence in everyday life and can interpret su
arbitrary dichotomies.

When we learn that model g_4 has an evidence ratio of 3
g_2, it means there is relatively little evidence in favor of m
here is an auditorium containing N people (let N be lar
Each person has a raffle ticket, except that a single person
The evidence ratio (relative likelihood) of Bob winning th
individual is 3. Clearly, Bob has an edge over any other
not strong. Of course, the probability that either Bob or
individual will win is small if N is large. However, the r
same, regardless of the value of N. In contrast, let Bob n
Then his relative likelihood of winning vs. any other indiv
is relatively strong evidence. Such evidence ratios are on
vs. another individual); nothing is to be inferred about B
other individual's chances) of winning the raffle outright.
relative to another individual's chances are quantified us
Finally, note that the probability of Bob winning, *given*
another single individual wins, is $100/(100 + 1) = 0.99$
model pairs (e.g., model g_4 vs. model g_2) are relative valu

Data analysis involves the proper tradeoff between bias a
ilarly, between underfitting and overfitting. The estimati
information is a natural and simple way to view model sel
set of candidate models, select that fitted model where info
mized. Proper model selection is reflected in good achieve
coverage for the parameters in the model (or for predicti
haps too much bias has been accepted in the tradeoff to g
a false sense of high precision. This represents the worst
a highly precise, but quite biased estimate. These ideas ha
in statistical thinking.

An information criterion (i.e., AIC, AIC$_c$, QAIC, and
rank the candidate models from best to worst and scale the
weights and evidence ratios. Often data do not support
clearly best for data analysis. Instead, suppose three mode
for best, while another, larger, set of models is clearly no
underfit or overfit). Such virtual "ties" for the best approx
be carefully considered and admitted. Poskitt and Trema
"portfolio of models" that deserve final consideration. Ch
that there may be more than one model that is to be regar

Ambivalence

The inability to ferret out a single best model is not a d
other selection criterion. Rather, it is an indication that
inadequate to reach such a strong inference. That is, the
concerning some effect or parametrization or structure.

In such cases, all the models in the set can be us
inferences: multimodel inference.

It is perfectly reasonable that several models would serv
in approximating the information in a set of data. Infere
there are sometimes competing models and the data do n
only one. The issue of competing models is especially r
model selection uncertainty into estimators of precision.
model has substantial support, some form of multimodel i
averaging) should be considered (Chapter 4). The follow
vide some important details that must be considered in a
research data.

2.11.1 *AIC Cannot Be Used to Compare Models*
Data Sets

Models can be compared using the various information cri
relative, expected K-L information, only when they have b
the same set of data. For example, if nonlinear regression m

AIC cannot be used to compare models where the data a
case (Model U) and grouped (e.g., grouped into histogram
(Model G).

Data Must Be Fixed

An important issue, in general, is that the data and their
must be fixed and alternative models fitted to this fixed d

Information criteria should not be compared across
because the inference is conditional on the data in hand.

2.11.2 Order Not Important in Computing AIC

The order in which the information criterion is computed o
is not relevant. Often, one may want to compute AIC_c, sta
model and proceed to simpler models with fewer paramete
to start with the simple models and work up to the more
many parameters; this strategy might be best if numeric
countered in fitting some high-dimensioned models. Th
here to proper interpretation, as opposed to the various hy
proaches where the order may be both arbitrary and the res
on the choice of order (e.g., stepup (forward) vs. stepdown
Section 3.4.6 provides an example).

2.11.3 Transformations of the Response Variabl

Model selection methods assume that some response varia
ject of interest. Assuming that the scientific hypotheses re
variable, then all the models must represent exactly this v
models in the set should all have the same response varia
of mistake is illustrated by the following example. An inve
in modeling a response variable y and has built 4 linear r
y, but during the model building, he decides to include a
that point he includes a model for $\log(y)$ as the fifth mod
information in such cases cannot be validly compared.
point, and often overlooked. In this example, one would f
model followed by the other 4 models, each having larg
on this result, one would erroneously conclude the impor
earity. **Investigators should be sure that all hypotheses
the same response variable (e.g., if the whole set of mo
log(y), no problem would be created; it is the mixing of
that is incorrect).**

$$g_1(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\frac{[y-\mu]^2}{\sigma^2}\right.$$

and another model,

$$g_2(y|\mu,\sigma) = \frac{1}{y\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\frac{[\log(y)-\right.$$

Another critical matter here is that all the components of ea be retained in comparing different probability distributi comparisons of different pdfs in this spirit in Section 6.7. requirement is not needed in cases like multiple regressio ance because all the comparisons are about the model stru to select) with an assumption of normal errors for every there is a global model and its associated likelihood, and to represent μ as a regression function.

In other cases, it is tempting to drop constants in the log they do not involve the model parameters. However, alte not have the same constants; this condition makes valid impossible. The simple solution here is to retain all th likelihood for all the models in the set.

2.11.4 Regression Models with Differing Error S

This issue is related to that in Section 2.11.3. A link betw of squares (RSS) and σ^2 from regression models with errors to the maximized log-likelihood value was provid This link is a special case, allowing one to work in an o regression framework for modeling and parameter estima to a likelihood framework to compute $\log(\mathcal{L}(\theta|data,\ mode$ quantities under an information-theoretic paradigm.

The mapping from $\hat{\sigma}^2$ to $\log(\mathcal{L}(\theta|data,\ model))$ is models in the set assume independent, normally distri uals) with a constant variance. If some subset of the lognormal errors, then valid comparisons across all the not possible. In this case, all the models, including those structures, should be put into a likelihood framework sin estimates of $\log(\mathcal{L}(\theta|data,\ model))$ and criteria such as A

Tests of null hypotheses and information-theoretic appro
used together; they are very different analysis paradign
mistake seen in the applied literature is to use AIC to rank t
and then "test" to see whether the best model (the alter
"significantly better" than the second-best model (the nu
procedure is flawed, and we strongly recommend agains
2001*c*). Despite warnings about the misuse of hypothesis t
et al. 2000, Cox and Reid 2000), researchers are still rep
trivial null hypotheses, while failing to report effect size a

Some authors state that the best model (say g_3) is *sig*
another model (say g_6) based on a Δ value of 4–7. Alte
one sees that model g_6 is *rejected* relative to the best mod
are poor and misleading. It seems best not to associate t
or *rejected* with results under an information-theoretic p
concerning the strength of evidence for the models in the s
using the evidence ratio (Section 2.10), as well as an a
adjusted R^2, and other model diagnostics or descriptive s

2.11.6 Null Hypothesis Testing Is Still Importan Strict Experiments

A priori hypothesis testing plays an important role when
(i.e., treatment and control groups being formally contra
design with random assignment) has been done and specif
hypotheses have been identified. In these cases, there is
statistical theory on testing of treatment effects in such
We certainly acknowledge the value of traditional testin
analysis of these *experimental* data. Still, the primary em
the size of the treatment effects and their precision; too ofte
regarding "significance," while the treatment and control
presented (Anderson et al. 2000 Cox and Reid 2000). Ne
are calling for estimates of effect size and associated preci
statistics, P-values, and "significance."

Akaike (1981) suggests that the "multiple comparison"
means should be viewed as a model selection problem, r
to one of the many testing methods that have been develo
1988). Here, a priori considerations would be brought t
and a set of candidate models derived, letting informat
aid in sorting out differences in treatment means—a refo
estimation, instead of on testing. An alternative approach i
effects modeling (Kreft and deLeeuw 1998).

model selection and associated parameter estimation. Th
the case where nuisance parameters are encountered in th
recapture or resighting probabilities in capture–recaptur
studies. Here, it is not always clear what either the nu
hypothesis should be in a hypothesis testing framework. I
potheses that are tested are naive or trivial, as Johnson
out with such clarity. Should we expend resources to fi
white? Is there any reason to test formally hypotheses such
of robins is the same in cities A and B"? Of course not!
assume that the number is different and proceed to estima
the difference and its precision: an estimation problem, n
testing problem.

2.11.7 *Information-Theoretic Criteria Are Not a*

The theories underlying the information-theoretic ap
hypothesis testing are fundamentally quite different.

Criteria Are Not a Test

Information-theoretic criteria such as AIC, AIC_c, an
"test" in any sense, and there are no associated concepts s
P-values or α-levels. Statistical hypothesis testing represe
and generally inferior, paradigm for the analysis of data i

**It seems best to avoid use of the word "signific
research results under an information-theoretic para**

The results of model selection under the two approache
similar with simple problems; however, in more complex s
candidate models, the results of the two approaches can b
Section 3.5). **It is critical to bear in mind that there is
to information-theoretic approaches to model selection
use of null hypothesis testing for model selection mus
hoc** (albeit a very refined set of ad hoc procedures in som

2.11.8 *Exploratory Data Analysis*

Hypothesis testing is commonly used in the early phases
analysis to iteratively seek model structure and understand
start with 3–8 models, compute various test statistics for
several of the better models each have a gender effect. Thu
are generated to include a gender effect, and more null
conducted. Then the analyst notes that several of these mo
time for some set of estimable parameters; thus more mo
are generated, and so on. While this iterative or sequent

We suggest that if the results are treated only as alternat
more confirmatory study to be conducted later, this mig
practice, particularly if other information is incorporate
stage. Still, the sequential and arbitrary nature of such testi
us wonder whether this is really a good exploratory tech
readily keys in on unique features of the sample data at ha
In any event, the key here is to conduct further investiga
on the "hunches" from the tentative exploratory work. Co
investigation has too often been ignored and the tentative '
published as if they were a priori results. Often, the auth
the post hoc activities that led to the supposed results.

We suggest that information-theoretic approaches mig
exploratory tool; at least key assumptions upon which the
are not terribly violated, and there is no arbitrary α lev
analysis using an information-theoretic criterion, instead
statistic, eliminates inferential problems in interpreting
but one must still worry about overfitting and spurious e
al. 2001b). The ranking of alternative models (the Δ_i and
useful in the preliminary examination of data resulting fro
on these insights, one could design a more confirmatory
issue of interest. The results of the pilot exploration should
While we do not condone the use of information theoretic a
data dredging, we suggest that it might be a more useful
testing in exploratory data analysis where little a priori kn
Data dredging has enough problems and risks without u
approach that carries its own set of substantial problems a

2.12 Some History and Further Insights

Akaike (1973) considered AIC and its information theoret
natural extension of the classical maximum likelihood prin
Fisher (1936) anticipated such an advance over 60 years a

> ... an even wider type of inductive argument may s
> veloped, which shall discuss methods of assigning f
> functional form of the population.

This comment was quite insightful; of course, we migh
A. Fisher! Akaike was perhaps kind to consider AIC an e
ML theory; he might just as well have said that classical li
a special application of the more general information theo
believed in the importance of information theory as a u
statistics.

that the negative of K- L information is Boltzmann's en
information has been called negative entropy or "negen
"disorder," while max entropy is maximum disorder or mi
Conceptually,

$$\text{Boltzmann's entropy} = -\log\left(\frac{f(x)}{g(x)}\right)$$

Then,

$$-\text{Boltzmann's entropy} = \log\left(\frac{f(x)}{g(x)}\right)$$

and

$$\begin{aligned}\text{K-L} &= \mathrm{E}_f(-\text{Boltzmann's entropy}) \\ &= \mathrm{E}_f\left(\log\left(\frac{f(x)}{g(x)}\right)\right), \\ &= \int f(x)\log\left(\frac{f(x)}{g(x)}\right)\,dx.\end{aligned}$$

Thus, minimizing the K-L distance is equivalent to m
tropy; hence the name *maximum entropy principle* (see J
1983a, 1985 and Bozdogan 1987, Jessop 1995 for further
However, maximizing entropy is subject to a constraint
information in the data. A good model contains the info
leaving only "noise." It is the noise (entropy or uncertaint
under the concept of the entropy maximization principle (
imizing K-L information then results in an approximatin
minimum amount of information in the data. Entropy ma
a model that maximizes the uncertainty, leaving only info
"maximally" justified by the data. The concepts are equiva
K-L distance (or information loss) certainly seems the m

The K-L information is *averaged* negative entropy, he
with respect to f. While the theory of entropy is a lar
readers here can think of entropy as nearly synonymous
randomness or disorder in physical systems.

Boltzmann derived the fundamental theorem that

entropy is proportional to − log(probab

Entropy, information, and probability are thus linked, allo
be multiplicative while information and entropies are a
was also derived by Shannon 1948). Fritz Hasenöhrl,
mann, Boltzmann's successor at Vienna University, and

mation concerning Boltzmann appears in Brush (1965, 19
insights into Akaike's career are found in Findley and Pa

2.12.2 A Heuristic Interpretation

After Akaike's innovative derivation of AIC, people noti
pretation that was both interesting and sometimes misle
in AIC,

$$\text{AIC} = -2\log(\mathcal{L}(\hat{\theta}|x)) + 2K,$$

is a measure of lack of model fit, while the second term
preted as a "penalty" for increasing the size of the model (
parsimony in the number of parameters). This heuristic e
do justice to the much deeper theoretical basis for AIC (i.
distance and information theory). The heuristic interpreta
ticians to consider "alternative" penalty terms, and this
productive (see Chapter 6). The so-called penalty term in
rather, it is the asymptotic bias-correction term. It is the
asymptotic estimator of relative, expected K-L informati
that had Akaike defined AIC $= -\log(\mathcal{L}(\hat{\theta}|x)) + K$, the mi
unchanged; some authors use this expression, but we wil
defined it.]

The heuristic view of the components of AIC clearly s
ance tradeoff and insight into how the principle of parsim
AIC (see Gooijer et al. 1985:316). Still, we recommend
estimate of the relative expected K-L information or dist
pairs (i.e., each g_i vs. f). Minimizing this relative, exp
vides an estimated best approximating model for that pa
the *closest* approximating model to f). The relative K-L
between information theory and the log-likelihood funct
element in AIC model selection.

2.12.3 More on Interpreting Information-Theore

Estimates of relative K-L information, the AIC difference
weights (w_i) provide a ranking of the models; thus the a
which fitted model is best, which are essentially tied for be
are clearly in an inferior class (and perhaps some that ar
class). These ranks are, of course, estimates based on the da
are quite useful (cf. Section 2.7 and Sakamoto et al. 1986
primary inference be developed using the model for whic
or the small number of models where there is an essential

model is > 9–10 AIC units from the minimum) there is l
uncertainty and the theoretical standard errors can be usec
in Sections 1.2.3 and 2.14). When the results of model sel
then methods described in Chapter 4 can be considered. A
of models and the identification of models that are nearly e
those that are clearly poor explanations for the data at ha
Hypothesis testing provides no general way to rank mod
that are nested.

One must keep in mind that there is often considerab
selection of a particular model as the "best" approximat
served data are conceptualized as random variables; th
different if another, independent set were available. It is
ability" that results in uncertain statistical inference fron
set being analyzed. While we would like to make infer
robust to other (hypothetical) data sets, our ability to
limited, even with procedures such as AIC, with its cros
ties, and with independent and identically distributed s
computer-intensive resampling methods may well furth
sessment of the uncertainty of our inferences, but it re
understand that proper model selection is accompanied by
of uncertainty. The bootstrap technique can allow insigh
tainty; this and other similar issues are the subject of so
chapters.

2.12.4 Nonnested Models

A substantial advantage in using information-theoretic cr
valid for nonnested models (e.g., Table 2.2). Of course, t
ratio tests are defined only for nested models, and this
substantial limitation in the use of hypothesis testing in r
ranking of models using AIC helps clarify the importance
1973:173); for example, some models for a particular data
and should not be used for inference.

A well-thought-out global model (where applicable) is
substantial prior knowledge is required during the enti
ment, including the clear statement of the question to b
collection of the data. This prior knowledge is then car
development of the set of candidate models (Section 1.2.4
ground science, the entire investigation should probably
very preliminary.

models, with some work being done in log-linear and c
(e.g., factor analysis) models. Bozdogan (1987) provides a
and some extensions. However, the number of published
examine the performance of AIC-selected models is quite
problem with the statistical literature as regards the evalua
the use of Monte Carlo methods using only very simple ge
a few large effects and no smaller, tapering effects. Furth
Carlo studies usually have a poor objective, namely, to e
criterion selects the simple generating model. We believe
point entirely with respect to real data analysis. Such evalu
even without regard for sample size (and often use AIC wh
been used).

In Monte Carlo studies it would be useful to generat
more realistic model with several big effects and a series
effects (Speed and Yu 1993). Then interest is refocused
a good approximating model and its statistical propertie
to select the simple, artificial model used to generate the
to select a best approximating model for the data at hand
the "true model" is at all complex, its use, with estimate
than true ones, would be poor for inference, even if it exist
form (but not parameter values) were known (e.g., Sakam
counterintuitive result occurs because the (limited) data w
to estimate all the unknown parameters in the "true model,"
result in a substantial loss of precision (see Figure 1.3B).

AIC reformulates the problem explicitly as a proble
of the true structure (probably infinite-dimensional, at le
sciences) by a *model*. Model selection then becomes a si
mization, where AIC (or more properly K-L information
to be minimized. AIC selection is objective and represe
paradigm to that of null hypothesis testing and is free f
levels, the multiple-testing problem, and the fact that son
might not be nested. The problem of what model to use
hypothesis testing problem (Akaike 1974). However, the
a simple comparison of models does not justify the compa
models (Akaike 1985 and Section 1.3.3). If one had 10
would be 1,024 possible models, even if interactions ar
terms are excluded. If sample size is $n \leq 1{,}000$, overfittin
certainty. It is simply not sensible to consider such a larg
because a model that overfits the data will almost surely re
of the problem has been lost. *Even in a very exploratory a
practice to consider all possible models; surely, some sci
to bear on such an unthinking approach* (otherwise, the sc
and the work could be done by a technician).

The bootstrap is a type of Monte Carlo method used freque
tics. This computer-intensive approach is based on resamp
data (Efron and Tibshirani 1993, Mooney and Duval 1993
first described by Bradley Efron (1979); thousands of pa
ten on the bootstrap, with various extensions and applica
decades, and it has found very wide use in applied prob
can be used for several purposes, particularly in the robus
pling variances or standard errors and (asymmetrical) co
has been used in the estimation of model selection frec
estimates of precision that include model selection uncer

The bootstrap has enormous potential for the biologis
skills; however, its computer intensive nature will continu
large problems. We believe that at least 1,000 bootstrap sa
many applications, and often 10,000 samples are needed
model selection. In extreme cases, reliable results could ta
time to apply the bootstrap to complex data analysis c
sample size and several dozen models, where the MLEs
be found numerically.

The fundamental idea of the model-based sampling the
tistical inference is that the data arise as a sample fro
probability distribution f. Uncertainties of our inferenc
if we can estimate f. The bootstrap method allows the c
sures of our inference uncertainty by having a simple e
f and sampling from this estimated distribution. In prac
empirical bootstrap means using some form of resampli
from the actual data x to generate B (e.g., $B = 1{,}000$ c
samples; a bootstrap sample is denoted as x_b, where (b
sample data consist of n independent units, and it then suf
random sample of size n, *with replacement*, from the n un
bootstrap sample. However, the nature of the correct boots
can be more complex for more complex data structures.

The set of B bootstrap samples is a proxy for a set of
samples from f (in reality we have only one actual sampl
expected from replicate real samples are inferred from th
by analyzing each bootstrap sample exactly as we first a
sample. From the set of results of sample size B we me
uncertainties from sample to (conceptual) population (Fi
applications it has been theoretically shown (e.g., Efron a
and Tibshirani 1993) that the bootstrap can work well fo
(n), but it is not generally reliable for small n (say 5, 10,
regardless of how many bootstrap samples B are used.
always successful in model selection (see Freedman et al

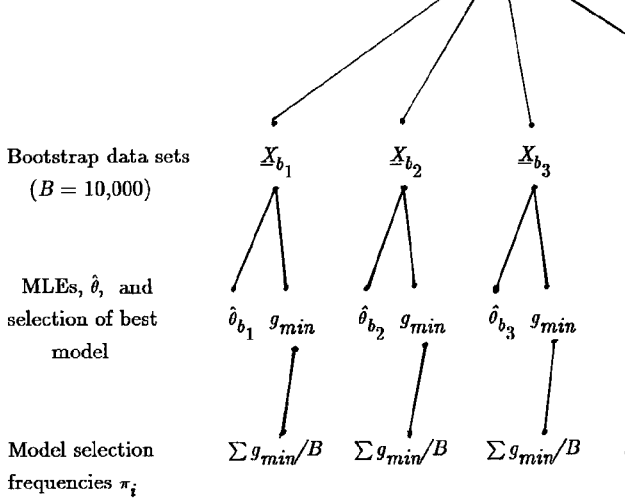


FIGURE 2.6. Diagram of the nonparametric bootstrap method as
(redrawn from Efron and Tibshirani 1993). The actual data set $\underline{X}$ i
ment, using the same sample size (n); this is done B times, to obta
$\underline{X}_b$. Maximum likelihood theory provides estimates of the parame
models i ($i = 1, 2, \ldots, R$) and the AIC-best model (denoted by m
its index stored for each of the bootstrap data sets. Finally, the r
frequencies (π_i) are computed as the sums of the frequencies whe
as best, divided by B. Of course, $\sum \pi_i = 1$.

2.13.1 Introduction

In many cases one can derive the sampling variance of an
eral likelihood theory. In other cases, an estimator may be
may not exist in closed form. For example, the finite rate c
(λ) can be derived from a Leslie population projection
of age-specific fecundity and age-specific, conditional su
Generally, λ cannot be expressed in closed form. The bo
variance estimation in such nonstandard cases.

Consider a sample of weights of 27 young rats ($n = 27$
Manly 1992),

57 60 52 49 56 46 51 63 49 57 59 54 56 59 57 52 52 61 59 53

The sample mean of these data is 54.7, and the standard d
cv $= 0.0824$. For illustration, we will estimate of the star
Clearly, this would be nonstandard; however, it represent
the bootstrap.

perhaps it would not appear in the subsample, or it coul
once. Similarly, there are 3 occurrences of the weight 57 i
perhaps the bootstrap sample would have, by chance, n
point here is that a random sample of size 27 is taken *wi*
the original 27 data values. This is the first bootstrap resa
this bootstrap sample, one computes $\hat{\mu} = \bar{x}$, the $\widehat{\text{se}}(\hat{\mu})$
$\text{cv} = \widehat{\text{se}}(\hat{\mu})/\hat{\mu}$, and stores that value of cv in memory.

Second, the whole process is repeated B times (whe
10,000 samples for this example). Thus, we generate 1
sets ($\text{b} = 1, 2, 3, \ldots, 10{,}000$) and from *each* of these w
and the cv and store the value of the cv.

Third, we obtain the estimated standard error of the
original sample by taking the standard deviation of the 1
responding to the 10,000 bootstrap samples). The proce
case, the standard error of the cv is 0.00922, or less than

Confidence intervals can be computed in the usual way
gives a 95% interval of (0.0640, 0.1009) for the rat data. Ho
distribution may be nonnormal and a more robust interva
Again, the bootstrap provides a simple approach. In this
$B = 10{,}000$ estimates of the cv in ascending order and se
cut off the lower and upper 2.5 percentiles. Thus, the resu
be asymmetric.

In the rat cv, the percentile bootstrap 95% confidence
0.0984). This interval is about the same width as in the trad
shifted a bit toward 0. Incidentally, the mean of the 10,00
was 0.0806 (compared to the actual sample cv of 0.0824).
usually adequate for the estimation of the sampling varian
ation; however, good estimates of percentile confidence i
$B = 10{,}000$ in complicated applications.

Just as the analysis of a single data set can have many
strap can be used to provide insight into a host of questio
each bootstrap sample one could compute and store the c
covariance matrix, goodness-of-fit values, the estimate
factor, the model selected, confidence interval width, a
Inference can be made concerning these quantities, based
the B bootstrap samples.

The illustration of the bootstrap on the rat data is cal
bootstrap, since no parametric distribution is assumed for
cess that generated the data. We assume only that the data i
were "representative" and that sample size was not small. T
strap is frequently used and allows assessment of bias and
of the parametric bootstrap will be illustrated by the estim
inflation factor $\hat{c}$.

$\chi^2_{gof}/\mathrm{df} = 3.2$. The investigators suspected some extrabi
are surprised by the large estimate of the variance infl
suspect that the estimate is high and decide to use a param
vestigate their suspicion. They realize that the program R
et al. 1987) can be used to do Monte Carlo simulations a
the goodness-of-fit statistics.

They input the MLEs from the real data into RELEA
parameters (ϕ_j and p_j) and use the numbers of new relea
as input. Then the amount of extrabinomial variation (i.e.
called EBV in RELEASE) is specified. In this illustrat
meaning no overdispersion. They then run 1,000 Monte
obtain the information on the estimated variance inflation
The average of these 1,000 values gives $\hat{\mathrm{E}}(\hat{c})$, and this can b
value used to generate the data. This result provides insigh
on what to do about possible overdispersion in their data.
investigators could conduct several such studies for a ra
whether $\mathrm{E}(\hat{c}|\mathrm{EBV}) = \mathrm{EBV}$ and assess any systematic bia
of EBV.

This bootstrap is parametric in that parameters were sp
from the MLEs from real data that were available) and
model to produce Monte Carlo data. The nonparametric
require parameters nor a model and relies on resampling

The bootstrap has been used in population biology to set
on the median and mean life span. It is conceptually simpl
widespread use in applied statistics. Biologists planning a
teaching should be familiar with the bootstrap. There is a
on the bootstrap; see Efron and Tibshirani (1993) for an
subject and a large list of references. Some valid applicat
are tricky (even multiple linear regression), so some care
complex settings!

2.13.2 The Bootstrap in Model Selection: The B

Consider the case where data (x) with sample size n are
models are under consideration, each representing some
of interest. Let $B = 10{,}000$ bootstrap data sets, each of si
resampling the data with replacement. MLEs of the param
could be computed for each bootstrap sample. Then AIC_c
for each of the 6 ($i = 1, 2, \ldots, 6$) models and the numb
(denote this by r^*, where r^* is the number of the best of t
associated AIC_c value stored for each of the 10,000 boot
10,000 such analyses, one has the bootstrap frequency

given by π_i = frequency/10,000 in this example. Of cou
or TIC could have been used to estimate the π_i.

Relative frequencies for model i being selected as the b
to the Akalke weights, but are not identical. There is no rea
data-based weights of evidence (as the set of w_i) to be the s
relative frequencies at which the models are selected by an
as being best. In general, likelihood provides a better me
weight of evidence about parameter values, given a mode
Royall 1997), and we think that this concept (i.e., evidenc
is best represented by the likelihood of a model) rightly
about a best model given an a priori set of models.

In our work we have not seen any particular advantage in
tion frequencies over the Akaike weights. Considering th
computer times required for the computation of the mod
cies, we prefer the Akaike weights in general. We presen
in Chapters 4 and 5.

We further elaborate on the interpretation of the Akai
conceptually different from the sampling-theory-based re
model selection. It has has been noted in the literature (
1994, Bozdogan 1987) that there is a Bayesian basis for int
weight w_i as being the probability that model g_i is the expe
given the data (for convenience we usually drop this "expe
just think of the K-L best model). Once we have accept
model g_i given the data $\mathcal{L}(g_i|x)$, then we can compute the ap
probability that model g_i is the K-L best model if we ar
prior probabilities on the models (note that some Bayes
this approach ad hoc since it is not the full Bayesian approa
must specify an a priori probability distribution $\tau_1, \ldots, \tau_R$
belief that fitted model g_i will be the K-L best model fo
model set. These probabilities τ_i must be specified indep
prior to) fitting any models to the data.

2.14 Return to Flather's Models

We now extend the example in Chapter 1 where 9 mod
accumulation curve for data from Indiana and Ohio were
(1992, 1996). The simple computation of AIC was don
regression output from program NLIN in SAS (SAS Ins
this case, apart from a constant that is the same over all n

$$\text{AIC} = n \cdot \log(\hat{\sigma}^2) + 2K,$$

number of parameters (K), AIC values, $\Delta_i = \text{AIC}_i - \text{AIC}_{min}$ valu
adjusted R^2 values for the Indian–Ohio Major Land Resource Are
each model; the order is not relevant. Here the models are show
the number of parameters (K). However, this is only a convenience
example in Table 1.1.

Model	Number of parameters[a]	AIC value	Δ_i
ax^b	3	227.64	813.12
$a + b\log(x)$	3	91.56	677.04
$a\left(x/(b+x)\right)$	3	350.40	935.88
$a(1 - e^{-bx})$	3	529.17	1114.65
$a - bc^x$	4	223.53	809.01
$(a + bx)/(1 + cx)$	4	57.53	643.01
$a(1 - e^{-bx})^c$	4	−42.85	542.63
$a\left(1 - [1 + (x/c)^d]^{-b}\right)$	5	−422.08	163.40
$a[1 - e^{-(b(x-c))^d}]$	5	−585.48	0

[a] K is the number of parameters in the regression model plus 1 for σ^2.

where $\hat{\sigma}^2 = \text{RSS}/n$ and K is the number of regression pa
σ^2). AIC values for the 9 models are given in Table 2.
clearly the best approximating model for these data. Valu
$\text{AIC}_{\text{min}} = \text{AIC}_i + 585.48$ are also given and allow the resu
interpreted. Here, the second- and third-best models ar
(corresponding to Δ_i values of 163.40 and 542.63, resp
these Δ values are very large, and the inference here is
is clearly the best of the candidate models considered fo
This conclusion seems to be born out by Flather (1992), s
this model based on a careful analysis of residuals for each
Mallows' C_p. The remaining question is whether a still bet
been postulated with 6 or 7 parameters and increased stru
**criteria attempt only to select the best model from th
available; if a better model exists, but is not offered a
the information-theoretic approach cannot be expec
new model.**

Adjusted R^2 values are shown in Table 2.2, and while
a measure of the proportion of the variation "explained,"
in model selection (McQuarrie and Tsai 1998). In the ca
the best 4 models all have an adjusted $R^2 \approx 0.99$, promp
(erroneously) that all 4 models are an excellent fit to th
of the Δ_i values shows that models 6, 7 and 8 are incred
model 9. The evidence ratio for the best model versus th

$\exp(-163.4/2)$

There are additional reasons why adjusted R^2 is poor in usefulness should be restricted to description.

2.15 Summary

Ideally, the investigator has a set of "multiple working h thought hard about the background science of the issue science of the matter, experience, and expertise are used set of candidate models, representing each of these hyp **important philosophical issues that must receive incre** research problem should be carefully stated, followed by c cerning the sampling or experimental design. Sample size issues should be considered fully before the data-gatherin

The basis for the information-theoretic approach to i inference is **Kullback–Leibler information**,

$$I(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x|\theta)}\right) dx.$$

$I(f,g)$ is the "information" lost when the model g is us full reality or truth f. An equivalent interpretation of $I($ from the approximating model g to full truth or reality terpretation, we seek to find a candidate model that min the candidate models. This is a conceptually simple, yet However, $I(f, g)$ cannot be used directly, because it rec full truth or reality and the parameters in the approximati

Akaike (1973), in a landmark paper, provided a way to pected $I(f, g)$, based on the empirical log-likelihood fun the maximized log-likelihood value was a biased estimate Kullback–Leibler information and that under certain con approximately equal to K, the number of estimable paran imating model g. His method, *Akaike's information crite* model selection to be firmly based on a fundamental th door to further theoretical work. He considered AIC to likelihood theory, the very backbone of statistical theory Takeuchi (1976) derived an asymptotically unbiased estir pected Kullback–Leibler information that applies in gene special conditions underlying Akaike's derivation of AIC for Takeuchi's information criterion) requires large samp elements of two $K \times K$ matrices in the bias-adjustment t an important conceptual advance and further justifies AIC

approaches to adjusting for this bias (the bias adjustmen
from the maximized log-likelihood) are summarized belo

Criterion	Bias adjustment term
AIC	K
AIC_c	$K + \frac{K(K+1)}{n-K-1}$
TIC	$\text{tr}(J(\theta)I(\theta)^{-1} \approx K.$

These information criteria are estimates of relative, expec
and are an extension of Fisher's likelihood theory. AIC
to compute, quite effective in many applications, and w
use. When count data are found to be overdispersed, appr
tion criteria have been derived, based on quasi-likelihoo
QAIC_c). If overdispersion is found in the analysis of cou
log-likelihood function must be divided by an estimate o
($\hat{c}$) to obtain the correct log-likelihood. Thus, investigator
data analysis have several powerful methods for selectin
making inferences from empirical data to the population o
In practice, one need not assume that the "true model" is in
(although this is sometimes mistakenly stated in the techn

The AIC differences (Δ_i) and Akaike weights (w_i) ar
ing and scaling the hypotheses, represented by models.
(e.g., w_i/w_j) help sharpen the evidence for or against th
hypotheses. All of these values are easy to compute and s
and interpret.

The principle of parsimony provides a philosophical ba
tion, K-L information provides an objective target based
AIC, AIC_c, QAIC_c, and TIC provide estimators of relativ
formation. Objective model selection is rigorously based
These methods are applicable across a very wide range o
ses and statistical models. We recommend presentation
appropriate information criterion (AIC, AIC_c, QAIC_c or
various models in research papers to provide full informa
evidence for each of the models.

Basic Use of the Information-Approach

3.1 Introduction

Model building and data analysis in the biological science pose that the investigator has some advanced educatio sciences, and statistics in particular. This requirement researcher has substantial knowledge of statistical null hy proaches. Such investigators, including ourselves over th often find it difficult to understand the information-theo because it is conceptually so very different from the testin familiar. Relatively speaking, the concepts and practical us theoretic approach are simpler than those of statistical hy much simpler than some of the Bayesian approaches to Laud and Ibrahim 1995 and Carlin and Chib 1995).

The prevailing philosophy has been to use some test or to select a model, from a set of models, that is somehow ticular sense. Inference is then entirely conditional on this believe that approach should be merely the beginning, a humble beginning at that. There is much more to the mod than this initial solution. Substantive information is con ences (Δ_i), since they are free from arbitrary (and unkn are directly interpretable in many cases. Both the Δ_i and allow scientific hypotheses, carefully represented by mode discrete likelihood of model i, given the data ($\mathcal{L}(g_i|x)$), way to assess the relative support for the alternative mode

dence for the second, third, and nth best model, irrespec
in the set. These methods go well beyond just the select
and are very useful in assessing the empirical evidence fo
applied scientific problems.

It will be made clear in the next two chapters that e
analysis and inference philosophies are only a midway poi
theoretic paradigm. As we have struggled to understand th
become clear to us that inference based on only a single be
atively poor for a wide variety of substantive reasons. Inst
favor multimodel inference: procedures to allow formal
from all the models in the set. These procedures are sim
interpret and are the subjects of Chapters 4 and 5. Such m
includes model averaging, incorporating model selection
timates of precision, confidence sets on models, and simp
relative importance of variables.

The examples below focus on the selection of a single be
will appear in the following chapters. However, many m
beyond this initial approach in terms of the evidence fo
set. Methods to assess model selection uncertainty (e.g.
and Akaike weights w_i) are illustrated and discussed.
relative likelihood of model i, given the data, provide
concerning inferences about the actual K-L best model. S
should be viewed as a halfway point in understanding t
theoretic approach where formal inferences are drawn fro

While the derivation of AIC (Chapter 7) lies deep in t
matical statistics, its application is quite simple. Our initia
multiple linear regression model of cement hardening and
in the model selection literature. The remaining examples
on more complex data sets and models. These examples
into real-world complexities and illustrate the ease and g
of AIC in model selection and inference. Several of these e
ued in later chapters as additional concepts and methods a
examples deal with survival models, since that has been
interests.

Given a model, likelihood inference provides a quantitat
strength of evidence in the data regarding the plausible valu
in the model (Royall 1997). Given a well-developed set
models, information-theoretic methods provide a quanti
the strength of evidence in the data regarding the plausib
is "best." Information criteria can be computed and inte
aid of subjective judgment (e.g., α-levels or Bayesian p
candidate models has been derived.

approach was the clarification of the importance of mod
for substantial, prior information on the system being stu

At some early point in the analysis of count data, th
the global model should be assessed using standard metho
should accompany continuous data (see Carrol and Rup
generally no concept of overdispersion in continuous da
residual variation should receive careful attention. One s
liers, highly leveraged points, symmetry, trends, and aut
residuals (McCullagh and Nelder 1989). There are many
procedures that should be used to aid in the modeling of th
If, after proper attention to the a priori considerations, th
fits poorly, then information-theoretic methods will selec
set of poor-fitting models. This undesirable situation pro
poor science that went into the modeling and definition of
models. Lack of fit of the global model should be a fla
more consideration must be given to the modeling, based
of the questions being asked and the design of the data
the effort must be classed as exploratory and very tentati
some data dredging, leading perhaps to some tentative mo
conclusions. Treated as the results of a pilot study, ther
collected and the analysis could proceed in a more confirm
the techniques we outline in this book.

Computer programs for likelihood methods nearly alwa
of the log-likelihood at its maximum, and the appropriate i
can be easily computed by hand, if necessary. Similarl
the MLE of σ^2 from standard output of LS programs a
information criteria from this estimate in most cases. W
packages currently print AIC, relatively few print the valu
and this is unfortunate (see Example 1 below, where A
because the ratio n/K is small).

3.2 Example 1: Cement Hardening Data

The first example is a small set of data on variables thoug
heat evolved during the hardening of Portland cement (Wo
649). These data represent a simple use of multiple linea
(see Section 1.2.2). This data set (the "Hald data") has b
authors (e.g., Hald 1952:635–649, Seber 1977, Daniel and
and Smith 1981:294–342 and 629–673, Stone and Brook
McCulloch 1993, Hjorth 1994:31–33, Ronchetti and Stau
Ibrahim 1996, and Sommer and Huggins 1996) and will i

($3CaO \cdot SiO_2$), x_3 = tetracalcium alumino ferrite ($4CaO \cdot Al_2O_3$ silicate ($2CaO \cdot SiO_2$)] are used to predict the dependent variabl evolved per gram of cement after 180 days of hardening.

x_1	x_2	x_3	x_4	y
7	26	6	60	78.5
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

important points. The data include 4 predictor variables an of 13 (Table 3.1). The predictor variables (as a percentag x_1 = calcium aluminate ($3CaO \cdot Al_2O_3$), x_2 = tricalcium s x_3 = tetracalcium alumino ferrite ($4CaO \cdot Al_2O_3 \cdot Fe_2O_3$), silicate ($2CaO \cdot SiO_2$), while the response variable is y = off during hardening per gram of cement after 180 day (1971) provide further details on these data for the inte *approximating model to use?*" is the primary focus of thi

The small size of the sample necessitates the use of however, we will present comparable values for AIC in th use an obvious notation for denoting what variables ar model. That is, if variables x_1 and x_3 are in a particular m as model {13}; each model has an intercept (β_0).

3.2.1 Set of Candidate Models

Because only 4 variables are available, the temptation is to models ($2^4 - 1 = 15$) involving at least one of the predictor the small sample size, we will consider this example as larg lacking any personal knowledge concerning the physics or hardening, we will consider the full set of models, includi {1234} with $K = 6$ parameters. While we generally adv consideration of all possible models of the x_i (but no int of the predictor variables), this approach will allow som

values for both AIC and AIC_c followed by the Akaike weights (w_i),
are ordered in terms of Δ_i for AIC_c.

Model	K	$\hat{\sigma}^2$	$\log(\mathcal{L})$	Δ_i AIC	Δ_i AI
{12}[1]	4	4.45	−9.704	0.4346	0.0
{124}	5	3.69	−8, 478	0.0000	3.1
{123}	5	3.70	−8.504	0.0352	3.1
{14}	4	5.75	−11.370	3.7665	3.3
{134}	5	3.91	−8.863	0.7528	3.8
{234}	5	5.68	−11.290	5.6072	8.7
{1234}	6	3.68	−8.469	1.9647	10.5
{34}	4	13.52	−16.927	14.8811	14.4
{23}	4	31.96	−22.519	26.0652	25.6
{4}	3	67.99	−27.426	33.8785	31.1
{2}	3	69.72	−27.586	34.2052	31.4
{24}	4	66.84	−27.315	35.6568	35.2
{1}	3	97.37	−29.760	38.5471	35.7
{13}	4	94.39	−29.558	40.1435	39.7
{3}	3	149.18	−32.533	44.0939	41.3

[1] Here, $\log(\mathcal{L}) = -n/2 \cdot \log(\hat{\sigma}^2) = -9.7039$, $AIC_{min} = -2\log(\mathcal{L}) + 2K =$
$AIC + \frac{2K(K+1)}{n-K-1} = 32.4078$.

others in the published literature (e.g., Draper and Smith
and Hoeting and Ibrahim 1996). We note, however, tha
only a single variable might have been excluded on a pri
cement involves a mixture of at least two compounds that
will extend this example in Chapter 4 to examine the issu
uncertainty and other issues.

3.2.2 Some Results and Comparisons

The use of AIC_c suggests model {12} as the best appro
these data (Table 3.2). The estimated regression coeffic
model are

$$\hat{E}(y) = 52.6 + 1.468(x_1) + 0.662(x_2)$$

where the estimated standard errors of the 3 estimated pa
model) are 2.286, 0.121, and 0.046, respectively (this re
with Hald 1952). The adjusted $R^2 = 0.974$ and the ML
AIC_c-selected model. The second-best model is {124},
units from the best model (Table 3.2). Other candidate mo
clearly many of the models represent poor approximations
(at least the models in Table 3.2 with Δ_i values > 10). No

global model (model {1234}).

The Akaike weight for the best model is not large, re
for the other models. The ratio of the weights for the be
4 next-best models ranges from only 4.8 to 7; this is not
model {12} is likely best if other replicate samples were a

Using a type of cross-validation criterion (Q_{cv}), Hjorth
model {124} with $K = 5$ for these data. Here, his result i

$$\hat{E}(y) = 71.6 + 1.452(x_1) + 0.416(x_2) - 0.2$$

where the estimated standard errors are 14.142, 0.117, 0
spectively. Model {124} has an adjusted $R^2 = 0.976$ and
and Smith (1981:325–327) used cross-validation and the l
selection criterion, which is quite similar to Q_{cv}, and also s
Note, had AIC been used, ignoring the ratio $n/K \approx 2$, moc
been selected (Table 3.2); AIC_c should be used if this ratio

Is there any basis to say that AIC_c selected a better a
than Hjorth's cross-validation procedure or AIC or the PR
is difficult to answer conclusively because truth is not kn
the regression coefficient on x_4 is not "significant" unde
pothesis testing scenario ($t = 1.36$, 9 df), and the estim
on the regression coefficient for x_2 increased by a factor
0.186 compared to model {12}. The adjusted R^2 statistics
model is 0.976 (vs. 0.974), but it has one additional param
coefficient between x_1 and x_3 was -0.824, while the co
and x_4 was -0.973. Just on the basis of this latter corr
wise to allow both x_2 and x_4 in the same model (if n w
only 13, perhaps there would be more support for includ
While not completely compelling, it would seem that A
better parsimonious model in this case. An additional, neg
is the computer-intensive nature of Hjorth's cross-validat
compared to the information-theoretic approach. With mor
sizes or more variables, or with more models to consider,
approaches may often become computationally too "cost

Draper and Smith (1981) used Mallows's C_p statistic and
{12}, in agreement with AIC_c (this might be fortuitous, be
ple version of C_p or Q_{cv} is available). They further point o
constant (approximately 98%) for any i; thus the $\boldsymbol{X'X}$ mat
is theoretically singular. Small rounding errors were ev
since the percentage data were expressed as integers, lea
barely nonsingular. At best, model {1234} would be a poo
ysis of these data. They also warn against the unthinkin
regressions and present a detailed analysis of forward, b

and retaining only x_1 and x_2. This represents an improveme
merely add new variables, without looking to see whether
has become redundant. Draper and Smith (1981) provid
of the various older model selection alternatives and offer
mendations (but do not discuss any of the information-th
They provide an intensive analysis of the cement data o
and include detailed computer output in two large append

Another analysis approach involves computation of th
nents on the (centered) $\boldsymbol{X'X}$ matrix and examination of th
for the 4 explanatory variables (see Draper and Smith 1
and Brooks 1990). The principal component eigenvalue
1.57607, 0.18661, and 0.00162. Approximately 95.3% o
is contained in the first 2 eigenvectors, while 99.96% is
vectors. These results certainly suggest that the global r
data (i.e., 4 predictor variables are redundant). In additi
that 2 predictor variables will suffice (given $n = 13$). C
of the percentage eigenvalues requires some judgment a
thermore, relatively few biologists are knowledgeable a
eigenvalues and eigenvectors. *We believe that the investig
stand the methods leading to the results of their work; this i
with some advanced methods*. Such understanding seems
the information-theoretic approaches.

One could ask whether there is a need for model selec
only 4 predictor variables (i.e., why not merely take the
parameters and use it for inference?). This simple strateg
as we illustrate here. First, note that this global model
relative to model {12}, and is therefore a poor approxim
data available. The estimates of parameters for the global

$$\hat{\mathrm{E}}(\hat{y}) = 62.4 + 1.551(x_1) + 0.510(x_2) + 0.102(x_3)$$

where the estimated standard errors, given this model, are 7
0.755, and 0.709, respectively. These standard errors are la
is nearly singular (the percentage coefficients of variatio
were 4.3, 8.2, and 6.9 under model {12}, compared to 11
respectively, under model {1234} (see Wood and Thom
regression coefficient for x_1 might be judged as "signific
testing sense, and the model is clearly overfit (see Figure
has an adjusted $R^2 = 0.974$ and $\hat{\sigma} = 1.918$. Surely a p
such as {12}, would better serve the analyst in this case.

Loss of precision is expected in using an overfit glob
there is also a nonnegligible probability that even the si
parameter may be incorrect in such cases. It seems some

If all the predictor variables are mutually orthogonal (u
selection is not quite as critical, and the global model
not be so bad. Orthognality arises in controlled experimer
and levels are *designed* to be orthogonal. In observation
high probability that some of the predictor variables wi
dependent. Rigorous experimental methods were just bei
the time these data were taken (about 1930). Had such d
widely available and the importance of replication under
have been possible to break the unwanted correlations an
and establish cause and effect.

With only a single data set, one could use AIC_c and s
for inference. However, if several other independent data
would the same model be selected? The answer is *perha*
generally there would be variation in the selected model fi
as there would be variation in parameter estimates over
the same model is used for analysis. The fact that other da
the use of other models leads us to the issue of model sel

Based on simulation studies, we are usually surprised by
there is in selecting a parsimonious model for a given pr
strably the case that in many real-world problems there
selection uncertainty. We generated 10,000 bootstrap sam
to estimate model selection uncertainty. The parameters
models shown in Table 3.2, were estimated and AIC_c wa
bootstrap sample. The following summary shows the rela
frequencies (π_i) from applying AIC_c (models not shown
to each of the 10,000 bootstrap samples. Here, $\hat{\pi}$ are the
lection probabilities. Also shown are the Akaike weights (
data:

Model	K	Bootstrap Sel. Freq. $\hat{\pi}_i$	Akaike weights w_i
{12}	4	0.5338	0.567
{124}	5	0.0124	0.118
{123}	5	0.1120	0.116
{14}	4	0.2140	0.107
{134}	5	0.0136	0.081
{234}	5	0.0766	0.007
{1234}	6	0.0337	0.003
{34}	4	0.0039	0.000.

As might be expected with such a small sample size, the s
varied substantially, and model {12} was selected as the be
of the bootstrapped samples. Model {14} was selected 21
that the simple correlation between variables x_2 and x_4 wa
it is a quite reasonable result that models {12} and {14} ar

size of 13 observations. Further results based on the boots
following chapter.

3.2.3 A Summary

In summary, the simple approach of using AIC_c appears t
parsimonious model as the basis for inference from the
AIC_c sharpens the inference about which parsimonious n
to AIC. A priori information could have resulted in few
and generally strengthened the process (note, that Hald (
only an analysis of x_1 and x_2 and presented the analysis
variables several pages later). It seems likely that model
variable might have been excluded from serious consider
must have been known about cement in the late 1920s. S
that Woods et al. (1932) had knowledge of the negative r
x_2 and x_4; after all, model {14} was their second-best mod
of both x_2 and x_4 in the same model (where the correlati
the over parametrized global model. An important feature
theoretic approach is that it provides a ranking of alternati
some inferences to be made about other models that mig
addition, the rankings suggest that some models that ren
models {24}, {1}, {13}, and {3} for the cement data). The A
to focus the evidence for or against the various models.
**carefully defining a small set of candidate models, bas
and what is known about the problem, cannot be over**

An investigator with, say, 10 explanatory variables ca
much from the data and a multiple linear regression ana
some substantial supporting science that can be used to hel
of models to consider. In this case, there would be $2^{10} =$
more if transformations or interaction terms were allow
would surely be a risk. The analysis, by whatever method
considered exploratory and the results used to design fu
leading to a more confirmatory analysis, based on some a p

3.3 Example 2: Time Distribution of an I
Added to a Simulated Ecosystem

This example concerns the addition of the insecticide DU
oratory system that simulates a pond of water. The orig
by Smith (1966) and his colleagues; our main reference f

Bliss and Neely note (1975) that the determination of
distribution of this chemical introduced into an ecosys
environmental issue. They go on to mention that "...
model describing each step of the process would be ex
is important, however, to try to find a suitable model
important chemical, physical, and biological phenomena
predict the long-term environmental consequences." Thi
is consistent with Akaike's and the one recommended h
used because it rests on systems of first-order different
parameters, given a model, are estimated by least squar
easily be used to compute AIC values to aid in selectio
approximating model.

The active ingredient of DURSBAN® is 0,0-Diethyl
pyridyl) phosphorothioate, which was labeled with radio
the pyridyl ring and added at a level of 1 mg/6 gal in a 10
Figure 3.1). This aquarium contained 2 inches of soil (
ter), plants (salvinia, anacharis, milfoil, and water cucumb
Samples of the various components were analyzed for ra
ferent time periods following the addition of DURSBAN
each time period yielded a sample size (n) of 36. The
in percentages from the crude radioactivity measureme

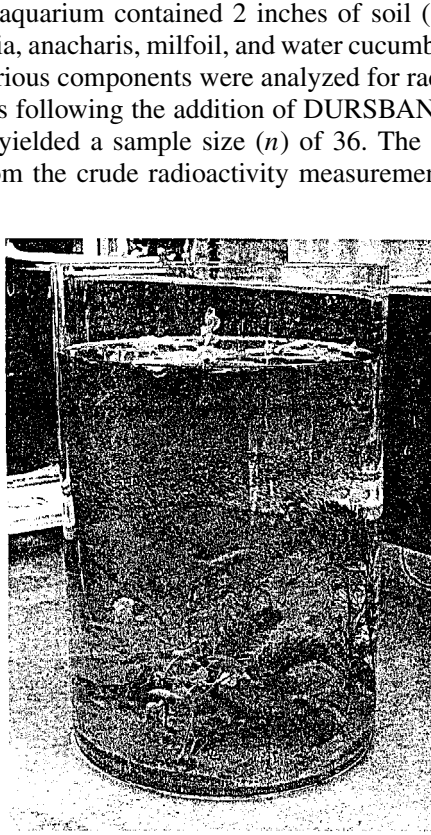


FIGURE 3.1. Glass aquarium used in the studies of DURSBAN

Time after DURSBAN® addition (hours)	Percent radioactiv	
	Fish	Soil & Plants
0	0	0
1.5	15.2	35.2
3.0	19.0	46.0
4.0	19.3	56.0
6.0	20.7	61.0
8.0	23.0	60.5
10.0	24.2	59.3
24.0	21.2	51.5
48.0	23.0	38.3
72.0	22.7	38.3
96.0	20.5	36.3
120.0	17.3	38.3

1975:150). The authors of the study assumed that the m
normally distributed, with zero means and a constant stan
(we take this to mean the actual measurement error of the

3.3.1 Set of Candidate Models

Blau and Neely (1975) had a great deal of knowledge ab
they exploited this in a priori model building. They begar
an equilibrium exists between DURSBAN® in the wate
components (B), and a direct uptake of the chemical by t
to their Model 1 (Figure 3.2), which was represented by a s
equations, where the rate parameters to be estimated are

$$dx_A(t)/dt = -k_1 x_A(t) + k_2 x_B(t) - k_3 x_C$$
$$dx_B(t)/dt = k_1 x_A(t) - k_2 x_B(t),$$
$$dx_C(t)/dt = k_3 x_A(t),$$

with initial conditions $x_A(0) = 100$, $x_B(0) = 0$, and $x_C(0$
of compartment model (Brown and Rothery 1993) and is
fields. Blau and Neely (1975) used $x_A(t)$, $x_B(t)$, and $x_C(t$
at time (t) of A, B, and C, respectively, with the restrictio

$$x_A(t) + x_B(t) + x_C(t) = 100.$$

They used nonlinear least squares to estimate model param
and their analytic methods were quite sophisticated. The
for this model were $\hat{k}_1 = 0.510, \hat{k}_2 = 0.800, \hat{k}_3 = 0.0093$

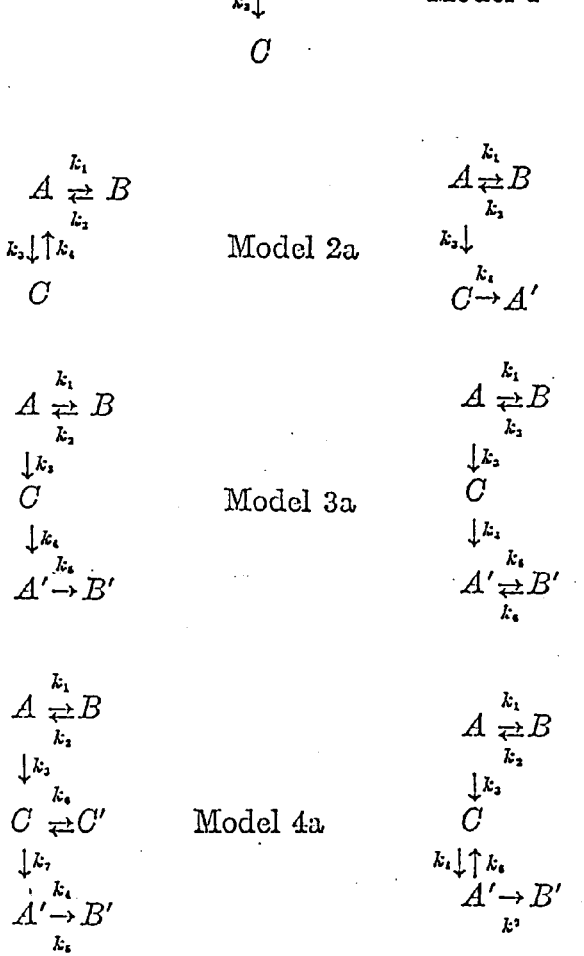


FIGURE 3.2. Summary of models used by Blau and Neely (DURSBAN® in a simulated pond ecosystem.

(this is their residual sum of squares (RSS) divided by n t σ^2); thus, $K = 4$ for this model.

Blau and Neely (1975) built six other models, each base of the system, but also based on examination of the residua (there are some inconsistencies here that we were unable will use the material from their paper). While some data d their main derivation of additional models seemed to s hypotheses about the processes. They were well aware

of-fit tests and separated "pure error" from the remaining re
selection was accomplished by statistical hypothesis te
tests) and examining the RSS. They found Model 4a (see
best and also found some support for Model 4b.

3.3.2 *Some Results*

Analysis of these data under an information-theoretic parad
Blau and Neely's (1975) Table II, since they provide value
for each of their seven models. Due to the relationships bet
and ML theory (see Section 1.2.2),

$$\log(\mathcal{L}(\underline{\hat{k}}, \hat{\sigma}^2 \mid data)) = -n/2 \cdot \log(\hat{\sigma}^2)$$

where $\hat{\sigma}^2 = \text{RSS}/n$. Then,

$$\text{AIC} = -2 \cdot \log(\mathcal{L}(\underline{\hat{k}}, \hat{\sigma}^2 \mid data)) + 2K$$

and

$$\text{AIC}_c = \text{AIC} + \frac{2K(K+1)}{n-K-1}.$$

These computations were done by hand on a simple calc
proximately 20 minutes. The results of this extended an
Table 3.4 and suggest that Model 4a is the best to use for infe
with Blau and Neely (1975). Only Model 4b is a compet
value of 7.611 and seems relatively implausible for these
The evidence ratio for model 4*a* vs. 4*b* is 0.978/0.022
strong support for 2-way transfer between the viscera and
the essential difference between models 4 and 4*b* is $c \leftrightarrows$

Carpenter (1990) used these data and seven models
Bayesian analysis with equal Bayesian prior probabilitie

TABLE 3.4. Summary of model selection statistics (the first three co
and Neely 1975). Statistics for the AIC_c-selected model are shown

Model	RSS	$\log(\mathcal{L}(\hat{k}, \hat{\sigma}^2 \mid data))$	K	AIC	AIC
1	5374	−90.105	4	188.209	189.4
2a	1964	−71.986	5	153.972	155.9
2b	848	−56.869	5	123.737	125.7
3a	208.3	−31.598	6	75.196	78.0
3b	207.9	−31.563	7	77.127	81.1
4a	**58.6**	**−8.770**	**8**	**33.540**	**38.8**
4b	79.4	−14.238	7	42.475	46.4

Model 4b a poor second. In this example, R ranged fr
the various methods might be expected to be in somewh
This example illustrates that it is often easy to perform
on complex systems, based on information provided in pu
analysis clearly shows that five of the seven models have
port, and inferences from these models would likely be po
third-best model (3a) has an evidence ratio of 3×10^9, wh
(1) has an evidence ratio of 2×10^{33}. Clearly, these mod
given the data available.

Blau and Neely's (1975) results are interesting, and w
best model. The evidence ratio for the second-best model
reasonable to base inference on just the best model in th
are often comfortable with the concept that inferences can
model; in a sense, the inference here is the model.

After a final model is chosen it is often wise to examin
standard methods. Such examination may reveal issues
study; in this sense, science never "stops."

Formal *statistical* inferences include the following: (1)
libration between DURSBAN and the soil and plant systen
by a shorter uptake of DURSBAN by the fish; (3) fish ter
excrete DURSBAN; (4) the liberated material (metabol
again taken up by the soil and plants; (5) fish have two co
cera and the flesh; (6) the final sink for DURSBAN is the
and (7) plants readily dissipate the metabolite as degra
H_2O. Estimates of the various transfer rates are given by
of precision are available as standard errors or confidence

3.4 Example 3: Nestling Starlings

We generated a set of Monte Carlo data to illustrate ma
cussed with a much more complicated example of an e
Thus, in a sense, the generating model is "truth"; we will
realism for the moment, but mitigate it by including many p
and a wide variety of tapering treatment effects. In additi
global model that has four fewer parameters than the ge
the generating model is not in the set of candidate mode
example contains many so-called nuisance parameters (
ties). This is the only example in Chapter 3 where "truth"
interesting insights can be gained from this knowledge. Th
is what parsimonious approximating model can be used f
will lead to valid inference about the structure of the sys

3.4.1 Experimental Scenario

We generated data to mimic the experiment conducted
(1988) (also see Burnham et al. 1987:343–348). The rese
to the survival effects of an organophosphate pesticide adn
European starlings (*Sturnus vulgaris*). We assume for illus
field experiment is designed using artificial nest boxes plac
island. Fledgling birds are assumed not to leave the islan
and early fall months when the experiment is conducted (
Nest boxes are monitored during the nesting season to d
hatching. All nestlings are leg-banded with uniquely num
following hatching, and half of those nestlings are ran
treatment group and the remaining birds assigned to a co
we will assume that 600 nestling starlings are banded and
box (i.e., the number of starlings originally released in
All nest boxes contain 4 young birds (thus 2 treatment
we assume these to be of nearly uniform size and age an
they move about and behave independently. Starlings ra
be in the treatment group receive an oral dose of pesticid
Birds in the control groups are given pure corn oil under ot
conditions. Colored leg bands provide a unique identificat
and therefore its group membership, on each weekly resig
collection will be assumed to begin after a 4-day period fo
for simplicity, we assume that no birds die due to handli
marking but before resighting efforts begin a week later. Su
potentially resighted during the following 9 weeks; sampl
island and is done on each Friday for 9 weeks. Thus, the
10 occasions; occasion 1 is the initial marking and release
9 resighting occasions.

The pesticide is hypothesized to affect conditional surv
parameters of interest) and resighting probabilities (the nu
however, the pesticide industry's position is that only m
are likely, while environmental groups suspect that there
(short-term) and chronic (long-term) effects on survival pro
that the resighting probabilities might also be affected by
the set of candidate models might span the range of the con
of course, one might design the experiment to include sever
released at different, independent locations (islands), and
the basis for empirical estimates of treatment effect and pre
et al. 1987 for a discussion of experiments of this general
focus on an example of the model selection issue and not

tional survival probability (ϕ) and resighting probability
and control (c) groups at week i:

$$\phi_{ti} = \phi_{ci} - (0.1)(0.9)^{i-1} \text{ for } i = 1, \ldots,$$
$$p_{ti} = p_{ci} - (0.1)(0.8)^{i-2} \text{ for } i = 2, \ldots,$$

using program RELEASE (Burnham et al. 1987). These
a smooth temporal tapering of effect size due to the treat
tional survival and resighting probabilities. That is, each w
pesticide is diminished. We used the initial per-week sur
probabilities for the control group as 0.9 and 0.8, respe
survival and resighting probabilities for the control grou
week (i.e., $\phi_{ci} \equiv \phi_c \equiv 0.9$ and $p_{ci} \equiv p_c \equiv 0.8$). The dat
3.5 for each treatment and control group.

3.4.3 Set of Candidate Models

Define ϕ_{vi} as the conditional probability of survival fo
($v = t$ for treatment and c for control) from week i to $i + 1$
as the conditional probability of resighting for treatment g
$i = 2$ to 10). The set of models that seem reasonable mig
no treatment effects (g_0), a model for an acute effect only
probability ($g_{1\phi}$), and a model for an acute effect on bo
probability and the first resighting probability (denote th
occurs at week 2) (model g_{2p}). This initial line of a priori
to three models:

Model	Parametrization
g_0	All $\phi_{ti} = \phi_{ci}$ and all $p_{ti} = p_{ci}$ (no treatme
$g_{1\phi}$	g_0, except $\phi_{t1} \neq \phi_{c1}$ (an acute effect on ϕ_1
g_{2p}	$g_{1\phi}$, except $p_{t2} \neq p_{c2}$ (acute effects on ϕ_1

Chronic effects might arise from starlings that are in
effects of the pesticide; these starlings might be more susc
(this would be revealed in lessened survival during the sum
be less active in foraging (this might be revealed in diffe
resighting compared to the control starlings, because sam
the summer period). Chronic effects, if they exist, might b
That is, one might expect chronic effects to diminish o
the starlings in the control group. Agreement is reached,
evidence, that chronic effects, if they exist, should not las
week.

10). The data given for each group (i) are the number of starlings
after last being released at time i. R_i = the number of birds relea
all of those released in weeks 2, ..., 9 were merely rereleased. Ea
$\left(R(i) - \sum_j m_{ij}\right)$ is modeled as a multinomial distribution with sa

Week	$R(i)$	Observed Recaptures for Treatmer $m(i, j)$					
		$j = 2$	3	4	5	6	7
1	300	158	43	15	5	0	
2	158		82	23	7	1	
3	125			69	17	6	
4	107				76	8	
5	105					67	2
6	82						5
7	81						
8	70						

Week	$R(i)$	Observed Recaptures for Control $m(i, j)$					
		$j = 2$	3	4	5	6	7
1	300	210	38	5	1	0	
2	210		157	20	8	2	
3	195			138	24	2	
4	163				112	24	
5	145					111	1
6	139						10
7	124						
8	115						

Define $S_i = \phi_{ti}/\phi_{ci}$ for $i = 1$ to 7 as the measure of
conditional survival probability, compared to the control
the control group will experience some mortality as the
here the interest is in any *additional* mortality incurred by
the pesticide treatment.) The parameters S_i ($i = 1, 2, \ldots,$
0.929, 0.943, 0.954, 0.964, and 0.971, respectively. Wit
effects, one expects $S_2 < S_3 < S_4 < \cdots < S_7 < 1$, as c
parameters above (of course, the unconstrained *estimates*
based on some approximating model, might not follow
Here, it seems reasonable to consider the presence of ch
additional impacts to the hypothesized acute effects. Thu
chronic effects on both conditional survival and resight

$g_{2\phi}$	g_{2p}, except $\phi_{t2} \neq \phi_{c2}$ (chronic effect o
g_{3p}	$g_{2\phi}$, except $p_{t3} \neq p_{c3}$ (chronic effect o
$g_{3\phi}$	g_{3p}, except $\phi_{t3} \neq \phi_{c3}$ (more chronic ef
g_{4p}	$g_{3\phi}$, except $p_{t4} \neq p_{c4}$ (more chronic ef
$\vdots$	
$g_{7\phi}$	All ϕ_{vi} and p_{vi} differ by treatment grou 7 weeks

This last candidate model ($g_{7\phi}$) allows chronic treatment e tional survival and resighting probabilities up through the in addition to the acute treatment effects on ϕ_{t1} and p_{t2}. T as our global model, and it has 30 parameters. The treat through the ninth week; thus, the generating model is not i models and has more parameters than the global model (

Model g_0 has 17 parameters, while model $g_{7\phi}$ has 30 p plest model would have a constant survival and resighting group ($g_{\phi,p}$) and thus no treatment or week effects on ei vival or resighting probabilities. This model would have (ϕ and p). Alternatively, a four-parameter model could all parameters to differ by treatment group (ϕ_t, ϕ_c, p_t, and relatively large sample size in this example, these models ple and unlikely to be useful based on initial biological might well exclude these from the set of candidate mode **biological support should not be included in the set o** However, as an example, we will include these simple m tion and note that these models might well be viewed as if the initial sample size released were 60 instead of 600.

The effective sample size in these product multinomial m of starlings released (or rereleased) at each week. [The e in these product multinomial models is a complicated is divert attention to this matter here, except to say that here in the context of AIC_c. Technical notes on this subject m KPB.] In this example, $n = 2{,}583$ releases (a resighting ing "recaptured and rereleased"). Because 600 starlings were released at week 1 (the nest boxes), the remaining resighted at least once. Because of the large effective sa AIC_c is unnecessary; however, if one chose always to u AIC, no problems would be encountered because AIC_c a n/K gets large.

A statistician on the research team suggests adding the possible tapering treatment effects on conditional su

insights concerning possible long-term chronic treatment
ploying a type of sine transformation on the parameters
used here. In this transformation, the parameter (θ, repres
assumed to be between 0 and 1) to be modeled as a fun
covariate (e.g., X) is replaced by the expression $(\sin(\alpha$
new parameters α and β are the intercept and slope para
in the covariate model. The transformation utilizes one-
to model increasing or decreasing sigmoid functions and
link function in generalized linear models. In particular,
submodel $g_{\sin p_t}$ were defined for the dynamics of starli
group:

$$g_{\sin \phi_t} \qquad \sin(\phi) = \alpha + \beta(\text{week}),$$
$$g_{\sin p_t} \qquad \sin(p) = \alpha' + \beta'(\text{week}).$$

These submodels each have only 2 parameters (intercepts
β and β') and assume that $\sin(\phi_t)$ or $\sin(p_t)$ is a linear fu
conditional survival of starlings in the treatment group wi
as the summer period progresses, eventually approximatin
the control group).

These above two submodels for the treatment group
four submodels below for the control group:

$g_{\phi_{ci}}$	ϕ is allowed to differ for each week; he ($i = 1, \ldots, 8$).
g_{ϕ_c}	ϕ is assumed constant across weeks.
$g_{p_{ci}}$	p is allowed to differ for each week; he ($i = 2, \ldots, 9$).
g_{p_c}	p is assumed constant across weeks.

For example, a model can be developed using $g_{\sin \phi_t}$ for co
the treatment group and model g_{ϕ_c} for the conditional su
group. This part of the model has 3 parameters; α, β, ϕ_c, p
tion of the resighting probabilities. Thus, one could cons
the treatment group and model $g_{p_{ci}}$ for the control group as
for the resighting probabilities. This would add the param
$\ldots$, p_{c10}, for a total of $K = 14$ parameters. As an illustr
rich mixture of candidate models in Table 3.6 (a set of 2
If this were a real situation, still other a priori models mig
carefully supported with biological reason. If this experi
only 60 nestlings, then several simple models should be in
high-dimensional models would be deleted. This set of 2
will serve as a first example where there is some substant

a pesticide. (All values are scaled by the additive constant −4,467.
best model.) Akaike weights (w_i) are also shown.

Model	AIC	No. Parameters	Δ_i
$g_{7\phi}$ (global)	4,495.409	30	27.
g_{7p}	4,493.619	29	25.
$g_{6\phi}$	4,491.649	28	23.
g_{6p}	4,489.889	27	22.
$g_{5\phi}$	4,491.679	26	23.
g_{5p}	4,491.929	25	24.
$g_{4\phi}$	4,490.199	24	22.
g_{4p}	4,489.029	23	21.
$g_{3\phi}$	4,489.629	22	21.
g_{3p}	4,492.619	21	24.
$g_{2\phi}$	4,501.809	20	34.
g_{2p}	4,517.019	19	49.
$g_{1\phi}$	4,523.489	18	55.
g_0	4,532.599	17	64.
$g_{\sin\phi_t,\phi_{ci},\sin p_t,p_{ci}}$	4,485.669	21	17.
$g_{\sin\phi_t,\phi_{ci},\sin pt,p_c}$	4,475.249	14	7.4
$g_{\sin\phi_t,\phi_c,\sin p_t,p_{ci}}$	4,479.359	14	11.
$\boldsymbol{g_{\sin\phi_t,\phi_c,\sin p_t,p_c}}$	**4,467.779**	**6**	**0.**
$g_{\sin\phi_t,\phi_{ci},p_{ti},p_{ci}}$	4,488.629	28	20.
$g_{\sin\phi_t,\phi_{ci},p_{ti},p_c}$	4,478.209	21	10.
$g_{\sin\phi_t,\phi_c,p_t,p_{ci}}$	4,484.699	13	16.
$g_{\sin\phi_t,\phi_c,pt,p_c}$	4,473.119	5	05.
$g_{\phi_t,\phi_c,p_t,p_c}$	4,770.479	4	302.
$g_{\phi,p}$	5,126.609	2	356.

3.4.4 Data Analysis Results

As one would expect with simulated data, they fit the r
generation; $g_{9\phi}$ ($\chi^2 = 35.5$, 36 df, $P = 0.49$). [A large lit
of-fit testing in this class of models exists (e.g., Burnh
Pollock et al. 1990); we will not pursue the details of su
data were simulated such that no overdispersion was pres
of the overdispersion factor c could be computed under t
from the results of the goodness-of-fit test, $\hat{c} = \chi^2/\text{df}$ =
global model $g_{7\phi}$ has fewer parameters than the generatin
these data well ($\chi^2 = 35.4$, 30 df, $P = 0.23$). The valu
model was 1.18, reflecting no overdispersion in this case,
(which is known to be true in this instance); after all, it,
of "truth." In practice, one cannot usually distinguish bet

multiplying these by 1.18 (or the standard errors by the sq
particular, one might consider using the modifications to A
2.5 (i.e., QAIC $= -2\log(\mathcal{L})/1.18 + 2K$). We will mer
a later point. The critical information needed for selectio
model and ranking and scaling the other models is shown

The interpretation of the 24 models for the experimenta
3.6) can be sharpened by examining the Akaike weights.
the AIC-selected model ($g_{\sin\phi_t,\phi_c,\sin p_t,p_c)}$ is 0.906, while th
($g_{\sin\phi_t,\phi_c,p_t,p_c}$) has a weight of 0.063 and the third-best mo
has a weight of 0.022. The sum of the weights for the 2
is less than 0.01. In this case, one is left with strong s
model, with fairly limited support for the second-best m
of best vs. second-best ≥ 14). The evidence ratio for the
about 41.2. Thus, the data support one model as convinc
seems to be little need to attempt model averaging or boo
4) to gain further robustness in inferences from these
models). In addition, the use of conditional standard er
model, will likely suffice. Note that bootstrapping in thi
very, very difficult. Software development would be a v
and computer time on a Pentium 1PC would likely take
the Akaike weights provide a distinct advantage in compl
this simulated starling experiment.

The model with the minimum AIC value was $g_{\sin\phi_t,\phi_c}$
parameters (α, β, α', β', ϕ_c, and p_c). Using estimates of
one can derive MLEs of the survival and resighting param
MLEs for the treatment survival probabilities were as fol

i	ϕ_{ti}	$\hat{\phi}_{ti}$	$\widehat{\text{se}}(\hat{\phi}_{ti})$
1	0.800	0.796	0.021
2	0.810	0.810	0.016
3	0.819	0.824	0.014
4	0.827	0.838	0.014
5	0.834	0.851	0.160
6	0.841	0.864	0.019
7	0.847	0.876	0.022
8	0.852	0.887	0.026
9	0.857	0.898	0.029

The survival parameter for the control group was 0.90, a
selected model was 0.893 ($\widehat{\text{se}} = 0.008$). These estimates
to the parameter values, and one can correctly infer the d
effect of the treatment on weekly survival probabilities.
the 9 estimates of survival probability for the treatment
derived from the MLEs of α and β in the submodel $\sin(\phi_t$

sine model estimates the acute and chronic effects of th
the conditional survival and resighting probabilities for b
group. The conditional survival and resighting probabili
control group were constant over weeks in this model, bu
in the treatment group. The AIC-selected model capture
of the generated process. Figure 3.3 illustrates the simila
values, the estimates from the global model ($g_{7\phi}$), and th
AIC-selected model in terms of the treatment effect, $1 -$

**Part of the reason that this analysis was successfu
reasoning that led to *modeling* the treatment effects, ra
estimate the week-specific treatment effects (i.e., the S_i**
modeling allowed substantial insight into the tapering, cl
case. Note: The two simplest models ($g_{\phi_t,\phi_c,p_t,p_c}$ with K
$K = 2$) were not at all plausible ($\Delta_i = 302.70$ and 356.13,
that these models would not normally have been considere
experiment, since they lacked any reasonable biological su
sample size involved. Of course, had sample size been ve
models might have been more reasonable to include in th

**If sample size is small, one must realize that relativel
is probably contained in the data (unless the effect size
and the data may provide few insights of much interest**
routinely err by building models that are far too complex fo

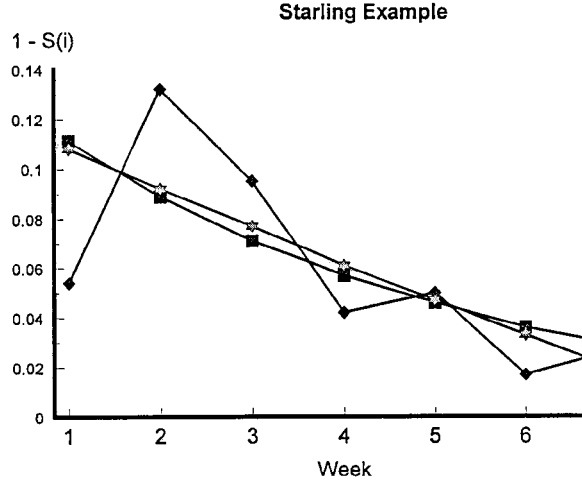


FIGURE 3.3. Treatment effect ($1 - S_i$, for week $i = 1, \ldots, 7$) fo
the generating models ($g_{9\phi}$) with 34 parameters, compared with es
eters from the global model ($g_{7\phi}$) with 30 parameters and the AIC
parameters.

before analysts get a feeling for modeling based on sam
known about the science of the problem of interest.

3.4.5 *Further Insights into the First Fourteen Ne*

If only the first 14 models (Table 3.6) had been defined a
concerning which model to use would have been far less cl
these 14 models is over 25 units from the AIC-selected m
not have been known. Second, 7 models have AIC values
best of the 14. Thus, some additional steps would be nece
model selection uncertainty into inference for these exp
analysis was based on just the first 14 models.

We now examine further the results that would have be
set of candidate models included just the first 14 models in
tial theory (e.g., the estimators exist in closed form) and
RELEASE, Burnham et al. 1987) exist for this sequenc
allowing the illustration of a number of deeper points. F
that these 14 models are clearly inferior to the models hy
treatment effects (a diminishing linear treatment effect
link function) for birds in the treatment group (e.g., the b
model g_{4p}, is 21.25 AIC units above the selected model an
compared to only 6 parameters for the AIC-selected mode
to the importance of a good set of candidate models. Se
chronic effects could not be identified by model g_{4p} (i.e., t
effects on survival in the later time periods, S_4, S_5, and
values provide clues that at the very least, models $g_{4\phi}$ (th
(therefore, S_5) are also somewhat supported by the data
models have AIC values within 1.17 and 0.86, respective
fact, models g_{3p} through model $g_{6\phi}$ have fairly similar A
and Figure 3.4). Unless the data uniquely support a particu
not take the resulting model as *the* answer for the issue a
that the particular data set can provide. Perhaps more tha
be considered for inference from the 14 models (Chapter

The program RELEASE (Burnham et al. 1987) allows ap
values of estimators and theoretical standard errors to be
models in this class (i.e., the 14 appearing at the top of Tab
allow insight into why the more minor chronic effects w
model g_{4p} (the model estimated to be the best among the

i	$1 - \mathrm{E}(\hat{S}_i)$	$\widehat{\mathrm{se}}(1 - \hat{S}_i)$	$(1 - \mathrm{E}(\hat{S}_i))/\widehat{\mathrm{se}}(1$
4	0.057	0.053	1.08
5	0.046	0.055	0.84
6	0.036	0.057	0.63

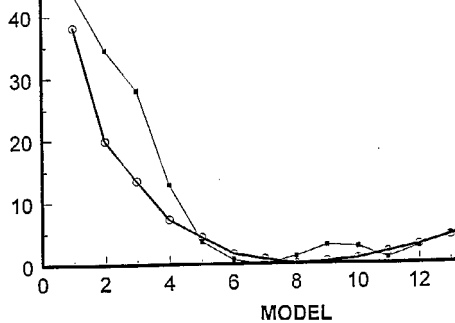


FIGURE 3.4. Estimated theoretical (heavy line) and sample Δ_i v
models used for the starling experiment. The estimated ($n = 50$,
theoretical, expected AIC values (shown as open circles) are minim
g_{5p}, while the realized AIC value from the sample data is minim
Table 3.6). Generally, there is good agreement between the theoret
here plotted as Δ_i values).

The expected treatment effect size (i.e., $1 - \mathrm{E}(\hat{S})$) was s
the standard errors were of a similar magnitude or larger
nal two columns above. The larger effects (i.e., S_1 and S_2
to identify; however, at some point, the effect size is too
rectly with confidence from the information contained i
Still, if one had only the first 14 models and had used A
g_{4p}, inference from the data in this example would have
able, but hardly optimal. The acute and larger chronic effe
convincingly identified. Comparison of AIC values for m
and g_{6p} ($K = 28$) would have provided reasonable evi
tended chronic treatment effects. Still, having to estimat
would lead to imprecise estimators, compared to those u
($g_{\sin\phi_t,\phi_c,\sin p_t,p_c}$). AIC, AIC_c, and QAIC_c are fundamental
a basis for a unified approach to the statistical analysis of
biological sciences. Further details concerning this class
vided by Anderson et al. (1994), Burnham et al. (1994), B
and b), and Anderson et al. (1998).

3.4.6 Hypothesis Testing and Information-Theor Have Different Selection Frequencies

At this point it is illustrative to examine briefly how in
selection compares to traditional approaches based on s
testing. Thus, Monte Carlo methods were employed to g

under model $g_{9\phi}$ with 34 parameters, which was parametrized to ref
effect on both conditional survival and resighting probabilities for

	Model	Hypothesis Testing Stepup	Stepdown	Stepwise	Informat AIC	A
1	g_0	0.6	0.0	0.0	0.0	
2	$g_{1\phi}$	17.4	0.0	13.9	0.0	
3	g_{2p}	14.8	0.3	13.5	0.2	
4	$g_{2\phi}$	26.8	1.5	26.3	1.2	
5	g_{3p}	16.3	2.8	16.9	2.7	
6	$g_{3\phi}$	14.6	6.9	16.1	6.8	
7	g_{4p}	5.9	7.5	7.1	8.5	
8	$g_{4\phi}$	2.7	11.9	3.8	13.5	
9	g_{5p}	0.8	10.3	1.3	12.0	
10	$g_{5\phi}$	0.2	13.3	0.7	14.3	
11	g_{6p}	0.0	10.9	0.2	11.3	
12	$g_{6\phi}$	0.0	12.9	0.2	11.3	
13	g_{7p}	0.0	10.2	0.1	8.8	
14	$g_{7\phi}$	0.0	11.5	0.1	9.5	

pendent samples (data sets) using the same methods as w
the original set of simulated data on nestling starlings. Th
the numbers released and all parameter values were ide
to generate the first set of data. Six methods were used t
inference: The first 3 methods involve well-known selecti
hypothesis testing (stepup or forward selection, stepdown
tion, and stepwise selection), each using $\alpha = 0.05$. Three i
methods were also used on each of the data sets: AIC, A
ing $\hat{c}$ as a variance inflation factor, estimated for each sim
results (Table 3.7 and Figure 3.5) show substantial differ
selection frequencies for the various methods.

Stepup selection, on average, selects model $g_{2\phi}$ with
average was 20.3 parameters selected). These results are
wise approach, which selects, on average, model g_{3p} with
the average number of parameters was 20.6). Given that
under model $g_{9\phi}$ with 34 parameters, these methods seer
simple models that miss most of the chronic treatment eff
ing resulted, on average, in model g_{6p} (mean 26.9 parame
quite different model selection frequencies than the othe
approaches. Of course, the selection frequencies would c
a different (arbitrary) α level (say, 0.15 or 0.01) had bee
treatment impacts differed (i.e., a different model used t

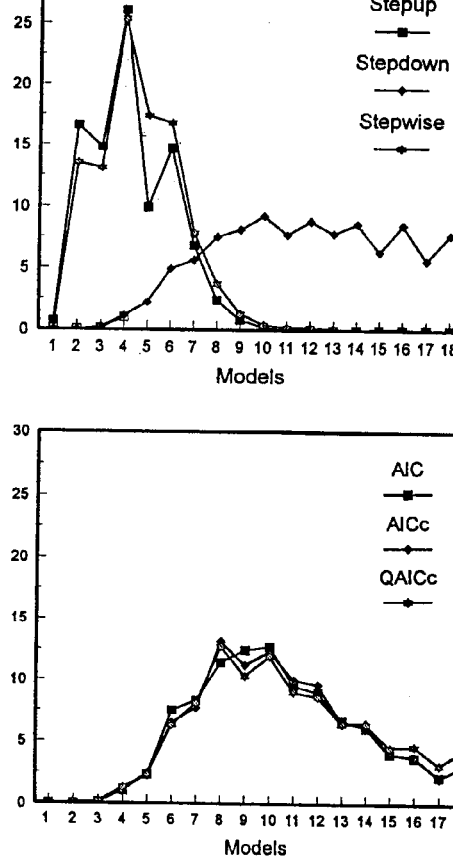


FIGURE 3.5. Model selection probabilities for three hypothesis t
and three information-theoretic approaches based on 50,000 Monte
starling data, generated under model $g_{9\phi}$.

The practical utility of hypothesis testing procedures is
model identification (Akaike 1981b:722).

In this example, AIC selection averaged 25.4 parame
model g_{5p}). Both AIC and the stepdown testing did reaso
ing the larger chronic effects. The differences between AI
are trivial, as one would expect from the large sample siz
ple (Table 3.7 and Figure 3.5). Even the use of QAIC_c r
difference in this example because, the estimated variance
near 1 ($\hat{c} = 1.18$).

In this simulated example. In addition, one must note th
tainty exists in model choice for all six approaches (Figu
the material in Chapter 4 particularly important, since th
certainty should be incorporated into estimates of precisi
estimators. In this example the statistical hypothesis testin
alternative to selection based on estimating the relative
general, we recommend strongly against the use of null h
model selection.

3.4.7 Further Insights Following Final Model Se

Selection of the best model and the relative ranking of all t
is objective, given a set of candidate models, and can be in
the aid of subjective judgment. The formal data-based sea
is a key part of the analysis. In the example, model $g_{9\phi}$ was
data. Thus "truth" is known and serves here as a basis fo
does not try to select the model that generated these data
which model is the "best approximating model" for anal
the sense of having the smallest K-L distance from appr
truth. Further information concerning the statistical proper
models in the $g_0, \ldots, g_{9\phi}$ class are given in Anderson et a

The starling example illustrates an ideal a priori strat
explore some potential realities *after* the analysis has be
point. We select model $g_{\sin\phi_t,\phi_c,\sin p_t,p_c}$ as the best (with 6
must also perhaps consider models

$$g_{\sin\phi_t,\phi_c,p_t,p_c} \qquad (\Delta_i = 5.34 \text{ with only 5 parame}$$
$$g_{\sin\phi_t,\phi_{ci},\sin p_t,p_c} \qquad (\Delta_i = 7.47, \text{ with 14 parameter}$$

before making some final inferences (at this time, th
dress the variance component due to model uncertainty
$g_{\sin\phi_t,\phi_{ci},\sin p_t,p_c}$ is somewhat inconsistent. Why would th
weekly variation in conditional survival for the birds in th
only smooth time trends in the treatment group? Perhap
lead to a thorough review of field methods in an effort to d
Perhaps this model should not have been in the set of ca
sidered, since it may be picking up random variation in the
Δ_i is 7.47, and this model has 8 additional parameters ov
model. It would seem that this model is a relatively poor o
though it might play a role in estimating the variance com
selection uncertainty.

After the analysis of the data to this point, suppose that o
bers asks about models where there is no treatment effe

be added to the set of candidate models and more AIC va
question brings up several points. First, if this suggestion w
ining the estimates of $p_{t2}, p_{t3}, p_{t4}, \ldots, p_{p7}$ vs. p_{c2}, p_{c3}, p_{c4}
that there seemed to be little difference between success
pairs, then this is a form of data dredging, and any subse
clearly detail the process by which the additional models w
**encourage full investigation of the data to gain all po
only want investigators to reveal the extent of any data
place**. Second, if that suggestion was made on conceptual
by studying the intermediate results, then the new class of
to the list, AIC computed, the Δ_i, w_i, and evidence ratio
and inferences made. However, in this second case, the t
what faulted for not considering the set of models more
place.

3.4.8 Why Not Always Use the Global Model for

Some might argue that the global model (or another mode
eters representing likely effects) should always be used fo
After all, this model has been carefully defined by exis
siderations and has all the effects thought to be reasonab
principle of parsimony and endure the various issues conc
tion, selection uncertainty, etc.? We can illustrate problem
using the starling data and the global model ($g_{7\phi}$) with 30
results are given here in detail for estimates of S_i for $i =$

i	$\hat{S}_i$	$\widehat{\text{se}}(\hat{S}_i)$	CI_L	CI_U
1	0.946	0.044	0.858	1.033
2	0.868	0.052	0.766	0.969
3	0.905	0.058	0.792	1.017
4	0.958	0.057	0.847	1.069
5	0.950	0.047	0.859	1.042
6	0.983	0.051	0.883	1.084
7	0.973	0.059	0.858	1.088

The poor precision is illustrated by the upper confidence i
6 of the 7 include the value of 1 (i.e., no treatment effe
limits would be truncated at 1.0). The average coefficien
$\hat{\phi}_{ti}$ under model $g_{7\phi}$ is 4.75% vs. 2.08% under the AIC-s

Attempts to select a properly parsimonious model f
rewards, primarily an approximating model that has a rea
tween bias and variance. The tradeoff between bias and va
of model selection where expected K-L information los

of model parameters that are unnecessarily imprecise (
Section 3.2.2), and this weakens the inferences made. In fa
to show the real patterns that can be validly inferred from
the smooth decrease in ϕ_{ti} and thus in S_i. Sometimes the
have 50, 100, or even 200 parameters, and this makes int
One cannot see patterns and structure, since there are s
most estimated with poor precision. Thus, some analysts
analyses of these estimated parameters in order to "see the
This has rarely been done correctly, since the estimators u
tial sampling correlations, making simple analysis results
better to embed the reduced model in the log-likelihood
information-theoretic criteria to select a simple, *interpre*
to the information in the data.

3.5 Example 4: Sage Grouse Survival

3.5.1 Introduction

Data from sage grouse (*Centrocercus urophasianus*) ba
Colorado, provide insights into hypothesis testing and i
criteria in data analysis. The example is taken from Zabl
tional details are found there. Here we will use data on
than 1 year old) and adult (birds more than 1 year old) ma
leks during the breeding season (first week of March thr
of May), from 1973 through 1987. Sage grouse are hu
nearly all of the band recoveries were from hunters who
banded bird and reported it to the Colorado Division of
time 1,777 subadult and 1,847 adult males were banded,
numbers of band recoveries were 312 and 270, respectiv
basic theory for modeling and estimation for these types
found in Brownie et al. (1985).

Two types of parameters are relevant here: S_i is the
probability relating to the annual period between banding
r_i is the conditional probability of a band from a bird bei
given that the bird died in year i. In the model building it i
as a subscript to denote age (subadult vs. adult) and t to de
(e.g., S_{a*t} denotes survival probabilities that vary by both
some models assume that r_i is a constant, thus resulting in
parameter S_{15}).

Year Banded	Number banded	Recoveries by hunting seas										
		73	74	75	76	77	78	79	80	81	82	8
1973	80	6	4	6	1	0	1	0	0	0	0	
1974	54		6	5	2	1	0	0	0	0	0	
1975	138			18	6	6	2	0	1	0	0	
1976	120				17	5	6	2	1	1	0	
1977	183					20	9	6	2	1	1	
1978	106						14	4	3	1	0	
1979	111							13	4	0	1	
1980	127								13	5	3	
1981	110									13	5	
1982	110										7	
1983	152											
1984	102											
1985	163											
1986	104											
1987	117											

		73	74	75	76	77	78	79	80	81	82	8
1973	99	7	4	1	0	1	0	0	0	0	0	
1974	38		8	5	1	0	0	0	0	0	0	
1975	153			10	4	2	0	1	1	0	0	
1976	114				16	3	2	0	0	0	0	
1977	123					12	3	2	3	0	0	
1978	98						10	9	3	0	0	
1979	146							14	9	3	3	
1980	173								9	5	2	
1991	190									16	5	
1982	190										19	
1983	157											
1984	92											
1985	88											
1986	51											
1987	85											

3.5.2 Set of Candidate Models

The biological objective of this study was to increase u
survival process of sage grouse. Zablan (1993) used mod
global model, with 58 parameters. For the purpose of this
the set of candidate models includes the following:

1	S, r	2	Constant S
2	S_t, r	16	Year-dependent S
3	S_a, r	3	Age-dependent S
4	S_{a+t}, r	17	Age- and year-dependent S, no
5	S_{a*t}, r	31	Age- and year-dependent S, in

Models with r year-dependent (t):

6	S, r_t	16	Constant S
7	S_t, r_t	29	Year-dependent S
8	S_a, r_t	17	Age-dependent S
9	S_{a+t}, r_t	30	Age- and year-dependent S, no
10	S_{a*t}, r_t	44	Age- and year-dependent S, in

Models with r age-dependent (a):

11	S, r_a	3	Constant S
12	S_t, r_a	17	Year-dependent S
13	S_a, r_a	4	Age-dependent S
14	S_{a+t}, r_a	18	Age- and year-dependent S, no
15	S_{a*t}, r_a	32	Age- and year-dependent S, in

Models with interaction terms (denoted by the $a * t$) allo
have its own set of time-dependent parameters. The addit
by "+") exclude interaction terms; e.g., for model S_{a+t}
difference between subadult and adult survival paramete
year estimates of survival probabilities for subadults and a
a logit scale and separated by β_0 (see below). A logit tran
made on S, and age (a) and year (as a dummy variable,
function,

$$S_{a+t} \text{ denotes } \text{logit}(S) = \beta_0 + \beta_1(a) + \beta_2(t_1) + \beta_3(t_2)$$

This approach is similar to logistic regression. Howeve
embedded in the log-likelihood function. Such models a
and can be biologically realistic. Models without interacti
parameters; for example, model S_{a+t}, r_a has 18 paramet
parameters for model S_{a*t}, r_a.

These 15 models plus models $\{S_a, r_{a*t}\}$ and $\{S_a, r_{a+t}\}$ a
$\{S_{a*t}, r_{a*t}\}$ seem like sound initial choices; however, furth
erations might lead one to exclude models with many pa
the relatively sparse data available (Table 3.8). We realize
of models would ideally be fine-tuned in a real-world appli
if a long-term increase or decrease in survival was hypo
introduce submodels for survival such as

$$\text{logit}(S) = \beta_0 + \beta_1(a) + \beta_2(T) + \beta_3(a *$$

survival probability in these years (say, $S_{\$}$) could have s
differentiate them from the survival probability in more n
assume that a great deal of thought has been put into the
of candidate models.

A primary interest in banding studies is often to estima
ities and assess what external covariates might influence
(1993) modeled sage grouse survival using 4 year-depe
covariates (cov_t): winter precipitation (wp), winter temp
precipitation (sp), and spring temperature (st) (she provi
initions of these variables; we will not need to note the s
Submodels with survival probabilities of the form

$$\text{logit}(S_t) = \beta_0 + \beta_1(\text{cov}_t) \quad \text{or} \quad \text{logit}(S_{t+a}) = \beta_0 +$$

could be constructed. Such submodels for survival have o
(an intercept and one slope coefficient for the first submo
1 age effect, and 1 slope coefficient for the second submo
some insights into biological correlates, which themselves

3.5.3 Model Selection

Zablan's analysis was done using the programs ESTIMA
(Brownie et al. 1985) and SURVIV (White 1983). Zabla
the global model $\{S_{a*t}, r_{a*t}\}$ fit the data well ($\chi^2 = 34.34$
and she computed a variance inflation factor from the g
$34.34/30 = 1.14$. Her calculations are in agreement wi
model $\{S_{a*t}, r_{a*t}\} = 87.85$, 80 df, $\hat{c} = 1.10$). There w
overdispersion; thus there was no compelling reason to use
sample size for parameter estimation in these surveys is th
of birds banded, which equaled 3,624 in this case. Zablan
58 parameters, giving the ratio $n/K = 3{,}624/58 = 62$; t
been safely used instead of AIC_c. We used the program
Burnham 1999, White et al. 2001) to compute MLEs of
their conditional covariance matrix, the maximized value
function, AIC_c, Δ_i, and w_i for each of the 17 candidat
weather covariate and 4 models with one of the weather c

AIC_c selected model $\{S_a, r_a\}$ with 4 parameters (Ta
models without a weather covariate on survival. This appr
sumes that conditional survival and reporting probabilitie
(subadult vs. adult), but constant over years. Here, the M
survival probability was 0.407 ($\widehat{\text{se}} = 0.021$), while the es
subadults was higher, at 0.547 ($\widehat{\text{se}} = 0.055$). The respe
cients of variation were 5.2 and 10.0. An inference here i
grouse survive at a higher rate than male adults; perhaps

Number	Model	log($\mathcal{L}$)	K	AIC$_c$	
Without environmental covariates:					
1	S, r	−2, 215.564	2	4435.13	
2	S_t, r	−2, 205.074	16	4442.30	
3	S_a, r	−2, 215.096	3	4436.20	
4	S_{a+t}, r	−2, 203.797	17	4441.76	
5	S_{a*t}, r	−2, 199.277	31	4461.11	1
6	S, r_t	−2, 204.893	16	4441.94	
7	S_t, r_t	−2, 194.611	29	4447.71	
8	S_a, r_t	−2, 204.526	17	4443.22	
9	S_{a+t}, r_t	−2, 193.633	30	4447.84	
10	S_{a*t}, r_t	−2, 188.531	44	4466.17	1
11	S, r_a	−2, 214.717	3	4435.44	
12	S_t, r_a	−2, 204.544	17	4443.26	
13	$\boldsymbol{S_a, r_a}$	**−2, 211.357**	**4**	**4430.72**	
14	S_{a+t}, r_a	−2, 204.544	18	4439.96	
15	S_{a*t}, r_a	−2, 196.065	32	4456.72	1
16	S_a, r_{a*t}	−2, 197.572	32	4459.73	1
17	S_{a*t}, r_{a*t}	−2, 174.557	58	4467.03	1

of breeding and increased predation on breeding males. probabilities (the $\hat{r}$) for first-year subadult birds were also for adult birds (0.227 ($\widehat{\text{se}} = 0.031$) and 0.151 ($\widehat{\text{se}} = 0$ The use of the AIC$_c$-selected model does not indicate tha dependent variation in the parameters, only that this var small in the sense of a bias–variance tradeoff and K-L in estimated next-best model ($\Delta_i = 4.41$) without a covaria with only two parameters, while the third-best model (Δ_i with three parameters. There is relatively little structure re this is not surprising, since the data are somewhat sparse

The AIC-selected model assumed that the survival and ties varied by age class. Thus we considered four models a linear function of age (subadult vs. adult) and one of th (wp, wt, sp, or st), while retaining the age-specific repor results were interesting and, at first, suggest that each of t models is nearly tied with the AIC$_c$-selected model {S_a, important point, illustrated below:

Number	Model	log($\mathcal{L}$)	K	AIC$_c$
With environmental covariates:				
18	S_{a+wp}, r_a	−2, 210.828	5	4431.6
19	S_{a+wt}, r_a	−2, 211.334	5	4432.6
20	S_{a+sp}, r_a	−2, 210.819	5	4431.6
21	S_{a+st}, r_a	−2, 210.802	5	4431.6

probabilities are related to temperature or precipitation. H
examination, it can be seen that the value of the maximiz
very similar to the best model in Table 3.9, without any c
inclusion of models with a covariate has not improved
to the data. The best model without a covariate has 4 p
the covariate models have 5 parameters. This differenc
parameters explains most of the difference in the AIC_c val
model and the 4 models with a covariate. Hence, upon
there is virtually no support for any of the covariates fr
leads to a point that is important in general.

Models Within Two Units of the Best M

Models having Δ_i within about 0–2 units of the bes
examined to see whether they differ from the best model
have essentially the same values of the maximized log-li
model.

In this case, the larger model is not really supported
rather is "close" only because it adds 1 parameter and ther
2 Δ_i units, even though the fit, as measured by the log-
not improved.

Further insights into the sage grouse data can be obtai
weights for the first 17 models in Table 3.9 (ignoring her
this case, the weight for the AIC_{min} model $\{S_a, r_a\}$ is 0.77
best model $\{S, r\}$ has a weight of 0.085 (evidence ratio
and fourth-best models had weights of 0.073 and 0.050, w
the other models were nearly zero (the sum of the Akaik
remaining models was < 0.02). The annual variation in
probabilities was small (temporal process variation $\hat{\sigma}_s = 0$
0.0279 for subadults); thus model $\{S_a, r_a\}$ seems reasonabl
4 all had fewer parameters than the AIC_{min} model. Thus, c
variances from those models were smaller than from th
addition, these three models had small Akaike weights. T
lead to some trust in the conditional sampling variances
as a reasonable reflection of the precision of the paramete

3.5.4 Hypothesis Tests for Year-Dependent Surv

Zablan (1993) computed a likelihood ratio test between r
$\{S_{a*t}, r_{a*t}\}$ (the global model) using the program BROWN
evidence of year-dependent survival ($\chi^2 = 46.78$, 26 df, P
MARK provides similar results ($\chi^2 = 46.03$, 26 df, P
allowed a fairly general structure on the reporting probab
seemed convincing and provided evidence that survival

for year-dependent survival probabilities (i.e., S_t vs. S_f) w
and the test result been even more "significant." However,
of reasoning, the test of $\{S_a, r_a\}$ vs. $\{S_{a*t}, r_a\}$ gives χ^2 =
$P = 0.336$. Still other testing strategies are possible, and
might be deemed the best.

Given a believed year-dependence in annual survival p
(1993) asked whether this variability was partially explain
covariates, with or without an age effect. However, she wa
lationship between annual survival probabilities and any c
using likelihood ratio tests (the smallest P-value for the
0.194). She used model $\{S_{a*t}, r_{a*t}\}$ (the global model) as a

3.5.5 Hypothesis Testing Versus AIC in Model S

An apparent paradox can be seen in the results for the male s
this allows us to further compare alternative paradigms of s
testing and AIC for model selection. The test between the t
and $\{S_{a*t}, r_{a*t}\}$ attempts to answer the questions, "Given
on the recovery probabilities, is there evidence that surv
addition to age) year-dependent?" The answer provided
or 0.009). But this answer is seemingly in contrast to the
AIC-selected model, where there is no hint of time-depend
(Table 3.9). The Δ_i values for models $\{S_a, r_{a*t}\}$ and $\{S_{a*t},$
30.400, respectively. AIC lends little support for a best a
that includes year-dependent survival or reporting probab

The answer to this paradox is interesting and importan
null hypothesis that $S_1 = S_2 = S_3 = \cdots = S_{14}$ for
obviously false, so why test it? This is not properly a hyp
(see Johnson 1995 and Yoccoz 1991 for related issues). Th

Model Interpretation

Sometimes, the selected model contains a parameter th
time, or areas, or age classes (i.e., $\theta = \theta_1 = \theta_2 = \cdots$
should not imply that there is *no* variation in this paramet
mony and its bias/variance tradeoff finds the actual variati
to be relatively small in relation to the information conta
data. It "costs" too much in lost precision to add estima
dividual θ_i. As the sample size increases, then at some
estimates of the individual parameters would likely be fa
Just because a parsimonious model contains a parame
across strata does not mean that there is no variation in
the strata.

not relate directly to the issue of what approximating mo
inference. One model selection strategy that has often b
is to do likelihood ratio tests of each structural factor (e.
for each of the parameters S and r) and then use a mode
that were "significant" at, say, $\alpha = 0.05$. However, the
would suggest that this strategy would lead to a model v
properties (i.e., small bias, good precision, and achieved
coverage at the nominal level).

Clearly, one must also worry about the multiple testin
the fact that many such tests would not be independent.
present, then likelihood ratio tests are not chi-square distri
the test statistics may not be chi-square distributed for
such as these, where sample sizes are far from asymptot
where many models contain nuisance parameters (the r_i
α-level is arbitrary as well. Many of the models in Table
thus likelihood ratio tests are not possible between these r
a certain lack of symmetry (this is again related to the d
null and alternative hypotheses and how this might rela
"best approximating model" (see Section 2.7.2). A very g
problem here is how the test results are to be incorporated
model for statistical inference. This problem becomes a
many (say, $>$ 8–10) candidate models. Using just the set
sage grouse data, one would have 136 potential likelihood
some of these models were not nested, prohibiting a test
With 136 (or even 36) test results there is no theory or u
what the best model should be and no rigorous, general wa
(e.g., which model is second-best? Is the second-best mo
as the best, or substantially inferior?). Finally, what is t
results are inconsistent, such as those found in Section 3.

The biological question regarding annual survival proba
be stated as, "How much did annual survival vary during
Has survival varied little over the 14 or 15 years, or h
variation in these annual parameters? Such questions are
not ones of hypothesis testing (see Franklin et al. 2002)
inquiry should be on the amount of variation among the po
$(S_1, S_2, \ldots, S_{14})$ for each of the two age classes; we will
deviation among these parameters by σ_s. Of course, if we
S_i, then $\hat{\sigma}_s = \left(\sum_{i=1}^{14}(S_i - \overline{S})^2/13\right)^{1/2}$.

We next ask why the AIC procedure did not pick up th
varied by year? The reason is simple; AIC attempts to se
approximating model for the observed data. In the sense
loss or a tradeoff between bias and variance, it was poor

els $\{S_a, r_{a*t}\}$ and $\{S_{a*t}, r_{a*t}\}$) to model the variation in S_t o
that the difference in AIC values for model $\{S_a, r_a\}$ versu
29.01, suggesting that model $\{S_a, r_{a*t}\}$ is highly overfit.
among survival probabilities are large enough to be inclu
model selection problem, not one of hypothesis testing.

Estimates of the 30 survival probabilities under mode
in Table 3.10. The average of the 15 estimates of adult
nearly the same as that from the 4-parameter model selec
However, the average percent coefficient of variation for
model $\{S_{a*t}, r_a\}$ compared to only 5.2 for $\hat{S}$ in the AIC_c-se
the AIC-selected model indicates that the best estimate of
particular year is merely $\hat{S}$ (from model $\{S_a, r_a\}$).

The situation was similar for subadult survival; the av
model $\{S_{a*t}, r_a\}$ was 0.548, compared to 0.547 for the A
The respective average percent coefficients of variation w
for models $\{S_{a*t}, r_a\}$ and $\{S_a, r_a\}$. In summary, 54 (or eve
rameters "cost too much" in terms of increased variability
Figure 1.3B and Table 3.10) and reflect substantial overfitt
cision illustrated in Table 3.9 for model $\{S_{a*t}, r_a\}$ was wo
$\{S_{a*t}, r_{a*t}\}$ was used; coefficients of variation were 35.6% f
30.5% for subadult survival. The model suggested by th
approach had 58 parameters, while the AIC_c-selected mod
eters. This illustrates the cost of using overparametrized
the results of hypothesis tests clearly show "significance"
survival (and reporting) probabilities. Models $\{S_a, r_{a*t}\}$ an
general for these data and lie far to the right of the bias–var
in Figure 1.2. Zablan recognized the problems in using m
commented, "While significant differences were found b
recovery rates of males and of both age classes, and betw
estimates had unacceptably wide confidence intervals."

3.5.6 A Class of Intermediate Models

The researcher could use the AIC_c-selected model $\{S_a, r_a$
of parameters and then proceed to obtain an estimate σ_s
age classes, using model $\{S_{a*t}, r_{a*t}\}$ or $\{S_{a*t}, r_{a+t}\}$ if desir
embed this parameter into the likelihood framework, allow
of σ_s; this becomes a "random effects" model. However,
"variance components," and consistent estimates of σ_s can
say, model $\{S_{a*t}, r_a\}$, following, for example, Anderson a
62–66); and Burnham et al. (1987:260–269). This appro
based on the simple partitioning of the total variance (

to first-year survival of subadults. The model assumes that the subad
second year after banding and thus have the same year-dependent s
birds banded as adults.

Year(i)	$\hat{S}_i$	Standard Error	95% Confidence I Lower	U
1	0.462	0.128	0.238	0
2	0.500	0.092	0.327	0
3	0.357	0.073	0.230	0
4	0.412	0.074	0.277	0
5	0.464	0.073	0.328	0
6	0.507	0.069	0.375	0
7	0.465	0.066	0.340	0
8	0.357	0.062	0.246	0
9	0.397	0.063	0.282	0
10	0.340	0.061	0.233	0
11	0.321	0.063	0.212	0
12	0.358	0.073	0.231	0
13	0.171	0.071	0.071	0
14	0.339	0.133	0.138	0
15	0.549	0.129	0.305	0
1	0.725	0.114	0.462	0
2	0.629	0.152	0.321	0
3	0.524	0.106	0.323	0
4	0.528	0.112	0.316	0
5	0.566	0.093	0.383	0
6	0.446	0.120	0.237	0
7	0.386	0.117	0.193	0
8	0.513	0.110	0.307	0
9	0.497	0.118	0.282	0
10	0.615	0.111	0.389	0
11	0.547	0.101	0.351	0
12	0.368	0.121	0.173	0
13	0.440	0.107	0.251	0
14	0.744	0.104	0.498	0
15	0.695	0.111	0.450	0

[a]Based on a back transformation of the interval endpoints on a logit scale (

additive components: the variance in the population par
conditional sampling variance (var($\hat{S}_i \mid model$)). The app
the true S_i are independently and identically distributed r
this case, these assumptions are weak and the effect somew
one has ML estimates of the sampling covariance matr
var($\hat{S}_i$) directly from the estimates $\hat{S}_i$; thus, by subtractio

The estimate of σ_s would provide some insight into
survival parameters; this would be done in the context
the S_i vary. Exact details of the optimal methodology w
afield; however, some unpublished results seem exciting
$\hat{\sigma}_s = 0.0426$, 95% likelihood interval [0, 0.106], and cv
survival probability and for the subadult data $\hat{\sigma}_s = 0.0$
interval [0, 0.129], and cv = 4.9% on S. Thus, one can i
variation in the true annual survival probabilities was fa
10%). Thus, the large variation in the estimates of annual s
under model $\{S_{a*t}, r_a\}$ (Table 3.9) is due primarily to samp
large estimated standard errors suggest. Additional details,
estimates of annual survival probabilities, appear in Chap
and White 2002).

Ideally, the number of parameters in the various candida
have large increments (see Section 2.7.2). In the grouse mo
the survival probabilities without year-dependent surviva
two (if age is included) parameters, while a model with yea
would have as many as 30 parameters (15 for each of the tw
differences in the number of parameters between certain c
not ideal, and one should consider intermediate models w
of candidate models. Zablan's (1993) various covariate
example; here an intercept and slope parameter on one of
introduce 2 parameters (3 with age) instead of 15 (30 w
with the hypothesis testing approach, AIC-selection show
covariate models to be essentially tied with the AIC-sel
(Table 3.9).

Hypothesis testing and AIC are fundamentally very di
model selection and in drawing inferences from a set of da
example, AIC_c tries to select a model that well approxim
in the data. That selected model then provides estimate
$S_1, S_2, \ldots, S_{15}$ for each age group in the sense of K-L in
bias versus variance tradeoff). That is, an estimate of avera
model $\{S_a, r_a\}$ (i.e., 0.407 for adults) would be used to estim
for adult grouse (hence, $\hat{S}_5 = 0.407$, 95% confidence inte
and this estimate would have better inferential properties t
$\{S_{a*t}, r_a\}$, whereby S_5 would be estimated using the year-
(see Table 3.10, where this estimate is given as 0.464 w
interval of 0.328 to 0.606). If inference about the condit
fifth year is made from the general model $\{S_{a*t}, r_{a*t}\}$, then t
and the precision is worse yet (95% confidence interval o
the last two cases, the precision is relatively poor (e.g., co
and C for further insights).

This example illustrates the use of information-theoretic
sis of count data displayed as multidimensional continge
Sakamoto 1982). Schoener (1970) studied resource utili
lizards of the genus *Anolis* on several islands in the Le
Caribbean. Here we use his data collected on *Anolis grah*
near Whitehouse, on Jamaica, as provided by Bishop et a
have been analyzed by Fienberg (1970), Bishop et al. (19
and Nelder (1989) and Qian et al. (1996), and the reader
the approaches given in these papers to that presented he

In his general studies of species overlap, Schoener (19
species of lizards in an area of trees and shrubs that had bee
The height (< 5 or ≥ 5 ft) and diameter (< 2 or ≥ 2 in) of
(sunny or shaded), and time of day (roughly early morni
afternoon) were recorded for each of the two species of li
taken on each individual only once per "census"; data we
lizard was disturbed, and the census route was varied co
observation period to the next. The data of interest for
summarized as a $2 \times 2 \times 2 \times 3 \times 2$ contingency table cor
(H), diameter (D), insolation (I), time of day (T), and sp
Table 3.11 (from McCullagh and Nelder 1989:128–135
observations of lizards appear in the table with 48 cell
here differ from those in the first edition due to errors (b
computing AIC_c.]

TABLE 3.11. Contingency table of site preference for two species o
and *A. opalinus* (denoted by g and o, respectively) on the island of J
1970).

			Time of Day (T);			
Insolation	Diameter	Height	Early morning		Midday	
(I)	(D) in	(H) ft	g	o	g	
Sunny	≤ 2	< 5	20	2	8	
		≥ 5	13	0	8	
	> 2	< 5	8	3	4	
		≥ 5	6	0	0	
Shaded	≤ 2	< 5	34	11	69	2
		≥ 5	31	5	55	
	> 2	< 5	17	15	60	3
		≥ 5	12	1	21	

pertise in lizard ecology and behavior, decisions concer
of candidate models will necessarily be somewhat contr
attention on modeling and model selection issues as an ex
on several inference issues. In reading Schoener's (1970)
ature he cited, it would seem that a model with all the r
D, I, T, and S) might serve as a starting point for mode
Several second-order interactions might be suspected, e.
and $I * T$. If the two species are partitioning their resource
$H * S$, $D * S$, $I * S$, and $T * S$ included should be reasonab
designed and data were collected, it was probably evident
was affected by several variables as well as some interacti
gest that a model with all main effects and second-order i
considered. Then issues remain concerning possible highe
On biological grounds, it might seem reasonable to consid
such as $H * D * I$, $H * D * T$, and $H * D * S$; or furthe
$H * T * S$, and $I * T * S$. Finally, the second-order term D
to be important; thus some models without this term were
use the short set of models in Table 3.12 for illustrative
in the late 1960s, T. W. Schoener and his colleagues, inclu
could have developed a better set of a priori candidate mo

3.6.2 Comments on Analytic Method

We used a loglinear model with Poisson errors following A
5 and pages 453–456), and the analysis was made conditio
frequencies, $\sum n_j$. Specifically, we used the SAS progra
1985). The likelihood is, $\mathcal{L}(\mu|n_j, model) = \prod_{j=1}^{48} \frac{\mu^{n_j} e^{-}}{(n_j)}$
log-likelihood for the global model can be expressed as

$$\log(\mathcal{L}) = \sum_{j=1}^{48}\left(n_j \cdot \log(\mu_j) - \mu_j\right) - \sum_{j=1}^{48} \log$$

where n_j is the number of observations in cell j with Pois
$j = 1, 2, \ldots, 48$. The purpose of the modeling is to put so
on the 48 means. Then one has the log-linear model lo
analysis of variance. Thus, β is the vector of effects and t
final term in the log-likelihood is a constant; thus SAS C
term, and the resulting log-likelihood is positive. AIC is co
manner, even though the AIC values are scaled by $\sum_{j=}^{48}$
such arbitrary, additive constants are not present in the
software packages allow ML estimates from discrete da
provide a number of relevant analysis options (see summ

if overdispersion is thought to exist in count data. We fi
comment, "In practice, we learn more from *estimating* de
than from *testing hypotheses* about their values."

3.6.3 Some Tentative Results

A model with only the five main effects might be a starti
the information in these data. Alternative models with v
interaction terms are shown in Table 3.12. None of the r
or third-order interactions are supported by the data; in
with any support is the model with merely the main e
in agreement with McCullagh and Nelder (1989), but d
Fienburg (1970) and Bishop et al. (1975).

Note that if another model were to be added, the Δ_i valu
to be recomputed (but not the AIC values). Such values are
to the minimum AIC model, given a set of candidate mo
derived from the data via a set of models are effectivel
set of models considered. Strict experimentation might be
additional insights into the issue of resource utilization in

If our original interest had been only on differences in
two species, we could examine a model with all five mai
S) vs. a similar model without a species effect (H, D, I, T
have been only 2 candidate models considered. The dif
two models is 163.70, indicating strongly that the two spe
resources differently. Alternatively, we could make a si
two models, but also include the relevant second-order in

H, D, I, T, S, H*D, H*I, H*T, H*S, D*I, D*T, D*S

vs.

H, D, I, T, H*D, H*I, H*T, D*I, D*T, I

Here, the difference Δ_i is 84.63 and again clearly inc
species are utilizing their habitat differently. Other altern
be pursued if more were known about the study design a

Bishop et al. (1975) performed a number of hypothesis t
and this resulted in a model with all the main effects pl
interaction terms H∗D, H∗S, D∗S, T∗S and the third-orde
study is considered to be merely exploratory, then one mi
wider class of candidate models, and many models with
found. However, the number of possible models would be
2,000, depending on what rules might be applied concer
lower-order effects if higher-order effects are included
with powerful computing equipment, an exhaustive stu

The model with the minimum AIC_c is shown in bold.

Model		log($\mathcal{L}$)	K	AIC
1	**All main effects, H D I T S**	**1,181.08**	**7**	**−2,34**
2	All main effects and second-order interactions	1,181.86	21	−2, 31
3	Base[(1)] but drop DT	1,180.52	19	−2, 32
4	Base[(1)] plus HDI, HDT and HDS terms	1,182.97	25	−2, 31
5	Base[(1)] plus HDI, HDS, HIT, HIS, HTS, and ITS	1185.75	30	−2, 30
6	Base[(1)] plus HIT, HIS, HTS, and ITS	1185.48	28	−2, 31
7	Base[(1)] plus HIS, HTS, and ITS	1184.01	26	−2, 31
8	Base[(1)] plus HIT, HIS, HTS, and ITS, but drop DT	1184.15	26	−2, 31
9	Base[(1)] plus HIT, HIS, and ITS, but drop DT	1183.29	24	−2, 31
10	Base[(1)] plus HIT and HIS, but drop DT	1182.18	22	−2, 31

[(1)] "Base" is a model with the five main effects plus all second-order intera

this $2 \times 2 \times 2 \times 3 \times 2$ table is nearly prohibitive (see Agr illustrates again the importance of a set of good a priori study is somewhat exploratory. Of course, data dredging co in finding a model that "fits" *these* data. However, the g usually stress an inference about the population or proc description. Thus, the results from intensive data dredgi as tenuous.

The rigorous analysis of multidimensional contingency lematic because of the large number of possible models. G (1978:19) suggest that conclusions drawn from continge only exploratory. While a good a priori set of models seem difficult to forecast higher-order interactions in many situ problematic issues are associated with contingency table lytic approaches exist (Manly et al. 1993), and alternative besides a log link and Poisson errors can be considered (s 1989 for additional information).

Solomon Kullback was born in 1907 in Brooklyn, New York. He g
College of New York in 1927, received an M.A. degree in mathem
pleted a Ph.D. in mathematics at the George Washington Universi
was known to all who knew him, had two major careers: one in t
(1930–1962) and the other in the Department of Statistics at George
(1962–1972). He was chairman of the Statistics Department from 1
professional life was spent in the National Security Agency, and r
this time is still classified. Clearly, most of his studies on informatio
ing this time. Many of his results up to 1958 were published in his
Theory and Statistics. Additional details on Kullback may be foun
and Anonymous (1997).

3.7 Example 6: Sakamoto et al.'s (1986)

Here we return briefly to the example used in Section 1.4
al.ś (1986) book. Ten data sets (each with $n = 21$) were
simple model

$$\mathbf{y = e^{(x-0.3)^2} - 1 + \epsilon.}$$

Sakamoto et al. used the simple polynomials from orde
candidate models to visually illustrate the concepts of u
(see Figure 1.4). Because the sample size ($n = 21$) is sm
dimension of the largest model in the set ($K = 7$), AIC_c sl
analysis of data in this example. AIC_c was computed fo
sets and then averaged for each of the 7 models. Then Δ_i
derived from these averages:

Quadratic	4	−29.63	2.88	0
Cubic	5	−26.80	5.71	0
4th-order	6	−23.79	8.72	0
5th-order	7	−19.26	13.25	0
$\min f(x) = 0.3$	**2**	**−32.51**	**0.0**	**0**

Clearly, the model based on the additional (hopefully a
that $f(x)$ is 0 at $x = 0.3$ is the best of the set (compa
results with the plots in Figure 1.4). The Akaike weigh
sharpen the inference and suggest that only the quadratic
is a competitor to the special model for these simulated
ratio between the best and second-best models is 0.77/0
this ratio between the best and third-best model is only
These comparisons are in line with the visual images in F
reinforce understanding of the information-theoretic quan
ample. More complex data and models defy simple plots a
thus K-L information and various information criteria be

3.8 Example 7: Models of Fish Growth

Shono (2000) presented a reanalysis of data on the gro
salmon (*Oncorhynchus masou*) from Kiso et al. (1992)
sented a comparison of model selection under AIC, AIC_c,
criterion, Schwarz 1978); we will focus on model selec
inferences made by using Δ_i, w_i, and evidence ratios.

Kiso et al. (1992) estimated the parameters of three sta
using length and age data over a period of 2 to 19 month
brief review of the information; further details can be found
and, particularly, in Shono (2000). The sample size was
but was large relative to the number of parameters in the h
model (where $K = 5$). The models for length $L(t)$ as a
months, were

$$\begin{aligned}
&\text{von Bertalanffy} && L(t) = L_\infty[1 - \exp\{-\kappa(t - t_\\
&\text{Gompertz} && L(t) = L_\infty \exp[-\exp\{-\kappa(t - \\
&\text{Logistic} && L(t) = \frac{L_\infty}{1 - \exp\{-\kappa(t - t_0)\}}.
\end{aligned}$$

The basic model parameters are L_∞ and κ. Each model
the exponent, and this was itself modeled in four ways t
concerning the seasonal pattern in growth. These submod

$$\text{Type 2} \quad F(t) = t + \frac{\theta_1}{2\pi}\sin 2\pi(t - t_1) + \frac{\theta_2}{4\pi}\sin 4$$

$$\text{Type 3} \quad F(t) = t + \frac{\theta_1}{2\pi}\sin 2\pi(t - t_1) + \frac{\theta_3}{6\pi}\sin 6$$

The unknown parameters in these submodels are θ_1, θ_2, a case, $\theta_i \geq 0$. The θ_i are the amplitudes of sine curves with and three cycles/year, respectively, and t_1 is the starting curves. They define Type 1 as a modified type where the once per year. Type 2 allows growth rate changes in a co cycles per year, whereas Type 3 allows growth rate chang of 1 and 3 cycles per year (Kiso et al. 1992:1780).

Each of the three growth models included four subm giving a total of 12 models. The results, taken partially 2, are given in Table 3.13. As shown by Shono (2000) the von Bertalanffy with the Type 1 seasonal effect. Hov the Δ_i values, we note that the Gompertz model with T essentially tied ($\Delta = 0.72$); the evidence ratio is $0.27/0.$ the Gompertz Type 3 ($\Delta = 1.13$) and the Gompertz Typ also close competitors. Three of the best four models in form. Even the worst of the top four models is still quit ratio with the best model is $0.27/0.13 = 2.1$. Clearly, t model, relative to the other types. In fact, all nine models seasonal pattern have Δ_i values < 6, suggesting that som important in salmon growth. Beyond this, there is consider uncertainty. This is a clear case where inference based only model is risky. Inference, including prediction, should pro twelve models or, at least, the nine models allowing seas addition, estimates of precision should allow for the high selection. These are subjects treated in Chapter 4.

3.9 Summary

The purpose of the analysis of empirical data is not to find not at all. Instead, we wish to find a best approximating data, and then develop statistical inferences from this mod model is *the* inference from the available data. We search, model," but rather for a *parsimonious model* giving an acc to the interpretable information in the data at hand. Data a question, "What level of model complexity will the data under- and overfitting are to be avoided. Larger data sets t

1992).

Formulae	Type	K	AIC_c	Δ_i
von Bertalanffy	Basic	2	135.82	117.26
Gompertz	Basic	2	50.56	32.00
Logistic	Basic	2	54.05	35.49
von Bertalanffy	1	4	18.56	0.00
Gompertz	1	4	19.28	0.72
Logistic	1	4	21.02	2.46
von Bertalanffy	2	5	21.03	2.47
Gompertz	2	5	20.08	1.52
Logistic	2	5	24.52	5.96
von Bertalanffy	3	5	21.30	2.74
Gompertz	3	5	19.69	1.13
Logistic	3	5	24.25	5.69

complex models, and the selection of the size of the model
between bias and variance.

The analysis of data under the information-theoretic app
simple. That is, the computational aspects are simple, and t
understand and interpret if one has the value of the maxi
($\log(\mathcal{L})$) or the residual sum of squares (RSS) for each mo
quantities are routinely printed by nearly any commercia
ware. Computation of AIC, AIC_c, or $QAIC_c$ from eithe
simple to the point that it can easily be done by hand. Si
of the differences (Δ_i), Akaike weights (w_i), and evidence
ial. The general approach is flexible enough to be used in
of practical situations in the life sciences. These are all a
back to Kullback–Leibler information and have a deep th
easy part of the information-theoretic approaches include
tional aspects and the clear understanding of these resul
evidence).

The hard part, and the one where training has been so
thinking about the science of the matter before data analys
collection. It has been too easy to collect data on a large nu
the hope that a fast computer and sophisticated software wi
tant things—the "significant" ones (the "just the numbers"
a major effort should be mounted to understand the natu
critical examination of the literature, talking with others
eral problem, and thinking deeply about alternative hypo
"test" dozens of trivial matters (is the correlation zero? is
treatment zero? are ravens pink?, Anderson et al. 2000), t
concerted effort to provide evidence on *meaningful* questio

Richard Arthur Leibler was born in Chicago, Illinois, on March
bachelor's and master's degrees in mathematics from Northwestern
in mathematics from the University of Illinois and Purdue Universi
in the Navy during WWII, he was a member of the Institute for
member of the von Neumann Computer Project 1946–1948. From
for the National Security Agency (1948–1958 and 1977–1980) an
Research Division of the Institute for Defense Analysis (1958-1977
Data Handling Inc., a consulting firm for the Intelligence Commun

to a discipline. This is the critical point: the common failu
tant science questions in a fully competent fashion. Think
synthesizing, and challenging the "known" lead to fresh hy
matical models to carefully reflect these alternative hypoth
science is likely to advance more quickly if the "hard part"
theoretic approaches are given much more weight. W
Bayesians might "second" the cry for more a priori thinkin
analysis begins. A rereading of Platt (1964) and delving int
often be good starting points in better understanding this pl

Data from the simulated starling experiment and the ban
ulation both illustrate moderate complexity. In both examp
on the global model would have been (needlessly) poor, i
cult to interpret. The global model for the sage grouse data
and the resulting MLEs had wide confidence intervals (Ta
of substantial sampling correlations (e.g., $\widehat{\text{corr}}(\hat{S}_i, \hat{S}_{i+1})$
Brownie et al. 1985), making it difficult to examine patter

The compartment models of BURSBAN—in a simple
of first-order differential equations, and they represent a hi
plexity. The real data on cement hardening and fish grov
examples but provide insights into the interpretation of e
sion of the analysis of the simulated data from Sakamoto
some comparisons with the graphical material in Section
of resource partitioning in *anolis* lizards must be consider
Burnham et al. (1996) give a comprehensive example de
survival probabilities of the Northern Spotted Owl (*Strix*
caused by widespread clear-cutting in national forests. Th
a large data set over eleven geographic areas in a politica
national importance.

Researchers often attempt to perform some further, sep
multiple linear or logistic regression) of the estimates f
model in an effort to understand the structure of the proc
into its behavior. However, such external analyses are not e
the estimates for the starling and grouse data have a mult
variance structure; the variances are unequal (not consta
estimates are dependent (not independent). When the ana
such further "external" analysis of the parameter estimate
is clear that a properly parsimonious, easily interpretable
achieved, and hence the analysis has partially failed.

We recommend carefully developing a set of candidat
the science of the issue (e.g., embed additive submodels fo
ities into the log-likelihood using an appropriate link func
MLEs under these models. Then one can focus on model
a properly parsimonious model(s) (the models in the set o
"close" to truth in the K-L information sense) that will
inference. In contrast, if only the high-dimensional globa
and estimates of parameters obtained by ML or LS, ther
analysis is virtually defeated, because a parsimonious int
may be impossible, patterns often cannot be found, and es
precise. Zablan (1993) used a global model with 58 param
of the sage grouse data and correctly observed, "... su
unacceptably wide confidence intervals."

Statistical analysis of empirical data should not be jus
given only a set of data (the numbers). The cement hard
3.2) have too frequently been analyzed without examinin
treatment of the data: the important a priori considerations
here. What was known about the chemistry of cement har
1930s, then at least by the 1960s)? For example, can ce
harden well with only a single ingredient? If not, this mig
single-variable models out of consideration. Given the ch

variables x_1 and x_3 is minimal. Surely, these conditions wer
could have affected the models in the set to be considered.
have known, or bothered to find out, that $\sum_{j=1}^{4} x_{ij}$ is a
the thirteen observations and therefore excluded the four-
consideration, based on these a priori grounds? Instead, un
have been the modus operandi, and "all possible models" h
tried. "Let the computer find out" is a poor strategy for res
bother to think clearly about the problem of interest and
The sterile analysis of "just the numbers" will continue t
for progress in the sciences.

Researchers often resort to using a computer progra
all possible models and variables automatically. Here,
computer will discover the important variables and relat
numbers approach" void of any thinking or science). Co
conducted stepwise linear regression analyses using AI
(SAS Institute 1988:786) in a study of elk (*Cervus ca*
They found that this approach "... provided results that
cally unrealistic, unstable due to multicolinearity, and ove
variables)." The literature is full of such failed studies; ju
used as a selection criterion does not mean that valid infere
The primary mistake here is a common one: the failure to
a priori models, each representing a *plausible* research hy

The presentation of results in scientific publications sh
used in arriving at a set of candidate models. The model se
ties to study design and the alternative research hypothe
sentation and discussion of the $\log(\mathcal{L})$ values, K, the app
criterion for each model, Δ_i, and w_i is recommended (
2001d). Evidence ratios should be presented with othe
allow a comprehensive assessment of the alternative hyp
mation allows the merits of each model to be contrasted.
data dredging was done following the formal analysis, th
clearly noted and the tentative insights from these activit

All Possible Models

Unthinking approaches have been the common modus
"all possible models" are frequently seen in the literature
find out" is a poor strategy and usually reflects the fact
did not bother to think clearly about the problem of inter
setting.

The sterile analysis of "just the numbers" will continue
for progress in the sciences.

Formal Inference From More Model: Multimodel Inference

4.1 Introduction to Multimodel Inference

Model selection is most often thought of as a way to selec
then inference is conditional on that model. However, i
approaches are more general than this simplistic concept
Given a set of models, specified independently of the samp
formal inferences based on the entire set of models. He
is on all the models in the set and this has several adv
does reinforce the importance of having a good set of
represent the scientific hypotheses of interest. Part of m
includes ranking the fitted models from best to worst, bas
and then scaling to obtain the relative plausibility of ea
by a weight of evidence (w_i) relative to the selected be
conditional sampling variance ($\text{var}(\hat{\theta}|x, g_i)$) from each m
weights (w_i), unconditional inferences about precision c
entire set of models. Model-averaged parameter estimat
unconditional sampling variances can be easily comput
uncertainty is a substantial subject in its own right, well b
of determining the best model.

By *unconditional*, we mean not conditional on any pa
ever, inference is still conditional on the full set of model
for spending the time to arrive at a good set of models, bas
or hypothesized about the science underlying the study
models should be small and well justified, at least for cor

in so doing? Second, how can model selection uncertain
incorporated into estimates of precision? Third, how can
tance of predictor variables in analyses such as linear or l
assessed? Finally, how can a confidence set of models b
K-L best model? More research is required to develop and
methods for these issues, but we provide several approa
Four examples are provided to illustrate the use of multip
formal inference.

4.2 Model Averaging

4.2.1 Prediction

Consider model-based inference for prediction, where R
ered, each having the parameter θ as the predicted value of
i allows an estimate of the parameter, θ_i. If one of the mo
K-L best (e.g., if its $w \geq 0.90$), then inference could pro
ditionally, on the selected best model. However, it is of
single model is clearly superior to some of the others in
dicted value ($\hat{\theta}$) differs markedly across the models (i.e.,
the models $i = 1, 2, \ldots, R$), then it is risky to base pr
selected model. An obvious possibility is to compute a w
the predicted value, weighting the predictions by the Aka

Model Averaging

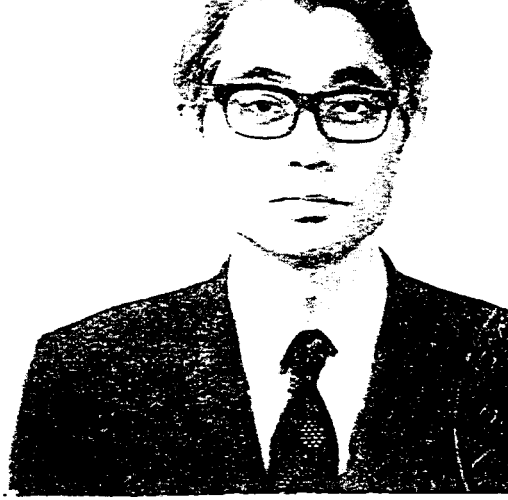
This concept leads to the model averaged estimates,

$$\bar{\hat{\theta}} = \sum_{i=1}^{R} w_i \hat{\theta}_i,$$

where $\bar{\hat{\theta}}$ denotes a model averaged estimate of θ. Alternati
is used to provide the estimated model selection frequ
averaging can be done using,

$$\bar{\hat{\theta}} = \sum_{i=1}^{R} \hat{\pi}_i \hat{\theta}_i.$$

This type of model averaging is useful for prediction p
where a particular parameter (e.g., γ an immigration pr
all the models in the set. Prediction is an ideal way to vie
because each model in a set, regardless of its parametriza
make a predicted value.



Hirotugu Akaike was born in 1927 in Fujinomiya-shi, Shizuoka-je
B.S. and D.S. degrees in mathematics from the University of Toky
spectively. He worked at the Institute of Statistical Mathematics for
its Director General in 1982. He has received many awards, prizes,
in theoretical and applied statistics (deLeeuw 1992, Parzen 1994)
"*Proceedings of the First US/Japan Conference on the Frontiers*
An Informational Approach (Bozdogan 1994) commemorated Prof
65th birthday. Bozdogan (1994) records that the idea of a connectio
Leibler discrepancy and the empirical log-likelihood function oc
morning of March 16, 1971, as he was taking a seat on a commute

4.2.2 Averaging Across Model Parameters

If one has a large number of closely related models, such a
based variable selection (e.g., all subsets selection), desig
model is unsatisfactory because that "best" model is often
is, the model estimated to be best would vary from data s
replicate data sets would be collected under the same un
this situation, model averaging provides a relatively m
inference.

The concept of inference being tied to all the models c
model selection bias effects on linear regression coeffic
subsets selection. For the linear regression coefficient β_j
dictor variable x_j there are two versions of model avera
the estimate $\hat{\tilde{\beta}}_j$ where β_j is averaged over all models in w

$$\hat{\beta}_j = \frac{\sum_{i=1}^{R} w_i I_j(g_i)\hat{\beta}_{j,i}}{w_+(j)},$$

$$w_+(j) = \sum_{i=1}^{R} w_i I_j(g_i),$$

and

$$I_j(g_i) = \begin{cases} 1 & \text{if predictor } x_j \text{ is in model } g \\ 0 & \text{otherwise.} \end{cases}$$

Here, $\hat{\beta}_{j,i}$ denotes the estimator of β_j based on model g_i. is merely the sum of the Akaike weights over all mod predictor variable j is explicitly in the model. Note, $w_+($ average value about whether variable x_j is in (or not in) Thus, $\hat{\tilde{\beta}}_j$ is a "natural" average to consider, as it only aver

where an unknown β_j parameter appears. Note, however, ignores evidence about models g_i wherein $\beta_{j,i} \equiv 0$.

An alternative way to average over linear regression m that variable x_j is "in" every model, it is just that in som sponding β_j is set to zero, rather than being considered un on model g_i being selected, model selection has the effect from zero (Section 1.6). Thus, a second model-averaged $\tilde{\tilde{\beta}}_j$, is suggested:

$$\tilde{\tilde{\beta}}_j = w_+(j)\hat{\tilde{\beta}}_j.$$

This $\tilde{\tilde{\beta}}_j$ actually derives from model averaging over all where x_j is not in a particular model, it is because $\beta_{j,i} \equiv$ the estimate $\hat{\beta}_{j,i}$. The resultant average is identical to w_+ $w_+(j)$ serves to *shrink* the conditional $\hat{\tilde{\theta}}_j$ back towards zer serves to ameliorate much of the model selection bias Investigation of this general idea and its extensions are ar

One point here is that while $\hat{\tilde{\beta}}_j$ can be computed ignorin those where x_j appears, $\tilde{\tilde{\beta}}_i$ does require fitting all R of Improved inference requires fitting all the a priori mod type of model averaging. When possible, one should use all the models, via model averaging and selection bias adju a "select the best model and ignore the others" strategy.

There are several advantages, both practical and philo averaging, when it is appropriate. Where a model averag used it often has reduced bias and, sometimes has better pr

motivating ideas). Bayesian model averaging is easy to
be difficult to implement in practice. Information-theoreti
averaging are easy both to understand and implement, e
large number of models, each with potentially many para

While there are many cases where model averaging is us
model averaging structural parameter estimates in some
models. While it is often appropriate to average slope
regression models, structural parameters in nonlinear mo

$$\mathrm{E}(y) = (a + b_x)/(1 + cx) \quad \text{or} \quad \mathrm{E}(y) = a(1 - [1$$

should not be averaged. For example, a weighted avera
models of any of the parameters a, b, c, or d would not be
model averaging the predicted expected response variab
value of x, across models, is advantageous in reaching a
is not conditional on only a single model.

It is important to realize that the expected value of the
mate, $\mathrm{E}(\hat{\tilde{\theta}})$, is not necessarily the same as θ from absolute t
sampling theory the estimator $\hat{\theta}$ ($\equiv \hat{\theta}_i$ for the selected m
by sample), arrived at in the two-stage process of model s
parameter estimation given the model, is by definition an u
$\mathrm{E}(\hat{\tilde{\theta}})$ as given by (4.1 or 4.2). Therefore, the unconditional
$\hat{\theta} \equiv \hat{\tilde{\theta}}$ is to be computed with respect to $\mathrm{E}(\hat{\tilde{\theta}})$. Any remain
$\hat{\tilde{\theta}}$ cannot be measured or allowed for in model selection u
part of the intent of having a good set of models and sou
to render this bias negligible with respect to the uncondit

Model-averaging ideas are well developed from the B
(see Madigan and Raftery 1994, Draper 1995, Raftery 199
Hoeting et al. 1999; Newman 1997 provides an applicatio
has not yet been commonly adapted into applied frequent
theoretical basis for these approaches and ideas appears i
Buckland et al. 1997 and the Bayesian references just ab

4.3 Model Selection Uncertainty

An understanding of statistical inference requires that one
that generates the sample data we observe. For a given
computer simulation study, data are observed on some pr
second, independent data set could be observed on the sa
under nearly identical conditions, the new data set would d
the first. Clearly both data sets would contain informatio

data observed. However, inferences must not be overly
to the (single) data set observed. That is, we would like
robust, with respect to the particular data set observed,
we tend to avoid problems associated with both underfi
(overinterpreting) the limited data we have. Thus, we wo
to make inferences about the process as if a large numb
were also available.

With only a single data set, one could use AIC, and s
for inference. However, if several other independent data
would the same model be selected? The answer is that
but generally, there would be variation in the selected m
data set, just as there would be variation in parameter esti
given that the same model is used for analysis. The fact
might suggest the use of other models leads us to model s
and hence another variance component that should be inc
precision of parameter estimates (Section 1.7).

If an analyst selects a model using AIC_c (or using so
such as cross-validation) and makes estimates of the sam
estimated parameter in that model, he invariably does s
selected model. The estimated precision will then likel
because the variance component due to model selection
omitted. The standard errors computed conditional on th
small, confidence intervals will be too narrow, and achie
below the nominal level. Chatfield (1995b) reviews this i
also see Rencher and Pun (1980), Chow (1981), Hurvi
Pötscher (1991), Goutis and Casella (1995), and Kabaila

This section presents a variety of methods that can be
the uncertainty associated with model selection, either w
model, or regarding uncertainty about selected variables; a
sures of unconditional precision (e.g., sampling variance,
confidence intervals) for parameter estimators, rather than
measures of sampling uncertainty conditional on a selecte
research is encouraged to better understand the properti
these approaches.

There are three general approaches to assessing model s
(1) theoretical studies, mostly using Monte Carlo simulat
bootstrap applied to a given set of data; and (3) utilizing t
ences (i.e., Δ_i) and model weights w_i from the set of mod
insights can be obtained about model selection and asso
by extensive Monte Carlo simulations of model selecti
and Tsai 1998). Use of the bootstrap and $\hat{\pi}_i$ values appl
gle data set; hence they represent our focus here. The bo
10,000 samples for reliable results, and it could take man

development and understanding.

Monte Carlo investigations generate 1,000 to 10,000 in
(sometimes, 100,000 or even a million samples are needed)
ating model. These data sets are then analyzed, the log-lik
to obtain the MLEs, and model selection is done to ide
for each sample. Finally, one can summarize resultant re
models selected and other information of interest, such a
as well as conditional and unconditional variances of p
over models. For results to apply fully to the K-L model
envisioned for real data, the generating model (which is t
lation study) should be complex and not contained in the
models, $g_1, \ldots, g_R$ (i.e., g_i or $g_i(\underline{x} \mid \underline{\theta})$). We present some
of Monte Carlo simulation results in Chapters 5 and 6.

Many, if not most, simulation studies on model selecti
these conditions (e.g., Wang et al. 1996). Rather, they a
because (1) a simple generating model is used (so no
only small K), (2) the set of models considered contains t
(i.e., contains "truth"), and (3) the model selection goal is
generating model (hence to select "truth"). None of these
of real data-analysis problems; hence we discount the resul
studies as appropriate guides to real-world model selectio
6 for more details on this common error).

The fundamental idea of the model-based sampling
statistical inference is that the data arise as a sample fro
probability distribution, f, and hence the uncertainties o
be measured if we can estimate f. There are ways to const
estimator of (in essence) f from the sample data. The fun
bootstrap method (Section 2.13) is that we compute meas
uncertainty from that estimated sampling distribution of

4.3.1 *Concepts of Parameter Estimation and Model Selection Uncertainty*

Statistical science should emphasize estimation of param
measures of estimator uncertainty. Given a correct mod
sumes $g = f$), an MLE is reliable, and we can compute
of its sampling variance and a reliable confidence interv
likelihood interval; see Royall 1997). If the model is se
pendently of the data at hand, and is a good approximati
large, then the estimated sampling variance is essentiall
appropriate confidence interval will essentially achieve it
This would be the case if we used only one model, deci

here), the selection process is expected to introduce an
sampling uncertainty into any estimated parameter; hence
sampling variances are too small: They are conditional on
reflect model selection uncertainty. One result is that co
intervals can be expected to have less than nominal cover

Consider a scalar parameter θ, which may be used in al
models considered, but is in the selected model, and therei
θ_i given model g_i. Here, the subscript i denotes the mode
with the understanding that this parameter means the sam
in which it appears. There is a conceptual true value of θ
However, the value of θ that we would infer, in the sense of
large sample size) from model g_i applied to the data, ma
model. Given model g_i, the MLE, $\hat{\theta}_i$, has a conditional sa
and hence a conditional sampling variance $\text{var}(\hat{\theta}_i \mid g_i)$. W
$\text{var}(\hat{\theta}_i \mid g_i)$ to be functionally and numerically identical to
notation is more traditional; we use the former notatio
emphasize the importance of assuming the model in its
other models are also being considered.

There is a concept of the true value of θ: It is the v
compute based on knowing truth, f, even though θ need
in f. To the extent a model, g_i, is wrong (i.e., $g_i \neq f$),
when this K-L best value is determined for θ under assu
is, even when data are generated by f, if those data a
model g_i, we will infer (for large n) that the value of
model this possible "bias" (i.e., $\theta_i - \theta$) is not of great co
be dominated by the conditional sampling standard erro
this domination is one feature of a "good" model). The
one source of model selection uncertainty into $\hat{\theta}$; that is,
models in an unknown manner. In many situations the m
something to us, and we will then take $\hat{\theta}_i$ derived only fro
g_i, as the most meaningful estimator of θ. This is what
done, and seems sensible, so much so that the alternative
seems at first strange, but is an alternative to getting an est
multiple models. Model averaging arises in a natural wa
the unconditional sampling variance of $\hat{\theta}_i$.

Another problem arising from model selection uncerta
where the estimate of sampling variance is derived from th
fitted model (e.g., multiple linear regression). In this case,
a negatively biased estimator of conditional sampling varia
related concepts are required to understand this issue. Fi
structure is Poisson, binomial, multinomial (includes cont

(e.g., for the binomial, the model parameter is p and the
variance is $\text{var}(\hat{p}) = \frac{p(1-p)}{n}$).

Underestimation of sampling variance due to structural
seem to be a serious problem in cases where such a theor
ance is known. In particular, this is true if we also use a vari
$\hat{c}$, applied for all models in the set of R models (so selecti
regards $\hat{c}$) to adjust for any modest structural lack of fit o

The second situation occurs often in regression model
sampling variance, σ^2, is functionally unrelated to the
there is no true replication. Then σ^2 must be estimated o
the fitted model. In this case there is neither true replicat
basis to infer σ^2, such as there is in models for count da
structural component of the model to the data, we will ge
hence estimated sampling standard errors of any $\hat{\theta}_i$ will be
likely to be a compensating increase in the factor $(\hat{\theta}_i - \theta$
at the level of the regressor values can eliminate this prob
not have such true replication.

The ideas of classical sampling theory can be used to c
sampling variance of $\hat{\theta}$ resulting from the two-stage p
selection, then (2) using $\hat{\theta} \equiv \hat{\theta}_i$ given that model g_i was s
process carried out many times, m, each time on an inde
sample j we get $\hat{\theta}_j$ as our estimator of θ. This conceptua
selected model in repetition j, but we do not need, hence a
indexed notation (such as $\hat{\theta}_{i,j}$) to denote both sample j a
given sample j.

The estimated unconditional sampling variance, $\widehat{\text{var}}(\hat{\theta})$
cates would be $\sum(\hat{\theta}_j - \overline{\hat{\theta}})^2/(m-1)$; $\overline{\hat{\theta}}$ is the simple avera
(hence the $\hat{\theta}_j$ have been averaged over selected models)
mator represents the total variation in the set of m valu
within and between-model variation is included. This s
can be partitioned into R subsets, one for each model wh
contains all the $\hat{\theta}$'s computed under selected model i. Th
from the ith subset of the $\hat{\theta}$ values an estimate of the c
variance of $\hat{\theta}$ when model g_i was selected. Formal mather
of partitioning the above $\widehat{\text{var}}(\hat{\theta})$ into R components and
to get a theoretical unconditional sampling variance gives
as a weighted combination of conditional variances, var
for variation among $\theta_1, \ldots, \theta_R$. The weights involved a
tion probabilities. Relevant formulas are given in the next
mention one more issue.

models. In this case the basis for unconditional inferenc
ter can be (but need not be) made based on just those m
parameter appears. An example is variable selection in li
y versus p regressors, $x_1, \ldots, x_p$ (plus an intercept). Th
models, but each regressor appears in only half of these
models). Thus if regressor variable x_j, hence parameter β_j
AIC best model, we could restrict ourselves to just that s
contain β_j in order to directly estimate the unconditiona
of $\hat{\beta}_j$. All the above (and below) considerations about co
ditional variances with regard to a particular parameter c
apply to just the subset of models that include the parame

We have emphasized models as approximations to tr
"wrong" is technically called model misspecification (se
the usual theoretical sampling variances of MLEs, $\text{var}(\hat{\theta}_i$
but only trivially so if the model is a good approximati
theory that gives the correct conditional (on the model) s
$\hat{\theta}_i$ in the event of model misspecification (Chapter 7 gives
However, the correct estimator of $\text{var}(\hat{\theta}_i \mid g_i)$ is then so
and variable (a type of instability) that it generally see
theoretical estimator supplied by the usual model-specifi
(which assumes that the model is correct). This simplif
especially defensible when done in conjunction with so
procedures intended to minimize both serious overfitting
believe AIC is suitable for this selection purpose and tha
consideration is thus to get reliable unconditional samp
confidence intervals) for MLEs after model selection.

4.3.2 *Including Model Selection Uncertainty in Sampling Variance*

We continue to assume that the scalar parameter θ is co
considered. This will often be the case for our full set of a p
and is always the case if our objective is prediction with th
as interpolation or extrapolation with a generalized linear
if our focus is on a model structural parameter that appea
our full set of models, then we can restrict ourselves to th
make the sort of inferences considered here about the para
the latter case we simply consider the relevant subset as t
under consideration.

In repeated (conceptual) samples there is a probability
model. Presentation of a defensible way to augment $\widehat{\text{var}}$
selection uncertainty involves the idea of model averag

$\theta = \sum_{i=1} \pi_i \theta_i$, and its estimator $\hat{\theta} = \sum_{i=1} \hat{\pi}$

In some theory development we use π_i rather than $\hat{\pi}_i$, but such as $\bar{\theta}$ and $\bar{\hat{\theta}}$.

The theoretical, unconditional sampling variance of $\hat{\theta}$ given by

$$\text{var}(\hat{\theta}) = \sum_{i=1}^{R} \pi_i \left[\text{var}(\hat{\theta}_i \mid g_i) + (\theta_i - \bar{\theta})^2 \right]$$

This result follows directly from frequentist sampling Section 4.3.1 that if we had m independent samples and selection to each sample to get $\hat{\theta}_j$, $j = 1, \ldots, m$, then an would be

$$\widehat{\text{var}}(\hat{\theta}) = \sum (\hat{\theta}_j - \bar{\hat{\theta}})^2 \;/\; (m-1).$$

Here, j indexes the sample that $\hat{\theta}_j$ came from (whatev whereas i indexes that $\hat{\theta}_i$ arose from model i (whatever th notation allows us to focus on different aspects of the mod without a notation so complex that it hinders understanding m become infinite, the above estimator of $\text{var}(\hat{\theta})$ converg unconditional sampling variance of $\hat{\theta}$. By first grouping t $\hat{\theta}$ values by model and then taking the needed limit as m

Readers less interested in the derivation of an estimator model-averaged estimate may want to skip to the followi $\text{var}(\hat{\theta}_i \mid g_i) + (\theta_i - \bar{\theta})^2$ is just the mean square error of $\hat{\theta}_i$ in one sense the unconditional variance of $\hat{\theta}$ is just an a error. Specifically,

$$\text{E}[(\hat{\theta}_i - \bar{\hat{\theta}})^2 \mid g_i] = \text{var}(\hat{\theta}_i \mid g_i) + (\theta_i - \bar{\theta}$$

and we recommend thinking of the above quantity as th of $\hat{\theta}_i$, given model i, when $\hat{\theta}_i$ is being used as an estima poration of model selection uncertainty into the variance new thinking like this. The matter arises again when we of covariance, $\text{E}[(\hat{\theta}_i - \bar{\theta})(\hat{\theta}_j - \bar{\hat{\theta}}) \mid g_i]$, that also allows uncertainty.

One might think to estimate the augmented samplin $\widehat{\text{var}}(\hat{\theta}_i \mid g_i) + (\hat{\theta}_i - \bar{\hat{\theta}})^2$. Such an estimator is not suppo and is likely to be both biased (it could be bias-correc able; however, we have not investigated this possible e

are estimated (they are not the major source of estimation v
ance estimator), we can evaluate $\mathrm{E}(\widehat{\mathrm{var}}(\hat{\theta}))$ to bias-corre
involves the sampling variance, $\mathrm{var}(\hat{\bar{\theta}})$, of the model-avera

$$\mathrm{E}(\widehat{\mathrm{var}}(\hat{\theta})) = \mathrm{var}(\hat{\theta}) + \sum \pi_i \, \mathrm{var}(\hat{\theta}_i \mid g_i) - \mathrm{v}$$

which leads to

$$\mathrm{var}(\hat{\theta}) = \mathrm{var}(\hat{\bar{\theta}}) + \sum \pi_i \mathrm{E}(\hat{\theta}_i - \hat{\bar{\theta}})^2.$$

It seems that we cannot avoid estimating $\hat{\bar{\theta}}$ and $\mathrm{var}(\hat{\bar{\theta}})$ ev
$\widehat{\mathrm{var}}(\hat{\theta})$ that we are seeking. First, note that efforts to evalu
are circular, thus useless. Anyway, at worst we would only
quantity without (much) bias, which we can clearly do. Se
that as one might expect, if our goal is to estimate $\bar{\theta}$, then
$\hat{\bar{\theta}}$ is to be preferred to $\hat{\theta}_i$ because it will have a smaller
However, given that our goal is to estimate θ, there is no
claim that $\hat{\bar{\theta}}$ is the superior estimator as compared to $\hat{\theta} \equiv$

From Buckland et al. (1997) we will take the needed v

$$\mathrm{var}(\hat{\bar{\theta}}) = \left[\sum_{i=1}^{R} \pi_i \sqrt{\mathrm{var}(\hat{\theta}_i \mid g_i) + (\theta_i - \bar{\theta})^2} \right]$$

with the estimator as

$$\widehat{\mathrm{var}}(\hat{\bar{\theta}}) = \left[\sum_{i=1}^{R} \hat{\pi}_i \sqrt{\widehat{\mathrm{var}}(\hat{\theta}_i \mid g_i) + (\hat{\theta}_i - \hat{\bar{\theta}})^2} \right]$$

Formula (4.5) entails an assumption of perfect pairwise
$\hat{\theta}_i - \hat{\bar{\theta}}$ and $\hat{\theta}_h - \hat{\bar{\theta}}$ for all $i \neq h$ (both i and h index mo
correlation of $\rho_{ih} = 1$ is unlikely; however, it will be hi
value of $\rho_{ih} = 1$ is conservative in that $\mathrm{var}(\hat{\bar{\theta}})$ computed
to be too large if this assumption is in error. Also, by just
into (4.6) a further upward bias to (4.5) results. Thus fr
$\widehat{\mathrm{var}}(\hat{\theta}) = \widehat{\mathrm{var}}(\hat{\bar{\theta}}) + \sum \pi_i (\hat{\theta}_i - \hat{\bar{\theta}})^2$ with $\widehat{\mathrm{var}}(\hat{\bar{\theta}})$ from (4.
much positive bias. Hence, we are now suggesting just us
from (4.6).

All simulations we have done so far, in various contexts
of this estimator:

$$\widehat{\mathrm{var}}(\hat{\bar{\theta}}) = \left[\sum_{i=1}^{R} \hat{\pi}_i \sqrt{\widehat{\mathrm{var}}(\hat{\theta}_i \mid g_i) + (\hat{\theta}_i - \hat{\bar{\theta}})^2} \right]$$

(4.6) and (4.7) is acceptable. Improved estimation of unc
variances under model selection may be possible. Howe
to give practical solutions to some problems under mod
expectation that improvements will be further explored.

The $\hat{\pi}_i$ in (4.7) (and the equivalent (4.6)) will usually be
weights, w_i. In general, $w_i \neq \pi_i$; rather, w_i can be consid
π_i, but (4.7) seems robust to slightly imprecise values o
natively, one can use the bootstrap estimates, $\hat{\pi}_i = b_i/B$;
one has bootstrap samples, the analytical formulas above

As a final part of this section we give some details
(4.5) in a more restricted context than was used in Buc
Specifically, we do not assume that the R models are rando
possible models. Rather, we just condition on the set of R m
provided; hence, inferences are conditional on just this set
each $\hat{\theta}_i$ is considered as an estimator of $\bar{\theta}$, and it is this con
critical to getting a variance formula that includes model s

Ignoring that the π_i in (4.5) need to be estimated, the
expressed as

$$\operatorname{var}(\hat{\bar{\theta}}) = \sum_{i=1}^{R} (\pi_i)^2 \mathrm{E}[(\hat{\theta}_i - \bar{\theta})^2 \mid g_i] + \sum_{h \neq i}^{R} \sum_{i}^{R} \pi_i \pi_h \Big[\mathrm{E}(\hat{\theta}_i - \bar{\theta})(\hat{\theta}_h - \bar{\theta}) \mid$$

From above we know that

$$\mathrm{E}[(\hat{\theta}_i - \bar{\theta})^2 \mid g_i] = \operatorname{var}(\hat{\theta}_i \mid g_i) + (\theta_i - \bar{\theta}$$

In order to coherently allow for model selection uncertainty
with the definition of a correlation, we must interpret the
this expression for $\operatorname{var}(\hat{\bar{\theta}})$ as

$$\mathrm{E}(\hat{\theta}_i - \bar{\theta})(\hat{\theta}_h - \bar{\theta}) \mid g_i, g_h) = \rho_{ih} \sqrt{\mathrm{E}[(\hat{\theta}_i - \bar{\theta})^2 \mid g_i] \mathrm{E}}$$

hence,

$$\mathrm{E}(\hat{\theta}_i - \bar{\theta})(\hat{\theta}_h - \bar{\theta}) \mid g_i, g_h) = \rho_{ih} \sqrt{[\operatorname{var}(\hat{\theta}_i \mid g_i) + (\theta_i - \bar{\theta})^2][\operatorname{var}(\hat{\theta}_h \mid g_h)}$$

$$\text{var}(\bar{\theta}) = \sum_{i=1}^{R} (\pi_i) \left[\text{var}(\hat{\theta}_i \mid g_i) + (\theta_i - \bar{\theta})^2 \right]$$

$$+ \sum_{h \neq i}^{R} \sum^{R} \pi_i \pi_h \rho_{ih} \sqrt{[\text{var}(\hat{\theta}_i \mid g_i) + (\theta_i - \bar{\theta})^2][\text{var}(\hat{\theta}$$

We have no basis to estimate the across-model correlation
(other than the bootstrap, but then we do not need theory fo
simplifies if we assume that all $\rho_{ih} = \rho$:

$$\text{var}(\hat{\bar{\theta}}) = (1 - \rho) \left[\sum_{i=1}^{R} (\pi_i)^2 \left[\text{var}(\hat{\theta}_i \mid g_i) + (\theta_i \right. \right.$$

$$+ \rho \left[\sum_{i=1}^{R} \pi_i \sqrt{\text{var}(\hat{\theta}_i \mid g_i) + (\theta_i - \bar{\theta})} \right.$$

From (4.8), if we further assume $\rho = 1$, then we get (4.5

$$\text{var}(\hat{\bar{\theta}}) = \left[\sum_{i=1}^{R} \pi_i \sqrt{\text{var}(\hat{\theta}_i \mid g_i) + (\theta_i - \bar{\theta})} \right.$$

Unconditional Variance Estimator

Then, using the Akaike weights (w_i) instead of the n
quencies (π_i) and using estimates instead of parameter
useful result,

$$\widehat{\text{var}}(\hat{\bar{\theta}}) = \left[\sum_{i=1}^{R} w_i \sqrt{\widehat{\text{var}}(\hat{\theta}_i \mid g_i) + (\hat{\theta}_i - \hat{\bar{\theta}})} \right.$$

where $\hat{\bar{\theta}}$ is model-averaged estimate (4.1).

This estimator of the unconditional variance can be
MLE $\hat{\theta}$ from the selected model or for the model averag

If only a subset of the R models in used, then the w_i n
based on just these models (thus these new weights must
If one has the estimated model selection frequencies (π_i)
the estimator

$$\widehat{\text{var}}(\hat{\bar{\theta}}) = \left[\sum_{i=1}^{R} \hat{\pi}_i \sqrt{\widehat{\text{var}}(\hat{\theta}_i \mid g_i) + (\hat{\theta}_i - \hat{\bar{\theta}}} \right.$$

is useful. In either case, $\widehat{\text{se}}(\hat{\bar{\theta}}) = \sqrt{\widehat{\text{var}}(\hat{\bar{\theta}})}$.

parameter estimators in a model, or (what is the same t
ditional covariance between two estimators as based on
model. A formula for the unconditional $\text{var}(\hat{\bar{\theta}})$ values is

$$\text{var}(\hat{\bar{\theta}}) = \left[\sum_{i=1}^{R} \pi_i \sqrt{\text{var}(\hat{\theta}_i | g_i) + (\theta_i - \hat{\bar{\theta}})^2} \right.$$

we need something useful for the analogous

$$\text{cov}(\hat{\bar{\theta}}_1, \hat{\bar{\theta}}_2),$$

$$\hat{\bar{\theta}}_1 = \sum_{i=1}^{R} \pi_i \hat{\theta}_{1i},$$

$$\hat{\bar{\theta}}_2 = \sum_{i=1}^{R} \pi_i \hat{\theta}_{2i}.$$

For a useful estimation formula the Akaike weight, w_i, w

We propose the formula below for $\widehat{\text{cov}}(\hat{\bar{\theta}}_1, \hat{\bar{\theta}}_2)$:

$$= \bar{r}_{1,2} \left[\sum_{i=1}^{R} w_i \sqrt{\widehat{\text{var}}(\hat{\theta}_{1i} | g_i) + (\hat{\theta}_{1i} - \hat{\bar{\theta}}_1)^2} \right] \left[\sum_{i=1}^{R} w_i \sqrt{\widehat{\text{var}} \right.$$

$$= \bar{r}_{1,2}\, \widehat{\text{se}}(\hat{\bar{\theta}}_1)\, \widehat{\text{se}}(\hat{\bar{\theta}}_2),$$

where $\bar{r}_{1,2}$ is

$$\bar{r}_{1,2} = \sum_{i=1}^{R} w_i \times r_{1,2|i},$$

and $r_{1,2|i}$ is the estimated sampling correlation betwee
model i. This model-conditional sampling correlation can
conditional-on-model-i sampling variance-covariance ma
large sample means).

Now if we are considering the difference $d = \hat{\bar{\theta}}_1 - \hat{\bar{\theta}}$
2001), then

$$\widehat{\text{var}}(d) = A^2 + B^2 - 2\bar{r}_{1,2} A B,$$

where

$$A = \sum_{i=1}^{R} w_i \sqrt{\widehat{\text{var}}(\hat{\theta}_{1i} | g_i) + (\hat{\theta}_{1i} - \hat{\bar{\theta}}_1)^2}$$

$$B = \sum_{i=1}^{R} w_i \sqrt{\widehat{\text{var}}(\hat{\theta}_{2i} | g_i) + (\hat{\theta}_{2i} - \hat{\bar{\theta}}_2)^2}$$

work and experience are needed with the formula versus si
formula performance.

4.3.3 Unconditional Confidence Intervals

The matter of a $(1-\alpha)100\%$ unconditional confidence int
ered. We have two general approaches: the bootstrap (see,
1997), or analytical formulas based on analysis results fr
set. The analytical approach requires less computing; her

The simplest such interval is given by the endpoints $\hat{\theta}_i \pm$
$\widehat{\text{se}}(\hat{\theta}_i) = \sqrt{\widehat{\text{var}}(\hat{\theta}_i)}$. One substitutes the model-averaged
the estimator used. A common form used and recommen
interval is $\hat{\theta}_i \pm t_{\text{df},1-\alpha/2}\, \widehat{\text{se}}(\hat{\theta}_i \mid g_i)$. When there is no moc
ignored, it is clear what the degrees of freedom (df) are f
here. For (4.7) it is not clear what the degrees of freedo
however, that we are focusing on situations where sample
that the normal approximation will be applicable. These
confidence interval are based on the assumption that $\hat{\theta}_i$ ha
distribution.

We will hazard a suggestion here; it has not been evalu
but a similar procedure worked in a different context. If f
we have degrees of freedom df_i for the estimator $\widehat{\text{var}}(\hat{\theta}_i \mid g$
small degrees of freedom one might try using the interval
where the adjusted standard error estimator is

$$\widehat{\text{ase}}(\hat{\theta}_i) = \sum_{i=1}^{R} \hat{\pi}_i \sqrt{\left(\frac{t_{\text{df}_i,1-\alpha/2}}{z_{1-\alpha/2}}\right)^2 \widehat{\text{var}}(\hat{\theta}_i \mid g_i) + (}$$

In cases where $\hat{\theta}_i \pm z_{1-\alpha/2}\, \widehat{\text{se}}(\hat{\theta}_i)$ is not justified by a nor
bution (as judged by the conditional distribution of $\hat{\theta}_i$), inte
coverage can be based on a transformation of $\hat{\theta}_i$ if a sui
is known. Log and logit transforms are commonly used
the context of general linear models. In fact, in general lin
tor parameter $\underline{\theta}$ will be linked to the likelihood by a set
$\underline{\theta} = \underline{W}(\underline{\beta})$. Then it is $\underline{\beta}$ that is directly estimated, and it i
the simple normal-based confidence limits on components
used. An interval constructed from a component of $\hat{\underline{\beta}}$ a
sampling variance ((4.5) applies) can be back-transform
the corresponding component of $\underline{\theta}$.

The above methods are justified asymptotically, or if
distribution applies to $\hat{\theta}$. However, "asymptotically" mea

there is no model selection is the profile likelihood interva
and Hsu 1999, Sprott 2000). We suggest here an adaptat
that widens the likelihood interval to account for model s

Let the vector parameter $\underline{\theta}$ be partitioned into the comp
and the rest of the parameters, denoted here by $\underline{\gamma}$. Then th
as a function of θ_i (the subscript denotes the model used)
by

$$\mathcal{PL}(\theta_i \mid \underline{x}, g_i) = \max_{\underline{\gamma}_i \mid \theta_i} \left[\mathcal{L}(\theta_i, \underline{\gamma}_i \mid \underline{x}, g_i) \right]$$

almost always $\mathcal{PL}(\theta_i \mid \underline{x}, g_i)$ has to be computed nume
profile deviance as

$$\mathcal{PD}(\theta_i) = 2 \left[\mathcal{PL}(\hat{\theta}_i \mid \underline{x}, g_i) - \mathcal{PL}(\theta_i \mid \underline{x},$$

The large sample profile likelihood interval ignoring m
tainty is the set of θ_i that satisfy the condition $\mathcal{PD}(\theta_i) \leq$
is the upper $1 - \alpha$ percentile of the central chi-squared
This interval is approximately a $(1 - \alpha)100\%$ confidence

We propose an interval that is a version of (4.9) adj
model selection uncertainty: the set of all θ_i that satisfy

$$\mathcal{PD}(\theta_i) \leq \left[\frac{\widehat{\text{var}}(\hat{\theta})}{\widehat{\text{var}}(\hat{\theta}_i \mid g_i)} \right] \chi^2_{1,1-\alpha}.$$

It suffices to solve (numerically) this inequality for the
endpoints, $\hat{\theta}_{i,\text{L}}$ and $\hat{\theta}_{i,\text{U}}$. In the event that we are not doing
seems logical to use the resultant confidence interval fro

All of the above was assuming that the parameter of inte
of the R models in the full set of models. Often this will no
there will be a subset of size $Q < R$ of the models in
θ occurs. Conceptually, the parameter θ does not occur
models, even as a value equal to zero. In this event we sug
above theory to just that subset of Q models. The $R - Q$
does not appear seem totally uninformative about the va
cannot play a direct role in inference about θ. The situati
selection as in linear all subset selection because there we
that every structural parameter is in each model, but so
$\theta \equiv 0$ in a model.

In the case that $Q = 1$ (θ is unique to one model in t
none of the above results can be used. In this case it s
not be a direct way to include model selection uncertainty
about the value of θ. An approach we can envision here is
conditional sampling variance estimator, $\widehat{\text{var}}(\hat{\theta} \mid g_i)$, by so

$$\frac{\widehat{\text{var}}(\hat{\theta})}{\widehat{\text{var}}(\hat{\theta}_i \mid g_i)}$$

and found that it can vary greatly by parameter. Thus, esti
inflation factor for $\hat{\theta}$ based on a different parameter in
ent models (from the one that model θ appears in) seem
Fundamentally, it is not clear that we should inflate the c
variance of a parameter unique to just one model in the se
all we can, and should, do in this case is note the uncert
that model is likely to be the K-L best model in the ful
use the model-specific conditional sampling variance fo
intervals for θ are then constructed based on just the on
appears (e.g., profile likelihood interval, or other parame
$\hat{\theta} \pm z_{1-\alpha/2}\, \widehat{\text{se}}(\hat{\theta}|g)$.

Bootstrap construction of an unconditional confidenc
rameter in the selected model is not fundamentally
bootstrap-based interval construction without model sel
much discussed in the statistical literature (see, e.g., E
1993, Mooney and Duval 1993, Hjorth 1994).

First one generates a large number, B, of bootstrap sa
and applies model selection to each bootstrap sample. A
original set are fit to each bootstrap sample, and one of t
selected as best. Only the estimated parameters from that
are kept in an output set of parameter estimates for each bo
the index of the selected model for each sample). Hence
not in the selected model g_i for bootstrap sample b, the va
for bootstrap sample b (the subscript b and $\hat{\theta}^*_b$ denote tha
bootstrap sample; the model used to get this $\hat{\theta}^*$ varies over
For any parameter in common over all models, there will
the output data set. In either case the variation in the outpu
values of $\hat{\theta}^*_b$, $b = 1, \ldots, m$ ($\leq B$), reflects both model sel
and within-model sampling variation of $\hat{\theta}$ given a model.

As noted in Section 4.2.2, the model selection frequenc
from these bootstrap results; but our focus here is on unc
uncertainty and confidence intervals (Efron and Tibshiran
of all m values of $\hat{\theta}^*_b$, $\overline{\hat{\theta}}^*$ is an estimator of the model-av
Hence the empirical variance of the set of m values of $\hat{\theta}^*_b$

$$\text{var}(\hat{\theta}^*) = \sum (\hat{\theta}^*_b - \overline{\hat{\theta}}^*)^2 \;/\; (m-1),$$

such an interval fails to make full use of the value of the
finding upper and lower interval estimates that allow for a
distribution for $\hat{\theta}$ under model selection.

The simple, direct bootstrap-based confidence interval
interval (Efron and Tibshirani 1993). Order the bootstrap v
est to largest and denote these ordered values by $\hat{\theta}^*_{(b)}$, b
$(1-\alpha)100\%$ confidence interval select the $\alpha/2$ lower and
centiles of these ordered bootstrap estimates as $\hat{\theta}_{\mathrm{L}}$ and $\hat{\theta}$
may not occur at integer values of b, but if m is large, it suf
and $\hat{\theta}_{\mathrm{U}} = \hat{\theta}^*_{(u)}$, where $l = [m \cdot \frac{\alpha}{2}]$ and $u = [m \cdot (1 - \frac{\alpha}{2})]$
possibly better, unconditional intervals after model selecti
strap are considered by Shao (1996) for regression prob
can be done with model averaging in a straightforward wa
$\hat{\bar{\beta}}$ or $\tilde{\bar{\beta}}$ and model averaging should be done for each boo
one obtains $\hat{\bar{\beta}}^*$ or $\tilde{\bar{\beta}}^*$ for each of the B bootstrap samples
the standard error and confidence intervals from these res

Note that B needs to be at least several hundred for the
begin to work well, and we recommend 10,000 (and at leas
the parameter of interest is in every model, then $m = B$, w
If the parameter is truly not in every model, m is random
be made larger to ensure that a sufficient sample size, m,
samples is obtained.

4.4 Estimating the Relative Importance o

Data analysis is sometimes focused on the variables to inclu
the selected model (e.g., important vs. unimportant). Varia
the focus of model selection for linear or logistic regressi
investigator uses stepwise analysis to arrive at a final mo
conclusion is drawn that the variables in this model are im
other variables are not important. While common, this i
among other issues, fails to fully consider model selecti
we provide simple methods to quantify the evidence for th
variable in the set.

Consider 10 models based on combinations of a numb
ables. Assume that the selected best model includes x_1
weight of only 0.3. There is considerable model selecti
and hence there would seem to be only weak evidence f
variable x_1 based on the selected best model. But one must

that situation would suggest that x_1 is a very important pr
of this importance is to sum the Akaike weights (or the b
subset of models that include variable x_1. This idea is a
to model selection whenever it is equated to variable se
nonlinear models of any type.

Consider the hypothetical example of three regressors,
search for the best of the eight possible models of the sim
type: $y = \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\epsilon$. The possible combi
that define the eight possible models are shown below, alo
Akaike weights w_i (a 1 denotes that x_i is in the model; othe

x_1	x_2	x_3	w_i
0	0	0	0.00
1	0	0	0.10
0	1	0	0.01
0	0	1	0.05
1	1	0	0.04
1	0	1	0.50
0	1	1	0.15
1	1	1	0.15.

While the selected best model has weight of only 0.
of 0.5 of being the actual K-L best model here, the sun
variable x_1 is 0.79. This is evidence of the importance of
the models considered. Variable x_2 was not included in the
but this should not suggest that it is of zero importance.
weight of evidence support is 0.35. Finally, the sum of the
predictor variable x_3 is 0.85. Thus the evidence for the im
x_3 is substantially more that just the weight of evidence
We can order the three predictor variables in this exampl
importance: x_3, x_1, x_2 with importance weights of 0.85,
with other methods recommended here, we see that we a
selection to go well beyond just noting the best model fro

Relative Variable Importance

Estimates of the relative importance of predictor varia
made by summing the Akaike weights across all the mod
variable j occurs. Thus, the relative importance of varia
the sum $w_+(j)$.

The larger the $w_+(j)$ the more important variable
other variables. Using the $w_+(j)$, all the variables can
importance.

The direction and magnitude of effect size should often
averaged estimates with appropriate measures of precisi

of all models that include the pair of variables. For the pai
of evidence for the importance of this pair is 0.19. For pai
of evidence for importance is 0.23, while for the pair x_1
evidence is 0.65 (compared to 0.5 for the selected mod
procedures apply when assessing the relative importance

When assessing the relative importance of variables us
it is important to achieve a balance in the number of mod
variable j. For example, in the numerical example abov
variables appeared in four models. This balancing puts ea
footing.

To summarize, in many contexts the AIC selected bes
some variables and exclude others. Yet this inclusion o
does not distinguish differential evidence for the importa
the model. The model weights, w_i or $\hat{\pi}_i$, summed over all
a given variable provide a better weight of evidence for th
variable in the context of the set of models considered.

4.5 Confidence Set for the K-L Best Mod

4.5.1 Introduction

There exists a concept of a confidence set for the K-L bes
data, just as there is a confidence interval for a parameter b
data. For a 95% confidence set on the actual K-L best m
not unique) approach is to sum the Akaike weights fron
until that sum is just ≥ 0.95; the corresponding subset o
confidence set on the K-L best model. In this example (as
indexed the models as 1 to 7 in order of decreasing weig
set is models {1, 2, 3, 4, 5}, which has sum of weights =
approach to a confidence set of models we are interpretin
as a posterior probability (i.e., given the data and the se
that model i is the K-L best model (see Section 6.4.5).
approach but it is easy to understand.

There is another approach to developing a confidence
on the idea of a Δ_i being a random variable with a sa
In particular, let index value *best* correspond to the actua
model in the set. There is always a K-L best model in the se
that ties might occur). It is thus model g_{best} that we sho
analysis; we just do not happen to know a priori the value
of conceptual interest is

$$\Delta_p = \mathrm{AIC}_{best} - \mathrm{AIC}_{min} .$$

confidence interval on θ. A pivotal quantity is one whose s
is independent of any unknown parameters, a t-distributed
The "p" in the Δ defined by (4.12) denotes that this Δ is
value rather than an actual Δ_i that we can compute from

It is not exact to consider $\Delta_p = \text{AIC}_{best} - \text{AIC}_{min}$ a
but it seems a useful approximation in some contexts. T
useful in is one of complex truth; tapering effect sizes;
being good approximations to truth, with full truth not
used; and a lot of nested sequences of models (as in the
example in Chapter 3). Monte Carlo studies on the abo
we have done many of these and results support the co
context, the sampling distribution of this Δ_p has substan
95th percentile of the sampling distribution of Δ_p is gene
10, and in fact generally less than 7 (often closer to 4 in sim
means that an alternative rule of thumb for an approxim
set on the K-L best model is the subset of all models g_i
value that is roughly in the range 4 to 7. In fact, the Δ
model is not competitive as a candidate for the K-L best n
is probably somewhere between 2 and 10 in many situati
is not large, while a $\Delta_i = 10$ is strong evidence against
K-L best model in the set of models considered, if samp
These guidelines, rough as they are, are useful.

We review this interpretation of evidence from the Δ
are independent, sample sizes are large, and models are n

Δ_i	Level of Empirical Support
0–2	Substantial
4–7	Considerably less
> 10	Essentially none

Models with $\Delta > 10$ represent very strong evidence that
K-L best model. The reader should not take these guideli
as there are situations to which they do not apply well (su
small sample size or dependent observations). Likewise, i
of models, these guidelines may not hold.

We had these guidelines well in mind when we encour
lines for the Bayes factor. The Bayes factor is a Bayesia
relative data-based likelihood of one model versus anot
out considering any priors on the set of models (Berger
Raftery 1996a); it is somewhat analogous to $\exp(-\frac{1}{2}\Delta_i)$.
1996b:165) presents a similar scale for interpretation of 2
evidence for the simpler model being considered.

A third reasonable basis for a confidence set on mo
likelihood-based inference (see e.g., Edwards 1992, Az

of relatively more plausible models. Thus our confidence
models for which the ratio

$$\frac{\mathcal{L}(g_i|x)}{\mathcal{L}(g_{min}|x)} > \quad \textbf{cutoff},$$

where the cutoff value might be $\approx 1/8$. Models where t
greater than 1/8 are in the confidence set and are deem
is no direct sampling theory interpretation required and
to the idea of the selected subset of models including
with a preset, known, long-run inclusion relative frequenc
procedure has the advantage that the cutoff remains uncha
or deletion of a model (of course, a new model will be
the confidence set). Thus, a confidence set based on the
desirable invariance property. In contrast, any change in
can alter the confidence set when summing the Akaike w

We have presented three approaches to finding a con
els: (1) base it directly on the Akaike weights, interpre
probabilities of each model being the actual best model
use a cutoff Δ_i motivated by the idea of the sampling
approximate pivotal, Δ_p (using, say, the 95th percentile
as the cutoff Δ); or (3) think in terms of relative likelih
min indexing the selected AIC best model) use a cutoff v
$\mathcal{L}(g_i|x)/\mathcal{L}(g_{min}|x) \equiv \exp(-\frac{1}{2}\Delta_i)$ is small, say 0.135 ($\Delta_i =$
or 0.050 ($\Delta_i = 6$). In general we favor this third approac

The use of intervals based purely on relative likelihood
by statistical theory (cf. Berger and Wolpert 1984, Edw
1996, Royall 1997), but rarely taught or used. Rather, mo
have been taught to think of confidence intervals in term
ability; hence they might feel more at home with metho
of which are motivated by the sampling theory idea of a
terval on a parameter. The approach based on simple ev
quite useful. More needs to be known about the properties
ods to construct a confidence set of models before we w
recommending just one approach.

4.5.2 Δ_i, Model Selection Probabilities, and the

For a given set of data we can estimate the sampling d
selection frequencies and the distribution of $\Delta_p = AI$
formula 4.12) using the bootstrap method. In this method
(unknown) K-L best model is played by the model selec
data analysis; denote that model by model g_{best}. For exa

$\Delta_p = \mathrm{AIC}^*_{best} - \mathrm{AIC}^*_{min}$, *best* does not change over boo
model producing AIC^*_{min} varies by bootstrap sample. How
model g_{best} in which case $\Delta^*_p = 0$. When it is not mode
AIC^*_{min}, then $\Delta^*_p > 0$.

The B bootstrap samples provide B values of Δ^*_p that
conditional on the data. The percentiles of the empirical
tion function of Δ^*_p provide the estimate of the percent
distribution of Δ_p, and hence provide a basis for a confic
best model for the actual data. For a $(1-\alpha)100\%$ confid
best model, order the $\Delta^*_{p,(b)}$ (smallest to largest) and find
$b = [(1-\alpha)B]$. For the actual data analysis results, the su
g_i having $\Delta_i \leq \Delta^*_{[(1-\alpha)B]}$ is the desired confidence set. Fo
the upper tail percentiles of Δ_p, B needs to be 10,000.

Other information can be gained from these bootstrap
selection uncertainty, in particular, the frequency of selec
models. Let b_i be the number of samples in which model
K-L best model. Then an estimator of the relative frequenc
in the given situation is $\hat{\pi}_i = b_i/B$. These estimated se
are useful for assessing how much sampling variation the
of the best model: they directly quantify model selectio
estimated selection probabilities are similar to, but not i
to, the Akaike weights, which also quantify strength of ev
selection uncertainty.

Also, for each bootstrap sample we can compute the A

$$w^*_i = \frac{\exp(-\frac{1}{2}\Delta^*_i)}{\sum_{r=1}^{R}\exp(-\frac{1}{2}\Delta^*_r)}$$

and then average these over the B samples to get $\overline{w}^*_i$. Co
$\overline{w}^*_i$, and $\hat{\pi}_i$ is informative as to the coherence of these me
provides information about the sampling uncertainty in r
theoretical measure of model selection sampling uncertai
unknown selection probabilities, $\pi_1 \ldots, \pi_R$. Either the $\hat{\pi}_i$
or the Akaike weights, w_i (we often prefer the latter becaus
computer-intensive calculations and they relate more di
evidence based on the data at hand), may be taken as the
uncertainty about model selection. Note that in this usage
we mean the structural form of the model (such as which v
vs. excluded) without consideration of the specific parame
each model. Parameter-estimation uncertainty is concept
(but influenced by) model selection uncertainty.

Consider a set of three models, in which models g_2 and g_3 a
a mistake was made in setting up the problem on the compu
and g_3 are 100% redundant in the set of models; the mod
only models g_1 and g_2. Assume $\Delta_1 = 0$ and $\Delta_2 = \Delta_3 =$
set of three models we get $\mathcal{L}(g_1)/\mathcal{L}(g_2) = \mathcal{L}(g_1)/\mathcal{L}(g_3$
for the correct set of two models, $\mathcal{L}(g_1)/\mathcal{L}(g_2) = 7.4$. Th
redundancy has not affected the Δ_i nor the likelihood evide
However, the (normalized) Akaike weights (Section 2.9)
set of two models, $w_1 = 0.881$ and $w_2 = 0.119$; wher
model redundancy, $w_1 = 0.787$ and $w_2 = w_3 = 0.106$.
model set we still have $w_1/w_2 = 7.4$ $(= w_1/w_3)$: likel
affected by model redundancy.

The difference between a w_1 of 0.881 and one of 0.787 is
our point is that this, clearly erroneous, model redundanc
set has affected the Akaike weights. The weights for th
redundant model included are not correct because the
twice (one time "disguised" as Δ_3). The effect on the we
here (but it could be), but they are wrong, and this could
adversely) calculations using the w_i (as $\hat{\pi}_i$), as for exampl
and unconditional variance calculations.

If the model redundancy was recognized, and we wa
should not), we could correct the situation by considering
having two subsets: Model g_1 is one subset; a second sub
g_2 and g_3. Given that we know that models g_2 and g_3 are
allocate prior weights, about which model is the expected
1/2 to each subset, and the 1/2 is further divided equally
subset. Thus, $\tau_1 = 0.5$, $\tau_2 = 0.25$, and $\tau_3 = 0.25$. Now w

$$w_i = \frac{\mathcal{L}(g_i|\underline{x})\tau_i}{\sum_{r=1}^{R} \mathcal{L}(g_r|\underline{x})\tau_r}$$

from Section 2.9 to compute correct Akaike weights for the
thus

$$w_1 \propto 1.0 \cdot \frac{1}{2}, \quad w_2 \propto 0.135335 \cdot \frac{1}{4}, \quad w_2 \propto 0.$$

The normalized (to add to 1) weights are 0.8808, 0.059
the sum of the weights for models g_2 and g_3 correctly add u
to be, for model averaging and unconditional sampling var
produce correct results when applied to the redundant set

This hypothetical example presumably would not oc
it serves to introduce the concept, and issue, of model re
of models considered. It is possible to have actual mode
is not careful in constructing the set of models consider

models structured into two or more subsets. Different sub
be specified in such a way that they have one key model i
adjustment terms are applied to get a sequence of models.
can be used with different types of adjustment terms. Sc
have this situation: The full set of models is given as tw
$\{g_1, g_2, g_3, g_4\}$ and $\{g_1, g_5, g_6, g_7\}$. If the full set of models
8 different models, then the redundancy of model g_1 is n
The situation can easily be rectified if it is recognized: eith
to 7, or compute the w_i from the Δ_i, for the models labele
differential priors, τ_i, as $\left\{\frac{1}{14}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}\right\}$ and $\left\{\frac{1}{14}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}\right\}$. It
there might be a partial model redundancy in the models

Failure to completely understand the models used car
dundancy. For example, a logistic regression may be us
model of the probability of success,

$$p(x) = \frac{\exp(a + bx)}{1 + \exp(a + bx)},$$

where the parameters are a and b. However, this mod
expressed as,

$$p(x) = \frac{1}{1 + \exp[-\{(x - d)/c\}]},$$

where $a \equiv d/c$ and $b = -1/c$. The second model is jus
different model representation and a 1-to-1 reparametriz
forms where included in the model set, total redundancy
avoid such model redundancy.

To further illustrate model redundancy we conside
capture–recapture data, obtained on k capture occasions
population size. The parameters of such models are popu
capture probabilities (denoted by p), by occasion, anima
One possible type of model is model g_b under which there
ent capture probabilities: for first capture or for recaptur
the case where animals have a behavioral response to first
factors affect capture probability ($K = 3$).

A different model (g_t) allows capture probabilities t
occasion; so we have $p_1, p_2, \ldots, p_k$ ($K = k + 1$). Thus, w
different models. However, the model under which captu
vary by time allows for many submodels ($2^k - k$ possibl
the most general case). Some example (sub) models are

g_{t1} : $p_1 = p_2$, other p_i all differ ($K = k$),
g_{t2} : $p_1 = p_2 = p_3$, other p_i all differ ($K =$
g_{t3} : all $p_i = p$ ($K = 2$),
g_{t4} : all p_i are different ($K = k + 1$).

model redundancy becomes irrelevant because g_b is over
model: The usual Akaike weight for that model is here
claim, however, that the correct weights here should be ba
as $\left\{\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{2}\right\}$ not $\left\{\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right\}$. If the Δ_i are $\{2, 0,$
model redundancy is irrelevant (redundancy is only betw
others; there is no redundancy in models 1 to 4 if model
the result for the Δ_i is $\{2, 2, 2, 2, 0\}$, then model redund
deal as regards the proper w_i. For wrong priors $\left\{\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right.$
under correct priors $\left\{\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{2}\right\}$, $w_5 = 0.73$.

By adding submodels of the general time-specific model
we dilute the absolute strength of evidence for model g_b as
weights; and we must use such absolute weights in certain f
averaging). Inasmuch as these added models deal only w
capture probabilities, they are all of a type (hence, redun
evidence *against* model g_b), so they unfairly "gang up
which is a totally different type of model.

The appropriateness of unequal priors if submodels
model, g_t, are included is justified here on a theoretical b
mented in the capture–recapture literature that there is no
as regards estimating N, of considering constrained ver
time-specific model. Thus, the original set of two models
be augmented as above. Hence, in the last example we sho
these two models, and they have Δ_1 and Δ_2 as $\{2, 0\}$.
$w_2 = 0.73$. A key point here is that when we did have mo
use of the unequal priors did produce the correct Akaik
think that model redundancy can be coped with analyti
modification of the otherwise equal model priors, τ_i.

Even more important than accepting model redundancy
ifying model priors, is to construct the set of models to b
there is no model redundancy. As the above example ill
knowledge about the correct formulation and use of mo
at hand should be utilized in defining the a priori set of
Another point worth repeating is that neither the Δ_i nor th
of the models will be affected by model redundancy. Thu
models based on all models with Δ_i less than some cut
safest type to use. Our ideas on the cutoff value to use c
the distribution of Δ_p, but only for situations with no mo
recommendations already made on this matter were so de

The concept and issue of model redundancy was brou
by S. T. Buckland (personal communication); the above
Professor Buckland suggested that the bootstrap would
solution to model redundancy as long as for a given bo
forced to select one best model. This seems reasonable; b

we must be able to recognize model redundancy in our
can do that (we can and should), we think that redundanc
If model redundancy operates at a more subtle level than
bootstrap would have an advantage. We are currently disi
there will be a model redundancy problem as regards Ak
as the set of models considered is carefully constructed. (N
issue would be helpful.)

4.7 Recommendations

If data analysis relies on model selection, then inferences
model selection uncertainty. If the goal is to get the best
parameters in common to all models (this includes predic
ing is recommended. If the models have definite, and diffe
as regards understanding relationships among variables,
standing that is sought, then one wants to identify the b
inferences based on that model. Hence, reported parame
then be from the selected model (not model averaged val
when selecting a best model, also note the competing m
their Akaike weights. Restricting detailed comparisons to
confidence set on models should often suffice. If a single
supported, $w_{min} \geq 0.9$, and competing models give altern
should be reported. It may occur that the basic inference
from all good models. However, this is not always the cas
based on a single best model may not be sound if suppo
model is weak (in all-subsets selection when $R > 1{,}00$
small, e.g., < 0.01).

We recommend that investigators compute and report
sures of precision based on (4.9) when inference is bas
unless the Akaike weight w_i for the selected model is lar
an unconditional confidence interval, often the form $\hat{\theta} \pm 2$
an interval of this type back-transformed from a function
via the link function in general linear models. If such a
clear deficiencies, or in general if the computation can b
profile likelihood intervals based on formulas (4.10) and

If interest is really just on some parameters in commo
we recommend using model-averaged parameter estima
sampling variance estimate to use is then (4.9). Again, ofte
will suffice for a confidence interval.

We think that these analytical procedures can suffice,
tially use the bootstrap to evaluate model selection uncer

The bootstrap provides direct, robust estimates of model
ties π_i, but we have no reason now to think that use of bo
model selection probabilities rather than use of the Akaike
superior unconditional sampling variances or model-ave
timators. The primary purpose of the bootstrap is to asse
inferences; therefore, we recommend that the point estima
MLEs from the selected model (not the bootstrap means)
very complex, where there may be no suitable analytical
tors of conditional (on model) sampling variances, the boo
to get conditional and unconditional measures of precisi
that more bootstrap samples be used than is commonly
for really reliable results, but even 400 would be better th
model selection uncertainty (no assessment has often bee

Be mindful of possible model redundancy. A carefull
a priori models should eliminate model redundancy pro
tral part of a sound strategy for obtaining reliable infere
to statistical tests of post hoc null hypotheses, the assoc
decisions concerning supposed "significance" are not val

The theory here applies if the set of models is a priori to
any models considered have been included after some an
model(s) are suggested by the data, then theoretical resu
formulas) might fail to properly apply (in principle, the b
used). Even for such data-driven model selection strate
assessing model selection uncertainty rather than ignorin

4.8 Cement Data

We return to the cement data of Section 3.2 to compare bo
model selection frequencies (π_i), Δ_i values, Akaike weig
ditional estimation of sampling variances. These quanti
in Table 4.1; the AIC_c-selected model is shown there in
seven models are not shown in Table 4.1 because they w
the 10,000 bootstrap samples (also, they have virtually ze
The three simple approaches shown in Table 4.1 provide
model selection uncertainty for this very small ($n = 1$
model {12} is indicated as the best by all approaches. H
model selection uncertainty is evident because that best
weight of only 0.57 and a bootstrap selection probability
proaches cast substantial doubt concerning the utility of t
models in Table 4.1. Model {34} is particularly unsupport
and $w_i = 0.0004$.

Δ_i and derived Akaike weights computed from the data.

Model	K	$\hat{\pi}_i$	Δ_i	w_i
{12}	**4**	**0.5338**	**0.0000**	**0.567**
{124}	5	0.0124	3.1368	0.118
{123}	5	0.1120	3.1720	0.116
{14}	4	0.2140	3.3318	0.107
{134}	5	0.0136	3.8897	0.081
{234}	5	0.0766	8.7440	0.007
{1234}	6	0.0337	10.5301	0.002
{34}	4	0.0039	14.4465	0.000

Evidence for the importance of each variable can be o
bootstrap and tallying the percentage of times that each
the AIC_c selected model (Section 4.4). For the 10,000 b
occurred in 93% of the models, followed by x_2 (76%), x_3 (
Again, this simple approach indicates the importance of x_1
and x_4. Similar evidence can be obtained by summing the
those models with a particular variable present. Using th
the relative support of the four variables is as follows: x
x_3 (21%), and x_4 (32%). Considering the small sample
bootstrap and Akaike weights seem to give similar result

Using the idea of the pivotal Δ_p (Section 4.5) to obta
tribution as Δ_p^*, we find that an approximate 90% confi
$\Delta_i < 8.75$, while a 95% set is achieved if $\Delta_i < 13.8$. Th
percentile values of Δ_i are quite extreme here because th
example is so small ($n = 13$).

Using the bootstrap selection frequencies ($\hat{\pi}_i$), models
represent an approximate 86% confidence set, while add
flects an approximate 94% confidence set of models. Us
(w_i), an approximate 90% confidence set includes model
and {14}. The Δ_i values suggest that the final three mod
little utility. These types of ranking and calibration mea
available under a hypothesis testing approach or cross-va

We now illustrate the computation of unconditional es
first for a parameter in common to all models. What if one
value $\hat{E}(Y_0)$, denoted for simplicity by $\hat{Y}_0$, given the valu
$x_3 = 10$, and $x_4 = 20$ (cf. Table 3.1)? The prediction un
models of Table 4.1 is shown in Table 4.2; we used PROC
SAS to easily compute predicted values and their conditi
$\widehat{se}(\hat{Y}_0 \mid g_i)$. Clearly, $\hat{Y}_0$ is high for model {234}, relative to t
estimated standard error for model {1234} is very high, a

model measures of precision are given for Y_0; Y denotes a model-av
and $\hat{Y}_0 - \overline{\hat{Y}}$ is the estimated bias in using a given model to estimate

Model	K	$\hat{Y}_0$	$\widehat{\text{se}}(\hat{Y}_0 \mid g_i)$	$\widehat{\text{var}}(\hat{Y}_0 \mid g_i)$	$(\hat{Y}_0 - \overline{\hat{Y}}_{\text{boo}}$
{12}	**4**	**100.4**	**0.732**	**0.536**	**4.26**
{124}	5	102.2	1.539	2.368	0.07
{123}	5	100.5	0.709	0.503	3.86
{14}	4	105.2	0.923	0.852	7.48
{134}	5	105.2	0.802	0.643	7.48
{234}	5	111.9	2.220	4.928	89.01
{1234}	6	101.6	5.291	27.995	0.74
{34}	4	104.8	1.404	1.971	5.45

because the $\boldsymbol{X}$ matrix is nearly singular. Both of these m
little support, as reflected by the small relative weights, s
under these fitted models is of little credibility.

The predicted value for the AIC_c-selected model is 100
conditional standard error of 0.73. However, this measu
underestimate because the variance component due to m
tainty has not been incorporated. Model averaging (4.2) r
value of 102.5 using the bootstrap estimated weights ($\hat{\pi}_i$)
Akaike weights (w_i) (4.1). The corresponding estimated
dard errors are 3.0 using the bootstrap-based weights and
weights. These unconditional standard errors are substan
conditional standard error of 0.73. In Monte Carlo studi
find that the unconditional standard errors better reflect th
the predicted value, and conditional confidence interval co
near the nominal level (Chapter 5).

Study of the final three columns in Table 4.2 above sho
in the model-specific predictions (i.e., the $\hat{Y}_0$) from the
$(\hat{Y}_0 - \overline{\hat{Y}}_{\text{bootstrap}})^2$ or $(\hat{Y}_0 - \overline{\hat{Y}})^2$) is substantial relative to the
conditional on the model (i.e., the $\widehat{\text{var}}(\hat{Y}_0 \mid g_i)$). Models {
exceptions because they overfit the data (i.e., more paramet
The Akaike weights are relatively easy to compute co
required to obtain the bootstrap estimates $\hat{\pi}_i$; w_i seem pref
in this example. That is, we perceive no advantage here
based results, compared to the Akaike-weight-based resul
for the computational cost of the bootstrap (we do not clai
based results are any worse, just not better).

The investigator has the choice as to whether to use the
the AIC_c-selected model (100.4) or a model-averaged pre

errors (3.6 for the bootstrap and 1.9 for Akaike weights), of weights makes no great difference. However, there is uncertainty associated with this data set, and we would model-averaged predictions (when prediction is the obje Akaike weights. Thus, we would use 101.6 as the pred unconditional standard error of 1.9. If the AIC_c-selected m strongly supported by the data, then we might suggest u based on that (best) model (i.e., $\hat{Y}_0 = 100.4$) combined wit of the unconditional standard error (1.9), based on the Ak

The selected model includes only regressor variables model the estimated partial regression coefficients and th dard errors are $\hat{\beta}_1 = 1.4683$ (conditional $\widehat{\text{se}} = 0.1213$ (conditional $\widehat{\text{se}} = 0.0459$). Each of these parameters appe To compute the estimate of unconditional sampling vari find each model containing β_1, its estimate and conditio model, and the model's Akaike weight:

Model	$\hat{\beta}_1$	$\widehat{\text{se}}(\hat{\beta}_1 \mid g_i)$	w_i
{12}	1.4683	0.1213	0.5670
{124}	1.4519	0.1170	0.1182
{123}	1.6959	0.2046	0.1161
{14}	1.4400	0.1384	0.1072
{134}	1.0519	0.2237	0.0811
{1234}	1.5511	0.7448	0.0029
{1}	1.8687	0.5264	0.0000
{13}	2.3125	0.9598	0.0000.

The first step is to renormalize the w_i so they sum to models. Here that sum is 0.9925 before renormalizing, so the renormalized w_i, but they are the weights to use in appl The model-averaged estimate of β_1 is 1.4561 (from 4.1).

$$\widehat{\text{var}}(\hat{\theta}) = \left[\sum_{i=1}^{8} w_i \sqrt{\widehat{\text{var}}(\hat{\theta}_i \mid g_i) + (\hat{\theta}_i - \hat{\bar{\theta}})} \right.$$

For example, the first term in the needed sum is 0.0 $\sqrt{(0.1213)^2 + (0.0122)^2}$. Completing the calculation $(0.1755)^2$, or an estimated unconditional standard erro compared to the conditional standard error given the selec

{12}	0.6623	0.0459	0.6988
{124}	0.4161	0.1856	0.1457
{123}	0.6569	0.0442	0.1431
{1234}	0.5102	0.7238	0.0035
{234}	−0.9234	0.2619	0.0089
{23}	0.7313	0.1207	0.0000
{24}	0.3109	0.7486	0.0000
{2}	0.7891	0.1684	0.0000

When all 16 models are considered, the Akaike weigh
models above add to 0.8114. However, to compute res
these eight models we must renormalize the relevant Ak
to 1. Those renormalization Akaike weights are what a
model-averaged estimator of β_2 is 0.6110, and the unc
standard error of $\hat{\bar{\beta}}_2$ is 0.1206 (compared to the condit
of 0.0459). It is important here to compute and use un
errors in all inferences after data-based model selection.
conservative) confidence intervals on β_1 and β_2, using resu
should be constructed based on a t-statistic with 10 df (t_1
two-sided 95% confidence interval). Such intervals here
well away from 0; for example, the 95% confidence inte
0.93.

We generated 10,000 bootstrap samples of these dat
selection to all 16 models fit to each bootstrap sample. The
(and common belief about the bootstrap) it should be ac
the unconditional standard error of an estimated partial r
based on the standard deviation of the set of realized est
over all bootstrap samples, b, wherein the selected mod
x_1. The results are given below, along with the average va
estimate over all relevant bootstrap samples:

	bootstrap results	
parameter	average	st. error
β_1	1.461	0.760
β_2	0.453	0.958
β_3	−0.420	1.750
β_4	−0.875	1.237

From the selected model, {12}, we get $\hat{\beta}_1 = 1.47$ and $\hat{\beta}_2 =$
unconditional standard errors of 0.18 and 0.12, respective
analytical methods using Akaike weights.

Based on the above, and other comparisons not given,
bootstrap failed here when all 16 models were allowed t
noted in Section 3.2, the full design matrix, $\boldsymbol{X}$, for this

eigenvalues sum to 95.5% of the total. Also, the pairwise
x_4 is $r = -0.973$. This information, to us, strongly justif
one to drop model {1234} from consideration and to dr
in which both x_2 and x_4 appear. Thus without any mode
should, reduce the 16 possible models to 12 by eliminating
{234}, and {1234}. These sorts of considerations should b
do not compromise an a priori (as opposed to explorato
strategy.

With the reduced set of 12 models we computed the $\Delta_$
10,000 new bootstrap samples (to get $\hat{\pi}_i$ and bootstrap es
the results below (models not shown were never selec
samples):

Model	K	$\hat{\pi}_i$	Δ_i	w_i
{12}	4	0.5804	0.0000	0.6504
{123}	5	0.1315	3.1720	0.1332
{14}	4	0.2340	3.3318	0.1229
{134}	5	0.0465	3.8897	0.0930
{34}	4	0.0076	14.4465	0.0005

Applying here the method of Section 4.4 based on the
weights, we get the relative importance for the four vari
(0.9995), x_2 (0.7836), x_3 (0.2267), and x_4 (0.2164). Us
Section 4.3.2, especially (4.7) with the above Akaike we
unconditional standard errors as $\widehat{\text{se}}(\hat{\beta}_1) = 0.18$ (for $\hat{\beta}_1 =$
0.12) and $\widehat{\text{se}}(\hat{\beta}_2) = 0.046$ (for $\hat{\beta}_2 = 0.66$, $\widehat{\text{se}}(\hat{\beta}_2 \mid g_{\{12\}}) =$
strap estimates of unconditional standard errors are 0.34 a
$\hat{\beta}_2$, respectively.

The two different methods (analytical Akaike weights
agree for $\widehat{\text{se}}(\hat{\beta}_2)$ but not for $\widehat{\text{se}}(\hat{\beta}_1)$. The resolution of this
on two items. First, the correlation in the data of x_1 an
second, the sample size is only $n = 13$. As a result, fitte
{134} are very unstable over bootstrap samples as regard
For example, the sampling standard deviation (this estim
the 1,315 bootstrap values of $\hat{\beta}_{1,b}$ that resulted when mode
by AIC_c was 0.65 (the average of the bootstrap estimates
theory-based estimate is $\widehat{\text{se}}(\hat{\beta}_1 \mid g_{\{12\}})) = 0.12$.

There are several points we wish to make with this ex
sensitive to having demonstrably poor models in the se
ered; thus it is very important to exclude models tha
The analytical method (vs. the bootstrap method) of asse
standard errors seems more stable, as regards having or e

seemed reasonable even with all 16 models considered (t
precise when only the 12 models were used). With the re
the bootstrap results are still suspect, but now only becau
small ($n = 13$). Monte Carlo evaluation and comparison
needed before definitive statements about reliability will

4.9 Pine Wood Data

We consider here an example of only two simple linear
neither one nested in the other. This example has been use
(1995) on Bayesian model choice using Markov chain M
The data also appear elsewhere, such as in Efron (1984).
4.3) can be considered a trivariate response vector $(y, x$
$n = 42$. Variable y is the measured strength of a piece of
measured density of that wood, and z is the measured den
for the measured resin content of the wood. The scientif
of x or z is a better predictor of the wood strength y, base
either $y = a + bx + \epsilon$ or $y = c + dz + \delta$ (ϵ or δ a
from the expected linear model structure). Residuals are t
distributed and homogeneous under either model. Scient
that wood density adjusted for resin content should be
wood strength, but it takes more time and cost to measure

TABLE 4.3. Pine wood strength data y, wood density x, and wo
resin content z (from Carlin and Chib (1995)); $n = 42$.

y	x	z	y	x	z	y
3040	29.2	25.4	2250	27.5	23.8	1670
2470	24.7	22.2	2650	25.6	25.3	3310
3610	32.3	32.2	4970	34.5	34.2	3450
3480	31.3	31.0	2620	26.2	25.7	3600
3810	31.5	30.9	2900	26.7	26.4	2850
2330	24.5	23.9	1670	21.1	20.0	1590
1800	19.9	19.2	2540	24.1	23.9	3770
3110	27.3	27.2	3840	30.7	30.7	3850
3160	27.1	26.3	3800	32.7	32.6	2480
2310	24.0	23.9	4600	32.6	32.5	3570
4360	33.8	33.2	1900	22.1	20.8	2620
1880	21.5	21.0	2530	25.3	23.1	1890
3670	32.2	29.0	2920	30.8	29.8	3030
1740	22.5	22.0	4990	38.9	38.1	3030

for model g_x (i.e., y vs. x) and $\hat{\sigma}^2 = 73{,}611$ for model g
$K = 3$. The AIC_c values are 493.97 for model g_x and 4
The latter model being the estimated K-L best model, w
predictor. The two Δ_i are 0 and 17.01 for models g_z and
corresponding Akaike weights are 0.9998 and 0.0002.

The context in which we developed guidelines about i
having more complexity and more models than here. Ther
be considered overwhelming evidence for the superiority
matter of interpretation of the strength of evidence is unce
for this model selection problem we recommend applying
as using the above analytical results. The bootstrap is quit
more so than in more complex model selection situation

Based on 10,000 bootstrap samples of the data in Ta
AIC model selection we obtained results as follows. The
distribution of Δ_p^* gives us estimated percentiles for the s
of Δ_p as 1.29 (95th percentile), 2.56 (96th percentile), 4.
9.84 (99th percentile), and 17 is at about percentile 99.85.
are thus consistent here with the bootstrap sampling distr
gives us a basis to interpret Δ_i as regards the plausibi
actually the K-L best model for the data. However, when w
selection relative frequencies from the bootstrap we find
93.8% of the 10,000 bootstrap samples. This is still stro
of variable z as the better predictor.

It is clear that we select, based on strong evidence,
$\text{E}(y) = c + dz$ as the better model. The estimates of the s
of this model, and their conditional standard errors, are
252.9), $\hat{d} = 183.3$ ($\widehat{\text{se}} = 9.3$). We do not have a way we w
to compute unconditional standard errors when a paramet
model (which d definitely is). However, when the eviden
of the selected model, such as here, it is reasonable to ac
only that model, hence act as if that model would always
data, in which case conditional standard errors apply. So h
the conditional standard errors as a measure of estimato
of thumb we will hazard the suggestion that if the select
weight ≥ 0.90, it is acceptable to use the conditional
exact value (i.e., 0.90) is not critical; the concept is that if
support the selected model strongly enough ($w_{min} \geq 0.9$
rule of thumb; also see Royall (1997), where a similar ru
conditional and unconditional standard errors will be near
like this if one is bothered by the issue here of using condi
data-based model selection, a modest simulation study ca
that issue, as well as other matters.

introduce one use of simulation here, namely generating
closely mimic the apparent nature of the real data. This
model selection in a case like these pine wood data when
generated the data. Generally, we would be against such
as having relevant applicability to AIC model selection is
the issue is clearly one of just deciding between two lin
complex data simulation models than used below do not

We proceed by considering $(y, x, z)'$ as a trivariate nor
with mean vector $\underline{\mu}$ and variance–covariance matrix Σ
diagonal matrix D has as its diagonal the marginal stand
x, and z, and C is the matrix of correlations

$$\begin{bmatrix} 1 & \rho_{yx} & \rho_{yz} \\ \rho_{xy} & 1 & \rho_{xz} \\ \rho_{yz} & \rho_{zx} & 1 \end{bmatrix}.$$

From the data we obtain $\hat{\underline{\mu}}' = (2992, 27.86, 26.79)$; the
standard deviations are 894.60 ($= \hat{\sigma}_y$), 4.4946 and 4.6475
estimates are $\hat{\rho}_{yx} = 0.9272$, $\hat{\rho}_{yz} = 0.9521$, and $\hat{\rho}_{xz} =$
a simulated observation mimicking the data we generat
standard normal random variables (i.e., normal(0,1)), say

$$(y, x, z)' = \underline{\mu} + DC^{0.5}\underline{v}$$

for some parameter choices "near" the estimated parame
software packages that will find the needed "square root"
MATLAB, Anonymous 1994).

The best model here is the one that has the smaller res
of y, given the predictor. Those true residual variances are
and $\sigma^2_{y|z} = \sigma^2_y(1 - \rho^2_{yz})$. Therefore, in the simulation,
predicting y is the one with the biggest correlation coeffic
What we cannot determine without simulation is perfor
model selection method.

We can tell from theory that only the values of ρ_{yx}, ρ_{yz},
selection performance, including the distribution of Δ_p
selection frequencies, hence Akaike weights. Hence, par
and D are irrelevant to that aspect of the problem (we migh
about values for $\underline{\mu}$ and D if we wanted a realistic evaluatio
bias on parameter estimators). Therefore, in the simulatio
set $\underline{\mu} = \underline{0}$ and $D = I$. Using these values results in $a =$
and $d = \rho_{yz}$, and knowing these as truth, we can infer relat
selection bias that might occur.

We generated 10,000 simulated observations for the
efficients being at their estimated values (as truth) and t

of $\rho_{yz} = \rho_{yx} = 0.825$ as in the real data and a final case w
correlation coefficients used in the simulation were set t
objective was to determine the relative frequency of m
(this is without loss of generality since $\pi_z + \pi_x = 1$) and
weight $\mathrm{E}(w_z)$ (also without loss of generality since $\mathrm{E}(w_x$
the 95th and 99th percentiles of Δ_p (denoted below by
Results are given below by assumed sets of correlation co
proportions have coefficients of variation of about 1%; th
coefficients of variation more like 2.5%, and this for 10,0

ρ_{yx}	ρ_{yz}	ρ_{xz}	π_z	$E(w_z)$	$\Delta_{p,0.95}$
0.927	0.952	0.958	0.97	0.96	0.0
0.927	0.952	0.900	0.90	0.89	4.6
0.927	0.952	0.980	0.99	0.99	0.0
0.900	0.925	0.958	0.92	0.90	1.8
0.900	0.925	0.900	0.84	0.83	7.6
0.950	0.950	0.950	0.50	0.50	18.3

In these cases it is clear that $\pi_z \approx \mathrm{E}(w_z)$, and that the sam
Δ_p is quite variable. This is a worse case as regards variabil
of Δ_p (only two models, and they are nonnested). We ca
case above were reality, we would expect to select the co
96% of all samples (for $n = 42$). These Monte Carlo r
faith in the usefulness of the bootstrap results based on the
faith in the strength of evidence deduced here from the d
and from $w_z = 0.9998$.

By looking at the more detailed results (not given here)
estimated parameters, $\hat{\theta}$, and their averaged estimated stan
given the selected model we can assess model selectio
estimates and standard errors. If case 1 above were truth, th
suggested there would be little model selection bias here w
and conditional standard errors applied (to be expected if
96% of the time). When model g_x was selected, no strong
estimators were suggested, but the sample size for this
$m = 314$.

There is another interesting question we can explore w
results. When a model is selected, right (g_z) or wrong (g_x
we then judge the weight of evidence to be strongly in f
Our interest in such a question is mostly focused on when
choice (we will not know this to be the case): Having pick
not the K-L best model, will the data appear strongly to
model as being best, or will the evidence be weak? For ca
the wrong choice with sampling probability only about

evidence in favor of the wrong model. Thus only 1% of
would be strongly misleading in this simulated scenario.
9,688 samples wherein model g_z was selected, 9,223 produ
($w_z > 0.9$) in favor of the selected model. Hence we ex
all samples (in this particular scenario) we would select th
do so with convincing evidence. Note however, that we
actual data whether it is one of the "1%" strongly mislead
say, again just for this simulated scenario, that the *estima*
that we have reached a correct conclusion for these pine

4.10 The Durban Storm Data

Linhart and Zucchini (1986:176–182) apply AIC to storm
the Botanical Gardens in Durban, South Africa. The detai
their Table 10.1. By seven-day periods in the year ("weel
1 January, they obtained the frequency of weeks with at l
occurring. For example, in 47 consecutive years of data, f
were 6 years with at least one storm event. The data are
definition of a storm: "a rainfall event of at least 30 mm i
and Zucchini 1986:176). We use here their period I data y
to 52), wherein for the first 22 weeks the sample size of y
weeks $23 \le i \le 52$, $n_i = 48$. Thus, under a fixed effects a
here is 2,474. The data are from January 1932 to Decemb
as did Linhart and Zucchini (1986), the minor matter of a
to have 8 days (such as 26 February to 4 March when
Listed in order $i = 1$ to 52, the data y_i are

6, 8, 7, 6, 9, 15, 6, 12, 16, 7, 9, 6, 8, 2, 7, 4, 4, 3, 3, 10
4, 0, 2, 0, 3, 1, 1, 5, 4, 3, 6, 1, 8, 3, 4, 6, 9, 5, 8, 6, 5, 7

Conceptually, there exists a probability p_i of a storm at t
Gardens in week i. Based on these data, what is a "good" e
That was the analysis objective of Linhart and Zucchini,
our objectives. Our other objective is to reliably assess the u
A simple estimator is $\hat{p}_i = y_i/n_i$; it is very nonparsimoni
and (most seriously) fails to be a smooth, hence informati
trends in the true p_i. We expect that anyone considering
strongly believe that the p_i would have a considerable de
as a function over the 52 weeks. Therefore, we want to f
for a not-large number of parameters represented by $\underline{\theta} =$

they construct a likelihood by treating the y_i as a set of in
random variables on sample sizes n_i for parameters p_i, a
model as

$$\text{logit}(p_i) = \log(p_i/(1-p_i)) = \sum_{j=1}^{K} \theta_j z$$

being some suitable linear model on $\underline{\theta}$, for known "covaria
this is a type of logistic regression (we consider theory for
in this situation in Section 7.6.6). Linhart and Zucchini u
series model for the z_{ji} and used TIC for model selection
essentially the same as AIC). We extend their example
model averaging; also, we compute unconditional confide
p_i. Here, K is the number of structural parameters in the m

The structure of the simplest model, model g_1, is given

$$\text{logit}(p_i) = \theta_1, \qquad i = 1, \ldots, 52.$$

For model g_2:

$$\text{logit}(p_i) = \theta_1 + \theta_2 \cos\left(\frac{2\pi(i-1)}{52}\right) + \theta_3 \sin\left(\frac{2\pi(i-1)}{52}\right)$$

For model g_3:

$$\text{logit}(p_i) = \theta_1 + \theta_2 \cos\left(\frac{2\pi(i-1)}{52}\right) + \theta_3 \sin\left(\frac{2\pi(i-1}{52}\right.$$
$$+ \theta_4 \cos\left(\frac{4\pi(i-1)}{52}\right) + \theta_5 \sin\left(\frac{4\pi(i-1)}{52}\right)$$

In general, the structure for model g_r (wherein $K = 2r$) i

$$\text{logit}(p_i) = \theta_1 + \sum_{j=1}^{r-1} \left[\theta_{2j} \cos\left(\frac{2j\pi(i-1)}{52}\right)\right.$$
$$\left. + \theta_{2j+1} \sin\left(\frac{2j\pi(i-1)}{52}\right)\right], \qquad i$$

Assuming conditional binomial variation and independe
likelihood for any model is

$$\mathcal{L}(\underline{\theta}) \propto \Pi_{i=1}^{52} (p_i)^{y_i} (1-p_i)^{n_i - y_i}.$$

Given the model for $\text{logit}(p_i)$ as a function of $\underline{\theta}$, say h
compute p_i as

$$p_i = \frac{1}{1 + \exp[-h_i(\underline{\theta})]}.$$

$\hat{c} = 1.4$), the df of the goodness of fit χ^2 is $52 - K + 1$.

Model	K^a	$\log(\mathcal{L})$	Δ-AIC	χ^2	P	
1	2	−863.24	62.66	131.4	0.000	2
2	4	−833.83	7.85	76.5	0.007	1
3	6	−829.17	2.53	69.3	0.019	1
4	8	−826.37	0.93	61.2	0.054	1
5	10	−823.91	**0.00**	55.6	0.094	1
6	12	−823.89	3.95	55.6	0.064	1
7	14	−823.40	7.04	54.7	0.049	1
8	16	−822.76	9.70	54.0	0.035	1
9	18	−822.47	13.11	53.8	0.022	1

[a]The number of structural parameters plus 1 for $\hat{c}$ for QAIC; the number of
example is $K - 1$.

The independence assumption may not be true, but it s
badly wrong. Similarly, the count y_i may not be the sum
neous Bernoulli events over the n_i years. Truth may corr
to having varying year-to-year weekly probabilities of a
to cope with these types of model inadequacies is to us
likelihood theory, hence to use a variance inflation factor
computed from the global model goodness-of-fit chi-squa
freedom df. Then we use QAIC, rather than AIC; also, c
variances based on assumed models are multiplied by $\hat{c}$ (

Following Linhart and Zucchini (1986) we consider sev
over which model uncertainty and model averaging are c
g_7, $K = 14$. We obtained MLEs for these models by usin
(SAS Version 6.12); it is easy to adapt PROC NLIN to pr
(see, e.g., Burnham 1989). In this example, it is not clea
should serve as the basis for estimation of the variance infla
several models were explored (Table 4.4), and estimates
stable at about 1.40. For each fitted model we also com
square goodness-of-fit statistic, its significance level (P-v
purpose of a more thorough consideration of model fit w
($K = 16$) and g_9 ($K = 18$). Table 4.4 gives basic resu
models: K, $\log(\mathcal{L})$, Δ-AIC, χ^2 goodness-of-fit, and corre
$\hat{c}$, and Δ-QAIC. The values of Δ-QAIC are for when mo
global model.

The $\log(\mathcal{L})$ values in Table 4.4 for models 1 through 7
Linhart and Zucchini (1986) in their Table 10.3 (they d
and g_9). The AIC-selected model has 9 parameters, and
the results of Linhart and Zucchini (on their page 182). T

nor their estimated conditional standard errors (standard li
applied to obtain the large-sample variance–covariance m
Rather, our goal is to estimate well the set of p_1 to p_{52},
parameters in common to all models.

4.10.2 Consideration of Model Fit

Before we accept the AIC-selected model, we must consid
model fits. Based on the results in Table 4.4, the global mod
fit to the data: $P = 0.049$. More importantly, $\hat{c} = 1.4$ on
greater than 1 that we should not accept results of AIC
require $c \approx 1$. Even the AIC-selected model has $\hat{c} = 1.$
even though this model is deliberately selected to fit w
issue further we fit two more models; models g_8 and g_9 als
If the problem was an inadequate structural model, we w
g_8 and g_9, compared to model g_7, to improve. The resul
strongly suggest that there is extrabinomial variation in th
a result is common for real count data such as these, as is
$1 < \hat{c} <\approx 2$).

However, before automatically resorting here to QAIC,
worth noting. The expected counts from the models fitted
(i.e., the data are sparse in the sense of being small co
for model g_5, $\hat{\mathrm{E}}(y_{26})$ to $\hat{\mathrm{E}}(y_{32})$ are about 1.5; these are th
expected count values here; the largest estimated expect
10. Perhaps even if the global model is structurally true, t
count values will invalidate the usual central chi-square nu
goodness-of-fit statistic.

We explored this matter by Monte Carlo methods (also c
bootstrap method). We generated data based on truth bei
selected model g_5. That is, independent y_i^* were generated
based on fitted model g_5. For each such data set we then
computed the chi-square goodness-of-fit statistic to see wh
was noticeably different from that of a central chi-squar
only a sample of 100 such generated data sets because w
big effect: for this situation is $c = 1$ or 1.29?

The answer was clear: If the model truly fits, then on
$\hat{c} = 1$; i.e., the usual null distribution holds well here desp
average of the 100 χ^2 goodness-of-fit values was 41.9 (th
The largest and smallest of the 100 values were 71.6 (P
($P = 0.997$); these are not unusual for a sample of 100
When each test statistic was converted to a P-value, the
fit a uniform (0, 1) distribution. Finally, the average of the
$\hat{c}$ was 0.98 ($\widehat{\mathrm{se}} = 0.21$). While the possibility remains that

based general statistical wisdom for real data supports the
accept that extra binomial variation often exists in coun
will use QAIC, not AIC, with $\hat{c} = 1.4$ as our basis for mo

The number of estimable parameters must be augmente
need to estimate the variance inflation factor (c); if mor
inflation factor is estimated, then the number of such esti
be included in K. If the estimate of c is close to 1, then
is necessary and K should not be increased. If one beli
concern about overdispersion and, therefore takes $c = 1$,
not be incremented.

When sufficient precision is used in the calculations, w
is the QAIC best model, although for practical purposes
tied for best (and model g_5 is almost as good, based on QA
a plot of the fitted $\hat{p}_i$ for both models g_3 and g_4. Also shown
95% confidence bands on p_i based on $\hat{p}_{L,i}$ and $\hat{p}_{U,i}$ for e
explain the calculation of these confidence intervals.

4.10.3 Confidence Intervals on Predicted Storm

Basically, $\hat{p}_{L,i}$ and $\hat{p}_{U,i}$ arise as back-transformed lower a
limits on logit(p_i). However, we used SAS PROC NLIN to
estimated MLE-based theoretical $\widehat{\text{se}}_t(\hat{p}_i \mid g)$ that is compu
The first step is then to form the correct (inflated) estim
$\sqrt{\hat{c}} \cdot \widehat{\text{se}}_t(\hat{p}_i \mid g) = 1.183\, \widehat{\text{se}}_t(\hat{p}_i \mid g) = \widehat{\text{se}}(\hat{p}_i \mid g)$. The inter
could be used. However, it is better to use here what is bas
back-transformed logit-based interval (Burnham et al. 19

$$\hat{p}_{L,i} = \frac{\hat{p}_i}{\hat{p}_i + (1 - \hat{p}_i)C},$$

$$\hat{p}_{U,i} = \frac{\hat{p}_i}{\hat{p}_i + (1 - \hat{p}_i)/C},$$

where

$$C = \exp\left[\frac{t_{\alpha/2,df}\, \widehat{\text{se}}(\hat{p}_i \mid g)}{\hat{p}_i(1 - \hat{p}_i)}\right]$$

(acceptable as long as $\hat{p}_i$ does not get too close to 0 or
bands in Figure 4.1 were computed in this manner and are
the model. We used $t_{\alpha/2}$, df $= 39$, because the df that app
$\hat{c}$; thus here df $= 39$.

Estimates of unconditional standard errors require the
the bootstrap), in this case based on Δ_i from QAIC (Tab
$w_1, \ldots, w_7$ are (from Section 2.6).

0.0000, 0.0833, 0.3149, 0.3149, 0.2465, 0.034

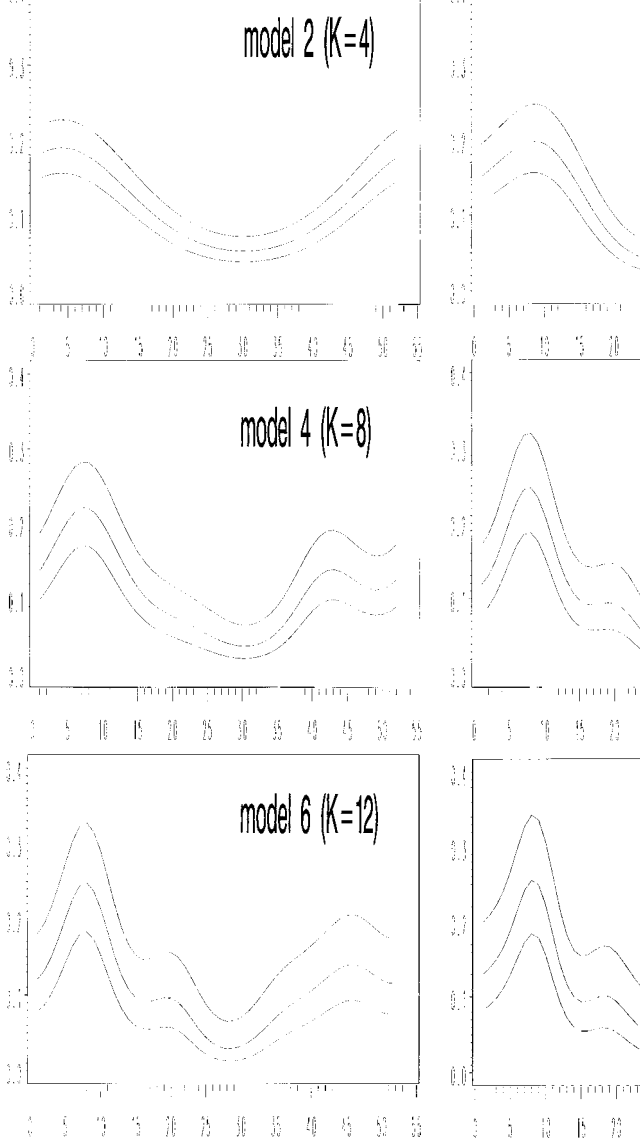


FIGURE 4.1. Plots of the predicted probability of one or more storm
from models g_2 to g_7 fitted to the Durban storm data from Linhart a
text for details). Also shown are approximate 95% confidence band
conditional on the model (see text for details).

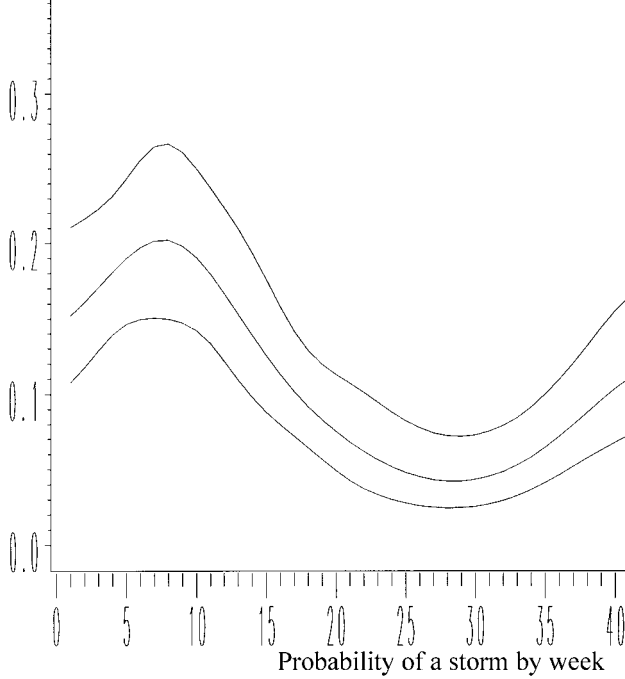


FIGURE 4.2. Plot of the model-averaged (from models g_2 to g_7) one or more storms per week, $\hat{p}_{a,i}$, from the Durban storm data fro (1986) (see text for details). Also shown are approximate 95% c these bands include model selection uncertainty.

Then for each week we find $\hat{p}_i$ under models g_1 to g_7 and a model-averaged $\hat{p}_{a,i}$. Next we apply (4.9) to obtain the un this unconditional standard error also applies to $\hat{p}_i$ from as well as to $\hat{p}_{a,i}$. In cases like this where the $\underline{\hat{\theta}}$ are not recommend that the $\hat{p}_i$ to use are the model-averaged va here we use the unconditional $\widehat{\text{se}}(\hat{p}_{a,i})$, and we base conf above formula for $\hat{p}_{L,i}$ and $\hat{p}_{U,i}$, however, for C based on $\widehat{\text{se}}$ $\hat{p}_{a,i}$, $\hat{p}_{L,i}$, and $\hat{p}_{U,i}$ are shown in Figure 4.2.

4.10.4 Comparisons of Estimator Precision

We now give some considerations about (estimated) sta ferent versions of the $\hat{p}_i$. Common practice would be to use the standard errors conditional on the model. In this c using model g_4 (by the slimmest of margins). We comput

$$r_i(\text{se}) = \frac{\widehat{\text{se}}(\hat{p}_i \mid g_4)}{\widehat{\text{se}}(\hat{p}_{a,i})}, \qquad i = 1, \ldots, 5$$

and

$$r(\overline{\text{se}}) = \frac{\sum \widehat{\text{se}}(\hat{p}_i \mid g_4)}{\sum \widehat{\text{se}}(\hat{p}_{a,i})}.$$

These ratios are less than 1 if the unconditional is larger
standard error (the notation used for these ratios has no s
just need to represent them somehow).

We obtained $0.78 \leq r_i(\text{se}) \leq 1.02$ and $r(\overline{\text{se}}) = 0.9$
unconditional standard errors are on average 1.11 times
that are conditional on the model, and hence ignore mod
we note that the average of the 52 values of $\widehat{\text{se}}(\hat{p}_{a,i})$ w
$\widehat{\text{se}}(\hat{p}_{a,i}) \leq 0.035$). This is good absolute precision; the act
the 52 confidence intervals was 0.084.

An alternative that avoids model selection is to use
parameter-saturated model is not very useful. Estimated
der this model are given by $\widehat{\text{se}}_s(\hat{p}_i) = \sqrt{\hat{p}_i(1-\hat{p}_i)/n_i}$
any $\hat{c}$ is used here since there is no basis on which
ance inflation factor given this model). We computed an
$\left[\sum \widehat{\text{se}}_s(\hat{p}_i)\right] / \left[\sum \widehat{\text{se}}(\hat{p}_{a,i})\right]$ and considered the separate
We obtained

$$\frac{\sum \widehat{\text{se}}_s(\hat{p}_i)}{\sum \widehat{\text{se}}(\hat{p}_{a,i})} = 2.31;$$

so on average the unconditional standard errors of the
were more precise by a multiplicative factor $0.433 = 1/2$
much less useful parameter-saturated model estimates. Al
$0 \leq \widehat{\text{se}}_s(\hat{p}_i) \leq 0.082$ (and a variance estimate of 0 is qu
$0.012 \leq \widehat{\text{se}}(\hat{p}_{a,i}) \leq 0.035$; thus the standard errors for
$\hat{p}_{a,i}$ are much more stable than is the case for the parame

Linhart and Zucchini (1986) used TIC for model sele
only difference between the two methods is the use of $\text{tr}(J$
estimating this trace term, which Linhart and Zucchini de
fact, in Section 7.6.6 we demonstrate by theory and exa
very near K unless the structural model is truly terrible.
it is better for count data (and generally simpler) to use t
than estimate this quantity. Linhart and Zucchini (1986:1
their Table 10.3. For models g_1 to g_7 the ratios of the est
K are 0.95, 0.97, 0.97, 0.98, 0.98, 0.98, 0.98. Not only a
to 1, but they are all less than 1 (see Section 7.6).

Young and Young (1998, 510–514) give as an example t
mortality of flour beetles (*Tribolium confusu*) caused by
hour exposure to gaseous carbon disulfide (CS_2); the data
Bliss (1935). Table 4.5 gives the basic data, as dose level
tested, and the number that died as an immediate causal resu
we take sample size here as total beetles tested, hence n
mortality rate (Table 4.5) increases with dosage in a rou
smooth, sigmoid form. It is typical to fit a parametric
smooth such data, hence to get a simple estimated dose
confidence bounds, and to allow predictions outside the
extrapolation beyond the range of applied dose levels re
response model; classically only one model was used. As
examine multimodel prediction of mortality at dose-level

A generalized linear models approach may easily, and a
to model the probability of mortality, π_i, as a function o
probability distribution assumed for the data is binomial:
at dose level x_i the response random variable y_i (number
be distributed as binomial(n_i, π_i). Therefore, the likeliho

$$\mathcal{L} = \prod_{i=1}^{8} (\pi_i)^{y_i} (1 - \pi_i)^{n_i - y_i}.$$

Within this setting a model means some parametric form
that $0 < \pi(x) < 1$ is maintained. Moreover, in the context
models there must be a nonlinear transformation (i.e., link
give a linear structural model in the parameters. There is
here: as dose increases, modeled mortality must not decre
ity). There are several commonly used forms for such a

TABLE 4.5. Flour beetle mortality at eight dose levels of CS_2 (f
1998, their Table 14.1).

Dose (mg/L)	Number of Beetles tested	killed	Observ mortality
49.06	49	6	0.12
52.99	60	13	0.22
56.91	62	18	0.29
60.84	56	28	0.50
64.76	63	52	0.83
68.69	59	53	0.90
72.61	62	61	0.98
76.54	60	60	1.00

We consider three commonly used generalized linear m
link functions: logistic, hazard, and probit (all are implem
LOGISTIC, SAS Institute Inc., 1985). The logistic mode

$$\pi(x) = \frac{1}{1 + e^{-(\alpha+\beta x)}},$$

with link function

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \text{logit}(\pi(x)) = \alpha + \beta$$

The hazard model and associated complementary log-log

$$\pi(x) = 1 - e^{-e^{(\alpha+\beta x)}},$$

and

$$\log[-\log(1-\pi(x))] = \text{cloglog}(\pi(x)) = \alpha$$

The cumulative normal model and associated probit link
NORMIT) are

$$\pi(x) = \int_{-\infty}^{\alpha+\beta x} \left[\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\right] dz \equiv \Phi(\alpha +$$

and

$$\Phi^{-1}(\pi(x)) = \text{probit}(\pi(x)) = \alpha + \beta x$$

Here, $\Phi(\cdot)$ denotes the standard normal cumulative prob
which does not exist in closed form.

Traditionally, the predictor variable x can be either dose
dosages range over an order of magnitude or more, henc
relatively near 0 compared to the largest dose, then log(dose
used because then $\hat{\pi}(0) = 0$ will apply. Otherwise, wi
above one risks getting $\hat{\pi}(0) > 0$, when in fact $\pi(0) = 0$
recording acute mortality over a time interval so short that
will occur. However, dose levels here are tightly clustered
either form of x is plausibly appropriate. One approach a
to use models with both forms of x, thus six models.

A priori we do not know whether the simple model st
suffice to fit the data. Thus one might be motivated to ex
model to be, say, $\text{link}(\pi(x)) = \alpha + \beta x + \gamma f(x)$ where $f($
(or log(dose) if x = dose). Clearly, there is no unique m
us consider six more models: the three link functions com
$\alpha + \beta x + \gamma x^2$, where x can be either dose or log(dose). On
without further thought. We do not recommend doing so
results for these 12 fitted models and then note further sim

AIC	Δ	weight	K	predictors	
366.536	0.000	0.19356	2	dose	
366.641	0.105	0.18366	2	logdose	
367.608	1.072	0.11325	3	dose,	$dose^2$
367.698	1.162	0.10826	3	logdose,	logdo
367.804	1.268	0.10267	3	logdose,	logdo
367.998	1.462	0.09318	3	dose,	$dose^2$
368.420	1.884	0.07546	3	dose,	$dose^2$
368.430	1.894	0.07508	3	logdose,	logdo
370.246	3.710	0.03028	2	dose	
371.313	4.777	0.01776	2	dose	
374.138	7.602	0.00433	2	logdose	
375.222	8.686	0.00252	2	logdose	

that, in fact, allow an a priori restriction to just the thre
link$(\pi(x)) = \alpha + \beta x$, for $x =$ dose.

Table 4.6 gives AIC results for the 12 generalized linea
above. The models are identifiable by the predictors they
function used. There is considerable model uncertainty,
Akaike weights. However, a first thing to note is that the ei
($\Delta < 2$) are "paired" on dose and log(dose), by link fun
predictors, in the sense that the two models of a pair hav
value. There is a logical reason for this, and it should be
which would have led to consideration of only 6 mod
log(dose) or dose.

For these data (Table 4.5) dosage is between 49 and 7
restricted interval log(dose) is almost perfectly linearly c
$\log_e(\text{dose}) = 1.35376 + 0.007001 \cdot \text{dose}$, with $r^2 = 0.99$
correlation justifies using either dose or log(dose), but n
for our models. In fact, to include here the six models ba
corresponding six models based on log(dose) is a form o
(Section 4.6), and we recommend against it. Thus, a priori
six models; it is our choice to keep things simple and just
on dose.

Of the remaining six models, a further a priori cons
eliminate the quadratic models such as logit$(\pi(\text{dose})) = \alpha$
These three models cannot be monotonic increasing in d
quadratic. They might fit the data well (they do), but they
at lower doses or decrease at high doses. Because the quad
be monotonic increasing in dose (unless $\hat{\gamma} = 0$ occurs), t
been used. A plot of a fitted model will reveal this fact, at
feel justified in eliminating that model.

366.536	0.000	0.80114	2	dose	
370.246	3.710	0.12534	2	dose	
371.313	4.777	0.07352	2	dose	

Table 4.7 shows the three models, of the 12, that ca
priori. The associated Akaike weights change, relative t
the model evidence ratios are invariant. For example, (
$10.90 = 0.19356/0.01776$ for the hazard vs. logistic mo
the predictor. Figure 4.3 shows plots of the three fitted m

Common analysis practice for these data would be t
models of Table 4.7: historically probit, but in recent de
recent analysis might have looked at all three models of
models of Table 4.6) and used AIC to select the best mo
have based inferences conditional on just that model. T
model uncertainty. Such uncertainty can be greater for ex
the range of doses used than at the actual doses. To il
inference (model averaging, here) we predict mortality a
($\hat{\pi}(40)$).

Because there is a causal relationship of dose and mo
beyond the data, while risky, is both reasonable and o
when extrapolation is done model deficiencies can be imp
monotonicity. Figure 4.4 gives plots of four fitted models
Table 4.7 and the third model in Table 4.6, the logit link on
latter model predicts increasing mortality as dose goes be
even though it provides a quite acceptable fit within the ran
Even if only discovered after the fact, we would use this
a basis for eliminating this model.

Table 4.8 gives the basic conditional and unconditional
models considered. The estimated best model has a weig
not overwhelming (the evidence ratio for the best vs. s
6.39). The point estimates of $\pi(40)$ from the three model
to 0.0308 and produce, at these extremes, nonoverlapping
(Table 4.8). This information should be reported if kno
critical. If a single best-point estimate is acceptable, use
$\hat{\pi}(40) = 0.0257$ with unconditional 95% confidence interv
Note that the unconditional standard error is 35% larger
standard error for the AIC best model. For a confidenc
based on the model-averaged $\hat{\pi}(40)(= 0.0257)$ and its un
error (0.01274), we used the logit-based interval of Burnh
(see also Section 4.10.3).

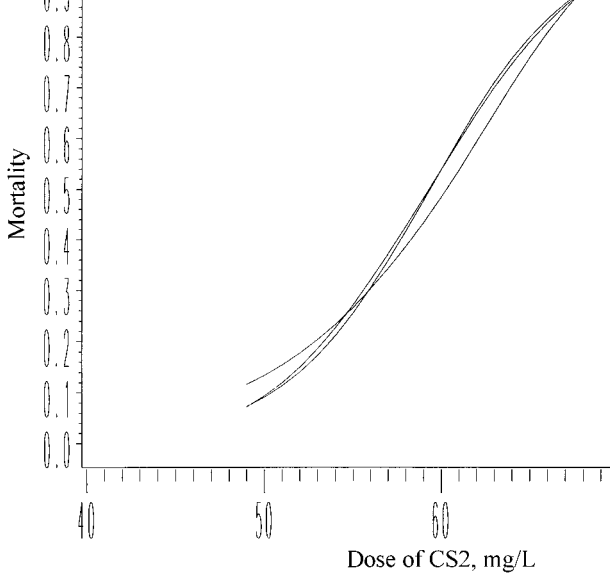


FIGURE 4.3. Fitted models of Table 4.7 plotted only over the range o
bit and logistic models are nearly identical compared to the slightly d
log-log model.

TABLE 4.8. Predicted mortality probability at dose 40mg/L; shown
on each of the three models and the unconditional model-averaged

Link function	Akalke weight	$\hat{\pi}(40)$	$\widehat{\text{se}}$	95% Confid lower
cloglog	0.8011	0.0308	0.00945	0.0168
probit	0.1253	0.0031	0.00226	0.0007
logit	0.0735	0.0085	0.00382	0.0035
model averaged:		0.0257	0.01274	0.0094

All three models used for Tables 4.7 and 4.8 fit the dat
no need for any overdispersion adjustment. A simple P
expected chi-square comparison suffices:

$$\chi^2 = \sum \frac{(O_j - \hat{E}_j)^2}{\hat{E}_j},$$

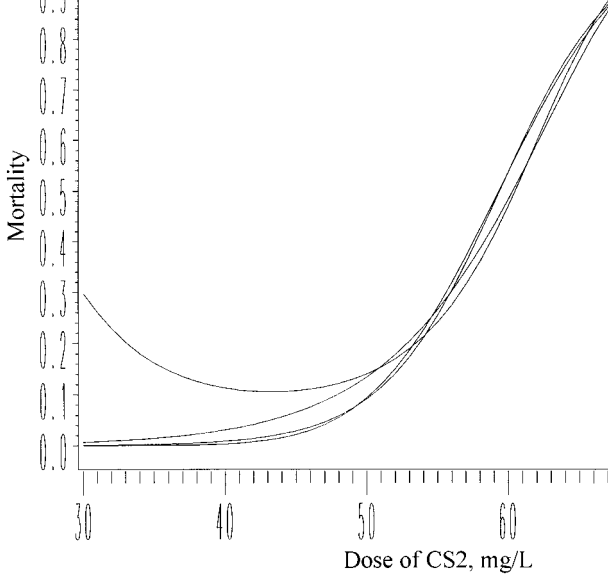


FIGURE 4.4. Four fitted models plotted beyond the range of dose l
tary log-log, probit and logit models of Table 4.7 (and Figure 4.3),
dose and dose^2 (Table 4.6), which wrongly predicts increasing mort
44 mg/L.

on 6 degrees of freedom ($= 8-2$ since each model has 2 es
For these binomially distributed data this chi-square stati

$$\chi^2 = \sum_{i=1}^{8} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}.$$

Goodness-of-fit results are as follows

model	χ^2	P
cloglog	3.49	0.74
probit	7.06	0.32
logit	7.65	0.26.

A possible outcome, not observed here, is the best n
worst model not fitting, say with $P = 0.01$ and $\hat{c} =$
worst-fitting model is not a global model, but is just an alt
dimension as the other models, we should attribute the fa
structural inadequacies, not overdispersion. Therefore, we
overdispersion adjustments). Overdispersion would result
three models, i.e., even the best model would not fit well.

dose caused the observed mortality, whereas for the stor
prediction. For that reason it sufficed with the storm data
function plus a series of increasingly general nested mode
a global model for that example. For the beetle data we c
that (1) the only predictor needed, or useful, is dose and
expected response must be imposed by any model. The is
thus reduced to one of an appropriate functional form, he
linear models framework, to what is the appropriate link
as a result, we have no global model, but rather several
natives for a best causal-predictive model (this situation
observational studies). We saw that this has implications
and overdispersion evaluation.

Moreover, an important issue here is prediction outsid
applied. It is well known that many models might fit the ol
occurred here, but give quite different extrapolated predicti
here. It is thus not a sound idea to pick a single model a
base extrapolated predictions on it when there is model u
ample illustrates how easy it is to compute model-average
the information from the fitted models, and that so doing
realistically cautious predictions.

4.12 Publication of Research Results

We provide an outline of how results under the informatio
might be presented in papers submitted for publication
Anderson et al. 2001d). The *Introduction* is the place
study objectives and note the degree to which the paper i
confirmatory.

Chamberlin's (1965) concept of "multiple working hyp
derlie the *Methods* section. This is the place to describe an
hypotheses and models in the set and how these relate spe
objectives. Ideally, one should be able to justify why a p
the set, as well as support the decision to exclude another
Avoid the routine inclusion of a trivial hypothesis or mod
all the models considered should have some reasonable
support. The *Methods* section should always provide suf
the reader can understand what was done.

A common mistake is the use of AIC rather than the se
AIC_c. Use AIC_c unless the number of observations is at lea
ber of explanatory variables (i.e., $n/K > 40$) for the most
model in the set. If using count data, some detail should

found, the log-likelihood should be computed as $\log(\mathcal{L})/\hat{c}$. When using QAIC or QAIC_c, one parameter must be
used in model selection, and the covariance matrix shou
$\hat{c}$. When using QAIC or QAIC_c, one parameter must be
of the number of estimated parameters in the model (K
the estimation of the overdispersion parameter c. When t
rion has been identified (AIC, AIC_c, or QAIC_c), it shoul
models in the set. The adequacy of the global model shou
when the response variable is continuous (e.g., regressio
of procedures to aid in this task and many general ways t
variation.

Discuss or reference the use of other aspects of the infor
proach, such as model averaging, confidence sets on mod
of the relative importance of variables. Define or referen
(e.g., K, Δ_i and w_i). Ideally, the variance component du
uncertainty should be included in estimates of precision
vs. conditional standard errors), unless there is strong ev
best model, say an Akaike weight (w_i) > 0.9.

For well-designed, strict experiments in which the num
tors is small and factors are orthogonal, use of the full m
appropriate (rather than considering more parsimonious m
tive is to assess the relative importance of variables, th
selected to provide a balance among variables (Section 4
the relative importance of variables can then be based on th
weights for each variable, across models, and these sums
Avoid the implication that variables not in the selected
model are unimportant. Give estimates of the important
fect size) and measures of precision (preferably a confider
evidence can be assessed using the Akaike weights (w_i)
Provide quantities of interest from the best model, or ot
$\hat{\sigma}^2$, coefficient of determination, estimates of model paran
dard errors, insights from an analysis of the residuals). In
evidence should be gathered, interpreted, and presented c

The *Results* section should be easy to report if the *Meth*
convincingly the hypotheses and associated models of in
of the value of the maximized log-likelihood function (log
estimated parameters (K), the appropriate selection crite
QAIC_c), the simple differences (Δ_i), and the Akaike wei
in the set (or at least the models with some reasonable l
where $\Delta_i < 10$). It is often helpful to report the models i
based on the differences (Δ_i); this makes it easy to see w
second best, and so on.

Do not include test statistics and *P*-values when usin
theoretic approach as this inappropriately mixes

The classical tests one would use for this purpose are inv
pairs have been ordered by an information criterion. Do
information-theoretic approaches are a test in any sens
terms such as significant and not significant, or rejecte
instead, view the results in a strength of evidence context

If some data dredging and modeling were done afte
then be sure that this is clearly explained when such re
in the *Discussion* section. It is important to separate analy
questions and hypotheses formed before examining the da
sequentially from examining the results of data analyses
tends to be more confirmatory, while the second approa
exploratory. In particular, if the analysis of data suggests
leading to an interesting hypothesis, then at this midway
tests or measures of precision remain valid. That is, an i
patterns or hypotheses as being an actual feature of the pop
interest are not well supported. Conclusions reached after r
of the results of prior analyses, while interesting, cannot be
degree of confidence as those from the more confirmatory
post hoc results, while somewhat likely to be spurious, may
hypotheses to be readdressed with a new, independent set
an important part of good science.

4.13 Summary

Model selection should not be considered just the search
Rather, the basic idea ought to be to make more reliable
the entire set of models that are considered a priori. This
and scale the set of models and determine, perhaps, a c
the R models, for the K-L best model. Parameter estimati
use of all the models when appropriate (e.g., model avera
and attempt to use unconditional variances unless the se
strongly supported (say its $w_{min} > 0.9$).

In general there is a substantial amount of model sele
many practical problems (but see the simulated starling e
pine wood data, for exceptions). Such uncertainty about w
(and associated parameter values) is the K-L best approxim
whether one uses hypothesis testing, information-theoretic
consistent criteria, cross-validation, or various Bayesian m
is a nonnegligible variance component for estimated para
prediction) due to uncertainty about what model to use,
should be included in estimates of precision.

bootstrap method. The simple computation of the relativ
model considered, given the data, and rescaling these mod
Akaike weights, w_i, is often effective and easy to understan
w_i provide a basis for a measure of relative support of the
what we can call model scaling. The use of evidence r
is effective and convenient.) The bootstrap, while compu
provides estimates of model selection probabilities, $\hat{\pi}_i$.

If there is substantial model selection uncertainty and
ance is estimated conditionally on the selected model, th
estimated parameters will likely be overestimated, and the
interval coverage will be below the nominal level (e.g.,
than 95%). Estimates of unconditional standard errors ca
ther Akaike weights or bootstrap selection probabilities. T
for the R models considered over all bootstrap samples n
compute unconditional standard errors of estimated para

Relatively little work has been done to understand and le
bias (Section 1.6). Shrinkage estimators such as

$$\tilde{\tilde{\theta}} = w_+ \hat{\tilde{\theta}}$$

deserve investigation. In particular, the estimation of the
further work.

Conditional inference can be relatively poor. There is a n
search on general methodology to incorporate model selec
variance component in the precision of parameter estimato
has potential in some applications where interest is concen
that appear (implicitly, at least) in every model in the set. Pr
be approached as a problem in (weighted) model averagin
are either Akaike weights (w_i) or bootstrap selection prob

The importance of a *small* number (R) of candidate mo
detailed analysis of the data, cannot be overstated. Small is
in contrast to what is not small: If the number of models co
sample size, hence $R > n$, then there is *not* a small number
This condition often occurs in the commonly used cas
possible linear models in variable selection. In that all-mo
has p variables, then $R = 2^p$; hence $R = 1024$ for $p = 1$
for $p = 20$. One should have R much smaller than n. MMI
increasingly important in cases where there are many m
the background science is lacking, so that needed a prio
deemed impossible, then the analysis should probably be c
exploratory.

Finally, investigators should explain what was actuall
selection (i.e., data analysis). Was it objective model selec

introducing new models into consideration as data and
the former case we recommend AIC (or its variants as n
case, if the strategy can be implemented as a computer alg
bootstrap to assess model selection uncertainty for such s
searches. If the model selection strategy cannot be represe
defined algorithm, one cannot determine model selection

Monte Carlo Insights and Exte Examples

5.1 Introduction

This chapter gives results from some illustrative exploratio of information-theoretic criteria for model selection and precision when there is model selection uncertainty. Th Chapter 4 are illustrated and additional insights are prov lation and real data. Section 5.2 utilizes a chain binomia some Monte Carlo evaluation of unconditional sampling confidence intervals, and model averaging. For this simul process is known and can be of relatively high dimens model and the models used for data analysis in this chain are easy to understand and have no nuisance parameters. parisons of AIC versus BIC selection and use achieved coverage as an integrating metric to judge the success of to inference.

Section 5.3 focuses on variable selection (equivalent to a in multiple regression for observational data assuming nor mogeneous errors. A detailed example of AIC_c data anal given, and it is shown how to extend the example to a re investigation of methodology. The same Monte Carlo meth erate additional specific illustrative results. We also discus to reduce a priori the number of variables. A discussion details of model selection bias and other ideas with emph and better approaches to, all-subsets selection.

One real example uses distance sampling data on kangaro
reader is encouraged to read these examples, because e
vey some general information about data analysis unde
approach.

In the following material it is important to distinguis
variances, or standard errors, that are conditional on one p
sus those that are unconditional, hence not based on just
Conditional measures of precision based on a restrictive n
noted by $\mathrm{var}(\hat{\theta}_j \mid \theta_j)$ or $\mathrm{se}(\hat{\theta}_j \mid \theta_j)$; however, they are more
$\mathrm{var}(\hat{\theta}_j \mid g)$ or $\mathrm{se}(\hat{\theta}_j \mid g)$. Corresponding unconditional va
by $\mathrm{var}(\hat{\theta}_j)$ or $\mathrm{se}(\hat{\theta}_j)$, and have (as needed) an added varia
to model selection uncertainty. Even such "unconditional"
do depend on the full set of models considered.

The "as needed" phrase is appropriate because som
design-based rather than model-based (see, e.g., Schreuder
6, and Edwards 1998). This applies, for example to the use
from a random sample; or here to the use of just $\hat{S}_i = n$
restrictive model assumed for this inference. However, mo
is usually necessary and can be very effective for parame
on complex data.

5.2 Survival Models

5.2.1 A Chain Binomial Survival Model

We consider here tracking a cohort of animals through
process of the cohort. Thus at time 1 there are a known nu
alive. Usually, we are thinking that these are all young
or born at time 1. One year later there are n_2 survivors; i
anniversary dates $i > 1$ there are n_i survivors. Eventuall
last anniversary year ℓ, $n_\ell > 0$ but $n_{\ell+1} = 0$. Given the i
the survival probability in year one (i.e., for the first year of
defined as $S_1 = \mathrm{E}(n_2 \mid n_1)/n_1$. In general, for year i, $S_i =$
the obvious general estimator is $\hat{S}_i = n_{i+1}/n_i$; this estim
any assumed model structure imposed on the set of basic
If n_i is large enough, then $\hat{S}_i$ is an acceptable estimator in
However, as the cohort dies out, n_i becomes small, and then
particular, in the year ℓ that the cohort dies out we always g
This is a terrible estimator of age ℓ survival probability.
problem, and to improve age-specific survival estimates f

The quantity n_1 is known at the study initiation, while t
$n_2, n_3, n_4, \ldots$, being not then known, are treated sequenti
variables and known ancillaries in the eventual data an
will generate, in the Monte Carlo study, n_2 as binomial(
survival process is sequential in time, once the survivors
are known, we can model n_3, given n_2, as binomial(n_2, S_2)
we know n_i; hence we can generate n_{i+1} as a binomial(n_i,
Thus we generate our probability model, hence likeliho
chain binomial model with conditional independence of
given the temporally preceding n_1 to n_{i-1}. The likelihood f
general possible model is

$$\mathcal{L}(\underline{S} \mid \underline{n}) = \prod_{i=1}^{\ell} \binom{n_i}{n_{i+1}} (S_i)^{n_{i+1}} (1 - S_i)^{n_i - n}$$

where $\underline{S}$ and $\underline{n}$ are vectors of the survival parameters and
tively. The underlying parameters explicitly in (5.1) are S
context a (restricted) model for the parameters is some i
such as $S_i \equiv S$ for all ages $i \geq 1$ (model g_1), or perhaps
only after age three; hence $S = S_4 = \cdots = S_\ell$, and S_1,
stricted (model g_4). If no structural restrictions are impose
general possible global model, g_i, for which the MLEs ar

$$\hat{S}_i = \frac{n_{i+1}}{n_i}, \quad i \leq \ell,$$

with conditional sampling variances

$$\operatorname{var}\left(\hat{S}_i \mid g_i\right) = \frac{S_i(1 - S_i)}{n_i}.$$

If such data were from moderately long-lived species
wolves, one might expect substantial differences in surv
age classes, hence $S_1 < S_2$ and so on for several age classe
be near-equal survival probabilities for adults until a decre
survival as the surviving animals approach old age (i.e.
there are some large age-specific effects in survival follow
ing effects (this conceptual model is not universally applic
does not apply to salmon). In real populations, the age-sp
would be confounded with annual environmental effects a
with some individual heterogeneity. We will not pursue
here and instead focus on some points related to condition
inference in model selection and some comparison of in
approaches versus dimension-consistent criteria (e.g., BIC
will illustrate the amount of variability in model selectio
sample size n_1.

Ritei Shibata was born in Tokyo, Japan, in 1949. He received a B
1973, and a Ph.D. in mathematics in 1981 from the Tokyo Institu
currently a professor in data science, Department of Mathematics
Yokohama, Japan. He has been a visiting professor at the Australi
University of Pittsburgh, Mathematical Science Research Institute (
toria University of Wellington, New Zealand. He is interested in sta
time series analysis, S software, and data and description (D&D).

Because of the paucity of data at older ages, some co
must be assumed, at least for the older ages, to get reliable
solution is to pool data beyond some particular age and as
ages have a constant survival probability. Thus, for anima
we assume that they all have the same survival probability
is not to say that truth is no longer age-specific, but rather
model of the survival process is not age-specific after some
is to do some modeling (e.g., logistic models) of the survi
do model selection. We will address both approaches in t

Model	K	Parameters	
1	1	$S \equiv S_i,$	$i \geq$
2	2	$S_1, S \equiv S_i,$	$i \geq$
3	3	$S_1, S_2, S \equiv S_i,$	$i \geq$
$\vdots$	$\vdots$	$\vdots$	
R	R	$S_1, S_2, \ldots, S_{R-1}, S \equiv S_i,$	$i \geq$

The pattern is obvious: Model g_r has r parameters wi probabilities for ages 1 to $r-1$, and constant annual sur age r and older. The global model is g_R; model g_r has K

This particular set of models is convenient for Mont because all MLEs exist in closed form. Let the tail sum of alive at and after age r be denoted by $n_{r,+} = n_r + n_{r+1} +$ the MLEs for model g_r ($r \leq \ell$) are

$$\hat{S}_i = \frac{n_{i+1}}{n_i}, \quad i = 1, \ldots, r-1,$$
$$\hat{S} = \frac{n_{r+1,+}}{n_{r,+}} = \frac{n_{r+1} + n_{r+2} + \cdots}{n_r + n_{r+1} + \cdots}.$$

Also, the likelihood for the r parameters given mode constrained by model g_r)

$$\mathcal{L}_r(\underline{S} \mid \underline{n}) = \left[\prod_{i=1}^{r-1} \binom{n_i}{n_{i+1}} (S_i)^{n_{i+1}} (1 - S_i)^{n_i - n_{i+1}}\right] \left[\binom{n_{r,+}}{n_r}\right.$$

5.2.2 An Example

The first example compares the performance of AIC_c a size $n_1 = 150$. BIC is the Bayesian information criterion developed independently and from somewhat differing vie (1978) and Akaike (1978, 1979). The form of BIC is

$$-2\log(\mathcal{L}) + K \cdot \log(n).$$

This is a type of criterion that Bozdogan (1987) calls "di since such criteria are often based on the assumption tha and this model is in the set of candidate models. Then, an estimate of the dimension of this true model with pro size increases asymptotically. Other interpretations weake "quasi-true models" (see Cavanaugh and Neath (1999)) a that BIC provides an estimate of the true model's dimensi

8 in some detail. Here, we only present some comparisons
exposure to BIC or SIC.

The true survival probabilities S_1, S_2, and so forth wer
0.8, and then beyond S_5, the age-specific annual survival
year (e.g., 0.784, 0.768, 0.753, 0.738, and so forth for S_6 a
the data-generating model the survival parameters vary s
to a maximum at ages 4 and 5, then simulating senescen
decrease in S_j. For each Monte Carlo sample the cohort is
animals are dead. For example, the simulated animal cou
$(n_i, i = 1, 2, \ldots, 16 = \ell + 1)$ were

150 (fixed), 74, 49, 34, 27, 21, 13, 9, 6, 6, 6, 4,

The set of models considered here is g_1 to g_{10}; thus $R =$
set of models within the global model g_{10}. Hence, the glob
separate estimates of S_1, S_2, $\ldots$, S_9 but a single "pooled
age 9. The results for 10,000 Monte Carlo samples are give
the two selection approaches yield substantial differences i
On average, AIC_c selected an approximating model with
while BIC selected a model with an average of 2.1 param

The thinking underlying BIC is that the true model ex
of candidate models; this condition is not met here: Tr
of ten models (the model closest to truth is model g_{10}).
attempts to estimate the dimension of the true model, a co
well-defined here because there are nominally an unbou

TABLE 5.1. Comparison of model selection relative frequencies
models g_1 to g_{10} and data from the chain binomial generating mo
to S_{10} as 0.5, 0.7, 0.75, 0.8, 0.8, 0.784, 0768, 0753, 0738, and 0
$S_{j+1} = 0.98 \cdot S_j$); results are for sample size $n_1 = 150$.

Model	K	AIC_c model selection		BIC model
		percent	cumul. %	percent
1	1	0	0	0.2
2	2	52.0	52.0	88.6
3	3	22.4	74.4	10.0
4	4	5.6	80.0	0.5
5	5	5.5	85.5	0.5
6	6	5.0	90.5	0.1
7	7	3.4	93.9	0
8	8	2.4	96.3	0
9	9	2.2	98.5	0
10	10	1.5	100.0	0

models used were g_1 to g_{10} (see text); these results are based on 10,
for sample size $n_1 = 150$.

Age i	S_i	$E(\hat{S}_i)$		$E(\widehat{se})$, AIC		$E(\widehat{se})$, BIC	
		AIC_c	BIC	cond.	unc.	cond.	un
1	0.500	0.500	0.501	0.041	0.041	0.041	0.0
2	0.700	0.711	0.732	0.039	0.053	0.029	0.0
3	0.750	0.751	0.749	0.035	0.053	0.026	0.0
4	0.800	0.772	0.751	0.033	0.052	0.026	0.0
5	0.800	0.769	0.750	0.032	0.050	0.026	0.0
6	0.784	0.760	0.749	0.032	0.048	0.026	0.0
7	0.768	0.754	0.749	0.032	0.048	0.026	0.0
8	0.753	0.750	0.749	0.032	0.048	0.026	0.0
9	0.738	0.745	0.749	0.032	0.047	0.026	0.0
10	0.723	0.740	0.749	0.032	0.047	0.026	0.0
Average							

specific survival rates. However, there will not be enough
all those parameters, and increasing n_1 does not change t

In sharp contrast, AIC_c attempts to select a parsimon
model as a basis for inference about the population sample
that full truth exists as a model, nor does it assume that
is in the set of candidates. AIC_c estimates relative expecte
distance and then selects the approximating model that is
truth (i.e., the model with the smallest value of AIC_c). Ba
$E(\Delta_i)$ (good to about two significant digits), it is model g_2
model here, on average, but only by a minute winning ma
Therefore, without loss of generality we can set $E(\Delta_2) =$
other values of $E(\Delta_i)$ compared to this minimum value. I
we have the $E(\Delta_i)$ as 25.0, 0.0, 0.09, 1.2, 2.0, 2.7, 3.6, 4.

Further results from the Monte Carlo study are presented
ically, we look at the properties of the estimated age-sp
under AIC_c and BIC model selection strategies. In Table
timated expected value of $\hat{S}_i$, where $\hat{S}_i$, by sample, is t
survival rate based on whatever model was selected as b
this estimator is computed for both AIC_c and BIC mode
For example, if model g_4 is selected, then $\hat{S}_i = n_{i+1}/n$
$\hat{S}_i = n_{5,+}/n_{4,+}$ for $i \geq 4$.

From Table 5.2 the estimators of age-specific surviva
nearly unbiased; however, bias was slightly smaller fo
models than for the BIC-selected models for 8 of the 9 a
differed. The expected estimated standard errors, condit

under both AIC_c and BIC selection are too small, since
for any model selection uncertainty.

Various comparisons are possible from these simulatio
empirical standard errors versus expected estimated stan
theory. However, confidence-interval coverage is an int
how well the methodology is performing. Therefore, w
confidence interval coverage for nominal 95% intervals, a
section) as $\hat{S} \pm 2\,\widehat{\text{se}}$. Use of this simple form may have resu
1 to 3% failure in coverage of S_i for older ages, but we ju
to the contrast of conditional versus unconditional covera

Conditional confidence interval coverage in Table 5.2 c
model selection is generally below the nominal level, ra
0.959 over S_1 to S_{10}. Adjusting the conditional standar
unconditional (unc.) using (4.9) provides much improved
95.0%), ranging from 0.906 to 0.977 (Table 5.2).

Achieved coverage using conditional standard errors f
tion averaged 77.8% across the 10 age classes (Table 5.2
BIC model selection (there is no theoretical basis to jus
proved the achieved coverage of true S_i for BIC model
(range 60.1 to 98.7%); coverage generally remained belo
(95%). Buckland et al. (1997) present results on a survival
to the one used here.

This example shows the large amount of uncertainty a
model selection when sample size is small ($n_1 = 150$)
model has tapering effects. This simple simulation exerci
that BIC selection cannot be recommended. It requires ve
to achieve consistency; and typically, BIC results in a
is underfit (e.g., biased parameter estimates, overestimat
achieved confidence interval coverage below that achiev
models). Conditional estimates of precision, under eithe
clude model selection uncertainty, and this is often an
Incorporating model selection uncertainty can bring ach
terval coverage up to approximately the nominal level
under AIC_c (using (4.7)) (this applies to AIC also).

The results in Table 5.2 are typical of many similar s
have examined (with different sample size, various sets o
models R). To make more use of this particular example
of true S_i, we obtained confidence interval coverage on
S_i, $i = 1$ to R, for AIC_c model selection for some addi
We also tabulated the 90th, 95th, and 99th percentiles of
is useful for interpretation of the Δ_i. As sample size incr

number of models considered, achieved confidence interval coverag
been averaged over conditional (cond.) and unconditional (unc.) in

Sample size	R	Best model	Confidence interval coverage, %		perce
n_1		K	cond.	unc.	0.90
100	10	2	84.4	95.5	4.0
250	10	3	77.8	93.3	4.8
500	10	3	75.5	92.2	7.6
1,000	10	8	78.6	92.8	5.9
1,000	12	8	76.0	90.6	6.4
10,000	10	10	86.2	93.0	1.0
10,000	20	13	78.3	88.9	5.4

model, should include more parameters. Thus one needs
size of the model set, i.e., R.

Unconditional confidence interval coverage is much su
coverage (Table 5.3). When R is large, such as 20 for n_1
(over S_1 to S_{20}) confidence interval coverage suffers as
extrapolated estimates of, say, S_{20} based on a fitted mod
g_{13}. The S_i have a shallow peak at $i = 4$ and 5, and this fea
a large sample size to detect reliably. Thus the theoretical
though increasing as sample size increases, stalls at mode
sample size is passed, after which that theoretically best
responsive to increasing sample size. It is this feature
percentiles of Δ_p to be so big (e.g., for $n_1 = 500$, a 95
and a 99th percentile of 16.8).

The other notable feature in Table 5.3 is found when n_1
10. The theoretically best AIC_c model is then model g_{10}, the
in the set considered. Model selection probabilities here in
$\pi_9 = 0.18$; hence, Δ_p is 0 in 80% of the samples and sm
of the samples.

Use of BIC is theoretically inappropriate for model sele
of Table 5.3 because the set of models used does not incl
ating model with fixed, small K and with all effects bein
increases. However, BIC performs best at large sample si
whether BIC performs well here when $n_1 = 10{,}000$ and
achieved coverage (for nominally 95% intervals) for BIC
these two cases is given below, along with average (ove

confidence interval coverage (nominally 95%) has been averaged
and unconditional (unc.) intervals on S_1 to S_R.

Sample size	R	Best model	Confidence interval coverage, %		perce
n_1		K	cond.	unc.	0.90
100	10	3	78.0	93.4	2.7
250	10	3	74.6	92.2	3.6
500	10	3	76.4	92.3	4.8
1,000	10	4	76.3	92.6	3.5
1,000	12	4	72.0	91.5	3.6
10,000	10	6	80.1	92.4	3.5
10,000	20	6	59.6	85.2	3.7

both BIC and AIC (AIC coverage is in Table 5.3):

		coverage, BIC		average MS	
n_1	R	cond.	unc.	BIC	
10,000	10	77.8	83.7	0.000245	0.0
10,000	20	45.9	51.5	0.003077	0.0

Average mean square error is lower for AIC, and confiden
is much better. Using BIC, model g_8 is selected with proba
or 0.36 ($R = 20$), and the probability of selecting one of r
is 0.96 ($R = 10$) or 0.93 ($R = 20$).

For some additional results and comparison to Table 5.3
5.3 results for a model wherein true $S_i = 0.5+0.3/i$. Table
interval coverage results and percentiles of Δ_p for this e
the monotonicity of the true S_i, the percentiles of Δ_p ar
being *about* 4 (90th percentile), 5 (95th percentile), and
Again, for the last case ($n_1 = 10{,}000$ and $R = 20$) the p
to averaging coverage over the intervals on all of S_1 to S
considerable extrapolations. Generally, AIC_c does not se
model g_{10}. If for this last case we look only at intervals on S
levels under AIC_c model selection are 80.7% (cond.) and

5.2.3 An Extended Survival Model

A second Monte Carlo example is given here to add additi
realism in the set of approximating models. This example
binomial generating model and associated parameters p
5.2.2; only the set of approximating models is different. R
for model g_r the constraint $S \equiv S_i$ for all $i \geq r$, the a

Model 1	$\text{logit}(S_i) = \alpha + \beta \cdot i,$
Model 2	$S_1, \text{logit}(S_i) = \alpha + \beta \cdot i,$
Model 3	$S_1, S_2, \text{logit}(S_i) = \alpha + \beta \cdot i,$
$\vdots$	$\vdots$
Model 10	$S_1, S_2, \ldots, S_9,$ and $\text{logit}(S_i) = \alpha + \beta \cdot i,$

Each model has an intercept and slope parameter for th
fitting of age-specific survival rates S_i for ages $i \geq r$. Mod
fits unconstrained age-specific survival rates for ages 1 to
g_r has $K = r + 1$ parameters.

The initial population size was $n_1 = 150$ animals, as be
were followed until the last one died. The true paramet
$S_2 = 0.7$, $S_3 = 0.75$, $S_4 = 0.8$, $S_5 = 0.8$, and for $r >$
The computer generated 10,000 independent repetitions u
analysis using the 10 approximating models. In some samp
could be fit because all animals died before reaching age 1
the last age at which $n_\ell > 0$. If $\ell - 2 \geq 10$, then all ten lo
were fit. Otherwise, only $\ell - 2$ models were fit (if $\ell -$
number of models that could be fit to a sample varied son
matter to the overall strategy of either selecting a best m
averaging, because for each fitted model it was always pos
derived $\hat{S}_i$ (and its estimated conditional variance) for any

Generally, the logistic approximating models were su
tion as compared to the models used in Section 5.2.2. T
model (under AIC_c selection) was g_3 ($K = 4$) with mod
($\text{E}(\Delta_4) = 0.17$ relative to setting $\text{E}(\Delta_3) = 0$). The mod
frequencies are shown in Table 5.5. In 84.5% of samples
or g_4 was selected. Model selection uncertainty in this e
that of the example in Section 5.2.2 (compare Tables 5.
the best model to use has 4 parameters, rather than 2, as w
models previously considered. **Use of a better set of m
informative inferences from the data.**

For the simulation underlying Table 5.5 results, all 10
in 9,491 samples. In 18 samples, only models g_1 to g_7 co
protocol noted above). Only models g_1 to g_8 were fitted
only up to model g_9 were fitted in 363 samples.

Table 5.6 summarizes the estimated expected values of
models and under model averaging (MA) using (4.2),
ing the Akaike weights. Also shown in Table 5.6 is the
of confidence intervals based on estimated unconditional
(i.e., using (4.9)). Model averaging provided a slightly
tor of conditional survival. However, both methods perfo

to increase to a maximum at ages 4 and 5 and then decrease slowly
Section 5.2.2 (see text for details). Model g_3 is the theoretically be
sample size $n_1 = 150$.

Model	Percent	Cumul. %
1	00.0	00.0
2	35.2	35.2
3	29.5	64.7
4	19.8	84.5
5	6.3	90.8
6	3.5	94.3
7	2.2	96.4
8	1.5	98.0
9	1.2	99.2
10	0.8	100.0

TABLE 5.6. Summary of Monte Carlo results, based on 10,000 s
survival estimation under AIC_c selection and model averaging (MA
had parameters S_i; 10 logistic models ($g_1, \ldots, g_{10}$) were fit to the
data; results are for sample size $n_1 = 150$.

Age	S_i	$E(\hat{S})$		Coverage	
i		AIC_c	MA	AIC_c	M
1	0.500	0.499	0.499	0.957	0.9
2	0.700	0.708	0.707	0.907	0.9
3	0.750	0.758	0.757	0.920	0.9
4	0.800	0.789	0.790	0.926	0.9
5	0.800	0.785	0.786	0.917	0.9
6	0.784	0.772	0.773	0.951	0.9
7	0.768	0.759	0.760	0.964	0.9
8	0.753	0.744	0.745	0.962	0.9
9	0.738	0.727	0.728	0.962	0.9
10	0.723	0.709	0.709	0.957	0.9
Average				0.942	0.9

The achieved confidence interval coverage for these inter
model selection uncertainty, is very close to the nominal
approaches; both approaches use the same estimate of un
error.

From Table 5.6 the unconditional confidence interva
over the intervals on S_1 to S_{10}, and over all 10,000 sampl
the strategy of inference based on the selected AIC_c bes

0.880. In the worst case, which was at age 3 (i.e., for S_3), C
was only 0.717, versus unconditional of 0.917.

In this example the simulation program also compute
error (MSE) for estimators of S_i at each age for AI
for model averaging, and for the simple (almost mode
age-dependent survival $\hat{S}_i = n_{i+1}/n_i$, $i = 1, 2, \ldots$
the model-averaged estimator was smaller than that of
timator (ranging from 11% to 25% smaller, except for
approaches had the same MSE). In this example, at least, r
ing Akaike weights, has important advantages. Both A
and model averaging are quite superior to the use of
tor of S_i. For example, at age nine, the MSEs for $\hat{S}_i$
selection, model averaging, and the simple estimator w
and 0.0186, respectively. This illustrates the advantage
modeling of the underlying survival probabilities (this is
learned from the starling experiment in Section 3.4).

We computed the pivotal quantities $\Delta_p = \text{AIC}_{best} - \text{AI}$
of the 10,000 samples to better understand these values f
this type. Some percentiles of the sampling distribution of
are shown below:

Percentile	Δ_p
50.0	1.20
75.0	2.09
80.0	2.29
85.0	3.11
90.0	4.21
95.0	6.19
97.5	8.33
98.0	8.80
99.0	10.63

Thus, in approximately 90% of the simulated data sets the
age, between AIC_c for the selected model, for that sample,
theoretically best model was ≤ 4.21. If we adopted the ide
imate 90% confidence set on what is the actual AIC_c theo
then in this example, fitted models that had Δ_i values be
be in that set, and hence might be candidates for some fur
making inferences from an individual data set.

To make further use of the generating model and set of ap
used here, we determined several quantities for differin
theoretically best model, conditional and unconditional c
model selection (as opposed to model averaging), and 9
percentiles of Δ_p. Regardless of this variation in sample siz

of the (approximate) pivotal Δ_β. These same results hold
generating models and fitted models that we have exami
methods. Often the conditional coverage averages are less
Table 5.7 (more like 70% to 85%).

The true S_i do not exactly fit any of the approximat
However, model g_4 turns out to be an excellent approx
least for $i \leq 15$ (we did not plan this to be the case). He
know S_1 to S_{15} exactly, but we needed a mathematical mod
numbers, we would likely fit and use model g_4. We woul
fitted model then provides a concise, yet almost exact, su
lack of fit is both trivial and statistically "significant" given
We do not generally have such huge sample sizes, but if w
concerns of small-sample-size statistics would not apply (p
then replaces statistical significance), and we might very
rather than model g_8 as a preferred approximation to tru
philosophical point: Even if we knew truth, we would of
said truth by a low-order parsimonious fitted model becau
this confers about understanding the basic structure of fu

5.2.4 *Model Selection if Sample Size Is Huge, or*

The results in Table 5.7 provide a motivation for us to r
sophical issues about model selection when truth is es
equivalently in statistical terms, when we have a huge

TABLE 5.7. Some Monte Carlo results (10,000 repetitions per samp
selection, for the same generating model and set of models g_i as use
size is n_1; R is the maximum number of models considered; K is s
AIC_c best model g_k; achieved confidence interval coverage (nor
averaged over conditional (cond.) and unconditional (unc.) interval

Sample size n_1	R	Best model K	Confidence interval coverage, % cond.	unc.	perce 0.90
50	10	2	84.8	92.6	3.8
100	10	3	86.2	94.1	3.7
150	10	3	86.0	94.2	4.2
200	10	4	86.1	94.2	3.6
500	10	4	88.3	94.6	2.7
1,000	12	4	90.4	95.4	2.6
10,000	15	4	90.3	95.5	4.9
100,000	15	8	86.4	93.3	5.0

theoretic statistical model selection, is g_8. However, at a sa
we find that (1) the parameter estimates under either mode
to at least two significant digits (so $\hat{S}_i \approx \mathrm{E}(\hat{S}_i)$ for practic
the difference (i.e., bias) under either model between E(
also trivial (but not always 0), hence ignorable for $i = 1,$

Therefore, at a sample size of 100,000, or if we literal
it is more effective from the standpoint of understanding
variation in the survival probabilities S_i to fit and use mo
$n_1 = 100{,}000$, or to truth if truth is in hand). In this situati
is gained by using model g_8 rather than model g_4. On th
less parsimonious, more complicated model g_8 has a cost
which we perceive the basic pattern in the true S_i. In essenc
that even if we knew the true S_i, we would be better ab
pattern of information therein by fitting (to the S_i) and rep
using that simple five-parameter model, rather than using e
actual S_i. (In practice one might find the simple truth here
$i \geq 5$).

The point is that once complex truth, as a set of parame
well, such as to two or three leading significant digits (w
what we can usually achieve with real sample sizes), we ma
model selection anymore. Instead, other criteria would be u
as is, or use a simple, parsimonious, interpretable model
truth when that model explains almost all the variation in
interpret, understand, and communicate one's model is im
models, and all numerical descriptions of reality should be
model if for no other reason than we will never know all d
parameter.

Another view of this matter is that model g_8 is actu
model g_4 once we can fit either model to truth. Either
would correspond to producing a correlation coefficien
true S_i and model-predicted survival probability. Once y
well with a simple model, the gains in understanding and
model, as truth, override that fact that the model is not exa
enough.

As a corollary of this philosophy, model goodness-of-fi
tests becomes irrelevant when sample size is huge. Instea
is to find an interpretable model that explains the inform
a suitable level of approximation, as determined by subj
ations. Unfortunately, in life sciences we probably neve
anywhere near this large; hence statistical considerations
fit to the data are important as well as subject-matter cons
a fitted model that is interpretable.

could) be based on statistical criteria. There is no theory
ministic model must fit deterministic truth in order for tha
Similarly, if truth as a probability distribution is known
approximating model, then K-L can be used, but there is
much information loss is tolerable. That is a subject-matt

5.2.5 *A Further Chain Binomial Model*

Here we consider a generating model with damped oscill
age increases. It is not intended to be considered as a biol
just a case of a complex truth against which we can exam
model selection. We let the initial population size be $n_1 =$
10,000 independent samples using the chain binomial da
with true survival probabilities as

$$S_i = 0.7 + (-1)^{i-1}(0.2/i),$$

where i is year of life ("age" for short). The data (n_1, n_2,
$n_{18} = 0$ but $n_{17} > 0$) for one repetition were

1000 (fixed), 890, 520, 414, 270, 202, 130, 93, 67, 50, 35

Animals were followed until all were dead. The models us
the logistic models g_i of Section 5.2.3. The model set w
$R = 15$). It is because sample size was 1,000 rather than
the size of the model set used. It is a general principle
increases, one can expect to reliably estimate more paran
the set of models considered should depend weakly on sa

Results for AIC_c and BIC model selection are very d
On average, AIC_c selected a model with 7.6 parameters, v
model on average with 5.1 parameters. The theoretically
terms of the expected K-L criterion, in this example is g
parameters. Under BIC selection 95% of the models sele
or 6 parameters. In contrast, under AIC_c selection appro
models selected had between 5 and 13 parameters, inclus

That BIC selection produces a more concentrated dis
models certainly seems to be an advantage. However, usef
focus on bias and precision of parameter estimates, hen
terval coverage (given that we have essentially the shorte
under the different model selection strategies). From Ta
tion produces average conditional and unconditional cov
92.7%, respectively. Under BIC selection we had average
conditional coverage of 73.9% and 78.5%, respectively.
(but not always) did achieve a smaller MSE in these examp
of Table 5.9, MSEs averaged over S_1 to S_{15} were 0.001

$(-1)^{i-1}(0.2/i)$, and the model set for data analysis as g_1 to g_{15}; size $n_1 = 1{,}000$.

Model i	K	AIC_c model selection		BIC model
		percent	cumul. %	percent
1	2	0	0	0
2	3	0	0	0
3	4	0.6	0.6	23.0
4	5	11.5	12.1	50.3
5	6	26.3	38.4	21.7
6	7	20.5	58.9	4.1
7	8	16.0	75.0	0.8
8	9	7.8	82.8	0.1
9	10	6.0	88.8	0
10	11	3.0	91.8	0
11	12	2.5	94.4	0
12	13	1.6	96.0	0
13	14	1.6	97.6	0
14	15	1.1	98.7	0
15	16	1.3	100.0	0

selection, whereas for AIC that average MSE was 0.002(this cost of smaller MSEs (by BIC) for the $\hat{S}_i$ is poor con erage; for S_1 to S_{15}, unconditional BIC coverage varied fr Table 5.9 for AIC coverage).

Comparison of the expected estimated standard errors i tude of the variance component due to model selection unc For example, for age 10, the expected conditional standar sus the expected unconditional standard error of 0.037. T results for ages 1 to 15.

We computed Δ_p values for each of the 10,000 repetitio there were damped oscillations in the S_i parameters. Son sampling distribution of Δ_p are shown below:

Percentile	Δ_p
50.0	1.7
75.0	3.0
80.0	3.5
85.0	4.2
90.0	5.7
95.0	7.9
97.5	10.1
98.0	10.9
99.0	13.3

the generating model allowed damped oscillations in survival as
for details). Conditional (cond.) and unconditional (unc.) standar
intervals were also evaluated.

Age i	S_i	$\mathrm{E}(\hat{S}_i)$	$\mathrm{E}(\widehat{\mathrm{se}}(\hat{S}_i))$		Cov
			cond.	unc.	cond.
1	0.900	0.900	0.009	0.009	0.950
2	0.600	0.600	0.016	0.016	0.955
3	0.767	0.767	0.018	0.018	0.951
4	0.650	0.651	0.023	0.024	0.904
5	0.740	0.734	0.023	0.027	0.821
6	0.667	0.678	0.025	0.031	0.701
7	0.729	0.716	0.024	0.032	0.671
8	0.675	0.690	0.024	0.034	0.711
9	0.722	0.706	0.025	0.035	0.767
10	0.680	0.693	0.028	0.037	0.838
11	0.718	0.699	0.030	0.039	0.871
12	0.683	0.691	0.035	0.043	0.887
13	0.715	0.692	0.040	0.048	0.900
14	0.686	0.685	0.046	0.054	0.902
15	0.713	0.681	0.052	0.061	0.908
Average					0.849

These results are consistent with percentile values for Δ_I
for other examples.

Table 5.10 gives expected Akaike weights $\mathrm{E}(w_i)$ and mo
bilities π_i from the 10,000 Monte Carlo samples for this e
results from this case of a large sample size ($n_1 = 1{,}00$
best), Table 5.10 also gives $\mathrm{E}(w_i)$ and π_i for $n_1 = 50$ (g_3
under the same data-generating model. For sample size
Table 5.10 that on average models g_4 to at least g_9 must be
ing inference about the population under AIC_c model se
weights give some support to models g_{10} and g_{11}, while
somewhat less support to these models.

There is a high degree of model selection uncertaint
particularly for the larger sample size. The underlying pro
lations in the S_i) is complicated, and the set of logistic ap
was relatively poor. Had some science been brought to
cial) problem, hopefully the set of approximating models
some models with at least some oscillating features. Th
uncertainty would likely have been greatly reduced.

Use of the bootstrap to estimate model selection probab
computationally intensive. Akaike weights can be easily c

allowing damped oscillations in survival as age increases; models
g_1 to g_{15}; two sample sizes are considered.

Model	$n_1 = 50$		$n_1 = 1{,}000$	
i	$\mathrm{E}(w_i)$	π_i	$\mathrm{E}(w_i)$	π_i
1	0.050	0.038	0.000	0.000
2	0.293	0.431	0.000	0.000
3	0.283	0.325	0.011	0.006
4	0.160	0.096	0.091	0.115
5	0.104	0.064	0.177	0.263
6	0.056	0.026	0.176	0.205
7	0.033	0.014	0.153	0.160
8	0.015	0.004	0.109	0.078
9	0.006	0.002	0.084	0.060
10	0.000	0.000	0.058	0.030
11	0.000	0.000	0.044	0.025
12	0.000	0.000	0.032	0.016
13	0.000	0.000	0.026	0.016
14	0.000	0.000	0.020	0.011
15	0.000	0.000	0.019	0.013

simple and effective alternative. From Table 5.10 we can
agreement between the expected Akaike weights and the m
abilities π_i. Note, however, that these are not estimates of t
that exact agreement is not expected. Akaike weights, nor
reflect the relative likelihood of each fitted model in the set
vide information about the relative support of the data for t
models. Finally, as all the examples in this section (and o
the weights w_i are very useful in model averaging and con
unconditional sampling variances, hence obtaining uncon
intervals that do substantially improve coverage after mo

5.3 Examples and Ideas Illustrated with Linear Regression

The model selection literature emphasizes applications in
series, often as selection of variables in regression; thi
as model selection. McQuarrie and Tsai (1998) is devo
Regression and Time Series Model Selection. Our explora
much with those of McQuarrie and Tsai (1998), because w
selection uncertainty and, in general, multimodel inferenc
focus on all-subsets model selection by presenting an ex

regression.

5.3.1 All-Subsets Selection: A GPA Example

We use an example based on four regressors (Table 5.11) 20 (a larger number of regressors makes it too demandin the full results). The example of Table 5.11 comes fro (1994). They use these example data extensively to illustr including all-subsets selection based on several criteria or the other ideas we use herein. Also, we note that the "realistic but not real" (H. Iyer, personal communication)

The full model to fit is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 +$$

This model, g_{15}, is also denoted in tables by {1234} becau variables are used (see, for example, Table 5.12). As ano notation, model g_6 uses only the predictors x_1 and x_3 and by $\{1 \cdot 3 \cdot\}$:

$$y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon.$$

Including the intercept-only model g_{16}, $\{\cdot \cdot \cdot \cdot\}$, there are 1 standard all-subsets approach. We use AIC_c for selectio models, and we also apply the bootstrap (10,000 sample full AIC_c model selection applied to all 16 models for ea

The selected best model (g_{11}) includes predictors x_1, x_2 However, support for g_{11} as the only useful model to use h its Akaike weight is just 0.454. Model g_5, $\{12 \cdot \cdot\}$, is also cr model g_6, $\{1 \cdot 3 \cdot\}$. The Akaike weights w_i and bootstra ities $\hat{\pi}_i$ agree to a useful extent, and both demonstrate uncertainty in this example. For a confidence set on mo compelled in practice to provide one in this sort of applic would use a likelihood ratio (evidence) criterion: the set which $w_{11}/w_i \leq 8$ (cf. Royall 1997). Thus, we have the This criterion is identical to using as a cutoff models for would suffice).

We can explore for this example the coherence (with the of such a rule by determining the bootstrap distribution $\Delta_p^* = \mathrm{AIC}_{c,11}^* - \mathrm{AIC}_{c,min}^*$. Here, $\mathrm{AIC}_{c,min}^*$ is the minimun given bootstrap sample, and $\mathrm{AIC}_{c,11}^*$ is the AIC_c value in sample for model g_{11}. We computed the 10,000 bootstra

First-year GPA	SAT		High School	
	math	verbal	math	Engli
y	x_1	x_2	x_3	x_4
1.97	321	247	2.30	2.63
2.74	718	436	3.80	3.57
2.19	358	578	2.98	2.57
2.60	403	447	3.58	2.21
2.98	640	563	3.38	3.48
1.65	237	342	1.48	2.14
1.89	270	472	1.67	2.64
2.38	418	356	3.73	2.52
2.66	443	327	3.09	3.20
1.96	359	385	1.54	3.46
3.14	669	664	3.21	3.37
1.96	409	518	2.77	2.60
2.20	582	364	1.47	2.90
3.90	750	632	3.14	3.49
2.02	451	435	1.54	3.20
3.61	645	704	3.50	3.74
3.07	791	341	3.20	2.93
2.63	521	483	3.59	3.32
3.11	594	665	3.42	2.70
3.20	653	606	3.69	3.52

give below some percentiles of this random variable:

percentile	Δ_p^*
50	1.00
80	3.21
90	3.62
95	5.18
98	7.00
99	8.79

Thus, we have a good basis to claim that the confidence
on the expected K-L best model is about a 95% confide
statistical inference, this result depends on sample size, and
if n increased.

Because model selection in regression is often thought o
tify the importance of each predictor variable, we next co
importance weights of each variable as the sum of the
for each model in which the predictor variable appears (

indicated by the set notation code such as {1234} for the full mod
model selection probabilities $\hat{\pi}_i$ are based on 10,000 bootstrap sam

Model i	Predictors used	AIC_c	Δ_i	w_i
11	$\{123\cdot\}$	−43.74	0.00	0.454
5	$\{12\cdot\cdot\}$	−42.69	1.05	0.268
6	$\{1\cdot3\cdot\}$	−40.77	2.97	0.103
15	{1234}	−39.89	3.85	0.066
12	$\{12\cdot4\}$	−39.07	4.67	0.044
13	$\{1\cdot34\}$	−38.15	5.58	0.028
1	$\{1\cdot\cdot\cdot\}$	−38.11	5.62	0.027
8	$\{1\cdot\cdot4\}$	−35.19	8.55	0.006
14	$\{\cdot234\}$	−32.43	11.30	0.002
10	$\{\cdot\cdot34\}$	−31.59	12.15	0.001
7	$\{\cdot23\cdot\}$	−29.52	14.21	0.000
9	$\{\cdot2\cdot4\}$	−25.87	17.86	0.000
3	$\{\cdot\cdot3\cdot\}$	−25.71	18.02	0.000
2	$\{\cdot2\cdot\cdot\}$	−23.66	20.07	0.000
4	$\{\cdot\cdot\cdot4\}$	−21.70	22.03	0.000
16	$\{\cdot\cdot\cdot\cdot\}$	−15.33	28.41	0.000

TABLE 5.13. Evidence for the importance of each regressor variabl
based on sums of Akaike weights $w_+(j)$ (denoted by $\sum w_i$ here) o
variable occurs and based on relative frequency of occurrence of the
model based on 10,000 bootstrap samples.

Predictor	Σw_i	$\Sigma \hat{\pi}_i$
x_1	0.997	0.965
x_2	0.834	0.722
x_3	0.654	0.658
x_4	0.147	0.148

computed the relative frequency of models selected cont
(hence, the same sum of the bootstrap-based $\hat{\pi}_i$). The resu
that at this sample size, x_4 is not important, whereas x_1 is
x_2 and x_3 are at least moderately important. Indeed, the
was g_{11}, $\{123\cdot\}$.

Standard theory gives us the MLEs for each β_j inclu
the estimated conditional standard error $\widehat{\text{se}}(\hat{\beta}_j \mid g_i)$. We g
and compute the model-averaged estimate of each regre

corresponding estimated unconditional standard error $\widehat{\text{se}}$
use as the estimated unconditional standard error of any
the AIC_c-selected best model. The needed formulas are gi
especially (4.9) with $\hat{\pi}_i = w_i$. The numerical inputs to th
given in Table 5.14 for each of the four parameters: Note
a model-averaged estimate of a parameter independently
a single best model. In Table 5.14 we show for each β_j
eight models it appears in, and the corresponding condit
and Akaike weight. These weights must be renormalized
(4.1) and (4.9). In Table 5.14 the normalizing constant is s
model-averaged estimate and its estimated unconditional
that for a given regression coefficient β_j, its estimate ca
by model. Part of this variation is due to model bias, as
5.3.5. For example, $\hat{\beta}_3$ is 0.18 ($\widehat{\text{se}} = 0.09$) for model g_{11},
model g_3, $\{\cdot\cdot 3\cdot\}$, we have $\hat{\beta}_3 = 0.51$ ($\widehat{\text{se}} = 0.12$). It is be
are correlated here that β_3 actually varies over the differen
inclusion/exclusion of other predictor variables), and this
contributes to model uncertainty about β_3 under model
averaging, of the set of $\hat{\beta}_3$ values in Table 5.14.

We also used the 10,000 bootstrap samples to estimate
parameter $\overline{\beta}_j$ and its unconditional sampling variation s
4. The results from the analytical approach using model s
weights, and the bootstrap results, are shown in Table
timated unconditional standard errors based on Akaike
averaging (i.e., from (4.9)) are each larger than the corres
standard error for the selected best model, g_{11}, $\{123\cdot\}$.
case for β_1 where based on analytical methods $\widehat{\text{se}}\big(\hat{\overline{\beta}}_1\big) =$
$\widehat{\text{se}}\big(\hat{\beta}_1 \mid g_{11}\big) = 0.000455$, and from the bootstrap, $\widehat{\text{se}}\big(\hat{\overline{\beta}}_1$
and the other bootstrap results are precise to essentially tw

The three different point estimates (bootstrap, AIC_c-bas
and best model) for each of β_1, β_2, and β_3 are quite similar
errors. However, this is not true for the two estimates of
whether the bootstrap result (point estimate 0.29, $\widehat{\text{se}} =$
result (point estimate 0.09, $\widehat{\text{se}} = 0.20$) is better in this exa

Note also that here the bootstrap-estimated condition
less than the estimated conditional standard errors for $\hat{\overline{\beta}}_2$ an
this example constitutes only a sample of size 1 as regards
and information-theoretic analytical methods. It will take
Carlo study to make a reliable general comparison of thes
assessing model selection uncertainty.

averaged β_j (4.1), its unconditional standard error (Section 4.3.2), weights over the relevant subset of models.

Model	Predictors	Results by model	
i	used	$\hat{\beta}_1$	$\widehat{\text{se}}(\hat{\beta}_1 \mid g_i)$
11	{123·}	0.002185	0.0004553
5	{12··}	0.002606	0.0004432
6	{1·3·}	0.002510	0.0004992
15	{1234}	0.002010	0.0005844
12	{12·4}	0.002586	0.0005631
13	{1·34}	0.002129	0.0006533
1	{1···}	0.003178	0.0004652
8	{1··4}	0.002987	0.0006357
$\hat{\bar{\beta}}_1$ and unc. se:		0.002368	0.0005350

Model	Predictors	Results by model	
i	used	$\hat{\beta}_2$	$\widehat{\text{se}}(\hat{\beta}_2 \mid g_i)$
11	{123·}	0.001312	0.0005252
5	{12··}	0.001574	0.0005555
14	{·234}	0.001423	0.0007113
15	{1234}	0.001252	0.0005515
12	{12·4}	0.001568	0.0005811
7	{·23·}	0.002032	0.0007627
9	{·2·4}	0.002273	0.0008280
2	{·2··}	0.003063	0.0008367
$\hat{\bar{\beta}}_2$ and unc. se:		0.001405	0.0005558

5.3.2 A Monte Carlo Extension of the GPA Exan

Simulation is a very useful way to gain insights into co tion issues. In particular, here we can assume that the five $(y, x_1, x_2, x_3, x_4)'$ is multivariate normal, MVN(μ, Σ) (t generate 10,000 independent simulated sets of data un model, and do full model selection to learn about select sues. Given the matrix Σ we can determine the true reg (and their approximate true conditional standard errors) regression models. The β_j (and other needed quantities model depend only on elements of Σ, so it suffices to set zero: $\mu = \underline{0}$. The needed 5×5 variance–covariance matri sample variance–covariance matrix from the GPA data. T will be under an assumed truth that is close enough to

i	used	$\hat{\beta}_3$	$\widehat{\text{se}}(\hat{\beta}_3 \mid g_i)$	
11	{123·}	0.1799	0.0877	
7	{·23·}	0.3694	0.1186	
6	{1·3·}	0.2331	0.0973	
15	{1234}	0.1894	0.0919	
3	{··3·}	0.5066	0.1236	
13	{1·34}	0.2474	0.0990	
14	{·234}	0.3405	0.1045	
10	{··34}	0.4171	0.1054	
$\hat{\bar{\beta}}_3$ and unc. se:		0.1930	0.0932	

Model	Predictors	Results by model		
i	used	$\hat{\beta}_4$	$\widehat{\text{se}}(\hat{\beta}_4 \mid g_i)$	
15	{1234}	0.08756	0.1765	
12	{12·4}	0.01115	0.1893	
13	{1·34}	0.17560	0.1932	
8	{1··4}	0.09893	0.2182	
14	{·234}	0.45333	0.1824	
10	{··34}	0.57902	0.1857	
9	{·2·4}	0.51947	0.2207	
4	{···4}	0.77896	0.2407	
$\hat{\bar{\beta}}_4$ and unc. se:		0.09024	0.1989	

TABLE 5.15. Bootstrap (10,000 samples) and Akaike weight-bas
example for model-averaged estimated regression coefficients ar
unconditional standard errors, which include model selection unce
the estimate from the selected model g_{11} and its estimated conditio

	Bootstrap results		w_i model-averaged		AIC
j	$\hat{\bar{\beta}}_j$	$\widehat{\text{se}}(\hat{\bar{\beta}}_j)$	$\hat{\bar{\beta}}_j$	$\widehat{\text{se}}(\hat{\bar{\beta}}_j)$	$\hat{\beta}_j$
1	0.00236	0.000652	0.00237	0.000535	0.002
2	0.00156	0.000508	0.00141	0.000556	0.001
3	0.2296	0.0684	0.1930	0.0932	0.179
4	0.2938	0.3056	0.0902	0.1989	

these GPA data to provide useful results and insights abou
GPA example.

This particular use of Monte Carlo simulation is also
ric bootstrap: We use as the generating model the para

the (nonparametric) bootstrap. The advantage of this pa
that we can specify true values of parameters and hence
interval coverage.

Symbolically, the full variance–covariance matrix is p
for model g_i:

$$\Sigma_i = \begin{bmatrix} \sigma_y^2 & \underline{c}' \\ \underline{c} & \Sigma_x \end{bmatrix}.$$

The marginal variance of the response variable y is σ_y^2. I
dictors x_j are in the regression model, the vector $\underline{c}$ (m
gives their covariances with y ($\mathrm{cov}(y, x_j)$). The varianc
of just the predictors considered (i.e., for any of the 15 m
intercept-only model, $\{\cdots\}$) is given by matrix Σ_x (m
true regression parameters, other than the intercept β_0, is

$$\underline{\beta}' = \underline{c}'(\Sigma_x)^{-1}$$

(we ignore β_0). The approximate sampling variance–cov
is given by $\sigma^2_{y\,|\,x}(\Sigma_x)^{-1}$, where

$$\sigma^2_{y\,|\,x} = \sigma_y^2 - \underline{c}'(\Sigma_x)^{-1}\underline{c}$$

is the true residual variance in the regression (a good
multivariate theory is Seber 1984).

The actual simulation process generates the rows of the
random, but then we condition on them in the regression
Conditionally on X, $\hat{\underline{\beta}}$ is unbiased; so it is also unconditionally
$\mathrm{E}(\hat{\underline{\beta}}') = \underline{c}'(\Sigma_x)^{-1}$. This same argument applies to $\hat{\sigma}^2_{y\,|\,x}$. I
ally (by sample), the variance–covariance matrix of $\hat{\underline{\beta}}$ i
$\mathrm{E}(X'X)^{-1} = (\Sigma_x)^{-1}$ holds only asymptotically as sample
for the simulations, $\sigma^2_{y\,|\,x}(\Sigma_x)^{-1}$ is only an approximation
variance–covariance matrix of $\hat{\underline{\beta}}$.

Rather than show Σ, we show the derived correlation ma
only:

	x_1	x_2	x_3	x_4
y	0.850	0.653	0.695	0.606
x_1		0.456	0.559	0.663
x_2			0.434	0.417
x_3				0.272

(the ordered x_i are SATmath, SATverbal, HSmath, HSer
x_i, x_j correlations are so high that we would need to eli

as g_{15} (as detailed in the text).

Model g_{15}, $\{1234\}$: $\sigma^2_{y\,|\,x} = 0.05692$

j	β_j	$\text{se}(\hat{\beta}_j \mid g_{15})$	$\text{cv}(\hat{\beta}_j \mid g_{15})$
1	0.002010	0.0005061	0.252
2	0.001252	0.0004776	0.381
3	0.1895	0.0796	0.420
4	0.0875	0.1528	1.745

Model g_{11}, $\{123\cdot\}$: $\sigma^2_{y\,|\,x} = 0.05785$

j	β_j	$\text{se}(\hat{\beta}_j \mid g_{11})$	$\text{cv}(\hat{\beta}_j \mid g_{11})$
1	0.002185	0.0004072	0.186
2	0.001312	0.0004698	0.358
3	0.1799	0.0784	0.436

Model g_5, $\{12\cdot\cdot\}$: $\sigma^2_{y\,|\,x} = 0.07307$

j	β_j	$\text{se}(\hat{\beta}_j \mid g_5)$	$\text{cv}(\hat{\beta}_j \mid g_5)$
1	0.002606	0.0004086	0.157
2	0.001574	0.0005121	0.325

Model g_1, $\{1\cdots\}$: $\sigma^2_{y\,|\,x} = 0.10759$

j	β_j	$\text{se}(\hat{\beta}_j \mid g_1)$	$\text{cv}(\hat{\beta}_j \mid g_1)$
1	0.003178	0.0004413	0.139

Model g_3, $\{\cdot\cdot 3\cdot\}$: $\sigma^2_{y\,|\,x} = 0.20001$

j	β_j	$\text{se}(\hat{\beta}_j \mid g_3)$	$\text{cv}(\hat{\beta}_j \mid g_3)$
3	0.5066	0.1173	0.117

is an important consideration. In addition, a principal-com
the covariance matrix of the predictor variables is a reaso

The data-generating model used here (g_{15}, $\{1234\}$),
data, has the pairwise correlations given above and th
$\sigma^2_{y\,|\,x} = 0.05692$ (from (5.5)). From (5.4) we compute
generating model g_{15}, and from $\sigma^2_{y\,|\,x}(\Sigma_x)^{-1}$ we compute t
ditional standard errors of the $\hat{\beta}_j$ under model g_{15}. We also
conditional coefficient of variation of each $\hat{\beta}_j$ (Table 5.16
tities are computable for any submodel fitted to the genera
shows these theoretical values for models $\{1234\}$, $\{123$
and $\{\cdot\cdot 3\cdot\}$.

Table 5.16 if we use model g_3, then $\mathrm{E}(\hat{\beta}_3) = 0.5066$ (cv
when all four predictors are included, $\beta_3 = E(\hat{\beta}_3) = 0.$
For model g_{11}, $\mathrm{E}(\hat{\beta}_3) = 0.1799$ (cv $= 0.436$). Results
when the specified model is always fit to the data; hence
selection occurs. When the inference strategy is to first s
on the data, then use it for inference, the properties of
inferences are affected (model selection biases and uncer

Examination of results in Table 5.16 demonstrates tha
sured by its coefficient of variation, precision of a paramete
as the number of other parameters in the models decrea
varies by model is because the predictors are correlated. Th
variation in $\mathrm{E}(\hat{\beta}_j)$) will get more pronounced if correlation
not occur if all predictors are uncorrelated with each othe
of fitted model g_i has little effect if that predictor is unim
by a large $\mathrm{cv}(\hat{\beta}_j \mid g_i)$. For example, x_4 (HSenglish) can be
model {1234}; hence one uses {123 $\cdot$ }. Indeed, in this M
the expected K-L best model is g_{11} (based on 10,000 sim

What we want to illustrate with simulation here (and i
some results under all-subsets AIC_c model selection in
focus on unconditional vs. conditional confidence interval
i.e., the value of β_j in the generating model g_{15} in Table
intervals we used $\hat{\beta} \pm 2\,\widehat{\mathrm{se}}$; hence, we ignored the issue of a
multiplier. This affects coverage a little, but the focus is rea
between conditional and unconditional coverage.

Second, we look at induced model selection bias in $\hat{\sigma}^2_y$
g_i, out of the 16 models fitted (especially at a small sam
AIC_c), tends to have a better fit for that data set, henc
sum of squares, than would occur on average if model g_i
Thus, data-based selection in regression will tend to resu
a little too well; as a result, we get $\mathrm{E}(\hat{\sigma}^2_{y\,|\,x}) < \sigma^2_{y\,|\,x}$ (th
on confidence interval coverage that is not correctable by
intervals). Confidence intervals depend on $\hat{\sigma}_{y\,|\,x}$; howev
$\sigma_{y\,|\,x}$ varies by model, so what we report to assess selectio
bias

$$\mathrm{RB} = \frac{\mathrm{E}(\hat{\sigma}_{y\,|\,x}) - \sigma_{y\,|\,x}}{\sigma_{y\,|\,x}}.$$

Other quantities of interest include the expected value of
estimator $\hat{\bar{\beta}}_j$ and unconditional interval coverage base
model selection variation, and percentiles of Δ_p. From the

50	0.9
80	3.1
90	3.6
95	5.1
98	7.3
99	9.4

When a parameter (hence $\hat{\beta}_j$) appeared in the selected
several quantities: the model-averaged estimate $\hat{\bar{\beta}}_j$, the un
error $\widehat{\text{se}}(\hat{\beta}_j)$, and three confidence intervals (nominally 95
interval (cond.) is based on $\hat{\beta}_j$ and its estimated condit
given the selected model. The unconditional interval (u
and $\widehat{\text{se}}(\hat{\beta}_j)$. The interval based on model averaging (MA)
Finally, it needs to be clearly understood that the covera
the true parameter from the actual data-generating model

In this example, the achieved coverage of the unconditio
than that of the conditional intervals (Table 5.17), especiall
80%; both coverage percentages increase by about 0.02 if a
interval is used). A source of lowered coverage comes fro
selection. For example, here model selection results in th
$0.00229 - 0.00201 = 0.00028$. This bias is important o
unconditional standard error, which is here $\text{se}(\hat{\beta}_1) =$ (
bias/se ratio is $\delta = 0.54$; this value of δ will result in a c
92.1% if coverage would be 95% at $\delta = 0$ (see Cochran

An unexpected result in this example is that the cond
β_2 and $\beta_3$was as high as (about) 0.9. The unconditional
improve and without exceeding 95% coverage.

The model selection bias induced in $\hat{\sigma}_{y|x}$ is negative
probability that the model will be selected (Table 5.18). S

TABLE 5.17. Expected values of estimators of β_j and confidenc
true β_j under AIC_c-based all-subsets model selection from the Mor
(10,000 samples) mimicking the GPA example; occurrence frequ
samples in which the selected model included the indicated β_j.

Occur.	j	β_j	$\text{E}(\hat{\beta}_j)$	$\text{E}(\hat{\bar{\beta}}_j)$	Achiev
freq.					cond.
9544	1	0.00201	0.00229	0.00232	0.801
7506	2	0.00125	0.00156	0.00157	0.906
6506	3	0.190	0.248	0.253	0.887

of the 10,000 values $\hat{\sigma}^2_{y|x}$ which the model is fit to every generated d
occurs).

Model i	Predictors used	$E(\hat{\sigma}^2_{y\|x} \mid g_i)$ no selection	π_i	S
11	{123·}	0.0576	0.3786	–
5	{12··}	0.0728	0.2730	–
6	{1·3·}	0.0803	0.1451	–
13	{1·34}	0.0762	0.0458	–
15	{1234}	0.0566	0.0389	–
1	{1···}	0.1076	0.0351	–
12	{12·4}	0.0728	0.0301	–
14	{·234}	0.1017	0.0216	–
10	{··34}	0.1272	0.0161	–
8	{1··4}	0.1063	0.0078	–
7	{·23·}	0.1420	0.0064	–
9	{·2·4}	0.1695	0.0012	–
2	{·2··}	0.2227	0.0001	–
3	{··3·}	0.2012	0.0001	–
4	{···4}	0.2446	0.0001	–

is the average value of $\hat{\sigma}^2_{y|x}$ when the model is fit to all 1
sets (compare to theoretical results in Table 5.16). The
under model selection is given by RB from (5.6). Good m
paradigm) do not correspond to very bad levels of RB. As
less acceptable (in terms of expected K-L value), it is sel
data are an unusually good fit to that model.

While general in their qualitative nature, these numeri
extreme for a sample size of 20 than would be the case at
In fact, for this generating model the selection bias in $\hat{\sigma}_{y|}$
size $n = 50$, and confidence interval coverage is nearly 9
interval. Even though model selection can induce biases
theoretic selection and associated unconditional inference
good and certainly better (for the sample size) than use
high-dimensional global model that includes all predictor

5.3.3 An Improved Set of GPA Prediction Model

An even better way to improve on the all-predictors glo
reduce one's models to a smaller set of a priori meaningf
by subject matter or logical considerations. Basically, this
transformations of the predictors (consideration of meani
also important) and dropping predictors that are very unli

different GPA prediction models based on simple derived
some sense.

The original four predictors are each just indices to gene
and they are measured with error. That is, a person's test
vary by circumstances (and luck), such as if they had a
exam. Viewing these predictor variables as just semicru
just compute a single averaged index? In so doing we av
verbal (English) ability, but grades in many courses depen
anyway. With a large sample size (say $n > 1{,}000$) it ma
regression fit calibrate the relative importance of the fou
with only 20 observations some combining of indices ma

The SAT and HS scores are on very different scales
ways to allow for this, such as first to normalize each p
have a mean of 0 and a standard deviation of 1 and then
adjusted predictors to get a total (*tot*) predictor index. To ci
nuisance we used geometric means to cope with the scale
new variables that replace the original four variables are

$$
\begin{aligned}
sat &= (x_1 \times x_2)^{0.5}, \\
hs &= (x_3 \times x_4)^{0.5}, \\
math &= (x_1 \times x_3)^{0.5}, \\
engl &= (x_2 \times x_4)^{0.5}, \\
tot &= (sat \times hs)^{0.5} = (math \times engl)^{0.5} = (x_1 \times x_2
\end{aligned}
$$

These variables are interpretable and seem just as adequat
variables as indices to first-year college GPA.

Next, we would not use, in an a priori analysis, any of the
The only linear regression models we would (did) conside
predictors are given below, in terms of predictor variabl
models have an intercept and σ^2). We numbered these as
order to compare results to the original 16 models:

model	K	variables included
g_{17}	3	*tot*
g_{18}	4	*sat hs*
g_{19}	4	*math engl*
g_{20}	3	*sat*
g_{21}	3	*hs*

We conceptualized these models before examining fit of th
to the GPA data, and no other derived models were consi

The AIC_c best model of the above is g_{17} (Table 5.19).
best in the full set of all 21 models (results are not shown
Table 5.19). In adding new models to an existing set, no

models originally considered (see text for details of the new predic

Model i	Predictors used	AIC_c	Δ_i	w
17	*tot*	−48.20	0.00	0.5
18	*sat hs*	−45.97	2.23	0.1
20	*sat*	−43.96	4.24	0.0
11	{123·}	−43.74	4.47	0.0
5	{12··}	−42.69	5.52	0.0
6	{1·3·}	−40.77	7.44	0.0
15	{1234}	−39.89	8.32	0.0
19	*math engl*	−39.23	8.97	0.0
...	...	...	...	
21	*hs*	−33.98	14.22	0.0
...	...	...	...	

need to be recomputed: Just reorder the full set from sma
The full set of Δ_i values may need to be recomputed if the
(as here, from g_{11} to g_{17}). Given the new set of Δ_i, rec
weights w_i.

The results in Table 5.19 illustrate that the best model
relative only to that set of models. Kullback–Leibler mod
provide an absolute measure of how good a fitted model
only in the set of 16 all-subsets models. Compared to m
(and the entire original set of 16 models) can almost be dis
for expected K-L best model for these data. Correspondi
that any model-based inference is conditional on the set o
The specifics of inferences and computable uncertaintie
the models formally considered.

The models used here are useful only for prediction; t
any causal process. Hence, we illustrate inclusion of mo
prediction based on models g_{17}–g_{21} (standard aspects of
given a fitted linear model are assumed here; see, e.g., Gra
As computed in Graybill and Iyer (1994:244), under mo
prediction of expected GPA at $x_1 = 730$, $x_2 = 570$, x
2.7 is $\hat{E}(y) = 3.185$ with conditional (on model) stand
comparison we note that given model g_{11}, {123·}, the c
are $\hat{E}(y) = 3.253$, $\widehat{se} = 0.102$.

The model-averaged predicted expected GPA is $\hat{\bar{E}}(y) =$
unconditional standard error of 0.11 (Table 5.20). These
using (4.1) and (4.9). To construct a confidence interval he
small degrees of freedom of $\hat{\sigma}^2_{y\,|\,x}$ we suggest that it suffic

g_{17}–g_{21} for the predictors $x_1 = 730$, $x_2 = 570$, $x_3 = 3.2$, and $x_4 =$

Model i	Predictors used	Δ_i	w_i	$\hat{\mathrm{E}}(y)$	$\widehat{\mathrm{se}}$
17	*tot*	0.00	0.685	3.016	
18	*sat hs*	2.23	0.224	3.095	
20	*sat*	4.24	0.082	3.177	
19	*math engl*	8.97	0.008	3.271	
21	*hs*	14.22	0.001	2.632	
			Weighted results:	3.056	

use $3.06 \pm t \times \widehat{\mathrm{se}}\big(\hat{\bar{\mathrm{E}}}(y)\big)$, where the multiplier $t = 2.10$ is fr
on 18 df. Model g_{17} has 18 df for $\hat{\sigma}^2_{y\,|\,x}$, and the weight on
0.685 (more sophisticated procedures will not make a
here). The model-averaged result is distinctly more precis
based on the fitted global model (standard errors of 0.108
averaged versus global model-based). Also, the inclusion
increases the standard error as compared to the result cond
($\widehat{\mathrm{se}} = 0.074$).

5.3.4 More Monte Carlo Results

The theory for Monte Carlo generation of regression da
gressors, was presented in Section 5.3.2. Using that appr
few more simulations. Our motivation was firstly to see w
occurred in using model averaging and unconditional co
did not), and secondly to see what biases might result fr
and what confidence interval coverage could be achieved

This is far from a full-scale simulation study because
the many factors to consider in the design of an all-subs
study. For example, we used only $m = 4$ predictors he
values of m need to be explored). Thus, the global model i
Given a sample size n, one generates a sample from the
MVN($\underline{\mu}$, Σ) generating model. Without loss of generalit
However, there are still (in general) $(m+1) \times (m+2)/2$ p
in Σ. To make this design problem tractable we used the
on the generating model: either

$$\underline{c}' = [0\ 0\ 0\ 0] \quad \text{or} \quad \underline{c}' = [0.8\ 0.6\ 0.4\ 0$$

$$\Sigma_x = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}.$$

The regression parameters of the generating model are giv

Another design factor was taken to be the regressio $\sigma^2_{y\,|\,x} = \sigma^2_y - \underline{c}'(\Sigma_x)^{-1}\underline{c}$. Given values for $\sigma^2_{y\,|\,x}$ (we used o the marginal variance of y, σ^2_y. These quantities (i.e., $\underline{c}$, Σ compute the full 5×5 variance–covariance matrix of (5 variance–covariance of $\hat{\underline{\beta}}$ is given by

$$\frac{\sigma^2_{y\,|\,x}}{n}(\Sigma_x)^{-1}.$$

We see that factors $\sigma^2_{y\,|\,x}$ and n are redundant in their eff variance. Therefore, for not-small sample sizes it is much fix n (say $n \geq 28 + K$, for the global model value of K $\sigma^2_{y\,|\,x}$ to gain precision, rather than to fix $\sigma^2_{y\,|\,x}$ and simul by increasing sample size. We did not do so here; it is st small actual n to explore small-sample-size effects.

Because of the choice of the form of Σ_x, the estimato all have the same conditional variance. In fact, for any ρ the constant diagonal element v of $(\Sigma_x)^{-1}$, because

$$\text{var}(\hat{\beta}_i \mid g_{15}) = v \times \sigma^2_{y\,|\,x} \;/\; n.$$

For $\underline{c}' = [0.8,\ 0.6,\ 0.4,\ 0.2]$ we simulated 10,000 san nation of $\rho = 0$, 0.2, 0.4, 0.6, and 0.8 crossed with $n =$ focused on $\hat{\beta}_1$, but looked at other parameters in a few c another set of 10,000 samples). To these cases 1 to 5 for and 7, as noted in Table 5.21. In total we looked at 29 simu

We tabulated some basic results (Table 5.22) wherein lection was applied to all 16 possible models (the labeli same as in Table 5.12). In particular, we tabulated the 9 percentiles of Δ_p. There is one variation here; with no n than the generating global model (g_{15}) if sample size gets converges on model g_{15}, and all percentiles of Δ_p go to effect (i.e., some degree of reduction in percentiles of Δ occur in real data analysis, so we flagged populations w fect is occurring. Our recommendations about interpreti no boundary effect occurs. For the 15 populations where occurred, the mean percentiles of Δ_p in Table 5.22 are 4.7 95th, and 99th, respectively).

200, 500, 1,000]; β_i are rounded to two decimal places.

Case	ρ	n-set	$\sigma^2_{y\mid x}$	v	β_1	β_2	
1	0	a	1	1.00	0.8	0.6	
2	0.2	a	1	1.09	0.69	0.44	
3	0.4	a	1	1.36	0.73	0.39	
4	0.6	a	1	1.96	0.93	0.43	−(
5	0.8	a	1	3.82	1.65	0.65	−(
6	0	a	1	1.00	0	0	
7	0	b	25	1.00	0.8	0.6	

Sample size also has an effect on the distribution of Δ effect for n greater than about 20. For case 7 in Table 5.22 size ($20 \leq n \leq 500$) is about 2 units at the 90th and 95th p 3 units at the 99th percentile. These are typical of sample observed.

There is considerable model selection uncertainty in populations (Table 5.22), for example, as indexed by ho probability π_k is even for the expected AIC_c best model model selection uncertainty used in Table 5.22 is simply a of models, of the 16, that have selection probabilities $\geq$ half of the possible models). For the all too typical app selection with 10 or more variables ($R \geq 1{,}024$) and no the selection process will be highly unstable (cf. Breim model is selected. That is, selection probabilities can be low, even for the actual K-L best model, and not exhibit result the selected model itself is not at all the basis for about the relative importance of the predictor variables, model provides reliable predictions.

A confidence interval for a parameter β_i was compu parameter was in the selected model, in which case point were the MLE $\hat{\beta}_i$ and the model-averaged $\hat{\bar{\beta}}_i$. Three typ computed: the classical conditional interval $\hat{\beta}_i \pm 2\,\widehat{\text{se}}(\hat{\beta}_i \mid g$ unconditional interval $\hat{\beta}_i \pm 2\,\widehat{\text{se}}(\hat{\bar{\beta}}_i)$, and the interval b averaged point estimate $\hat{\bar{\beta}}_i \pm 2\,\widehat{\text{se}}(\hat{\bar{\beta}}_i)$. One result was th interval types two and three was barely different, but wa the interval $\hat{\bar{\beta}}_i \pm 2\,\widehat{\text{se}}(\hat{\bar{\beta}}_i)$ (ratio of coverages: 0.995). Henc only for this latter interval (Table 5.23). We focused on (i.e., parameter index 1); Table 5.23 gives coverage result of a difference parameter (β_3 or β_4).

column six is the number of models for which $\pi_i \geq 0.01$, also given

Case	ρ	n	Best Model	π_k best	# π_i ≥ 0.01	Percentiles 90	95
1	0	20	11	0.243	13	5.37	6.74
1	0	50	15	0.368	4	2.55	2.59
1	0	50	15	0.378	4	2.56	2.59
1	0	100	15	0.668	2	1.81	2.12
2	0.2	20	5	0.342	12	4.65	6.51
2	0.2	20	5	0.334	12	4.58	6.41
2	0.2	50	11	0.264	7	2.67	3.78
2	0.2	100	11	0.462	4	2.22	3.55
3	0.4	20	1	0.398	12	7.03	9.35
3	0.4	50	12	0.353	8	3.35	4.29
3	0.4	100	12	0.611	6	2.17	3.36
4	0.6	20	8	0.227	12	6.38	8.73
4	0.6	50	12	0.544	8	2.47	3.65
4	0.6	100	12	0.743	4	1.74	2.89
5	0.8	20	8	0.499	10	5.43	7.68
5	0.8	50	15	0.305	4	3.32	4.03
5	0.8	50	15	0.308	4	3.24	4.03
5	0.8	100	15	0.595	4	2.04	2.23
5	0.8	100	15	0.597	4	2.03	2.23
6	0	20	16	0.594	11	3.43	5.11
6	0	50	16	0.535	11	3.35	4.96
6	0	100	16	0.518	11	3.41	4.90
7	0	20	16	0.493	11	4.55	6.47
7	0	50	1	0.180	13	5.62	7.48
7	0	100	5	0.131	16	5.11	6.90
7	0	200	11	0.143	15	4.08	5.20
7	0	200	11	0.140	15	4.10	5.20
7	0	500	11	0.391	8	3.42	4.93
7	0	1000	15	0.380	4	2.00	2.04

Confidence interval coverage is affected by bias in eithe
or its standard error estimator. Therefore, we tabulated i
biases for the interval based on the model-averaged estin
In this context bias is important only in relation to standar
5.23 shows,

$$\delta = \frac{\mathrm{E}(\hat{\bar{\beta}}_i) - \beta_i}{\mathrm{E}(\widehat{\mathrm{se}}(\hat{\bar{\beta}}_i))},$$

coverage is trivial for $|\delta| \leq 0.25$ and ignorable for $|\delta| \leq$

The other factor that can affect coverage is bias in th
hence we show the ratio

$$\text{se-}r = \frac{\text{E}\left(\widehat{\text{se}}\left(\hat{\bar{\beta}}_i\right)\right)}{\text{MC-se}\left(\hat{\bar{\beta}}_i\right)}$$

in Table 5.23. Here, MC-se($\hat{\bar{\beta}}_i$) is the actual achieved stan
the Monte Carlo samples (out of 10,000) wherein $\hat{\bar{\beta}}_i$ is c
se-$r = 1$ is desirable. Coverage would be reduced (other
if se-r becomes much less than 1. The effect on cover
$0.9 \leq$ se-$r \leq 1.1$. For all these results, the relevant samp
"Freq." in Table 5.23: the number of samples wherein $\hat{\bar{\beta}}_i$ i
the parameter is in the AIC_c-selected model.

The biggest surprise was the high achieved coverage of
ditional) confidence interval (Table 5.23). When that cov
example, case 2, $\rho = 0.2$, $n = 20$, unconditional covera
it was because of severe bias in either the point estimator
estimator. Moreover, these biases are clearly a form of n
and they occurred when the reference parameter was infrec
infrequent selection of any model containing β_i), which it
predictor variable x_i being unimportant at the given sam
the matter, however, not all instances of small frequency
including β_i resulted in deleterious effects on coverage (
$\rho = 0.8, n = 50$, unconditional coverage of 0.925 on β_3).
if a predictor variable was important (as judged by high s
its unconditional (and conditional) coverage was always

In all 29 simulated populations the unconditional int
greater than or equal to the conditional coverage and prov
erage when the conditional coverage was less than 0.95
5, $\rho = 0.8$, $n = 20$, conditional and unconditional cov
sus 0.937 on β_1). For the 23 populations where the paran
to the selected model, the average conditional and uncon
interval coverage was 0.930 versus 0.947 (and $\overline{\delta} = 0.2$
improvement in coverage is not dramatic, but is generally

For the other six populations, the bias in coverage is s
strong biases in point estimates or their standard errors, a
se-r ($\overline{\text{abs}(\delta)} = 0.64$ and $\overline{\text{se-}r} = 0.63$). These biases are a d
selection, i.e., they are model selection bias. In those case
model is not the expected AIC_c best (i.e., expected K-L
it includes a variable x_s that is rarely included in the s

of such selected models is denoted by "Freq."; coverage is for the
interval and the interval based on the model-averaged estimator $\widehat{\widehat{\beta}}_i$;
of δ and se-r.

Case	ρ	n	Parm. index	Freq.	Coverage	
					cond.	MA
1	0	20	1	8864	0.932	0.948
1	0	50	1	9994	0.943	0.948
1	0	50	4	4337	0.934	0.934
1	0	100	1	10000	0.948	0.950
2	0.2	20	1	8309	0.923	0.947
2	0.2	20	3	2106	0.694	0.755
2	0.2	50	1	9970	0.945	0.955
2	0.2	100	1	10000	0.944	0.950
3	0.4	20	1	8026	0.908	0.945
3	0.4	50	1	9912	0.918	0.947
3	0.4	100	1	9999	0.932	0.944
4	0.6	20	1	8178	0.905	0.957
4	0.6	50	1	9953	0.905	0.942
4	0.6	100	1	10000	0.937	0.945
5	0.8	20	1	9175	0.858	0.937
5	0.8	50	1	10000	0.918	0.945
5	0.8	50	3	3572	0.913	0.925
5	0.8	100	1	10000	0.930	0.946
5	0.8	100	3	6055	0.957	0.960
6	0	20	1	1220	0.396	0.495
6	0	50	1	1450	0.643	0.677
6	0	100	1	1501	0.658	0.676
7	0	20	1	2014	0.726	0.759
7	0	50	1	3624	0.904	0.913
7	0	100	1	5594	0.952	0.955
7	0	200	1	7911	0.967	0.969
7	0	200	4	2225	0.785	0.796
7	0	500	1	9828	0.969	0.969
7	0	1000	1	9998	0.956	0.956

inference on β_s (an unimportant variable) can be very m
resultant model selection bias for $\hat{\beta}_s$. Fortunately, this sce
almost by definition, since it is a case of x_s being comm
the selected model. Also, even then inference on an impo
generally sound in these simulations and others we have
ence on an unimportant β_s was slightly (but not strongly)
unconditional confidence interval was considered.

reference to linear regression and so-called variable select
the most used and misused application area of model sel
every conceivable type of model (variable) selection me
been tried in the context of having m predictors and u
regression models (see, for example, Hocking 1976, Dra
Henderson and Velleman 1981, Breiman and Freedman
Miller 1990, Hjorth 1994, Breiman 1995, Tibshirani 1996,
However, almost always the statistical literature approach
it is only a matter of "just-the-numbers" data analysis met
fact there is always a subject-matter scientific context, and
of sample size, that must be brought into the problem, and
an enormous difference as compared to any naive mode
that does not consider context, prior knowledge, and sam

Secondly, there is always a goal of either (1) selectin
should include ranking competitor models) because one
of the relationships (presumably causal) between $\underline{x}$ (inc
and y, or (2) prediction of $E(y \mid \underline{x})$ at values of $\underline{x}$ (predicto
(prediction of $E(y \mid \underline{x})$ for $\underline{x}$ in the sample can be consid
estimation). These goals really are different. That is, if t
model selection uncertainty, then selecting the best mode
using it for goal (2), prediction, is not optimal.

We recommend that prior to any data analysis full cons
how the problem (i.e., set of models) should be structured
means dropping variables that cannot reasonably be rela
tion, or cannot reasonably be causally related at detectabl
the sample size. From the literature and our experience,
too reluctant to drop clearly irrelevant variables and othe
considerations based on logic and theory. This is the "mea
is easy to measure and let the computer sort it out" syndr
work. Even a good exploratory analysis needs input of inv
reach useful results. In this regard we quote Freedman et

> A major part of the problem in applications is the curse o
> there is a lot of room in high-dimensional space. That is
> need model specifications tightly derived from good th
> expect statistical modeling to perform at all well in an e
> sisting of large, complicated data sets and weak theor
> at present that describes many applications.

An important a priori aspect is to consider reducing the
dent variables by functionally combining them into a small
variables. This may be as simple as computing, by observ
some of the predictor variables (such as in the GPA exam

density as mass per unit volume in physics of animals per
rates of all sorts, and so forth). Consider also any bounds
often college GPA is bounded on 0 to 4; hence, we do no
be able to make a prediction of 4.2. We could model GP
link function and rescale predictions by 4.

**As a rule of thumb, the maximum number of stru
to allow in a regression (or other univariate) model s**
not possible to reliably estimate anything like $n/2$ (or n
"noisy" data. Mistakenly, models of such size are often fit
selected based on an invalid criterion such as minimum
an inappropriate to the situation criterion like adjusted R^2
correctly adjusts for either small sample size or large K).

To illustrate some of this thinking we consider anoth
example of variable selection (Hocking 1976; see also Hen
1981 for the actual data; we did not read this latter paper
below): automobile gas mileage (y) as MPG (miles per ga
independent variables (there is clearly causation involve
not so we use the term predictor variable). Note that y
variable. The 10 x_i are:

1. Engine shape (straight or V)
2. Number of cylinders (4, 6, or 8)
3. Transmission type (manual or automati
4. Number of transmission speeds
5. Engine size (cubic inches)
6. Horsepower
7. Number of carburetor barrels
8. Final drive ratio
9. Weight
10. Quarter mile time

The data arose from testing 32 ($= n$) different types o
standardized conditions. We independently generated a p
(it would be better to get an automotive engineer involved).
at the dependent variable (either as y alone, or as y versu
given that one has decided that the analysis will be only t
a response variable y based on $x_1, \ldots, x_m$ as predictor va
the models are all conditional), one next reduces the numb
as much as possible by logical and subject-matter consi
resultant reduced set of predictors, we recommend lookin
of those independent variables to be assured that there ar
having an extremely high pairwise correlation.

A more comprehensive examination would be a principa
value evaluation of the design matrix X. Such results ar
(1976) for the full set of 10 predictors (but the data are n

eigenvalues add to about 90% of the total of all 10 eige
explained variation is often all we can hope for without o
sample size 32.

The x_i are highly intercorrelated in this observational
design: Big cars have bigger engines; are more likely to
cylinders; are therefore more likely to have a V-engine c
For such observational studies if the issue of interest is
substantial inference problems (see Draper and Smith
some sage cautionary comments on such problems with c

One of us (KPB, who is automotively challenged) pr
Because $n = 32$, do not include more than three structu
mileage is strongly dependent on car weight, so always in
an intercept will be used here, this leaves room for only o
a first thought, then, consider the nine models

$$y = \beta_0 + \beta_1 x_9 + \beta_i x_i + \epsilon, \qquad i \neq 9$$

plus $y = \beta_0 + \beta_1 x_9 + \epsilon$. However, bearing in mind the inte
these variables, consider dropping some on a priori groun
x_i based just on a high correlation unless it is extreme such
then there is a variable redundancy problem (near coline
(near) colinearity problems; and do eliminate variables b
reasoning, and experience.

As a type of thought experiment (because the data di
experiment) consider whether engine shape (x_1) is causa
Do we really think that if all car features were held fixed
engine is a straight or V8 that there would be any effect
either not at all, or at a level we will never care about a
except with an experiment and a huge sample size. Concl
x_1 (we surmise that it was recorded because it is easy to
not justification for including a variable). **Recommend
experiments in conjunction with observational studies**

The same reasoning leads KPB to drop variable 2: n
Again the thought is that if all else (horsepower, total cy
etc.) were fixed, would just number of cylinders (as 4,
MPG? And again, no; at least not in these data. Variable
be retained on fundamental grounds. Quarter mile time is
derived variable; it might predict MPG well, but it is n
Instead, variable 10 might itself be well predicted based
Conclusion: drop x_{10}. Thus KPB would consider only mo
always included and at most one of variables 3–8 (seven m
different from an all-subsets selection over $10^m = 1{,}024$ m
of thinking can, and should, always be brought to bear on
problem.

that horsepower itself is a derived variable but one that et
some extent horsepower might replace variables 5 and 7
important than variables 3, 4, 6, and 8. Thus we have the
that the best two-variable model might be based on x_9 a
was not best, but nearly tied for best with a less-interpr
priori thinking would justify objective selection of these t
important.

Another issue is the suitability of the linear model for
bounded below by zero but weight can be unbounded, a
predict negative MPG. Surely, over a big enough weight ra
is curvilinear, such as $\mathrm{E}(y \mid x_9) = \beta_0 e^{-\beta_1 x_9}$ or $\mathrm{E}(1/y \mid x_9) =$
MPG only by convention; hence a priori KPB would fit
but based on an inverse link function to MPG. One could j
linear models and select a best model. However, AIC_c is
from models fit to y versus models fit to $1/y$.

Less time was afforded to this exercise by DRA (who
motively knowledgeable), who independently put forth t
Both include weight (x_9);

$$y = \beta_0 + \beta_1 (x_9)^2 + \beta_2 z + \epsilon$$

and

$$y = \beta_0 + \beta_1 (x_9)^2 + \beta_2 x_{10} + \epsilon.$$

The variable z is a derived variable meant to reflect the
several variables on MPG:

$$z = \frac{x_2 \times x_5}{x_6}.$$

Similarly, x_{10} is used here as a predictor that summarizes
ability of the car to consume fuel.

Considerations like these based on reasoning and theo
about before data analysis if reliable uncertainty bounds
an inference made after model selection. One can alway
exploratory analyses after the a priori analysis. We just rec
the two processes, because results of exploratory analyses
reliable inferences in the sense that the data cannot both
(the model) and then reliably affirm the inferential uncer
(the same model). There is a saying from the USA western
first, ask questions later." This strategy often precludes c
formation. Similarly, "compute first, then create models"
think later") is also not a strategy for making reliable infe
be a model that describes the data very well (because it c
is a poor model as an inference to independent data from

5.3.6 *Discussion*

A variety of comments and opinions are given here, some
fit well elsewhere in Section 5.3. There is no particular or
comments and opinions.

The Monte Carlo simulations of Sections 5.3.2 and 5
violate our general philosophy that the actual data-genera
(should) in reality be more general than the global model
for data analysis (an expanded vector of predictors $\underline{x}_T$ wo
true generating model; the global model does not use all
This is only partly true. The part that is not true is think
generated the data under the global model, no more gener
ally apply. In fact, the residual variation of the global reg
is a confounding of average (with respect to f) model
arising from the differences $\mathrm{E}(y \mid \underline{x}, G) - \mathrm{E}(y \mid \underline{x}_T, f)$, p
plained residual variation σ_ϵ^2 (it might be 0) under f. Thu
of a conceptually more general data-generating model an
the global model. Hence, it is more economical simply to
an assumed global model.

One way in which this lacks generality is that the num
components of the true parameter vector $\underline{\beta}_T$ that are in the
global $\underline{\beta}$ may not exactly equal their counterpart comp
This would affect confidence interval coverage, which s
the appropriate components of $\underline{\beta}_T$, not to $\underline{\beta}$. This seems lik
initial Monte Carlo studies intended to explore basic mod

The more important lack of generality relates to how we
totic sequence of models as sample size increases. Classi
model or set of models fixed, independent of sample size.
with reality, because as sample size grows we will inclu
the data and in our models (e.g., in the GPA example, e
school type, university attended, major, student age, and
reason we should simply have a larger global generating
we used in simulations here, and include more factors so
selected model can grow without the arbitrary bound of
only four predictors. In a sense the issue becomes one of
ing" up against a bound (i.e., large π_i for the generating m
increases, because this feature of data analysis is often uni
tional studies. This problem is solved simply by having a
global generating model, and it is then still acceptable (
have that generating model also as the global model for d

It is well known that selecting a best model from a set o
can lead to important biases in parameter estimates and a

selecting a good model, because all selection criteria inv
seeking a fitted model with a relatively small residual su
AIC_c will not entirely protect one from this possible bi
e.g., Table 5.18; the relative bias in $\hat{\sigma}^2_{y\,|\,x}$ is not high for th
and this example is for a small sample size). The bias i
more infrequently a model is selected because for such p
selected only when they fit a sample unusually well. In the
of a study (hence, sample size is then a given) the best w
bias in $\hat{\sigma}^2_{y\,|\,x}$ is to keep the candidate set of models small.

When the predictors are intercorrelated and model sel
selection tends to induce a bias in the estimators of regre
selected predictors. The less important a predictor x_i is, t
be selected, and then when selection occurs, both of the a
$\hat{\beta}_i$ or $\hat{\bar{\beta}}_i$, conditional on the model, tend to be biased awa
let $E(\hat{\beta}_i \mid g_r$ always) denote the expected value of $\hat{\beta}_i$ unde
model is always fit to the data. Let $E(\hat{\beta}_i \mid g_r$ selected) de
expectation of $\hat{\beta}_i$ when model g_r is selected, as by AIC_c. If
0, then we usually find that $E(\hat{\beta}_i \mid g_r$ selected$) > E(\hat{\beta}_i \mid g_r$
$E(\hat{\beta}_i \mid g_r$ always$) < 0$, then we find that $E(\hat{\beta}_i \mid g_r$ selected)
The strength of the bias depends mostly on the importanc
measured by its overall selection probability (and that pro
sample size and goes to 1 as n goes to infinity if $|\beta_i| > 0$

Consider Table 5.17, which gives Monte Carlo result
mimicking the GPA data. Using that information, and ex
computed the percent relative bias of $\hat{\beta}_i$, $PRBias(\hat{\beta}_i)$ bel
to model selection. The reference value for computing b
of β_i from the data-generating model, not the parameter v
conditionally to model g_r when model g_r is always fit to t
frequency of occurrence of the given parameter (i.e., predi
model is denoted by $Pr\{x_i\}$:

i	$Pr\{x_i\}$	$PRBias(\hat{\beta}_i)$
1	0.954	14%
2	0.751	25%
3	0.650	31%
4	0.162	267%

A 31% relative bias in conventional estimators due to mo
be of concern (let alone 267%, but x_4 is not in the K-L be

We have looked at this issue for other models and sample
selection, and it is quite clear that this aspect of model s
mators is, as above, strongly related to the importance of t
selection bias is less for a predictor always included in the

presence of better predictors they correlate with) and the
a few of those unimportant predictors will end up in the se
that happens, all the model selection biases operate in a d
think that the selected variables are important ("signifi
testing terms). The best way to reduce this risk of mislead
a small list of carefully considered candidate variables. (T
selecting one or more unimportant variables; just have a
conceived variables and a small sample size; see, e.g., Free
et al. 1988, 1990).

Two undesirable, but mutually exclusive, properties
strategies particularly relevant to all-subsets regression fo
are worth noting here: overfitting the data or overfitting
strategy is to always fit and use the global model, you w
the model (i.e., include unnecessary variables). This appr
avoid subjectively tailoring the model to the data, but you p
inflate standard errors of all the $\hat{\beta}_i$. This loss of precisio
all the estimates are worthless. Thus, usually one is forc
model selection with multivariable observational data (it
a priori considerations).

If you use a subjective selection procedure of first fitti
examining the results (e.g., residual plots, r-squares, lever
formations of variables) in search of a better model ba
synthetic criterion of your own choosing, you probably
Thus, you will include in model structure what are reall
of the data, thereby possibly biasing $\hat{\sigma}^2_{y|x}$ quite low and
"noise" as real structure. The resultant model may becom
tion of the particular data at hand than a valid inference
samples have their nearly unique peculiarities as well as
that would show up in all, or most, samples you might get
uation at hand. Inference is about correctly identifying th
of samples. When you overfit the data, you mistakenly inc
ture uncommon data features that would not be found in
that might arise.

When you have a large number of models for a much
variables (like $R = 1{,}024$, $m = 10$, all-subsets) and a
fit and considered for selection, you run a high risk tha
overfit the data. The use of AIC_c reduces this risk (beca
looks at model fit penalized by a function of model size
n; but there is no built-in "penalty" for having a huge
does not eliminate it for all-subsets selection: Some deg
remains. For this reason, and the instability of all-subsets s
to properly evaluate model selection uncertainty under a
and use inference rather than just use the selected best m

is, either the bootstrap failed (the clients data, Section 4
some peculiar results that the theoretical approach did no
ample, Section 5.3.1). Two small studies that evaluated th
of assessing aspects of model reliability after model sel
expressed pessimism that the bootstrap would always be a
the task (Freedman et al. 1988, Dijkstra and Veldkamp 198
we can now offer is that for even moderately high-dimens
$m \geq 7$, hence $R > 100$) one should not blithely think that
a reliable way to assess model selection uncertainty for a
the method needs more study.

In fairness, it can also be said that AIC_c and associated
here need more evaluation for their performance under a
However, a more basic issue is whether or not ever to do
(especially when the number of predictors is large) when
a single best model and ignoring all other models. The pr
instability of what model is selected (cf. Brieman 1996) an
selection biases. Model instability arises when all model se
(i.e., the π_i) are low. For large R (hence if m is at all larg
the expected K-L best model might have selection proba
so the set of supposed important regressor variables, as ju
best model, can vary dramatically over samples (An ins
given in Chapter 6 for $R = 8{,}191$ models).

At a fundamental level the question of variable selectio
a question of the strength of evidence in the data for the
predictor variable. If the problem is thought of this wa
selection as such is an illegitimate discretization of what o
of estimating continuous parameters (the regression coef
model). The flaw in using model selection is then just li
hypothesis testing that makes a problem a reject-or-not
ought to be approached as an evaluation of strength of e
P-values rather than strict reject-or-not procedures is also
a flawed methodology; see, e.g., Harlow et al. 1997, Sel
strongly recommend against doing all-subsets selection wi
is identifying a single "best" model: Promoting this practic
and is a failing of statistical science.

We believe that the only defensible reason for fitting all-
models should be to obtain the full set of Akaike weights,
are based on the full set of models as mediated by thei
weights (i.e., model averaging). The selected best model
subset (of R) of the predictor variables. Unless the Akaike
model is very high (say $w_k \geq 0.9$), we maintain that it is t
infer that one has found *the* important predictors, and th
selected are unimportant. As noted above, it is not properly

value is only with respect to prediction, reliable causal i
made from just the data alone when those data are from an

We are led to believe that the only legitimate application
fitting with purely observational data (and then only aft
of the number of predictors, as discussed in Section 5.3.
prediction in this context we recommend model averaging
is made with each model, and the Akaike weights are
weighted average of these predictions. We do not know
averaging, but we have seen it only in the Bayesian litera
Bayesian-based model weights (see, e.g., Madigan and R
1995, Hoeting et al. 1999, Raftery et al. 1997, Hoeting et

As noted by Breiman (1996), selection of a best model
is inherently unstable in its outcome. The solution prop
produce stabilized inferences is a type of model averag
perturbed sets of the data (such as bootstrap samples cr
model in each case, and produce some sort of averaged inf
for stabilized inference, is a sort of reverse strategy: Keep t
as is, but find for each fitted model its Akaike weight; then
as some form of weighted average over all the models.

Interest in regression parameter estimates in conjuncti
all-subsets model fitting will no doubt continue. Perhaps
need for this (we are not convinced). Motivated by thi
ommendation to use model-averaged predictions, we de
relate such prediction to parameter estimation. This led to
ideas and issues that we will outline here. These are issue
Chapter 6, but still need additional research.

The model-averaged prediction (estimate) of $\mathrm{E}(y \mid \underline{x})$ is

$$\hat{\bar{\mathrm{E}}}(y \mid \underline{x}) = \sum_{r=1}^{R} w_r \hat{\mathrm{E}}(y \mid \underline{x}, g_r).$$

We define an indicator function for when a predictor is

$$I_i(g_r) = \begin{cases} 1 & \text{if predictor } x_i \text{ is in model} \\ 0 & \text{otherwise.} \end{cases}$$

For model g_r the value of β_i is denoted here by $\beta_{i,r}$. One
averaged parameter estimator is

$$\hat{\bar{\beta}}_i = \frac{\sum_{r=1}^{R} w_r I_i(g_r) \hat{\beta}_{i,r}}{\sum_{r=1}^{R} w_r I_i(g_r)} = \frac{\sum_{r=1}^{R} w_r I_i(g}{w_+(i)}$$

$$w_+(i) = \sum_{r=1}^{R} w_r I_i(g_r).$$

we simply set $\beta_{i,r} = 0$. Thus a new estimator, denoted by

$$\tilde{\bar{\beta}}_i = w_+(i)\hat{\bar{\beta}}_i.$$

This is just $\hat{\bar{\beta}}_i$ shrunk toward zero by the amount $(1 - w_+($
found, based on empirical results, that we could also cor
estimator as

$$\tilde{\bar{\beta}}_i \equiv \hat{\bar{\beta}}_i - (1 - w_+(i))\hat{\bar{\beta}}_i = \hat{\bar{\beta}}_i - \widehat{\text{model select}}$$

that is, (5.8) is our original model-averaged estimator adj
model selection bias. Certainly, the term $(1 - w_+(i))\hat{\bar{\beta}}_i$
estimator of model selection bias, but it is a usable estima

Then we realized that $\tilde{\bar{\beta}}_i$ is of fundamental importance
averaged prediction can be expressed as

$$\hat{\bar{\mathrm{E}}}(y \mid \underline{x}) = \hat{\bar{\beta}}_0 + \sum_{i=1}^{m} w_+(i)\hat{\bar{\beta}}_i x_i = \hat{\bar{\beta}}_0 + \sum_{i=1}^{m}$$

If we accept $\hat{\bar{\beta}}_i$ as the appropriate naive estimate of β_i g
ference, then heuristically, the above suggests that predic
shrinkage toward zero of each parameter's estimate by a m
eter's unimportance $(= 1 - w_+(i))$. The value of shrinkag
in statistics (see, e.g., Copas 1983, Tibshirani 1996); her
line of thought worth pursuing.

Thus we have compelling reasons to want to replace, a
regression, the conditional estimator of (5.7) by the unc
of (5.8). This would allow us to ignore the issue of what
simply make inferences from the full set of models as re
or prediction. In our limited Monte Carlo evaluation of thi
that $\tilde{\bar{\beta}}_i$ is less biased by model selection than is $\hat{\bar{\beta}}_i$.

An unresolved matter is a simple, yet reliable, estima
variance of $\tilde{\bar{\beta}}_i$ and an associated confidence interval for
oretical sampling variance formula for a model averag
here:

$$\text{var}\left(\tilde{\bar{\beta}}_i\right) = \left[\sum_{r=1}^{R} \pi_r \sqrt{\text{var}\left(\hat{\beta}_{i,r} | g_r\right) + \left(\beta_{i,r} -\right.}\right.$$

However, what we need is an estimator of $\text{var}\left(\tilde{\bar{\beta}}_i\right)$. Wh
interest, say θ, appears in every model (hence, θ_r in mod

$$\widehat{\text{var}}\left(\bar{\theta}\right) = \left[\sum_{r=1} w_r \sqrt{\widehat{\text{var}}\left(\hat{\theta}_r|g_r\right) + \left(\hat{\theta}_r - \bar{\theta}\right.}\right.$$

This variance estimator makes no allowance for the uncer
inherent in $\hat{\bar{\theta}}$ because the weights $w_r(=\hat{\pi}_r)$ are random
because these weights are positive and sum to 1, a decre
compensated for by an increase in one or more other we
that the result is an acceptably stable variance estimator ev
sampling variation of the weights.

Applying the above variance estimator to $\tilde{\bar{\beta}}_i$ the result, e
is

$$\widehat{\text{se}}(\tilde{\bar{\beta}}_i) = \sum_{r=1}^{R} w_r \sqrt{\widehat{\text{var}}\left(\hat{\beta}_{i,r}|g_r\right) + \left(\hat{\beta}_{i,r} - \tilde{\bar{\beta}}_i\right)^2} + |\tilde{\bar{\beta}}$$

Now the sampling variation in the weights matters ver
formance of this variance estimator. Another way to
note that $\tilde{\bar{\beta}}_i = w_+(i)\hat{\bar{\beta}}_i$, and the variance of the co
reliably estimated (because one renormalizes the Akail
subset of models that contain β), but now we need a for
$(w_+(i))^2\,\widehat{\text{var}}(\hat{\bar{\beta}}_i) + (\hat{\bar{\beta}}_i)^2\,\widehat{\text{var}}(w_+(i))$. We do not know $\widehat{\text{va}}$
simulations we know that it is not trivial (i.e., cannot be i

Despite seeking one, we do not yet know a reliable ana
$\widehat{\text{se}}(\tilde{\bar{\beta}}_i)$. Moreover, if we had one, there is a second issue as
interval on β_i under this unconditional model averaging. T
$\tilde{\bar{\beta}}_i \pm 2\,\widehat{\text{se}}(\tilde{\bar{\beta}}_i)$ is not justified, in general, because the sam
$\tilde{\bar{\beta}}_i$ can be *very* skewed (i.e., quite nonnormal) when $\text{E}(w_+$

While finding a reliable estimator of $\text{var}(\tilde{\bar{\beta}}_i)$ is worthw
have the issue that the sampling distribution of $\tilde{\bar{\beta}}_i$ can b
distribution. Hence, the only reliable approach to frequent
seems to be the bootstrap. To apply the bootstrap to this
one must compute this estimator in each of the B bootstra
will obtain $\tilde{\bar{\beta}}^*_{i,b},\ b = 1, \ldots, B$, and then determine the p
interval, and, if desired, the usual bootstrap-based estima

We again consider the Monte Carlo evaluation of the
Results in Table 5.17 were extended to β_4 and $\tilde{\bar{\beta}}_i$ to comp
We need to be clear on what was done in this new Mon
predictor variable x_i was in the selected AIC_c best mode

these estimates is $E(\beta_i)$ under model selection, the expe
size is $10{,}000 \times \pi_+(i)$ for $\pi_+(i)$ the probability that x_i is
model. In contrast, $\tilde{\bar{\beta}}_i$ was computed for every Monte C
$E(\tilde{\bar{\beta}}_i)$. The results are below; $\tilde{\bar{\beta}}_i$ has the better performan

i	$\pi_+(i)$	β_i	$E(\hat{\beta}_i)$	$E(\tilde{\bar{\beta}}_i)$
1	0.95	0.00201	0.00229	0.0021
2	0.75	0.00125	0.00156	0.0011
3	0.65	0.18945	0.248	0.1577
4	0.16	0.08752	0.321	0.0588

5.4 Estimation of Density from Line Tra

5.4.1 Density Estimation Background

Animal inventory and monitoring programs often focus
population density (i.e., number per unit area). Buckland
provide the theory and application for field sampling a
using line transects. We will illustrate several aspects of
in the face of model selection uncertainty using line tra
by Southwell (1994) on the eastern grey kangaroo (*Mac*
Wallaby Creek, in New South Wales, Australia. The p
(Laake et al. 1994) was written for the analysis of line trans
model selection, and has an option for bootstrapping the sa
et al. 1997 for a similar example). Thus, line transect samp
DISTANCE will be used to provide some deeper insight
selection uncertainty and will serve as another comprehe

In line transect sampling, the estimator of density (D)

$$\hat{D} = \frac{n}{2wL\hat{P}},$$

where n (= 196 in this example) is the number of objec
78) transects of total length L (= 88.8 km) and width
unconditional probability of detection, within the strips
focus of the estimation of animal density is on the probabi
this is defined as

$$P = \frac{\int_0^w g(x)dx}{w},$$

where $g(x)$ is the detection function (i.e., the probability o
an animal is at perpendicular distance x from the line).

in detectability among animals being surveyed. The detec
estimated from perpendicular distances taken from the tra
ject detected. Assumptions required in line transect sampl
and theory are given in Buckland et al. (2001).

Substituting the expression for P into the estimator of $
the w and $1/w$, gives

$$\hat{D} = \frac{n}{2L \int_0^w \hat{g}(x)dx}.$$

Thus, the essence of data analysis here is to find a good a
for $g(x)$, the detection function. Buckland et al. (2001) re
the general form

$$g(x) = key(x)[1 + series(x)].$$

The key function alone may be adequate for modeling $g(x)$
size is small or the distance data are easily described by a s
one or more adjustment terms must be added to achieve a
for the data. For the purposes of this example we chose
models of the above form. Each of these models provides
unique, basis for data analysis in this example.

5.4.2 *Line Transect Sampling of Kangaroos at W*

Eastern grey kangaroos often occur in family groups; thus
animal density is the product of the estimated number of gr
group size. In this example we will focus only on estim
groups of this species of kangaroo. We define the set o
for this example in Table 5.24. The analysis theory of lin
has been the subject of a great deal of work since abou
of candidate models is relatively well based in this exa
DISTANCE (Laake et al. 1994) was used to compute
parameters in the key functions (σ, a, or b, in Table 5
expansions (the a_j in Table 5.24), and choose, using A
among the four.

5.4.3 *Analysis of Wallaby Creek Data*

The results of the initial analysis of these data are given in
was selected using AIC and provides an estimated densit
km^2 (conditional se = 1.00 and conditional cv = 10.12
produce relatively similar point estimates of D for these
value was 3.86 (model 3). In this example, the estimated lo

half-normal key function.

Model	$g(x)$	K	Key functio
1	$\left\{\frac{1}{w}\right\}\left\{1+\sum_{j=1}^{2} a_j \cos\left(\frac{j\pi x}{w}\right)\right\}$	2	uniform
2	$\left\{e^{-x^2/(2\sigma^2)}\right\}\left\{1+\sum_{j=2}^{3} a_j H_{2j}\left(\frac{x}{\sigma}\right)\right\}$	3	half-no
3	$\left\{1-e^{-(x/a)^{-b}}\right\}\left\{1+\sum_{j=2}^{3} a_j \left(\frac{x}{w}\right)^{2j}\right\}$	4	hazard
4	$\left\{e^{-x^2/(2\sigma^2)}\right\}\left\{1+\sum_{j=2}^{3} a_j \cos\left(\frac{j\pi x}{w}\right)\right\}$	3	half-no

TABLE 5.25. Summary statistics for the line transect data on eas
Wallaby Creek, New South Wales, Australia (from Southwell 19
model is shown in bold; $\hat{D}_i$ is used to clarify that $\hat{D}$ is based on mo

Model	K	$\log\left(\mathcal{L}(\hat{\theta})\right)$	AIC	Δ_i	$\exp(-\Delta_i/2)$	w_i
1	**2**	**−1,021.725**	**2,047.449**	**0.000**	**1.000**	**0.499**
2	3	−1,021.546	2,049.092	1.643	0.440	0.220
3	4	−1,021.654	2,051.307	3.858	0.145	0.072
4	3	−1,021.600	2,049.192	1.743	0.418	0.209

number of model parameters, and estimated density are si
however, the estimated conditional sampling variances
almost 3. In this case, all four models contain the sam
equivalently, P); thus, model averaging (4.1) and (4.9) sho
either analysis, estimates of unconditional variances and a
intervals should be used in making inferences about popu

Using the Akaike weights w_i and the conditional
$\widehat{\text{var}}(\hat{D}_i|g_i)$ for each model, we computed the model-a
density $\bar{\hat{D}} = 10.19$ and an estimate of its unconditiona
$\widehat{\text{var}}(\bar{\hat{D}}) = 1.51$ (4.9). Hence, the (unconditional) standar
and its cv is 12.06%. This unconditional cv is slightly h
10.12% conditional on the AIC-selected model. Inferen
tially the same here whether based on the model-averaged
the density estimate from the AIC-selected model but u
the unconditional estimate of 1.51. In either case the ach

5.4.4 *Bootstrap Analysis*

The most obvious advantages of using Akaike weights as
estimated unconditional sampling variances (and $\widehat{\overline{D}}$) are s
However, the bootstrap method also can be used to make
ences; the bootstrap is especially useful in complex situ
for analytical variances, even given the models, is lackir
such analytical theory to compare to the bootstrap results

We used the program DISTANCE to draw and analyz
based on transects as the sampling unit (thus, there were 7
the bootstrap), and thereby compute lower and upper co
as well as an estimated unconditional sampling variance
10,000 bootstrap samples; we present first the results from
Then we examine the variability inherent here in a "me
samples, based on the 10 sets of 1,000 samples each f
1,001–2,000, and so forth.

The resultant density estimates, by model, and the mod
cies are shown in Table 5.26. The mean of the estimat
bootstrap samples, 10.39, is quite close to the estimate b
weights and (4.1) and (4.9) (10.19, $\widehat{\text{se}} = 1.23$). Based on
$\hat{D}^*$, the bootstrap estimate of the unconditional standard e
and $\overline{\hat{D}^*}$) is 1.48. The model selection relative frequencie
procedure are similar to, but do not exactly match, the
is expected). However, the results are close for the favore
weight $w_1 = 0.50$ (Table 5.25) and from the bootstrap
5.26).

5.4.5 *Confidence Interval on D*

There are several options for setting a confidence interva
estimated density and its estimated unconditional samp
there is the usual procedure that assumes that the samplir
estimator is approximately normal. Hence, an approxim
interval is based on

$$\hat{D} \pm 2\,\widehat{\text{se}}(\hat{D}),$$

where $\widehat{\text{se}}(\hat{D})$ (= 1.23 from Section 5.4.3) is the estimate
ditional standard error. For this example, $9.88 \pm 2 \times 1.$
(7.42, 12.34).

sampling distribution for an estimator θ in any context wh
is strictly positive, and for fixed sample size the cv($\hat{\theta}$) ten
of the actual value of θ. Then one computes lower and up
Burnham et al. 1987)

$$D_L = \hat{D}/C \quad \text{and} \quad D_U = \hat{D}C,$$

where

$$C = \exp\left[t_{\alpha/2,df}\sqrt{\log[1 + (\text{cv}(\hat{D}))^2]}\right]$$

The confidence level is $1 - \alpha$; $t_{\alpha/2,df}$ is the upper $1 - \alpha/$
the t-distribution on df degrees of freedom. The degrees
of the estimated var($\hat{D}$). For an approximate 95% interval,
it suffices to use 2 in place of $t_{0.025,df}$.

For this example $\hat{D} = 9.88$ (from the AIC-selected m
tional cv($\hat{D}$) $= 0.124$, and thus $C = 1.28$. Therefore, i
on the AIC-selected model, the approximate 95% confid
to 12.65. If we base inference on the model-averaged
(which increasingly strikes us as the preferred approach)
$\hat{\overline{D}} = 10.19$, again with standard error estimate 1.23, he
approximate 95% confidence interval 8.01 to 12.96. Th
would provide a point estimate of $\hat{\overline{D}}$, hence the correspo
terval is more comparable to the analytical results for mo
the results based on the selected single best model.

A third option is to use the bootstrap to produce a robust
for example, based on the percentile method (Efron and T
and Tu 1995). Here the 10,000 values of $\hat{D}^*_b$ generated in p

TABLE 5.26. Summary of results from 10,000 bootstrap samples
for eastern grey kangaroos at Wallaby Creek, New South Wales, Au
1994): Empirical means of the $\hat{D}^*$ by selected model and overall, s
and selection frequencies.

Model	$\hat{\overline{D}}^*$	Standard error estimate	Selectio frequen
1	9.97	1.10	4,529
2	10.63	1.41	2,992
3	10.92	2.34	1,239
4	10.75	1.38	1,240
All	10.39	1.48	10,000

confidence interval endpoints on D: $D_{(250)} \leq D \leq D_{(97$
were $\hat{D}^*_{(250)} = 7.88 \leq D \leq 13.78 = \hat{D}^*_{(9750)}$.

The interval lower bounds from the three methods are
upper bounds. Results from the bootstrap in this example
selection uncertainty than the results based on use of Aka
2.9); we rectify this matter in Section 5.4.6 below. In g
good analytical approach or the bootstrap, achievement of
interval coverage is likely if a good model is selected, if
certainty has been incorporated into an estimate of the un
error, and if nonnormality has been accounted for.

It can be problematic to identify a correct unit of d
bootstrap resampling. Aside from this fundamental iss
conceptually simple and can effectively handle model s
if computer software exists or can be written. The prog
lows bootstrapping in the context of distance sampling
In contrast, bootstrapping the experimental starling data
have been nearly impossible. Specialized software devel
case would be prohibitive; and the computer time require
in weeks. In these cases, we recommend use of Akaike
the estimate of an unconditional standard error, and then
analytical confidence interval procedure.

5.4.6 Bootstrap Samples: 1,000 Versus 10,000

The $B = 10{,}000$ bootstrap samples were partitioned, in
generated, into 10 sets of 1,000 samples per set, and estim
on a per-set basis. The results are given in Table 5.27. Bef
results we need to establish our goals for precision of
computation (estimate, actually) of quantities such as

$$\widehat{\text{se}}^*(\hat{D} \mid B) = \sqrt{\frac{\sum(\hat{D}^*_b - \overline{\hat{D}^*})^2}{B-1}}.$$

The true bootstrap estimate of the standard error of $\hat{D}$
actually the limit of $\widehat{\text{se}}^*(\hat{D} \mid B)$ as B goes to infinity. We den
by $\widehat{\text{se}}(\hat{D})$; however, this bootstrap standard error need not
number as the analytically computed standard error of $\hat{D}$ (
same notation). For any value of B we have $\widehat{\text{se}}^*(\hat{D} \mid B) = \hat{s}$
goes to 0 quickly as B gets large and $\text{var}(\epsilon) = \phi/B$ (ϕ unkn
The goal in selecting B should be to ensure that $\sqrt{\phi/B}$ is
value of $\widehat{\text{se}}(\hat{D})$. Our preference is to achieve a bootstrap un

two significant digits. That is, we target a large enough
result for $\widehat{\text{se}}(\hat{D})$ (or whatever is being computed) is near
two significant digits over all bootstrap samples of size
should be 100, we want to be assured that generally our bo
between about 99 and 101. This does not seem like too m
for; yet even this precision may require in excess of 10,00
it is rarely achieved with $B = 1{,}000$.

Now consider the variation exhibited in Table 5.27 in b
π_1, $\overline{D}$, unconditional standard error of $\hat{D}$, percentile conf
points (95%), and the interval width, $\hat{D}_{\text{U}} - \hat{D}_{\text{L}}$. Only $\hat{D}$
our precision criterion for $B = 1{,}000$. However, we do no
get $\overline{\hat{D}}^*$: We already have $\hat{D}$, from the best model and $\overline{\hat{D}}$ fro
It is the other quantities in Table 5.27 that we use the
compute. We find (empirically or theoretically) that $\hat{\pi}_1$
generally within 0.42 to 0.48; this does not meet our prec
ilarly, none of $\widehat{\text{se}}(\hat{D})$, the confidence interval bounds, or
meet our (modest) precision criterion when $B = 1{,}000$. B
over the 10 sets of samples in Table 5.27 we estimate tha
the percent coefficients of variation on the bootstrap estim
$\text{cv}(\hat{\pi}_1) = 0.005$, $\text{cv}(\widehat{\text{se}}(\hat{D})) = 0.007$, $\text{cv}(\hat{D}_{\text{L}}) = 0.004$
and $\text{cv}(\hat{D}_{\text{U}} - \hat{D}_{\text{L}}) = 0.01$. Thus here $B = 10{,}000$ is not
produce bootstrap-computed quantities reliable to (almost
When using the bootstrap, think in terms of $B = 10{,}000$.

5.4.7 Bootstrap Versus Akaike Weights: A Lesso

The estimated unconditional standard error of $\hat{D}$ is 1.23 b
formula and use of the Akaike weights. However, base
intensive bootstrap method we obtained 1.48 for the esti
standard error of $\hat{D}$. The bootstrap method is telling us th
certainty in our density estimator than our analytical (i.e.,
accounts for. We perceived a need to resolve this issue. Un
the wrong approach: We assumed that the bootstrap res
and tried to find out why. It is not wrong, but we mention
before giving the correct resolution of this matter.

The correct analytical variance of $\hat{D}$, given a model, i
tal line length L (88.85 km) and has two parts: $\text{var}(n/L$
$\text{var}(\hat{P})$. The $\text{var}(\hat{P})$ component is conditional on n (196)
and kangaroo locations may not be independent within
units used here as the basis for the bootstrapping are th

South Wales, Australia (from Southwell 1994). $\hat{\pi}_1$ is the selectio
g_1; standard errors and (percentile) confidence intervals for D are u
include model selection uncertainty. Results for "All" are based on

Set	$\hat{\pi}_1$	$\overline{\hat{D}^*}$	$\widehat{\mathrm{SE}}(\hat{D})$	95% Conf. Int.	
1	0.478	10.39	1.47	7.77	13.64
2	0.412	10.41	1.52	7.71	13.84
3	0.473	10.43	1.42	7.98	13.56
4	0.418	10.37	1.49	7.84	13.73
5	0.442	10.40	1.49	8.00	13.80
6	0.410	10.37	1.48	8.03	13.82
7	0.461	10.40	1.50	7.73	13.90
8	0.447	10.40	1.53	7.92	13.82
9	0.448	10.39	1.48	7.84	13.75
10	0.540	10.30	1.44	7.92	13.47
All	0.453	10.39	1.48	7.88	13.78

ments (78 of them). The length of these segments varies
In generating a bootstrap sample the value of g_b^* is not
over bootstrap samples b. Instead, g_b^* varies considerably.
varies substantially over bootstrap samples. Might these
incorporated into the bootstrap samples result in an inflate
We investigated this issue very intensively for this exa
that the bootstrap estimate of $\mathrm{se}(\hat{D})$ was acceptable here.
theoretically computed unconditional $\mathrm{se}(\hat{D})$ did not accou
in $\hat{D}$ (even though it accounts for all model selection unc

The resolution of the matter also turned out to be simp
to consider the need for a variance inflation factor $\hat{c}$. Th
above was based on theoretical formulae under ML estima
However, this variance is underestimated if important as
assumption of independence of detections within a line seg
may be spatio-temporal variation in true detection proba
distance x; there may be errors in recording detection dis
are). All these problems lead to more variance than theo
can adjust the theoretical $\widehat{\mathrm{se}}(\hat{D})$ to allow for these sourc
way analogous to what the bootstrap does). The simplest
$\sqrt{\hat{c}} \cdot \widehat{\mathrm{se}}(\hat{D})$ as our theory-based unconditional standard er

When all models considered are subsets of one globa
QAIC, and variance inflation, comes from the goodness
model: $\hat{c} = \chi^2/\mathrm{df}$. However, here we have four models, b

of freedom, and $\hat{c}$ for these four models, as well as the A
on use of AIC:

Model	χ^2	df	$\hat{c}$	w_i
1	25.11	17	1.48	0.499
2	23.73	16	1.48	0.220
3	24.66	15	1.64	0.072
4	23.40	16	1.46	0.209

The weighted average of $\hat{c}$, weighted by w_i, is 1.49. Here
$\hat{c}$ from the selected model or this weighted average. It
here; hopefully, this would be the usual situation in distan
we use here $\hat{c} = 1.48$, df $= 17$ from model g_1.

We should, however, have been using QAIC rather th
Akaike weights might then change (along then with othe
ble 5.25 we obtain $-2\log(\mathcal{L})$ for each model and thus
$(-2\log(\mathcal{L})/\hat{c}) + 2K$ and the associated weights w_i:

Model	QAIC	Δ_i	w_i
1	1,384.71	0.00	0.511
2	1,386.47	1.76	0.211
3	1,388.61	3.90	0.073
4	1,386.54	1.83	0.205

The differences between the Akaike weights based on A
here trivial (this is because of a large sample size her
weights with each $\hat{D}$ from Table 5.25 gives a model-aver
the original result was 10.19. For an unconditional standa
QAIC-derived w_i we get 1.23 (the same as with AIC-bas
stay with the originally computed $\hat{\bar{D}} = 10.19$. In this exa
of using QAIC is to make us realize that we need to use
factor with our theoretical standard errors.

The quick way to adjust the theoretical unconditional
compute $\sqrt{\hat{c}} \cdot \text{se}(\hat{D}) = \sqrt{1.48} \cdot 1.23 = 1.22 \cdot 1.23 = 1.50$
result for the unconditional standard error of $\hat{D}$ was 1.4
of $\sqrt{\hat{c}} \cdot \text{se}(\hat{D})$ is not the correct formula (we have used it
epistemological value). Rather, one should adjust each th
to be $\hat{c} \cdot \widehat{\text{var}}(\hat{D}_i \mid g_i)$ and then apply (4.9), which here bec

$$\widehat{\text{se}}(\hat{D}) = \sum_{i=1}^{4} w_i \sqrt{\hat{c} \cdot \widehat{\text{var}}(\hat{D}_i \mid g_i) + (\hat{D}_i -$$

The two approaches will give almost identical results
$(\hat{D}_i - \hat{\bar{D}})^2$ are small relative to $\hat{c} \cdot \widehat{\text{var}}(\hat{D}_i \mid g_i)$, as they are

by the bootstrap (this may be a coincidence).

The bootstrap method to obtain the unconditional sta
rameter estimator will, if done correctly, automatically in
uncertainty in that standard error. Estimation based on th
formulae, for models that do not automatically estimat
variation, will not automatically include overdispersion va
what theory assumes. Thus in these cases we must alway
to include an empirical variance inflation factor $\hat{c}$ in our c

5.5 Summary

Model-based data analysis is very important, as illustrated
book, and as demonstrated by the much improved estim
precision, less bias) for the chain binomial survival data
5.2. For example, rather than try to separately estimate su
age one should produce smoothed estimates of these p
suitable parametric models. For such observational data
other examples here—GPA example, gas mileage data,
would rarely, if ever, know a priori the single best model to

However, in all such cases the investigator can and shou
a small set of suitable candidate models for data analys
creating a meaningful, reduced number of predictor varia
$QAIC_c$-based model selection can be very effective at pr
the models based on Akaike weights. If it makes sense to
(if the models mean something as alternative scientific o
nations), one can use the expected K-L best model to draw
in mind that the selection of that model as best is itself
pling standard errors of estimated parameters can and sh
selection uncertainty.

If the models are only a means to the end of "smoothin
case for prediction, then we recommend computing mod
eter (prediction) estimators and their unconditional samp
based on the Akaike weights. Monte Carlo methods showe
worked well for the chain binomial models; unconditior
val coverage is generally close to the nominal 95%, while
conditioned on the selected best model may achieve only
Monte Carlo studies in this chapter also show that there
uncertainty but that the Akaike weights are effective at m
tainty. The sampling distribution of Δ_p was examined for
we found that generally, for a small number of candidate m
corresponds to at least the 95th percentile and more oft

For reliable results from simulation we recommend at
Carlo samples at each set of conditions used to generate
for the bootstrap also: For the results to be stable to two s
must often use at least 10,000 bootstrap samples. Too n
these simulation methods do not use enough replications.

We do not recommend the dimension-consistent criteri
model selection in the biological sciences or medicine wh
set of well thought out candidate models. Such criteria are
information, are based on poor assumptions, and perfor
sample size is quite large. We do not recommend using an
testing for model selection.

The choice of models to examine is important. The
amples demonstrated that a class of logistic models pro
than a model class that assumed constant survival rate af
GPA example demonstrates that in variable-selection p
considerations can lead to much better models than unth
selection.

In Section 5.3 we note that model selection bias occurs
Regression coefficient estimators $\hat{\beta}_i$ are biased away from
able x_i is included in the model only when that variable se
(i.e., when $\hat{\beta}_i$ is sufficiently different from 0). The less i
the more biasing effect model selection has on $\hat{\beta}_i$. Estimat
$\hat{\sigma}^2_{y|x}$ is negatively biased by model selection. The use of
other methods, provides some protection against both mo
The best way to minimize model selection bias is to re
models fit to the data by thoughtful a priori model for

Usually, selection of a best model is needed if scienti
the goal. However, often it is better to think in terms o
ence using the full set of models, rather than selecting
basing inferences on that single model. This is especiall
variable selection as practiced in regression, because the
is highly variable. Model averaging is then particularly us
the relative importance of a variable as the sum of the Ak
models in which that variable appears and examination of
regression parameters.

Erroneous results have stemmed from the frequent mi
simulation in judging various model selection approaches
generating model has had a few parameters (very often < 8
no or few tapering effects, and the objective has been to
method most often chooses the generating model. This co
the spirit of BIC and is a sterile exercise as regards real
hence there is no reason why results would apply to real

amples we have examined, such intervals have excellent
There are surely cases where this simple approach does n
we have not found any during our investigations.

Advanced Issues and Deeper I

6.1 Introduction

Much of this chapter is new material not in the first edition
moved from other chapters because we judged it to be n
helpful on a first reading of introductory ideas. In either
material here did not fit well in a logical, linear progres
ideas about K-L-based model selection and multimodel in
firmatory orientation with a relatively small set of models
ordering to the sections of this chapter; they can be read i

We consider $R < 100$, or perhaps even $R < 200$, as
because many classical variable selection analyses, or a
consider thousands, tens of thousands, or even millions or
$R \gg n$). We consider the analysis as exploratory rather tha
the number of models exceeds the sample size, which us
real thought has been expended on the issue of meaningful
The detailed properties of model selection, and subseque
the number of models considered is huge are not well stu
gargantuan amount of computing required. To illustrate iss
Section 6.2 looks in some detail at a published all-subsets
regression example with sample size 252 (n) and 13 pr
include the no-effects null model, hence $R = 8{,}191$ mod

Another subject in this chapter is an overview of selec
proaches, followed by a more detailed contrasting of BIC
all current model selection criteria fall into two classes,

hope to cast some light on this matter. Another issue is
random coefficient (effects) models. This is a rapidly deve
importance and promise. A partially related issue is dete
size for a data set. The issue of "the" sample size is ofte
there is not a single sample size for complex data structures
correspond to noninteger effective sample size.

Also delved into here is goodness-of-fit for count data
els; essentially, this is about estimating c for overdispersi
general handling of overdispersion is considered when
overdispersion parameter can be estimated and used. T
at formulas for Bayesian model averaging, for the intere
portance of a small-sample version of AIC is discussed an
(as we have defined it). Another subject considered is co
when the assumed probability distributions are different;
there is a single "error" distribution (e.g., normal or mu
model structural aspects vary.

6.2 An Example with 13 Predictor Variab 8,191 Models

6.2.1 Body Fat Data

In this example multiple regression is used to predict per
on predictors that are easily measured. The data are fro
males, ages 21 to 81. A key reference is Johnson (1996
journal (http://www.amstat.org/publications/jse/toc.html)
able on the web in conjunction with Johnson (1996). The
data were generously supplied by Dr. A. Garth Fisher, E
Research Center, Brigham Young University, Provo, Uta
permission to freely distribute the data and use them for
poses." Reference to the data is also made in Penrose et al
have also been used in Hoeting et al. (1999), a seminal
model averaging.

We take the response variable as $y = 1/D$; D is mea
(observed sample minimum and maximum are 0.9950 and
weight, lower body density means more body fat becaus
as muscle and bone. The reciprocal of body density is
related to percent body fat; however, there is no agreem
experts on the parameters of that calibration, which is
$1/D$ as our response variable. Measuring body density re
time-consuming underwater weighing method. For each

Sampling aspects for the study were not stated, we suspect
were a self-selected sample (i.e., volunteers) from the Pi
that this sample was obtained and processed in a short time
1980s.

We consider aspects of five possible approaches, orde
desirable, in our opinion:

1) fit the full (i.e., global) model only;
2) select one model by standard stepwise selection from
possible simple regression models, then ignore selecti
3) select the best model using AIC_c and consider selectio
4) do full multimodel inference, such as model-averaged
8,191 models,
5) first reduce in number and refine the predictors base
logic, to a set of meaningful derived variables, then
fewer models).

We then explore using the nonparametric and parametric
compare aspects of selection under AIC and BIC. Our m
example is to demonstrate how much model selection unce
the model set is huge, and how this uncertainty is reduce
set of models is used.

6.2.2 *The Global Model*

Table 6.1 shows basic results of fitting the global regre
$\beta_0 + \sum \beta_i x_i + \epsilon, \epsilon \sim \text{normal}(0, \sigma^2)$. Hoeting et al. (1999)
model checking showed this to be an acceptable model
also that correlations among the predictors are strong, but
entirely positive, and range from -0.245 (age & height)
hips). The design matrix is of full rank.

The absolute value of the usual t-test statistic (Wald v
sion coefficient is 1/|cv|. Hence, in Table 6.1 any parame
would be considered "significant" at the $P = 0.05$ level.
unambiguously suggests dropping knee (x_g), chest (x_5), a
the global model. The issue of other predictors that one mi
by the strong correlations among the predictors. Howeve
strong suggestion that not all 13 predictors need to be in
is motivated to apply formal model selection.

6.2.3 *Classical Stepwise Selection*

We used SAS PROC REG, at its defaults, for stepwise v
only a few steps a model was selected, thus giving, perha

$R^2 = 0.7428$.

Variable		$\hat{\beta}_i$	$\widehat{\text{se}}(\hat{\beta}_i \mid g)$	|c
INTERCEPT		0.873844	0.04594	0
X1	age	0.000109	0.00007	0
X2	weight	−0.000215	0.00013	0
X3	height	−0.000163	0.00037	2
X4	neck	−0.000971	0.00049	0
X5	chest	−0.000106	0.00021	2
X6	abdomen	0.002036	0.00019	0
X7	hips	−0.000432	0.00030	0
X8	thigh	0.000525	0.00030	0
X9	knee	0.000024	0.00051	21
X10	ankle	0.000571	0.00046	0
X11	biceps	0.000492	0.00036	0
X12	forearm	0.000923	0.00041	0
X13	wrist	−0.003649	0.00110	0

TABLE 6.2. Selected predictors, hence selected best model, for s
methods applied to the body fat data.

Selection	Indices of predictor variables sel								
Stepwise		2		4	6				
Forward	1	2		4	6	7	8		10
Backward			3			7		9	10
Mallows C_p	1	2		4	6		8		
AIC	1	2		4	6	7	8		
AIC_c		2		4	6				
BIC		2			6				

pression of confidence in the selected model. Common be
rather than forward or backward, selection is the best of t
lection methods. We give all three results in Table 6.2, plus
C_p, AIC, AIC_c, and BIC.

There is substantial variation in the best model by me
stepwise versus forward and backward methods. Even th
AIC and AIC_c differ by four predictors. However, this r
because we know that we should use AIC_c in deference t
$n/K = 252/13 \ll 40$. This example shows that using AI
makes a difference even with $n = 252$ and global K
parameters plus σ^2). Mallows's C_p does not select the s
Some literature erroneously claims that these are identica
ear models; in general they give similar, but not identica
expected, BIC is more conservative than AIC.

with 8,191 models, there will be considerable model sel
is unconscionable not to evaluate this uncertainty and use
unconditional inference about importance of predictors.
justified in saying that the selected predictors are impo
selected are not important (this is the same false dichoto
hypothesis testing). Even if we can agree on a selection
model selection uncertainty, and we consider this next.

6.2.4 *Model Selection Uncertainty for AIC_c and*

We assume that most readers are now aware of the Bayesia
rion (BIC, Schwarz 1978, Hoeting et al. 1999): BIC $= -$
$\log(n) \cdot K$ (whereas AIC $= -2 \log \mathcal{L}(\hat{\theta}|data, g) + 2 \cdot K$. C
model j, $\Delta\text{BIC}_j = \text{BIC}_j - \text{BIC}_{min}$ (BIC_{min} is the mini
R models). Also, the same structural formula that gives
from ΔAIC, is used with ΔBIC to give the (posterior) pro
$g_1, \ldots, g_R$. More information about BIC appears in Secti

For this data analysis we quantify model selection u
ΔAIC_c and ΔBIC, but mostly by Akaike weights and po
bilities (for BIC). Table 6.3 shows Akaike weights, w_j, fo
and also shows weights and models at a few other rank
all 8,191 models ordered by largest to smallest w_j. Th
($r^2 = 0.733$) has a weight of only 0.010738; the other mo
weights (the 8,191 weights sum to 1). The next best model
The weight of evidence is essentially identical for both the
ratio is 1.0025). Table 6.3 shows a few other evidence rat
model ranked 52nd by AIC_c was ranked 1st by BIC. Pl
include those ranked first to 176th, or even to 642nd.

Another way to determine a confidence set on models is
where the sum of these *ordered* Akaike weights is some va
(this method is not the best one in general, but it is use
All we care about here is how large such confidence s
models have some plausibility (as opposed to models w
near certainty)? The number of models in such sets unde
shown below:

Σw_j	# of models	ER
0.900	649	$w_1/w_{649} =$ 34
0.950	876	$w_1/w_{876} =$ 68
0.990	1449	$w_1/w_{1449} =$ 413
0.999	2266	$w_1/w_{2266} =$ 3579

are shown.

Model order j	w_j	Δ_j	
1	0.010738	0.00000	
2	0.010711	0.00496	
3	0.010333	0.07690	
4	0.009752	0.19273	
5	0.009459	0.25371	
6	0.009104	0.33008	ER: $w_1/w_6 = 1$
.	...	...	
36	0.003924	2.01	
52	0.003349	2.33	ER: $w_1/w_{52} =$
176	0.001450	4.00	
642	0.000321	7.02	ER: $w_1/w_{642} =$
1103	0.000071	10.00	ER: $w_1/w_{1103} =$
8191	1.3E-72	322.01	

There is no computational impediment here to using all model-averaged results, but if we were to use Occam's and Raftery 1994), we would want $\sum w_j$ at least 0.95, a Thus, model-averaged inferences require here using on of the possible models. This is not at all like using j mated as) best model and erroneously thinking that th basis for reliable inferences in a repeated sampling (or B (the Akaike weights approximate repeated sampling-bas frequencies).

One type of model-averaged inference is the variable weight $w_+(i)$, which is the sum of the Akaike weights fo models in which predictor i occurs. Table 6.4 shows these weights and the variables included in the six top-ranked (Table 6.3 shows w_j for these top six models).

Results for BIC, analogous to those for AIC_c in Table 6 6.5. The model ranked 12th by BIC is the model ranke defining operational (frequentist) property of BIC is that to infinity, the posterior probability of a single model goo both the sampling context and model set to be fixed, in this example the BIC best model has associated posterio 0.14; hence there is again substantial model selection u window for $\Sigma \Pr(\text{model } j) = 0.999$ includes 1,611 mode whether approached in a K-L or Bayesian context there is uncertainty here, and inferences, after selection, should re

It is beyond our intended use of this example to use model-averaged prediction under AIC (see Hoeting et al

			AIC$_c$ top mode			
$w_+(i)$	variable	i	1	2	3	4
0.495	age	1	0	1	1	1
0.933	weight	2	1	1	1	1
0.314	height	3	0	0	0	0
0.652	neck	4	1	1	1	1
0.283	chest	5	0	0	0	0
1.000	abdomen	6	1	1	1	1
0.445	hips	7	0	1	0	0
0.588	thigh	8	0	1	1	1
0.293	knee	9	0	0	0	0
0.448	ankle	10	0	0	0	0
0.600	biceps	11	1	0	1	0
0.828	forearm	12	1	1	1	1
0.976	wrist	13	1	1	1	1
		$K =$	8	10	10	9

TABLE 6.5. Some posterior model probabilities for BIC, based on a
most to least probable, and the cumulative probabilities.

Model order j	Pr(model j)	ΣPr
1	0.13930	0.13930
2	0.08980	0.22911
3	0.05681	0.28591
4	0.03829	0.32420
5	0.03488	0.35908
6	0.03118	0.39027
.	...	...
12	0.01484	0.50689
158	0.00060	0.90010
292	0.00023	0.95001
757	0.00003	0.99001
1611	0.00000	0.99900
8191	9.4E-69	1.00000

model-averaged prediction with these data). Moreover, we
simply accepting these 13 predictors and using either t
doing stepwise selection in the first place. Rather, we enc
approach (noted as #5 in Section 6.2.1): First reduce the
hence models, based on theory and/or logic to a set of
variables related to y, then do multimodel inference.

tially, no analysis at all was done. Rather, one of us (KPI
matter intermittently over several weeks and decided to
ables below. The actual data were not studied before this
done; i.e., none of the above model fitting was done until a
below were decided on. A knowledgeable health-trained
even better at generating derived variables and suitable n
not consider improved model forms beyond linear regres

Weight and height jointly ought to be very important fo
but not as separate predictors in linear regression; they s
together. Allometric relationships are common in biolog
body might have a nearly constant ratio of some function
In many animals mass tends to be proportional to the cube
species that proportionality is often very stable. Hence, v
$z_1 = \log(\text{weight})/\log(\text{height})$ would be very stable for th
body and thus variation in this derived variable would be
with variation in body fat (hence, 1/density).

Additional considerations led to five more derived varia
ply information about different dimensions of the predicti
set of six, ordered as considered most to least important, a
tionale for each, and the predicted sign (+ or −) of each r
are given below:

$z_1 = \frac{\log(\text{weight})}{\log(\text{height})}$	based on ideas of allometry (+)
$z_2 = \frac{\text{abdomen}}{\text{chest}}$	beer gut factor (+)
$z_3 = \frac{(\text{knee}*\text{wrist}*\text{ankle})^{\frac{1}{3}}}{\text{height}}$	heavyset or light-boned (−)
$z_4 = \left[\frac{\text{biceps}*\text{thigh}*\text{forearm}}{\text{knee}*\text{wrist}*\text{ankle}}\right]^{\frac{1}{3}}$	fleshiness index (+)
$z_5 = \text{age}$	standardized by mean and stand
$z_6 = \text{age}^2$	based on standardized age (+).

Part of the thinking here is that because the response v
percent body fat, only ratios of body measurements should
z_1 the most important predictor seemed as if it ought to be
but only relative to some other body size metric. Based o
shapes of men (and experience with his own measureme
z_2. Next there are issues of genetic variation in being sligh
z_3: For a given weight, more bony is less fat), and there i
(hence z_4). Finally, it seemed reasonable that age might
given z_1 to z_4. Percent body fat would tend to increase w
twenties, but asymptotically, so not exactly linear on age.
(derived) predictors, and therefore 63 possible models (w

(via feedback) from the data analysis whether our reasc
not. This feedback aspect of data analysis is critical in the

Table 6.6 gives some basic results about model selectio
AIC_c, for the fat data with these six predictors. Notewo
AIC_c best model ($r^2 = 0.659$) is also the best model und
and BIC. The number of models in the confidence set wi
to different values is shown below:

$\sum w_j$	# of models	ER
0.900	3	$w_1/w_3 =$ 2.8
0.950	4	$w_1/w_4 =$ 5.9
0.990	5	$w_1/w_5 =$ 63.2
0.999	7	$w_1/w_7 =$ 178.3

Thus, we need to consider only a few models here, not hun
Table 6.7 shows the top six models and the variable-impo

TABLE 6.6. Akaike weights w_j and Δ AIC_c values for some of th
most to least supported by the data, based on the z_i predictors.

Model order j	w_j	Δ_j
1	0.48867	0.000
2	0.23627	1.453
3	0.17745	2.026
4	0.08316	3.542
5	0.00773	8.294
6	0.00287	10.278
.	...	...
63	2.0E-58	264.258

TABLE 6.7. Variable relative-importance weights and predictors
models (1 if included, 0 otherwise) based on the z_i and K for that

			AIC_c top mode				
z_i	description	$w_+(i)$	1	2	3	4	5
z1	wt/ht	1.000	1	1	1	1	1
z2	gut	1.000	1	1	1	1	1
z3	bony	0.323	0	1	0	1	0
z4	fleshy	0.986	1	1	1	1	0
z5	age	1.000	1	1	1	1	1
z6	age*age	0.264	0	0	1	1	0
		$K =$	6	7	7	8	5

	$\hat{\beta}_i$	$\widehat{\text{se}}(\hat{\beta}_i \mid g)$	|cv|
z_1	0.18693	0.03714	0.199
z_2	0.14404	0.01717	0.119
z_3	0.04520	0.05828	1.290
z_4	0.00554	0.00168	0.303
z_5	0.00310	0.00070	0.227
z_6	0.00011	0.00053	4.785

We did not intuit the correct sign for predictor z_3, but it is n
The other signs we predicted correctly, but we can discou
z_6 as it is relatively unimportant here ($w_+(6) = 0.264$). Th
importance of the z_i is 2, 1, 5, 4, 3, 6, with z_3 and z_6 havir

Overall we think that the a priori considerations here
and influential than just the purely statistical model selec
is always possible to do exploratory analysis after any a
not vice-versa). In fact, we did some final exploratory an
of having only one or two simple predictors and a high m
deal of probing we ended up with a very competitive 1-pr
on $z_e =$ abdomen/height ($r^2 = 0.682$), which gave the re

parameter	estimate	$\widehat{\text{se}}(\hat{\beta} \mid g)$	|cv|
β_0	0.8259	0.00528	0.0064
β_e	0.0924	0.00398	0.0432

This type of *post hoc* hypothesis-generating analysis is a
it is reported for what it is: strictly exploratory, hypoth
confirmatory.

6.2.6 *Bootstrap Evaluation of Model Uncertaint*

We now return to the original 13 predictors problem and ou
this extended example: to illustrate the extent of model un
is a huge number of models ($R \gg n$). We explore the b
model selection probabilities (π) for these selection me
want to know how well the estimated selection probabiliti
weights (or for BIC, the posterior model probabilities; how
this is not a fair question).

For each of several model selection methods we create
strap samples, all of size 252, from the data. In addition
relative frequencies we want to know how many differe
possible, ever get selected (this number depends weakly
For AIC_c and BIC we also looked at the sampling distribu
ΔBIC.

These 1,200 models were listed in rank order, most to
lected. Table 6.8 shows a few records from this list. Mo
whether predictor x_1 to x_{13} (in that order) is in or out of the
0101010000111 denotes the model with predictors 2, 4, 6
models selected as best, by method, with the actual data a
6.8.

From Table 6.8 we see that even the most commonly s
stepwise selection ($\hat{\pi} = 0.0215$) has a very low select
model selected by the stepwise method from the actual d
This bootstrap assessment corroborates that there is con
lection uncertainty. Corresponding bootstrap-based asses
BIC are in Tables 6.9 and 6.10. Applying AIC_c selection
samples, 1,233 distinct models were selected with $\hat{\pi} = 0.0$
model for the actual data. This $\hat{\pi}$ compares well to the Ak
Applying BIC selection to 10,000 bootstrap samples, the
models selected with $\hat{\pi} = 0.0891$ for the BIC best mode
From the data the posterior model probability for the BIC
Whatever one thinks of 0.09 versus 0.14, the comparison
posterior probabilities, which are conditional on the data
be comparable to predata random variable frequencies.

For AIC_c model selection the bootstrap assessment o
matches well to the Akaike weights. However, when we loo
based estimate of the sampling distribution of the ΔAI
different from our earlier assessments of this distributi
was 9.1; maximum was 49.3. This maximum depends we

TABLE 6.8. Model ranks and selection frequencies for a few of th
selected by the stepwise method applied to 10,000 bootstrap sam
data; the models selected by different methods with the actual data

Rank	Model	Frequency	
1	1101011100011	215	AIC
2	0101010000111	184	AIC_c & step
3	0101010001111	141	
4	1101011101011	141	
.	...	.	
16	1100010100011	89	
17	0100010000011	87	BIC
18	1101011100111	87	
.	...	.	
21	0101011100111	70	
22	1101010100111	63	C_p
23	0100010100011	62	

selected by AIC, AIC_c, and BIC with the actual data are also indic

Rank	Model	Frequency	
1	1101011100011	218	AIC
2	1001011100011	151	
3	1011011100011	149	
4	1101011101011	148	
5	1101010100011	145	
6	0101010000111	130	AIC_c
7	1101010101111	125	
8	0101011100011	112	
.	...	.	
53	1011110000111	36	
54	0100010000011	34	BIC
55	1001011101011	34	

TABLE 6.10. Model ranks and selection frequencies for a few of
selected by BIC applied to 10,000 bootstrap samples from the bo
selected by AIC_c and AIC with the actual data are also indicated.

Rank	Model	Frequency	
1	0100010000011	891	BIC
2	0100010000101	689	
3	0100010001101	470	
4	0010010000001	388	
5	0100010000001	359	
.	...	.	
23	0100010100001	90	
24	0101010000111	87	AIC_c
25	1000010000001	87	
.	...	.	
142	1101011100011	10	AIC

tivated looking at the same sampling distribution informat
the guidelines for interpreting Δ are the same for K-L c
erage $\Delta\text{BIC} = 7.3$, maximum was 49.2. Bootstrap-based
percentiles:

Percentile	ΔAIC_c	ΔBIC
0.50	8.0	6.2
0.90	17.5	15.3
0.95	20.8	18.6
0.99	27.5	25.3

These sampling results are not consistent with what we
number of models is small, such as $R < 100$ (and certainl

bootstrap).

6.2.7 *Monte Carlo Simulations*

We assumed that the measurements y and $\underline{x} = (x_1, \ldots, x_1$ be suitably modeled as multivariate normal with a varianc taken to be the observed variance-covariance matrix (this the parametric bootstrap). This full variance-covariance as below; $\underline{c}$ is 13×1 and Σ_x is 13×13:

$$\Sigma = \begin{bmatrix} \sigma_y^2 & \underline{c}' \\ \underline{c} & \Sigma_x \end{bmatrix}.$$

The global model is now also the generating model unc of true regression parameters is given by $\underline{\beta}' = \underline{c}'(\Sigma_x)^{-1}$ limited purposes, to set $\beta_0 = 0$ and generate the data in observation $\underline{x}$ is generated from the marginal MVN($\underline{0}$, Σ_x and $y = E(y|\underline{x}) + \epsilon$, where ϵ is a normal random va and variance $\sigma^2_{y|x} = \sigma_y^2 - \underline{c}'(\Sigma_x)^{-1}\underline{c}$. More details, and p approach are given in Sections 4.3.6, 5.3.2, and 5.3.4.

The approximate theoretical standard error for each $\hat{\beta}$ (see Section 5.3.2). Hence, we computed the "effect siz $\lambda_i = \beta_i/\text{se}(\hat{\beta}_i|\text{global } g)$; this is essentially the mean for $\hat{\beta}$ a normal(λ_i, 1) random variable. These λ_i values are

i	λ_i	i	λ_i	i	λ
1	1.687	6	11.123	10	1
2	−1.727	7	−1.484	11	1
3	−0.453	8	1.783	12	2
4	−2.048	9	0.048	13	−3
5	−0.509				

There clearly are tapering effects, and the only trivial circumference. The actual average ordering of predicto importances may not match the ordering by $|\lambda_i|$ becau nature of the predictors.

We generated 10,000 independent samples of size 252 a BIC model selection. Our interest is in regard to, first, the fr of models selected: Do those relative frequencies match and results from the (nonparametric) bootstrap? Second, based sampling distributions of ΔAIC_c and ΔBIC ma bootstrap. The answers are yes; there was no substantial the bootstrap and Monte Carlo approaches. Some summary frequencies are given in Tables 6.11 (AIC_c) and 6.12 (for

of the body fat data; the models selected by AIC_c and BIC with
indicated.

Rank	Model	Frequency	
1	1101010100011	197	
2	1101011100011	177	
3	1101011101011	165	
4	1101010101011	152	
5	0101010000111	149	AIC_c
.	...	.	
9	1101010100111	132	
10	1101010101111	119	
.	...	.	
21	1100011101011	72	
22	0100010000011	70	BIC
23	0100010000111	70	

TABLE 6.12. Model ranks and selection frequencies for a few of
selected by BIC applied to 10,000 Monte Carlo samples that mimic
of the body fat data; the models selected by AIC_c and BIC with
indicated.

Rank	Model	Frequency	
1	0100010000011	1063	BIC
2	0100010000101	852	
3	0100010000001	371	
4	0100010001011	349	
5	010100000100	331	
.	...	.	
9	0101010000010	221	
10	0100010100011	212	
.	...	.	
19	0100010001001	112	
20	0101010000111	111	AIC_c
21	1001011100011	109	

The sampling distribution percentiles for the 10,000
from the Monte Carlo samples are below:

Percentile	ΔAIC_c	ΔBIC
0.50	5.6	5.5
0.90	12.8	14.9
0.95	15.3	18.6
0.99	21.2	26.6

are generally a little smaller than those from the bootst
inference we only have a sample of size 1). However, both
results show larger percentiles of Δ_p than what we have se
of models is very much smaller than 8,191. We believe tha
here generally apply when R is so large.

The bootstrap simulation relative frequencies of model
to the Akaike weights. However, we noticed that the sam
Δ_p was stretched to the right. We wondered whether this re
the bootstrap in this case. Therefore, we then did the Mon
to verify the bootstrap; both approaches gave about the sa
had to reconsider the distribution of Δ_p because the guide
interpreting Δ as regards inferential evidence strength ab
questioned, at least to the extent those guidelines were p
sampling distribution ideas. Our inferential guidelines are
as those for ΔBIC (Raftery 1996a), and therefore the sam
ΔBIC is here also out of line with those guidelines.

The resolution of this concern is that we need to realize
basis for inference about model selection uncertainty un
theoretic approach is the model likelihood $\mathcal{L}(g_i|\text{data})$ an
it (evidence ratios and Akaike weights). This is analogou
based on posterior model probabilities for BIC in a Bay
der both of these approaches inference is conditional on
likelihood, rather than being justified by ideas of samplin
we are justified in retaining our guidelines as being usefu
be interpreted strictly in a sampling distributional framew

6.2.8 Summary Messages

The first general point illustrated by this example is tha
selection uncertainty should be expected when the numbe
large, such as under many instances of all subsets (i.e.,
This example has a good sample size ($n = 252$) relative to
of predictor variables (13) for such applications, and still
models. With so many models we find here that the selecte
a very small Akaike weight (0.010738), and is essentially
and third-place models (Table 6.3). Moreover, a confidenc
easily includes over 100 models. Any all-subsets applicatio
with R far exceeding n can be expected to have such extr
uncertainty wherein even the best model has a very small

When all the models have very low weights, such as
ferential credibility for any single model regarding what
predictor variables. It is foolish to think that the variables
model are "the" important ones and the excluded variable

of variable importance are needed (Brieman 2001). The
variable i, $w_+(i)$ (Table 6.4), provide a model-averaged m
importance of each predictor variable. They are relative, n
the baseline value that corresponds to no predictive value
not at $w_+(i) = 0$, but at some value > 0 (randomization r
to estimate this baseline value, see Section 6.9.8).

Even if prediction is the goal, it is foolish to think th
model has any special credibility when its Akaike wei
Rather, model-averaged prediction should be used (this is
literature, see e.g., Brieman 1996, 2001, regarding "bag
forests"). For a vector of predictors each fitted model yie
and the model-averaged prediction is $\hat{\bar{y}} = \sum w_j \hat{y}_j$. Fo
implies the best measure of the absolute importance for

the model-averaged partial regression coefficient for that
5.3.6).

If there is so little inferential weight for the best mo
selection been considered to be so useful? Because the b
terion) model gives good in-sample prediction, relative to
measured by the coefficient of determination, r^2. Howeve
ally the same, r^2 is achieved by many competitor models.
here has $r^2 = 0.733$. However, in the confidence set of 87
by the sum of the ordered (large to small) w_j being $= 0.9$
maximum r^2 are 0.718 and 0.742. Any of these 876 mo
tially the same average in-sample predictability, but each
subset of predictor variables. This phenomenon of many n
els as judged by r^2 values is acute when there are many
are strongly intercorrelated, which is the usual case in va

A second point we make here is that stepwise model
be used. Almost any thoughtful model selection will fin
has an r^2 above 0.7. Even stepwise selection, ad hoc thou
lead to a model with decent r^2, relative to what is possib
same model as AIC. So why not use stepwise selection?
no theoretical basis for stepwise selection, as regards any
(2) there is no simple way to compute model (inference) w
of stepwise selection and, as practiced, no such model in
provided by stepwise selection (unless one resorts to t
never seems to be done); this is a major failing. (3) step
the subjective appearance of much less model uncertainty
only a small number of all possible models are fit, and n
for stepwise selection lists only a few (perhaps < 10) m
that were fit to the data. As a result, the user is mislea
model selection uncertainty exists. (4) as practiced, stepw

that the selected model was the one and only a priori mod
results of stepwise selection depend nontrivially on the
to enter and drop predictors; there is no theory for thes
McQarrie and Tsai 1998, 427–429 about stepwise selecti

A third point follows from the results of the bootstra
evaluation of model selection uncertainty and the samplin
ΔAIC values. This is one of the few cases where we have
for the number of models, R, over several hundred, as
is a several dozen or fewer. The guidelines we have giv
large Δ did not hold up with $R = 8{,}191$; the same guidel
for BIC differences and they also did not hold up here. I
being big here, hence discounting the model with the bigg
like $\Delta = 20$ is "big." However, the model selection relati
these simulations were very consistent with Akaike wei
data analysis of all 8,191 models. Thus, using these weig
evidence ratios) as the basis for inference about model sele
by the simulations.

We are quite convinced now, from all of our research
matter, that the w_j are valid and useful inferential statistic
Conversely, rigorous inference should not be based on the s
of the ΔAIC, even though rough guidelines on this matter
R is small. It is a principle that sampling variation acros
is not the same as inferential uncertainty (as reflected in t
models, or parameters in models, given the single data s
types of "variation" are often similar, but when they diff
be based on the likelihood.

A fourth point is the advantage of reducing a priori th
to consider, especially by reducing the number of predi
can best be done by thoughtful creation of meaningful de
dropping meaningless predictors (or ones whose usefulnes
This idea was illustrated in Section 6.2.5. Instead of 8,19
up with only 63 models; of these 63 only 4 had substantia
and the best model had $w = 0.489$, with $r^2 = 0.682$.
greater interpretability and logic underlying these model
out-of-sample predictions better than those from the br
approach (alas, we have no other data with which to test

A fifth point is that after the a priori analyses one is fr
more judgment-based analyses and model selection, as l
about the inferences one makes: a priori versus rangi
ploratory to reckless data dredging. For example, it is clear
Table 6.7) that variables z_3 and z_6 are useless predictors (
on the "objective" methodology and our understanding o
So one might chose to drop them and just use the AIC_c

that health professionals can use to quickly predict percen
you can do the sort of uninhibited exploration we did her
analyses, of really simple models to end up with a linear r
the one derived predictor, z_e = abdomen/height (r^2 = 0
models were considered in arriving at this one, so it sho
without testing it by application to a new set of data. I
consider this model as being the result of reckless data d
restricted ourselves to single predictor models.

6.3 Overview of Model Selection Criteria

There is a variety of model selection methods. However, fr
of statistical performance of a method, and intended con
are only two distinct classes of methods: These have be
and *consistent*. We will characterize these two classes in
introducing other model selection criteria.

Under the frequentist paradigm for model selection one
main approaches: (I) optimization of some selection crite
potheses, and (III) ad hoc methods. One has a further cla
(1) criteria based on some form of mean squared error (
Mallows 1973) or mean squared prediction error (e.g., P
(2) criteria that are estimates of K-L information or dist
the special cases AIC, AIC_c, and QAIC_c), and (3) criteri
estimators of K, the dimension of the "true model" (e.g., B
(2) and (3) in the following material.

6.3.1 Criteria That Are Estimates of K-L Inform

AIC, AIC_c, and QAIC_c are estimates of the relative K-L di
$f(x)$ and the approximating model $g(x)$. These criteria w
concept that truth is very complex and that no "true mode
that it was immaterial to the argument). Thus, one could on
with a model, say $g(x)$. Given a good set of candidate mo
could estimate which approximating model was best (amo
considered, given the data and their sample size). Linhart
speak of "approximating families" of models. Hurvich and
that these criteria select the best finite-dimensional appr
large samples when truth is infinite-dimensional. The ba
seems reasonable in the biological sciences.

When sample sizes are quite large, there are other criter
offer advantages in model selection and inference (e.g.,
These criteria specifically allow for "misspecification" o

literature (but see Konishi and Kitagawa, 1996).

Takeuchi (1976) provides a general derivation from
AIC. An intermediate result indicated that a selection
when the candidate models were not particularly close a
He derived TIC (Takeuchi's information criterion) for mo
a more general bias adjustment term to allow $-2\log_e(\mathcal{L})$
an asymptotically unbiased estimate of relative K-L,

$$\mathrm{TIC} = -2\log(\mathcal{L}) + 2\cdot\mathrm{tr}\left(J(\theta)I(\theta)^{-1}\right)$$

The matrices $J(\theta)$ and $I(\theta)$ involve first and second mixe
of the log-likelihood function, and "tr" denotes the matrix
is an approximation to TIC, where $\mathrm{tr}\left(J(\theta)I(\theta)^{-1}\right) \approx K$. T
excellent when the approximating model is quite "good" a
when the approximating model is poor. One might consid
and worry less about the adequacy of the models in the se
consideration involves two issues that are problematic. Fi
worry about the quality of the set of approximating mode
this is not something to shortcut. Second, using the expan
term in TIC involves estimation of the elements of the ma
(details provided in Chapter 7). Shibata (1989) notes tha
these two matrices can cause instability of the results of n
that the matrices are of dimension $K \times K$). If overdispers
data, then the log-likelihood could be divided by an estimat
factor, given QTIC. In most practical situations, AIC and
approximations to relative K-L information.

Linhart and Zucchini (1986) proposed a further genera
(1993) proposed a network information criterion (NIC)
training samples in neural network models. Konishi and K
gest even more general criteria for model selection and pro
into AIC and TIC and their derivation. Shibata (1989) deve
criterion, based on the theory of penalized likelihoods. H
called RIC for "regularized information criterion." We w
methods, since they would take us too far afield from our
they do not have the direct link with information theory a
relative K-L distance. However, we note that almost no
to evaluate the utility of these extensions in applied probl
of these criteria must be reserved for problems where the
large and good estimates of the elements of the matrices (
bias adjustment term are available.

Mallows's C_p (Mallows 1973, 1995) statistic is well
selection, but limited to LS regression problems with norm
C_p lacks any direct link to K-L information. Atilgan (19
tionship between AIC and Mallows's C_p, shows that un

linear regression assumptions hold, the two criteria seer
same model and rank the contending models in the sam
not equivalent. We have not found a small-sample versi
be useful when the sample size is small compared to the
variables (like AIC_c) (see Fujikoshi and Satoh 1997). R
(1994) provide a robust version of C_p (also see Sommer
Of course, adjusted R^2 has been used in classical multi
analysis; however, it has very poor performance (see e.g.,
1998).

6.3.2 *Criteria That Are Consistent for K*

This section deals with a class of criteria used in mod
"consistent" or "dimension-consistent" and with how thes
those that are estimates of Kullback–Leibler information.
been developed, based on the assumptions that an exactly
that it is one of the candidate models being considered,
selection goal is to select the *true* model. Implicit is the a
is of fairly low dimension (i.e., $K = 1$–5 or so) and th
generating (true) model, is fixed as sample size increases.
derived to provide a consistent estimator of the order or d
"true model," and the probability of selecting this "true mo
sample size increases. Bozdogan (1987) provides a nice re
"dimension-consistent" criteria. The best known of the "d
criteria" was derived by Schwarz (1978) in a Bayesian c
BIC for Bayesian information criterion (or occasionally
information criterion); it is simply

$$\text{BIC} = -2\log(\mathcal{L}) + K \cdot \log(n).$$

BIC arises from a Bayesian viewpoint with equal prior
model and very vague priors on the parameters, given the
purpose of the BIC-selected model was often simple pred
scientific understanding of the process or system under
estimator of relative K-L.

Rissanen (1989) proposed a criterion that he called m
length (MDL), based on coding theory, another branch o
(see also Yu 1996, Bryant and Cordero-Braña 2000). Whi
its justification are difficult to follow without a strong ba
theory, his result is equivalent to BIC. Hannan and Qui
criterion (HQ) for model selection whereby the penalty te

$$c \cdot \log(\log(n)),$$

practice. Bozdogan (1987) proposed a criterion he called
"consistent" and F denoting the use of the Fisher informa

$$\text{CAICF} = -2\log(\mathcal{L}) + K\{\log(n) + 2\} + \log$$

where $\log|I(\hat{\theta})|$ is the natural logarithm of the determin
Fisher information matrix. He has recently advanced a sc
terion based on a notion of complexity (ICOMP, Bozdo
CAICF nor ICOMP is invariant to 1-to-1 transformation
and this feature would seem to limit their application. AI
TIC are invariant to 1-to-1 transformations.

We question (deny, actually) the concept of a simple
biological sciences (see the Preface) and would surely t
even if a "true model" existed, it might be included in
models! If an investigator knew that a true model existe
the set of candidate models, would she not know which
little philosphical justification for these criteria in the b
medical sciences, although they have seen frequent ap
few people seem to be aware of the differences in the ba
for these dimension-consistent criteria relative to criteria
K-L information. The dimension-consistent criteria are d
objective than those addressed by criteria that are estimat

People have often (mis) used Monte Carlo methods
criteria, and this has been the source of confusion in so
Rosenblum 1994). In Monte Carlo studies, one *knows* th
and often considers it to be "truth." The generating model
and it is included in the set of candidate models. In the a
lated data, attention is (mistakenly) focused on what crite
this true model (e.g., Bozdogan 1987, Fujikoshi and Sato
Chen 1997). Under this objective, we would suggest the u
consistent criteria in this artificial situation, especially if
model was quite low (e.g., $K = 3$–5), or the residual vari
small, or the sample size was quite large. However, this
is far from that confronted in the analysis of empirical d
sciences. Monte Carlo studies to evaluate model selectio
analysis of real data must employ generating models wit
effect sizes and substantial complexity. Such evaluations
selection of a best approximating model and ranking of th
the notion that the true (in this case, the generating) mode
be discarded.

Research into the dimension-consistent criteria has oft
model with only a few large effects. More realistic model
of tapering effects have been avoided. In addition, the bas

addition of new geographic field sites or laboratories, the in
years, and the inclusion of new animals with genetic variat
Thus, as substantial increases in sample size are achieved, t
in the model also increases. The data-generating model d
as $n \rightarrow \infty$. We have found that the dimension-consist
poorly in open population capture–recapture models eve
K is small, but the parameters reflect a range of effect si
1998).

Notwithstanding our objections above, the sample sizes
the benefits of dimension-consistent estimation of model
very, very large by any usual standard. In the examples w
have substantial residual variances) we have seen the ne
in the thousands or much more before the consistent cr
to the "true model" with a high probability. In cases wh
was very large, say 100,000, one might merely examine t
decide on the parametrization, with little regard for the pri
(given the assumption that the true model is being sought,
candidates). It should be emphasized that these dimensio
are not linked directly to K-L information and are "informa
in the weakest sense. Instead, their motivation veered to c
of the order (K) of the supposed "true model" by employin
terms (but see Section 2.12.2).

When sample size is less than very large for realistic
data, these dimension-consistent criteria tend to select und
the attendant large bias, overestimated precision, and ass
inference. Umbach and Wilcox (1996:1341) present the re
simulations conducted under the BIC-type assumptions. F
100,000, AIC performed better than BIC in terms of the se
with the "correct" set. The two criteria were tied at sa
However, even at that large sample size, BIC selected the
79% of the cases; this is still far from selecting the correct
1. While these criteria might be useful in some of the pl
engineering, we suspect that they have relatively little uti
and social sciences or medicine. Findley (1985) notes th
can be an undesirable property in the context of selecting

6.3.3 Contrasts

As Reschenhofer (1996) notes, regarding criteria that are
K-L information vs. criteria that are dimension consistent
employed in the same situations, which is in contrast to th
been designed to answer different questions" (also see Pöt

AIC_c, QAIC, $QAIC_c$, and TIC) are reasonable for the analy
The dimension-consistent criteria (e.g., BIC, MDL, HQ, C
might find use in some physical sciences where a simple tr
and where sample size is quite large (perhaps thousands
or more). Still, we question whether this true model wo
candidate models. Even in cases where a simple true m
contained in the set of candidates, AIC might frequently h
properties than the dimension-consistent criteria.

Still other, somewhat similar criteria have been derive
1994a, b, and Stoica et al. 1986 for recent reviews). A la
methods have appeared, including the lasso (Tibshirani 1
strap (Breiman 1992), the nonnegative garrote (Breiman 1
quasi-deviance (Qian et al. 1996), various Bayesian me
and Chen 1997) including the use of Gibbs sampling (Ge
1993). Some of these approaches seem somewhat ad hoc,
ficult to understand, interpret, or compute. Often the met
for example, several are applicable only to regression-ty
not pursue these methods here, since they take us too far

In summary, we recommend the class of information-t
are estimates of relative K-L information such as AIC, AI
the selection of a parsimonious approximating model for
for sample sizes that occur in practice. If count data are
persed, then QAIC and $QAIC_c$ are useful. If large sample
TIC might offer an improvement over AIC or AIC_c. How
vestigations suggest that the simpler criteria perform as
we examined (Chapter 7).

6.3.4 *Consistent Selection in Practice: Quasi-tr*

The original motivation for a consistent model selection
the idea that the true (i.e., data-generating) model is in th
is, or may be, nested within some overly general models
size goes to infinity we want to select that true model wit
scenario is also based on the condition that one can increa
large while keeping the data generating context fixed: N
may enter as sample size n increases. This sort of sampl
can be done in Monte Carlo computer simulation, but not i
More formally, the frequentist motivation for BIC is idea

Assume that we have a nested sequence of models, g_1 t
model, g_t, is neither the first nor last model. The addition
nally in models $g_i, t < i \leq R$ are actually not needed. The
regression based on predictors x_1 to x_R where $y = \beta_0 + \beta$
and x_{t+1} to x_R have zero correlation with y and with all

parametrized. Therefore, the unique, lowest dimension tr
is g_t. This is the model that consistent criteria must select
n gets large. BIC will do this, and the inferred posterior p
BIC will also go to 1 as n gets large (this holds for any
but it suffices to restrict ourselves to BIC).

The inferential model "weights" from BIC selection ha
as the Akaike weights, but may be interpreted as probab
given the data, the model set, and the prior model probab
model. Define BIC differences as $\Delta\text{BIC}_i = \text{BIC}_i - \text{BIC}_n$
BIC_{min} is the minimum BIC value over all models and it c
By context we mean that the index *min* may differ for AI
do not complicate the notation to distinguish these two po

Under BIC the posterior model probabilities are given

$$\Pr\{g_i\} = \frac{\exp\left(-\frac{1}{2}\Delta\text{BIC}_i\right)}{\sum_{r=1}^{R}\exp\left(-\frac{1}{2}\Delta\text{BIC}_r\right)}.$$

If there is a true model, g_t, in the set then $\Pr\{g_t\}$ goe
infinity; and of course $\Pr\{g_i\}$ goes to 0 for all other model
sizes may be required). For model g_t the Kullback-Leibl
$I(f, g_t) = 0$. However, when model, g_t is nested in any
structures, in the model set considered, then for those o
have $I(f, g) = 0$. Hence, from the standpoint of its sel
must mean the smallest dimension representation of this tr
with smallest K that has $I(f, g) = 0$. For the nested mod
to zero for all $i > t$ as well as for all $i < t$, while $\Pr\{g_t\}$ g
large-sample behavior of BIC that we are to look for if t
true model in our set, and it is this pattern of posterior pro
be taken as evidence for the fitted model, $\hat{g}_t$, as being true.
fallacy here that we need to make very clear.

Whereas this is the asymptotic behavior BIC and $\Pr\{g_i$
model is in the set and is nested in (unnecessarily) more
sample size is quite large, it is also what will happen if I
model g_j that g_t is nested in (i.e., the added parameters
0), $I(f, g_t) = I(f, g_j)$. However, asymptotically BIC ac
on relative distances $I(f, g_t) - I(f, g_j)$, not absolute. O
in K-L distances are estimable. Thus, even if we have the
size, we cannot infer that model g_{min} selected by BIC i
$\Pr\{g_{min}\}$ is 1, or nearly 1. This is a type of nonidentifiabi
rather than parameters in models. Model g_{min} may be ver
it may not even have the correct form or the correct predi

The best that BIC can do asymptotically is identify
general versions of the apparent true model in which g_{mi}

asymptotically by both BIC and AIC. Thus, the concerns
theoretically based on the idea that for a subset of the
effects exist: All models in that subset have the identical
this is the minimum of the K-L information loss over a
Then BIC is a criterion that selects from this subset of mo
smallest dimension K.

We need a name for this property of a selected model t
"true" (as explained above) even though it is not the true
it a quasi-true model. In practice, consistent model selec
an inference that a quasi-true model has been found if the
$\Pr\{g_{min}\}$ virtually 1 and that model is nested in more ge
set. We do not need this concept of a quasi-true model if
selected; we would not be inclined to think that it is trut
theoretic approach also does not need this concept, esp
practice we expect tapering effects in our model set rathe
K-L information loss values.

To make these concepts less abstract we give a si
$x_1, \ldots, x_6$ be independent normal(0, 1) random variables
Given the x_i let $y = 100 + 15x_1 + 10x_2 + 5x_3 + 3z + \epsilon$,
normal(0, 1) random variable. Let the model set consider
regression models for response variable y:

model	predictors
1	x_1
2	x_1, x_2
3	x_1, x_2, x_3
4	x_1, x_2, x_3, x_4
5	x_1, x_2, x_3, x_4, x_5
6	$x_1, x_2, x_3, x_4, x_5, x_6$

Although y depends on z, y is uncorrelated with x_4, x_5 an
and 6, $\beta_4 = \beta_5 = \beta_6 = 0$. In this model set, model 3 is
The inferential properties of BIC, as regards model 3, ar
model 3 was the true model.

Another set of six models was also considered: The 6 m
x_4 replaced by z. Denote these as models $1z$ to $6z$. Model
to models $1z$, $2z$, $3z$. Now the true model ($4z$) is in this s
is nested in models $5z$ and $6z$. Table 6.13 presents Akaik
used) and BIC model probabilities from one random san
for each power of 10, $n = 10$ to 1 million.

In Table 6.13 consider results in the first model set for
BIC at $n = 100{,}000$ (also perhaps at $n = 10{,}000$ and o
$\Pr(g_3) = 1$. It would be careless, and wrong, to infer fro
model 3 is true. But this pattern of inferential statistics

	Inference by sample size								
Model	10		100		1,000		10,000		
set	w	Pr	w	Pr	w	Pr	w	Pr	w
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	100	49	66	89	27	87	57	99	6
4	0	35	23	10	44	12	26	1	2
5	0	12	8	1	16	1	12	0	9
6	0	4	3	0	12	0	5	0	4
$1z$	0	0	0	0	0	0	0	0	0
$2z$	0	0	0	0	0	0	0	0	0
$3z$	77	1	0	0	0	0	0	0	0
$4z$	23	51	57	83	55	96	41	97	2
$5z$	0	23	32	15	22	4	42	3	4
$6z$	0	25	11	2	23	0	17	0	3

occur if model 3 is true. In applying model selection with
but using the second model set the inferences change dram
$n \geq 100$ model 3z ($\equiv$ model 3) is ruled out; it gets an in
of 0. This is a very different inference than for the first r
is that statistically we can infer only that a best model (by
been selected, never that it is the true model. Yet the initial
underlying BIC was that it would select the true model, o
given a large enough sample size (it seemed to be implic
course the true model was in the set—where else would it
in Section 6.4.2 what criterion BIC is optimizing and ho
and posterior probabilities for BIC should be interpreted.

A few more comments on the example used here. W
cated Monte Carlo simulations, and results were as
such results are both extensive and mostly irrelevant to
section. We do note below the average value of $\Pr\{g_3$
$\Pr\{g_4|$ models $1z$ to $6z\}$. The number of Monte Carlo s
shown:

# M.C. samples	n	models 1 to 6 $\bar{P}r\{g_3\}$	models 1 $\bar{P}r\{g$
10,000	10	0.2664	0.30
1,000	100	0.8067	0.81
1,000	1,000	0.9294	0.92
100	10,000	0.9694	0.96
10	100,000	0.9947	0.98
1	1,000,000	0.9989	0.99

10,000 (or better yet 100,000).

The summary messages of this subsection: **Truth and t**
statistically identifiable from data. BIC selection prod
justifies only an inference that we have the quasi-true m
and this strained concept requires that there be a subset of
identical K-L distances (an unlikely event, it seems). Conv
very large sample sizes. So we think that in practice BIC c
frequentists want it to do in the unrealistic, idealized con
not consistent (an asymptotic property), which is the jus
given for recommending against AIC. The Bayesian pers
is more general. Finally, to argue, after selection, that yo
true model (an oxymoron) you must argue a priori that the
model set; true models are not statistically identifiable.

6.4 Contrasting AIC and BIC

6.4.1 A Heuristic Derivation of BIC

The derivation of BIC holds both the model set and the c
true) model fixed as sample size goes to infinity. It is a
model set contains the true (generating) model, then BIC
with probability 1 to that generating model as $n \to \infty$
probability of that model goes to 1), even if the generatin
some too-general set of models. The literature has not be
the derivation of BIC requires the true model to be in the
his derivation Schwarz (1978) interprets the prior probab
being the probability that model g_j is the true model (hen
probabilities are to be interpreted this same way). Howe
Neath (1999) make it clear that the derivation of BIC d
assumption about the true model being in the set of model
between AIC and BIC is the $\log(n)$ in BIC (and not in AI
needed for idealized asymptotic consistency.

So a question is, why does the $\log(n)$ arise in derivi
to understand the answer to this question, thinking it v
the issue of the role of the "true model." It was evident t
interpretations about prior probabilities are irrelevant in de
result. As used, BIC assumes equal prior probability for
easily adapted to allow any model priors. Because the deri
of any aspect of the priors on the models, its derivation and
nothing about how we should interpret model prior and po
The Bayesian literature we have seen simply refers to "the
g_j," without clarifying what these probabilities mean. We

The critical quantity to be approximated is the margin
data:

$$\int \left[\prod_{i=1}^{n} g(x_i|\theta) \right] \pi(\theta) d\theta.$$

(Section 6.4.4 gives basic formulas for the Bayesian ap
BIC in context.) The parameter θ has dimension K, as do
function of θ, the product in $g(\cdot)$ under the integral is th
we can write it symbolically as

$$\int [\mathcal{L}(\theta|x, g)] \pi(\theta) d\theta,$$

where x represents the data. Under general regularity co
size increases the likelihood function "near" the MLE, $\hat{\theta}$
the probability distribution of $\hat{\theta}$), can be well approximat

$$\mathcal{L}(\theta|x, g) = \mathcal{L}(\hat{\theta}|x, g) \mathrm{e}^{-\frac{1}{2}(\theta-\hat{\theta})' V(\hat{\theta})^{-1} (\theta-$$

Here, $V(\hat{\theta})$ is the (estimated) $K \times K$ variance–covariance
This form of the likelihood is related to the fact that the s
of the MLE becomes multivariate normal as sample size
$\hat{\theta}$ converging to a fixed value θ_0 (see Section 7.1). As reg
there is no requirement that g be the true model. It suffic
$I(\hat{\theta}) J(\hat{\theta})^{-1} I(\hat{\theta})$ (same I, J as used in TIC, Section 7.3
model, $I \equiv J$ and $V^{-1} = I$. Nevertheless, for a rando
$V(\hat{\theta})^{-1} = n V_1(\hat{\theta})^{-1}$ where the matrix $V_1(\cdot)$ is independen
$V_1(\hat{\theta})^{-1}$ converges to $V_1(\theta_0)^{-1}$.

Now we consider the needed integral, which is approx

$$\mathcal{L}(\hat{\theta}|x, g) \int \mathrm{e}^{-\frac{1}{2}(\theta-\hat{\theta})' V(\hat{\theta})^{-1} (\theta-\hat{\theta})} \pi(\theta)\, d\theta$$

As n goes to infinity the approximation becomes exact, th
trates near $\hat{\theta}$ (which is converging to θ_0) and the prior is
(over the space where $\hat{\theta}$ has any substantial probability
treat $\pi(\theta)$ as a constant. Alternatively, in the spirit of h
we can just directly use the improper prior $d\theta$. The neede
related to the underlying multivariate normal distribution
because we know the needed normalizing constant:

$$\int (2\pi)^{-K/2} \| V(\hat{\theta})^{-1} \|^{1/2} \mathrm{e}^{-\frac{1}{2}(\theta-\hat{\theta})' V(\hat{\theta})^{-1} (\theta-\hat{\theta})}\, d$$

$$\int \left[\prod_{i=1}^{n} g(x_i|\theta)\right] \pi(\theta)\, d\theta \approx \mathcal{L}(\theta|x, g) \left[(2\pi)^{K/2}\|V($$

$$= \mathcal{L}(\hat{\theta}|x, g) \left[(2\pi)^{K/2}\|nV$$

and by a property of the determinant, $\|nV_1(\hat{\theta})^{-1}\| \equiv n^K \|V$ the approximation

$$\int \left[\prod_{i=1}^{n} g(x_i|\theta)\right] \pi(\theta)\, d\theta \approx \mathcal{L}(\hat{\theta}|x, g) \left[(2\pi)^{K/2} n^{-K/2}\right.$$

Taking -2 times the log of the right hand side above, we BIC criterion:

$$-2\log(\mathcal{L}(\hat{\theta}|x, g)) + K\log(n) - K\log(2\pi) - \log$$

The literature drops the last two terms of the expressio because, asymptotically, they are dominated by the ter well as by the log-likelihood term (which is of order n).

We now see that the $\log(n)$ term arises because of the qui feature of marginalization over θ (i.e., integrating out θ) matical requirement in the derivation of BIC that the mo the model set does not need to contain the true model. H nothing in the foundation or derivation of BIC that addre tradeoff, and hence addresses parsimony as a feature of B This is not a strike against BIC because this tradeoff is a not explicitly invoked in Bayesian statistics. But we are le basis to know what sort of parsimony the BIC model sele Simulation studies of this question have been done, but th paring AIC and BIC performance depend on the nature o model (such as having many tapering effects or not), or set contains the generating model, on the sample sizes co objective: select the true model or select the K-L best ap One can simulate situations where either BIC or AIC is Thus, it is the unknown context and intent (i.e., true mod their use that is critical for deciding which method is "co

6.4.2 A K-L-Based Conceptual Comparison of A

The motivation for this section was to clarify what obj meant, in the Bayesian-oriented literature about BIC, by the model," or similar such vague phrases. We precede objective K-L discrepancy, as opposed to allowing "mo be subjective, hence meaningless (to us). Here are two ex

denotes your probability that the jth model is the most
commonly the literature simply refers to the probability
no clarity about what this means as regards a model bein
"appropriate" by some unspecified criterion.

From its operating characteristics we know that BIC's r
is that it asymptotically will select, with probability 1, the
true model is in the set. However, such convergence in a s
sense to a single model does not, and cannot, logically mea
(Section 6.3.4). In fact, as sample size $n \to \infty$, the mo
is consistent for the quasi-true model in the model set. W
quasi-true model below.

For a set of R models the Kullback-Leibler "distance
truth is denoted $I(f, g_r)$ (Section 2.1.3). If $g_r \equiv g_r(x|$
denote a parametric family of models with $\underline{\theta} \in \Theta$, Θ bei
space, then g_r is the family member for the unique $\underline{\theta}_0 \in$
closest to truth in K-L distance (see Section 7.2). For ou
also assume the models are indexed worst (g_1) to best, i.e
$I(f, g_2) \geq \cdots \geq I(f, g_R)$. Let Q be the tail-end subset o
by $\{g_r, r \geq t, 1 \leq t \leq R | I(f, g_{t-1}) > I(f, g_t = \cdots = I$
because $t = R$ is allowed, in which case the K-L best mod
is unique. For the case when Q contains more than one mo
we assume the models g_t to g_R are ordered such that K_t
(in principle $K_t = K_{t+1}$ could occur).

The set Q contains models that are all equally good app
distance, to truth f. However, we can further distinguish t
eter space dimension, and we must prefer the smallest
$t < R$, and $K_t < K_{t+1}$ holds, then model g_t is the unique
the R models. As a matter of inference from data, BIC mo
sistent for this quasi-true model, which is not absolute truth
(In principle, there might not be a unique quasi-true mod

Both AIC and BIC model selection actually depend on
$I(f, g_i) - I(f, g_j)$, not on absolute K-L values. Only t
estimable. For a random sample we can write $I(f, g_i)$
$I_1(f, g_i)$ being for $n = 1$ is a constant as regards sample siz
$I(f, g_j) = n(I_1(f, g_i) - I_1(f, g_j))$. Because the MLE $\hat{\underline{\theta}}$
basic convergence properties of AIC and BIC for large n
approximations such as

$$\mathrm{AIC}_i - \mathrm{AIC}_j \approx 2n[I_1(f, g_i) - I_1(f, g_j)] + (K_i -$$

$$\mathrm{BIC}_i - \mathrm{BIC}_j \approx 2n[I_1(f, g_i) - I_1(f, g_j)] + (K_i -$$

model.

In the case of tapering effects, so $t = R$,

$$2n(I_1(f, g_i) - I_1(f, g_R)) > 0, \quad i < R$$

Hence, as $n \to \infty$ all these differences diverge to ∞ at a ra
Also, the magnitude of these differences dominates the "pe
at best only grow at a rate proportional to $\log(n)$. Therefor
BIC-selected model will converge to model g_R with certa

The case of a nontrivial quasi-true model (i.e., $t < R$) p
sively) corresponds to model g_t nested in models g_i, $i >$
case here. The relevant differences are

$$\begin{aligned}
\mathrm{AIC}_i - \mathrm{AIC}_t &\approx 2n[I_1(f, g_i) - I_1(f, g_t)] + (K_i - K_t)2 \\
\mathrm{AIC}_i - \mathrm{AIC}_t &\approx -\chi_i^2 + (K_i - K_t)2, \\
\mathrm{BIC}_i - \mathrm{BIC}_t &\approx 2n[I_1(f, g_i) - I_1(f, g_t)] + (K_i - K_t)\log \\
\mathrm{BIC}_i - \mathrm{BIC}_t &\approx -\chi_i^2 + (K_i - K_t)\log(n),
\end{aligned}$$

Here, χ_i^2 is a central chi-square random variable on $K_i - K_t$
For all $i < t$ the differenes $\mathrm{AIC}_i - \mathrm{AIC}_t$ and $\mathrm{BIC}_i - \mathrm{BIC}$
$n \to \infty$, with probability 1, hence model g_t is always s
g_1 to g_{t-1}. For all $i > t$ the differences $\mathrm{BIC}_i - \mathrm{BIC}_t$
$n \to \infty$, with probability 1, as long as $K_i > K_t$ because
to ∞. Hence, if there is a nontrivial quasi-true model th
frequentist sampling theory) of the model being the one s
to 1 for a big enough sample size.

By contrast, for $i > t$ we only have $\mathrm{E}(\mathrm{AIC}_i - \mathrm{AIC}_t)$
is > 0 but is independent of n. While these expected
positive, the actual AIC values are random variables wit
that AIC does not select model g_t with certainty in this h
of $t < R$. However, even if this were the situation, Shibata
shows that there is a sense in which use of AIC leads t
estimates and predictions as $n \to \infty$.

It is clear to us, as argued for above, that the Bayesian "
g_i" used in conjunction with BIC can, and must, be interpr
Mathematically it is

$$\Pr\{g_i | \, data\} = \frac{\exp(-\frac{1}{2}\Delta\mathrm{BIC}_i)}{\sum_{r=1}^{R} \exp(-\frac{1}{2}\Delta\mathrm{BIC}_r)}$$

This posterior probability assumes equal prior probabiliti
models and is conditional on that set of R models. Conce
must be interpreted as the probability that model g_i is the
the set of R models. This model will generally be unique

if inference from data is to have credibility.

Finally, given the necessity of this interpretation on the
of model g_i, then logically one must interpret the model
the same way, as the prior probability (it can be degree of b
the quasi-true model in the set of R models. If one wants to
that model g_i might be the true model, then one must be
the true model is already in the model set (see e.g., Wasser
of possible one-to-one transformations of models the fo
model dimension, K_t, which is unique, hence the dimensi

In summary, K-L distance is fundamental to understa
of both AIC and BIC model selection. Both selection cri
and applied, without assuming the true model is in the
the defining characteristic of BIC (i.e., what is it trying t
asymptotically in relation to the concept of a quasi-true m
seeks to select only a best model at a given sample size; "
an expected estimated K-L criterion which serves to recog
trade-off in model selection. For AIC, "best" varies with n
i.e., the quasi-true model, does not depend on n. Howe
BIC-selected model approaches its target "best" model
of the model ordering imposed here by the $I(f, g_r)$. Ho
AIC and BIC performance depends on the performance cri
model or best model), the assumptions they make (usually
simulation studies) about the underlying K-L values, $I(f$
and the sample sizes considered. Failure to properly re
factors and issues has led to much confusion in the mode
about AIC versus BIC.

6.4.3 Performance Comparison

A generally accepted measure of model and model selec
predictive mean square error (MSE). We evaluated the pred
and BIC using Monte Carlo simulated data that mimics th
in Section 6.2. Thus, the simulated variable y has the pr
fat data example (such as tapering effects) and is generat
using 13 predictors (the $\underline{x}$) as described in Section 6.2.
comparsion involved generating a random sample of size
$\underline{x}$, generating E(y) ($\equiv$ here to E($y|\underline{x}$)) for each given $\underline{x}$, a
$y = \mathrm{E}(y) + \epsilon$ for the first 252 observations. Model selec
of size $n = 252$; $\underline{x}_{253}$, and E(y_{253}) were set aside.

Initially, classical model selection was done; hence we
best model selected under AIC_c and under BIC for each sim
using these selected best models and $\underline{x}_{253}$, E(y) ($\equiv$ E($y|\underline{x}$

the 10,000 values of $(\mathrm{E}(y_{\mathrm{AIC}_c}) - \mathrm{E}(y))^2$ and $(\mathrm{E}(y_{\mathrm{BIC}}) -$
BIC, respectively.

The model selection procedure with the smallest MSE
dure. The results of this simulation were a MSE of 5.6849
7.6590×10^{-6} for BIC. Thus the ratio of predictive MS
estimated as 0.74. Both MSEs have a coefficient of variatio
statistical comparison is based on the paired nature of the c
at the 10,000 values of $diff = (\hat{\mathrm{E}}(y_{\mathrm{BIC}}) - \mathrm{E}(y))^2 - (\hat{\mathrm{E}}(y_{\mathrm{AIC}_c}$
of *diff* was 1.9741×10^{-6} with a standard error of 0.0908
hence, a 95% confidence interval on E(*diff*) of 1.8×10^{-6}
results about the sample distribution of the 10,000 values
and $(\hat{\mathrm{E}}(y_{\mathrm{AIC}_c}) - \mathrm{E}(y))^2$ are listed below:

performance	sample percentiles $\times 10^6$				
measure	5	10	50	90	95
$(\hat{\mathrm{E}}(y_{\mathrm{AIC}_c}) - \mathrm{E}(y))^2$	0.019	0.089	2.5	14.9	22.
$(\hat{\mathrm{E}}(y_{\mathrm{BIC}}) - \mathrm{E}(y))^2$	0.029	0.110	3.3	20.5	29.

This shows that the predictions based on the AIC_c
stochastically closer to the true $\mathrm{E}(y)$ values than are the
BIC-selected model (as opposed to the result for MSE
BIC-produced outliers).

In 5.7% of the simulated sample $diff = 0$ occurred; al
AIC_c and BIC selected the same model. In the other 94.3
0 occurred in 57.3% of cases. Thus, overall by our pe
AIC_c-model selection performed as well as or better tha
samples.

Simple linear regression of $\hat{\mathrm{E}}(y)$ on $\mathrm{E}(y)$ was also done
intercept (β_0) and slope (β_1) parameter by each model
The results are listed below, with standard errors in paren

procedure	$\hat{\beta}_0 \times 10^6$	$\hat{\beta}_1$	
AIC_c	−8.83(23.8)	0.9936(0.0016)	0
BIC	12.80(27.6)	0.9818(0.0019)	0

These results suggests that prediction based on the AIC
much closer to being unbiased than prediction based
model. It is interesting that the smaller prediction MSE o
at the expense of greater prediction bias.

A second set of 10,000 simulations was done wherein $\hat{\mathrm{E}}($
on model averaging the $\hat{\mathrm{E}}_r(y)$, $r = 1, \ldots, R = 8{,}191$ d
one for each fitted model. Under such a model-averaged
model weights vary by selection (i.e., weight generation)
of $\hat{\mathrm{E}}_r(y)$. The results of this simulation example of mo

MSEs have a coefficient of variation of 1.5%. Other aspect
prediction-based model averaging are about the same as tho
best-model strategy.

A final noteworthy comparison is that the MSE valu
smaller under model averaging, as shown below, for MSE

method	model averaged	best model	ratio
AIC_c	4.8534	5.6849	0.85
BIC	5.8819	7.6590	0.77

For this simulation scenario (which mimics real da
prediction beats the traditional best model approach, and

Producing the above simulation results for the best-m
9 hours of CPU time on a dedicated 1.9 GHz speed co
RAM and 80G hard drive (and about 14 hours clock tim
aged example took 12 hours of CPU time and 19 hours
same computer. Thus, whereas, extensive simulation studi
needed, it will be very computer intensive.

We also examined a few aspects of AIC_c versus BIC m
the simulation scenario of the example in Section 3.4. The
on a real capture–recapture experiment, and the simulated
data. The reader would benefit from scanning parts of Se
Section 3.4.6, before reading the next two paragraphs. In
shows the results of AIC model selection on a set of 14 m
on 50,000 Monte Carlo data sets. BIC model selection
these 14 models. The key point we make here is that the
between the models selected by AIC and BIC.

BIC selection in the face of tapering treatment effect
of "only" 2,500 (but still quite less than ∞), and a ger
34 parameters that was not (quite) in the set of candidat
poorly, as theory would suggest. BIC selection most freq
$g_{2\phi}$ (26.8%), followed by model $g_{1\phi}$ (24.5%), and mode
selection frequencies fell rapidly for models g_{3p}, $g_{3\phi}$,
9.8, 3.1, and 1.3, respectively). The BIC selected mod
underfit and would have poor confidence interval covera
didate models were to be expanded to include the gene
then as sample size increased, BIC should select model
one. The initial number of nestlings required for BIC to s
model ($K = 34$) with probability approaching one in th
plex example is approximately 108,000 birds (instead of
example).

Clearly, it would often be impossible to find and ban
one small island for a particular year; it might be quite u

several years. However, in so doing, other factors becom
conceptualization of truth must include obvious factors su
in addition to slightly less obvious factors such as techn
resighting probabilities and islands with differing vegeta
resighting probabilities. The "year" effect is not so much
year, but a host of covariates (most unmeasured) that affe
resighting probabilities in complex, nonlinear ways acr
there is individual heterogeneity that is substantial (e.g., w
growth rate, dispersal distance). Thus, the concept of tru
very complex. To think that such reality exists as an exac
useful; to think that such a true model is included in the set
seems absurd.

The primary foundations of the BIC criteria do not ap
sciences and medicine and the other "noisy" sciences. R
sample size is increased by orders of magnitude in biolog
the target "true model" sought by BIC increases in size
simple fact is a violation of the assumptions that form the ba
this is allowed under the AIC-type criteria.

Many other published works have compared AIC or A
selection. Some, but not all, of these studies note the reali
For example Hjorth (1994:46) says "... the asymptotic
measures is theoretically correct, but in the model select
be skeptical against asymptotic results when we are analy
50 or even 500 observations. The asymptotics may requi
ten more data before a reasonable accuracy is achieved,
possible." McQuarrie and Tsai (1998) report on extensive
of model selection methods. A parsimonious summary of
on pages 410–411 of that book.

6.4.4 *Exact Bayesian Model Selection Formulas*

The analytical formulae for Bayesian model selection ar
to help put BIC in context. Implementation of these for
computer-intensive methods such as Markov chain Mont

For structural model g_i the likelihood of θ_i (a $K_i \times 1$ vec
denoted as $g_i(x|\theta_i)$ ($\equiv \mathcal{L}(\theta|x, g)$ as used in Section 6.4.1).
for θ_i is denoted $\pi_i(\theta_i)$. A key quantity needed is the mar

$$g_i(x, \pi_i) = \int g_i(x|\theta_i)\pi_i(\theta_i)\, d\theta_i.$$

This is taken (by us) as the likelihood of model structure
the prior on θ_i. In essence this quantity is what was appr
6.4.1 to obtain the BIC criterion. The posterior distribut

$$m_i(\theta_i|x, \pi_i) = \frac{g_i(x, \pi_i)}{g_i(x, \pi_i)}.$$

Let the prior probability for model structure g_i be deno
relates only to the structure of model g_i, not its paramete
should be interpreted as the prior probability that model
model in the set of R models. The posterior probability
quasi-true model in the set of R models is given by

$$m_i(x, g_i, \pi_i) = \frac{g_i(x, \pi_i)p_i}{\sum_{r=1}^{R} g_r(x, \pi_r)p_r}.$$

From data analysis alone one cannot conclude that a m
generating model even if it gets $m_i = 1$. There is an ide
here. All we can know from the data is that a model is th
i.e., for a large enough sample size we can infer that the
in the model set that has a smaller K-L discrepancy, even
question is nested in some other models. In order to infe
we need to believe (i.e., *know*) a priori that the true mode
(we just do not know a priori which model it is). This see
it is justifiable to believe that there is always a single best
set, if we have a suitable criterion for what is a best mode

The posterior odds ratio is informative:

$$\frac{m_i(x, g_i, \pi_i)}{m_j(x, g_j, \pi_j)} = \left[\frac{g_i(x, \pi_i)}{g_j(x, \pi_j)}\right]\left[\frac{p_i}{p_j}\right].$$

The ratio $g_i(x, \pi_j)/g_i(x, \pi_j)$ is called the Bayes factor; i
information-theoretic evidence ratio. Often, BIC is motiva
a simple approximation to the Bayes factor, inasmuch as
be difficult to compute exactly. The prior odds ratio is
posterior odds ratio solely by the Bayes factor.

6.4.5 *Akaike Weights as Bayesian Posterior Mod*

For a large sample size a good approximation to the Baye
probability can be based on BIC, provided one is willing t
model probabilities (useful background Sections here ar
and 6.4.4). However, the expression BIC can be used mor
model priors. Let p_i be the prior probability placed on

[1] A note to the reader: At the time we were checking the secon
this book (late March 2002) we found relationships of such importa
to include them. The results are this section.

$$\Pr\{g_i|data\} = \frac{\exp\left(-\frac{1}{2}\Delta\text{BIC}_i\right)p_i}{\sum_{r=1}^{R}\exp\left(-\frac{1}{2}\Delta\text{BIC}_r\right)p_r}$$

To get Akaike weights we use the model prior

$$p_i = B \cdot \exp\left(\tfrac{1}{2}\Delta\text{BIC}_i\right) \cdot \exp\left(-\tfrac{1}{2}\Delta\,\text{AIC}_i\right)$$

B is a normalizing constant (p_i simplifies, as will be shown)

$$\exp\left(-\tfrac{1}{2}\Delta\text{BIC}_i\right) \cdot \exp\left(\tfrac{1}{2}\Delta\text{BIC}_i\right) \cdot \exp\left(-\tfrac{1}{2}\Delta\,\text{AIC}_i\right) =$$

hence, with this prior probability distribution on models

$$\Pr\{g_i|data\} = \frac{\exp\left(-\frac{1}{2}\Delta\text{BIC}_i\right)p_i}{\sum_{r=1}^{R}\exp\left(-\frac{1}{2}\Delta\text{BIC}_r\right)p_r} = \frac{\exp\left(-\right.}{\sum_{r=1}^{R}\exp\left(\right.}$$

which is the Akaike weight for model g_i.

Moreover, this model prior is actually simple and not d
as shown by simplifying how it is expressed (there is an
the formula for the normalizing constant):

$$\begin{aligned} p_i &= B \cdot \exp\left(\tfrac{1}{2}\Delta\text{BIC}_i\right) \cdot \exp\left(-\tfrac{1}{2}\Delta\,\text{AI}\right. \\ &= B \cdot \exp\left(\tfrac{1}{2}[\Delta\text{BIC}_i - \Delta\,\text{AIC}_i]\right) \\ &= C \cdot \exp\left(\tfrac{1}{2}[\text{BIC}_i - \text{AIC}_i]\right) \\ &= C \cdot \exp\left(\tfrac{1}{2}K_i\log(n) - K_i\right) \end{aligned}$$

and

$$C = \frac{1}{\sum_{r=1}^{R}\exp\left(\frac{1}{2}K_r\log(n) - K_r\right)}.$$

The result easily generalizes to AIC_c and formally to
Bayesian would not handle overdispersion by simply
Formally, the Akaike weights from AIC_c are Bayesi
probabilities for the model prior

$$p_i = C_c \cdot \exp\left[\frac{1}{2}K_i\log(n) - \frac{nK_i}{n-K_i-1}\right.$$

$$C_c = \frac{1}{\sum_{r=1}^{R}\exp\left[\frac{1}{2}K_r\log(n) - \frac{nK_r}{n-K_r-1}\right]}.$$

We will call this the K-L model prior. Because BIC is for
bility of the result (i.e., AIC_c as Bayesian) can be questio
it does apply for large n and small or large K_i.

Do not be confused by the $\log(n)$ appearing in both
model prior. BIC actually arises in the context of obtai
approximation to the Bayes factor, a quantity that is un

Rather, the $\log(n)$ in BIC arises owing to the reformulati
assuming vague priors on the parameters in model g_i (Se
the BIC formula is used with any prior p_i that is not a fun
then the posterior will converge (as $n \to \infty$) to the quasi-
of models.

When the K-L model prior is used with BIC so that we g
result, the interpretation of the model probabilities is not
Whereas the quasi-true model g_i is the "target" of BIC (
sample size), the target model of AIC is the model that
estimated K-L information loss. From Section 7.3 (7.18),
that minimizes

$$E_{\underline{y}}\left[I(f, g(\cdot|\hat{\underline{\theta}}(\underline{y})))\right] = \text{ constant } - E_{\underline{y}}E_{\underline{x}}\left[\log[g\right.$$

For simplicity we call this target model the K-L best m
sample size and it is essentially a fitted model as oppos
$g_t(x|\underline{\theta}_o)$ for BIC, which is the model of smallest dimensio
imum $I(f, g_i)$. Thus as a Bayesian result we must interpr
w_i and a model prior p_i as the probability that model g_i is

The target K-L best model has a variance-bias trade-o
model. Such a trade-off depends on both sample size and n
to be estimated. Still, we did not expect that the model pri
in a Bayesian context had to depend on n and K: That pri
be independent of these values (which are known prior to
easy to numerically explore the K-L model prior so we
numerical examples.

There are now two ways to compare AIC and BIC. One
tist framework of looking at sampling measures of perfor
predictive mean square error and confidence interval cove
is to consider AIC as Bayesian and think about and com
prior to the K-L model prior in conjunction with knowi
models for selection are different, and the interpretation o
is different, for AIC versus BIC. Such comparisons must
of what we are assuming about the information in the data
estimation and the models as approximations to some co
generating distribution ("truth"). It is useful to think in
$|\theta|/\text{se}(\hat{\theta})$. We would assume few or no effects are truly ze
meaningful, informative data and thoughtfully selected pr
We assume tapering effects: some may be big (values lik
are only 2, 1 or 0.5, or less. We assume we can only estima
eters reliably; m might be 20 or as small as 10. These ideas
of savvy model priors, with properties like the K-L prior
and K. The K-L prior is a particularly important savvy m

bility. However, the interpretation of what the probability
is different for AIC versus BIC. Also, to use the Bayes f
provided by BIC in the context required by AIC implies th
prior probability on model g_i be an increasing function o
function of K (i.e., a savvy prior). The implicit BIC prior c
in the information-theoretic context.

6.5 Goodness-of-Fit and Overdispersion

Overdispersion of count data, relative to a theoretical m
with to obtain valid inferences. A more sophisticated app
advocated here is to incorporate one or more variance para
the parametric model, hence into the likelihood. As noted
this is a desirable approach that can be quite flexible a
is not the simple omnibus approach that QAIC and QA
Lindsey's ideas here would take us too far afield from our
jective. Instead, we present a generalization of QAIC to a
overdispersion parameter. This, however, assumes that th
by some factor, or factors, that allow partitioning the da
of overdispersion, c, can vary by data subset. Before de
sues, we consider a strategy for obtaining $\hat{c}$ when there i
model.

6.5.1 Overdispersion $\hat{c}$ and Goodness-of-Fit: A

When there is a global model we can usually compute fro
$\hat{c}$. The logic is that the global model is theoretically the
because all other models are special cases of the global
cases cannot, on average, fit the data better than the globa
atic overdispersion exists in the data, then the goodness-o
average exceed its degrees of freedom (as shown by a s
will be true even if the global model is structurally adequ
find $\hat{c}$ meaningfully > 1 we do not know whether this is a
sion or inadequate model structure. In the end this distinc
operationally if we cannot generalize the global model.
model is not adequate, this must be accounted for, and it
We prefer to say that the lack of fit problem is due to ove
then the global model is structurally adequate.

However, the use of $\hat{c}$ (when $c > 1$ is clearly indicat
inference for count data (such as inflated standard erro
means that the inferences are based on empirical residua

tomatically reflect both lack of model structural adequa
stochastic variation about "truth." Thus using $\hat{c}$ we are co
ferences because importance of a structural data feature (i
judged against the totality of unexplained stochastic vari
lack of fit, just as with models for continuous data.

With no global model, there is the question of how to
$\hat{c}$. For example, in the dose-response example of Section
distinct a priori models, all on an equal footing. None is n
all have the same number of parameters. We must com
(and $\hat{c}$) for each model, because there is no theoretical b
is the best-fitting model. If the best-fitting model has a
use $c = 1$. If even the best-fitting model is a poor fit (
use $\hat{c}$ from that model. The logic here is that overdisper
show up in each goodness-of-fit test, but so will inadequ
If at least one model is adequate, then on average, its go
only overdispersion, and that would, on average, be the
more than one model fits the data (no overdispersion, c
chance get $\hat{c} < 1$. Fortunately, this is not a problem, b
$c = 1$.

The general strategy to obtain $\hat{c}$ is as follows. Partition
into s subsets, where each subset (size R_i, $i = 1, \ldots, s$
global model. Thus, one gets s ($\leq R$) subglobal models
subsets of each other. Compute $\hat{c}_i$ for each subglobal mo
data are used in fitting each subglobal model. Even if one
global models structurally fits the data, if there is problema
will stochastically inflate all of the goodness-of-fit tests. O
structurally fits the data, its $\hat{c}$ value estimates actual ove
smallest $\hat{c}$ as the estimate: $c = 1$ if for that goodness-of-
a guideline). Otherwise, use the computed $\hat{c}$.

We do not see a simple alternative to this strategy, altho
potential biases arising from taking the minimum of a
smaller is s, the better, as this minimizes potential selec
such bias is not a big concern here, for the reason that the
in each calculation of goodness-of-fit. As a result, the $\hat{c}$
pairwise correlated and this reduces selection bias (whic
of for selection over independent random variables). Of
be small degrees of freedom for a goodness-of-fit statisti
other subglobal models have much larger degrees of freed
freedom for goodness-of-fit leads to less reliable $\hat{c}$. Howev
model allows ample, or similar degrees of freedom for $\hat{c}$ a
or 4) this strategy should work. Assessing goodness-of-fi
$\hat{c}$ (in particular), requires some judgment. If nothing else
as to when we use $c = 1$, i.e., when we judge that the mo

strictive for many instances of count data. Such data can hav
due to being collected across factors such as gender, age
ments, and so forth. The degree of overdispersion may v
For example, it is easy to imagine that in a survival study,
for males versus females that a common $\hat{c}$ should not be

Assume that the data are naturally partitioned into
sets, hence symbolically $data = \cup_{v=1}^{V} data_v$. Then th
log-likelihood is a simple sum of separate log-likelihoods

$$\log \mathcal{L}(\theta|data, g) = \sum_{v=1}^{V} \log \mathcal{L}(\theta|data_v, g$$

We let θ represent the full vector of structural parameters, c
may be in common over different parts of the likelihood.
data has likelihood component denoted by $\mathcal{L}_v(\theta|data_v, g)$
the dispersion parameter for $data_v$ be c_v. The appropriate
is actually

$$\sum_{v=1}^{V} \frac{\log \mathcal{L}_v(\theta|data_v, g)}{c_v}.$$

The overdispersion parameters must be estimated under th
same approaches apply here as are discussed in Section
apply for each subset of the data. The overall goodness-of-
sum of the separate chi-square statistics on their summed
However, if even by this overall result we judge the glo
should check whether any separate component, $data_v$, cle
can use some $c_v = 1$ if warranted, while other $\hat{c}_v > 1$ are

There are some caveats. In theory, the V data subsets sh
hence without correlations across subsets. However, one c
overdispersion is correlation structure in the data, which
both within and between subsets. Hence, it would be bes
that subsets are independent. Still, even if there are weak
data subsets, if the c_v are quite variable, this generalized
to be preferred to the simple case of only a single $\hat{c}$. We
the data partitioning is a priori to data analysis. In a se
any model, even though it is based on factors that may al
construction.

Determination of the data partition, hence also the va
based on subject-matter knowledge. Strive to have V s
data, partition on factors such a sex (i.e., males vs. female
vs. adults), especially if the sexes and ages have different b
sure to use the total number of parameters as $K = p+V$, w

The QAIC formula is not changed by this data partitio

$$\text{QAIC} = \sum_{v=1}^{V} \frac{-2 \log \mathcal{L}_v(\hat{\theta}|data_v, g)}{\hat{c}_v} + 2$$

However, its small sample version, QAIC_c is problema involving reciprocals in n_v should be used, but these t partitioning of K over v. We considered the issue of a but more complex, formula versus the value of just using formula, hence minimizing the number of formulae one now we suggest using just

$$\text{QAIC}_c = \sum_{v=1}^{V} \frac{\log \mathcal{L}_v(\hat{\theta}|data_v, g)}{\hat{c}_v} + 2K + \frac{2K}{n -}$$

When $V = 1$ we can ignore the estimate $\hat{c}$ while finc is, we can work directly with $\log \mathcal{L}(\theta|data, g)$. Moreover, nominal empirical variance-covariance matrix $\hat{\Sigma}$ (as the in directly from $\log \mathcal{L}(\theta|data, g)$, and then the appropriate matrix of $\hat{\theta}$ is taken as $\hat{c}\hat{\Sigma}$. Also, the degrees of freedom (variance estimates is the df of $\hat{c}$. However, for $V > 1$ we with the quasi log-likelihood. To find the MLE, we must

$$\sum_{v=1}^{V} \frac{\log \mathcal{L}_v(\theta|data_v, g)}{\hat{c}_v}$$

over θ. The likelihood equations to solve are

$$\sum_{v=1}^{V} \left[\frac{1}{\hat{c}_v}\right] \left[\frac{\partial \log \mathcal{L}_v(\theta|data_v, g)}{\partial \theta_i}\right] = 0, \qquad i =$$

The elements of the $K \times K$ appropriate Hessian are

$$\sum_{v=1}^{V} \left[\frac{1}{\hat{c}_v}\right] \left[\frac{\partial^2 \log \mathcal{L}_v(\theta|data_v, g)}{\partial \theta_i \partial \theta_j}\right] = 0, \qquad i, j$$

Each $\hat{c}_v$ has associated degrees of freedom $\text{df}_v = n_v -$ the number of structural parameters in the global model f this global model there should be no parameters in comm Hence, another complication is determing the appropriat hence the variance of any component of $\hat{\theta}$. As an exact simply add as $\text{df} = \Sigma \text{df}_v$. However, to keep it simple w this summed df for the applicable degrees of freedom.

A single $\hat{c}$ should often suffice, and should not be qu more complicated approaches: Favor parsimony even in

be ignored (e.g., 1.3 versus 1.7) given that the $\hat{c}_v$ are o
sufficient df_v the levels of differences to be concerned
versus 2, or 1.4 versus 2.8. Bear in mind that exact mod
overdispersion is not as important as having at least a ba
$V = 1$) for overdispersion, as by quasi log-likelihood.

6.5.3 Model Goodness-of-Fit After Selection

Often the set of models under consideration contains a mo
global model), in which case we recommend assessing
model to the data (preferably before commencing with
the global model fits, as by some standard goodness-of-fi
selected model will fit the data. We think that this is true
selection (but we are not sure). If the global model doe
one might decide that the lack of fit is not of concern and
or $QAIC_c$. (In fact, if the global model does not fit, bu
K-L–based model selection, you must use QAIC or QAI

There is a philosophy under which one would want to
are using BIC, even when the context is such that AIC
clear that BIC selects more parsimonious models than AI
to model fit; if the AIC selected model fits, will the BIC
the same data also fit? This is a question we have never s
literature, and we do not know the answer.

The paper by Leroux (1992) motivated our interest i
oux (1992) reports the observed versus expected count f
automobile accident data ($n = 9{,}461$). A pure Poisson
two mixture models are fit to the data. BIC selects the t
ture model ($K = 3$), while AIC selects the three-compo
($K = 5$). Model selection tends to lead to overly optim
model fit. Hence, model selection may result in optimistic
tical model fit for the selected model. However, a goodne
to the global model will not be biased, because no selectio
curred. The usual chi-square goodness-of-fit procedure a
selected in Leroux (1992) (this entails some pooling of s
$\chi^2 = 1.11$ (1 df, $P = 0.2921$) for the AIC-selected mode
df, $P = 0.0031$) for the BIC-selected model.

Whereas the goodness-of-fit $P = 0.0031$ is small, the
size here, and perhaps therefore it is acceptable to use f
that statistically is not a good fit to the data. We think tha
acceptable, but it must be argued for on a case-by-case b
ticians have consistently cautioned about drawing infer
that does not fit the data. We should not ignore the issu
selection procedures systematically select models that do

data at the nominal α levels used in goodness-of-fit tests.
to understand this general issue of the fit, and assessing
data after model selection.

6.6 AIC and Random Coefficient Models

6.6.1 Basic Concepts and Marginal Likelihood

Parameters are sometimes considered as "random effects"
as random coefficients; see Longford 1993). In the simp
are parameters all of the same type (e.g., survival rates
consider the K elements of $\underline{\theta}$ as random variables. Thus,
we conceptualize $\theta_1, \ldots, \theta_K$ as independent random variab
variance σ^2. Now the inference problem could be entirely
fixed population-level parameters μ and σ^2. However, th
directly write down is for $\underline{\theta}$ as if the elements of $\underline{\theta}$ were the
direct inference interest. In using the likelihood $\mathcal{L}(\underline{\theta})$ we a
how $\theta_1, \ldots, \theta_K$ may have arisen from some process or son
population. The likelihood $\mathcal{L}(\underline{\theta})$ is appropriate for wher
"fixed effects." This is a valid approach if we interpret the
θ_K as deterministic. It is then possible to fit this global mo
lihood methods and also fit simpler models based on dete
on $\underline{\theta}$, such as $\theta_i \equiv \mu$, where the likelihood is $\mathcal{L}(\mu)$.

However, we may also want to consider an intermedi
only the two parameters μ and σ^2, where we regard θ a
with mean μ and variance σ. Thus while we directly have
$\mathcal{L}(\mu)$ (hence models g_K and g_1), we also want the likeli
for the two-parameter model g_2. The parameter σ^2 in mod
possible stochastic nature of the θ_i. Model g_K allows arbi
θ_i, but this freedom costs us $K-2$ extra parameters comp
the unexplained variation in these K parameters is substa
with them being considered as exchangeable random v
select model g_2 rather than model g_K. Model g_2 with o
μ and σ^2, parsimoniously allows for variation in the θ_i (s
does not do parsimoniously and model g_1 does not do at

Whereas the likelihood $\mathcal{L}(\mu)$ is a special case of eithe
the conceptually intermediate model g_2 is not mathematic
model between models g_K and g_1 in the simple sense o
ministically constrained version of the global model g_K.
cannot be fit by standard likelihood methods based only
likelihood $\mathcal{L}(\underline{\theta})$ and deterministic constraints. These ra
random-effects) models are different from other models

Our focus here is on how we can compute a valid AIC fo
models, such as this case of having only the parameters μ
our only obvious starting point is model g_K and its likel
likelihood approach to stochastic parameters is to postulat
bution, hence a model, for the random variable θ, say $h(\theta$
the needed (proper) likelihood $\mathcal{L}(\mu, \sigma^2)$ based on the ma

$$g(\underline{x} \mid \mu, \sigma^2) = \int g(\underline{x} \mid \underline{\theta}) h(\underline{\theta} \mid \mu, \sigma^2) d\underline{\theta}$$

thus $\underline{\theta}$ has been integrated out. Considering $g(\underline{x} \mid \mu, \sigma^2)$
parameters given the data, we have $\mathcal{L}(\mu, \sigma^2) = g(\underline{x} \mid \mu, \sigma$
model g_2 is computed based on $g(\underline{x} \mid \mu, \sigma^2)$, which we ca
the integral in (6.1).

A more informative way to think about random-coeffi
sometimes a few parameters, defined by deterministic con
explain all the variation in the much larger set $\theta_1, \ldots, \theta_K$,
(like 2 or 3). We might have $K = 10, 15$, or 20 (or more). If
of the same type of parameter (perhaps for 20 years or sites)
is some "explainable" (i.e., consistent, simple, and unde
low-level pattern, such as a linear trend, to the variation
However, to be consistent with our philosophy of models
the actual values of the 20 parameters will not perfectly fit
(a two-parameter linear trend). There will be unignorabl
residual variation in the parameters; the modeling issue is
this residual variation we can detect with the data. If the
behave like *iid* random variables, then random-coefficient
effective tools for data analysis when there are large nu
types of parameters. The explainable variation is fit by st
structural models, and the unexplainable (not smooth) va
is swept into σ^2.

Conceptually, we still think of a parsimonious struct
on $\underline{\theta}$, but one allowing homoskedastic residuals. For exa
reason to try the constrained model structure $\underline{\theta} = X\underline{\psi}$ (
regressors in matrix X), but we think that this model wou
θ_i even if we could apply it to those exact θ_i. If the model
(relative to the size of components of $\underline{\psi}$), we could safely
interpretation of $\underline{\theta}$ and define any new structural model b
and directly get the likelihood for the new parameters $\underline{\psi}$

$$\mathcal{L}(\underline{\psi}) = \mathcal{L}(\underline{\theta} \mid \underline{\theta} = X\underline{\psi});$$

$\underline{\psi}$ would have only, say, 1 to 4 components.

It is often no more reasonable to assume that $\underline{\theta} = X$
servable $\underline{\theta}$ than it is to assume that $Y = \underline{Z}'\underline{\theta}$ is exact for th

and model $h(\underline{\theta} \mid \underline{\psi})$ imposed on $\underline{\theta}$ in terms of fixed para
conceptualize $\underline{\theta}$ as a normal random variable with $\underline{\theta} = X\underline{\psi}$
variance–covariance matrix of $\underline{\delta}$ as $\sigma^2 I$. This serves to defi
we compute the actual parsimonious reduced model for the

$$g(\underline{x} \mid \underline{\psi}, \sigma^2) = \int g(\underline{x} \mid \underline{\theta}) h(\underline{\theta} \mid \underline{\psi}, \sigma^2) d$$

It is more likely that we will be faced with mixed mod
some random parameters in one or more of our models
explicit we extend our notation by partitioning the generi
into two parts: $\underline{\theta} = (\underline{\alpha}', \underline{\beta}')'$ with $\underline{\alpha}$ fixed and $\underline{\beta} = (\beta_1, \ldots$
of course may choose to consider $\underline{\beta}$ as fixed for some m
will have a model imposed on $\underline{\beta}$ in terms of a distributio
variable, hence $h(\underline{\beta} \mid \underline{\psi}, \sigma^2)$. The needed marginal distrib

$$g(\underline{x} \mid \underline{\alpha}, \underline{\psi}, \sigma^2) = \int g(\underline{x} \mid \underline{\alpha}, \underline{\beta}) h(\underline{\beta} \mid \underline{\psi}, \sigma^2$$

Further generalizations are possible, but for our purpose
the simple case, hence (6.1) and issues of likelihoods for f
models,

$$\mathcal{L}(\underline{\psi}) = \mathcal{L}(\underline{\theta} \mid \underline{\theta} = X\underline{\psi}),$$

versus random-effects reduced models, hence

$$\mathcal{L}(\underline{\psi}, \sigma^2) = g(\underline{x} \mid \underline{\psi}, \sigma^2),$$

where computing (6.3) requires the multidimensional in
by (6.1). The MLE $\hat{\underline{\psi}}$ from these two models (6.2 vs. 6.3
the same value even if $\hat{\sigma}^2 > 0$ occurs ($\hat{\sigma}^2$ might be zero
likelihood functions will differ. Therefore, AIC values f
will be different, and it is important to consider the rand
well as its more restrictive fixed-effects version (wherein
in sets of models fit to data. Also, the variation represent
interest in its own right. This is the case in Section 3.5.6
coefficient model is indirectly fit to real sage grouse surv
the process variation, σ^2, from a set of annual survival ra

One classical basis for the concept and use of random eff
experiments where levels of some factor are selected rand
population, for example animals from a herd, farms in
cultivars from a population of cultivars (see, e.g., Cox and
then a well-defined inference to a real population. Howeve
models need not have this "random selection" feature, esp
involved. The set of K annual survival probabilities, $S_1, \ldots$

modeling is conceptual and refers to the envisioned residua
level smoothing model imposed on the parameters, such a
or even just $S_i = \alpha + \epsilon_i$. If the imposed smoothing (
explainable variation in the survival parameters over time
residuals $\epsilon_i, \ldots, \epsilon_K$ to have the properties of exchangeab
Hence, we can treat them as independent, identically distri
and $\text{var}(\epsilon) = \sigma^2$. It is these residuals, relative to some l
that are the random effects. Therefore, there is no conc
that the years (hence the S_i) be in any way selected at ran
population of years.

In collapsing the problem to $g(\underline{x} \mid \underline{\psi}, \sigma^2)$ we are restric
$\underline{\psi}$ and σ^2, hence ignoring the original individual $\theta_1, \ldots,$
"have our cake and eat it too," that is, get estimators of $\underline{\theta}$
This can be accomplished using shrinkage estimators. S
$\tilde{\underline{\theta}}$ arise in both Bayesian and frequentist theories (see, e.g
1975, Morris 1983, Longford 1993, Casella 1995, and Car
Shrinkage estimators, $\tilde{\underline{\theta}}$, for random-coefficient models s
$\text{E}(\underline{\delta}) = \underline{0}$, $\text{E}(\delta\delta') = \sigma^2 I$ can be based on the MLE of $\hat{\underline{\theta}}$
such a way that the residuals from direct simple linear reg
reproduce the estimate of σ^2 computed in obtaining $\tilde{\underline{\theta}}$. Th
as saying that we may be able to find a suitable proxy
likelihood of the fitted model in (6.3), $\mathcal{L}(\hat{\underline{\psi}}, \hat{\sigma}^2)$, by use
evaluated at such a shrinkage estimator, hence via $\mathcal{L}(\tilde{\underline{\theta}})$.

It would be a considerable advantage if such random
could be fit without ever computing the integral in (6.1
a practical approach to getting a nearly correct AIC v
$g(\underline{x} \mid \underline{\psi}, \sigma^2)$ yet based on $\log(\mathcal{L}(\tilde{\underline{\theta}}))$. As it is, the shrinkage
matic way to fit what amounts to model $g(\underline{x} \mid \underline{\psi}, \sigma^2)$, thus
of $\underline{\theta}$ subject to the stochastic "constraint" inherent in the
model, without making distributional assumptions.

If we do have to compute the integrals as in (6.1), it
using the recent developments from Bayesian methods; se
Smith (1990), Zeger and Karim (1991), and Carlin and C

6.6.2 *A Shrinkage Approach to AIC and Random*

Making inferences about all the random and fixed effects
dom effects model can be accomplished by Bayesian me
1995), empirical Bayes methods (Carlin and Louis 1996), c
age methods (Tibshirani 1996, Royle and Link 2002). Ho
has formal model selection been considered in conjunc

methods (Burnham in review, Burnham and White 2002).
section is to give an overview of one way AIC can be exte
simple random effects based only on the likelihood $\mathcal{L}(\underline{\alpha},$
not straightforward because we need to consider the θ_i as r
in this likelihood they are technically to be considered a
not proposed that this is the ultimate methodology we sho
model selection with random effects. But it is a practica
one factor is a random effect and sample size is not small

Fixed effects inference based on the likelihood $\mathcal{L}(\underline{\alpha}, \underline{\theta}$
focuses on the MLE $\underline{\hat{\theta}}$, which is considered conditional on
dimensional vector. This MLE has conditional sampling
matrix W (it may depend on $\underline{\theta}$) which applies to $\epsilon = \underline{\hat{\theta}} - \underline{\theta}$
g, and hence $\mathcal{L}(\underline{\alpha}, \underline{\theta})$, with a random-effects model wh
$\mathrm{E}(\underline{\delta}) = \underline{0}$, $\mathrm{VC}(\underline{\delta}) = \sigma^2 I$; $\underline{\beta}$ is an r-dimensional vector (r
that an unconditional structural model applicable to the (ot
MLE is

$$\underline{\hat{\theta}} = X\underline{\beta} + \underline{\delta} + \underline{\epsilon}, \qquad \mathrm{VC}(\underline{\delta} + \underline{\epsilon}) = D = \sigma^2 I$$

VC means variance-covariance matrix. Parameters $\underline{\beta}$, as
effects. In practice, $\hat{\mathrm{E}}_{\underline{\theta}}(W) = \hat{W}$ from standard likelihood

From generalized least square theory, for σ^2 known, the
estimator of $\underline{\beta}$ is

$$\underline{\hat{\beta}} = (X' D^{-1} X)^{-1} X' D^{-1} \underline{\hat{\theta}}.$$

Assuming normality of $\underline{\hat{\theta}}$ (approximate normality suffices)
ual sum of squares $(\underline{\hat{\theta}} - X\underline{\hat{\beta}})' D^{-1} (\underline{\hat{\theta}} - X\underline{\hat{\beta}})$ has a central chi-
on $k - r$ degrees of freedom. Therefore, a method of mom
is obtained by solving the equation

$$k - r = (\underline{\hat{\theta}} - X\underline{\hat{\beta}})' D^{-1} (\underline{\hat{\theta}} - X\underline{\hat{\beta}}).$$

Under random effects for inference about $\underline{\theta}$ we use sh
not the MLE (see e.g., Burnham in review). Shrinkage is a
smoothing. Computing the shrinkage estimator requires t

$$H = \sigma D^{-1/2} = \sigma \left(\sigma^2 I + \hat{E}_{\underline{\theta}}(W)\right)^{-1/2} = \left(I + \frac{1}{\sigma^2}\right.$$

evaluated at $\hat{\sigma}$. Then $\underline{\tilde{\theta}} = H(\underline{\hat{\theta}} - X\underline{\hat{\beta}}) + X\underline{\hat{\beta}}$. Move inf
$H + (I - H)AD^{-1}$, where $A = X(X' D^{-1} X)^{-1} X'$. The
matrix such that $\underline{\tilde{\theta}} = G\underline{\hat{\theta}}$. This shrinkage estimator is s
squares of the shrunk residuals (i.e., $\underline{\tilde{\theta}} - X\underline{\hat{\beta}}$), divided b

model from $\underline{\theta}$ using only the fixed effects likelihood eval

Let ℓ be the dimension of $\underline{\alpha}$. Then as a fixed effects m
$\text{AIC} = -2\log\mathcal{L}(\hat{\underline{\alpha}}, \hat{\underline{\theta}}) + 2K$. The random effects log-likelih

$$\log\mathcal{L}(\tilde{\underline{\alpha}}, \tilde{\underline{\theta}}) \equiv \log\mathcal{L}(\hat{\underline{\alpha}}(\tilde{\theta}), \tilde{\underline{\theta}}) = \max_{\underline{\alpha}}[\log\mathcal{L}($$

Reoptimizing over α at fixed $\tilde{\underline{\theta}}$ is necessary. The dimens
space to associate with this random effects model is K_{re},

$$K_{re} = \text{tr}(G) + \ell.$$

We note that $1 \leq \text{tr}(G) \leq k$. This corresponds to the fa
effects model for $\theta_1, \ldots, \theta_k$ is intermediate between a
variation among $\theta_1 = \cdots = \theta_k$ is unstructured and u
no-effects model wherein $\theta_1 = \cdots = \theta_k$.

AIC for the random effects model on $\underline{\theta}$ is $-2\log\mathcal{L}(\tilde{\underline{\alpha}}, \tilde{\underline{\theta}})$
small sample version is

$$\text{QAIC}_c = \frac{-2\log\mathcal{L}(\tilde{\underline{\alpha}}, \tilde{\underline{\theta}})}{\hat{c}} + 2K_{re} + 2\frac{K_{re}(K}{n + K}$$

If it is justified to take $\hat{c} = 1$, then the above becomes AI
of this methodology is given in Burnham and White (20
inference formulae not given here.

Examples are given in Burnham (in review). In partic
recovery example for which $k = 41$ (units are years); θ is
probability. The fixed effects model means fitting 40 m
just a mean (μ) of the θ_i merely to "capture" average a
might be better represented by a single parameter σ^2, whi
own right. The full model also requires 42-band recovery
Results for three models (the r_i are unrestricted) are belo
but justified because sample size was 42,015 birds bande

Model	K	ΔQAIC	Akaike weight	survival
$\theta_{\mu,\sigma}$	73.26	0.00	0.9984	random tim
θ_t	83	12.87	0.0016	fixed-time
θ	43	100.11	0.0000	time-con

Summary points we want to make from this section
generalized to random-effects models, ultimately probabl
given here. Second, the correct measure of parameter dime
effects is the trace of the associated projection (smoothing
of computing K_{re} is not unique). Extensions of AIC to oth
not simple, fixed effects) models often require computing
a matrix.

stood for models that correspond to fixed-effects likeliho
in Section 6.6.2 it is applicable to, and has been developed
effects models. Extensions of AIC-type model selection
are with respect to a best approximating model, not a tru
model inference to other contexts are mostly state-of-the-
successfully, Our point here is that AIC does not dead
effects models. It has a wider spectrum of applications, so
mention here, but without details.

Data smoothing by semi- or non-parametric methods,
area in statistics. Hurvitch et al. (1998) provide versions
ing parameter selection on nonparametric regression (see
2001). For data vector $\underline{y}$ Hurvitch et al. (1998) note tha
correspond to $\hat{\underline{y}} = H\underline{y}$ for a smoothing matrix H that
as a function of a smoothing constant; this constant is no
gle traditional parameter. Moreover, for several AIC-type
selectors "Each of these selectors depends on H through
be interpreted as the effective number of parameters used
..." (Hurvitch et al. 1998:273). Indeed, the role of the
parameters is taken here by $\mathrm{tr}(H)$ and then, generally,
cause there is also one variance parameter, σ^2. There are
for generalized smoothing and generalized cross-validati
alized additive models (Hastie and Tibshirani 1990:49, 15
the trace of a smoother matrix takes the role of the num
structural parameters.

Other general statistical-modeling methodologies to
model selection has been extended include generalized e
(Pan 2001a,b). The method of Pan also serves to extend K-
modeling. Robust regression is sometimes approached
deviations (also called L1 regression, a type of quantile r
and Tsai (1990a) give the small-sample AIC for L1 reg
it as L1cAIC. The formula for L1cAIC is very differen
ever, one of their conclusions was that both AIC_c and L1
Tsai 1990a:263) "... provide good model selections in
a linear regression model with double exponential errors
recommendation that AIC_c is useful in general. Recent
model selection for robust and nonparametric regressio
Shi and Tsai (1998, 1999), Hurvich and Tsai (1998), and
(1999).

A general approach to K-L model selection when the mo
effects remains elusive. However, the recent deviance in
(DIC) approach within a Bayesian framework may pr
(Spiegelhalter et al. 2002). As with AIC the intent of DI

MCMC approach to model fitting. Heuristically, the bas
to minimize the posterior expected value of (relative) Kul
mation loss (see also Thabane and Haq 1999, regarding
general likelihood approach to fitting and selecting among
random effects (without integrating out the random effec
within the framework of h-likelihood (Lee and Nelder 19
Research along this line seems especially worthwile. Son
sults in Section 6.6.2 combined with ideas from the DIC ap
that a likelihood solution may be possible (i.e., AIC for ge
models).

6.7 Selection When Probability Distribut by Model

6.7.1 Keep All the Parts

Most model selection focuses on questions about mode
plainable variation in data, within the context of a single
distribution. For example, if y is the response variable an
predictors x_1 to x_6, we may assume that the structural a
is some functional form for E(y|predictors based on the
transformed with impunity. The (so-called) error distri
the residual $\epsilon = y - \mathrm{E}(y|\text{predictors})$, i.e., unexplained v
Leibler-based model selection allows the error distribution
For example, we can compare models wherein ϵ has a no
models wherein ϵ has a lognormal distribution. However
when so doing: No component part of either probability
safely be dropped in forming the likelihoods. Also, the co
based on y for some models and $\log(y)$ for other mode
2.11.3). We can contrast models for y as normal versus
must be done in a correct way.

Denote the model structural aspects by $\mu_i = \mathrm{E}(y_i|\text{pred}$
depend functionally on some smaller number of paramete
mal probability distribution, variance homogeneity, and
likelihood, as $\prod g_1(y_i|\mu_i, \sigma)$, is

$$\mathcal{L}_1 = \left[\frac{1}{\sqrt{2\pi}}\right]^n \left[\frac{1}{\sigma}\right]^n \exp\left[-\frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \mu$$

As long as every model considered is concerned just w
given the assumed normality of y, then we can drop fr
involving 2π. This is a general principle: If the likelih

and comparisons based on relative expected K-L distanc
siblings) remain valid.

As another example assume that you think that the pro
for y should be a special case of the gamma: $g_2(y|\mu) =$
Then,

$$\mathcal{L}_2 = \left[\prod_{i=1}^{n} y_i\right]\left[\prod_{i=1}^{n} \frac{1}{\mu_i}\right]^2 \exp\left[-\sum_{i=1}^{n} \frac{y_i}{\mu_i}\right.$$

and as long as this is the global likelihood, the leading te
dropped. However, to compare g_1 to g_2 no parts of either

Keep All the Parts to Compare Different Dis

To compare two models g_1 and g_2 that are based on d
distributions (those models may have the same or diff
the μ_i) we must keep in $\mathcal{L}_1$ and $\mathcal{L}_2$ all component par
underlying probability distributions.

If in both (6.4) and (6.5) we dropped the leading term,
resultant AIC_1 to AIC_2 is invalid: it gives meaningless r
this is because we would have confounded real model "effe
arising just because $\log\left[\frac{1}{\sqrt{2\pi}}\right]^n$ and $\log\left[\prod_{i=1}^{n} y_i\right]$ are diff

There is a link here to comparing models on data tra
also cannot be done directly. For example, rather than co
mal distribution models for y versus $\log(y)$, we must co
wherein y can have either a normal or log-normal distrib
to do this is use software for generalized linear models,
recommend. Under GLM (McCullagh and Nelder 1989)
are transformed, rather it is the parametric part of the mod
and linked with different assumed "error" distributions.

6.7.2 A Normal Versus Log-Normal Example

To compare a model based on the normal distribution (g_1
log-normal distribution (g_2) we have to use the log-norma

$$g_2(y|\theta,\sigma) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{[\log(y) -}{\sigma^2}\right.$$

(Again, a GLM approach would be doing this for you
to know the correct probability density form for y bei
likelihood, in general, is now

$$\mathcal{L}_2 = \left[\prod_{i=1}^{n} \frac{1}{y_i}\right]\left[\frac{1}{\sqrt{2\pi}}\right]^n\left[\frac{1}{\sigma}\right]^n \exp\left[-\frac{1}{2}\frac{\sum_{i=1}^{n}[\log(y}{\sigma^2}\right.$$

be compared based on likelihoods (6.4) and (6.6) for nor
distributions, respectively. It will generally make more
compare the same structural models; hence the comparis
with a plausible probability model for the data. (Note,
primary focus of data analysis is the question of a suitab
then the issue of what to use for the error distribution part
major matter, as long as the distribution chosen is not a terr

We provide results of a small Monte Carlo study to selec
log-normal model, based on likelihoods in (6.4) and (6.6)
model wherein the μ_i are constant, and hence $\mu_i \equiv \mu$. The
from either the normal or log-normal model, with the sca
at 1 (results are scale invariant). Because both alternative
issues of using AIC or AIC_c (or even BIC) are moot. Tabl
in terms of selection relative frequency of the normal dis
expected (i.e., average) value of the Akaike weight for
Results in Table 6.14 are accurate to the two decimal plac

This likelihood-based discrimination between the norm
distributions performs well (Table 6.14). The two distri
at small values of $E(y)$ (which actually should be inte
ing $E(y/\sigma)$). However, these distributions are increasin
increases. Correspondingly, the two distributions can be
given sufficient sample size, if they are meaningfully diffe

TABLE 6.14. Results from 10,000 Monte Carlo trials to select betw
and log-normal models, variance$(y) = 1$ for both distributions; π is
wherein the normal model was selected, $E(w)$ is the average of the
normal model, given as a percentage.

					$E(y)$, generating model: n		
	5		10		25		
n	π	$E(w)$	π	$E(w)$	π	$E(w)$	π
10	58	54	53	51	52	50	5
50	77	71	65	57	56	51	5
100	88	83	72	63	59	53	5
500	100	100	92	88	71	62	6

					$E(y)$, generating model: log		
	5		10		25		
n	π	$E(w)$	π	$E(w)$	π	$E(w)$	π
10	38	44	44	49	48	50	4
50	21	29	34	43	43	49	4
100	12	17	27	36	41	47	4
500	0	1	9	13	29	38	3

identical distributions).

The π reflect sampling variation, whereas the E(w) re
certainty. Sampling variation and inferential uncertainty a
different in general, and numerically not identical in Ta
some cases. Notice that the average inferential uncertai
sampling variation. That is, E(w) is always intermediat
$100\pi\%$ in Table 6.14.

It is also worth noting again that these results are equiva
arising from AIC or BIC (because $K = 2$ for both models
entitled (pragmatically) to interpret w_i, for large sample
that model g_i is the K-L best model, but not that it is
further illustrate this last point about inferring truth, we
from the negative exponential distribution $g_3 = \exp(-y/\lambda$
between the normal and log-normal models. The paramete
parameter so results are invariant to its value. Hence, w
model selection results for different sample sizes. As wit
are for selection of the normal model as the best approxi

n	π	E(w)
5	24	32
10	17	22
20	10	12
50	2	3
100	0	0

For n greater than about 100 the result is selection of the
essentially with no inferential uncertainty. Of course, tha
erate the data: Selecting a model with (inferential) certaint
the model is truth.

6.7.3 *Comparing Across Several Distributions: A*

Lindsey and Jones (1998) gives an example based on obs
per cubic millimeter of blood. This is a type of leukocyte c
immune system. The data are from 20 patients in remiss
disease (considered "treatment" here) and 20 patients as
(their Table I, shown here in Table 6.15). The interest is in th
of this white cell count between the two groups of patients.
deviations are $\bar{y}_T = 823$, $\bar{y}_C = 522$, $s_T = 566$, and $s_C = 2$
to use as the basis of parametric inference is problematic a
in favor of distribution-free inference, except that the sa
across patients for this type of count is of interest in its o

Standard practice has been to assume a normal distribut
test, or preferably, a point estimate and a confidence interv

in remission from a non-Hodgkin's disease.

Treatment				Control		
171	397	795	1212	116	375	440
257	431	902	1283	151	375	503
288	435	958	1378	192	377	675
295	554	1004	1621	208	410	688
396	568	1104	2415	315	426	700

TABLE 6.16. AIC results for the T_4 blood cell count data (Table 6
Jones (1998, Table II), "difference" means a treatment effect, henc
more explanation.

	No difference				
Model	K	AIC	Δ	K	
normal	2	608.8	22.8	3	
log-normal	2	590.1	4.1	3	
gamma	2	591.3	5.3	3	
inverse Gaussian	2	590.0	4.0	3	
Poisson	1	11652.0	11066.0	2	10
negative binomial	2	589.2	3.2	3	

the t-distribution. However, one might postulate several p
as the basis of the model and then use AIC to compute a
for the suitability of each model. As long as we compute
based on the complete probability distribution (i.e., no
approach is valid, as noted in Lindsey and Jones (1998). T
by Lindsey and Jones (their Table II) for models he consid
in Table 6.16. The parameters of these models are either
$E(y) = \mu$, or are a dispersion parameter such as for the
Within a distribution there is a pair of models: either $\mu_T =$
or $\mu_T \neq \mu_C$. We will show aspects of calculation of three
6.16.

For the Poisson model when $\mu_T = \mu_C = \mu$ the proba
given by

$$g(y|\mu) = \frac{e^{-\mu}\mu^y}{y!}.$$

Hence the likelihood is to be taken here as (T is the sum

$$\mathcal{L}(\mu) = \prod_{i=1}^{40} \frac{e^{-\mu}\mu^{y_i}}{y_i!} = \left[\prod_{i=1}^{40} \frac{1}{y_i!}\right] \left[e^{-40\mu}\right.$$

$$= \exp(-195150.88)\left[e^{-40\mu}\mu^{26905}\right]$$

The huge term involving factorials could be ignored (di
our only model and inference were just about μ (given t
we would care only about the ratios $\mathcal{L}(\mu)/\mathcal{L}(\hat{\mu})$. But w
parts to make a comparison of this model (i.e., probabil
different probability distribution. The MLE is $\hat{\mu} = 269$
$\log \mathcal{L}(\hat{\mu}) = -5824.9478$ and AIC $= 11651.896$, rounde
6.16.

Consider the normal model for the no-difference cas
use is given by (6.4). The MLEs are $\hat{\mu} = 672.625$ and
$(464.57)^2$; $\log \mathcal{L}$ at the MLEs is given by

$$\log \mathcal{L} = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\sigma}^2) - \frac{n}{2} = -3$$

Hence, here AIC $= -2\log \mathcal{L} + 4 = 608.8$.

For the normal model with a difference by treatment
dispersion parameter σ^2, the likelihood is

$$\mathcal{L} = \left[\frac{1}{\sqrt{2\pi}}\right]^{40}\left[\frac{1}{\sigma}\right]^{40}\exp\left[-\frac{1}{2}\frac{\sum_{i=1}^{20}(y_{T_i} - \mu_T)^2 + \sum}{\sigma^2}\right.$$

The MLEs are $\hat{\mu}_T = \bar{y}_T = 823$, $\hat{\mu}_C = \bar{y}_C = 522$ and $\hat{\sigma}^2 =$
$193151.45 = (439.49)^2$ (SS denotes sum of squares). Th
this model, evaluated at the MLEs, reduces to the same
difference model case, but $\hat{\sigma}^2$ is different:

$$\log \mathcal{L} = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\sigma}^2) - \frac{n}{2} = -3$$

AIC $= 606.364$.

To consider the question of the best approximating m
make comparisons that confound structural and stocha
nents. So in Table 6.16 we must compare within mode
columns). The correct comparison is automatic when we u
by structural model pairs. For example, to compare the no
distributions the two evidence ratios are

$$11499 = \frac{e^{-4.1/2}}{e^{-22.8/2}}, \quad \text{forcing } \mu_T = \mu_C,$$

$$7332 = \frac{e^{-2.6/2}}{e^{-20.4/2}}, \quad \text{allowing } \mu_T \text{ and } \mu_C \text{ to}$$

Either one of these ratios is overwhelming evidence agai
bution, relative to the log-normal, as a best approximating

In practice, to address the distribution issue we would
the difference-allowed cases. There it is clear that the be
binomial, but the log-normal (evidence ratio in favor of th
is 3.7), gamma (2.7), and inverse Gaussian (3.0) are pl
Akaike weights for this subset of four models (ordered a
0.14, 0.18, 0.17, 0.51.

A final observation here is that inference about the "tre
much affected by choice of model, for models that fit wel
is a terrible fit). For the other model pairs, evidence ratios
are

model	ER for an effect
normal	3.3
log-normal	2.1
gamma	2.7
inverse Gaussian	2.5
negative binomial	5.0

How would we analyze these data for an effect? First, w
the no-effect models. Second, we would reparametrize e
from μ_T and μ_C to $\delta = \mu_T - \mu_C$ and $\gamma = \mu_T + \mu_C$. Thi
each reparametrized model leads to $\hat{\delta}$, $\widehat{\text{se}}(\hat{\delta}|g)$, and Akaike
the six models (those w_i are 0.000, 0.138, 0.186, 0.169, 0
we would base inference about δ on model averaging (no
requires substantially more data).

6.8 Lessons from the Literature and Othe

6.8.1 Use AIC$_c$, Not AIC, with Small Sample Siz

It is far too common that papers examining AIC, by itself
fail to use AIC$_c$ when the latter must be used because the nu
at least for some models considered, is not small relative
example, Chatfield (1996) considered model selection iss
series example with $n = 132$ and $R = 12$ a priori designate
ranged from 6 to 61. Overall we commend the paper; howe
example the conclusion that AIC performed poorly is mis
poorly; but it is well known, documented, and commente
on K-L–based model selection that in such an example it
AIC$_c$, not AIC (e.g., Sakamota et al. 1986, Bozdogan 198
1989, Hurvich et al. 1990).

$$n - K - 1$$

Consider for $n = 132$ the effect of the bias-correction ter
for these two models with likelihoods denoted by $\mathcal{L}_6$ and

K	AIC	AIC_c
6	$-2\log(\mathcal{L}_6) + 12$	$-2\log(\mathcal{L}_6) + 12.6$
61	$-2\log(\mathcal{L}_{61}) + 122$	$-2\log(\mathcal{L}_{61}) + 230$

The difference here between AIC and AIC_c is huge for K
greatly affect which model is selected.

We present in Table 6.17 the results that Chatfield (199
sented as regards K-L–based model selection versus BIC.
use Δ_i values, not absolute values of these model selec
show for comparison the ΔAIC values implicitly used b
sults in Table 6.17 are based on the results in Table 1 of C
nature of the models need not concern us, so we label th
keep them in the same order as used in Table 1 of Chat
in Table 6.17 for the AIC_c criterion do have here the inte
described in Section 2.6. Those interpretations are not tru
the Δ_i derived here from AIC, because for large K relat
biased an estimator of the expected K-L distance.

As noted by Chatfield, in this example AIC and BIC l
selected models. However, AIC_c (which must be used her
acceptable results. In fact, the Akaike weights here (see T
only four fitted models have any plausibility in this set

TABLE 6.17. The ΔAIC_c that must be used for K-L model sele
considered in Table 1 of Chatfield (1996), and corresponding Δ
also, the Akaike weights based on AIC_c.

Model	K	ΔAIC_c	ΔAIC	ΔBIC	w_i
1	6	4.5	68.0	**0.0**	0
2	11	**0.0**	62.0	13.0	0
3	21	2.5	58.3	47.0	0
4	9	86.1	148.8	92.2	0
5	17	98.6	157.4	130.9	0
6	41	155.5	181.4	246.0	0
7	11	83.7	145.7	96.7	0
8	21	94.7	150.5	139.3	0
9	13	0.5	61.6	20.0	0
10	25	10.1	62.0	65.7	0
11	31	15.9	60.3	86.8	0
12	61	43.9	**0.0**	139.8	0

definitely eliminates 8 of the 12 models, because relativ
the set, they are extremely implausible. Also, most of the w
put on models with low K. Based on comparing AIC to B
concluded that BIC was a better criterion than AIC. Th
justified in that here one must compare AIC_c to BIC for a
of K-L information-theoretic model selection versus BIC

6.8.2 *Use AIC_c, Not AIC, When K Is Large*

Leirs et al. (1997) report the analysis of an extensive set o
data from Tanzania on the rat *Mastomys natalensis*. Th
data analysis was to examine factors potentially importa
survival dynamics of the species. The data were collect
1986 and February 1989 by live trapping on a 1 ha grid
positions (several traps per position). There were three c
trapping each month (hence 29 primary trapping periods
of 6,728 captures of 2,481 individual animals. We take
size to be the latter, i.e., $n = 2{,}481$. Leirs et al. (1997)
six a priori models to represent how environmental (rain
density factors might affect survival probabilities (S) an
subadults maturing to adults (ψ). Capture probabilities (p
of parameters in these models. Data analysis was by ML m
capture–recapture models (see, e.g., Brownie et al. 1993,
1995) with incorporation of covariates for rainfall and pop
goodness-of-fit of the global model was quite acceptable

The global model used by Leirs et al. (1997) allows ful
poral variation in all model parameters (hence S and p v
and ψ varies by time). In their Table 1, this is model g_1 w
eters. Their model g_2 is the most restricted model: no tem
parameters (S and p vary by age only, subadult versus ad
one maturation probability parameter, ψ). Model 3 allow
ties to vary by time, but S and ψ are not time-varying. Mo
structured time variation in the three classes of parameter
complex, models are based on population dynamics mod
general capture–recapture model. Model 4 has temporal
as functions only of population density (internal factors
regulation); Model 5 has functions only of rainfall (i.e.,
population regulation); Model 6 has temporal paramete
tions of both population density and rainfall. Leirs et al. (
AIC_c) for model selection. In Table 6.18 we present the
their analyses, as well as results for AIC_c, which we co
weights.

Model	K	AIC results Δ_i	w_i	AI Δ
1 Global	113	0.0	0.99	0
2 No effects	5	540.7	0.00	529
3 No dynamics	49	207.4	0.00	198
4 Density effects	52	205.8	0.00	197
5 Rainfall effects	55	25.9	0.00	17
6 Rainfall and Density	64	9.7	0.01	2

Even though the sample size is large here (2,481), t
model with 113 parameters means that AIC_c should be us
be used for all models). The term added to AIC to get
$(2 \times 113 \times 114)/2367 = 10.9$, is not trivial. Clearly, using
a different interpretation of the relative evidence for model
using AIC, Leirs et al. (1997) opted to select model g_6 (
model and therefore to infer that there were population dyn
could be substantially explained only by both external (r
(population density) factors. They worried some about
Nichols, personal communication); they did not need to.
should be done here, model g_6 is a tenable model.

6.8.3 *When Is AIC_c Suitable: A Gamma Distribu*

The K-L approach to model selection is exact, philosophic
sample size (i.e., it is not intrinsically asymptotic). Howe
tion in the face of truth being unknown means some degr
to the target model selection criterion, usually large-samp
get a practical estimator of this criterion. The simplest
asymptotic results, which produces TIC; AIC is a practic
implementation of TIC. Simulation studies and experien
these "large sample" formulas will perform very poorly
or when n is small. Useful insights to small sample ver
selection are obtained by deriving exact versions of the t
various assumed-true models (Section 7.4). The results h
pressible as AIC+ one or more terms of the form $h(K)/(n$
functions $h(\cdot)$ and $m(\cdot)$.

The term added to AIC to get AIC_c is just $2K(K+1)/$
not unique, AIC_c is especially compelling as an omnibus
of AIC because essentially it requires only that the lik
proportional to a normal distribution. This will be a good
at quite small sample sizes if the sample elements are (

needed adjustment term to AIC for the gamma distribut
$2K(K+1)/(n-K-1)$. We strongly recommend using a
of AIC when K is not large relative to n; use AIC_c unl
known.

Theoretically, no small sample adjustment term is neede
for example regression when the residual variation, σ^2, i
estimated). However, performance of model selection wi
in this case by use of AIC_c. The added term basically pre
from reaching and exceeding n, which must somehow be

The theoretical bias of AIC (Section 7.2) is given by

$$2\left(\mathrm{E}_x \log[\mathcal{L}(\hat{\theta}(x)|x)] - \mathrm{E}_x\mathrm{E}_y \log[\mathcal{L}(\hat{\theta}(x)|y)]\right)$$

For the normal distribution model as truth, and with $g =$
is

$$\frac{2K(K+1)}{n-K-1}.$$

We evaluated a simpler form of (6.7) for $f = g$ under th
Monte Carlo methods to compare those results to (6.8). Tal
of these Monte Carlo evaluations; however, we first need
details of what was done.

The gamma model probability density function is

$$g(x|\alpha,\beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, \qquad 0 < x, 0 < \alpha,$$

Here $\mathrm{E}(x) = \alpha\beta$ and $\mathrm{var}(x) = \alpha\beta^2$. Because β is a scale p
invariant to the value of β. Thus it sufficed to just set $\beta =$
(we still must estimate β from the data).

The gamma distribution is in the exponential family. The
7.5 a simpler alternative to (6.7) is

$$2\mathrm{tr}[\mathrm{COV}(\hat{\theta}, S)] - 2K,$$

where $\mathrm{COV}(\hat{\theta}, S)$ is here a 2×2 matrix and θ and S are
The canonical form for the gamma distribution is

$$g(x) = \exp\left[x\left[-\frac{1}{\beta}\right] + \log(x)(\alpha) - \log(x) - \log(\Gamma(\alpha\right.$$

Hence, $\theta_1 = -1/\beta$ and $\theta_2 = \alpha$ are a 1-1 transformation. Th
statistic, S (in 6.9), has the components $\sum x$ and $\sum \log(x$

For given α, a random sample of size n was generat
distribution, and then S was computed and $\hat{\theta}$ found by
methods. This was repeated for 100,000 independent tria

	Sample Size, n			
α	10	20	50	100
0.25	2.78	1.06	0.34	0.16
0.50	2.03	0.80	0.28	0.14
0.75	1.93	0.77	0.28	0.13
1.00	1.83	0.76	0.25	0.13
5.00	1.74	0.66	0.26	0.14
10.00	1.73	0.71	0.27	0.13
25.00	1.68	0.69	0.28	0.14
50.00	1.71	0.69	0.24	0.13
100.00	1.75	0.71	0.27	0.13
From (6.8)	1.71	0.71	0.26	0.12

was found, and finally (6.9) was computed. Results, and l
α, are shown in Table 6.19, along with the value of (6.8) fo
value based on (6.9) is reliable to (almost) two decimal pl
from Table 6.19 that AIC_c is very adequate for use with the
(at $K = 2$), except for when both α and n are both smal
$\alpha = 0.25$ and $n = 10$ is quite extreme in that the underly
is very skewed, hence the likelihood is not near to having a
form at sample size 10. However, even for $n = 10$ the ap
good at $\alpha = 1$, which corresponds to the negative expone

More research is desirable on the issue of small-sample
on the general suitability of AIC_c. However, this example
have examined support AIC_c as generally suitable unless
ability distribution (for a single sample) is extremely nonn
terms of being strongly skewed.

6.8.4 *Inference from a Less Than Best Model*

We continue with some ideas, exemplified by the exampl
Leirs et al. (1997), about inference from other than the K-L
circumstances this is justified, especially if (1) the model
notation for the second-best AIC model), used for infere
the best model, $g_{(1)}$; and (2) the unexplained "effects" in
by the additional parameters added to model $g_{(2)}$ to gen
small relative to the explained effects represented by mod
this assumes a parametrization of the models as $g(\underline{x} \mid \underline{\theta}$
with model $g_{(2)}$ arising under the imposed constraint $\underline{\theta}_1 =$

We elaborate these ideas further using the Leirs et al. (1
AIC_c best model (g_1) was not interpretable in its entirety, l
model (g_6) was interpretable, and because $\Delta_6 = 2.3$, that
model for the data. Moreover, that second-best model is n

additive to (and preferably orthogonal to) the 64 parame
best fitted model. Therefore, their results provide overwhe
set of models used) for the joint importance of rainfall an
as at least good predictors of the observed population
and capture probabilities (if not outright support for a
variables).

In choosing to make inferences based on model $g_{(2)}$
ignoring model $g_{(1)}$ (their model 1), Leirs et al. (1997)
that they cannot interpret the meaning of the additional
constitute the difference between their best and second-be
not in any way invalidate inference from the second-best m
where $g_{(2)}$ is nested within $g_{(1)}$. This sort of argument ho
models are nested.

The only pressing concern here, in ignoring the best
the 49 "effects" defining the difference here between the
models, is the issue of the relative magnitude of the tw
analysis of variance terms this issue is about the partition
of effects represented by the difference in their fitted mo
a sum of squares for effects of g_6 versus g_2 plus a sum o
of g_1 versus g_6. Analogous to ANOVA, we can use here
(ANODEV) (see, e.g., McCullagh and Nelder 1989, Sk
accomplish a useful partition.

In this example, ANODEV proceeds as follows to
importance of the ignored effects left unexplained in m
explained effects in model g_6. First, some baseline "n
needed; here that baseline is model g_2 of Leirs et al. (199
$g_2 \subset g_6 \subset g_1$ and corresponding values of K: 5, 64, and
proceeds by obtaining the log-likelihood values and con
of total deviance of model g_2 versus g_1 as

$$\left[2\log(\mathcal{L}(\hat{\underline{\theta}} \mid g_1) - 2\log(\mathcal{L}(\hat{\underline{\theta}} \mid g_2)\right] = \left[2\log(\mathcal{L}(\hat{\underline{\theta}} \mid g_6) - 2\right.$$

$$+ \left[2\log(\mathcal{L}(\hat{\underline{\theta}} \mid g_1)\right.$$

The result here is $756.8 = 649.0 + 107.8$. The above three b
are also interpretable as likelihood ratio test statistics on

The above partitions a measure of the magnitude of the
on 108 df) represented by fitted model g_1 into a measure of
by model g_6 alone (649.0, on 59 df), plus the additiona
(107.8, on 49 df) explained by the added 49 parameters
g_1 from model g_6. Based on this partition we can defin

$$R^2 = \frac{}{2\log(\mathcal{L}(\hat{\underline{\theta}} \mid g_1)) - 2\log(\mathcal{L}(\hat{\underline{\theta}} \mid g_2))} = \frac{}{756.}$$

The interpretation is that 86% of the total structural inform
ter variation in model g_1 is contained in model g_6. Thus, i
potentially interpretable effects has been lost by making i
on model g_6 (i.e., the second-best AIC_c model), rather tha
However, that other 14% of information was left as not
judged to be real information, as evidenced by AIC_c sele
the best model, but ignoring it does not invalidate the in
model g_6.

Clearly, the addition to model g_6 of all the structure
additional 49 parameters (to get model g_1) does, for the da
K-L best-fitted model. However, in principle there is some
between models g_6 and g_1, that adds far fewer than 49 pa
produce an even smaller AIC_c than model g_1. Such an add
extract additional useful information from the data; it mi
random-effects model, or some interaction effect of rai
density. The situation faced here is, essentially, considere
3.5.6, and 6.9.3, where we point out that if there are two
in the other and differing by a large number of paramete
then anomalies can arise in data analysis based on K-L m

In general, there are situations where choosing to ma
on other than the AIC_c best model can be justified. Howe
not satisfied if the AIC_c best model has many additional p
to the model one uses for the basis of inference. If we fi
situation, it suggests that we did not think hard enough a
of models, because we probably left out at least one goo
a posteriori (to the initial data analysis) model building
done; just admit, then, which models were a priori and w
by initial data analyses.

6.8.5 Are Parameters Real?

Consideration of what is a parameter seems important,
focused entirely on parametric models. With only one cl
regard a parameter as a hypothetical construct. Hence, a pa
embodiment of a concept and does not have the reality of
variable. As such, a parameter in a statistical setting is (u
virtually essential, conceptual abstraction based on the f
of the expected value of a measurable variable that is n
There also needs to be a large number of actual occurren
measurable variable, or at least a well-defined concept
number of occurrences. Then the concept of an average

one, or a few, simple measurements. There is no instrumen
to record the exact value of a parameter used in a sta
exception occurs in measurement error models where the
real but becomes the parameter of interest because each re
is recognized to be imprecise at a nonignorable level of i

We go a step further and recognize two classes of para
models: (1) parameters that appear in the log-likelihood; t
have any associated physical or biological reality; and (2)
above that are directly related to expectations of measurabl
variables. The second class of parameters are tied to me
need not appear in the likelihood (they often do appear).

As an example, consider the analysis of cohort surviv
resented by examples in Section 5.2. The age-specific
parameters S_r cannot be directly measured (such as the
can be). However, the concept represented by S_r has clear
measurable event: survival of an animal over a defined tim
can be repeated based on a sample of animals (from a larg
infinite, population of animals). These survival probabili
the second class of parameters above. To provide both a u
of a set of age-specific survival probabilities, $\{S_r\}$, and p
parsimonious estimation of this set of parameters from l
cal science adopts smooth, deterministic parametric mat
("models" for short) such as

$$S_r = \frac{1}{1 + \exp[-(\theta_1 + \theta_2 \cdot r + \theta_3 \cdot r^2)}$$

(as emphasized in this book, we should not pretend that
holds). The parameters θ_1, θ_2, and θ_3 appear in the likeli
These parameters are in our first class of parameters ab
not have any direct physical or biological reality. In thi
very useful in making parsimonious predictions of the S_r
derived parameters based on the interpretable and parsi
model. Often, interpretability is as important as parsimon
that the two criteria of model usefulness are compleme
conflict. (Interpretability is a subject-matter criterion, no
we have not focused on it here).

The relationship of a parameter to prediction and exp
themselves concepts) is straightforward in a simple linea

$$E(y \mid x) = \beta_0 + \beta_1 x.$$

If we can measure the values of y when separately x and
be able to control x), then

$$\beta_1 = E(y \mid x + 1) - E(y \mid x).$$

y can be discovered by direct measurement. As Mayr (1
are often the driving force in science, much more so tha
ies. The concept of parametric models in statistical scien
powerful force.

6.8.6 *Sample Size Is Often Not a Simple Issue*

Students are introduced to statistical data using the concep
n of a single response variable, y, and possibly some exp
x. This simple data structure fails to convey the possible
especially when that structure is not describable by a sin
example, there may be a sample n_s of subjects, and on e
measurements are taken at n_t time points. To then claim tha
$n = n_s \times n_t$ can be misleading; it is only defensible under
inference model. However, if subjects are treated as rando
is one sample size for subjects and a different sample siz
within subjects. This latter case creates difficulties in the p
sample size and thus the computation of AIC_c.

A related example is the Durban storm data (Section 4.
sample size as all 2,474 weeks in which a storm event co
one might argue that the sample size should be the numbe
or maybe n should be taken as 52, the number of weeks
makes some sense because if we knew the weekly storm pr
then $n = 52$ would be correct (we would still use a mode
Results of the Durban storm analysis are slightly differe
and hence use QAIC_c. There is no definitive basis to kno
the "correct" sample size.

For some data structures there may be two (or more) d
In line transect sampling (simple designs) there will be
lines (often k ranges 10 to 30) and a sample of n detecte
lines, with total length L. For modeling the detection func
sample size, whereas k is the relevant sample size for es
variation of encounter rate, n/L. For capture–recapture
n_s for the number of distinct animals captured once or m
potential number of recapture events. There are logical
of either of these as the sample size to use for AIC_c (
choice now). Further thought suggests that the sample si
the survival parameters (under fixed-effects modeling)
could be n_s, while for the capture probability parameter
be n_r. In general, one can envision different sample size
different subsets of the parameters in models. Certainly
parameters are treated as random effects.

issue. Under a Bayesian approach Spiegelhalter et al. (20
model selection statistic called DIC that seems to behave
with complex sample-size structure. A similar approach
under likelihood-based inference.

The issue of sample size can be complex and has imp
use as AIC_c and $QAIC_c$. We do not pursue solutions he
issue as a future research area.

6.8.7 *Judgment Has a Role*

We have seen published statements to the effect that forma
ing automated or mathematical) model selection methods
presumably because they do not allow for judgment and th
an inference based on a single model. Apparently, a prop
ing would want to select a model by some ill-defined app
and possibly do some form of ill-defined multimodel inf
with such thinking, even though we value and respect t
in science and data analyses. To paraphrase the first sen
in Stewart-Oaten (1995), statistical analyses are based o
ous, formal mathematical-statistical methods and judgme
matter knowledge and a (hopefully) deep understanding
limitations of the formal data analysis methods used. The e
ematical ("objective") methods with the admission that j
for data analysis and interpretation of results.

Judgment does play an important role in data analysis.
the decisions on the general approach and the specific a
will use. Also, a great deal of judgment may be required
priori set of models to be considered. However, once thi
is established, then we maintain that a formal and well-g
ically and mathematically) model selection methodology
these R models. The result will be at least an objectively
selected best model, and preferably, model credibility w
full multimodel inference. This level of inferential mode
be achieved by objective, criterion-based methods such a
interval estimation of a parameter given a model surely
argue that inference should be just a matter of judgment.

Given a model, the statisticians insist on exactly this
founded, objective criterion, or approach, that is preci
mathematics and that, for given data, gives the same
all who do the analysis. After various such formal analy
conclusions and recommendations may indeed be the res
ment to the situation. But at least the formal results can b
of objectivity. Just as we insist on a well-founded obje

multimodel inference). There is room for subsequent pr
about a final best model or inference but only after the re
selection methods have been presented and defended as r

The other aspect of the critique of formal-objective (i.e
selection that is sometimes heard is that it forces one to s
This is a valid concern; it is also now mute. Formal m
methods will greatly reduce the need for what are diffi
means to cope with the obvious inadequacies of methods
single best model.

The real issue about the use of judgment is not whet
but rather who is qualified to exercise good judgment,
they) know they are qualified? Good judgment can be ta
However, it takes a lot of training and experience to achi
(as opposed to say medicine), we do not get frequent a
about whether our judgments (decisions) are good ones.

6.9 Tidbits About AIC

The section contains miscellaneous ideas and results
elsewhere, but are worth understanding.

6.9.1 Irrelevance of Between-Sample Variation

Likelihood-based inference, including AIC-based model s
cerned with relative evidence about alternatives, conditiona
For AIC the technical nature of the inference is about co
relative K-L information loss, as bias-adjusted log-likeliho
ferent models; all such comparisons must be based on a
neither meaningful nor valid to compare individual AI
values between different samples. However, people have s
data and noted the sample-to-sample variation in the valu
fitted model and concluded that our guidelines about AI
could not be correct. This "insight" is wrong because su
variation for a single model (rather than for Δ) is totally i
of inference about alternative models given the data. The s
to a randomized complete block (RCB) experiment wher
are estimated only from within-block differences.

This error is common enough that we illustrate it here.
linear model $y_i - \underline{x}_i'\underline{\beta} + \epsilon_i, \epsilon \sim$ normal$(0, \sigma^2)$ where $\underline{\beta}$ h
rameters. For a random sample under this model, conditio

$$-2\log(\mathcal{L}) = n\log\left[2\pi\sigma^2\left(\frac{\quad}{n}\right)\right.\frac{}{(n-p)\sigma}$$

where RSS is the residual sum of squares. As a random v
$\mathrm{RSS}/(n-p)\sigma^2$ is central chi-square on $\nu = n - p$ degr
random variable

$$-2\log(\mathcal{L}) = n\log(\chi^2_\nu) + n\log\left[2\pi\sigma^2\left(\frac{n-p}{n}\right.\right.$$

Thus, ignoring the additive constant, the sample-to-sam
maximized -2 log-likelihood is easy to simulate in this c
variation depends strongly on sample size. For example, t
or 500, and 10,000 Monte Carlo repetitions of this mode
below for $-2\log(\mathcal{L})$:

sample	n	
%-tile	100	500
1	412.6	3,021.2
10	429.3	3,056.1
25	438.9	3,075.3
50	449.0	3,096.5
75	458.7	3,117.7
90	467.8	3,136.9
99	482.5	3,168.5

The "absolute" variation over samples for a given model
than our guidelines, such as a $\Delta \geq 10$ being generally stro
the model with the bigger AIC. But this sampling variatio
model is meaningless and misleading. The only relevant
ence about alternative models are differences over models
The variation of such a difference is not related to, and
from, the (irrelevant) sampling variation of a single AIC
treatment effect in an RCB experiment can be inferred
variation of a given treatment.

A related misconception arises because AIC values ca
above. Sometimes authors and analysts have AIC values
and 5020 for three models under consideration and concl
are a short distance apart and "one model is nearly as goo
This is a poor interpretation and is probably influenced
size that contributes to the fact that AIC values in this c
range. The focus of attention must always be on the differ
the Δ_j and the associated Akaike weights, w_j, and the ra
the models based on these w_j.

relationship between the G-statistic for goodness-of-fit
distance. The G-statistic is usually written as

$$G = 2\sum_{j=1}^{k} O_j \log\left(\frac{O_j}{E_j}\right),$$

where O_j is the observed count and E_j is the expectatio
model. Under mild conditions, G is asymptotically distril
under the null hypothesis that the model is an adequat
data. Such G-statistics are additive, whereas the more t
goodness-of-fit test statistic

$$\text{Pearson} = \sum_{j=1}^{k} \left((O_j - E_j)^2 / E_j\right)$$

is not. The K-L distance for discrete data is written as

$$I(f, g) = \sum_{i=1}^{k} p_i \log\left(\frac{p_i}{\pi_i}\right)$$

and is almost identical in form to the G-statistic.

Given a sample of count data $n_1, \ldots, n_k$ ($n = \sum$
correspond to the observed relative frequencies. Denote th
probabilities under the approximating model by $\hat{\pi}_j(\theta)$; t
the discrete case, we have $0 < p_i < 1$, $0 < \pi_i < 1$, and t
sum to 1, as do their estimators. Then $I(\hat{f}, \hat{g})$ can be rew

$$\sum_{j=1}^{k} (n_j/n) \log\left(\frac{n_j/n}{E_j/n}\right).$$

Now K-L distance between these (estimated) distribution

$$\frac{1}{n}\sum_{j=1}^{k} (n_j) \log\left(\frac{n_j}{E_j}\right),$$

or

$$\frac{1}{n}\sum_{j=1}^{k} O_j \log\left(\frac{O_j}{E_j}\right).$$

Thus, the G-statistic and K-L information differ by a cons
i.e., in this context, $G = 2n \cdot I(\hat{f}, \hat{g})$. Similar relationship
information expectations of likelihood ratio statistics for c
a likelihood ratio test (LRT) for discrete data). Thus, the L
related to the K-L distance.

The use of the information-theoretic criteria in model se
different from that of hypothesis testing, and this is an
understand. These differences can be illustrated by consid
candidate models, each successive model differing by on
g_i is the null model with i parameters, and model g_{i+j} is
$i + j$ parameters. Model i is nested within model $i + j$;
tests (LRT) can be used to compare the null model with a
models g_{i+j}, where $j \geq 1$. Thus, if model g_i has 12 para
g_{i+1} has 13, model g_{i+2} has 14, and so on.

This concept of a set of nested models is useful in il
ferences between AIC versus LRT for model selection. F
AIC value for each of the models is exactly the same; thu
has more support than any other model. Second, in each
hypothesis be model g_i and assume that it is an adequate
Then, we entertain a set of alternative hypotheses, models
hypothesized to offer a "significantly" better explanation
g_i (the null) is tested individually against the $j \geq 1$ alter
set. The first test statistic (g_i versus g_{i+1}) here is assumed
χ^2 with 1 df, while the second test statistic (g_i versus g_{i+2}
distribution with 2 df, and so on. The following relations

$$\begin{aligned} \text{AIC}_i &= -2\log(\mathcal{L}_i) + 2i, \\ \text{AIC}_{i+j} &= -2\log(\mathcal{L}_{i+j}) + 2(i+j), \\ \text{LRT} &= -2\big(\log(\mathcal{L}_i) - \log(\mathcal{L}_{i+j})\big) \text{ with} \end{aligned}$$

Then, in general,

$$\text{LRT} = \text{AIC}_i - \text{AIC}_{i+j} + 2j.$$

Now, for illustration of a point about the difference betwe
model selection, assume

$$\text{AIC}_i \equiv \text{AIC}_{i+j}\,.$$

If this boundary condition were to occur (where K-L
indifferent to the model), then we would have,

$$\text{LRT} = 2j \text{ on } j \text{ degrees of freedom.}$$

Now, a difference of 1 df between g_i and g_{i+1} correspond
with 1 df, and a P-value of 0.157 (Table 6.20). Similarly,
($j = 4$) between g_i and g_{i+4} corresponds to a χ^2 value
of 0.092. If the degrees of freedom is less than about 7 (a
then hypothesis-testing methods support the null model
alternative models ($g_{i+1}, g_{i+2}, g_{i+3}, \ldots$) (Table 6.20). This
with AIC selection, where in this example all the models ar

number of estimable parameters differs by j (after Sakamoto et al.

j	χ^2	P
1	2	0.157
2	4	0.135
3	6	0.112
4	8	0.092
5	10	0.075
6	12	0.062
7	14	0.051
8	16	0.042
9	18	0.035
10	20	0.029
15	30	0.012
20	40	0.005
25	50	0.005
30	60	0.001

Test results change in this scenario when there are mo
tional parameters in the alternative model (Table 6.20). I
(g_i) is rejected with increasing strength since the alternat
creasing number of parameters. For example, the likelih
versus g_{i+10} has 10 df, $\chi^2 = 20$, and $P = 0.029$. More str
versus g_{i+30}, which has 30 df, $\chi^2 = 60$, and $P = 0.001$, e
value is the same for all the models (the null and the vari
these cases (i.e., > 8 parameters difference between the
model), the testing method indicates increasingly strong s
with many parameters and strong rejection of the simpl
Sakamoto 1991 and Sakamoto and Akaike 1978:196 for a
this issue).

More extreme differences between the two approache
letting $\mathrm{AIC}_i = \mathrm{AIC}_{i+j} - x$ for x in the range of about 0 t
to work with the Δ_{i+j} values; then relative to the selected
g_{i+j} is x. If $x = 4$, the choice of model g_i is compe
of nested models, as judged by AIC. For comparison,
$2j - x$. Let $x = 4$ and $j = 20$; then the LRT statistic
$P = 0.0154$. Most would take this P-value as compelling
of model g_{i+j}. Thus, AIC can clearly support the simple m
can clearly support model g_{i+j} with 20 additional parame
this dilemma is entirely a matter of which the model sele
sound theoretical basis: Information criteria based on K-
likelihood ratio testing does not.

and may even believe that AIC "loses power" as the diff
between models increases beyond about 7. [Note: The con
no utility in the information-theoretic approach because it
way.] Akaike (1974) noted, "The use of a fixed level of
comparison of models with various numbers of paramete
it does not take into account the increase of the variabi
when the number of parameters is increased." The α-lev
to sample size and the degrees of freedom if hypothesis t
how used as a basis for model selection (see Akaike 19
However, the α-level is usually kept fixed, regardless of sa
of freedom, in the hypothesis testing approach. This pra
α-level constant corresponds to asymptotically inconsist
pothesis testing. For example, if the null hypothesis is tru
say, 0.05), then even as the degrees of freedom approach ∞
probability of rejecting the null hypothesis, even with near
The inconsistency is that statistical procedures in this si
converge on truth with probability 1 as $n \to \infty$.

6.9.4 *A Subtle Model Selection Bias Issue*

Consider having 10 independent one degree-of-freedom
random variables, denoted as x_1 to x_{10}. Let z be the mi
$\{x_i, i = 1, \ldots, 10\}$. Because of the selection of z as a min
variables, z is not distributed as central chi-square on 1 df.
tically smaller ("biased") compared to a central chi-square
1 df. The selection process induces what may be conside
few authors have expressed concern for a similar sort of s
as a random variable; they are both right, yet mostly wro
selection as such. By right, we mean that if one focuse
K-L-based criterion, T (Section 7.2), to be estimated, the
fied (good) model we can indeed have $\mathrm{E}(\hat{T}) = T$. Thus a
nearly an unbiased estimator of T (which is related to relat
loss) if only one model is considered. However, if we hav
compute $\hat{T}$ (i.e., AIC) for each model and then select the s
a bias by this selection process as regards $\hat{T}$ for the selec

Technically, just assume $\mathrm{E}(\hat{T}_j) = T_j$, where the exp
possible samples. If model j produces the minimum $\hat{T}_r$
and we can define $\mathrm{E}(\hat{T}_j | min = j)$. The selection bias tha
mistakenly worry about occurs because $\mathrm{E}(\hat{T}_j | min = j) <$

However, the situation of AIC is not really analogous to
example for two reasons. First, the same data are used to
(i.e., each $\hat{T}$); this induces a strong positive correlation in t

then the $10\,x_i$ would all be equal, so selecting their minimu
a bias; the selected variable would always be just the origin
random variable.

Second, we do not care about a single AIC; we only care
ΔAIC. Most of any bias induced by selection (i.e., order
will drop out of these differences. In particular, let the
values) smallest to largest be AIC_1 to AIC_{10}. Then we ca
correct (not biased) for models close to each other in
producing on average small Δ, such as expected for Δ
Here, one model was selected as best and the other as
expect that the selection biases are very similar, and hence
of the difference. Moreover, it is highly relevant that the sa
fitting both (all) models; only the models differ, not the
Again, this means potential selection bias in an individual
concept. Rather, the possible selection bias noted herein st
out of the differences, ΔAIC. Because only these differen
possible biasing of a single AIC by selection of the minim
is an almost irrelevant issue. Possible selection bias of th
K-L differences is relevant, but should be quite small for
other in K-L information loss, especially if R is not large

6.9.5 *The Dimensional Unit of AIC*

The dimensional unit of AIC is the unit of $-\log(g(x|\theta))$,
bility distribution, not a pdf (which is for a continuous ra
a proper probability distribution. Thus, g is strictly for a d
able. From a philosophical point of view all random variab
possible values increment by some minimal step size, δ.
a countable number of possible values. For example, we
is continuous, but we can only measure it to some numbe
four digits (and at best to the nearest discrete atom of m
properly be considered a discrete random variable. For co
it as continuous, which allows both the use of models b
probability and all the convenience and power of calculus

The reason for this observation is simply that likeli
rives from probability, not pdf's (see Lindsey 1999b). Th
$g(x|\theta) \equiv \Pr\{\tilde{X} = x|\theta\}$, where $\tilde{X}$ is the random variabl
not correct, we prefer to use the briefer form. However,
issues about the dimensional unit of $-\log(\mathcal{L}(\theta|x)) = -$
In particular, the dimensional unit of x is irrelevant to t
of $\Pr\{\tilde{X} = x|\theta\}$, which is *always* probability. The prob
$\tilde{X} = x$ is invariant to the measurement units used fo

units being negative log-probability. But Claude Shannon
founder of information theory, established that $-\log(\text{prob}$
mathematical representation of information (and K-L is
g as a model of f). The intrinsic unit for $-\log(\mathcal{L})$ is inf
units of AIC are always information.

There does not seem to be an established name for a
Claude Shannon died in February 2001. It has now been p
of information be called the "Shannon." Regardless of the
here is that the units of the data are irrelevant when we c
(which then become the units); and likelihood, properly, is
of data.

More clarification is needed. A continuous random var
as a convenient way of dealing with the situation when
very small relative to the range of $\tilde{X}$, for example, a r
$\delta = 0.0001$. The range could be 0 to infinity, as long a
big values of $\tilde{X}$ drops off sufficiently fast for large x.
the pdf for a (conceptual) continuous random variable,
approximation $\Pr\{x|\theta\} \approx g(x|\theta)\delta$. For sufficiently small
is excellent and justifies using $g(x|\theta)\delta$ for $\mathcal{L}(\theta|x) = \Pr\{$.
drops out of all uses of the likelihood as a measure of rel
parameter values given the model, or about models given

Thus, for models based on the presumption of continuo
$-\log(\Pr\{x|\theta\}) = -\log(g(x|\theta)) - \log(\delta)$ is technically r
suffices to use $-\log(\mathcal{L}) = -\log(g(x|\theta))$ for purposes of
computed, negative log-likelihood has information as its
However, $-\log(g(x|\theta))$ does not have this unit of dimen
missing $\log(\delta)$. This carries over to AIC. Because of such
stants possibly left out of the likelihood (i.e., dropped f
confusion over the inferential dimensional unit of AIC: it is
justified because the only inferential way we use AIC is in
ner, as Δ AIC. The dimensional unit of Δ is information,
additive constants common to, but left out of, each underl

It is easy to lose sight of these deep matters when consi
$-\log \mathcal{L}$ for normal distribution-based models. We should

$$-\log(\mathcal{L}(\hat{\theta}|x, g) = \frac{n}{2}\log\left(\frac{\text{RSS}}{n}\right) + C($$

where RSS is the residual sum of squares for the fitted mo
C depends only on the assumed normality of "errors,"
If all models considered assume normality and variance
$C(g)$ is identical over all R models, and thus it drops o

When we bypass all of this important information about
ally is, we risk confusion about what is the dimensional uni
AIC. For example, for the normal distribution one might e
dimensional unit is log(dimension of x^2). In fact this is
because we are now dealing with units in − log-probabili

6.9.6 AIC and Finite Mixture Models

Finite mixture models (see e.g., McLachlan and Peel 2000
statistical models, however, they are nonstandard in many
ular, the usual likelihood ratio test statistic is not applicab
about parameter values under the null hypothesis being or
parameter space (McLachlan and Peel 2000: 185–186).
caused concerns about the use of AIC for model selection
models are in the model set. People wonder, should the "2
ified, just as the likelihood ratio test for mixtures must be
considered the issue enough to provide a preliminary opin
AIC does not need to be modified for mixture models. H
shallow level), this is because AIC is not a test. At a deep
AIC to mixture models we encounter two issues: (1) usi
for K, in light of failure of parameter estimability when th
on a parameter space boundary, and (2) model redundanc
estimability failure occurs, and must be dealt with (redun
dropped from the model set). Thus, it is an aspect of how A
be modified.

Consider an example of a strictly positive response var
component mixture negative exponential model be usefu
1988). The mixture model is

$$g_3(x) = \pi \left[\frac{1}{\lambda_1} e^{-x/\lambda_1} \right] + (1-\pi) \left[\frac{1}{\lambda_2} e^{-x/} \right.$$

$0 < \pi < 1, 0 < \lambda_i$, and $\lambda_1 \neq \lambda_2$. The model of (6.10) ca

$$g_3(x) = \pi g_1(x) + (1-\pi) g_2(x),$$

where

$$g_i(x) = \frac{1}{\lambda_i} e^{-x/\lambda_i}, \qquad i = 1, 2.$$

The nominal number of parameters in model g_3 is $K =$
g_1, or g_2, $K = 1$. The problem with fitting mixture mod
that the MLE will be on a boundary if either $\hat{\pi} = 0$ (or 1
and then the model is no longer a mixture. In the first ca
estimable. In the second case π is not estimable; $\hat{\pi}$ can ta

At a minimum, model selection here would have two f
equivalent to either g_1 or g_2, i.e., no subscript is neede
model always has $K = 1$. When the likelihood based on
maximized with $\hat{\pi}$, $\hat{\lambda}_1$, and $\hat{\lambda}_2$ not on boundaries, only th
and compute the usual AIC or AIC_c for the mixture mode
boundary, then the fitted mixture model actually collapse
in that case fitted models g_3 and g are redundant (Sectio
selection occurs because the only fitted model one has is

If model g_3 is not estimable, for the data at hand, it is a
an AIC for it as if $K = 3$. Such an AIC would actually
log-likelihood identical to that for model g but would be f
the AIC for model g. However, model selection or mul
meaningless in this situation where the mixture model in f
$\hat{g}_3$ becomes $\hat{g}$.

The ideas here generalize if the mixing models are of
the mixture model has more than two components. In the
have $g_3 = \pi g_1 + (1 - \pi)g_2$, with models g_1 and g_2 bein
(e.g., negative exponential and half-normal) with K_1 and
model g_3 will collapse to either model g_1 or g_2 if $\hat{\pi}$ is on a b
model redundancy occurs, and there really are only mode
the mixture model is estimable does one have three mod
$K_3 = 1 + K_1 + K_2$ for model g_3.

If the same type of basic model is used in a three (o
mixture, the possible complications increase. But the pr
the full mixture model cannot be fit because the MLE of
is on a boundary, then the model set must be adjusted for
redundancy. Also, the true nature of the reduced-compone
be recognized and its correct K determined. For exam
mixture, when fit to the data, might collapse to a 2-compo
may have already been a model in the model set). Thus
models and AIC-model selection entails some traps that

The above does not address AIC and mixture models a
One way to examine theory is to numerically compare th
AIC_c, as properly used for mixture models, to the theoret
estimating (Section 7.2), e.g., $target = -2E_{\hat{\theta}}E_{\underline{x}}[\log(g_3($
are with respect to the actual data-generating distributio
were "on track" we did some Monte Carlo evaluations w
is (6.10) and data are generated under either (6.10) or
were 25, 50, 100, and 200, with 25,000 or 50,000 Monte
sufficed to fix $\lambda_1 = 1$ and vary λ_2 over 1, 5, 10, 15, 20. Th
over 0 to 0.5 by 0.1. The resulting comparisons, as $\delta = |E($
quite good for $n = 100$ and 200, and (to us) acceptable at

the generating distribution was a mixture of a simple
(mean and maximum δ are over the set of values used for

generating distribution	n	δ mean	δ maximum
simple	25	1.63	1.79
	50	1.10	1.31
	100	0.82	0.97
	200	0.55	0.90
mixture	25	0.66	1.32
	50	0.29	0.66
	100	0.24	0.53
	200	0.23	0.57

6.9.7 *Unconditional Variance*

The formula for estimating the unconditional variance c
parameter estimate is a derived result (Section 4.3.2, denc
4.9):

$$\widehat{\text{var}}\left(\hat{\bar{\theta}}\right) = \left[\sum_{i=1}^{R} w_i \sqrt{\widehat{\text{var}}(\hat{\theta}_i|g_i) + \left(\hat{\theta}_i - \hat{\bar{\theta}}\right.}\right.$$

A corresponding formula is given in Section 4.3.2 f
covariance, $\widehat{\text{cov}}(\hat{\bar{\theta}}, \hat{\bar{\tau}})$. However, in obtaining (6.11) we a
wise conditional (on the models) correlation of estimator
models. Thus, just on that ground there might be weak
theoretical exploration of an alternative to (6.11).

A better, but not overwhelming in practive, motivation
underlying (6.11), as for example using linear models an
with predictors x_1 to x_p. Thus, model g_i uses some subset
and $\theta_i = \text{E}(y|\underline{x}, g_i)$ is the appropriate linear combination
parameters under model g_i. Because all the models cons
have (Section 5.3.6)

$$\hat{\bar{\theta}} = \bar{\tilde{\beta}}_0 + x_1\bar{\tilde{\beta}}_1 + \cdots + x_p\bar{\tilde{\beta}}_p = \underline{x}'\bar{\underline{\tilde{\beta}}}$$

(see also Section 4.2.2 regarding $\bar{\tilde{\beta}}_i$). Using (6.11) a
ing formula for unconditional covariances we can obta
variance–covariance matrix for $\bar{\underline{\tilde{\beta}}}$, say $\hat{\Sigma}$; $\hat{\Sigma}$ is very co
$\widehat{\text{var}}\left(\hat{\theta}_i|g_i\right)$, hence (6.11), is simple. An alternative for $\widehat{\text{var}}$
however, because of the nonlinearities involving the weig

One heuristic approach is to note what is used with BIC averaged posterior is a mixture distribution. Similarly pseudo-likelihood can be defined in the K-L framework of each model-specific likelihood. Either motivation lead possible alternative to (6.11) as

$$\widehat{\widehat{\text{var}}}(\hat{\bar{\theta}}) = \sum_{i=1}^{R} w_i \left[\widehat{\text{var}}(\hat{\theta}_i | g_i) + (\hat{\theta}_i - \hat{\bar{\theta}})^2 \right.$$

This formula is linear in the weights. Using the Cauchy–Sc can show that $\widehat{\text{var}}\left(\hat{\bar{\theta}}\right) \leq \widehat{\widehat{\text{var}}}\left(\hat{\bar{\theta}}\right)$, hence (6.12) actually yiel We emphasize that (6.12) is not a derived result in the framework. We have not studied this matter further; how unconditional variance and covariance for AIC are subj more research.

6.9.8 *A Baseline for $w_+(i)$*

The measure $w_+(i)$ of the relative importance of variable Section 4.2.2. We envision it as applied mostly when t variables (thus, $w_+(i)$ for $i = 1, \ldots, p$), and the R mode subsets of variables, such as just main-effect terms like x_i models), or these terms plus interaction-type terms, such p is small). One important point is that the interpretati within both the context of the set of models and predictors the model set or the set of predictors, and $w_+(i)$ can chang that $w_+(i)$ is not expected to be 0, even if x_i has no contex at all. Rather, in this case $\text{E}(w_+(i)) > 0$, regardless of san reason these summed weights give only a relative import

A randomization method can be used to estimate th $w_+(i)$ if x_i has no predictive value. We denote this unkno $w_{0+}(i)$. The data structure is an n by 1 response-variable c the full design matrix X, which is n by $p+1$ if the models (they usually do). Based on this data structure all R mod model AIC is obtained; Akaike weights are obtained, an computed.

To estimate $w_{0+}(i)$ there is one variation on this scena permute the n values of x_i that are in column $i+1$ of other columns unaltered. Then proceed in the usual way w compute what is nominally $w_+(i)$, except what you get is This computation is quite easy, especially if one already ha for the data. A slight alteration in code changes the bootstr to generate a sample in terms of the needed randomly

permutations of predictor variable i, see text for details.

i	variable	$w_+(i)$	$\hat{w}_{0+}(i)$
1	age	0.50	0.31
2	weight	0.93	0.31
3	height	0.31	0.29
4	neck	0.65	0.29
5	chest	0.28	0.29
6	abdomen	1.00	0.30
7	hips	0.45	0.31
8	thigh	0.59	0.31
9	knee	0.29	0.30
10	ankle	0.45	0.31
11	biceps	0.60	0.31
12	forearm	0.83	0.29
13	wrist	0.98	0.31

$x_{i1}, \ldots, x_{in}$. Obviously, the random permutation renders y
on average; the permuted x_i and all x_j, $j \neq i$, are also un

We have tried this methodology; it seems it could be u
drawback. A single permutation sample is not enough; $\hat{w}_0$
from permutation sample to sample. We suggest doing a
more might be needed. Also the random variable $\hat{w}_{0+}(i)$ ca
distribution. Hence, rather than average the resultant sam
we suggest using the sample median as the single best $\hat{w}_0$

An example of estimating baseline values for the $w_+(i)$ i
One hundred independent permutation samples were use
variable i, as described above, and the sample median v
Because of the correlations among these predictors we th
distinctly vary, but they did not do so here. Perhaps the sy
set combined with the randomization process will alway
$w_{0+}(i)$ is the same for all x_i. From Table 6.21 we judg
and in this context of usage the predictors height, chest
knee circumference have essentially zero importance. T
define a measure of absolute variable importance base
$w_+(i) - \hat{w}_{0+}(i)$. Similar ideas appear in Breiman (2001). C
of these methods and ideas is possible and worthwhile.

Another informative idea for a baseline here is to loo
the full set of $w_+(1), \ldots, w_+(p)$ by leaving the X matrix
permuting the elements of vector y. Then fit all models, g
(the reader might think formal null hypothesis test here
intention). If each $w_+(i)$ and $\hat{w}_{0+}(i)$ pair are about the s
no predictability of y by the entire set of x_i. We have seen

e.g., $p = 50$ and $n = 100$.

For the body fat data example we also randomly permute
$\underline{y}$ vector to estimate the $w_{0+}(i)$. This is a much faster appr
the median of the randomization sample was used as th
samples the results varied from 0.30 to 0.34 (mean of 0.3
the results varied from 0.31 to 0.34 (mean of 0.32). For ei
or y, we note that $w_{0+}(i)$ will depend upon p, and in ge
nothing special about 0.31 or 0.32.

We conclude that more research on these ideas seems

6.10 Summary

This chapter is a more in-depth examination of some as
model selection; included are some comparisons to other m
ods. In particular, AIC is contrasted to BIC to better und
both methods. Those results are mostly in Sections 6.3
their importance we will start this summary with referer
and compare AIC and BIC.

The derivation of BIC (Section 6.4.1) can be done wit
that the set of models contains the true model. Thus, n
(or Bayesian, in general) model selection methods require
validity, or use that the true data-generating model is in the
consideration. Moreover, in the commonly assumed BIC
generating model and a fixed model set) as sample size ge
that selection converges with probability 1 to a single mo
valid to infer that the selected model is truth (see e.g., Se

Rather, the model selected by BIC converges to the mo
dimension (i.e., the minimum K) in the subset (of size
all have the identical minimum Kullback–Leibler distance
6.4.2). Denote this model, which BIC selection converges
the value of $\underline{\theta}$ that minimizes K-L information loss $I(f, g$
g_b of models. Model $g_b(\underline{x}|\underline{\theta}_o)$ is the model with both the
dimension K_i and for which $I(f, g_i(\cdot|\underline{\theta}_o))$ is minimized o

In reality with real data we expect the model that min
to be unique in the model set. But in many, if not most, si
of model selection the set of models used includes the da
and has it nested in some overly parametrized models.
being a subset of the R models that all have the same K-L
the generating ("true") model. BIC is then consistent f
that generating model, which has the smallest dimension
Thus the dimension-consistent property of BIC is motiva

analysis the BIC target model cannot validly be inferred to
merely the model with minimum dimension that is nearest
by K-L distance.

Whereas both the context and the target model for B
of sample size, the context for AIC is sample-size specif
is unchanging, under the information–theoretic approach
models is allowed (in fact, assumed) to grow if n increase
an order of magnitude). It is not realistic to let n go to in
the set of candidate models fixed (as in BIC) because sub
means both more information and more factor levels, fac
study. It then follows that more parameters need to be,
estimated. In this spirit the AIC target model g_a depend
is the model for which $E_{\hat{\theta}}[I(f, g_i(\cdot|\hat{\underline{\theta}}))]$ (expectation is
minimized over $i = 1, \ldots, R$.

Thus the AIC target model (the K-L best model) is gen
the BIC target model. Because both g_a and the model
specific it is not logical to compare g_b to the model that g
$n \to \infty$. Furthermore, the model that is g_a at huge sam
different from the realized model g_a at actual n.

Section 6.4.5 provides a milestone result: It shows that
as a Bayesian model selection criterion. The key is to u
with a quite different prior probability distribution on the
Bayesian context BIC assumes a uniform prior distributi
The K-L model prior that yields AIC is proportional to ex
can be generalized for QAIC_c. One consequence of this re
for interpreting the Akaike weights as a posterior probabi
the model set. Hence, $w_i =$ the probability that model g_i is

Another result of Sections 6.3 and 6.4 is a clear unders
tific meaning we must associate with prior and posterior
Saying "p_i is the probability of model g_i" we must be refe
ity that this model is the target model of the model selecti
target models, g_a and g_b noted above, are different for A
neither target is necessarily truth f). Now a Bayesian b
of AIC and BIC is to argue for or against their respecti
in general to understand the implications of those priors.
$p_i = 1/R$. The prior for AIC makes p_i an increasing fun
and a decreasing function of the number of estimable para
Such a prior corresponds to the idea that we expect the nu
we can reliably estimate to depend on n and K_i, and tha
information in the data.

The alternative way to understand and compare select
frequentist terms of their actual performance and expecte

$n = 252$ ($R^2 = 0.191$), the AIC_c selected model has an R^2
A 95% confidence set on the K-L best model includes
Thus it is totally unacceptable to say the best model has a
In particular, it is absurd to interpret the included variab
ones. This example illustrates the point that with high dim
typically be the case that even the best model will have v
support. The extensive body fat example of Section 6.2 h
Section 6.2.8; we direct the reader there rather than repea

A few theoretical simulations were done based on th
in Section 6.2 to examine predictive mean square erro
BIC best-model selection strategy and for model averag
The predictive mean square error (MSE) for AIC_c was
BIC for the best model strategy and 17% lower under mo
model averaging was superior to the traditional best-mo
AIC_c (MSE 15% lower) and BIC (23% lower). Ongoing re
model-averaged inferences are generally superior in all su

The remainder of Chapter 6 (Sections 6.5 to 6.9) has
Section 6.5 shows that overdispersion adjustment can be
far more than one $\hat{c}$. The key idea is to partition the data
overdispersion estimates by these data subsets. Partitioni
year, area, treatment, and so forth. When this is done, the c
the number of different $\hat{c}$ values used. Issues of goodness-
with estimation of overdispersion factors. Therefore, a
these issues is given in Section 6.5.1, including suggestio
no global model. We also note the matter of goodness-o
model, assuming there is a global model and that it fits
is, does the selected model then also fit the data? This
research. Ongoing work (not included here) has shown
(global model fits) BIC can select a model that is in fact a
but AIC virtually never does so. It is not clear it either, bot
operating characteristics should be of any concern.

Almost all model selection literature is only really appl
parameters as fixed effects. However, the range of applic
expanded to random effects. It can also be expanded to oth
ations, such as generalized estimation equations. Section 6
about these matters, especially application of AIC to mod
ple random effects. These extensions are in an early sta
What is important is that AIC can be extended beyond the
ML approaches of this book. An AIC-like Bayesian proc
applied in general to Bayesian hierarchical models.

Another seeming nonstandard application occurs wher
the set based on different probability distributions. For
want to compare models for the data based on a gamma

the standard formula for AIC. That section gives some h
and details.

In Section 6.8.3 we use numerical methods to evalua
sample bias of AIC when the data are from a gamma distri
this result to the AIC bias-correction term used to get A
by assuming a normal distribution). The result is basically
ally an adequate small sample version of K-L model sele
distribution for the situation studied ($n \geq 10$). This is in li
confirming the general usefulness of AIC_c even when the
normal distribution.

This book focuses on the formal (i.e., objective) aspect
and multimodel inference. Whereas we do not offer muc
to apply professional judgments in the course of data an
the substantial role that judgment plays. We therefore rec
try to be clear in their work about what supports the co
from data. What is the quantifiable evidence; upon what
assessment of evidence rely; are the persons making the
to be making these judgments?

In Section 6.9.5 we consider the dimensional unit of A
in the Shannon sense that negative $\log(probability)$ ma
terizes information. Although the data have associated u
those units are lost, in a sense, as soon as one interpret
the data by using a likelihood, which properly has units o
the units for $-\log(\mathcal{L})$ are information, regardless of the
data. It also follows that the dimensional interpretation
loss when using model $\hat{g}_i$ to approximate model $\hat{g}_{min}$ (the
model).

Model selection has been applied to finite mixture mo
model is nonstandard and the likelihood ratio test must
with mixture models. This motivates a concern that theo
for AIC might need changing to apply to selection of fi
Our thought (Section 6.9.6) at this time is that the formu
need to be modified for use with mixture models (heuristi
not a test, such as a LRT is). Instead, the big issue with fi
is model redundancy that arises when a mixture model
data, i.e., the fitted mixture model actually collapses to a
mixture models it is critical to properly deal with this mo
Section 6.9.6).

We did some theoretical evaluation of whether AIC_c a
target for a two-component mixture negative exponential
accounting for model redundancy. It did quite well at n
acceptable even for $n \geq 50$. Research on an improved sma

The relative importance of a predictor variable x_i can be
This number is relative, not absolute, because it will be g
predictor x_i has no predictive value at all in the given cont
for $w_+(i)$ can be estimated by computer-intensive data pe
One can then better judge what predictors or factors are
the data at hand. The method is simple, but computer int
Section 6.9.8.

There is a lot of material in this chapter; some of it we h
here. Many sections probe issues about AIC without fully
thereby suggest additional research areas. A couple of b
follow: The body fat example shows that one can expe
selection uncertainty with all subsets selection applied t
and illustrates ways to deal with this uncertainty. Those
types of multimodel inference. A seminal result is that Al
same extent as BIC and shows that the difference is all i
tion over the model set (i.e., model probabilities). Final
how Bayesian model probabilities must be interpreted for
interpretations are different.

Statistical Theory and Numeri Results

This chapter contains theory and derivations relevant to K formation-theory–based model selection. We have tried chapters of this book readable by a general audience, esp dents in various fields. Hence, we have reserved this chapt material we believe should be made available to statistici biologists. For many, it will suffice to know that this theory encourage researchers, especially if they have some mat training, to read and try to understand the theory given understanding provides a much deeper knowledge of m based model selection in particular, and of some general m also.

The material given here is a combination of our distillati of the existing literature and what we feel are clarificatio the existing theory. In the former case we have not draw from any one source; hence there is no particular referen these derivations. We have not indicated what results m the literature about the estimation of expected K-L inform this is sometimes not clear even to us.

7.1 Useful Preliminaries

The sole purpose of this section is to provide a summary o concepts, and mathematical background needed to produc

in that it establishes much of the notation and conventions
7.2 and beyond in this chapter.

As a model selection criterion, it is clear what AIC is:
for a model with K estimated parameters, $\hat{\underline{\theta}}$ being the M
eters, computed from the data $\underline{x}$, under an assumed mod
However, we need more detailed notation than just $\mathcal{L}(\hat{\underline{\theta}})$, a
we need to alternate between the likelihood and the pdf i
model. Therefore, without loss of generality we take th
$\mathcal{L}(\underline{\theta} \mid \underline{x}) = g(\underline{x} \mid \underline{\theta})$ by simply then interpreting g as a func
instead of using this convention we had constantly switch
$g(\underline{x} \mid \underline{\theta})$ and $\mathcal{L}(\underline{\theta} \mid \underline{x})$, that would be more confusing than
the single notation $g(\underline{x} \mid \underline{\theta})$. This dual usage of the nota
noted; the reader must follow the mathematics with an e
being made at any point.

A second dual usage of notation for the random variable
$\underline{x}$ denotes the data (as a random variable), and sometimes $\underline{x}$
of integration, always with respect to $f(\underline{x})$, under an inte
dimensional space). Because we are dealing with random v
is usually denoted in terms of the statistical expectatio
operator is just an integral. At times we must have both a
and, separately, data, say $\underline{y}$. But the notation for data versu
is arbitrary and sometimes must be switched back and for
It becomes impossible always to use $\underline{x}$ for a variable of i
data; hence, we do not try to do so, and instead we often
even though at other times $\underline{x}$ is an integrand variable and $\underline{y}$
however, the data, no matter how denoted ($\underline{x}$ or $\underline{y}$, or othe
from truth $f(\cdot)$, not from $g(\cdot \mid \underline{\theta})$ (when $f \neq g$); this is a
point.

AIC has been motivated, justified, and derived in a vari
example, Akaike 1973, Sawa 1978, Sugiura 1978, Chov
Shibata 1989, Bozdogan 1987), but these derivations are o
difficult to follow. Here we give a general derivation in son
being rigorous about all required conditions (they are not
do note where approximations are made. The data have
and the general result is justified for "large" n. That is, t
asymptotically as $n \to \infty$. Also, the integrals and expecta
an n-dimensional sample space, although that fact is not f
notation used.

The most general approach to deriving AIC uses the Tay
to second order. An elementary introduction to the Tayl
Peterson (1960) (or any introductory calculus book); a mor
including results for real-valued multivariable functions,

to $\underline{\theta}$ is given below:

$$h(\underline{\theta}) = h(\underline{\theta}_o) + \left[\frac{\partial h(\underline{\theta}_o)}{\partial \underline{\theta}}\right]' [\underline{\theta} - \underline{\theta}_o] + \frac{1}{2}[\underline{\theta} - \underline{\theta}_o]' \left[\frac{\partial^2 h(\underline{\theta}_o)}{\partial \underline{\theta}^2}\right.$$

($\underline{\theta}$ and $\underline{\theta}_o$ are just two different points in the space over wh
Here, Re represents the exact remainder term for the qu
expansion; the exact nature of Re is known (see Aposto
proximations for the error that results from ignoring Re
heuristic value only, we can claim that an approximation
of order

$$O(\|\underline{\theta} - \underline{\theta}_o\|^3).$$

Here, for any vector argument $\underline{z} - \underline{w}$,

$$\|\underline{z} - \underline{w}\| = \sqrt{\sum_{i=1}^{K} (z_i - w_i)^2}$$

denotes the Euclidean distance between the two points in
space. Thus, the order of the approximation error is the cu
distance between $\underline{\theta}$ and $\underline{\theta}_o$. This is quite a simplification
makes the point that the error of approximation is quite s
is small.

The notation $O(x)$ denotes an unspecified (but po
function of the scalar argument x that satisfies the con
approximately equal to cx for small x, where c is a co
goes to 0 at least at a linear rate in x as x gets near 0. In
quadratic approximation to $h(\underline{\theta})$ "near" $\underline{\theta}_o$ is arbitrarily
nearer to $\underline{\theta}_o$ for $h(\cdot)$ a suitably smooth and bounded func

In (7.1) the notation

$$\left[\frac{\partial h(\underline{\theta}_o)}{\partial \underline{\theta}}\right]$$

denotes a $K \times 1$ column vector of the first partial deri
respect to $\theta_1, \ldots, \theta_K$, evaluated at $\underline{\theta} = \underline{\theta}_o$; hence,

$$\left[\frac{\partial h(\underline{\theta}_o)}{\partial \underline{\theta}}\right] = \left[\begin{array}{c} \frac{\partial h(\underline{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial h(\underline{\theta})}{\partial \theta_K} \end{array}\right]_{|\,\underline{\theta}=\underline{\theta}_o} .$$

$$\left[\frac{\partial^2 h(\underline{\theta}_o)}{\partial \underline{\theta}^2}\right] \quad \left[\frac{\partial^2 h(\underline{\theta}_o)}{\partial \theta_i \partial \theta_j}\right]_{|\underline{\theta}=\underline{\theta}_o}$$

denotes the $K \times K$ matrix of second mixed partial deri
respect to $\theta_1, \ldots, \theta_K$, evaluated at $\underline{\theta} = \underline{\theta}_o$. This matrix
Hessian of $h(\underline{\theta})$.

The expansion in (7.1) when terminated at the quadr
approximation to $h(\underline{\theta})$. In this deterministic case, as indic
of approximation is related roughly to the cube of the
between $\underline{\theta}$ and $\underline{\theta}_o$. For a sufficiently small distance, this
approximation. For the cases of interest, $h(\cdot)$ will be a log
based on a probability distribution. One special value o
needed in these expansions is the large-sample (hence ap
value of the MLE $\hat{\underline{\theta}}$; that is, $\mathrm{E}(\hat{\underline{\theta}}) \approx \underline{\theta}_o$ for large n (the e
relation to K-L information will be given below). The ap
often of order $1/n$, denoted by $O(1/n)$. This notation me
approximation in $\mathrm{E}(\hat{\underline{\theta}}) \approx \underline{\theta}_o$ is less than or equal to a cor
sample size for large sample sizes (the constant might ev

Stronger statements about large-sample limits are po
as sample size $n \to \infty$, $\hat{\underline{\theta}} \to \underline{\theta}_o$ with probability 1, a
approximation given by (7.1) is quite good. In this case (

$$h(\hat{\underline{\theta}}) = h(\underline{\theta}_o) + \left[\frac{\partial h(\underline{\theta}_o)}{\partial \underline{\theta}}\right]' [\hat{\underline{\theta}} - \underline{\theta}_o] + \frac{1}{2}[\hat{\underline{\theta}} - \underline{\theta}_o]' \left[\frac{\partial^2 h(\underline{\theta}_o)}{\partial \underline{\theta}^2}\right]$$

Now the error of approximation in (7.3) is stochastic, b
generally on the order of $1/n$ with probability going to
the added "p" notation of the form $O_p(\cdot)$. The exact size
of approximation in expansions like (7.3) is not known
negligible for large sample sizes, subject to mild regular
same type needed to ensure that the MLE is well behave
Lehmann 1983).

In the context of parametric MLE the standard approach
data are generated by one specific member of a family of
of models, denoted here by $g(\underline{x} \mid \underline{\theta})$, is a set of probability
by an unknown parameter that may be estimated by any va
space Θ. By assumption, truth corresponds to one specific
of $\underline{\theta}$, which we could for clarity denote by $\underline{\theta}_o$. One wou
comes from; it simply exists as (unknown) truth. Thus ev
that the known model structure of g is true, there is still a f
of an underlying unknown truth to the problem of inferenc
cannot know, metaphysically, where this truth $\underline{\theta}_o$ comes f

When we acknowledge that g is just a model of truth,
specified, the issue arises as to what unique parameter in Θ

given the concept of a fixed underlying unknown truth, as
of this essential conceptualization of the inference probl
that the data arose from some deep truth, denoted without
f. Now, one cannot usefully ask where truth f comes fro
physical sense that one cannot ask where $\underline{\theta}_o$ comes from u
that $g(\underline{x} \mid \underline{\theta}_o)$ is truth, but we just do not happen to know

Given this essential framework of f as truth (rather than
as truth) we can, and must, ask whether there is a unique
class of models $g(\cdot \mid \underline{\theta})$ that best describes the data. Her
models $g(\cdot \mid \underline{\theta})$, $\underline{\theta} \in \Theta$, is there a unique $\underline{\theta}_o$ that the ML
is this $g(\cdot \mid \underline{\theta}_o)$ a best model in some sense? In fact, th
samples) estimating a unique parameter value that we wi
this parameter value that indexes our target model under
(we will say more on this below).

Approached theoretically, ignoring issues of data and es
proximating model g in the class of models considered, un
K-L information measure, is simply the model that produ
L value over Θ. Hence we look for a unique value of $\underline{\theta}$
denote by $\underline{\theta}_o$, that provides the K-L best approximating
is the solution to the optimization problem

$$\min_{\underline{\theta} \in \Theta} [I(f, g)] = \int f(\underline{x}) \log \left(\frac{f(\underline{x})}{g(\underline{x} \mid \underline{\theta}_o)} \right)$$

Clearly, $g(\underline{x} \mid \underline{\theta}_o)$ is the best model here, and this serves, i
as a target $\underline{\theta}_o$ given f and given the class of models g.
below, the MLE of $\underline{\theta}$ under model g is estimating $\underline{\theta}_o$.

Given the assumed regularity conditions on the model, $\underline{\theta}$
equations

$$\frac{\partial}{\partial \underline{\theta}} \int f(\underline{x}) \log \left(\frac{f(\underline{x})}{g(\underline{x} \mid \underline{\theta}_o)} \right) d\underline{x} = \underline{0}.$$

Rewriting (7.4) using that $\log(a/b) = \log(a) - \log(b)$, w

$$\frac{\partial}{\partial \underline{\theta}} \int f(\underline{x}) \log(f(\underline{x})) - \frac{\partial}{\partial \underline{\theta}} \int f(\underline{x}) \log(g(\underline{x} \mid \underline{\theta}$$

Because $\underline{\theta}$ is not involved in $f(\cdot)$, the first term of the ab
term (ignoring the minus sign) can be written as

$$\int f(\underline{x}) \left[\frac{\partial}{\partial \underline{\theta}} \log(g(\underline{x} \mid \underline{\theta})) \right]_{\mid \underline{\theta} = \underline{\theta}_o} d\underline{x} = \mathrm{E}_f \left[\left[\frac{\partial}{\partial \underline{\theta}} \log(g(\underline{x}$$

$$\mathrm{E}_f\left[\frac{\partial}{\partial \underline{\theta}} \log(g(\underline{x} \mid \underline{\theta}_o))\right] = \underline{0}.$$

The well-known asymptotic consistency property of convergence of means of *iid* random variables allow, (7.5), another interpretation of $\underline{\theta}_o$. If $\underline{x}$ represents an n from pdf $f(\underline{x}) \equiv \prod_{i=1}^{n} f(x_i)$ and we consider the $g(\underline{x} \mid \underline{\theta}) \equiv \prod_{i=1}^{n} g(x_i \mid \underline{\theta}_o)$, then for every n we have the K (expressed as a mean, without loss of generality)

$$\frac{1}{n}\left[\sum_{i=1}^{n} \frac{\partial}{\partial \underline{\theta}} \log(g(x_i \mid \hat{\underline{\theta}}))\right] = \underline{0}.$$

As $n \to \infty$ two limits are approached with probabili convergence). The sequence of MLEs $\hat{\underline{\theta}}(n)$ (adding no MLE as a function of sample size), converges to some has to converge to $\underline{\theta}_o$, because the means on the left-han likelihood equations converge (as n gets large) to their exp must all equal 0. Under suitable regularity conditions, (7.5 the unique value of $\underline{\theta} = \underline{\theta}_o$. Hence, the sequence $\hat{\underline{\theta}}(n)$ m surely to $\underline{\theta}_o$, which is the K-L minimizer (see, e.g., White

Some deep ideas and philosophy are involved in the abo ular, we have the distinction that unknown truth $f(\underline{x})$ im the numerical values on the (often only) conceptual, but $\underline{\theta}$ of interest to us. Yet $f(\underline{x})$ is not a mathematical func model $g(\underline{x} \mid \underline{\theta})$ is a mathematical function of $\underline{\theta}$, because th but interpretable, hence useful to consider, and varies over space. Even if we think that $g(\cdot)$ represents truth, this is single point $\underline{\theta}_o$ in the parameter space (in a frequentist p tics). Hence, in this case we would be saying that $f(\underline{x}) \equiv$ is a single fixed point; thus even in this context $\underline{\theta}$ is not Therefore, in this or any case, $f(\underline{x})$ is not a function of $\underline{\theta}$.

$$\frac{\partial}{\partial \underline{\theta}} \int f(\underline{x}) \log(f(\underline{x})) d\underline{x} = \underline{0}.$$

The derivation of AIC occurs in the context of prob and expectations of functions of random variables. Such types of integrals, but the notation and "machinery" of sta are more convenient to use here than the explicit notation particular aspect of this matter that needs to be noted is changing the order of taking two expectations of the form $\underline{x}$ and $\underline{y}$ denote random variables. The function $h(\cdot, \cdot)$ is a calculus of integrals as linear operators, $\mathrm{E}_{\underline{x}}\mathrm{E}_{\underline{y}}[h(\underline{x}, \underline{y})] =$

Another aspect of preliminaries concerns (7.2). If $h(\cdot)$ i
$\log(g(\underline{x} \mid \underline{\theta}))$, then (7.2) is

$$\left\{ \frac{\partial^2 \log(g(\underline{x} \mid \underline{\theta}))}{\partial \theta_i \partial \theta_j} \right\}_{\mid \underline{\theta} = \underline{\theta}_o},$$

which is related to the Fisher information matrix

$$\mathcal{I}(\underline{\theta}_o) = \mathrm{E}_g \left\{ -\frac{\partial^2 \log(g(\underline{x} \mid \underline{\theta}))}{\partial \theta_i \partial \theta_j} \right\}_{\mid \underline{\theta} = \underline{\theta}_o}$$

(expectation here is with respect to $g(\cdot)$). If $g(\cdot)$ is the true
it is if f is a special case of g, or if $g = f$), then the
covariance matrix Σ of the MLE is (for large samples) Σ
is, $\Sigma = \mathrm{E}(\hat{\underline{\theta}} - \underline{\theta}_o)(\hat{\underline{\theta}} - \underline{\theta}_o)'$ is $[\mathcal{I}(\underline{\theta}_o)]^{-1}$. If g is not the
may be less general than the true model, or otherwise dif
in general we must expect that $\Sigma \neq [\mathcal{I}(\underline{\theta}_o)]^{-1}$. In fact,
take expectations with respect to f, not g. Hence, we def

$$I(\underline{\theta}_o) = \mathrm{E}_f \left\{ -\frac{\partial^2 \log(g(\underline{x} \mid \underline{\theta}))}{\partial \theta_i \partial \theta_j} \right\}_{\mid \underline{\theta} = \underline{\theta}_o}$$

In the case that $f = g$ or f is a special case of g, then and
$\mathcal{I}(\underline{\theta}_o) = I(\underline{\theta}_o)$. We will not generally make this distinctio
to whether the situation allows $\mathcal{I}(\underline{\theta}_o) = I(\underline{\theta}_o)$ or not. It is
however, to be always cognizant of whether the expectatio
$I(\underline{\theta}_o)$ is with respect to f or g.

Additional notation useful here is the empirical, but un

$$\hat{I}(\underline{\theta}_o) = \left\{ -\frac{\partial^2 \log(g(\underline{x} \mid \underline{\theta}))}{\partial \theta_i \partial \theta_j} \right\}_{\mid \underline{\theta} = \underline{\theta}_o}.$$

For simpler notation we will use

$$I(\underline{\theta}_o) = \mathrm{E}_f \left[-\frac{\partial^2 \log(g(\underline{x} \mid \underline{\theta}_o))}{\partial \underline{\theta}^2} \right],$$

which means exactly the same as (7.7), and hence simp
$\hat{I}(\underline{\theta}_o)$ is

$$\hat{I}(\underline{\theta}_o) = -\frac{\partial^2 \log(g(\underline{x} \mid \underline{\theta}_o))}{\partial \underline{\theta}^2}.$$

It is obvious that $\mathrm{E}_f[\hat{I}(\underline{\theta}_o)] = I(\underline{\theta}_o)$. When $\underline{x}$ is a random
$\hat{I}(\underline{\theta}_o)$ converges to $I(\underline{\theta}_o)$ as $n \to \infty$. We can express this

$$\hat{I}(\underline{\theta}_o) = I(\underline{\theta}_o) + Re, \text{ and usually } Re \text{ is } O($$

$$\hat{I}(\underline{\hat{\theta}}) = -\frac{\partial^2 \log(g(\underline{x} \mid \underline{\hat{\theta}}))}{\partial \underline{\theta}^2}.$$

Because $\underline{\hat{\theta}}$ is the MLE under the model $g(\underline{x} \mid \underline{\theta})$, $\underline{\hat{\theta}}$ converg
and hence $\hat{I}(\underline{\hat{\theta}})$ converges to $I(\underline{\theta}_o)$. Thus, $\hat{I}(\underline{\hat{\theta}}) \approx I(\underline{\theta}_o$
approximation is at worst $O(1/\sqrt{n})$, and in most commo
be $O(1/n)$. If we could determine the analytical form (u
of data) of $I(\underline{\theta}_o)$, an alternative estimator would be $I(\underline{\hat{\theta}})$,
at the MLE; $I(\underline{\hat{\theta}})$ is often not the same as $\hat{I}(\underline{\hat{\theta}})$. Note also
used estimator $\mathcal{I}(\underline{\hat{\theta}})$ (i.e., (7.6) evaluated at the MLE) is
as either $I(\underline{\hat{\theta}})$ or $\hat{I}(\underline{\hat{\theta}})$ and may not converge to $I(\underline{\theta}_o)$.

There are two ways to compute the Fisher information m
when $f = g$. This additional material, and more, is nee
given here. Because the model is a probability distributio

$$\int g(\underline{x} \mid \underline{\theta}) d\underline{x} = 1,$$

and therefore (under the same mild regularity conditions

$$\int \frac{\partial g(\underline{x} \mid \underline{\theta})}{\partial \underline{\theta}} d\underline{x} = \underline{0}.$$

Next, we use in the above the result

$$\frac{\partial \log(g(\underline{x} \mid \underline{\theta}))}{\partial \underline{\theta}} = \frac{1}{g(\underline{x} \mid \underline{\theta})} \left[\frac{\partial g(\underline{x} \mid \underline{\theta})}{\partial \underline{\theta}} \right]$$

and hence we get

$$\int g(\underline{x} \mid \underline{\theta}) \left[\frac{\partial}{\partial \underline{\theta}} \log(g(\underline{x} \mid \underline{\theta})) \right] d\underline{x} = \underline{0}$$

Now take the partial derivative vector of (7.9) with resp
this derivation uses the chain rule of differentiation and
algebraic results:

$$\int g(\underline{x} \mid \underline{\theta}) \left[\frac{\partial}{\partial \underline{\theta}} \log(g(\underline{x} \mid \underline{\theta})) \right] \left[\frac{\partial}{\partial \underline{\theta}} \log(g(\underline{x} \mid \underline{\theta})) \right]' d\underline{x}$$
$$+ \int g(\underline{x} \mid \underline{\theta}) \frac{\partial^2 \log(g(\underline{x} \mid$$

(O is a $K \times K$ matrix of zero elements). We can rewrite

$$\mathrm{E}_g \left[\left[\frac{\partial}{\partial \underline{\theta}} \log(g(\underline{x} \mid \underline{\theta})) \right] \left[\frac{\partial}{\partial \underline{\theta}} \log(g(\underline{x} \mid \underline{\theta})) \right]' \right] = \mathrm{E}_g \left[-$$

$$\mathrm{E}_g\left[\left[\frac{\partial}{\partial\underline{\theta}}\log(g(\underline{x}\mid\underline{\theta}))\right]\left[\frac{\partial}{\partial\underline{\theta}}\log(g(\underline{x}\mid\underline{\theta}))\right]\right]$$

We will denote the left-hand side of the above by $\mathcal{J}(\underline{\theta})$; h

$$\mathcal{J}(\underline{\theta}) = \mathrm{E}_g\left[\left[\frac{\partial}{\partial\underline{\theta}}\log(g(\underline{x}\mid\underline{\theta}))\right]\left[\frac{\partial}{\partial\underline{\theta}}\log(g(\underline{x}\mid\underline{\theta})\right.\right.$$

Thus, $\mathcal{I}(\underline{\theta}) = \mathcal{J}(\underline{\theta})$, but the expectations underlying t with respect to $g(\underline{x}\mid\underline{\theta})$, not with respect to to $f(\underline{x})$. Onc is that the inverse Fisher information matrix may not correct conditional variance–covariance matrix of the M misspecified.

What we need more than (7.11) is

$$J(\underline{\theta}) = \mathrm{E}_f\left[\left[\frac{\partial}{\partial\underline{\theta}}\log(g(\underline{x}\mid\underline{\theta}))\right]\left[\frac{\partial}{\partial\underline{\theta}}\log(g(\underline{x}\mid\underline{\theta})\right.\right.$$

We can expect $J(\underline{\theta}) = \mathcal{J}(\underline{\theta})$ only when $f = g$, or f is Although $\mathcal{I}(\underline{\theta}) = \mathcal{J}(\underline{\theta})$, there is no such general equalit $J(\underline{\theta})$ when g is only an approximation to f, hence when between f and g, $I(f, g)$, is > 0. Heuristically, however equalities of the sort $I(\underline{\theta}_o) \approx J(\underline{\theta}_o)$, $\mathcal{I}(\underline{\theta}_o) \approx I(\underline{\theta}_o)$, and $\mathcal{J}$ $I(f, g) \approx 0$, hence when a good approximating model is

There is a large-sample relationship among $I(\underline{\theta}_o)$, $J(\underline{\theta}_o)$ knowing, and perhaps should be used more:

$$I(\underline{\theta}_o)\Sigma = J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1},$$

and hence

$$\Sigma = [I(\underline{\theta}_o)]^{-1}J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1},$$

where Σ is the true large-sample variance–covariance ma derived from model g when f is truth. It suffices to deriv is (7.13) that we will use more directly in deriving AIC.

Expanding the likelihood equations evaluated at $\underline{\theta}_o$ as series about the MLE, we have

$$\frac{\partial}{\partial\underline{\theta}}\log(g(\underline{x}\mid\underline{\theta}_o)) \approx \frac{\partial}{\partial\underline{\theta}}\log(g(\underline{x}\mid\hat{\underline{\theta}})) + \left[\frac{\partial^2\log(g(\underline{x}\right.$$

The MLE satisfies

$$\frac{\partial}{\partial\underline{\theta}}\log(g(\underline{x}\mid\hat{\underline{\theta}})) = \underline{0};$$

$$\frac{\partial}{\partial \underline{\theta}} \log(g(\underline{x} \mid \underline{\theta}_o)) \approx \left[\frac{\partial^2}{\partial \underline{\theta}^2} \right] (\underline{\theta}$$

$$= \hat{I}(\hat{\underline{\theta}})(\hat{\underline{\theta}} - \underline{\theta}_o) \approx I(\underline{\theta}_o)(\hat{\underline{\theta}}$$

From the above we get

$$[I(\underline{\theta}_o)]^{-1} \left[\frac{\partial}{\partial \underline{\theta}} \log(g(\underline{x} \mid \underline{\theta}_o)) \right] \approx (\hat{\underline{\theta}} - \underline{\theta}$$

Transpose (7.15) and use that transposed result along w
derive

$$[I(\underline{\theta}_o)]^{-1} \left[\frac{\partial}{\partial \underline{\theta}} \log(g(\underline{x} \mid \underline{\theta}_o)) \right] \left[\frac{\partial}{\partial \underline{\theta}} \log(g(\underline{x} \mid \underline{\theta}_o)) \right]' [I(\underline{\theta}_o)]^{-}$$

Now take the expectation of the above with respect to $f($

$$[I(\underline{\theta}_o)]^{-1} J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1} \approx \mathrm{E}_f(\hat{\underline{\theta}} - \underline{\theta}_o)(\hat{\underline{\theta}} - \underline{\theta}$$

hence, we have (7.14) as a large-sample result.

The above likelihood-based results under either a true da
or under model misspecification (i.e., truth is f, the mode
the statistical literature. For very rigorous derivations see

To take expectations of the quadratic forms that are in e
we will need to use an equivalent expression of that form

$$\underline{z}' A \underline{z} = \mathrm{tr}\left[A \underline{z} \underline{z}' \right].$$

Here "tr" stands for the matrix trace function, the sum of th
of a square matrix. The trace function is a linear operator;
quadratic is a stochastic variable in $\underline{z}$, its expectation can

$$\mathrm{E}_{\underline{z}}\left[\underline{z}' A \underline{z} \right] = \mathrm{tr}\left[\mathrm{E}_{\underline{z}}\left[A \underline{z} \underline{z}' \right] \right].$$

If A is fixed (or stochastic but independent of $\underline{z}$), then $\mathrm{E}_{\underline{z}}$
If $\underline{z}$ has mean $\underline{0}$ (such as $\underline{z} = \hat{\underline{\theta}} - \mathrm{E}(\hat{\underline{\theta}})$), then $\mathrm{E}_{\underline{z}}[\underline{z}\underline{z}'] =$
covariance matrix of $\underline{z}$; hence then

$$\mathrm{E}_{\underline{z}}\left[\underline{z}' A \underline{z} \right] = \mathrm{tr}\left[A \Sigma \right].$$

If A is stochastic but independent of $\underline{z}$, then we can use

$$\mathrm{E}_A \mathrm{E}_{\underline{z}}\left[\underline{z}' A \underline{z} \right] = \mathrm{tr}\left[\mathrm{E}_A \mathrm{E}_{\underline{z}}\left[A \underline{z} \underline{z}' \right] \right] = \mathrm{tr}\left[\mathrm{E}_A(A)$$

A final aspect of notation, and of concepts, reemphasiz
start of this section: The notation for a random variable
arbitrary in taking expectations over the sample space. W
for such an expectation (i.e., integration) is the model u
its form, its assumptions, its parameters, and the distribu

data-analysis framework. Because an expectation is a ty
defined space, the result of the integration is not depen
used in the integrand. Thus

$$\begin{aligned} \mathrm{E}_f\left[\log(g(\underline{x} \mid \hat{\underline{\theta}}(\underline{x})))\right] &= \int f(\underline{x}) \log(g(\underline{x} \mid \hat{\underline{\theta}}(\\ &\equiv \int f(\underline{y}) \log(g(\underline{y} \mid \hat{\underline{\theta}}(\\ &= \mathrm{E}_f\left[\log(g(\underline{y} \mid \hat{\underline{\theta}}(\underline{y}))\right. \end{aligned}$$

Changing notation for the integrand (i.e., $\underline{x}$ to $\underline{y}$) has no
this type of useful notation change is required in derivat
in places, at a conceptual level, we recognize two indepen
have two notations, $\underline{x}$ and $\underline{y}$. In fact, these derivations ar
quentist properties of data-analysis methods, but there is
being used in these derivations. Rather, in these theore
possible "data" are just points in an n-dimensional samp
accordance with some true probability distribution $f(\cdot)$.

7.2 A General Derivation of AIC

We now give a general conceptual and then mathematic
starting from K-L information for the best approximatin
of models $g(\underline{x} \mid \underline{\theta})$:

$$I(f, g(\cdot \mid \underline{\theta}_o)) = \int f(\underline{x}) \log\left(\frac{f(\underline{x})}{g(\underline{x} \mid \underline{\theta}_o)}\right)$$

Note that while for the model we do not know $\underline{\theta}$, the tar
value for the class of models is appropriately taken as $I($
(i.e., 7.16), because the parameter value we will be estima
the expanded notation in (7.16), so we can represent I
in general, on the unknown parameter value, given the m
$I(f, g)$ does not involve any data, nor any value of $\underline{x}$, since
out.

Given that we have data $\underline{y}$ as a sample from $f(\cdot)$, the l
to find the MLE $\hat{\underline{\theta}} = \hat{\underline{\theta}}(\underline{y})$ and compute an estimate of $I($

$$I(f, g(\cdot \mid \hat{\underline{\theta}}(\underline{y}))) = \int f(\underline{x}) \log\left(\frac{f(\underline{x})}{g(\underline{x} \mid \hat{\underline{\theta}}(\underline{y}))}\right.$$

approaches, both will lead to the same basis for AIC (the
from K-L to AIC).

If we could find the $\underline{\theta}_o$ that minimizes K-L (for a given
that our target for a perfect model would be $I(f, g) = 0$.
how good any model is relative to this absolute value
change when we have only an estimate of $\underline{\theta}$. Even if our
(miraculously) truth, hence $g(\underline{x} \mid \underline{\theta}_o) = f(\underline{x})$, our estim
equal $\underline{\theta}_o$ almost surely for continuous parameters and distr
for some discrete distributions the equality would be wi
Any value of $\hat{\underline{\theta}}(\underline{y})$ other than $\underline{\theta}_o$ results in $I(f, g(\cdot \mid \hat{\underline{\theta}}(\underline{y}$
Thus, even if we had the correct model structure, becaus
$\underline{\theta}$ we should think in terms of the (essentially estimated
average, a value > 0. This motivates us to revise our ide
must be as a measure of perfect agreement of fitted mode

In the context of repeated sampling properties as a
we would expect our estimated K-L to have on averag
$\mathrm{E}_{\underline{y}}\left[I(f, g(\cdot \mid \hat{\underline{\theta}}(\underline{y})))\right]$. We should therefore readjust our i
the model to be not the minimizing of $I(f, g(\cdot \mid \underline{\theta}_o))$ (give
larger value, on average, given by

$$\mathrm{E}_{\underline{y}}\left[I(f, g(\cdot \mid \hat{\underline{\theta}}(\underline{y})))\right] > I(f, g(\cdot \mid \underline{\theta}_o))$$

(and repeating ourselves because it is an important point:
are with respect to f regardless of the notation for random
such as $\underline{x}$, $\underline{y}$, or $\hat{\underline{\theta}}$). Thus, given the reality that we must
adopt the criterion

"select the model g to minimize $\mathrm{E}_{\underline{y}}\left[I(f, g(\cdot \mid \hat{\underline{\theta}}(\underline{y}$

Hence our goal must be to minimize the expected value o
estimated K-L information value. (If we could compute
each model, we could stay with the goal of minimizing
curious we note here that the large-sample difference is $\mathrm{E}_{\underline{y}}$
$I(f, g(\cdot \mid \underline{\theta}_o)) = \frac{1}{2}\,\mathrm{tr}\left[J(\underline{\theta}_o)I(\underline{\theta}_o)^{-1}\right]$, which does not dep
n.)

Rewriting the basis of this new target to be minimized,

$$\mathrm{E}_{\underline{y}}\left[I(f, g(\cdot \mid \hat{\underline{\theta}}(\underline{y})))\right] = \int f(\underline{x}) \log(f(\underline{x}))d\underline{x} - \mathrm{E}_{\underline{y}}\left[\int f(\underline{x}\right.$$

hence

$$\mathrm{E}_{\underline{y}}\left[I(f, g(\cdot \mid \hat{\underline{\theta}}(\underline{y})))\right] = \text{constant} - \mathrm{E}_{\underline{y}}\mathrm{E}_{\underline{x}}\left[\log[g(\underline{x} \mid\right.$$

value given by (7.18). In most of our writing here about t
much simpler just to say that we are selecting an estimat
model by use of AIC.

There is a second, less compelling, approach that we ca
K-L to AIC: Start with

$$I(f, g(\cdot \mid \underline{\theta}_o)) = \text{constant} - \mathrm{E}_{\underline{x}}\left[\log(g(\underline{x} \mid$$

and see whether we can compute (or estimate) $\mathrm{E}_{\underline{x}}\left[\log(g($
Taylor series expansions. As will be made evident belov
result

$$\mathrm{E}_{\underline{x}}\left[\log(g(\underline{x} \mid \hat{\underline{\theta}}(\underline{y})))\right] \approx \mathrm{E}_{\underline{x}}\left[\log(g(\underline{x} \mid \hat{\underline{\theta}}(\underline{x})))\right] - \frac{1}{2}\,\mathrm{tr}$$
$$- \frac{1}{2}(\hat{\underline{\theta}}(\underline{y}) - \underline{\theta}_o)' I(\underline{\theta}_o)(\hat{\underline{\theta}}(\underline{y})$$

On the right-hand side above, the only component that a
estimated or computed (in any useful way) is the quad
$(\hat{\underline{\theta}}(\underline{y}) - \underline{\theta}_o)$ (and it is pointless therein to use $\underline{\hat{\theta}}_o = \hat{\underline{\theta}}(\underline{y})$
expectation of both sides above with respect to $\underline{y}$, we ge
estimate:

$$\mathrm{E}_{\underline{y}}\mathrm{E}_{\underline{x}}\left[\log(g(\underline{x} \mid \hat{\underline{\theta}}(\underline{y})))\right] \approx \mathrm{E}_{\underline{x}}\left[\log(g(\underline{x} \mid \hat{\underline{\theta}}(\underline{x})))\right] - \mathrm{tr}$$

Thus, either line of derivation demonstrates that we
objective from model selection based on minimum K-L v
g, to selecting the model with estimated $\underline{\theta}$ based on min
K-L information measure. It is still the case that only a re
be found based on $\mathrm{E}_{\underline{y}}\mathrm{E}_{\underline{x}}\left[\log(g(\underline{x} \mid \hat{\underline{\theta}}(\underline{y})))\right]$ as the target obj
maximized; the constant $\mathrm{E}_{\underline{x}}[f(\underline{x})\log(f(\underline{x}))]$ cannot be co

Only some of the literature is clear that AIC model s
the concept of minimizing the expected K-L criterion
(see, e.g., Sawa 1978, Sugiura 1978, Bozdogan 1987 (pag
Milhaud 1994). It is the relative value of this criterion that
set of models. That is, we want to estimate without bias, as
criterion (denoted below by T for target) for each appro
value of

$$T = \int f(\underline{y}) \left[\int f(\underline{x}) \log(g(\underline{x} \mid \hat{\underline{\theta}}(\underline{y})))d\underline{x} \right.$$

The change from conceptual model selection based on mi
model selection based on maximizing an estimate of T i

Sometimes the criterion given by (7.19), hence AIC, is i
cept of Akaike's predicative likelihood $\mathrm{E}_p[\log(\mathcal{L}(\hat{\theta}))] = \mathrm{E}$
$\equiv T$, which has a heuristic interpretation in terms of
independent random variables $\underline{x}$ and $\underline{y}$. However, the qu
tion by maximizing $\hat{T}$ (or minimizing $-2\hat{T}$), does aris
approach to the problem of model selection without ever
cross-validation.

In a slightly simplified, but obvious, notation, the K-L–b
problem is now to find a useful expression for, and estima

$$T = \mathrm{E}_{\underline{\hat{\theta}}}\mathrm{E}_{\underline{x}}\left[\log(g(\underline{x} \mid \underline{\hat{\theta}}))\right],$$

where it is understood that the MLE $\underline{\hat{\theta}}$ is based on sa
expectations are for $\underline{x}$ and $\underline{y}$ (hence $\underline{\hat{\theta}}$) both with resp
because T is also a double expectation based, conceptually
samples that AIC-based model selection is asymptotically
validation (see, e.g., Stone 1977); cross-validation is a w
model selection.

Step 1 is an expansion of the form (7.3) applied to log
for any given $\underline{x}$:

$$\log(g(\underline{x} \mid \underline{\hat{\theta}})) \approx \log(g(\underline{x} \mid \underline{\theta}_o)) + \left[\frac{\partial \log(g(\underline{x} \mid \underline{\theta}_o))}{\partial \underline{\theta}}\right.$$
$$+ \frac{1}{2}[\underline{\hat{\theta}} - \underline{\theta}_o]' \left[\frac{\partial^2 \log(g(\underline{x} \mid \underline{\theta}_o))}{\partial \underline{\theta}^2}\right.$$

Truncation at the quadratic term entails an unknown degr
(but it is an error of approximation that goes to zero as
(7.21) to (7.20) we first take the expected value of (7.21)

$$\mathrm{E}_{\underline{x}}\left[\log(g(\underline{x} \mid \underline{\hat{\theta}}))\right] \approx \mathrm{E}_{\underline{x}}\left[\log(g(\underline{x} \mid \underline{\theta}_o))\right] + \mathrm{E}_{\underline{x}}\left[\frac{\partial \log(g($$
$$+ \frac{1}{2}[\underline{\hat{\theta}} - \underline{\theta}_o]' \left[\mathrm{E}_{\underline{x}}\frac{\partial^2 \log(g(\underline{x} \mid \underline{\theta}_o))}{\partial \underline{\theta}^2}\right.$$

The vector multiplier of $[\underline{\hat{\theta}} - \underline{\theta}_o]$ in the linear term above
as (7.5). It is just that for clarification $\mathrm{E}_{\underline{x}}$ is used to mean
of the random variable $\underline{x}$ (and keep remembering that $\underline{\hat{\theta}} \equiv$
of $\underline{x}$). Therefore, upon taking this expectation, the linear
is, (7.5) applies:

$$\mathrm{E}_{\underline{x}}\left[\frac{\partial \log(g(\underline{x} \mid \underline{\theta}_o))}{\partial \underline{\theta}}\right] = \underline{0}.$$

$$\mathrm{E}_{\underline{x}}\left[\log(g(\underline{x}\mid\hat{\underline{\theta}}))\right] \approx \mathrm{E}_{\underline{x}}\left[\log(g(\underline{x}\mid\underline{\theta}_o))\right] - \frac{1}{2}[\hat{\underline{\theta}} - \underline{\theta}_o]' I($$

Now we can take the expectation of (7.23) with respect t
where the trace function is used, yielding

$$\mathrm{E}_{\hat{\underline{\theta}}}\mathrm{E}_{\underline{x}}\left[\log(g(\underline{x}\mid\hat{\underline{\theta}}))\right] \approx \mathrm{E}_{\underline{x}}\left[\log(g(\underline{x}\mid\underline{\theta}_o))\right] - \frac{1}{2}\,\mathrm{tr}\left[I(\underline{\theta}_o)\mathrm{E}_{\hat{\underline{\theta}}}\right.$$

The left-hand side above is T from (7.20), and $\mathrm{E}_{\hat{\underline{\theta}}}\left[[\hat{\underline{\theta}} - \underline{\theta}\right.$
the correct large-sample theoretical sampling variance o
the expectation herein is taken with respect to truth f, n
Thus we have

$$T \approx \mathrm{E}_{\underline{x}}\left[\log(g(\underline{x}\mid\underline{\theta}_o))\right] - \frac{1}{2}\,\mathrm{tr}\left[I(\underline{\theta}_o)\Sigma\right.$$

Step 2 starts with the realization that we have not yet de
a relationship between T and $\mathrm{E}_{\underline{x}}\left[\log[g(\underline{x}\mid\hat{\underline{\theta}}(x))]\right]$, whic
of the actual log-likelihood at the MLE. We now do a se
time of $\log(g(\underline{x}\mid\underline{\theta}_o)$ about $\hat{\underline{\theta}}(x)$, treating $\underline{x}$ as the sample
the MLE of $\underline{\theta}$ for this $\underline{x}$. This procedure is acceptable, bec
is an expected value, which means taking an integral ove
in the sample space. Therefore, it does not matter what
these sample points: $\underline{x}$ or $\underline{y}$. Applying the Taylor series
(but with the roles of $\hat{\underline{\theta}}$ and $\underline{\theta}_o$ switched; also note well t
we obtain

$$\log(g(\underline{x}\mid\underline{\theta}_o)) \approx \log(g(\underline{x}\mid\hat{\underline{\theta}})) + \left[\frac{\partial\log(g(\underline{x}\mid\hat{\underline{\theta}}))}{\partial\underline{\theta}}\right.$$

$$+ \frac{1}{2}[\underline{\theta}_o - \hat{\underline{\theta}}]'\left[\frac{\partial^2\log(g(\underline{x}\mid\hat{\underline{\theta}}))}{\partial\underline{\theta}^2}\right.$$

The MLE $\hat{\underline{\theta}}$ is the solution of, hence satisfies, the equatio

$$\frac{\partial\log(g(\underline{x}\mid\hat{\underline{\theta}}))}{\partial\underline{\theta}} = \underline{0}.$$

Therefore, the linear term in (7.25) vanishes. Taking the
we can write

$$\mathrm{E}_{\underline{x}}\left[\log(g(\underline{x}\mid\underline{\theta}_o))\right] \approx \mathrm{E}_{\underline{x}}\left[\log(g(\underline{x}\mid\hat{\underline{\theta}}))\right] - \frac{1}{2}\,\mathrm{tr}\left[\mathrm{E}_{\underline{x}}\left[\hat{I}(\hat{\underline{\theta}})\right]\right.$$

See (7.8) for $\hat{I}(\hat{\underline{\theta}})$, the Hessian of the log-likelihood evalu

$$\mathrm{E}_{\underline{x}}\left[\hat{I}(\hat{\underline{\theta}})\right][\underline{\theta}_o - \hat{\underline{\theta}}][\underline{\theta}_o - \hat{\underline{\theta}}]' \approx \left[I(\underline{\theta}_o)\right]\left[\mathrm{E}_{\underline{x}}[\underline{\theta}_o - \hat{\underline{\theta}}\right.$$
$$= \left[I(\underline{\theta}_o)\right]\left[\mathrm{E}_{\underline{x}}[\hat{\underline{\theta}} - \underline{\theta}\right.$$
$$= \left[I(\underline{\theta}_o)\right]\Sigma.$$

The approximation made in (7.27) is often good to $O(1/n$ However, there are circumstances where the approximat good, and the overall approximation in (7.27) is equival $I(\underline{\theta}_o)$ after first writing the approximation

$$\mathrm{E}_{\underline{x}}\left[\hat{I}(\hat{\underline{\theta}})\right][\underline{\theta}_o - \hat{\underline{\theta}}][\underline{\theta}_o - \hat{\underline{\theta}}]' \approx \left[\mathrm{E}_{\underline{x}}\left[\hat{I}(\hat{\underline{\theta}})\right]\right]\left[\mathrm{E}_{\underline{x}}[\underline{\theta}_o -\right.$$
$$= \left[\left[I(\underline{\theta}_o)\right]\right]\Sigma$$

to arrive at the same result as (7.27). In any case, (with sample size, but the overall error involved in thi $\mathrm{E}_{\underline{x}}\left[\hat{I}(\hat{\underline{\theta}})\right][\underline{\theta}_o - \hat{\underline{\theta}}][\underline{\theta}_o - \hat{\underline{\theta}}]'$ is hard to assess, in genera visited below for the exponential family of distributions, to be there a good approximation to $O(1/n)$.)

Using either (7.27) or (7.28), along with (7.26), we hav

$$\mathrm{E}_{\underline{x}}\left[\log(g(\underline{x}\,|\,\underline{\theta}_o))\right] \approx \mathrm{E}_{\underline{x}}\left[\log(g(\underline{x}\,|\,\hat{\underline{\theta}}(\underline{x})))\right] - \frac{1}{2}\,\mathrm{tr}\left[I\right.$$

Recall (7.24):

$$T \approx \mathrm{E}_{\underline{x}}\left[\log(g(\underline{x}\,|\,\underline{\theta}_o))\right] - \frac{1}{2}\,\mathrm{tr}\left[I(\underline{\theta}_o)\Sigma\right.$$

Substituting (7.29) into (7.24) we have a key result th literature:

$$T \approx \mathrm{E}_{\underline{x}}\left[\log(g(\underline{x}\,|\,\hat{\underline{\theta}}(\underline{x})))\right] - \mathrm{tr}\left[I(\underline{\theta}_o)\Sigma\right.$$

The literature usually presents not (7.30), but rather an a form based on (7.13):

$$T \approx \mathrm{E}_{\underline{x}}\left[\log(g(\underline{x}\,|\,\hat{\underline{\theta}}(\underline{x})))\right] - \mathrm{tr}\left[J(\underline{\theta}_o)[I(\underline{\theta}_o\right.$$

The notation $\hat{\underline{\theta}}(\underline{x})$ rather than just $\hat{\underline{\theta}}$ is used above onl on the right-hand side of (7.31) only one random variab can be taken to refer to the actual data. From (7.30) or that a criterion for model selection (i.e., a nearly unbiase structurally of the form

$$\hat{T} \approx \log(g(\underline{x}\,|\,\hat{\underline{\theta}})) - \widehat{\mathrm{tr}}\left[I(\underline{\theta}_o)\Sigma\right],$$

Simple, direct estimation of Σ from one sample is not po
is only one $\hat{\underline{\theta}}$ available (a bootstrap estimator of Σ is pos
$J(\underline{\theta}_o)$ and $I(\underline{\theta}_o)$ are directly estimable from the single s
(7.33), but not (7.32), requires a parametrization wherein
whence its inverse exists. There is no loss in generality if w
probability distribution models have fully identifiable pa
are of full rank.

The maximized log-likelihood $\log(g(\underline{x} \mid \hat{\underline{\theta}}))$ in (7.31) is
tor of its own expectation $\mathrm{E}_{\underline{x}}[\log(g(\underline{x} \mid \hat{\underline{\theta}}))]$ (but is biased a
Hence, the only problem left is to get a reliable (low, or n
the trace term, or at least an estimator with small mean sq
best model to use is the one with the largest value of $\hat{T}$, bec
duce a model with the smallest estimated expected K-L d
of convention the criterion is often stated as that of minim

$$-2\log(g(\underline{x} \mid \hat{\underline{\theta}})) + 2\,\widehat{\mathrm{tr}}\left[J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1}\right.$$

If f is a subset of g (i.e., if $g = f$ or f is contained w
of nested models), then $I(\underline{\theta}_o) \equiv \mathcal{I}(\underline{\theta}_o) = \mathcal{J}(\underline{\theta}_o) = J(\underline{\theta}_o$
$\mathrm{tr}\left[I(\underline{\theta}_o)\Sigma\right] = K$. Even if g is just a good model (i.e., a g
for f, the literature supports the idea that our best estimat
$\widehat{\mathrm{tr}}\left[I(\underline{\theta}_o)\Sigma\right] = K$ (Shibata 1989).

When the model is too restrictive to be good, the term
be much inflated (compared to this same term for a "go
will not select that model. In this case having a good e
term should not matter. The practical key to making AI
assumed $\widehat{\mathrm{tr}}\left[I(\underline{\theta}_o)\Sigma\right] = K$) work is then to have some go
considered, but not too many good, but over parametrized,
model we mean one that is close to f in the sense of havin
in which case such "closeness" also means that the use
is itself a parsimonious estimator. This matter of estimati
and closeness of g to f is explored further in Section 7.
7.6 derivations, and the above ideas in this paragraph, tha
which is seen as a special case of (7.34):

$$\mathrm{AIC} = -2\log(g(\underline{x} \mid \hat{\underline{\theta}})) + 2K.$$

The generalization given by (7.34) leads to Takeuchi's
criterion (TIC) for model selection (Shibata 1989). The r
that we might use the bootstrap to compute $\widehat{\mathrm{tr}}\left[I(\underline{\theta}_o)\Sigma\right]$ a
the TIC criterion via

$$-2\log(g(\underline{x} \mid \hat{\underline{\theta}})) + 2\,\widehat{\mathrm{tr}}\left[I(\underline{\theta}_o)\Sigma\right]$$

Kei Takeuchi was born in 1933 in Tokyo, Japan, and graduated in 1
of Tokyo. He received a Ph.D. in economics in 1966 (Keizaigaku Ha
interests include mathematical statistics, econometrics, global envir
tory of civilization, and Japanese economy. He is the author of many
statistics, and the impacts of science and technology on society. He
on the Faculty of International Studies at Meiji Gakuin University
University of Tokyo (recent photograph).

or even use more exact forms for the trace term. These id
in the next section.

First, however, there is one more crucial point on whic
clear: It is not required that truth f be in the set of model
AIC model selection. Many derivations of AIC are quite m
the assumption (often implicitly, hence without realizing
$f \subset g$). Such derivations lead directly to AIC, hence by
general result of (7.33), which does not require $f \subset g$. C
then it is possible to see how a proper philosophy of h
approximating models to complex truth in conjunction w
choice of $\widehat{\text{tr}}\left[I(\underline{\theta}_o)\Sigma\right] = K$ justifies use of AIC.

There are a few odds and ends worth considering at this
the result

$$\mathrm{E}_{\underline{y}}\left[I(f, g(\cdot \mid \underline{\hat{\theta}}(\underline{y})))\right] - I(f, g(\cdot \mid \underline{\theta}_o)) = \frac{1}{2}\,\mathrm{tr}\left[J(\underline{\theta}\right.$$

$$\underline{y}\left[\underline{x}[\log(g(\underline{x}\mid\underline{\theta}_o))]\right] \qquad \underline{x}[\log(g(\underline{x}\mid\underline{\theta}$$

Now substitute (7.23) for $\mathrm{E}_{\underline{x}}[\log(g(\underline{x}\mid\hat{\underline{\theta}}(\underline{y})))]$ in the abov

$$\frac{1}{2}\mathrm{E}_{\underline{y}}\left[\hat{\underline{\theta}}(\underline{y})-\underline{\theta}_o\right]' I(\underline{\theta}_o)[\hat{\underline{\theta}}(\underline{y})-\underline{\theta}_o\Big],$$

which becomes $\frac{1}{2}\operatorname{tr}\left[I(\underline{\theta}_o)\Sigma\right]=\frac{1}{2}\operatorname{tr}\left[J(\underline{\theta}_o)I(\underline{\theta}_o)^{-1}\right]$.

It should be almost obvious (and it is true) tha $\operatorname{tr}\left[J(\underline{\theta}_o)I(\underline{\theta}_o)^{-1}\right]$, does not depend upon sample size. Ra els it is about equal to K (these matters are explored sections below). In stark contrast, quantities such as the pected log-likelihood, and both of K-L $I(f, g(\cdot\mid\underline{\theta}_o))$ an $\mathrm{E}_{\underline{y}}\left[I(f, g(\cdot\mid\hat{\underline{\theta}}(\underline{y})))\right]$ increase linearly in sample size n. sample sizes, and K/n small, the ratio

$$\frac{\mathrm{E}_{\underline{y}}\left[I(f, g(\cdot\mid\hat{\underline{\theta}}(\underline{y})))\right]}{I(f, g(\cdot\mid\underline{\theta}_o))}$$

is essentially 1 even though the difference between expe is > 0. Thus, on an absolute scale TIC and AIC (when t model selection are producing the model estimated to p K-L model from the set of models considered if sample s is small.

The reason that the criterion for practical model selectio minimum K-L to minimum expected K-L as $\mathrm{E}_{\underline{y}}\big[I(f, g($ we must estimate $\underline{\theta}$ by the model-based MLE. This seem fact has deep ramifications. It is why the K-L–based conc the start of this section) virtually forces us to adopt $\mathrm{E}_{\underline{y}}\big[I$ minimized, hence T, i.e., (7.20), to be maximized.

In this regard there is a nominally puzzling result: If w as

$$I(f, g(\cdot\mid\underline{\theta}_o)) = \text{constant} - \mathrm{E}_{\underline{x}}\left[\log(g(\underline{x}\mid\underline{\theta}\right.$$

and no actual data in hand, hence no estimate of $\underline{\theta}$, we Taylor series expansion of $\log(g(\underline{x}\mid\underline{\theta}_o))$ about what woul any value of the variable of integration $\underline{x}$ (which is *not* da expectation over the sample space of the random variable

$$\mathrm{E}_{\underline{x}}\left[\log(g(\underline{x}\mid\underline{\theta}_o))\right]=\mathrm{E}_{\underline{x}}\left[\log(g(\underline{x}\mid\hat{\underline{\theta}}(\underline{x})))\right]-\frac{1}{2}\operatorname{tr}\left[J\right.$$

The above would suggest that K-L model selection maximizing $\log(g(\underline{x}\mid\hat{\underline{\theta}}(\underline{x})))-\frac{1}{2}\operatorname{tr}\big[\hat{J}(\underline{\theta}_o)\hat{I}(\underline{\theta}_o)^{-1}\big]$; it can

tion has already been integrated over the sample space, a
no $\underline{x}$ and no data involved in K-L. Data cannot be manu
series expansion on a random variable. Thus, the intrigu
matically correct, but conceptually wrong for what we a
hence misleading.

7.3 General K-L–Based Model Selection

7.3.1 Analytical Computation of TIC

There are other alternatives to estimation of relative K
that try to provide a data-based estimator of the trace te
are computationally much more intense, and the result
trace term can be so variable, and may have its own bi
tionable whether such approaches are worth applying (
huge). Takeuchi (1976) proposed TIC (see also Shibata 19
Kitagawa 1996): Select the model that minimizes (7.34)
tors of $J(\underline{\theta}_o)$ and $I(\underline{\theta}_o)$, hence getting an estimator of tr $\left[J\right.$
estimator of $I(\underline{\theta}_o)$ is (7.8), the empirical Hessian:

$$\hat{I}(\underline{\theta}_o) = \hat{I}(\hat{\underline{\theta}}) = -\frac{\partial^2 \log(g(\underline{x} \mid \hat{\underline{\theta}}))}{\partial \underline{\theta}^2}.$$

General estimation of $J(\underline{\theta}_o)$ relies on recognizing the s
on n independent units of information. In the simplest ca
as an *iid* sample, $x_1, \ldots, x_n$. It is required only that the sa
as having n conditionally independent components so tha
can be computed as the sum of n terms; hence we have

$$\log(g(\underline{x} \mid \hat{\underline{\theta}})) = \sum_{i=1}^{n} \log(g_i(x_i \mid \hat{\underline{\theta}})).$$

For the *iid* sample case, $g_i(x_i \mid \hat{\underline{\theta}}) \equiv g(x_i \mid \hat{\underline{\theta}})$. Using here
basic sample-size one pdf and for the probability distribu
full sample of size n is a minor abuse of notation. How
the reader will understand the meaning of the formulas a
minimize notation to facilitate comprehension of concept

A general estimator of $J(\underline{\theta}_o)$ for TIC can be derived fr

$$J(\underline{\theta}_o) = \mathrm{E}_f \left[\left[\frac{\partial}{\partial \underline{\theta}} \log(g(\underline{x} \mid \underline{\theta}_o)) \right] \left[\frac{\partial}{\partial \underline{\theta}} \log(g(\underline{x}\right.\right.$$

$$J(\underline{\theta}_o) = \mathrm{E}_f \left[\left[\sum_{i=1} \frac{}{\partial \underline{\theta}} \log(g(x_i \mid \underline{\theta}_o)) \right] \left[\sum_{i=1} \frac{}{\partial \underline{\theta}} \log \right.$$

$$= \sum_{i=1}^{K} \mathrm{E}_f \left[\frac{\partial}{\partial \underline{\theta}} \log(g(x_i \mid \underline{\theta}_o)) \right] \left[\frac{\partial}{\partial \underline{\theta}} \log(g(x_i \mid \right.$$

Therefore, we are led to use

$$\hat{J}(\underline{\theta}_o) = \sum_{i=1}^{K} \left[\frac{\partial}{\partial \underline{\theta}} \log(g(x_i \mid \hat{\underline{\theta}})) \right] \left[\frac{\partial}{\partial \underline{\theta}} \log(g(x_i \mid \hat{\underline{\theta}} \right.$$

A general version of TIC can be defined based on (7.36) a
Shibata 1989:222):

$$\mathrm{TIC} = -2 \log(g(\underline{x} \mid \hat{\underline{\theta}})) + 2 \operatorname{tr} \left[\hat{J}(\underline{\theta}_o)[\hat{I}(\underline{\theta}_o) \right.$$

One selects the model that produces the smallest TIC. Be
is for each model an asymptotically unbiased estimator of I
−constant, the underlying optimization criterion is that we
on average (over the set of models) minimizes this expec
loss. For large n this expected criterion is almost the sam
criterion $I(f, g)$ − constant; thus using (7.38), we are
selecting the K-L best model of the set of models, and
whether or not f is in the model set.

The estimator $\hat{J}(\underline{\theta}_o)$ converges to $J(\underline{\theta}_o)$, and $\hat{I}(\underline{\theta}_o)$ co
TIC is asymptotically unbiased (i.e., consistent) as a selec
minimum expected K-L model. In practice this estimator
so variable (and is not unbiased), even for large n, that it
use the parsimonious "estimator" $\widehat{\operatorname{tr}}\left[J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1}\right] = k$
(we will consider the matter further in later sections). Th
appropriate if we have done a good job of specifying our
which to select a best-fitting model.

7.3.2 *Bootstrap Estimation of TIC*

The primary value of the bootstrap method herein is to as
uncertainty based on applying an analytical model sele
AIC, AIC_c, QAIC_c, or TIC based on formulas (7.33), (7.34
ever, a second and quite different use of the bootstrap can
bootstrap method to estimate directly the quantity $T =$ E
the K-L best model is the one that maximizes $\hat{T}$. Variatio
volve more direct bootstrap estimation of the key quan

From (7.24) and (7.26) (wherein $\underline{\theta}$ denotes $\underline{\theta}(\underline{x})$) we d

$$T \approx \mathrm{E}_{\underline{x}}\left[\log(g(\underline{x} \mid \hat{\underline{\theta}}))\right] - \frac{1}{2}\operatorname{tr}\left[I(\underline{\theta}_o)\Sigma\right] - \frac{1}{2}\operatorname{tr}\left[\mathrm{E}_{\underline{x}}\left[\hat{I}(\hat{\underline{\theta}})\right]\right.$$

Hence, a model selection criterion can be based on

$$\hat{T} = \log(g(\underline{x} \mid \hat{\underline{\theta}})) - \frac{1}{2}\operatorname{tr}\left[\hat{I}(\underline{\theta}_o)\hat{\Sigma}\right] - \frac{1}{2}\operatorname{tr}\left[\hat{\mathrm{E}}_{\underline{x}}\left[\hat{I}(\hat{\underline{\theta}})\right][\underline{\theta}\right.$$

Additional approximations applied to (7.39), or to the ba
to

$$\hat{T} = \log(g(\underline{x} \mid \hat{\underline{\theta}})) - \operatorname{tr}\left[\hat{I}(\underline{\theta}_o)\hat{\Sigma}\right],$$

which could also be the basis for a bootstrap estimator (a

We assume that the sample structure allows a meaningfu
procedure (easily done in the *iid* sample case). Let a b
denoted by $\underline{x}^*$ with corresponding bootstrap MLE $\hat{\underline{\theta}}^*$. Th
second partial derivatives will have to be determined e
numerically. To avoid more notation, we do not index th
but rather just note that needed summations are over B b

In the bootstrap estimators, the MLE $\hat{\underline{\theta}}$ plays the role of
estimators of $I(\underline{\theta}_o)$, Σ, and $\mathrm{E}_{\underline{x}}\left[\hat{I}(\hat{\underline{\theta}})\right][\underline{\theta}_o - \hat{\underline{\theta}}][\underline{\theta}_o - \hat{\underline{\theta}}]'$ are

$$\hat{I}(\underline{\theta}_o) = -\frac{1}{B}\left[\sum_B \frac{\partial^2 \log(g(\underline{x}^* \mid \hat{\underline{\theta}}))}{\partial \underline{\theta}^2}\right],$$

$$\hat{\Sigma} = \frac{1}{B}\left[\sum_B [\hat{\underline{\theta}}^* - \hat{\underline{\theta}}][\hat{\underline{\theta}}^* - \hat{\underline{\theta}}]'\right],$$

$$\hat{\mathrm{E}}_{\underline{x}}\left[\left[\hat{I}(\hat{\underline{\theta}})\right][\underline{\theta}_o - \hat{\underline{\theta}}][\underline{\theta}_o - \hat{\underline{\theta}}]'\right]$$

$$= \frac{1}{B}\left[\sum_B \left[-\frac{\partial^2 \log(g(\underline{x}^* \mid \hat{\underline{\theta}}^*))}{\partial \underline{\theta}^2}\right][\hat{\underline{\theta}}\right.$$

These estimators mimic the expectation over f, because th
f, the bootstrap resamples the sample, and under any mod
of $\underline{\theta}_o$ is the MLE $\hat{\underline{\theta}}$ (note that $\underline{\theta}_o$ varies by model g). One sh
bootstrap samples with every model in the set of models
is made.

The above suffices to compute TIC as

$$\mathrm{TIC} = -2\log(g(\underline{x} \mid \hat{\underline{\theta}})) + 2\operatorname{tr}\left[\hat{I}(\underline{\theta}_o)\hat{\Sigma}\right.$$

based model selection, base the estimation of its second a
on (7.40), (7.41), and (7.42); or in a form analogous to AIC
selection criterion to minimize is

$$-2\log(g(\underline{x}\mid\hat{\underline{\theta}})) + \operatorname{tr}\left[\hat{I}(\underline{\theta}_o)\hat{\Sigma}\right] + \operatorname{tr}\hat{\mathrm{E}}_{\underline{x}}\left[\left[\hat{I}(\hat{\underline{\theta}})\right][\underline{\theta}_o - \hat{\underline{\theta}}\right.$$

It may well be that (7.43), i.e., TIC, would suffice and (
estimator of $-2T$.

Recent work on this use of the bootstrap to find $\hat{T}$ fo
selection is found in Ishiguro, et al. (1997), Cavanaugh an
Shao (1996) and Chung et al. (1996). Shibata (1997a)
general context, theoretical properties of many alternative
the bootstrap to estimate the needed model selection crite
there is no unique way to do this bootstrapping to estim
model selection criterion, but that all reasonable bootstrap
asymptotically equivalent to TIC. This use of the bootstrap
bypassing concerns about all approximations used to get
this apparent advantage, Shibata (1997a, page 393) concl
reason to use the bootstrap this way to compute $\hat{T}$. It pro
a simple nonbootstrap computation of $\hat{T}$ (in particular, A

It should thus be clear that there are two very differ
bootstrap in model selection. Not much used is the cas
estimate of T for each model based on the full set of boo
more common (and more useful) use of the bootstrap in m
to accept some easily computable model selection criteri
then to apply that criterion to all models considered for all t
created (and tabulate results like frequency of selection
use of the bootstrap leads to information about inferenc
model selection. [There is also a large literature on use of
non–K-L–based model selection; see, e.g., Breiman 199
Hjorth 1994; Linhart and Zucchini 1986; and Shao 1996.

7.4 AIC_c: A Second-Order Improvement

7.4.1 Derivation of AIC_c

The results above are completely general, and as such do
the more specific results in the literature. In particular,
variate linear structural model with homogeneous, normal
conditional on any regressor variables, we can get the re

$$\mu_i = \mathrm{E}(x_i \mid \underline{z}) = \sum_{j=1} z_{ij}\beta_j, \qquad i = 1, \ldots$$

More specifically (but without explicitly denoting th "regressors" $\underline{z}_i$),

$$x_i = \sum_{j=1}^{K-1} z_{ij}\beta_j + \epsilon_i, \qquad i = 1, \ldots, n$$

where the ϵ_i are *iid* normal$(0, \sigma^2)$. There are thus K para (σ^2 is the Kth one), and $g(\underline{x} \mid \underline{\theta})$ is given by the multivari distribution (MVN); I is the $n \times n$ identity matrix. If we then we can derive the AIC_c results of Hurvich and Tsa means that either g is the true data-generating "model," or f distribution and structural form as *model g* but with one o set to 0 (hence there are superfluous parameters). The sup serve only to increase K; hence the simplest way to get A regression model g and assume that $f \equiv g$. The derivati some detail because of the importance of AIC_c.

Matrix notation is simpler to use, and hence $\underline{X} = Z\underline{\beta}$ Without loss of generality we assume that Z (n by $K - 1$ likelihood is

$$g(\underline{x} \mid \underline{\theta}) = \left[\frac{1}{\sqrt{2\pi}}\right]^n \left[\frac{1}{\sigma^2}\right]^{n/2} \exp\left[-\frac{1}{2}\frac{(\underline{X} - Z\underline{\beta})'}{\sigma}\right.$$

and we are here taking $f \equiv g$. Ignoring additive consta the log-likelihood can be taken as

$$\log(g(\underline{x} \mid \underline{\theta})) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\frac{(\underline{X} - Z\underline{\beta})'(\underline{X}}{\sigma^2}$$

The MLEs are well known here:

$$\hat{\underline{\beta}} = (Z'Z)^{-1}Z'\underline{X},$$
$$\hat{\sigma}^2 = \frac{(\underline{X} - Z\hat{\underline{\beta}})'(\underline{X} - Z\hat{\underline{\beta}})}{n}.$$

Therefore,

$$\log(g(\underline{x} \mid \hat{\underline{\theta}}(\underline{x}))) = -\frac{n}{2}\log(\hat{\sigma}^2) - \frac{1}{2}\frac{(\underline{X} - Z\hat{\underline{\beta}})'(}{\hat{\sigma}^2}$$

hence, the maximized log-likelihood is

$$\log(g(\underline{x} \mid \hat{\underline{\theta}}(\underline{x}))) = -\frac{n}{2}\log(\hat{\sigma}^2) - \frac{n}{2}$$

our target

$$T = \mathrm{E}_{\underline{x}}\mathrm{E}_{\hat{\underline{\theta}}(\underline{y})}\left[\log(g(\underline{x} \mid \hat{\underline{\theta}}(\underline{y})))\right],$$

where $\underline{x}$ and $\underline{y}$ are two independent random samples of
evaluation here we actually use the specified form of the
take expectations with respect to $f \equiv g$). Hence, we want
is used here)

$$T = \mathrm{E}_{\underline{x}}\mathrm{E}_{\hat{\underline{\theta}}(\underline{y})}\left[\log(g(\underline{x} \mid \hat{\underline{\theta}}(\underline{y})))\right]$$
$$= \mathrm{E}_{\hat{\underline{\theta}}(\underline{y})}\mathrm{E}_{\underline{x}}\left[-\frac{n}{2}\log(\hat{\sigma}_y^2) - \frac{1}{2}\frac{(\underline{X} - Z\hat{\underline{\beta}}_y)'}{\hat{\sigma}}\right.$$

The order of integration was reversed for the right-hand
first task is to evaluate

$$\mathrm{E}_{\underline{x}}\left[(\underline{X} - Z\hat{\underline{\beta}}_y)'(\underline{X} - Z\hat{\underline{\beta}}_y)\right]$$
$$= \mathrm{E}_{\underline{x}}\left[((\underline{X} - Z\underline{\beta}) + (Z\underline{\beta} - Z\hat{\underline{\beta}}_y))'((\underline{X} - Z\underline{\beta}) +\right.$$
$$= \mathrm{E}_{\underline{x}}\left[(\underline{X} - Z\underline{\beta})'(\underline{X} - Z\underline{\beta})\right] + \mathrm{E}_{\underline{x}}\left[2(Z\underline{\beta} - Z\hat{\underline{\beta}}\right.$$
$$+ \mathrm{E}_{\underline{x}}\left[(Z\underline{\beta} - Z\hat{\underline{\beta}}_y)'(Z\underline{\beta} - Z\hat{\underline{\beta}}_y)\right]$$
$$= \mathrm{E}_{\underline{x}}\left[(\underline{X} - Z\underline{\beta})'(\underline{X} - Z\underline{\beta})\right] + \left[2(Z\underline{\beta} - Z\hat{\underline{\beta}}_y)'\right.$$
$$+ \left[(Z\underline{\beta} - Z\hat{\underline{\beta}}_y)'(Z\underline{\beta} - Z\hat{\underline{\beta}}_y)\right].$$

The middle term above vanishes because $\mathrm{E}_{\underline{x}}(\underline{X}) = Z\underline{\beta}$.
three terms above is identical to $\mathrm{E}_{\underline{x}}(\underline{\epsilon}'\underline{\epsilon}) = n\sigma^2$. So we ha

$$\mathrm{E}_{\underline{x}}\left[(\underline{X} - Z\hat{\underline{\beta}}_y)'(\underline{X} - Z\hat{\underline{\beta}}_y)\right] = n\sigma^2 + \left[(Z\underline{\beta} - Z\hat{\underline{\beta}}_y)\right.$$

Using this partial result we have

$$T = \mathrm{E}_{\hat{\underline{\theta}}(\underline{y})}\left[-\frac{n}{2}\log(\hat{\sigma}_y^2)\right] - \frac{1}{2}\mathrm{E}_{\hat{\underline{\theta}}(\underline{y})}\left[\frac{n\sigma^2 + \left[(Z\underline{\beta} - Z\hat{\underline{\beta}}_\right.}{\hat{\sigma}_y^2}\right.$$

The first term above does not need to be evaluated further
leading term in the expected log-likelihood. Also, at this p
designation of $\underline{\theta}$ as being based on sample $\underline{y}$. The designat
just dummy arguments in integrals. Consequently, in the

$$T = \mathrm{E}_{\underline{\hat{\theta}}}\left[-\frac{n}{2}\log(\hat{\sigma}^2)\right] - \frac{1}{2}\mathrm{E}_{\underline{\hat{\theta}}}\left[\frac{}{\hat{\sigma}^2}\right.$$

Now we make use of another well-known result in t
Under a linear model structure with errors as *iid* normal(
and $\hat{\sigma}^2$ are independent random variables. Therefore, the
term in (7.45) partitions into two multiplicative parts, as

$$T = \mathrm{E}\left[-\frac{n}{2}\log(\hat{\sigma}^2)\right] - \frac{1}{2}\mathrm{E}_{\underline{\hat{\beta}}}\left[n\sigma^2 + \left[(Z\underline{\beta} - Z\underline{\hat{\beta}})'(Z\underline{\beta} -\right.\right.$$

As a next step, rewrite the needed expectation of the q
above as

$$\mathrm{E}\left[(Z\underline{\beta} - Z\underline{\hat{\beta}})'(Z\underline{\beta} - Z\underline{\hat{\beta}})\right] = \mathrm{tr}\left[(Z'Z)\mathrm{E}\left[(\underline{\hat{\beta}} - \underline{\beta}\right.\right.$$

The expectation on the right-hand side above, i.e., E[($\underline{\hat{\beta}}$ -
sampling variance–covariance matrix of $\underline{\hat{\beta}}$, which is know
Thus, for the $K-1$ square identity matrix I,

$$\mathrm{E}\left[(Z\underline{\beta} - Z\underline{\hat{\beta}})'(Z\underline{\beta} - Z\underline{\hat{\beta}})\right] = \mathrm{tr}[\sigma^2 I] = \sigma^2($$

Putting it all together to this point in the derivation, we

$$T = \mathrm{E}\left[-\frac{n}{2}\log(\hat{\sigma}^2)\right] - \frac{1}{2}\left[(n+K-1)\sigma^2\right]\mathrm{E}_{\hat{\sigma}^2}\left[\right.$$

To finish the process we relate $\hat{\sigma}^2$ to a central chi-squar
namely χ^2_{df} on $n-(K-1)$ degrees of freedom, df. These
known in statistical theory:

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-K+1}.$$

So we now rearrange (7.46) to be

$$T = \mathrm{E}\left[-\frac{n}{2}\log(\hat{\sigma}^2)\right] - \frac{1}{2}\left[(n+K-1)n\right]\mathrm{E}\left[\frac{}{n}\right.$$

$$T = \mathrm{E}\left[-\frac{n}{2}\log(\hat{\sigma}^2)\right] - \frac{n}{2}(n+K-1)\mathrm{E}\left[\frac{1}{\chi^2_{n-K}}\right.$$

Yet another known exact result is

$$\mathrm{E}\left[\frac{1}{\chi^2_{df}}\right] = \frac{1}{\mathrm{df}-2}$$

(assuming df >2).

$$T = \mathrm{E}\left[-\frac{n}{2}\log(\hat{\sigma}^2)\right] - \frac{n}{2}(n+K-1)\left[\frac{}{n-K-}\right.$$

This result is exact. No approximations were made in its
it applies only to the particular context of its derivation,
constraint $f \subseteq g$. Some more simplification of (7.47):

$$\begin{aligned} T &= \mathrm{E}\left[-\frac{n}{2}\log(\hat{\sigma}^2)\right] - \frac{n}{2}\left[\frac{n+K-1}{n-K-1}\right] \\ &= \mathrm{E}\left[-\frac{n}{2}\log(\hat{\sigma}^2)\right] - \frac{n}{2}\left[1+\frac{2K}{n-K-}\right. \\ &= \mathrm{E}\left[-\frac{n}{2}\log(\hat{\sigma}^2)\right] - \frac{n}{2} - \frac{nK}{n-K-1} \\ &= \mathrm{E}\left[-\frac{n}{2}\log(\hat{\sigma}^2) - \frac{n}{2}\right] - \frac{nK}{n-K-1} \end{aligned}$$

The term above within the expectation operator is t
likelihood. Thus we have

$$\begin{aligned} T &= \mathrm{E}\left[\log(g(\underline{x} \mid \hat{\underline{\theta}}(\underline{x})))\right] - \frac{nK}{n-K-1} \\ &= \mathrm{E}\left[\log(g(\underline{x} \mid \hat{\underline{\theta}}(\underline{x})))\right] - \frac{(n-K-1+K}{n-K-} \\ &= \mathrm{E}\left[\log(g(\underline{x} \mid \hat{\underline{\theta}}(\underline{x})))\right] - K - \frac{K(K+1)}{n-K-1} \end{aligned}$$

If we convert this to an AIC result, we have, as an exact r

$$\begin{aligned} -2T &= -2\mathrm{E}\left[\log(g(\underline{x} \mid \hat{\underline{\theta}}(\underline{x})))\right] + 2K + \frac{2K(}{n-} \\ &= \mathrm{E}(\mathrm{AIC}) + \frac{2K(K+1)}{n-K-1} = \mathrm{E}(\mathrm{AIC}_c). \end{aligned}$$

This result thus motivates use of the term $2K(K+1)$
small-sample-size bias-correction term added to AIC. T
fixed-effects linear model with normal errors and constan
Under different sorts of models, a different small-sampl
arises (the matter is explored some in the next subsection)
given by (7.48) seems useful in other contexts, especially
also large relative to n. Without exception, if sample size
of "AIC_c" is required for good model selection results,
(7.48) unless a more exact small-sample correction to AI

so for multivariate) models, with $g = f$ to facilitate a de
deriving a small-sample adjustment to AIC, will lead to
terms. This section is just a brief elaboration of this idea.

The simplest case to present arises for a situation an
ANOVA, but we let the within-subgroup variance differ
This can be generalized to having m subsets of data, eac
and the full model is as used in Section 7.4.1 above, bu
$\{\underline{\theta}\} = \{\underline{\beta}_i, \sigma_i^2, i = 1, \ldots, m\}$ (this might be a global m
Let each parameter subset be of size K_i; hence $K = K_1 +$
be almost obvious, after some thought, that for this situati
correction to AIC is

$$-2T = -2\mathrm{E}\left[\log(g(\underline{x} \mid \underline{\hat{\theta}}(\underline{x})))\right] + 2\sum_{i=1}^{m}\left[K_i + \frac{2K}{n_i}\right.$$

hence,

$$\mathrm{AIC}_c = -2\log(g(\underline{x} \mid \underline{\hat{\theta}}(\underline{x}))) + 2K + \sum_{i=1}^{m}\left[\frac{2K_i(}{n_i -}\right.$$

The reason that there are m "correction" terms is that w
different variance parameters. One can thus envision many
the form of AIC_c must differ from that of the simple norr
only one estimated σ^2.

Another informative exact calculation of the bias term, T
is obtained for the case of the model and truth being the one
exponential distribution (hence $K = 1$):

$$g(x \mid \lambda) = \frac{1}{\lambda}e^{-x/\lambda}.$$

For an *iid* sample from $g(x \mid \lambda)$, let $S = x_1 + \cdots + x_n$. Th

$$\log(g(\underline{x} \mid \lambda)) = -n\log(\lambda) - S/\lambda.$$

The MLE is $\hat{\lambda} = S/n$, so

$$\log(g(\underline{x} \mid \hat{\lambda})) = -n\log(\hat{\lambda}) - n.$$

The target to be unbiasedly estimated is

$$T = \mathrm{E}_{\underline{x}}\mathrm{E}_{\underline{y}}\left[-n\log(\hat{\lambda}) - S_{\underline{y}}/\hat{\lambda}\right],$$

where the sum $S_{\underline{y}}$ is based on an independent sample of siz
on sample $\underline{x}$. It is easy to evaluate the above T to be

$$T = \mathrm{E}\left[\log(g(\underline{x} \mid \hat{\lambda}))\right] + n - n^2\lambda\mathrm{E}\left[\frac{1}{S}\right.$$

to derive the exact result, expressed in "AIC" form (i.e., t

$$-2T = -2\mathrm{E}(\log(g(\underline{x} \mid \hat{\lambda}))) + 2 + \frac{2}{n-}$$

Recall that here $K = 1$, so the corresponding total bias-c the AIC_c form would be $2 + 4/(n-2)$. The point is that th would be $2K$ plus a small-sample correction term that wou the model assumed. It is reasonable to think that this sma term should be $O(1/n)$.

Theoretically, when $f \subset g$ the error in using K as to $\hat{T} = \log(g(\underline{x} \mid \hat{\underline{\theta}}))$ is always $O(1/n)$, and Hurvich an like a good general choice. There is, however, considerab on improved bias terms for AIC_c-type criteria. In this reg research opportunities is that of when the random variable Sugiura 1978, Shibata 1997b), such as Poisson, binomial, also logistic regression), because then we can get parame the value 0. This creates a problem in evaluating the theo selection criterion because we encounter the need to comp is not defined (see, e.g., Burnham et al. 1994). AIC is small-sample properties are now more problematic, as bias-correction term needed to define an AIC_c. Operati AIC-based model selection for count-type data need m sample sizes.

7.5 Derivation of AIC for the Exponentia of Distributions

A generalization of normality-based models is found in th of distributions. The realizations that (1) many common a tical analyses are based on exponential family models, and regression is in the exponential family and leads to exact results (i.e., AIC_c) motivated us to show the derivation of A restricted but very useful case. The canonical representati family pdf involves sums of functions of the sample value denote these sums by S_j.

A suitable canonical representation for the exponential distributions is

$$\begin{aligned} g(\underline{x} \mid \underline{\theta}) &= \exp\left[\left[\sum_{j=1}^{K} S_j \theta_j\right] + H(\underline{\theta}) + G \right. \\ &= \exp\left[\underline{S}'\underline{\theta} + H(\underline{\theta}) + G(\underline{S})\right]. \end{aligned}$$

K-element vector of sufficient statistics is $\underline{S} = (S_1, \ldots, S$

In the canonical representation of (7.50) the parameter
1-to-1 transformation of another K-dimensional paramet
There is no loss of generality in allowing any such 1-to-1
will revisit this matter and show why it is so at the end of

Our goal is to evaluate

$$T = \mathrm{E}_{\underline{x}}\mathrm{E}_{\underline{y}}\left[\log(g(\underline{x} \mid \hat{\underline{\theta}}_y))\right] = \mathrm{E}_{\underline{x}}\mathrm{E}_{\underline{y}}\left[\underline{S}'_x\hat{\underline{\theta}}_y + H(\hat{\underline{\theta}}_y) +$$

Here, $\underline{S}_x$ and $\hat{\underline{\theta}}_y$ are thought of as based on independent s
also simplified the notation, now using $\hat{\underline{\theta}}_y$ rather than $\hat{\underline{\theta}}(\underline{y}$

Formula (7.51) above can be rewritten as

$$\begin{aligned}
T &= \mathrm{E}_{\underline{x}}\mathrm{E}_{\underline{y}}\left[(\underline{S}_x - \underline{S}_y + \underline{S}_y)'\hat{\underline{\theta}}_y + H(\hat{\underline{\theta}}_y) + G(\underline{S}_x)\right] \\
&= \mathrm{E}_{\underline{x}}\mathrm{E}_{\underline{y}}\left[(\underline{S}_x - \underline{S}_y)'\hat{\underline{\theta}}_y + \underline{S}'_y\hat{\underline{\theta}}_y + H(\hat{\underline{\theta}}_y) + G(\underline{S}_x)\right. \\
&= \mathrm{E}_{\underline{x}}\mathrm{E}_{\underline{y}}\left[\underline{S}'_y\hat{\underline{\theta}}_y + H(\hat{\underline{\theta}}_y) + G(\underline{S}_x)\right] + \mathrm{E}_{\underline{x}}\mathrm{E}_{\underline{y}}\left[(\underline{S}_x -\right. \\
&= \left[\mathrm{E}_{\underline{y}}(\underline{S}'_y\hat{\underline{\theta}}_y + H(\hat{\underline{\theta}}_y)) + \mathrm{E}_{\underline{x}}(G(\underline{S}_x))\right] + \mathrm{E}_{\underline{y}}\left[(\mathrm{E}_{\underline{x}}(\right.
\end{aligned}$$

The interchangeability of integration arguments now is u
sible because both expectations are with respect to f;
$\mathrm{E}_{\underline{y}}(G(\underline{S}_y))$. Also, for simplicity we will use $\mathrm{E}_{\underline{x}}(\underline{S}_x) = \mathrm{E}(\underline{S}$

$$T = \mathrm{E}_{\underline{y}}\left[\underline{S}'_y\hat{\underline{\theta}}_y + H(\hat{\underline{\theta}}_y)) + G(\underline{S}_y)\right] + \mathrm{E}_{\underline{y}}\left[(\mathrm{E}(\underline{S})\right.$$

Changing the argument from y to x in the first part abov
the result for this exponential family case, we have

$$\begin{aligned}
T &= \mathrm{E}_{\underline{x}}(\log(g(\underline{x} \mid \hat{\underline{\theta}}))) + \mathrm{E}_{\underline{y}}\left[(\mathrm{E}(\underline{S}) - \underline{S}_y)'\right. \\
&\equiv \mathrm{E}_{\underline{x}}(\log(g(\underline{x} \mid \hat{\underline{\theta}}))) - \mathrm{E}_{\underline{y}}\left[(\underline{S}_y - \mathrm{E}(\underline{S}))'\right.
\end{aligned}$$

Formula (7.52) is an exact result and clearly shows the b
from $\mathrm{E}_{\underline{x}}(\log(g(\underline{x} \mid \hat{\underline{\theta}})))$ to get T:

$$\text{Bias} = \mathrm{E}_{\underline{y}}\left[(\underline{S}_y - \mathrm{E}(\underline{S}))'(\hat{\underline{\theta}}_y - \underline{\theta}_*)\right].$$

The notation used here is $\mathrm{E}(\hat{\underline{\theta}}_y) = \underline{\theta}_*$ to denote the exac
MLE for the given sample size n and model g; $\underline{\theta}_o \approx$
equality.

$$\text{Bias} = \mathrm{E}\left[(\underline{S} - \mathrm{E}(\underline{S}))'(\hat{\underline{\theta}} - \underline{\theta}_*)\right],$$

$$\text{Bias} = \text{tr}\,\mathrm{E}\left[(\hat{\underline{\theta}} - \underline{\theta}_*)(\underline{S} - \mathrm{E}(\underline{S}))'\right] = \text{tr}\left[\text{COV}\right.$$

Hence for the exponential family *an exact result* is

$$T = \mathrm{E}_{\underline{x}}(\log(g(\underline{x} \mid \hat{\underline{\theta}}))) - \text{tr}\left[\text{COV}(\hat{\underline{\theta}}, \underline{S})\right.$$

(something similar appears in Bonneu and Milhaud 1994)
of covariance elements, $\text{COV}(\hat{\underline{\theta}}, \underline{S})$, can be approximat
methods. If the exact covariance matrix can be found, th
result for the needed bias term above (Hurvich and Tsai
such an exact evaluation for the normal distribution cas
may seem not very useful because it seems to apply o
form of the exponential family. This is not true; the matte
canonical result will be addressed below.

Before further evaluation of the bias term, we conside
Hessian. First,

$$\log(g(\underline{x} \mid \underline{\theta})) = \underline{S}'\underline{\theta} + H(\underline{\theta}) + G(\underline{S}),$$

so

$$\frac{\partial \log(g(\underline{x} \mid \underline{\theta}))}{\partial \underline{\theta}} = \underline{S} + \frac{\partial H(\underline{\theta})}{\partial \underline{\theta}},$$

$$\frac{\partial^2 \log(g(\underline{x} \mid \underline{\theta}))}{\partial \underline{\theta}^2} = \frac{\partial^2 H(\underline{\theta})}{\partial \underline{\theta}^2},$$

and thus

$$I(\underline{\theta}_o) = \mathrm{E}_f\left[-\frac{\partial^2 \log(g(\underline{x} \mid \underline{\theta}_o))}{\partial \underline{\theta}^2}\right] = -\frac{\partial^2 H}{\partial}$$

It follows that the MLE satisfies

$$\underline{S} = -\frac{\partial H(\hat{\underline{\theta}})}{\partial \underline{\theta}}.$$

It is worth noting here that $\underline{\theta}_o$ satisfies

$$\mathrm{E}_f(\underline{S}) = -\frac{\partial H(\underline{\theta}_o)}{\partial \underline{\theta}}.$$

This is an exact result, whereas $\mathrm{E}(\hat{\underline{\theta}}) \approx \underline{\theta}_o$ is (in general)

The formula for $J(\underline{\theta}_o)$, based on (7.12), becomes

$$J(\underline{\theta}_o) = \mathrm{E}_f\left[\underline{S} - \mathrm{E}_f(\underline{S})\right]\left[\underline{S} - \mathrm{E}_f(\underline{S})\right]'$$

by (7.56) is

$$\hat{J}(\underline{\theta}_o) = \frac{n}{n-1}\left[\sum_{i=1}^{n}\left[\underline{s}_i - \overline{\underline{s}}\right]\left[\underline{s}_i - \overline{\underline{s}}\right]'\right.$$

Returning now to the evaluation of the bias term, a firs
expansion gives us

$$-\frac{\partial H(\hat{\underline{\theta}})}{\partial \underline{\theta}} \approx -\frac{\partial H(\underline{\theta}_o)}{\partial \underline{\theta}} - \frac{\partial^2 H(\underline{\theta}_o)}{\partial \underline{\theta}^2}(\hat{\underline{\theta}} - \underline{\theta}$$

hence

$$\underline{S} \approx \mathrm{E}(\underline{S}) + I(\underline{\theta}_o)(\hat{\underline{\theta}} - \underline{\theta}_o), \qquad O_p(1/\sqrt{}$$

Inserting (7.58) into the exact result (7.53) as well as also
mate $\underline{\theta}_*$ (inasmuch as we are now replacing an exact result
result anyway), we have

$$\begin{aligned}
\text{Bias} &\approx \mathrm{E}\left[\left[I(\underline{\theta}_o)(\hat{\underline{\theta}} - \underline{\theta}_o)\right]'(\hat{\underline{\theta}} - \underline{\theta}_o)\right] \\
&= \mathrm{E}\left[(\hat{\underline{\theta}} - \underline{\theta}_o)' I(\underline{\theta}_o)(\hat{\underline{\theta}} - \underline{\theta}_o)\right] \\
&= \mathrm{E}\,\mathrm{tr}\left[I(\underline{\theta}_o)(\hat{\underline{\theta}} - \underline{\theta}_o)(\hat{\underline{\theta}} - \underline{\theta}_o)'\right] \\
&= \mathrm{tr}\left[I(\underline{\theta}_o)\mathrm{E}\left[(\hat{\underline{\theta}} - \underline{\theta}_o)(\hat{\underline{\theta}} - \underline{\theta}_o)'\right]\right] = \mathrm{tr}\left[\right.
\end{aligned}$$

Thus we have shown that in this common case of an
model,

$$T \approx \mathrm{E}_{\underline{x}}(\log(g(\underline{x} \mid \hat{\underline{\theta}}))) - \mathrm{tr}\left[I(\underline{\theta}_o)\Sigma\right]$$

(the approximation is to $O(1/n)$). Note that this derivatio
any problems like those in approximation (7.28) in the g
AIC in Section 7.2.

These results can be extended to any parametrized for
family model, because then we just have a 1-to-1 transform
$\underline{\beta}$ via some set of K functions, denoted here by $\underline{W}(\underline{\theta}) = \underline{\beta}$
and let Σ_θ and Σ_β be the variance–covariance matrices for
two parametrizations. An expected matrix of mixed secon
as per (7.55) exists for the $\underline{\beta}$ parametrization; denote it by
Jacobian of $\underline{W}$, evaluated at $\underline{\theta}_o$, be

$$J_w = \left\{\frac{\partial W_i(\underline{\theta}_o)}{\partial \theta_j}\right\}.$$

Then

$$J_w \Sigma_\theta J_w' = \Sigma_\beta,$$

Both the K-L–based target and the expected log-likeliho
to-1 parameter transformations, so this must also be true fo
correction. That is, any likelihood and MLE-based mode
ought to be invariant to 1-to-1 reparametrizations of the
the case here:

$$\begin{aligned}\mathrm{tr}(I(\underline{\beta}_o)\Sigma_\beta) &= \mathrm{tr}\left[(J_w')^{-1}I(\underline{\theta}_o)(J_w)^{-1}J_w\Sigma_\theta J_w'\right.\\ &= \mathrm{tr}\left[(J_w')^{-1}I(\underline{\theta}_o)\Sigma_\theta J_w'\right]\\ &= \mathrm{tr}\left[I(\underline{\theta}_o)\Sigma_\theta J_w'(J_w')^{-1}\right] = \mathrm{tr}\left[I(\right.\end{aligned}$$

Note, however, that if we were to estimate this trace term,
perform better under some parametrizations than under o

One last point here: It is certainly still true that

$$\mathrm{tr}\left[I(\underline{\theta}_o)\Sigma\right] = \mathrm{tr}\left[J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1}\right].$$

So an alternative to (7.59) is

$$T \approx \mathrm{E}_{\underline{x}}(\log(g(\underline{x}\,|\,\hat{\underline{\theta}}))) - \mathrm{tr}\left[J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1}\right]$$

This could be directly proven here, based on the simple r

$$\frac{\partial \log(g(\underline{x}\,|\,\underline{\theta}))}{\partial \underline{\theta}} = \underline{S} - \mathrm{E}(\underline{S})$$

and (7.58) to derive $[I(\underline{\theta}_o)]^{-1}J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1} = \Sigma$.

For TIC we can use $\hat{J}(\underline{\theta}_o)$ from (7.57) and from (7.55)

$$\hat{I}(\underline{\theta}_o) = -\frac{\partial^2 H(\hat{\underline{\theta}})}{\partial \underline{\theta}^2},$$

getting an estimator of $\mathrm{tr}\left[J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1}\right]$ that can be u
variance) even if the parametrization of interest (and used
$\underline{\beta}_o = \underline{W}(\underline{\theta}_o)$, not $\underline{\theta}$.

Working with exponential family cases facilitates som
ation of both $\mathrm{tr}\left[J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1}\right]$, relative to the value K, a
the estimator $\mathrm{tr}\left[\hat{J}(\underline{\theta}_o)[\hat{I}(\underline{\theta}_o)]^{-1}\right]$. These topics, and others
next section.

7.6 Evaluation of $\mathrm{tr}(J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1})$ and

The general derivation of a K-L–based model selection
(7.31) and hence (7.33). By "a general derivation," we m

L²-based model selection does not appear to be AIC. Rati
bias correction term subtracted from the expected maximiz
get T is $\mathrm{tr}(J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1})$ (Takeuchi 1976). In deriving th
and need not be, any assumption that any of the candida
truth. However, in general we know with certainty that $\mathrm{tr}(J$
only when $J(\underline{\theta}_o) = I(\underline{\theta}_o)$ (this is sufficient but not neces
equality is certain, in general, only when f is a special ca
model g equals or is a generalization of "truth." This co
to expect, so how good is the approximation $\mathrm{tr}(J(\underline{\theta}_o)[I$
the truth is more general than the model, but the model is
tion to truth? We make here some limited, but useful, pr
Extensive theory, simulation studies, and experience (e.g
chini 1986:176–182, especially results such as in their Ta
to give us full confidence in when we can expect reliabl
versus when we might have to use TIC. Below, we establi
for some models within the exponential family of distribu

7.6.1 Comparison of AIC Versus TIC in a Very S

We consider two simple one-parameter distributions:
tial and half-normal. For each distribution we can c
$\mathrm{tr}(J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1})$, assuming either that the distribution is
or that the other distribution is truth f and the given distr
(so $f \neq g$). We will also examine the estimators of the tra
in TIC model selection and contrast TIC selection with AI
two distributions as models. This is a convenient situatio
because both distributions are in the exponential family.

For the negative-exponential distribution let $S = x_1 +$

$$g(\underline{x} \mid \lambda) = \frac{1}{\lambda^n} e^{-S/\lambda},$$
$$\log(g(\underline{x} \mid \lambda)) = -n \log(\lambda) - S/\lambda,$$
$$\frac{\partial}{\partial \lambda} \log(g(\underline{x} \mid \lambda)) = -n/\lambda + S/\lambda^2,$$

so $\hat{\lambda} = S/n = \overline{x}$; also here $\mathrm{E}(x) = \lambda$.

Direct verification yields the following results:

$$I(\lambda) = \frac{n}{\lambda^2},$$
$$\hat{I}(\hat{\lambda}) = \frac{n}{\hat{\lambda}^2},$$

$$J(\lambda) = \sum_{i=1}^{} \left[\frac{}{\lambda} + \frac{}{\lambda^2} \right] = \left[\sum_{i=1}^{} \frac{}{\lambda^4} \right]$$

The true $J(\lambda) = \hat{\mathrm{E}}_f(\hat{J}(\lambda))$, and $\hat{J}(\hat{\lambda})$ is the empirical e
we assume that the negative-exponential model is truth,
(because $\mathrm{E}(x^2) = 2\lambda^2$),

$$J(\lambda) = \frac{n}{\lambda^2}.$$

Clearly, if $f = g$, then here $\mathrm{tr}[J(\lambda)[I(\lambda)]^{-1}] = 1$. T
estimator of this trace is

$$\mathrm{tr}[\hat{J}(\hat{\lambda})[\hat{I}(\hat{\lambda})]^{-1}] = \frac{1}{n\overline{x}^2} \left[\sum_{i=1}^{n} (x_i)^2 \right] -$$

or

$$\mathrm{tr}[\hat{J}(\hat{\lambda})[\hat{I}(\hat{\lambda})]^{-1}] = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n\overline{x}^2}.$$

This estimator of the trace is the same as $\frac{n-1}{n}(\widehat{\mathrm{cv}})^2$ (whe
denominator of the sample s^2 as per convention), which
quite variable. The trace estimator of (7.60) does converge
$f = g$. This trace estimator is scale-invariant, so we can
tional properties by Monte Carlo methods with a single ru
at each n, and for each case of truth being either the negat
value of λ can be used) or half-normal distribution (any σ
case where the negative-exponential model is truth, the s
a revised, nearly unbiased, version of (7.60):

$$\widehat{\mathrm{tr}}[J(\lambda)[I(\lambda)]^{-1}] = \frac{n}{n-1}(\widehat{\mathrm{cv}})^2.$$

The estimated mean and standard deviation of the estim
based on (7.60) and (7.61) are given below, based on 100

	Eq. (7.60)		Eq. (7.61)	
n	mean	st.dev.	mean	st.dev.
20	0.90	0.37	1.00	0.41
50	0.96	0.26	1.00	0.27
100	0.98	0.19	1.00	0.20
500	1.00	0.09	1.00	0.09

Notice the substantial standard deviation of either trace es
of large variability in the trace estimator that has given the
tine use of TIC. We also see that the direct estimator has bi

Consider now the half-normal distribution

$$f(x \mid \sigma^2) = \sqrt{\frac{2}{\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right]$$

Under this distribution $\mathrm{E}(x^2) = \sigma^2$, and direct integration

$$\mathrm{E}(x) = \sigma\sqrt{\frac{2}{\pi}}.$$

If the half-normal distribution is truth and the negative-ex
then λ_o as a function of σ^2 is found from

$$\mathrm{E}_f(x) = \lambda_o = \sigma\sqrt{\frac{2}{\pi}}.$$

For example, with $\sigma^2 = 1$, $\lambda_o = 0.79788$ is the K-L best

Note the usage and concepts here: We denote the K-
λ_o to distinguish that the corresponding negative-exponen
that based on λ_o) is the K-L best negative-exponential d
the model for the underlying truth. By denoting this valu
emphasizing that all we have is the K-L best negative-ex
it may be a poor model; it certainly may not be truth.

The above expectations producing $J(\lambda)$ and $I(\lambda)$ were
negative-exponential as g (ignoring what f might be), but
those expectations with respect to f as half-normal. Dire
$I(\lambda_o) = n/\lambda_o^2$, whereas

$$\begin{aligned} J(\lambda_o) &= n\mathrm{E}_f\left[\frac{x}{\lambda_o^2} - \frac{1}{\lambda_o}\right]^2 \\ &= I(\lambda_o)\left[\frac{\sigma^2}{\lambda_o^2} - 1\right] \\ &= I(\lambda_o)\left[\frac{\pi}{2} - 1\right]. \end{aligned}$$

Hence, when f is half-normal and g is negative-exponen

$$\mathrm{tr}\left[J(\lambda_o)[I(\lambda_o)]^{-1}\right] = \frac{\pi}{2} - 1 = 0.570$$

This trace term is not very close to 1, the number AIC
big relative difference (i.e., 0.5708 versus 1) results be
exponential model is a very poor approximation to the half
Note that this trace term is < 1 (K is 1 here).

$$\hat{\sigma}^2 = \frac{\sum_{i=1}(x_i)^2}{n},$$

$$I(\sigma^2) = \frac{n}{2\sigma^4},$$

$$\hat{J}(\sigma^2) = \sum_{i=1}^{n}\left[\frac{-1}{2\sigma^2} + \frac{(x_i)^2}{2\sigma^4}\right]^2$$

$$= I(\sigma^2)\left[\frac{1}{2n}\right]\sum_{i=1}^{n}\left[\frac{(x_i)^2}{\sigma^2} - 1\right]^2$$

The key part of the empirical estimator $\hat{J}(\hat{\sigma}^2)$ is again a s variation, but here it is for the variable x^2. Denote this $\hat{c}$ can use the notation

$$\hat{J}(\hat{\sigma}^2) = \hat{I}(\hat{\sigma}^2)\left[\frac{n-1}{2n}\right]\left[\widehat{\mathrm{cv}}(x^2)\right]^2.$$

Thus, for the half-normal distribution being the model, t the trace is

$$\mathrm{tr}[\hat{J}(\hat{\sigma}^2)[\hat{I}(\hat{\sigma}^2)]^{-1}] = \left[\frac{n-1}{2n}\right]\left[\widehat{\mathrm{cv}}(x^2)\right.$$

If truth is the half-normal model, this quantity will conv (7.62) is scale-invariant. For the half-normal model as tru mean and standard deviation of (7.62) by Monte Carlo m samples at each n. This led to a nearly unbiased version o

$$\widehat{\mathrm{tr}}[J(\sigma^2)[I(\sigma^2)]^{-1}] = \frac{1}{2}\left[\frac{n}{n-1}\right]^2\left[\widehat{\mathrm{cv}}(x^2\right.$$

that we then also used in the simulations. The results are

	Eq. (7.62)		Eq. (7.63)	
n	mean	st.dev.	mean	st.dev.
20	0.87	0.39	1.01	0.45
50	0.94	0.30	1.00	0.32
100	0.97	0.23	1.00	0.23
500	0.99	0.11	1.00	0.11

The main point from the above is how variable the trace

To complete a set of analytical results (useful for vali results) we computed the value of $\mathrm{tr}[J(\sigma_o^2)[I(\sigma_o^2)]^{-1}]$ model when truth is the negative-exponential distribu $\sigma_o^2 = \mathrm{E}(x^2) = 2\lambda^2$, because the expectation of x mus spect to the negative-exponential distribution. We find tha

$$J(\sigma_o^2) = I(\sigma_o^2)\left[\frac{1}{2}\right]\mathrm{E}\left[\frac{\chi^2}{\sigma_o^2} - 1\right]^2 = 2.5$$

Thus when truth is the negative-exponential, and the mo
terrible model in this case), $\mathrm{tr}[J(\sigma_o^2)[I(\sigma_o^2)]^{-1}] = 2.5$ (no
Done the other way around we had the trace < 1. It tur
function under model misspecification can be either above
by situation, and in some situations the trace function ca
a misspecified model.

We can now compare AIC versus TIC model selectio
and truth are negative-exponential or half-normal.
For the negative-exponential model:

$$\mathrm{AIC} = 2n[\log(\overline{x}) + 1] + 2,$$

$$\mathrm{TIC} = 2n[\log(\overline{x}) + 1] + 2\frac{n-1}{n}(\widehat{\mathrm{cv}}(x)$$

$$\mathrm{TICu} = 2n[\log(\overline{x}) + 1] + 2\frac{n}{n-1}(\widehat{\mathrm{cv}}(x)$$

For the half-normal model (using $\hat{\sigma}^2 =$ mean of the x_i^2):

$$\mathrm{AIC} = n\left[\log\left(\hat{\sigma}^2\right) + 1 - \log(2/\pi)\right] + 2,$$

$$\mathrm{TIC} = n\left[\log\left(\hat{\sigma}^2\right) + 1 - \log(2/\pi)\right] + \left[\frac{n-1}{n}\right]$$

$$\mathrm{TICu} = n\left[\log\left(\hat{\sigma}^2\right) + 1 - \log(2/\pi)\right] + \left[\frac{n}{n-1}\right]$$

In both cases here TICu means just that the estimator o
is almost unbiased, as opposed to the direct, biased, plug
needed coefficients of variation.

Table 7.1 shows some results. The point of this brief con
something about AIC versus TIC in a simple setting, espe
they would give greatly different results. The context here
two models are compared. Moreover, one or the other m
data-generating distribution (i.e., truth). We did not cons
so the only possible criterion to use to compare perform
TIC is rate of selection of the true model. We did not wis
any serious evaluation of AIC versus TIC under full-blow
of complex truth, and a set of approximating models, whe
of evaluation is how well a selection procedure does at se
model (technically, we would be selecting the expected K

Several inferences supported by Table 7.1, and by al
examined for this situation, surprised us. For the cas
exponential model is true, the selection results based on

average percent correct is based on equal weighting of the two cas

sample size, n	selection criterion	truth: negative expon.	truth: half-normal	ave: per co
20	AIC	64	85	
	TIC	73	77	
	TICu	75	75	
50	AIC	82	92	
	TIC	87	87	
	TICu	87	87	
100	AIC	93	97	
	TIC	95	95	
	TICu	95	95	
500	AIC	100	100	1
	TIC	100	100	1
	TICu	100	100	1

as good or better than those under AIC. The improvemer
at small sample sizes, wherein an "AIC_c" should be used
for the case that the half-normal model is true, the select
AIC were uniformly as good or better than those under
bias-correction of the trace estimator makes no real diffe
know a priori which (if either) model was true. If we
percent-correct selection based on the idea that we hav
justify any weighting other than a 50:50 weighting of thes
average, no advantage at all for TIC over AIC. Clearly,
extent to which these results would generalize.

7.6.2 *Evaluation Under Logistic Regression*

Logistic regression is used often, therefore we illustrate t
exponential family model, and we explore the above tra
model. Let x_i be a Bernoulli random variable with true pr
1 (and probability $1 - \mu_i$ of being 0). For a sample of n ind
analysis on some assumed model for the μ_i. In order to d
model we adopt the notation for the model as $p_i \equiv p_i(\underline{\theta}$
imposed on these p_i, as a function of a K-dimensional
The relevant pdf, or likelihood (the same notation continu

$$g(\underline{x} \mid \underline{\theta}) = \prod_{i=1}^{n} (p_i)^{x_i} (q_i)^{1-x_i}.$$

We assume that known covariates $\underline{z}_i$, as $K \times 1$ column ve
with each observation, x_i, and an explanatory structural r

$$p_i = \frac{1}{1 + e^{-\underline{z}_i'\underline{\theta}}}, \text{ or } q_i = \frac{e^{-\underline{z}_i'\underline{\theta}}}{1 + e^{-\underline{z}_i'\underline{\theta}}},$$

which is equivalent to

$$\log [p_i/(1 - p_i)] = \underline{z}_i'\underline{\theta}.$$

A modest amount of algebra gives the result

$$g(\underline{x} \mid \underline{\theta}) = \exp \left[\left[\sum_{i=1}^{n} (x_i \underline{z}_i) \right]' \underline{\theta} + \left[\sum_{i=1}^{n} \left(- \log \left(1 \right. \right. \right.$$

which is in the canonical form of the exponential family

$$H(\underline{\theta}) = \sum_{i=1}^{n} \left(- \log \left(1 + e^{\underline{z}_i'\underline{\theta}} \right) \right)$$

and

$$\underline{S} = \sum_{i=1}^{n} (x_i \underline{z}_i) = \sum_{i=1}^{n} \underline{s}_i$$

($G(\underline{\theta}) = 0$). We will need the true expectation of $\underline{S}$:

$$\mathrm{E}_f(\underline{S}) = \left[\sum_{i=1}^{n} (\mu_i \underline{z}_i) \right].$$

Also, from (7.64), $H(\underline{\theta}) = \sum_{i=1}^{n} \log(1 - p_i)$ is an equiva

Two key quantities we need are

$$I(\underline{\theta}) = -\frac{\partial^2 H(\underline{\theta})}{\partial \underline{\theta}^2}$$

(see 7.55) and $-\partial H(\underline{\theta})/\partial \underline{\theta} = \mathrm{E}_f(\underline{S})$. Some straightforwa
to the results

$$-\frac{\partial H(\underline{\theta})}{\partial \underline{\theta}} = \sum_{i=1}^{n} p_i \underline{z}_i$$

and

$$I(\underline{\theta}) = -\frac{\partial^2 H(\underline{\theta})}{\partial \underline{\theta}^2} = \sum_{i=1}^{n} p_i q_i \underline{z}_i \underline{z}_i'.$$

with ith diagonal element $p_i q_i$, and all $n \times K$ matrix Z $\underline{z}_i'$. Then

$$\mathrm{E}_f(\underline{S}) = Z'\underline{\mu},$$
$$-\frac{\partial H(\underline{\theta})}{\partial \underline{\theta}} = Z'\underline{P}(\underline{\theta}),$$

and

$$I(\underline{\theta}) = Z'V_p Z.$$

The MLE $\hat{\underline{\theta}}$ is found by setting (7.65) to $\underline{S}$ and solving the r equations for $\underline{\theta}$, hence solving $\underline{S} = Z'\underline{P}(\hat{\underline{\theta}})$. The true par applies here, given truth $\underline{\mu}$ and the model, is found by equations but with $\underline{S}$ replaced by its true expectation, hen

$$Z'\underline{\mu} = Z'\underline{P}(\underline{\theta}_o),$$

or

$$Z'(\underline{\mu} - \underline{P}(\underline{\theta}_o)) = \underline{0}.$$

In partly nonmatrix notation, we solve

$$\sum_{i=1}^{n}(\mu_i - p_i(\underline{\theta}_o))\underline{z}_i = \underline{0}.$$

If truth $\underline{\mu}$ is not given exactly by the assumed model e $\underline{\mu} = \underline{P}(\underline{\theta}_o)$ will not hold even though the above equation solution in $\underline{\theta}_o$, just as the MLE equations will have a unic

To proceed we also need to know the general formula fo we have

$$J(\underline{\theta}_o) = \mathrm{E}_f\left[\left[\underline{S} - Z'\underline{P}(\underline{\theta}_o)\right]\left[\underline{S} - Z'\underline{P}(\underline{\theta}_o\right.\right.$$

In partly nonmatrix form this formula is

$$J(\underline{\theta}_o) = \mathrm{E}_f\left[\sum_{i=1}^{n}(\underline{s}_i - p_i\underline{z}_i)\right]\left[\sum_{i=1}^{n}(\underline{s}_i - p_\right.$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}_f(\underline{s}_i - p_i\underline{z}_i)(\underline{s}_j - p_j\underline{z}_j)'$$

Here, using $\underline{s}_i = x_i\underline{z}_i$ and $\mathrm{E}_f(x_i) = \mu_i$ the above become

$$J(\underline{\theta}_o) = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}_f(x_i - p_i)(x_j - p_j)[\underline{z}_i\underline{z}_j']$$

$$= \sum_{i=1}^{n} \left[(\mu_i(1-\mu_i)) + (\mu_i - p_i)^2 \right] [\underline{z}_i \underline{z}_i'] + \sum_{i=1}^{n} \sum_{i \neq j}^{n} (\mu_i - p_i)(\mu_j - p_j)[\underline{z}_i \underline{z}_j'].$$

Completing the square in the trailing term above, we get

$$J(\underline{\theta}_o) = \sum_{i=1}^{n} \left[(\mu_i(1-\mu_i)) + (\mu_i - p_i)^2 \right] [\underline{z}_i \underline{z}_i'] + \left[\sum_{i=1}^{n} (\mu_i - p_i(\underline{\theta}_o)) \underline{z}_i \right] \left[\sum_{i=1}^{n} (\mu_i - p$$
$$- \sum_{i=1}^{n} \left[(\mu_i - p_i)^2 \right] [\underline{z}_i \underline{z}_i'].$$

The middle term of the above is zero because of the equa
the third term cancels with part of the first term, so we ha

$$J(\underline{\theta}_o) = \sum_{i=1}^{n} \mu_i(1-\mu_i)[\underline{z}_i \underline{z}_i'],$$

or in pure matrix terms,

$$J(\underline{\theta}_o) = Z' V_\mu Z.$$

Here, V_μ is an $n \times n$ diagonal matrix with ith diagonal
Contrast (7.67) to $I(\underline{\theta}_o) = Z' V_p Z$.

It is easy, but not very informative, now to write

$$\operatorname{tr}\left[J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1} \right] = \operatorname{tr}\left[(Z' V_\mu Z)(Z' V_p Z)^{-1} \right]$$
$$= K + \operatorname{tr}\left[(Z'(V_\mu - V_p)Z)(Z \right.$$

The above makes it easier to realize that the trace term is
$\mu_i(1-\mu_i) = p_i(\underline{\theta}_o)(1 - p_i(\underline{\theta}_o))$. However, these equalit
yet we can still get $\operatorname{tr}\left[J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1} \right] = K$; hence this lat
with a model that does not match truth, i.e., where $g \subset f$ v

The above results are totally general, so they apply to
w replicate observations are taken at each of $j = 1, \ldots$
The total sample size is then $n = r * w$, but we will h
values of μ_j to specify for truth and only r values of p
der any model. Hence, to gain some insights here we us
$\log\left[p_j/(1-p_j) \right] = a + bj$ for $j = 1, \ldots, r$, with w re
at each j. Thus $K = 2$, $\underline{\theta} = (a, b)'$, and $\underline{z}_j = (1, j)'$. I

j. However, the results so derived apply reasonably well
$n = r * w$ would be "large," say 100 or more (given $K =$
the trace term under this use of a simple logistic regress
only specify a set of $\mu_1, \ldots, \mu_r$, solve

$$\sum_{i=1}^{r} (\mu_i - (a_o + b_o i))\underline{z}_i = \underline{0}$$

for $\underline{\theta}_o = (a_o, b_o)'$, and compute $J(\underline{\theta}_o)$ (7.67), $I(\underline{\theta}_o$
$\text{tr}\left[J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1}\right]$ (also denoted by "bias"). In doing this
of μ that were near to fitting the logistic structural model,
a $\underline{p}(\underline{\theta})$ vector that fit the model, then perturbing some (
by starting with $\mu_j = j/(r+1)$, which is not a logistic
model but is not too far from fitting such a structural mod

With at most modest deviation of truth from any actua
gression model structure we found that the trace term valu
(between about 1.8 and 2, sometimes going a little above 2
case of truth being the simple linear model ($\mu_j = j/(r$
varied monotonically from 1.98 at $r = 5$, to 1.91 at $r =$
some results for $r = 10$, based on truth being perturb
logistic model $\text{logit}(p_j) = 3.0 - 0.5j$. The first line of Ta
μ_j computed from this model (scaled by 1,000).

What one can see in Table 7.2 (and other computations
this) is that the true μ_j have to be here a very poor appro
simple logistic model before the trace term deviates much
if the data seem at all well fit by a logistic model, then the
opposed to any attempted estimation of the trace) seems
is especially important here because V_μ, hence $J(\underline{\theta}_o)$, ca
all unless there is replication at each $\underline{z}_i$, and there would n
such replication; this condition rarely occurs with logistic

Formula (7.68) was corroborated by direct Monte Ca
target bias (trace term) for a few cases in Table 7.2. The
and hence most direct, way to do this is to evaluate using
of

$$\text{bias} = \text{E}_{\underline{x}}\text{E}_{\hat{\underline{\theta}}(\underline{y})}\left[\log(g(\underline{x} \mid \hat{\underline{\theta}}(\underline{y})))\right] - \text{E}_{\underline{x}}\left[\log(g(\underline{x} \mid \underline{\theta}$$

Hence for one Monte Carlo replicate (generating *iid* $\underline{x}$ an

$$\widehat{\text{bias}} = \log(g(\underline{x} \mid \hat{\underline{\theta}}(\underline{y}))) - \log(g(\underline{x} \mid \hat{\underline{\theta}}(\underline{x})$$

Averaged over many samples (m), if large sample size n
$\widehat{\text{bias}}$ will equal $\text{tr}\left[J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1}\right]$.

For many models the first term on the right-hand side of
in $\underline{x}$, so we can analytically take the expectation with re

$3 - 0.5j$), values of μ_j are shown, scaled by 1,000; case one (i.e., t
the logistic model, but none of the other cases are a perfect fit to th
The results are reasonably applicable if w is at least 10 or 20.

μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	μ_9
924	881	818	731	622	500	378	269	182
900	900	900	900	622	500	378	269	182
924	881	970	757	725	530	410	231	186
900	944	898	773	666	522	349	264	176
864	993	971	821	564	514	366	236	150
894	990	870	826	583	457	372	252	216
924	720	818	796	749	632	492	188	260
924	881	768	831	622	400	378	269	282
924	952	650	838	638	448	278	189	204
874	874	874	558	558	558	558	190	190
924	881	818	731	400	600	378	269	182
924	881	818	731	300	700	378	269	182
924	881	818	731	622	500	378	269	182
924	881	568	831	622	100	378	269	432
924	881	818	731	200	800	378	269	182
674	981	818	731	622	500	128	669	182
924	881	818	731	622	500	900	900	900

logistic example we thus get, expressed in basic form,

$$\widehat{\text{bias}} = \sum_{i=1}^{n} \mathrm{E}_{\underline{y}} \Big[n_i \mu_i \log(p_i(\hat{\underline{\theta}})) + n_i(1-\mu_i) \log$$

$$- \sum_{i=1}^{n} \mathrm{E}_{\underline{y}} \Big[y_i \log(p_i(\hat{\underline{\theta}})) + (n_i - y_i) \log$$

Here the MLE $\hat{\underline{\theta}}$ is based on data $\underline{y}$.

For direct Monte Carlo evaluation we used replicate
noted above with the same number of replicates w for each
we generated a large number m of independent samples
model, fit the model-based MLE to each sample, compute
by sample, and got its average and empirical standard err

All this is quite obvious; where we are going here is th
Carlo evaluation is poor in the sense of needing a huge
The problem is that as w (i.e., n) increases, the number o
get a small standard error (like 0.005) on the estimate bia
the variance of $\widehat{\text{bias}}$, for one sample, increases with increas
that variance can be quite large. For a large sample size n (

example, for the case in Table 7.2 where the trace is comp
$w = 100$ (hence sample size $n = 1{,}000$ Bernoulli trials) f
set of 10,000 ($= m$) Monte Carlo samples we got averag
an estimated standard error of 0.135. Other runs verifie
one million Monte Carlo samples to get a standard error
of about 0.014 with $w = 100$ in this example. But we mi
for the trace formula to apply exactly; for $w = 1{,}000$ and
samples we got the average $\widehat{\text{bias}} = 1.982$ with $\widehat{\text{se}} = 0.42$
is the reverse of what we expect; i.e., we expect to get incr
the same number, m, of Monte Carlo samples) as sample s
reverse phenomenon occurs here because the expected diff
involved in direct computation of $\widehat{\text{bias}}$ (i.e., (7.70)) is con
sample size n, but the variance of each of those two likelih
proportional to n and the two terms are not highly correla
size n increases, the precision of the estimated bias, give
Monte Carlo samples (m), actually decreases. So to eva
brute-force approach, the adequacy of the trace term ap
sample sizes, it takes a huge number of Monte Carlo sam

With models that are in the exponential family there i
to do exact Monte Carlo evaluation of the bias that mus
the maximized log-likelihood for exact K-L based mode
(7.54) is an exact result for any sample size:

$$\text{bias} = \text{tr}\left[\text{COV}(\underline{\hat{\theta}}, \underline{S})\right].$$

While $\underline{\theta}$ and $\underline{S}$ are only for the canonical form of the n
apply for any parametrization of the assumed model beca
of the result to 1-to-1 transformations of $\underline{\theta}$ (see end of Se
alternative Monte Carlo evaluation is simply to take for eac
computed MLE and minimal sufficient statistic and, fro
of size m, estimate the covariances $\text{cov}(\hat{\theta}_i, S_i)$, $i = 1, \ldots$
these K estimates. The result is $\widehat{\text{bias}}$, and this approach is
For the same case in Table 7.2 (i.e., trace $= 1.774$), usin
hence $n = 1{,}000$) and 10,000 Monte Carlo samples we g
its $\widehat{\text{se}} = 0.019$ using the covariance approach. Based on
there was a clear suggestion that $w = 100$ was not quite
trace (7.68) to apply reliably to three digits (it was then re
Using $w = 1{,}000$ and 10,000 Monte Carlo samples we
$\widehat{\text{se}} = 0.017$. This result held up on more study: (7.68) see
for $w = 1{,}000$ (which here meant $n = 10{,}000$).

As another example consider the last case in Table 7.2,
($=$ bias). Using $w = 100$, for one run of 10,000 Monte Ca
such run made) we got the direct result based on (7.70) as

exponential family, Monte Carlo or bootstrap evaluation
trace term should be based on (7.54).

It is worth noting a basis for the estimated standa
$\widehat{\text{tr}}\left[\text{COV}(\underline{\hat{\theta}}, \underline{S})\right]$. For the point estimate, use all the sin
compute means; then for component i,

$$\widehat{\text{cov}}(\hat{\theta}_i, S_i) = \frac{\sum_{j=1}^{m}(\hat{\theta}_{i,j} - \overline{\hat{\theta}}_i)(S_{i,j} - \overline{S}_i}{m-1}$$

and

$$\widehat{\text{bias}} = \sum_{i=1}^{K} \widehat{\text{cov}}(\hat{\theta}_i, S_i).$$

However, to estimate the standard error we must partition t
say into 25 equal-sized subsets (for $m = 10{,}000$ then e
400). Compute by the above formulas $\widehat{\text{bias}}_s$ for each sub
the standard error of $\widehat{\text{bias}}$ from these 25 independent esti
will almost equal $\widehat{\text{bias}}$, but will not be equal due to nonlin

The standard error of $\widehat{\text{bias}}$ from this covariance app
function of data sample size n because of how the produ
It suffices to consider the product $(\hat{\theta}_i - \theta_{o,i})(S_i - \text{E}_f(S_i))$
i). This product has variance virtually independent of n b
converges (in n) at rate proportional to $1/\sqrt{n}$, while the se
at rate proportional to $\sqrt{n}$. As a result, the standard erro
based bias (hence trace) estimator is almost independent
is much better behavior (as a function of n) than the s
estimator of bias based directly on the likelihood functio
(i.e., (7.69)) also requires more calculations beyond first

7.6.3 *Evaluation Under Multinomially Distribut*

We here assume that we have count data $n_1, \ldots, n_r$ that su
n. Truth is the multinomial distribution mult$(n, \mu_1, \ldots,$
bilities μ_i summing to 1, and $0 < \mu_i < 1$. To know trut
only need to know the true μ_i (assuming that the count
distributed; they could have overdispersion, which violat
We might totally fail to know how these true probabilitie
relation to any explanatory variables, or what would hap
defined in some other way. Thus, deeper truth may exist
tion, but it is irrelevant to model selection purposes once
to a particular multinomial setting.

not to be considered, since then the fitted model matches
being a perfect match to the data). The theory in Section
used; note that here

$$\log(g(\underline{n} \mid \underline{\theta})) = \sum_{i=1}^{r} n_i \log(p_i(\underline{\theta})).$$

First, $\underline{\theta}_o$ is determined as the solution to (7.5), which here

$$\sum_{i=1}^{r} \frac{\mu_i}{p_i(\underline{\theta}_o)} \frac{\partial p_i(\underline{\theta}_o)}{\partial \underline{\theta}} = \underline{0}.$$

In (7.71) if we replace μ_i by n_i, we have the likelihood
can treat the μ_i as data and find $\underline{\theta}_o$ by MLE methods. Ec
MLE when the data are replaced by their true expected va

Second, applying (7.7) we directly get

$$I(\underline{\theta}_o) = n \left[\sum_{i=1}^{r} \frac{\mu_i}{[p_i(\underline{\theta}_o)]^2} \left(\frac{\partial p_i(\underline{\theta}_o)}{\partial \underline{\theta}} \right) \left(\frac{\partial p_i(}{\partial \underline{\theta}} \right. \right.$$
$$\left. \left. - n \left[\sum_{i=1}^{r} \frac{\mu_i}{p_i(\underline{\theta}_o)} \left(\frac{\partial^2 p_i(\underline{\theta}_o)}{\partial \underline{\theta}^2} \right) \right] \right].$$

Finally, applying the definition in (7.12), we have

$$J(\underline{\theta}_o) = \mathrm{E}_f \left[\sum_{i=1}^{r} \frac{n_i}{p_i(\underline{\theta}_o)} \left(\frac{\partial p_i(\underline{\theta}_o)}{\partial \underline{\theta}} \right) \right] \left[\sum_{i=1}^{r} \frac{n_i}{p_i(\underline{\theta}_o)} \right.$$

The evaluation of $J(\underline{\theta}_o)$ does take some algebra and kno
nomial distribution, but it is mostly a straightforward exe
the result:

$$J(\underline{\theta}_o) = n \left[\sum_{i=1}^{r} \frac{\mu_i}{[p_i(\underline{\theta}_o)]^2} \left(\frac{\partial p_i(\underline{\theta}_o)}{\partial \underline{\theta}} \right) \left(\frac{\partial p_i(}{\partial \underline{\theta}} \right. \right.$$

Define the matrix A as

$$A = n \left[\sum_{i=1}^{r} \frac{\mu_i}{p_i(\underline{\theta}_o)} \left(\frac{\partial^2 p_i(\underline{\theta}_o)}{\partial \underline{\theta}^2} \right) \right],$$

and we have $I(\underline{\theta}_o) = J(\underline{\theta}_o) - A$. Furthermore, if the
$\mu_i = p_i(\underline{\theta}_o)$, and A reduces to the null matrix; hence ther

Using these results we can write

$$\mathrm{tr}\left[J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1} \right] = \mathrm{tr}\left[(J(\underline{\theta}_o) - A + A)[I(\right.$$

For a long time a nagging question for us was whether t
always be either $> K$ or $< K$ when the model did not
yet the model is logically known to be simpler than tr
some general sense). Stated differently, if the Kullback–
is positive, i.e., K-L $= I(f, g) > 0$, then must tr $\left[J(\underline{\theta}_o)\right.$
maybe $< K$) always occur when the model is some form
(hence the model can be said to approximate, but not equa
is no, as was indicated by the logistic regression exampl
however, a more convincing answer is given here: The trac
or $< K$ and there need be no consistency as to which v
question also explored below is, If tr $\left[J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1}\right] = K$
That answer is also no.

Because the cell probabilities sum to 1, the sum of the
partials is the null matrix, O, of all zeros (the vector of fi
to a null vector). Therefore, an equivalent expression for

$$A = n\left[\sum_{i=1}^{r} \frac{\mu_i - p_i(\underline{\theta}_o)}{p_i(\underline{\theta}_o)}\left(\frac{\partial^2 p_i(\underline{\theta}_o)}{\partial \underline{\theta}^2}\right)\right.$$

The weights in this linear combination of second partial
must be either identically zero (hence K-L is 0), or son
some positive. This would suggest that A might not alw
sign, unless the second partials are very strangely relate
truth. But a more detailed case is need to get an example,
use $K = 1$ for an example, such as by using a binomial m

Let us further assume that the data arise from n indepe
integer random variable, y, taking values 0 to $r - 1$. Th
the frequency counts n_i of times $y = i - 1$. A very simp
probabilities, μ_i, is thus to assume that this underlying m
binomial random variable. This corresponds to imposing
multinomial cells, without loss of generality, and thus the
probability is

$$p_i(\theta) = \binom{r-1}{i-1}\theta^{i-1}(1-\theta)^{r-i}, \qquad i = 1$$

Thus we have, as our model, an assumed underlying binor
$y \sim \text{bin}(r - 1, \theta)$ and a random sample of size n of th
In fact, y $(= 0, 1, \ldots, r - 1)$ has the distribution given
true distribution. We will need the functions below, involv
partial derivatives:

$$P1_i = \frac{1}{p_i(\theta)}\left(\frac{\partial p_i(\theta)}{\partial \theta}\right) = \frac{(i-1) - \theta(r-1)}{\theta(1-\theta)},$$

$$- \left[\frac{(i-1)(1-2\theta) -}{(\theta(1-\theta}$$

We solve (7.71), which is $\sum \mu_i P1_i = 0$, to find θ_o same process as finding an MLE (again, the only tricky a assumed here):

$$\theta_o = \frac{\sum_{i=1}^{r} \mu_i (i-1)}{r-1}.$$

This θ_o is the true expected value of $y/(r-1)$ regardless of We compute (7.72) as $n \sum \mu_i (P1_i)^2$:

$$J(\theta_o) = n \left[\sum_{i=1}^{r} \mu_i \left(\frac{(i-1) - \theta_o(r-1)}{\theta_o(1-\theta_o)} \right)^2 \right] \equiv n\mathrm{E}_f$$

We find matrix A as $n \sum \mu_i P2_i$:

$$A = J(\theta_o) - \frac{n(r-1)}{\theta_o(1-\theta_o)};$$

hence

$$I(\theta_o) = \frac{n(r-1)}{\theta_o(1-\theta_o)}.$$

It is now easy to find the trace:

$$\mathrm{tr}\left[J(\theta_o)[I(\theta_o)]^{-1} \right] = \sum_{i=1}^{r} \mu_i \frac{[(i-1) - \theta_o(r}{(r-1)\theta_o(1 -}$$

For the case of $\mu_i = p_i(\theta_o)$, then (7.73) is 1 (this can b hence using $\mu_i \equiv p_i(\theta_o) + (\mu_i - p_i(\theta_o))$ in (7.73) we obt

$$\mathrm{tr}\left[J(\theta_o)[I(\theta_o)]^{-1} \right] = 1 + \sum_{i=1}^{r} (\mu_i - p_i(\theta_o)) \frac{[(i-1)}{(r-1}$$

whereupon it should be essentially obvious that the tern can be either positive or negative. However, we will give mostly for $r = 3$ because this is the smallest r we can u here, and small r is desirable when we need to display tru

Our model is thus bin$(2, \theta)$; hence $p_1 = (1-\theta)^2$, p $p_3 = \theta^2$. The approach is to specify the μ_i and compute and from (7.73), for $r = 3$,

$$\mathrm{tr}\left[J(\theta_o)[I(\theta_o)]^{-1} \right] = \frac{4\mu_1(\theta_o)^2 + \mu_2(1-2\theta_o)^2 + 4\mu_}{2\theta_o(1-\theta_o)}$$

cases of truth (the μ_i) in relationship to the K-L best appr
model. Case one exactly fits a binomial model. Cases two
ribly by even the K-L best approximating binomial mode
$\theta_o = 0.5$. In what is below, "Trace" means the value comp
$\theta_o = 0.5$, and K-L is the Kullback–Leibler information c
truth and the best approximating binomial model ("Bia
below):

μ_1	μ_2	μ_3	Trace	K-L	Bias-l
0.25	0.50	0.25	1.000	0.000	1.00
0.05	0.90	0.05	0.200	0.368	0.19
0.45	0.10	0.45	1.800	0.368	1.81

(we note that in this situation the trace term (7.73) seems t
by 2). Clearly, this bias-correction trace term can be eithe
than 1 when the model does not match truth. This is bec
variance of y can be either larger or smaller than the
variance for y implied by the K-L best-fitting binomial m

We build on this example by doing an exact Monte Ca
expected log-likelihood and the K-L–based target mode
to verify the asymptotic derivation of the bias as being th
above, "Bias-MC" denotes the results (accurate to two
sample size $n = 200$, from one million Monte Carlo sam
bias that the trace term measures based on asymptotic the

For a truth that cannot be well approximated here by a
clear that the trace (equation 7.74) can be far from 1. Rat
example for models that are arbitrarily poor (like cases tw
we should consider models that are closer to truth, becaus
the term $-2\log(\mathcal{L})$ will prevent the selection of really poo
those models a choice between the use of K or $\widehat{\text{trace}}$ is irr
models has some good candidates.

So we looked at one set of cases where a binomial mo
wrong to use. We chose a θ, generated $p_1 = (1-\theta)^2$, p
$p_3 = \theta^2$, then perturbed these cell probabilities to get a
to a binomial model by setting $\mu_i = p_i + \epsilon_i$, where $\epsilon_i =$
uniform$(-h, h)$. Inadmissible sets of μ_i were not generate
μ_2, and μ_3, (7.74) was evaluated; thus, this is not a Monte
we use Monte Carlo methods only as a convenience in ge
μ_i that are close to a binomial model.

For $\theta = 0.5$ and $h = 0.1$ (and 1,000 generated sets of t
lowing results for the trace given by (7.74): min $= 0.$
and mean $= 0.996$. These results support practical us
$\widehat{\text{tr}}\left[J(\theta_o)[I(\theta_o)]^{-1}\right]$. However, it is fair to ask about estim

general, the true trace value is known for these cases, the mean a
(7.75) are given based on one million samples.

μ_1	μ_2	μ_3	n	Trace	Mean	S
0.25	0.50	0.25	50	1.000	0.990	
0.25	0.50	0.25	100	1.000	0.995	
0.25	0.50	0.25	200	1.000	0.997	
0.04	0.32	0.64	200	1.000	0.997	
0.20	0.55	0.25	50	0.897	0.887	
0.20	0.60	0.20	50	0.800	0.790	
0.30	0.40	0.30	50	1.200	1.190	
0.30	0.45	0.25	50	1.098	1.088	
0.05	0.90	0.05	200	0.200	0.199	
0.45	0.10	0.45	200	1.800	1.799	

here (hence using TIC), as can be done by plugging $\hat{\theta}_o$
(7.74); after simplification,

$$\widehat{\mathrm{tr}}\left[J(\theta_o)[I(\theta_o)]^{-1}\right] = \frac{4\hat{\mu}_1(\hat{\theta}_o)^2 + \hat{\mu}_2(1-2\hat{\theta}_o)^2 + 4\hat{\mu}}{2\hat{\theta}_o(1-\hat{\theta}_o)}$$

A small Monte Carlo evaluation of this estimator was
it was badly biased or highly variable. Variables in this st
and sample size n. Results, given in Table 7.3 based on
are the theoretical trace value, and the mean and standar
evaluated by simulation, accurate to two decimal places
trace is 1, then the binomial distribution is truth; otherwis

From Table 7.3 it appears that the trace estimator has g
is reasonable to consider using TIC rather than AIC; at l
of the two seems worth doing here. For the sets of true
we compared AIC to TIC for the binomial model ($K =$
model, R) and the parameter-saturated general model (
corresponding maximized likelihoods be $\mathcal{L}_R$ and $\mathcal{L}_G$. He

$$\begin{aligned}
\mathrm{AIC}_R &= -2\log(\mathcal{L}_R) + 2,\\
\mathrm{AIC}_G &= -2\log(\mathcal{L}_G) + 4,\\
\mathrm{TIC}_R &= -2\log(\mathcal{L}_R) + 2\,\widehat{\mathrm{tr}}\left[J(\theta_o)[I(\theta_o)]^{-1}\right],\\
\mathrm{TIC}_G &= -2\log(\mathcal{L}_G) + 4,
\end{aligned}$$

where $\widehat{\mathrm{tr}}\left[J(\theta_o)[I(\theta_o)]^{-1}\right]$ is given by (7.75). Because there
here, and because we want to keep matters simple, we just
methods based on how often they selected the same m
they selected the correct data-generating model. Results
Monte Carlo samples, which suffices here to get stand

generating model varied (R for the binomial, G for the multinom
on 10,000 samples; column AIC (or TIC) denotes the proportion
TIC) selected the correct data-generating model; column "Match"
selected the same model whether or not it was the data-generating
details).

μ_1	μ_2	μ_3	n	Truth	AIC	TIC
0.25	0.50	0.25	50	R	0.83	0.84
0.25	0.50	0.25	100	R	0.85	0.84
0.25	0.50	0.25	200	R	0.84	0.84
0.04	0.32	0.64	200	R	0.84	0.83
0.20	0.55	0.25	50	G	0.69	0.76
0.20	0.60	0.20	50	G	0.57	0.53
0.30	0.40	0.30	50	G	0.46	0.53
0.30	0.45	0.25	50	G	0.24	0.28
0.05	0.90	0.05	200	G	1.00	1.00
0.45	0.10	0.45	200	G	1.00	1.00
				means	0.73	0.75

for estimated proportions. "Truth" denotes the correct ge
"AIC" and "TIC" columns denote the proportion of sam
methods selected the correct model. "Match" denotes the p
in which both methods selected the same model, regardle
was.

We looked at many more results than are given above
TIC in this limited context; there was then no change fr
obvious conclusion: no meaningful difference in perfor
versus TIC. More study is surely warranted; this limited
spirit that maybe something dramatic would result. It di
conclusion (based on all the considerations we have don
just those of this section), it seems that simplicity strongl
over TIC.

We return to an interesting theoretical question posed ab
if $f = g$, then $\text{tr}\left[J(\theta_o)[I(\theta_o)]^{-1}\right] = K$. However, if $\text{tr}\left[J$
does this mean $f = g$?

A counterexample shows that the assertion is false; henc
wherein truth is more complex than the models used for a
is appropriate to use (as opposed to TIC, which would th
just estimating K).

For this multinomial context we have shown that $I(\underline{\theta}_o)$
siderations here are for any value of r). So if $A = O$ (i.
$I(\underline{\theta}_o) = J(\underline{\theta}_o)$ and $\text{tr}\left[J(\theta_o)[I(\theta_o)]^{-1}\right] = K$ regardless o

$$A = n\left[\sum_{i=1}^{r} \frac{\mu_i}{p_i(\underline{\theta}_o)}\left(\frac{\partial^2 p_i(\underline{\theta}_o)}{\partial \underline{\theta}^2}\right)\right];$$

therefore, if all second partial derivatives of the model c
zero, we do get $A = O$. This will occur for any linea
probabilities; that is, $p_i(\underline{\theta}) = \underline{x}_i'\underline{\theta}$ for a set of known ve
such models are discouraged because they can generate
out of range.

As an example we revert to the case of $r = 3$ and use
$\mu_1 = \mu_3 = \theta/2$ and $\mu_2 = 1 - \theta$; so $\log(g(\underline{n} \mid \underline{\theta})) = (n_1 + n_3)\log(\theta/2) + n_2 \log(1-\theta)$. Here, $\theta_o = 1 - \mu_2$. Upon computing $I(\underline{\theta}_o)$
basic definitions, we do in fact get $I(\underline{\theta}_o) = J(\underline{\theta}_o) = n/[\theta_o$
of the values of the μ_i. This means that here is a situation
AIC rather than TIC is the correct selection procedure ev
does differ from truth (i.e., $f \subseteq g$ is not true, yet this co
cited as always required for the theoretical validity of AI

For the case of general r and the binomial model w
investigate this trace term and AIC versus TIC. But even
to derive any insights from it directly, and numerical met
all we really need are usable computational formulas to
estimate $\text{tr}[J(\theta_o)[I(\theta_o)]^{-1}]$. We can get the needed formul
model for multinomial data. First, we can find $\hat{\underline{\theta}}_o$ by solv

$$\sum_{i=1}^{r} \frac{n_i/n}{p_i(\underline{\theta}_o)} \frac{\partial p_i(\underline{\theta}_o)}{\partial \underline{\theta}} = \underline{0},$$

which we do anyway, since this is just our MLE of $\underline{\theta}$
model. We do have to compute the set of first and secon
of the model cell structures evaluated at the MLE, but ev
numerically. Thus we can get, hence use and explore, TI

$$\hat{I}(\underline{\theta}_o) = n\left[\sum_{i=1}^{r} \frac{n_i/n}{[p_i(\hat{\underline{\theta}}_o)]^2}\left(\frac{\partial p_i(\hat{\underline{\theta}}_o)}{\partial \underline{\theta}}\right)\left(\frac{\partial p_i(\underline{\theta}}{\partial \underline{\theta}}\right.\right.$$
$$\left.\left. - n\left[\sum_{i=1}^{r} \frac{n_i/n}{p_i(\hat{\underline{\theta}}_o)}\left(\frac{\partial^2 p_i(\hat{\underline{\theta}}_o)}{\partial \underline{\theta}^2}\right)\right]\right.,$$
$$\hat{J}(\underline{\theta}_o) = n\left[\sum_{i=1}^{r} \frac{n_i/n}{[p_i(\hat{\underline{\theta}}_o)]^2}\left(\frac{\partial p_i(\hat{\underline{\theta}}_o)}{\partial \underline{\theta}}\right)\left(\frac{\partial p_i(\underline{\theta}}{\partial \underline{\theta}}\right.\right.$$

Clearly, we can also compute theoretical values of thes
postulated truth and model. Such studies would be inform
the intention of this book.

so under the philosophy we espouse here. Models used
not truth; full truth is very complex; one's analytic goal
best approximating fitted model. deLeeuw concludes that
(essentially, compelling) explicit model selection criteri
particular, he says (deLeeuw 1988:132), "This gives a justi
AIC, even if the model is not true." On an important relate
worth quoting deLeeuw (1988:136–137): "The independe
example, which is at the basis of most work in statisti
falsified. As we have seen, the independence assumption
with a particular framework of replication, for which we ha
it is relevant or not."

7.6.4 Evaluation Under Poisson-Distributed Da

The purpose of this subsection is to see whether a result fo
data extends to the Poisson-distributional case. We assum
of observed Poisson counts with unknown means μ_i (=
the data are Poisson distributed). The model for these me
$\ldots, r$. Some results:

$$\log(g(\underline{n} \mid \underline{\theta})) = \sum_{i=1}^{r} \left[-\lambda_i(\underline{\theta}) + n_i \log(\lambda_i(\underline{\theta}\right.$$

$\underline{\theta}_o$ is determined as the solution to

$$\sum_{i=1}^{r} \left[\frac{\mu_i}{\lambda_i(\underline{\theta}_o)} - 1\right] \frac{\partial \lambda_i(\underline{\theta}_o)}{\partial \underline{\theta}} = \underline{0};$$

$$I(\underline{\theta}_o) = \left[\sum_{i=1}^{r} \frac{\mu_i}{[\lambda_i(\underline{\theta}_o)]^2} \left(\frac{\partial \lambda_i(\underline{\theta}_o)}{\partial \underline{\theta}}\right) \left(\frac{\partial \lambda_i(\underline{\theta}\right.\right.$$

$$- \left[\sum_{i=1}^{r} \left(\frac{\mu_i}{\lambda_i(\underline{\theta}_o)} - 1\right) \left(\frac{\partial^2 \lambda_i(\underline{\theta}_o)}{\partial \underline{\theta}^2}\right.\right.$$

$$J(\underline{\theta}_o) = \left[\sum_{i=1}^{r} \frac{\mu_i}{[\lambda_i(\underline{\theta}_o)]^2} \left(\frac{\partial \lambda_i(\underline{\theta}_o)}{\partial \underline{\theta}}\right) \left(\frac{\partial \lambda_i(\underline{\theta}\right.\right.$$

We define the $K \times K$ matrix B as

$$B = \sum_{i=1}^{r} \left(\frac{\mu_i}{\lambda_i(\underline{\theta}_o)} - 1\right) \left(\frac{\partial^2 \lambda_i(\underline{\theta}_o)}{\partial \underline{\theta}^2}\right);$$

then $I(\underline{\theta}_o) = J(\underline{\theta}_o) - B$ and $\text{tr}\left[J(\theta_o)[I(\theta_o)]^{-1}\right] = K$ -
matrix B is zero, then regardless of how much μ_i and $\lambda_i(\underline{\theta}_o$
of these values (hence $I(f, g) > 0$ occurs), we still have t

7.6.5 *Evaluation for Fixed-Effects Normality-Ba Models*

The fixed-effects linear model based on n *iid* normally di so common that it seems almost mandatory that we conside under this model for some tractable "truth." The model i multivariate-normal($\underline{0}, \sigma^2 I$), and without loss of general matrix X is assumed of full rank. Truth has a structural c (which can be estimated by $\underline{Y}$) and a stochastic compon distributed in some unknown way, the properties of which without strong assumptions. If the model is truth, then $\underline{\mu} = X\underline{\beta}$) and distributional assumptions of the model are

In reality, the ϵ_i may not be independent, may not be id and may not be normally distributed. In fact, they may not e some or all ϵ_i are zero with probability 1. In this latter cas tic; that is, there is some sufficiently complex computing formula, with many covariates) such that if we knew that predict $\underline{Y}$ with certainty (measurement error would bec fore this level of model accuracy was reached). Hence, fo cannot, for cases of real data, evaluate K-L–based model of continuous random variables.

We can, however, derive informative results under gen that are better approximations to reality (by assumption be used for data analysis. Therefore, we assume here th $\underline{\epsilon}$, $\underline{\epsilon} \sim$ multivariate-normal($\underline{0}, \tau^2 I$), where τ^2 may be z even drop the full distributional assumption, as we will because the relevant evaluations require only the first four distribution. More generally, results could be gotten und $\underline{\epsilon} \sim$ multivariate-normal($\underline{0}, \Sigma$) for given Σ, but the more will suffice.

We first need basic notation and results: $\underline{\theta}$ denotes the K for $\underline{\beta}$ a $(K-1) \times 1$ vector of the structural parameters. W the parameter to estimate. The model pdf for the data is

$$g(\underline{y} \mid \underline{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\underline{Y} - X\underline{\beta})'(\underline{Y} -$$

and we take, without loss of generality,

$$\log(g(\underline{y} \mid \underline{\theta})) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\underline{Y} - X\underline{\beta})'$$

$$\frac{\partial \log(g(\underline{y} \mid \underline{\theta}))}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\underline{Y} - X\underline{\beta})'(\underline{Y}$$

As per theory, we take the expectations of (7.76) and (7 what is here just assumed truth, f (evaluation under abso impossible) to get the equation

$$\frac{1}{\sigma_o^2} X'(\underline{\mu} - X\underline{\beta}_o) = \underline{0}$$

from (7.76), and then from (7.77) we derive

$$\begin{aligned} -\frac{n}{2\sigma_o^2} + \frac{1}{2(\sigma_o^2)^2} \mathrm{E}_f(\underline{Y} - \underline{\mu} + \underline{\mu} - X\underline{\beta}_o)'(\underline{Y} - \underline{\mu} & \\ = -\frac{n}{2\sigma_o^2} + \frac{1}{2(\sigma_o^2)^2} \left[\mathrm{E}_f(\underline{\epsilon}'\underline{\epsilon}) + (\underline{\mu} - X\underline{\beta}_o)' \right. & \\ = -\frac{n}{2\sigma_o^2} + \frac{1}{2(\sigma_o^2)^2} \left[n\tau^2 + \|\underline{\mu} - X\underline{\beta}_o\|^2 \right] = & \end{aligned}$$

It is now a simple matter to find

$$\begin{aligned} \underline{\beta}_o &= (X'X)^{-1}X'\underline{\mu}, \\ \sigma_o^2 &= \tau^2 + \frac{\|\underline{\mu} - X\underline{\beta}_o\|^2}{n}. \end{aligned}$$

These parameter values define the vector $\underline{\theta}_o = (\underline{\beta}_o', \sigma_o^2)'$. F that lack-of-fit variation, from the assumed structural mo of residual (unexplained) variation.

To find $I(\underline{\theta}_o)$ we need the expected second mixed parti (7.77), as follows:

$$\mathrm{E}_f \left[-\frac{\partial^2 \log(g(\underline{y} \mid \underline{\theta}_o))}{\partial \underline{\beta}^2} \right] = \frac{1}{\sigma_o^2} X'X,$$

$$\mathrm{E}_f \left[-\frac{\partial^2 \log(g(\underline{y} \mid \underline{\theta}_o))}{\partial \sigma^2 \partial \sigma^2} \right] = -\frac{n}{2(\sigma_o^2)^2} + \frac{1}{(\sigma_o^2)^3} n\sigma_o^2$$

and

$$\mathrm{E}_f \left[-\frac{\partial^2 \log(g(\underline{y} \mid \underline{\theta}_o))}{\partial \underline{\beta} \partial \sigma^2} \right] = \frac{1}{\sigma_o^2} X'(\underline{\mu} - X\underline{\beta}_o) = \underline{0}$$

$$I(\underline{\theta}_o) = \begin{bmatrix} \frac{1}{\sigma_o^2} X'X & \underline{0} \\ \underline{0}' & \frac{n}{2(\sigma_o^2)^2} \end{bmatrix}.$$

The evaluation of $J(\underline{\theta}_o)$ is harder, and more depende
Evaluation of $I(\underline{\theta}_o)$ required only the second moment of f
of $J(\underline{\theta}_o)$ also requires third and fourth moments. Both deriv
on the independence of the ϵ_i. The upper left $(K-1)\times($
$J(\underline{\theta}_o)$ is

$$\mathrm{E}_f\left[\frac{1}{(\sigma_o^2)^2} X'(\underline{Y} - X\underline{\beta}_o)(\underline{Y} - X\underline{\beta}_o)'X\right] = \frac{\tau}{(\sigma}$$

The last $(K-1)\times 1$ column vector (of the first $K-1$ ro

$$\mathrm{E}_f\left[\frac{1}{\sigma_o^2} X'(\underline{Y} - X\underline{\beta}_o)\left[-\frac{n}{2\sigma_o^2} + \frac{1}{2(\sigma_o^2)^2}(\underline{Y} - X\underline{\beta}_o)'\right.\right.$$

Making use of $X'(\underline{\mu} - X\underline{\beta}_o) = \underline{0}$, and some algebra
above to

$$\mathrm{E}_f\left[\frac{1}{2(\sigma_o^2)^3} X'(\underline{Y} - \underline{\mu})\left[(\underline{Y} - \underline{\mu})'(\underline{Y} - \underline{\mu}) + 2(\underline{Y} - \underline{\mu}\right.\right.$$

and then to

$$\mathrm{E}_f\left[\frac{1}{2(\sigma_o^2)^3}\left[X'(\underline{Y} - \underline{\mu})[(\underline{Y} - \underline{\mu})'(\underline{Y} - \underline{\mu})] + 2X'(\tau^2\right.\right.$$

$$= \mathrm{E}_f\left[\frac{1}{2(\sigma_o^2)^3}\left[X'(\underline{Y} - \underline{\mu})[(\underline{Y} - \underline{\mu})'(\underline{Y} - \underline{\mu})]\right.\right.$$

Now write the needed expectation in terms of the hypothet
are *iid* $N(0, \tau^2)$; hence

$$\mathrm{E}_f\left[(\underline{Y} - \underline{\mu})[(\underline{Y} - \underline{\mu})'(\underline{Y} - \underline{\mu})]\right] = \mathrm{E}_f\left[\underline{\epsilon}'\left[\sum_{i=}^{n}\right.\right.$$

The jth element of this vector is $\mathrm{E}_f\big((\epsilon_j)^3 + \sum_{i\neq j}\epsilon_j(\epsilon_i)^2\big)$
the mutual independence is $\mathrm{E}_f(\epsilon_j)^3$. Because the ϵ_j are ass
distributed, their third central moment is 0. Hence, we ha

$$\mathrm{E}_f\left[\underline{\epsilon}'\left[\sum_{i=1}^{n}(\epsilon_i)^2\right]\right] = \underline{0},$$

and the desired part of $J(\underline{\theta}_o)$ is $\underline{0}$.

$$J_{KK}(\underline{\theta}_o) = \mathrm{E}_f\left[-\frac{n}{2\sigma_o^2} + \frac{1}{2(\sigma_o^2)^2}(\underline{Y} - X\underline{\beta}_o)'(\underline{Y}$$

Several straightforward steps reduce the above to

$$J_{KK}(\underline{\theta}_o) = \frac{1}{(2\sigma_o^2)^2}\left[-n^2 + \frac{1}{(\sigma_o^2)^2}\mathrm{E}_f\left[(\underline{Y} - X\underline{\beta}_o)'(\underline{Y}\right.\right.$$

Define the ith row vector of X as $\underline{x}_i'$. Then

$$\begin{aligned}
&\mathrm{E}_f\left[(\underline{Y} - X\underline{\beta}_o)'(\underline{Y} - X\underline{\beta}_o)\right]^2 \\
&\quad = \mathrm{E}_f\left[\sum_{i=1}^{n}(y_i - \underline{x}_i'\underline{\beta}_o)^2\right]^2 \\
&\quad = \mathrm{E}_f\left[\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - \underline{x}_i'\underline{\beta}_o)^2(y_j - \underline{x}_j'\underline{\beta}_o)^2\right] \\
&\quad = \mathrm{E}_f\left[\sum_{i=1}^{n}(y_i - \underline{x}_i'\underline{\beta}_o)^4 + \sum_{i\neq j}(y_i - \underline{x}_i'\underline{\beta}_o)^2(y\right.
\end{aligned}$$

By virtue of mutual independence, the expectation of the above is easily found, giving

$$\begin{aligned}
&\mathrm{E}_f\left[(\underline{Y} - X\underline{\beta}_o)'(\underline{Y} - X\underline{\beta}_o)\right]^2 \\
&\quad = \mathrm{E}_f\left[\sum_{i=1}^{n}(y_i - \underline{x}_i'\underline{\beta}_o)^4\right] + (n\sigma_o^2)^2 - \sum_{i=1}^{n}\left[\tau^2 +\right.
\end{aligned}$$

For the case $\tau^2 = 0$ note that $\underline{Y} = \underline{\mu}$, and so the a $\mathrm{E}_f\left[(\underline{Y} - X\underline{\beta}_o)'(\underline{Y} - X\underline{\beta}_o)\right]^2 = (n\sigma_o^2)^2$; hence

$$J_{KK}(\underline{\theta}_o) = \frac{1}{(2\sigma_o^2)^2}\left[-n^2 + \frac{1}{(\sigma_o^2)^2}(n\sigma_o^2)^2\right]^2$$

It is thus clear that if $\tau^2 = 0$, then $J(\underline{\theta}_o) = O$.

The next steps are valid only if $\tau^2 > 0$. Let

$$\sqrt{\lambda_i} = \frac{\mu_i - \underline{x}_i'\underline{\beta}_o}{\tau}$$

and

$$z_i = \frac{y_i - \mu_i}{\tau}.$$

$$\mathrm{E}_f\left[\sum_{i=1}^{n}(y_i - \underline{x}_i\underline{\beta}_o)\right] = \tau^4\left[\sum_{i=1}^{n}\mathrm{E}_f\left[z_i + \nu\right]\right.$$

The needed expectation is now easily found because it is ju
of a normal random variable with a nonzero mean; or it
a function of the first four moments of a standard norm
We find it easier to note that the needed expectation is th
noncentral chi-square random variable on 1 df and noncer
The result is

$$\mathrm{E}_f\left[\sum_{i=1}^{n}(y_i - \underline{x}_i'\underline{\beta}_o)^4\right] = \tau^4\left[\sum_{i=1}^{n}\left[3 + 6\lambda_i +\right.\right.$$

Now, by carefully constructing the full result from all th
simplifying it, we get

$$J_{KK}(\underline{\theta}_o) = \frac{n}{2(\sigma_o^2)^2}\left[\frac{2\tau^2\sigma_o^2 - \tau^4}{(\sigma_o^2)^2}\right].$$

While derived only for $\tau > 0$, the above result can also be
case of $\tau^2 = 0$.

Finally,

$$J(\underline{\theta}_o) = \begin{bmatrix} \dfrac{\tau^2}{(\sigma_o^2)^2}X'X & \underline{0} \\ \underline{0}' & \dfrac{n}{2(\sigma_o^2)^2}\left[\dfrac{2\tau^2\sigma_o^2 - \tau}{(\sigma_o^2)^2}\right. \end{bmatrix}$$

The result we sought can now be found:

$$\mathrm{tr}\left[J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1}\right] = \frac{\tau^2}{\sigma_o^2}\left[K + 1 - \frac{\tau^2}{\sigma_o^2}\right.$$

If model equals truth, we have $\underline{\mu} = X\underline{\beta}$, so that $\sigma_o^2 = \tau^2$ (
and the trace term equals K. By continuity in τ^2 we n
trace term as K when $\tau^2 = 0$ if the model is true. How
philosophical issues and problems associated with a truth
we will consider only the situation wherein even for truth
unexplainable uncertainty. In particular, if true replication
iment (or study), we suggest that it is most useful to con
the variance within true replicates (assuming variance ho
a definition of convenience, as even truth can be at diff
are mostly interested in structural truth of our models in th
irreducible uncertainty inherent in data for finite sample s

Surprisingly enough, we see from (7.79) that for thi
and context, $\mathrm{tr}\left[J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1}\right] < K$ under a misspeci

R^2, hence use of AIC (rather than TIC) seems acceptabl
average, on the side of parsimony. Moreover, estimatio
(hence, TIC) seems very problematic, since τ^2 cannot be
lacking true replication (cf. Linhart and Zucchini 1986
call true replication in an experiment provides only an
definition, if we restrict our concept of truth to what w
the design structure and independent variables used in th
Philosophically, we might be able to predict some (hence
(hence $\tau^2 = 0$) of the observed differences among repli
knew ultimate truth.

This example can be easily generalized; that is, we
model and generalize truth somewhat. Whereas we ass
normal$(0, \tau^2)$, the only way this entered the derivations
central moments of ϵ. If we retain the *iid* assumption, we ca
results; note that we retain $\mathrm{E}(\epsilon) = 0$ with loss of generalit
asymmetric distribution for ϵ; we will not do so: We assu
all we need is the fourth moment of ϵ, which we will exp
form as

$$\gamma = \mathrm{E}(\epsilon^4)/[\mathrm{E}(\epsilon^2)]^2.$$

For assumed normal truth, $\gamma = 3$. For f as a logistic distri
a Laplace distribution, $\gamma = 6$; and for a uniform$(-h, h)$ d
The last two are extreme cases; one might think that γ li
the range 2 to 4.

Redoing the derivations for this more general way of
straightforward; the results are the same for $I(\underline{\theta}_o)$; but fo

$$J(\underline{\theta}_o) = \begin{bmatrix} \dfrac{\tau^2}{(\sigma_o^2)^2} X'X & \underline{0} \\ \underline{0}' & \dfrac{n}{2(\sigma_o^2)^2} \left[\dfrac{2\tau^2\sigma_o^2 + \tau^4 \left(\frac{\gamma-}{2}\right.}{(\sigma_o^2)^2} \right. \end{bmatrix}$$

For the trace function we get

$$\mathrm{tr}\left[J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1}\right] = \frac{\tau^2}{\sigma_o^2}\left[K + 1 + \frac{\tau^2}{\sigma_o^2}\left[\frac{\gamma - 1}{2} -\right.\right.$$

Hence, for these fixed-effects linear models assuming
of structural misspecification appears in σ_o^2 as manifest
However, the effect of error distribution misspecification
moment, γ (assuming symmetric errors). We have stressed
models wherein K will not be trivially small; from (7.80)
large the effect of error distribution misspecification upo
becomes trivial. In contrast, the effect of structural mis

misspecification of the error distribution becomes trivial i
use of the approximation $\mathrm{tr}\left[J(\underline{\theta}_o)[I(\underline{\theta}_o)]^{-1}\right] = K$, hence t
of AIC rather than TIC. These musings seem likely to a
linear models.

One last point: How good is use of $[I(\underline{\theta}_o)]^{-1}$ for the
matrix of $\underline{\hat{\theta}}$ under model misspecification here (ignoring s
that is)? As was shown (and is known in general) in Sec
asymptotic variance–covariance matrix is $V(\underline{\hat{\theta}}) = [I(\underline{\theta}$
Thus, here, $V(\underline{\hat{\beta}}) = \tau^2(X'X)^{-1}$, and

$$V(\hat{\sigma}^2) = \frac{2(\sigma_o^2)^2}{n}\left[\frac{2\tau^2\sigma_o^2 + \tau^4\left(\frac{\gamma-1}{2} - 2\right.}{(\sigma_o^2)^2}\right.$$

If a good structural fit has been achieved (so $\tau^2 \approx \sigma_o^2$), th

$$V(\hat{\sigma}^2) = \frac{2(\sigma_o^2)^2}{n}\left[\frac{\gamma-1}{2}\right].$$

The drastic bias induced by a γ not near 3 (but assumed
one to use an estimator of γ, and with the same estimator
AIC. This approach can be recommended only weakly,
estimator of the fourth moment is so highly variable.

7.7 Additional Results and Considerations

7.7.1 Selection Simulation for Nested Models

The detailed stochastic characteristics of the model selec
be studied mostly by Monte Carlo simulation methods. C
if the only completely general approach is to specify a data
and a set of models to be fit to each generated sample
samples and do all the calculations associated with model
This is very useful, but is not a study of properties of mode
in the abstract. Rather, each application has some underly
and type of truth (as generated data), and may require ext
that are peripheral to the heart of the model selection pro

An exception arises if we restrict ourselves to a single c
els and to selection methods based on log-likelihood differ
models. This scenario includes simulation of AIC, AIC_c,
ratio testing-based methods (and can be easily adapted to
QAIC_c model selection). All we need to generate are the i
trally distributed log-likelihood chi-square random variab

but no parameters need actually be postulated or estimat
that we cannot simultaneously study properties of parame
important restriction here is that we can correctly generat
variables only for what would be a single chain of nest
vantage of the method is speed and generality; this allow
properties of some model selection procedures.

At the heart of this procedure we have (conceptual)
and g_{i+j}; model g_i is nested in model g_{i+j}, and the diff
parameters is j. The method can be developed in gener
give it, and use it, with $j = 1$. That is, our conceptual se
$g_1 \subset g_2 \subset \cdots \subset g_R$, and each incremented model (i.e.
only 1 added parameter. We assume that large sample theo
estimation conceptually underlie such Monte Carlo simu
the usual likelihood ratio test statistic has, in general, a n
distribution on 1 df; denote that random variable by $\chi_1^2(\lambda_i$
parameter for model g_i versus g_{i+1} is λ_i.

For our set of R models we have $R - 1$ noncentrality
λ_{R-1}. These λ_i would be functionally related to the true da
the model structures assumed $g_i(\underline{x} \mid \cdot)$, and the specific pa
that specify the actual best approximating model in eac
However, we will be able to bypass all of those specificati
method below. We do need to be able to interpret sets
would reflect the failure of at least model g_i to perfectl
had $\lambda_{i-1} > 0$ and $\lambda_i = \cdots = \lambda_{R-1} = 0$, we would inter
being the true data-generating model (we will ignore pa
invalidate this interpretation). Also, it is possible to have r
for example, $\lambda_1 = 0$ but $\lambda_2 > 0$. Then both models g_1 and g
bad approximations to truth, because model g_3 improves
to truth compared to model g_2, but g_2 does not improve ove
situation we would have the set of λ_i monotonically decr
is that of which model provides the AIC best model, i.e
best model, when parameter estimation occurs. We conti
how to do simulation in this context.

Let the fitted model log-likelihoods be $\log(\mathcal{L}_i)$. From b

$$2\log(\mathcal{L}_{i+1}) - 2\log(\mathcal{L}_i) \sim \chi_1^2(\lambda_i).$$

Let the number of parameters in the simplest model be K
the above as

$$-(\mathrm{AIC}_{i+1} - 2K_1 - 2i) + (\mathrm{AIC}_i - 2K_1 - 2i + 2$$

or

$$\mathrm{AIC}_i - \mathrm{AIC}_{i+1} \sim \chi_1^2(\lambda_i) - 2.$$

$$\mathrm{AIC}_i = \mathrm{AIC}_{i+1} + \chi_1^2(\lambda_i) - 2.$$

Thus for purposes of a simulation study, if we know λ_i generate AIC_i. We just need to be able to generate a nonce random variable (there are routines for this, such as CIN

Based on (7.82), we can do a backwards recursive gen AIC_1 starting with $i = R - 1$, given a value for AIC_R. the formula

$$\mathrm{AIC}_i = \mathrm{AIC}_R + \sum_{j=i}^{R-1} (\chi_1^2(\lambda_j) - 2);$$

we just need a value for AIC_R. Because everything we c model selection depends only on the relative differences, s or Δ_i or Δ_p, it suffices to set $\mathrm{AIC}_R = R$ (any constan one has advantages) for every sample of AICs genera corresponds to a realization of a set of independent $\chi_1^2(\lambda_i$

We have used this approach to do simulation studies of m AIC and other likelihood-based methods. To evaluate B $\mathrm{AIC}_i - 2i + i \log(n)$. Because n should now vary, one n noncentrality parameters on a per-unit sample size bas compute $\lambda_{i,n} = n\lambda_{i,1}$. The $\lambda_{i,1}$ should be very small, but is arbitrary. To mimic AIC_c selection, use

$$\mathrm{AIC}_{c,i} = \mathrm{AIC}_i + 2\frac{K_i(K_i+1)}{n - K_i - 1},$$

where $K_i = K_1 + i$. Now one must specify K_1, the nu envisioned in model g_1, as well as use $\lambda_{i,n} = n\lambda_{i,1}$.

To mimic QAIC model selection is a little more invo value of c (variance inflation factor; $c \geq 1$) and its df (i vary over samples). Generate $\hat{c} = \chi^2_{\mathrm{df}}(\mathrm{df}(c-1))/\mathrm{df}$ fo $\mathrm{df}(c-1)$ is the noncentrality parameter for this chi-squa Then use

$$\mathrm{QAIC}_i = \mathrm{QAIC}_R + \sum_{j=i}^{R-1} \left[\frac{\chi_1^2(c - 1 + c \cdot \lambda_j)}{\hat{c}} - 2 \right],$$

and $\mathrm{QAIC}_R = R$. Also,

$$\mathrm{QAIC}_{c,i} = \mathrm{QAIC}_i + 2\frac{K_i(K_i+1)}{n - K_i - 1}.$$

A few results are given below using this simulation appr into model selection. For the most part, however, it is not extensive tables of simulation results. We encourage inte

From this setup which allows simulating model selectio
we can also compute theoretical expected AIC *differences*,
expected AIC best model exactly. From (7.83) we get, for r
by just one parameter,

$$\mathrm{E}(\mathrm{AIC}_i) = \mathrm{E}(\mathrm{AIC}_R) + \sum_{j=i}^{R-1} (E[\chi_1^2(\lambda_j)] -$$

$$\mathrm{E}(\mathrm{AIC}_i) = \mathrm{E}(\mathrm{AIC}_R) + \sum_{j=i}^{R-1} (\lambda_j - 1).$$

Let $\lambda_{i+} = \lambda_i + \cdots + \lambda_{R-1}$, $i = 1, \ldots, R-1$, and $\lambda_{R+} =$

$$\mathrm{E}(\mathrm{AIC}_i) = (\mathrm{E}(\mathrm{AIC}_R) - R) + (\lambda_{i+} + i$$

Then compute the set of values V_i given by

$$V_i = \lambda_{i+} + i, \quad i = 1, \ldots, R,$$

find their minimum, $V_{\min}$, and then compute

$$\mathrm{E}(\Delta_i) = V_i - V_{\min}.$$

As an example, if $R = 10$ and (in order) we have λ_i as 2,
0.4, and 0.2, then the $\mathrm{E}(\Delta_i)$ are, in order, 22.5, 21.5, 16.5
0.8, and 1.6. Thus the best expected AIC selected model is
variances can also be computed, but nothing directly usefu
like Δ_p.

7.7.2 Simulation of the Distribution of Δ_p

The random variable $\Delta_p = \mathrm{AIC}_{best} - \mathrm{AIC}_{min}$ was intro
(Section 4.5). For a set of models indexed $i = 1, \ldots,$
size n, and a conceptually well-defined repeated samplin
a sample space), we let model g_{best} represent the best m
fit under the AIC selection criterion. Monte Carlo simul
determine this actual best model (sometimes theory suffi
we are not saying that model g_{best} is truth; it is just that
must be the best model, on average, to use for all possible
the truth that model g_{best} represents.

For each simulation-generated sample we can compu
AIC_{min}. This AIC_{min} and the value of AIC_{best} vary by sa
value of *best* is fixed for all samples; for example, model g
best model to use (hence *best* $= 4$). If model g_{best} is se
sample, then $\Delta_p = 0$; otherwise, $\Delta_p > 0$. We can com
distribution, hence percentiles, of this pivotal under the

for R and $\lambda_1, \ldots, \lambda_{R-1}$; if we use AIC_c, we also need R

Results about Δ_p under AIC model selection, for a few $\lambda_i = 0$, are feasible to show. In this case model g_1 is the model ($k = 1$ in Δ_p). This scenario is clearly at odds w applies to real data analysis (all models for data analysis are to truth). However, it can be used as a benchmark for p the extent that this situation is too simple, it may serve o on the percentiles of the cumulative distribution functio nested models, or for real problems where there is substa of the models considered.

For the case of $R = 2$ and large sample size, the q $(0 < q < 1)$, $\Delta_{p,q}$, is

$$\Delta_{p,q} = \max\{0, (\chi^2_{1,q} - 2)\}$$

(easily derivable from 7.81). Here, $\chi^2_{1,q}$ is the qth percen square random variable on 1 df. For example, $\chi^2_{1,0.95} = 3$ 1.84. We used Monte Carlo simulation to determine some values of $R > 2$. One million samples were used for eac independent runs of 250,000 samples, so we can estimat below for $R > 2$ have a cv of about 0.5%:

	percentiles of Δ_p			
R	80%	90%	95%	99%
2	0.00	0.71	1.84	4.63
3	0.11	1.37	2.67	5.77
4	0.35	1.71	3.33	6.40
5	0.49	1.93	3.40	6.86
10	0.75	2.34	3.97	7.61
20	0.82	2.47	4.15	8.05

We have done many of these simulations to find the for sets of noncentrality parameters wherein $\lambda_i > 0$; the somewhat larger as compared to the case where all $\lambda_i =$ $R = 10$ and $\lambda_1, \ldots, \lambda_9$ be 2, 6, 10, 6, 3, 1.5, 0.8, 0.4, 0.2. the set of models (actually, g_R could be truth; we cannot on 20,000 Monte Carlo samples (two sets of 10,000), we the expected K-L (i.e., AIC) best model. The averages of t (rescaled so their minimum is 0), in order, are 22.6, 21.5, 0.2, 0.8, 1.6 (reliable to $\pm 0.1 = 2$ se); compare these valu expectations from the end of Section 7.7.1: 22.5, 21.5, 1 0.2, 0.8, and 1.6.

Based on $\Delta_p = AIC_7 - AIC_{min}$ over these 20,000 samp of Δ_p are 3.3 (80%), 4.6 (90%), 6.4 (95%), 10.6 (99%) From these sorts of simulations, and others with explici

sequences, a model g_i for which $\Delta_i > 4$ is implausible a
model structure, and $\Delta_i \approx 7$ is strong evidence against m
being the K-L best model (and $\Delta_i \geq 10$ is very strong evi
g_i). The Akaike weights provide a refined interpretation o

7.7.3 *Does AIC Overfit?*

The conceptual framework underlying valid use of AIC is
infinitely many parameters. Overfitting is often defined in
there is a simple true model, with a finite number of para
model is in the set of models considered. Then if the tru
nested within the selected model structure, the selected mo
One has estimated more parameters than are in the true m

This simplistic concept of overfitting does not apply in
lection framework. However, there is a best expected K
the model we should use as our basis for data analysis.
is nested within the selected model, might we claim that
overfit model? We decline to use this definition because
ability in the model selected. If we miss the target model
(or, what is the same, that the structure of the selected m
same as that of the target model), we should not say that
is philosophically the same as being concerned that a pa
may sometimes give a point estimate far away from the
are upset only by cases where $\hat{\theta}$ is quite far from θ. But
with suitably small probability, we consider $\hat{\theta}$ as an accep
argument becomes somewhat circular at this point becau
accustomed ourselves to being satisfied if $\hat{\theta}$ is within a
hence $\hat{\theta}$ is unacceptable with a probability of only about 0
titioners of null hypothesis testing typically are willing t
facto) a 0.05 probability of type I error (and probably a
probability in most applications).

Something similar should apply to possible overfitting
lection (underfitting is at best a minor concern with AIC m
need some idea of how far from the actual K-L best model
before it is regarded as an overfit model. We have to allow
do well to select a model within, say, 1 or 2 parameters o
model. In contrast, if the selected model has 10 or 20 pa
the target model, we think that most people will agree that
We can use simulation (in some cases theory exists, see S
and Yu 1993) to find the probability distribution of the se
say $\widehat{best}$, for a nested sequence of models g_1 to g_R and ac

Here is a typical example of what can happen. For a
models with $R = 10$ and the λ_i as 2, 6, 10, 6, 3, 1.5, 0
best model was found to be g_7. We extend this to R be
with all additional $\lambda_i = 0$ (and assume that sample size
This means that the true data-generating model is model
g_{30} are overfit, if selected, in the sense that they do conta
structure. The K-L best model remains model g_7 even fo
give the model selection frequencies, in model order g_1 to
Monte Carlo samples:

R						g_7							
10:	11	5	50	746	2282	2635	**1924**	1161	735	451			
20:	16	3	20	405	1557	2399	**1970**	1373	816	542	274	192	125 9
30:	15	3	23	407	1689	2351	**1954**	1360	800	522	237	155	121 9
													11

The long-tailed nature of the distribution of selected
AIC when there are many "big" models (models with
parameters) containing the K-L best model structure. For
we have a probability of about 0.01 of selecting a mod
parameters (models g_{19} to g_{30}), hence having estimated 1
parameters. For both $R = 20$ and 30 there is about a
selecting a model with five or more unneeded parameters
higher). In general, if we say that we can accept a procedu
of its cases a bit misleading, then we should not be upset
the K-L best model by about 5 or more parameters with pr
0.06 if many such too-general models are in the set of mo

Note that these results are effectively for very large samp
was used, not AIC_c. For not-large n, use of AIC_c would
the long tail at $R = 20$ and 30 (say $n = 100$). Hence, th
worst-case scenarios.

Here is a worst-case scenario for one linear sequenc
All $\lambda_i = 0$, so model g_1 is the K-L best model. Based on
Carlo samples, the estimated model selection probabilities
error $\leq 0.05\%$) are below:

R	$\hat{\pi}_i \times 100\%$, in order $i = 1, 2, \ldots$												
3	78.7	13.3	8.0										
4	76.0	12.5	6.7	4.8									
5	74.4	12.0	6.4	4.1	3.1								
10	71.8	11.4	5.8	3.5	2.4	1.7	1.2	.9	.7	.6			
20	71.2	11.2	5.7	3.4	2.3	1.6	1.1	.8	.6	.5	.4	.3	.2 .2

(For this case of a single nested sequence of models, these
ties are known theoretically for large n and all $\lambda_i = 0$; see
1989). The long tail is disturbing, yet with probability abo

of realistic AIC model selection, we claim that this is a typ
sample AIC (about 0.06 probability of overfitting by five
parameters if such general, overparametrized models are in
Given models with a large amount of unneeded structur
can select overfitted models, but the probability of a serio
arguably less than the total error probabilities (type I plus t
hypothesis testing.

What will reduce the probability of getting a badly
AIC_c, which helps considerably when sample size, relati
Otherwise, the only recommendation we have to avoid the
a much overfit AIC-selected model is to be very thoughtf
set of models considered. In particular, do not casually i
a great many parameters more than you think are really n
variable selection this would mean do not simply consid
regressor variable and include it for possible selection. If
having large numbers of variables that have no explanat
a $\lambda \approx 0$), and that leaves you with a small but real proba
model with many worthless variables.

7.7.4 Can Selection Be Improved Based on All t

Given the potentially long-tailed nature of K-L–based m
pends on the set of models), it seems natural to ask whe
information in the full set of Δ_i values that would allow
cases where we have selected a very overfit model. If so
selection to a better model based on information in the ent
We have explored this matter for a single series of nested
ered in the above three subsections). The idea was that p
the Δ_i would be like that below in the event of selecting a
(line one is model number i; line two is Δ_i):

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
30	10	5	0.1	1	2.5	2	1.5	2.7	3	0.8	2.3	1	0	2.

Here, AIC has selected model g_{14}, but we might suspect t
to use is g_4 ($\Delta_4 = 0.1$ and none of models g_5 to g_{13} have
Δ_i). So we could change our choice to model g_4.

There is no theory to help here regarding properties of
Δ_R. So we looked at a large number of simulated results f
worst-case scenario of all $\lambda_i = 0$. Hence, the simplest m
true data-generating model and the K-L best model. Tab
$R = 10$, some selected cases of $\Delta_1, \ldots, \Delta_{10}$ (a case is one
selection results, for a large sample size). Cases 1, 2, and
model g_1 is selected in about 72% of all samples here. Cas

selected), A is atypical but not rare (11% of cases, model g_2 select
of all cases, and for these, overfit models are selected.

	Δ_i							
pattern	$i = 1$	2	3	4	5	6	7	
1 T	**0.0**	1.7	3.7	5.0	6.1	8.1	9.5	1
2 T	**0.0**	0.5	0.1	2.8	4.7	6.6	8.5	
3 T	**0.0**	0.6	0.5	2.5	0.8	2.8	4.3	
4 A	1.5	**0.0**	0.6	1.5	3.2	3.6	5.5	
5 R	7.5	2.0	2.7	0.8	1.5	**0.0**	3.5	
6 R	9.5	10.4	9.6	7.6	5.6	5.1	**0.0**	
7 R	0.2	1.1	2.0	3.6	1.4	**0.0**	2.7	
8 R	2.0	2.2	0.8	2.8	2.5	**0.0**	1.9	
9 R	1.1	1.8	3.8	5.4	7.3	5.5	3.3	
10 R	1.1	3.1	0.2	1.7	2.2	4.2	**0.0**	
11 R	7.9	9.0	11.0	6.2	1.3	**0.0**	2.0	
12 R	0.1	2.1	3.3	1.5	1.1	2.9	3.9	
13 R	1.1	1.8	3.8	5.4	7.3	5.5	3.3	
14 R	10.4	12.2	8.8	8.5	8.6	6.3	**0.0**	

of samples wherein model g_2 is selected (about 11% of sa
are rare (models g_6 to g_{10} selected; it happens here in abou
In these latter cases we would say that an overfit model w
case 12 might we feel justified in rejecting model 10 in
the selected model, but even there nothing in the 10 val
think that this decision is particularly justified. In the oth
nothing in the nature of the pattern of the ten Δ_i that give
that we are justified in selecting a model as best other than
(for which $\Delta_k = 0$). That is, changing our selected mode
not better, results: We cannot tell based on the data. In fac
11, 14, and perhaps 5) the "evidence" in the set of Δ_i seen
the correctness of the AIC best model.

We also simulated the performance of various ad hoc r
selection procedures to change the selection to a more p
Such a change was done if the AIC-selected model had
a more parsimonious model that had Δ_i very near zero (
exact algorithms tried varied, but none of those ad hoc
meaningful difference to the overall AIC model selection
Also, we did not visually perceive any useful information
Δ_i. While more work could be done along these lines, we
it would be fruitful. Basically, if the data "lie" to you (i
selected because the sample is atypical), there are no diag
from that sample to that tell you that it has "lied."

which informative results can be computed, for a certa
gression under constant error variance. Specifically, we a
linear regression based on orthogonal regressors, but our
a subset of the regressors (the global model may include
$z_1, \ldots, z_m$ be *iid* normal(0,1). Independently, let ϵ be no
response variable x, based on the regressors, is given by t

$$x = \beta_o + \sum_{j=1}^{m} \beta_j z_j + \epsilon,$$

so

$$\mathrm{E}(x \mid \underline{z}) = \beta_o + \sum_{j=1}^{m} \beta_j z_j.$$

Results below are scale-invariant in terms of the regress
z_j occur in the models only as a product. To keep nota
Section 7.7.1, we define here the base model, g_1, to be
with δ assumed as normal(0, σ_1^2). The number of parame
$K_1 = 3$. The normal assumption is true, but $\sigma_1^2 \neq \sigma^2$. In
model g_r is

$$x = \beta_o + \sum_{j=1}^{r} \beta_j z_j + \delta,$$

with δ assumed as normal(0, σ_r^2); the number of paramet

We can consider the sequence of nested models for r
The ordering of regressors is arbitrary, but is used in the fo
nested models are considered sequentially (as per the theo
It is then convenient, but not required, to specify the regre
satisfy $|\beta_j| > |\beta_{j+1}|$. Doing so yields insights into AIC m
easily (such a structured situation is also considered in Sp

From the point of view of insights to be gained, the assu
regressors is not restrictive if $R = m$, because any regres
transformed into the case of orthogonal regressors, for e
to regression on principal components. However, if we
$R < m$, the orthogonality assumption is restrictive, beca
our observed regressors orthogonal to the regressors not ob
realistic of real data). For this reason there is no advantage
with $R < m$.

The regressors are random variables. Therefore, to g
we take an additional expectation over certain matrices t
random (row) vector $\underline{z}_r = (1, z_1, \ldots, z_r)$. For a sample of
matrix notation is $\underline{x} = Z\underline{\beta} + \underline{\delta}$, so matrices such as $Z'Z$ an

by their expectations with respect to the random $\underline{z}_r$ (for for large n, $\mathrm{E}(Z'Z)^{-1} \approx I$). However, the nonlinearities some theoretical formulas below are only large-sample ap

Under this scenario of regressor independence we deter noncentrality parameters of Section 7.7.1 that apply, for la the nested sequence of models defined here (g_1 to g_R):

$$\lambda_i = n \log\left[1 + \frac{(\beta_{i+1})^2}{(\beta_{i+2})^2 + \cdots + (\beta_m)^2 + \sigma^2}\right], \qquad i$$

$$\lambda_{m-1} = n \log\left[1 + \frac{(\beta_m)^2}{\sigma^2}\right].$$

Given this context and the theory in Section 7.7.1 we proximate) expected Δ_i values for AIC and AIC_c, h (approximate) theoretical expected K-L optimal model. V actual sets of Δ_i values. The needed partial sums of nonc are

$$\lambda_{i+} = n \log\left[1 + \frac{\sum_{j=i+1}^{m} (\beta_j)^2}{\sigma^2}\right], \qquad i = 1, \ldots,$$

The value of exploring this situation is that we can also theoretical quantities that can be related, or compared, to A Under any of the models the regression coefficient estima the true parameters (because all regressors are orthogo expected in general). Under model g_r the value of σ_r^2 is

$$\sigma_r^2 = (\beta_{r+1})^2 + \cdots + (\beta_m)^2 + \sigma^2.$$

In notation used elsewhere in Chapter 7, the above σ_r^2 is The usual (conditional on the model) parameter cv's und

$$\mathrm{cv}(\hat{\beta}_r) = \frac{\beta_r}{\mathrm{se}(\hat{\beta}_r \mid g_r)} = \frac{\sqrt{n}\beta_r}{\sigma_r}, \qquad r = 1,$$

However, rather than compute the above cv's it is informa the cv's under the global model, $\sqrt{n}\beta_r/\sigma$, and observe the cv's versus what parameters are included or excluded fr optimal model.

We can also determine the overall mean square erro model. Minimum MSE is generally accepted as a good model selection; here we have

$$\mathrm{MSE}(Z) = \sum_{j=1}^{n} \mathrm{E}_f\left[\hat{\mathrm{E}}(x_j \mid \underline{z}_r) - \mathrm{E}(x_j \mid \underline{z}_m\right.$$

here over $\mathcal{L}$ to get the unconditional result

$$\text{MSE} = \sigma_r^2(r+1) + \sum_{j=r+1}^{m} n(\beta_j)^2.$$

Sometimes model selection is based on minimum MSE response variable for a single additional (independent of s gressors. Then this average mean square error of predictio g_r is MSEP $=$ MSE $+n\sigma^2$. Thus, it suffices to consider o C_p implements minimum MSE model selection for regres 1995).

The K-L–based target criterion T (7.20) can be deter We express that result as $-2T$ for direct comparison to MSE-based results, and we will label it here as KL. Hen model g_r is exactly

$$\text{KL} = n\text{E}[\log(\chi^2_{n-r-1}(0))] + n\log\left[\frac{\sigma_r^2}{n}\right]$$
$$+ \frac{n}{n-r-3}\left[(r+1) + \left(\frac{n\sigma^2}{\sigma_r^2}\right) + \sum_{j=r+1}^{m}\right.$$

In MSE (7.88) the term $\sum_{j=r+1}^{m} n(\beta_j)^2$ is bias squared gressors. Note that in MSE these components of bias are a to theoretical precision of the excluded $\hat{\beta}_j$, whereas in K of excluded regressors is "judged" relative to the theore $\hat{\beta}_j$ have. That is, the comparable term reflecting bias is i $(\beta_j/\sigma_r)^2$, not just $(\beta_j)^2$. This feature of the KL criterion s to us than just optimizing on pure bias versus variance as

Most insights based on all of these results will need to c examples and simulation. Because of the possible volume sidering all the variables here), we leave such computing reader. We have done a lot of computing and simulation these formulas and Section 7.7.1 results. One result (kn about C_p versus AIC) is that K-L–based model selection about the same as selection based on minimum theoretic termine this by computing the theoretical criteria KL and is convenient to rescale the R values of those criteria to at zero. Table 7.6 gives such results for one case: $R =$ $\beta_i = (0.6)^{i-1}$, and σ taking several values in the range 0.

We have not undertaken a detailed analysis of these c gression, let alone for this case of orthogonal regressors. results are representative of cases we have looked at in K-L best model has either the same number of parameters

of σ; results for both KL and MSE have been rescaled so that their
clearly indicating the theoretical optimal model under these criteria

	$\sigma = 1.0$		$\sigma = 0.5$		$\sigma = 0.25$		$\sigma = 0.$	
r	KL	MSE	KL	MSE	KL	MSE	KL	M
1	8.01	14.08	27.17	16.66	57.72	17.58	142.67	1
2	1.70	3.76	11.16	5.59	33.52	6.32	113.80	
3	**0.00**	0.56	2.75	1.64	15.09	2.18	85.50	
4	0.59	**0.00**	**0.00**	0.33	4.34	0.68	58.46	
5	2.26	0.42	0.28	**0.00**	0.20	0.17	34.20	
6	4.52	1.21	2.01	0.04	**0.00**	0.02	15.33	
7	7.20	2.13	4.50	0.21	1.74	**0.00**	4.10	
8	10.29	3.10	7.52	0.43	4.48	0.03	**0.00**	
9	13.83	4.09	11.03	0.67	7.90	0.09	0.46	
10	17.90	5.08	15.09	0.91	11.92	0.14	3.33	

parsimonious than the theoretically best model under mi
expected models selected under operational criteria such
differ slightly from these results (but less so as n gets larg

7.7.6 *AIC_c and Models for Multivariate Data*

The derivation of large-sample AIC in Section 7.2 *does* a
independent multivariate observations, each with p noni
nents. The small-sample improvement of AIC that applie
$p = 1$) linear models with homogeneous normal residuals
ply in the corresponding multivariate case. This problem
Fujikoshi and Satoh (1997). They focused on selection of
is, inclusion or exclusion of the same set of possible regres
of the p regressions. They assume that a general $p \times p$
matrix Σ applies for the residual vector of each observati
with k regressors (this may include an intercept) there
parameters. Each model also includes $p(p+1)/2$ unkno
Thus, $K = (k \cdot p) + p(p+1)/2$.

For their data analysis context Fujikoshi and Satoh (19
AIC, CAIC in their notation, analogous to the univaria
result (their formula 7) can be expressed as follows:

$$\text{CAIC} = \text{AIC} + 2\frac{K(k+1+p)}{n-k-1-p}.$$

The univariate case corresponds to $p = 1$, and then K
general result in (7.90) reduces to the univariate AIC_c.

other multivariate modeling problems we do not know wh
Our key point here is that the univariate result for AIC_c c
multivariate setting.

The form of (7.90), by virtue of including variable
restricted context considered by Fujikoshi and Satoh (
ing k from (7.90) we hypothesize a generalization of
corresponding multivariate applications:

$$AIC_c = AIC + 2\frac{K(K+v)}{np - K - v}.$$

In (7.91) v is the number of distinct parameters used in
Σ; $1 \leq v \leq p(p+1)/2$. Note that the count K include
is correct for the univariate case wherein $v = p = 1$. I
seems reasonable until a derivation is published for the ne
of AIC_c to multivariate applications.

The emphasis of this section is the generalization of
applications. Many multivariate analysis methods, such as
sion, analysis of variance or covariance, are done in a leas
while assuming a multivariate normal model such as MV
be full rank). Given that least squares is used, the software
value of the maximized log-likelihood. However, the resi
and cross products matrix (SSCP) is nearly always prov
software packages and the MLE of Σ is $\hat{\Sigma} = SSCP/n$. Fu
mized log-likelihood is proportional to $-(n/2)\log(|\hat{\Sigma}|)$, w
determinant of $\hat{\Sigma}$. Hence, as long as all models considered
normal residuals, we may use

$$AIC = n \times \log(|\hat{\Sigma}|) + 2K$$

and

$$AIC_c = n \times \log(|\hat{\Sigma}|) + 2K + \frac{2K(K+$$
$$n - K -$$

In the univariate case SSCP is just the residual sum of squa
and the MLE of σ^2 is RSS/n. Thus the multivariate
univariate case. Theory for the multivariate case is su
(1984:61).

If the determinant |SSCP| is directly available, one coul
$n^{-p}|\ SSCP\ |$. However, because $\log(|\hat{\Sigma}|) = -p \times \log(n)$
p and n are constants we can just as well take, for examp

$$AIC = n \times \log(|\ SSCP\ |) + 2K.$$

that drops out of all inferential uses of AIC.

7.7.7 *There Is No True TIC_c*

Fujikoshi and Satoh (1997) also consider a small sampl
essence they want to extend AIC_c to TIC_c. However, AIC
1989) arises by computing the exact value of the target mod
(7.20) for a linear model with constant normally distribute
condition (assumption) that this model is the true data-ge
is derived without any assumption that truth, f, is the sa
That derivation can be justified only for large sample sizes
sample, an exact version of TIC would require us to spec
the distribution f (i.e., specify truth). Even if we could
in general, or at all, the result would depend upon assu
truth. Thus no defensible, general small-sample analytical
a TIC_c) seems possible.

The issue did not escape the attention of Fujikoshi and
they did (and they knew it) was to assume that the linear,
global model defined by using all available regressors cont
as an unknown submodel. Thus the true model is, by as
of models considered and is a special case of the global
assumptions Fujikoshi and Satoh (1997) derived an analy
sample size, for the target criterion of (7.20). Their formu
beyond $-2\log(\mathcal{L})$, that must be estimated from the data, a
nents that are simple functions of known n and K, and tha
estimable only by virtue of the strong assumptions made.
the same small-sample-size adjustments with TIC as we u
ing so may be a good idea; but we cannot find truly gener
adjustments for TIC.

7.7.8 *Kullback–Leibler Information Relationshi*
Information Matrix

The Fisher information matrix is defined by (7.6) for any

$$\mathcal{I}(\underline{\theta}) = E_g\left[-\frac{\partial^2 \log(g(\underline{x}\,|\,\underline{\theta}))}{\partial\theta_i\,\partial\theta_j}\right].$$

In taking this expectation it is assumed that the true data
$g(\underline{x}\,|\,\underline{\theta})$ (hence the underlying integration is with respect t
when this one particular member (i.e., g at $\underline{\theta}_o$) of the se
for fixed structure, and any $\underline{\theta} \in \Theta$, is the generating mod

$$I(\underline{\theta}_o) = \mathrm{E}_f\left[-\frac{\partial^2 \log g(\underline{x} \mid \underline{\theta})}{\partial\theta_i \partial\theta_j}\right],$$

in general, $I(\underline{\theta}_o) \neq \mathcal{I}(\underline{\theta}_o)$. Moreover, $\mathcal{I}(\underline{\theta}_o)^{-1}$ is guarant
sample variance–covariance matrix of the MLE $\hat{\underline{\theta}}$ only
For a value of $\underline{\theta} \in \Theta$ that is near the K-L minimizing
quadratic approximation to the K-L difference is

$$I(f, g(\cdot \mid \underline{\theta})) - I(f, g(\cdot \mid \underline{\theta}_o)) \approx \frac{1}{2}(\underline{\theta} - \underline{\theta}_o)' I(\underline{\theta}_o$$

For the case of $f = g$ we get the result

$$I(g(\cdot \mid \underline{\theta}_o), g(\cdot \mid \underline{\theta})) \approx \frac{1}{2}(\underline{\theta} - \underline{\theta}_o)' \mathcal{I}(\underline{\theta}_o)(\underline{\theta} -$$

Thus if one member of the set of models $g(\underline{x} \mid \underline{\theta})$, $\underline{\theta}$
generating distribution, then the approximate K-L informa
from using a nearby distribution as the approximating
quadratic form in the Fisher information matrix. This is
given the definitions of both $\mathcal{I}(\underline{\theta})$ and $I(g(\cdot \mid \underline{\theta}_o), g(\cdot \mid \underline{\theta}))$
show that the two underlying concepts of "information" ar
different, concepts. It was in the 1920s that Fisher chose t
matrix of second mixed partials of a probability distribu
There is no relationship to information theory, which is
mostly since Shannon's pioneering work in the late 1940
deals with logs of probabilities. The Fisher information m
relates to the precision of ML estimators. The Kullback
1951 was a result of their attempt to understand and explai
by "information" in relation to sufficiency (personal co
Leibler).

7.7.9 *Entropy and Jaynes Maxent Principle*

In Section 2.9 we noted that the Akaike weights w_i can be
Bayesian approach based on prior probabilities τ_i. To cho
abilities in a manner philosophically consistent with the
suggest resorting to the use of the Jaynes maximal entro
ple (Jaynes 1957, 1982, Jessop 1995). This principle aris
theory. The maxent principle says that if we must completel
ity distribution with only partial knowledge about momer
of that distribution, then we should choose the distributio
uninformative with regard to missing information. This m
the distribution that has maximal entropy subject to any inf
we can justify, such as constraints based on data. Mathem

justify (for example, constraints about the mean and/or
tribution). The result is a distribution that conveys no inf
what we explicitly build into it. If the only constraint w
prior probabilities sum to 1, then the maxent distribution i

That we can justify this uninformative prior for the mo
mation theory is yet another example of how deeply info
theory underlie statistical model selection. We will not di
tive of exploring information-theoretic data-based model
we recommend that interested readers pursue some of t
here on the subject. An introductory reference that ties to
of statistics and information theory, including the Jayne
is Jessop (1995). A nontechnical reference is Lucky (1
cal perspective see Yockey (1992), while Cover and Th
very thorough overview of information theory. Short, h
treatments are given by Wehrl (1978) and Ullah (1996).

As a general comment we emphasize the extensive
tent of information and entropy theory, and how these
many scientific and technical areas (from Boltzmann to Ei
Kullback–Leibler, for example). There is thus a deep foun
Leibler information measure and a firm basis for its use in
other aspects of statistics. K-L is not just another (of many
of discrepancies between probability distributions; it is u
data-based model selection in science when truth is ver
"noisy," and models can be only approximations to truth.

In saying that this theoretical foundation for use of K-L
we would liken it to the theoretical basis for the importan
(≈ 2.7183) in mathematics. It is not at all obvious why such
number should universally be the basis for logarithms
most of mathematics and science. But just as with K-L in
compelling, deep reason, not easily perceived, for the imp

7.7.10 *Akaike Weights w_i Versus Selection Prob*

The model selection probabilities can be expressed as expe
random variables that are a function of the sample data:

$$M_i(\underline{x}) = \begin{cases} 1 \text{ if model } i \text{ is selected by Al} \\ 0 \text{ otherwise.} \end{cases}$$

By definition, $\mathrm{E}(M_i(\underline{x})) = \pi_i$. We assume no ties for the
The Akaike weights (see Section 2.9) defined by

$$w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^{R} \exp(-\frac{1}{2}\Delta_r)}$$

$0 < \gamma < \infty$,

$$w_k(\gamma) = \frac{1}{1 + \sum_{r \neq k} \exp(-\gamma \Delta_r)}$$

and

$$w_i(\gamma) = \frac{\exp(-\gamma \Delta_i)}{1 + \sum_{r \neq k} \exp(-\gamma \Delta_r)}, \quad i \neq$$

In the limit as γ goes to infinity we have the result

$$\lim_{\gamma \to \infty} w_i(\gamma) = M_i(\underline{x}),$$

whence

$$\lim_{\gamma \to \infty} \mathrm{E}(w_i(\gamma)) = \pi_i$$

(the implied interchange of limits will be valid here). Th
erally be the case that $\mathrm{E}(w_i(0.5)) \equiv \mathrm{E}(w_i) \neq \pi_i$; also, I
unrelated.

This result does not rule out $\mathrm{E}(w_i) \approx \pi_i$, which sin
sometimes a useful approximation. Moreover, use of the *s*
as an estimator for the *set* of selection probabilities seem
where such $\hat{\pi}_i$ are needed. (Research could be done to fin
on the Akaike weights).

7.8 Kullback–Leibler Information Is Alw

It is not obvious that the Kullback–Leibler discrepancy,

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

is strictly nonnegative for any possible $g(x)$. Here we reduc
model g to just $g(x)$ rather than $g(\underline{x} \mid \underline{\theta})$. Also, the possib
nature of f and g is not emphasized in the proofs in this

Rigorous proofs exist that $I(f, g) \geq 0$ and that $I(f,$
if $g(x) \equiv f(x)$ for all x. Here we give a valid, but not
$I(f, g) \geq 0$. We do so for both the case of continuous
case of discrete distributions such as the Poisson, binom
wherein

$$I(f, g) = \sum_{i=1}^{k} p_i \log\left(\frac{p_i}{q_i}\right).$$

the $q_1, \ldots, q_k$ constitute the approximating probability
model). Hence, here f and g correspond to the p_i and q_i

In the first case, both $f(x)$ and $g(x)$ must be valid prob
hence satisfy $f(x) \geq 0$, $g(x) \geq 0$ and both integrate to 1:

$$\int f(x)dx = 1, \qquad \int g(x)dx = 1.$$

The exact limits of integration need not be specified he
same for both f and g. Moreover, without loss of gener
$f(x) > 0$, $g(x) > 0$; hence the ratio $f(x)/g(x)$ is neve
undefined (but may be taken as ∞). For the discrete case
$0 < q_i < 1$, and

$$\sum_{i=1}^{k} p_i = 1, \qquad \sum_{i=1}^{k} q_i = 1.$$

We consider first the case of continuous probability di
to one line of proof is to define a new function

$$h(x) = \frac{g(x) - f(x)}{f(x)};$$

thus,

$$\frac{g(x)}{f(x)} = 1 + h(x).$$

The lower bound on $h(x)$ is -1, because for any x over
performed, $g(x)$ can be arbitrarily close to 0. The upper b
thus $-1 < h(x) < \infty$. Note also that $\log(a) = -\log(1/a$

$$\begin{aligned} I(f, g) &= \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx \\ &= -\int f(x) \log\left(\frac{g(x)}{f(x)}\right) dx \\ &= 0 - \int f(x) \log\left(\frac{g(x)}{f(x)}\right) dx \\ &= \int f(x)h(x)dx - \int f(x) \log\left(\frac{g(}{f(}\right. \end{aligned}$$

The last step above uses the fact that

$$\begin{aligned} 0 = \int f(x)h(x)dx &= \int f(x)\frac{g(x) - f(x}{f(x)} \\ &= \int \big(g(x) - f(x)\big)dx \end{aligned}$$

Returning now to the main proof, we have

$$\begin{aligned} I(f,g) &= \int f(x)h(x)dx - \int f(x)\log\left(\frac{g(x)}{f(x)}\right. \\ &= \int f(x)h(x)dx - \int f(x)\log\bigl(1+ \\ &= \int f(x)\bigl[h(x) - \log\bigl(1+h(x)\bigr)\bigr]dx \\ &= \int f(x)t(h(x))dx, \end{aligned}$$

where $t(h(x)) = h(x) - \log\bigl(1+h(x)\bigr)$. We do not need to values of $t(h(x))$. Nor do we need to consider $t(\cdot)$ as also, x may be univariate or multivariate. It suffices to c $h - \log(1+h)$, hence $t(h)$, over the full range of h, $-$ possible by varying x. All we care about is some basic asp namely that is it strictly nonnegative. It is.

Calculus can be used to show that $t(h) \geq 0$, and that $t($ minimum, and for any $h \neq 0$, then $t(h) > 0$. A simple heu to plot $t(h)$ over, say, $-1 < h \leq 5$, and check $t(h)$ at a fe (Figure 7.1, and $t(10) = 7.6021$, $t(20) = 16.9555$, $t(1000) = 993.0913$). Given that $t(x) \geq 0$ for all x,

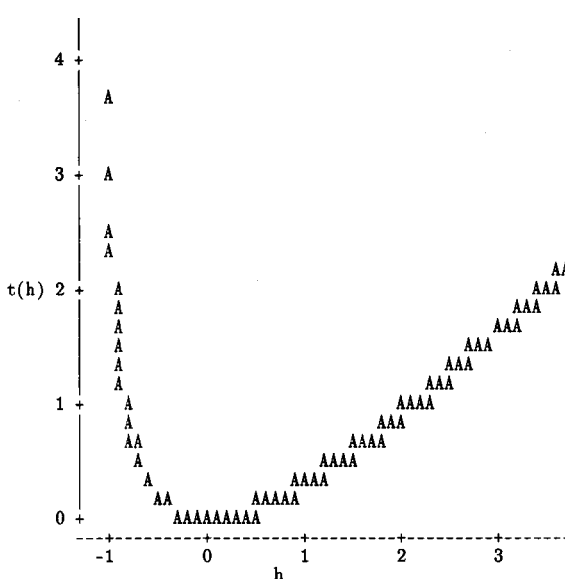


FIGURE 7.1. Plot of the function $t(h)$ near (

$$I(f,g) = \int f(x)\log\left(\frac{f(x)}{g(x)}\right)dx = \int f(x)t($$

The calculus proof that $t(x) \geq 0$ makes use of the first an
of the function $t(h) = h - \log(1+h)$:

$$t'(h) = \frac{h}{1+h},$$
$$t''(h) = \frac{1}{(1+h)^2}.$$

The set of critical points (which includes minima, max
points) of $t(h)$ consists of the solutions to $t'(h) = 0$ plus th
$(-1$ and infinity). In this case the unique solution is $h =$
that $h(x) = 0$ could occur for more than one value of x
extremum is deduced from $t''(0) = 1$, which, because it is
$h = 0$ is a minimum of the function $t(h)$ (and by unique
it is the only minimum). Therefore, for all h (and hence
Also, from these results, $t(h(x))$ is a convex function.

Deeper mathematical theory is required to prove that
$f(x) = g(x)$ for all x (in the relevant range of integratio
if $f(x) = g(x)$, then $I(f,g) = 0$. Part of the "deeper" n
to says that when $f(x)$ is a continuous probability den
$t(x) \geq 0$, then

$$I(f,g) = \int f(x)t(x)dx = 0$$

if and only if $t(x) = 0$ for all x in the range of integra
seems reasonably intuitive, so we will not belabor the po

$$h(x) - \log\big(1+h(x)\big) \equiv 0, \quad \text{for all } x$$

or

$$h(x) = \log\big(1+h(x)\big),$$

and finally,

$$e^{h(x)} = 1 + h(x).$$

The standard series expansion for e^h can be used here, w

$$1 + h(x) + \sum_{i=2}^{\infty} \frac{1}{i!}[h(x)]^i = 1 + h(x)$$

or

$$\sum_{i=2}^{\infty} \frac{1}{i!}[h(x)]^i = 0.$$

set of x values $\mathcal{N}$ for which

$$\int_{\mathcal{N}} f(x)dx > 0,$$

then we would have

$$\int g(x)dx < \int f(x)\,dx,$$

which cannot be true. Thus, because both $f(x)$ and $g(x)$ ar
functions, $h(x) \leq 0$ for all x, which implies that we mu
$h(x) = 0$ for all x.

Now we consider (in less detail) the discrete case

$$I(f, g) = \sum_{i=1}^{k} p_i \log\left(\frac{p_i}{q_i}\right),$$

such that $0 < p_i < 1$, $0 < q_i < 1$ for all i, and $\sum_{i=1}^{k} p_i$
For fixed k, this $I(f, g)$ is a function of $k - 1$ variables, w
$q_1, q_2, \ldots, q_{k-1}$ (the p_i distribution is considered fixed he

$$h_i = \frac{q_i - p_i}{p_i}, \qquad i = 1, \ldots, k,$$

whence

$$\frac{q_i}{p_i} = 1 + h_i, \qquad -1 < h_i < \infty, \qquad i = 1$$

As in the continuous case,

$$\sum_{i=1}^{k} p_i h_i = \sum_{i=1}^{k} (q_i - p_i) = 0,$$

so we can derive

$$I(f, g) = \sum_{i=1}^{k} p_i\big(h_i - \log(1 + h_i)\big) = \sum_{i=1}^{k} p$$

It was proved above that $t(h) \geq 0$; thus it must be that eve
$I(f, g) \geq 0$.

It is clear that if $p_i = q_i$ for all i, then $I(f, g) = 0$. A
Then it must be that $t(h_i) = 0$ for all i (otherwise, $I($
Therefore, we must have

$$e^{h_i} = 1 + h_i, \qquad i = 1, \ldots, k.$$

The set of indices $\{1, \ldots, k\}$ can be partitioned into two se
$h_i < 0$ for i in $\mathcal{N}$ and $h_i \geq 0$ for i in $\mathcal{P}$. For i in $\mathcal{P}$

Next, note that

$$
\begin{aligned}
0 &= \sum_{i=1}^{k} p_i h_i \\
&= \sum_{i=1}^{k} (q_i - p_i) \\
&= \sum_{i \text{ in } \mathcal{N}} (q_i - p_i) + \sum_{i \text{ in } \mathcal{P}} (q_i - p_i) \\
&= \sum_{i \text{ in } \mathcal{N}} (q_i - p_i).
\end{aligned}
$$

But i in $\mathcal{N}$ means $h_i < 0$, or $q_i - p_i < 0$, which would
sum would be strictly < 0, which is a contradiction. This
that the set $\mathcal{N}$ is empty: There cannot be any $h_i < 0$ if
$I(f, g) = 0$, then $f \equiv g$ (i.e., $p_i = q_i$, for all i in the disc

7.9 Summary

Most of this chapter is quite technical; we will try to provi
mary of key points or results. Sections 7.1 through 7.6 p
mathematical theory for K-L information-theoretic mode
eral theory is given (Sections 7.1–7.3) along with sever
cases (Sections 7.4–7.5) and some specific exploration
(Section 7.6). In particular, a very detailed derivation of
tion 7.2, along with the relationship to AIC. Then Section
not exhaustive) examination of the issue of whether we ca
AIC_c) rather than TIC; the results strongly support use of
ceptable, relative to TIC, but actually preferable. Section
provides simple (though not general) methods to explo
model selection that are operationally based on the log-li
oretical results are also given. Sections 7.7.6–7.7.10 give
results and considerations that do not fit elsewhere. Sectio
less technical than Sections 7.1–7.6, and we urge you to
the general insights therein. Section 7.8 gives a proof of

There are several rigorous derivations from Kullback–
leading to various information-theoretic criteria: The m
derivation leads to Takeuchi's (1976) information criteri
derivation of AIC_c is given in detail in Section 7.4. It is also
unique small-sample version of AIC, but AIC_c is recomme
(that could change in the future, especially for discrete di
7.5 gives a derivation of AIC for the exponential family o

from K-L information is much more exact, relies on few
is easier to understand.

The fact that such derivations exist is important to know
explanations are very detailed because the theory underly
based on K-L information is important to have clearly st
standing. Such understanding of the theory puts one in a
to accept use of the information-theoretic criteria and und
and weaknesses.

While Kullback–Leibler information is the logical basis
model selection, it turns out we must use *expected* (over
information as the quantity of interest when model para
mated. This, of course, is the reality of actual data analy
variance tradeoff and the associated model parsimony ach
model selection is an important byproduct of the approach
tions make it clear that K-L–based model selection doe
explicit objective of meeting the principle of parsimony
Rather, it is a natural consequence of data-based K-L mod
bias–variance tradeoff happens. In fact, it is because the
mated parameters, must minimize *expected* K-L informa
cross-validation property of AIC.

The detailed derivations make it clear that use of infor
teria in the analysis of real data is not based on the existen
or the notion that such a true model is in the set of candidat
Literature contrary to this point is mistaken.

Model selection attempts to establish some rigorous ba
parsimony in the model(s) used for inference. The relation
made clear, and investigations were undertaken to show tha
proxy for TIC. It seems poetic that AIC can be thought o
implementation of the more general TIC. The trace term
about equal to K for "good" models and does not depend
sample size is large. Some insights are provided to help i
relationship between $\mathrm{tr}[J(\underline{\theta}_o)I(\underline{\theta}_o)^{-1}]$ and K in a variety o
Evaluations were conducted for logistic regression, multi
count data, and normal regression models. In all cases we
term of TIC is very close to being K as long as the model st
error distribution are not drastically different from truth.
not the true data-generating model, the trace term was not
or $< K$. Rather, the matter is unpredictable; the model
and still the trace term can be any of $= K$, $> K$, or $<$
examined, however, if the model was less general than t
case), we predominantly found $\mathrm{tr}[J(\underline{\theta}_o)I(\underline{\theta}_o)^{-1}] < K$. Thu
then often lead to slightly more parsimonious models

If the set of models contains one or more good model
will tend not to be selected (or even ranked high), bec
be relatively large for a poor model and this term will do
value for that model, hence rendering the issue of use of TI
moot. As noted above, for good models use of AIC is acce
small samples and even for large samples if values of K b
to sample size. More research on such second-order impr
especially for discrete random variables.

Monte Carlo methods seem to be the only tool to asse
aspects of model selection and methods to incorporate mo
tainty. In some cases, asymptotic results can be obtained,
of little interest or practical use. We present some quick v
model selection using Monte Carlo simulation in the case
of models. For that same context, the theoretical expect
can be easily found, and this is explored in some detail
models with normal errors. The issue of AIC overfitting
plored. Extreme overfitting can occur, but the probability
and one way to minimize the problem is to keep the set o
small. Doing searches over "all possible models" (e.g., a
increases the risk of overfitting. In linear regression it se
tion is very similar to model selection based on minimum
course, in the analysis of real data we cannot do selection
theoretical MSE).

Summary

This book covers some philosophy about data analysis, interface between mathematical statistics and informatic practical statistical methodology useful in the applied sci we present a general strategy for modeling and data analys challenging examples from our fields of interest, provide not to do, and suggest some areas needing further theor We side with the fast-growing ranks that see limited util hypothesis testing. Finally, we provide references from t on these subjects for those wishing to study further.

Conceptually, there is information in the observed da express this information in a compact form via a "model resents a scientific hypothesis and is then a basis for mak the process or system that generated the data. One can vie mation in data as a change in "coding" like a change in l or emotion expressed in one language (e.g., French) lo when expressed in another language (e.g., Russian). A g only a finite, fixed amount of information. The (unachiev selection is to attain a perfect 1-to-1 translation such th lost in going from the data to a model of the information are only approximations, and we cannot hope to perfectl ized goal. However, we can attempt to find a model of th the sense that the model loses as little information as pos leads directly to Kullback–Leibler information $I(f, g)$: when model g is used to approximate full reality f. We model that minimizes K-L information loss. Because we

information-theoretic criteria (e.g., AIC, AIC_c, QAIC, or
model allows the efficient and objective separation or filtr
from *noise*. In an important sense, we are not really tryin
instead, we are trying to model the *information* in the dat

**While we use the notation f to represent truth or f
the existence of a "true model" in the life sciences**.
be the process (truth) that generates the sample data w
to make inferences about truth, while realizing that full
be beyond us when we have only sample data. Data ana
thought of as an attempt to identify f; instead, we must
good approximations to truth and from which therefore
inferences concerning truth. We do not want merely to de
a model that has a very large number of parameters; instea
data to aid in the selection of a parsimonious model that all
to be made about the system or process under study. A p
representing a well-defended scientific hypothesis, aids i
of the system of interest.

Relatively few statistics books provide a summary of the
yet fewer provide an effective, unified strategy for data an
where there is substantial complexity. The breadth of th
covered here makes a summary difficult to write. Underg
casionally ask the professor, "What is important for me t
examination?" The professor is typically irritated by such
the student should realize that it is *all* important! Indeed,
Akaike's pioneering work is that it *is* all important. The in
paradigm is a *package*; each of the package's contents is
but it is the integration of the contents that makes for an ef
consistent strategy, and a practical and powerful methodol
package that has been so frequently left out is the critical
generation, and modeling *before* examination of the data;
thinking should occur prior even to data collection. Thi
the science of the issue formally enters the overall "analy
Burnham 1999a).

The information-theoretic methods we present can be u
best model that can be used in making inferences from en
often portrayed in the literature in this simple manner. T
is much richer than this simplistic portrayal of model sele
**In fact, an emphasis of this second edition is multimod
MMI has several advantages; all relate to the broa
selection uncertainty**. One can easily rank alternative m
from best to worst using the convenient differences Δ_i
each model, given the data [i.e., $\mathcal{L}(g_i \mid data)$], can be easily

a subset of good models. Evidence ratios are useful for
support of one model versus another, given the data; su
irrespective of other models in the set.

Model selection uncertainty can be easily quantified us
(the bootstrap is an alternative). Estimates of this compone
be incorporated into unconditional estimates of precision
ods. For many problems (e.g., prediction) model-averag
and we treat this important issue in Chapters 4–5. Thus, w
formal inference from all models in the set.

For those who have scanned through the pages of this
surprise at the general lack of mathematics and formu
7 being the exceptions). That has been our intent. The
information-theoretic methods is relatively simple. They
stand and use ("low tech"), while the underlying theory
Chapter 7). As we wrote the book and tried to understa
papers (see Parzen et al. 1998) we found the need to delv
that are generally philosophical. The science of the proble
into modeling *before* one begins to rummage through the
In some critical respects, applied statistics courses are fa
tics as an integral part of scientific discovery, with little
model selection methods or their importance, while succ
teaching null hypothesis testing methods and data analys
the assumption that the model is both true and given. Se
note, "The standard approach in teaching—stressing the
a p value while warning against its misinterpretation—
abysmal failure." It seems necessary to greatly reduce the r
(Anderson et al. 2001b and d).

8.1 The Scientific Question and the Colle

The formulation of the research question is crucial in inve
plex systems and processes in the life sciences. A goo
question is a mistake all too often seen in the published l
better than a poor answer to a poor question. Investigators
readdress the importance and quality of the question to be
scientific hypotheses, represented by models, must have a
the table.

A careful program of data collection must follow from th
Particular attention should be placed on the variables t
interesting covariates. Observational studies, done well,

come from more formal experimentation (i.e., importa
tors are controlled or balanced, experimental units are ra
treatment and control groups with adequate replication),
al. (1980), Gail (1996), Beyers (1998), and Glymour (1
philosophies. Valid inference must assume that these ba
have been carefully planned and conducted. Before one
general questions must be answered in the affirmative:

Are the study objectives sound, relevant, and ac
Has there been proper attention to study design and labora

8.2 Actual Thinking and A Priori Modeli

Fitting models, each representing a scientific hypothesi
important in many biological, ecological, and medical
statistical inferences about the system of interest are made
parsimonious model of the observational or experimenta
see this activity increase as more complicated scientific an
are addressed. In particular, a priori modeling becomes in
as several data sets are collected on the same issue by diff
at widely differing field sites over several years.

We recommend much more emphasis on thinking! Lea
for a while, giving time to think hard about the overall pr
information is contained in the published literature, even
what related to the issue at hand? What nonlinearities a
might be predicted? What interactions are hypothesized t
two or more variables be combined to give a more mea
analysis? Should some variables be dropped from consid
should be encouraged with the people in the field or labor
to the data collection. What parameters might be simila
data sets)? Model building should be driven by the underly
sue combined with a good understanding of mathematical
important conceptual phase might take several days or e
this seems far more time than is often spent under curren

**Biologists generally subscribe to the philosophy of
hypotheses" (Chamberlain 1890, Platt 1964, Mayr 199
form the basis for the set of candidate models to be c**
Model building can begin during the time that the a prio
being sorted out. Modeling must carefully quantify the s
interest. Often it is effective to begin with the global mo
some lower-dimensional models. Others may favor a b
The critical matter here is that one arrives, eventually, at

however, every attempt should be made to keep the number small.

Critical Thinking

Our science culture does not do enough to regularly c
critical thinking. This failure has slowed the scientific c

We fail to fault the trivial content of the typical ecolog

There is a need for more *careful thinking* (than is usu
better balance between scientific hypotheses, data, and a

Chamberlin's concept of *multiple working hypotheses*,
100 years ago, has a deep level of support among scienc
thought the method led to "certain distinctive habits of r
value in education." Why has this principle not become
than the rare exception, in so many fields of applied scie

Platt (1964) noted that years and decades can be wast
unless one thinks carefully in advance about what the r
conclusive experiments would be.

With the information-theoretic approach, there is no
hypothesis, or a statistical hypothesis test, or an arbitrary
able power, or the multiple testing problem, or the fact th
hypothesis is nearly always *obviously* false in the first pla
plication of statistical hypothesis testing arbitrarily classi
meaningless categories of "significant" and "nonsignifica
has little to contribute to the advancement of science (An
We recommend that researchers stop using the term "sign
overused, uninformative, and misleading. The results of m
on estimates of expected (relative) Kullback–Leibler info
different from the results of some form of statistical hyp
the simulated starling data, Section 3.4, or the sage grous

So, investigators may proceed with inferential or confir
if they feel satisfied that they can objectively address two

Was the set of candidate models derived a p
What justifies this set?

The justification should include a rationale for models bo
cluded from the set. A carefully defined set of models
information-theoretic methods are used to select the si
the entire set of models is used to reach defensible infe
known about the system under study that a large numbe
included in the candidate set, then the analysis should pro
only exploratory (if models are developed as data analysis
exploratory and risky). **One should check the fit or ade**
model using standard methods. If the global model is

tem under study and the data collected. There is no substitute for hard thinking at this point (Platt 1964).

8.3 The Basis for Objective Model Selec

Statistical inference from a data set, *given a model*, is supported by a very large amount of theory. Theorists and p tinely employing this theory, either likelihood or least squ of problems in the applied sciences. The most compellin *model to use*?" Valid inference must usually be based on a model, but which one?

Akaike chose the celebrated Kullback–Leibler discrim as a basis for model selection. This is a fundamental quanti has earlier roots in Boltzmann's concept of *entropy*, a cr of nineteenth-century science. The K-L distance betwee and model g is defined for continuous functions as the in

$$I(f, g) = \int f(x) \log \left(\frac{f(x)}{g(x \mid \theta)} \right) dx,$$

where log denotes the natural logarithm and f and g are *n* bility distributions. Kullback and Leibler (1951) develope "information theory," thus the notation $I(f, g)$ as it rela tion" lost when model g is used to approximate truth f. an approximating model that loses as little information equivalent to minimizing $I(f, g)$ over the models in th considered to be fixed. An interpretation equivalent to m that we seek an approximating model that is the "shortest Both interpretations seem useful and compelling.

The K-L distance can be written equivalently as

$$I(f, g) = \int f(x) \log(f(x))\, dx - \int f(x) \log(g($$

The two terms on the right in the above expression are sta with respect to f (truth). Thus, the K-L distance (above) a difference between two expectations,

$$I(f, g) = \mathrm{E}_f[\log(f(x))] - \mathrm{E}_f[\log(g(x \mid$$

each with respect to the true distribution f. The first expecta is a constant that depends only on the unknown true dis treating this unknown term as a constant, only a measure

$I(f, g) - \text{constant} = -\mathrm{E}_f[\log(g(x \mid \theta)$

or

$I(f, g) - \text{constant} = -\mathrm{E}_f[\log(g(x \mid \theta)$

Thus, the term $\big(I(f, g) - \text{constant}\big)$ is a *relative* distance model g. This provides a deep theoretical basis for model compute or estimate $\mathrm{E}_f[\log(g(x \mid \theta))]$.

Akaike (1973, 1974, 1985, 1994) showed that the estimating relative K-L information was

$$\mathbf{E}_y\mathbf{E}_x[\mathbf{log}(\boldsymbol{g}(\boldsymbol{x}|\hat{\boldsymbol{\theta}}(\boldsymbol{y})))],$$

where y and x are independent random samples from th and both statistical expectations are taken with respect to ble expectation, both with respect to truth f, is the targe approaches based on K-L information.

8.4 The Principle of Parsimony

Parsimony is the concept that a model should be as sim respect to the included variables, model structure, and nu Parsimony is a desired characteristic of a model used fo usually visualized as a suitable tradeoff between square of parameter estimators (Figure 1.3). Parsimony lies betw derfitting and overfitting (Forster and Sober 1994, Fors K-L information is a fundamental basis for achieving p modeling.

The concept of parsimony has a long history in the sc expressed as "Occam's razor": shave away all that is unn is to make things "as simple or small as possible." Par represents a tradeoff between bias and variance as a funct of the model (K). A good model is a proper balance betw overfitting, given a particular sample size (n). Most mod are based on the concept of a squared bias versus varianc of a model from a set of approximating models must em parsimony. These philosophical issues are stressed in th some experience and reconsideration to reach a full un importance.

Roots of Theory

As deLeeuw (1992) noted, Akaike found a form
tween Boltzmann's entropy and Kullback–Leibler info
paradigms in information and coding theory) and maxin
dominant paradigm in statistics).

This finding makes it possible to combine estimatio
val estimation) and model selection under a single theo
optimization.

Akaike's (1973) breakthrough was the finding of an estir
relative K-L information, based on a bias-corrected maxi
value. His estimator was an approximation and, under
asymptotically unbiased. He found that

estimated expected (relative) K-L information ≈

where $\log(\mathcal{L}(\hat{\theta}))$ is the maximized log-likelihood value an
estimable parameters in the approximating model (this i
term). Akaike multiplied through by -2 and provided A
criterion (AIC)

$$\mathbf{AIC} = -2\log(\mathcal{L}(\hat{\theta})) + 2K.$$

Akaike considered his information-theoretic criterion an e
likelihood theory. Conceptually, the principle of parsimon
added "penalty" (i.e., $2K$) while minimizing AIC.

Assuming that a set of a priori candidate models has be
then AIC is computed for each of the approximating mod
model where AIC is minimized is selected as best for the er
This is a simple, compelling concept, based on deep the
(i.e., K-L information). Given a focus on a priori issues, m
scientific hypotheses, and model selection, *the inference i*
In a sense, parameter estimates are almost byproducts of
This inference relates to the estimated best approximatio
information seems to be contained in the data.

Important refinements followed shortly after the pioneer
Most relevant was Takeuchi's (1976) information criterion
provided an asymptotically unbiased estimate of relative
mation. TIC is little used, since it requires the estimation o
first and second partial derivatives of the log-likelihood fu
tical use hinges on the availability of a relatively large sar
AIC can be viewed as a parsimonious version of TIC. A
was motivated by Sugiura's (1978) work, and resulted in
Hurvich and Tsai (1989, 1990b, 1991, 1994, 1995a and

$$\mathbf{AIC}_c = -2\log(\mathcal{L}(\hat{\theta})) + 2K + \frac{2K(K+}{(n-K-}$$

where n is sample size The final bias-correction term van with respect to K (and AIC_c becomes AIC), but the addit tant if n is not large relative to K (we suggest using AI alternatively, always using AIC_c).

A third extension was a simple modification to AIC ar persed count data (Lebreton et al. 1992). A variance computed from the goodness-of-fit statistic, divided by dom, $\hat{c} = \chi^2/\ \mathrm{df}$. The value of the maximized log-lik divided by the estimate of overdispersion to provide a pr log-likelihood. These criteria are denoted by QAIC and derived from quasi-likelihood theory (Wedderburn 1974)

$$\mathbf{QAIC} = -[2\log(\mathcal{L}(\hat{\theta}))/\hat{c}] + 2K,$$

and

$$\begin{aligned}\mathbf{QAIC}_c &= -[2\log(\mathcal{L}(\hat{\theta}))/\hat{c}] + 2K + \frac{2K(K}{n-K}\\ &= \mathbf{QAIC} + \frac{2K(K+1)}{n-K-1}.\end{aligned}$$

When no overdispersion exists, $c = 1$, and the formulas fo reduce to AIC and AIC_c, respectively. There are other, more to account for overdispersion in count data, but this sim quite satisfactory. Methods are given in Chapter 6 to allow of the data to have partition-specific estimates of overdisp number of estimable parameters (K) must include the n of c. Thus, if males and females have different degrees o these are to be estimated from the data, then K must inclu these estimates.

AIC is often presented in the scientific literature in an the bias-correction term K (the so-called penalty term) w yet, perhaps, is that AIC is often given without reference to with Kullback–Leibler information. Such shallow presenta have had very negative effects, and have misled many int is a whole class of selection criteria that are "information 6). Criteria such as AIC, AIC_c, QAIC, and TIC are es (relative) Kullback–Leibler distance and are useful in the in the "noisy" sciences.

Because only *relative* K-L information can be estimat
information criteria, it is convenient to rescale these value
with the minimum AIC (or AIC_c or TIC) has a value of 0
criterion values can be rescaled as simple differences,

$$\begin{aligned}\Delta_i &= AIC_i - AIC_{min} \\ &= \hat{E}_{\hat{\theta}}[\hat{I}(f, g_i)] - \min \hat{E}_{\hat{\theta}}[\hat{I}(f, g_i)]\end{aligned}$$

While the value of minimum $\hat{E}_{\hat{\theta}}[\hat{I}(f, g_i)]$ is not known (onl
we have an estimate of the size of the increments of i
the various models compared to the estimated best mo
the minimum $E_{\hat{\theta}}[\hat{I}(f, g_i)]$). The Δ_i values are easy to
a quick comparison and ranking of candidate models
in computing Akaike weights. As a rough rule of thu
Δ_i within 1–2 of the best model have substantial sup
ceive consideration in making inferences. Models havi
4–7 of the best model have considerably less support,
$\Delta_i > 10$ have either essentially no support and mig
further consideration or at least fail to explain some s
variation in the data. If the observations are not indepen
as such) or if the sample size is quite small, or if th
number of models, then the simple guidelines above ca
hold.

There are cases where a model with $\Delta_i > 10$ might st
ularly if the sample size is very large (e.g., see Section
let model A, with year-specific structure on one of the pa
model in the set ($\Delta_A = 0$) and model B, with less struct
year-specific parameters, have $\Delta_B = 11.4$. Assume that a
didate set were derived prior to data analysis (i.e., no dat
model A is able to identify important variation in a parame
is important. However, in terms of understanding and ge
based on the data, it might sometimes be justified to us
B, because it may seem to "capture" the important fixe
and B should both be detailed in any resulting publicat
ing and interpretation might be enhanced using model B
information in the data would be (intentionally) lost. Su
could be partially recovered by, for example, using a rand
(see Section 3.5.5) to estimate the mean of the time-effec
variance of its distribution.

The principle of parsimony provides a philosophical ba
tion; Kullback–Leibler information provides an objective
fundamental theory; and the information criteria (particu
provide a practical, general methodology for use in data

in the set of candidates (although this is sometimes stated technical literature). [We note that several "dimension-con been published that attempt to provide asymptotically un tent") estimates of the dimension (K) of the "true mode only estimates of K-L information in a strained way, are assumption sets, and often perform poorly (even toward tive) unless a very large sample size is available (or wh small, such as in many problems in the physical sciences mend these dimension-consistent criteria for the analysi life sciences.]

8.7 Scaling Alternative Models

The information-theoretic approach does more than me model is best for making inference, given the set of a pri and the data. The Δ_i allow a ranking of the models from the worst; the larger the Δ_i, the less plausible is model i. **not reasonable to expect to be able to make inferences model; biology is not simple; why should we hope for from a single model?** The information-theoretic paradi for examination of alternative models and, where appropr inference from more than one model (MMI).

The simple transformation $\exp(-\frac{1}{2}\Delta_i)$ results in the (di model i, given the data $\mathcal{L}(g_i|x)$. These are functions in $\mathcal{L}(\theta|x, g_i)$ is the likelihood of the parameters θ, given t model (g_i). These likelihoods are very useful; for exampl for model i versus model j is merely

$$\mathcal{L}(g_i|x)/\mathcal{L}(g_j|x).$$

It is convenient to normalize these likelihoods such that t

$$w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^{R}\exp(-\frac{1}{2}\Delta_r)},$$

and interpret these as a weight of evidence. Akaike (e.g., A 1980, and 1981b; also see Kishino 1991 and Buckland et these values, and we have found them to be simple and v dence ratio of model i versus model j is then just w_i/w the ratio of the likelihood $\mathcal{L}(g_i|x)/\mathcal{L}(g_j|x)$. Drawing on Ba interpret w_i as the estimated probability that model i is the the data at hand, given the set of models considered (see

through here was to consider priors on models that are
and K (we call this class of model priors "savvy," i.e., sh
then AIC and AIC_c fall out as a strictly Bayesian result. I
Bayesian derivative, it is compelling to interpret the Aka
erior model probablilities. While many (objective) Bayes
with the use of a defuse or noninformative prior on mo
a uniform prior on a model parameter), use of such defu
(such as $1/R$) may have poor properties or unintended co
some priors on models may be uninformative, but not i
the Bayesian derivation of AIC (or AIC_c) and BIC diffe
on models. However, these criteria are fundamentally dif
substantive ways. In this book we place an emphasis on t
and AIC_c as bias-corrected estimates of Kullback–Leibler
this seems so much more objective and fundamental.

The w_i are useful as the "weight of evidence" in favor of
actual K-L best model in the set. The bigger the Δ_i, the sm
the less plausible is model i as being the best approximati
is conditional on both the data and the set of a priori mod

Alternatively, one could draw B bootstrap samples (B sh
rather than 1,000), use the appropriate information crite
model for each of the B samples, and tally the proportion
the ith model was selected. Denote such bootstrap-selec
$\hat{\pi}_i$. While w_i and $\hat{\pi}_i$ are not estimates of exactly the same
closely related and provide information concerning the u
model for use. The Akaike weights are simple to compute
weights are computer-intensive and not practical to con
(e.g., the simulated starling experiment, Section 3.4), b
bootstrap repetitions must be drawn and analyzed.

Under the hypothesis-testing approach, nothing can ge
ranking or scaling models, particularly if the models were
least squares problems one could turn to adjusted R^2 value
of models, but other kinds of models cannot be scaled u
very poor) approach (see the analogy in Section 2.5).

8.8 MMI: Inference Based on Model Ave

Rather than base inferences on a single selected best mo
set of models, we can base our inferences on the entire
averaging. The key to this inference methodology is the A
if a parameter θ is in common over all models (as θ_i in m
is prediction, by using the weighted average we are basin

$$\theta = \sum_{i=1}^{R} w_i \theta_i,$$

or

$$\hat{\bar{\theta}} = \sum_{i=1}^{R} \hat{\pi}_i \hat{\theta}_i.$$

This approach has both practical and philosophical advanta
averaged estimator can be used, it appears to have better p
bias compared to $\hat{\theta}$ from the selected best model.

If one has a large number of closely related models, su
based variable selection (all-subsets selection), designat
model is unsatisfactory, because that estimated "best" mo
from data set to data set. In this situation model-averaging
much more stabilized inference. The concept of inferen
the models can be used to reduce model selection bias e
coefficient estimates in all-subsets selection. For the re
associated with predictor x_j we use the estimate $\hat{\bar{\beta}}_j$, wh
regression coefficient β_j averaged over all models in whi

$$\hat{\bar{\beta}}_j = \frac{\sum_{i=1}^{R} w_i I_j(g_i) \hat{\beta}_{j,i}}{w_+(j)},$$

$$w_+(j) = \sum_{i=1}^{R} w_i I_j(g_i),$$

where i is for model $i = 1, \ldots, R$, j is for predictor vari

$$I_j(g_i) = \begin{cases} 1 \text{ if predictor } x_j \text{ is in model } g \\ 0 \text{ otherwise.} \end{cases}$$

Conditional on model g_i being selected, model selecti
biasing $\hat{\beta}_{j,i}$ away from zero. Thus a new estimator, denote

$$\tilde{\bar{\beta}}_i = w_+(i) \hat{\bar{\beta}}_i.$$

Investigation of this idea, and extensions of it, is an ope
point here is that while $\hat{\bar{\beta}}_j$ can be computed ignoring m
ones x_j appears in, $\tilde{\bar{\beta}}_i$ does require fitting all R of the a p

8.9 MMI: Model Selection Uncertainty

At first, one might think that one could use an informatio
approximating model that was "close" to truth (remembe

been specified a priori as the only model considered. Ac
would not be terrible, since at least one would have a re
lected objectively, based on a valid theory and a priori
approach would often be superior to much of current pra
case where the best model has an Akaike weight $> 0.$
considering only this model, and the usual measures of p
on this selected model, is that this tends to overestimate
(1992) calls the failure to acknowledge model selection
scandal." [We might suggest that the widespread use of s
testing and blatant data dredging in model selection repres
In fact, there is a variance component due to model selec
should be incorporated into estimates of precision such t
ditional (on the selected model). While this is a research
development, several useful methods are suggested in th
will surely appear in the technical literature in the next f
additional Bayesian approaches.

The Akaike (w_i) or bootstrap (π_i) weights that are us
models can also be used to estimate unconditional precis
in the parameter θ over R models (model g_i, for $i = 1, \ldots$

$$\widehat{\mathbf{var}}(\bar{\hat{\theta}}_i) = \left[\sum_{i=1}^{R} w_i \sqrt{\widehat{\mathbf{var}}(\hat{\theta}_i \mid g_i) + (\hat{\theta}_i - \bar{\hat{\theta}}} \right.$$

$$\widehat{\mathbf{var}}(\bar{\hat{\theta}}_i) = \left[\sum_{i=1}^{R} \pi_i \sqrt{\widehat{\mathbf{var}}(\hat{\theta}_i \mid g_i) + (\hat{\theta}_i - \bar{\hat{\theta}}} \right.$$

These estimators, from Buckland et al. (1997), include a
tional sampling variance, given model g_i (denoted by $\hat{v}$
incorporate a variance component for model selection u
These estimators of unconditional variance are also appro
one wants a model-averaged estimate of the parameter w
models.

Chapter 4 gives some procedures for setting confidence
model selection uncertainty, and it is noted that achieved
coverage is then a useful measure of the utility of methods
selection uncertainty into inference. Only a limited aspect
can be currently handled. *Given* a set of candidate models
lection method, we can assess selection uncertainty. The u
the set of models cannot be addressed; we lack a theory for
lack good, general guidelines for defining the a priori set o
papers to appear on these scientific and philosophical issu

Inference on the importance of a variable is similarly impr
on all the models. If one selects the best model and says th
are the important ones and the other variables are not imp
naive, unreliable inference. We suggest that the relative im
x_j be measured by the sum of the Akaike weights over
that variable appears:

$$w_+(j) = \sum_{i=1}^{R} w_i I_j(g_i).$$

Thus again, proper inference requires fitting all the mod
type of model-averaging. A certain balance in the number
model j, must be achieved. When possible, one should
on all the models, via model-averaging and selection bias
than risk making inference based only on the model esti
and, often, ignoring other models that are also quite good

8.11 More on Inferences

Information-theoretic methods do not offer a mechanical,
to science. While these methods can certainly be misuse
thinking as models are developed to represent the multiple
that must be the focus of the entire study. A central theme
attention to the need to ask better scientific questions in
(Platt 1964). Rather than test trivial null hypotheses, it is
questions relating to well-defined alternative hypotheses
achieved, a great deal more hard thinking will be required

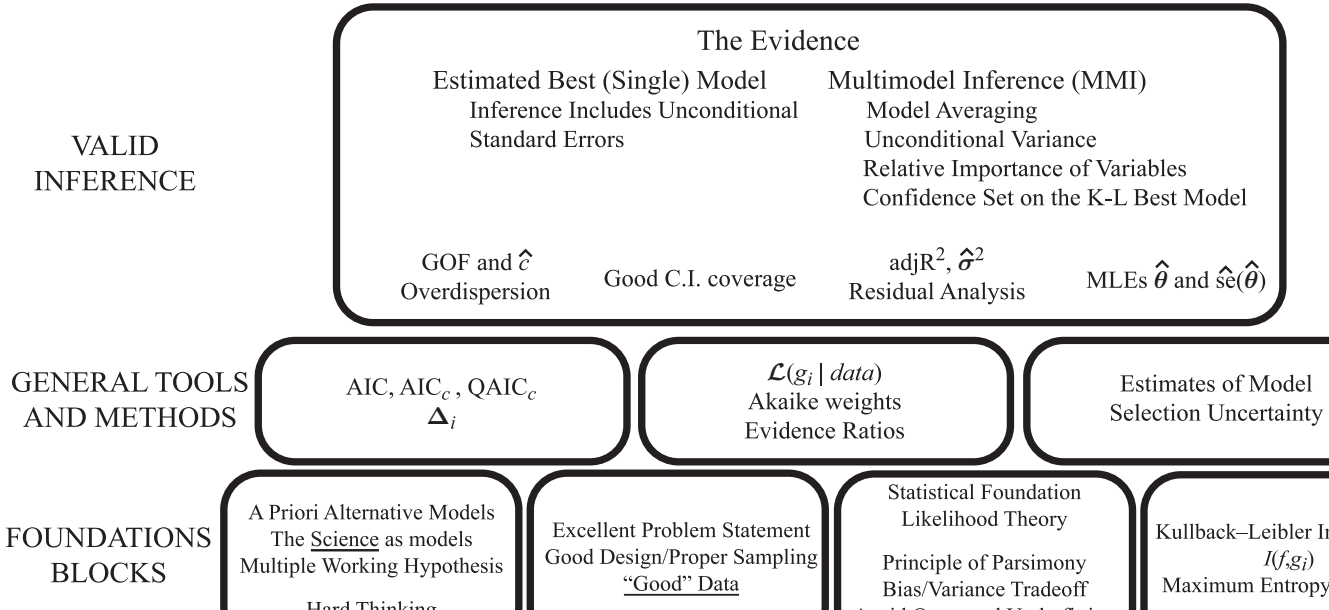
There needs to be increased attention to separating thos
on a priori considerations from those resulting from some
ing. White (2000:1097) comments, "Data snooping is a d
be avoided, but in fact is endemic."

Essentially no justifiable theory exists to estimate precis
ses, for those still so inclined) when data dredging has tak
(mis)used is for a priori analyses, assuming that the model
to the data). A major concern here is the finding of effects a
are actually spurious where inferences are made post hoc
Anderson et al. 2001b). This glaring fact is either not u
tioners and journal editors or is simply ignored. Two typ
include (1) an iterative approach, in which patterns and
after initial analysis are "chased" by repeatedly building ne
effects included and (2) analysis of "all possible models.
poor approach to making inferences about the sampled p

The information theoretic paradigm avoids statistical
ing concepts and focuses on relationships of variables (v
the estimation of effect size and measures of its precisio
primarily in the context of making inferences from a si
or making robust inference from many models (e.g., usi
based on Akaike weights). Data analysis is a process of l
are supported by the data and the degree of complexity c
the set. Often, models other than just the estimated best mo
information. Evidence ratios and confidence sets on mo
inferences on all, or several of the best, models in the set. In
approaches should not be used unthinkingly; a good set
is essential, and this involves professional judgment and
scientific hypotheses into the model set.

When the analysis of data has been completed und
theoretic approach, one should gather and report on the tot
at hand. The primary evidence might be the selected mod
estimates and appropriate measures of precision (includ
ponent for model selection uncertainty.) The ranks of ea
and the Akaike weights should be reported and interpret
parameter estimates are often important, particularly for
ratios, confidence sets on the K-L best model, and a rankin
portance of predictor variables are often useful evidence
quantities such as adjusted R^2 and $\hat{\theta}^2$ should be reported
model. The results from an analysis of residuals for the s
also be important to report and interpret. Every effort sho
and objectively report on all the evidence available. If some
ing post hoc activities, this should be clearly stated in publ
8.1 provides a simplistic graphical representation of the in
approach. The point of Figure 8.1 is to reinforce some
(bottom building blocks) and the practical tools and meth
blocks) that rest on these foundations. If these are used
tively, one can hope to provide compelling evidence allow
The weakest link seems often to be the left block on th
deeply about the science problem and the alternative hyp

It seems worth noting that K-L information and MMI c
types of conflict resolution where data exist that are cer
resolution of the conflict (Anderson et al. 1999, 2001*c*).
take us too far afield; however, as Hoeting et al. (1999) n
context), "Model averaging also allows users to incorpora
models in the estimation process; thus model averaging m
of scientists a better estimation method than the traditiona
to get the committee to agree on a best model."


The Evidence
Estimated Best (Single) Model
Inference Includes Unconditional Standard Errors
Multimodel Inference (MMI)
Model Averaging
Unconditional Variance
Relative Importance of Variables
Confidence Set on the K-L Best Model
VALID INFERENCE
GOF and $\hat{c}$
Overdispersion
Good C.I. coverage
adjR^2, $\hat{\sigma}^2$
Residual Analysis
MLEs $\hat{\theta}$ and $\hat{se}(\hat{\theta})$
GENERAL TOOLS AND METHODS
AIC, AIC$_c$, QAIC$_c$
Δ_i
$\mathcal{L}(g_i \mid data)$
Akaike weights
Evidence Ratios
Estimates of Model Selection Uncertainty
FOUNDATIONS BLOCKS
A Priori Alternative Models
The Science as models
Multiple Working Hypothesis
Excellent Problem Statement
Good Design/Proper Sampling
"Good" Data
Statistical Foundation
Likelihood Theory
Principle of Parsimony
Bias/Variance Tradeoff
Kullback–Leibler I
$I(f,g_i)$
Maximum Entropy

At a conceptual level, reasonable data and a *good* model
of "information" from "noise." Here, information relate
relationships, estimates of model parameters, and comp
Noise then refers to the residuals; variation left unexplai
information extracted from the data to make proper infer

Summary

We want an approximating model that minimizes infor
and properly separates noise (noninformation, or entro
information. The philosophy for this separation is the prin
the conceptual target for such partitioning is Kullback–L
and the tactic for selection of a best model is an informa
AIC, AIC_c, $QAIC_c$, or TIC). The notion of data-based r
resulting inference is a very difficult subject, but we do kn
uncertainty about the selected model can often be expe
incorporated into estimates of precision.

Still, model selection (in the sense of parsimony) is the
analysis. In using the more advanced methods presented h
can be thought of as a way to compute Akaike weights. T
more models in the set as a way to make robust inferences f
More research is needed on the quantification of model u
of the plausibility of alternative models, ways to reduce n
and ways to provide effective measures of precision (witho
on a given model). Confidence intervals with good achie
a goal of inference following data-based model selection.

Information-theoretic methods are relatively simple to u
tical to employ across a very wide class of empirical situ
disciplines. The information-theoretic approach unifies p
and model selection under an optimization framework,
Leibler information and likelihood theory. With the excep
the methods are easy to compute by hand if necessary (as
the MLEs, maximized log-likelihood values, and $\widehat{\text{var}}(\hat{\theta}_i \mid$
models). Researchers can easily understand the informatio
presented here; we believe that it is *very* important that res
the methods they employ.

References

Agresti, A. (1990). *Categorical data analysis*. John Wiley and Son

Aitkin, M. (1991). Posterior Bayes factors (with discussion). *Journa*
Society, Series B, **53**, 111–143.

Akaike, H. (1973). Information theory as an extension of the maxim
Pages 267–281 *in* B. N. Petrov, and F. Csaki, (eds.) *Second Inter*
Information Theory. Akademiai Kiado, Budapest.

Akaike, H. (1974). A new look at the statistical model identificatio
Automatic Control AC **19**, 716–723.

Akaike, H. (1976). Canonical correlation analysis of time series and t
criterion. Pages 27–96 *in* R.K. Mehra, and D.G. Lainiotis (eds.
Advances and Case Studies. Academic Press, New York, NY.

Akaike, H. (1977). On entropy maximization principle. Pages 27
(ed.) *Applications of statistics*. North-Holland, Amsterdam, The

Akaike, H. (1978a). A new look at the Bayes procedure. *Biometrik*

Akaike, H. (1978b). A Bayesian analysis of the minimum AIC p
Institute of Statistical Mathematics **30**, 9–14.

Akaike, H. (1978c). On the likelihood of a time series model. *The S*

Akaike, H. (1979). A Bayesian extension of the minimum AIC proc
model fitting. *Biometrika* **66**, 237–242.

Akaike, H. (1980). Likelihood and the Bayes procedure (with discus
J.M. Bernardo, M.H. De Groot, D.V. Lindley, and A.F.M. Smith (e
University Press, Valencia, Spain.

Akaike, H. (1981b). Modern development of statistical methods Eykhoff (ed.) *Trends and progress in system identification*. Perg

Akaike, H. (1983a). Statistical inference and measurement of ent G.E.P. Box, T. Leonard, and C-F. Wu (eds.) *Scientific inferen robustness*. Academic Press, London.

Akaike, H. (1983b). Information measures and model selection. *Institute* **44**, 277–291.

Akaike, H. (1983c). On minimum information prior distributions. *Statistical Mathematics* **35A**, 139–149.

Akaike, H. (1985). Prediction and entropy. Pages 1–24 *in* A.C. Atki (eds.) *A celebration of Statistics*. Springer, New York, NY.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika* **52**, 31

Akaike, H. (1992). Information theory and an extension of the max ple. Pages 610–624 *in* S. Kotz, and N.L. Johnson (eds.) *Breakth* 1. Springer-Verlag, London.

Akaike, H. (1994). Implications of the informational point of view statistical science. Pages 27–38 *in* H. Bozdogan (ed.) *Engineerin tions*. Vol. 3, Proceedings of the First US/Japan Conference on th Modeling: An Informational Approach. Kluwer Academic Pub Netherlands.

Akaike, H., and Nakagawa, T. (1988). *Statistical analysis and con* KTK Scientific Publishers, Tokyo. (English translation by H. Aka

Allen, D.M. (1970). Mean square error of prediction as a criterion *Technometrics* **13**, 469–475.

Amari, S. (1993). Mathematic methods of neurocomputing. Pages Nielson, J.L. Jensen, and W.S. Kendall (eds.) *Networks and probabilistic aspects*. Chapman and Hall, New York, NY.

Anonymous. (1997). *The Kullback Memorial Research Conference* The George Washington University, Washington, D.C.

Anderson, D. R. (2001). The need to get the basics right in wildli *Society Bulletin* **29**, 1294–1297.

Anderson, S., Auquier, A., Hauck, W.W., Oakes, D., Vandaele, (1980). *Statistical methods for comparative studies*. John Wiley

Anderson, D.R., and Burnham, K.P. (1976). *Population ecology effect of exploitation on survival*. U.S. Fish and Wildlife Servic No. 128.

Anderson, D.R., and Burnham, K.P. (1999a). General strategies analysis of ringing data. *Bird Study* **46** (suppl.), S261–270.

Anderson, D.R., Burnham, K.P., and White, G.C. (1994). Al overdispersed capture–recapture data. *Ecology* **75**, 1780–1793.

Anderson, D.R., Burnham, K.P., and White, G.C. (1998). Compar for model selection and statistical inference from capture–recap *Applied Statistics* **25**, 263–282.

Anderson, D.R., Burnham, K.P., Franklin, A.B., Gutierrez, R.J., F R.G., White, G.C., and Shenk, T.M. (1999). A protocol for conflic empirical data related to natural resource controversies. *Wildl* 1050–1058.

Anderson, D.R., Burnham, K.P., and Thompson, W.L. (2000). N problems, prevalence, and an alternative. *Journal of Wildlife Ma*

Anderson, D.R., and K.P. Burnham. 2001*a*. Commentary on model *the Ecological Society of America* **82**, 160–161.

Anderson, D.R., Burnham, K.P., Gould, W.R., and Cherry, S. (2 finding effects that are actually spurious. *Wildlife Society Bulleti*

Anderson, D.R., Burnham, K.P., and White, G.C. (2001*c*). Kullbacl resolving natural resource conflicts when definitive data exist. *W* **29**, 1260–1270.

Anderson, D.R., Link, W.A., Johnson, D.H. and Burnham, K.P. (2 presenting results of data analyses. *Journal of Wildlife Manageme*

Apostol, T.M. (1957). *Mathematical analysis: a modern approach* Addison-Wesley Publishing Co., Inc. Reading, MA.

Armitage, P. (1957). Studies in the variability of pock counts. *J* 564–581.

Atilgan, T. (1996). Selection of dimension and basis for density e of dimension, basis and error distribution for regression. *Commu Theory and Methods* **25**, 1–28.

Atkinson, A.C. (1978). Posterior probabilities for choosing a regres **65**, 39–48.

Atkinson, A. C. (1980). A note on the generalized information c model. *Biometrika* **67**, 413–18.

Atmar, W. (2001). A profoundly repeated pattern. *Bulletin of the America* **82**, 208–211.

Augustin, N.H., Mugglestone, M.A., and Buckland S.T. (1996). An the spatial distribution of wildlife. *Journal of Applied Ecology* **3**

Azzalini, A. (1996). *Statistical inference based on the likelihoo* London.

Barron, A., Rissanen, J., and Yu, B. (1998). The minimum descrip
coding and modeling. *IEEE Transactions on Information Theor*

Bartlett, M.S. (1936). Some notes on insecticide tests in the labo
Journal of the Royal Statistical Society, Supplement **1**, 185–194

Bedrick, E.J., and Tsai, C-L. (1994). Model selection for multivar
samples. *Biometrics* **50**, 226–231.

Berger, J.O., and Pericchi, L.R. (1996). The intrinsic Bayes factor
prediction. *American Statistical Association* **91**, 109–122.

Berger, J.O., and Wolpert, R.L. (1984). The likelihood principle. *I*
Statistics Monograph 6.

Berk, R.H. (1966). The limiting behavior of posterior distributi
incorrect. *Annals of Mathematical Statistics* **37**, 51–58.

Bernardo, J.M., and Smith, A.F.M. (1994). *Bayesian theory*.
Chichester, UK.

Berry, D.A. (1988). Multiple comparisons, multiple tests and dat
perspective. *Bayesian Statistics* 3. Clarendon, Oxford, UK.

Berryman, A.A., Gutierrez, A.P., and Arditi, R. (1995). Credible, p
predator–prey models—a reply to Abrams, Gleeson, and Sarne
1985.

Beyers, D.W. (1998). Causal inference in environmental impact
North American Benthological Society **17**, 367–373.

Bliss, C.I. (1935). The calculation of the dosage-mortality curve. *A*
22, 134–167.

Bickel, P., and Zhang, P. (1992). Variable selection in nonpara
categorical covariates. *Journal of the American Statistical Assoc*

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discret*
theory and practice. The MIT Press, Cambridge, MA.

Blansali, R.J. (1993). Order selection for linear time series models
in T. S. Rao (ed.) *Developments in time series analysis*, Chapma

Blau, G.E., and Neely, W.B. (1975). Mathematical model building
determine the distribution of DURSBAN® insecticide added to
Pages 133–163 *in* A. Macfadyen (ed.) *Advances in Ecological Res*
London.

Bollen, K.A., and Long, J.S. (1993). *Testing structural equations*.

Boltzmann, L. (1877). Über die Beziehung zwischen dem Hauptsa
nischen Wärmetheorie und der Wahrscheinlichkeitsrechnung res
das Wärmegleichgewicht. *Wiener Berichte* **76**, 373–435.

Box, G.E.P. (1967). Discrimination among mechanistic models. *Te*

Box, G.E.P. (1976). Science and statistics. *Journal of the Americar* **71**, 791–799.

Box, G.E.P., and Jenkins, G.M. (1970). *Time series analysis: fo* Holden-Day, London.

Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978). *Statistics for exp* and Sons, New York, NY, USA.

Box, G.E.P., Leonard, T., and Wu, C-F. (eds.) (1981). *Scientific in and robustness*. Academic Press, London.

Box, J.F. (1978). *R. A. Fisher: the life of a scientist*. John Wiley an

Boyce, M.S. (1992). Population viability analysis. *Annual Re Systematics* **23**, 481–506.

Bozdogan, H. (1987). Model selection and Akaike's information crit theory and its analytical extensions. *Psychometrika* **52**, 345–370

Bozdogan, H. (1988). A new model-selection criterion. Pages 599- *Classification and related methods of data analysis*. North-Hollar Amsterdam, The Neterlands.

Bozdogan, H. (1994). Editor's general preface. Pages ix–xii *in* H. Bo *ing and Scientific Applications*. Vol. 3, Proceedings of the First U the Frontiers of Statistical Modeling: An Informational Appro Publishers, Dordrecht, the Netherlands.

Breiman L. (1992). The little bootstrap and other methods for dim regression: *X*-fixed prediction error. *Journal of the American Sta* 738–754.

Breiman, L. (1995). Better subset regression using the nonnegative **37**, 373–384.

Breiman L. (1996). Heuristics of instability and stabilization in mod *of Statistics* **24**, 2350–2383.

Breiman, L. (2001). Statistical modeling: the two cultures (with *Science* **16**, 199–231.

Breiman, L., and Freedman, D.F. (1983). How many variables regression equation? *Journal of the American Statistical Associa*

Brisbin I.L., Jr., Collins, C.T., White, G.C., and McCallum, D.A. (for the analysis and interpretation of growth data: the shape of th 552–554.

Brockwell, P.J., and Davis, R.A. (1987). *Time series: theory and me* New York, NY.

Broda, E. (1983). *Ludwig Boltzmann: man, physicist, philosopher* (
Ox Bow Press, Woodbridge, Connecticut.

Brown, D. (1992). A graphical analysis of deviance. *Applied Statis*

Brown, D., and Rothery, P. (1993). *Models in biology: math*
computing. John Wiley and Sons. New York, NY.

Brown, L.D., Cai, T.T., and DasGupta, A. (2001). Interval esti
proportion. *Statistical Science* **16**, 101–135.

Brown, P.J. (1993). *Measurement, regression, and calibration*. Cl
UK.

Brownie, C., Anderson, D.R., Burnham, K.P., and Robson, D.S. (198
from band recovery data–a handbook. 2nd ed. U.S. Fish and Wi
Publication 156.

Brownie, C., Hines, J.E., Nichols, J.D., Pollock, K.H., and Hestbech
recapture studies for multiple strata including non-Markovian
Biometrics **49**, 1173–1187.

Brush, S.G. (1965). *Kinetic theory*. Vol. 1 Pergamon Press, Oxford

Brush, S.G. (1966). *Kinetic theory*. Vol. 2 Pergamon Press, Oxford

Bryant, P.G., and Cordero-Braña, O.I. (2000). Model selectio
description length principle. *The American Statistician* **54**, 257–

Buckland, S.T. (1982). A note on the Fourier series model for anal
Biometrics **38**, 469–477.

Buckland, S.T. (1984). Monte Carlo confidence intervals. *Biometri*

Buckland, S.T., Anderson, D.R., Burnham, K.P., and Laake, J.L. (19
estimating abundance of biological populations. Chapman and

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borch
L. (2001). *An introduction to distance sampling*. Oxford Univers

Buckland, S.T., Burnham, K.P., and Augustin, N.H. (1997). Mod
part of inference. *Biometrics* **53**, 603–618.

Buckland, S.T., and Elston, D.A. (1993). Empirical models for th
wildlife. *Journal of Applied Ecology* **30**, 478–495.

Burman, P. (1989). A comparative study of ordinary cross-va
validation and repeated learning-testing methods. *Biometrika* **76**

Burman, P., and Nolan, D. (1995). A general Akaike-type criterion
robust regression. *Biometrika* **82**, 877–886.

Burnham, K.P. (1988). A comment on maximum likelihood estima
of distributions. *Biometrical Journal* **30**, 379–384.

NY.

Burnham, K.P. (in review). Basic random effects models in ringing data. *Ecological and Environmental Statistics.*

Burnham, K.P., and Anderson, D.R. (1992). Data-based selection o ical model: the key to modern data analysis. Pages 16–30 *in* D.R. Barrett (eds.) *Wildlife* 2001: *Populations*. Elsevier Scientific Pub

Burnham, K.P., and Anderson, D.R. (2001). Kullback-Leibler inf strong inference in ecological studies. *Wildlife Research* **28**, 111

Burnham, K.P., Anderson, D.R., White, G.C., Brownie, C., and Pollo *and analysis methods for fish survival experiments based on relea* Fisheries Society, Monograph **5.**

Burnham, K.P., Anderson, D.R., and White, G.C. (1994). Evalu Leibler discrepancy for model selection in open population cap *Biometrical Journal* **36**, 299–315.

Burnham, K.P., White, G.C., and Anderson, D.R. (1995a). Model of capture–recapture data. *Biometrics* **51**, 888–898.

Burnham, K.P., Anderson, D.R., and White, G.C. (1995b). Selecti tion capture–recapture models when capture probabilities are he *Applied Statistics* **22**, 611–624.

Burnham, K.P., Anderson, D.R., and White, G.C. (1996). Meta-ana Northern Spotted Owl. *Studies in Avian Biology* **17**, 92–101.

Burnham, K.P., and White, G.C. (2002). Evaluation of some rando applicable to bird ringing data. *Journal of Applied Statistics* **29**,

Bystrak, D. (1981). The North American breeding bird survey, Ralph, and J.M. Scott (eds.) *Estimating numbers of terrestrial Biology* **6**.

Carlin, B., and Louis, T. (2000). *Bayes and empirical Bayes method* ed.) Chapman and Hall, London.

Carlin, B.P., and Chib, S. (1995). Bayesian model choice via Mar methods. *Journal of the Royal Statistical Society*, Series B **57**, 4

Carlin, B.P., and Louis, T.A. (1996). *Bayes and empirical Bayes me* Chapman and Hall, London.

Carpenter, S.R. (1990). Large-scale perturbations: opportunities **71**, 2038–2043.

Carrol, R.J., and Ruppert, D. (1988). *Transformation and weighting* and Hall, New York, NY.

Carrol, R., Ruppert, D., and Stefanski, L. (1995). *Measurement err* Chapman and Hall, London.

Caswell, H. (2001). *Matrix population models: construction, anal*
Sinauer Associates, Inc., Publishers, Sunderland, MA.

Cavanaugh, J.E., and Neath, A.A. (1999). Generalizing the der
information criterion. *Communication in Statistics—Theory and*

Cavanaugh, J.E., and Shumway, R.H. (1997). A bootstrap variant
model selection. *Statistica Sinica* **7**, 473–496.

Chamberlain, T.C. (1890). The method of multiple working hyp
(Reprinted 1965, Science **148**, 754–759.

Chamberlin, T. C. (1965). (1890) The method of multiple working h
754-759. (reprint of 1890 paper in *Science*).

Chatfield, C. (1991). Avoiding statistical pitfalls (with discussion
240–268.

Chatfield, C. (1995a). *Problem solving: a statistician's guide*. Se
and Hall, London.

Chatfield, C. (1995b). Model uncertainty, data mining and statistic
the Royal Statistical Society, Series A **158**, 419–466.

Chatfield, C. (1996). Model uncertainty and forecast accuracy. *Jou*
495–508.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J.G. (2000). *Monte Car*
computation. Springer, New York, NY.

Cherry, S. (1998). Statistical tests in publications of The Wildlife S
Bulletin **26**, 947–953.

Chib, S., and Jeliazkov, I. (2001). Marginal likelihood from the Me
Journal of the American Statistical Association **96**, 270–281.

Chow, G.C. (1981). A comparison of the information and posterio
model selection. *Journal of Econometrics* **16**, 21–33.

Chung, H-Y., Lee, K-W., and Koo, J-A. (1996). A note on boo
criterion. *Statistics and Probability Letters* **26**, 35–41.

Clayton, D., and Hills, M. (1993). *Statistical models in epidemiol*
Press, Oxford, UK.

Clayton, M.K., Geisser, S., and Jennings, D. (1986). A comparison
tion procedures. Pages 425–439 *in* P. Goel, and A. Zellner (eds.)
decision. Elsevier, New York, NY.

Cochran, W.G. (1963). *Sampling techniques*. 2nd ed., John Wiley a
NY.

Cohen, E.G.D., and Thirring, W. (eds.) (1973). *The Boltzmann*
applications. Springer-Verlag, New York, NY.

Collopy, F., Adya, M., and Armstrong, J.S. (1994). Principles for e
lidity: the case of information systems spending forecasts. *Inform*
5, 170–179.

Conner, M.M., McCarty, C.W., and Miller, M.W. (2000). Detection
estimates of chronic wasting disease prevalence in mule deer. *Jou*
36, 691–699.

Conner, M.M., White, G.C., and Freddy, D.J. (2001). Elk moveme
season hunting in northwest Colorado. *Journal of Wildlife Mana*

Cook, T.D., and Campbell, D.T. (1979). *Quasi-experimentation: de*
for field settings. Houghton Mifflin Company, Boston, MA.

Cook, R., Cook, J.G., Murray, D.L., Zager, P., Johnson, B.K., and
Development of predictive models of nutritional condition fo
Journal of Wildlife Management **65**, 973–987.

Copas, J.B. (1983). Regression, prediction and shrinkage (with dis
Royal Statistical Society, Series B, **45**, 311–354.

Cover, T.M., and Thomas, J.A. (1991). *Elements of information*
Sons, New York, NY.

Cox, D.R. (1990). Role of models in statistical analysis. *Statistical*

Cox, D.R. (1995). The relation between theory and application in sta
Test **4**, 207–261.

Cox, D.R., and Reid, N. (2000). *The theory of the design of expe*
Hall/CRC, Boca Raton, FL.

Cox, D.R., and Snell, E.J. (1989). *Analysis of binary data*. 2nd ed., C
York, NY.

Craven, P., and Wahba, G. (1979). Smoothing noisy data with spli
the correct degree of smoothing by the method of generalized cros
Mathematics **31**, 377–403.

Cressie, N.A.C. (1991). *Statistics for spatial data*. John Wiley and

Cutler, A., and Windham, M.P. (1994). Information-based validity
analysis. Pages 149–170 *in* H. Bozdogan (ed.) *Engineering and*
vol. 2, Proceedings of the First US/Japan Conference on the
Modeling: An Informational Approach. Kluwer Academic Pub
Netherlands.

Daniel, C., and Wood, F.S. (1971). *Fitting equations to data*. Wiley-
NY.

de Gooijer, J.G., Abraham, B., Gould, A., and Robinson, L. (1985). M
the order of an autoregressive-moving average process: a survey.
Review **53**, 301–329.

deLeeuw, J. (1992). Introduction to Akaike (1973) information the
the maximum likelihood principle. Pages 599–609 *in* S. Kotz, a
Breakthroughs in statistics. Vol. 1. Springer-Verlag, London.

Dempster, A.P. (1971). Model searching and estimation in the logic c
on the Foundations of Statistical Inference. University of Wate
Canada.

Dempster, A.P. (1997). The direct use of likelihood for significanc
Computing **7**, 247–252.

Desu, M.M., and Roghavarao, D. (1991). *Sample size methodolog*
New York, NY.

Dijkstra, T.K. (ed). (1988). *On model uncertainty and its statistic*
Notes in Economics and Mathematical Systems, Springer-Verla

Dijkstra, T.K., and Veldkamp, J.H. (1988). Data-driven selectio
bootstrap. Pages 17–38 *in* T.K. Dijkstra (ed.) *On model uncer*
implications. Lecture Notes in Economics and Mathematical Sy
New York, NY.

Draper, D. (1995). Assessment and propagation of model uncert
Journal of the Royal Statistical Society, Series B **57**, 45–97.

Draper, N.R., and Smith, H. (1981). *Applied regression analysis*. Sec
and Sons, New York, NY.

Eberhardt, L.L. (1978). Appraising variability in population stud
Management **42**, 207–238.

Eberhardt, L.L., and Thomas, J.M. (1991). Designing enviro
Ecological Monographs **61**, 53–73.

Edwards, A.W.F. (1976). *Likelihood: an account of the statistical c*
its application to scientific inference. Cambridge University Pre

Edwards, A.W.F. (1992). *Likelihood: expanded edition*. The Joh
Press, Baltimore, MD.

Edwards, A.W.F. (2001). Occam's bonus. *In* A. Zellner, H.A. Keuze
(eds.), *Simplicity, inference and modelling*, pgs. 128–132. Camb
Cambridge, UK.

Edwards, D. (1998). Issues and themes for natural resources trend
Ecological Applications **8**, 323–325.

Efron, B. (1979) Bootstrap methods: another look at the jackknif
7, 1–26.

Efron, B. (1983). Estimating the error rate of a prediction rule: i
validation. *Journal of the American Statistical Association* **78**, 3

Efron, B. (1986). How biased is the apparent error rate of a predict
American Statistical Association **81**, 461–470.

Efron, B., and Morris, C. (1975). Data analysis using Ste
generalizations. *Journal of the American Statistical Association*

Efron, R., and Gong, G. (1983). A leisurely look at the bootstrap, t
validation. *The American Statistician* **37**, 36–48.

Efron, B., and Tibshirani, R.J. (1993). *An introduction to the bootst*
London.

Ellison, A.M. (1996). An introduction of Bayesian inference for
environmental decision-making. *Ecological Applications* **6**, 103

Feder, M., Merhav, N., and Gutman, M. (1992). Universal p
sequences. *IEEE Transactions on Information Theory* **38**, 1258–

Fienberg, S.E. (1970). The analysis of multidimensional continge
419–433.

Fildes, R., and Makridakis, S. (1995). The impact of empirical a
series analysis and forecasting. *International Statistics Review* **6**

Findley, D.F. (1985). On the unbiasedness property of AIC for exact
stochastic time series models. *Journal of Time Series Analysis* **6**

Findley, D.F. (1991). Counterexamples to parsimony and BIC. *A*
Statistical Mathematics **43,** 505–514.

Findley, D.F., and Parzen, E. (1995). A conversation with Hirot
Science **10**, 104–117.

Finney, D.J. (1971). *Probit analysis*. 3rd. ed. Cambridge University

Fisher, R.A. (1922). On the mathematical foundations of theoretical
of London. *Philosophical Transactions* (Series A) **222**, 309–368

Fisher, R.A. (1936). Uncertain inference. *Proceedings of the Am*
and Sciences **71,** 245–58.

Fisher, R.A. (1949). A biological assay of tuberculins. *Biometrics*

Flack, V.F., and Chang, P.C. (1987). Frequency of selecting no
regression analysis: a simulation study. *The American Statistici*

Flather, C.H. (1992). Patterns of avian species-accumulation rates
landscapes. Ph.D. dissertation. Colorado State University. Fort C

Flather, C.H. (1996). Fitting species-accumulation functions and
use impacts on avian diversity. *Journal of Biogeography* **23**, 155

Ford, E.D. (2000). *Scientific method for ecological research*. Cam
Cambridge, UK.

Forster, M.R. (2000). Key concepts in model selection: performan
Journal of Mathematical Psychology **44**, 205–231.

Forster, M.R. (2001). The new science of simplicity. *In* A. Zell
and M. McAleer (eds.), *Simplicity, inference and modelling*, p
University Press, Cambridge, UK.

Forster, M.R., and Sober, E. (1994). How to tell when simpler, m
hoc theories will provide more accurate predictions. *British Jour
Science* **45,** 399–424.

Franklin, A.B., Anderson, D.R., and Burnham, K.P. (2002). Estima
and variation in avian survival probabilities using random effe
Applied Statistics **29**, 267–287.

Franklin, A.B., Shenk, T.M., Anderson, D.R., and Burnham, K.P. (
selection: the alternative to null hypothesis testing. Pages 75–90
Franklin (eds.) *Modeling in Natural Resource Management*. Isl
D. C.

Freedman, D.A. (1983). A note on screening regression equations. *Ti*
37, 152–155.

Freedman, D. (1999). From association to causation: some rem
statistics. *Statistical Science* **14**, 243–258.

Freedman, D.A., Navidi, W., and Peters, S.C. (1988). On the impa
in fitting regression equations. Pages 1–16 *in* T.K. Dijkstra (ed.
and its statistical implications. Lecture Notes in Economics and
Springer-Verlag, New York, NY.

Fujikoshi, Y, and Satoh, K. (1997). Modified *AIC* and C_p in multiv
Biometrika **84**, 707–716.

Gail, M.H. (1996). Statistics in action. *Journal of the American St
1–13.

Gallant, A.R. (1987). *Nonlinear statistical models*. John Wiley and

Gammerman, D. (1997). *Markov Chain Monte Carlo*. Chapman a

Garthwaite, P.H., Jolliffe, I.T., and Jones, B. (1995). *Statistical i
London.

Gause, G.F. (1934). *The struggle for existence*. Williams and Wilk

Geisser, S. (1975). The predictive sample reuse method with appl
American Statistical Association **70**, 320–328.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-based app
marginal densities. *Journal of the American Statistical Associat

Gelfand, A., and Dey, D.K. (1994). Bayesian model choice:
calculations. *Journal of the Royal Statistical Society*, Series B, **5**

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *B*
Chapman and Hall, London.

George, E.I., and Foster, D.P. (2000). Calibration and empirical B
Biometrika **87**, 731–748.

George, E.I., and McCulloch, R.E. (1993). Variable selection via C
of the American Statistical Association **88**, 881–889.

Gerard, P.D., Smith, D.R., and Weerakkody, G. (1998). Limits
analysis. *Journal of Wildlife Management* **62**, 801–807.

Gilchrist, W. (1984). *Statistical modelling*. Chichester, Wiley and S

Gilks, W. R., Richardson, S., and Spiegelhalter, D.J. (1996). *Marko*
practice. Chapman and Hall, London.

Glymour, C. (1998). Causation. Pages 97–109 *in* S. Kotz (ed.), *En*
sceinces. John Wiley and Sons, New York, NY.

Gochfeld, M. (1987). On paradigm vs. methods in the study of gro

Gokhale, D.V., and Kullback, S. (1978). *The information in con*
Dekker, New York, NY.

Golub, G.H., Health, M., and Wahba, G. (1979). Generalized cross
for choosing a good ridge parameter. *Technometrics* **21**, 215–22

Goodman, S.N. (1993). *p* values, hypothesis tests, and likelihoo
demiology of a neglected historical debate (with discussion)
Epidemiology **137**, 485–501.

Goodman, S.N., and Berlin, J.A. (1994). The use of predicted co
planning experiments and the misuse of power when interpre
Internal Medicine **121**, 200–206.

Goutis, C., and Casella, G. (1995). Frequentist post-data inference.
Review **63**, 325–344.

Granger, C.W.J., King, M.L., and White, H. (1995). Comments on te
and the use of model selection criteria. *Journal of Econometrics*

Graybill, F.A., and Iyer, H.K. (1994). *Regression analysis: con*
Duxbury Press, Belmont, CA.

Greenhouse, S.W. (1994). Solomon Kullback: 1907–1994. *Ins*
Statistics Bulletin **23**, 640–642.

Guiasu, S. (1977). *Information theory with applications*. McGraw-

Guisan, A., and Zimmermann, N.E. (2000). Predictive habitat
ecology. *Ecological Modelling* **135**, 147–186.

Ha, I.D., Lee, Y., and Song, J-K. (2001). Hierarchical likelihood app
Biometrika **88**, 233–243.

Hald, A. (1952). *Statistical theory with engineering applications*. New York, NY.

Hald, A. (1998). *A history of mathematical statistics*. John Wiley,

Hand, D.J. (1994). Statistical strategy: step 1. Pages 1–9 *in* P. Chees (eds.) *Selecting models from data*. Springer-Verlag, New York,

Hand, D.J. (1998). Data mining: statistics and more? *The Americ* 118.

Hand, D.J., Blunt, G., Kelly, M.G., and Adams, N.M. (2000). Data n *Statistical Science* **15**, 111–131.

Hannan, E.J., and Quinn, B.G. (1979). The determination of the orc *Journal of the Royal Statistical Society*, Series B **41**, 190–195.

Hansen, M.H., and Yu, B. (2001). Model selection and the principle length. *Journal of the American Statistical Association* **96**, 746–

Harlow, L.L., Mulaik, S.A., and Steiger, J.H. (eds.) (1997). *What if th tests?* Lawrence Erlbaum Associates, Publishers, Mahwah, NJ.

Hasenöhrl, F. (ed.) (1909). *Wissenschaftliche Abhandlungen*. 3 Vol

Hastie, T.J., and Tibshirani, R.J. (1990). *Generalized additive mod* London.

Haughton, D. (1989). Size of the error in the choice of a model to fit c family. *Sankhya*, Series A **51**, 45–58.

Hayne, D. (1978). Experimental designs and statistical analyses. Pa (ed). *Populations of small mammals under natural conditions*. P in Ecology, University of Pittsburgh, Vol 5.

Henderson, H., and Velleman, P. (1981). Building multiple regressi *Biometrics* **37**, 391–411.

Heyde, C.C. (1997). *Quasi-likelihood and its application: a gene parameter estimation*. Springer-Verlag, New York, NY.

Hilborn, R., and Mangel, M. (1997). *The ecological detective: c data*. Princeton University Press, Princeton, NJ.

Hjorth, J.S.U. (1994). *Computer intensive statistical methods: val and bootstrap*. Chapman and Hall, London.

Hobson, A., and Cheng, B-K. (1973). A comparison of the information measures. *Journal of Statistical Physics* **7**, 301–310

Hocking, R.R. (1976). The analysis and selection of variable *Biometrics* **32**, 1–49.

Hoenig, J.M., and Heisey, D.M. (2001). The abuse of power: the pe calculations for data analysis. *The American Statistician* **55**, 19-

Hoeting, J.A., Madigan, D., Raftery, A.E., and Volinsky, C.T. (averaging: a tutorial (with discussion). *Statistical Science* **14**, 38

Hosmer, D.W., and Lemeshow, S. (1989). *Applied logistic regressi* and Sons, New York, NY.

Howard, R.A. (1971). *Dynamic probabilistic systems*. John Wiley a

Hurvich, C.M., Simonoff, J.S., and Tsai, C-L. (1998). Smoothing nonparametric regression using an improved Akaike information *Royal Statistical Society*, Series B, **60**, 271–293.

Hurvich, C.M., and Tsai, C-L. (1989). Regression and time series m samples. *Biometrika* **76**, 297–307.

Hurvich, C.M., and Tsai, C-L. (1990a). Model selection for le regression in small samples. *Statistics and Probability Letters* **9**,

Hurvich, C.M., and Tsai, C-L. (1990b). The impact of model selecti regression. *The American Statistician* **44**, 214–217.

Hurvich, C.M., and Tsai, C-L. (1991). Bias of the corrected AIC regression and time series models. *Biometrika* **78**, 499–509.

Hurvich, C.M., and Tsai, C-L. (1994). Autoregressive model sele using a bias-corrected version of AIC. Pages 137–157 *in* H. Bozd *and Scientific Applications*. Vol. 1, Proceedings of the First U the Frontiers of Statistical Modeling: An Informational Appro Publishers, Dordrecht, the Netherlands.

Hurvich, C.M., and Tsai, C-L. (1995a). Relative rates of converge selection criteria in linear regression. *Biometrika* **82**, 418–425.

Hurvich, C.M., and Tsai, C-L. (1995b). Model selection for ext models in small samples. *Biometrics* **51**, 1077–1084.

Hurvich, C.M., and Tsai, C-L. (1996). The impact of unsuspecte model selection in linear regression. *Statistics and Probability L*

Hurvich, C.M., Shumway, R., and Tsai, C-L. (1990). Improved e Leibler information for autoregressive model selection in small s 709–719.

Ibrahim, J.G., and Chen, M-H. (1997). Predictive variable select linear model. *Biometrics* **53**, 465–478.

Inman, H.F. (1994). Karl Pearson and R. A. Fisher on statistical tests *Nature*. *The American Statistician* **48**, 2–11.

Irizarry, R.A. (2001). Information and posterior probability criteria local likelihood estimation. *Journal of the American Statistical A*

Jaffe, A.J., and Spirer, H.F. (1987). *Misused statistics: straight ta*
Marcel Dekker, Inc., New York, NY.

James, F.C., and McCulloch, C.E. (1990). Multivariate analysis in e
panacea or Pandora's box? *Annual Reviews of Ecology and Syst*

Jaynes, E.T. (1957). Information theory and statistical mechanic
620–630.

Jaynes, E.T. (1982). On the rationale of maximum-entropy meth
IEEE **70**, 939–952.

Jaynes, E.T. (in prep.). *Probability theory: the logic of science*. Cam
Cambridge, UK.

Jeffreys, H. (1948). *Theory of probability*. Oxford University Press

Jeffreys, H. (1973). *Scientific inference*. 3rd ed. Cambridge Unive
UK.

Jessop, A. (1995). *Informed assessments: an introduction to in*
statistics. Ellis Horwood, London.

Jevons, W. S. (1874, 1877). The principles of science. MacM
Kingdom.

Jiménez, J.A., Hughes, K.A., Alaks, G., Graham, L., and Lacy, R.C. (
study of inbreeding depression in a natural habitat. *Science* **266**,

Johnson, D.H. (1995). Statistical sirens: the allure of nonparametr
2000.

Johnson, D.H. (1999). The insignificance of statistical significa
Wildlife Management **63**, 763–772.

Johnson, J.W. (1996). Fitting percentage of body fat to simple body
of Statistics Education **4** (e-journal).

Johnson, N.L., and Kotz, S. (1970). *Continuous univariate distribu*
Company, New York, NY.

Johnson, N.L., and Kotz, S. (1992). *Univariate discrete distribu*
Interscience Publication, New York, NY.

Jones, D., and Matloff, N. (1986). Statistical hypothesis testing in
in terms. *Journal of Economic Entomology* **79**, 1156–1160.

Judge, G.C., and Yancey, T. (1986). *Improved methods of inference*
Holland, Amsterdam, the Netherlands.

Kabaila, P. (1995). The effect of model selection on confidence
regions. *Econometric Theory* **11**, 537–549.

Kapur, J.N., and Kesavan, H.K. (1992). *Entropy optimization princ*
Academic Press, London.

Kass, R.E., and Raftery, A.E. (1995). Bayes factors. *Journal of t*
Association **90**, 773–795.

Kass, R.E., and Raftery, A.E. (1995). "Bayes factors." *Journal of t*
Association **90**, 773–795.

Kass, R.E., and Wasserman, L. (1995). A reference Bayesian test fo
its relationship to the Schwarz criterion. *Journal of the Americar*
90, 928–934.

Kishino, H., Kato, H., Kasamatsu, F., and Fujise, Y. (1991). Det
and estimation of population characteristics from field survey
feasibility study of the Southern Hemisphere minke whales. *A*
Statistical Mathematics **43**, 435–453.

Kiso, K., Akamine, T., Ohnishi, S., and Matsumiya, Y. (1992). Mat
of the growth of sea-run and fluviatile forms of the female masu
masou in rivers of the southern Sanriku district, Honshu, Japan. *N*
58, 1779–1784.

Kittrell, J.R. (1970). Mathematical modelling of chemical reactors
Engineering **8**, 97–183.

Knopf, F.L., Sedgwick, J.A., and Cannon, R.W. (1988). Guild structu
relative to seasonal cattle grazing. *Journal of Wildlife Managem*

Konishi, S., and Kitagawa, G. (1996). Generalized information crit
Biometrika **83**, 875–890.

Kooperberg, C., Bose, S., and Stone, C.J. (1997). Polychotomous re
American Statistical Association **92**, 117–127.

Kreft, I., and deLeeuw, J. (1998). *Introducing multilevel model*
Thousand Oaks, CA, USA.

Kuhn, T.S. (1970). *The structure of scientific revolutions*. 2nd ed.
Press, Chicago, IL

Kullback, S. (1959). *Information theory and statistics*. John Wile
NY.

Kullback, S. (1987). The Kullback-Leibler distance. *The American S*

Kullback, S., and Leibler, R.A. (1951). On information and
Mathematical Statistics **22**, 79–86.

Laake, J.L., Buckland, S.T., Anderson, D.R., and Burnham, K.P. (1
guide. Version 2.1. Colorado Cooperative Fish and Wildlife R
State University, Fort Collins, CO.

Lahiri, P. (ed.) (2001). Model selection. Institute of Mathematical S
No. 38.

Larimore, W.E., and Mehra, R.K. (1985). The problem of overfitting

Laud, P.W., and Ibrahim, J.G. (1995). Predictive model selection *Statistical Society*, Series B **57**, 247–262.

Laud, P.W., and Ibrahim, J.G. (1996). Predictive specification of p in variable selection. *Biometrika* **83**, 267–274.

Leamer, E.E. (1978). *Specification searches: ad hoc inference with* John Wiley and Sons, New York, NY.

Lebreton, J-D., Burnham, K.P., Clobert, J., and Anderson, D.R. (19 and testing biological hypotheses using marked animals: a unif studies. *Ecological Monograph* **62**, 67–118.

Lee, Y., and Nelder, J.A. (1996). Hierarchical generalized linear *Royal Statistical Society*, Series B **58**, 619–678.

Lehmann, E.L. (1983). *Theory of point estimation*. John Wiley and

Lehmann, E.L. (1990). Model specification: the views of Fisher developments. *Statistical Science* **5**, 160–168.

Leirs, H., Stenseth, N.C., Nichols, J.D., Hines, J.E., Verhagen, R., an Stochastic seasonality and nonlinear density-dependent factors r in an African rodent. *Nature* **389**, 176–180.

Leonard, T., and Hsu, J.S.J. (1999). *Bayesian methods: an analy interdisciplinary researchers*. Cambridge University Press, Cam

Leroux, B.G. (1992). Consistent estimation of a mixing distribution. **20**, 1350–1360.

Lewis, P. (1998). Maximum likelihood as an alternative to parsimony using nucleotide sequence data. Pages 132–163 *in* D. Soltis, P. So *Molecular systematics of plants II*. Kluwer Publishing, Boston,

Liang, K-Y, and McCullagh, P. (1993). Case studies in binary dis 623–630.

Lindley, D.V. (1986). The relationship between the number of facto ment. Pages 459–470 *in* P.K. Goel, and A. Zellner (eds.) *Bayesian techniques*. Elsevier Science Publishers, New York, NY.

Lindsey, J.K. (1995). *Modeling frequency and count data*. Oxford U UK.

Lindsey, J.K. (1996). *Parametric statistical inference*. Oxford Scien Oxford, UK.

Lindsey, J.K. (1999*a*). On the use of corrections for overdispersion 553–561.

Lindsey, J.K. (1999*b*). Some statistical heresies. *The Statistician* **4**

Lindsey, J.K., and Jones, B. (1998). Choosing among generalized l
medical data. *Statistics in Medicine* **17**, 59–68.

Linhart, H. (1988). A test whether two AIC's differ significantly. *S*
Journal **22**, 153–161.

Linhart, H., and Zucchini, W. (1986). *Model selection*. John Wile
NY.

Longford, N.T. (1993). *Random coefficient models*. Oxford Univ
York, NY.

Longford, N.T. and Nelder, J.A. (1999). Statistics versus statistical s
process. *Statistics in Medicine* **18**, 2311–2320.

Lucky, R.W. (1991). *Silicon dreams: information, man, and mach*
New York, NY.

Ludwig, D. (1989). Small models are beautiful: efficient estimators
Pages 274–284 *in* C. Castillo-Chavez, S.A. Levin, and C.A. Shoer
cal approaches to problems in resource management and epidem
London.

Lunneborg, C.E. (1994). *Modeling experimental and observationa*
Belmont, CA.

Lytle, D.A. (2002). Flash floods and aquatic insect life-history e
multiple models. *Ecology* **83**, 370–385.

Madigan, D., and Raftery, A.E. (1994). Model selection and accou
tainty in graphical models using Occam's window. *Journal of t*
Association **89**, 1535–1546.

Madigan, D.M., Raftery, A.E., York, J.C., Bradshaw, J.M., and Almo
gies for graphical model selection. Pages 91–100 *in* P. Chessen
(eds.) *Selecting models from data: AI and statistics IV*. Springe
in Statistics **89**.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics* **12**,

Mallows, C.L. (1995). More comments on C_p. *Technometrics* **37**,

Manly, B.F.J. (1991). *Randomization and Monte Carlo methods in*
Hall, New York, NY.

Manly, B.F.J., McDonald, L.L., and Thomas, D.L. (1993). *Resourc*
statistical design and analysis for field studies. Chapman and Ha

Manly, B.F.J. (1992). *The design and analysis of research studies*
Press, Cambridge, UK.

Manly, B.F.J. (2001). *Statistics for environmental science and man*
Hall, London.

Marshall, J.R. (1990). Data dredging and noteworthiness. *Epidemi*

MATLAB® (1994) *High-performance numerical computations and*
The MathWorks, Inc., Natick, MA.

Maurer, B.A. (1998). Ecological science and statistical paradigms: a
279, 502–503.

Mayr, E. (1997). *This is biology: the science of the living world*
Harvard University Press. Cambridge, MA.

McBride, G.B., Loftis, J.C., and Adkins, N.C. (1993). What do sig
us about the environment? *Environmental Management* **17**, 423-

McCullagh, P., and Nelder, J.A. (1989). *Generalized linear models*
Hall, New York, NY.

McCullagh, P., and Pregibon, D. (1985). Discussion comments on
and Efron. *Annals of Statistics* **13**, 898–900.

McLachlan, G.J., and Peel, D. (2000). *Finite Mixture Models*. Joh
New York, NY.

McQuarrie, A.D. (1999). A small-sample correction of the Schwa
criterion. *Statistics and Probability Letters* **44**, 79–86.

McQuarrie, A.D.R., and Tsai, C-L. (1998). *Regression and time*
World Scientific Publishing Company, Singapore.

Mead, R. (1988). *The design of experiments: statistical principles fo*
Cambridge University Press, New York, NY.

Miller, A.J. (1990). *Subset selection in regression*. Chapman and H

Mooney, C.Z., and Duval, R.D. (1993). *Bootstrapping: a nonp*
statistical inference. Sage Publications, London.

Moore, D.F. (1987). Modelling the extraneous variance in the pre
variation. *Journal of the Royal Statistical Society* **36**, 8–14.

Morgan, B.J.T. (1992). *Analysis of quantal response data*. Chapma

Morgan, B.J.T. (2000). *Applied stochastic modelling*. Arnold Press

Morris, C.N. (1983). Parametric empirical Bayes inference: theory a
of the American Statistical Association **78**, 47–65.

Mosteller, F., and Tukey, J.W. (1968). Data analysis, including stati
E. Aronson (eds.) *Handbook of Social Psychology*, Vol. 2. Ad
MA.

Myers, R.A., Barrowman N.J., Hutchings, J.A., and Rosenberg, A
dynamics of exploited fish stocks at low populations levels. *Scie*

Naik, P.A., and Tsai, C-L. (2001). Single-index model selections. *B*

Nester, M. (1996). An applied statistician's creed. *Applied Statistic*

Nichols, J.D., and Kendall, W.L. (1995). The use of multi-strata ca
to address questions in evolutionary ecology. *Journal of Applied*

Nishii, R. (1988). Maximum likelihood principle and model selecti
is unspecified. *Journal of Multivariate Analysis* **27**, 392–403.

Noda, K., Miyaoka, E., and Itoh, M. (1996). On bias correction of
criterion in linear models. *Communications in Statistics—Theory*
1857.

Norris, J.L., and Pollock, K.H. (1995). A capture–recapture model
behavioural response. *Environmental and Ecological Statistics* **2**

Norris, J.L., and Pollock, K.H. (1997). Including model uncertainty
in multiple capture studies. *Environmental and Ecological Statis*

O'Connor, M.P., and Spotila, J.R. (1992). Consider a spherical lizar
approximations. *American Zoologist* **32**, 179–193.

O'Connor, R.J. (2000). Why ecology lags behind biology. *The Scie*

O'Hagan, A. (1995). Fractional Bayes factors for model compar
Journal of the Royal Statistical Society, Series B **57**, 99–138.

Olden, J.D., and Jackson, D.A. (2000). Torturing data for the sake
are our regression models? *Ecoscience* **7**, 501–510.

Otis, D.L., Burnham, K.P., White, G.C., and Anderson, D.R. (197
from capture data on closed animal populations. *Wildlife Monog*

Pan, W. (1999). Bootstrapping likelihood for model selection with
of Computational and Graphical Statistics **8**, 687–698.

Pan, W. (2001*a*). Akaike's information criterion in generalized
Biometrics **57**, 120–125.

Pan, W. (2001*b*). Model selection in estimating equations. *Biometr*

Parzen, E. (1994). Hirotugu Akaike, statistical scientist. Pages 25–
Engineering and Scientific Applications. Vol. 1, Proceedings of t
ference on the Frontiers of Statistical Modeling: An Informatio
Academic Publishers, Dordrecht, the Netherlands.

Parzen, E., Tanabe, K., and Kitagawa, G. (eds.) (1998). *Selected pap*
Springer-Verlag Inc., New York, NY.

Pascual, M.A., and Kareiva, P. (1996). Predicting the outcome
experimental data: maximum likelihood and Bayesian approache

Peirce, C.S. (1955). Abduction and induction. Pages 150–15
Philosophical writings of Peirce. Dover, New York, NY.

Peters, R.H. (1991). *A critique for ecology*. Cambridge University
USA.

Peterson, T.S. (1960). *Elements of calculus*(2nd ed.). Harper Broth

Platt, J.R. (1964). Strong inference. *Science* **146**, 347–353.

Pollock, K.H., Nichols, J.D., Brownie, C., and Hines, J.E. (1990).
capture–recapture experiments. *Wildlife Monographs*. **107**, 1–97

Pope, S.E., Fahrig, L., and Merriam, H.G. (2000). Landscape
metapopulation effects on leopard frog populations. *Ecology* **81**

Posada, D., and Crandall, K. (1998). MODELTEST: testing the mod
Bioinformatics **14**, 817–818.

Posada, D., and Crandall, K. (2001). Selecting models of nucleotid
cation of human immunodeficiency virus 1 (HIV-1). *Molecular*
18, 897–906.

Poskitt, D.S., and Tremayne A.R. (1987). Determining a portfol
models. *Biometrika* **74**, 125–137.

Pötscher, B.M. (1991). Effects of model selection on inference.
163–185.

Qian, G., Gabor, G., and Gupta, R.P. (1996). Generalized linear
predictive least quasi-deviance criterion. *Biometrika* **83**, 41–54.

Qin, J., and Lawless, G. (1994). Empirical likelihood and genera
Annals of Statistics **22**, 300–325.

Quinn, J.F., and Dunham, A.E. (1983). On hypothesis testing in
American Naturalist **122**, 22–37.

Raftery, A.E. (1995). Bayesian model selection in social resea
Sociological Methodology **25**, 111–195.

Raftery, A. (1996a). Approximate Bayes factors and accounting f
generalized linear regression models. *Biometrika* **83**, 251–266.

Raftery, A. (1996b). Hypothesis testing and model selection. Pages
S. Richardson, and D.J. Spiegelhalter (eds.), *Markov chain M*
Chapman and Hall, London.

Raftery, A., Madigan, D.M., and Hoeting, J. (1993). *Model selec*
model uncertainty in linear regression models. Technical Repo
of Statistics, University of Washington, Seattle, WA.

Raftery, A.E., Madigan, D., and Hoeting, J.A. (1997). Bayesian mo
regression models. *Journal of the American Statistical Associati*

Rawlings, J.O. (1988). *Applied regression analysis: a research*
Belmont, CA.

Renshaw, E. (1991). *Modelling biological populations in space* University Press, Cambridge, UK.

Reschenhofer, E. (1996). Prediction with vague prior knowledg *Statistics—Theory and Methods* **25**, 601–608.

Reschenhoffer, E. (1999). Improved estimation of the expe discrepancy in case of misspecification. *Econometric Theory* **15**

Rexstad, E. (2001). Back cover of T.M. Shenk and A.B. Franklin, *in natural resource management*. Island Press, Washington, D.C

Rexstad, E.A., Miller, D.D., Flather, C.H., Anderson, E.M., Hupp, J. (1988). Questionable multivariate statistical inference in wildlife studies. *Journal of Wildlife Management* **52**, 794–798.

Rexstad, E.A., Miller, D.D., Flather, C.H., Anderson, E.M., Hupp, J. (1990). Questionable multivariate statistical inference in wildlife studies: a reply. *Journal of Wildlife Management* **54**, 189–193.

Ripley, B.D. (1996). *Pattern recognition and neural networks*. Cam Cambridge, UK.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Wo Computer Science, Vol 15. Singapore.

Rissanen, J. (1996). Fisher information and stochastic complexity *Information Theory* **42**, 40–47.

Robert, C.P., and Casella, G. (1999). Monte Carlo statistical method York, NY.

Roecker, E.B. (1991). Prediction error and its estimation for s *Technometrics* **33**, 459–468.

Ronchetti, E., and Staudte, R.G. (1994). A robust version of Mall *American Statistical Association* **89**, 550–559.

Rosenblum, E.P. (1994). A simulation study of information theore sical hypothesis tests in one factor ANOVA. Pages 319–346 *Engineering and Scientific Applications*. Vol. 2, Proceedings of t ference on the Frontiers of Statistical Modeling: An Informati Academic Publishers, Dordrecht, the Netherlands.

Roughgarden, J. (1979). *Theory of population genetics and ev introduction*. Macmillan Publishing Company, New York, NY.

Royall, R.M. (1997). *Statistical evidence: a likelihood paradig* London.

Royle, J.A., and Link, W.A. (2002). Random effects and shrinkage recapture methods. *Journal of Applied Statistics* **29**, 329–351.

Sakamoto, Y. (1991). *Categorical data analysis by AIC*. KTK Scien
Japan.

Sakamoto, Y., and Akaike, H. (1978). Analysis of cross classified
the Institute of Statistical Mathematics Part B **30**, 185–197.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). *Akaike informa*
KTK Scientific Publishers, Tokyo, Japan.

Santer, T.J., and Duffy, D.E. (1989). *The statistical analysis of d*
Verlag, New York, NY.

SAS Institute Inc. (1985). SAS® language guide for personal compu
SAS Institute Inc., Cary, NC.

SAS Institute. (1988). SAS/STAT®user's guide. Edition 6.03. SAS

Sauerbrei, W., and Schumacher, M. (1992). A bootstrap resamplin
building: application to the Cox regression model. *Statistics in M*

Sawa, T. (1978). Information criteria for discriminating among
models. *Econometrica* **46**, 1273–1291.

Scheiner, S.M., and Gurevitch, J. (eds.) (1993). *Design and*
experiments. Chapman and Hall, London.

Schmidt, B.R., and Anholt, B.R. (1999). Analysis of survival
common toads. *Amphibia-Reptilia* **20**, 97–108.

Schoener, T.W. (1970). Nonsynchronous spatial overlap of liza
Ecology **51**, 408–418.

Schreuder, H.T., Gregoire, T.G., and Wood, G.B. (1993). *S*
multiresource forest inventory. John Wiley and Sons, New York,

Schwarz, G. (1978). Estimating the dimension of a model. *Annals o*

Sclove, S.L. (1987). Application of some model-selection criteri
multivariate analysis. *Psychometrika* **52**, 333–343.

Sclove, S.L. (1994a). Small-sample and large-sample statistical m
Pages 31–39 *in* P. Cheeseman, and R.W. Oldford (eds.) *Selec*
Springer-Verlag, New York, NY.

Sclove, S.L. (1994b). Some aspects of model-selection criteria. Page
(ed.) *Engineering and Scientific Applications*. Vol. 2. Proceeding
Conference on the Frontiers of Statistical Modeling: An Informati
Academic Publishers, Dordrecht, the Netherlands.

Seber, G.A.F. (1977). *Linear regression analysis*. John Wiley and S

Seber, G.A.F. (1984). *Multivariate observations*. John Wiley and S

Seber, G.A.F., and Wild, C.J. (1989). *Nonlinear regression*. John
York, NY.

Severini, T.A. (2000). *Likelihood methods in statistics*. Oxford U
UK.

Shannon, C.E. (1948). A mathematical theory of communication
Journal **27**, 379–423 and 623–656.

Shao, J. (1993). Linear model selection by cross-validation. *Journa
Association* **88**, 486–494.

Shao, J. (1996). Bootstrap model selection. *Journal of the Americal*
91, 655–665.

Shao, J. (1997). An asymptotic theory for linear model selection.
264.

Shao, J, and Tu, D. (1995). *The jackknife and bootstrap*. Springer-

Shefferson, R.P., Sandercock, B.K., Proper, J., and Beissinger, S
dormancy and survival of a rare herbaceous perennial using
Ecology **82**, 145–156.

Shenk, T.M., and Franklin, A.B., (eds.) (2001). *Modeling in natural*
Island Press, Washington, D. C.

Shi, P., and Tsai, C-L. (1998). A note on the unification of the Akaik
Journal of the Royal Statistical Society, Series B, **60**, 551–558.

Shi, P., and Tsai, C-L. (1999). Semiparametric regression and mod
Statistical Planning and Inference **77**, 341–349.

Shibata, R. (1976). Selection of the order of an autoregressiv
information criterion. *Biometrika* **63**, 117–26.

Shibata, R. (1980). Asymptotically efficient selection of the order of
parameters of a linear process. *Annals of Statistics* **8**, 147–164.

Shibata, R. (1981). An optimal selection of regression variables.
Correction (1982). **69**, 492.

Shibata, R. (1983). A theoretical view of the use of AIC. Pages 237
(ed.) *Time series analysis: theory and practice*. Elsevier Scien
Holland, the Netherlands.

Shibata, R. (1986). Consistency of model selection and parameter
141 *in* J. Gani, and M.B. Priestly (eds.) *Essays in time series and a
of Applied Probability*, Special Volume 23A.

Shibata, R. (1989). Statistical aspects of model selection. Pages 2
(ed.) *From data to model*. Springer-Verlag, London.

Shibata, R. (1997a). Bootstrap estimate of Kullback–Leibler
selection. *Statistica Sinica* **7**, 375–394.

Shimizu, R. (1978). Entropy maximization principle and selectio
toregressive Gaussian process. *Annals of the Institute of Stati*
263–270.

Shono, H. (2000). Efficiency of the finite correction of Akaike
Fisheries Science **66**, 608–610.

Silverman, B.W. (1982). Algorithm AS 176: kernel density estimati
transform. *Applied Statistics* **31**, 93–99.

Silvey, S.D. (1975). *Statistical inference*. Chapman and Hall, Lond

Simonoff, J.S., and Tsai, C-L. (1999). Semiparametric and additiv
an improved AIC criterion. *Journal of Computational and Grapl*

Skalski, J.R., Hoffman, A., and Smith, S.G. (1993). Testing the sig
and cohort-level covariates in animal survival studies. Pages 9–2
P.M. North (eds.) *Marked individuals in the study of bird popula*
Basel, Switzerland.

Skalski, J.R., and Robson, D.S. (1992). *Techniques for wildlife inv*
analysis of capture data. Academic Press, New York, NY.

Smith, G.N. (1966). Basic studies on DURSBAN® insecticide. *Do*

Smith, S.C., Skalski, J.R., Schlechte, J.W., Hoffman, A., and Cass
Statistical Survival Analysis of Fish and Wildlife Tagging Studies
Sciences, University of Washington, Seattle, WA.

Sober, E. (1999). Instrumentalism revisited. *Crítica* **31**, 3–39.

Sober, E. (2001). Instrumentalism, parsimony and the Akaike fra
the Philosophy of Science Association (in press).

Sommer, S., and Huggins, R.M. (1996). Variables selection using th
C_p. *Applied Statistics* **45**, 15–29.

Soofi, E.S. (1994). Capturing the intangible concept of information.
Statistical Association **89**, 1243–1254.

Southwell, C. (1994). Evaluation of walked line transect counts fo
density. *Journal of Wildlife Management* **58**, 348–356.

Speed, T.P., and Yu, B. (1993). Model selection and prediction: no
of the Institute of Statistical Mathematics **1**, 35–54.

Spiegelhalter, D.J., Best, N.G., and Carlin, B.P., and van der Linc
measures of model complexity and fit. *Journal of the Royal Stat*
64, 1–34.

Sprott, D.A. (2000). *Statistical inference in science*. Springer Serie

Steel, M., and Penny, D. (2001). Parsimony, likelihood and the role
phylogenetics. *Molecular Biology and Evolution* **17**, 839–850.

Sterling, T.D., Rosenbaum, W.L., and Weinkam, J.J. (1995). Publica
the effect of the outcome of statistical tests on the decision to pub
American Statistician **49**, 108–112.

Stewart-Oaten, A. (1995). Rules and judgments in statistics: three
2001–2009.

Stigler, S.M. (1986), *The history of statistics*. Harvard University P

Stoica, P., Eykhoff, P., Janssen, P, and Söderström, T. (1986). Mod
cross-validation. *International Journal of Control* **43**, 1841–187

Stone, M. (1974). Cross-validatory choice and assessment of stati
discussion). *Journal of the Royal Statistical Society*, Series B **39**

Stone, M. (1977). An asymptotic equivalence of choice of model
Akaike's criterion. *Journal of the Royal Statistical Society*, Serie

Stone, C.J. (1982). Local asymptotic admissibility of a generaliza
selection rule. *Annals of the Institute of Statistical Mathematics*

Stone, M., and Brooks, R.J. (1990). Continuum regression: cross
constructed prediction embracing ordinary least squares, partial
cipal components regression (with discussion). *Journal of the R*
Series B **52**, 237–269.

Stromborg, K.L., Grue, C.E., Nichols, J.D., Hepp, G.R., Hines,
(1988). Postfledging survival of European starlings exposed to
insecticide. *Ecology* **69**, 590–601.

Sugiura, N. (1978). Further analysis of the data by Akaike's inform
finite corrections. *Communications in Statistics, Theory and Me*

Takeuchi, K. (1976). Distribution of informational statistics and a cr
Suri-Kagaku (Mathematic Sciences) **153**, 12–18. (In Japanese).

Taub, F.B. (1993). Book review: Estimating ecological risks. *Ecol*

Taubes, G. (1995). Epidemiology faces its limits. *Science* **269**, 164

Thabane, L., and Haq, M.S. (1999). On Bayesian selection of the
using the Kullback–Leibler divergence measure. *Statistica Neer*

Thompson, S.K. (1992). *Sampling*. Wiley, New York, NY.

Thompson, M.E. (1997). *Theory of sample surveys*. Chapman and

Thompson, W.L., and Lee, D.C. (2000). Modeling relationships b
attributes and snorkel counts of chinook salmon and steelhead p
Journal of Fisheries and Aquatic Sciences **57**, 1834–1842.

Tibshirani, R. (1996). Regression shrinkage and selection via the las
Statistical Society, Series B **58**, 267–288.

Tong, H. (1994). Akaike's approach can yield consistent order deter
in H. Bozdogan (ed.) *Engineering and Scientific Applications*. Ve
First US/Japan Conference on the Frontiers of Statistical Mode
Approach. Kluwer Academic Publishers, Dordrecht, the Netherl

Tukey, J.W. (1980). We need both exploratory and confirmatory. *Th*
34, 23–25.

Turchin, P., and Batzli, G.O. (2001). Availability of food and the
arvicoline rodents. *Ecology* **82**, 1521–1534.

Ullah, A. (1996). Entropy, divergence and distance measures with ec
Journal of Statistical Planning and Inference **49**, 137–162.

Umbach, D.M., and Wilcox, A.J. (1996). A technique for measu
useful features of birthweight distributions. *Statistics in Medicin*

Venter, J.H., and Snyman, J.L.J. (1995). A note on the generalized c
in linear model selection. *Biometrika* **82**, 215–219.

Ver Hoef, J.M. (1996). Parametric empirical Bayes methods for
Ecological Applications **6**, 1047–1055.

Wade, P.R. (2000). Bayesian methods in conservation biology. *Co*
1308–1316.

Walters, C.J. (1996). Computers and the future of fisheries. Pages 22
and E. Mcksness (eds.) *Computers in fisheries research*. Chapm

Wang, C. (1993). *Sense and nonsense of statistical inference: co*
subtlety. Marcel Dekker, Inc., New York, NY.

Wang, P., Puterman, M. L., Cockburn, I., and Le, N. (1996). Mi
models with covariate dependent rates. *Biometrics* **52**, 381–400.

Wasserman, L. (2000). Bayesian model selection and model
Mathematical Psychology **44**, 92–107.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, general
the Gauss–Newton method. *Biometrika* **61**, 439–447.

Wehrl, A. (1978). General properties of entropy. *Reviews of Moder*

Weiner, J. (1995). On the practice of ecology. *Journal of Ecology* **8**

Weisberg, S. (1985). *Applied linear regression*. 2nd ed. Wiley, New

Wel, J. (1975). Least squares fitting of an elephant. *Chemtech* Feb.

Westfall, P.H., and Young, S.S. (1993). *Resampling-based multipl*
methods for p-value adjustment. John Wiley and Sons, New Yor

White, G.C. (1983). Numerical estimation of survival rates fr
biotelemetry data. *Journal of Wildlife Management* **47**, 716–728

White, G.C., Anderson, D.R., Burnham, K.P., and Otis, D.L. (19
and removal methods for sampling closed populations. Los Alam
LA-8787-NERP, Los Alamos, NM.

White, G.C., and Burnham, K.P. (1999). Program MARK-su
populations of marked animals. *Bird Study* **46**.

White, G.C., Burnham, K.P., and Anderson, D.R. (2001). Advanc
MARK. *in* R. Fields (ed.) *Integrating People and Wildlife fo*
Proceedings of the Second International Wildlife Management
Society, Bethesda, MD.

White, H. (1994). *Estimation, inference and specification analysis*
Press, Cambridge, UK.

White, H. (2000). A reality check for data snooping. *Econometric*

Williams, B.K., Nichols, J.D., and Conroy, M.J. (2002). *Analysis an*
populations: modeling, estimation, and decision making. Acad
CA.

Williams, D.A. (1982). Extra-binomial variation in logistic linear m
31, 144–148.

Wood, S.N. and Thomas, M.B. (1999). Super-sensitivity to structu
Proceedings of the Royal Society **266**, 565–570.

Woods, H., Steinour, H.H., and Starke, H.R. (1932). Effect of c
cement on heat evolved during hardening. *Industrial and Eng*
1207–1214.

Ye, J. (1998). On measuring and correcting the effects of data mini
Journal of the American Statistical Association **93**, 120–131.

Yockey, H.P. (1992). *Information theory and molecular biology*.
Press.

Yoccoz, N.G. (1991). Use, overuse, and misuse of significance tests
and ecology. *Bulletin of the Ecological Society of America* **72**, 1

Young, L.J., and Young, J.H. (1998). *Statistical ecology*. Kluwe
London, UK.

Yu, B. (1996). Minimum description length principle: a review. In
Conference on Information Sciences and Systems, Princeton Un

Yu, B. (1999). Coding and model selection: a brief tutorial on the
description length. *Statistical Computing and Graphics* **9**, 1, 27-

Zablan, M.A. (1993). Evaluation of sage grouse banding program i
M.S. thesis, Colorado State University, Fort Collins, CO.

Zellner, A., Keuzenkamp, H. A., and McAleer, M., (eds.) (2001). *S*
modelling: keeping it sophisticatedly simple. Cambridge Unive
UK.

Zhang, P. (1992a). Inferences after variable selection in linear regres
79, 741–746.

Zhang, P. (1992b). On the distributional properties of model selec
the American Statistical Association **87**, 732–737.

Zhang, P. (1993a). Model selection via multifold cross-validation.
299–313.

Zhang, P. (1993b). On the convergence rate of model selection crite
Statistics, Part A—Theory and Methods **22**, 2765–2775.

Zhang, P. (1994). On the choice of penalty term in generalized FPE
in P. Cheeseman, and R.W. Oldford (eds.) *Selecting models from*
New York, NY.

Zucchini, W. (2000). An introduction to model selection. *Jou*
Psychology **44**, 41–61.

Index