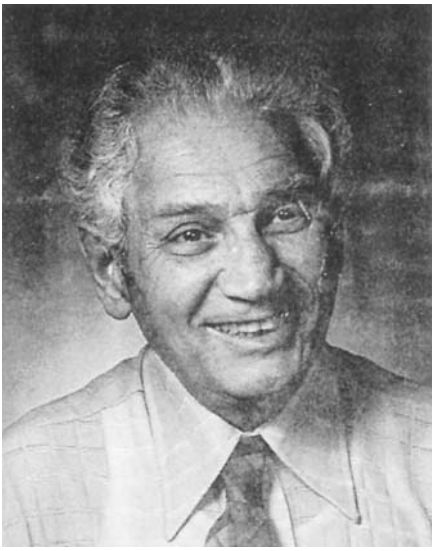
3

Information Theory and Entropy



Solomon Kullback (1907–1994) was born in Brooklyn, New York, USA, and graduated from the City College of New York in 1927, received an M.A. degree in mathematics in 1929, and completed a Ph.D. in mathematics from the George Washington University in 1934. Kully as he was known to all who knew him, had two major careers: one in the Defense Department (1930–1962) and the other in the Department of Statistics at George Washington University (1962–1972). He was chairman of the Statistics Department from 1964–1972. Much of his professional life was spent in the National Security Agency and most of his work during this time is still classified. Most of his studies on information theory were done during this time. Many of his results up to 1958 were published in his 1959 book, "*Information Theory and Statistics*." Additional details on Kullback may be found in Greenhouse (1994) and Anonymous (1997).

When we receive something that decreases our uncertainty about the state of the world, it is called *information*. Information is like "news," it informs. Information is not directly related to physical quantities. Information is not material and is not a form of energy, but it can be stored and communicated using material or energy means. It cannot be measured with instruments but can be defined in terms of a probability distribution. Information is a decrease in uncertainty.

This textbook is about a relatively new approach to empirical science called "information-theoretic." The name comes from the fact that the foundation originates in "information theory"; a set of fundamental discoveries made largely during World War II with many important extensions since that time. One exciting discovery is the ability to actually quantify *information* and this has led to countless breakthroughs that affect many things in our daily lives (e.g., cell phone and global positioning system technologies). One might think of information theory as being things like coding and encrypting theory and signal transmission, but it is far more general than these subjects.

Allowing "data analysis" to hook up with information theory has had substantial advantages and statistical scientists are still trying to exploit this combination. The concepts and practical use of the information-theoretic approach are simpler than that of hypothesis testing, and much easier than Bayesian approaches to data analysis.

Before proceeding further, I want to summarize the necessary "setting." This setting will set the tone for all of the following material. I will assume the investigator has a carefully considered science question and has proposed R hypotheses (the "multiple working hypotheses"), all of which are deemed plausible. A mathematical model (probability distribution) has been derived to well represent each of the R science hypotheses. Estimates of model parameters (θ) and their variance–covariance matrix ($\hat{\Sigma}$) have been made under either a least squares (LS) or maximum likelihood (ML) framework. In either case, other relevant statistics have also been computed (adjusted R^2, residual analyses, goodness-of-fit tests, etc.). Then, under the LS framework, one has the residual sum of squares (RSS), while under a likelihood framework, one has the value of the log-likelihood function at its maximum point. **This value (either RSS or max log($\mathcal{L}$)) is our starting point and allows answers to some of the relevant questions of interest to the investigator, such as:**

- Given the data, which science hypothesis has the most empirical support (and by how much)?
- What is the ranking of the R hypotheses, given the data?
- What is the probability of, say, hypothesis 4, given the data and the set of hypotheses?
- What is the (relative) likelihood, say, of hypothesis 2 vs. hypothesis 5?
- How can rigorous inference be made from all the hypotheses (and their models) in the candidate set? This is multimodel inference.

3.1 Kullback–Leibler Information

The scope of theory and methods that might be classed as "information theory" is very large. I will focus primarily on Kullback–Leibler information and this comes from a famous paper by Soloman Kullback and Richard Leibler published in 1951. Their work was done during WWII and published soon after the termination of the war.

Kullback–Leibler Information

In the context of this book, Kullback–Leibler (K–L) information is a function denoted as "*I*" for information. This function has two arguments: *f* represents full reality or "truth" and *g* is a model. Then, **K–L information *I*(*f, g*)** is the

"information" lost when the model *g* is used to approximate full reality, *f*.

An equivalent, and very useful, interpretation of *I*(*f, g*) is the

"distance" from the approximating model *g* to full reality, *f*.

Under either interpretation, we seek to find a candidate model that minimizes *I*(*f*, *g*), over the hypothesis set, represented by models.

Thus, if one had a set of five hypotheses, each represented by a model, *I*(*f*, *g*) would be computed for each of the five. The model with the smallest information loss would be the best model and, therefore, would represent the best hypothesis. The model *g* has its parameters given; there is no estimation and no data involved at this point (this will change as we go forward).

Alternatively, one could interpret the model with the smallest *I*(*f*, *g*) value as being "closest" to full reality. Thus, when a "best model" is mentioned, the "best" will stem from the concept of the smallest information loss or a model being closest to full reality. This is a conceptually simple, yet powerful, approach. The idea of a "distance" between a model and full reality seems compelling.

Kullback–Leibler information is defined by the unpleasant-looking integral for continuous distributions (e.g., the normal or gamma):

$$I(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x \mid \theta)}\right) \mathrm{d}x.$$

K–L information is defined as the summation for discrete distributions (e.g., Poisson, binomial, or multinomial):

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

Here, there are k possible outcomes of the underlying random variable; the true probability of the ith outcome is given by p_i, while the $\pi_1,\ldots,\pi_k$ constitute the approximating probability distribution (i.e., the approximating model). In the discrete case, we have $0 < p_i < 1$, $0 < \pi_i < 1$, and $\sum p_i = \sum \pi_i = 1$. Hence, here f and g correspond to the p_i and π_i, respectively. In the following material, we will generally think of K–L information in the continuous case and use the notation f and g for simplicity.

Some readers might start to "lose it" thinking that they must compute K–L information loss for each model in the set. It turns out that *I*(*f*, *g*) cannot be used

directly because it requires knowledge of full reality (f) and the parameters (θ) in the approximating models, g_i; we will never have knowledge of these entities in real problems. We will see that K–L information can be easily *estimated*, without advanced mathematics (although the derivation is very deeply mathematical). This estimation requires data relevant to the science question.

Kullback–Leibler information is the most fundamental of all information measures in the sense of being derived from minimal assumptions and its additivity property. It can be viewed as a quantitative measure of the *inefficiency* of assuming a model g when truth is f. Again, one wants to select a model from the set that minimizes inefficiency. While the Kullback–Leibler distance can be conceptualized as a "distance" between f and g, strictly speaking this is a measure of "discrepancy." It is not a simple distance because the measure from f to g is not the same as the measure from g to f – it is a "directed" or "oriented" distance.

3.2 Linking Information Theory to Statistical Theory

We usually think that "data analysis" is tied in with the subject of "statistics." How are statistical principles linked with information theory, and K–L information in particular? This linkage was the genius of Hirotugu Akaike in an incredible discovery first published in 1973.

A glimpse into the linkage between information and entropy and their relationship to mathematical statistics is given below; a full and technical derivation appears in Burnham and Anderson (2002:Chap. 7). I urge people to wade through this to gain a notion of the derivation. In particular, when there are unknown parameters to be estimated from data, the criterion must change. This change is introduced in the derivation to follow:

Akaike's main steps started by using a property of logarithms (i.e., $\log(A/B) = \log(A) - \log(B)$) to rewrite K–L information as

$$I(f,g) = \int f(x)\log(f(x))\mathrm{d}x - \int f(x)\log(g(x\,|\,\theta))\mathrm{d}x.$$

Both terms on the right-hand side are statistical expectations (Appendix B) with respect to f (truth). Thus, K–L information can be expressed as

$$I(f,g) = E_f[\log(f(x))] - E_f[\log(g(x\,|\,\theta))],$$

each expectation with respect to the true distribution f. This last expression provides insights into the derivation of AIC. Note that no approximations have been made, no parameters have been estimated and there are no data at this point; K–L information has merely been re-expressed.

The first expectation is a constant that depends only on the conceptual true distribution and it is not clearly known. However, this term is constant across the model set. In other words, the expectation of $[\log(f(x))]$ does not change

from model to model; it is a constant. Thus, we are left with only the second expectation,

$$I(f,g) - C = -E_f[\log(g(x \mid \theta))].$$

The constant term (C) can be made to vanish in a subsequent step (Chap. 4). The question now is if we can somehow compute or estimate $E_f[\log(g(x|\theta))]$. The short answer is no as the criterion or target must be altered to achieve a useful result and this will require data.

Kullback–Leibler information or distance $I(f, g)$ is on a true ratio scale, where there is a true zero. In contrast, $-E_f[\log(g(x|\theta))] = -\int f(x)\log(g(x|\theta))dx$ is on an interval scale and lacks a true zero, because of the constant (above). A difference of magnitude D means the same thing anywhere on the scale. Thus, $D = 10 = 12 - 2 = 1012 - 1002$; a difference of 10 means the same thing anywhere on the interval scale. Then, $10 = V_1 - V_2$, regardless of the size of V_1 and V_2. A large sample size magnifies the separation of research hypotheses and the models used to represent them. Adequate sample size conveys a wide variety of advantages in making valid inferences (e.g., improved estimates of $E_f[\log(g(x|\theta))]$).

3.3 Akaike's Information Criterion

Akaike introduced his information-theoretic approach in a series of papers in the mid-1970s as a theoretical basis for model selection. He followed this pivotal discovery with several related contributions beginning in the early 1980s and classified these as falling under the *entropy maximization principle*. This world class discovery opened the door for the development of relatively simple methods for applied problems, ranging from simple to quite complex, but based on very deep theories – entropy and K–L information theory on the one hand and Fisher's likelihood theory (see Appendix A) on the other.

Akaike's (1973) seminal paper used Kullback–Leibler information as a fundamental basis for model selection and recognized model parameters must be estimated from data and there is substantial uncertainty in this estimation. The estimation of parameters represents a major distinction from the case where model parameters are assumed to be known. Akaike's finding of a relation between the K–L information and the maximized log-likelihood has allowed major practical and theoretical advances in model selection and the analysis of complex data sets. deLeeuw (1992) said it well, "Akaike found a formal relationship between Boltzmann's entropy and Kullback–Leibler information (dominant paradigms in information and coding theory) and maximum likelihood (the dominant paradigm is statistics)."

Akaike's next step was stymied as no way could be found to compute or estimate the second term, $E_f[\log(g(x|\theta))]$. However, the expectation of this quantity led to a major breakthrough. Data enter the derivation and allow parameter estimates ($\hat{\theta}$)Akaike found that he could not estimate K–L, but he could estimate the expectation of K–L information. This second expectation is over the data (denote these data as y)

$$E_f[\log(g(x|\,\hat{\theta}))],$$

where the estimates $\hat{\theta}$ are based on the data (y).

The Modified Target

Akaike showed that the critical issue became the estimation of

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

This double expectation, both with respect to truth f, is the target of all model selection approaches based on K–L information. This notation makes it clear that the first (outer) expectation is over the data (y) and these data allow estimates of the unknown model parameters. Thus, we now have modified the target of relevance here due to the need for data to estimate model parameters. The proper criterion for model selection relates to the *fitted* model. The modification required is *expected* K–L information; Akaike called this a "predictive likelihood."

Akaike realized that this complex entity was closely related to the log-likelihood function at its maximum. However, the maximized log-likelihood is biased upward as an estimator of this quantity. Akaike found that, under certain conditions, this bias is approximately equal to K, the number of estimable parameters in the approximating model. This is an asymptotic (meaning as sample size increases to infinity) result of fundamental importance.

Thus under mild conditions, an asymptotically unbiased estimator of

$$\boldsymbol{E_y E_x\,[\log(g(x|\,\hat{\theta}(y)))] = \log(\mathcal{L}(\hat{\theta}|\text{data}) - K.}$$

This stunning result links expected K–L information to the maximized log-likelihood ($\log(\mathcal{L})$) corrected for bias. The important linkage is summarized as,

negentropy = K–L information and E(K–L information) = $\log(\mathcal{L})$ – K

thermodynamics ***information theory*** ***statistics***

Akaike's final step defined "*an information criterion*" (AIC) by multiplying both terms through by –2 ("taking historical reasons into account"). Thus, *both* terms in $\log(\mathcal{L}(\hat{\theta}|\,data)) - K$ were multiplied by –2 to get

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

This has become known as *Akaike's Information Criterion* or AIC. AIC has a strong theoretical underpinning, based on entropy and expected Kullback–Leibler information. Akaike's inferential breakthrough was finding that the maximized log-likelihood could be used to estimate the expected (averaged) K–L distance between the approximating model and the true generating mechanism. The expectation of the logarithm of $f(x)$ drops out as a constant across models, independent of the data.

In practice, one computes AIC for each of the models in the set and then selects the model that yields the smallest value of AIC for inference. One justifies this selection because that selected model minimizes the information lost when approximating full reality by a fitted (i.e., parameters estimated from the data using, for example, ML or LS methods) model. Said another way, that selected model is "closest" to full reality, given the data. This approach seems a very natural, simple concept; select the approximating model that is closest to the unknown reality.

It might be argued that I should have merely defined $l = \log(\mathcal{L}(\theta|\text{data,model}))$; then $\text{AIC} = -2l + 2K$, making the criterion appear more simple. While this may have advantages, I believe the full notation works for the reader and helps in understanding exactly what is meant. The full notation, or abbreviations such as $\log(\mathcal{L}(\theta|x,g_i))$, makes it implicit that the log-likelihood is a function of (only) the parameters (θ); while the data (x) and model (g_i, say multinomial) must be *given* (i.e., known). These distinctions become more important when we introduce the concept of a likelihood of a model, given the data: $\mathcal{L}(g_i|\text{data})$ in Chap. 4. Both concepts are fundamental and useful in a host of ways in this book and the notation serves an important purpose here.

3.3.1 The Bias Correction Term

Correction of estimators for bias has a long history in statistics. The usual estimator of the variance is a ready example

$$\text{variance} = \frac{\sum (x_i - \hat{\mu})^2}{n-1},$$

where the subtraction of 1 from the sample size (n) in the denominator corrects for a small sample bias (note that as n gets large the bias correction becomes unimportant). The bias correction term (K = the number of estimable parameters), above, is a special case of a more general result derived by Takeuchi (1976) and described in Sect. 3.9.1. AIC is a special case of Takeuchi's Information Criterion (TIC) and is, itself, a parsimonious approach to the estimation of expected K–L information.

3.3.2 Why Multiply by −2?

Akaike multiplied the bias-corrected log-likelihood by −2 for "historical reasons." It is a well-known statistical result that −2 times the logarithm of the

ratio of two maximized likelihood values is asymptotically chi-square distributed under certain conditions and assumptions (this is the likelihood ratio test). The term −2 occurs in other statistical contexts, and so it was not unreasonable that Akaike performed this simple operation to get his AIC. Three points frequently arise and I will note these here.

First, the model associated with the minimum AIC remains unchanged had the bias-corrected log-likelihood (i.e., $\log(\mathcal{L}) - K$) been multiplied by −0.17, −51.3, −3.14159, or any other negative number. Thus, the minimization is not changed by the multiplication of *both* terms by any negative constant; Akaike merely chose −2. Second, some investigators have not realized the formal link between expected K–L information and AIC and believed, then, that the number 2 in (*only*) the second term in AIC was somehow arbitrary and that other multipliers should also be considered. This error has led to considerable confusion in the technical literature; $-K$ is the asymptotic bias correction and is not arbitrary. Akaike chose to work with $-2\log(\mathcal{L})$, rather than $\log(\mathcal{L})$; thus the term $+ 2K$ is theoretically correct for large sample size. As long as *both* terms (the log-likelihood and the bias correction term) are multiplied by the same negative constant, the model where the criterion is minimized is unchanged and there is nothing arbitrary. Third, $-2\log(\mathcal{L})$ is termed "deviance" in mathematical statistics. People with a statistical background immediately interpret deviance as a way to quantify lack of fit and they then view AIC as simply "deviance + $2K$." I suspect that this was Akaike's thinking when he multiplied through by −2; that is simply, "deviance penalized by $2K$ to correct for asymptotic bias."

3.3.3 Parsimony is Achieved as a by-Product

AIC is linked directly to the estimation of expected K–L information. The derivation itself was not based on the concept of parsimony. It was after Akaike's elegant derivation of AIC that people noticed a heuristic interpretation that was interesting and allowed insight into how parsimony is enforced with AIC. The best model is closest to full reality and, therefore, the goal is to find the model where AIC is smallest. The first term (the deviance) in AIC

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

is a measure of lack of model fit, and can be made smaller by adding more parameters in the model g_i. Thus, for a fixed data set, the further addition of parameters in a model g_i will allow it to fit better. However, when these added parameters must be estimated (rather than known or "given"), further uncertainly is added to the *estimation* of expected K–L information or distance. At some point, the addition of still more estimated parameters will have the opposite effect and the estimate of expected K–L information will increase because "noise" is then being modeled as if it were structural. The second term in AIC ($2K$) then functions as a "penalty" for adding more parameters in the model.

Thus, the penalty terms (2*K*) gets larger as more parameters are added. One can see that there is a tension between the deviance and the penalty term as the number of parameters is increased – a trade-off.

Without a proper penalty term the best model would nearly always be the largest model in the set, because adding more and more parameters to be estimated from the fixed amount of data would be without "cost" (i.e., no penalty). The result would be models that are overfit, have low precision, and risk spurious effects because noise is being modeled as structure.

This heuristic explanation does not do justice to the much deeper theoretical basis for AIC (i.e., the link with expected K–L information). However, the advantage of adding parameters and the concomitant disadvantage of adding still more parameters suggests a trade-off. This is the trade-off between bias and variance or the trade-off between underfitting and overfitting that is the Principle of Parsimony (see Sect. 2.4). Note that parsimony was not a condition leading to AIC, instead parsimony appears almost as a by-product of the end result of the derivation of AIC from expected K–L information.

Inferences for a given data set are conditional on sample size. We must admit that if much more data were available, then further effects could probably be found and supported. "Truth" is elusive; model selection tells us what inferences the data support, not what full reality might be. Full reality cannot be found using a finite data set.

3.3.4 Simple vs. Complex Models

Data analysis involves the critical question, "how complex a model will the data support?" and the proper trade-off between underfitting and overfitting. This dilemma has had a long history in the analysis of data and model based inference. As biologists, we think certain variables and structure must be in a 'good model' often without recognition that putting in too many variables and too much structure introduces large uncertainties, particularly when sample size is relatively small or even moderate. In addition, interpretability is often decreased as the number of parameters increases.

As biologists, we have a strong tendency to want to build models of the information in the data that are too complex (overfit). This is a parsimony issue that is central to proper model selection. One cannot rely on intuition to judge a proper trade-off between under- and overfitting, a criterion based on deep theory is needed. Expected K–L information and AIC provide the basis for a rigorous trade-off. This seems a very natural, simple concept; *select the fitted approximating model that is estimated, on average, to be closest to the unknown full reality, f.*

Ideal model selection results in not just a good fitting model, but a model with good out-of-sample prediction performance. This is a tall order. The selected model should have good achieved confidence interval coverage for the estimators in the model and small predictive mean squared errors (PMSE).

3.3.5 AIC Scale

As defined, AIC is strictly positive. However, during an analysis, it is common to omit mathematical terms that are constant across models and such shortcuts can result in negative values of AIC. Computing AIC from regression statistics often results in negative AIC values. This creates no problem, one just identifies the model with the smallest value of AIC and declares it is the model estimated to be the best. This fitted model is estimated to be "closest" to full reality and is a good approximation for the information in the data, relative to the other models considered. For example,

Model	AIC
g_1	1,400
g_2	1,570
g_3	1,390
g_4	1,415.

One would select model g_3 as the basis for inference as it has the smallest AIC value; meaning that it is estimated to be closest to full reality. Because these values are on a relative scale, one could subtract, say, 2,000 from each and have the following rescaled AIC values: −600, −430, −610, and −585. The rank of each model is not changed by the rescaling; the ranks, in each case remain g_3 (best), g_1, g_4, and g_2 (worst). I have seen AIC values that range from −80,000 to as high as 340,000 in different scientific applications. It is not the absolute size of the AIC value, it is the *relative* values, and particularly the differences, that are important (Chap. 4).

3.4 A Second-Order Bias Correction: AICc

Second-Order Bias Correction: AICc

Akaike derived an asymptotically unbiased estimator of expected K–L information; however, AIC may perform poorly if there are too many estimated parameters in relation to the size of the sample. A second-order variant of AIC has been developed and **it is important to use this criterion in practice:**

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

where n is sample size. This can be rewritten as

$$\mathbf{AICc} = -2\log(\mathcal{L}(\hat{\theta})) + 2\mathbf{K} + \frac{2K(K+1)}{n-K-1}$$

or equivalently

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

AICc merely has an additional bias correction term. If n is large (asymptotic) with respect to K, then the second-order correction is negligible and AICc converges to AIC. AICc was derived under Gaussian assumptions and is weakly dependent on this assumption. Other model-specific assumptions can be made and this might be worthwhile in data analysis where there are severe controversies or consequences (Burnham and Anderson 2002: Chap. 7). ***The use of AICc is highly recommended in practice; do not use just AIC***.

3.5 Regression Analysis

Least squares regression is a very useful approach to modeling. Here, model selection is often thought of as "variable selection." It is easy to move from regression statistics such as the residual sum of squares (RSS) to the log-likelihood function at its maximum point; this allows one to use AICc. Note, LS and ML provide exactly the same estimates of the β_j in linear models; however, the estimates of the residual variance σ^2 can differ appreciably if sample size is small.

Mapping the RSS into the Maximized Log-Likelihood

The material to this point has been based on likelihood theory (Appendix A) as it is a very general approach. In the special case of LS estimation ("regression") with normally distributed errors, and apart from a constant, we have

$$\log(\mathcal{L}) = -\frac{n}{2}\cdot\log(\hat{\sigma}^2).$$

Substituting this expression, AICc for use in LS models can be expressed as

$$\text{AICc} = n\log(\hat{\sigma}^2) + 2K\left(\frac{n}{n-K-1}\right),$$

where $\hat{\sigma}^2 = \sum \hat{\varepsilon}_i^2 / n$ (the MLE) and $\hat{\varepsilon}_i$ are the estimated residuals for a particular candidate model.

A common (but minor) mistake is to take the LS estimate of σ^2 from the computer output, instead of the ML estimate (above). In regression models, K is the total number of estimated parameters, including the intercept and σ^2. The value of K is sometimes computed incorrectly as either β_0 or σ^2 are mistakenly ignored in obtaining K. AICc is easy to compute from the results of LS estimation in the case of linear models. It is not uncommon to see computer software that computes simple AIC value incorrectly; few packages provide AICc; however, this can be computed easily manually.

Given a set of candidate models g_i, with parameters to be estimated from the observed data, the model which minimizes the predictive expectation is "closest" to full reality (f) and is to be preferred as a basis for inference. AICc allows an estimate as to which model is best for a given data set; however, a different best model might be selected if another (replicate) data set was available. These are stochastic biological processes, often with relatively high levels of complexity, we must admit to uncertainty and try to quantify it. This condition is called "model selection uncertainty." We must also admit that if much more data were available, then further effects could probably be found and supported. "Truth" is elusive; proper model selection helps us understand what inferences the data support.

AICc attempts to reformulate the problem explicitly as a problem of *approximation* of the true structure (probably infinite dimensional) by a *model*. Model selection then becomes simply finding the model where AICc is minimized. I will show in a later chapter that model selection is much more than this.

AICc selection is objective and represents a very different paradigm to that of null hypothesis testing and is free from the arbitrary α levels, the multiple testing problem, and the fact that many candidate models are not nested. The problem of what model to use is inherently not a null hypothesis testing problem.

The fact that AIC allows a simple comparison of models does not justify the comparison of all possible models. If one had 10 variables, then there are 1,024 possible models, even if interactions and squared or cubed terms are excluded. If sample size is $n \leq 1{,}000$, overfitting is almost a certainty. It is simply not sensible to consider such a large number of models because an overfit model will almost surely result and the science of the problem has been lost. Even in a very exploratory analysis it seems like poor practice to consider all possible models; surely some science can be brought to bear on such an unthinking approach. I continue to see papers published where tens of thousands or even millions of models are fit and evaluated; this represents a foolish approach and virtually guarantees spurious effects and absurdities.

As a generally useful rule, when the number of models (R) exceeds the sample size (n), one is asking for serious inferential difficulties. I advise people to think first about their set of *a priori* science hypotheses; these will typically be relatively few in number. A focus on models is the result of computer software that is very powerful, but unthinking.

3.6 Additional Important Points

3.6.1 Differences Among AICc Values

Often data do not support only one model as clearly best for data analysis (i.e., little or no model selection uncertainty). Instead, suppose three models are essentially tied for best, while another subset of models is clearly not appropriate (either under- or overfit). Such virtual "ties" for the estimated best model

must be carefully considered and admitted. The inability to ferret out a single best model is not a defect of AICc or any other selection criterion, rather, it is an indication that the data are simply inadequate to reach such a strong inference.

It is perfectly reasonable that several models would serve nearly equally well in approximating the information in a set of data. Inference must admit that there are sometimes competing hypotheses and the data do not support selecting only one. Large sample sizes often reduce close ties among models in the set. The issue of competing models is especially relevant in including model selection uncertainty into estimators of precision and model averaging (Chap. 5).

Consider studies of *Plasmodium* infection of children in tropical Africa and data from two different sites have been modeled and fitted. The best model for the eastern site has AIC = 104, whereas the best model for the western site has AIC = 231. Are the models better for the western site? Perhaps, however, just the fact that the best model for the western site has a larger AIC value is *not* evidence of this. AIC values are functions of sample size and this precludes comparing AIC values across data sets.

3.6.2 Nested vs. Nonnested Models

The focus should be on the science hypotheses deemed to be of interest. Modeling of these hypotheses should not be constrained to only models that are nested. AICc can be used for nonnested models and this is an important feature because likelihood ratio tests are valid only for nested models. The ranking of models using AICc helps clarify the importance of modeling.

3.6.3 Data and Response Variable Must Remain Fixed

It is important that the data are fixed prior to data analysis. One cannot switch from a full data set to one where some "outliers" have been omitted in the middle of the analysis. It would be senseless to evaluate two hypotheses using data x and the remaining four hypotheses using a somewhat different data set. The fixed nature of the data is implied in the shorthand notation for models: $g(\theta|\text{data})$, the model as a function of the known parameters (θ), given the (fixed) data (x).

Some analyses can be done on either the raw data or some grouping of the raw data (e.g., histogram classes). In such cases, one must be consistent in performing the analysis on one data type or the other, not a mixture of both types. Any grouping of the raw data loses some information, thus grouping should be carefully considered.

If Y is the response variable of interest, it must also be kept fixed during the analysis. One cannot evaluate models of Y and then switch to models of $\log(Y)$ or $\sqrt{Y}$. Having a mix of response variables in the model set is an "apples and oranges" issue. Such changes make the AICc values uninterpretable; more importantly, the science problem is muddied. For example, presence–absence data on some plant species cannot be compared to counts of that plant on a series of plots.

3.6.4 AICc is not a "Test"

Information-theoretic approaches do not constitute a statistical "test" of any sort (see Appendix C). There are no test statistics, assumed asymptotic sampling distributions, arbitrary α-levels, P-values, and arbitrary decision about "statistical significance." Instead, there are numerical values that represent the scientific evidence (Chaps. 4 and 5), often followed by value judgments made by the investigators and perhaps others.

It is poor practice to mix evidence from information-theoretic approaches with the results of null hypothesis testing, even though this is a common mistake in the published literature. One sees cases where the models are ranked using AICc and then a "test" is carried out to see if the best model is "significantly better" than the second-best model. This is seriously wrong on several different technical levels and I advise against it. Null hypothesis testing and information-theoretic results are like oil and water; they do not mix well.

3.6.5 Data Dredging Using AICc

Ideally science hypotheses and their models are available prior to data analysis and, ideally, prior to data collection. These *a priori* considerations led to a confirmatory result. Following that, I often encourage some *post hoc* examination of the data using hypotheses and models suggested by the *a priori* results. Such after-thoughts are often called "data dredging." I do not condone the use of information-theoretic criteria in data dredging, even in the early phases of exploratory analysis. For example, one might start with 8–10 models, compute AICc for each, and note that several of the better models each have a gender effect. Based on these findings, another 4–7 models are derived to include a gender effect. After computing AICc for these models, the analyst notes that several of these models have a trend in time for some parameter set; thus more models with this effect are derived, and so on. This strategy constitutes traditional data dredging but using an information theoretic criteria instead of some form of test statistic or visual inspection of plots of the intermediate results. I recognize that others have a more lenient attitude toward blatant data dredging. I think investigators should understand the negative aspects of data dredging and try to minimize this activity.

3.6.6 Keep all the Model Terms

It is good practice to retain all the terms in the log-likelihood in order for AICc to be comparable across models. This is particularly important for nonnested models (e.g., the nine models of Flather, Sect. 3.9.6) and in cases where different error distributions are used (e.g., log-normal vs. gamma). If several computer programs are used to get the MLEs and the maximum $\log(\mathcal{L})$, then one is at risk that some terms in one model were dropped, while these terms were not dropped in other models. This is a rather technical issue: Burnham and Anderson (2002, Sect. 6.7) provide some insights and examples.

3.6.7 Missing Data

A subtle point relates to data sets where a few values of the response variable or predictor variables are missing. Such missing values can arise for a host of reasons, including loss, unreadable recording, and deletion of values judged to be incorrect. If a value or two are missing from a large data set, perhaps no harm is done. However, if the missing values are numerous at all then more careful consideration is called for. In particular, if some values for covariates are missing, this can also lead to important issues, including the fact that some software may either stop or give erroneous results. There are *ad hoc* routines for assigning "innocent" values to be used in place of the missing values; these could be considered. There are a variety of Bayesian "imputation" techniques that have merit; these are far beyond the scope of this text. The real moral here is to collect data with utmost care and in doing so, avoid issues with missing data.

3.6.8 The "Pretending Variable"

Putting aside the second-order correction for bias for a moment, AIC is just $-2\log(\mathcal{L}) + 2K$ or deviance $+ 2K$. The addition of each new parameter suffers a "penalty" of 2. Now, consider the case where model A has K parameters and model B has $K + 1$ parameters (i.e., one additional parameter). Occasionally, we find that model B is about 2 units from model A and thus, we would view model B as a good model – it *is* a good model. Problems arise when the two AIC values are about 2 units apart but the deviance is little changed by the addition of a variable or parameter in model B. In this case, the additional variable does not contribute to a better fit, instead, it is a "good" model only because the bias correction term is only 2 (i.e., 2×1). This should not be taken as evidence that the new parameter (and the variable it is associated with) is important. The new parameter is only "pretending" to be important; to confirm this, one can examine the estimate of the parameter (perhaps a regression coefficient β) and its confidence interval. However, the real clue here is that the deviance did not change and this is an indication that the model fit did not improve.

I will call this issue a "pretending variable" as a noninformative variable enters as one additional parameter and therefore incurs only a small "penalty" of about 2, but does not increase the log-likelihood (or decrease the deviance). Is this model (B) a good model? YES. Can we take this result to further imply that the added variable is important? NO. Thus, scientists must examine the table of model results to be sure that added variables increase the log-likelihood values. Pretending variables may arise for any models i and j in the set where the difference in AIC values increase by about 2. Less commonly, a model (call it C) will add two parameters and its added penalty is 4 (still a decent model). However, unless there is a change in the log-likelihood, the two new variables or parameters are only "pretending" to be important. Finally, when using AICc,

the values are a bit different from 2 or 4; still a focus on the log-likelihood or deviance is advised, to be sure that the fit has improved.

3.7 Cement Hardening Data

The computation of AICc from the cement hardening data from Sect. 2.2.1 is shown in the table below:

Model	K	$\hat{\sigma}^2$	$\log(\mathcal{L})$	AICc	Rank
{mean}	2	208.91	−34.72	74.64	5
{12}	4	4.45	−9.704	32.41	1
{12 1∗2}	5	4.40	−9.626	37.82	2
{34}	4	13.53	−16.927	46.85	3
{34 3∗4}	5	12.42	−16.376	51.32	4

PROC REG (SAS Institute 1985) was used to compute the residual sum of squares (RSS), the LS estimates of the β parameters, and the standard errors of the parameter estimates for each model. The MLE of the residual variance is $\hat{\sigma}^2 = \text{RSS}/n$, where the sample size (n) = 13. AICc = $n\log(\hat{\sigma}^2) + 2K + 2K(K + 1)/(n - K - 1)$ was used. The calculations can be illustrated using the information from the {mean} model above where $K = 2$. The MLE of the residual variance is 208.91, thus the first term in AICc is 13 log(208.91) = 69.444, the second term is 2·2 = 4, and the third term is (2·2·3)/(13 − 3) = 1.2. Summing the three terms leads to 74.64. The computations are easy but the reader should compute a few more entries in the table above to be sure they understand the notation and procedure. Note, log-likelihood values are usually negative, while AICc values are generally positive.

3.7.1 Interpreting AICc Values

AICc is an estimator of expected K–L information and we seek the fitted model where the information loss is minimal. Said another way, we seek the model where the estimated distance to full reality is as small as possible; this is the model with the smallest AICc value. The model that is estimated to be closest to full reality is referred to as the "best model." This best model is {12} from the table above, namely,

$$E(Y) = \beta_0 + \beta_1(x_1) + \beta_2(x_2)$$

with four parameters ($K = 4 = \beta_0, \beta_1, \beta_3$, and σ^2).

The {mean} model in the table is just the mean and variance of the response variable, thus only two parameters are estimated, β_0 and σ^2. The model notation {12 1∗2} denotes the model

$$E(Y) = \beta_0 + \beta_1(x_1) + \beta_2(x_2) + \beta_3(x_1 * x_2)$$

where an interaction term is introduced. This model is estimated to be the second best model; however, one can quickly see that the interaction is not an important predictor by examining the MLEs and their standard errors:

Parameter	MLE	$\hat{se}(\hat{\beta})$
β_1	1.275	0.581
β_2	0.636	0.091
β_3	**0.004**	**0.012**

Note that the estimated standard error on $\hat{\beta}_3$ is three times larger than the MLE, thus it certainly seems to be unimportant. This result allows an important, but subtle, point to be made here. Let us ask two questions from the information above. First, is the interaction model {12 1*2} a relatively good model? The answer is YES, this can be seen from the table of AICc values. Second, does this answer imply that the interaction term is an important predictor? The answer is *no*; to judge its importance one needs to examine the standard error of the estimate and a confidence interval (i.e, –0.020 to 0.028 for β_3). This interval was computed as $\hat{\beta}_3 \pm 2 \times \hat{se}(\hat{\beta}_3)$ and is essentially centered on zero and fails to support the hypothesized importance of β_3.

Another thing to note is that the scale (i.e., the size, Sect. 3.3.5) of the AICc values is unimportant. One cannot look at the AICc value for the third model (37.82) and judge weather it is "too big" or not "big enough." These AICc values have unknown constants associated with them and are functions of sample size. It is the relative values of AICc that are relevant. In fact, we will see in Chap. 4 that it is the *differences* in AICc values that become the basis for extended inferences.

Of course AICc allows a quick ranking of the five hypotheses, represented by the five models. The ranks are (from estimated best to worst): g_2, g_3, g_4, g_5, and g_1 for this simple example. Models {34} and {34 3*4} are poor and the mean-only model is very poor in the rankings. The ability to rank science hypotheses is almost always important; however, it will be seen that far more information can be gained using methods introduced in the following chapter.

3.7.2 What if all the Models are Bad?

If all five models are essentially worthless, AICc will still rank them; thus, one must have some way to measure the worth of the best model or the global model. In regression, a natural measure of the worth of a model is the adjusted R^2 value. In other contexts, one can use a method outlined by Nagelkerke (1991) for a likelihood-based analysis. In this case of cement hardening, the model estimated to be the best in the set was model {12} with an adjusted $R^2 = 0.974$. This suggests that the best model is quite good, at least for these data.

If this best model and its MLEs were used with a new, replicate data set, one would find that the adjusted R^2 would be substantially lower than 0.974. Adjusted R^2 exaggerates our notions about the out-of-sample predictive ability of models fit to a given data set. The derivation of AICc is based on a predictive likelihood and this attempts to optimize performance measures such as predictive mean squared error. Thus, AICc attempts to deal with out-of-sample prediction by its very derivation. Even model {34 3*4} had an adjusted $R^2 = 0.921$. R^2 is a descriptive statistic and should not be used for formal model selection (it is very poor in this regard). A likelihood version of "R^2" is given in Appendix A and is useful when the analysis has been done in a likelihood framework.

The danger of having all the models in the set be useless arises most often in exploratory work where little thought went into hypothesizing science relationships or data collection protocols. I have seen a number of habitat–animal association models where the best model in the set had an R^2 value around 0.06, certainly indicating that more work is needed. In such cases, the rankings of the models carry little meaning.

Generally some assessment of the worth of the global model is suggested. This assessment might be a goodness-of-fit test, residual analysis, adjusted R^2, or other similar approach. If a global model fits, AICc will not select a more parsimonious model that fails to fit. Thus, it is sufficient to check the worth and fit of the global model. Often it is appropriate to provide an R^2 value for the best model in reports or publications.

Another approach relies on including a "null" model in the set to evaluate the worth of particular hypotheses or assumptions. Consider a study of growth in tadpoles where density is a hypothesized covariate. One could evaluate a model where growth depends on density and another model where growth is independent of density. This procedure, used carefully, might allow insights as to the worth of models in the set. Details for such evaluations are given in Chap. 4.

3.7.3 Prediction from the Best Model

One goal of selecting the best model is to use it in making inferences from the sample data to the population. This is model based inductive inference and prediction is one objective. In this case, prediction would come from the model structure:

$$E(Y) = \beta_0 + \beta_1(x_1) + \beta_2(x_2),$$

where x_1 = calcium aluminate and x_2 = tricalcium silicate. The least squares estimates of β_1 and β_2 allow predictions to be made from the fitted model:

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

The adjusted R^2 for this model is 0.974 suggesting that prediction is expected to be quite good until one realizes that the out-of-sample prediction performance might be poor with a sample size of 13 and the fitting of four parameters. This issue will be further addressed in Chap. 5.

3.8 Ranking the Models of Bovine Tuberculosis in Ferrets

The computation of AICc from the tuberculosis data allows a ranking of the five science hypotheses and these are shown in the table below:

Hypotheses	K	$\log(\mathcal{L})$	AICc	Rank
H_1	6	−70.44	154.4	4
H_2	6	−986.86	1,987.2	5
H_3	6	−64.27	142.1	3
H_4	6	−45.02	103.6	1
H_5	6	−46.20	105.9	2

AICc for the model corresponding to H_5 is computed as

$$\text{AICc} = -2\log(\mathcal{L}(\hat{\theta})) + 2K + \frac{2K(K+1)}{n-K-1} = -2(-46.20) + 2\cdot 6 + (2\cdot 6\cdot 7)/(62-5) = 92.4 + 12 + 1.474 = 105.9.$$

Because $K = 6$ for all five models in this example, AIC and AICc would select the same model.

Parameter estimates for these models were MLEs and there were no estimates of residual variance $\hat{\sigma}^2$; instead, the maximized value of the log-likelihood was available directly. Here it is easy to compute AICc, given the number of estimable parameters (six for each model here), the sample size ($n = 62$), and the value of the maximized log-likelihood function (tabled above) for each of the five models. The reader is asked to verify the computation for a few entries in the table to be sure they understand the issues. Note, too, because a likelihood approach was used here, there is no statistics strictly analogous to an R^2 value in the usual (i.e., least squares) sense (but see Nagelkerke (1991) for a useful analog. A final technical note is that there is often no unique measure of "sample size" for binomial outcomes such as these; Caley and Hone were conservative in using $n = 62$ in this case.

Empirical support favors H_4, the dietary-related hypothesis as the best of the five hypotheses. Ranking hypotheses from best to worst was H_4, H_5, H_3, H_1, and H_2. Clearly, H_2 (transmission during mating and fighting from the age

of 10 months when the breeding season starts) seems very poor relative to the other hypotheses. Such observations and interpretations will be made more rigorous in Chap. 4. At this time it might be reasonable to begin to wonder about the ranking of hypotheses if a different data set of the same size was available for analysis; would the rankings be the same? This issue is termed "model selection uncertainty" and will turn out to be very important.

3.9 Other Important Issues

3.9.1 Takeuchi's Information Criterion

I mention Takeuchi's information criterion (TIC) but it is rarely used in practice. However, it is important in the general understanding of information criteria. Akaike's derivation assumed, at one step, an expectation over the model (not full reality). This has lead to the assumption that AIC was based on a true model being in the set; although Akaike clearly stated otherwise in his papers.

Takeuchi (1976), in a little known paper in Japanese, made the first published derivation clearly taking all expectations with respect to full reality. Takeuchi's TIC is an asymptotically unbiased estimate of expected K–L and does not rest in any way on the assumption that a "true model" is in the set. TIC is much more complicated to compute than AICc because its bias adjustment term involves the estimation of the elements of two $K \times K$ matrices of first and second partial derivatives, $J(\theta)$ and $I(\theta)$, the inversion of the matrix $I(\theta)$, and then the matrix product. TIC is defined as

$$\text{TIC} = -2\log(\mathcal{L}(\hat{\theta})|\text{data}) + 2\text{tr}(J(\theta)I(\theta)^{-1}),$$

where "tr" is the matrix trace operator. Unless sample size is *very* large, the estimate of $\text{tr}(J(\theta)I(\theta)^{-1})$ is often numerically unstable; thus, its practical application is nil (I have never seen TIC used in application). However, it turns out that a very good estimate of this messy term is merely K or $K + K(K+1)/(n-K-1)$, corresponding to AIC and AICc. Thus, it can be seen that AIC and AICc represent a *parsimonious* approach to bias correction! That is, rather than trying to compute estimates of all the elements in two K by K matrices, inverting one, multiplying the two, and computing the matrix trace, just use K or $K + K(K+1)/(n-K-1)$, as these are far more stable and easy to use. [In fact, if f was assumed to be in the set of candidate models, then for that model $tr(J(\theta)I(\theta)^{-1}) \equiv K$. If the set of candidate models includes any decent models, then $\text{tr}(J(\theta)I(\theta)^{-1})$ is approximately K for those models.]

It is important to realize that the deviance term nearly always dwarfs the "penalty" term in AICc or TIC. Thus, poor fitting models have a relatively large deviance and, thus, the exact value of the penalty term is not critical in many cases.

3.9.2 Problems When Evaluating Too Many Candidate Models

A common mistake is to focus on models without full consideration of the all important science hypotheses. Armed with too little science thinking and computer software that allows "all possible" models to be fit to a hapless data set, one is ready to find a wide variety of effects that are spurious. This is a subtle but important point and there is a large statistical literature on this matter. The entire fabric of the investigation breaks down in many exploratory studies where sample size might be only 35–80 and there are 15–20 explanatory variables, leading to about 33,000 or 1,050,000 models, respectively. In these cases, one may expect substantial overfitting and the finding of many effects that are actually spurious (Freedman 1983; Flack and Chang 1987; Anderson 2001). One useful rule of thumb is when the sample size is smaller than the number of models (i.e., $n < R$), then the analysis must be viewed as only exploratory (see Burnham and Anderson 2002:267–284). If one thinks as Chamberlin suggested, the focus will be on the science issues and multiple working hypotheses. Then develop models to represent these hypotheses, keeping an eye on the science and less so on countless models that can be run easily by sophisticated (but unthinking!) software. Good application can expect $n \gg R$.

Hoeting et al. (2006) provide an example of geostatistical modeling of whiptail lizards in southern California. There were 37 predictor variables available, leading to 1.4×10^{11} possible models. They were able to reduce the number of variables to six which resulted in a tractable 160 models. Three of these were judged to be good models and involved similar variables.

3.9.3 The Parameter Count K and Parameters that Cannot be Uniquely Estimated

Often there are some parameters in a model that are not uniquely estimable from the data and these should not both be counted in K. Such "nonidentifiability" can arise due to inherent confounding (e.g., the estimators of survival and sampling probabilities, S_{t-1} and f_t, respectively, in certain band recovery models of Brownie et al. 1985). In such cases, the correct value of K counts the product $S_{t-1} f_t$ as a single parameter (not two parameters). Here, it is the *estimators* $\hat{S}_{t-1}$ and $\hat{f}_t$ that are confounded, not the parameters themselves.

Smith et al. (2005) provide another example of nonidentifiablity in their study of entomological inoculation rates and *Plasmodium falcipraum* infection in children in Africa. Their best model was

$$\mathrm{PR} = 1 - \left(1 + \frac{b\varepsilon}{rk}\right)^{-k},$$

where PR = parasite ratio, b = transmission efficiency, ε = annual entomological inoculation rate, r = inverse of the expected time to clear an infection and $1/k$ = the coefficient of variation of the population infection rate. They found that b and r were exactly collinear, only the ratio b/k was relevant or identifiable. Thus, K would be 3 in this case: (b/r), ε, and $1/k$. If an error distribution was included, then K would be increased by 1. Nonestimability and nonidentifiability are common issues in some life science problems.

Sometimes a parameter is estimated on a boundary and this can be confusing. If the parameter being estimated is a probability (e.g., a transition probability of moving from state i to state j, say ψ_{ij}), then it may be that the MLE $\hat{\psi}$ is either 0 or 1 (i.e., on a boundary). Often the estimated standard error is 0, with a confidence interval of 0 width. In such cases, this parameter estimate must enter the count for K, even though some software may indicate such a parameter was "not estimated." Here, the parameter was estimated, it just happened that the most likely estimate was 0 or 1 (a boundary) and it should be counted in K.

Another technical point is the case where the iterative numerical procedure fails to "converge" in likelihood-based estimation. This condition is important and is nearly always noted on the output by the software. Until convergence is obtained or the specific situation understood, the analysis for that model should not go forward (i.e., the maximum of the log-likelihood function has not been found). Often, the failure to converge is due to the log-likelihood surface (see Appendix A) being nearly perfectly flat over some region in the parameter space. Thus, repeated tries to find the exact maximum point can fail. Alternatively, the log-likelihood surface might have more than a single mode, making valid inference more difficult (but there are many ways to address this problem).

3.9.4 Cross Validation and AICc

Basing AICc on the expectation (over $\hat{\theta}$) of $E_x[\log(g(x|\hat{\theta}(y)))]$ provides the criterion with a cross validation property for independent and identically distributed samples (Stone 1974, 1977). Golub et al. (1979) show that AIC asymptotically coincides with generalized cross validation in subset regression (also see review by Atilgan 1996). These are important results for application and are another by-product of Akaike's predictive likelihood. The practical utility of these findings suggest that computer-intensive cross validation results will average about the same result as just using AICc.

3.9.5 Science Advances as the Hypothesis Set Evolves

Evolution importantly involves time and information. Consider an investigator with $R = 5$ good, plausible science hypotheses, a mathematical model representing each of the five, and a set of relevant data from a proper collection scheme. Upon completion of the analysis using an information-theoretic

approach, it may become clear that two of the hypotheses have virtually no empirical support; their *likelihoods* (Chap. 4) are perhaps 3,000 or 6,600 to one of having utility.

At this point, one wants the hypothesis set to "evolve" allowing rapid progress in learning and understanding the system under study. First, the set is now reduced to three plausible alternatives (i.e., the two hypotheses lacking empirical support can be dropped from further consideration). Second, perhaps the three remaining hypotheses can be refined or their models can be made a better reflection of the intended hypothesis. Third, more hard thinking and consideration might lead to the introduction of one or more new hypotheses into the set. At this point, new data are collected and the process is repeated.

There is some art involved in this evolution. For example, if a large amount of new data can be anticipated, one must be careful and not discard some intricate hypotheses with high dimensioned models because such models might find support with a much larger data set. Often a scientist might prefer a more simple model if it predicts well, has parameters that are directly related to the system, and captures the main effects. Thus, there is some flexibility to use a model other than that estimated to be best for some inferences. An important aspect of science is that it never stops; each step (the set continually evolves) tends to lead to new and better understanding. Some steps might go "backward" for awhile, but science has a way of correcting missteps.

3.10 Summary

The crucial, initial starting point for advancement in the life sciences is a set of "multiple working hypotheses" defined prior to data analysis. These are the result of a determination to address the background science of the issue at hand. Following this important step, the science of the matter, experience, and expertise are used to define an *a priori* set of candidate models, representing the hypotheses. **These are important philosophical issues that must receive increased attention.** The research problem should be carefully stated, followed by careful planning concerning the sampling or experimental design. Sample size and other planning issues should be considered fully before the data gathering program begins. Information-theoretic procedures are not for rectifying poor science questions or resurrecting bad data.

Of course, hypotheses and models not in the set remain out of consideration. AICc can be useful in selecting the best model in the set; however, if all the models are very poor, AICc will still select the one estimated to be best and rank the rest. However, even that relatively best model will be poor in an absolute sense. Thus, every effort must be made to assure that the set of hypotheses and models is well founded.

A good model separates "information" from noise or noninformation. We are not trying to model the data; instead we are trying to model the information in

the data. We are trying to use the data at hand to make inferences about the process that generated the data and to make good out-of-sample predictions.

The underlying basis of AIC is (heuristically) a model that minimizes

$$E_{\hat{\theta}}(I(f, g(\cdot \mid \hat{\underline{\theta}}))).$$

This is the K–L distance or information loss, given the model is fit to the data (in the sense that parameters are estimated from the data). When faced with data and unknown model parameters, the target changes to *expected* K–L information and is based on the fitted model.

The Principle of Parsimony provides a conceptual guide to model selection, while expected K–L information provides an objective criterion, based on a deep theoretical justification. AICc provides a practical method for model selection and associated data analysis and are estimates of expected K–L information. AIC, AICc, and TIC represent extensions of classical likelihood theory, are applicable across a very wide range of scientific questions, and AICc is quite simple to use in practice.

I advise that the theories underlying the information theoretic approaches and hypothesis testing are fundamentally quite different. AICc is not a "test" in any sense and there are no associated concepts such as test power or α-levels; statistical hypothesis testing represents a very different paradigm. The results of model selection under the two approaches might happen to be similar with simple problems and a large amount of data; however, in more complex situations, with many candidate models and less data, the results can be quite different.

3.11 Remarks

Guiasu (1977) and Cover and Thomas (1991) provide an overview of the broad field of information theory for those wanting to read more. Akaike's main results on this issue appeared in 1973, 1974, and 1977, but these are for the statistically and mathematically gifted. His broader and more contextual papers appeared in 1981a and b, 1985, 1992, and 1994 and these are more readable by mortals. Many of Akaike's collected works were published by Parzen et al. (1998) and insights into Akaike's career are found in Findley and Parzen (1995).

Cohen and Thirring (1973) and Broda (1983) give a full account of Boltzmann's life and science contributions. It is said that Boltzmann was the nineteenth century's greatest scientist. Gallager (2001) and Golomb et al. (2002) provide information on Claude Shannon's life and contributions to information theory. It is said that Shannon's Master of Science thesis is the most famous or well-known thesis ever written. Claude Shannon wanted to go into genetics and his Ph.D. dissertation (never published) was on genetics. Like Boltzmann, Shannon was working far beyond existing science frontiers of the time.

Pronunciation is important; Akaike is pronounced with an accent on the "ka" and the "i" is pronounced like an "e" – AKAeke. Leibler is pronounced with the accent on the "i" while the "e" is silent – LIbler.

Akaike (1973) considered his information criterion to be a natural extension of R. A. Fisher's likelihood theory. It is of historic interest that Fisher (1936) anticipated such an advance when he wrote,

> "*an even wider type of inductive argument may some day be developed, which shall discuss methods of assigning from the data the functional form of the population.*"

Zellner's book (Zellner et al. 2001) and particularly Forster's chapter make for interesting reading about modeling and inference (also see Jessop 1995 and Wallace 2004). Some authors view K, the asymptotic bias correction term in AIC, as a measure of "complexity." Perhaps no harm is done in viewing it this way; however, it does not need to be so defined. I doubt if our word "complexity" can be quantified in a satisfactory way as a single number or quantity. I view K as merely an asymptotic bias correction term.

A technical point: Given a parametric structural model, there is a unique value of θ that, in fact, minimizes K–L information $I(f, g)$. This (unknown) minimizing value of the parameter depends on truth f, the model g through its structure, the parameter space, and the sample space (i.e., the structure and nature of the data that can be collected). In this-sense there is a "true" value of θ underling ML estimation (let this value be θ_o). Then θ_o is the absolute best value of θ for model g; actual K–L information loss is minimized at θ_o. If one somehow knew that model g was, in fact, the K–L best model, then the MLE $\hat{\theta}$ would estimate θ_o. This property of the model $g(x|\theta_o)$ as the minimizer of K–L, over all possible θ, is an important feature involved in the derivation of AIC or AICc (Burnham and Anderson 2002:Chap. 7).

Another technical point concerns f the conceptual full reality. At a high level of abstraction we consider entities such as random variables and probability distributions. These are intellectual ways of thinking and understanding. Such abstraction carries over the notion of full reality which I denote as f. This symbol relates to the *concept* of the best "model" of full reality. There are no unknown parameters; reality may not even be parameterized. We parameterize models in an effort to understand full reality, f.

Some computer software use the expression $2\log(\mathcal{L}) - 2K$ as "AIC" and then the objective is to maximize this across models. While this is not incorrect, it is certainly confusing and thus statements such as "bigger is better" must be displayed to help the user from getting to worst model and thinking it is the best model! I recommend against this practice; AIC has a clear definition and I think it is best to use it.

A colleague wrote his explanation for the "pretending variable" issue. Consider two models, (1) $E(Y) = \beta_o + \beta_1(X_1)$ and (2) $E(Y) = \beta_o + \beta\,(X_1) + \beta_2$ (independent random variable). Both models will have essentially the same deviance because of the addition of only a "noise" variable. The models will

differ by one parameter; the first model has $K = 3$, whereas the second model has $K = 4$, hence, Δ for the second model will be bigger that the first model by two. The clue here is that the deviance did not change with the addition of another variable and its parameter.

The nonparametric bootstrap can be used in model selection; this was investigated by Burnham and Anderson (2002) and in general we found the performance of the analytic approach to be as good as if not better. Given the computer-intensive nature of the bootstrap, we have not given this approach much more attention or interest. Still, it is a general approach and might find use in some cases.

One can find in the published literature that AIC is only for nested models; this statement is incorrect. Likewise, other literature states that AIC is only for nonnested models. This, too, is incorrect. The general derivation (e.g., Takeuchi 1976 or Burnham and Anderson 2002:Chap. 7) makes no restriction concerning nestedness.

The methods outlined in this book apply to virtually all problems where a likelihood exists (Lahiri 2001). In addition, there are general information-theoretic approaches for models well outside the likelihood framework (Qin and Lawless 1994; Ishiguro et al. 1997; Hurvich et al. 1998; Pan 2001a,b). There are now model selection methods for generalized estimation equations, kernel methods, martingales, nonparametric regression, and splines. Thus, methods exist for nearly all classes of models we might expect to see in the theoretical or applied life sciences.

Richard Leibler explained to me (about 1997) that many people thought their 1951 paper was a direct result from the war effort. Instead, the motivation for that (now) famous paper was to provide a rigorous definition of what Fisher meant by the word "information" in relation to his "sufficient statistics." Indeed, they showed that all the "information" in the data was contained in sufficient statistics, given the model; just as Fisher had alleged. Few people realized the importance of the 1951 paper; they got no reprint requests for their paper for many years! Also interesting, Leibler had never realized that K–L information was the negative of Boltzmann's entropy.

The K–L information or distance has also been called the K–L discrepancy, divergence, and number – I will treat these terms as synonyms, but tend to use *information* or *distance* in the material here (see Ullah 1996 for applications). Later, Kullback (1987) preferred the term *discrimination information*. Kullback served as head of the Statistics Department at George Washington University from 1964–1972 where he had a profound impact. He believed that information theory provides a unification of known results, leads to generalizations and the derivation of new results, and offers a unifying principle in statistics.

The second-order bias correction (leading to AICc) stems from Suguira's (1978) work and several follow-up papers by Hurvich et al. While these papers are not theoretical contributions on the same scale as Akaike's papers, they are very important in application. One should not use AIC in standard

application; people should be using AICc, the second-order version of AIC (or derive new results if a specific distribution is required, see Burnham and Anderson 2002:Sect. 7.4.2).

It must be noted that Rissanen (1989, 1996) has derived a sophisticated model selection theory based on information and coding theory. His approach is very different, both conceptually and mathematically, than that presented in this book. His initial contribution was MDL for minimum description length and he has extended this in later publications and books (Rissanen 2007). The MDL approach does not require prior distributions on parameters or models and many people would see this as an advantage in science issues. The MDL result was the same form as BIC (Appendix D), but later theory expands on this result. I will not go further into Rissanen's work as it is quite technical unless one has the required background in coding theory. I note only that this interesting class of "information-theoretic" alternatives exists.

Akaike (1973, 1974) used what he called a predictive log-likelihood in deriving his information criterion; this has advantages and properties that are still not well recognized in the literature. Full discussion of his approach is technical and I will not provide more than a few insights here (see Akaike 1973, 1987:319, 1992; Bozdogan 1987; Sakamoto 1991; deLeeuw 1992; and Burnham and Anderson 2002:Chap. 7). His approach involves a statistical expectation based on a different, independent sample. It is this second expectation over a conceptually independent "data set" that provides AIC with a cross validation property (see Tong 1994; Stone 1977). Akaike's predictive log-likelihood is

$$E_p[\log(\mathcal{L}(\hat{\theta}))] = E_f E_f[\log(\mathcal{L}(\hat{\theta}_y)|x)].$$

Thus, $E_f E_f[\log(\mathcal{L}(\hat{\theta}_y)|\ x)]$ is the "target" of estimation; under certain conditions, $\log(\mathcal{L}(\hat{\theta})) - K$ is an estimator of this target when sample size is large (asymptotically). The expectation over both the data x and the estimated parameters $\hat{\theta}$ are taken with respect to the true $f(x)$. This expectation addresses the technical issue of parameter uncertainty. Zucchini (2000) provides a nice introduction to model selection using a well chosen example that helps understanding. Konishi and Kitagawa (2007) provide a technical review of these issues and introduce another extension.

It is not easy to see why including a great many models in the candidate set is poor practice. One sidesteps this issue if they concentrate on science hypotheses first, and then think hard about a good model to represent each hypothesis. It is the availability of software to "run practically everything in sight" that leads to this confusing issue. Zucchini (2000) provides several figures to illustrate the dangers of evaluating an excessive number of models.

Some software packages offer a "stepwise AIC" as an option for model selection (often termed variable selection in regression analysis). This is hardly in

the spirit of the information-theoretic approach and I strongly recommend against it. Such *ad hoc* procedures strongly encourage an "all possible models" approach that seems counter to good science. Good science has to be more about hard thinking and developing what seem to be plausible hypotheses; then proceed to build models as a way to evaluate the strength of evidence for these *a priori* hypotheses.

Shannon Entropy

Claude Shannon, working during the 1940s, is often regarded as the father of information theory. Shannon (1948) justified entropy for discrete variables with discrete and finitely many outcomes as

$$H = -\sum P_i \log P_i,$$

where P_i is the probability of outcome i. He approached this by positing three conditions that information (in the context of probability) should satisfy. He then proved that H was the unique solution that satisfied the conditions. The entropy of a probability distribution is

$$H = -\int p(x) \log p(x) \mathrm{d}x,$$

where $p(x)$ denotes the probability density with respect to the measure dx. Ecologists toyed with computing entropies in the early 1970s (an endeavor that Shannon termed the "bandwagon" in an editorial in the *Transactions of Information Theory*). While hundreds of papers presented entropies in leading ecological journals, most people now believe that this avenue produced little of value. In actuality, the definition of information was designed to help communication engineers send messages, rather than to help people understand the *meaning* of messages.

Goldman (1953) considers information to be the *difference* between our uncertainty before and after receiving a message. In this thinking, information is not an absolute quantity as implied from H, but is seen as a *change* in uncertainty. Let q_i be the probability of the ith event before receiving the message and p_i be the revised probability after receipt of the message. The change in the uncertainty is

$$[-\log(q_i)] - [-\log(p_i)] = \log(p_i) - \log(q_i) = \log(p_i/q_i).$$

If the message received indicates that the ith event is certain, then $p_i = 1$ and $\log(p_i) = 0$, resulting in a change in information of $-\log(q_i)$. Jessop (1995) terms this "surprisal." Taking the expectation

$$E[\log(p_i / q_i)] = \sum p_i \log(p_i / q_i)$$

and is the discrete version of K–L information! Kullback–Leibler information is an extension of Shannon's contribution and is sometimes called a "relative entropy" (Hobson and Cheng (1973)). The K–L information between models (probability distributions) is a *fundamental quantity* in science and information theory and is the logical basis for model selection.

Boltzmann's Entropy

Ludwig Boltzmann, working in the late 1800s, originally defined entropy in thermodynamics, demonstrated the second law of thermodynamics (e.g., there could not be a perpetual motion machine), and proved the irreversibility of entropy. Entropy is "disorder," max entropy is maximum disorder or minimum information. While the theory of entropy is a large subject by itself, readers here can think of entropy as nearly synonymous with uncertainty.

Conceptually, Boltzmann's entropy is $-\log(f(x)/g(x))$ and taking its expectation one gets

$$E_f\left(-\log\left(\frac{f(x)}{g(x)}\right)\right)=\int f(x)\log\left(\frac{f(x)}{g(x)}\right)\mathrm{d}x,$$

which is K–L information (see Good 1979). It is fascinating that Kullback–Leibler information is equal to the negative of Ludwig Boltzmann's entropy. Thus, minimizing the K–L information or distance is equivalent to maximizing the entropy; hence the name *maximum entropy principle* (Jaynes 1957).

Maximizing entropy is subject to a constraint – the model of the information in the data. A good model contains the information in the data, leaving only "noise." It is the noise (entropy or uncertainty) that is maximized under the concept of the *Entropy Maximization Principle*. Minimizing K–L information then results in an approximating model that loses a minimum amount of information in the data. Entropy maximization results in a model that maximizes the uncertainty, leaving only information (the model) "maximally" justified by the data. The concepts are equivalent, but minimizing K–L distance (or information loss) certainly seems the more direct approach. In summary,

– entropy = K – L information

and K–L information is often referred to as negative entropy or negentropy.

Boltzmann's discoveries concerning entropy are seen as the zenith of nineteenth century science. Of course, K–L information was derived along very different lines than entropy; the mutual convergence is striking and suggests something very fundamental. K–L information is *averaged* entropy, hence

the expectation with respect to f. Then, $-E$(entropy) = K–L information. Boltzmann derived the fundamental theorem that,

entropy is proportional to log(probability).

Entropy, information, and probability are thus linked, allowing probabilities to be multiplicative while information and entropies are additive.

3.12 Exercises

1. Cox (2006:117) states, "The relevance of automatic model selection depends strongly on the objectives of the analysis, for example as to whether it is for explanation or for empirical prediction." By "automatic model selection" I think he means criteria such as AICc, BIC, TIC, etc. Can examples be found where an investigator might need to use, say, AICc for prediction, but another criteria (or an entirely different approach?) for explanation (given the data are fixed)? What theory might bear on his statement? What practical advice might be given as to how to approach model selection when the main objectives of the analysis might vary? Discuss this with colleagues and see if the premise has merit.
2. In ecology increased diversity is often associated with ecotones. In a sense, Akaike was at a science ecotone when he saw a way to relate information theory and statistical theory in his AIC. Can you think of other parallels of this nature? What might this say about coursework to be taken by an exceptional Ph.D. student?
3. Akaike found an analytic expression for the asymptotic bias when the maximized log($\mathcal{L}$) was used as an estimator of expected K–L information; this bias correction was simply K, the number of estimated parameters in the model. Give other examples of estimators in your field where bias adjustments have been found.
4. AICc is simple to compute and understand, but it rests on very deep statistical theory. This makes it an ideal science tool. Give other examples where this is the case.
5. The data on hardening of Portland cement had four predictor variables; this leads to $2^4-1 = 15$ models. If all 2- and 3-way interactions would have been added, how many models would there be? What is the danger here in focusing on the models during data analysis?
6. Traditional statistics provided judgments about "significance" and this is related to some predefined, but arbitrary α-level. Such terms and dichotomies are shunned under the information-theoretic approach. Discuss and attempt to reconcile your thoughts on this matter of fixed dichotomies.

7. Examine a recent issue of a journal in your field of interest. Can you find a well written paper that carefully sets out several working hypotheses before data analysis? In some subdisciplines, such papers can be easily found. Once having found such a paper, what approach did the authors use as a measure of "strength of evidence" for and against the science hypotheses?
8. Atmar (2001) wrote a fitting obituary of Claude Shannon that makes interesting reading. He also references Dawkins (1986:111–112):

   A few years ago, if you asked almost any biologist what was special about living things as opposed to nonliving things, he would have told you about a special substance called protoplasm. Protoplasm wasn't like any other substance; it was vital, vibrant, throbbing, pulsating, "irritable" (a schoolmarmish way of saying responsive).… When I was a school boy, elderly textbook authors still wrote of protoplasm, although, by then, they really should have known better. Nowadays you never hear or see the word. It is as dead as phlogiston and the universal aether. There is nothing special about the substances from which living things are made. Living things are collections of molecules, like everything else. What is special is that these molecules are put together in much more complicated patterns than the molecules of nonliving things, and this putting together is done by following programs, sets of instructions for how to develop, which the organisms carry around inside themselves. Maybe they do vibrate and throb and pulsate with "irritability," and glow with living warmth, but these properties all emerge incidentally. What lies at the heart of every living thing is not a fire, not a warm breath, not a "spark of life." It is information, words, instructions. It you want a metaphor, don't think of fires and sparks and breath. Think, instead, of a billion discrete, digital characters carved in tablets of crystal. If you want to understand life, don't think about vibrant, throbbing gels and oozes, think about information technology.

   This thinking is certainly exciting – evolution and life are about information! Think hard about this and discuss it with colleagues and instructors. Is evolution so much about information? Where might these concepts lead us in the life sciences?

9. Assume you have some data on a well-defined science issue and the models for the four hypotheses are complimentary log–log models for a binary response variable. You have $n = 19$ and the global model has $K = 6$ parameters and AIC has been used as the first step in providing measures of strength of evidence for the four hypotheses. What is the issue that might be of concern here? Why?
10. Your new student questions the concern about models with "too many" parameters that must be estimated from the data. You speak of overfitting but he insists that biology is complex and some simple models are not "realistic." Prepare a clear response to help him understand this issue.

11. Recompute the information at the beginning of Sect. 3.7 using AIC. Provide your interpretation of any differences you encounter. What is the "moral" of this example?
12. You have just been hired by a government laboratory that has access to a very large amount of data from a Superfund site in Georgia. The questions were well formed, data collection was quite sophisticated, and sample sizes were very large by any usual standard. You are to work in a team situation and the team members have been educated and experienced in a variety of relevant disciplines. Some members of the team want to do an analysis using AIC, while others have heard about TIC and they favor this approach. They look to you for advice and council. What do you tell them? Why?
13. The bovine tuberculosis study by Caley and Hone (Sect. 3.8) is interesting in many ways. For example, they collected data by gender (also across five sites) and gender was a variable in all their models. A reviewer with expertise in *mustelids* claims that gender is unimportant in disease transmission and should not have been in the models (for parsimony reasons, if no other). Using AICc, how could you determine if the deletion of gender was better than models including gender? Be specific but concise.