5

Multimodel Inference

Hirotugu Akaike was born in 1927 in Fujinomiya-shi, Shizuoka-jen in Japan. He received B.S. and D.S. degrees in mathematics from the University of Tokyo in 1952 and 1961, respectively. He worked at the Institute of Statistical Mathematics for over 30 years, becoming its Director General in 1982. He has received many awards, prizes, and honors for his work in theoretical and applied statistics (de Leeuw 1992; Parzen 1994). This list includes the Asahi Prize, the Japanese Medal with Purple Ribbon, the Japan Statistical Society Award, and the 2006 Kyoto Prize. The three volume set entitled "*Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*" (Bozdogan 1994) was to commemorate Professor Hirotugu Akaike's 65th birthday. He is currently a Professor Emeritus at the Institute, a position he has held since 1994 and he received the Kyoto Prize in Basic Science in March, 2007.

For many years it seemed logical to somehow select the best model from an *a priori* set (but many people ran "all possible models") and make inductive inferences from that best model. This approach has been the goal, for example, in regression analysis using AIC, Mallows' (1973) C_p or step-up, step-down, or stepwise methods. Making inferences from the estimated best model seems logical and has served science for the past 50 years.

It turns out that a better approach is to make inferences from all the science hypotheses and their associated models in an *a priori* set. Upon deeper reflection, perhaps it is not surprising that one might want to make inferences from all the models in the set, but it is very surprising that this expanded

approach is so conceptually simple and easy to compute. Approaches to making formal inferences from all (or, at least, many) models is termed *multimodel inference*.

Multimodel inference is a fairly new issue in the statistical sciences and one can expect further advances in the coming years. At this time, there are four aspects to multimodel inference: model averaging, unconditional variances, gauging the relative importance of variables, and confidence sets on models.

5.1 Model Averaging

There are theoretical reasons to consider multimodel inference in general and model averaging in particular. I will not review these technicalities and instead offer some conceptual insights on this interesting and effective approach.

First, it becomes clear from the model probabilities (w_i) and the model likelihoods ($\mathcal{L}(g_i|\text{data})$) that there is relevant *information* in models ranked below the (estimated) best model. Although there are exceptions (e.g., Flather's data in Sect. 3.9.6), there is often a substantial amount of information in models ranked second, third, etc. (e.g., the dipper data in Table 4.2). If there is information in the model ranked second, why not attempt to use it in making inferences? Multimodel inference captures this information by using all, or several of, the models in the set.

Second, most models in the life sciences are far from full reality. With fairly adequate sample sizes, we might hope to find first- and perhaps second-order effects and maybe low-order interactions. Our models are often only crude approximations: unable to get at the countless smaller effects, nonlinearities, measured covariates, and higher order interactions. For example, even the best of the Flather models with an R^2 value of 0.999 is hardly close to full reality for the system he studied (i.e., dozens of species of birds, across several states and years, with data taken from a wide variety of observers). So, we might ask why we want to base the entire inference on the (estimated) best model when there is usually uncertainty in the selection as to the "best" model. Instead, perhaps inference should be based on a "cloud" of models, weighted in some way such that better models have more influence than models that are relatively poor. Here, rough classification of "good" and "poor" models is empirically based using model probabilities and model likelihoods. These lines of reasoning lead to what has been termed *model averaging*. Although the general notion has been in the statistical literature for many years, Bayesians have championed this approach in the past 10–15 years; however, their approach is computationally quite different.

Third, *ad hoc* rules such as "Δ_i greater than 2" become obsolete as all the models are used for inference. A model-averaged regression equation attempts to display the smaller effects and one can judge their usefulness against the science question of interest.

5.1.1 Model Averaging for Prediction

The best way to understand model averaging is in prediction. Consider the case where one has four well-defined science hypotheses ($H_1,\ldots, H_4$) and these are represented by four regression models ($g_1,\ldots, g_4$). The data are analyzed using standard least squares methods to obtain the parameter estimates ($\hat{\underline{\beta}}$, $\hat{\sigma}^2$) and their covariance matrix ($\hat{\Sigma}$). For a given set of values of the predictor values (say, $\mathbf{x_1 = 4.1, x_2 = 3.3, x_3 = 0.87, and\ x_4 = -4.5}$), one can use any of the models to make a prediction ($\hat{Y}$) of the response variable Y, for example,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1\left(x_1\right) + \hat{\beta}_3\left(x_3\right) + \hat{\beta}_4\left(x_4\right)$$

specifically,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1\left(4.1\right) + \hat{\beta}_3\left(0.87\right) + \hat{\beta}_4\left(-4.5\right)$$

using the values given above (note that this model does not include x_2). Of course, the MLEs of the β_j would have to be available before this model could be used for prediction of the response variable (Y).

Assume that this is the best model and has a model probability of 0.66. The second-best model excludes x_1 and is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_3\left(x_3\right) + \hat{\beta}_4\left(x_4\right) \quad \text{or} \quad = \hat{\beta}_0 + \hat{\beta}_3\left(0.87\right) + \hat{\beta}_4\left(-4.5\right)$$

and has a model probability of 0.27. The β_i parameters differ somewhat by model; that is, β_2 in the models above are different. I trust that notation to make this difference explicit is not needed as it would clutter the simple issue (i.e., the numerical value of $\hat{\beta}_3$ in the best model is almost surely different from the value of $\hat{\beta}_3$ in the second model). These and the other two models are summarized by rank as

Model i	$\hat{Y}_i$	Model probability w_i
1	67.0	0.660
2	51.7	0.270
3	54.5	0.070
4	71.1	<0.001

The predicted values can vary substantially by model; the β_j typically vary less from model to model. Prediction based on the best model might be risky as it has only two-thirds of the model probability and prediction from this model is somewhat higher than the other two models having the remaining one-third of the weight. Thus, the urge to make inference concerning predicted values using all four models (multimodel inference).

In this case, model-averaged prediction is a simple sum of the model probability for model i (w_i) times the predicted value for model i (Y_i)

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

where $\hat{\bar{Y}}_i$ denotes a model-averaged prediction of the response variable Y and $R = 4$ in this example. This is merely

$$0.660 \times 67.0 + 0.270 \times 51.7 + 0.070 \times 54.5 + 0.000 \times 71.1 = 61.994$$

or 62 for practical purposes. The model-averaged prediction downweights the prediction from the (estimated) best model for the fact that the other two models provide lower predicted values and they carry one-third of the weight. The fourth model has essentially no weight and does not contribute to the model-averaged prediction. Model-averaging prediction is trivial to compute; it is a simple weighted average. In the past, people did not know what to use for the weights and an unweighted average made little sense. The proper weights are just the model probabilities and $\hat{\bar{Y}}$ is fairly robust to slight differences in the weights. For example, bootstrapped weights (see Burnham and Anderson 2002:90–94) or weights from BIC (Appendix E) are different somewhat from those defined here; however, these differences often make relatively little change in the model-averaged estimates).

Model Averaging Predictions

Model averaging for prediction is merely a weighted average of the predictions ($\hat{Y}$) from each of the R models.

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

where the subscript i is over models. The weights are the model probabilities (w_i) and the predicted values (Y_i) from each model are given a particular setting of the predictor variables.

Typically, all the models in the set are used in such averaging. If one has good reason to delete one or more models, then the model weights must be renormalized such that they sum to 1.

I have been asked if some of the poorest models have "bad" information and, thus, should be excluded in the averaging. Lacking absolute proof, I have recommended against this. For one thing, if a model is that poor, it receives virtually no weight. I generally prefer using all the models in the *a priori* set in model averaging as this keeps unwanted subjectivity out of the process. Prediction is the ideal way to understand model averaging because each model can always be made to make an estimate of the response variable, given settings of the predictor variables.

5.1.2 Model Averaging Parameter Estimates Across Models

There are many cases where one or more model parameters are of special interest and each of the models has the parameter(s). In such cases, model averaging may provide a robust estimate.

Model Averaging Parameters within Models

An obvious approach to gain robust estimates of model parameters is to model average the estimates ($\hat{\theta}$) from each of the models *i*, where i = 1, 2, ..., R,

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

where $\bar{\hat{\theta}}$ is the model averaged estimate of θ (θ is used to denote some generic parameter of interest)

I will use the data on European dippers from Sect. 4.8 to illustrate averaging of estimates of model parameters. I found the MLEs and model probabilities for the two models to be the following:

Model	MLE	Model probabilities w_i
$\{\phi,p\}$	0.5602	0.08681
$\{\phi_n,\phi_f,p\}$ (n)	0.6071	0.91319
(f)	0.4688	

Model averaging for the survival probability in normal years is just

$$\bar{\hat{\phi}} = \sum_{i=1}^{2} w_i \hat{\phi}_i,$$

or

$$0.08681 \times 0.5602 + 0.91319 \times 0.6071 = 0.6030,$$

and the model-averaged estimate of the survival probability in flood years is

$$0.08681 \times 0.5602 + 0.91319 \times 0.4688 = 0.4767.$$

Note that the MLE for both flood and nonflood years in model $\{\phi,p\}$ is merely $\phi = 0.5602$, because this model does not recognize a flood year as being different from any other year. In this particular example, the estimates changed little because the best model had nearly all of the weight (91%). If the best model has a high weight (e.g., 0.90 or 0.95), then model averaging will make little difference as the other models contribute little to the average because they have virtually no weight (an exception would be the case where the estimates for models with little weight are very different than the other estimates). In other words, there is relatively little model selection uncertainty in this case and the model-averaged estimate is similar to the numerical value from the best model.

There are many cases where the parameter of interest does not appear in all the models and this important case is more complicated and is the subject of Sect. 6.3. Although not as easy to conceptualize or compute, there are ways to cope with the difficult issue of model selection bias (Sect. 6.3).

The investigator must decide if model averaging makes sense in a particular application. In general, I highly recommend model averaging in cases

where there is special interest in a parameter or subset of parameters. It should be clear that the parameter being considered from model averaging must mean the same thing in each model. Thus, if there is special interest in a disease transmission probability λ, then this parameter must be consistent in its meaning across models.

The models used by Flather have parameters a, b, c, and d but these mean very different things from one model to another. For example, consider three of his models:

$$E(Y) = ax^b,$$

$$E(Y) = a + \log(x), \text{and}$$

$$E(Y) = a\left(1 - [1 + (x/c)^d]^{-b}\right).$$

The parameter a, for instance, has quite different roles in these three models and it makes no sense to try to average the estimates $\hat{a}$ across models.

5.2 Unconditional Variances

Sampling variances are a measure of precision or repeatability of an estimator, *given* the model and the sample size. In the model above

$$E(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_3(x_3) + \hat{\beta}_4(x_4).$$

The variance of the estimator $\hat{\beta}_3$ is said to be "conditional" on the model. In other words, *given* this model, one can obtain a numerical value for var($\hat{\beta}_3$) using least squares or maximum likelihood theory. The fact is, another variance component should be included – a variance component due to *model selection uncertainty*. In the real world, we are not given *the* model and thus we cannot condition on the single best model resulting from data based selection. Instead, we use information-theoretic criteria or Mallows *Cp* or *ad hoc* approaches such as stepwise regression to *select* a model from a set of models. Usually, there is uncertainty as to the best model selected in some manner. That is, if we had other replicate data sets of the same size and from the same process, we would often find a variety of models that are "best" across the replicate data sets. This is called *model selection uncertainty*." Model selection uncertainty arises from the fact that the data are used both to select a model and to estimate its parameters. Much has been written about problems that arise from this joint use of the data. The information-theoretic approaches have at least partially resolved these issues (the same can be said for several Bayesian approaches).

Thus, a proper measure of precision or repeatability of the estimator $\hat{\beta}_3$ must include both the usual sampling variability (i.e., given the model) and a measure of model selection uncertainty (i.e., the model to model variability in the estimates of $\hat{\beta}_3$). This issue has been known for many years; statisticians have noted that using just the sampling variance, given (i.e., "conditional on") the model represents a "quiet scandal." Ideally,

$$\begin{aligned}\text{var}\left(\hat{\theta}\right) &= \text{sampling variance given a model} + \\ &\quad \text{variation due to model selection uncertainty.} \\ &= \text{var}\left(\hat{\theta}_i|\text{g}_i\right) + \sum\left(\hat{\theta}_i - \bar{\hat{\theta}}\right)^2.\end{aligned}$$

The last term captures the variation in the estimates of a particular parameter θ across models. If estimates of θ vary little from one model to another, this term is small and the main variance component is the usual one, var($\hat{\theta} \mid g$), which is the variance of $\hat{\theta}$, conditional (given) on the model g_i. However, as is often the case, $\hat{\theta}$ varies substantially among models and if this term is omitted, the estimated variance is too small, the precision is overestimated, and the confidence intervals are too narrow.

Building on these heuristics, we can think of the above equation and take a weighted sum across models as

$$\mathbf{var}\left(\bar{\hat{\boldsymbol{\theta}}}\right) = \sum_{i=1}^{R} \boldsymbol{w}_i \left\{ \mathbf{var}\left(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{g}_i\right) + \left(\hat{\boldsymbol{\theta}}_i - \bar{\hat{\boldsymbol{\theta}}}\right)^2 \right\}$$

where the final term is the model-averaged estimator. This variance is called "unconditional" as it is not conditional on a *single* model; instead, it is conditional on the set of models considered (a weaker assumption). [Clearly, the word "unconditional" was poorly chosen; we just need to know what is meant by the term.] Note that model averaging arises as part of the theory for obtaining an estimator of the unconditional variance. The theory leading to this form has some Bayesian roots.

Note that the sum of the two variance components is weighted by the model probabilities. If the best model has, say, $w_{\text{best}} > 0.95$, then one might ignore the final variance component, $\left(\hat{\theta}_i - \bar{\hat{\theta}}\right)^2$, because model selection uncertainty is nil. In such cases, model averaging is not required and the usual conditional variance should suffice.

There are sometimes cases where the investigator wants to make inference about a parameter from only the best model; here an unconditional variance and associated confidence intervals could still be used with the expected advantages.

Unconditional Variance Estimator

An estimator of the variance of parameter estimates that incorporates both sampling variance, given a model, and a variance component for model selection uncertainty is

$$\mathbf{var}\left(\bar{\hat{\boldsymbol{\theta}}}\right) = \sum_{i=1}^{R} \boldsymbol{w}_i \left\{ \mathbf{var}\left(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{g}_i\right) + \left(\hat{\boldsymbol{\theta}}_i - \bar{\hat{\boldsymbol{\theta}}}\right)^2 \right\},$$

where $\bar{\hat{\theta}}$ is the model averaged estimate, w_i are the model probabilities, and g_i is the ith model. This expression is also useful in a case where one is interested in a proper variance for $\hat{\theta}$ from only the best model (not the model-averaged estimator).

Of course $\text{se}\left(\bar{\hat{\boldsymbol{\theta}}}\right) = \sqrt{\text{var}\left(\bar{\hat{\boldsymbol{\theta}}}\right)}$, and if the sample size is large, a 95% confidence interval can be approximated as $\bar{\hat{\theta}} \pm 1.96 \times \text{se}\,\bar{\hat{\theta}}$.

Use of this approach provides estimators with good achieved confidence interval coverage. I highly recommend this approach when data are used to both select a model and estimate its parameters (i.e., the usual case in applied data analysis).

5.2.1 Examples Using the Cement Hardening Data

We return to the cement hardening data to provide useful insights into model selection uncertainty for this small (n = 13) data set. Material presented in Sect. 4.5 indicated that only two models (models {12} and {12 1*2} had any discernible empirical support. This is almost too simple to illustrate the points I want to make clear; therefore, I will use an extended example using these data from Burnham and Anderson (2002:177–183).

Model averaging the β_i parameters and obtaining the unconditional variances was done using, strictly for illustration, all possible subsets, thus there are $2^4-1=15$ models. Clearly, model {12} is indicated as the best; however, substantial model selection uncertainty is evident because that best model has a model probability of only 0.567 (Table 5.1).

TABLE 5.1. Model probabilities for 15 nested models of the cement hardening data.

Model i	w_i	Model i	w_i
{12}	0.5670	{23}	0.0000[a]
{124}	0.1182	{4}	0.0000
{123}	0.1161	{2}	0.0000
{14}	0.1072	{24}	0.0000
{134}	0.0810	{1}	0.0000
{234}	0.0072	{13}	0.0000
{1234}	0.0029	{3}	0.0000
{34}	0.0004		

[a]Values shown are zero to at least five decimal places.

TABLE 5.2. Quantities used to compute the unconditional variance of the predicted value for the cement hardening data.

Model i	K	$\hat{Y}$	w_i	var($\hat{Y} \mid g_i$)	$(\hat{Y}_i - \bar{\hat{Y}})^2$
{12}	4	100.4	0.5670	0.536	1.503
{124}	5	102.2	0.1182	2.368	0.329
{123}	5	100.5	0.1161	0.503	1.268
{14}	4	105.2	0.1072	0.852	12.773
{134}	5	105.2	0.0810	0.643	12.773
{234}	5	111.9	0.0072	4.928	105.555
{1234}	6	101.6	0.0029	27.995	0.001
{34}	4	104.8	0.0004	1.971	10.074

The computation of unconditional estimates of precision for a predicted value is simple because every model i can be made to provide a prediction ($\hat{Y}$). We consider prediction where the predictor values are set at $\boldsymbol{x_1 = 10, x_2 = 50}$, $\boldsymbol{x_3 = 10, x_4 = 20}$. The prediction under each of the eight models is shown in Table 5.2 (the last seven models were dropped here as they have virtually no weight). Clearly, $\hat{Y}$ is high for model {234}, relative to the other models. The estimated standard error for model {1234} is very high, as might be expected because the X matrix is nearly singular. Both of these models have relatively little support, as reflected by the small model probabilities, and so the predicted values under these fitted models are of little credibility.

Table 5.2 predicts the response variable Y for the cement hardening data. $\hat{Y}$ is a predicted expected response based on the fitted model (given the predictor values, $x_1 = 10, x_2 = 50, x_3 = 10$, and $x_4 = 20$), conditional on the model. Measures of precision are given for $\hat{Y}$; $\hat{\bar{Y}}$ denotes a model-averaged predicted value, and $(\hat{Y}_i - \hat{\bar{Y}})^2$ is the model-to-model variance component when using model i to estimate Y. For example, for model {12} and $i = 1$, $(\hat{Y}_1 - \hat{\bar{Y}})^2$.

The remaining seven models had essentially no weight and are not shown.

The predicted value for the AICc-best model is 100.4 with an estimated conditional variance of 0.536. However, this measure of precision is an underestimate because the variance component due to model selection uncertainty has not been incorporated. Model averaging results in a predicted value of 101.6. The corresponding estimated unconditional standard error is 1.9. The unconditional standard error is substantially larger than the conditional standard error 0.73. Although it is not a virtue that the unconditional standard error is larger than has been used traditionally, it is a more honest reflection of the actual precision. If only a conditional standard error is used, then confidence intervals are too narrow and achieved coverage will often be substantially less than the nominal level (e.g., 95%).

Study of the final two columns in Table 5.2 shows that the variation in the model-specific predictions (i.e., $\hat{Y}$) from the weighted mean (i.e., $(\hat{Y}_i - \hat{\bar{Y}})^2$) is substantial relative to the estimated variation, conditional on the model (i.e., the var($\hat{Y} \mid g_i$)). Models {124} and {1234} are exceptions in this case.

The investigator has the choice as to whether to use the predicted value from the AICc-selected model (100.4) or a model-averaged prediction (101.6). In this example, the differences in predicted values are small relative to the unconditional standard errors (1.9); thus, here the choice makes little difference. However, there is often considerable model uncertainty associated with real data and I would suggest the use of the model-averaged predictions. Thus, I would use 101.6 as the predicted value with an unconditional standard error of 1.9. If the best model was more strongly supported by the data (i.e., $w_i > 0.95$), then I might suggest the use of the prediction based on that (best) model (i.e., $\hat{Y}$=100.4) and use the estimate of the unconditional standard error (1.9).

TABLE 5.3. Quantities needed to make multimodel inference concerning the parameter β_1 for the cement hardening data.

Model i	$\hat{\beta}_1$	var($\hat{\beta}_1 \mid g_i$)	w_i
{12}	1.4683	0.01471	0.5670
{124}	1.4519	0.01369	0.1182
{123}	1.6959	0.04186	0.1161
{14}	1.4400	0.01915	0.1072
{134}	1.0519	0.05004	0.0811
{1234}	1.5511	0.55472	0.0029
{1}	1.8687	0.27710	<0.0001
{13}	2.3125	0.92122	<0.0001

Model {1} has only x_1 in it, allowing a "straight shot" at $\hat{\beta}_1$; however, it is a very poor model, relative to the others. When other variables are included in a model with x_1, the parameter estimates change due to the fact that the x_1 is correlated with the other predictor variables (see Table 5.3). This is frustrating as the biologist wants a simple answer to this simple question. So, what is a robust estimate of the regression slope on x_1? Model averaging is one approach to answer this question.

The computation of the model-averaged estimates of β_1 and its unconditional sampling variance is illustrated in Table 5.3 (recall that the variance is the square of the standard error).

The other seven models do not have β_1 in them and are not shown above.

Note the variation in the estimates of β_1 across models – this model-to-model variation is model selection uncertainty and needs to be reflected in estimates of precision. The model-averaged estimate,

$$\hat{\bar{\beta}}_1 = \sum_{i=1}^{R} w_i \hat{\beta}_{1i} = 1.4561, \text{ while the estimate of } \beta_1 \text{ from model } \{1\} \text{ is } 1.8687.$$

The unconditional variance of this model-averaged estimate of β_1 is obtained by using the formula above, expressed in terms of β_1 (instead of the generic parameter θ),

$$\text{var}\left(\hat{\bar{\beta}}_1\right) = \sum_{i=1}^{R} w_i \left\{ \text{var}\left(\hat{\beta}_{1i} \mid g_i\right) + \left(\hat{\beta}_{1i} - \hat{\bar{\beta}}_1\right)^2 \right\},$$

where i reflects the model and "1" is the parameter of interest, β_1. For example, the first term in the needed sum is

$$w_1 \times \left\{ \hat{\text{v}}\text{ar}\left(\beta_{11} \mid g_1\right) + \left(\hat{\beta}_{11} - \hat{\bar{\beta}}_1\right)^2 \right\}$$

$$0.5670 \times \{(0.0147) + (0.0122)^2\} = 0.00842.$$

Note that the final term is 1.4683 – 1.4561 = 0.0122; then this is squared. Completing the rest of the calculations, we get var($\hat{\beta}_1$) = 0.0308, or an estimated unconditional standard error on $\hat{\bar{\beta}}_1$ of 0.1755. This compares to the

TABLE 5.4. Quantities needed to make multimodel inference concerning the parameter β_2 for the cement hardening data.

Model	$\hat{\beta}_2$	$\text{se}(\hat{\beta}_2 \mid g_i)$	w_i
{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	<0.0001
{24}	0.3109	0.7486	<0.0001
{2}	0.7891	0.1684	<0.0001

conditional standard error of 0.1213, given the selected model. This difference is the "quiet scandal" because model selection uncertainty (a component of variance) had been omitted (ignored) in the conditional approach.

Model selection uncertainty is more clearly present when examining the model-to-model variation in the estimator of β_2 (Table 5.4). Note that the estimate of β_2 for model {234} is negative; this is due to the high correlations among the predictor variables.

The Akaike weights (w_i) for just the eight models above add to 0.8114. However, to compute results relevant to just these eight models, we must renormalize the relevant model probabilities to add to 1. Those renormalization probabilities are given in Table 5.4. The model-averaged estimator of β_2 is 0.6110 (compared to 0.6623 for the best model and 0.7891 for model {2}) and the unconditional estimated standard error of $\bar{\hat{\beta}}_2$ is 0.1206. The conditional standard error for the best model was 0.0459, while the unconditional standard error was 0.1206 – reflecting the high degree of model selection uncertainty. This estimate attempts to provide a robust estimate of the "slope" on the variable x_2 without regard to other variables in the model.

It is important to compute and use unconditional standard errors in inferences after data based model selection. Note also that confidence intervals on β_1 and β_2, using results from model {12}, should be constructed based on a t-statistic with 10 df ($t_{10,0.975} = 2.228$ for a two-sided 95% confidence interval). Such intervals here will still be bounded well away from 0; for example, the 95% interval for β_2 is 0.39–0.93.

5.2.2 *Averaging Detection Probability Parameters in Occupancy Models*

Ball et al. (2005) used occupancy models (MacKenzie et al. 2006) to evaluate a habitat model for the Palm Springs ground squirrel (*Spermophilis tereticaudus chlorus*) in the Coachella Valley, California, in 2002. I will use this paper to illustrate several things, in addition to model averaging and unconditional variances.

Ball et al. (2005) were interested in the evaluation of a habitat model developed by a Scientific Advisory Committee (SAC) as part of a multispecies Habitat Conservation Plan. Interest was focused on mesquite (*Prosopis glandulosa*), which was not common, thought to be excellent habitat, and in decline, while creosote (*Larrea tridentata*) was much more common but of somewhat questionable value compared to the squirrel.

Ball et al. (2005) developed 15 models based on squirrel occupancy (ψ) and squirrel detectability (p). They justified their choice of *a priori* hypotheses based largely on the literature, knowledge of the biology of the species, and on management concerns and this required considerable thought (Doherty, personal communication). A pilot study was conducted and these results used to improve survey design and data-gathering protocols. Four preserves were sampled using systematic samples with random starts. Over 1,900 points were sampled in the initial field session. Occupancy was modeled as a constant ($\cdot$) or as a function of individual vegetation (mesquite, creosote, or desert scrub) or substrate types (dune and hummock) or combinations. Detection probability was modeled as a function of these same variables and sampling time (t); sampling was done on three occasions ($t = 1, 2, 3$). Our entree into this issue will be Table 5.5 where 15 models are summarized. One thing to notice at first glance is that virtually 100% of the model probability is tied to just three models and the science hypotheses that they represent.

Estimates of the constant detection probability from the best model was 0.216 but varied from 0.156 to 0.256 for other models in the set. Model averaging was done and the results are shown in Table 5.6.

TABLE 5.5. Model selection statistics for 15 models of ground squirrel occupancy (ψ) and detection probability (p) from Ball et al. (2005). The models are shown by rank.

	Hypothesis/model	log(L)	AICc	K	Δ_i	w_i
1	$\psi_{\text{mdh,cdh,sdh,ae}}, p_{\cdot}$	–229.27	468.56	5	0.00	0.58
2	$\psi_{\text{mdh,cdh,sdh,ae}}, p_{\text{mdh = cdh,ae}}$	–229.11	470.26	6	1.70	0.25
3	$\psi_{\text{mdh,cdh,sdh,ae}}, p_{t}$	–228.49	471.05	7	2.48	0.17
4	$\psi_{\text{mdh = cdh,sdh,ae}}, p_{\cdot}$	–235.32	478.66	4	10.09	<0.01
5	$\psi_{\text{mdh = cdh,ae}}, p_{\cdot}$	–236.97	479.96	3	11.40	<0.01
6	$\psi_{\text{m,}dist\text{om}}, p_{\cdot}$	–236.00	480.01	4	11.45	<0.01
7	$\psi_{\text{mdh = cdh,sdh,ae}}, p_{\text{mdh = cdh,ae}}$	–235.31	480.65	5	12.09	<0.01
8	$\psi_{\text{mdh = cdh,sdh,ae}}, p_{t}$	–234.93	481.91	6	13.35	<0.01
9	$\psi_{\text{mdh = cdh,sdh,ae}}, p_{\text{mdh = cdh,ae}}$	–236.96	481.95	4	13.38	<0.01
10	$\psi_{dist\text{om}}, p_{t}$	–235.25	482.54	6	13.98	<0.01
11	$\psi_{\text{mdh = cdh,ae}}, p_{t}$	–236.57	483.17	5	14.60	<0.01
12	$\psi_{\cdot}, p_{\text{mdh = cdh,ae}}$	–242.19	490.17	3	21.83	<0.01
13	$\psi_{\cdot}, p_{\cdot}$	–278.61	561.23	2	92.67	<0.01
14	$\psi_{\cdot}, p_{t}$	–278.36	564.74	4	96.18	<0.01
15	$\psi_{\text{m.distom.}}, p_{\text{mdh = cdh.ae}}$	–279.66	569.34	5	100.78	<0.01

M mesquite; *C* creosote; *D* dune; *H* hummock; *S* desert scrub (e.g., *Atriplex* spp.); *AE* all else; *DisToM* distance to mesquite.

TABLE 5.6. Summary of model averaging for detection probability for the ground squirrel data from Ball et al. (2005).

	Estimated		
Model number	Model probability	Detection probability	Standard error
1	0.576	0.216	0.044
2	0.247	0.228	0.051
3	0.166	0.256	0.064
4	0.004	0.161	0.050
5	0.002	0.162	0.050
6	0.002	0.195	0.040
7	0.001	0.156	0.061
8	0.001	0.178	0.062
9	0.001	0.156	0.061
10	0.001	0.231	0.058
11	<0.001	0.180	0.063
12[a]	<0.001	0.209	0.063
Weighted average		0.225[b]	0.049[c]
Unconditional standard error[d]			0.052

[a]The remaining models had virtually no weight and are not shown. The results are shown to only three places.

[b]The weighted average was based on

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

[c]The first entry is a weighted average of the conditional standard errors, while the unconditional standard error includes a variance component for model selection uncertainty.

[d]The unconditional standard error was based on

$$\text{var}\left(\hat{\bar{p}}\right) = \sum_{i=1}^{R} w_i \left\{ \text{var}\left(\hat{p}_i \mid g_i\right) + \left(\hat{p}_i - \hat{\bar{p}}\right)^2 \right\}.$$

Approximately 11% of the variation stems from model selection uncertainty and is a small proportion in this example. Note that the weighted average of the conditional standard errors (0.049) is larger than the conditional standard error for the best model (0.044).

Careful examination of the log-likelihood values suggests that model 2 is a good model only because it has one additional parameter (thus a "penalty term" of approximately 2); however, the fit did not improve. That is, the log-likelihood value for the best model was −229.27, whereas this value for the second-best model was −229.11. This finding leads to the conclusion that the structure on the detection probability ($p_{\text{MDH=CDH,AE}}$) is without support. The two estimates of detection probability under the second-best model are similar (0.228 vs. 0.170) and their confidence intervals overlap entirely (0.144–0.342 vs. 0.059–0.403). This is an example of a "pretending variable" (see Sect. 3.6.8). One should check to be sure that there has been a change in the log-likelihood values to avoid the "pretending variable problem."

If model 2 is removed from the set for some *post hoc* reason, the model probabilities for the first- (1) and the second-best (formerly 3) models change to 0.764 and 0.221, respectively, from 0.576 and 0.166 (see Sect. 3.7.1). One should always examine the log-likelihood or deviance to be sure that the addition of a new parameter or variable improves the fit.

Now we consider the issue of vegetation and substrate type on the occupancy parameter ψ: variables MDH and CDH. Model 3 allows these variables to operate independently, whereas model 8 enforces the equality constraint, MDH = CDH and requires one less parameter to be estimated. All other structural aspects of these two models are the same. Evidence in issues such as this can be provided using an evidence ratio, $E_{3,8} = 0.28938/0.00126 = 230$. Thus, the evidence is strong (my value judgment) that the constraint represents a poor hypothesis. Such evidence ratios do not depend on other models in or out of the set and are useful in contrasting two models that have differing parameterizations.

Because of the interest in mesquite, the relationship between occupancy and the distance to the nearest mesquite was quantified. This was done by computing the evidence ratio between models 6 and 13 and this showed substantial evidence in favor of a relationship ($E_{6,13} = 118$ divided by essentially 0). The estimated slope of the relationship from model 6 was –0.00075 with se = 0.00014 and 95% confidence interval of (–0.0010, –0.00048), further confirming the importance of mesquite to this species of ground squirrel. [Do not be fooled by the low numerical value of the estimated slope (–0.00075). It is very small because its associated variable was large (a distance). The importance of this variable is revealed by the relatively small standard error of 0.00014 and the fact that the coefficient of variation is 19%.]

A final aspect of Ball et al.'s (2005) work was to examine the evidence for the SAC habitat model that had been proposed for the management of this ground squirrel. Model $\{\psi, p\}$ most closely represented the proposed habitat SAC model; however, it was ranked third to the last with a model probability of $e^{92.67/2} = 1.33 \times 10^{-20}$. One must conclude that the SAC model was very poor as a basis for management decisions.

5.3 Relative Importance of Predictor Variables

In some cases, research is in an early descriptive stage and model selection and valid inference may be constrained by lack of knowledge, small sample size, high dimensionality of the predictor variables, high degree of multicolinearity, and high variability. In such cases, it may be judicious to gain insight into the more important variables from analyzing data from a pilot study and then attempt to collect high quality data on these (few) more important variables. Proper *a priori* hypothesizing and modeling might then focus better on a few variables thought to be important, rather than tackling data on all the variables. This seems like a useful way to approach exploratory data analysis.

5.3.1 Rationale for Ranking the Relative Importance of Predictor Variables

Consider a small team of researchers interested in both understanding relationships and making predictions about a response variable Y, based on measurements of 15 predictor variables ($x_1, x_2, \ldots, x_{15}$). There are, in this case, $2^{15} - 1 = 32{,}767$ possible models, excluding interactions or transformations such as quadratic terms. Rather than gearing up for a huge computer run, the team decides to generate a reasonable subset of variables that seem most important. Thus, an ability to rank the relative importance of the predictor variables might be useful. Then, further research could chase understanding and prediction based on a few variables that rank high. This is not the only approach to making scientific progress under the severe constraints noted, but it is an interesting alternative.

Such ranking can be done with ease if one has some experience with spreadsheets and has a statistical software package that can unthinkingly run "all possible models." In general, I do not recommend running all the models; this is a special case where every variable must be put on an equal footing with the rest for the ranking to be interpretable. Running all the models is an easy way to achieve the balance (fairness) in ranking the relative importance of the predictors. As with AICc model selection, there is no guarantee that any of the predictors are good in some absolute sense; we are merely going to rank them.

5.3.2 An Example Using the Cement Hardening Data

The cement hardening data will serve as a handy example with four predictor variables, thus $2^4 - 1 = 15$ possible models. Step 1 is to list all 15 models and their associated model probability (Table 5.7).

Ranking, step 2, is done by merely selecting all the models where x_i appears and summing up the associated model probabilities. Thus, let $i = 1$, then predictor variable x_1 appears in the following eight models: {1}, {12}, {13}, {14}, {123}, {124}, {134}, and {1234}. The sum of these eight model probabilities

TABLE 5.7. Summary of the model probabilities for the cement hardening data.

Model	Probability[a]	Model	Probability[a]
{1}	0.0000	{24}	0.0000
{2}	0.0000	{34}	0.0004
{3}	0.0000	{123}	0.1161
{4}	0.0000	{124}	0.1182
{12}	0.5670	{134}	0.0811
{13}	0.0000	{234}	0.0072
{14}	0.1072	{1234}	0.0029
{23}	0.0000		

[a]Shown to four decimal places.

is 0.9925. The process is repeated for the eight models where x_2 appears, namely models {2}, {12}, {23}, {24}, {123}, {124}, {234}, and {1234}.

The results for the four predictor variables are summarized as

Variable	Sum	Rank
1	0.9925	1
2	0.8114	2
3	0.2077	4
4	0.3170	3

In this small example, one might want to focus further work on the two predictors that are ranked high. Note that the use of all possible models gave each variable an equal footing; each variable was in exactly eight models and the sums were all based on eight entries. The method has utility even when R is fairly large (e.g., 20) as standard software can compute the models in a few hours. Then one needs to capture the relevant statistics, compute AICc, the Δ_i values, and model probabilities, and use a spreadsheet to compute the simple summations.

This ranking procedure will never see heavy use but it is a method worth knowing about when faced with exploratory phases of investigation where dimensionality is high. This ranking approach is most appealing for hypotheses that can be well represented by linear or logistic regression models. The ranking tries to break correlations between and among the predictor variables by having a variable appear on its own and then together with all the other variables. This is an opportunity to determine, via the model probabilities, which variables are related to the response variable and which variables appear to be related, but only because they are correlated with another predictor.

Ad hoc procedures such as stepwise regression give a false impression of "importance." Once the algorithm has stopped adding and deleting predictor variables, one might have the final fitted model (where there are 13 predictor variables),

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1(x_1) + \hat{\beta}_6(x_6) + \hat{\beta}_7(x_7) + \hat{\beta}_{12}(x_{12}).$$

Then, one is led to the (incorrect) conclusion that variables x_1, x_6, x_7, and x_{12} are "important" in terms of the response variable. One is compelled to believe that these variables *must* be important because, after all, they are in the final model. Conversely, the remaining variables, x_2, x_3, x_4, x_5, x_8, x_9, x_{10}, x_{11}, and x_{13} are *surely* not important; after all, if they had been important, they would have been in the final model. Because of the intercorrelations among the predictor variables, such seemingly obvious dichotomies are surprisingly false. Stepwise algorithms do not even identify the second-best model; in fact, no ranking of models is possible using these techniques. Insights such as the one above tend to motivate the use of models beyond just the one estimated to be the best one. There is often substantial *information* in models ranked 2, 3,…, R and it is easy to use this information to allow better inferences to be made from the evidence available. Model based inference ought to be about making inference from all the models in most scientific work. Ranking of the importance of predictor variables is one facet of multimodel inference.

5.4 Confidence Sets on Models

Bayesians define an interval in a manner in which they can assert that a single interval contains the parameter with a certain probability (e.g., 0.95). These are often called credible intervals and are easy to understand, whereas the frequentist confidence interval falls back on notions of repeated sampling and the long run coverage at the nominal level (e.g., 0.95). In the real world, there is often little numerical difference between the two types of intervals, but there are worthwhile philosophical differences.

Now consider a set of five science hypotheses represented by five models and a data set with sample size 145. Given this set and the fixed sample size, one of the five models is *the* best model in a K–L sense; we just do not know which one it is (much like an unknown parameter). We can estimate which model is best and the probability that each model i is that actual best model (the model probabilities w_i). This thinking leads the way to consider a confidence set on models.

Caley and Hone's (2002) data on bovine tuberculosis in ferrets can be used to illustrate the concept. From Sect. 4.6, we had

Hypothesis	Model	Model probability
H_1	g_1	<0.0001
H_2	g_2	<0.0001
H_3	g_3	<0.0001
H_4	g_4	0.7595
H_5	g_5	0.2405

Summation of the last two probabilities gives 1.0 in this example. We can say probabilistically that the best K–L model is either g_4 or g_5 (*given* the model set) with virtual certainty in this case. The model set (g_4, g_5) constitutes an approximate 100% confidence set on models. This concept is sometimes useful as an aid in comprehending the meaning of the evidence.

A second example comes from Linhart and Zucchini's (1986) book and deals with prediction and smoothing of weekly data on storm frequency at a botanical garden in Durban, South Africa. They used a combination of logistic regression and Fourier series terms to model weekly storm frequency across 47 consecutive years. The Fourier series terms enter in pairs (sines and cosines); thus, models are nested and K jumps by 2 from model to model. I will avoid other complications (e.g., overdispersion) and provide a summary of the quantities available:

Model i	Δ_i	Model probability
1	39.04	<0.0001
2	1.64	0.2141
3	**0.00**	**0.4861**
4	1.49	0.2305

(continued)

5	3.98	0.0665
6	10.59	0.0024
7	17.24	<0.0001
8	24.51	<0.0001
9	33.29	<0.0001

Model 3 with six parameters is the best with probability 0.4861 flanked by model 2 and model 4, with probabilities 0.2141 and 0.2305, respectively. The sum of these three probabilities gives 0.93, giving an approximate 95% confidence set. Including the probability for model 5 gives a 99.7% confidence set. These sets are only approximate but allow the notion of confidence intervals for parameters to be extended to confidence sets for models. Here, models other than 2, 3, and 4 lie outside a 93% confidence set. Clearly, models 1, 7, 8, and 9 lie well outside either set.

The use of confidence sets on models is occasionally useful, particularly when the models are nested. These sets can help understand subsets of models that have reasonable plausibility.

5.5 Summary

The idea of making formal inductive inferences from an array of *a priori* models is compelling. Given a choice of using one model where there is uncertainty concerning its rank and using all the models in the set, I think people would prefer the latter. Multimodel inference seems generally desirable. The curious thing is that multimodel inference is computationally easy. In the future, it seems likely that additional approaches will be developed to allow inference from multiple models.

5.6 Remarks

A good discussion of model averaging is given by Hoeting et al. (1999). Their paper is written in a Bayesian setting, but the review of the general approach is good reading.

Anthony et al. (2006) provide the results of an enormous research program on the Northern Spotted Owl (*Strix occidentalis caurina*). The analysis of these data was done under a multimodel inference paradigm due partly to the litigious nature of the long-term controversy over cutting old growth forests and its effects on spotted owls and other conservation concerns. Model averaging can have substantial values when the science issue involves controversy (see Hoeting et al. 1999; Anderson 2001).

Chatfield's (1995b) extensive paper is excellent on several important, but perhaps subtle, issues; in particular, problems that arise when using the data to both select a model and then make inferences from that selected model.

These concepts are largely handled by the information-theoretic approaches for many classes of problems.

Breiman (1992) offered the term "quiet scandal" when estimates of precision are presented without a variance component for model selection uncertainty.

Burnham and Anderson (2002:Chap. 5) provide the results from a number of MC simulation studies showing the poor confidence interval coverage of estimators when model selection uncertainty is ignored. They simulated binomial data (10,000 replicates) from a simple age-specific survival model with ten age classes with sample size of 150 subjects. The model set allowed estimates of survival probability up to some age, whereas the remaining age classes were pooled, as is often done when the number of survivors dwindles. Inference was made from the best model and confidence interval coverage was poor when only the sampling variance was used as a measure of precision: mean coverage was 81.3%, ranging from a low of 63.0% to 95.9%. In contrast, when a variance component for model selection uncertainty was added, coverage averaged 95.0%, ranging from 90.6% to 97.7%.

Various approaches to model selection began to appear in the technical literature since computers became available in the 1960s. Prior to that, one was happy to obtain the MLEs and covariance matrix for a single model as calculations were laborious and had to be done by hand. Procedures such as stepwise regression filled an important void and saw heavy use. Only in the past 15 years have people begun to ask about the statistical properties of the selected model (be it from stepwise, Mallows' *Cp*, AICc, or whatever). It became clear that the estimators used in the selected model had confidence interval coverage below the nominal level. This limitation was caused because model selection uncertainty was not embedded into estimates of precision (Chatfield (1995b) covers this issue and provides insights into problems with data dredging).

Statistical software packages could be much more useful if they treated sets of models, given a data set, rather than treating individual models in isolation. I am aware of only two major software packages that take this approach: program MARK (White and Burnham 1999) and Distance 5.2 (Thomas et al. 2006). Both of these packages are freeware; however, neither is general purpose statistical package.

Predictor variables in linear and nonlinear regression are often correlated and this has its consequences. The cement hardening data will serve as an example where the variables x_2 and x_4 had a correlation coefficient of –0.973. One advantage of using AICc is that both of the variables are retained in the analysis. Thus, models {12}, {14}, and {124} were three of the best fourmodels, with model probabilities of 0.567, 0.107, and 0.118, respectively. Although x_4 is not the best of the pair, perhaps this variable is very much less expensive to measure; thus, it should not be lost from the results. There are several ways to handle correlated variables, including a simple geometric mean of the members of the pair, thus reducing two variables to one. If several similar variables have high correlations, one can perform a principal components analysis (PCA) and hope that most of the variation is contained in the first 1–2 components; however, issues of interpretability often arise.

Ranking the relative importance of variables is a more sound way to try to identify important variables from a large set. In the past, people have used

some sort of statistical test to sequentially weed out "nonsignificant" variables and this approach has poor properties (the multiple testing problem to mention only one issue).

I hope the reader is gaining an appreciation for how bad *ad hoc* procedures such as stepwise regression can be, even in routine situations where the assumptions are fairly well met. Although stepwise methods are still being taught routinely, they are a poor basis for model based inference in a linear models setting (see McQuarrie and Tsai 1998).

Guidelines have been published outlining the quantities that should often appear in publications (Anderson et al. 2001b). In the ground squirrel example, the issue surrounding model 2 could not have been uncovered had the value of the log-likelihood (or the deviance) not been published.

5.7 Exercises

1. The first exercise in Chap. 4 dealt with the data in bill lengths in Darwin's finches. Would you employ model averaging the estimates of β_1 in this case? Why? Why not? Would you do any model averaging in this example? Should model selection uncertainty be incorporated into estimates of precision in this example?
2. Review Table 5.5 from the study of Palm Springs ground squirrels. Your colleague provides you with the evidence ratio $E_{5,8} = 2.65$. Write a concise paragraph explaining the biology implied by this result.
3. When faced with many predictor variables in linear or logistic regression one must often try to reduce the dimensionality by various means. One approach has been to perform a principal components analysis (PCA) on the X matrix. Then, the regression is on PCI, PC2,…, rather than on the original variables x_1, x_2,…. What are the advantages and disadvantages to this approach? (advanced question)
4. A nonparametric bootstrap might be used in model selection. Outline this approach in a one-page report and offer a critique. (advanced question)
5. We learned in Chap. 2 that information was additive. How might this fact be exploited using the Δ_i values? (advanced question)
6. Review the paper by van Buskirk and Arioli (2002) and consider ways in which their model set might evolve to the next level, given their results.
7. Bortz and Nelson (2006) studied HIV infection dynamics that surely gives some insight into modeling complex system using state-of-the-art quantitative methods. Readers with a background in various types of differential equations and mixed effects modeling should read this paper.

    a. What is gained by thinking that the "penalty term" in AIC, AICc, and TIC is a measure of "complexity"?
    b. They seem to favor an information criterion termed ICOMP. Can you determine the rationale for this choice?
    c. Is it not surprising that K is so small for the models they evaluate?. Why might this be?