# Ch. 4

# AIC and model selection[1]

***"There are only two mistakes one can make along the road to truth; not going all the way, and not starting."***
**-- Buddha**

***"If you are out to describe the truth, leave elegance to the tailor."***
**-- Albert Einstein**

## Questions to ponder:

- *What does a "model" describe, in the context of parameter estimation?*
- *Do we know which model is the best model to use?*
- *What is AIC, and how is AIC calculated?*
- *Should I use $AIC_c$ or AIC?*
- *How do I interpret AIC values?*
- *What is model averaging?*
- *Do I always need to model-average my parameter estimates?*
- *It seems like everyone uses a different method for model averaging. What should I do?*

## Start here, but keep going

A discussion of model comparison is critical to the background section before we begin to explore the various types of parameter estimation processes. Here, we will provide a cursory introduction. To continue your journey, we point you towards the gold standard for ecological investigations and model inference: Burnham and Anderson (2002). And, we heartily recommend another small primer, Anderson (2007) that picks up where we leave off with an easy-to-understand approach.

This chapter has two parts: first, an introduction to the theory and concept of AIC and model comparison. Then, we finish with a discussion of the process (which can often be complex!) of selecting the "best model" and how to present your parameter estimates to the world.

---

[1] *With thanks for content to Evan Cooch and Gary White.*

## Part One: Multi-model inference

A person does not need to use multi-model inference to estimate the value of a parameter. In fact, in all of the examples used in this primer to this point, we have looked at one model and estimated the value of a parameter based on that model.

But, ecology is complex. Most biologists estimate parameters (e.g., survival, density, movement probabilities) in the context of a question. Does survival vary between genders or ages? Does density vary among my study plots? Which management scenario results in the highest probability of nest survival for mallard ducks (*Anas platyrhynchos*)?

To answer these questions, ecologists have long-used an approach of alternative hypotheses (Chamberlin 1965). We use **"model"** in many ways (see Chapter 1), and one way is to describe a concept (or hypothesis, in this case) for the truth with respect to values of parameters for a population. One "model" might suggest that survival is the same among all animals, while another "model" would propose that survival varies by gender. Still another "model" might suggest that survival, instead, varies by age of the individual (e.g., juvenile and adult categories).

The statistical approach to alternate hypotheses regarding demographic parameters has changed over the years—the earliest, simplest form was to compare a null model with one alternative model in which a parameter varied according to categories of animals. For, example, do annual survival rates of northern bobwhite (quail [*Colinus virginianus*], Figure 4.1) differ by gender (male and female categories)? Perhaps age (juvenile and adult categories)? We could lump our sample into data obtained from males and data obtained from females, and we could estimate a survival rate for each group. Hines and Sauer (1989) provided a comparison using program CONTRAST to assess whether the resulting estimates of survival (and its associated estimate of variance) provided evidence of a difference. If there was no evidence for variation of survival according to gender, the null model of "no difference" would be supported. At that time in history, the ability to make simple comparisons was a huge step forward in our analyses. Today, it seems very simple (*which, we remind the reader, does not mean it is useless!*).

One problem with the approach used by CONTRAST was that it was limited to categorical variables (e.g., gender or age) to describe differences in a parameter. What if a model suggested a continuous variable might be important—for example, might the survival of quail change with respect to body mass (small birds have lower survival than large birds?)? Program CONTRAST couldn't handle that comparison.

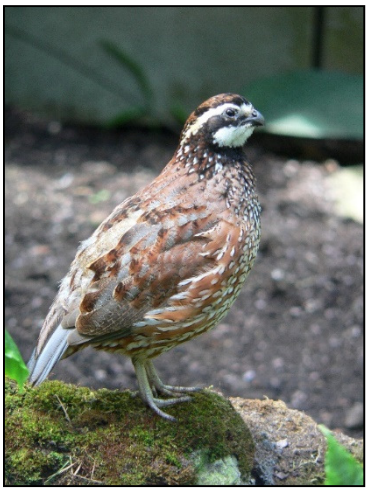

**Figure 4.1:** *Northern bobwhite (Colinus virginianus). Photo by BS Thurner Hof, available in the public domain.*

As a result, since the mid-1990's, ecologists have developed easy-to-use methods to allow exploration of categorical (discrete) or continuous effects on the value of parameters. And, during this time, ecologists have increased their use of an analysis framework known as multi-model inference, which uses model comparison as its basic structure. The general idea is that a team of scientists can put forward two or more models that they

believe to represent alternative, biological hypotheses about the way the world "works" with respect to the value of the parameter(s) they are estimating. Each model, individually, is constructed. The values of the parameters are obtained. Then, the models are compared with each other to determine which model best represents variation in the value of parameters that is supported by the data.

So, for example, a team of quail biologists could submit the following models about annual survival of quail in northern Florida:

| **Model** | **Description** |
|---|---|
| Null model | The entire population has the same probability of annual survival |
| Gender model | Males and females vary in their probability of annual survival |
| Age model | Juveniles (<1 year) and adults vary in their probability of annual survival |
| Mass model | Mass (g) of an individual affects the probability of annual survival |

So, the biologists have now created four models and are ready to compare them. Before we let them continue with their comparison, we need to pause and think about the philosophical framework.

In Chapter 2, we discussed inference methods as ways in which scientists seek "truth" about their biological systems. So, philosophically, there is some "truth" about our population, and we are trying to get as close as we can to the "truth." But, we have also acknowledged that ecologists do not know (and may never know) the complete truth about their biological systems. Thus, you might ask: *"How will we ever know which of our models is the best model, if we don't know the truth?"*

And, it is true—if we knew "truth", it would be easy to compare the results from our parameter estimation to "truth". It seems fairly easy to understand that a side-by-side comparison of all of our models to "truth" would show which model is "best" at describing the system. In fact, statisticians have named the theoretical distance between any model and the full reality, or truth, the **Kullback-Leibler distance** (Figure 4.2). In information theory, the size of that distance provides information, termed

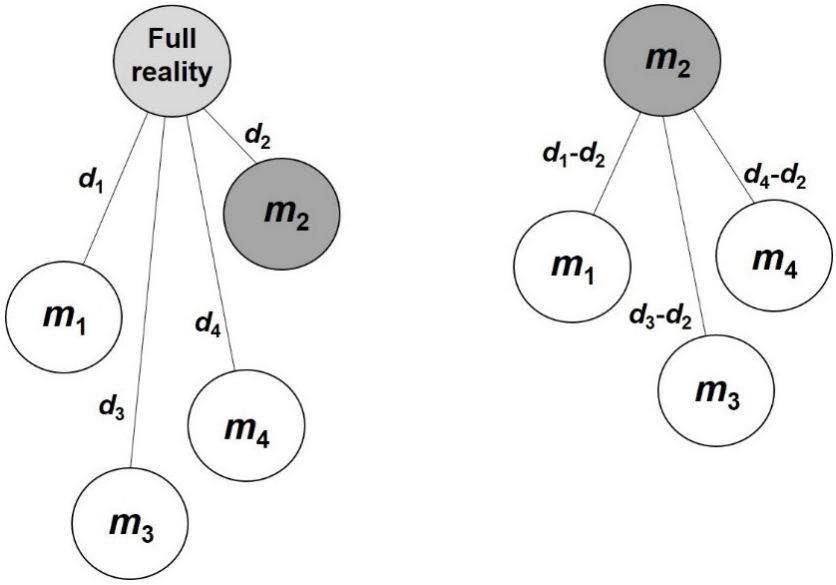


**Figure 4.2:** *The theoretical distances ($d_i$, Kullback-Leibler information) between full reality and competing models ($m_i$) are shown at left. The estimated distances from the best model ($m_2$) to other models provide the conceptual framework for AIC values and ⊥AIC for model comparisons (after Cooch and White 2014).*

*Kullback-Leibler information*. We could use that information to compare the relative merits of our models.

*But, we don't know the truth.*

Luckily, statisticians have also theorized that if such "real" distances between models and truth exist for our set of models, then we can estimate the distance from all models to the best model. To understand this, consider two cities (City A and City B) that exist near a sacred mountain with a sacred population of mountain goats (*Oreamnos americanus*) at its top in a faraway land. Both of the cities can see the mountain (and on a day with good visibility, the goats). And, citizens of both cities know that City A is closer to the mountain. There is direct evidence of which city is closest to the holy site (Figure 4.3).

Our story takes a turn for the worse when a massive earthquake occurs. The mountain disappears, the sacred mountain goats all perish, and only a few people in City A and City B survive. Over generations, the cities rebuild. Eventually, a shopkeeper in City A finds a book with stories of a sacred mountain that once stood near the cities. The book contains the phrase, "Blessed will be the city that is closest to the sacred mountain…" But, the rest of the book was destroyed in the earthquake, and no one knows which city is closest. Without a reference point, it is impossible to know which city was the blessed city. Or…is it?

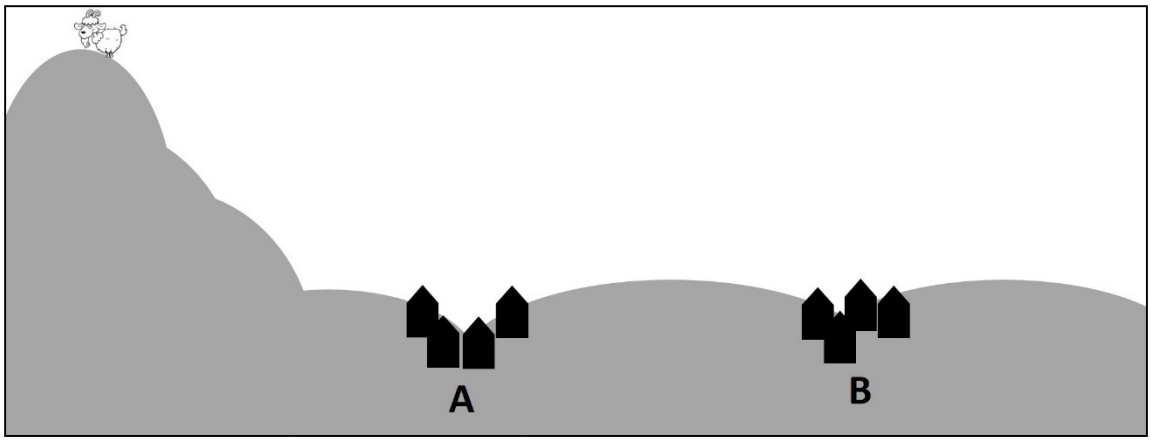


**Figure 4.3:** *Relative distances from City A and City B to the sacred mountain, the similarly sacred sanctum for a population of mountain goats.*

## Akaike's Information Criterion to the rescue

Many readers will confirm that it is impossible to read many ecological papers that deal with parameter estimation without seeing a reference to **Akaike's Information Criterion**, or **AIC**. And, it is not difficult to determine—from the ecological papers that use AIC values—that AIC is a method of model comparison. That is, AIC allows us to compare our models in attempt to select the best model among our model set.

What exactly is Akaike's Information Criterion? AIC was developed by the Japanese statistician, Akaike, as an unbiased estimator of the "relative, expected" Kullback-Leibler distance. Thus, AIC is a method that identifies which model is closer to the truth that appears to be expressed by the data from your sample.

So, without knowing the location of the mountain, the AIC model comparison concept allows us to use evidence to learn about City A and City B (Figure 4.3). Perhaps, after searching through all known diaries of residents of the two cities, we learn that more diaries from City A mention the mountain. And, more holy relics are found in City A. As evidence builds, we become more and more certain that City A was closest to the mountain. So, we use evidence to determine that A was most likely the closest city to the original site of the sacred mountain without knowing where the mountain once stood. That's spooky, but it is pretty useful! And, that is essentially what the AIC method does—it assesses evidence that is gathered in our sample.

Conceptually, Akaike's method addresses a trade-off between model fit (reduced bias) and the variance of the estimate, and this trade-off is important (Figure 4.4). On one hand, we want a model that 'fits' (adequately explains) the variation we believe is in our sample data. Such an approach would favor a more complex model—one with more parameters—to attempt to explain variation. For example, we might believe that survival of quail is affected by gender and age of the bird. So, we'd need a complicated model to estimate survival for juvenile males, juvenile females, adult males, and adult females (4 parameters estimated).

On the other hand, as we estimate more parameters with the same pool of data, we create more variance components and more uncertainty about the value of each parameter (we address why this is true with an example below). We like precision, so we would want to have a simple model with fewer parameters and lower variance. A model that uses all the data to estimate one survival rate (no age or gender effects) would be the simplest and would give a survival rate with a lower standard error. AIC addresses that trade-off to determine if we have enough evidence to support a more complex model.

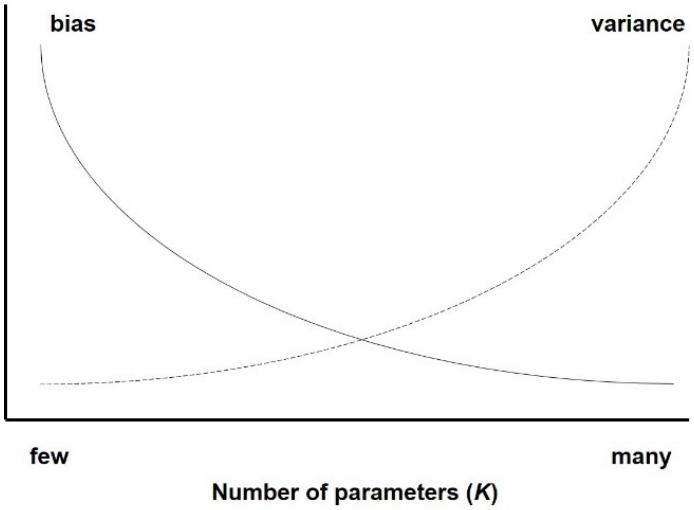


**Figure 4.4:** *Conceptual trade-off between model fit (high bias = poor fit) and uncertainty for a parameter estimate, as expressed by variance of parameters in the model. Models with few parameters may have more bias than highly parameterized models, but models with many parameters may have high estimates of variance for the parameter values (after Cooch and White 2014). Y-axis is scaled "low" (at origin) to "high" for both bias and variance.*

## AIC: "fit" versus variance components

To understand the idea of complexity in a model, relative to evidence to support such a model, let's step away from estimation of survival rates and think about a population's trend over time. Let's suppose a biologist gathers information on population size from surveys every year, resulting in data points displayed in Figure 4.5.

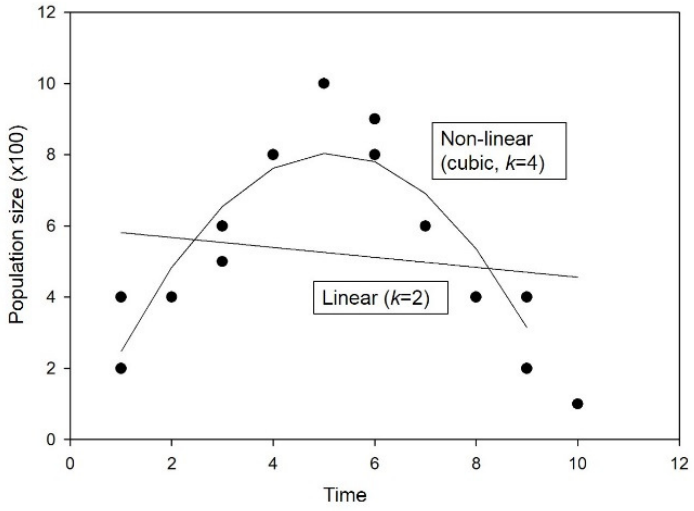


**Figure 4.5:** *Changes in population size over time (dots) with a comparison of two explanatory models: a simpler linear model and a more complex non-linear model.*

The biologist wants to find the best description of the trend in this data. It is possible that there is only a linear trend through time, but the biologists visually inspects the raw data and sees the possibility that the population has increased and is now decreasing (Figure 4.5).

The linear trend can be quantified with a regression model that estimates an intercept ($B_0$) and a slope ($B_1$): $Y = B_0 + B_1(x)$

The non-linear trend could be quantified with a "cubic" model that allows flexibility in the line by adding parameters: $Y = B_0 + B_1(x) + B_2(x^2) + B_3(x^3)$

Thus, we have a simple description of the trend (the linear model, with 2 parameters—the intercept and a slope) that would be easier to estimate and would have fewer variance components. But, it really doesn't look like it 'fits' the data collected by the biologist, does it?

The alternative is a complex model with 4 parameters. It really looks like it "fits" the data much better. But, is there evidence to support the complexity? We can select the best model by assessing the AIC values for each model.

**AIC does not tell you which model "fits the best"!**

When reporting the results of a model comparison using AIC, avoid the statement *"the model that best fit the data was…"!*

AIC assesses both **fit** and **complexity**. Thus, it is possible that a more complicated model "fits" your data better that the one selected by AIC as the best!

The best way to report the model selected by AIC is: *"The model with the most support, given our data, was…"*

## Calculation of AIC values

AIC is calculated with two simple components, which represent the tradeoff between fit and variance:

$$AIC = -2\ln(L) + 2K$$

where:

*L = likelihood for a model under consideration (this describes the "fit" and should be familiar to the reader from our discussion in Chapter 3), and*

*K = number of parameters in the model (this describes the complexity and number of variance components)*

The reader can see that as more parameters are added to the model (as *K* gets larger), the value for AIC should increase. Hence, the two portions of the equation to obtain the AIC are working against each other.

In the 'game' of using AIC for model selection, the idea is to select the model for which the AIC value is the minimum. If properly applied, the AIC method results in the selection of the best approximating model. That is, of the models under consideration, AIC will select the best approximation of "full reality", or truth, as evidenced by the data that is provided in the sample.

## Correction for small sample size?

Sample size can affect the performance of the AIC statistic. Specifically, AIC may perform poorly if there are too many parameters relative to the size of the sample. *We should emphasize that this does not affect the estimate of the parameter (e.g., survival), but it does affect the ranking of the models in competition.*

Thus, many software packages (e.g., SAS proc's, R packages, program MARK) allow the user to examine a version of AIC, corrected for small sample size (**$AIC_c$**):

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1}$$

One might ask: *"So, how <u>small</u> does my sample need to be to use $AIC_c$ rather than AIC?"*

The answer is—always use $AIC_c$. And, we suggest this not because you should be worried about your sample size in all situations, but rather because $AIC_c$ and AIC become the same as the sample size becomes large. That is, as you increase *n* in the correction component of the formula for $AIC_c$, the correction component quickly goes to 0.

In Figure 4.6, we show how the AIC correction for a model with 3 parameters (*K*=3) becomes insignificant when the sample goes above 50. So, for samples of more than 50, the value of $AIC_c$ is approximately the same as the value of AIC. If you always use the $AIC_c$ value in your

comparison, you won't have to worry about when to switch to AIC…the $AIC_c$ just 'becomes' AIC as the sample increases.

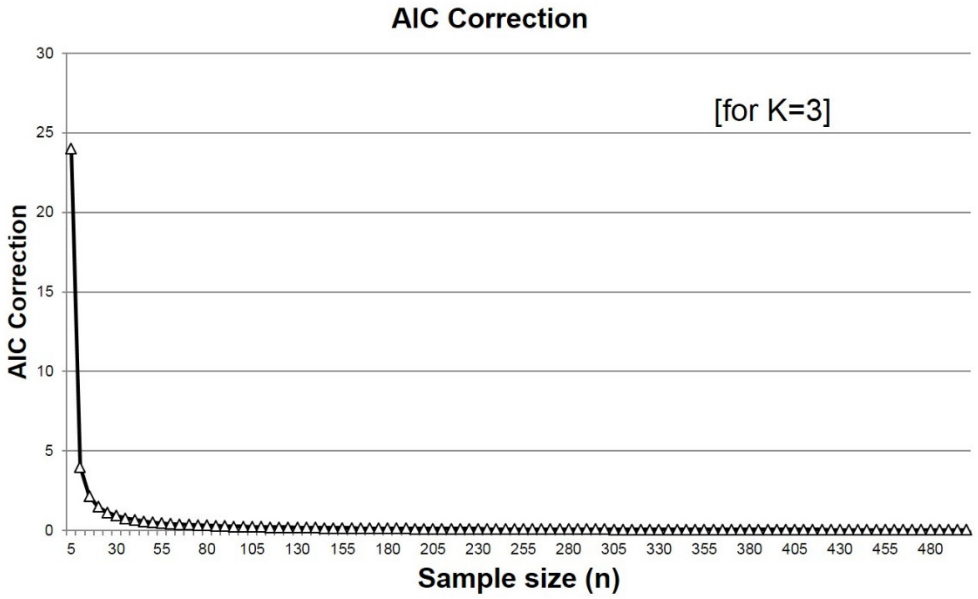


**Figure 4.6:** *The magnitude of the correction applied to AIC values to create AICc values (corrected for small sample size) as sample size increases from 5 to 500 for a model with 3 parameters.*

## Quail example—working through a basic comparison with $AIC_c$

Let's return to our quail survival example that we have used previously in this chapter. We can stipulate the following models for a simple analysis:

| **Model** | **Description** | **Survival parameters estimated** | **K** |
|---|---|---|---|
| Model 1 | All animals have the same probability of monthly survival | $S(.)$ | $K = 1$ |
| Model 2 | Males and females have distinct survival rates | $S_m$ and $S_f$ | $K=2$ |

Let's consider a situation in which we radio-mark 100 adult quail: 50 males and 50 females. All animals are marked and released at the same time, and one month later, the animals are checked using telemetry. We obtain the following data:

*20 of the 50 males are still alive*
*35 of the 50 females are still alive*

We are using 'known fate survival' (the topic of Chapter 9), and it is fairly easy to estimate that the monthly survival probability for males is $S_m$=0.40 and $S_f$=0.70 for females. If we pool all animals together, we can see that 55 of the 100 quail survived, so survival of all adults is $S_a$=0.55.

Those rates appear fairly different for males and females, but do we have enough evidence to support that assessment? *Is survival equal for male and female quail?*

To use model comparison with $AIC_c$ to answer our question, we need to splice the information into the equation for the $AIC_c$, and we can see that we need the values of the model likelihoods (L) and the number of parameters (K).

If you made it through Chapter 3 (Maximum Likelihood Estimation) alive, you should be able to set up this likelihood in your sleep! Let's give it a shot…

For our simple model, Model 1, we need to estimate one survival rate ($S_a$, adult survival). The likelihood is derived by using our males and our females as two samples of the same statistical population (adults). The males had a sample of 50 "trials", with 20 successes. The females had a sample of 50 "trials" with 35 successes. This can be written as:

$$L(S_a \mid 20,35) = \binom{50}{20} S_a^{\,20}(1-S_a)^{(30)} \times \binom{50}{35} S_a^{\,35}(1-S_a)^{(15)}$$

If we work out the math for this (using the estimates of $S_a$ = 0.55), we get an estimate for the likelihood,

$$L = 0.0001.$$

Then, we can calculate

$$\ln L = -8.8876.$$

Thus,

$$AIC = -2\ln(L) + 2K$$
$$AIC = -2(-8.8876) + 2(1)$$
$$\mathbf{AIC = 19.7752}$$

Above, we encouraged you to always report $AIC_c$, with the correction for small sample size, so we need to adjust AIC to $AIC_c$. We need to know values for *K* and *n*; in this model, we have 1 parameter (*K*=1) and our sample size is 100 animals (*n*=100). So, our correction is:

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1}$$

$$AIC_c = 19.7752 + \frac{2 \cdot 1(1+1)}{100-1-1}$$

$$AIC_c = 19.7752 + \frac{4}{98} = 19.7752 + 0.0408 = 19.8160$$

As expected, with a large sample, the $AIC_c$ (19.81) is very similar to the value for AIC (19.78).

Our more complex model, Model 2, is going to give us two separate survival rates ($S_m$ and $S_f$: male and female survival). The trials and successes are still the same. We write this likelihood as:

$$L(S_m, S_f \mid 20,35) = \binom{50}{20} {S_m}^{20}(1-S_m)^{(30)} \times \binom{50}{35} {S_f}^{35}(1-S_f)^{(15)}$$

If we work out the math for this (using the estimates of $S_m = 0.4$ and $S_f = 0.7$), we get an estimate for the likelihood,

$$L = 0.0140.$$

Then, we can calculate

$$\ln L = -4.2676.$$

Thus,

$$AIC = -2\ln(L) + 2K$$
$$AIC = -2(-4.2676) + 2(2)$$
$$\mathbf{AIC = 12.5352}$$

Here, we calculate the adjusted $AIC_c$ with $K = 2$ and $n = 100$: the adjustment = 0.1237, and **$AIC_c$ = 12.6589**.

When we compare the two values for $AIC_c$, we see that the $AIC_c$ for Model 2 is less ($AIC_c$=12.7) than for Model 1 (AIC = 19.8). Thus, we have evidence from our sample that survival rates do vary for quail between males and females. *Model 2 (sex-specific survival) is closer to reality than is Model 1 (no difference in survival).*

We should note that **our assessment is only as complete as the set of models that we compared**. For example, we did not include age or mass in our assessment. So, our model comparison and the inferences that are derived from it are constrained to the models considered. To investigate a more complete truth about quail survival, we would most likely consider a broader range of models that assess more than a simple gender comparison.

## Putting it together: AIC comparison statistics

We can use our newly-calculated AIC values to calculate a set of useful statistics that can help us interpret our results. To what degree is the sex-specific survival model in the bobwhite quail example above, better than the simpler model that suggests survival is the same between sexes?

We can start by calculating **$\Delta AIC_c$**, which is the difference between the $AIC_c$ value for any model, *i*, in our model set and the top-ranked model (Figure 4.2, the model with the minimum $AIC_c$ value, or $AIC_{c\_0}$):

$$\Delta AIC_c = AIC_{c\_i} - AIC_{c\_0}$$

To interpret the relative support for a given model we need to calibrate the assessment of that model relative to the entire set of models under consideration. We use the ΔAIC to calculate **AIC weights** ($w_i$) to provide an index of this evidence—the likelihood of a model, given the model set. We calculate the model weight ($w_i$) for the *i*th model in the set of *n* models as:

$$w_i = \frac{\exp\left(-\frac{1}{2}\cdot\Delta AIC_{c\,i}\right)}{\sum_{i=1}^{n}\exp\left(-\frac{1}{2}\cdot\Delta AIC_{c\,i}\right)}$$

The next statistic that we may calculate uses the model weights to express the **model likelihood** (ML, or, literally, "how likely") for a given model. The model likelihood is the ratio of the model weight for model *i*, compared to the top-ranked model's weight ($w_0$):

$$ML_i = \frac{w_i}{w_0}$$

We interpret the model likelihood value as the strength of evidence of this model, relative to other models under comparison.

We can place these statistics, calculated for our bobwhite quail example, in a model comparison table that would be suitable for use in a publication (where k is the number of parameters for model *i*):

| Model | $AIC_c$ | $\Delta AIC_c$ | $w_i$ | Model likelihood | *k* |
|---|---|---|---|---|---|
| S(sex-specific) | 12.659 | 0 | 0.973 | 1.000 | 2 |
| S(.) | 19.826 | 7.167 | 0.027 | 0.028 | 1 |

Now, we can interpret our results, fully. The top-ranked model has approximately 97% of the model weight in our simple model set of two models. The second-ranked model only has 3% of the model weight, and the top-ranked model is 36 times more likely to be the best model than the second-ranked model (1/0.028=36).

We recommend using **"evidence-based language"** when describing the results of model comparison. Here, there is "strong evidence" for the top-ranked model. If model weights were more similar, our evidence would start to fade. We might use phrases like "good evidence",

"some evidence", "little evidence", or "no evidence" to describe our certainty that males and females have different survival rates, based on model results.

As a rough guideline, Burnham and Anderson (2002, p. 70) provide the following recommendations for interpretation of AIC-based model comparisons:

| $\Delta AIC_i$ | Level of empirical support for model *i* |
|---|---|
| 0–2 | "Substantial" |
| 4–7 | "Considerably less" |
| > 10 | "Essentially none" |

Thus, our example of bobwhite quail survival is an easy interpretation—we have strong evidence as the support for the null model, *S*(.), is "considerably less" than support for the sex-specific model.. But, what if our results were different? What if we had weaker evidence that the sex-specific model was the best? What if there was "substantial" support for the null model, *S*(.)? Let us consider a *different outcome* from this model comparison, to explore some details of model selection:

| Model | $AIC_c$ | $\Delta AIC_c$ | $w_i$ | Model likelihood | *k* |
|---|---|---|---|---|---|
| S(sex-specific) | 12.659 | 0 | 0.630 | 1.000 | 2 |
| S(.) | 13.600 | 1.065 | 0.370 | 0.587 | 1 |

Now, the weights of the two models are more similar. The ΔAIC for the second-ranked model is <2.0, and the table above suggests that our second-ranked model would have "substantial support", even though it was ranked second. And, our model likelihood suggests that our top-ranked model is not even twice as likely to be the best model as our second-ranked model (1.000/0.587=1.7). So, it is very difficult for us to distinguish between our two models.

In such a situation, model selection theory provides an approach that we could use to select the best model—the concept of **parsimony.** *Parsimony suggests that, given equal explanatory value, we should select the simplest explanation.* We measure simplicity as the number of parameters, so the *S*(.) model (with 1 survival rate and 1 parameter) would be selected as the best model. Another way to view this situation is that the sex-specific survival model had not accumulated enough evidence to distance itself from the simpler model, and thus we have very little evidence to suggest that quail have sex-specific survival rates.

## Part Two: Thoughts on model selection (when life gets messy)

In fact, model comparison/selection can become very difficult when one model is not a clear 'best model'. Should we consider all models plausible if they are within 2 ΔAIC from the best model? It is not unusual for ecologists to throw up their hands in frustration when they receive results such as those shown in the table below:

| Model | $AIC_c$ | $\Delta AIC_c$ | $w_i$ | Model likelihood | $k$ | -2ln(L) |
|---|---|---|---|---|---|---|
| Model 1 | 187.30 | 0 | 0.40 | 1.000 | 2 | 183.3 |
| Model 2 | 188.50 | 1.2 | 0.22 | 0.549 | 3 | 182.5 |
| Model 3 | 188.60 | 1.3 | 0.21 | 0.522 | 2 | 184.6 |
| Model 4 | 189.10 | 1.8 | 0.16 | 0.407 | 6 | 177.1 |

However, we can find a way forward when we find ourselves in this situation. Let's review what we know about our results, keeping in mind the formula for AIC:

$$AIC = -2\ln(L) + 2K$$

- *Model 1:* was ranked the top model, and it is a simple model with only 2 parameters.
- *Model 2:* we can see that this model's 'fit' (-2lnL) was actually a bit better (lower value) than Model 1. But, the AIC formula penalizes (with the term 2**K*) this model by 2 AIC for the one additional parameter (*K*=3), compared to Model 1. Thus, Model 2 fits a little better than Model 1, but not enough to make up for the complexity.
- *Model 3:* this model's 'fit' is not as good (larger value for -2lnL) as Model 1, and it has exactly the same number of parameters as Model 1 (*K*=2).
- *Model 4:* this model's 'fit' is also better (smaller value for -2lnL) than Model 1. But, AIC penalizes Model 4 a whopping +8 for four additional parameters (*K*=6), relative to Model 1. The better fit is not enough to make up for the complexity of the model.

Neither Model 2 nor Model 4 have improved their likelihood *enough* to counter the addition of additional parameters. And, although Model 4 has a better fit (as measured by (-2lnL), it is not sufficiently better to merit the addition of 4 more parameters relative to Model 1. As such, we can suggest that the additional parameters in Model 2 and Model 4 are *uninformative parameters*. The additional parameters are not helpful to describe reality, as judged by the evidence we have collected (our data).

With reference to our assessment of Model 2, Burnham and Anderson (2002, p. 131) wrote:

> *"Models having $\Delta_i$ ($\Delta$AIC) within about 0-2 units of the best model should be examined to see whether they differ from the best model by 1 parameter and have essentially the same values of the maximized log-likelihood as the best model. In this case, the larger model is not really supported or competitive, but rather is 'close' only because it adds 1 parameter and therefore will be within 2 $\Delta_i$ units, even though the fit, as measured by the log-likelihood is not improved."*

Arnold (2010) provided a useful summary of ways forward for ecologists when we encounter this situation of models that appear to have uninformative parameters (many models within 2 ΔAIC of the top model). Arnold (2010) evaluated the merits of five options:

- Full reporting when model sets are small
- Model averaging
- Assessment of confidence intervals to evaluate evidence
- Assessment of relative variable importance
- Discarding models with uninformative parameters

In the next few sections of this chapter, we will follow Arnold's (2010) thoughts with some examples of our own.

Normally, the practice of **full reporting** is possible only when the model set is small (perhaps fewer than 10 models). In this situation, we would create a table that shows all 10 models, and we would report on the levels of uncertainty (e.g., confidence intervals for parameter estimates) and we would evaluate the strength of the evidence. We could interpret the AIC values for each model in light of the number of parameters. We might use the concept of **parsimony**, for example, to do a complete assessment of our models with a mind to use the simplest explanation of reality if evidence does not accumulate for more complex examples. An example of full reporting is the detailed evaluation of Models 1, 2, 3, and 4 that we provide above.

Ecologists may also use a process known as **model averaging** to incorporate **model uncertainty** into the estimates of parameters. This approach focuses on the estimate of parameters obtained from individual models in the data set and uses a weighted average across the estimates from more than one model. In our quail example, to obtain a model-averaged value for males in the population, we could use a weighted average (weighted by the model weights, $w_i$) of the male survival rate from the sex-specific model and the pooled survival that would be attributed to males (and also females) in the S(.) model. *However, we note that model averaging is not recommended in our first set of quail results, because we are highly certain (based on model weights) that survival is sex-specific.*

The approaches to model averaging are numerous, and the reader is encouraged to read more about model averaging before conducting the process. The first decision in model-averaging is to *decide which of your models to include in the average.* Some ecologists, for example, model average across the set of models that are within 7 ΔAIC of the top model. Other ecologists select the model set for averaging that, as a group, make up 90-95% of the cumulative model weight—in our example above, we would use all four models by either set of rules. Rehme et al. (2011) provide a review of the complex set of approaches that have been used by biologists to establish a model set for model averaging.

However, we agree with Arnold (2010) when he states that if models with uninformative parameters were removed prior to model averaging, the top model might often have 80-90% of the model weight, and we would often have no need to model average. The argument to avoid model averaging when possible is that the SE calculated for our model-averaged parameter is always larger than the standard SE estimated using maximum likelihood estimation for the parameter in an individual model. The SE becomes larger, because we incorporate **model**

**uncertainty** into the SE when we model average—and thus, we incorporate more error. Although this does account for uncertainty in the search for truth, we'd like to avoid inflating variance estimates when we can—especially if the manner in which we constructed our models added to the confusion. However, in the spirit of presenting all sides of the argument (we told you this could be complicated, right?!), we note that other quantitative scientists are firmly supportive of model averaging (e.g., Burnham and Anderson 2002, Lukacs et al. 2010, Doherty et al. 2012).

Two methods for **model averaging** are used when model certainty is low. **Unconditional model averaging** is used to calculate a weighted estimate by considering all models, *R*, in the model set, regardless of whether they contain the covariate of interest, $\beta_j$.

$$\overline{\beta}_j = \sum_{n=1}^{R} w_i \hat{\beta}_{j,i}$$

**Conditional model averaging** also calculates a weighted estimate for a covariate, but the average is *conditioned* on whether the covariate, $\beta_j$, appears in the model—only such models and their summed weights (denominator, below) are used for averaging.

$$\overline{\beta}_j = \frac{\sum_{n=1}^{R} w_i \hat{\beta}_{j,i}}{\sum_{n=1}^{R} w_i}$$

Let us consider an assessment of survival that includes 5 models. Each model contains covariates that describe potential effects of time (*t*), vegetative cover, and age of the individual on survival:

| **Model** | $\beta_{cover}$ | $w_i$ | **Unconditional** | **Conditional** |
|---|---|---|---|---|
| 1: S(cover, t) | 2.16 | 0.54 | X | X |
| 2: S(t) | - | 0.35 | X | |
| 3: S(cover) | 2.57 | 0.05 | X | X |
| 4: S(age) | - | 0.04 | X | |
| 5: S(age, t) | - | 0.02 | X | |

If we decide to obtain a model-averaged value for the cover covariate, $\beta_{cover}$, we have two choices. The table shows the value for the covariate of cover ($\beta_{cover}$) for the two models in which the covariate is present (Model 1: $\beta_{cover}$ =2.16; Model 3: $\beta_{cover}$ 2.57). Unconditional model averaging will use the model weights and the values for $\beta_{cover}$ to average across all five models in our model set. **Note that when "cover" is not present in the model, unconditional model averaging uses $\beta_{cover}$ =0.**

The **unconditional** model averaged estimate:

$$\begin{aligned}\overline{\beta}_{cover} &= (2.16*0.54)+0+(2.57*0.05)+0+0\\ &= 1.17+0+0.13+0+0\\ &= 1.30\end{aligned}$$

Alternatively, a **conditional** model averaging approach would only use the models for which $\beta_{cover}$ is included, and the sum of the weights for those models:

$$\begin{aligned}\overline{\beta}_{cover} &= \frac{(2.16*0.54)+(2.57*0.05)}{0.54+0.05}\\ &= \frac{1.17+0.13}{0.59}\\ &= 1.30/0.59\\ &= 2.20\end{aligned}$$

It may seem odd to use $\beta_{cover}=0$ for models in which $\beta_{cover}$ does not exist. But, proponents of unconditional model averaging suggest that if a model such as S(age) hypothesizes that age has an effect on survival, we also have an unwritten hypothesis: *the S(age) model suggests that cover does NOT affect survival.* Thus, $\beta_{cover}=0$ by default in the S(age) model.

As your authors, we will admit a personal bias against unconditional model averaging. It seems fairly obtuse, to us, to go to the trouble of finding an unbiased, maximum likelihood estimate for a covariate value, only to modify (*bias?!*) that estimate through an averaging process that could vary tremendously because of the number of models and the structure of the models submitted for consideration. *We tend to favor conditional model averaging, if model averaging is necessary.* **Above all, we encourage you to consider your analysis (<u>will you model average?</u>) before you construct your model set.**

## Building model sets: plan ahead

As an example of how the structure of the set of models submitted in an analysis can affect the results when model averaging, let us consider the following two sets of models:

**Model Set 1:**
S(age)
S(t)
S(gender)
S(.)

**Model Set 2:**
S(age)
S(t)
S(gender)
S(.)
S(age+t)
S(age+gender)
S(gender+t)

If we were interested in a model-averaged estimate of the covariate for "gender" in this analysis, which model set is better planned for the use of model averaging? We would submit that the second set is best constructed for averaging. Doherty et al. (2012) also suggest 'balanced' sets of models for use in model averaging.

In fact, it would seem that a model averaged estimate for 'gender' is nonsensical in Model Set 1, as there is only one model that includes gender. By default, there is no other model to average across—and if unconditional model averaging is used, the estimate will be affected by the $\beta_{gender} = 0$ found in the other three models.

In Model Set 2, "gender" appears in three models—we could average across three things, logically.

We will also note that the first set of models has the underlying hypothesis that only age *OR* time *OR* gender affect survival, but not more than one effect. And, if both age and gender have an effect on survival (a common occurrence in nature), both the age and gender model should be highly supported…but neither will separate itself as the 'best model'. In truth, there are two good, single-factor models [S(age) and S(gender)]. Such a dynamic would be documented, most likely, by the use of the second set of models, which includes models that propose two influences on survival [e.g., S(age+gender)].

## AIC values are only one portion of your results: don't forget to look at your estimates!

Thus far, we have only looked at the **model rankings** to document the strength of evidence found in our analyses. But, of course we can also look within each model to assess the parameter estimates and the variance—often expressed as a **95% confidence interval**. If the confidence intervals for male and female survival overlap, that is good evidence that males and females do not have distinct survival rates. However, if confidence intervals do NOT overlap, it is good evidence to suggest that males and females do have distinct survival rates.

However, Arnold (2010) pointed out that ecologists may find evidence that appears to conflict when they compare AIC rankings with information from 95% confidence intervals of parameter estimates. For example—in some instances, the AIC rankings may indicate that a sex-specific model (or other type of more complex model) is ranked higher than a null model with no effect. The $\Delta$AIC for the null model may be >2.0, leading us to believe that we should favor the sex-specific model. But, when we look at the 95% confidence intervals for male and female survival estimates, the intervals overlap—suggesting that there is not an effect of sex on survival. Thus, the AIC comparison and the 95% confidence interval are sending conflicting messages. If this hasn't happened to you yet in your career, it will happen to you at some point!

We encourage the reader to evaluate Arnold's (2010) thoughtful analysis of this situation in detail, but we can summarize the explanation. We know from the information early in this chapter that AIC provides a +2 penalty for adding an additional parameter. When we use $\Delta$AIC>2.0 as our guide to tell us when one model is better than another, this is equivalent to using an $\alpha$-level of 0.15, rather than $\alpha = 0.05$.

As we know, the 95% confidence interval is based on $\alpha = 0.05$. So, it does not make sense to evaluate parameters at the 95% confidence interval level—we should, in fact, use an 85% confidence interval for our comparison of male and female survival rates. An 85% confidence interval is "tighter" (narrower) than a 95% confidence interval. And, this explains why we sometimes see AIC comparisons that suggest survival rates are different between groups, yet the 95% CI's overlap for the groups.

If you use software for your analysis that allows you to set the confidence intervals provided, we suggest that you follow Arnold's (2010) advice to create 85% confidence intervals. However, some software packages (e.g., PRESENCE, MARK) do not allow this flexibility as of this writing.

## Cumulative weight: more evidence

Another strategy to assess the value of a covariate in a more complex model analysis, especially when model certainty is low, is to analyze the **relative importance** of each variable. In this case we determine the cumulative weight of all models in which the parameter is found—the logic is that if a variable is important, most of the models in which it appears should have high model weights, and thus the cumulative weight for that variable will be high.

As Burnham and Anderson (2002) suggest, we must take care to not create implausible hypotheses and miss the point of "thoughtful" model sets. Thus, we must think about how we construct models for comparison. Specifically, model sets must be constructed in "balanced" fashion—with balanced combinations of variables that have biological meaning.

| Model | Weight |
|---|---|
| Grass | 0.35** |
| Day + Grass | 0.25** |
| Gender | 0.12 |
| Hormone | 0.10 |
| Grass + Gender | 0.08** |
| Null | 0.05 |
| Grass + Forb | 0.05** |

In the model set above, the cumulative weight for the effect of grass cover = 0.73 (the sum of all weights marked with "**"). Is grass cover an important variable? We hardly know—because grass appeared in all but three models. The effect of day and forb cover only appear in one model. That's not a fair comparison, no matter how you think about it.

| **Model** | **Weight** |
|---|---|
| Grass | 0.35** |
| Day + Grass | 0.25** |
| Null | 0.12 |
| Hormone | 0.10 |
| Day | 0.08 |
| Hormone + Day | 0.05 |
| Hormone + Grass | 0.05** |

The second set of models, above, is balanced with one-factor and two-factor models. Each variable, above, appears in three models. We find that the cumulative weights for grass = 0.65 (i.e., 0.35+0.25+0.05 = 0.65), while cumulative weight for day = 0.38 and hormone = 0.20. Thus, we can conclude—much more legitimately—that grass is an important factor to describe variation in the parameter of interest.

Another approach is to avoid situations with high model uncertainty through careful planning and the use of **step-down model selection**. We can use a step-by-step process to eliminate models that are not useful, which will limit the number of models under consideration in the final model comparison. Lebreton et al. (1992) described this process as "step-down" analysis, and they recommended it in situations when model sets were prone to be unmanageably large if an "all possible combinations" approach was taken (see Chapter 13 for an example of a robust design analysis that evaluated models with 150-500 parameters). In fact, program Distance (Buckland et al. 2001; Chapter 19) uses a similar approach to step through the selection of the best model to describe detection patterns in surveys.

If your analysis has a small number of parameters and associated covariates to estimate, it seems reasonable to follow Doherty et al. (2012) who recommended an "all combinations" model strategy. Such an approach provides balanced sets of models that can be used in model averaging and it avoids ad hoc strategies of model selection. In fact, the simulation study by Doherty et al. (2012) suggested that ad hoc strategies ran the risk of inflating the importance of variables that were weakly correlated with the parameters of interest.

Here is an example to explain the "step-down" model selection approach that can simplify your model set. Let us assume that we have a goal to assess the effects of age, sex, vegetation density (of the current location) on the survival of the northern bobwhites we used in a previous example. Sex and age are categorical variables (male/female and juvenile/adult), but, vegetation density is a continuous variable. We might postulate that vegetation density has a linear effect on survival (as vegetation density increases, survival increases). However, we also might postulate that vegetation density has a non-linear effect on survival—perhaps survival is low

when quail use sparsely vegetated areas and also low when the quail use very densely vegetated areas.

Thus, we develop two models that describe the possible effect of vegetation density on survival:

Linear: $S = B_0+B_1(\text{veg. density})$
Non-linear: $S = B_0+B_1(\text{veg. density})+B_2(\text{veg. density}^2)$

To use step-down model comparison, we would first use AIC to compare our two vegetation density models to see which best describes variation in survival. This step is exactly like the process used by program Distance to select the shape of detection functions for surveys (Buckland et al. 2001; see Chapter 19).

Once a model is selected (e.g., linear), that model is put forward with the other models for a final comparison:

Model 1 (sex): $S=B_0+B_1(\text{male})$
Model 2 (age): $S=B_0+B_1(\text{adult})$
Model 3 (vegetation): $S=B_0+B_1(\text{veg. density})$
Model 4 (null, no effect): $S=B_0$

In this simple model set, our final list of models is only reduced by one model through step-down model selection. But, you can imagine a more complicated scenario with multiple covariates that might benefit from assessing linear and non-linear effects in a first step, before the final selection process. In Chapter 19, we will discuss program Distance (Thomas et al. 2010), and this software uses a 3-step, step-down model selection process to find the best model to describe the decline in detection probability with distance from a transect or point.

To conclude our suggestions for model comparison under low levels of certainty, we suggest (*following* Arnold 2010):

- <u>If you have small sets of models:</u> use full reporting, avoid model averaging "at all costs", and provide interpretation with 85% confidence intervals and assessment of the number of parameters and relative model fit (likelihood).
- <u>If you have model sets with many variables:</u> use balanced sets of models, and report the relative importance using cumulative model weights. Use sequential, or step-down modeling to remove uninformative parameters and/or models.
- <u>For all circumstances:</u> try to avoid model averaging if at all possible. If model averaging must be used, construct your model set to prepare for the potential use of model averaging. *Do not make model averaging a last-minute thought!* And, construct models sets that are conducive to proper application of model averaging. Consider conditional model averaging to avoid the loss of information from estimates derived from unbiased maximum likelihood estimation methods.

## Conclusion

AIC is a method that allows comparison of the support provided to multiple models by our data. As the investigative biologist, we want a good, descriptive model that informs us about our population. But, we also value simplicity, because simplicity allows us to limit the variance components in a model. Parsimony suggests that, given equal explanatory value (fit, as measured by likelihood), we select the simplest explanation. AIC evaluates the trade-off between model fit and variance components to suggest which model is closest to reality. Model selection is, conceptually, a simple idea; but the realities of model selection can be very messy! We encourage you to evaluate the various information provided by your analyses (model rankings and weights, presence of uninformative parameters, cumulative model weights, and estimates of coefficients from models) to make sense of model selection and parameter estimation. Above all, plan ahead when constructing models for analysis.

## References

Anderson, D. R. 2007. Model based inference in the life sciences: a primer on evidence. Springer Science & Business Media.

Arnold, T. W. 2010. Uninformative parameters and model selection using Akaike's Information Criterion. The Journal of Wildlife Management 74: 1175-1178.

Buckland, S. T., D. R. Anderson, K. P. Burnham, J. L. Laake, D. L. Borchers, and L. Thomas. 2001. Introduction to Distance Sampling: Estimating Abundance of Biological Populations. Oxford University Press, New York.

Burnham, K. P., and D. R. Anderson. 2002. Model Selection and Multimodel Inference. Springer-Verlag, New York

Chamberlin, T. C. 1965. The method of multiple working hypotheses. Science 148(3671): 754-759.

Doherty, P. F., G. C. White, and K. P. Burnham. 2012. Comparison of model building and selection strategies. Journal of Ornithology 152: 317-323.

Hines, J.E., and J.R. Sauer. 1989. Program CONTRAST: A General Program for the Analysis of Several Survival or Recovery Rate Estimates. US Fish & Wildlife Service, Fish & Wildlife Technical Report 24, Washington, DC.

Lebreton, J. D., K. P. Burnham, J. Clobert, and D. R. Anderson. 1992. Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. Ecological Monographs 62: 67-118.

Lukacs, P. M., K. P. Burnham, and D. R. Anderson. 2010. Model selection bias and Freedman's paradox. Annals of the Institute of Statistical Mathematics 62: 117-125.

Rehme, S. E., L. A. Powell, and C. R. Allen. 2011. Multimodel inference and adaptive management. Journal of Environmental Management 92: 1360-1364.

Thomas, L., S. T. Buckland, E. A. Rexstad, J. L. Laake, S. Strindberg, S. L. Hedley, J. R. B. Bishop, T. A. Marques, and K. P. Burnham. 2010. Distance software: design and analysis of distance sampling surveys for estimating population size. The Journal of Applied Ecology, 47: 5–14.

## For more information on topics in this chapter

Conroy, M. J., and J. P. Carroll. 2009. Quantitative Conservation of Vertebrates. Wiley-Blackwell: Sussex, UK.

Cooch, E., and G. White. 2014. Chapter 4: Building and Comparing Models. *In* Program MARK: a gentle introduction, 12th edition, Cooch, E. and G. White, eds. Online: http://www.phidot.org/software/mark/docs/book/pdf/chap4.pdf

Williams, B. K., J. D. Nichols, and M. J. Conroy. 2002. Analysis and management of animal populations. Academic Press, San Diego.

## Citing this primer

Powell, L. A., and G. A. Gale. 2015. Estimation of Parameters for Animal Populations: a primer for the rest of us. Caught Napping Publications: Lincoln, NE.

*A biologist prepares to release a channel catfish (Ictalurus punctatus) with individually numbered t-bar tags (Floy type) near its dorsal fin on the Red River near Selkirk, Manitoba, Canada. Photo provided by Stephen Siddons, University of Nebraska-Lincoln (used with permission).*