does not happen (e.g., Cronin and Strong 1993; Rosenheim and Mangel 1994). This tradition has evolved in part because rate-maximizing models are easy to use and in part because for so many years there were no feasible alternatives. Our analysis suggests that dynamic state-variable models are exceedingly more likely than egg-independent models.

Our conclusions are similar to those reached by Rosenheim and Rosen using logistic regression (Hosmer and Lemeshow 1989; Collett 1991). Both logistic regression and the bootstrap methods that we used assume that the data are "representative" of the natural world. This concern usually arises when one creates bootstrap samples, but the same is true for logistic regression. The tradeoff is this. When using the bootstrap methods for model comparison, we make no assumptions about how uncertainty enters the system. The disadvantage is that we are unable to make accurate probability statements. When using logistic regression, we make distributional assumptions about the uncertainty and from these we are able to make probability statements. The disadvantage is that the assumption is made and control of the analysis is turned over to the computer, rather than having the ecological detective at the helm.

In this chapter, because we ignored the details of how uncertainty enters into the behavioral processes, we used the sum of squares to compare essentially qualitative predictions of models. In subsequent chapters, we make more assumptions about the nature of the uncertainty and are able to make more precise quantitative comparisons.

CHAPTER SEVEN

The Confrontation: Likelihood and Maximum Likelihood

OVERVIEW

The method of sum of squares can be used to find the best fit of a model to the data under minimal assumptions about the sources of uncertainty. Furthermore, goodness-of-fit profiles and bootstrap resampling of the data sets allow us to make additional inferences about the competition between different models. All of this can be done without assumptions about how uncertainty enters into the system. However, there are many cases in which the form of the probability distributions of the uncertain terms can be justified. For example, if the deviations of the data from the average very closely follow a normal distribution, then it makes sense to assume that the sources of uncertainty are normally distributed.

In such cases, we can go beyond the sum of squares and use the methods of maximum likelihood, which are discussed in this chapter. The likelihood methods discussed here allow us to calculate confidence bounds on parameters (something we could not do with the sum of squares), and to test hypotheses in the traditional manner. In addition, likelihood forms the foundation for Bayesian analysis, which is discussed in Chapter 9.

In this chapter, we use the probability distributions discussed in Chapter 3 to (i) find parameters of a given model that provide the best fit to the data (called maximum likelihood estimation), (ii) compare alternative hypotheses (by using the likelihood ratio test or its generalization to non-

nested models), and (iii) calculate confidence bounds (using the method of the likelihood profile). We now introduce these methods.

LIKELIHOOD AND MAXIMUM LIKELIHOOD

For any of the probability distributions considered in Chapter 3, the probability of observing data Y_i, given a particular parameter value p, is

$$\Pr\{Y_i|p\}. \tag{7.1}$$

The subscript on Y_i indicates that there are many possible outcomes (for example, $i = 1,2, \ldots I$), but only one parameter p. For example, suppose that Y_i follows a Poisson distribution with rate parameter r. Then in one unit of time we predict that $Y_i = k$ with probability

$$\Pr\{Y_i = k \mid \text{rate parameter} = r\} = \frac{e^{-r}r^k}{k!}. \tag{7.2}$$

This expression is also the probability of the "data" given the "hypothesis," where the "data" are k events in one unit of time and the "hypothesis" is that the rate parameter is r. When confronting models with data, we usually want to know how well the data support the alternative hypotheses. That is, after collection, the data are known but the hypotheses are still unknown. We ask, "Given these data, how likely are the possible hypotheses?"

To do this, we introduce a new symbol to denote the "likelihood" of the data given the hypothesis:

$$\mathcal{L}\{\text{data} \mid \text{hypothesis}\} \quad \text{or} \quad \mathcal{L}\{Y|p_m\}. \tag{7.3}$$

Note the subtle shift in going from Equation 7.1 to Equation 7.3: Y has no subscript because there is only one observation, but now the parameter is subscripted because there are alternative parameters (hypotheses); for example, we might have $m = 1, 2, \ldots, M$.

The key to the distinction between likelihood and probability is that with probability the hypothesis is known and the data are unknown, whereas with likelihood the data are known and the hypotheses unknown. In general, we assume that the likelihood of the data, given the hypothesis, is proportional to the probability Equation 7.1 (Edwards 1992), so the likelihood of parameter p_m, given the data Y, is

$$\mathcal{L}\{Y|p_m\} = c \ \Pr\{Y|p_m\}. \tag{7.4}$$

Also, in general, we are concerned with relative likelihoods because we mostly want to know how much more likely one set of hypotheses is relative to another set of hypotheses. In such a case, the value of the constant c is irrelevant and we set $c = 1$. Then the likelihood of the data, given the hypothesis, is equal to the probability of the data, given the hypothesis. Note that although it must be true that if the parameter p is fixed $\Sigma_{i=1}^{I} \Pr\{Y_i|p\} = 1$, when the data Y are fixed, the sum over the possible parameters $\Sigma_{m=1}^{M} \mathcal{L}\{Y|p_m\}$ need not even be finite, let alone equal to 1. It may be helpful to think of likelihood as a kind of unnormalized probability.

For example, suppose that the data were $k = 4$ events in one unit of time. For the Poisson model, Equation 7.2, the likelihood is

$$\mathcal{L}\{4|r\} = \frac{e^{-r}r^4}{4!}. \tag{7.5}$$

If the data were six events in one unit of time, then

$$\mathcal{L}\{6|r\} = \frac{e^{-r}r^6}{6!}. \tag{7.6}$$

By plotting the likelihood as function of r (Figure 7.1a), we get a sense of the range of parameters for which the observations are probable. When looking at this figure, remember that the comparisons are within a particular value of the data and not between different values of the data. For

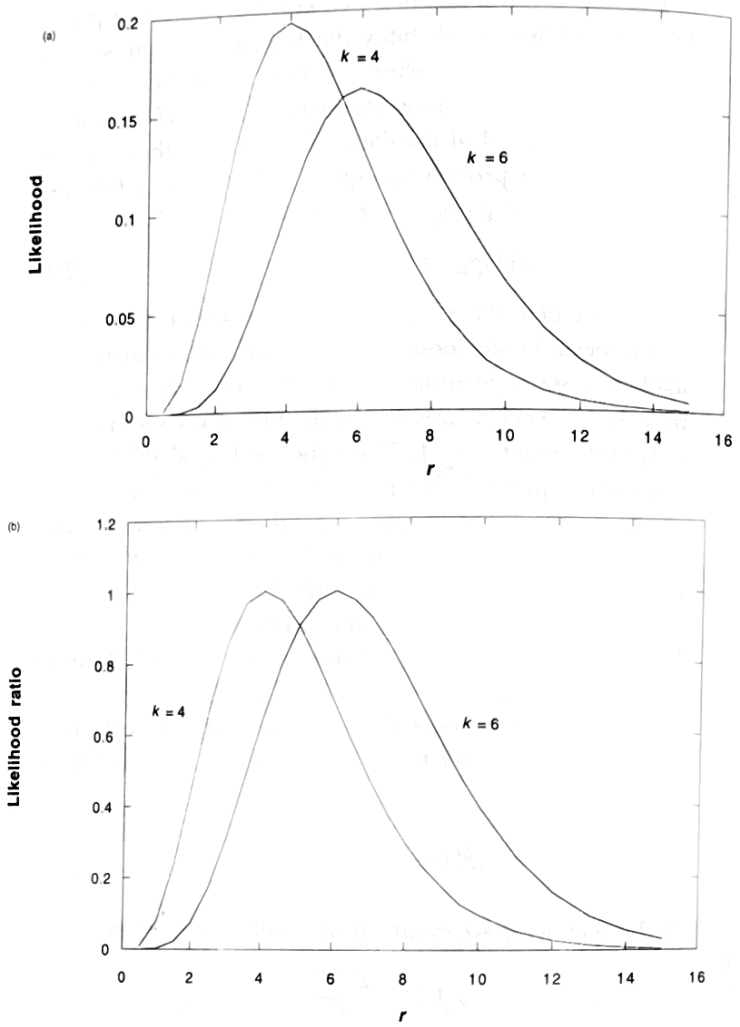
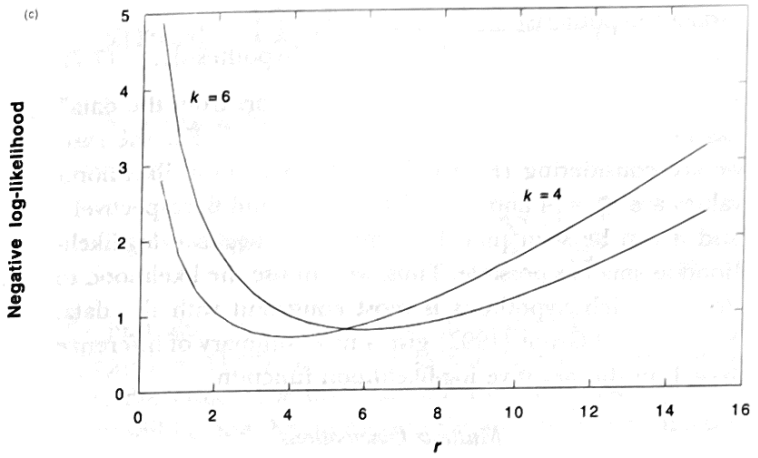


FIGURE 7.1. (a) The likelihood $\mathcal{L}\{k|r\} = e^{-r}r^k/k!$ for $k = 4$ and 6. (b) The likelihood ratio $\mathcal{L}\{k|r\}/\mathcal{L}\{k|r^*\}$, where r^* is the value of the parameter that maximizes the likelihood, for $k = 4$ and 6. (c) The negative log-likelihoods.



this reason, it is often helpful to scale the likelihoods relative to the parameter value that makes the likelihood as large as possible (Figure 7.1b). For example, when $k = 4$, we see that the most likely value of the parameter is $r = 4$, and that values of r in the range [2,7] are at least half as likely as the most likely parameter. Similarly, when $k = 6$, the most likely value of the parameter is 6 and values of r in the range [4,10] are at least half as likely as the most likely parameter. The parameter that makes the likelihood as large as possible is called the *maximum likelihood estimate (MLE)*.

Because likelihoods may be very small numbers, the tradition is to use the logarithm of the likelihood, called the log-likelihood, for comparisons. This is also called the *support function* (Edwards 1992).

In analogy to the sum of squares, we use the negative of the logarithm of the likelihood, so that the most likely value of the parameter is the one that makes the negative log-likelihood as small as possible:

$$\mathbf{L}\{\text{data} \mid \text{hypothesis}\} = -\log(\mathcal{L}\{\text{data} \mid \text{hypothesis}\}). \quad (7.7)$$

Then the hypothesis with the most "support from the data" has the smallest value of $\mathbf{L}\{\text{data} \mid \text{hypothesis}\}$. For the case we are considering (Figure 7.1c), the maximum likelihood values are $r^* = 4$ and $r^* = 6$ for $k = 4$ and 6, respectively, and it can be seen that these make the negative log-likelihood as small as possible. Thus, we can use the likelihood to decide which hypothesis is most consistent with the data. Schnute and Groot (1992) give a nice summary of inference based on the negative log-likelihood function.

Multiple Observations

We often have multiple observations of different types of data. Since likelihoods are determined from probabilities, the likelihood of a set of independent observations is the product of the likelihoods of the individual observations. Thus,

$$\mathcal{L}\{Y_1,Y_2,Y_3|p\} = \mathcal{L}\{Y_1|p\}\,\mathcal{L}\{Y_2|p\}\,\mathcal{L}\{Y_3|p\}, \quad (7.8)$$

and since logarithms are additive, the negative log-likelihoods add:

$$\mathbf{L}\{Y_1,Y_2,Y_3|p\} = \mathbf{L}\{Y_1|p\} + \mathbf{L}\{Y_2|p\} + \mathbf{L}\{Y_3|p\}. \quad (7.9)$$

Thus, likelihood allows the inclusion of different types of information in a single framework. If a model predicts several different types of observations, we can use likelihood to determine the extent to which the model is consistent with all of the observations.

Maximum Likelihood and Sum of Squares May Be the Same

One interesting feature of the normal distribution is that the negative log-likelihood and the sum of squares will be minimized at the same values of the parameters. To see this, we begin with the likelihood for n observations $\{Y_i\}$ which follow a normal distribution with mean m and variance σ^2:

$$\mathcal{L}\{Y|m,\sigma\} = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(Y_i - m)^2}{2\sigma^2}\right). \quad (7.10)$$

The negative log-likelihood is

$$\mathbf{L}\{Y|m,\sigma\} = n[\log(\sigma) + \tfrac{1}{2}\log(2\pi)] + \sum_{i=1}^{n} \frac{(Y_i - m)^2}{2\sigma^2}. \quad (7.11)$$

To find the value of m that minimizes $\mathbf{L}$, notice that $n[\log(\sigma) + (1/2)\log(2\pi)]$ does not depend on m. Therefore, the value of m that minimizes the negative log-likelihood will be one that minimizes the sum on the right-hand side, which is the square deviation between the predicted (m) and observed (Y_i) values. Many of the familiar problems in regression and analysis of variance assume normal distributions, and therefore the estimated parameters will be the same using sum of squares or maximum likelihood.

Calculating Averages Using Maximum Likelihood

As an easy introduction to how to use maximum likelihood, let us consider the following set of data. Suppose that the heights (in cm) of ten people are 171, 168, 180, 190, 169, 172, 162, 181, 181, and 177. Also assume that we know that height is normally distributed with standard deviation 10 cm. Therefore the likelihood of any individual height Y, if the true mean of the population is m, is

$$\mathcal{L}\{Y|m\} = \frac{1}{10\sqrt{2\pi}} \exp\left(-\frac{(Y - m)^2}{200}\right), \quad (7.12)$$

and the negative log-likelihood for 10 of the ten heights is

$$\mathbf{L}\{Y|m\} = n[\log(10) + \tfrac{1}{2}\log(2\pi)] + \sum_{i=1}^{n} \frac{(Y_i - m)^2}{200}. \quad (7.13)$$

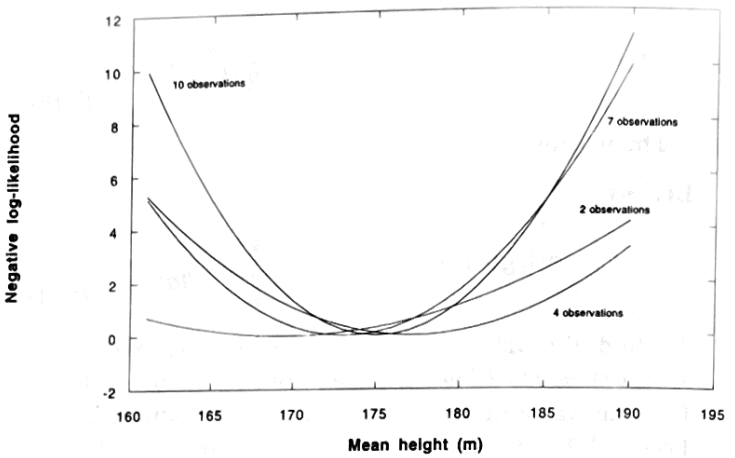


FIGURE 7.2. The negative log-likelihood (scaled so that the minimum is at 0) for the average height of the population when 2, 4, 7, or 10 observations are used.

Figure 7.2 shows the negative log-likelihood for different values of m for the data set using the first 2, 4, 7, and finally all 10 observations. In all cases, the minimum **L** has been subtracted from the **L** so that they are all plotted with 0.0 as a minimum. When we use only two data points, the curve is very flat, that is, the alternative hypotheses about m have similar likelihoods. As the number of data points used increases, the negative log-likelihood becomes steeper, which indicates that we have more confidence in our knowledge about m. Later in this chapter, we show how to find confidence intervals from **L**.

DETERMINING THE APPROPRIATE LIKELIHOOD

At this point, you may ask, "Given data and hypotheses, what likelihood function should I use?" If you find yourself in this position, then you have not completely specified the model. In particular, you may have a deterministic model but not a stochastic one, because a fully specified stochastic

model contains a hypothesis about the way in which randomness enters into the system. If you have not done so, you should return to Chapter 3 and formulate hypotheses about the stochastic components of your model. Ask questions such as: Is there process uncertainty? If so, what kind of distribution is appropriate? Is there observation uncertainty? If so, what kind of distribution is appropriate?

This choice is often made on first principles from the basic distributions described in Chapter 3. For instance, when dealing with simple proportions, the binomial distribution naturally might occur. Data that fall into several possible categories can be described by a multinomial distribution. Counts of rare events could be Poisson or negatively binomially distributed. Quantities that result from the sum of events are often normally distributed, and quantities that result from a series of multiplicative probabilities frequently are log-normal.

You may be able to use the data to distinguish between different probability models for the stochasticity in your system. Different probability models can be thought of as competing hypotheses in exactly the same way that different parameter values are competing hypotheses. Remember that the model consists not only of the deterministic equations, but also of the assumptions about randomness. More simply, examine the residuals, as we did in Chapter 4, to see if there is a systematic pattern to the difference between the model and the data. For example, if the residuals are symmetrically distributed, the normal distribution may be appropriate, but strong skewness in residuals suggests a log-normal distribution.

Observation and Process Uncertainty

To illustrate the distinction between observation and process uncertainty, imagine a population growth process. If there is only observation uncertainty, then the population dynamics (births and deaths) will be deterministic, but we are unable to accurately estimate population abundance.

Observation uncertainty does not propagate in time. If we underestimate the population in one year, it does not affect the population the next year (the organisms do not know if we overcount or undercount them). As long as our observation uncertainties are independent from year to year, we will be just as likely to overestimate or underestimate the population next year.

If we have process uncertainty but not observation uncertainty, then we estimate population size perfectly (as in many laboratory populations), but the processes of birth and death have random components. If the process uncertainty reduces population size in one year (due to poor births or survival), then the population will be smaller the next year; process uncertainty will propagate over time.

Suppose that we observe a system in which the variable Y depends linearly on the independent variable X. We might begin by writing

$$Y = p_0 + p_1 X + W. \tag{7.14}$$

In this equation, p_0 and p_1 are the parameters to be determined from the data, and W is the process uncertainty (for simplicity, we will not subscript the variables by time or observation number in this section). Now let us explicitly recognize that the observed values of the independent and dependent variables, X_{obs} and Y_{obs}, respectively, also involve observation uncertainty by writing

$$\begin{aligned} Y_{obs} &= Y + V_1, \\ X_{obs} &= X + V_2, \end{aligned} \tag{7.15}$$

where V_1 and V_2 are the observation uncertainties. We combine Equations 7.14 and 7.15 as

$$\begin{aligned} Y_{obs} &= p_0 + p_1 X + W + V_1 \\ &= p_0 + p_1 (X_{obs} - V_2) + W + V_1 \\ &= p_0 + p_1 X_{obs} + Z, \end{aligned} \tag{7.16}$$

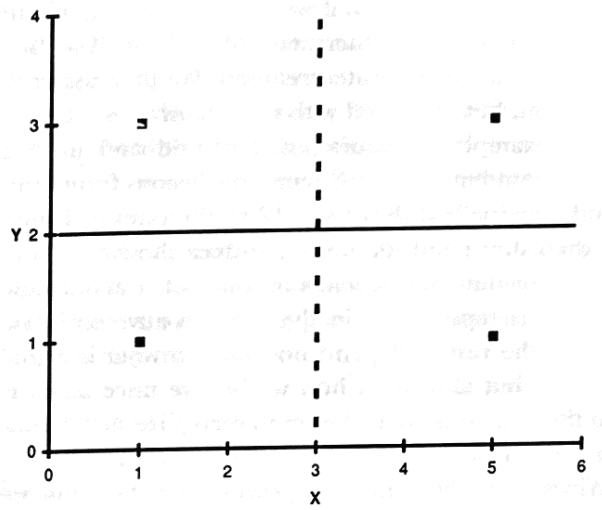


FIGURE 7.3. The four "observations" represent a possible set of data relating Y to X. The horizontal line is the interpretation if we believe that X is measured perfectly but that there is process uncertainty. The vertical line is the intepretation if we believe that there is no process uncertainty, but that X is measured imperfectly.

where $Z = W + V_1 - p_1 V_2$ is the "total uncertainty." This is the regression equation usually encountered in statistics books, where it is typically assumed that X is observed perfectly and that Y is subject to process uncertainty.

Why should one think about the sources of uncertainty, particularly to separate process and observation uncertainty, when it is possible to use the last line of Equation 7.16 and ignore the issue entirely? Schnute (1987) illustrates the importance of thinking about the sources of uncertainty. Suppose we have four measurements (Figure 7.3). If we believe that there is no observation uncertainty ($V_1 = V_2 = 0$) but only process uncertainty, then the horizontal line is the appropriate interpretation of the data. In such a case, we assert that Y is independent of X, but because of process uncertainty we observe different values for Y at different values

of X. On the other hand, if we believe that the only uncertainty occurs with the observation of X ($V_1 = W = 0$), then the vertical line is the interpretation. We then assert that X is constant, but measured with uncertainty.

This example, of course, is contrived and most of us would not attempt to draw many conclusions from four data points, especially if they looked like the ones in Figure 7.3. On the other hand, the example does show how our *interpretation* of the data depends on our belief about how randomness is represented in the data. In any comparison of models, the results depend not only on what is actually in the data, but also upon how we believe uncertainty enters into the data. It is always better to recognize such limitations from the outset.

When only observation or process uncertainty is present, we can estimate the amount of variation from the data. For example, in a standard linear regression (Equation 7.16) we assume no observation uncertainty and usually estimate the slope, the intercept, and the variance of the process uncertainty. However, when X is measured imprecisely, it is impossible to estimate the variances of both the observation and process uncertainties. In particular, if both observation and process uncertainty are present, we must either specify the variance of one of the two, or we must specify the ratio of the variances (Schnute 1987). However, even once we specify one of the variances or the ratio of variances, the joint estimation of observation and process uncertainty is computationally difficult and frequently ambiguous. We recommend the following:

1. Whenever possible, conduct independent experiments to determine the magnitude of observation and process uncertainties so that you will not have to estimate these from the data used in the comparison of models.
2. If possible, eliminate observation uncertainty by good experimental design or instrumentation.

3. Compare models and/or estimate parameters using the alternative, extreme assumptions of no observation uncertainty or no process uncertainty.
4. If there is little difference between your conclusions using the different assumptions in step 3, you can stop worrying about the issue. If, however, there are major differences in the results of the analysis depending on the assumption in step 3, you must either delve deeply into the statistical literature on the subject (Schnute 1987 is a good starting point) or redesign the experiments and try again.

Likelihoods for Observation and Process Uncertainty

While simultaneous estimation of process and observation uncertainty can be complex, the special cases in which only one is present can be analyzed in a straightforward manner.

We begin with a general deterministic model for Y, based on independent variables X and parameters p,

$$Y_{\text{det}} = f(X,p) \tag{7.17}$$

where $f(X,p)$ is assumed to be known. Now assume that the observed value of Y depends on the deterministic value and the process uncertainty W, so that

$$Y_{\text{obs}} = Y_{\text{det}} + W. \tag{7.18}$$

The deviation D between the observed and predicted (deterministic) values of the dependent variable is

$$D = Y_{\text{obs}} - Y_{\text{det}} = W. \tag{7.19}$$

Thus, the probability distribution of the deviation is exactly the same as the probability distribution W. For example, if W is normally distributed, the negative log-likelihood (using t as a subscript for individual observations of X and Y) is

$$\mathbf{L}_t = \log(\sigma) + \frac{1}{2}\log(2\pi) + \frac{[Y_{\text{obs},t} - f(X_t,p)]^2}{2\sigma^2}. \tag{7.20}$$

Now assume that X is measured imprecisely but that Y is measured exactly. In that case, the statistically interesting questions involve the value of X, which is related to Y through the inverse function

$$X = f^{-1}(Y,p). \tag{7.21}$$

For example, if $Y = pX$, then

$$f^{-1}(Y,p) = \frac{Y}{p}. \tag{7.22}$$

That is, the inverse function involves "solving for x in terms of y." This cannot always be done explicitly, and in some cases—involving nonlinear functions—the inverse function may not exist at all.

The observed value of X is then

$$X_{\text{obs}} = f^{-1}(Y,p) + V, \tag{7.23}$$

where V is the observation uncertainty. Given Y, we calculate the predicted value of X (remember the model is deterministic), and the difference between the observed X and the predicted value from the inverse model is the value of the observation uncertainty. For example, if V were normally distributed with mean 0 and variance σ^2, the negative log-likelihood would be

$$\mathbf{L}_t = \log(\sigma) + \frac{1}{2}\log(2\pi) + \frac{[X_{\text{obs},t} - f^{-1}(Y_t,p)]^2}{2\sigma^2}. \tag{7.24}$$

Linear regression models are a special case of this analysis in which there is a straightforward inverse model. For a linear regression,

$$Y_{\text{det}} = f(X,p) = p_1 + p_2X, \tag{7.25}$$

the inverse model is

$$f^{-1}(Y,p) = \frac{Y - p_1}{p_2}. \tag{7.26}$$

An ecological example of the linear regression Equation 7.25 is the simple model of population dynamics with survival (s) and births (b), process (W_t), and observation uncertainties (V_t) that are normally distributed with mean 0 and variance σ_w or σ_v, respectively:

$$\begin{aligned} N_{t+1} &= sN_t + b + W_t, \\ N_{\text{obs},t} &= N_t + V_t. \end{aligned} \tag{7.27}$$

When there is only process uncertainty, N_t is measured perfectly and the only stochastic element affects N_{t+1}. The negative log-likelihood is

$$\mathbf{L}_t = \log(\sigma_W) + \frac{1}{2}\log(2\pi) + \frac{(N_{t+1} - b - sN_t)^2}{2\sigma_W^2}. \tag{7.28}$$

only process uncertainty

On the other hand, if we assume only observation uncertainty, we use the inverse function method to write

$$N_{\text{obs},t} = -\frac{b}{s} + \frac{1}{s}N_{t+1} + V_t, \tag{7.29}$$

and the negative log-likelihood is

$$\mathbf{L}_t = \log(\sigma_V) + \frac{1}{2}\log(2\pi) + \frac{[N_{\text{obs},t} + b/s - (1/s)\,N_{\text{obs},t+1}]^2}{2\sigma_V^2}. \tag{7.30}$$

only obs. uncertainty

The likelihoods in Equations 7.28 and 7.30 refer to only a single time period. The natural next question is: What should be done when time periods are linked?

Considerations for Dynamic Models

The ecological detective often deals with observations that are a time series about the state of the system and perturbations to the system. Such time series commonly arise in wildlife, fisheries, and forestry. When the data are a time series, the model must perforce be a dynamic one in which

the state of the system at a given time is linked with its values at previous times. In this section, we shall illustrate the special considerations that arise in such a case. To illustrate the ideas, we use the discrete logistic equation

$$N_{t+1} = N_t + rN_t\left(1 - \frac{N_t}{K}\right). \tag{7.31}$$

In this equation, N_t is the population size at time t, r is the maximum possible per capita growth rate, and K is the carrying capacity. We can include additions or removals (C_t) from the population to obtain

$$N_{t+1} = N_t + rN_t\left(1 - \frac{N_t}{K}\right) - C_t. \tag{7.32}$$

Next, we must specify the nature of the observation and process uncertainty for this model. When the logistic model is used in practice, it is commonly (although certainly not universally) assumed that both observation and process uncertainties are log-normally distributed. This means, for example, that we assume that the observation is

$$N_{\text{obs},t} = N_t V,$$

$$V = \exp\left(Z\sigma_V - \frac{\sigma_V^2}{2}\right), \tag{7.33}$$

where Z is normally distributed with a mean of zero and a standard deviation of 1, and σ_V is the standard deviation of the observation uncertainty (see Equation 3.68 ff. to justify the formulation).

Process uncertainty is included in a similar manner:

$$N_{t+1} = W_t\left[N_t + rN_t\left(1 - \frac{N_t}{K}\right) - C_t\right],$$

$$W_t = \exp\left(Z\sigma_W - \frac{\sigma_W^2}{2}\right). \tag{7.34}$$

A scenario described by this model might be the discovery of a previously unfished resource, its overexploitation, and

subsequent reductions in catch to correct the problem. To describe this situation, we could use the Monte Carlo method to generate data in ten time periods (starting with an unperturbed population), allow harvesting of half of the population at times 3, 4, and 5 (the overexploitation), and reduce the harvest rate to almost zero for the last four time periods (the "management action").

Assuming the parameters $r = 0.5$, $K = 1000$, $\sigma_W = 0.1$, and $\sigma_V = 0.1$, a pseudocode is:

Pseudocode 7.1

1. Input values of the parameters r, K, σ_W, and σ_V.
2. Set the initial value of population at K.
3. Calculate population size next year based on the logistic equation with process uncertainty, harvesting half of the population at times 3, 4, and 5.
4. Calculate the observed population at each time period.
5. Repeat steps 3 and 4 for ten years.

The Monte Carlo method provides a trajectory of population size over time (Figure 7.4). Assuming only observation uncertainty means that we should use Equation 7.33. The deviation between the observed and true values of the logarithm of population size is

$$D_t = \log(N_{\text{obs},t}) - \log(N_t) + \frac{\sigma_V^2}{2}, \tag{7.35}$$

and using Equation 7.33,

$$D_t = [\log(N_t) + \log(V)] - \log(N_t) + \frac{\sigma_V^2}{2}$$

$$= \log(V) + \frac{\sigma_V^2}{2}$$

$$= \left(Z\sigma_V - \frac{\sigma_V^2}{2}\right) + \frac{\sigma_V^2}{2} = Z\sigma_V. \tag{7.36}$$

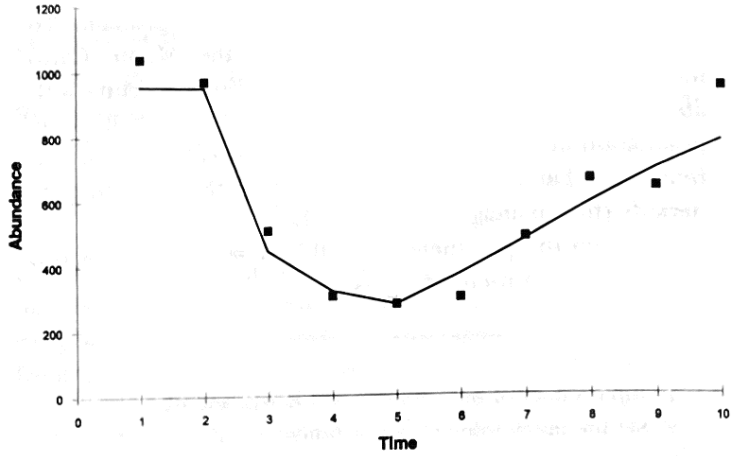


FIGURE 7.4. The Monte Carlo data (squares) for the logistic model Equation 7.34. The harvest rate is 50% during periods 3, 4, and 5, and 0.01 at other times. The line shows the best fit of the model assuming observation uncertainty. The estimated parameters are $r = 0.47$ and $K = 960$.

Thus, D_t is normally distributed with mean 0 and variance σ_V^2, so that the likelihood of a deviation of size d_t is

$$\mathcal{L} = \frac{1}{\sqrt{2\pi\sigma_V^2}} \exp\left(-\frac{d_t^2}{2\sigma_V^2}\right), \tag{7.37}$$

and the negative log-likelihood for the observation at time t is

$$\mathbf{L}_t = \log(\sigma_V) + \frac{1}{2}\log(2\pi) + \frac{d_t^2}{2\sigma_V^2}. \tag{7.38}$$

This is analogous to Equation 7.28. The negative log-likelihood for all of the data (across multiple periods) is the sum of the $\mathbf{L}_t$ from Equation 7.38

Given the data and particular values of r, K, and σ_V, we can evaluate the likelihood of that set of parameters. Alternatively, we can select the parameters that make the negative log-likelihood as small as possible and call these the "best-fit" parameters. A pseudocode to do these calculations is:

Pseudocode 7.2

1. Input data values for observed population size.
2. For specified values of r and K, systematically search over individual r and K values and generate predicted deterministic population sizes using Equation 7.32.
3. Calculate the deviation at each time period using Equation 7.36.
4. Calculate the negative log-likelihood of the deviations using Equation 7.38.
5. Sum the $\mathbf{L}_t$ over t to obtain the negative log-likelihood for the combination of r and K in question.
6. See which values of r and K lead to the smallest total likelihood.

By implementing this pseudocode, we predict a deterministic trajectory of the population conditioned on the parameters of the model and the starting population size (Figure 7.4). We assumed that the population is initially at carrying capacity; if one does not know that $N_0 = K$, the starting population size must also be estimated.

If there is only process uncertainty, the dynamic model becomes

$$\begin{aligned} N_{\text{obs},t} &= N_t, \\ N_{t+1} &= W_t\{N_{\text{obs},t} + rN_{\text{obs},t}[1 - (N_{\text{obs},t}/K)] - C_t\}, \\ W_t &= \exp\left(Z\sigma_W - \frac{\sigma_W^2}{2}\right). \end{aligned} \tag{7.39}$$

The deviation is defined in a similar manner:

$$\begin{aligned} D_t &= \log(N_{t+1}) - \log(N_{\text{obs},t}) + \frac{\sigma_W^2}{2} \\ &= \log(W_t) - \frac{\sigma_W^2}{2} = Z\sigma_W. \end{aligned} \tag{7.40}$$

The key difference between the deviations in Equations 7.36 and 7.40 is that in Equation 7.40 the predicted value depends on the *observed* value in the previous time period, rather than on the *predicted* value in the previous time period. The negative log-likelihood for a single period is analogous to Equation 7.38:

$$\mathbf{L}_t = \log(\sigma_W) + \frac{1}{2}\log(2\pi) + \frac{d_t^2}{2\sigma_W^2}. \qquad (7.41)$$

Once again, we can find the values of r and K that give the best fit to the data. To do so, we need a different pseudocode:

Pseudocode 7.3

1. Input the data values for observed population size.
2. For specified values of r and K, generate predicted population sizes using Equation 7.39.
3. Calculate the deviation at each time period using Equation 7.40.
4. Calculate the negative log-likelihood of deviations using Equation 7.41.
5. Sum L_t across t to obtain the negative log-likelihood for the combination of r and K in question.
6. See which values of r and K lead to the smallest total likelihood.

The results (Figure 7.5) show that assuming only one kind of uncertainty provides a reasonably good fit to the data, although clearly neither of these models is "correct." This is gratifying, since the two assumptions that we considered are the "extremes" that bracket the true situation. As a general rule, when the data are informative, the assumption about how uncertainty enters does not matter greatly. In practice, the assumptions of only observation uncertainty or only process uncertainty have specific strengths and weaknesses. For instance, in order to use the assumption of pro-

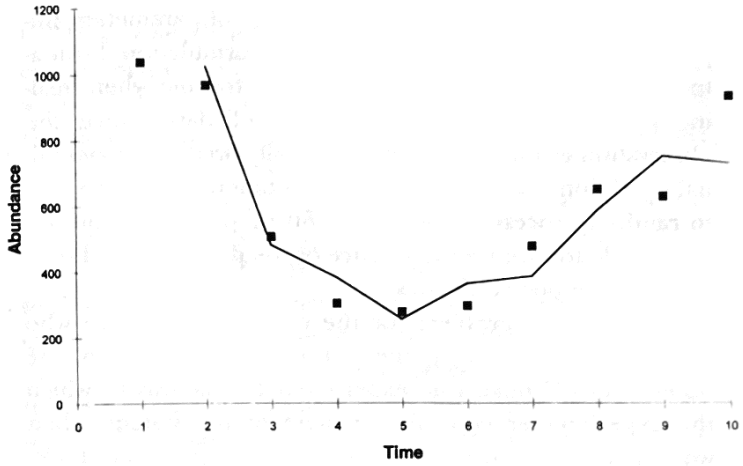


FIGURE 7.5. The same Monte Carlo data as in Figure 7.4 and the fit of the model assuming process uncertainty. The estimated parameters are $r = 0.44$ and $K = 1023$.

cess uncertainty, we should observe each state variable at each occasion; otherwise the computation of the predicted value at future times becomes much more complex. In contrast, the assumption of only observation uncertainty makes no specific requirements about how much of the state can be observed, nor how often it is observed. The likelihood can be calculated from a single observation at any time. The major limitation of the observation uncertainty assumption is the need to specify the starting state. For example, above we assumed that $N_0 = K$. If we did not have this additional information, we would have to estimate an additional parameter, N_0.

The importance of the starting condition is accentuated when the model exhibits chaotic behavior, since the time trajectory of a chaotic model is highly sensitive to the starting conditions. In practice (Adkison 1992), estimators based on observation uncertainty cannot be used in chaotic models. Many models, including the discrete logistic, can

exhibit chaotic behavior over some range of parameters, implying that particular care is needed in formulation. Estimators based on observation tend to have trouble when dealing with long, complex time series of data. Since the observation estimator is deterministically predicted from initial conditions, if the time series has numerous changes due to random processes, observation-fitting procedures are often unable to capture the essence of the dynamics, and thus may provide poor estimates.

An additional problem for the ecological detective who works with time series is the lack of independence of the observations. Unlike true experimental situations in which the experimenter controls the state of the system, when working with time series the most we can hope for are informative perturbations. The data from one time to the next are not independent, and biases in parameters may be introduced. In practice, it is rarely possible to calculate a bias correction factor, and we recommend the use of Monte Carlo simulations to explore the sensitivity of results to the time series bias. Such simulations can be accomplished by taking the parameters estimated from the data, using them as "true" values in a Monte Carlo model, generating a few hundred data sets, and then seeing how accurately one can estimate the "true" parameters.

MODEL SELECTION USING LIKELIHOODS

We are now ready to consider the resolution of the contest between different models for the same phenomenon, arbitrated by the data, using likelihood as the criterion. Imagine a number of models M_1, M_2, . . . , in which model M_i has parameters p_{i1},p_{i2}, . . . , and that we have determined the best-fit values of the parameters. In most situations, a model will rarely win the contest outright, but rather each additional experiment or observation changes our relative belief in competing models. The treatment of relative

belief is covered in Chapter 9 on Bayesian methods. However, in many applications, and for many scientific journals, we must decide a winner in the contest; that is, we must choose which model appears to be "best" given the available data.

In the discussion that follows, we shall use the words "model" and "hypothesis" interchangeably. The first principle we use is that of likelihood, which quantifies how consistent a particular hypothesis is with the observations. As a general rule, the best model is the one that has the highest likelihood. When we have many competing hypotheses with the same number, of parameters, the hypothesis with the highest likelihood is the "best" one. For example, in a regression model, different slopes and intercepts are competing hypotheses, and the slope and intercept that have the highest likelihood are the best estimates of the true slope and intercept. An interesting evolutionary application of model selection using likelihood is the work of Sanderson and Donoghue (1994) in a study of the origin of angiosperms.

The Likelihood Ratio Test for Nested Models

Commonly, the competing models do not have the same number of parameters, and a model with more parameters has an intrinsic advantage in being able to fit data. How do we referee a contest between unequal competitors, for example, between a model with one parameter and a model with two parameters? Here we rely on a second principle, known as the likelihood ratio test. The likelihood ratio test is based on the following result from theoretical statistics (Kendall and Stewart 1979, 240 ff.). Imagine two nested models, A and B, in which B is the more complicated model. That is, model B has more parameters and collapses to model A when some of them are set equal to 0. Denote the data by Y and the negative log-likelihoods of the data, given the models, by $\mathbf{L}\{Y|M_A\}$ and $\mathbf{L}\{Y|M_B\}$. We assume that

the more complicated model fits the data better, so that $\mathbf{L}\{Y|M_A\} > \mathbf{L}\{Y|M_B\}$.

The result of statistical theory is that

$$\mathcal{R} = 2[\mathbf{L}(Y|M_A) - \mathbf{L}(Y|M_B)] \tag{7.42}$$

has a chi-square distribution (refer to Chapter 3), with the degrees of freedom equal to the difference in the number of parameters between models B and A. Because the right-hand side of Equation 7.42 involves log-likelihoods, $\mathcal{R}$ is the ratio of the logarithm of the likelihoods, and this procedure is called the likelihood ratio test.

It is perhaps easiest to understand how Equation 7.42 is used for the case of comparing the likelihood associated with a maximum likelihood estimate (MLE) parameter with the likelihood for other values of the parameters. We replace $\mathbf{L}\{Y|M_B\}$ with $\mathbf{L}\{Y|p_{MLE}\}$ and $\mathbf{L}\{Y|M_A\}$ with $\mathbf{L}\{Y|p\}$, where p is another value of the parameter. The difference $\mathcal{R}(p)$ now has a chi-square distribution with one degree of freedom, because we have one fitted parameter. If we plot the probability that $\mathcal{R}(p)$ is less than z as p varies, we obtain a function that is symmetric around p_{MLE} and which is zero when $p = p_{MLE}$ (Figure 7.6). The thin parabolic line is the difference in log-likelihood between p_{MLE} and p. The thick funnel-shaped line is the probability that the χ^2 random variable is less than $z = (p - p_{MLE})^2$. This plot rises to 1 as the difference between p and p_{MLE} increases. We construct confidence intervals by noting that $\Pr\{\chi^2 < 3.84\} = 0.95$. Consequently, if model B has one more parameter than model A, twice the difference in negative log-likelihoods must be greater than 3.84 for model B to be significantly better at the 0.05 level. We construct the confidence intervals by drawing a horizontal line at the desired confidence level (e.g., 95%) and seeing where the line intersects the χ^2 probability curve. In the case of Figure 7.6, we see that this occurs at p values of roughly 1 and 9. The likelihood ratio test allows us to examine models of increasing complexity to

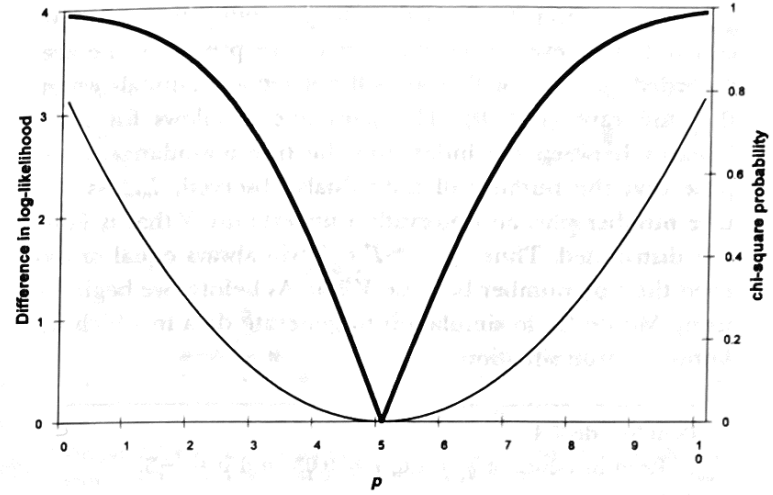


FIGURE 7.6. The relationship between the negative log-likelihood and the χ^2 value used in the likelihood ratio test. The thin line is the difference in negative log-likelihoods between the best-fit parameter ($p_{MLE} = 5$) and other values of the parameter. The thick line is the probability that the χ^2 random variable is less than the deviation p.

determine if the more complex model provides a significantly better fit.

An Ecological Scenario. To illustrate the use of likelihood for model selection, consider a model (Schnute 1987) relating the number of animals recorded by observers in a survey (an index of abundance I) to the true abundance D by

$$I = \max\left\{0, \frac{p + qD}{1 + rD}\right\}, \tag{7.43}$$

where p, q, and r are parameters. We obtain a series of nested models by setting one or all of the parameters equal to 0. In the simplest case, when $r = p = 0$, the index is proportional to the number of animals present with constant of proportionality q,

$$I = qD. \tag{7.44}$$

The parameter p allows for the possibility that we may conclude that even when no animals are present some are recorded ($p > 0$), or that we will not see any animals when they are rare ($p < 0$). The parameter r allows for nonlinearity between the index and the true abundance. Suppose that the number of individuals observed, I_{obs}, is the true number plus an observation uncertainty V that is Poisson distributed. Thus, $I_{obs} = I + V$ will always equal or exceed the true number because $V \geq 0$. As before, we begin by using Monte Carlo simulation to generate data in which we know the true situation:

Pseudocode 7.4

1. Read in values of $q = 1.0$, $r = 0.03$, and $p = -3$.
2. Set $D = 1$.
3. Calculate the deterministic values from Equation 7.43.
4. Calculate the actual observation by adding a Poisson distributed term to the result from step 3.
5. Increment the value of D by 1 and repeat steps 3 and 4 until $D > 20$.

The squares in Figure 7.7 are the data that result from this pseudocode (Table 7.1). The dashed line is the true relationship between the index and abundance. As is typical of Poisson processes (in which the variance is equal to the mean), there is more variability at higher expected values of the index. There are four possible models:

Model A: Only q determines the relationship (i.e., p and r are assumed to be equal to zero) between D and I

Model B: The parameters q and p determine the relationship between D and I

Model C: The parameters q and r determine the relationship between D and I

Model D: All three parameters determine the relationship between D and I

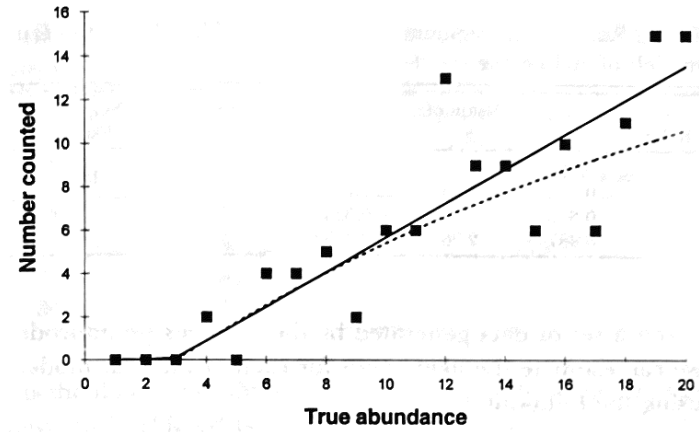


FIGURE 7.7. One set of data (squares) generated from Pseudocode 7.4 with $q = 1$, $r = 0.03$, and $p = -3$. The dashed line shows the true relationship and the solid line shows the linear model fit to the data.

TABLE 7.1. Data generated from Pseudocode 7.4.

Density	Index from Equation 7.43	Number observed
1	0	0
2	0	0
3	0	0
4	0.89	2
5	1.74	0
6	2.54	4
7	3.31	4
8	4.03	5
9	4.72	2
10	5.38	6
11	6.02	6
12	6.62	13
13	7.19	9
14	7.75	9
15	8.28	6
16	8.78	10
17	9.27	6
18	9.74	11
19	10.19	15
20	10.63	15

TABLE 7.2. Parameters and negative log-likelihoods for the four models of abundance.

Model	Value of: q	p	r	Number of parameters	Negative log-likelihood
A	0.586	—	—	1	42.47
B	0.793	−2.29	—	2	38.38
C	0.393	—	−0.023	2	40.92
D	0.987	−2.96	0.0157	3	38.22

Given a set of data generated by the previous pseudocode, we can estimate the likelihoods for each of the four models using the following pseudocode:

Pseudocode 7.5

1. Input the data as in Table 7.1 and starting values for the parameters q, p, and r.
2. Specify which model is to be used to make predictions.
3. Compute the likelihood as follows:
   a. Cycle from $D = 1$ to 20.
   b. Calculate the predicted abundance I_{pre} using Equation 7.43.
   c. Calculate the negative log-likelihood of observing I_{obs} given I_{pre} and add this negative log-likelihood to the total negative log-likelihood.
   d. Repeat steps a–c for each data point.
4. Sum negative log-liklihood for each data point.
5. Repeat steps 2–4 for each model.

We then combine the likelihood calculation with a nonlinear-function minimization routine to calculate the best estimates for each model (Table 7.2). Model B reduces the negative log-likelihood by over four units by adding one parameter. Since twice the difference in likelihoods must be at least 3.84 for the models to differ at the 0.05 level, the difference in the log-likelihoods between model A and model B is clearly significant. Models B and C have the same number of parameters, so model B is clearly preferred. Model D

TABLE 7.3. Number of times in one hundred Monte Carlo trials that each of the four abundance models was selected.

Model	Parameters	Number of times selected with one hundred Monte Carlo data sets
A	q only	14
B	q and p	79
C	q and r	0
D	q, p, and r	7

fits the data better than model B, but the difference in negative log-likelihood is very small, and not significant according to the likelihood ratio test. Therefore we conclude that for this set of data model B is the "best."

In this particular case (for the data shown in Figure 7.7), the estimation procedure failed to detect the nonlinearity between the index abundance and real abundance but did detect the non-zero intercept. When we repeat this procedure with many different Monte Carlo–generated sets of data, we find quite frequently that model A is preferred (Table 7.3).

Akaike Information Criterion (AIC) for Non-nested Models

The likelihood ratio test provides a simple and powerful format for comparing alternative models, but requires that the models being compared be nested, that is, the more complex model reduces to the simpler model by setting parameters equal to 0. When dealing with non-nested models, the Akaike information criterion (AIC) is normally used (Akaike 1973; Sakamota et al. 1986). Whereas the likelihood ratio test is based on an inferential criterion, the AIC is based on an optimization criterion (Akaike 1985, 1992; de-Leeuw 1992).

The AIC for model M_i with p_i parameters is

$$A_i = \mathbf{L}(Y|M_i) + 2p_i. \tag{7.45}$$

The model selection criterion is that the best model is the one that has the lowest AIC. By adding 2 to the negative log-likelihood for every free parameter, we are "penalizing" the goodness of fit in a way that is similar to the likelihood ratio test. We compare models by looking at differences in the AIC and are once again implicitly using a form of the likelihood ratio test, although the AIC is considered valid when using non-nested models.

Sakamoto et al. (1986) describe an alternative to the AIC called the Bayesian information criterion or BIC (Schwarz 1978). Hongzhi (1989) proposed an analog of the AIC for use with the sum of squares. The proposal is to use $\log(SSQ_k) + 2k/n$ as the analog of Equation 7.45, where SSQ_k is the residual sum of squares for the model with k parameters, and n is the number of points. Anderson et al. (1994) evaluate the performance of the AIC for model selection in capture-recapture data. Matsumiya (1990), Hiramatsu and Kitada (1991), and Hiyama and Kitahara (1993) provide examples of the use of the AIC in fisheries problems.

Which Criterion to Use?

The AIC is appropriate for non-nested models but for nested models either the likelihood ratio or the AIC can be used. As a note of caution, when using the Poisson or multinomial likelihoods and if the data are overdispersed, the likelihood ratio test or the AIC will be biased, and the analysis of deviance (McCullagh and Nelder 1989) is appropriate.

ROBUSTNESS: DON'T LET OUTLIERS RUIN YOUR LIFE

Our colleague David Fournier once said, "The problem with likelihood is that some observations are just too unlikely." That is, some outliers will dominate the likelihood, and the fitting procedures often go to great lengths to make predictions closer to the outlier so that the total likelihood will not be too low.

"Robust estimation" has two meanings (Huber 1981). First, what happens if the assumption of normally distributed uncertainty is not appropriate, which is often the case for ecological data sets? Second, how does one deal with one or two data points that are highly irregular (greatly deviate from the pattern suggested by the other data)? We already discussed one approach when we considered the goodness of fit provided by the sum of squares. In that case, we noted that the use of the square of the deviation between the observed and predicted data points is implicitly based on an assumption of normally distributed uncertainty, but that other measures of deviation such as absolute value (or even fractional powers of the absolute value) could be used just as easily. Most of these have the effect of reducing the penalty which the outliers contribute to the sum of the deviations.

Another approach (Press et al. 1986, 539 ff.) is to weight the data points in the sum of squares or the likelihood. For example, one could use Tukey's "biweight"

$$\begin{aligned} \omega(e) &= \text{weight assigned to uncertainty of size } e \\ &= \left(1 - \frac{e^2}{c^2}\right)^2 \quad \text{if } |e| < c, \\ &= 0 \quad \text{if } |e| > c, \end{aligned} \tag{7.46}$$

where c is a constant chosen by the user (Press et al. 1986, 542). (For normally distributed uncertainty, the appropriate value of c is 6.0). This weighting function actually decreases as e increases, and is consonant with the idea that outliers might be caused by something other than the actual ecological processes being studied. For example, to modify the simple sum of squares

$$\mathcal{S}(A_{est}, B_{est}, C_{est}) = \sum_{i=1}^{n} (Y_{pre,i} - Y_{obs,i})^2,$$

we use

$$\mathcal{S}(A_{\text{est}},B_{\text{est}},C_{\text{est}}) = \sum_{i=1}^{n} \omega(e_i)\ (Y_{\text{pre},i} - Y_{\text{obs},i})^2. \qquad (7.47)$$

where $e_i = Y_{\text{pre},i} - Y_{\text{obs},i}$. One way to think about outliers is that for any data point there is a probability p_{model} that the point arose from the model that you are considering and a probability $1 - p_{\text{model}}$ that it arose from a process other than the one specified in the model. Then the likelihood of a particular point is really $p_{\text{model}}\mathcal{L}(\text{data|model}) + (1 - p_{\text{model}})\mathcal{L}(\text{data|alternative processes})$. In general, we assume that $p_{\text{model}} = 1$. To use this approach, one needs to begin to specify what the alternative processes are; in effect, one must specify alternative models (Schnute 1993; Schnute and Hilborn 1993).

BOUNDING THE ESTIMATED PARAMETER: CONFIDENCE INTERVALS

We must always be aware that the most likely parameters are almost certainly not the real parameters of the underlying process, but rather depend on the data. How do we determine reasonable bounds for the estimated parameter? In this section we explore two approaches to quantifying uncertainty about parameter values.

Likelihood Profile

Hudson (1971) provides an especially simple method for determining a confidence bound for the case in which (i) we consider a model with only one parameter and (ii) the log-likelihood function is a unimodal function of the parameter. Hudson's method is a special case of the general technique of the likelihood profile. Using the likelihood ratio test (the theory relies, once again, on the asymptotic relationship followed by the differences in log-likelihood), the

95% confidence interval is the range of parameters for which the log-likelihood is within 1.92 of the maximum value of the log-likelihood. Thus, for example, to find the confidence interval for the Poisson rate parameter for the negative log-likelihoods shown in Figure 7.1c, we draw a horizontal line at the minimum negative log-likelihood plus 1.92 (the critical value of c^2 with one degree of freedom divided by 2) and look for the intersections of that line and the curve. Those intersection points give the limits of the confidence interval.

An Ecological Scenario. Suppose that we are involved in the control of mites that attack pistachios and have decided that if fewer than 10% of the nuts are attacked, the mite is being controlled. We want to determine the proportion infested (f) by sampling nuts. If the true level of infestation is f and we sample S nuts, the number I that are infested follows a binomial distribution:

$$\Pr\{I = i|f\} = \binom{S}{i} f^i(1 - f)^{S-i}. \qquad (7.48)$$

If we view this as the likelihood of values of f, given S and i, the negative log-likelihood is

$$\mathbf{L}\{S,i|f\} = -i\log(f) - (S - i)\log(1 - f) + J, \qquad (7.49)$$

where J denotes terms that do not depend on f and can therefore be ignored. Setting the derivative of $\mathbf{L}\{S,i|f\}$ with respect to f equal to 0 leads us to the MLE value

$$f_{\text{MLE}} = \frac{i}{S}. \qquad (7.50)$$

We use the likelihood ratio test to determine the approximate 95% confidence interval for f by finding the value of f such that the log-likelihood $\mathbf{L}\{S,i|f\} - \mathbf{L}\{f_{\text{MLE}}|S,i\} = 1.92$. Furthermore, we can do this with a sequential sampling scheme, as in the following pseudocode:

Pseudocode 7.6

1. Set $S = 0$, $i = 0$.
2. Input the number of nuts sampled and the number of sampled nuts that were infested. Replace S by S plus the number of sampled nuts and i by i plus the number of infested nuts in the current sample.
3. Find the MLE value $f_{MLE} = i/S$. Find the negative log-likelihood associated with this MLE from Equation 7.49.
4. Find the value of f_b such that

$$\mathbf{L}\{f_b|S,i\} = \mathbf{L}\{f_{MLE}|S,i\} + 1.92.$$

If this value of $f \leq 0.1$, declare the mite under control. Otherwise return to step 2.

A typical set of results using this pseudocode would be these.

Sample number	Current sample	Infested nuts	Total sample	Total infested	f_{MLE}	f_b
1	20	2	20	2	0.1	0.283
2	20	1	40	3	0.075	0.186
3	20	1	60	4	0.067	0.151
4	20	0	80	4	0.05	0.114
5	20	0	100	4	0.04	0.092

Note that after the first sample, the MLE is already 0.1, but the boundary of the 95% confidence interval for the true value of f is 0.283, so that we must continue sampling. It is only at sample 5, for which the MLE is 0.04 and the boundary of the confidence interval is 0.092, that we can declare the mite under control. Now, of course, had we sampled one-hundred nuts at the start and found four of them infested, we would draw the same conclusion as was done after the fifth sample. The advantage of the sequential sampling scheme, using likelihood, is that we might be able to stop even sooner.

The likelihood profile can be extended for situations in which the model has more than one parameter. For example, in the abundance model Equation 7.43, the best model had two free parameters, q and p. In such a case, we might want to know about the confidence intervals for q and p, either separately or together.

To conduct a likelihood profile for a system with parameters $p_1, p_2, \ldots, p_m$, one varies one (or more) parameter(s) systematically and computes the values of the other parameters that maximize the likelihood. It has the same function as a goodness-of-fit profile, giving information concerning how the parameters depend on each other, and how sensitive the likelihood is to the systematically varied parameter (Venzon and Moolgavkar 1988).

For example, suppose that the random variables $X_1, \ldots, X_n$ are normally distributed with mean m and standard deviation σ. The negative log-likelihood is then

$$\mathbf{L} = n[\log(\sigma) + \frac{1}{2}\log(2\pi)] + \sum_{i=1}^{n} \frac{(X_i - m)^2}{2\sigma^2}, \tag{7.51}$$

from which we determine the maximum likelihood estimates,

$$m_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and}$$

$$\sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (X_i - m_{MLE})^2. \tag{7.52}$$

A likelihood profile is appropriate for a situation where we are interested in one parameter but not particularly interested in the other. If the parameter of interest is the mean, we systematically search over values of m and instead of σ_{MLE}, we compute the profile standard deviation

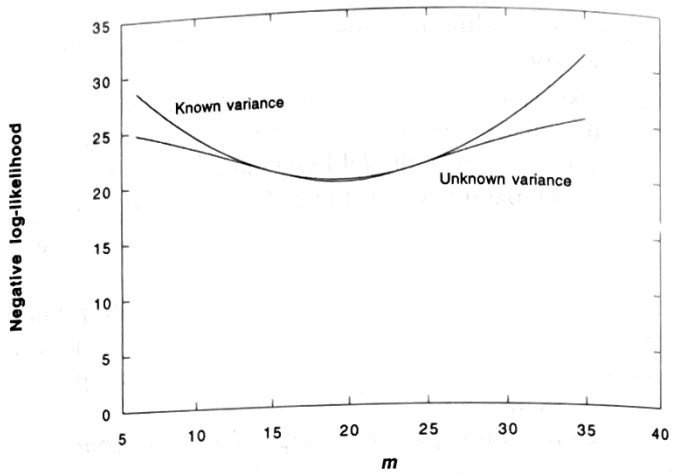


FIGURE 7.8. The negative log-likelihood for the mean m of a normal distribution when the variance is known, and the profile likelihood in which the variance is specified once the mean is given. Note that both the negative log-likelihood and likelihood profile find the mean, but that the likelihood profile is shallower (more uncertainty) when the variance is unknown.

$$\sigma^2_{\text{pro}} = \frac{1}{n}\sum_{i=1}^{n} (X_i - m)^2. \tag{7.53}$$

For example, if the data are 27.7286, 16.4676, 21.1222, 27.6477, 10.4809, and 13.9685 (generated from $m = 20$ and $\sigma = 8$), plots of the negative log-likelihood and likelihood profile find the true mean (Figure 7.8), but admitting that the standard deviation is unknown leads to a shallower negative log-likelihood and consequently to a wider confidence interval.

An Ecological Scenario. To find the likelihood profile for q for the abundance model Equation 7.43, we find the values of p and r that maximize the likelihood for each possible value of q (or, in reality, a grid search over q), as in the following pseudocode:

Pseudocode 7.7

1. Input the lower and upper bounds, and the step size of q to search.
2. Set q fixed at the lower bound.
3. Choose one of:
   - Option *a*. Calculate the negative log-likelihood by using true values of r and p.
   - Option *b*. Minimize the negative log-likelihood by searching over possible values of r and p (the true likelihood profile).
4. Plot or table the values of q and the negative log-likelihood.
5. Increment q and repeat steps 3 and 4.

This algorithm allows for two cases. First (step 3, option *a*), we fix the other parameters (r and p) at their true values (known because we have used Monte Carlo data) and examine the likelihood in q. This will demonstrate how much more we would know about q if the values of r and p were known. That is, instead of the MLE values, we use the true values of the other parameters. The results (Figure 7.9, dashed line) are quite impressive. The confidence interval for q is very narrow. Second (step 3, option *b*), we find the r and p that maximize the likelihood as q is systematically varied; this is the likelihood profile. The results (Figure 7.9, solid line) are discouraging. We can fit the data very well (i.e., the negative log-likelihood is small) with very large values of q. For example, the dashed line in Figure 7.10 shows the fit obtained when $q = 10$, $p = -40$, and $r = 1.58$. This curve is very similar to the true relationship, but clearly the individual parameter values are far from the true values (recall that a similar phenomenon occurred in Chapter 5). The confidence bound on q is enormous. In effect, admitting uncertainty in p and r means that we know nothing about the value of the individual parameter q.

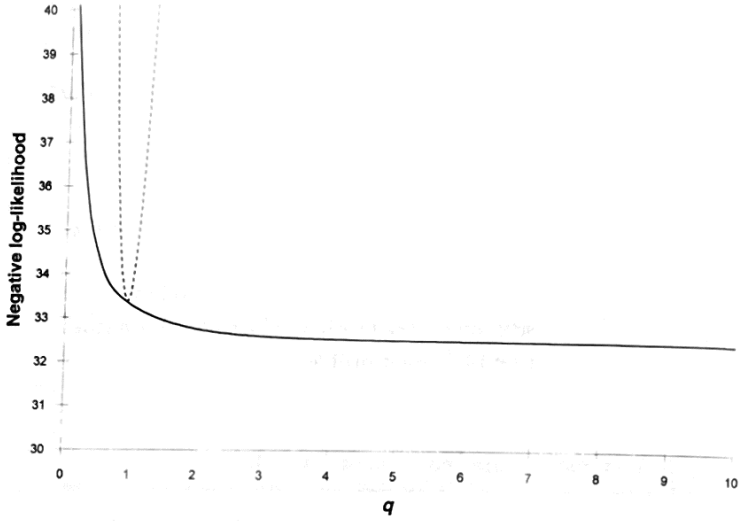


FIGURE 7.9. Likelihood profiles of q when p and r are estimated parameters (solid line) and when p and r are fixed at their true values (dashed).

THE BOOTSTRAP METHOD

In Chapters 5 and 6, we used the bootstrap method to resample data sets for model comparison. Here we extend its use for understanding the uncertainty about parameter values. The bootstrap method can be used to find confidence intervals and variances of models of any complexity by intense computation (Efron and Tibshirani 1991, 1993). As before, the bootstrap method involves generation of new data sets by sampling the original data with replacement. We begin with a set of N observations $\{Y_i, \ldots, Y_N\}$. We generate a large number of new data sets $\{Y_{\text{boot}}(i)\}$ by sampling from Y_{obs} with replacement and then generate a large number of bootstrap data sets. For each bootstrap data set we obtain an estimate of the parameters of interest and estimate the variances of the parameters from the variances of the bootstrap estimates.

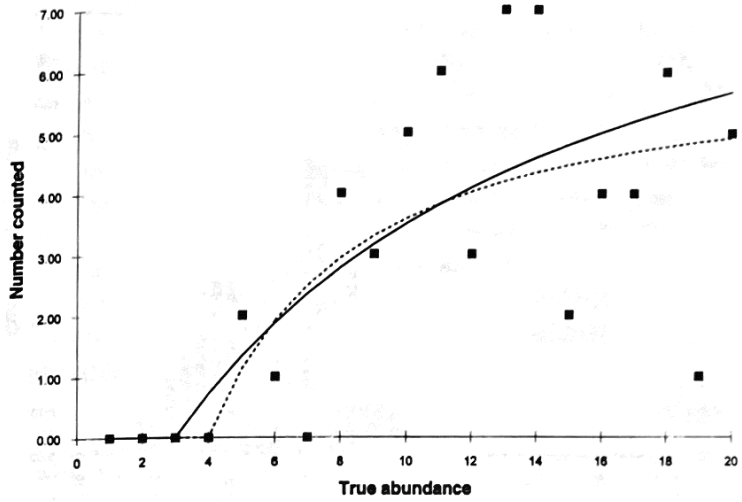


FIGURE 7.10. Data generated by the Monte Carlo method for the abundance model Equation 7.43. The true relationship is shown by the solid line, and a model with $q = 10$, $p = -40$, and $r = 1.58$ is shown by the dashed line.

Suppose that there is just one parameter, that we generate B bootstrap data sets, and that $p_{\text{boot},i}$ is the parameter estimate from the i^{th} bootstrap data set. We first set

$$\hat{p}_{\text{boot}} = \sum_{i=1}^{B} p_{\text{boot},i}/B. \tag{7.54}$$

We estimate the variance by

$$\hat{\sigma}_p = \frac{1}{B-1} \sum_{i=1}^{B} (\hat{p}_{\text{boot}} - p_{\text{boot},i})^2. \tag{7.55}$$

Returning once again to the abundance model Equation 7.43, we might want to use the bootstrap method to estimate the variance of the parameter q. This can be done using a pseudocode such as:

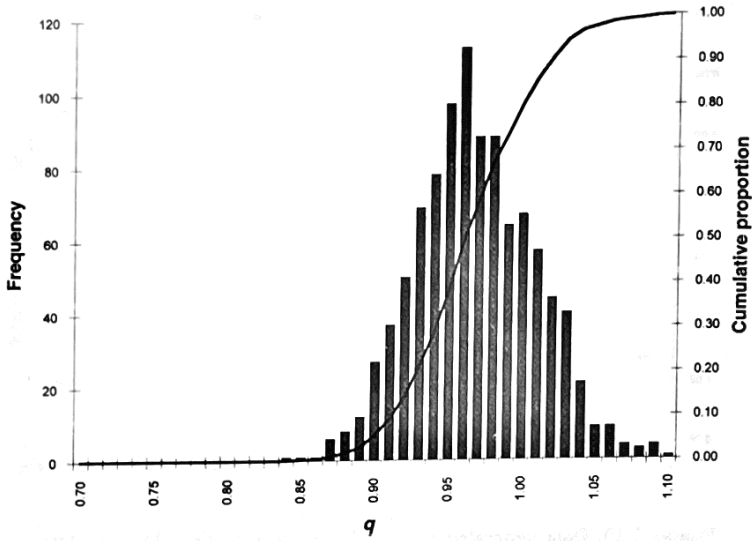


FIGURE 7.11. The distribution of estimates of q from one thousand bootstrap replicates. The solid line is the cumulative distribution function.

Pseudocode 7.8

1. Read in observed densities and index of abundance from Table 7.1
2. Set $r = 0$, $p = 0$.
3. Generate a bootstrap data set by sampling with replacement from the data twenty pairs of D_i and $I_{\text{obs},i}$.
4. Obtain the maximum likelihood estimate of q from the bootstrap data.
5. Repeat steps 3–5 1000–10 000 times.
6. Plot the frequency distribution of the estimated q values.

The output of a program based on this algorithm is a frequency distribution of estimates of q (Figure 7.11). Given a variance estimate from Equation 7.55, we can calculate the confidence bounds in the usual manner using normal distribution theory, or we can use the empirical frequency distri-

bution of the bootstrap estimates. In the latter case, the bootstrap provides a link between the likelihood methods in this chapter and the Bayesian methods of Chapter 9.

The bootstrap method as described here is often called the non-parametric bootstrap. A refinement, based on some knowledge of the ecological system, is to assume a distribution for the uncertainty; instead of resampling the data we add a random term to the predicted data based on the assumed distribution. That is, we now generate bootstrap data sets by taking the i^{th} observation $Y_{\text{pre},i}$ and adding a random variable E to it:

$$Y_{\text{boot},i} = Y_{\text{pre},i} + E, \tag{7.56}$$

where E is drawn from the assumed distribution. In principle, this should be "better" because we are incorporating more knowledge about the system into the methods of estimation. We leave it to you to modify the previous pseudocode for the case in which E has a Poisson distribution. Doing this leads to a different frequency distribution of bootstrap estimates (Figure 7.12)

Bootstrapping is a computationally intensive procedure, but it can be used on models that have dozens or even hundreds of parameters. Obtaining an estimate for large models may take minutes or even hours. It is not unknown for bootstrap runs to take several days on desktop computers, and obtaining a 99% confidence interval requires about 10 000 bootstrap samples (Efron and Tibshirani 1991, 1993).

LINEAR REGRESSION, ANALYSIS OF VARIANCE, AND MAXIMUM LIKELIHOOD

The statistical tools learned in introductory courses in biometrics were designed in an age when computation was difficult (Efron and Tibshirani 1991), but things are different today. We now show that they can be performed using the methods of maximum likelihood and the likelihood ra-

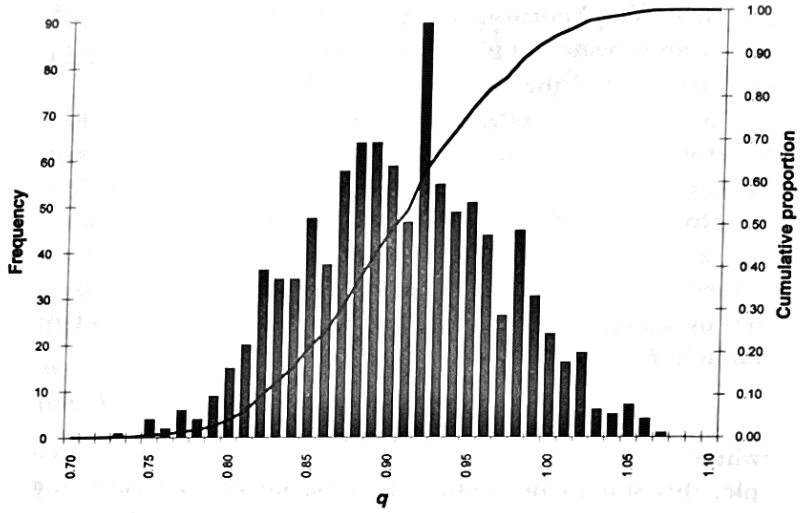


FIGURE 7.12. Estimates of q from one thousand replicates of the parametric bootstrap.

tio test but in a numerically intensive manner, thus taking advantage of modern computing technologies.

It is easier to understand statistics within the unifying concept of likelihood rather than thinking of regression, analysis of variance, and contingency tables as intellectually separate subjects.

Regression as a Problem of Maximum Likelihood

The linear regression model is

$$Y_i = a + bX_i + Z_i, \tag{7.57}$$

where the parameters a and b are to be determined and Z_i is normally distributed with mean 0 and variance σ^2. Proceeding as before, the negative log-likelihood is

$$\mathbf{L} = n[\log(\sigma) + \frac{1}{2}\log(2\pi)] + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - a - bX_i)^2. \tag{7.58}$$

A nonlinear search over a, b, and σ can be used to minimize the negative log-likelihood. However, the maximum likelihood estimates of a and b are solutions of the linear equations

$$\sum_{i=1}^{n} Y_i = na_{\text{MLE}} + b_{\text{MLE}}\sum_{i=1}^{n} X_i,$$

$$\sum_{i=1}^{n} X_iY_i = a_{\text{MLE}}\sum_{i=1}^{n} X_i + b_{\text{MLE}}\sum_{i=1}^{n} X_i^2, \tag{7.59}$$

found by taking the derivative of the likelihood with respect to a or b and setting it equal to zero.

Note that these are independent of the variance, which we estimate by

$$\sigma^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - a_{\text{MLE}} - b_{\text{MLE}}X_i)^2. \tag{7.60}$$

A two-dimensional confidence interval on a and b is found by searching over all values of a and b that provide likelihoods within a specified value of the minimum negative log-likelihood. For example, for a 95% confidence interval, we use a chi-square distribution with two degrees of freedom, for which the critical value is 6.0. Thus, we contour all negative log-likelihoods that are three greater than the best value.

On the other hand, we might be interested in a single parameter, say b, and not at all interested in the other parameter, so that a likelihood profile on b is appropriate. We first specify b in the negative log-likelihood and then compute that value of a that minimizes the negative log-likelihood for that value of b. This can be done from Equation 7.59:

$$a_{\text{pro}} = \frac{\sum_{i=1}^{n} Y_i - b\sum_{i=1}^{n} X_i}{n}. \tag{7.61}$$

Since this is now a one-parameter confidence bound, the critical chi-square value is 3.84, so values of negative log-likelihood that are 1.92 greater than the minimum are in the 95% confidence interval.

To illustrate these ideas, we generated data from the model $Y_i = 1 + 2X_i + Z_i$, with $\sigma = 5$. A typical set of ten data points is:

X_i	Y_i
1	7.830 37
2	2.792 27
3	7.701 37
4	13.779 8
5	5.050 55
6	9.230 33
7	3.452 11
8	11.952 8
9	23.855 9
10	22.088 5

for which $a_{\text{MLE}} = 1.77$, $b_{\text{MLE}} = 1.641$, $\sigma_{\text{MLE}} = 5.69$, and the minimum negative log-likelihood is 30.5738.

The 95% confidence contour for both parameters (Figure 7.13) is an ellipse with a negative correlation between the estimated values of a and b. The data allow a to be large, but then b must be small, and vice versa. The likelihood profile on b (Figure 7.14) considerably tightens the confidence region.

A good ecological detective will recognize that there are other models, such as

$$Y_i = k + Z_i \qquad \text{(average value model)},$$

$$Y_i = a + bX_i + cX_i^2 + Z_i \qquad \text{(quadratic regression model)}. \quad (7.62)$$

We encourage you to compute the negative log-likelihoods for these other models with one and three parameters, re-

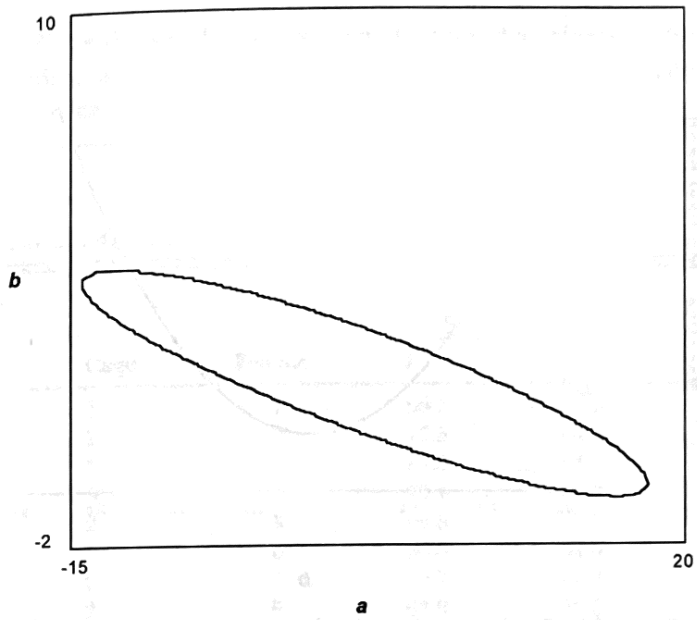


FIGURE 7.13. The 95% confidence region, determined by maximum likelihood analysis, for the parameters a and b of the linear regression model.

spectively, and compare the results with the regression model that we analyzed (two parameters). Which model would you choose on the basis of a likelihood criterion?

Regression methods also usually report the "proportion of variance explained by the model." Here, likelihood methods provide little additional information. However, Bayesian methods tell us that we should not attempt to "explain variation"; instead, we should construct posterior probability densities and ask about the shape of those distributions. After reading Chapter 9, we encourage you to rethink this analysis from a Bayesian perspective. What kind of priors would you choose for a and b?

Finally, we encourage you to experiment with a situation in which we do not know the underlying model. Sokal and

Since this is now a one-parameter confidence bound, the critical chi-square value is 3.84, so values of negative log-likelihood that are 1.92 greater than the minimum are in the 95% confidence interval.

To illustrate these ideas, we generated data from the model $Y_i = 1 + 2X_i + Z_i$, with $\sigma = 5$. A typical set of ten data points is:

X_i	Y_i
1	7.830 37
2	2.792 27
3	7.701 37
4	13.779 8
5	5.050 55
6	9.230 33
7	3.452 11
8	11.952 8
9	23.855 9
10	22.088 5

for which $a_{\text{MLE}} = 1.77$, $b_{\text{MLE}} = 1.641$, $\sigma_{\text{MLE}} = 5.69$, and the minimum negative log-likelihood is 30.5738.

The 95% confidence contour for both parameters (Figure 7.13) is an ellipse with a negative correlation between the estimated values of a and b. The data allow a to be large, but then b must be small, and vice versa. The likelihood profile on b (Figure 7.14) considerably tightens the confidence region.

A good ecological detective will recognize that there are other models, such as

$$Y_i = k + Z_i \qquad \text{(average value model)},$$

$$Y_i = a + bX_i + cX_i^2 + Z_i \qquad \text{(quadratic regression model)}. \quad (7.62)$$

We encourage you to compute the negative log-likelihoods for these other models with one and three parameters, re-

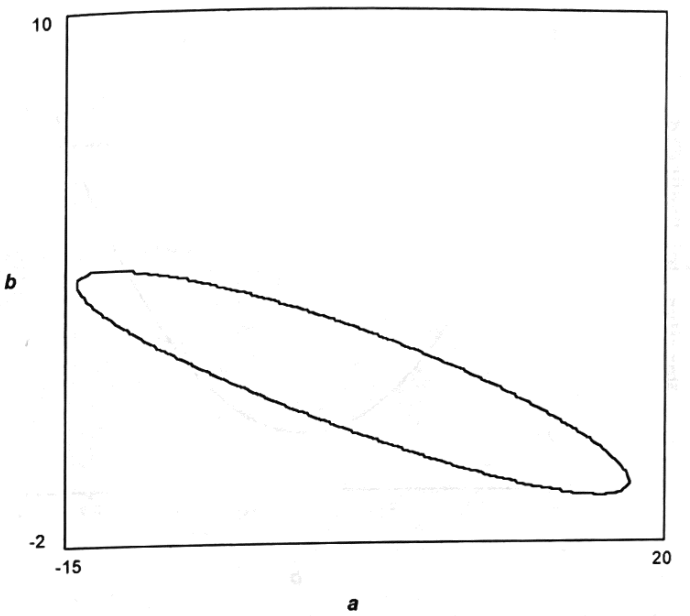


FIGURE 7.13. The 95% confidence region, determined by maximum likelihood analysis, for the parameters a and b of the linear regression model.

spectively, and compare the results with the regression model that we analyzed (two parameters). Which model would you choose on the basis of a likelihood criterion?

Regression methods also usually report the "proportion of variance explained by the model." Here, likelihood methods provide little additional information. However, Bayesian methods tell us that we should not attempt to "explain variation"; instead, we should construct posterior probability densities and ask about the shape of those distributions. After reading Chapter 9, we encourage you to rethink this analysis from a Bayesian perspective. What kind of priors would you choose for a and b?

Finally, we encourage you to experiment with a situation in which we do not know the underlying model. Sokal and

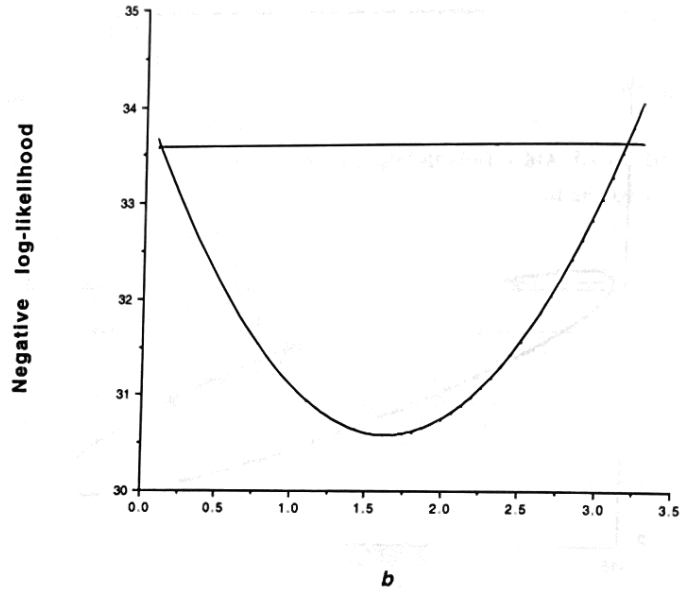


FIGURE 7.14. The likelihood profile of the parameter b and the 95% confidence region (below the solid line) in the linear regression model.

Rohlf (1969) report experiments in which twenty-five individual flour beetles were starved for six days at nine different humidities. The data are:

Relative humidity (%)	Average weight loss (mg)
0	8.98
12	8.14
29.5	6.67
43	6.08
53	5.9
62.5	5.83
75.5	4.68
85	4.2
93	3.72

Since the weight loss shows a clear trend with relative humidity, a linear regression model might be appropriate. What can you conclude about these data?

Analysis of Variance by Maximum Likelihood

TABLE 7.4. Mosquito wing lengths (Sokal and Rohlf 1969).

		Left wing length measurement	
Cage	Female	First	Second
1	1	58.5	59.5
1	2	77.8	80.9
1	3	84.0	83.6
1	4	70.1	68.3
2	5	69.8	69.8
2	6	56.0	54.5
2	7	50.7	49.3
2	8	63.8	65.8
3	9	56.6	57.5
3	10	77.8	79.2
3	11	69.9	69.2
3	12	62.1	64.5

We now show how a traditional analysis of variance can be performed using maximum likelihood theory. Sokal and Rohlf (1969) describe an experiment in which twelve field-caught mosquito pupae were reared in three different cages, four mosquitoes to each cage. When the mosquitoes hatched, the left wing of each mosquito was measured twice (Table 7.4). The observations are thus the wing length L_{ij} for female i on observation j, and the cage in which female i is reared, c_i. We postulate three different models:

$$\begin{aligned} L_{ij} &= K + Z_{ij} \quad \text{(model A)}, \\ L_{ij} &= D_{c_i} + Z_{ij} \quad \text{(model B)}, \\ L_{ij} &= F_i + Z_{ij} \quad \text{(model C)}. \end{aligned} \tag{7.63}$$

In each model, Z_{ij} is normally distributed. The alternatives are (i) the observations are normally distributed about some constant (K) (model A); (ii) there is a different average length (D_{c_i}) within each cage (model B); or (iii) there is a different average length (F_i) for each individual fly (model C).

The likelihoods for the three models are

$$\mathcal{L}_A = \prod_{i=1}^{12}\prod_{j=1}^{2} \frac{1}{\sigma_A\sqrt{2\pi}} \exp\left(-\frac{[L_{ij}-K]^2}{2\sigma_A{}^2}\right),$$

$$\mathcal{L}_B = \prod_{i=1}^{12}\prod_{j=1}^{2} \frac{1}{\sigma_B\sqrt{2\pi}} \exp\left(-\frac{[L_{ij}-D_{c_i}]^2}{2\sigma_B{}^2}\right),$$

$$\mathcal{L}_C = \prod_{i=1}^{12}\prod_{j=1}^{2} \frac{1}{\sigma_C\sqrt{2\pi}} \exp\left(-\frac{[L_{ij}-F_i]^2}{2\sigma_C{}^2}\right). \tag{7.74}$$

In principle, each model has a different standard deviation. When computing the negative log-likelihoods for the three models (Table 7.5), model A requires two parameters (the global mean and the standard deviation); the standard deviation can be obtained analytically. Model B requires four parameters, a mean for each cage, and a standard deviation. Finally, model C requires a mean for each of the twelve flies and a standard deviation.

TABLE 7.5. Analysis of variance by maximum likelihood for the mosquito data.

Model	Number of parameters	Negative log-likelihood	Chi-square probability*
A (Average)	2	89.32	—
B (Cage effect)	4	85.42	0.02
C (Female effect)	13	28.90	~0.0000

*Used to compare models A and B (with two degrees of freedom) and models B and C (with nine degrees of freedom).

When comparing models A and B, the negative log-likelihood is reduced by about four by adding two additional parameters. Twice the difference in the likelihood between model A and model B is 7.8. The chi-square probability of a change in 7.80 with two degrees of freedom is about 0.02, so the significance of the difference is borderline (significant at 0.05 but not at 0.01). Comparing models B and C, however, we find a considerable reduction in the negative log-likelihood and an associated chi-square probability that is essentially zero. We therefore conclude that there are differences between females and that model C is preferred.