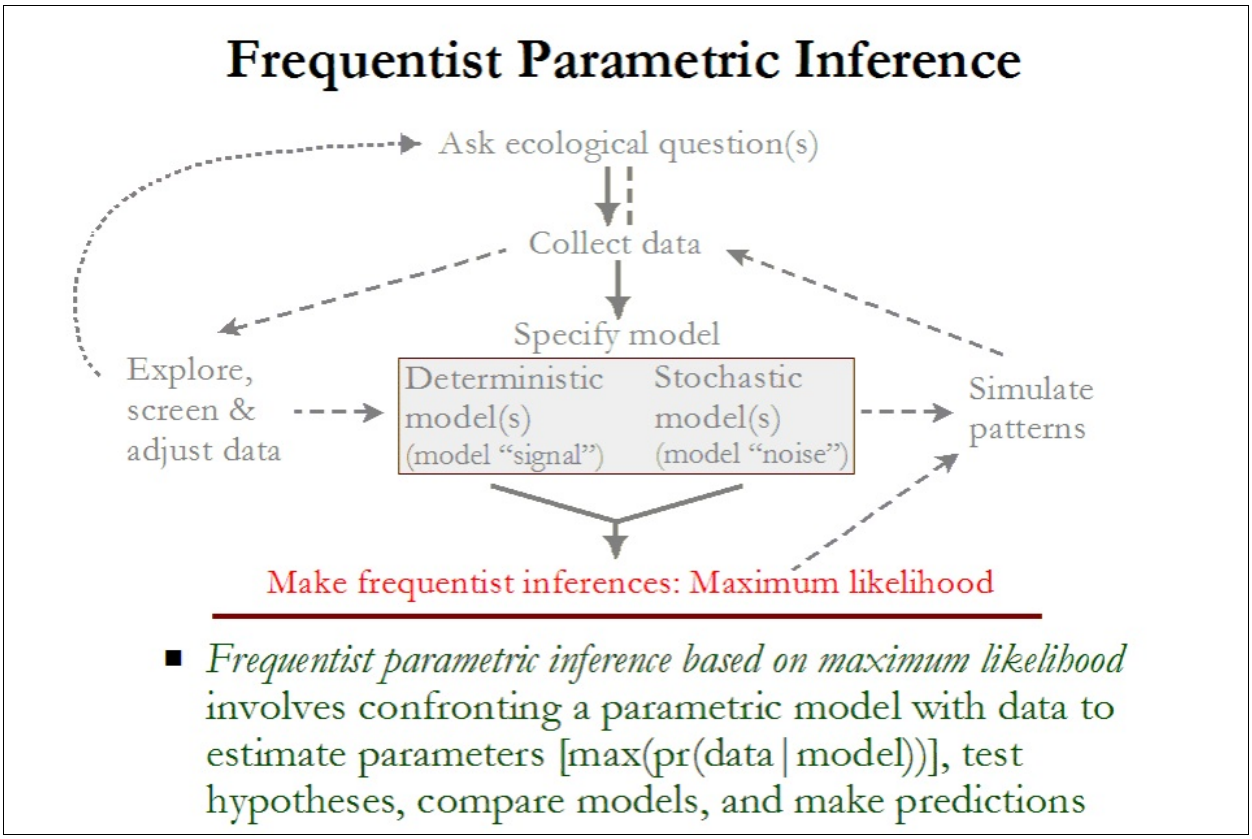
Analysis of Environmental Data
Chapter 8 Conceptual Foundations:
Maximum Likelihood Inference



1. Frequentist parametric inference based on maximum likelihood

The method of ordinary least 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, bootstrap resampling of the data set, Monte Carlo randomization procedures allows us to make additional inferences. 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 least squares approach and use frequentist parametric inference methods based on maximum likelihood, which we discuss in this chapter. The likelihood methods discussed here allow us to calculate confidence bounds on parameters directly without resampling the original data set, and to test hypotheses in the traditional manner (i.e., without resorting to Monte Carlo randomization procedures). In addition, likelihood forms the foundation for Bayesian analysis, which we discuss in the next chapter.

Frequentist Parametric Inference...

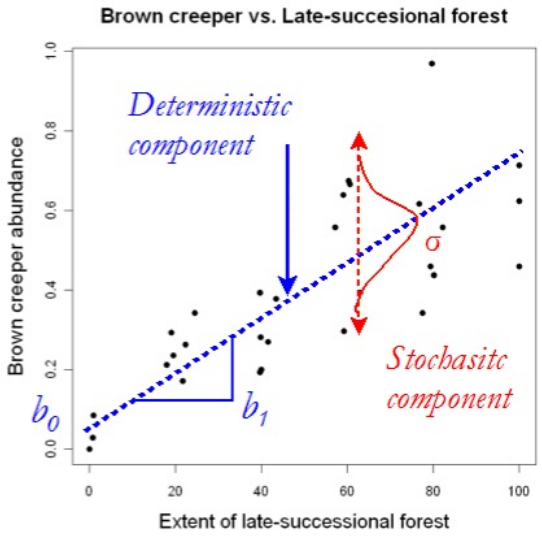
What is the statistical model?

- Do brown creepers increase in relative abundance with increasing extent of late-successional forest?

Statistical Model:

$Y \sim \text{Normal}(a + bx, \sigma)$

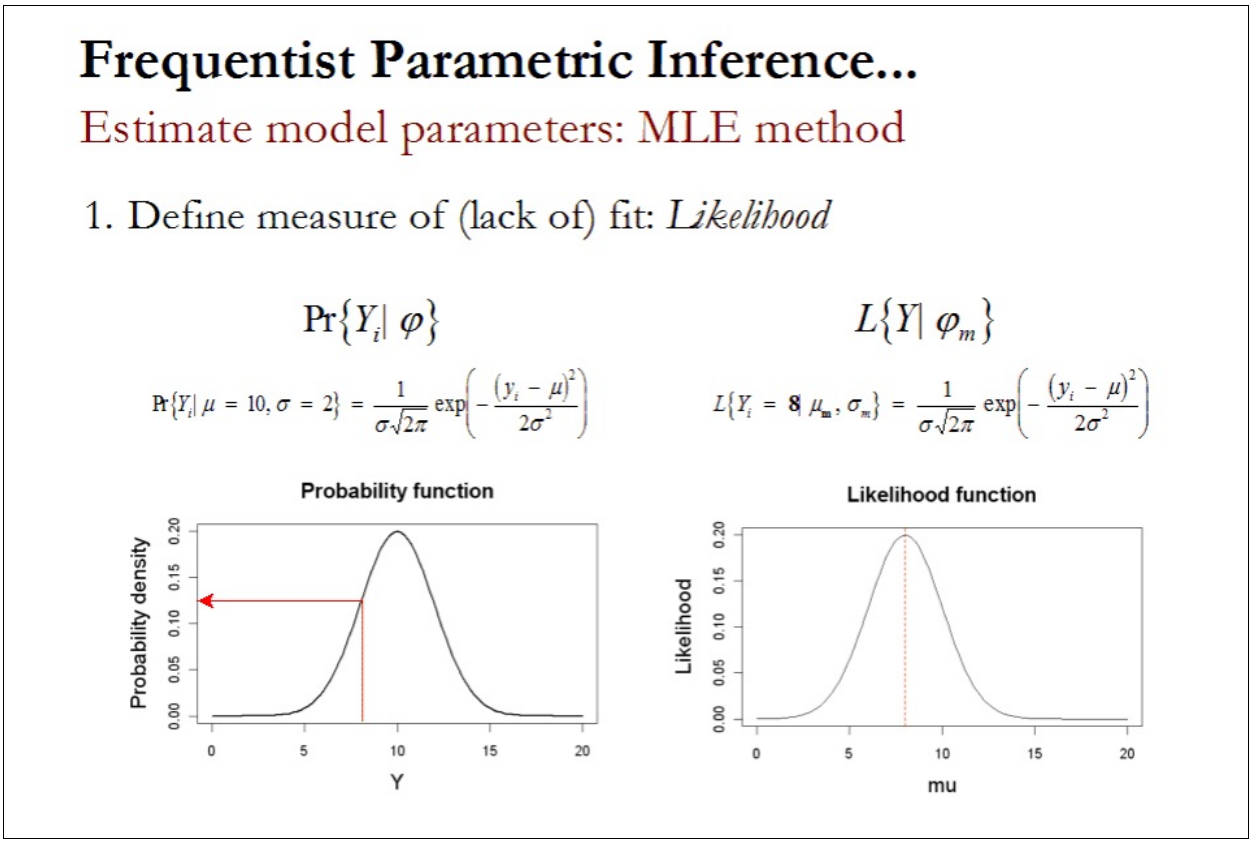
Parameters



2. The parametric statistical model

Given a question, the first step in model-based inference is to propose a statistical model that we wish to confront the data with. In a parametric approach, we need to specify both the deterministic component and the error component, although in some simple cases we may only need to specify the error.

Example: Let's continue with the familiar brown creeper example. Here, we are proposing a linear model for the deterministic component (i.e., a linear relationship between brown creeper abundance and late-successional forest extent) and normally distributed errors for the stochastic component (for simplicity).



3. Parameter estimation: maximum likelihood

The next step is to fit the model; i.e., estimate the model parameters. What we need is an objective method of computing parameter estimates from the data that are in some sense the 'best' estimates of the parameters for these data and this particular model. In this frequentist parametric inference framework, we call the "best" estimates the *maximum likelihood estimates* of the parameters because they are the parameter values that make the observed data the most likely to have happened. To find the maximum likelihood estimates, we need to define an objective measure of fit that we can maximize (or a measure of 'lack of fit' that we can minimize). Our new measure of fit is called the Likelihood and it works as follows.

Likelihood:

For any of the known probability distributions (see earlier chapter), the probability of observing data Y_i, given a (possibly vector-valued) parameter value φ, is:

$$\Pr\{Y_i|\varphi\}$$

The subscript on Y_i indicates that there are many possible outcomes (for example, $i = 1, 2, \ldots I$), but only one value of the parameter φ. For example, suppose that Y_i follows a normal distribution with the parameters mean = μ and standard deviation = σ. Then for any observation we predict that Y_i =

y_i with probability:

$$\Pr\{Y_i = y_i | \mu, \sigma\} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right)$$

Note, the right side of the equation is just the formula for the probability density function (pdf) for the normal distribution. Recall that the pdf gives the probability of any particular outcome given values of the parameter(s). We can also express this as the probability of the 'data' (treated as random) given the 'hypothesis' (treated as fixed), where the 'data' is a single observation of Y_i (= y_i) and the 'hypothesis' is that the mean = μ and standard deviation = σ.

However, when confronting models with data, we usually want to know how well the data support the alternative hypotheses, where hypotheses represent different values of the parameters. That is, after data collection, the data are known (fixed) but the hypotheses (parameter values) are still unknown. We ask, "given these data, how likely are the possible hypotheses (parameter values)?" To do this, we introduce a new symbol to denote the "likelihood" of the data given the hypothesis:

$$\mathcal{L}\{Y | \varphi_m\}$$

Note the subtle but important shift between equations: Y has no subscript here because there is only one observation (the observed value of Y_i), but now the parameter is subscripted because there are alternative parameter values (hypotheses); for example, we might have $m = 1, 2, ..., M$.

Frequentist Parametric Inference...

Estimate model parameters: MLE method

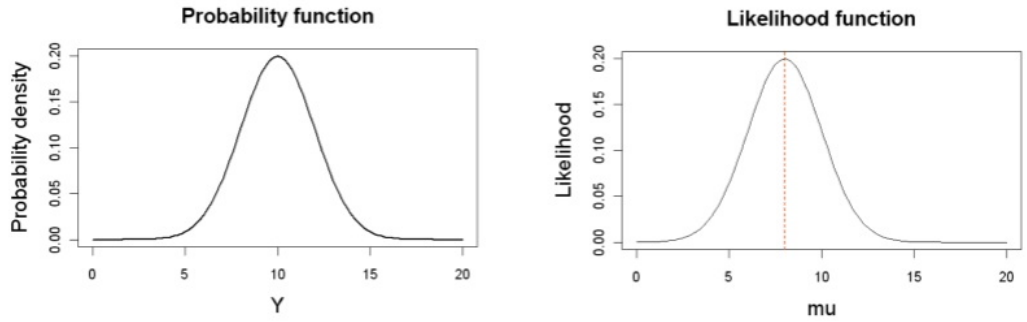
1. Define measure of (lack of) fit: *Likelihood*

$$\Pr\{Y_i|\,\varphi\}$$

$$\sum_{i=1}^{I} \Pr\{Y_i|\,\varphi\} = 1$$

$$L\{Y|\,\varphi_m\} \equiv \Pr\{Y|\,\varphi_m\}$$

$$\sum_{m=1}^{M} L\{Y|\,\varphi_m\} \neq 1$$



The key to the distinction between likelihood and probability is that with probability the hypothesis (parameter value) 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 of the data, given the hypothesis, so the likelihood of parameter φ_m given the data Y, is:

$$\mathcal{L}\{Y|\,\varphi_m\} = c \cdot \Pr\{Y|\,\varphi_m\}$$

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 equivalent to the probability of the data, given the hypothesis:

$$\mathcal{L}\{Y|\,\varphi_m\} \equiv \Pr\{Y|\,\varphi_m\}$$

Note that since likelihood is not equal to probability, but as far as we are concerned it is equivalent to probability, we replace the equal sign with the symbol for equivalence – three bars instead of two. Also note that although it must be true that if the parameter φ is fixed:

$$\sum_{i=1}^{I} \Pr\{Y_i|\,\varphi\} = 1$$

when the data Y are fixed, the sum over the possible parameter values

$$\sum_{m=1}^{M} \mathcal{L}\{Y \mid \varphi_m\} \neq 1$$

need not be finite, let alone equal to 1. Thus, likelihood is not equal to probability. Nevertheless, it may be helpful to think of likelihood as a kind of unnormalized probability.

The likelihood function:

By plotting the likelihood ($\mathcal{L}$) as a function of φ_m – the likelihood function – we can get a sense of the range of parameter values for which the observations are probable. When looking at the likelihood function, remember that the comparisons are within a particular value of the data and not between different values of the data. The likelihood function depicts the likelihood of the data given alternative values of the parameter(s). The value of the parameter(s) that gives the maximum likelihood is the "best" estimate of the parameter – because it makes the data the most likely outcome.

Frequentist Parametric Inference...

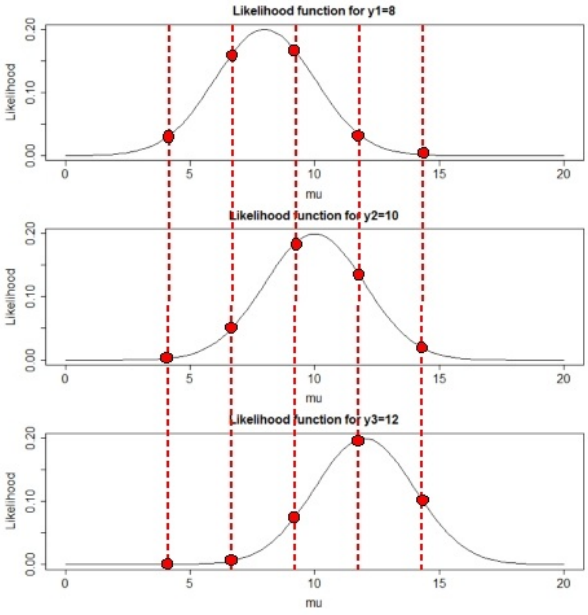
Estimate model parameters: MLE method

1. Define measure of (lack of) fit: *Likelihood*

$$L\{Y_i = 8|\mu_m, \sigma_m\} = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right)$$

$$L\{Y_i = 10|\mu_m, \sigma_m\} = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right)$$

$$L\{Y_i = 12|\mu_m, \sigma_m\} = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right)$$



The likelihood as given above is for a single observation $Y=y_i$. However, we usually have multiple observations in a data set, $Y = y_1, y_2, ..., y_n$. Since likelihoods are determined from probabilities, the likelihood of a set of independent observations is the product of the likelihoods of the individual observations:

$$\mathcal{L}\{Y_1, Y_2, Y_3|\,\varphi_m\} = \mathcal{L}\{Y_1|\,\varphi_m\} \cdot \mathcal{L}\{Y_2|\,\varphi_m\} \cdot \mathcal{L}\{Y_3|\,\varphi_m\}$$

We can visualize this graphically by creating a separate likelihood curve for each observation, where we plug in the single value of $Y=y_i$ and evaluate the likelihood function across the range of possible parameter values. Thus, in the figure above, the top likelihood curve shows the curve for $y_1=8$. Note, when $y_1=8$ the likelihood is greatest when the mean is also 8, which makes sense because with the normal distribution the mean will always be the most likely value. The middle likelihood curve shows the curve for $y_1=10$, and note that it is highest when the mean is 10. Similarly, the bottom likelihood curve shows the curve for $y_1=12$, and again it is highest when the mean is 12.

Now, when we want to determine the likelihood for the entire dataset, we simply evaluate each of the curves for a fixed value of *mu* to compute the individual likelihoods, and then multiply the results to get the joint likelihood of the entire dataset. So for a *mu* of say 4, we compute the likelihoods for each of the observations (shown by the red dots with the vertical line intersects the likelihood curves) and take the product. We do the same for each possible value of *mu* and plot the overall result as our likelihood curve for the entire dataset.

Frequentist Parametric Inference...

Estimate model parameters: MLE method

1. Define measure of (lack of) fit: *Likelihood*

Likelihood of single observation ≡ probability

$$L\{Y_1|\,\varphi_m\} = L\{Y_1|\,\varphi_m\}$$

Likelihood of multiple observations ≡ product of probabilities

$$L\{Y_1, Y_2, Y_3|\,\varphi_m\} = L\{Y_1|\,\varphi_m\} \cdot L\{Y_2|\,\varphi_m\} \cdot L\{Y_3|\,\varphi_m\}$$

Log-likelihood of multiple observations ≡ sum of *log* probabilities

$$LL\{Y_1, Y_2, Y_3|\,\varphi_m\} = LL\{Y_1|\,\varphi_m\} + LL\{Y_2|\,\varphi_m\} + LL\{Y_3|\,\varphi_m\}$$

Because likelihoods may be very small numbers, the product of the likelihoods can get very small very fast and can exceed the precision of the computer. Because we only care about the relative likelihood, the convention is to work with the log-likelihoods instead, because the log of a product is equal to the sum of the logs. Therefore, we can take of the logarithm of each individual likelihood and add them together and get the same end result, i.e., the parameter value that maximizes the likelihood ($\mathcal{L}$) is equal to the parameter value that maximizes the log-likelihood (L):

$$L\{Y_1, Y_2, Y_3|\,\varphi_m\} = L\{Y_1|\,\varphi_m\} + L\{Y_2|\,\varphi_m\} + L\{Y_3|\,\varphi_m\}$$

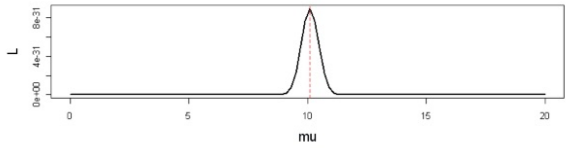
In addition, in analogy to ordinary least squares, we use the negative of the log-likelihood so that the most likely value of the parameter is the one that makes the negative log-likelihood as small as possible. In other words, the maximum likelihood estimate is equal to the minimum negative log-likelihood estimate. Thus, like sums of squares, negative log-likelihood is really a "badness"-of-fit criterion. Even though in practice we almost always find the minimum negative log-likelihood estimates, we usually still refer to the estimates as the maximum likelihood estimates – since they are equivalent anyways.

Frequentist Parametric Inference...

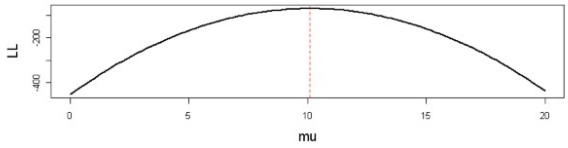
Estimate model parameters: MLE method

2. Find *maximum likelihood estimates* (MLE):

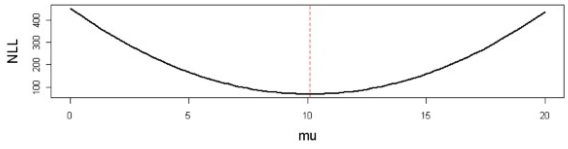
- Maximize the *likelihood*, or



- Maximize the *log-likelihood*, or



- Minimize the *negative log-likelihood*



Maximum likelihood estimation:

Given our goodness(or badness)-of-fit measure, our next step is to find the values of the parameters that give us the best fit – the so-called maximum likelihood estimators. By convention, we usually minimize the negative log-likelihood function, but the solution is the same if we were to maximize the likelihood or log-likelihood functions. Note, the figures shown here, the y-axis changes but the optimum value of the parameter (x-axis) does not.

Frequentist Parametric Inference...

Estimate model parameters: MLE method

Brown creeper example: $Y \sim \text{Normal}(a + bx, \sigma)$

Likelihood

Data

$$L\{Y|\, b_0, b_1, \sigma\} = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(y_i - b_0 - b_1 x_i)^2}{2\sigma^2} \right)$$

Normal pdf

Substitute for mean μ

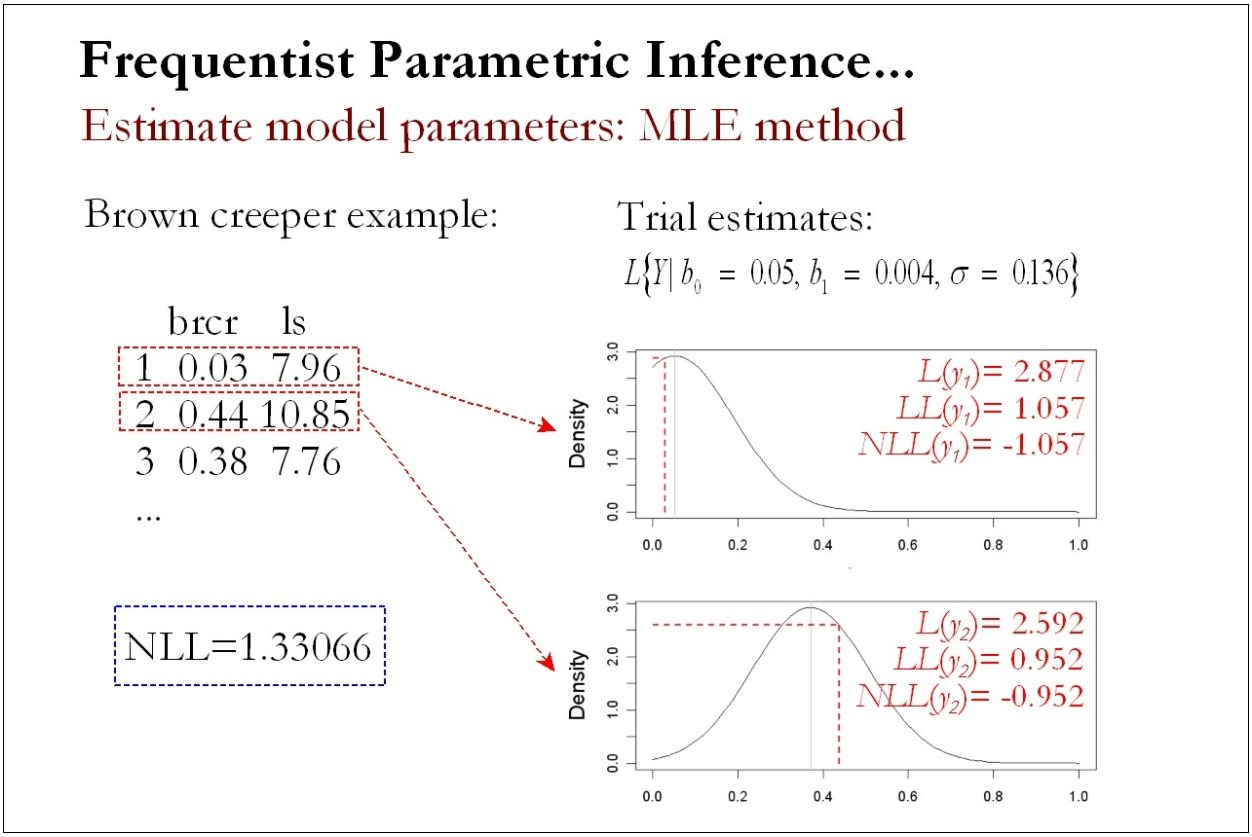
Negative log-likelihood

$$NLL\{Y|\, b_0, b_1, \sigma\} = n\left[\log(\sigma) + \frac{1}{2}\log(2\pi)\right] + \sum_{i=1}^{n} \frac{(y_i - b_0 - b_1 x_i)^2}{2\sigma^2}$$

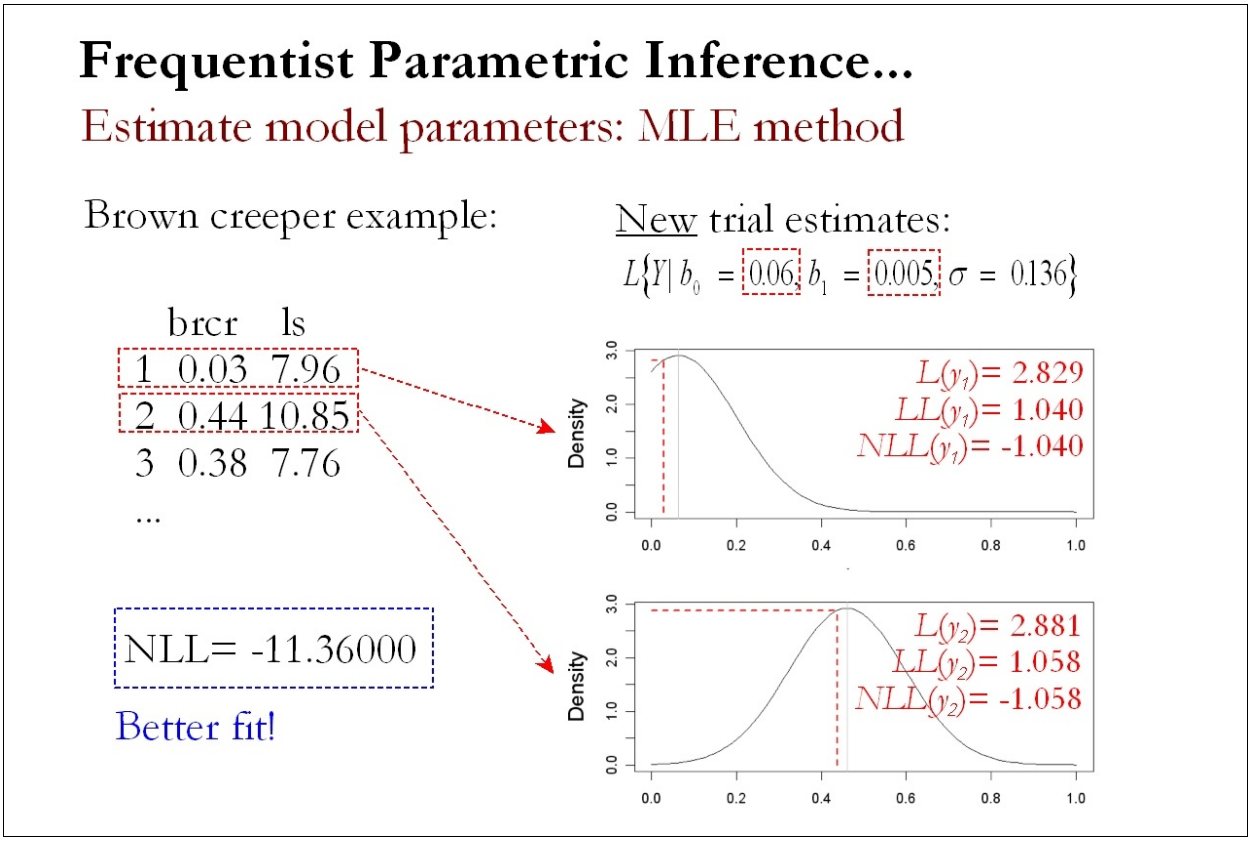
MLE: analytically or numerically find values of b_0, b_1 and σ that minimize negative log-likelihood function

Example: Let's see how maximum likelihood estimation works for our linear model example. First, we have to create the appropriate likelihood function for our model. In this case, given our choice of normal errors, the likelihood of each observation is based on the probability density function for the normal distribution. However, because we are assuming a linear relationship between x (%late-successional forest) and y (brown creeper abundance), we replace the mean in the normal equation with the linear model $b_0 + b_1 x$. The likelihood of the entire data set is simply the product of the likelihoods of each observation, assuming that they are independent observations – which we will assume here. The negative log-likelihood of the data set is negative of the sum of the log-likelihoods of each observation, which can be simplified further as shown.

The best estimates of our model parameters are those that minimize our measure of lack of fit (*NLL*). Like with ordinary least squares, the solution can be found either numerically or analytically. In this case, given our simple model and choice of normal errors, an analytical solution exists. However, in more complex models or models with non-normal errors, a numerical solution is needed. Note that in our case, the solution is found by finding values of b_0 and b_1 that minimize the sums of squared residuals, which is the same solution found through ordinary least squares.



Let's take the first observation y_1, and calculate the likelihood given a trial set of values for the model parameters. Note, the likelihood can be obtained by plugging in the right numbers for this observation into the likelihood function. Graphically, this is equivalent to reading off the probability density for the observed value of y (0.03) given a normal pdf with a mean equal to the predicted value of *y* obtained from the linear equation and a standard deviation equal to the trial value for this parameter. Note, the x-axis of the pdf is cut-off at 0 here since we can never observe a value of brown creeper abundance less than 0; however, in reality the normal pdf extends to the left into negative territory. The likelihood for the first observation is 2.877. The log-likelihood log(2.877) is 1.057 and the negative log-likelihood is -1.057. We repeat this process for the second observation and get a negative log-likelihood of -0.952. If we repeat this process for each observation and sum the results, we get a negative log-likelihood for the data set of 1.33066. Is this a good fit or bad fit? We cannot say in absolute terms since likelihood (and negative log-likelihood) does not have a probabilistic interpretation. We can only say by comparison to other fits of the model.

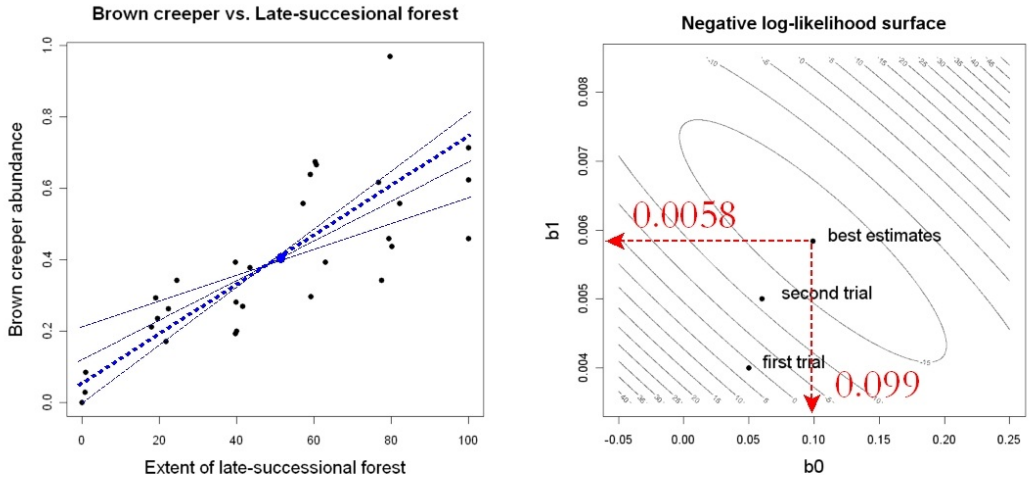


Let's try another set of values for the parameters and recalculate the negative likelihood of the data set. Here we increased the value of the intercept and slope parameters, but held the standard deviation the same, and we got a much smaller negative log-likelihood. Remember, smaller negative log-likelihood means a better fit, so we did much better with these new parameter estimates.

Frequentist Parametric Inference...

Estimate model parameters: MLE method

Brown creeper example: MLE by numerical optimization



Numerical optimization essentially involves trying new values for the parameters until we find values that minimize the negative log-likelihood of the data set. This essentially involves shifting the fit of the linear model around until the "best" fit is found. Best in this case is defined by the values that minimize the negative log-likelihood of the data, but this can be shown to be equivalent to minimizing the sums of squared residuals since the errors are assumed to be normally distributed.

If we allow b_0 to vary between -0.05-0.25 and b_1 to vary between 0.0035-0.0085, and then for each combination of values recalculate the negative log-likelihood, we can plot the results as a goodness (or badness)-of-fit surface. The surface depicts the value of *NLL* for every combination of parameter values evaluated. The lowest point on this surface represents the combination of parameter values that minimizes *NLL*, our badness-of-fit metric. The contours are not close enough near the bottom of the surface to estimate precisely where the minimum is, but the computer tells us that they are b_0=0.099 and b_1=0.0058.

Frequentist Parametric Inference...

Estimate model parameters: MLE method

Pros and Cons of MLE Estimation:

- Requires assumptions about the error – it is a parametric method
- Likelihoods can be computed for any parametric model – not restricted to normal errors
- If the errors are normally distributed, then MLE and OLS estimates are virtually identical
- MLE is the basis for most modern ecological modeling – but see Bayesian

Pros and cons of maximum likelihood estimation:

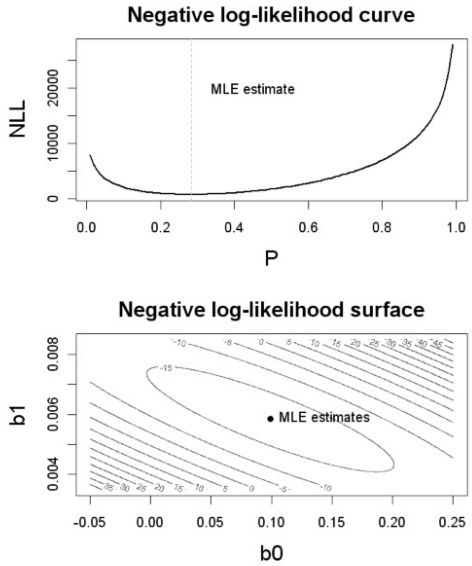
- Maximum likelihood estimation is a parameter procedure; thus, it requires that we make assumptions about the stochastic component of the model.
- The good news is that maximum likelihood solutions can be found for just about any parametric model – it's not restricted to normally distributed errors – assuming that the negative log-likelihood function can be derived, which gets increasingly difficult with increasingly complex problems. However, at least theoretically, a negative log-likelihood function can be specified for any problem.
- If the errors are normally distributed, then the maximum likelihood estimates and ordinary least squares estimates are virtually identical. There may be differences in some estimates but these are usually trivial, especially if sample sizes are large.
- Maximum likelihood estimation is the basis for most modern ecological modeling – along with Bayesian estimation, so it behooves us to become very familiar with the approach.

Frequentist Parametric Inference...

Confidence intervals for model parameters

Likelihood curves and surfaces:

- Depict the likelihood (usually negative log-likelihood) as a function of parameter values: *curves* for one parameter, *surfaces* for two parameters
- Each point on the curve or surface corresponds to a goodness(badness)-of-fit to the data
- The shape of the curve or surface reflects the precision of our estimates



4. Confidence intervals

Thus far, we have used maximum likelihood estimate to get point estimates, but these are generally useless without measures of uncertainty. We really want to know the uncertainty associated with the parameter estimates.

Likelihood curves and surfaces:

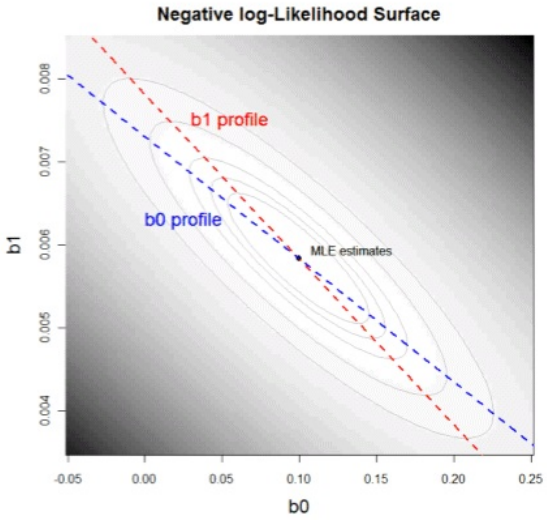
The most basic tool for understanding how likelihood depends on one or more parameters is the likelihood curve or likelihood surface, which is just the likelihood (usually the negative log-likelihood) plotted as a function of parameter values. The *likelihood curve* is plotted for a single parameter, whereas the *likelihood surface* shows the likelihood as a function of two parameters. Each point on the curve or surface corresponds to a goodness(badness)-of-fit to the data, and the shape of the curve or surface reflects the precision of our estimates. Intuitively, it makes sense that if the value of the likelihood degrades rapidly as we move away from the point estimate (the minimum negative log-likelihood), then we should probably have high confidence in our point estimate and the corresponding confidence interval should be small. On the other hand, if the curve or surface degrades slowly, then it means that we should have lower confidence in our point estimate.

Frequentist Parametric Inference...

Confidence intervals for model parameters

Likelihood profiles:

- Depict "ridgelines" in parameter space showing the minimum negative log-likelihood for particular values of a *single* parameter
- To calculate a profile for a particular parameter, we set the focal parameter in turn to a range of values, and for each value optimize the NLL with respect to all other parameters



Likelihood profiles:

If we want to deal with models with more than two parameters, or if we want to analyze a single parameter at a time, we have to find a way to isolate the effects of one or more parameters while still accounting for the rest. The preferred way to do this is calculate *likelihood profiles*, which represent "ridgelines" in parameter space showing the minimum negative log-likelihood for particular values of a single parameter. To calculate a likelihood profile for a focal parameter, we have to set the focal parameter in turn to a range of values, and for each value optimize the likelihood with respect to all of the other parameters.

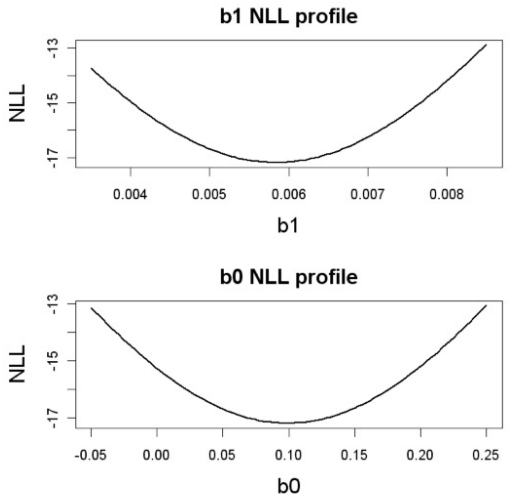
In the brown creeper example shown here, the likelihood profile for the slope b_1 and intercept b_0 are plotted on the likelihood surface for b_1 and b_0. For example, the likelihood profile for b_1 was obtained by systematically varying its value between 0.0035-0.0085 and for each value finding the value of the other parameters b_0 and σ that minimized the negative log-likelihood. On the two-parameter likelihood surface shown here, this is equivalent to finding where a horizontal transect across the surface at each value of b_1 intercepted the minimum negative log-likelihood, which is at the point at which the horizontal transect is tangent to a contour line (i.e., the point at which the contour line is perfectly horizontal). This is the corresponding value of b_0 that optimized the negative log-likelihood.

Frequentist Parametric Inference...

Confidence intervals for model parameters

Likelihood profiles:

- Depict "ridgelines" in parameter space showing the minimum negative log-likelihood for particular values of a *single* parameter
- To calculate a profile for a particular parameter, we set the focal parameter in turn to a range of values, and for each value optimize the NLL with respect to all other parameters



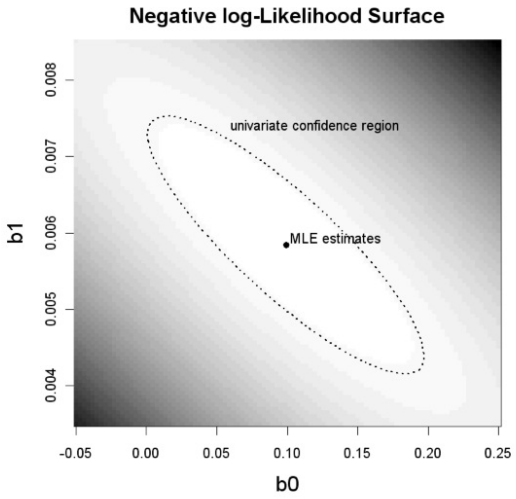
It is perhaps easiest to visualize the likelihood profile as a one-dimensional curve in which the minimum negative log-likelihood across all other parameters is plotted against fixed values of the focal parameter. The profiles shown here depict how the degradation in the model fit as b_1 and b_0 vary systematically over a reasonable range of values. The maximum likelihood estimate is the point where the negative log-likelihood curve is minimum and the shape of the curves depict how much confidence we should have in each of the estimates.

Frequentist Parametric Inference...

Confidence intervals for model parameters

Profile confidence intervals:

- On a negative log-likelihood curve or surface, the steeper and narrower the valley (i.e., the faster the fit degrades as we move away from the best fit), the more precisely we can estimate parameters
- *Rule of thumb*: include parameter values within 2 negative log-likelihood units of the minimum



On a negative log-likelihood curve or surface, the steeper and narrower the valley (i.e., the faster the fit degrades as we move away from the best fit), the more precisely we can estimate parameters. Thus, the likelihood profile contains information that we can use to create confidence intervals for our parameters. In addition, since the negative log-likelihood for a set of independent observations is the sum of the individual negative log-likelihoods, adding more data makes the likelihood curves steeper, which means that our confidence in the estimates gets greater – which makes sense.

It makes sense to determine confidence limits by setting some upper limit on the negative log-likelihood and declaring that any parameters that fit the data at least that well are within the confidence limits. The steeper the likelihood surface, the faster we reach the limit and the narrower are the confidence limits. Since we care only about the relative fit of different models and parameters, the limits should be relative to the minimum negative log-likelihood. A common rule of thumb is to include parameter values within 2 negative log-likelihood units of the minimum, which corresponds to all fits that gave likelihoods within a factor of $e^2 \approx 7.4$ of the maximum. As shown, a univariate confidence region based on the minimum plus 2 rule is plotted on the likelihood surface for b_1 and b_0. However, this approach lacks a frequentist probability interpretation – there is no corresponding p-value. This deficiency may actually be an advantage, since it makes dogmatic null-hypothesis testing impossible.

Frequentist Parametric Inference...

Confidence intervals for model parameters

Profile confidence intervals based on Likelihood ratio test (LRT):

- Take likelihood function (L) and find maximum likelihood

$$L\left\{\hat{Y}|\ \hat{\varphi}_1, \hat{\varphi}_2, \ldots, \hat{\varphi}_m\right\}$$

- Fix one parameter in turn to a range of values and in turn find maximum likelihood with respect to all others (L_r) – call this the restricted model

$$L_r\left\{\hat{Y}|\ \varphi_1^*, \hat{\varphi}_2, \ldots, \hat{\varphi}_m\right\}$$

Fix focal parameter

- Twice the negative log of the likelihood ratio, called deviance, is approx chi-squared distributed with r degrees of freedom

$$deviance = -2\log\left(L_r/L\right) = 2(NLL_r - NLL) \sim \chi_r^2$$

r = difference in #parameters
r = 1 in this case

Likelihood Ratio Test:

If we want confidence intervals with a *p*-value interpretation, we can use the differences in log-likelihoods (corresponding to ratios of likelihoods) in a frequentist approach called the *Likelihood Ratio Test* (LRT). Take some likelihood function and find the overall best (maximum likelihood) value:

$$L\left\{\hat{Y}|\ \hat{\varphi}_1, \hat{\varphi}_2, \ldots, \hat{\varphi}_m\right\}$$

Now fix one (or more) of the parameters, say φ^*, and maximize with respect to the remaining parameters, and called this the maximum likelihood of the restricted (or reduced or nested) model:

$$L_r\left\{\hat{Y}|\ \varphi_1^*, \hat{\varphi}_2, \ldots, \hat{\varphi}_m\right\}$$

The likelihood ratio test says that twice the negative log of the likelihood ratio, called the *deviance*, is approximately χ^2 ("chi-squared") distributed with r (difference in # parameters between full and reduced models; 1 in this case) degrees of freedom. This is equivalent to twice the difference in the negative log-likelihoods between the restricted and original model:

$$deviance = -2\log\left(L_r/L\right) = 2(NLL_r - NLL) \sim \chi_r^2$$

Frequentist Parametric Inference...

Confidence intervals for model parameters

Profile confidence intervals based on Likelihood ratio test (LRT):

- Univariate confidence intervals:

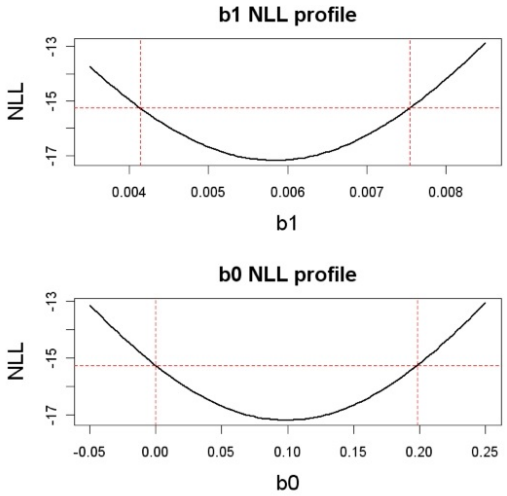
$$NLL + \chi^2_1(1-\alpha) / 2$$

$\alpha = 0.05$ for 95% ci

degrees of freedom =1

$$\chi^2_1(1-0.05) / 2 = 1.92$$

$$(qchisq(0.95, df = 1)) / 2 = 1.92$$



The definition of the likelihood ratio test echoes the definition of the likelihood profile, where we fix one parameter and minimize the negative log-likelihood with respect to all the other parameters: $r = 1$ in the definition above. Thus, for *univariate* confidence limits we cut off the likelihood profile at:

$$NLL + \chi^2_1(1-\alpha) / 2$$

where a is our chosen type I error level (e.g., 0.05 for a 95% confidence interval). The cutoff is a one-tailed test, since we are interested only in differences in likelihood that are larger than expected under the null hypothesis. Note, the degrees of freedom r is one for a univariate confidence interval since we are fixing only one parameter. The 95th quantile of χ^2 distribution with 1 degree of freedom equals 3.84, divided by 2 equals 1.92. Consequently, the univariate confidence interval is almost identical to the "rule-of-thumb" confidence interval of plus 2 negative log-likelihood units from the minimum.

Frequentist Parametric Inference...

Confidence intervals for model parameters

Profile confidence intervals based on Likelihood ratio test (LRT):

- Bivariate confidence intervals:

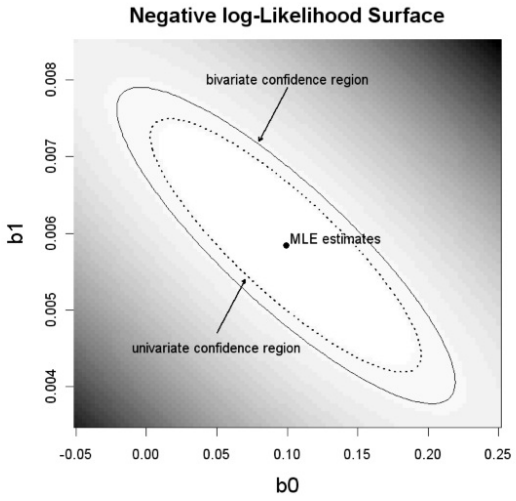
$$NLL + \chi^2_2(1-\alpha) / 2$$

$\alpha = 0.05$ for 95% ci

degrees of freedom =2

$$\chi^2_2(1-0.05) / 2 = 3.00$$

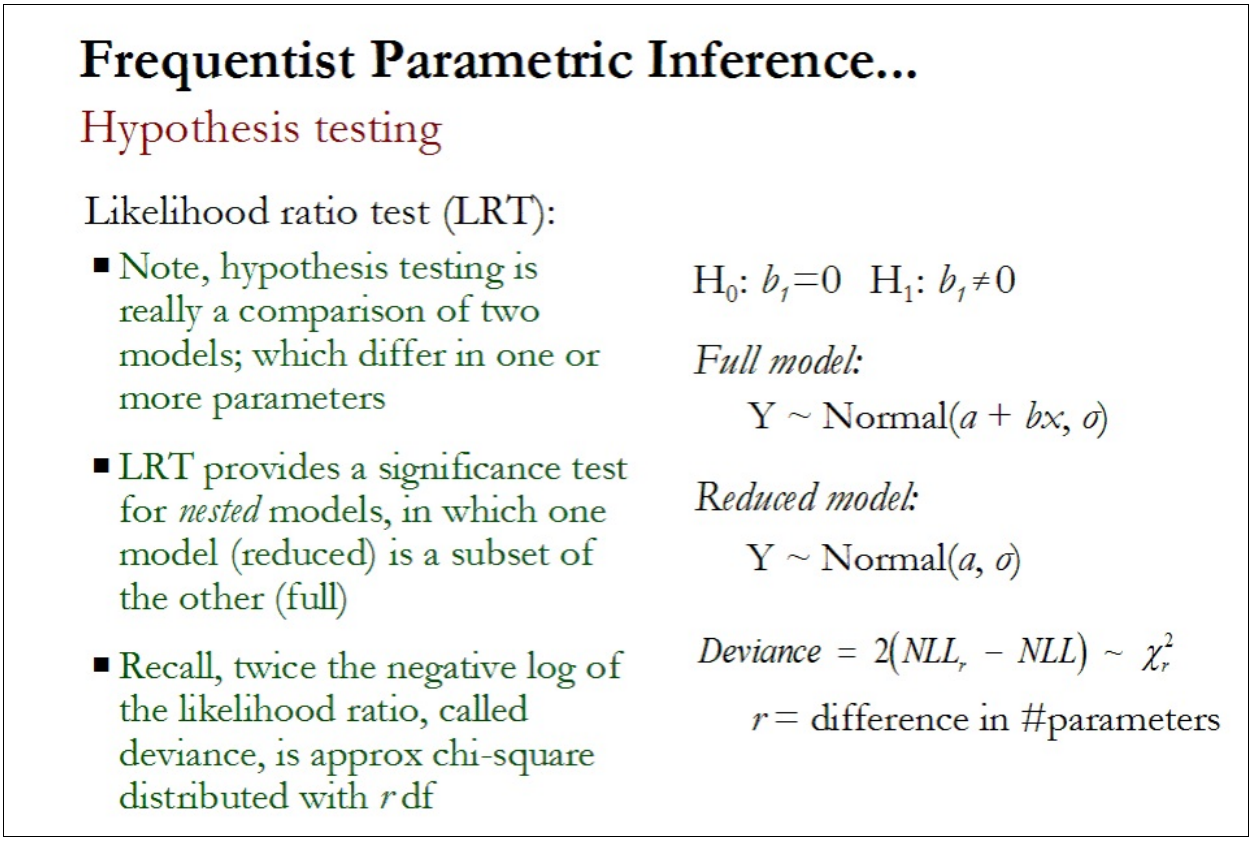
$$(qchisq(0.95, df = 2)) / 2 = 3.00$$



What if we want to establish confidence limits on two parameters simultaneously? We need *bivariate* confidence limits instead of univariate confidence limits. For bivariate confidence limits we cut off the likelihood profile at:

$$NLL + \chi^2_2(1-\alpha) / 2$$

which looks just like the univariate confidence limit except that now instead of 1 degree of freedom we have 2. The 95th quantile of χ^2 distribution with 2 degrees of freedom equals 5.99, divided by 2 equals roughly 3.00. Not surprisingly, the bivariate confidence region is larger than the univariate confidence region since we need to account for the uncertainty in two parameters instead of one. In the figure shown here, the bivariate confidence region is depicted on the negative log-likelihood surface for b_1 and b_0 along with the smaller univariate confidence region, although the later is not really appropriate for this figure, because it applies to a single parameter at a time, but it illustrates that univariate confidence intervals are smaller than the bivariate confidence regions.



5. Hypothesis testing

Now that we have fit the model; i.e., we found the maximum likelihood estimates of the parameters that make the data the most likely, the next step in a classical frequentist framework is to test whether the model is statistically significant. In a likelihood framework, hypothesis testing is really just a comparison of the likelihood (or negative log-likelihood) of two models which differ in one or more parameters. With adherence to the goal of parsimony (also called "Occam's razor"), we generally want to chose the simplest model that can explain the data even though we know the world is more complex. Hypothesis testing, and model selection in general, approaches typically go beyond parsimony to say that a more complex model must be not just better than, but a specified amount better than, a simpler model before we will accept it over the simpler model. If the more complex model doesn't exceed a threshold of improvement in fit, we typically reject it in favor of the simpler model. Model complexity also affects our predictive ability. The more complex we make a model, the better we are able to explain the data in hand. However, when we attempt to make predictions with the model, we may fail to make accurate predictions. This can happen because the model is so fine-tuned to the data set in hand that it is no longer useful for anything but explaining the data in hand. We call this phenomenon "overfitting", because the model is so overfit to the data that it loses its predictive ability. So how can we tell when we are overfitting real data?

Frequentist Parametric Inference...

Hypothesis testing

Likelihood ratio test (LRT):

- Note, hypothesis testing is really a comparison of two models; which differ in one or more parameters
- LRT provides a significance test for nested models, in which one model (reduced) is a subset of the other (full)
- Recall, twice the negative log of the likelihood ratio, called deviance, is approx chi-squared distributed with r degrees of freedom

H_0: $b_1=0$ H_1: $b_1 \neq 0$

$NLL = -17.17989$

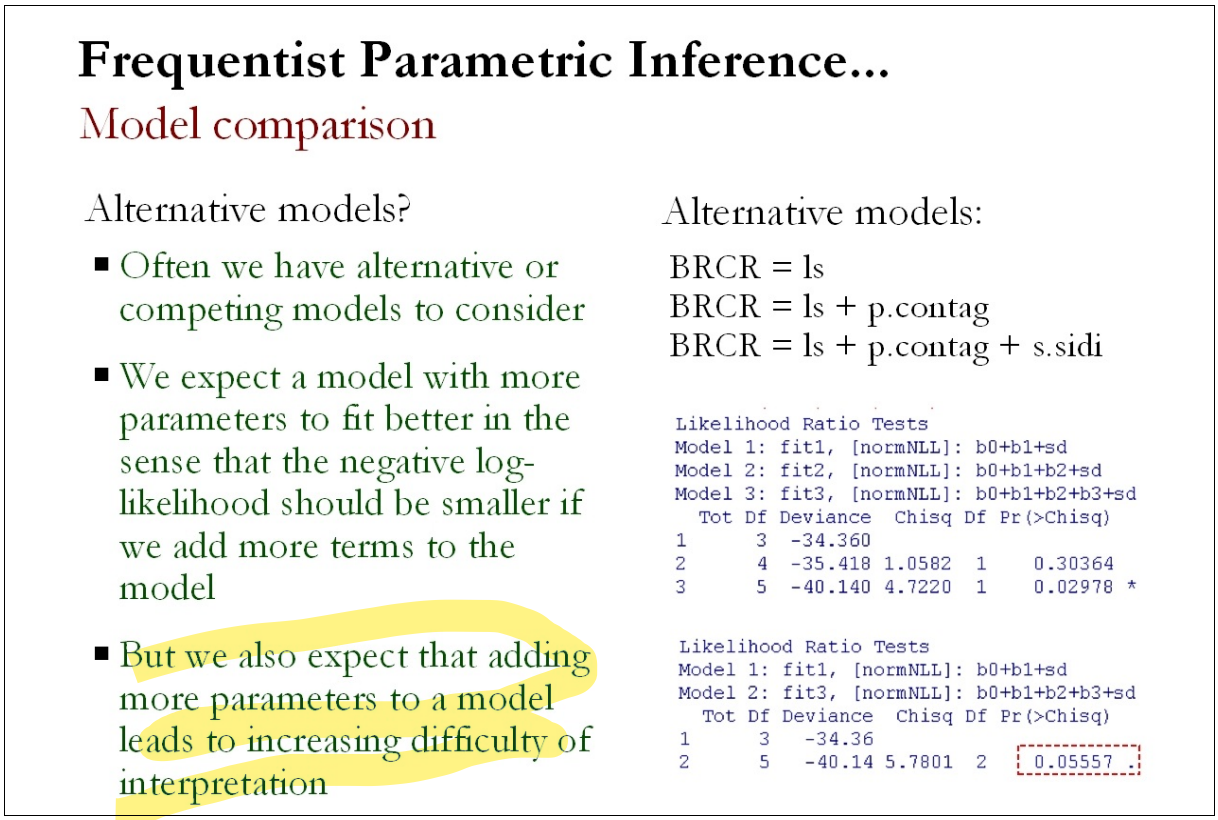
$NLL_r = -2.815654$

$Deviance = 28.72847$

p-value = 1-pchisq(28.73, df=1)

p-value = 8.33e-08

We can use the *Likelihood Ratio Test* (LRT), which we used before to find confidence intervals and regions, to choose models in certain cases. A simpler model (with fewer parameters) is nested in another, more complex model (with more parameters) if the complex model reduces tot he simpler model by setting some parameters to particular values (often zero). For example, a constant model, $y = b_0$, is nested in the linear model, $y = b_0 + b_1x$, because setting $b_1 = 0$ makes the linear model constant. The LRT provides a significance test for nested models. Recall that twice the negative log of the likelihood ratio of the nested models, deviance, is approximately χ^2 distributed with r degrees of freedom. In our linear model example, we can test the null hypothesis that $b_1 = 0$, which is equivalent to testing whether the linear model is significantly better than the constant (intercept only) model. Twice the difference between the negative log-likelihood of the constant model and the negative log-likelihood of the linear model, or deviance, is equal to 28.73, which is distributed χ^2 with 1 degree of freedom (since the difference in number of parameters is 1). A deviance this large or larger would be expected almost never ($p<0.0001$) under the null model; i.e., if the constant model were true. So we can reject the null hypothesis in favor of the linear model as being a significantly better fit.



6. Model comparison

The LRT hypothesis test that we just described involves testing the significance of a parameter by framing it as a comparison of two nested models, where the model without the parameter (reduced) is compared to the model with the parameter (full). However, this is just a special case of comparing two alternative or competing models that differ in parameters. Often we have multiple alternative or competing models that we want to consider. We expect a model with more parameters to fit better in the sense that the negative log-likelihood should be smaller if we add more terms to the model. But we also expect that adding more parameters to a model leads to increasing difficulty of interpretation. So how do we compare a model with m parameters to a model with p parameters?

Let's say that we wish to consider three competing models of increasing complexity:

Model 1: BRCR = ls
Model 2: BRCR = ls + p.contag
Model 3: BRCR = ls + p.contag + s.sidi

Because these models are nested in terms of the explanatory variables (i.e., model 1 is nested with model 2, which is nested within model 3), we can use the LRT to compare pairs of increasingly complex models. The results shown here indicate that model 2 is not a significant improvement over model 1 (p=.304) and that we should therefore accept the simpler model 1 over the more complex model 2. However, model 3 is a significant improvement over model 2 (p=0.030), but only mildly significantly better than model 1 (p=0.056).

Frequentist Parametric Inference...

Model comparison

Information criteria:

- Compares all models at once and does not require nested models
- Based on the expected "distance" between a particular model and the "true" model
- Penalized goodness(badness)-of-fit criterion – penalty for additional parameters – differs among criteria

Akaike Information Criterion (AIC):

$$AIC = 2NLL + 2m$$

m = #parameters

Small sample AIC (n/m<40):

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

Information criteria:

The LRT approach can work well when all of the models are nested, but even so involves a series of pairwise comparisons which can make interpreting the results more complex. One way to avoid a plethora of pairwise model comparisons is to select models based on *information criteria*, which compare all candidate models at once and do not require nested alternatives. These relatively recent alternatives to LRT are based on the expected distance (quantified in a way that comes from information theory) between a particular model and the "true" model. In practice, all information-theoretic methods reduce to finding the model that minimizes some criterion that is the sum of a term based on the likelihood (usually twice the negative log-likelihood) and a *penalty term* which is different for different information criteria.

The *Akaike Information Criterion*, or AIC, is the most widespread information criterion and is defined as:

$$AIC = 2NLL + 2m$$

where m is equal to the number of model parameters. As with all information criteria, small values represent better overall fits; adding a parameter with a negligible improvement in fit penalizes the AIC by 2 log-likelihood units, which is similar to the significance threshold for the LRT test with 1 degree of freedom. For small sample sizes (n), such as when $n/m < 40$, there is a finite-size correction to AIC:

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

Frequentist Parametric Inference...

Model comparison

Information criteria:

- Not suitable for significance testing (like LRT)
- Rule of thumb:
  - ΔAIC<2 : equivalent
  - ΔAIC 4-7 : distinguishable
  - ΔAIC>10 : definitely different
- AIC model weight: relative likelihood of a model – "probability" of the model – given the data

Alternative models:

BRCR = ls
BRCR = ls + p.contag
BRCR = ls + p.contag + s.sidi

AIC_c Table:

	dAICc	df	weight
fit3	0.0	5	0.434
fit1	0.2	3	0.392
fit2	1.8	4	0.174

$$W_i = \frac{e^{(-0.5*\Delta AIC_i)}}{\sum_{i=1}^{m} e^{(-0.5*\Delta AIC_i)}}$$

There are other information criteria besides AIC, but they all work on the same principle. Importantly, information criteria do not allow frequentist significance tests based on the estimated probability of getting more extreme results in repeated experiments. Some would claim this is an advantage. With information criteria, we cannot say that there is a significant difference between models; a model with a lower AIC for example is better, but there is no *p*-value associated with how much better it is. Instead, there are some commonly used rules of thumb: models with information units less than 2 apart, delta AIC for example, are more or less equivalent; those with 4-7 information units difference are clearly distinguishable; and models with >10 information units difference are definitely different. The model with the lowest information units is the "best" model, but those with differences of <10 are worth considering. One way to approach this situation is with AIC weights, which give the relative likelihood of a model. Model weights are based on the delta AIC values and are often interpreted as giving the "probability" of the model given the data – which is a Bayesian like interpretation of the support for a particular model, although weights are not true probabilities and should not be confused with Bayesian posterior probabilities (discussed in the next chapter).

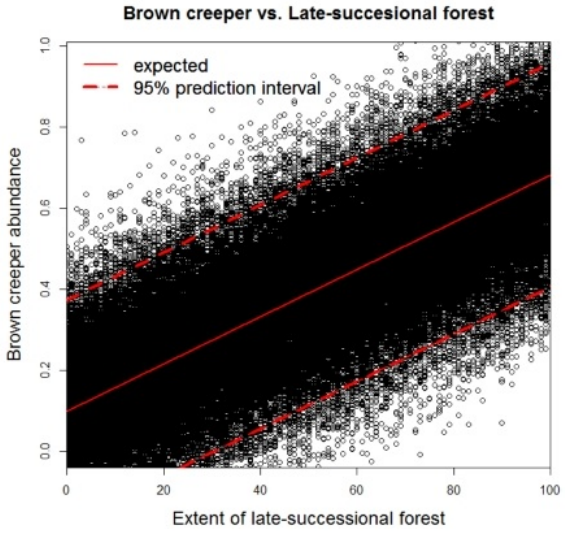
In the example shown here, model 3 is selected as the "best" model based on AIC corrected for small sample size, but model 1 and model 2 are both nearly as good since they are within 2 AIC units of the best model. Model weights indicate that there is almost as much weight of evidence in favor of either model 3 as model 1, but that model 2 has much less support.

Frequentist Parametric Inference...

Predictions

Parametric predictions:

- *Point estimates*... apply the fitted deterministic model to new values of x
- *Interval estimates*... one approach is to use the fitted statistical model to simulate new values and then construct quantile intervals from the predicted values



7. Predictions

The final goal of statistical inference is to make predictions. In many cases, once we confirm that our model is a good one, say by confirming that it is significantly better than the null model (e.g., of no relationship between x and y) or that it is the best among competing models considered, we might want to use this model to predict values for future observations or for sites not sampled.

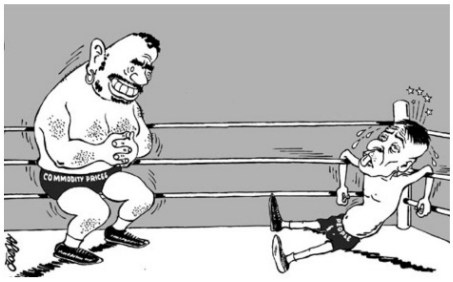
Point estimates for predictions are relatively straightforward, simply plug in the new values for the independent (predictor) variables into the fitted model equation for the deterministic component to get the expected values. Interval estimates for our predictions are more difficult and there are many nuanced approaches for deriving them for different situations. However, one relatively straightforward approach is to simply use the fitted statistical model to simulate new values and then construct quantile intervals from the predicted values (as shown in the figure here). If we are willing to assume that the errors are independent and identically normally distributed, then we can construct a prediction interval as described previously in the chapter on nonparametric inference based on ordinary least squares.

Frequentist Parametric Inference...

Pros and Cons of Maximum Likelihood

- Parametric method --requires assumptions about the error distribution
- Stronger statistical inference as a result of above
- Likelihood can be difficult to figure out for complex models
- Basis for much of modern statistical inference

$Y \sim \text{Normal}(a + bx, \sigma^2)$



8. Pros and cons of maximum likelihood inference

The frequentist parametric inference framework based on maximum likelihood methods is powerful but not without some drawbacks.

1. *Parametric method*... It is a parametric approach and therefore it requires assumptions about the error distribution.
2. *Stronger statistical inference*... As a consequence of number one above, in general the inferences from a parametric procedure are stronger than from a nonparametric procedure. This is because the parametric statistical model is a complete description of the underlying population, and we can conceive of the model as a data-generating mechanism for the population.
3. *Likelihood specification*...One of the biggest challenges confronting users of maximum likelihood methods is the specification of the likelihood function, or negative log-likelihood function, for complex models. With simple models this is usually not a problem, as there are built in functions in R for simple and even moderately complex models. However, as models increase in complexity, it becomes the responsibility of the user to write the likelihood function and this can be extremely challenging for most ecologists, even those with a good background in statistics.
4. *Modern statistical inference*... Maximum likelihood is the basis for much of modern statistical inference in ecology, although the Bayesian framework is rapidly gaining in popularity. Nevertheless, the likelihood function is used in both maximum likelihood and Bayesian methods, so it behooves us to master our understanding of likelihood-based methods.