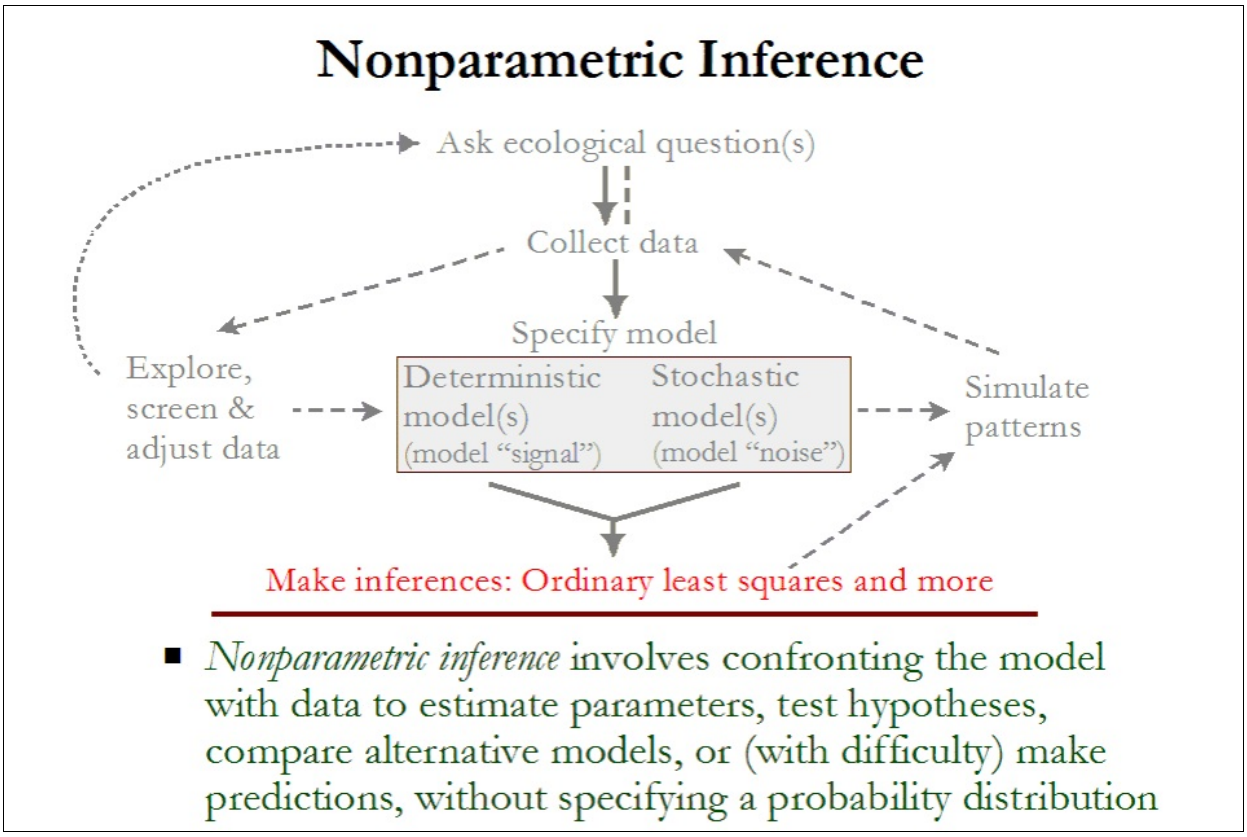
Analysis of Environmental Data

Chapter 7. Conceptual Foundations:

Nonparametric Inference : Ordinary Least Squares and More



1. Nonparametric inference

In the previous chapters we introduced all the ingredients needed to define a statistical model – mathematical functions to describe the deterministic patterns and probability distributions to describe the stochastic patterns – and described how to use these ingredients to simulate simple environmental systems. The final steps of the modeling process are estimating parameters from data, optionally testing null hypotheses (i.e., do the parameter estimates differ significantly from pre-specified values), comparing alternative models, and making predictions for observations not yet collected – the stuff of statistical inference. In this chapter, we will describe methods for inference in a nonparametric framework; i.e., when it is not possible to specify a known probability distribution for the stochastic component of the model. Since nonparametric methods don't assume any particular underlying distribution, they do not involve making an explicit link to the underlying population from which the sample was drawn. Instead, they base all inferences on the sample itself. Consequently, some refer to nonparametric methods such as ordinary least squares as a "noninference" framework, since technically one is not able to infer characteristics of the population without making an assumption that the sample was drawn from a particular population. Nevertheless, it is perhaps more useful to think of nonparameteric approaches such as ordinary least squares as an inference framework, but one with only weak inferential power compared to parametric methods.

Nonparametric Inference...

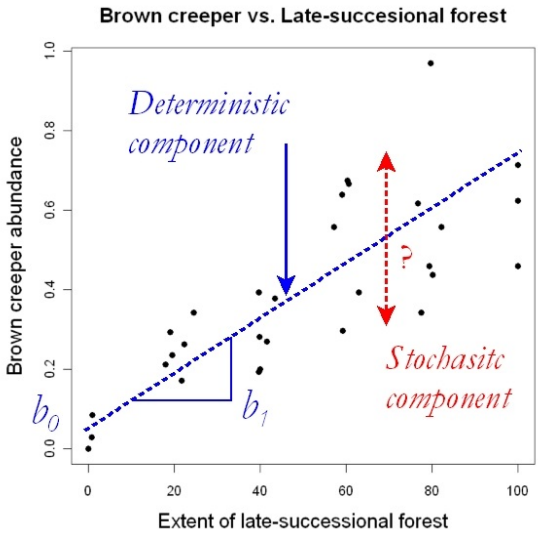
What is the statistical model?

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

Statistical Model:

$$y_i = b_0 + b_1 x_i + e_i$$

Parameters Unspecified



2. The (nonparametric) statistical model

Given the question, the first step in model-based inference is to propose a statistical model that we wish to confront the data with. In a nonparametric approach we need only to specify the deterministic component, since the error component does not need formal specification. Note, this does not mean that model contains no error, only that its distribution is not specified in the model.

Example: Let's begin with a now familiar question: do brown creepers increase in relative abundance with increasing extent of late-successional forest in the Oregon Coast Range? The data represent the relative abundance of brown creepers (# per unit area per unit effort) across 30 subbasins in the Oregon Coast Range which vary in their composition of late-successional forest. Given this question, a logical deterministic model would be a linear relationship between brown creeper abundance and late-successional forest extent, since this is the simplest (i.e., most parsimonious) and yet meaningful model that we can propose. A nonlinear model might also be appropriate, but we would need more information a priori to propose the form of such a relationship. Note, we could do some exploratory data analysis with the data in hand to see if a nonlinear relationship is apparent and then select an appropriate mathematical function based on the patterns observed – a phenomenological approach, but let's avoid doing so to avoid possible criticism for data-dredging.

Nonparametric Inference...

Estimate model parameters: OLS method

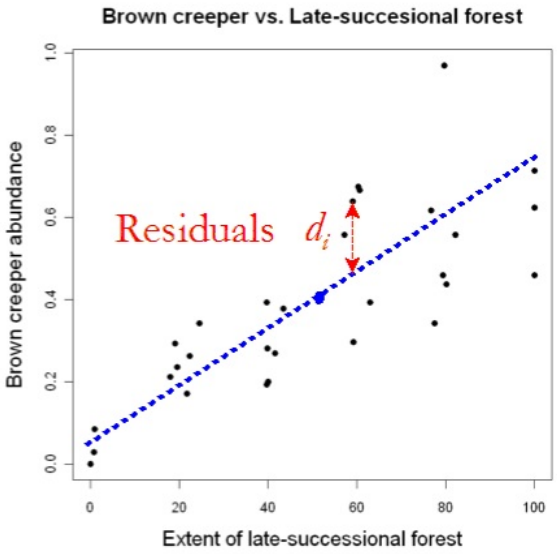
1. Define measure of (lack of) fit:

$$d_i = y_i - \hat{y}_i$$

$$\hat{y}_i = b_0 + b_1 x_i$$

$$d_i = y_i - b_0 - b_1 x_i$$

$$L(Y_i | b_0, b_1) = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$



3. Parameter estimation: Ordinary Least Squares

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. First, we need to define an objective measure of fit that we can maximize (or a measure of 'lack of fit' that we can minimize) in order to find the parameters that fit the data the best. The method of Ordinary Least Squares (or sums of squares) finds the parameters φ that minimize the sum of the squared residuals (hence the name "least squares" or "ordinary least squares"), where the 'residuals' are defined as the vertical differences between the data (points in scatterplot) and the fitted model (line in scatterplot).

Example: Let's see how this works for our linear model example. Each residual is a distance, d_i, between a data point, y_i, and the value predicted by the fitted model, $\hat{y}_i$, evaluated at the appropriate value of the explanatory variable, x_i:

$$d_i = y_i - \hat{y}_i$$

Now, we replace the predicted value $\hat{y}_i$ by its formula, noting the change in sign:

$$\hat{y}_i = b_0 + b_1 x_i$$

$$d_i = y_i - b_0 - b_1 x_i$$

Finally, our measure of lack of fit is the sum of the squares of these distances:

$$L(Y_i | b_0, b_1) = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

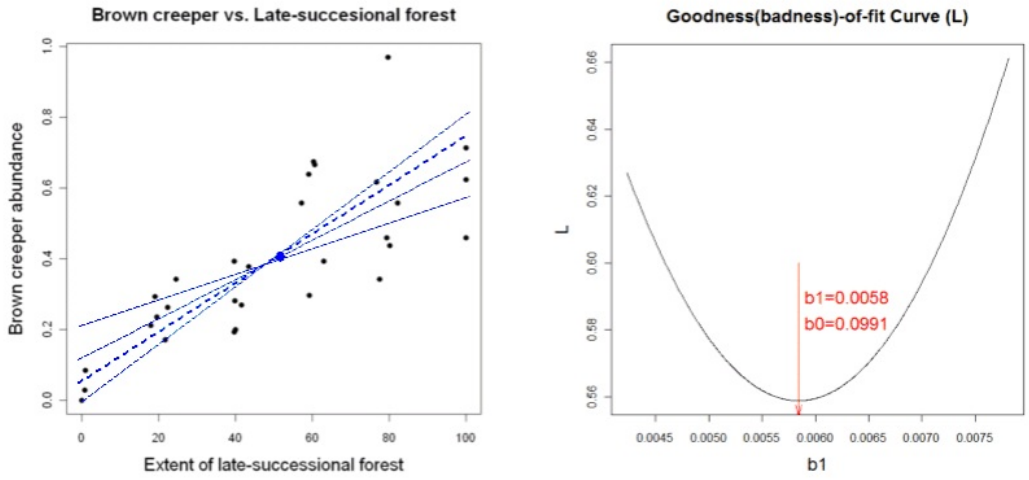
where $L(Y_i | b_0, b_1)$ refers to the likelihood of the data Y_i given values of the parameters b_0 and b_1, where the likelihood is defined here as the sums of squared deviations.

Nonparametric Inference...

Estimate model parameters: OLS method

2. Find estimates that minimize $L(Y_i | b_0, b_1)$

- Numerical solution



The best estimates of our model parameters are those that minimize our measure of lack of fit (L). So how do we find the parameters that minimize L? We can think of OLS as working as follows. The best fit line is defined as passing through the point defined by the mean value of x ($\bar{x}$) and the mean value of y ($\bar{y}$) – the centroid of the data. Imagine that the straight line through this point is pivoted until the sum of squared residuals L is minimum. This solution can be achieved numerically by trying lots of values of the parameters and choosing the set that minimizes the objective function. If we compute the lack of fit L for each combination and plot the result, we produce a goodness(badness)-of-fit curve. The values of b_0 and b_1 that minimize this curve are the best estimates of our model parameters.

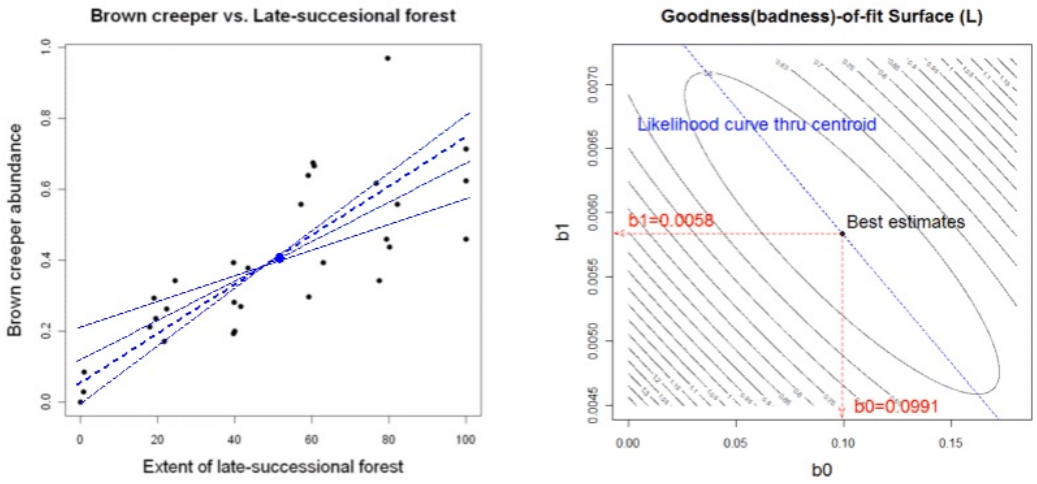
Example: In our example, if we allow b_1 to vary between 0.0045-0.0072 and then set b_0 to the value that forces the line thru the centroid, and then for each combination of values recalculate L, we can plot the results as a goodness (or badness)-of-fit curve. The curve depicts the value of L for every combination of parameter values evaluated; i.e., each possible line thru the centroid. The lowest point on this curve represents the combination of parameter values that minimizes L, our badness-of-fit metric. In this case we can see that best estimate of b_1=0.0058 and thus b_0=0.0991.

Nonparametric Inference...

Estimate model parameters: OLS method

2. Find estimates that minimize $L(Y_i | b_0, b_1)$

- Numerical solution



While in the case of linear regression it may make sense to force the line thru the centroid, in other cases we may not want to impose such constraints. In the more general case, we want to find the parameters that minimize *L*? We can think of OLS as working as follows. The best fit line is found by exploring the entire parameter space by trying all possible combinations of intercept and slope values – without forcing the line thru the centroid. Again, this solution can be achieved numerically by trying lots of values of the parameters and choosing the set that minimizes the objective function. If we compute the lack of fit *L* for each combination and plot the result, we produce a goodness(badness)-of-fit surface. The values of b_0 and b_1 that minimize this surface are the best estimates of our model parameters.

Example: In our example, if we allow b_0 to vary between 0-0.18 and b_1 to vary between 0.0045-0.0072, and then for each combination of values recalculate *L*, we can plot the results as a goodness (or badness)-of-fit surface. The surface depicts the value of *L* for every combination of parameter values evaluated. The lowest point on this surface represents the combination of parameter values that minimizes *L*, our badness-of-fit metric. The contours are not close enough near the bottom of the surface to see precisely where the minimum is, but we can compute it to be exactly b_0=0.0991 and b_1=0.0058.

Nonparametric Inference...

Estimate model parameters: OLS method

2. Find estimates that minimize $L(Y_i | b_0, b_1)$

- Analytical solution

$$\frac{dL(Y_i|b_0,b_1)}{db_1} = -2\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i)$$

Set to zero and solve for b_1

Sums of squares & products:

$$SSXY = \sum_{i=1}^{n} (x_i - \bar{x}_i)(y_i - \bar{y}_i)$$

$$SSX = \sum_{i=1}^{n} (x_i - \bar{x}_i)^2$$

$$SSY = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

$$b_1 = \frac{SSXY}{SSX}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Example:

$b_1 = 0.0058$
$b_0 = 0.0991$

However, ideally we want an analytical solution that gives the best estimates directly. With our simple linear model, it turns out that we can use calculus to find the solution, but this is not always the case with more complex models. In this case, all we need to do is find the derivative of *L* with respect to the slope (b_1), set this equal to zero and solve for b_1:

$$\frac{dL(b_0, b_1)}{db_1} = -2\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i)$$

We will not derive the solution here (see most introductory stats books). But briefly, to find the solution, we need to compute the following 'corrected sums of squares and products'; although only the first two below are needed for this purpose, but the third will be used later:

$$SSXY = \sum_{i=1}^{n} (x_i - \bar{x}_i)(y_i - \bar{y}_i)$$

$$SSX = \sum_{i=1}^{n} (x_i - \bar{x}_i)^2$$

$$SSY = \sum_{i=1}^{n} \left(y_i - \bar{y}_i \right)^2$$

The estimates of the slope, b_1, and intercept, b_0, are:

$$b_1 = \frac{SSXY}{SSX}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Example: In our example, if we substitute the corresponding values into the equations above, we get estimates of b_0=0.0991 and b_1=0.0058, which are the roughly same values we obtained using the numerical approach above.

Nonparametric Inference...

Estimate model parameters: OLS method

Pros and Cons of OLS Estimation:

- No assumptions about the error required
- Squared deviations make analytical solutions easier
- If the errors are normally distributed, then the sums of squares is identical to other methods of estimation
- No a priori justification for using the squared measure of deviation, which has an accelerating penalty

Pros and cons of ordinary least squares estimation

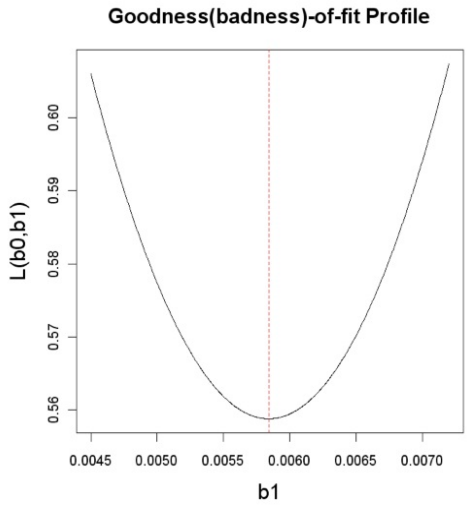
- No assumptions about the stochastic component of the model are required, making this an extremely flexible method.
- If one is trying to find an analytical solution, squared deviations are good because the derivatives are easily found.
- If the errors are normally distributed, then the sums of squares is identical to other methods of estimation (such as maximum likelihood, discussed later).
- The squared measure of deviation has an accelerating penalty: a deviation that is twice as large contributes four times as much to the sums of squares. There is no a priori reason to choose such a measure.

Nonparametric Inference...

Confidence intervals for model parameters

What is a confidence interval?

- *Interval* estimate of the uncertainty associated with each of the estimated parameters; in other words, the precision of our estimate
- "Were this procedure to be repeated on multiple samples, the calculated confidence interval (which would differ for each sample) would encompass the true population parameter say 95% of the time"



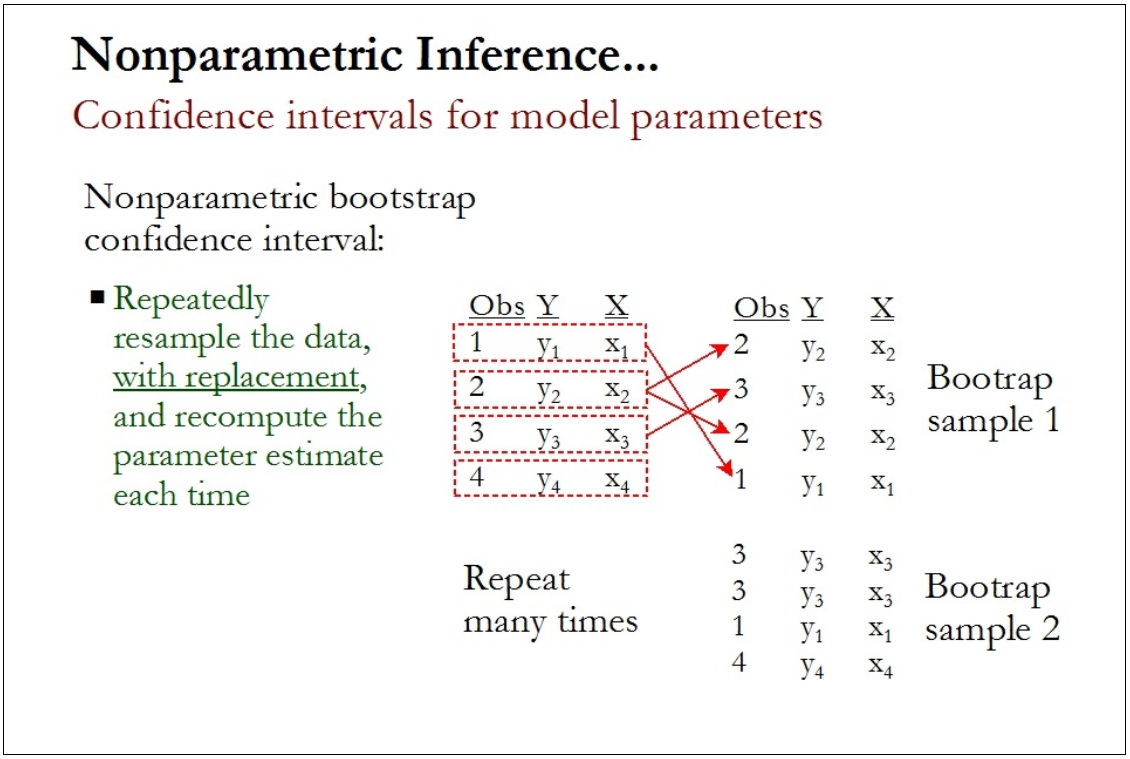
4. Confidence intervals

Thus far, our parameter estimation has focused on so-called 'point' estimates of the parameters. For a frequentist, a point estimate is our best estimate of the true, fixed underlying value in the population, or the expected (or average) value of our sample estimate under hypothetical repeated sampling. However, usually we are interested in knowing more than just the point estimate. We might also want to determine the unreliability or uncertainty associated with each of the estimated parameters; in other words, the precision of our estimate. Our estimate may be the most precise of all the possible estimators, but if its value still varies widely under repeated sampling, it will not be very useful for inference. If repeated sampling produces an estimate that is very consistent, then it is precise and we can be confident that it is close to the true parameter (assuming that is also unbiased). Thus, we usually want to compute 'interval' estimates as well as point estimates.

In the classical frequentist framework the 'confidence interval' represents our uncertainty in the fixed (true) but unknown value of the parameter in the underlying population. That is, we assume that there exists a fixed, true value of the parameter in the population and that our sample-based estimate of the parameter is merely a random outcome. Thus, the greater the random variability in the system and the smaller the sample size, the less likely it is that any single sample-based estimate of the parameter will be close to the true value of the parameter. The confidence interval represents this uncertainty. In the classical frequentist context, the confidence interval can have a number of different interpretations, but most often it is interpreted in terms of hypothetical repeated samples:

"were this procedure to be repeated on multiple samples, the calculated confidence interval (which would differ for each sample) would encompass the true population parameter say 95% of the time." Importantly, the confidence interval is not a probability statement about the true parameter value; rather, it is interpreted as follows. If we were to repeatedly sample the population and each time compute a 95% confidence interval, 95% of the time our confidence intervals would be correct, in that they would contain the true value of the parameter. Conversely, 5% of the time they would be wrong. All we can do is hope that our original confidence interval is one of the lucky ones and contains the true value. In this sense, the 'probability' associated with confidence intervals is interpreted as a long-run frequency under hypothetical repeated sampling. Note, this is one of the major differences between the frequentist and Bayesian approaches, as we shall discuss later.

What we need is an objective method of estimating a confidence interval for each of the parameters. It turns out that if we are willing to assume a normal error distribution for the stochastic part of the model, we can usually compute a confidence interval analytically. However, with a nonparametric model, we make no assumptions about the shape of the error distribution and therefore cannot analytically compute a confidence interval. We need to find a different solution. One clever solution is to resample the data in order to generate an empirical sampling distribution for each parameter and then compute the confidence interval directly from this distribution. This method is known as bootstrapping, or the nonparametric bootstrap. Below we will briefly discuss both approaches.



Nonparametric bootstrap confidence interval:

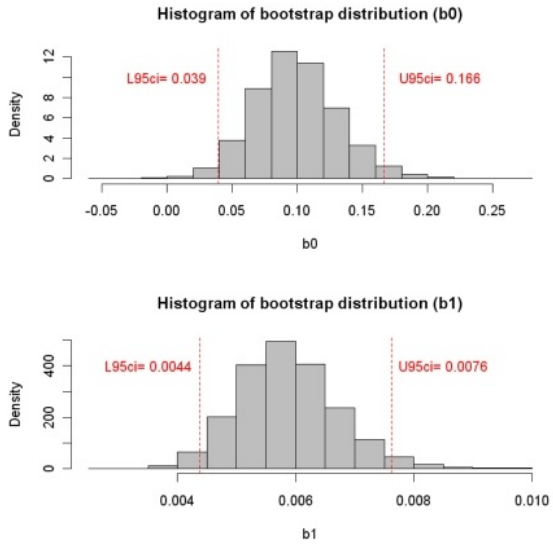
You have probably heard the old saying about "pulling yourself up by your own bootlaces". That is where the term 'bootstrap' comes from. It is used here in the sense of getting 'something for nothing'. The idea is very simple. You have a single sample of *n* measurements, but you can sample from this in very many ways, so long as you allow some values to appear more than once, and other samples to be left out (i.e., sampling with replacement). All you do is draw observations at random from the original sample, replacing each observation after it is selected so that it has the same chance of being drawn on the next draw. By doing this repeatedly, you can create a new data set by resampling the original data set.

Nonparametric Inference...

Confidence intervals for model parameters

Nonparametric bootstrap confidence interval:

- Repeated sampling of the data, with replacement, to empirically generate the sampling distribution of the estimate
- Quantiles of the bootstrap distribution give the specified confidence interval



The bootstrap has many applications in statistics, but by far its most important use involves calculating non-parametric confidence intervals for parameter estimates. In this context, the bootstrap simulates the frequentist concept of obtaining estimates from repeated similar experiments. It substitutes resampling of one data set for repeated experiments. Here, we simply create a bootstrap sample, compute the parameter estimate, create another bootstrap sample, compute the parameter estimate, and repeat this process over and over again, as many as say 10,000 times. The end result is 10,000 bootstrap estimates of the parameter. Each bootstrap sample is like a new sample of the population. Consequently, the distribution of bootstrap estimates reflects the sampling variability we might expect in our point estimate of the parameter if we were to repeatedly sample the population. We simply take the 2.5% and 97.5% quantiles of the bootstrap distribution to compute a 95% confidence interval.

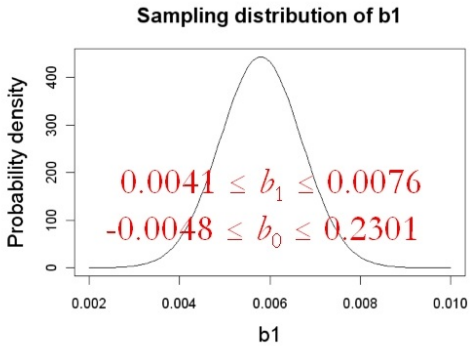
Example: 95% bootstrap confidence intervals: b_0= [0.039,0.166] and b_1=[0.0044,0.0076]

Nonparametric Inference...

Confidence intervals for model parameters

Parametric confidence interval:

- Calculate the *standard error* of the parameter estimate and multiply it by the appropriate value of Student's *t* and then subtract this interval from, and add it to, the parameter estimate to get the corresponding confidence interval



s^2 = Error variance

$$se_{b1} = \sqrt{\frac{s^2}{SSX}} \qquad 95\%CI = b_1 \pm t_{0.025,n-2} se_{b1}$$

$$se_{b0} = \sqrt{\frac{s^2 \sum x^2}{n \cdot SSX}} \qquad 95\%CI = b_0 \pm t_{0.025,n-2} se_{b0}$$

Parametric confidence interval:

Now let's assume that we are willing to assume that the errors are independent and identically distributed according to a normal distribution. In other words, let's assume that the errors (or residuals) are normally distributed about the mean (the expected value from the deterministic part of the model) and the variance in the errors is constant over the full range of x – the explanatory variable. Given these assumptions, a parametric confidence interval can be constructed. Unfortunately, the details of how this done is beyond the scope of this chapter (but see separate primer on confidence intervals), but we will briefly describe how a parametric confidence interval can be computed here.

First we calculate the standard error of the parameter estimate and then multiply it by the appropriate value of Student's *t,* and then subtract this interval from, and add it to, the parameter estimate to get the corresponding confidence interval. Note, the *standard error* is simply the standard deviation in the parameter estimate – that is, the standard deviation in the sampling distribution of the estimate if we were able to generate repeated estimates from resampling the population – and is typically given by the standard deviation of the model error divided by the square root of the sample size, so that the standard error increases with increasing variance and decreases with increasing sample size. There are extra components to the standard error, however, which are specific to the uncertainty of a slope or an intercept. Specifically, the standard error for the slope increases with increasing error variance s^2 and declines with increasing sample size n and the range of x values (as

measured by SSX):

$$se_{b1} = \sqrt{\frac{s^2}{SSX}}$$

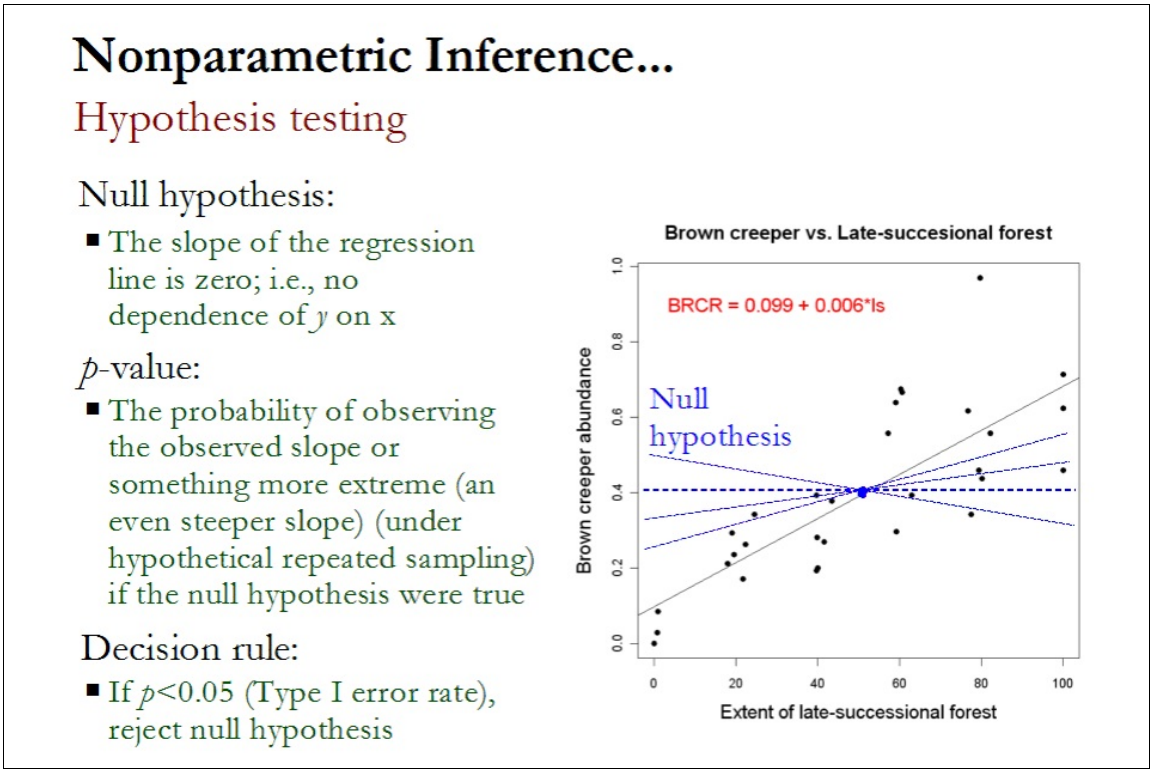
The uncertainty of the estimated intercept increases with increasing variance and declines with increasing sample size. As with the slope, uncertainty declines with increasing sample size n and the range of x values (as measured by SSX). Uncertainty in the estimate of the intercept also increases with the square of the distance between the origin and the mean value of x (as measured by $\sum x^2$):

$$se_{b0} = \sqrt{\frac{s^2 \cdot \sum x^2}{n \cdot SSX}}$$

The confidence interval for each parameter is obtained by subtracting from, and adding to, each parameter estimate an interval which is the standard error times Student's t with the appropriate error degrees of freedom. In this case, the appropriate value of t is given by the 0.975 quantile of the t distribution with n-2 degrees of freedom (= 2.048). Thus, the 95% confidence interval for b_1 is:

$$b_1 \pm t_{0.025,n-2} se_{b1}$$

Example: 95% parametric confidence intervals: b_0= [-0.0048, 0.2301] and b_1=[0.0041, 0.0076]



5. Hypothesis testing

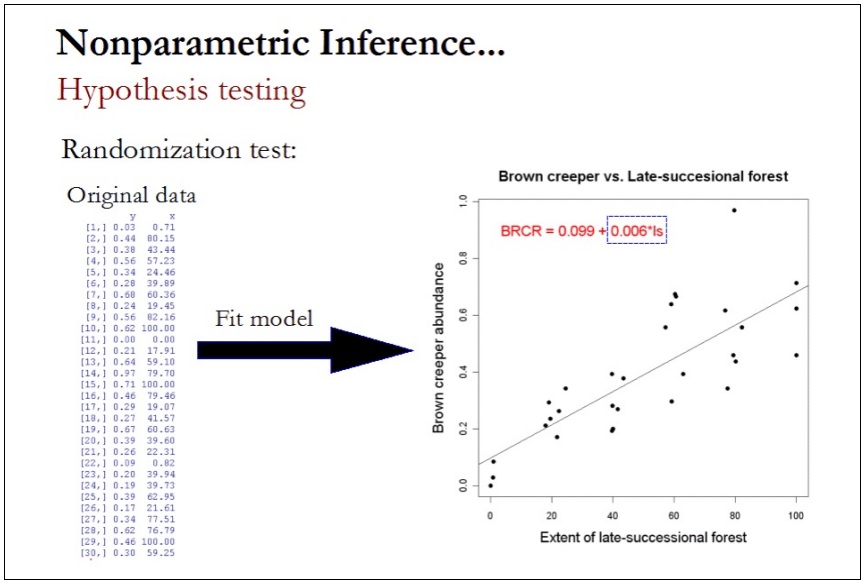
Now that we have fit the model (i.e., have parameter estimates that make the data the most likely based on our goodness-of-fit metric), the next step in a classical frequentist framework is to test whether the model is statistically significant. To do this, we have to first establish a *null hypothesis*, which is usually stated as the absence of a real relationship (e.g., the slope of the regression line is zero; i.e., no dependence of y on x). Then, we can calculate a p-value, which is the probability of observing our data (the observed slope) or something more extreme (an even steeper slope) (under hypothetical repeated sampling) if the null hypothesis were true. Frequentists also typically set up a decision rule for rejecting the null hypothesis, which involves a priori establishing a critical p-value, usually $p<0.05$, below which the null hypothesis will be rejected. This decision rule establishes the rate at which we are willing to accept making a Type I error – rejecting the null hypothesis when it is true. If we reject the null hypothesis with a $p=0.05$, by definition we will be wrong 5% of the time.

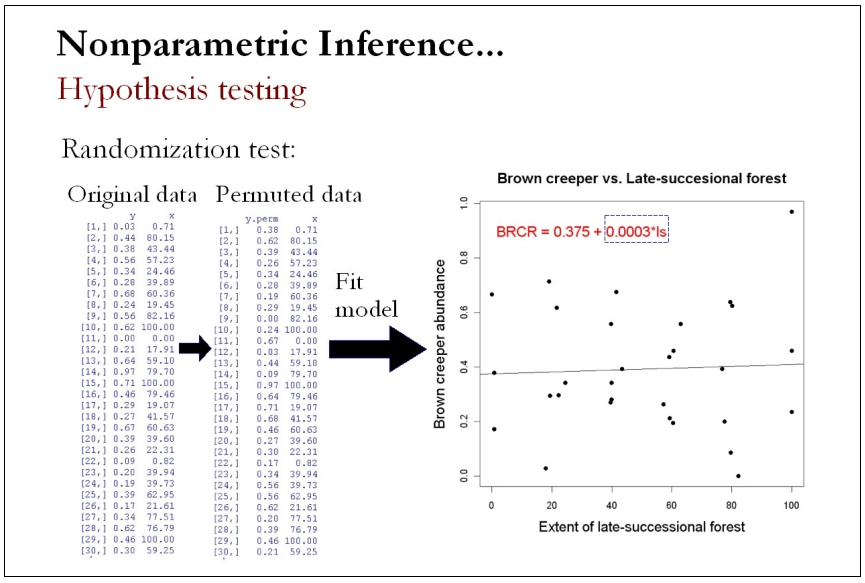
There are at least a couple of different ways to compute a p-value for the null hypothesis of no slope ($b_1=0$). The conventional approach involves computing the relevant test statistic, the Student's t or F-ratio (described below) in this case, for the observed data and comparing this value to the corresponding probability distribution. This approach is *parametric* because a theoretical probability distribution (t or F in this case) is assumed to represent the probability distribution of the test statistic. However, in our *nonparametric* framework we are not willing to assume any particular probability distribution. Thus, we need to find a different solution. Once again a clever solution was found in data resampling. Below we will briefly discuss both approaches.

Nonparametric randomization test of significance:

The bootstrap method described above offers a powerful means of quantifying uncertainty in parameter estimates when parametric approaches are suspect and, as such, it can be used to construct hypothesis tests. For example, does the 95% bootstrap confidence interval contain zero? If not, we can reject the null hypothesis that the parameter is equal to zero with high confidence. An alternative to the bootstrap for testing null hypotheses is the *Monte Carlo randomization test*. Like the bootstrap, the idea is quite simple and can be readily applied in almost any circumstance.

Briefly, the procedure involves repeatedly resampling the original data after removing any real structure (i.e., randomizing the data) to generate an empirical distribution of the test statistic under the null hypothesis of no real structure. There is no need to assume an underlying theoretical (i.e., parametric) distribution because the distribution is generated empirically through resampling the original data. The basic approach is very simple. All we do is randomly shuffle some or all of the data, taking care not to produce observations that fall outside the domain of the original data. The intent is to remove real structure from the data, and this is usually accomplished by shuffling just one of the variables, although this can vary depending on the context of the test. After generating a random permutation of the data, we calculate the test statistic of interest.



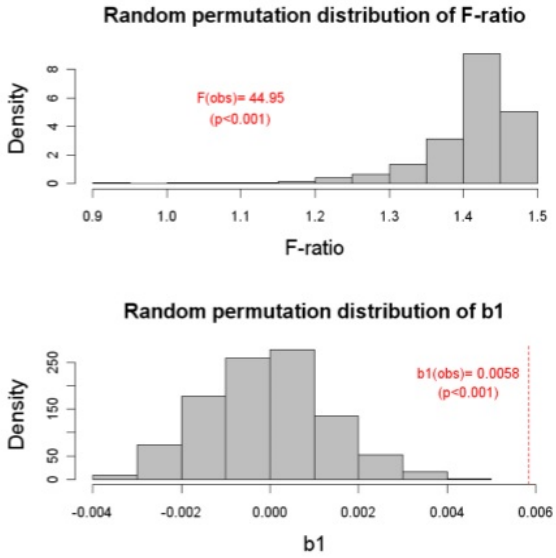


Nonparametric Inference...

Hypothesis testing

Randomization test:

- Repeat permutation and parameter estimation process many times
- Compare observed test statistic (e.g., *F*-ratio) to the permutation distribution and compute proportion of distribution larger than observed



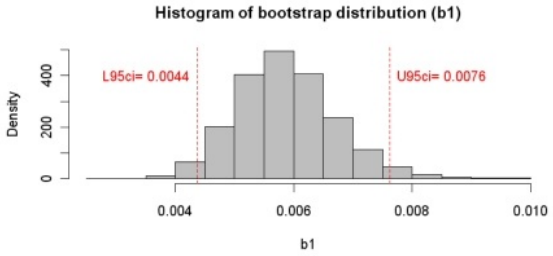
We repeat the process over and over, usually 1,000-10,000 times, and the permutation distribution of the statistic provides an empirical distribution of the test statistic under the null hypothesis of no real structure. We compare the original test statistic to the permutation distribution to compute an exact *p*-value. In this example, let's say that we computed the *F*-ratio as our test statistic. The *F*-ratio is the ratio between two variances, in this case the variance explained by the regression model (numerator) and the error (or unexplained) variance (denominator), and it is an objective measure of how well the model fit the data, since the larger the *F*-ratio the more of the variance in the data is explained by the model. More on the computation of the *F*-ratio below.

Note, the randomization test procedure can be applied to a test statistic, such as the *F*-ratio, or more directly to any model parameter of interest. For the latter, we might compute the random permutation distribution of the slope parameter *b1* and compare our original slope estimate to this distribution to compute an exact *p*-value, as shown here in the figure.

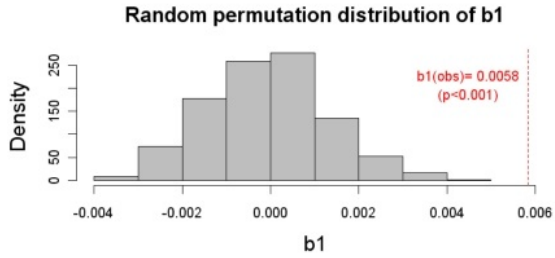
Nonparametric Inference...

Bootstrap versus randomization procedures

- *Bootstrap*... repeated resampling of the original data with replacement to generate the sampling distribution of the test statistic under the alternative hypothesis, used for interval estimation!



- *Randomization*... repeated resampling of the original data after removing real structure via randomization to generate the sampling distribution under the null hypothesis, used for hypothesis testing!



Note the difference between the randomization test procedure and the bootstrap. With the bootstrap, the original data structure is maintained; that is, the individual observation vectors are left intact. We simply draw intact observation vectors randomly (with replacement) to represent the randomness in the sampling process. With the Monte Carlo randomization procedure, the original data structure is destroyed by randomly shuffling some of the data. In this manner, the bootstrap generates a distribution of the test statistic under the *alternative* hypothesis (of real structure), which is used to construct a confidence interval for the statistic, whereas the Monte Carlo randomization procedure generates a distribution of the test statistic under the *null* hypothesis (of no real structure) which can be used directly to compute a *p*-value and conduct a hypothesis test.

Example: Based on the results of the nonparametric randomization test procedure shown here, we can conclude that the probability of observing our data if the null hypothesis of no relationship between x and y were true is less than 0.001. Hence, had we established a decision rule for rejecting the null hypothesis at the $p<0.05$, clearly we would reject the null hypothesis.

Nonparametric Inference...

Hypothesis testing

Parametric F (variance ratio) test:

Variance decomposition:

$$SSY = SSR + SSE$$

$$Total = SSY = \sum_{i=1}^{n} \left(y_i - \bar{y}_i\right)^2$$

$$Error = SSE = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_i\right)^2$$

$$Regression = SSR = SSY - SSE$$

Analysis of variance table

Source	Sums of squares	df	Mean squares	F	p-value
Regression	0.897	1	0.897	44.95	<0.001
Error	0.559	28	0.020		
Total	1.456	29			

Error variance (s^2)

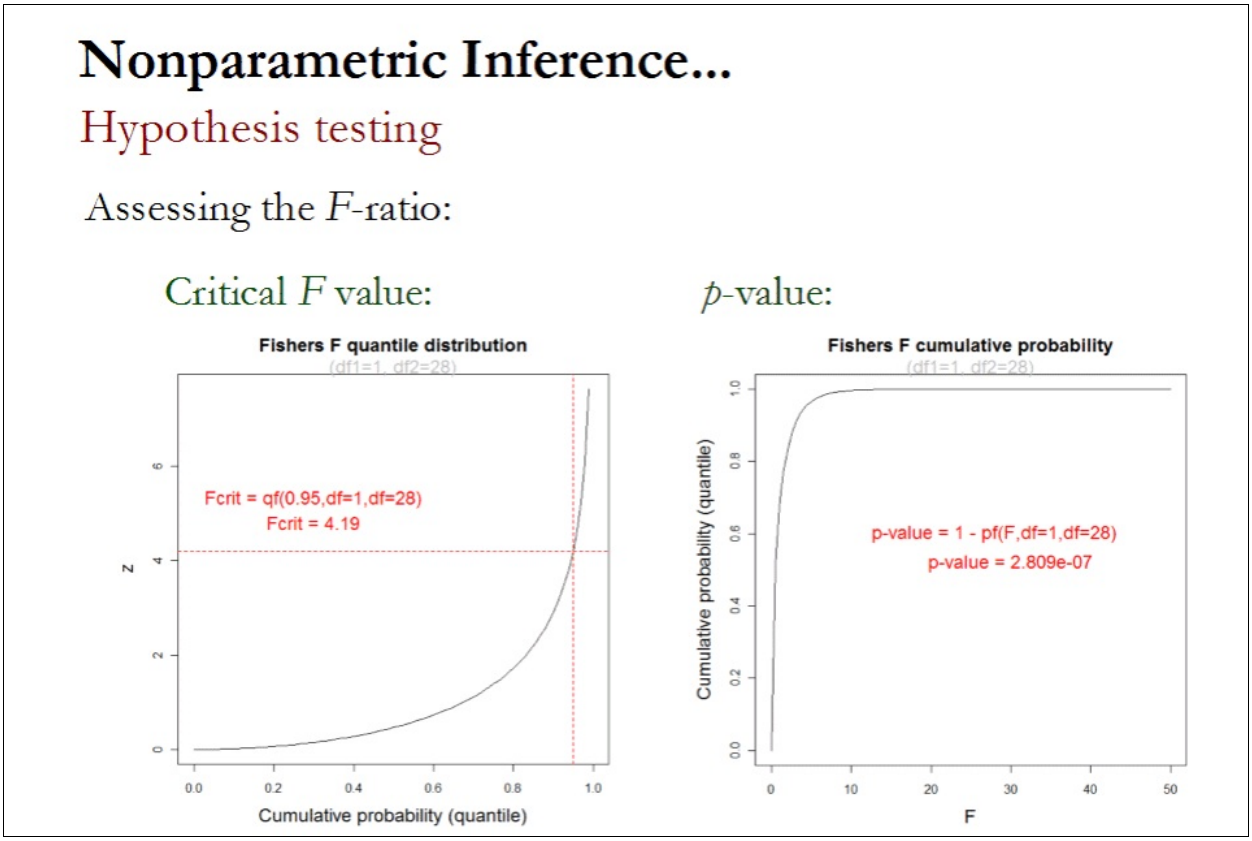
Parametric test of significance:

For the moment, let's assume that we are willing to assume that the errors are independent and identically normally distributed. In other words, let's assume that the errors (or residuals) are normally distributed about the mean (the expected value from the deterministic part of the model) and the variance in the errors is constant over the full range of x – the explanatory variable. Given these assumptions, there are a couple of different ways to test the same null hypothesis. I will describe what is called the 'analysis of variance' approach. The idea is relatively simple: we take the total variation in *y*, SSY, and partition it into components that tell us about the explanatory power of the model. The variation that is explained by the model is called the regression sums of squares, SSR, and the unexplained variation is called the error sums of squares, SSE. Then, SSY = SSR + SSE. We already computed SSY (above). All we need to compute is either SSR or SSE and we have all three quantities by subtraction. SSE is equal to the sums of squares of the deviations of the data points from the fitted model, which is exactly our lack of fit measure (*L*) that we defined above.

Now that we have all of the sums of squares, we need to think about the *degrees of freedom*. Degrees of freedom tell us about how much data (number of observations) we have relative to the number parameters we have. We need to have more observations than parameters at a minimum, and the more observations we have relative to parameters, the better off we should be, right? So we need to adjust our sums of squares to reflect our degrees of freedom. We had to estimate one parameter, the

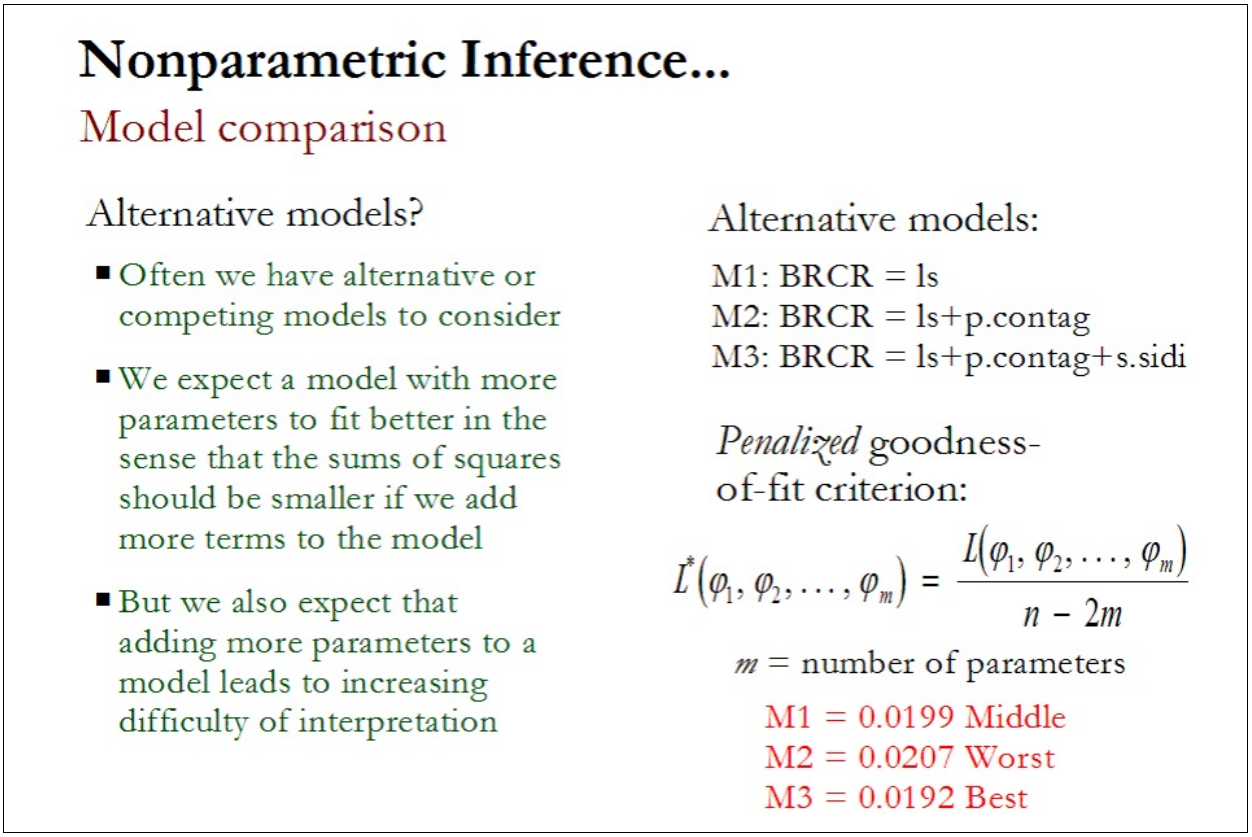
overall mean, $\bar{y}$, before we could calculate SSY, so the total degrees of freedom are n-1. The error sums of squares SSE was calculated after two parameters had been estimated from the data (the intercept and slope), so the degrees of freedom are n-2. Finally, the regression model added just one parameter, the slope b_1, compared with the null model, which has only the intercept b_0, so there is one regression degree of freedom. Note, the regression and error degrees of freedom always add up to the total degrees of freedom; this is always true in any analysis of variance table.

Now that we have the sums of squares and their degrees of freedom, next we need to compute the mean squares simply by dividing the sums of squares by their degrees of freedom. Now we have two variances: the regression variance (regression mean squares) and the error variance (error mean squares). Finally, we can work out the F-ratio, which is a ratio between variances. In this case, we divide the regression variance by the error variance. To test whether the F-ratio is significantly large to reject the null hypothesis, we compare the observed F with the critical value of F, expected by chance alone if the null hypothesis were true.



There are two ways to assess the *F*-ratio. One way is to compare it with the critical value of *F*. For this, we can use the quantiles of the *F* distribution to find the critical value of *F* for a specified quantile or probability of observing an *F* this large or larger. If our observed *F* is larger than this critical value of *F*, we can be confident in rejecting the null hypothesis. The other way, which is much better than working rigidly with a specified uncertainty level (e.g., $p<0.05$), is to ask what is the probability of getting a value of *F* as big as the one we observed or bigger if the null hypothesis is true. For this, we use the cumulative probability distribution. In either case, it is clear in our example that we should reject the null hypothesis as there is almost no chance we would have observed an *F* of 44.9 with 1 numerator degree of freedom and 28 denominator degrees of freedom if the null hypothesis (no slope) were true.

Example: Based on the results of the parametric testing procedure shown here, we can conclude that the probability of observing our data if the null hypothesis of no relationship between x and y were true is less than 0.001. Hence, had we established a decision rule for rejecting the null hypothesis at the $p<0.05$, clearly we would reject the null hypothesis.



6. Model comparison

Up till now, we have assumed a single statistical model and then estimated parameters assuming that we knew or believed this model to be correct. Usually, however, we are not that lucky because we do not know that the model is correct. More often than not, we have alternative models that we would like to consider. Ideally, we would like a method for determining which model has the strongest support in the data. First, we need an objective measure of how well each model fits the data. Fortunately, we already have a goodness(badness)-of-fit measure, the sums of squares. We can confront each of the models with the data and estimate the parameters that give the minimum sums of squares for each model. The model with the minimum sums of squares should be our best model, right? This would be true if each of the models contained the same number of parameters, but is it fair to compare a model with 2 parameters to a model with 3 or 4 parameters? We expect a model with more parameters to fit better in the sense that the sums of squares 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?

Suppose that a model with m parameters has the sums of squares $L(\varphi_1, \varphi_2,...,\varphi_m)$, which will generally decrease as m increases. However, it makes sense to penalize the introduction of additional

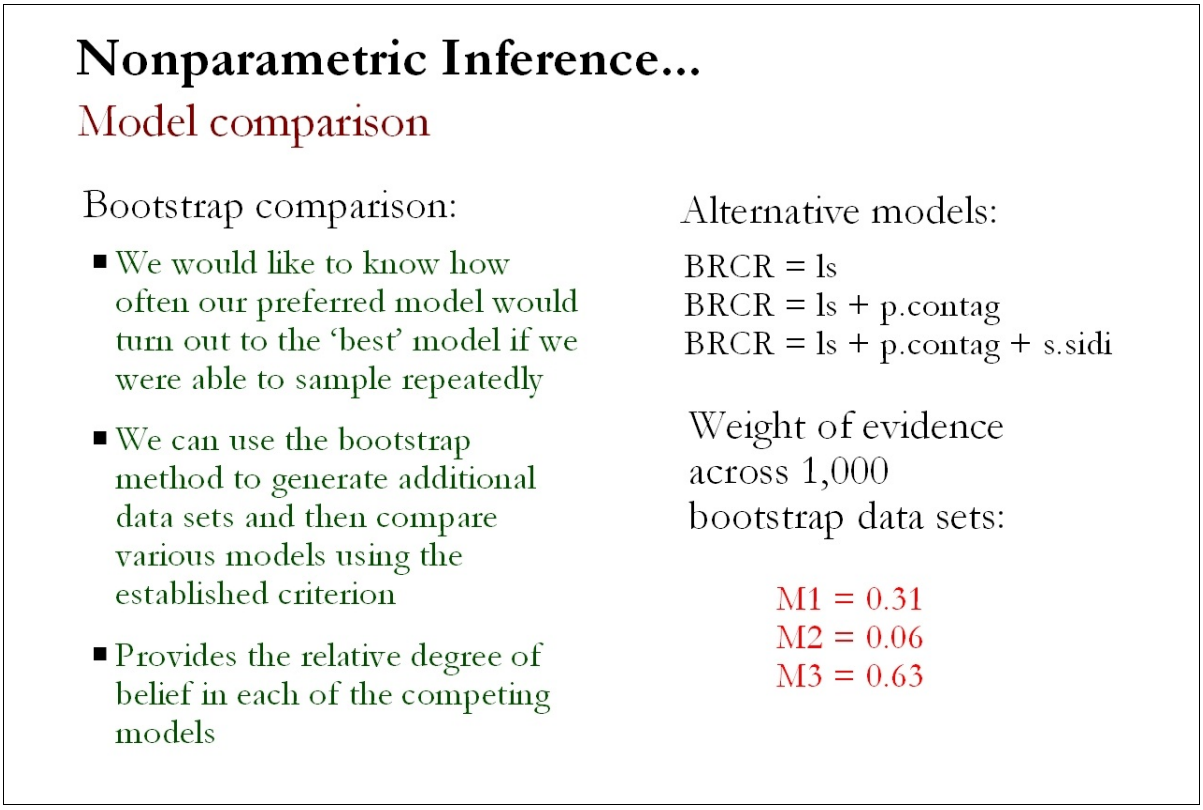
parameters. There are a number of ways in which this can be done. The simplest comparison replaces the sums of squares L by the following:

$$L^*(\varphi_1, \varphi_2, \ldots, \varphi_m) = \frac{L(\varphi_1, \varphi_2, \ldots, \varphi_m)}{n - 2m}$$

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

Model 1: BRCR = ls (%late-successional forest)
Model 2: BRCR = ls + p.contag (patch contagion; a measure of habitat aggregation)
Model 3: BRCR = ls + p.contag + s.sidi (stand Simpson's diversity index; a measure of landscape compositional diversity)

There are a couple of things about these competing models to note. First, they are all simple linear models, but of increasing complexity. Second, they are nested models in terms of the explanatory variables; i.e., model 1 is nested with model 2, which is nested within model 3. Based on the penalized goodness(badness)-of-fit criterion above, the models rank from best to worst as follows: model 3 = 0.0192; model 1 = 0.0199; model 2 = 0.0207. Thus, we can conclude that model 3 is the best model given the data. Importantly, we cannot say that model 3 is truly the best model for this system, only that it is the best among those considered here – there will always be other models that we have not considered.



The model that minimizes $L^*(\varphi_1, \varphi_2,...,\varphi_m)$ should be our preferred model, right? In general, we might ask how the preferred model would act with other data sets. In a sense, we would like to know how often our preferred model would turn out to the best model if we were able to sample the population repeatedly. Once again, the bootstrap method is useful in this context. We can use the bootstrap method to generate additional data sets and then compare various models using the criterion above. This kind of comparison gives us a sense of how confident we should be with the model that wins the competition arbitrated by the actual data set. This kind of comparison brings us closer to a Bayesian viewpoint. Almost all environmental models can be built with differing levels of complexity; it is easy to add additional parameters. The sums of squares provides a way of quantifying the support the data offers for each model. However, when we want to chose a "best" model, then we need a criterion such as the one above. The choice of a best model implies that in some way we reject the others and accept the best one. A Bayesian would, instead, want to assign relative degrees of belief to each of the competing models. The comparison of models with bootstrap data sets lets us mimic the Bayesian approach.

Example: We constructed 1000 data sets via bootstrap with replacement. For each data set we computed the penalized goodness(badness)-of-fit criterion above and declared the model with the smallest value the 'best' model. We computed the proportion of times each model was declared the 'best' model, with the following result: model 1 = 0.310; model 2 = 0.062; model 3 = 0.628. Thus, there is fairly compelling evidence that model 3, the most complex model, is the 'best' model. However, it might be more useful to say that the weight of evidence is 63% in favor of model 3, 31% in favor of model 1 and only 0.6% in favor of model 2.

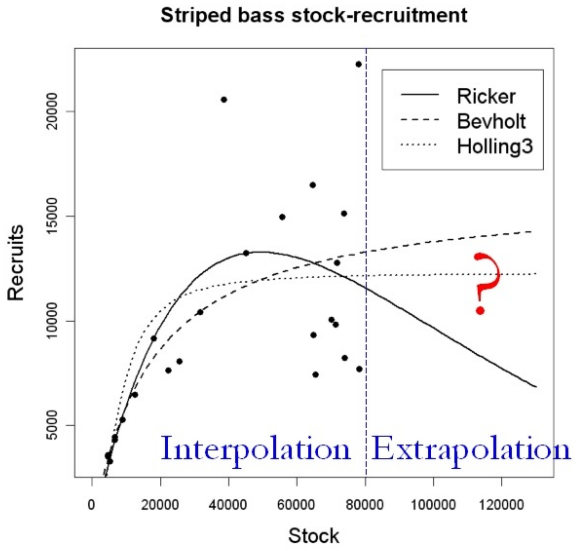
Nonparametric Inference...

Predictions

Kinds of predictions:

- *Interploation*...which is prediction within the measured range of the data
- *Extrapolation*... which is prediction beyond the measured range of the data

Extrapolation requires careful choice of model and thus should be done with extreme caution



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. There are two kinds of prediction, and these are subject to very different levels of uncertainty:

- *Interpolation* – which is prediction within the measured range of the data.
- *Extrapolation* – which is prediction beyond the measured range of the data.

Extrapolation is far more problematical than interpolation, and model choice is a major issue. Choice of the wrong model can lead to wildly different predictions beyond the measured range of the data, even if the predictions within the range are similar among different models. Thus, extrapolation should be done with extreme caution. For either purpose, we can use the fitted deterministic model to predict point estimates (i.e., expected values) of new observations, but without a formal model for the stochastic component of the model, it is difficult to provide confidence (prediction) limits on those predictions. That is to say, we can easily give a point estimate for the prediction – which is the expected value from the deterministic model, but it is much more difficult to give an interval estimate for the prediction.

Nonparametric Inference...

Predictions

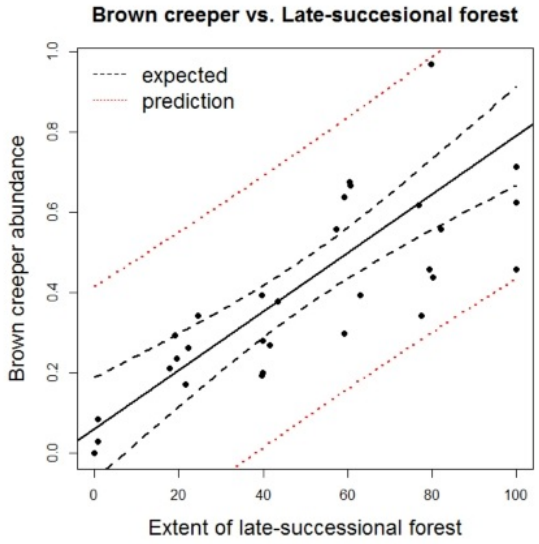
Parametric predictions:

- *Point estimates*... apply the fitted deterministic model to new values of x
- *Interval estimates*... Calculate the standard error for a predicted value and construct as before

s^2 = Error variance

$$se_{\hat{y}} = \sqrt{s^2\left[1 + \frac{1}{n} + \frac{\left(x - \bar{x}\right)^2}{SSX}\right]}$$

$$95\%PI = \hat{y} \pm t_{0.025,\, n-2} se_{\hat{y}}$$



Parametric prediction interval:

For the moment, let's assume that we are willing to assume that the errors are independent and identically normally distributed as before. Given these assumptions, a parametric prediction interval can be constructed for predicted values. The idea is relatively simple: we calculate the standard error for a predicted value (not the parameter estimate, as before) and multiply it by the appropriate value of Student's *t* and then subtract this interval from, and add it to, the predicted value (i.e., the predicted point estimate or expected value) to get the corresponding prediction interval. Note, the standard error for a predicted value in regression is computed a little differently than the standard error for a parameter estimate. Specifically, the standard error for a predicted value increases with the square of the difference between mean x and the value of x at which the prediction is made – reflecting increasing uncertainty as we get farther from the mean of the data. As with the standard error of the slope parameter estimate, the wider the range of x values (as measured by SSX) and the bigger the sample size, *n*, the lower the uncertainty:

$$se_{\hat{y}} = \sqrt{s^2\left[1 + \frac{1}{n} + \frac{\left(x - \bar{x}\right)^2}{SSX}\right]}$$

The prediction interval for each predicted value is obtained by subtracting from, and adding to, each predicted value an interval which is the standard error times Student's *t* with the appropriate error degrees of freedom. In this case, the appropriate value of *t* is given by the 0.975 quantile of the *t* distribution with *n*-2 degrees of freedom (= 2.048). Thus, the 95% prediction interval for *y* given the value of *x* at which the prediction is made is:

$$\hat{y} \pm t_{0.025,n-2} se_{\hat{y}}$$

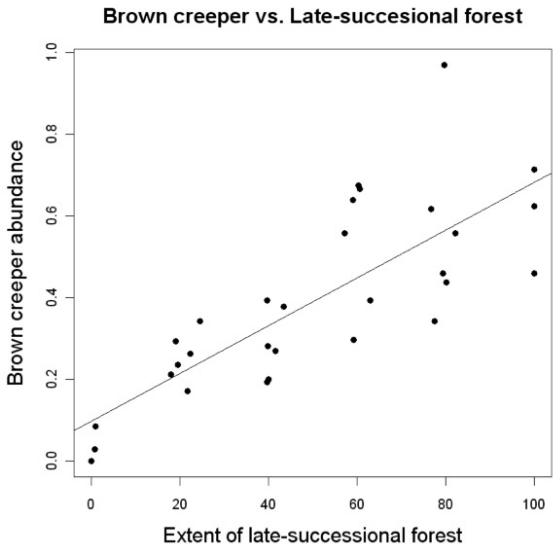
Example: For the example data set, we can compute the 95% prediction interval for a range of x values and plot the result as a prediction band on the original scatter plot, as shown.

Nonparametric Inference...

Predictions

Nonparametric predictions:

- *Point estimates...* apply the fitted deterministic model to new values of x
- *Interval estimates...* much more difficult to do; requires complex bootstrap procedure?



Nonparametric bootstrap prediction interval:

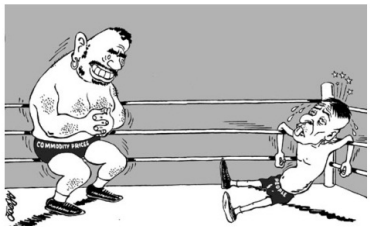
Not surprisingly, nonparametric approaches employing procedures such as the bootstrap have been developed for computing prediction intervals for cases where we can make no assumptions about the shape of the error distribution. Unfortunately, these approaches are poorly described and rather complex. Therefore, we will not attempt to delve into the methods for nonparametric prediction. Suffice it to say that it is simple to predict point estimates from a nonparametric model – simply apply the fitted deterministic model, but to provide an estimate of uncertainty on those predictions requires a sophisticated approach that goes beyond the scope of this treatment.

Nonparametric Inference...

Considerations

- Requires minimal assumptions about the error
- Weaker statistical inference as a result of above
- Reliance on data resampling to solve many of the inference problems

$$y_i = b_0 + b_1 x_i + e_i$$



Population → Data → Sample1, Sample2, Sample3

8. Considerations with nonparametric inference

The nonparametric framework for statistical inference can be very useful, but it has some serious drawbacks that proponents of parametric inference are quick to point out.

1. *Requires minimal assumptions about the error*... The major appeal of nonparametric inference is that it requires fewer assumptions about the model than parametric approaches. In particular, nonparametric inference does not require a particular probability distribution to be specified for the stochastic component of the model.

2. *Weak statistical inference*... As a consequence of number one above, in general the inferences from a nonparametric procedure are weaker than from a parametric procedure. This is because the statistical model is a less complete description of the underlying population. And the less that is known or assumed about the underlying population, the less one can say about it from any sample data set. Without specifying the error model, it is more difficult to conceive of the statistical model as a data-generating mechanism and more difficult to simulate new data.

3. *Reliance on data resampling methods*... As a consequence of number one above, nonparametric approaches typically invoke data resampling methods such as the bootstrap and Monte Carlo randomization to solve many of the problems of statistical inference. These procedures are computationally intensive methods that can often serve as effective substitutes for the corresponding parametric method. Interestingly, these resampling methods come as close as we can to imitating the frequentist approach to statistical inference, which is based on the idea of hypothetical repeated sampling. The bootstrap, for example, does just that – repeatedly draws samples by resampling the original data. The major difference is that in the true frequentist approach, we would resample the population not the original data set.