ECO 602 Analysis of Environmental Data

FALL 2019 – UNIVERSITY OF MASSACHUSETTS

DR. MICHAEL NELSON

Today's Agenda

- More nonparametric (and parametric) OLS inference
- Resampling methods:
  - Bootstrapping
  - Monte Carlo sampling
- ANOVA table partitioning variance
- Wilcoxon signed rank test
- Predictions

Today’s Agenda

- Confidence interval visualization
- More nonparametric (and parametric) OLS inference
- Resampling
- T tests and Wilcoxon signed-rank tests
- Maximum Likelihood preview

Visualizing confidence intervals

https://seeing-theory.brown.edu/frequentist-inference/index.html#section2

Today's Agenda

Confidence interval visualizations

Seeing theory

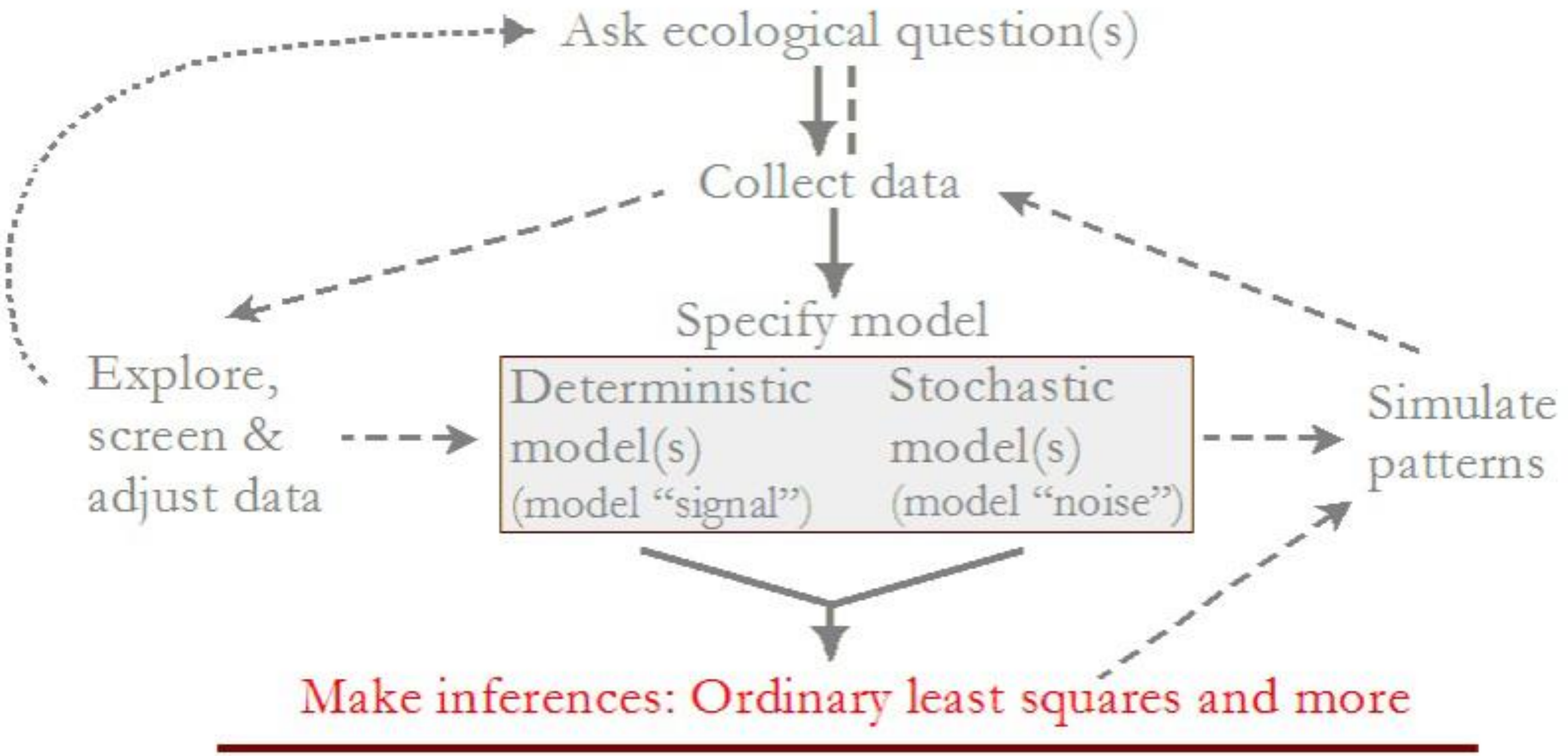
Model para m cCIs

Hypotheses of slope vs intercept

Nonparametric Inference

- Much of the mathematical hardware is similar to parametric inference.
- Main difference: no attempt to guess a theoretical distribution for the population.
- Main consequence: weaker inference
- ‘Nonparametric’ refers to the lack of an explicit stochastic model for the population.
- We usually calculate statistics in nonparametric inference!

Nonparametric Inference



- *Nonparametric inference* involves confronting the model with data to estimate parameters, test hypotheses, compare alternative models, or (with difficulty) make predictions, without specifying a probability distribution

Landscape of Statistical Methods…

The basic statistical model:

Y = deterministic part + stochastic part

- Univariate
- Multivariate

- Linear
- Nonlinear
- Smoothed

- Distribution
- Heterogeneity
- Autocorrelation
- Multiple levels
- Random noise

Nonparametric Inference…

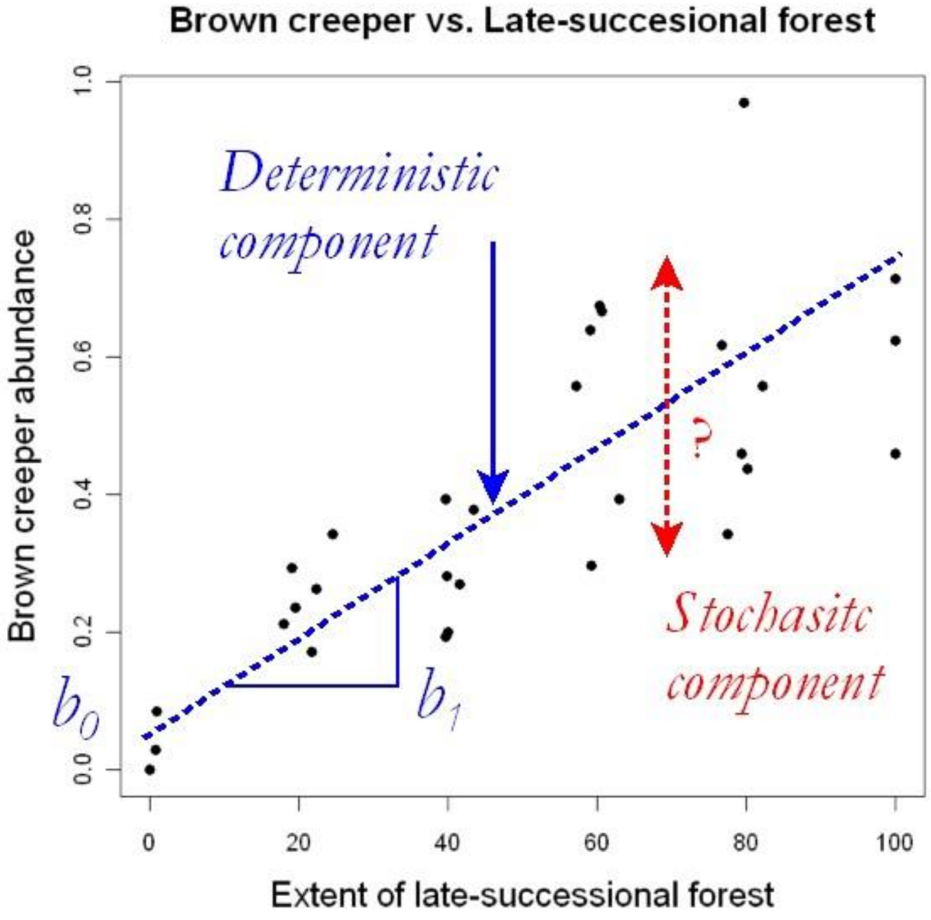
What is the statistical model?

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

Statistical Model:

$$y_i = \boxed{b_0} + \boxed{b_1} x_i + e_i$$

Parameters Unspecified



Nonparametric Inference...

Estimate model parameters: OLS method

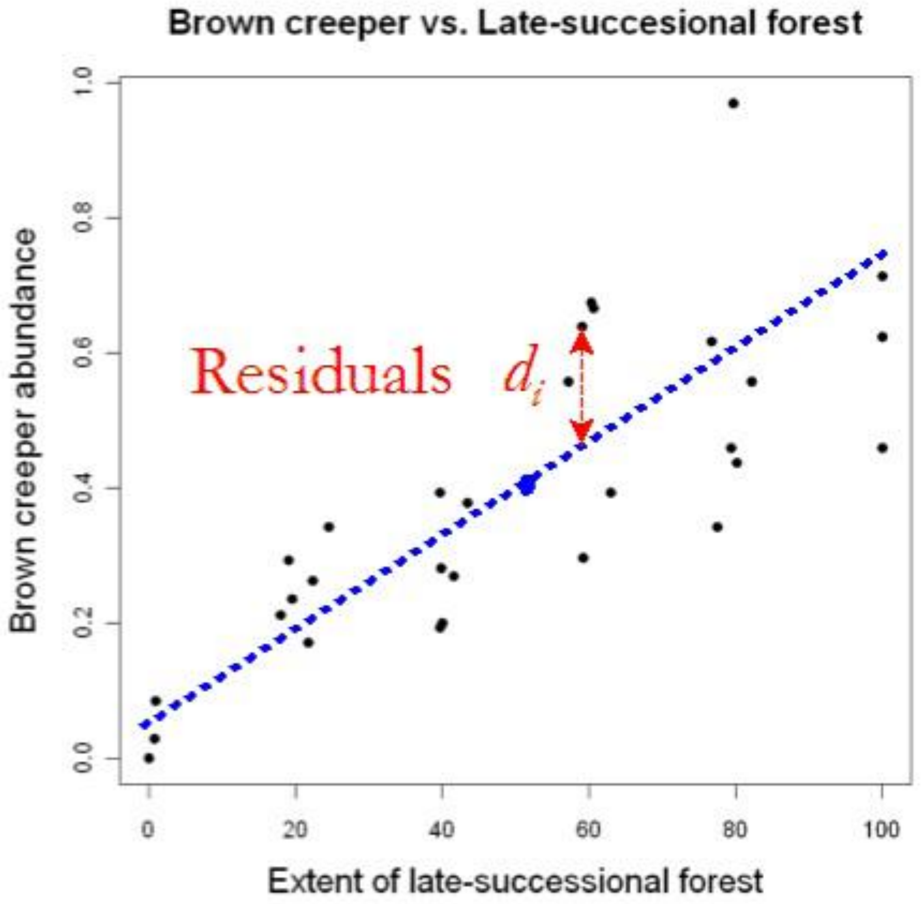
1. Define measure of (lack of) fit:

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

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

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

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



How wo we minimize our lack-of-fit function?

We might require that the best fit line passes through the point at the mean of x and mean of y.

We could try to guess values of slope and intercept parameters that minimize lack-of-fit.

But…

If we do this, we've constrained our possibilities

How wo we minimize our lack-of-fit function?

We might **require** that the best fit line passes through the point at the mean of x and mean of y.

If we do this, we've constrained our possibilities!

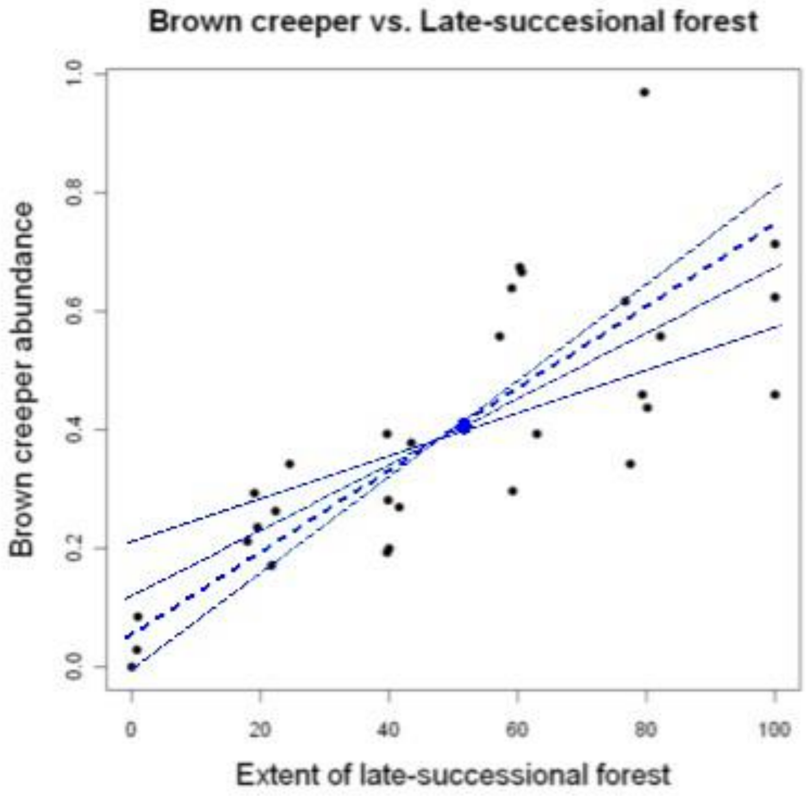
How many parameters are actually free with this requirement?

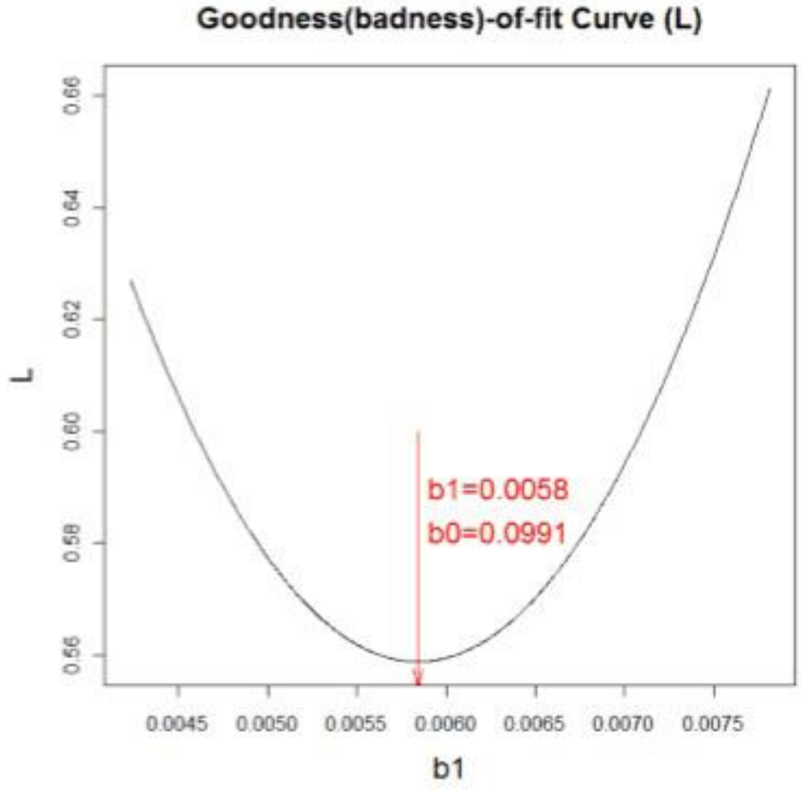
Nonparametric Inference...

Estimate model parameters: OLS method

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

- Numerical solution





Seems reasonable, but what could go wrong?

How certain are we about the mean values of the predictor and response?

In the sample?

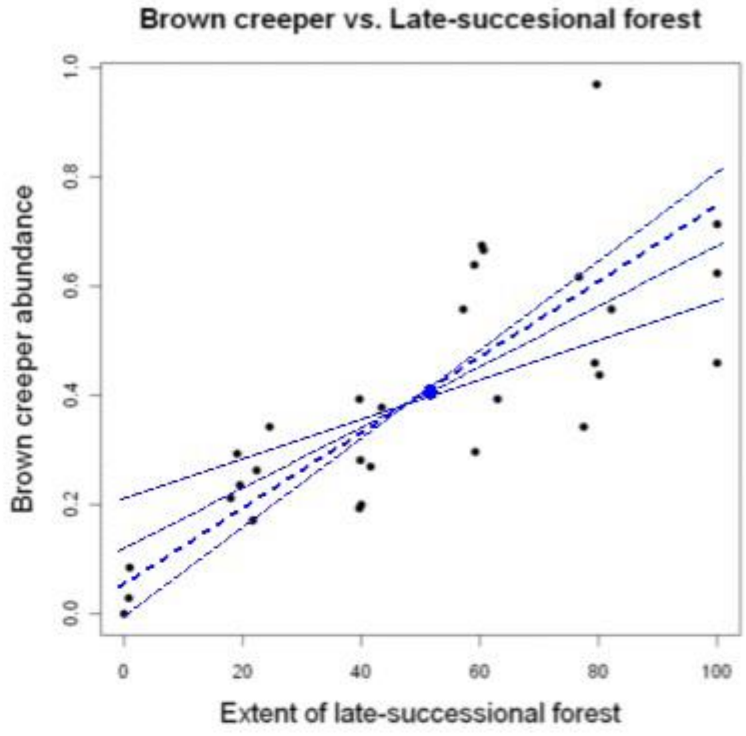
In the population?

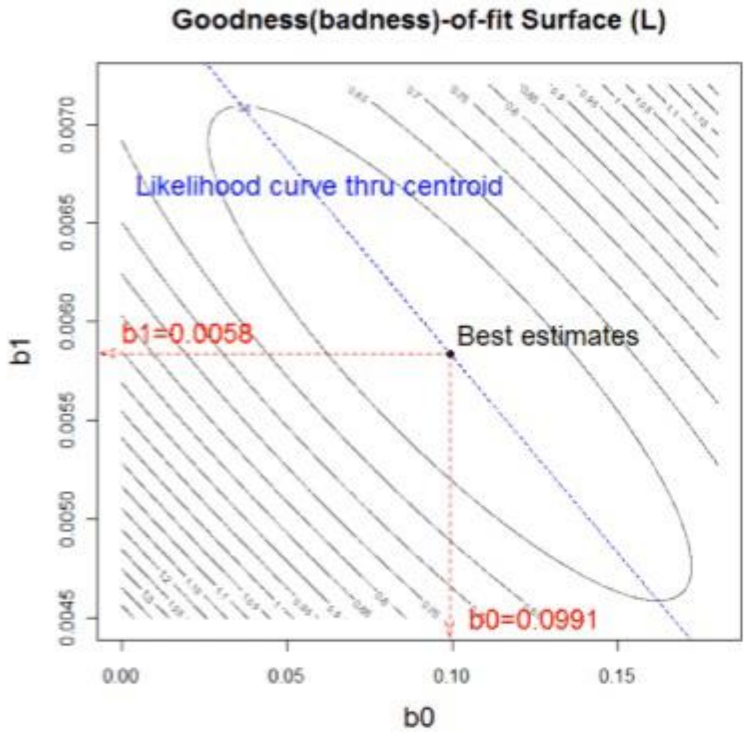
Nonparametric Inference...

Estimate model parameters: OLS method

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

- Numerical solution





What if we didn’t require passage through the centroid?

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:

$b_1 = 0.0058$
$b_0 = 0.0991$

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

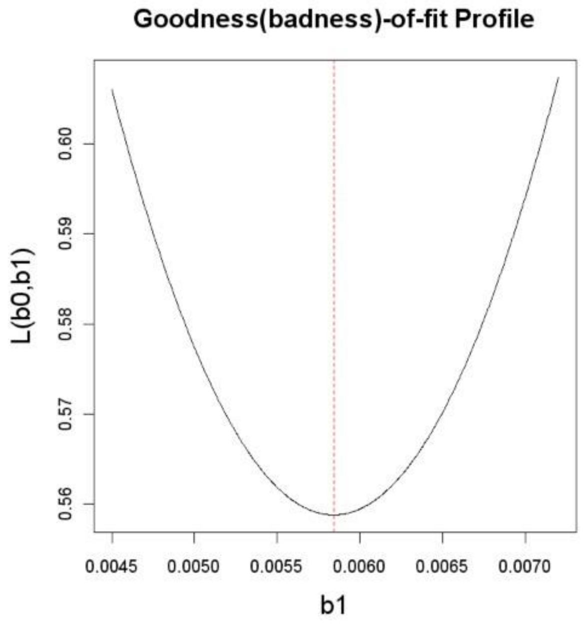
Nonparametric confidence intervals

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"



What kind of confidence interval have we considered?

Remember: Standard Errors can be calculated for any statistic (parameter estimate).

What about a CI for a slope or intercept?

Resampling Methods

Create new samples from our existing sample.

It sounds like cheating, but….

Remember our random sampling scheme?

- Nonparametric inference can't help us if we use a poor sampling design.

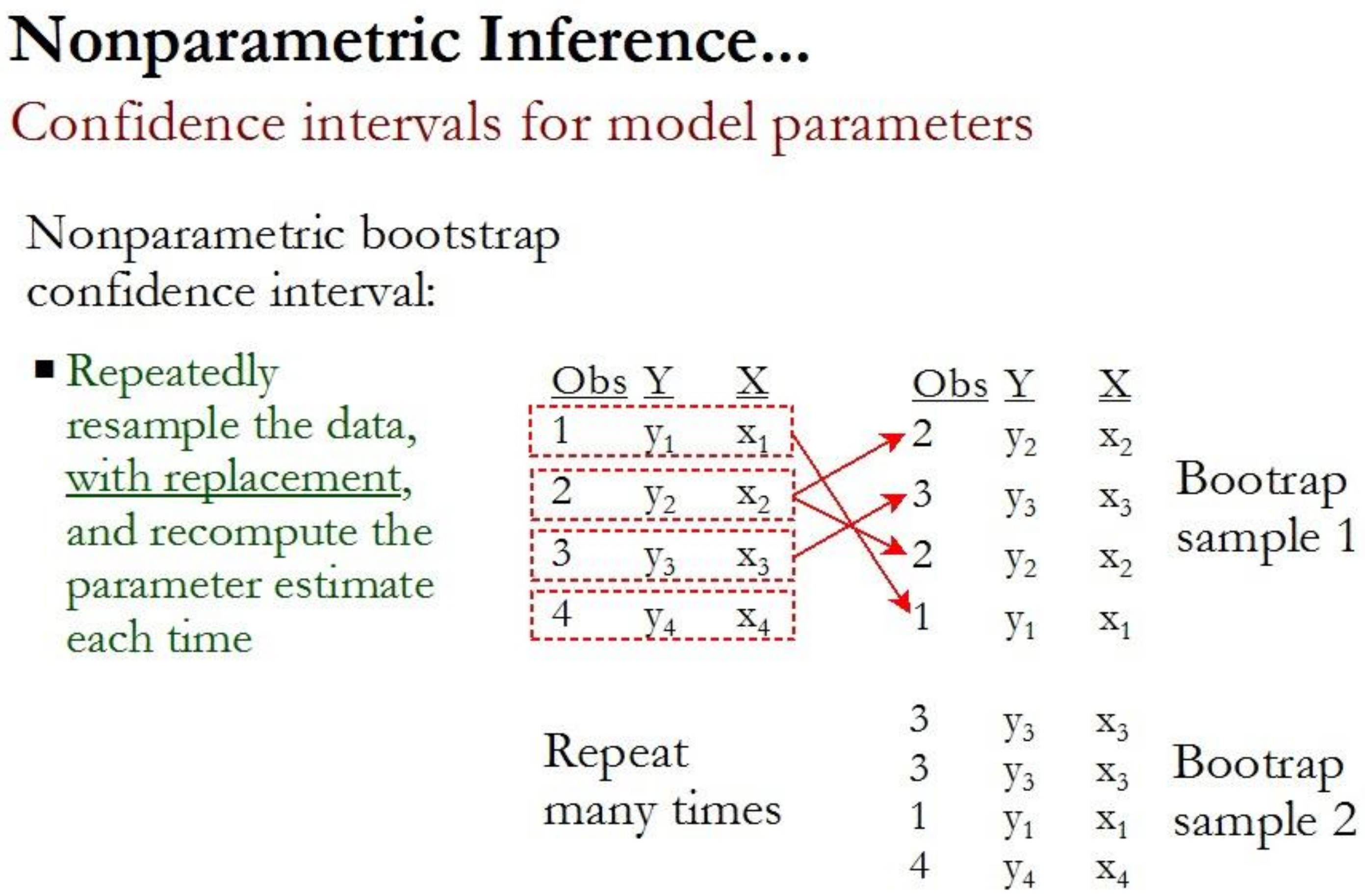
Bootstrapping and Monte Carlo methods

What to resample?

What does resampling mean?

Should we use replacement?

Nonparametric CI: Bootstrapping



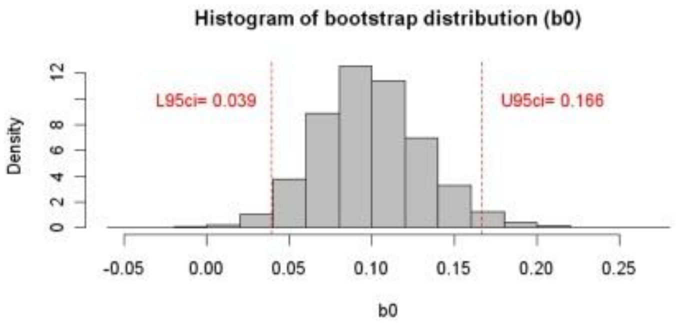
(Re)sampl**ing** distributions of estimates

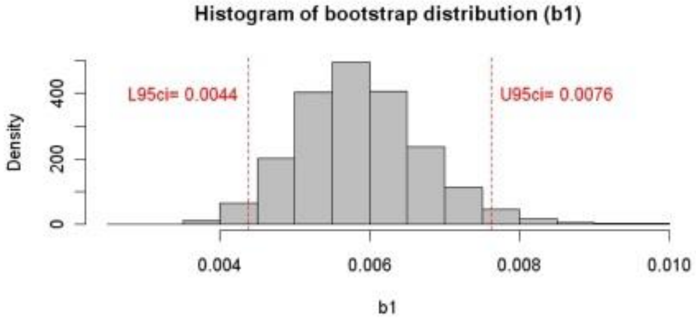
Nonparametric Inference...

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





Parametric slope/intercept CIs

If we use parametric inference, we can often* find a closed-form solution for parameter estimates and standard errors/confidence intervals!

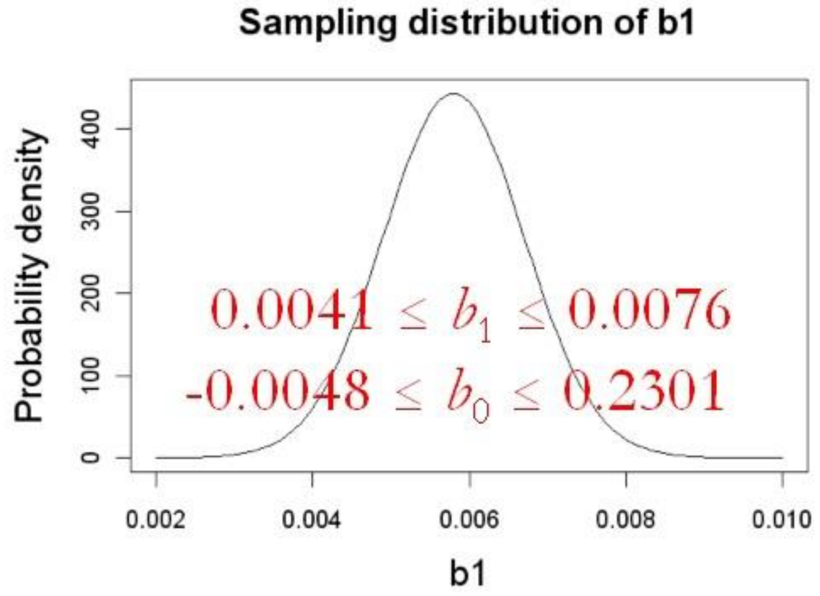
But we often cannot find analytical solutions for the models we actually want to use!

~~No~~nparametric 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



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

$$95\%CI = b_1 \pm t_{0.025,n-2}\, se_{b1}$$

s^2 = Error variance

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

$$95\%CI = b_0 \pm t_{0.025,n-2}\, se_{b0}$$

Nonparametric Inference...

Hypothesis testing

Null hypothesis:

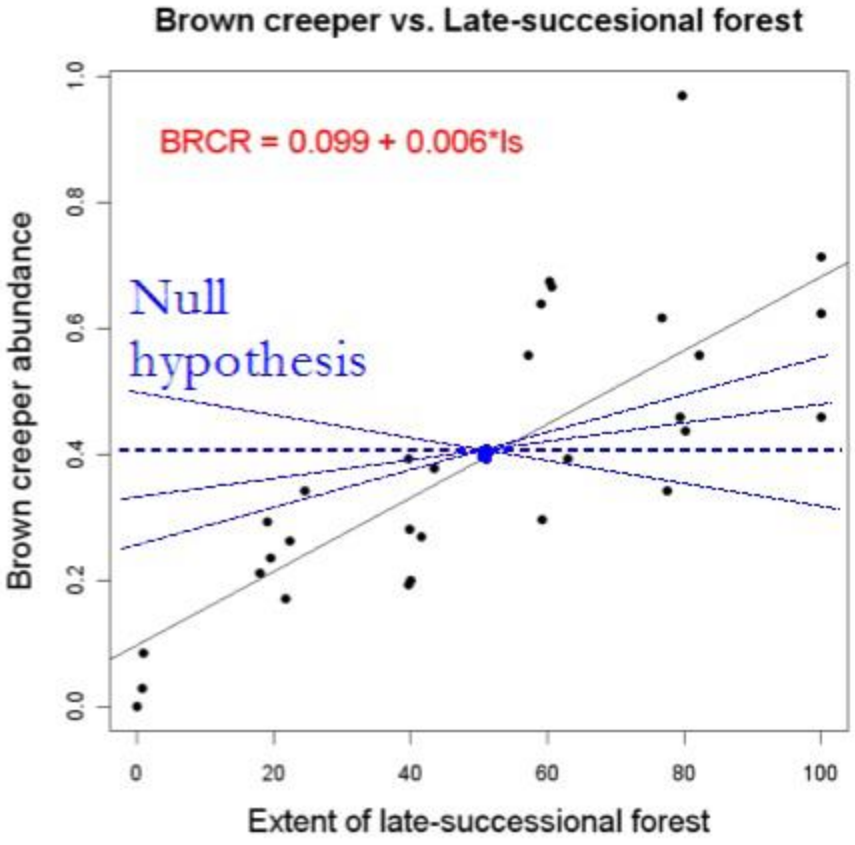
- The slope of the regression line is zero; i.e., no dependence of *y* on x

p-value:

- The probability of observing the observed slope or something more extreme (an even steeper slope) (under hypothetical repeated sampling) if the null hypothesis were true

Decision rule:

- If $p<0.05$ (Type I error rate), reject null hypothesis



Resampling x and y

Monte Carlo resampling

Sample predictor/response variables separately.

Readings say "remove structure"

Null hypothesis: MC resampling

Alternative hypothesis: Bootstrapping

Nonparametric Inference…

Hypothesis testing

Randomization test:

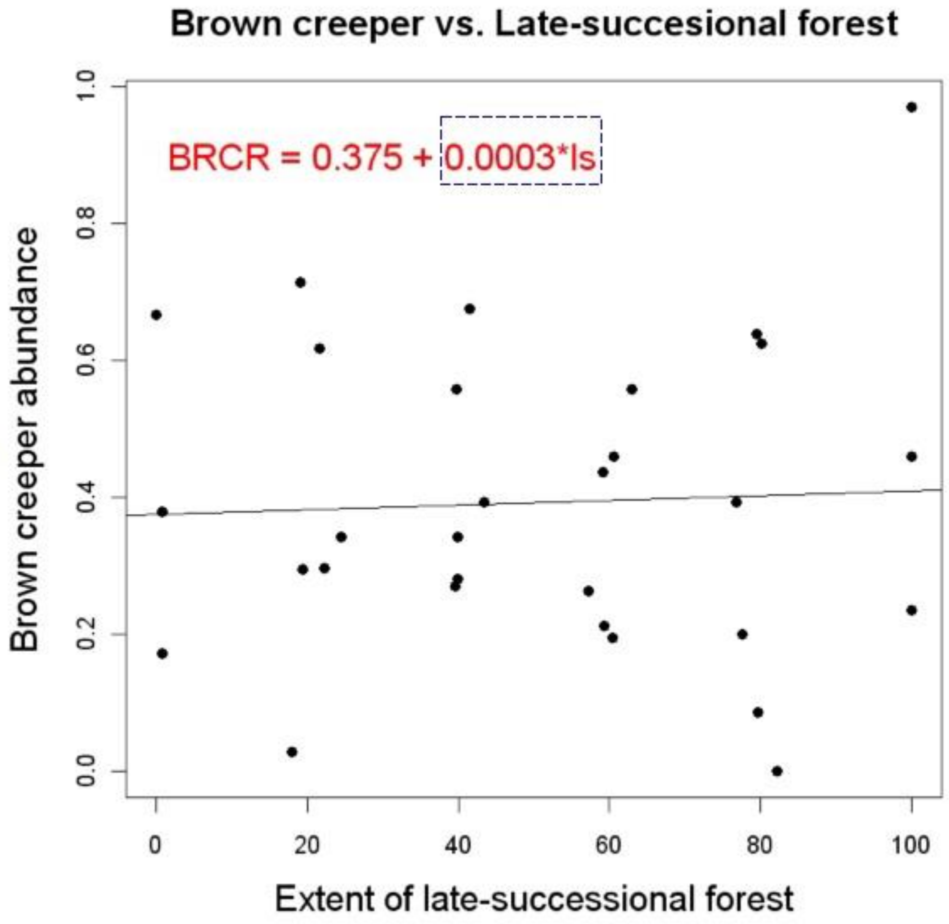
Original data

	y	x
[1,]	0.03	0.71
[2,]	0.44	80.15
[3,]	0.38	43.44
[4,]	0.56	57.23
[5,]	0.34	24.46
[6,]	0.28	39.89
[7,]	0.68	60.36
[8,]	0.24	19.45
[9,]	0.56	82.16
[10,]	0.62	100.00
[11,]	0.00	0.00
[12,]	0.21	17.91
[13,]	0.64	59.10
[14,]	0.97	79.70
[15,]	0.71	100.00
[16,]	0.46	79.46
[17,]	0.29	19.07
[18,]	0.27	41.57
[19,]	0.67	60.63
[20,]	0.39	39.60
[21,]	0.26	22.31
[22,]	0.09	0.82
[23,]	0.20	39.94
[24,]	0.19	39.73
[25,]	0.39	62.95
[26,]	0.17	21.61
[27,]	0.34	77.51
[28,]	0.62	76.79
[29,]	0.46	100.00
[30,]	0.30	59.25

Permuted data

	y.perm	x
[1,]	0.38	0.71
[2,]	0.62	80.15
[3,]	0.39	43.44
[4,]	0.26	57.23
[5,]	0.34	24.46
[6,]	0.28	39.89
[7,]	0.19	60.36
[8,]	0.29	19.45
[9,]	0.00	82.16
[10,]	0.24	100.00
[11,]	0.67	0.00
[12,]	0.03	17.91
[13,]	0.44	59.10
[14,]	0.09	79.70
[15,]	0.97	100.00
[16,]	0.64	79.46
[17,]	0.71	19.07
[18,]	0.68	41.57
[19,]	0.46	60.63
[20,]	0.27	39.60
[21,]	0.30	22.31
[22,]	0.17	0.82
[23,]	0.34	39.94
[24,]	0.56	39.73
[25,]	0.56	62.95
[26,]	0.62	21.61
[27,]	0.20	77.51
[28,]	0.39	76.79
[29,]	0.46	100.00
[30,]	0.21	59.25

Fit model



Nonparametric Inference...

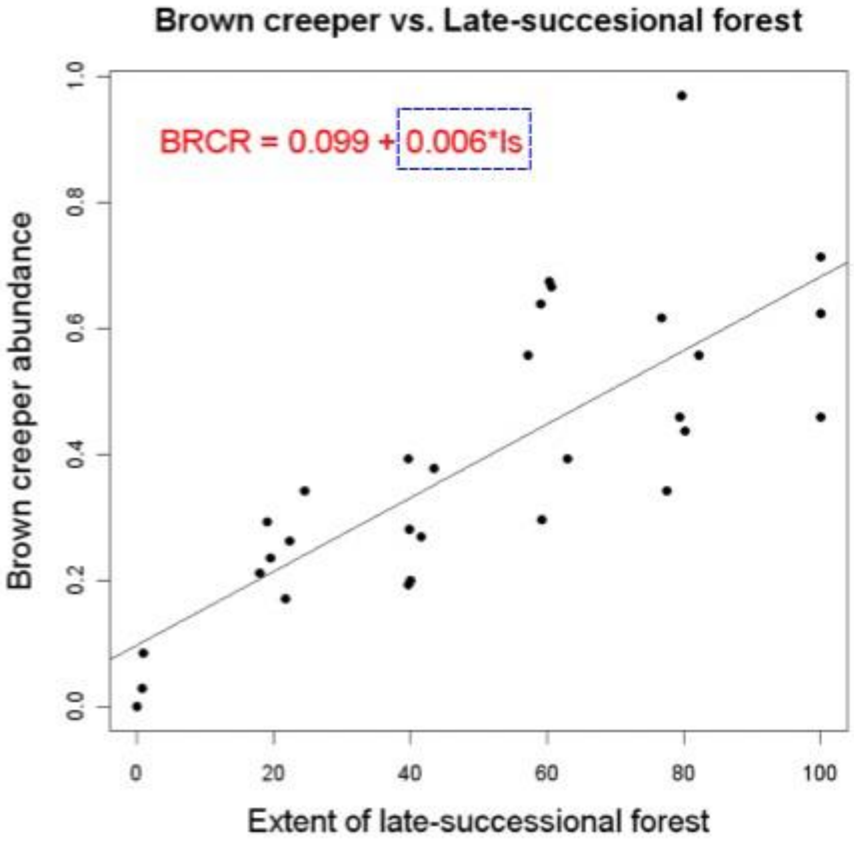
Hypothesis testing

Randomization test:

Original data

```
         y      x
 [1,] 0.03   0.71
 [2,] 0.44  80.15
 [3,] 0.38  43.44
 [4,] 0.56  57.23
 [5,] 0.34  24.46
 [6,] 0.28  39.89
 [7,] 0.68  60.36
 [8,] 0.24  19.45
 [9,] 0.56  82.16
[10,] 0.62 100.00
[11,] 0.00   0.00
[12,] 0.21  17.91
[13,] 0.64  59.10
[14,] 0.97  79.70
[15,] 0.71 100.00
[16,] 0.46  79.46
[17,] 0.29  19.07
[18,] 0.27  41.57
[19,] 0.67  60.63
[20,] 0.39  39.60
[21,] 0.26  22.31
[22,] 0.09   0.82
[23,] 0.20  39.94
[24,] 0.19  39.73
[25,] 0.39  62.95
[26,] 0.17  21.61
[27,] 0.34  77.51
[28,] 0.62  76.79
[29,] 0.46 100.00
[30,] 0.30  59.25
```

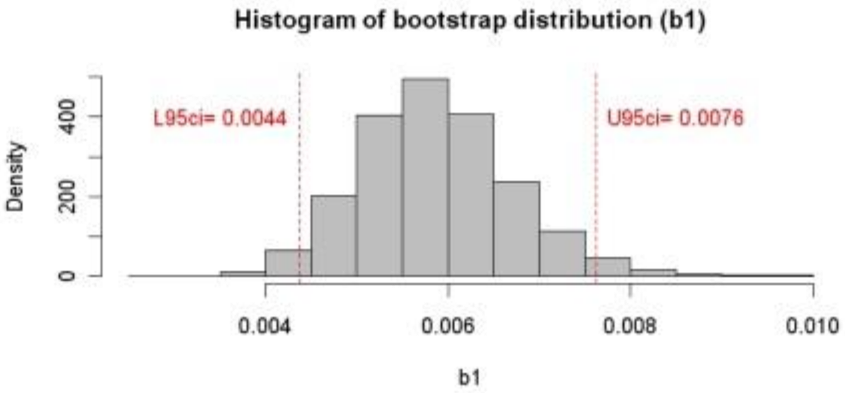
Fit model



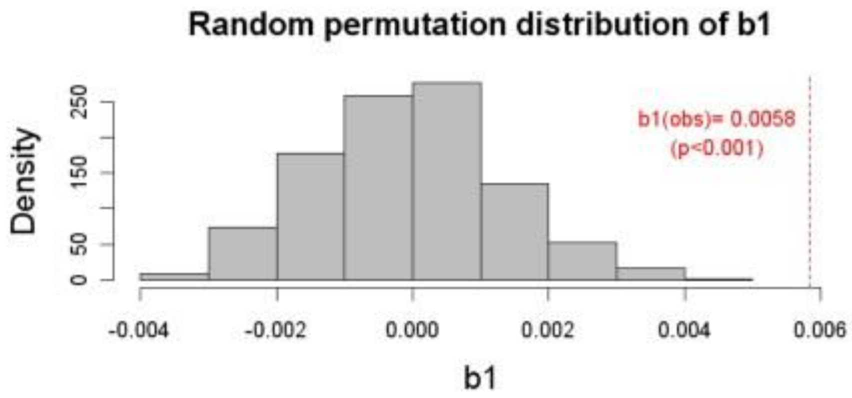
Nonparametric Inference...

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!



Bootstrap visualization

https://seeing-theory.brown.edu/frequentist-inference/index.html#section2

Prediction is harder for nonparametric!

Parametric is easy.

Nonparametric requires resampling or simulation methods.

Nonparametric Inference...

Predictions

Parametric predictions:

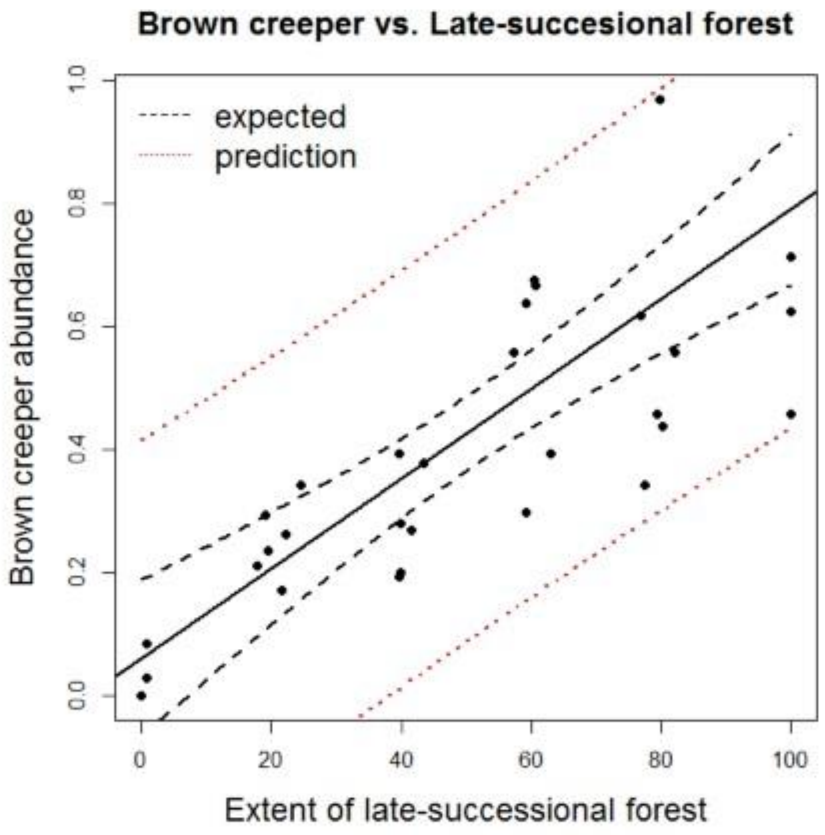
- *Point estimates*... apply the fitted deterministic model to new values of x

- *Interval estimates*... Calculate the standard error for a predicted value and construct as before

s^2 = Error variance

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

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

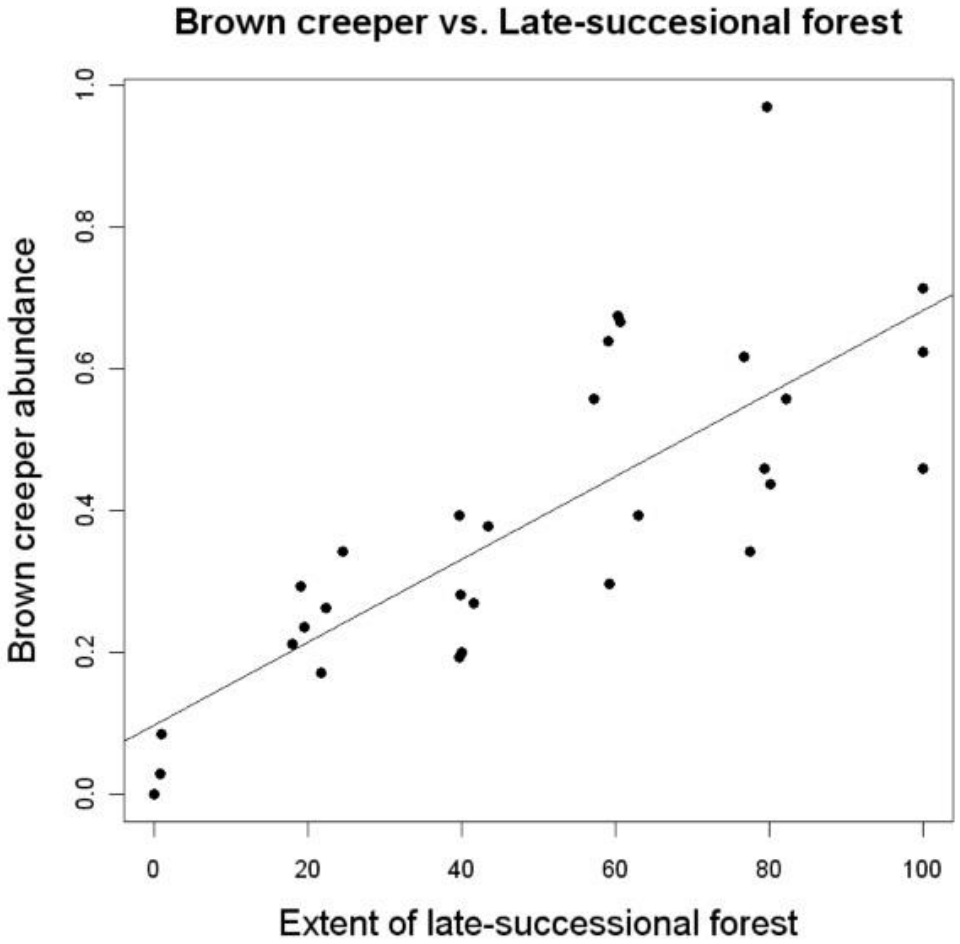


Nonparametric Inference...

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?



Simpler nonparametric test: Wilcoxon signed rank-sum test

Kind of like a nonparametric analogue of a paired 2-sample t-test

Test two populations, often repeated measurements, for significant difference.

Useful when assumptions of independence, normality, etc are invalid.

Common examples:

- Grades for same students in 3rd and 4th graders.
- Drug efficacy, measured before and after

Wilcoxon rank-sum

For each paired observation:

-1 if 2nd observation smaller, otherwise 1

What would we expect if there was no difference?

Wilcoxon rank-sum

Null hypothesis is that sums are the same in the before and after populations.

Likelihood

Shared element between frequentist and Boolean paradigms.

Likelihood: Likelihood function

What combination of parameters make our observed data most likely?

Frequentist Parametric Inference...

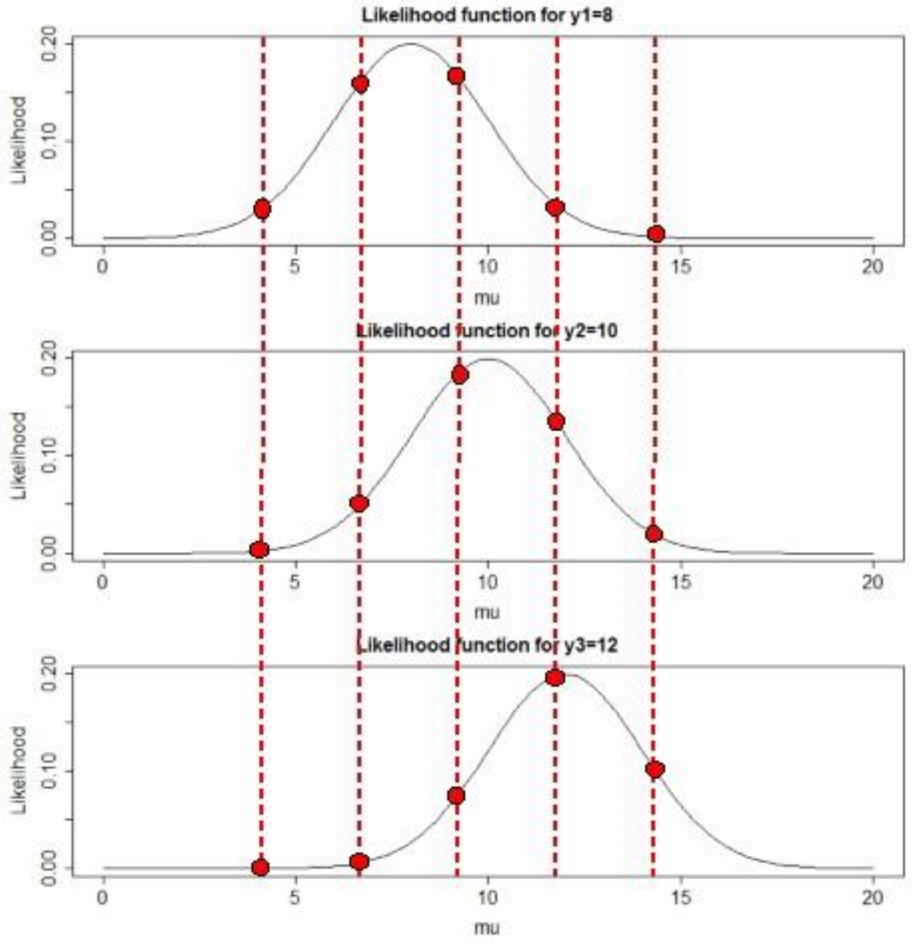
Estimate model parameters: MLE method

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

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

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

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



Likelihood: Likelihood function

OLS likelihood function is pretty easy.

Most likelihood functions are not!

For next time:

Finish Likelihood chapter (McGarigal ch. 9)

Start Bayesian (McGarigal ch. 10)