Data transformations

Analysis of Ecological and Environmental Data

QERM 514

Mark Scheuerell

17 April 2020

Goals for today

- Identify possible transformations of the response when your errors have unequal variance or are skewed
- Understand how to use common transformations and make inference from the resulting model
- Understand that there are alternatives to transformation that we will use later

Why would you transform?

We have made a number of assumptions about our models, which include

- the distribution of the errors (IID)
- linear relationship(s) between the response and predictor(s)

What can we do when these assumptions are not met?

What can you transform?

It's possible to transform both sides of our models to

- achieve constant variance (y)
- correct for skewness (y)
- linearize the relationship (y, x)

Types of transformations

The most common form is where $y' = y^{\lambda}$

and $\lambda > 1$ (powers)

or $0 < \lambda < 1$ (roots)

For example

$\lambda = 2 \Rightarrow y' = y^2$

$\lambda = \frac{1}{2} \Rightarrow y' = \sqrt{y}$

Types of transformations

One can also use inverses where $y' = y^{-\lambda}$

and $\lambda > 1$ (powers)

or $0 < \lambda < 1$ (roots)

For example

$\lambda = 2 \Rightarrow y' = \frac{1}{y^2}$

$\lambda = \frac{1}{2} \Rightarrow y' = \frac{1}{\sqrt{y}}$

Box-Cox transformation

The Box-Cox transformation is a popular method for stabilizing the variance of errors

It is defined as

$$y' = \frac{y^\lambda - 1}{\lambda}$$

for all $y > 0$

Box-Cox transformation

More specifically, because

$$\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} = \log(y)$$

we instead use

$$y' = \begin{cases} \frac{y^\lambda - 1}{\lambda} \text{ if } \lambda \neq 0 \\ \log(y) \text{ if } \lambda = 0 \end{cases}$$

Box-Cox transformation

How does one choose λ?

By using *profile likelihoods* (which we will see in a later lecture)

(We'll use the `boxcox()` function in the **MASS** package)

Box-Cox transformation

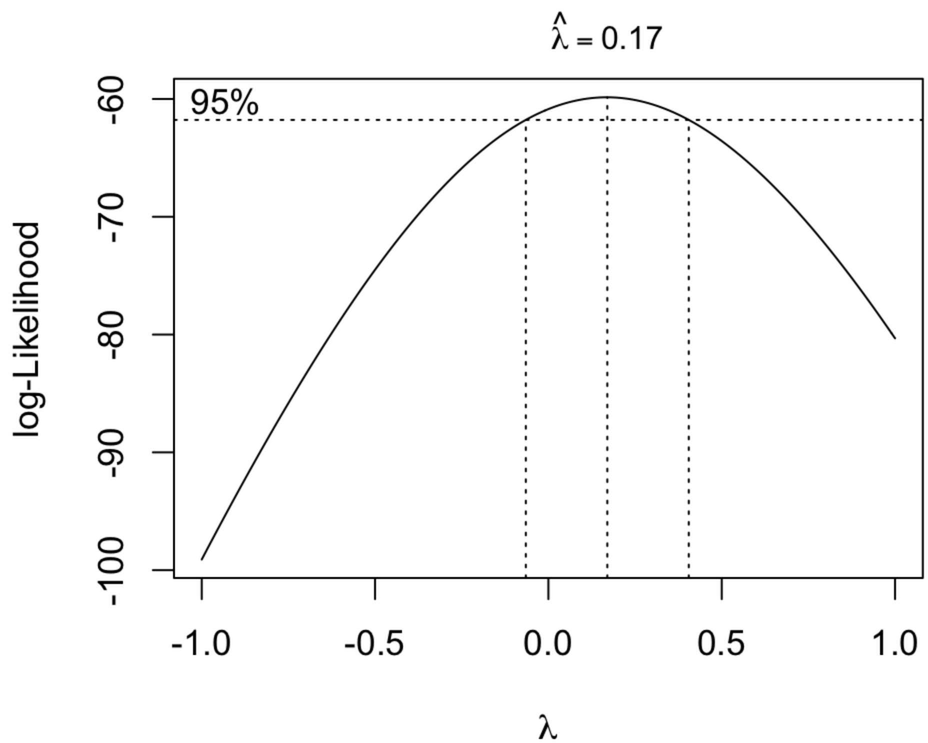
An example

Let’s return to the plant data from the Galapagos Archipelago where we modeled diversity as a function of island area

```
## get data
data(gala, package = "faraway")
## fit regression model
mm <- lm(Species ~ Area, gala)
## estimate lambda
MASS::boxcox(mm)
```

Box-Cox transformation

Here is the result of calling `boxcox(mm)`



Box-Cox transformation

After transformation, how do we interpret $\lambda = 0.17$?

Box-Cox transformations work well, but sometimes we can do better with an approximation to λ

Box-Cox transformation

General considerations

- The Box–Cox method gets upset by outliers

For example, if $\hat{\lambda} = 5$ there is little rationale for such an extreme transformation

Box-Cox transformation

General considerations

- The Box–Cox method gets upset by outliers
- If some $y_i < 0$, we can add a constant to all the y

This works if the constant is small, but it’s a “hack”

Box-Cox transformation

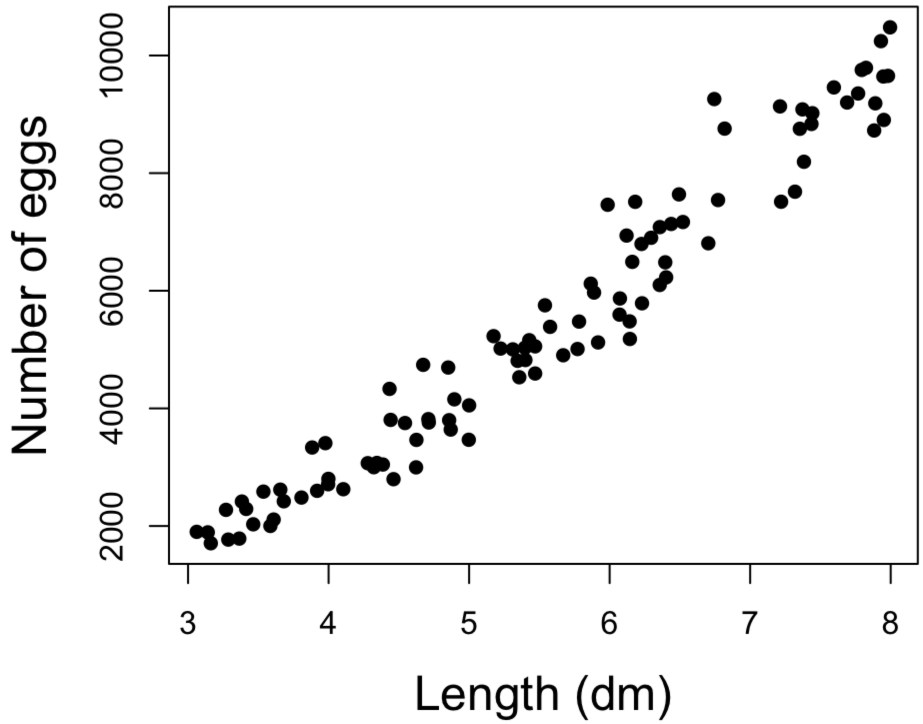
General considerations

- The Box–Cox method gets upset by outliers
- If some $y_i < 0$, we can add a constant to all the y
- If the range in y is small, then the Box–Cox transformation will not have much effect

Recall that linear models work well for *local* non-linear functions

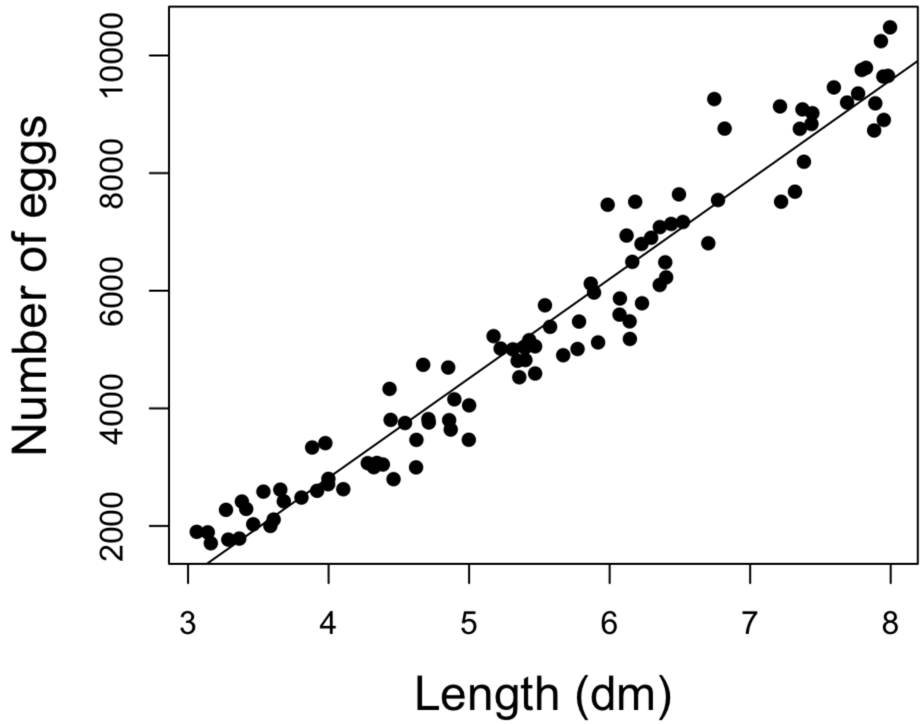
Alternative to Box-Cox

Consider the fecundity of a fish versus its length



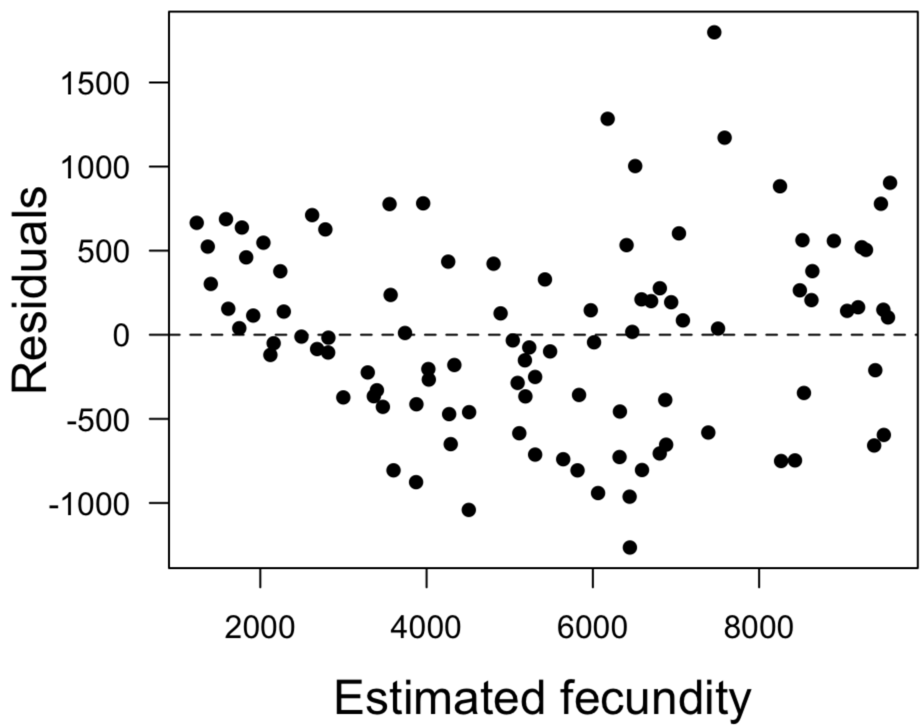
Alternative to Box-Cox

Here's the fit from a linear regression



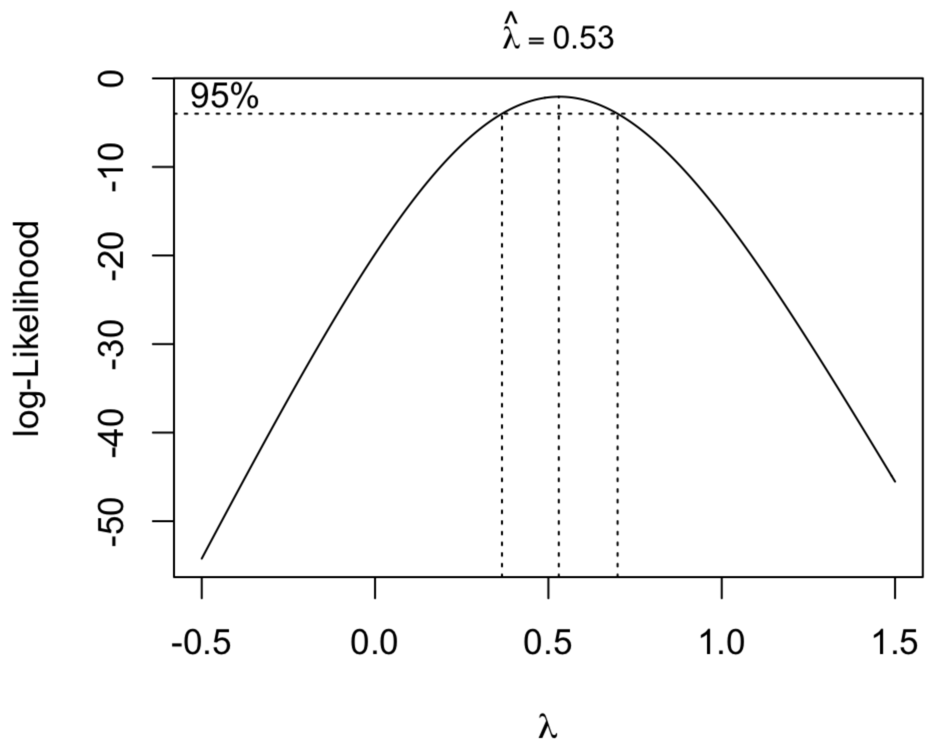
Alternative to Box-Cox

And here are the residuals from the fitted model



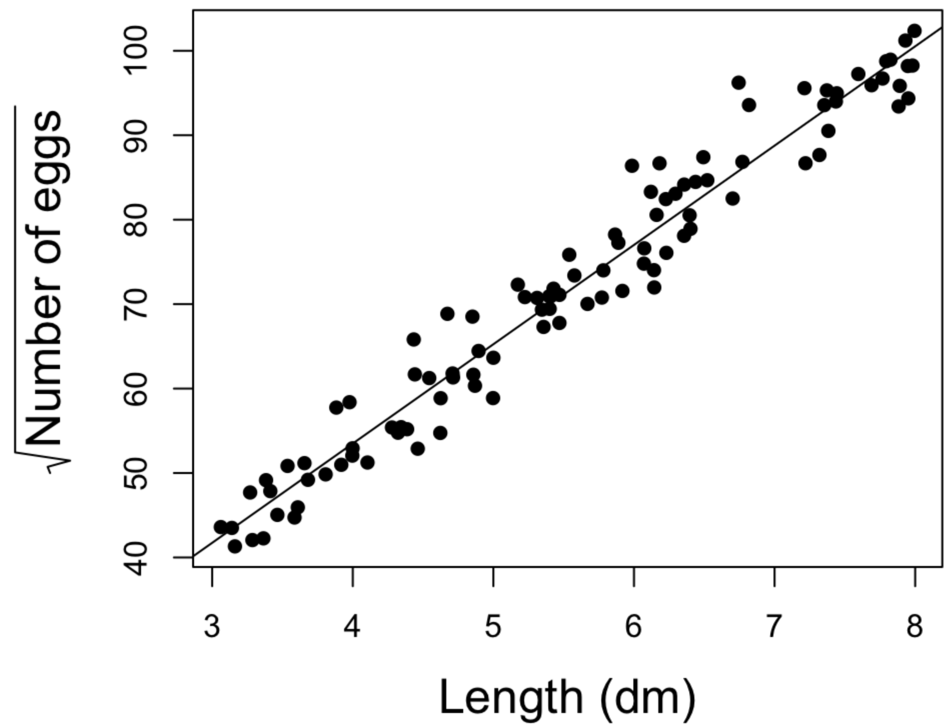
Alternative to Box-Cox

This $\hat{\lambda}$ is really close to 0.5 (ie, a square root transform)



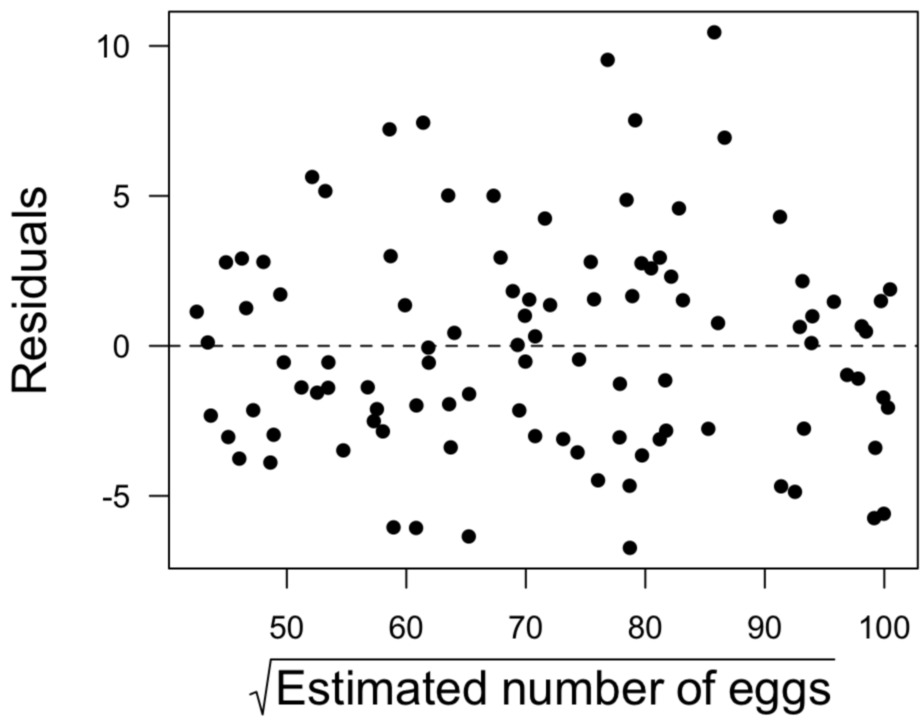
Square root transformation

Here's the fit from a linear regression to $\sqrt{y}$



Square root transformation

And here are the residuals from the fitted model



Predictions from a transformed model

Using `predict()` will give fits on the transformed scale

```
## expected sqrt(fecundity) for length = 5 dm
predict(ms, data.frame(ll = 5), interval = "confidence")
```

```
##        fit      lwr      upr
## 1 65.25383 64.48932 66.01835
```

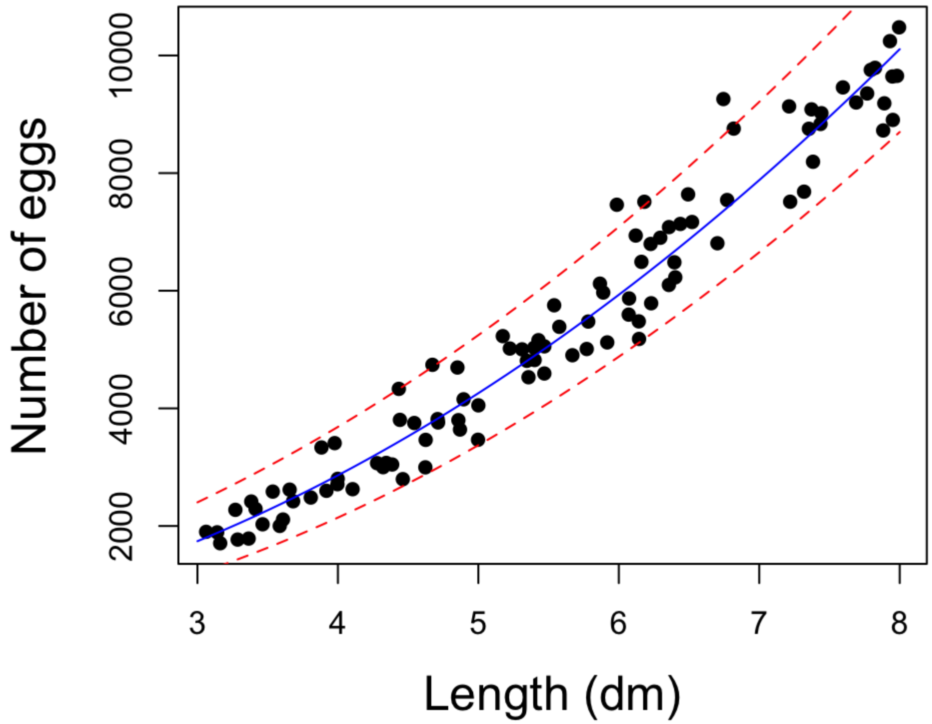
Predictions from a transformed model

We need to incude the back-transformation on `predict()`

$$
\begin{aligned}
\sqrt{\hat{y}_i} &= x_i\hat{\boldsymbol{\beta}} \\
&\Downarrow \\
\hat{y}_i &= (x_i\hat{\boldsymbol{\beta}})^2
\end{aligned}
$$

Back-transformed fit

Here's the fit and prediction interval on the natural scale



Transformed polynomials

Think back to an early lecture where we transformed a nonlinear polynomial into a linear model

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i}^2 + \epsilon_i$$

$$\Downarrow$$

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 z_{2,i} + \epsilon_i$$

$$z_{2,i} = x_{2,i}^2$$

Transformed polynomials

Polynomials are an easy way to model nonlinearities in data, such as

- Seasonal effects on primary productivity
- Temperature effects on growth of poikilotherms

Ecological data

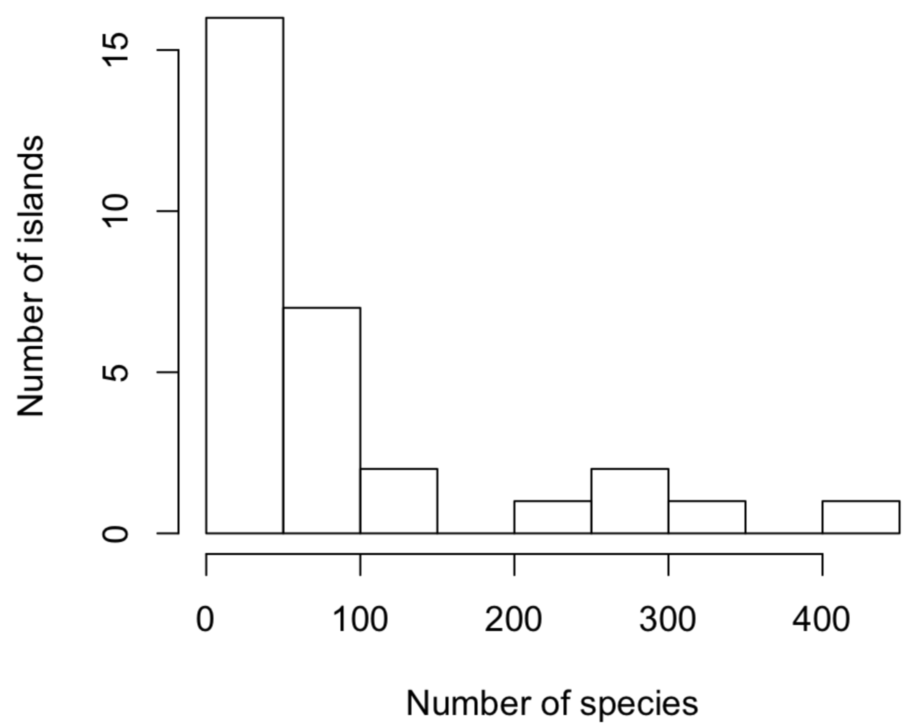
Many ecological observations only take positive values $(y > 0)$

- length or mass or fecundity
- species counts/density
- latency periods for infectious diseases

The distributions of these data also tend to be “long-tailed”

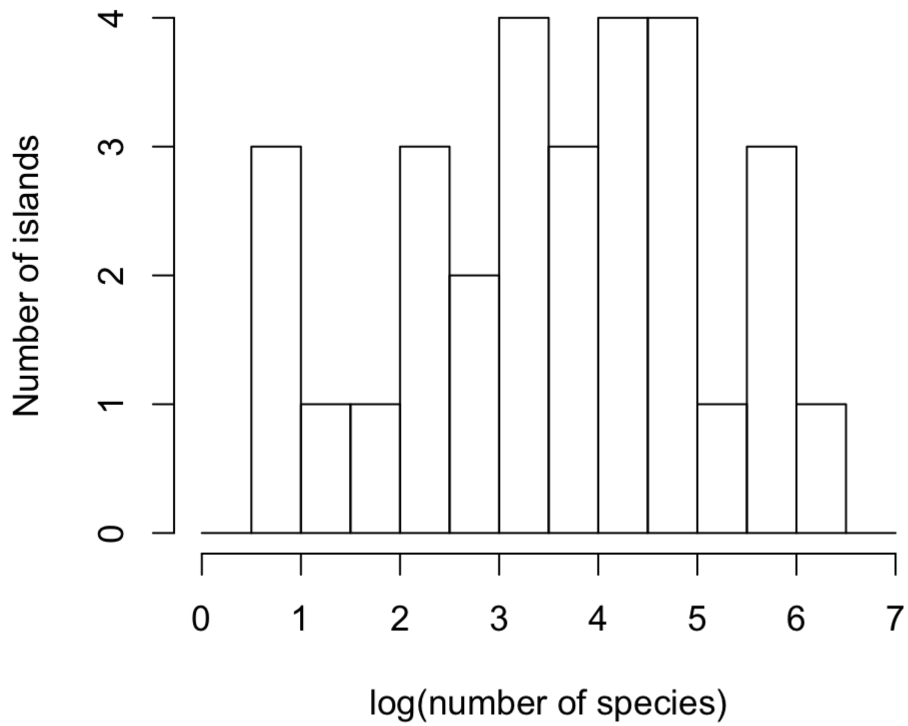
Long-tailed data

Distribution of plant diversity data in the `gala` dataset



Long-tailed data

These long-tailed data often follow a log-normal distribution



Log transformation

A log-transformation is a really common way to deal with ecological data that are constrained to be positive

$$y_i = \exp(\beta_0 + \beta_1 x_i + \epsilon_i)$$
$$\Downarrow$$
$$\log(y_i) = \beta_0 + \beta_1 x_i + \epsilon_i$$

Log-log transformation

Consider allometric scaling laws in ecology of the form

$$y_i = \alpha x_i^{\beta} \epsilon_i$$

For example, body mass as a function of length

Log-log transformation

Log-log transformations are an easy way to linearize power models

$$m_i = \alpha l_i^{\beta} \epsilon_i$$

$$\Downarrow$$

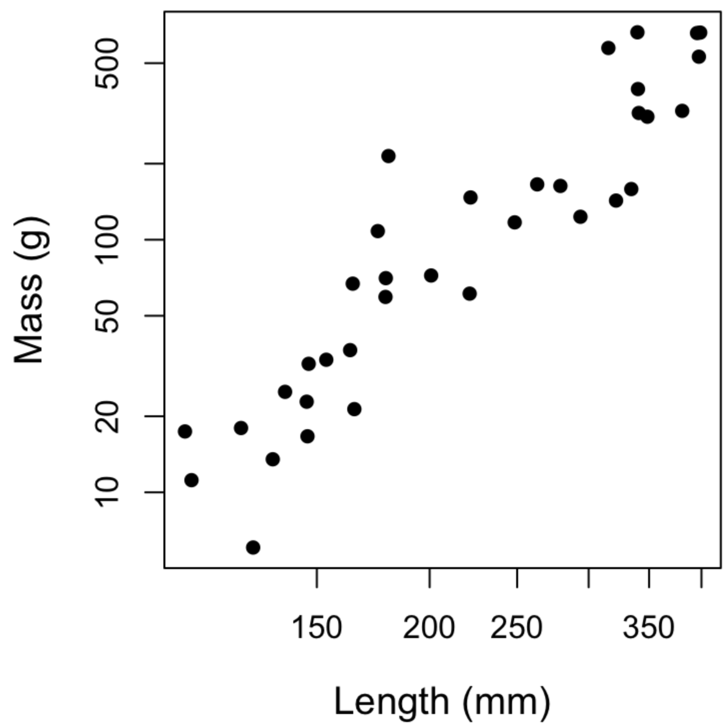
$$\log(m_i) = \log(\alpha) + \beta \ \log(l_i) + \log(\epsilon_i)$$

$$\Downarrow$$

$$y_i = \alpha' + \beta x_i + \epsilon_i'$$

Linear model for size of fish

The response and predictor are linear on the log-log scale



Summary

- Box-Cox is good to help ID a power/root, but the transformed variable can be hard to interpret
- $\sqrt{y}$ is good for equalizing variance
- $\log(y)$ is good for skewed data
- $\log(y+a)$ with a small relative to the data is good for skewed data with some 0’s
- We will see later that there are model alternatives to transformations (GLMs)