

# Regression: Transformations

Grinnell College

October 6, 2025

# Four Assumptions for the Linear Model

- Independence
  - ▶ Observations aren't correlated with each other
- Errors are normally distributed with mean 0
  - ▶ Want a random scattering above and below the horizontal line at 0
- Homoskedasticity
  - ▶ The general spread of the residuals should be constant through the whole graph
  - ▶ Eg We don't want a megaphone shape pattern in the graph
- X and y's relationship is linear

# General Plan

Can we do anything with violated assumptions?

- Yes, transformations!
- Logarithms
  - ▶ (FORMAT: Response Var.'s Scale - Explanatory Var.'s Scale)
  - ▶ Linear-Log
  - ▶ Log-Linear
  - ▶ Log-Log
- Power transformations
  - ▶ Square Root (generally of  $y$ )
  - ▶ Square (generally of  $x$ )
- Waaaay more that we can't get into

# Independence

Sometimes data is temporally (or spatially) related

- Eg my chess rating over time
- Have techniques for this
  - ▶ Back shift operator: subtract this years ave. September temp from last year's ave. September temp
  - ▶ Moving average: our current prediction is the (weighted?) average of the last few time points
  - ▶ Auto-Regressive: Observations are correlated with their neighbors

Sometimes it's physically related

- 6 green onions grown in the same small flower pot
- Have techniques for this as well
  - ▶ Mixed Models (not in this class)
  - ▶ Side step the issue (average over the six green onions)

# Non-Normality

Sometimes data just behaves differently

- CHECK FOR LURKING VARIABLE

- ▶ A variable not yet looked at could drive strange behavior

- I guess 3 heads out of four coin flips; I can only underguess by 1, overguess by 3 (not symmetric!)

- ▶ Logistic Regression

- Discrete data with very few non-0 numbers

- ▶ Discrete distribution over normal, eg Poisson Distribution
- ▶ Zero-Inflated Regression Model

- It's just weird

- ▶ Bootstrap (later) or simulations (also later)

# Heteroskedasticity

A common violation of our assumptions is....heteroskedasticity!

# Heteroskedasticity

A common violation of our assumptions is....heteroskedasticity!

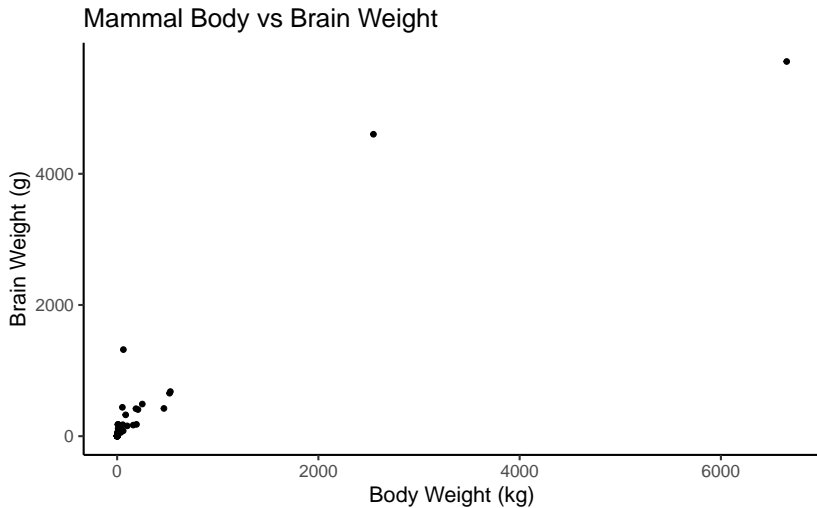
- It's common the st. dev. of your residuals changes depending on where you are at in the predicted vs residual graph
- Usually (but not always) scientifically expected
- St. Dev. usually (but not always) balloons as the predictions increase

Is there a way to transform the data into a more “usable” form?

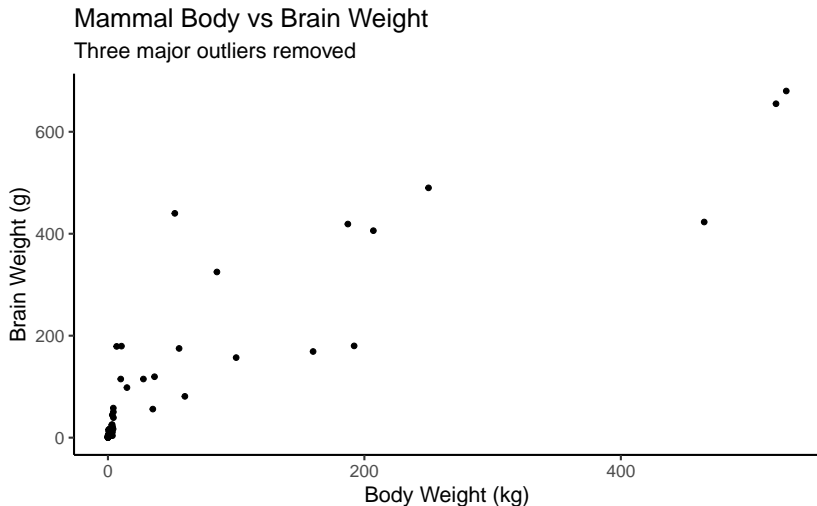
## A BRIEF MESSAGE FROM OUR SPONSORS



# Example



## Example: Continued



# Goal

We want data in a linear form, randomly scattered around a line with constant standard deviation (spread).

- We need a monotonic function that takes in (positive) numeric data
- and makes really large numbers not that large
- while keeping the small numbers relatively small
- and we want to be able to “back transform” (work backwards to get the original data).

Ideas?

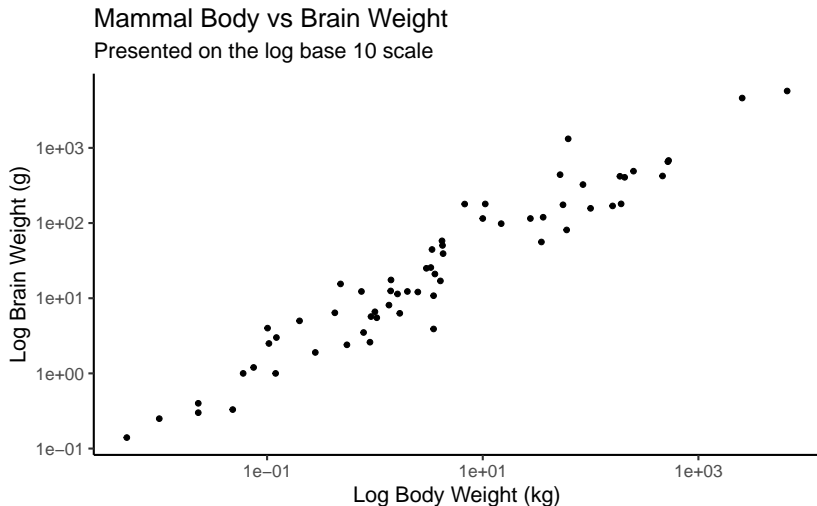
# Goal

We want data in a linear form, randomly scattered around a line with constant standard deviation (spread).

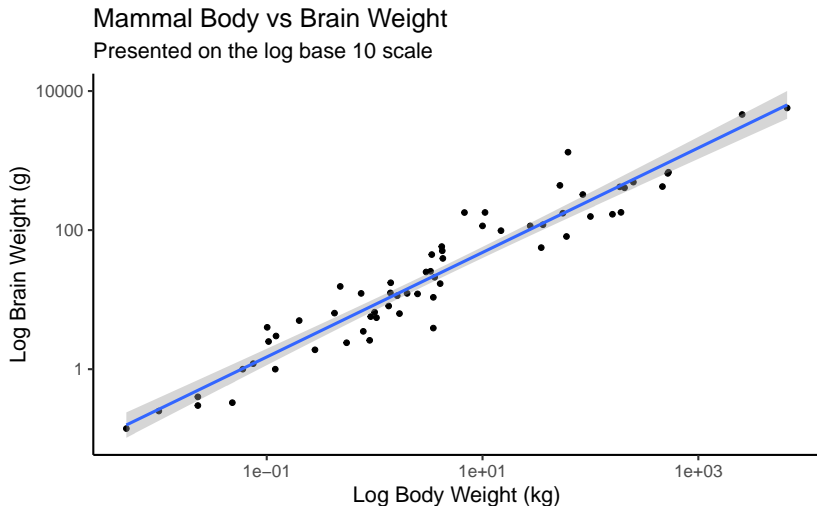
- We need a monotonic function that takes in (positive) numeric data
- and makes really large numbers not that large
- while keeping the small numbers relatively small
- and we want to be able to “back transform” (work backwards to get the original data).

Ideas? The logarithm does this, and so does the (positive) square root

## Example: Continued Some More



## Example: Predictions on Log-Log Scale



# Back Transforming

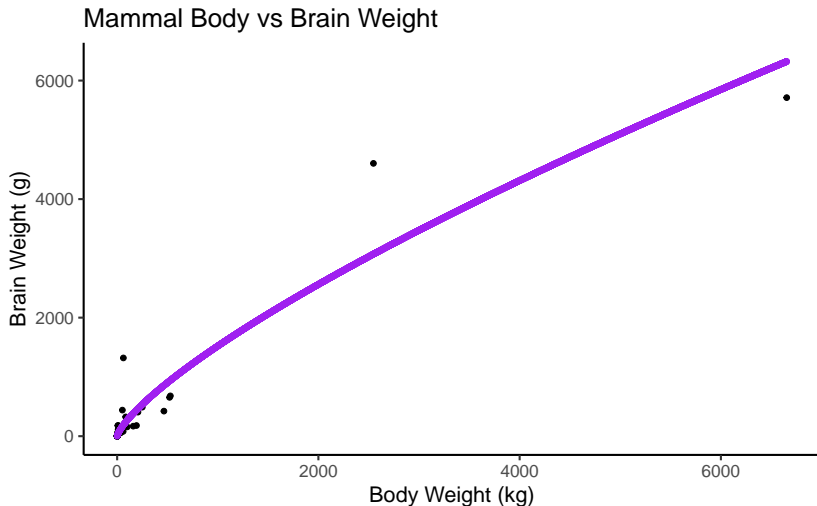
**Back Transforming** is applying a function to an already transformed variable to undo the transformation.

- ▶ Eg taking the square root of a squared variable
- ▶ Eg Undoing the log function by exponentiation
- ▶ Sounds more intimidating than it is

Usually back transformed values do well but not always

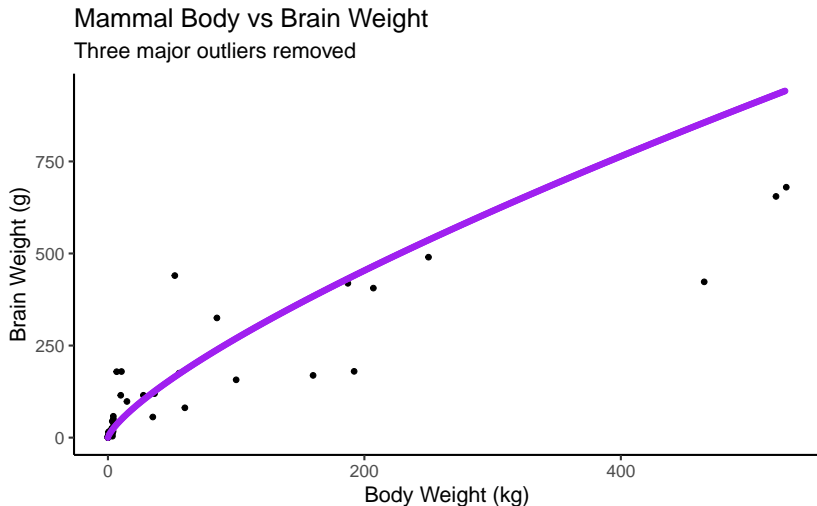
- Estimated spread (on the linear scale) balloons as you go up.
- Being off by .1 log units...
  - ▶ when the log value is low it's not bad (eg  $10^{.5} = 3.16$  vs  $10^{.6} = 3.98$ )
  - ▶ when the log value is high is bad (eg  $10^2 = 100$ ,  $10^{2.1} = 126$ )

## Example: Predictions on Linear-Linear Scale





## Example: Predictions on Linear-Linear Scale



# So what just happened?

Instead of

$$y = \beta_0 + \beta_1 X + e$$

we fit

$$\log(y) = \beta_0 + \beta_1 \log(X) + e$$

Raw Value	Log <sub>10</sub> Value
.2 (=5 <sup>-1</sup> )	-.698
.5	-.301
1	0
5	.698
500	2.698

# What's the catch?

Well.....

- We are modeling the log of the response,  $\log(y)$ , and not the response,  $y$ .
- Best-fit-line is guaranteed “best” only on the scale modeled
- And we are now fitting the median.....math is weird

Q: Wait, what??

# Interpretations Background

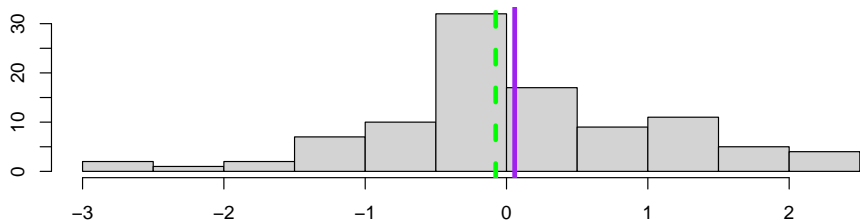
Well.....we are modeling the log of the response,  $\log(y)$ , and not the response,  $y$ .

- We are assumign the errors are bell shaped on the log scale
  - ▶ The distribution will be skewed when we back transform
  - ▶ But the median remains the same
- Best-fit-line is guaranteed “best” only on the scale modeled
- And we are now fitting the median.....math is weird

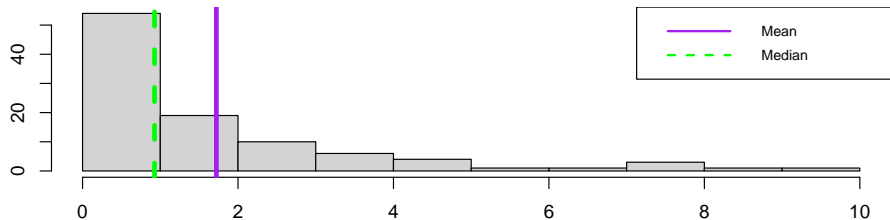
Q: Wait, what??

# Graphic Explanation

## Histogram of log\_normal



## Histogram of line\_skewed



# Estimated Log-Log Regression Equation

$$\log(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 * \log(X)$$

$$\hat{y} = \exp(\hat{\beta}_0 + \hat{\beta}_1 * \log(X))$$

$$\hat{y} = \exp(\hat{\beta}_0) * \exp(\hat{\beta}_1 * \log(X))$$

# Log-Log Model's Interpretation for $\hat{\beta}_0$

Two Interpretations:

---

BETTER:

When the log of the explanatory variable is 0, we expect the median response to be  $\exp(\hat{\beta}_0)$

---

WORSE:

When the log of the explanatory variable is 0, we expect the (mean/median) of the log of the response to be  $\hat{\beta}_0$

# Slope Interpretation Primer

$$\log(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 * \log(X)$$

$$\hat{y} = \exp(\hat{\beta}_0 + \hat{\beta}_1 * \log(X))$$

$$\hat{y} = \exp(\hat{\beta}_0) * \exp(\hat{\beta}_1 * \log(X))$$

---

$$y_{new}^{\hat{}} = \exp(\hat{\beta}_0) * \exp(\hat{\beta}_1 * \log(1.10 * X))$$

$$y_{new}^{\hat{}} = \exp(\hat{\beta}_0) * \exp(\hat{\beta}_1 * \log(X)) * \exp(\hat{\beta}_1 * \log(1.10))$$

$$y_{new}^{\hat{}} = \exp(\hat{\beta}_0) * \exp(\hat{\beta}_1 * \log(X)) * 1.10^{(\hat{\beta}_1)}$$

$$y_{new}^{\hat{}} = \hat{y} * 1.10^{(\hat{\beta}_1)}$$



# Log-Log Model's Interpretation for $\hat{\beta}_1$

## Two Interpretations

---

### BETTER:

When the the explanatory variable increases by 10% we expect the median response to increase by a multiplicative factor of  $1.1^{\hat{\beta}_1}$

---

### WORSE:

When the log of the explanatory variable increases by 1, we expect the (mean/median) of the log of the response to increase by  $\hat{\beta}_1$

## Example Continued

$$\text{Predicted log(Brain Weight)} = 2.13 + .75 * \log(\text{Body Weight})$$

$\beta_0$ : For a mammal with a log body weight of 0 (= the body weight of the animal is 1kg), then the predicted (median) weight of it's brain is  $e^{2.14} = 8.499\text{g}$

$\beta_1$ : If the body weight of the animal increases by 10%, then the predicted (median) brain weight increases by a multiplicative factor of  $1.1^{.75} = 1.074$

## Example Continued

Arctic Fox has a body weight of 3.385kg, what is the predicted median weight for its brain?

$$\text{Predicted log(Brain Weight)} = 2.13 + .75 * \log(\text{Body Weight})$$

## Example Continued

Arctic Fox has a body weight of 3.385kg, what is the predicted median weight for its brain?

$$\text{Predicted log(Brain Weight)} = 2.13 + .75 * \log(3.385)$$

$$\text{Predicted log(Brain Weight)} = 3.044$$

$$\text{Predicted Brain Weight} = e^{(3.044)} = 20.999g$$

$$\text{Residual} = \text{Obs.} - \text{Predicted} = 44.50 - 20.999 = 23.501$$

Estimate Log-Linear Regression Equation:

$$\log(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 * X$$

$$\hat{y} = \exp(\hat{\beta}_0 + \hat{\beta}_1 * X)$$

$$\hat{y} = \exp(\hat{\beta}_0) * \exp(\hat{\beta}_1 * X)$$

$\hat{\beta}_0$ : Interpretation is the same as for the log-log model

$\hat{\beta}_1$ : For a one unit increase in  $X$  we expect the median response to increase by a multiplicative factor of  $\exp(\hat{\beta}_1)$

Estimate Linear-Log Regression Equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * \log(X)$$

I don't know of any special interpretations to this one....useful if your x-axis is stretched out waaay too far

# What's the point?

Our goal is to achieve a graph that is linear between whatever is on the x-axis and whatever is on the y-axis with a random scattering points above and below the line (4 assumptions)

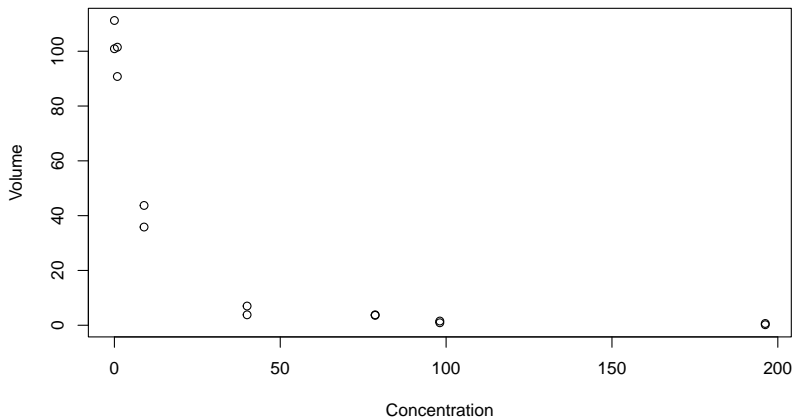
- We have techniques help us a way to achieve these goals
- The  $\log()$  transformation is popular to fix ballooning heteroskedasticity
  - ▶ Comes at a cost
  - ▶ No “best” properties on the human scale
  - ▶ The spread of our predictions balloons
  - ▶ Comments can be restricted to medians, not means
- Transformations in general can help us with this goal

# Other Transformations

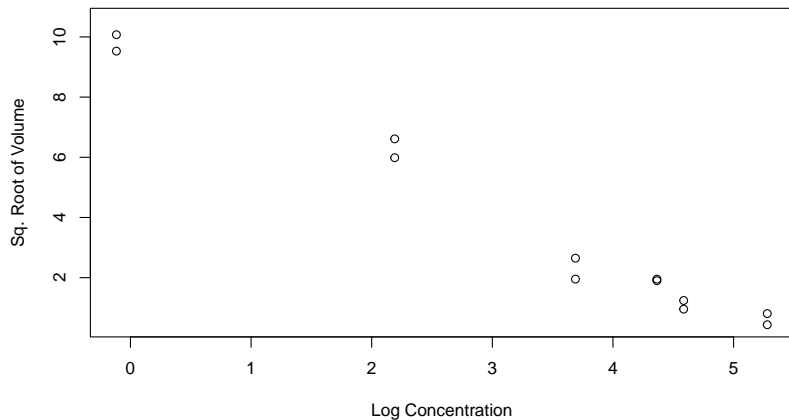
- Not uncommon for  $X$  to be replaced with  $X^2$
- The response can be raised to a power
  - ▶ Often square root or cubic root
  - ▶ Box-Cox Transformation gives a formula to optimize what power to raise the response to
  - ▶ Useful if the units give an indication (see next slide)
- Generalized linear models (GLM's) is a way to deal with the non-normal, quirky data via a different distribution
  - ▶ Eg Logistic and Poisson regression
  - ▶ Eg Exponential family
- Finally, non-parametric methods such as ranking the data
  - ▶ Calling Spearman's Correlation....



# Algae Blooms



# Algae Blooms: Transformed



# Next Time

How do we make a linear model when we have categories for our explanatory variable?