Advanced Topics in Regression

November 2025

WARNING

Today is suppose to be a fun day so these notes are not nearly comprehensive.

They are more to show you some of the more interesting things we can do with regression and to tie up some loose strings

I'm skipping over assumptions for all this but the three big ones (random, IID, large n) are still here in minor variations

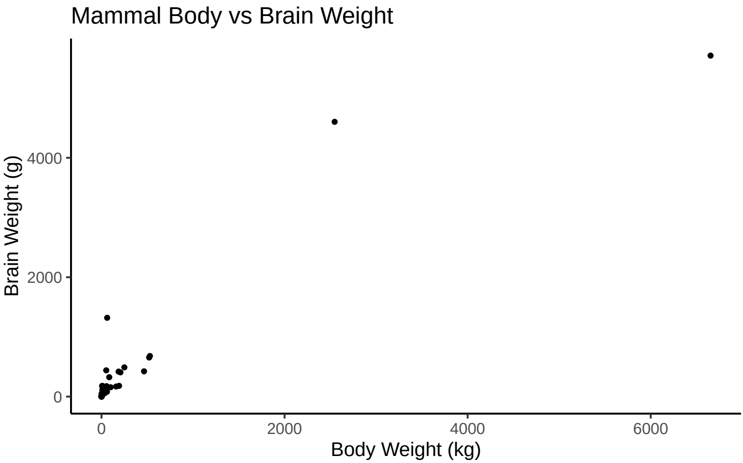
Starting Point: Why log(y)?

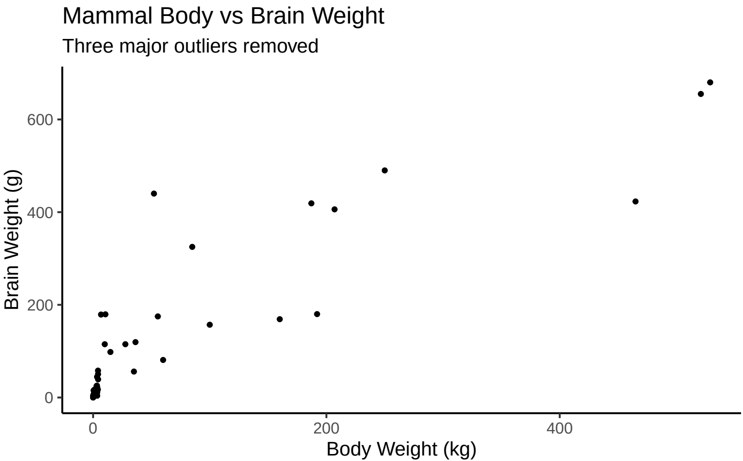
When do we take the log of the response variable y?

Starting Point: Why log(y)?

When do we take the log of the response variable y?

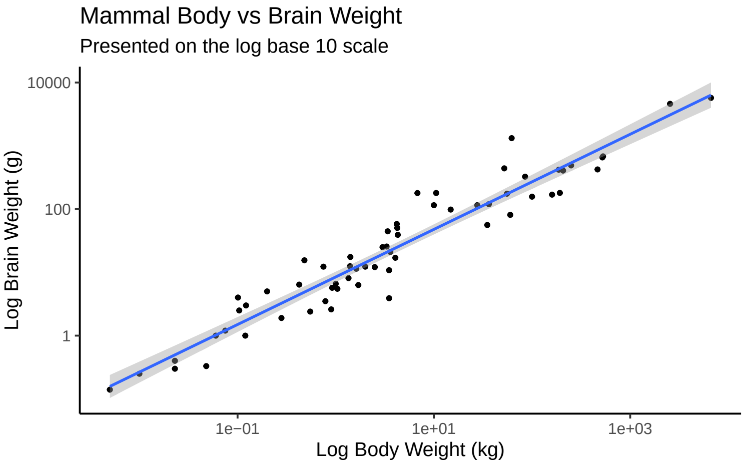
When our assumptions are violated and we need to try something new

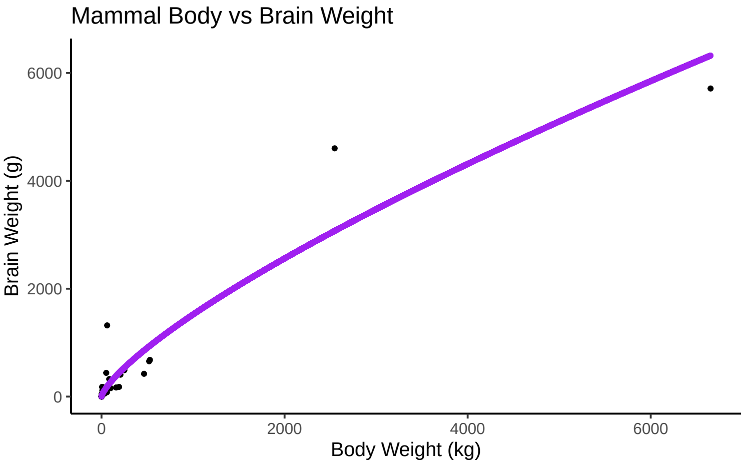

Mammal Body vs Brain Weight
Brain Weight (g)
4000
2000
0
Body Weight (kg)
0
2000
4000
6000

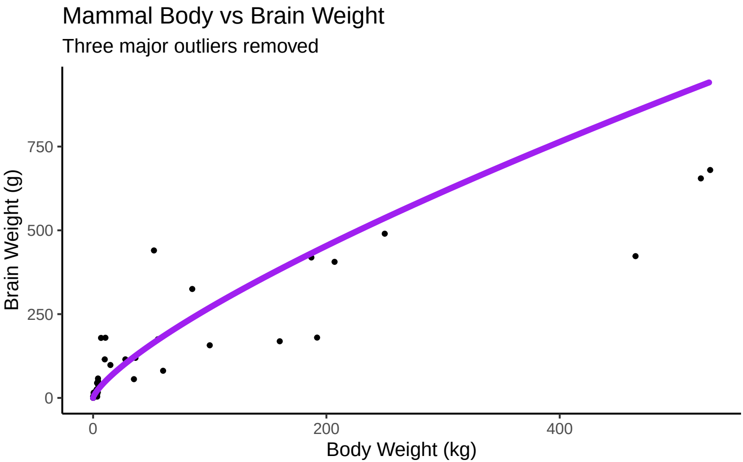


Mammal Body vs Brain Weight

Three major outliers removed


Mammal Body vs Brain Weight
Presented on the log base 10 scale
10000
100
1
Log Brain Weight (g)
1e−01
1e+01
1e+03
Log Body Weight (kg)


Mammal Body vs Brain Weight
Brain Weight (g)
0
2000
4000
6000
Body Weight (kg)
0
2000
4000
6000


Mammal Body vs Brain Weight
Three major outliers removed
Brain Weight (g)
750
500
250
0
0
200
400
Body Weight (kg)

Concept

We can work well with straight lines in the form of...

$$\hat{y} \ = \ \beta_0 \ + \ \beta_1 x$$

where $\hat{y}$ is our response variable and x is our explanatory variable

---

Our goal is to find some way to transform the response and/or the explanatory variable to get the model into this form

Going back to the mammal example, without the log transform on y our model looks like...

$$\hat{y} \ = \ e^{\beta_0 \ + \beta_1 x}$$

With the transformation we get

$$\hat{log}(y) \ = \ \beta_0 \ + \beta_1 x$$

Strategy

We start with a function in the form of....

$$\hat{y} \; = \; f(\beta_0 \; + \; \beta_1 x)$$

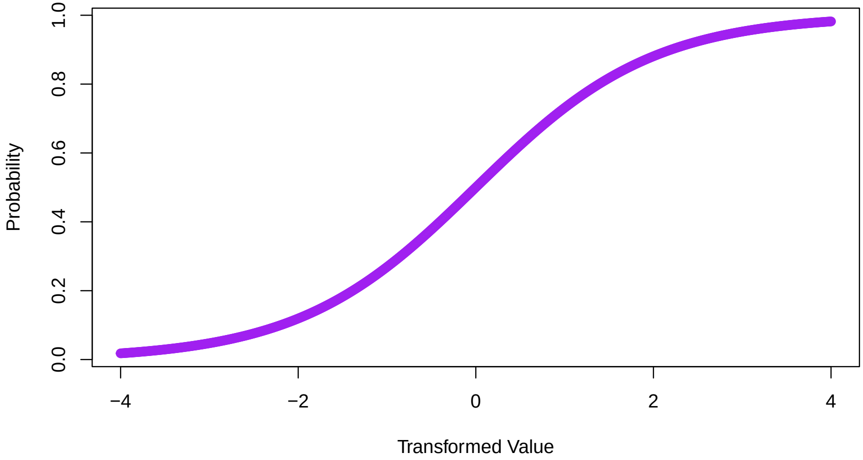
for some function f(.)

We then apply a back transform f^{-1}(.) to "undo" the function f(.) asdf

$$f^{-1}(\hat{y}) \; = \; \beta_0 \; + \; \beta_1 x$$

The earlier example had $\mathrm{f(a) = e}^{a}$ such that $\mathrm{f}^{-1}\mathrm{(a) = \log(a)}$

Example: Logistic Regression



Example: Logistic Regression

Logistic Regression is regression (line-fitting) where the resonse is either a proportion or a probability

- Response is betwnen 0 and 1
- Predictions are also between 0 and 1
  - Seems obvious but not true for other techniques.....
- Nothing to do with normality (initially)
- Explanatory variable can be numeric or categorical (indicator variables ftw!)
- Always check for overdisperion!!!
  - In R inside the glm() fucntion your use quasibinomial
- Checking residuals is verrry difficult

Agrettsi's "Categorical Data Analysis" is an excellent reference book even and I'd recommend downloading it for your future self

Example: Logistic Regression

Still using this form....

$$\hat{y} \ = \ f(\beta_0 \ + \ \beta_1 x)$$

The below equation fits the previous line

$$\hat{y} \ = \ \frac{1}{1 \ - \ e^{\beta_0 \ + \ \beta_1 x}}$$

And this one is the back transform that turns it linear

$$ln(\frac{\hat{}p}{1 - p}) \ = \ \beta_0 \ + \ \beta_1 x$$

which hopefully has normal errors

GLM's

Both linear and logistic regression are special types of regression called *generalized linear models*

- What shape do you think your response has around it's mean?
  - Ie what do you think the spread will be like around the mean?
- Linear models assumes normal
- Logistic assumes binomial
- Poisson regression assumes a poisson, etc...
- Usually does things to heterogenity
  - Eg log for fan shapes

GLM's offer us different shapes to choose for our model instead of normally distributed with equal variance

GLM Advice

A few things...

- Understand that a lot of the results depend on asymptotics, that is they assume an infinite sample size
- Small samples can be unreliable but difficult to know how small is small
- Often times can rationalize which one you want based on how the data is collected
    - eg For 12 beans sprouted out of 20 planted my knee jerk reaction to logistic
- Residuals...
    - Almost each distribution has it's own quirky variance
        - eg proportion's $\frac{p(1-p)}{n}$
    - So residuals are standardized by an estimation of their variance
    - Rule of thumb for continuous distributions: You want 95% of your residuals to be within $\pm$ 2

Interceptless Model

Previously in the lego lab you guys saw what happens when we add variables to the model

Turns out we can remove variables as well including the intercept, β_0

What effects will this have on the model?

Interceptless Model

Previously in the lego lab you guys saw what happens when we add variables to the model

Turns out we can remove variables as well including the intercept, β_0

What effects will this have on the model?

The best-fit-line will be forced to go through the origin (0,0). Usually this messes up our line but sometimes it's useful......

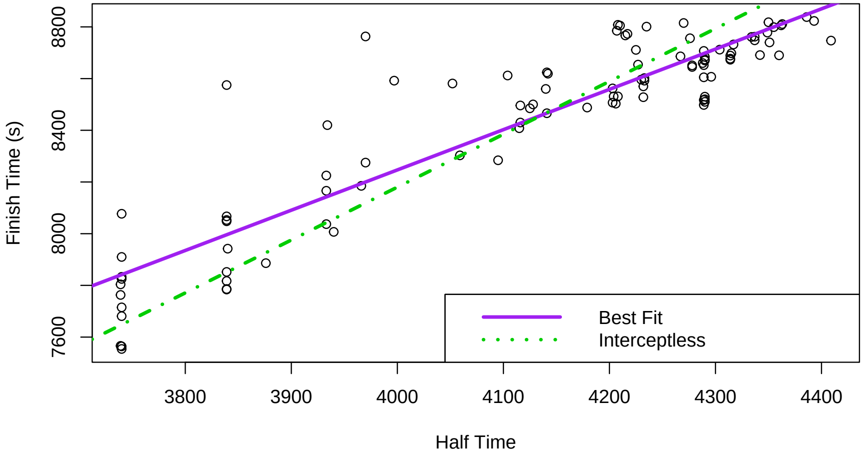
Interceptless Model: Example

Boston Marathon Lab:

- We predicted finish time based on how long it took to run half the race
- When we did our slope was 1.558
- This implies if it takes 1000 seconds to get to the midway point, it'll take 1558 seconds to finish the race
- Waaay counter-intuitive
- Having an intercept in our model makes things go weird so instead....

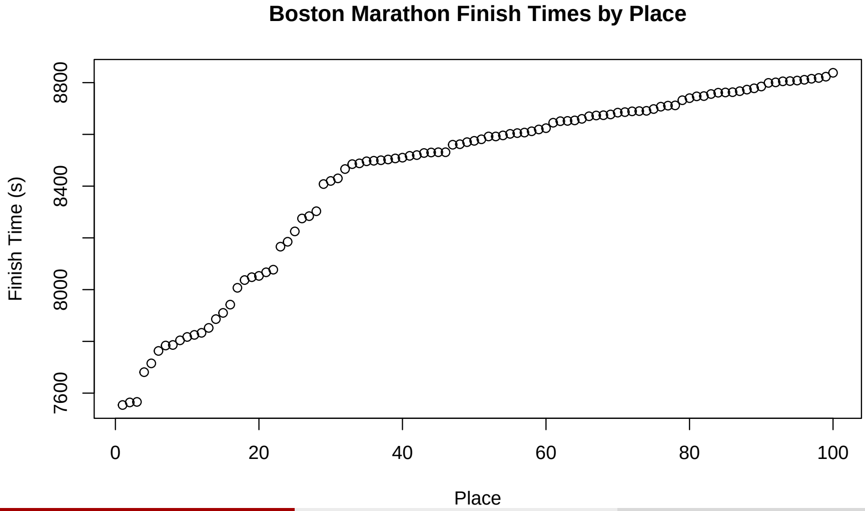
Interceptless Model: Example

Unclear if the interceptless model is an improvement buuuut the slope is now 2.045 which makes a lot more sense

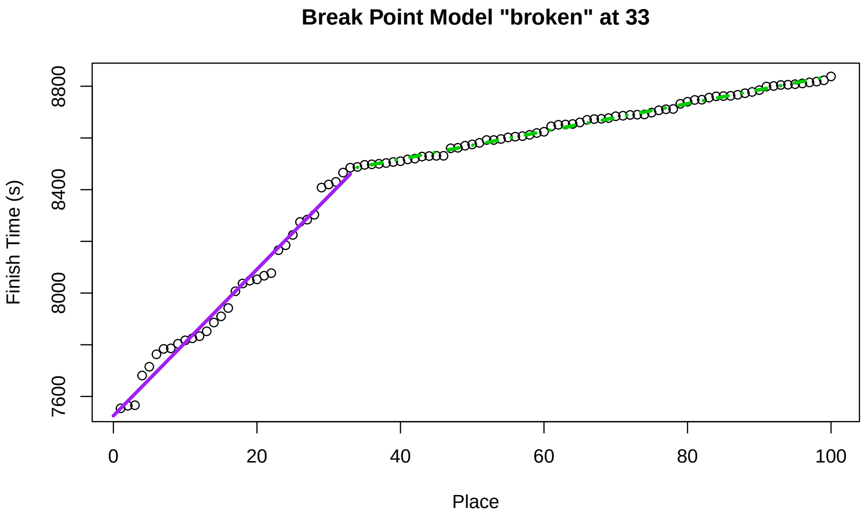


Break Point Model

Special type of multiple regression where we have a “break point” on the x-axis where are line changes



Break Point Model



Break Point Model

Several things to be aware of...

- We can allovw the different lines to have different variances (spreads) as well
- WARNING: It is very tempting to use these to overcome all problems
  - Usually that's masking the real underlying scientific cause you should try to find instead
  - Like duct tape...an ugly solution that can be used in a lot of places it shouldn't be used
- Best is if there was some a priori known incident to change how the population behaves
  - Eg Life expectancy of Native Americans along the East Coast pre-/post-Jamestown's settlement

Polynomials

We also saw the linear-log model earllier with

$$\hat{y} \;=\; \beta_0 \;+\; \beta_1 log(x)$$

---

which can generalize to...

$$\hat{y} \;=\; \beta_0 \;+\; \beta_1 f(x)$$

For example...

$$\hat{y} \;=\; \beta_0 \;+\; \beta_1 x^2$$

Polynomials

These come up a whole lot in nature

- Think about fishing over the course of a year
  - Catching fish will peak in summer time
  - And be pretty lower for fall through spring
    - Even ice fishing making a subpopulation can be put in the model
- Only useful if your explanatory variable is numeric (ie not categorical)
- Each new term allows your best fit line ot have one more “turn”
  - Second-order polynomial $(1 + x + x^2)$ is a parabola/ a u-shape
  - Fourth order Polynomial $(1 + x + x^2 + x^3 + x^4)$ let's you draw an “m”

Many scientific processes don't work in just straight lines but instead are polynomials or apporoximated by polynomials

Polynomials

Hierarchy Principle: If a higher order term is in the model (eg x^5) then all lower order terms must be in the model (eg x^4, x^3, etc...)

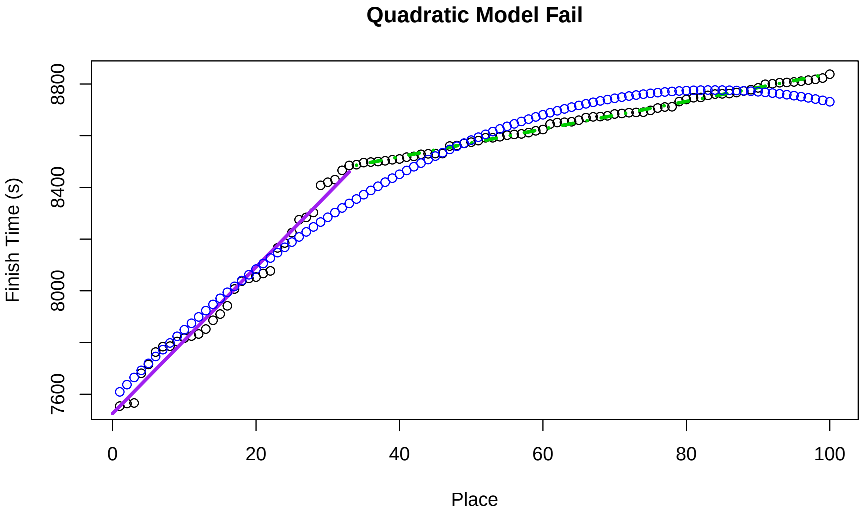
$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2$$

Advice from my (other) advisor: You will never know if there is a higher order term if you don't check

- Run the model with a quadratic term and without it
- Compare residuals to see if there's an improvement
- Also can consult p-values
- For some reason a plurality of scientists have a phobia of this

(Usually the hierarchy principle is invoked with respect to interactions (if AB then both A and B) but a square is an interaction of a variable with itself)

Polynomials



f