Lesson 26

Evaluating SLR models; Data Transformation

Using SLR models and data transformations

CSCI 3022

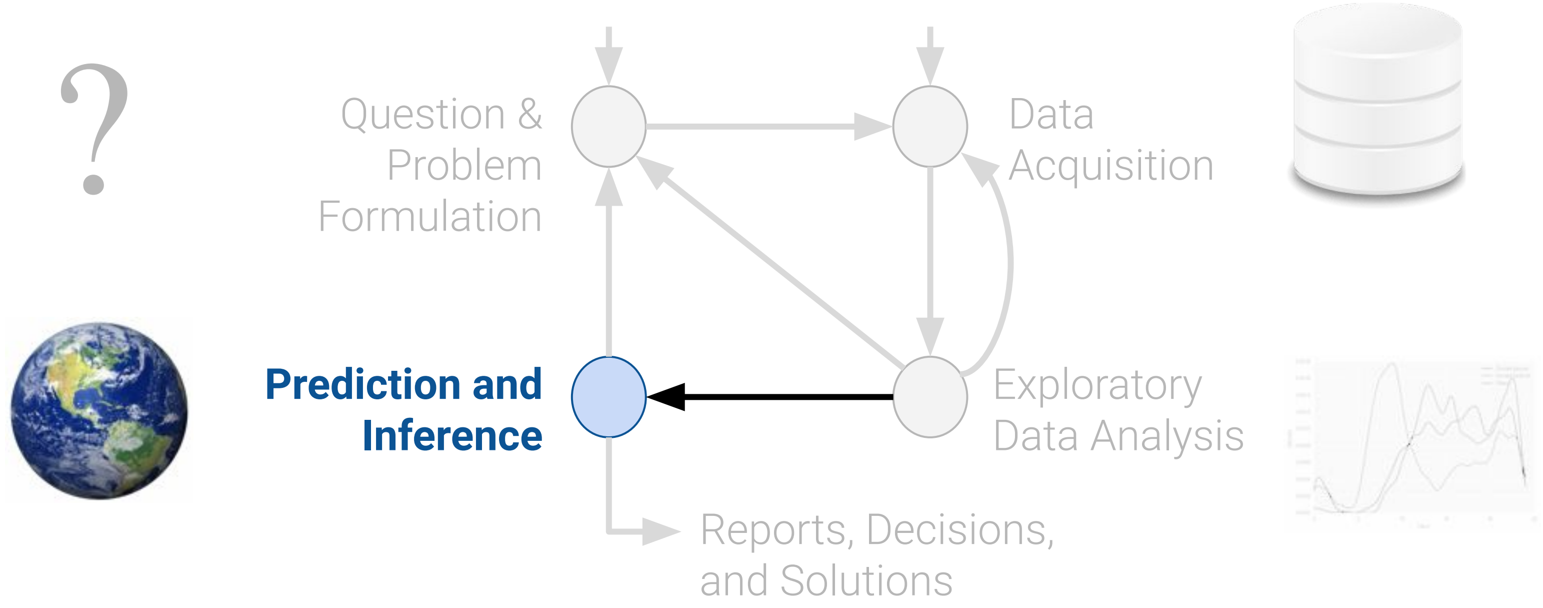
Maribeth Oscamou

Content credit: Acknowledgments

Course Logistics: 12th and 13th Weeks At A Glance

Mon 4/8	Tues 4/9	Wed 4/10	Thurs 4/11	Fri 4/12
Attend & participate in class	TA NB Discussion 5pm-6pm via Zoom **Project Part 1 Released**	Attend & participate in class	**HW 10 Due 11:59pm MT**	Attend & participate in class NO Quiz!
Mon 4/15	Tues 4/16	Wed 4/17	Thurs 4/18	Fri 4/19
Attend & participate in class	TA NB Discussion 5pm-6pm via Zoom **Project Part 2 Released**	Attend & participate in class	**Project Part 1 Due: 11:59pm MT**	Attend & participate in class Quiz 8

Plan for Rest of Semester: Modeling


Question & Problem Formulation
Data Acquisition
Prediction and Inference
Exploratory Data Analysis
Reports, Decisions, and Solutions

(today)

Modeling I: Different models, loss functions

Modeling II: Simple Linear Regression

Modeling III: Multiple Linear Regression

Today's Roadmap

- Finish Lesson 25:
  - Inference in SLR models
- Evaluating Simple Linear Regression Model
- Data Transformations to Fit Models

Summary of the 3 Models We've Learned So Far:

	Model	Estimate	Unique?
Constant Model + MSE	[illegible]	$\hat{\theta} = \mathbf{mean}(y)$	**Yes**. Any set of values has a unique mean.
Constant Model + MAE	[illegible]	$\hat{\theta} = \mathbf{median}(y)$	**Yes**, if odd. **No**, if even. Return average of middle 2 values.
Simple Linear Regression + MSE	$\hat{y} = \theta_0 + \theta_1 x$	$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$ $\hat{\theta}_1 = r \frac{\sigma_y}{\sigma_x}$	**Yes**. Any set of non-constant* values has a unique mean, SD, and correlation coefficient.

The Modeling Process

1. Choose a model ✓

How should we represent the world?

$$\hat{y} = \theta_0 + \theta_1 x$$

SLR model

2. Choose a loss function ✓

How do we quantify prediction error?

$$L(y, \hat{y}) = (y - \hat{y})^2$$

Squared loss

3. Fit the model ✓

How do we choose the best parameters of our model given our data?

$$\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1 x))^2$$

MSE for SLR

4. Evaluate model performance

How do we evaluate whether this process gave rise to a good model?

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x \quad \begin{cases} \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} \\ \hat{\theta}_1 = r\dfrac{\sigma_y}{\sigma_x} \end{cases}$$

Evaluating the Model: RMSE, Residual Plot

Evaluating Models

What are some ways to determine if our model was a good fit to our data?

1. **Visualize data, compute statistics:**

   Plot original data.
   Compute means, standard deviations.
   If we want to fit a linear model, compute correlation r.

2. **Performance metrics:**

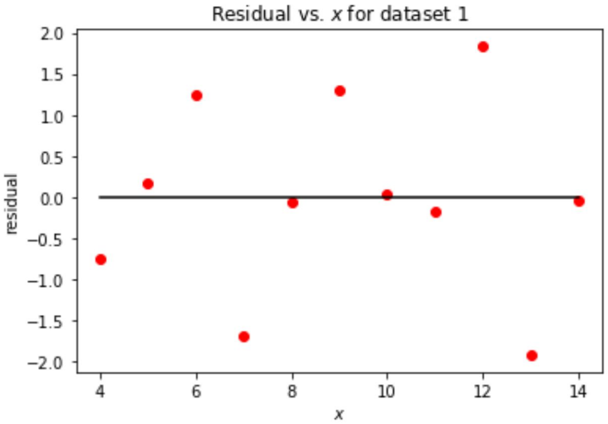
   Root Mean Square Error (RMSE)

   $$\boxed{\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}$$

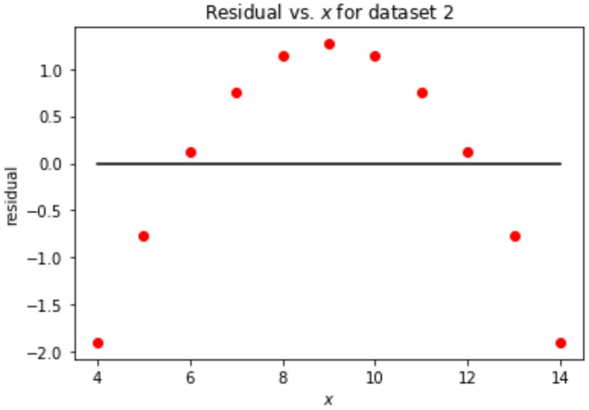
   - It is the square root of MSE, which is the average loss that we've been minimizing to determine optimal model parameters.
   - RMSE is in the same units as y.
   - A lower RMSE indicates more "accurate" predictions (lower "average loss" across data)

3. **Visualization:**

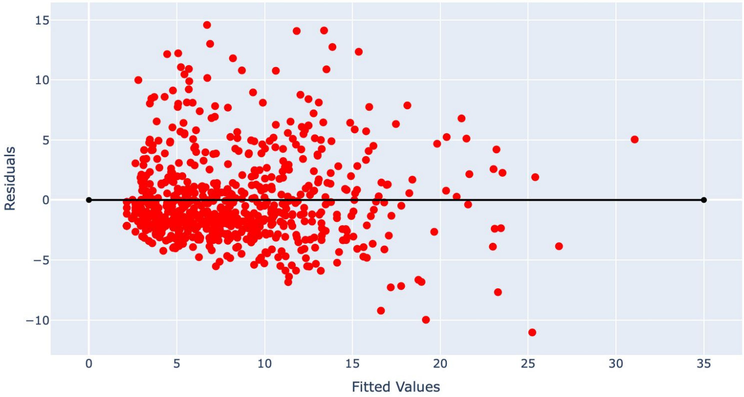
   Look at a residual plot of $e_i = y_i - \hat{y}_i$ to visualize the difference between actual and predicted values.

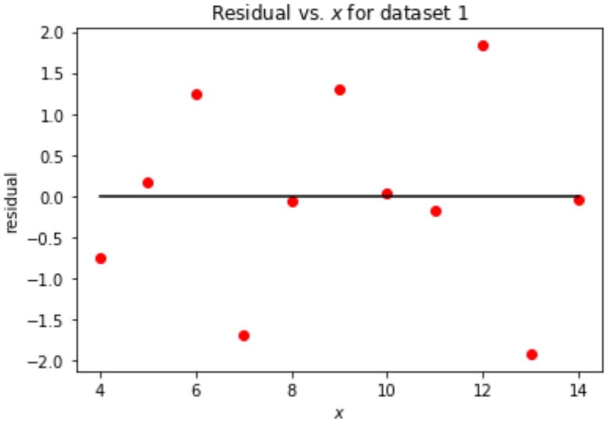


Residual vs. x for dataset 1

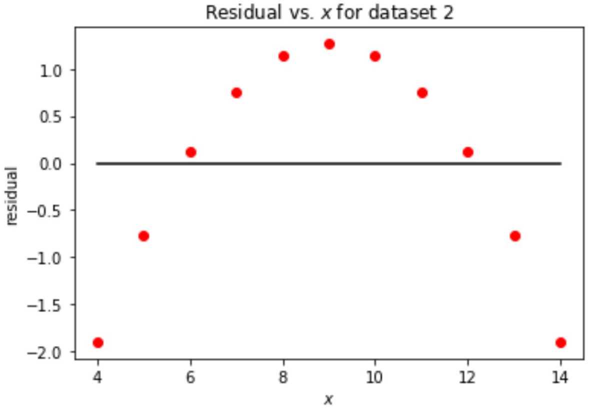


Residual vs. x for dataset 2

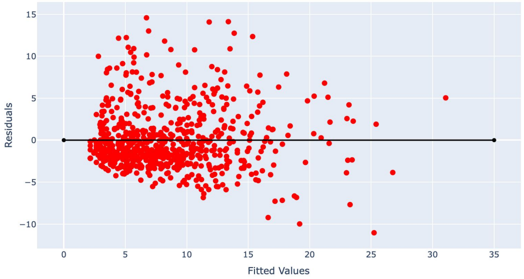




No pattern, even spread.



Clear quadratic relationship in the residuals Should try a transformation of the data.



No clear relationship, but uneven spread. Should be careful when usin linear model to make predictio

Discussion Question.

Suppose you have two datasets A and B.

Both datasets each have the same mean of x, mean of y, SD of x, SD of y, and r value.

$$\bar{x} = 9, \bar{y} = 7.501$$

$$\sigma_x = 3.162, \sigma_y = 1.937$$

$$r = 0.816$$

True or False:

A). Both datasets must the be same (i.e. any data point in A must also be in B and vice versa)

B). Both datasets must have the same regression line

Demo Slides

Four Mysterious Datasets + Least Squares

Ideal model evaluation steps, in order:

1. **Visualize original data, compute statistics**
2. **Performance Metrics**
   For our simple linear least square model, use RMSE (we'll see more metrics later)
3. **Residual Visualization**

It is tempting to only look at step 2.
But you need to always visualize!!!!

Visualize, Then Quantify!

Anscombe's quartet refers to the following four sets of points on the right.

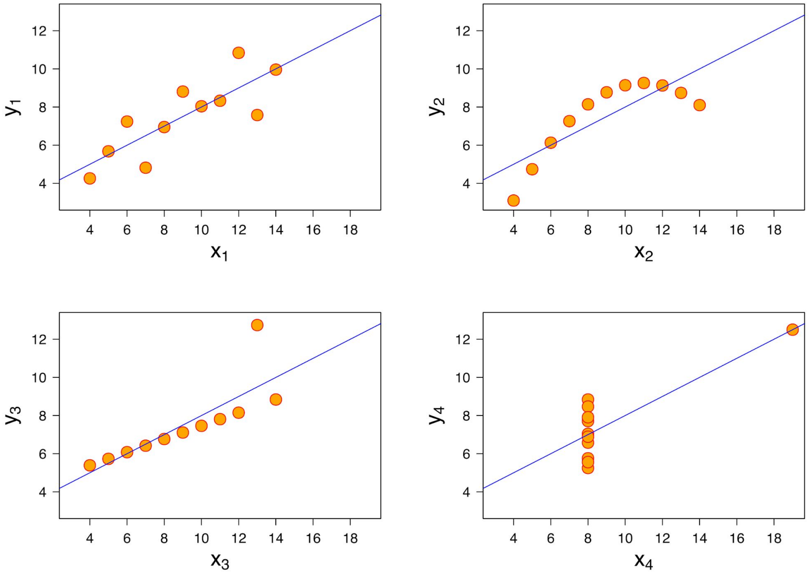
- They each have the same mean of x, mean of y, SD of x, SD of y, and r value.
- Since our optimal Least Squares SLR model only depends on those quantities, they all have the **same regression line**.

$$\bar{x} = 9, \bar{y} = 7.501$$

$$\sigma_x = 3.162, \sigma_y = 1.937$$

$$r = 0.816$$

However, only one of these four sets of data makes sense to model using SLR.



Before modeling, you should always visualize your data first!

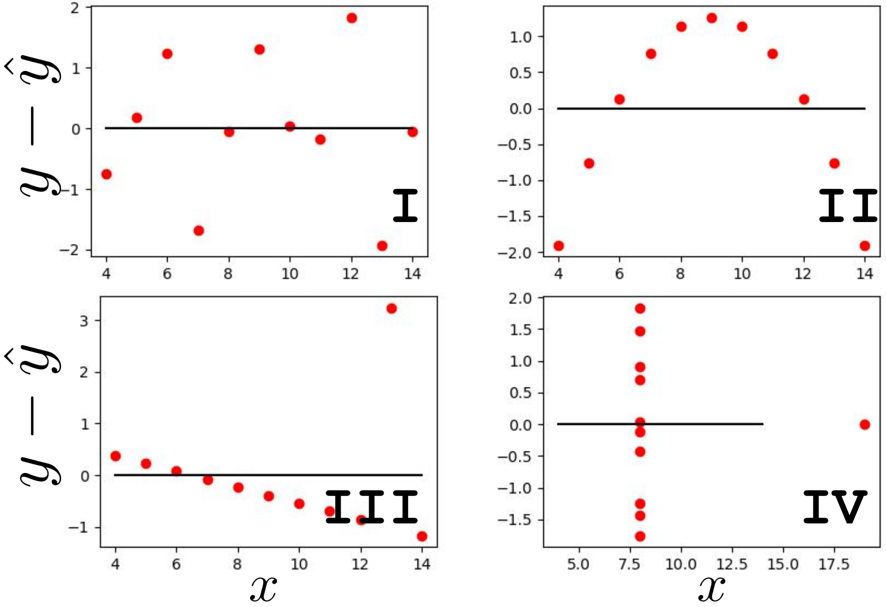
Four Mysterious Datasets + Least Squares

Ideal model evaluation steps, in order:

1. Visualize original data,
   Compute Statistics
2. Performance Metrics
   For our simple linear least square model, use RMSE (we'll see more metrics later)
3. **Residual Visualization**

4 datasets could have similar aggregate statistics but still be wildly different:

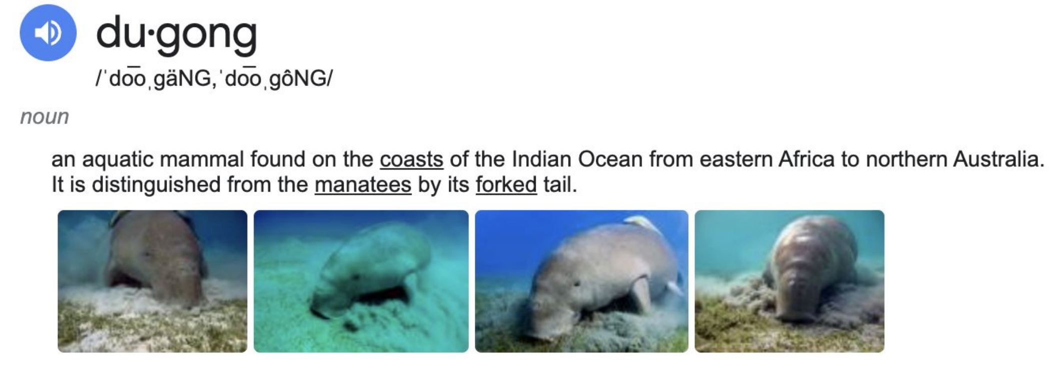
```
x_mean : 9.00, y_mean : 7.50
x_stdev: 3.16, y_stdev: 1.94
r = Correlation(x, y): 0.816
ahat: 3.00,    bhat: 0.50
RMSE: 1.119
```



The residual plot of a good regression shows no pattern.

Example:

Suppose we wanted to predict dugong ages.



du·gong

/ˈdo͞oˌgäNG,ˈdo͞oˌgôNG/

noun

an aquatic mammal found on the coasts of the Indian Ocean from eastern Africa to northern Australia. It is distinguished from the manatees by its forked tail.

Compare

[Data] Comparing Two Different Models, Both Fit with MSE

Suppose we wanted to predict dugong ages.

[image source]

Constant Model

$$\hat{y} = \theta_0$$

Data: Sample of ages.

$$\mathcal{D} = \{y_1, y_2, \ldots, y_n\}$$

Simple Linear Regression

$$\hat{y} = \theta_0 + \theta_1 x$$

Data: Sample of (length, age)s.

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

Compare

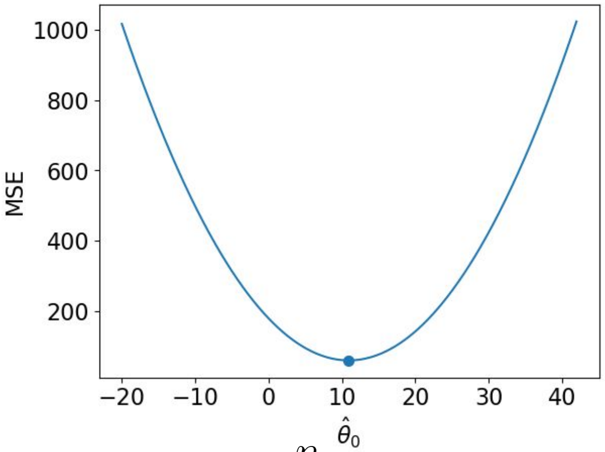
[Loss] Comparing Two Different Models, Both Fit with MSE

Constant Model

$$\hat{y} = \theta_0$$

$\hat{\theta}_0$ is **1-D**.

Loss surface is **2-D**.



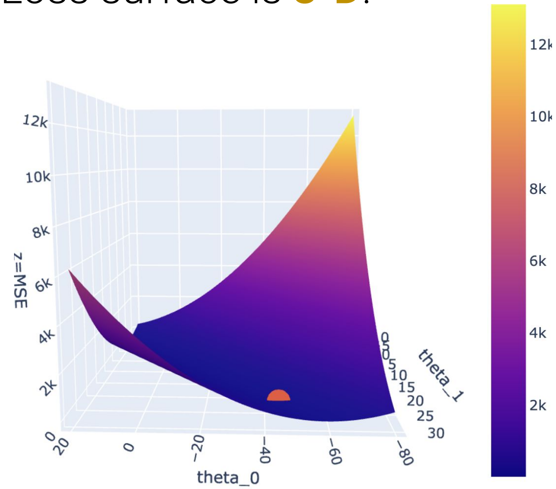
$$\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_0)^2$$

Simple Linear Regression

$$\hat{y} = \theta_0 + \theta_1 x$$

$\hat{\theta} = [\hat{\theta}_0, \hat{\theta}_1]$ is **2-D**.

Loss surface is **3-D**.



$$\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1 x))^2$$

Compare

See notebook for code

[Fit] Comparing Two Different Models, Both Fit with MSE

Constant Model

$$\hat{y} = \theta_0$$

RMSE: **7.72**

Simple Linear Regression

$$\hat{y} = \theta_0 + \theta_1 x$$

RMSE **4.31**

Interpret the RMSE (Root Mean Square Error):

- Constant error is **HIGHER** than linear error
- Constant model is **WORSE** than linear model (at least for this metric)

**In general, the RMSE will always decrease when you add new features ** (if you are using the same data to train both models).

[Fit] Comparing Two Different Models, Both Fit with MSE

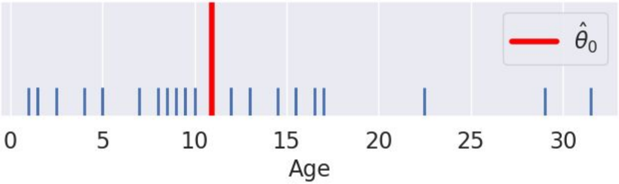
Compare

Constant Model

$$\hat{y} = \theta_0$$

RMSE: **7.72**

Predictions on a **rug plot**.

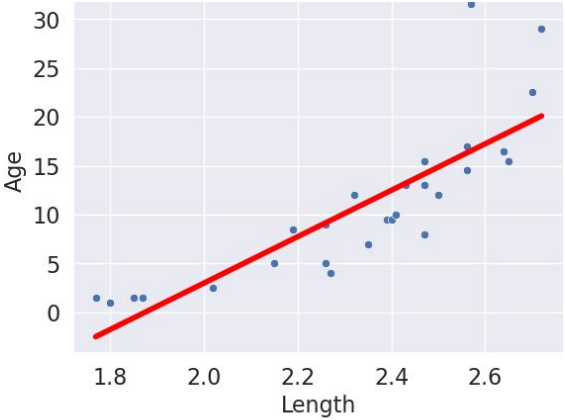


Simple Linear Regression

$$\hat{y} = \theta_0 + \theta_1 x$$

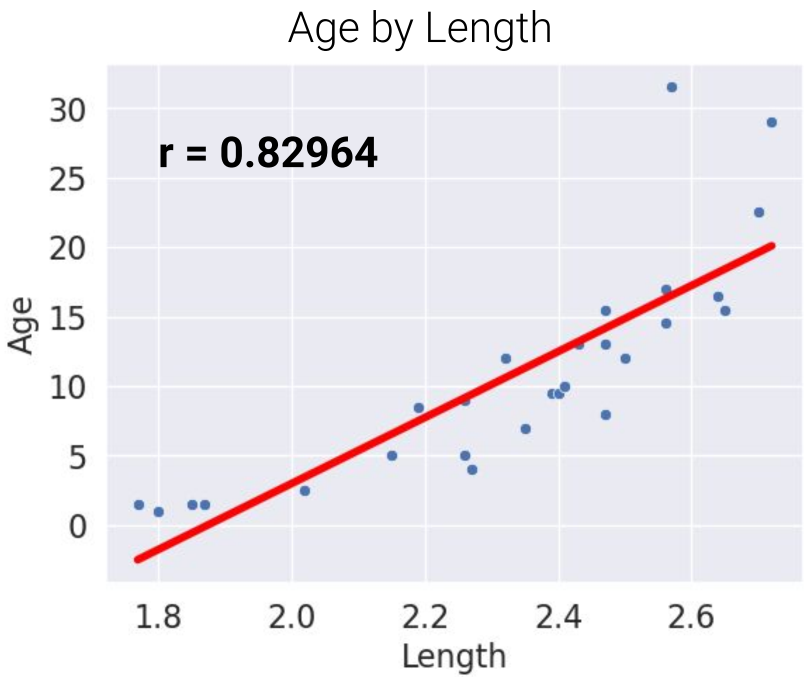
RMSE **4.31**

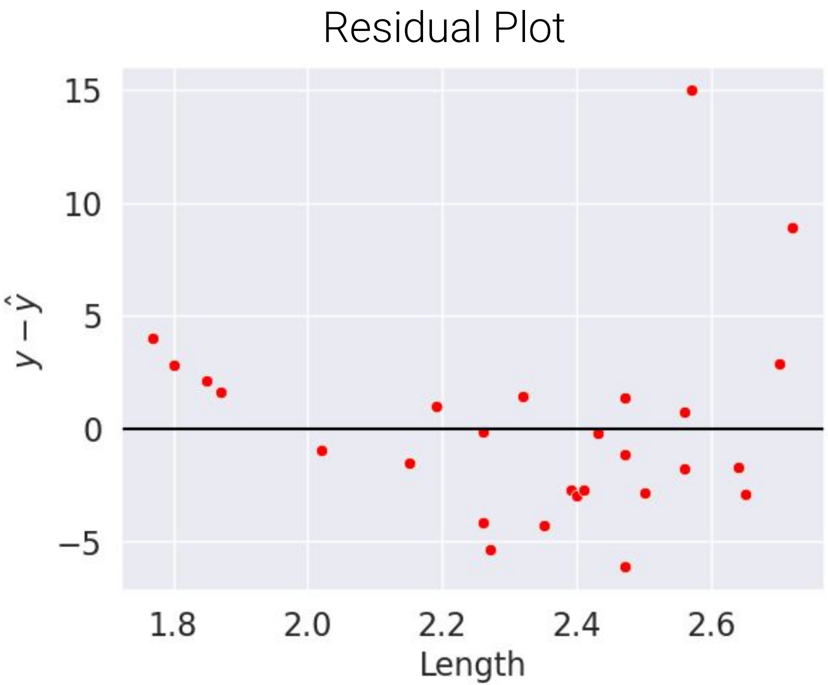
Predictions on a **scatter plot**.



Not a great linear fit visually?
We'll come back to this…

Least Squares Regression with Dugongs





Residual plot shows a clear pattern! On closer inspection, the scatter plot **curves upward**.

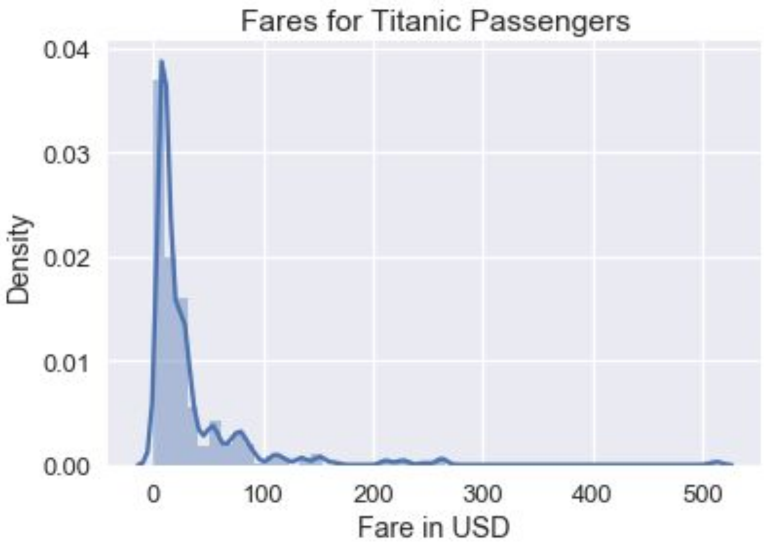
Q: How can we fit a curve to this data with the tools we have?

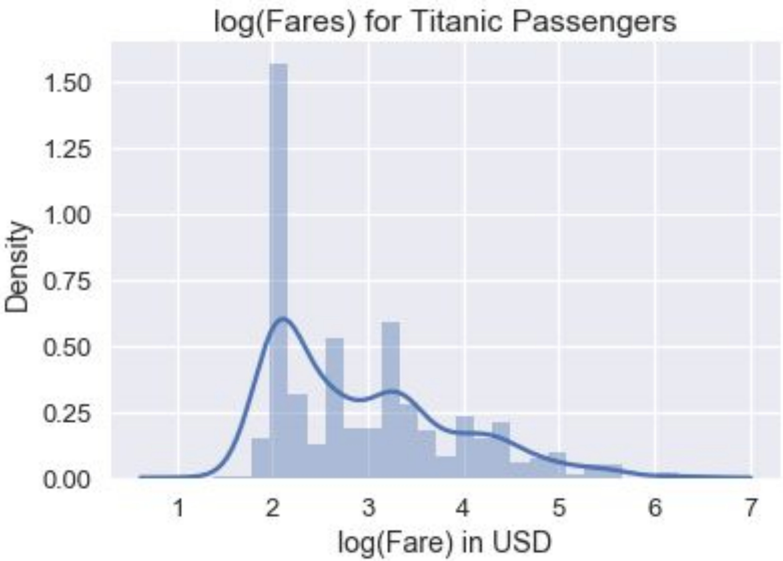
A: **Transform the Data**.

Transformations to Fit Linear Models

Transformations to Fit Linear Models

Transforming data can reveal patterns

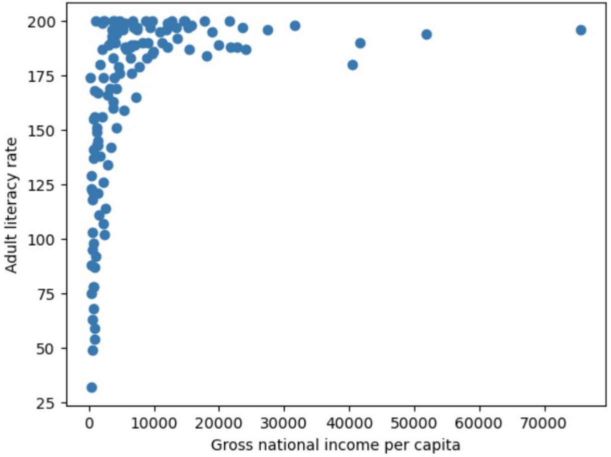


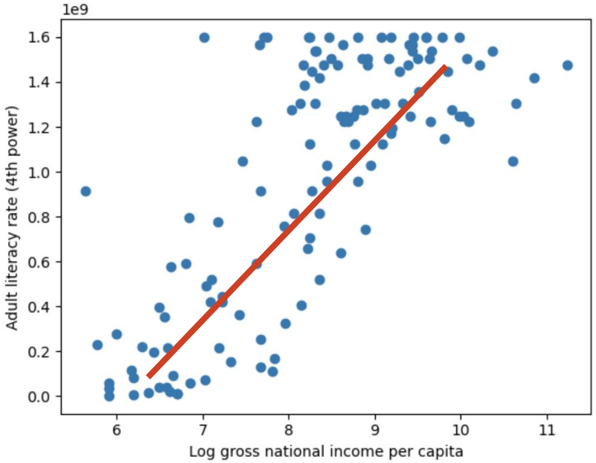


When a distribution has a large dynamic range, it can be useful to take the log.

Linearization

When applying transformations, we often want to **linearize** the data – rescale the data so the x and y variables share a linear relationship.





Why?

- Linear relationships are simple to interpret – we know how to work with slopes and intercepts to understand how two variables are related.
- We can then build linear models

Log of y-values

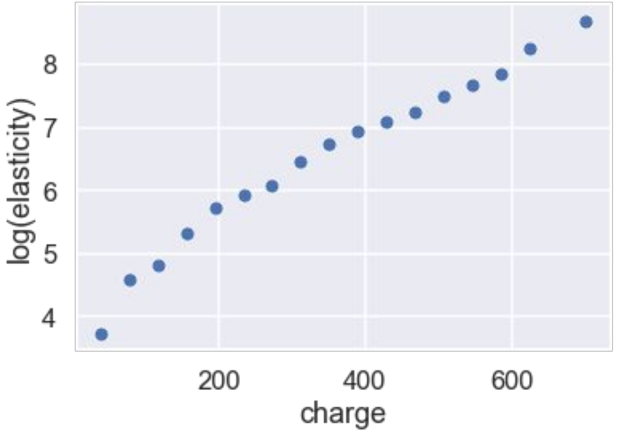
If we take the log of our y-values and notice a linear relationship, we can say (roughly) that

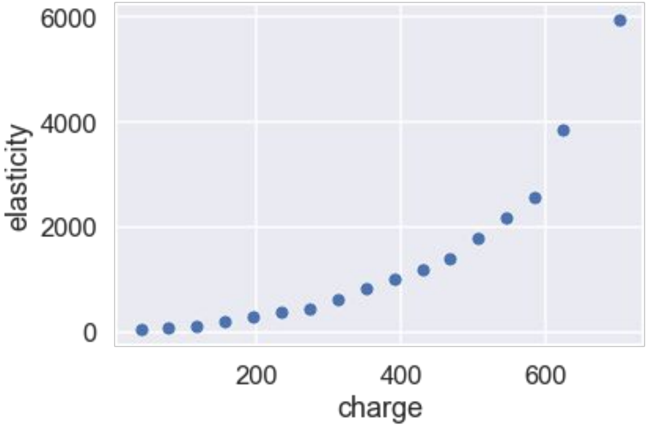
$$\log y = ax + b$$

Working backwards:

$$\log y = ax + b$$

This implies an ________________ relationship in the original plot.





Log of y-values

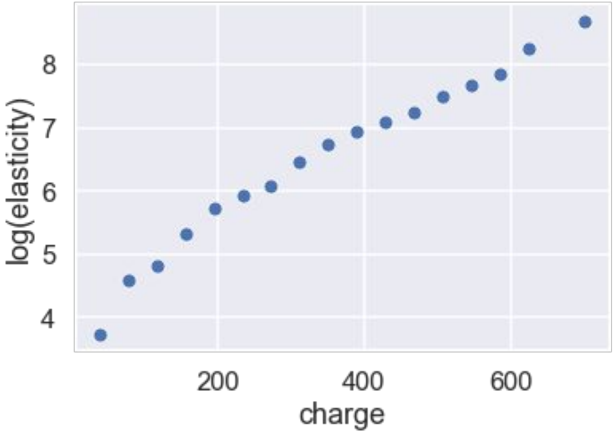
If we take the log of our y-values and notice a linear relationship, we can say (roughly) that

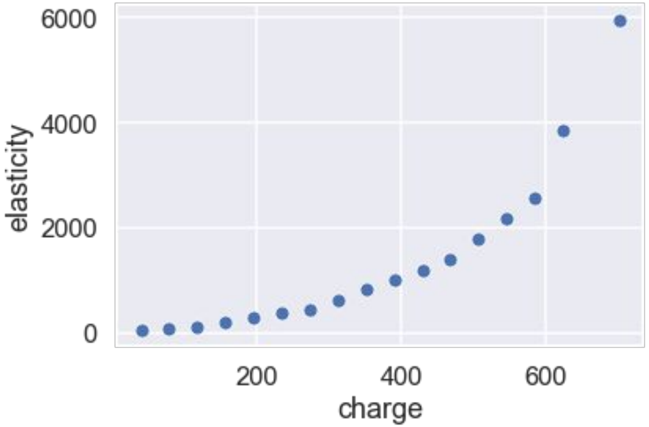
$$\log y = ax + b$$

Working backwards:

$$\begin{aligned}\log y &= ax + b \\ y &= e^{ax+b} \\ y &= e^{ax}e^{b} \\ y &= Ce^{ax}\end{aligned}$$

This implies an **exponential** relationship in the original plot.





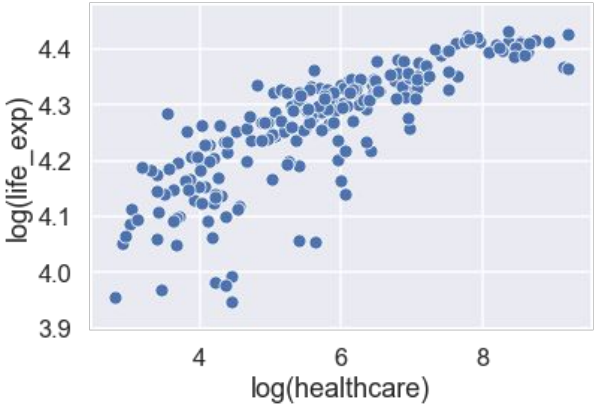
Log of both x and y-values

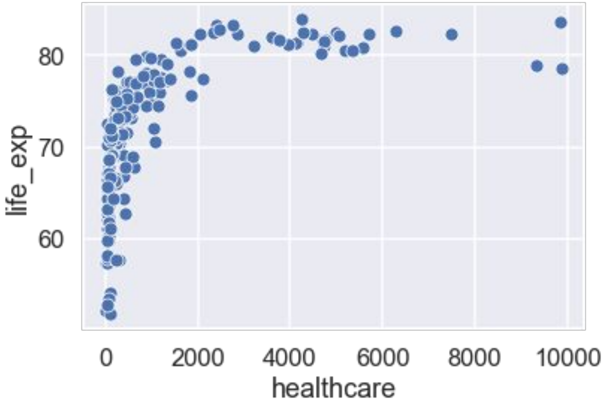
If we take the log of both axes and notice a linear relationship, we can say (roughly) that

$$\log y = a \cdot \log x + b$$

Working backwards:

This implies a ____________ relationship in the original plot (a one-term _______________)





Log of both x and y-values

If we take the log of both axes and notice a linear relationship, we can say (roughly) that

$$\log y = a \cdot \log x + b$$

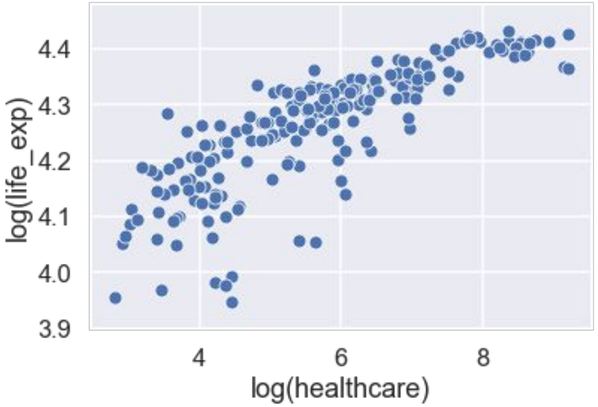
Working backwards:

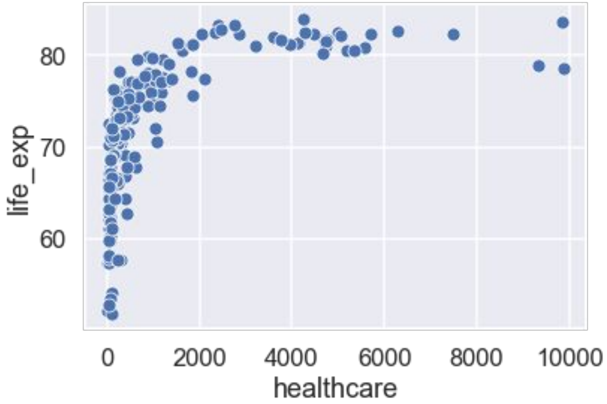
$$y = e^{a \cdot \log x + b}$$

$$y = Ce^{a \cdot \log x}$$

$$y = Cx^a$$

This implies a **power** relationship in the original plot (a one-term **polynomial**)





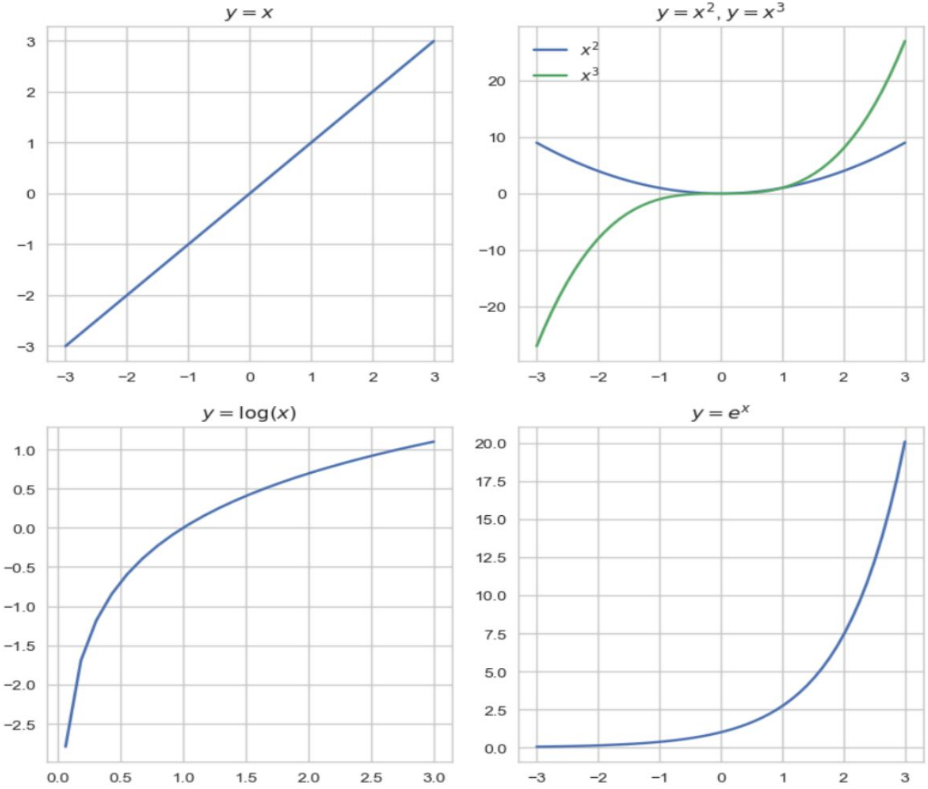
Log transform as a “Swiss army knife”

$$y = a^x \rightarrow \log(y) = x\log(a)$$
$$y = ax^k \rightarrow \log(y) = \log(a) + k\log(x)$$

Properties of logarithms make them very powerful!

Basic functional relations

Knowing the general shapes of polynomial, exponential, and logarithmic curves (regardless of base) will go a long way.

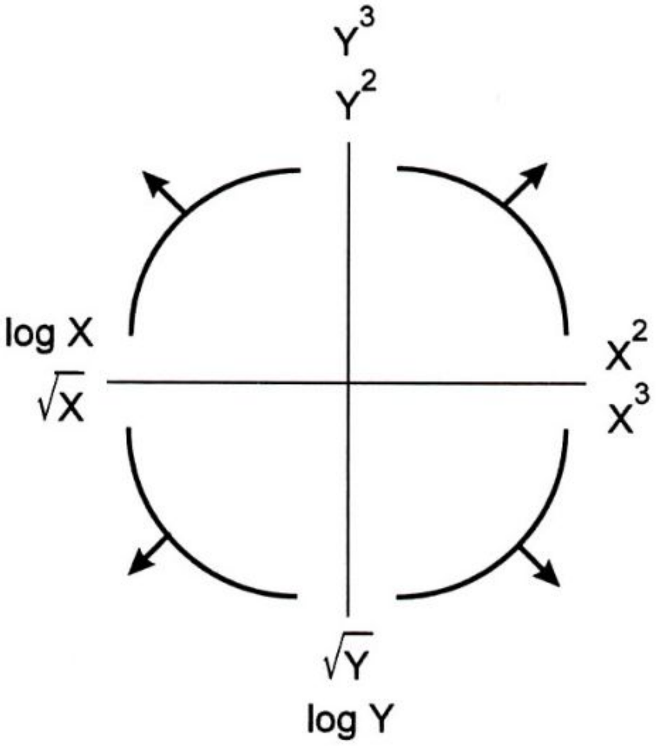


Tukey-Mosteller Bulge Diagram

The **Tukey-Mosteller Bulge Diagram** is a guide to possible transforms to try to get linearity.

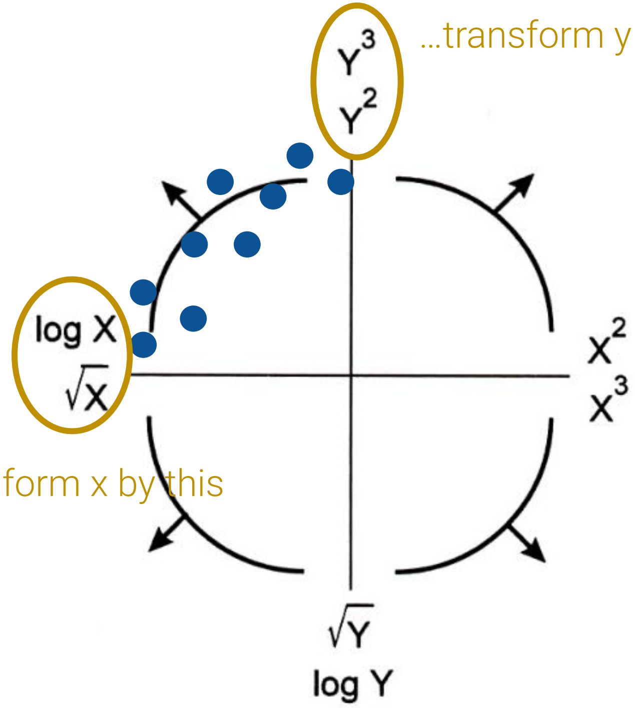
- A visual summary of the reasoning we just worked through.
- sqrt and log make a value "smaller".
- Raising to a value to a power makes it "bigger".
- There are multiple solutions. Some will fit better than others.

You should still understand the *logic* we just worked through to decide how to transform the data. The bulge diagram is just a summary.



Tukey-Mosteller Bulge Diagram

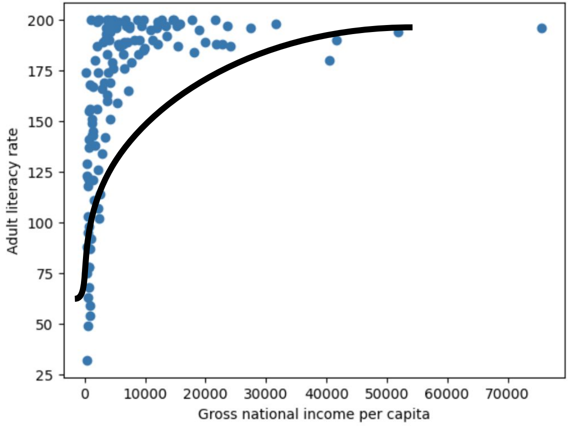
If the data bulges like this…



…transform y by this

…or transform x by this

Could transform y by y^2, y^3



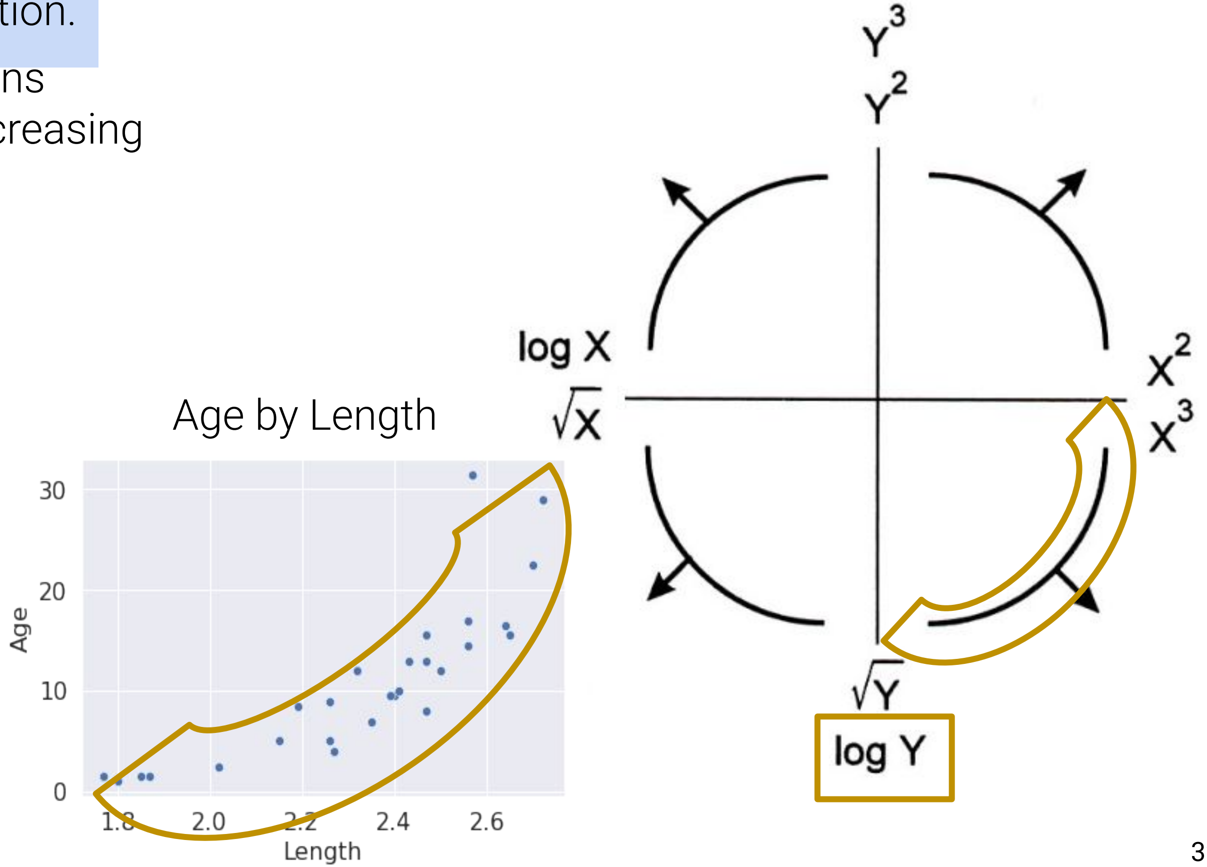
OR: Could transform x by log(x), sqrt(x)

Tukey-Mosteller Bulge Diagram

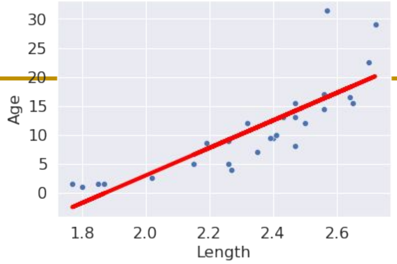
If your data “bulges” in a direction, transform x and/or y in that direction.

- Each of these transformations equates to increasing or decreasing the scale of an axis.
- Roots and logs make a value “smaller”.
- Raising to a power makes a value “bigger”.

There are multiple solutions! Some will fit better than others.



Transforming Dugongs



Suppose we do a log(y) transformation

Notice that the resulting model is still **linear in the parameters** $\theta = [\theta_0, \theta_1]$:

$$\widehat{log(y)} = \theta_0 + \theta_1 x$$

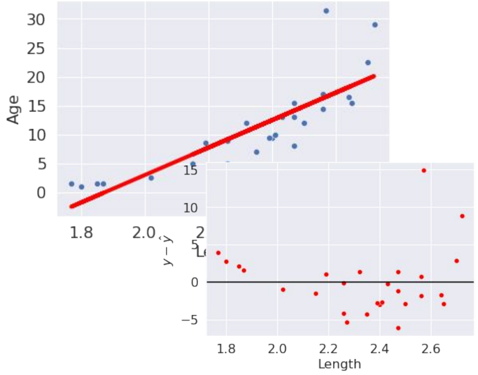
In other words, if we apply the variable transform $z = \log(y)$:

$$\hat{z} = \theta_0 + \theta_1 x$$

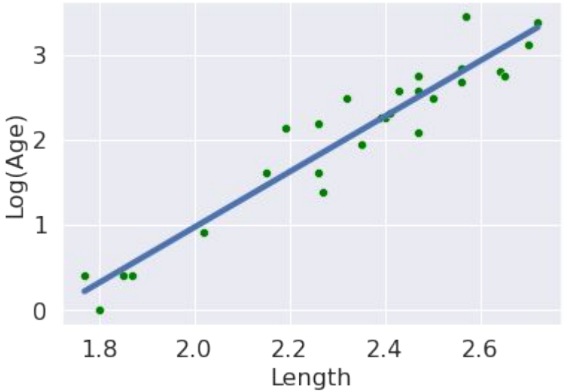
$$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)^2$$

$$\hat{\theta}_0 = \bar{z} - \hat{\theta}_1 \bar{x} \qquad \hat{\theta}_1 = r\frac{\sigma_z}{\sigma_x}$$

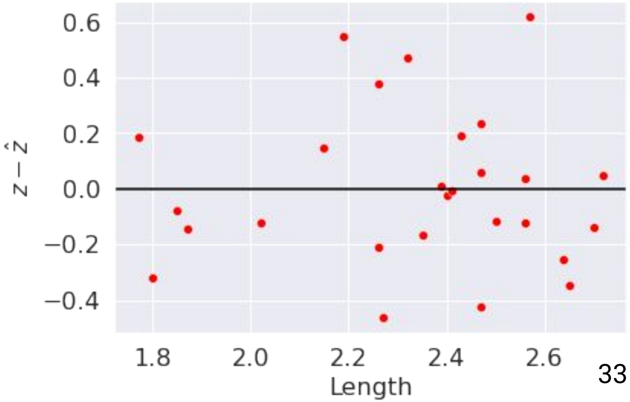
Original (Age by Length)



Log(Age) by Length



Residual Plot

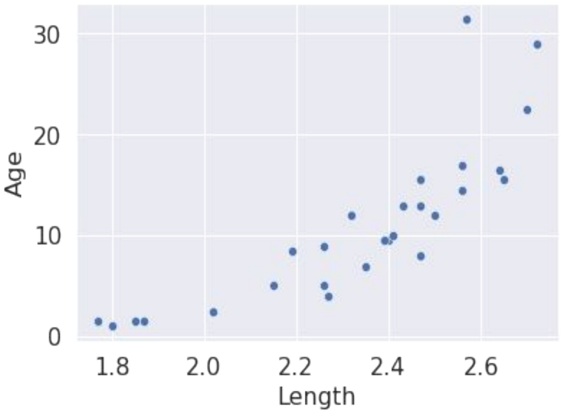


Fit a Curve using Least Squares Regression

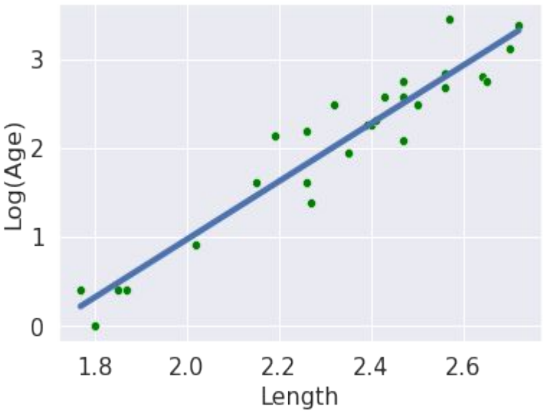
$$z = \log(y)$$

$$y = e^z$$

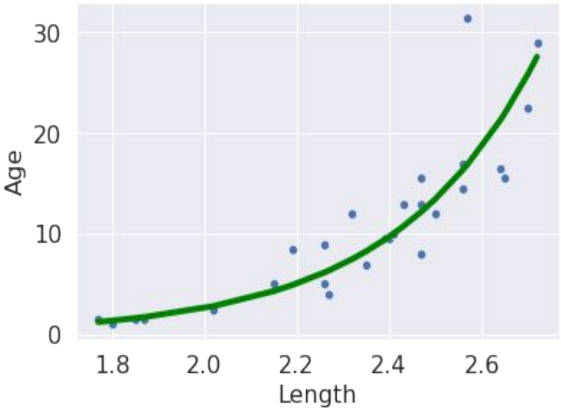
Age by Length



Log(Age) by Length



Age by Length

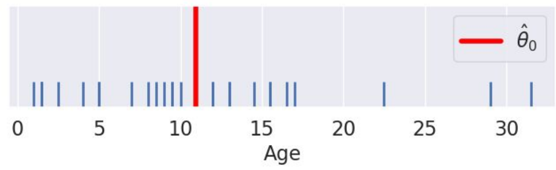


[Fit] Comparing Three Different Models, Both Fit with MSE

Constant Model

$$\hat{y} = \theta_0$$

RMSE: **7.72**



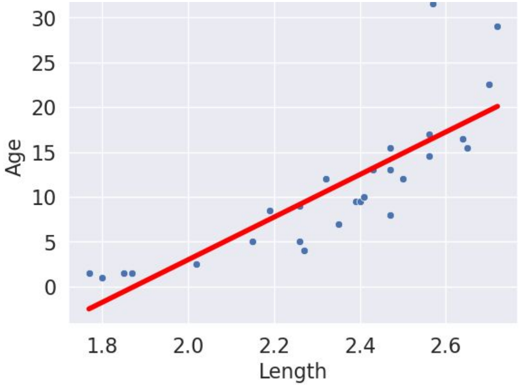
Compare

See notebook for code

Simple Linear Regression

$$\hat{y} = \theta_0 + \theta_1 x$$

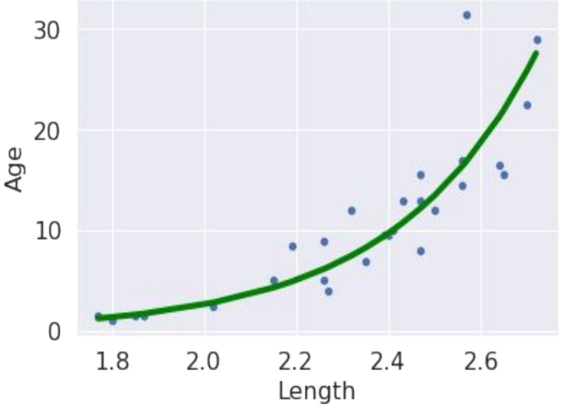
RMSE **4.31**



Log Transformation then Simple Linear Regression:

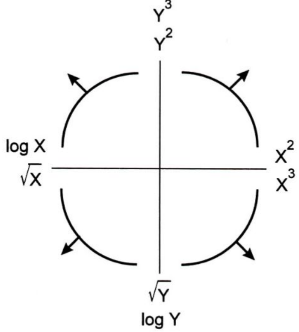
$$\hat{y} = e^{\theta_0 + \theta_1 x}$$

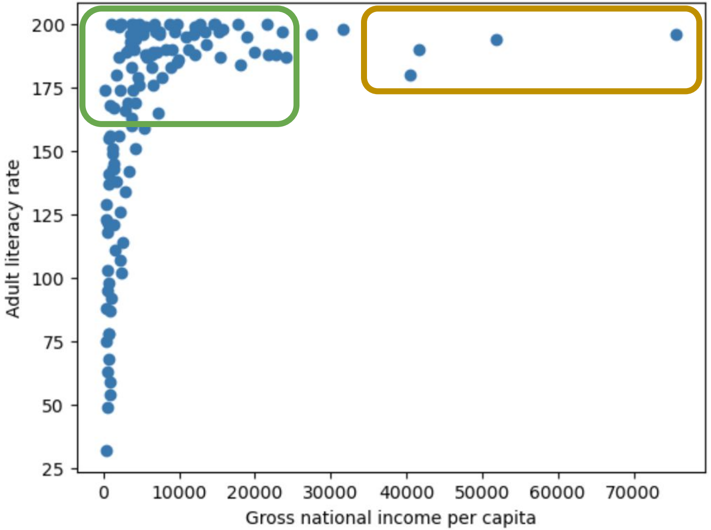
RMSE **3.75**



More Practice: Applying Transformations

What makes this plot non-linear?





1. A few large outlying x values are distorting the horizontal axis.

2. Many large y values are all clumped together, compressing the vertical axis.

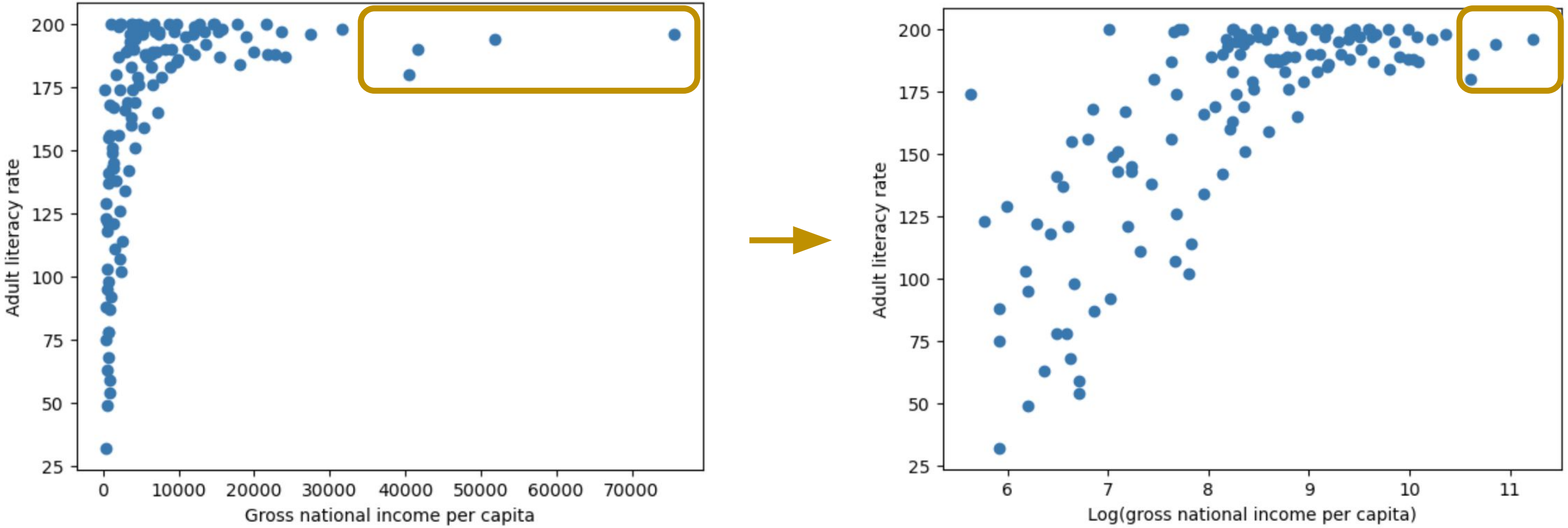
Applying Transformations

What makes this plot non-linear?

1. A few large outlying x values are distorting the horizontal axis.

Resolve by log-transforming the x data:

- Taking the log of a large number decreases its value significantly.
- Taking the log of a small number does not change its value as significantly.



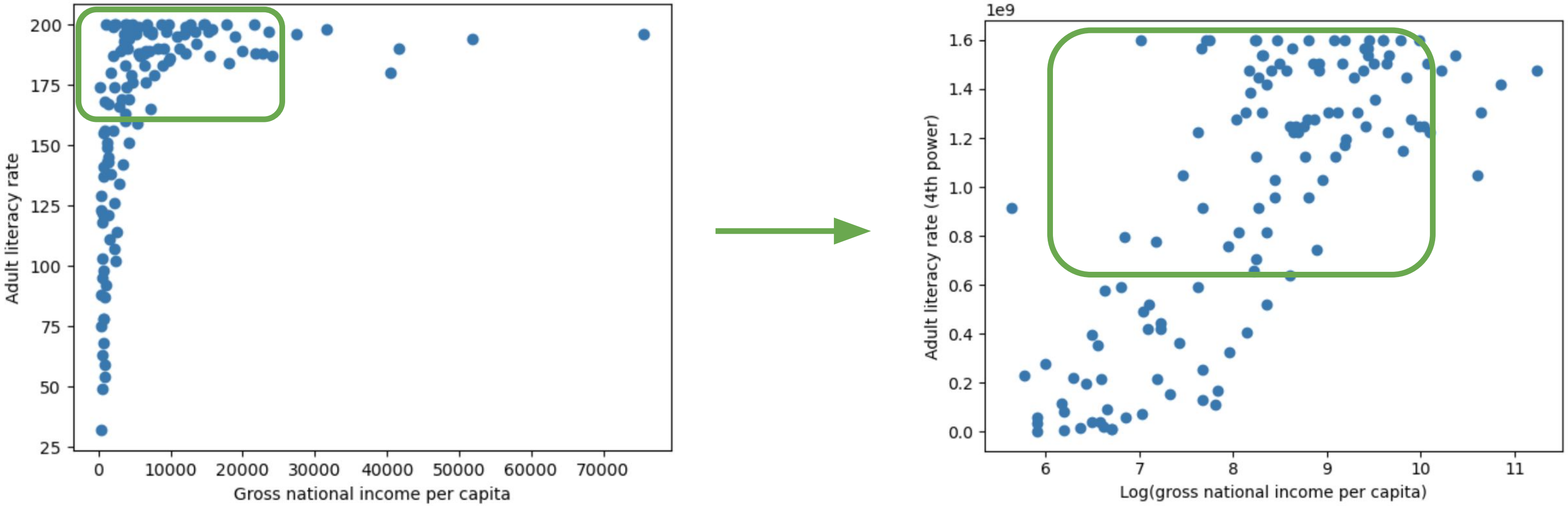
Applying Transformations

What makes this plot non-linear?

2. Many large y values are all clumped together, compressing the vertical axis.

Resolve by power-transforming the y data:

- Raising a large number to a power increases its value significantly.
- Raising a small number to a power does not change its value as significantly.



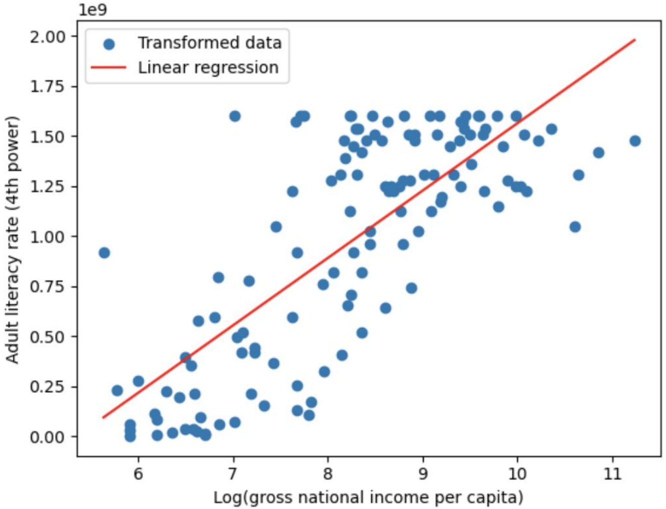
Interpreting Transformed Data

Now, we see a linear relationship between the transformed variables.

This tells us about the underlying relationship between the *original* x and y!

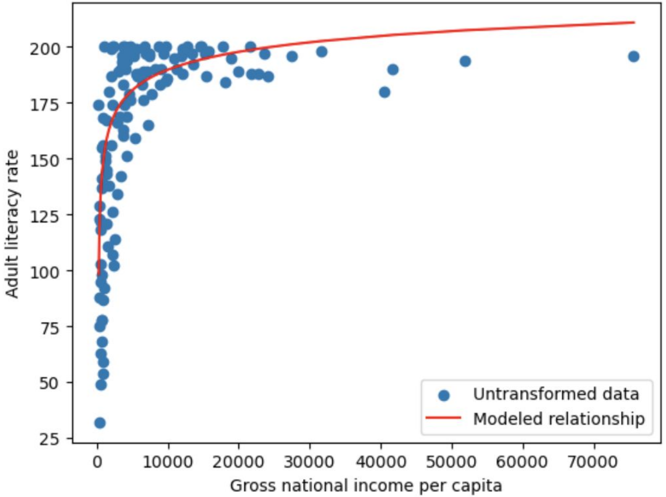
$$y^4 = m(\log x) + b \longrightarrow y = [m(\log x) + b]^{1/4}$$

y^4



$\log x$

y



x