Lesson 27

Simple Linear Regression

Understanding the usefulness of models and the simple linear regression model

CSCI 3022

Maribeth Oscamou

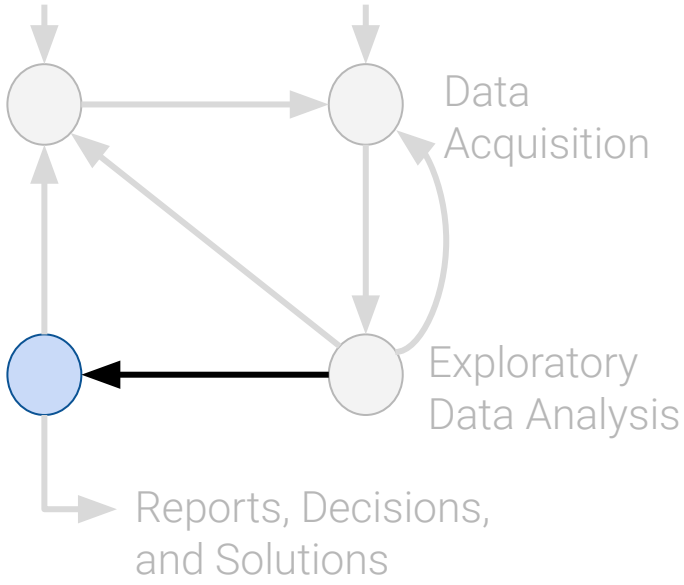
Content credit: Acknowledgments

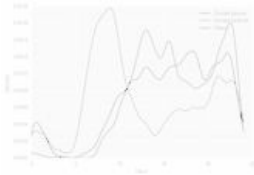
Announcements

- HW 10 Due Thursday: Corrections made Monday to otter grader for several questions. Please make sure you are using v2
- TA session on Lab 10 (Confidence Intervals and Climate Modeling) posted on Canvas
- Quiz 9 Friday:
  - Scope: HW 9,
    - HW 9, nb 9
    - Lesson 21 (A/B Testing, Permutation Tests & Causality)
    - Lesson 22 (Hypothesis Test Errors)

- Exam 2 next Friday
  - Scope: Cumulative. Questions will focus on concepts/topics from
    - Lessons 17-Lessons 25 (i.e. HW 7-10, nb 7-10)

Plan for Rest of Semester: Modeling





(today)

Modeling I: Different models, loss functions	→	Modeling II: Simple Linear Regression, linearization	→	Modeling III: Multiple Linear Regression

Today's Roadmap

- What is a Model?
- Relationships between Quantitative Variables
- Simple Linear Regression

The Modeling Process

1. Choose a model	How should we represent the world?
2. Choose a loss function	How do we quantify prediction error?
3. Fit the model	How do we choose the best parameters of our model given our data?
4. Evaluate model performance	How do we evaluate whether this process gave rise to a good model?

Models in CSCI 3022

In CSCI 3022, we'll treat a model as some mathematical rule to describe the relationships between variables.

Dataset

x	y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

Observation

(x_i, y_i)

- Independent variable
- **Input**
- **Feature**
- **Attribute**

- Dependent variable
- **Output**
- **Outcome**
- **Response**

Prediction

If we use x to predict y, the predictions are denoted as

$$\hat{y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$$

Models

Some models we will see in the next few lectures:

$$\hat{y}_i = \theta_0 + \theta_1 x_i$$

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

$$\hat{y}_i = x_i^\top \theta$$

→ Parametric models

Models in CSCI 3022: Parametric Models

Parametric models are described by a few **parameters** $(\theta_0, \theta_1, \text{etc.})$

- No one tells us the parameters: the data informs us about them.
- The x, y values are **not** parameters because we directly observe them.
- Sample-based **estimate** of parameter θ is written as $\hat{\theta}$
- Usually, we pick the parameters that appear "**best**" according to some criterion we choose.

Today, we'll be using **Simple Linear Models:**

Model parameter(s)

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

Any linear model with parameters $\theta = [\theta_0, \theta_1]$

Estimated parameter(s), "best" fit to data in some sense

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$$

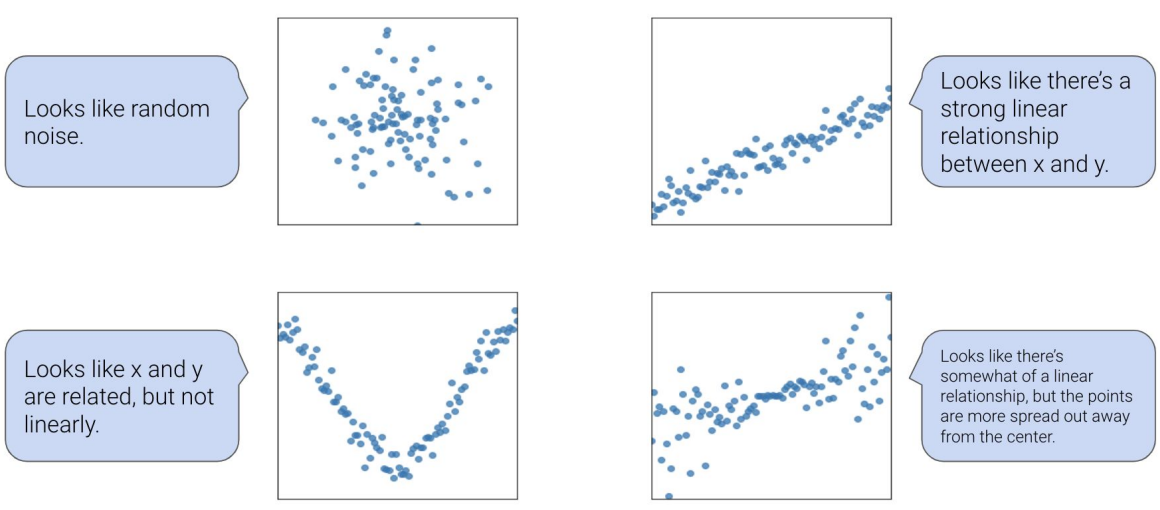
The "best" fitting linear model with parameters $\hat{\theta} = [\hat{\theta}_0, \hat{\theta}_1]$

Relationships between Quantitative Variables

- **Relationships between Quantitative Variables**

Exploring relationships between two variables

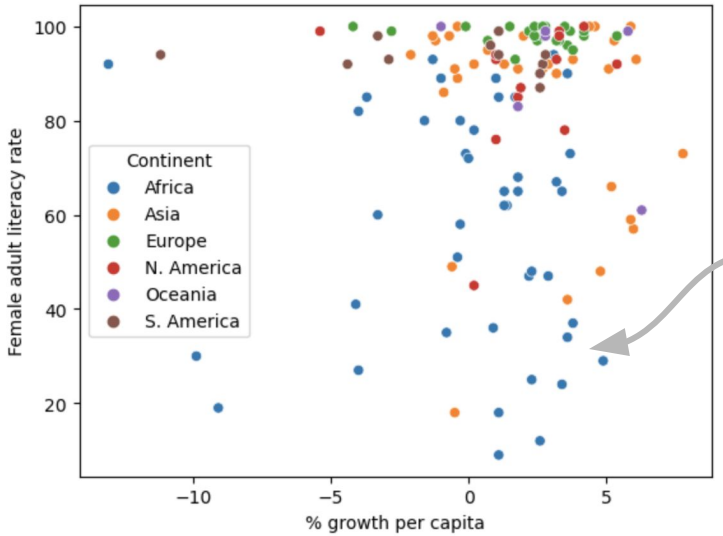
- The constant model we saw in the last lesson was only able to capture the distribution of a single variable. It was a **summary statistic**.
- More commonly, we create models that try to explain the relationships between **multiple variables** (which we will now denote with x and y).



Visualizing Relationships Between 2 Variables: Scatter Plots

Scatter plots are used to reveal relationships between two quantitative variables.

- Plot one quantitative continuous variable on the x-axis, and second quantitative continuous variable on the y-axis.
- Each scatter point represents one datapoint in the dataset.



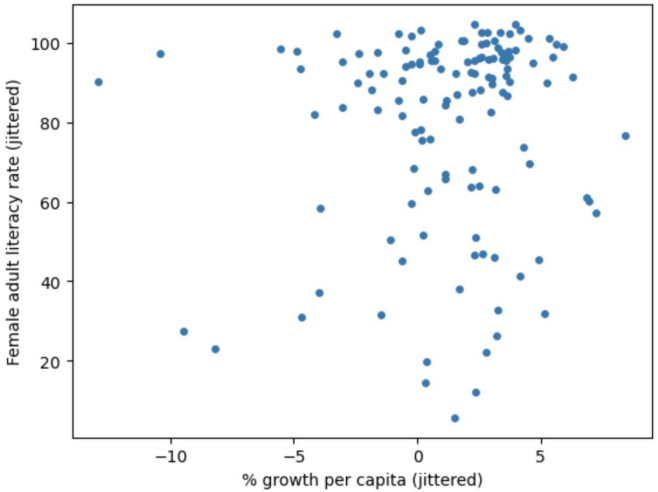

```
plt.scatter(x_values, y_values)
```

```
sns.scatterplot(data=df, x="x_column", \
                y="y_column", hue="hue_column")
```

Overplotting

The plot on the previous slide suffered from **overplotting** – scatter points all stacked on top of one another are difficult to see.

Jittering: adding a small amount of random noise to all x and y values to slightly move each scatter point. Main trends are still present, but individual datapoints are easier to distinguish.



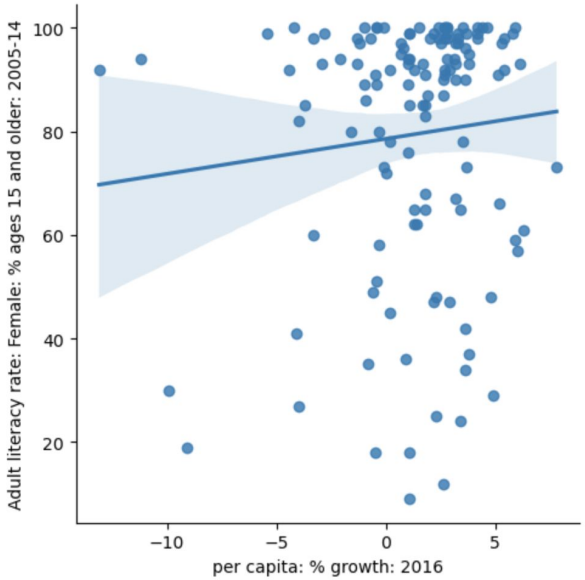

```
x_noise = np.random.uniform(-1, 1, len(wb))
y_noise = np.random.uniform(-5, 5, len(wb))

plt.scatter(wb['% growth'] + x_noise, \
            wb['Literacy rate: Female'] + y_noise, \
            s=15);
```

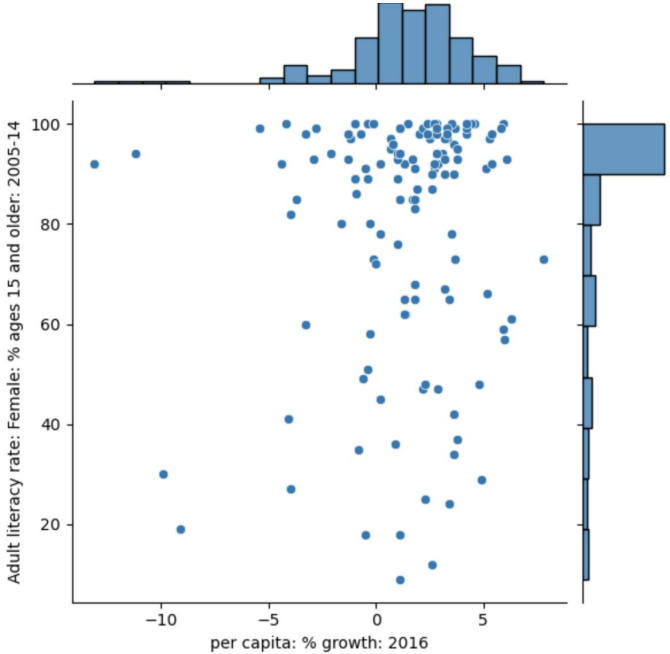
Decreasing point size also helps. `s` specifies the marker size in Matplotlib.

Scatter Plot Alternatives

Seaborn includes several built-in functions for making more complex scatter plots.




```
sns.lmplot(data=df, \
x="x_column", y="y_column")
```




```
sns.jointplot(data=df, \
x="x_column", y="y_column")
```

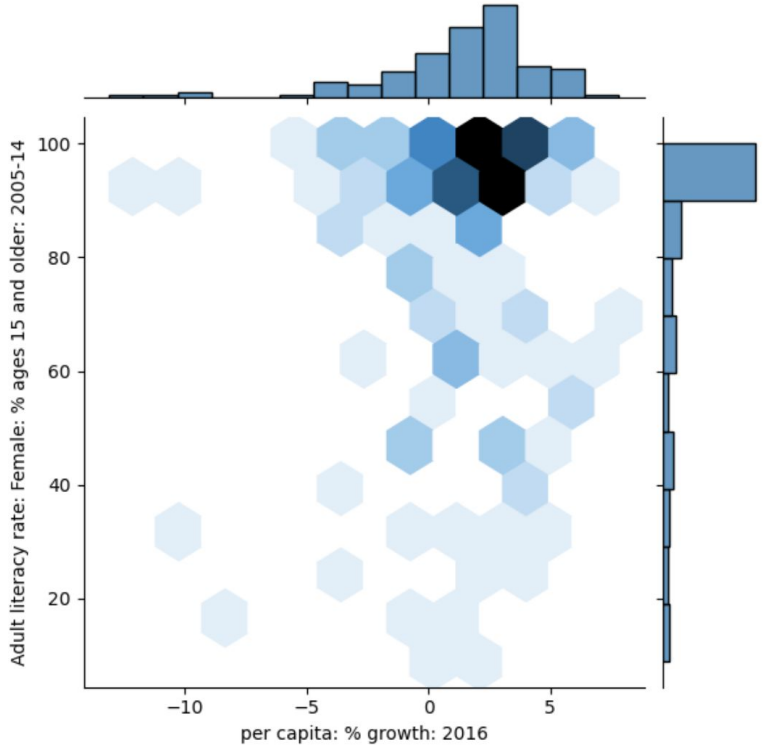
Hex Plots

Rather than plot individual datapoints, plot the *density* of their joint distribution.

Can be thought of as a two dimensional histogram.

- The xy plane is binned into hexagons.
- More shaded hexagons typically indicate a greater density/frequency = more datapoints lie in that spot

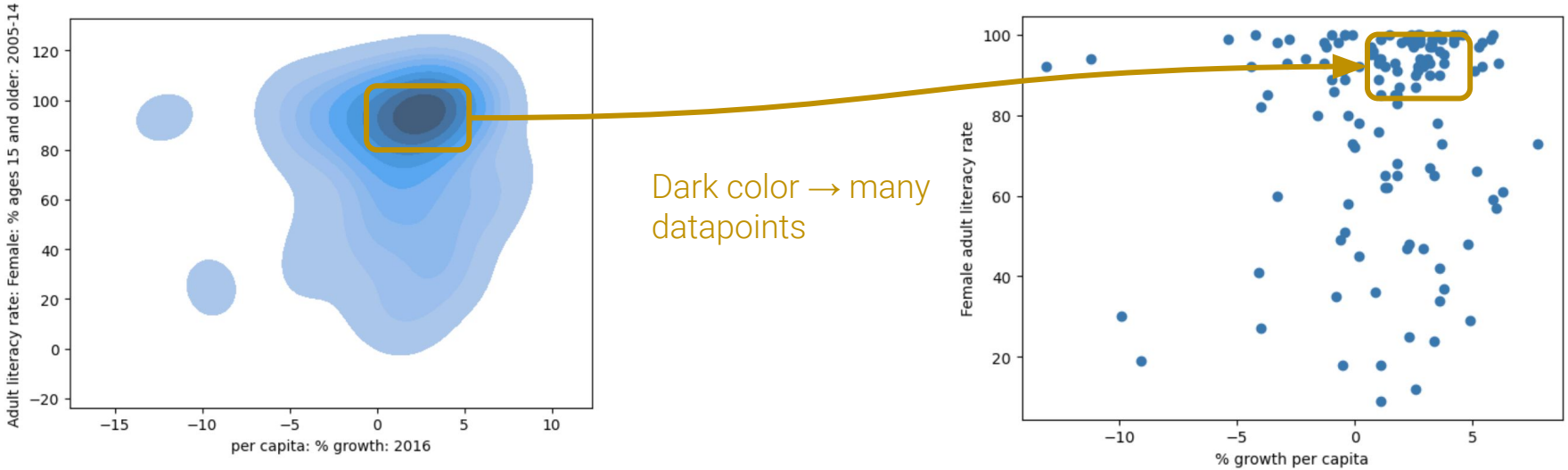
```
sns.jointplot(data=df, x="x_column", \
y="y_column", kind="hex")
```



Contour Plots

2-dimensional version of a KDE plot.

Similar to a topographic map – contour lines represent an area that has the *same density* of datapoints throughout. Darker colors indicate more datapoints in the region.




```
sns.kdeplot(data=df, x="x_column", y="y_column", fill=True)
```

Correlation: Recall Correlation of RV X and Y:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

What if we have a set of data instead?

Define the following:

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

means $\bar{x}, \bar{y}$ standard deviations σ_x, σ_y

The **correlation** r is the **average** of the **product** of x and y, both measured in standard units.

$$r = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right)$$

- x_i in standard units: $\frac{x_i - \bar{x}}{\sigma_x}$
- r is also known as Pearson's correlation coefficient.
- Side note: **covariance** is $\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = r \sigma_x \sigma_y$

Correlation

> The **correlation** r is the average of the product of x and y, both measured in standard units.

Define the following:

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

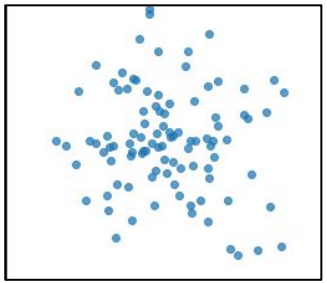
means $\bar{x}, \bar{y}$ standard deviations σ_x, σ_y

$$r = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right)$$

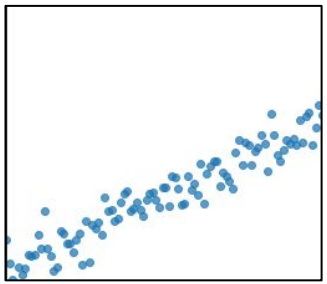
- x_i in standard units: $\frac{x_i - \bar{x}}{\sigma_x}$
- r is also known as Pearson's correlation coefficient.
- Side note: **covariance** is $\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = r \sigma_x \sigma_y$

Correlation measures the strength of a **linear association** between two variables.

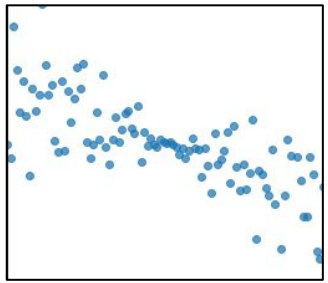
$$|r| \leq 1$$



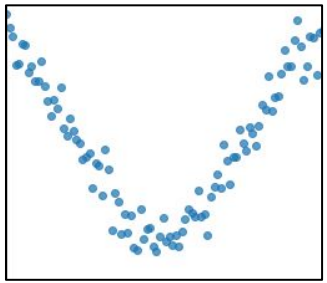
r = −0.121



r = 0.951



r = −0.723



⚠ **r = 0.056**

Simple Linear Regression

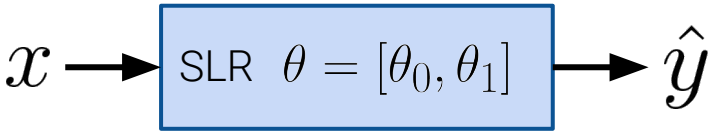
- **The Modeling Process**
  - **Choose a Model**
  - Choose a Loss Function
  - Fit the Model
  - Evaluate the Model

Simple Linear Regression Model (SLR)

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

SLR is a **parametric model**, meaning
we choose the "best" **parameters**
for slope and intercept based on data.

- We often express θ as a single parameter vector.
- x is **not** a parameter! It is input to our model.



- Note that the true relationship between x and y is usually non-linear.
  This is why $\hat{y}$ (and not y) appears in our **estimated** **linear model** expression.

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?

3. Fit the model

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

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$$

4. Evaluate model performance

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

Reflect

Loss Functions

- **The Modeling Process**
  - Choose a Model
  - **Choose a Loss Function**
  - Fit the Model
  - Evaluate the Model

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?

3. Fit the model

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

4. Evaluate model performance

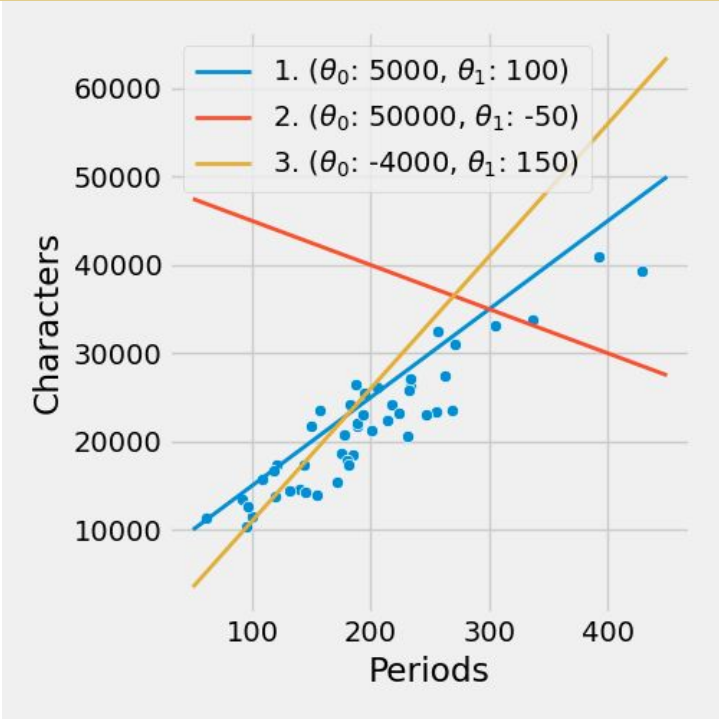
How do we evaluate whether this process gave rise to a good model?

Which 〰 is best?

Based on your interpretation of the data, which are the "optimal parameters" for this linear model?

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

$$\hat{\theta}_0 = ? \quad \hat{\theta}_1 = ?$$



Which of these lines matches the data better?

A). Blue B) Red C) Yellow

We only had 3 values to choose from to find the optimal parameter. In practice, our parameter domain is all reals, i.e., $\theta = [\theta_0, \theta_1] \in \mathbb{R}^2$

For every chapter of the novel *Little Women*, Estimate the **# of characters** $\hat{y}$ based on the **# of periods** x in that chapter.

L2 and L1 Loss for SLR

1357201

Squared Loss (L2 Loss)

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

For an SLR model $\hat{y} = \theta_0 + \theta_1 x$

$$L(y, \hat{y}) = (y - (\theta_0 + \theta_1 x))^2$$

Absolute Loss (L1 Loss)

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

For an SLR model $\hat{y} = \theta_0 + \theta_1 x$

$$L(y, \hat{y}) = |y - (\theta_0 + \theta_1 x)|$$

Why don't we directly use residual error as the loss function? $e = (y - \hat{y})$

- Doesn't work: big negative residuals shouldn't cancel out big positive residuals!

Which loss function is better: L1 or L2?

L2 penalizes larger residuals more.

Empirical Risk is Average Loss over Data

We care about how bad our model's predictions are for our entire data set, not just for one point.

A natural measure, then, is of the **average loss** (aka **empirical risk**) across all points.

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

$$\hat{R}(\theta) = \frac{1}{n} \sum_i L(y_i, \hat{y}_i)$$

The colloquial term for average loss depends on which loss function we choose.

L2 loss → **Mean Squared Error (MSE)**

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

L1 loss → **Mean Absolute Error (MAE)**

$$\frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

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?

The combination of model + loss that we focus on today is known as **least squares regression**.

Fit the Model

- **The Modeling Process**
  - Choose a Model: SLR
  - Choose a Loss Function
  - **Fit the Model**
  - Evaluate the Model

Day 2 Announcements

- Exam 2 next Friday
  - Scope: Cumulative. Questions will focus on concepts/topics from
    - Lessons 17-Lessons 25 (i.e. HW 7-10, nb 7-10)
- In Class Review on Wednesday. Post questions on Google Doc linked on Piazza

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?

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

We want to find $\hat{\theta}_0, \hat{\theta}_1$ that minimize this **objective function**.

4. Evaluate model performance

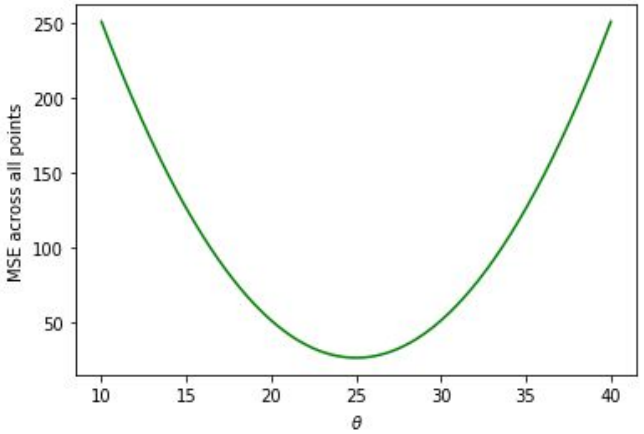
How do we evaluate whether this process gave rise to a good model?

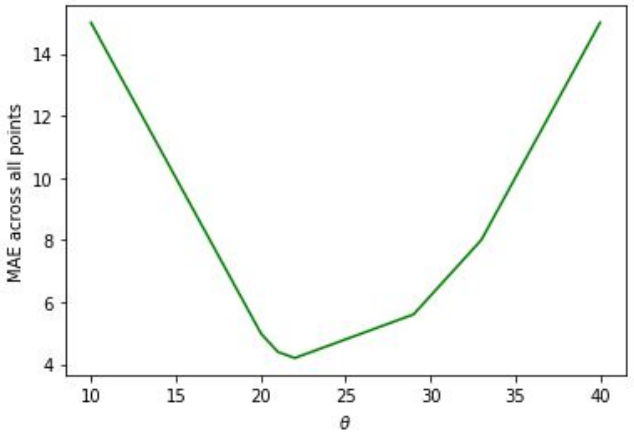
Visualizing loss surfaces

On the right, we have the plots of the **loss surfaces** for the constant model (from last lecture).

- Top: squared loss (so average loss = MSE).
  - The y-axis shows the MSE for each value of theta on the x-axis.
- Bottom: absolute loss (so average loss = MAE).

The simple linear regression model has two parameters, a and b (or equivalently, θ_0 and θ_1). This means the loss surface will be **3D!**





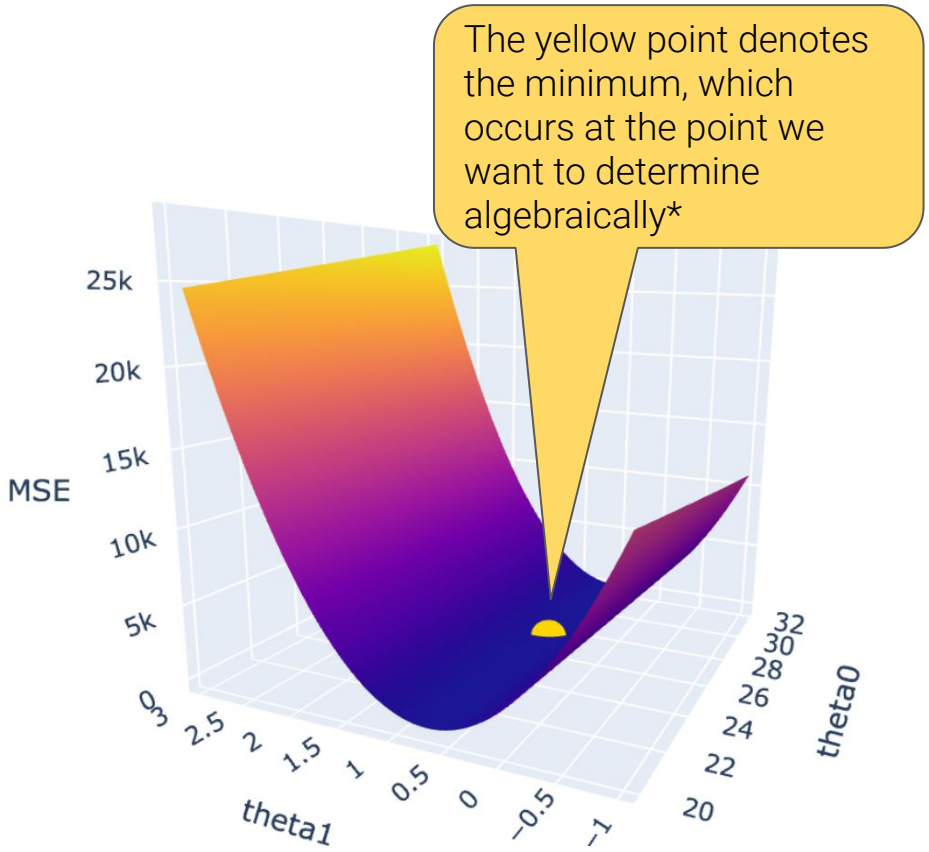
Visualizing loss surfaces: SLR

Here, we have 3 axes.

- One for θ_0.
- One for θ_1.
- One that tells us the mean squared error on our dataset, using the model $\hat{y} = \theta_0 + \theta_1 x$.

The loss surface is nice and smooth

Let's look at a **demo** of this in code.



Step 1 of 2: Fix θ_1 and minimize with respect to θ_0

Derivation video: link

1. Rewrite the function:

$$\sum_{i=1}^{n}(y_i - \hat{\theta}_0 + \theta_1 x_i))^2 = \sum_{i=1}^{n}(y_i - \theta_1 x_i - \theta_0))^2$$

2. Differentiate with respect to θ_0:

$$\frac{\partial}{\partial \theta_0}\sum_{i=1}^{n}(y_i - \theta_1 x_i - \theta_0)^2 = \sum_{i=1}^{n}\frac{\partial}{\partial a}(y_i - \theta_1 x_i - \theta_0)^2$$

Derivative of sum is sum of derivatives

$$= \sum_{i=1}^{n} 2(y_i - \theta_1 x_i - \theta_0)(-1)$$

Chain rule

$$= -2\sum_{i=1}^{n}(y_i - \theta_1 x_i - \theta_0)$$

Simplify constants

3. Set equal to 0:

$$0 = -2\sum_{i=1}^{n}(y_i - \theta_1 x_i - \theta_0)$$

4. Finally, rearrange and solve for $\hat{\theta}_0$:

$$0 = -2\sum_{i=1}^{n}(y_i - \theta_1 x_i - \theta_0)$$

$$= \sum_{i=1}^{n}(y_i - \theta_1 x_i - \theta_0)$$

$$= \sum_{i=1}^{n} y_i - \theta_1\sum_{i=1}^{n} x_i - n\theta_0$$

Pull out scalars

$$n\theta_0 = \sum_{i=1}^{n} y_i - \theta_1\sum_{i=1}^{n} x_i$$

$$\hat{\theta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \theta_1 \cdot \frac{1}{n}\sum_{i=1}^{n} x_i$$

Divide by n

$$\hat{\theta}_0 = \bar{y} - \theta_1\bar{x}$$

Step 2 of 2: Plug in $\hat{\theta}_0$ and minimize with respect to θ_1

Our expression for $\hat{\theta}_0$: $\theta_0 = \bar{y} - \theta_1 \bar{x}$

1. Plug in $\hat{\theta}_0$ to our objective function:

$$\sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1 x_i))^2 = \sum_{i=1}^{n}(y_i - (\bar{y} - \theta_1 \bar{x} + \theta_1 x_i))^2$$

$$= \sum_{i=1}^{n}(y_i - \bar{y} - \theta_1(x_i - \bar{x}))^2$$

2. Differentiate with respect to θ_1

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta_1}(y_i - \bar{y} - \theta_1(x_i - \bar{x}))^2$$

Derivative of sum is sum of derivatives

$$= \sum_{i=1}^{n} 2 \cdot (y_i - \bar{y} - \theta_1(x_i - \bar{x})) \cdot (-1) \cdot (x_i - \bar{x})$$

Chain rule

$$= -2 \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y} - \theta_1(x_i - \bar{x}))$$

Simplify constants

3. Set equal to 0:

$$0 = -2 \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y} - \theta_1(x_i - \bar{x}))$$

4. Finally, rearrange and solve for $\hat{b}$:

$$0 = \left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - \theta_1(x_i - \bar{x})^2\right]$$

Distribu-tive prop.

$$= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - \theta_1 \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Separate sums

Add in constants

$$= n\sigma_x\sigma_y \left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)\right) - \theta_1 n \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= nr\sigma_x\sigma_y - \theta_1 n\sigma_x^2$$

Definitions of r, σ_x

$$\theta_1 \sigma_x^2 = r\sigma_x\sigma_y$$

$$\hat{\theta}_1 = \frac{r\sigma_x\sigma_y}{\sigma_x^2} = r\frac{\sigma_y}{\sigma_x}$$

$$\boxed{\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x}}$$

The Regression Line

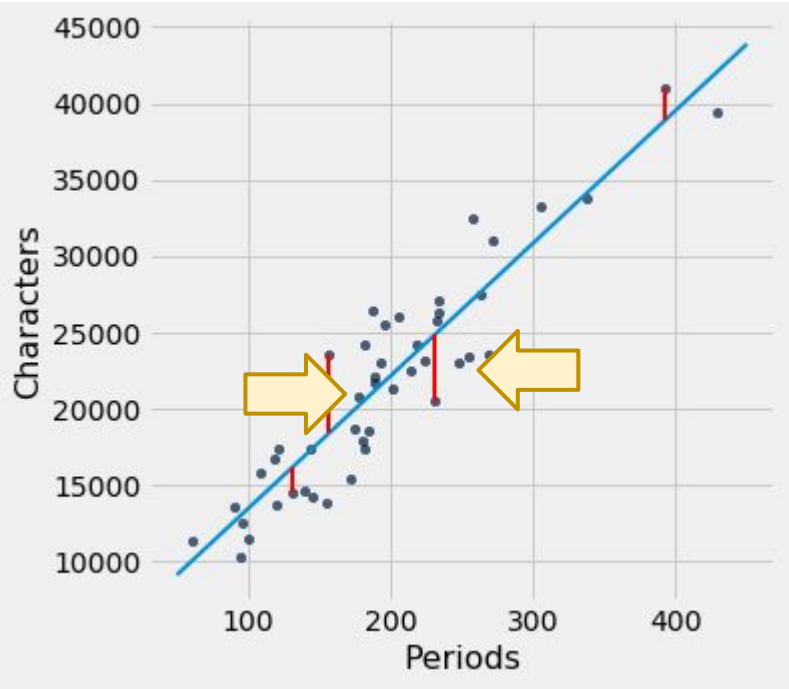
> The **least squares linear regression model** is the unique straight line that minimizes the **mean squared error** of estimation among all straight lines.

$$\text{slope} = r \times \frac{\text{SD of } y}{\text{SD of } x}$$

$$\text{intercept} = \text{average of } y - \text{slope} \times \text{average of } x$$

$$\text{regression estimate} = \text{intercept} + \text{slope} \times x$$

$$\boxed{\text{residual}} = \text{observed } y - \text{regression estimate}$$



For every chapter of the novel *Little Women*, Estimate the **# of characters** $\hat{y}$ based on the **number of periods** x in that chapter.

The Regression Line

The **least squares linear regression model** is the unique straight line that minimizes the **mean squared error** of estimation among all straight line

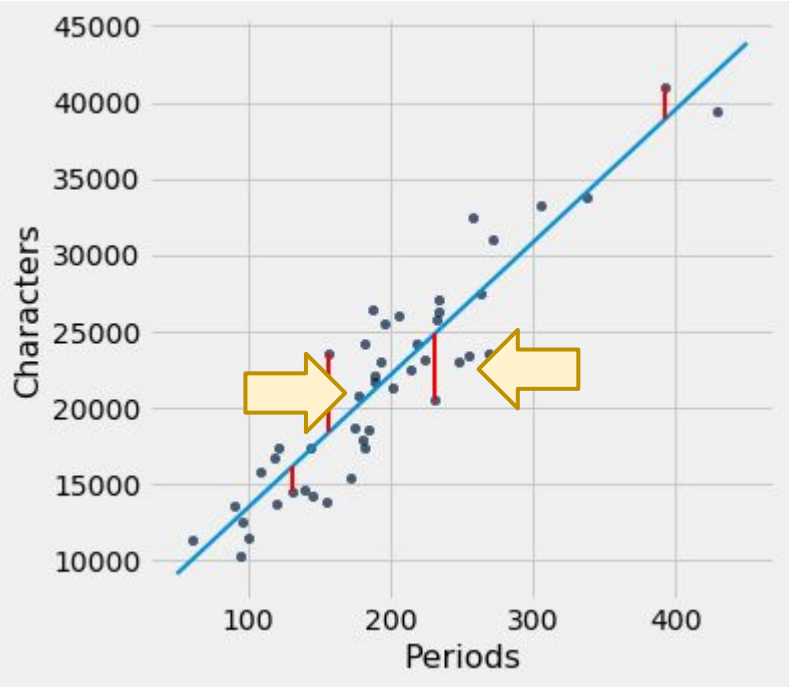
correlation

$$\text{slope} = r \times \frac{\text{SD of } y}{\text{SD of } x}$$

$$\text{intercept} = \text{average of } y - \text{slope} \times \text{average of } x$$

$$\text{regression estimate} = \text{intercept} + \text{slope} \times x$$

$$\text{residual} = \text{observed } y - \text{regression estimate}$$



For every chapter of the novel *Little Women*, Estimate the **# of characters** $\hat{y}$ based on the **number of periods** x in that chapter.

Expressing the Regression Line Mathematically

The **least squares linear regression model** is the unique straight line that minimizes the **mean squared error** of estimation among all straight lines.

Define the following:

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

$\bar{x}, \bar{y}$ means; σ_x, σ_y standard deviations;

r correlation coefficient

$$\hat{y} = \theta_0 + \theta_1 x$$ **regression line**

1. slope $= r \cdot \dfrac{\text{SD of } y}{\text{SD of } x}$

2. intercept $=$ average of y $-$ slope $\cdot$ average of x

3. residual $=$ observed value $-$ regression estimate

?

Rewrite each expression using math notation.

Expressing the Regression Line Mathematically

> The **regression line** is the unique straight line that minimizes the **mean squared error** of estimation among all straight lines.

Define the following:

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

$\bar{x}, \bar{y}$ means; σ_x, σ_y standard deviations;

r correlation coefficient

$$\hat{y} = \theta_0 + \theta_1 x \quad \textbf{regression line}$$

1. slope $= r \cdot \dfrac{\text{SD of } y}{\text{SD of } x}$

2. intercept $=$ average of y $-$ slope $\cdot$ average of x

3. residual $=$ observed value $-$ regression estimate

?

$$\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x}$$

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$$

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

Error for the i-th datapoint

The Regression Line in Standard Units

Recall the least squares linear regression model is defined as:

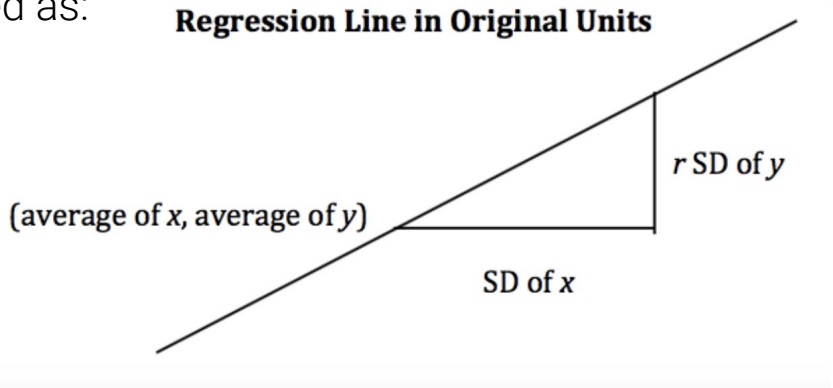
$$\hat{y} = \left(\frac{r\sigma_y}{\sigma_x}\right) \times x + \left(\bar{y} - \frac{r\sigma_y}{\sigma_x}\bar{x}\right)$$

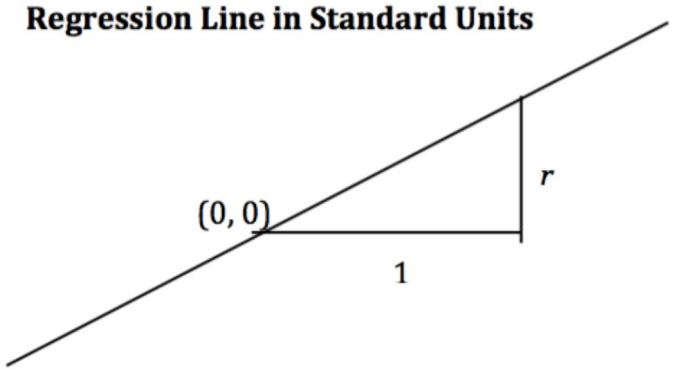
$$\hat{y} = \sigma_y \times r \times \frac{x - \bar{x}}{\sigma_x} + \bar{y}$$

$$\frac{\hat{y} - \bar{y}}{\sigma_y} = r \times \frac{x - \bar{x}}{\sigma_x}$$

When the variables x and y are measured in standard units, the regression line for predicting y based on x has slope r passes through the origin and the equation will be:

$$\hat{y} = r \times x \ [both\ measured\ in\ standard\ units]$$





Discussion Question

A course has a midterm (average 70; standard deviation 10) and a really hard final (average 50; standard deviation 12)

If the scatter diagram comparing midterm & final scores for students has an oval shape with correlation 0.75, then...

a). What do you expect the average final score would be for students who scored 90 on the midterm?

Poll:

A). 65 B). 85 C). 70 D). 68 E). none

b). What about 60 on the midterm?

Interpreting SLR: Slope

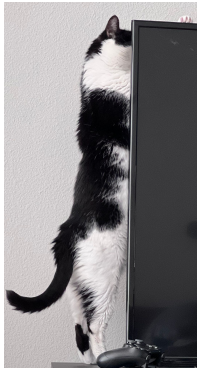
Interpreting the Least Squares Linear Regression Model

You may sometimes hear the prediction task defined as: "**regressing** y on x."

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

Suppose we fit a model that predicts a cat's weight (in pounds) given its length (in inches).

```
predicted weight = 2 + 0.5 * length
```



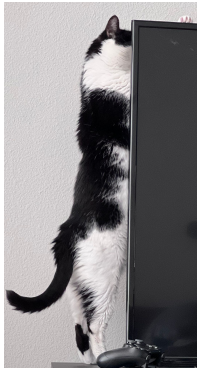
Interpreting the Least Squares Linear Regression Model

You may sometimes hear the prediction task defined as: "**regressing** y on x."

Suppose we fit a model that predicts a Chihuahua's weight (in pounds) given its length (in inches).

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

```
predicted weight = 2 + 0.5 * length
```



Interpreting the slope?

By definition, the slope measures the increase in y (pounds) for a 1 unit increase in x (1 inch).

1. Does this mean that if a cat in the dataset grows 1 inch, we estimate that they will get 0.5 pounds heavier? What does it actually mean?

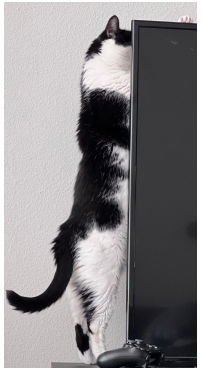
Interpreting the Least Squares Linear Regression Model

You may sometimes hear the prediction task defined as: "**regressing** y on x."

Suppose we fit a model that predicts a Chihuahua's weight (in pounds) given its length (in inches).

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

```
predicted weight = 2 + 0.5 * length
```



Interpreting the slope?

By definition, the slope measures the increase in y (pounds) for a 1 unit increase in x (1 inch).

1. Does this mean that if a cat in the dataset grows 1 inch, we estimate that they will get 0.5 pounds heavier? What does it actually mean? **No!**

- The model we created shows **association**, not causation.
- The data we collected is a snapshot of several cats at one instance of time (**cross-sectional**), not snapshots of cats over time (**longitudinal**).

Slope interpretation: If two cats have a 1 inch height difference, their estimated weight difference is 0.5 lbs.

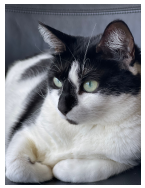
Interpreting the Least Squares Linear Regression Model

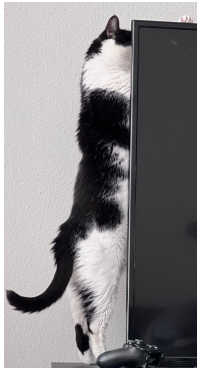
You may sometimes hear the prediction task defined as: "**regressing** y on x."

Suppose we fit a model that predicts a Chihuahua's weight (in pounds) given its length (in inches).

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

```
predicted weight = 2 + 0.5 * length
```





Interpreting the slope?

By definition, the slope measures the increase in y (pounds) for a 1 unit increase in x (1 inch).

1. Does this mean that if a cat in the dataset grows 1 inch, we estimate that they will get 0.5 pounds heavier? What does it actually mean? **No!**

Predicting on wildly different data?

Domestic shorthair range from 8-10 pounds, and 13-16 inches in length.
Maine Coon range from 10-25 pounds, and 19-40 inches in length.

2. Should we use this model to predict the weight of all cat breeds? **No!**

Lesson 27 - Day 3:

Announcements

- Exam 2 Friday
  - Scope: Cumulative. Questions will focus on concepts/topics from
    - Lessons 17-Lessons 25 (i.e. HW 7-10, nb 7-10)
    - Bring:
      - Buff One Card
      - Crib Sheet (2 sided)
      - Calculator (optional)
- In Class Review on Wednesday. Post questions on Google Doc linked on Piazza

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?**

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

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 column means, standard deviation.
   If we want to fit a linear model, compute correlation r.

2. **Performance metrics:**

   Root Mean Square Error (RMSE)

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

   - 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 y values. Should look like an unassociated blob for linear relations.
- Look at a plot of predicted vs actual y values.

Residuals

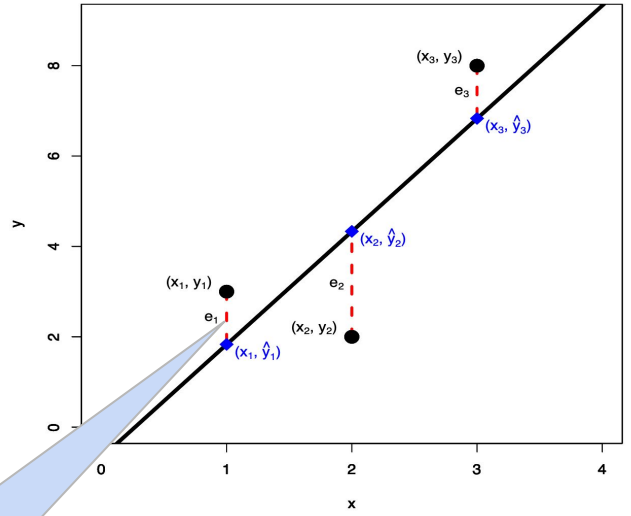
Residuals are defined as being the difference between an actual and predicted value, in the regression context.

- We use the letter e to denote residuals. The residual i is

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

- The MSE of a model is equal to the mean of the squares of its residuals:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$



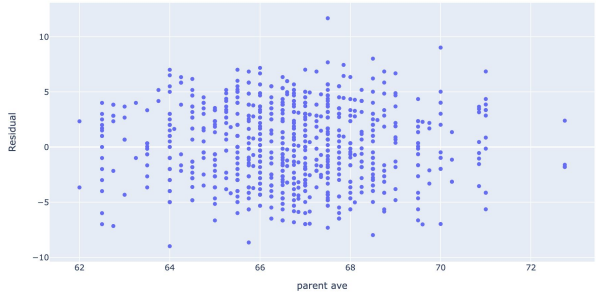
The red dotted lines represent residuals.

Residual Plot

Residuals are defined as being the difference between an actual and predicted value, in the regression context.

- We use the letter e to denote residuals. The residual i is

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

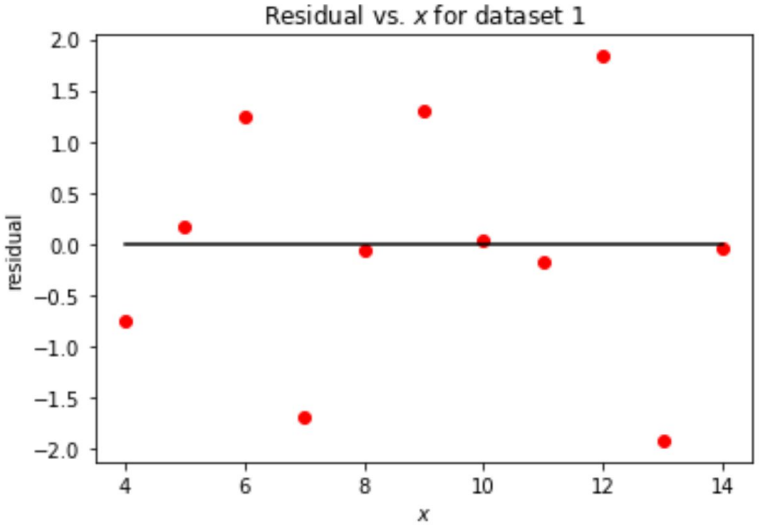


Residual plots can tell us about the quality of our model.

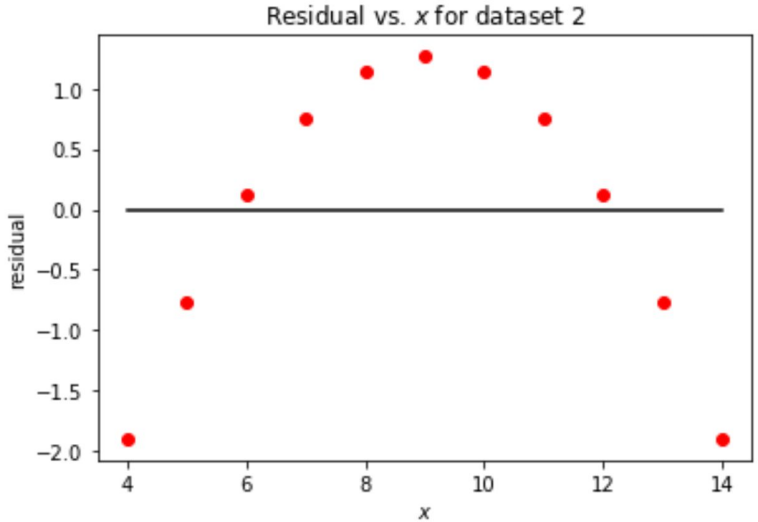
- In the **simple linear regression** case, with only one independent variable, we typically plot residuals vs. x.
- More generally (for multiple linear regression) a residual plot is of **residuals vs. fitted values**.

Properties:

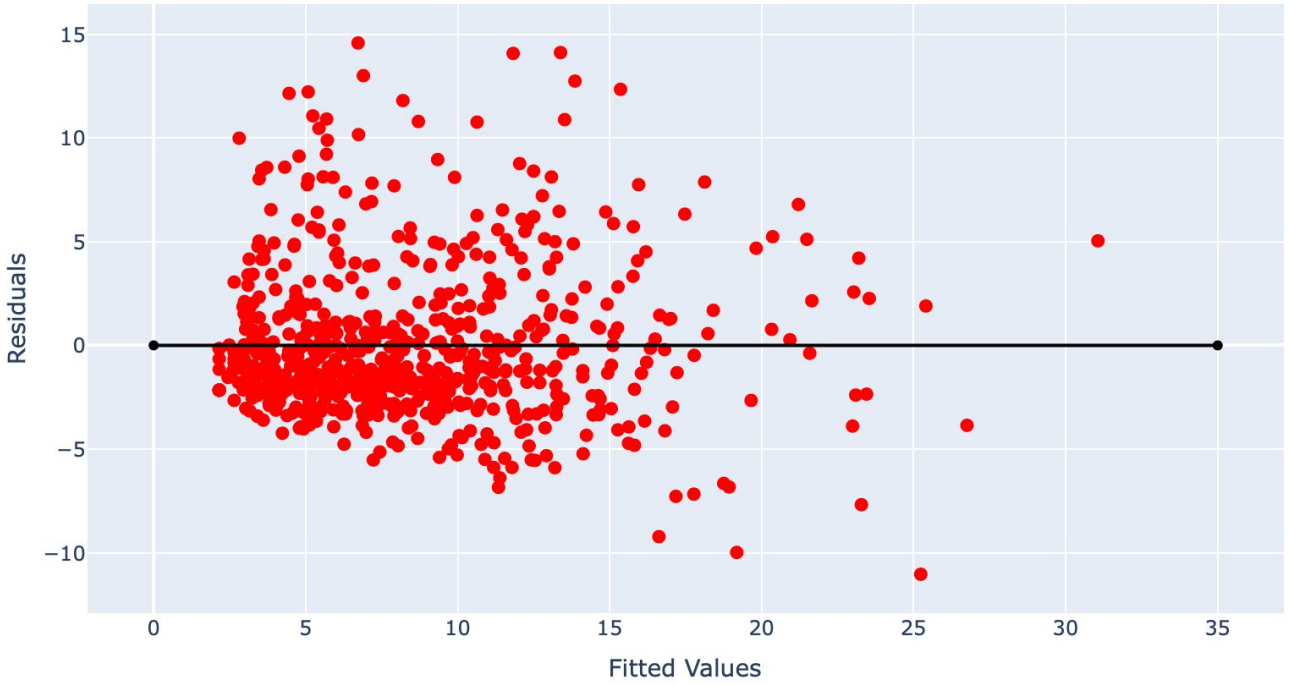
- A good residual plot has no pattern (Should look like an unassociated blob). This means that our model represents the relationship in the data well.
  - Look for curves, trends, changes in spread, outliers, or any other patterns - it is a sign that transformations or additional variables could help.
- A good residual plot also has a similar vertical spread throughout the entire plot.
  - If this is not the case, the accuracy of the predictions is not reliable.



No pattern, even spread.



Clear quadratic relationship in the residuals.



No clear relationship, but uneven spread.

Properties when our model has an intercept term

For all linear models with an **intercept term**, the **sum of residuals is zero**.

$$\sum_{i=1}^{n} e_i = 0$$

- This is the real reason why we don't directly use residuals as loss.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$$

- This is also why positive and negative residuals will cancel out in any residual plot where the (linear) model contains an intercept term, even if the model is terrible.

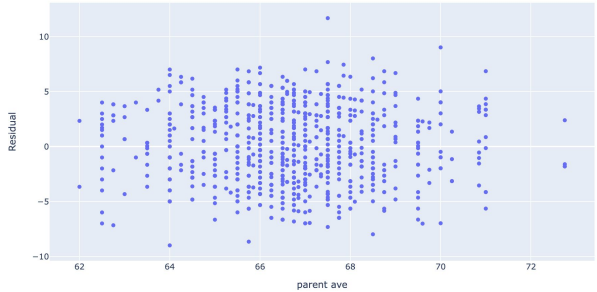
It follows from the property above that for linear models with intercepts, the average predicted y value is equal to the average true y value.

$$\bar{y} = \overline{\hat{y}}$$

These properties are true when there is an intercept term, and not necessarily when there isn't.

Properties of residuals

- Residuals from a linear regression (with an intercept) always have
  - Zero mean
    - (so rmse = SD of residuals)
  - Zero correlation with x
  - Zero correlation with the fitted values



- These are all true no matter what the data look like
  - Just like deviations from mean are zero on average

Does a unique solution always exist?

	Model	Estimate	Unique?
Constant Model + MSE		$\hat{\theta} = \textbf{mean}(y)$	**Yes**. Any set of values has a unique mean.
Constant Model + MAE		$\hat{\theta} = \textbf{median}(y)$	**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.

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!

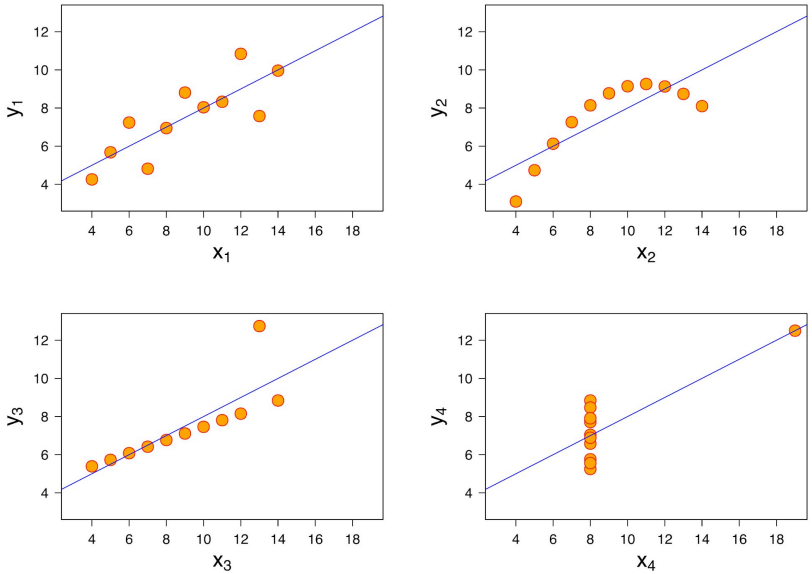
Anscombe's quartet refers to the following four sets of points on the right.

- 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!

Visualize, Then Quantify!

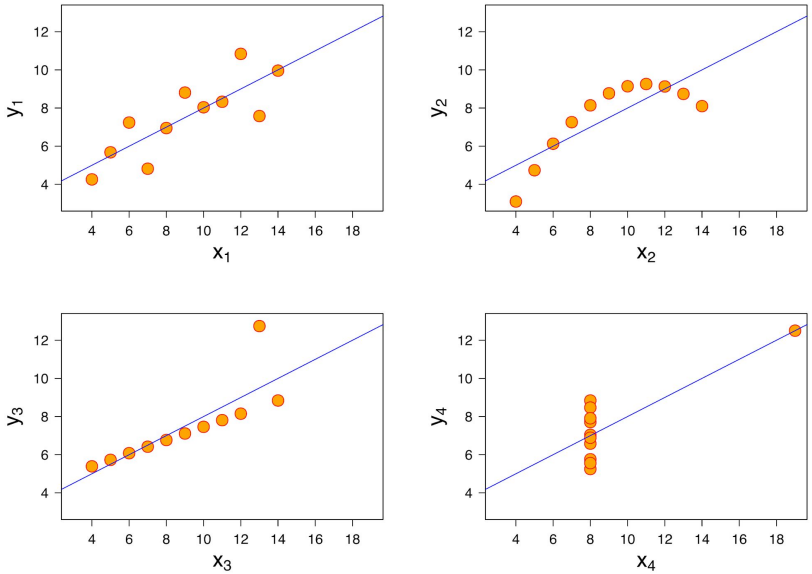
Anscombe's quartet refers to the following four sets of points on the right.

- 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!

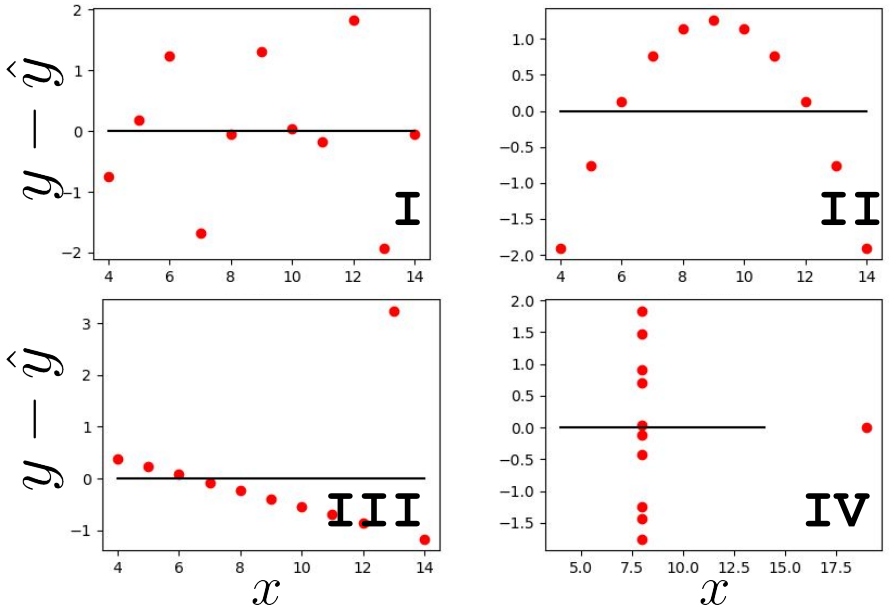
Four Mysterious Datasets + Least Squares

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

/ˈdo͞o ˌgäNG,ˈdo͞o ˌgôNG/

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.

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

Compare

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

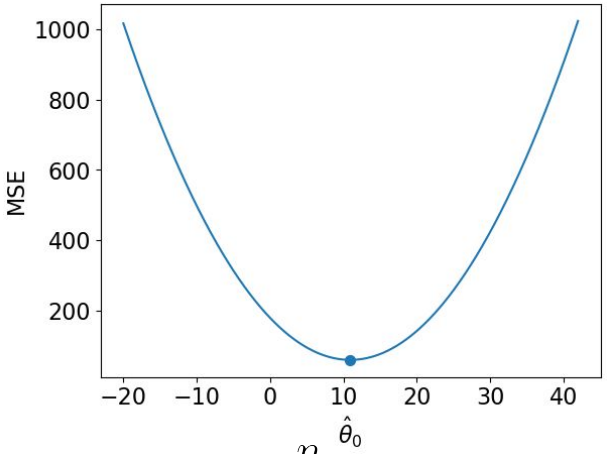
[Loss] Comparing Two Different Models, Both Fit with MSE

Constant Model

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

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

Loss surface is **2-D**.



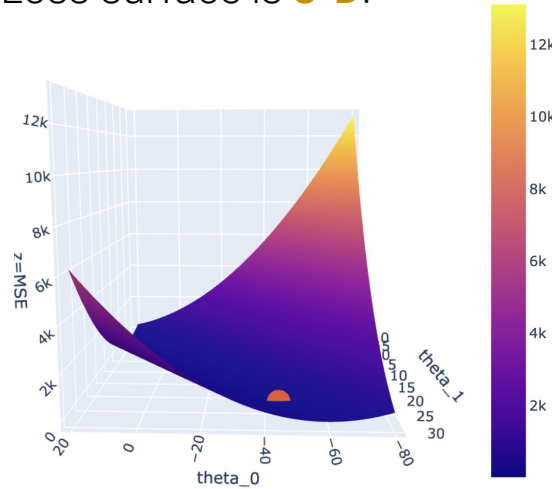
$$\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_0)^2$$

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**.



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

[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**

Compare

See notebook for code

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 terms ** (if you are using the same data for both models).

Compare

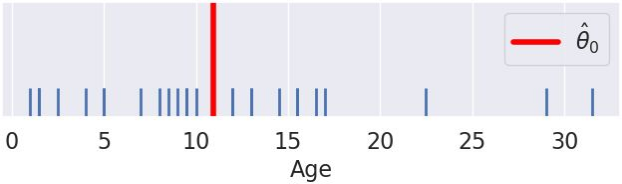
[Fit] Comparing Two Different Models, Both Fit with MSE

Constant Model

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

RMSE: **7.72**

Predictions on a **rug plot**.

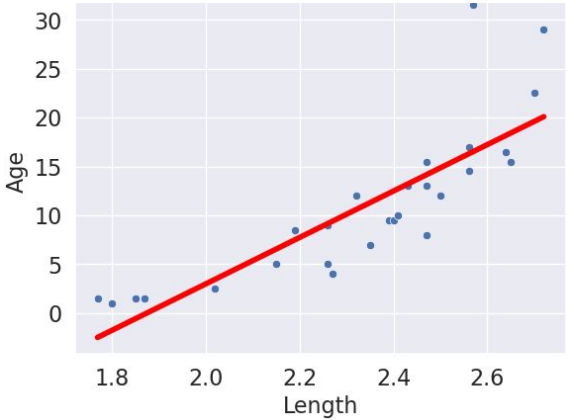


Simple Linear Regression

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

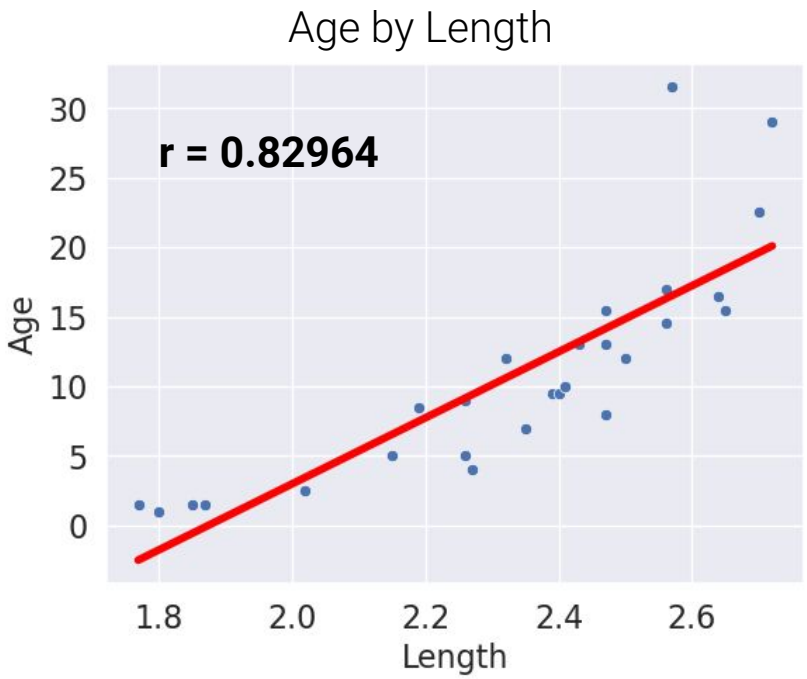
RMSE **4.31**

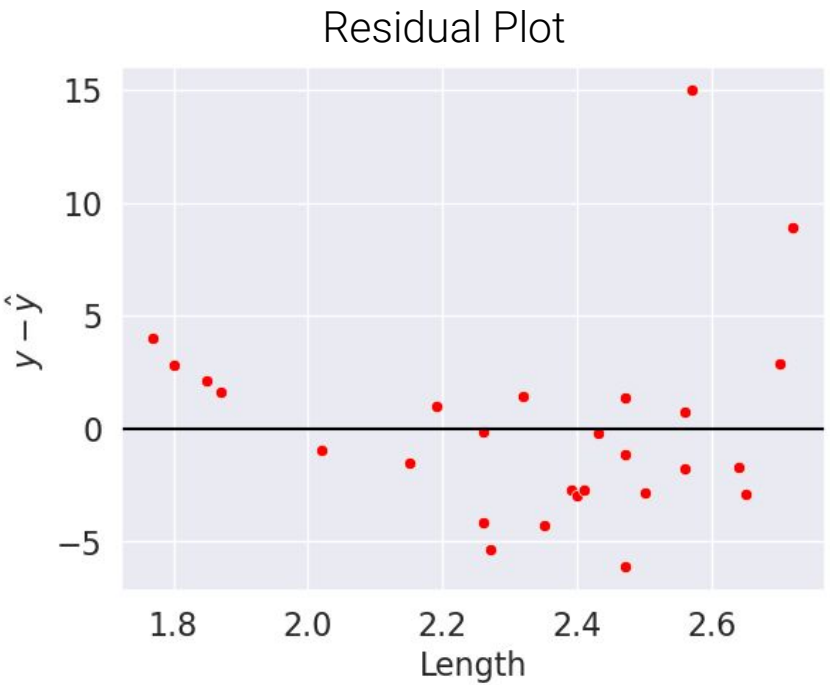
Predictions on a **scatter plot**.



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**.

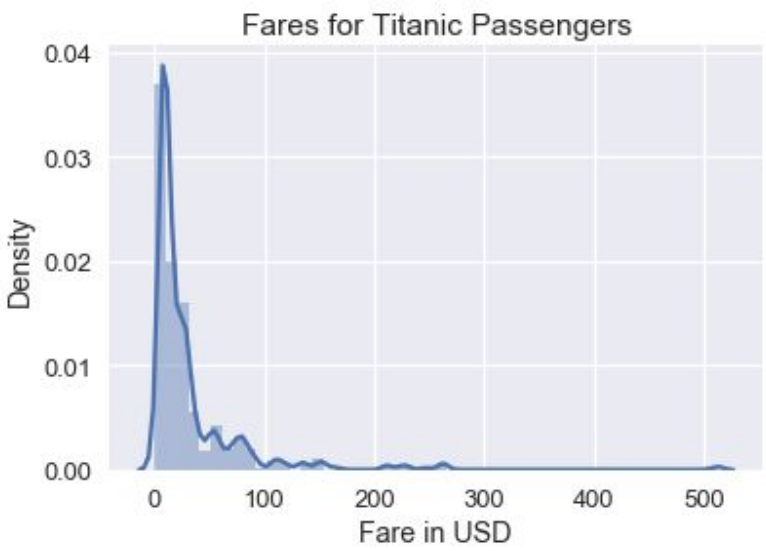
Q: How can we fit a curve to this data with the tools we have?

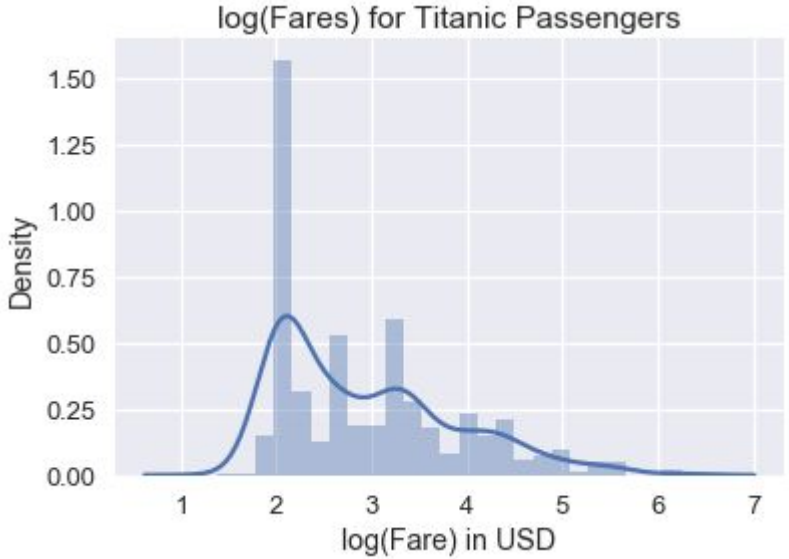
A: **Transform the Data**.

Transformations to Fit Linear Models

Transformations to Fit Linear Models

Transforming data can reveal patterns

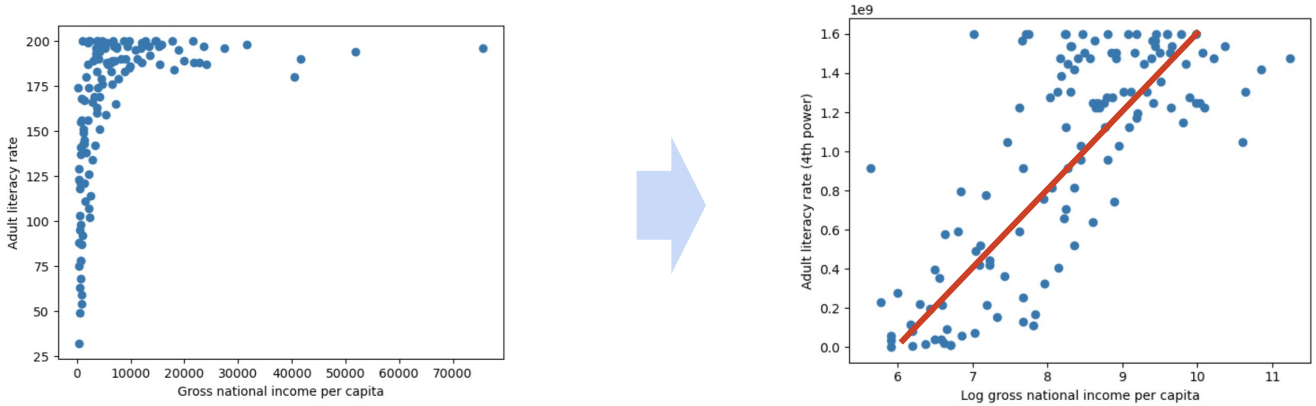




When a distribution has a large dynamic range, it can be useful to take the log.

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

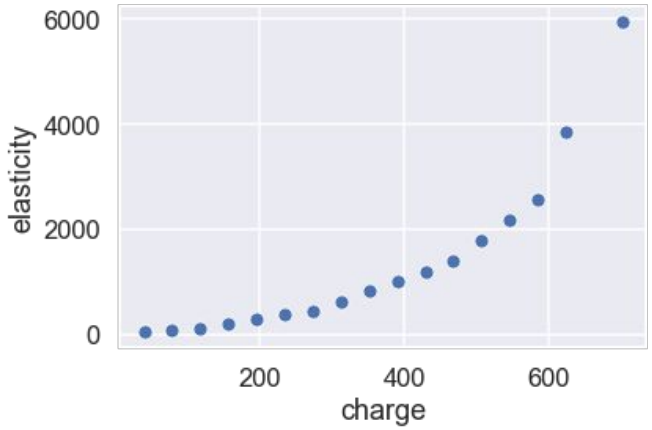
If we take the log of our y-values and notice a linear relationship, we can say (roughly) that

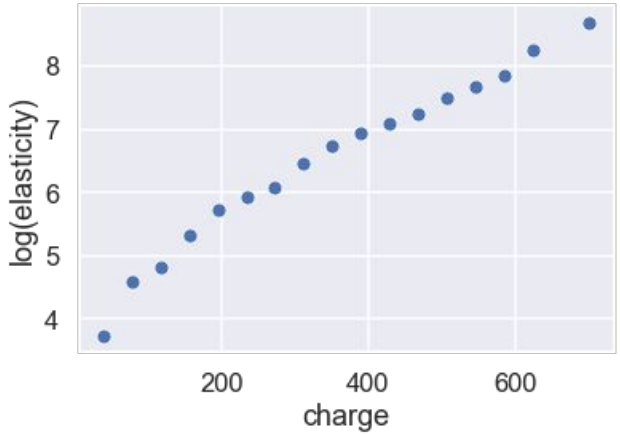
$$\log y = ax + b$$

Working backwards:

$$\log y = ax + b$$
$$y = e^{ax+b}$$
$$y = e^{ax}e^{b}$$
$$y = Ce^{ax}$$

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

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

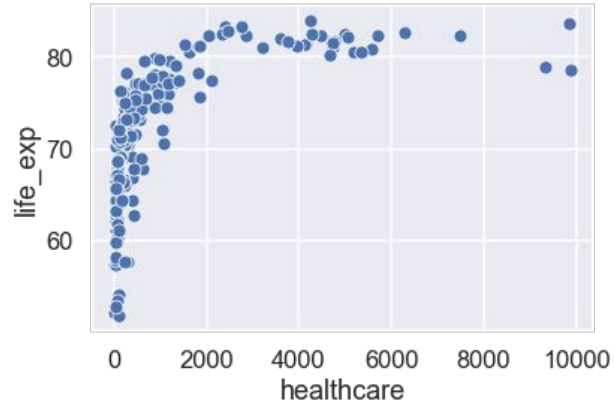
Working backwards:

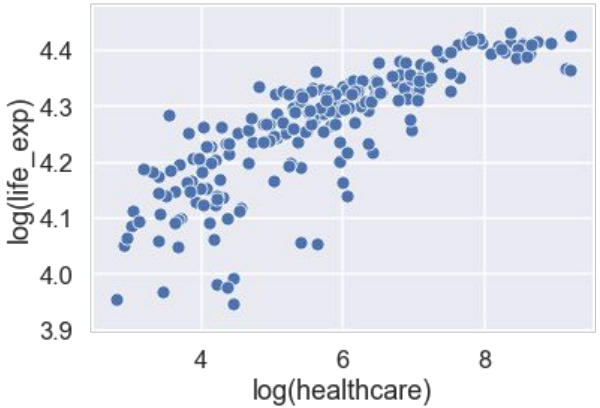
$$y = e^{a \cdot \log x + b}$$

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

$$y = C x^a$$

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





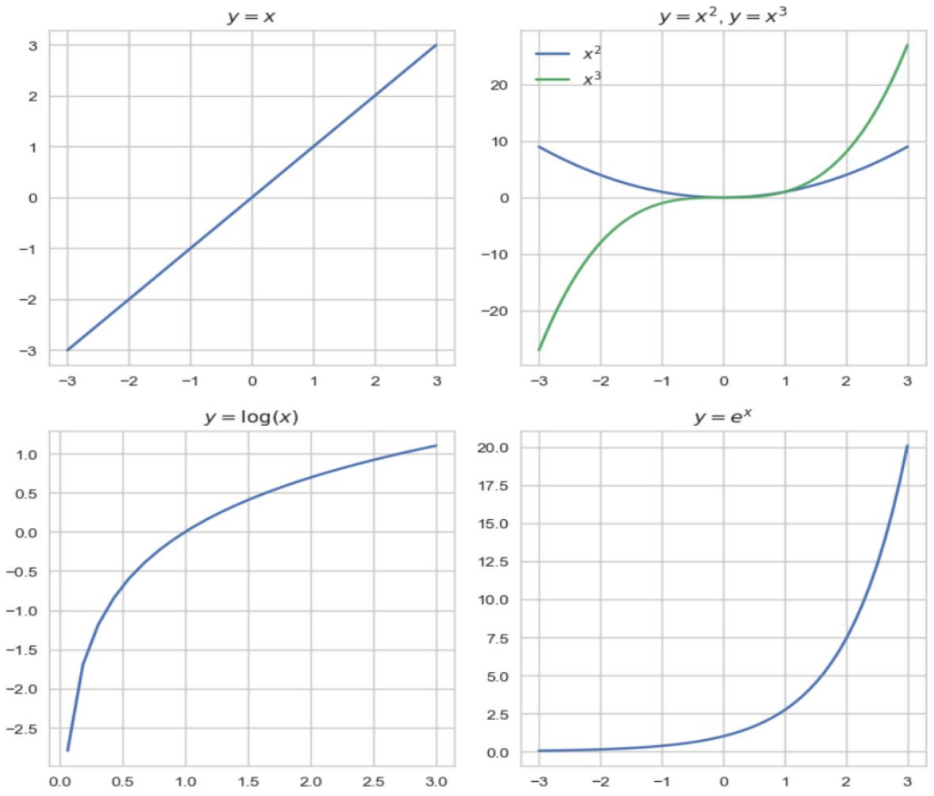
Log transform as a “Swiss army knife”

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

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.

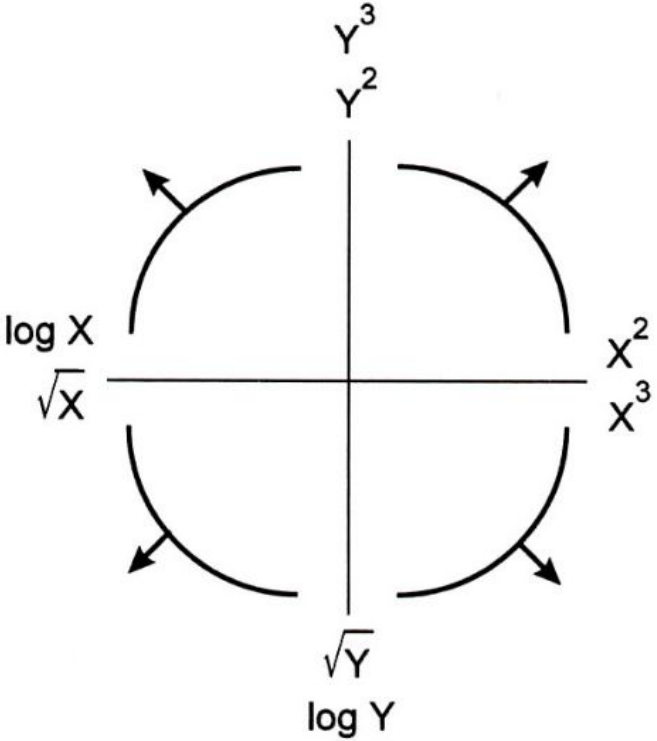


Tukey-Mosteller Bulge Diagram

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

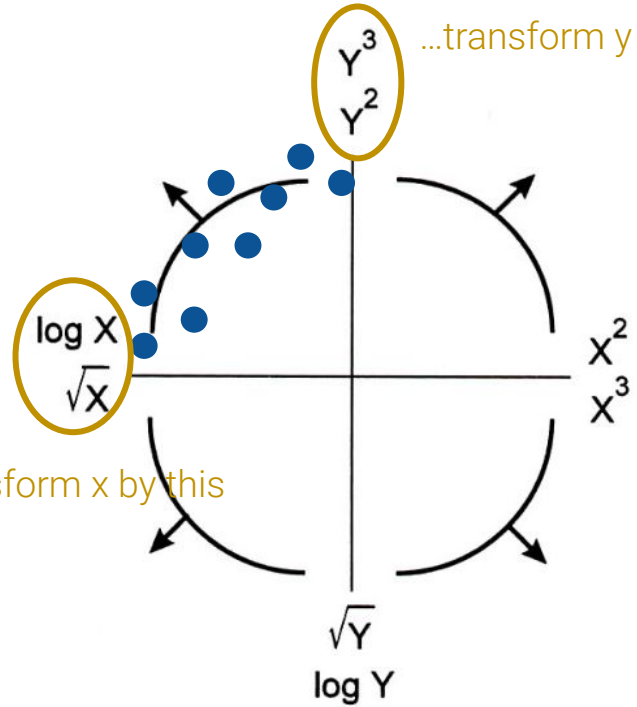
- A visual summary of the reasoning we just worked through.
- sqrt and log make a value "smaller".
- Raising to a value to a power makes it "bigger".
- There are multiple solutions. Some will fit better than others.

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

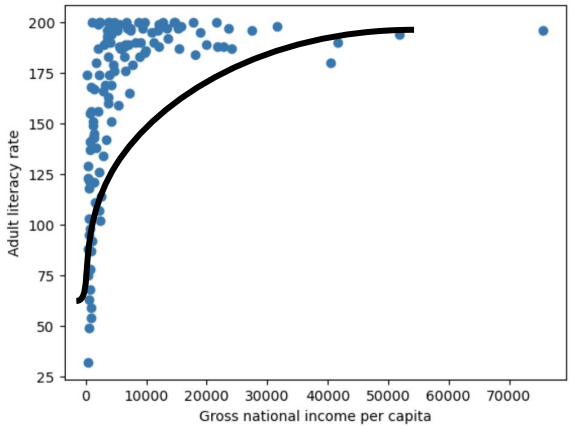


Tukey-Mosteller Bulge Diagram

If the data bulges like this…



Could transform y by y^2, y^3



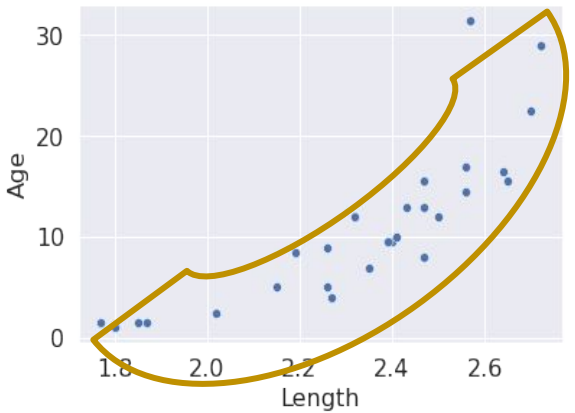
OR: Could transform x by log(x), sqrt(x)

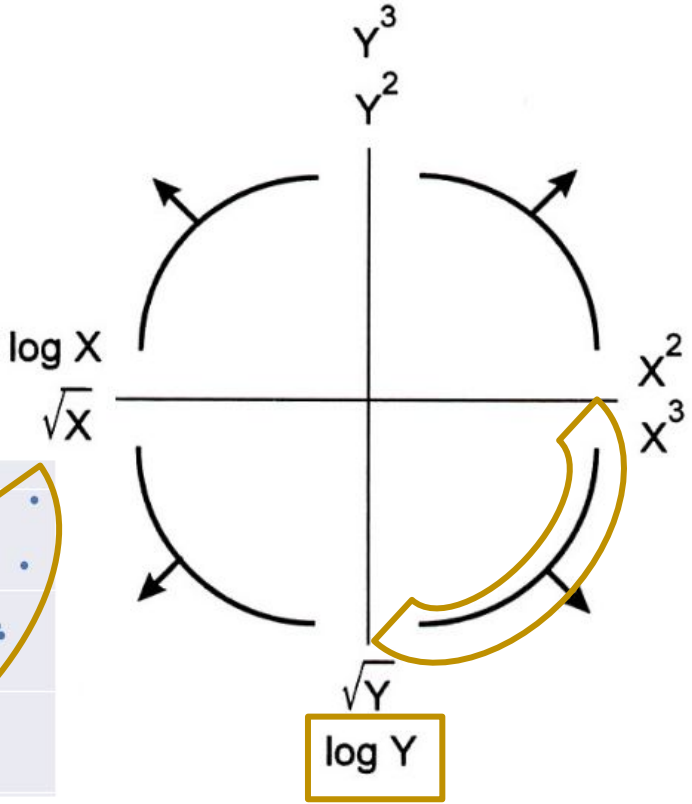
Tukey-Mosteller Bulge Diagram

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

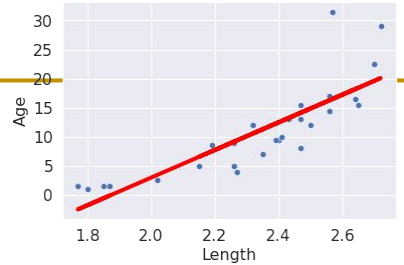
- Each of these transformations equates to increasing or decreasing the scale of an axis.
- Roots and logs make a value “smaller”.
- Raising to a power makes a value “bigger”.

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** $\theta = [\theta_0, \theta_1]$: $\widehat{log(y)} = \theta_0 + \theta_1 x$

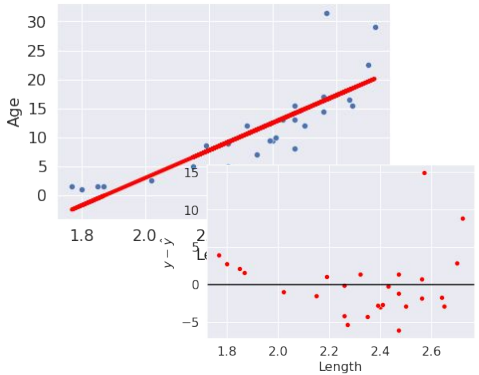
In other words, if we apply the variable transform $z = \log(y)$:

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

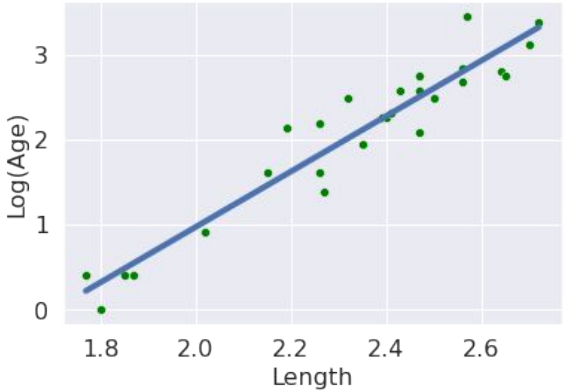
$$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)^2$$

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

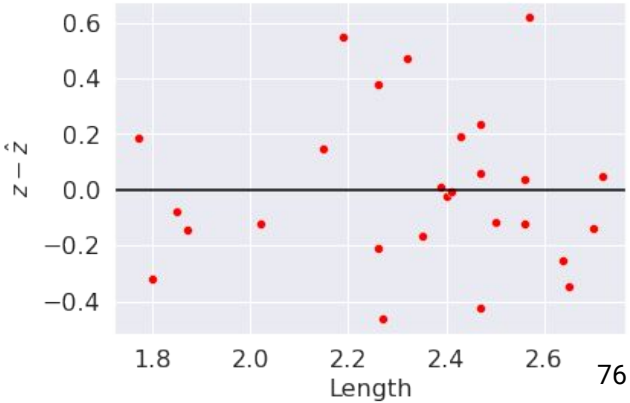
Original (Age by Length)



Log(Age) by Length



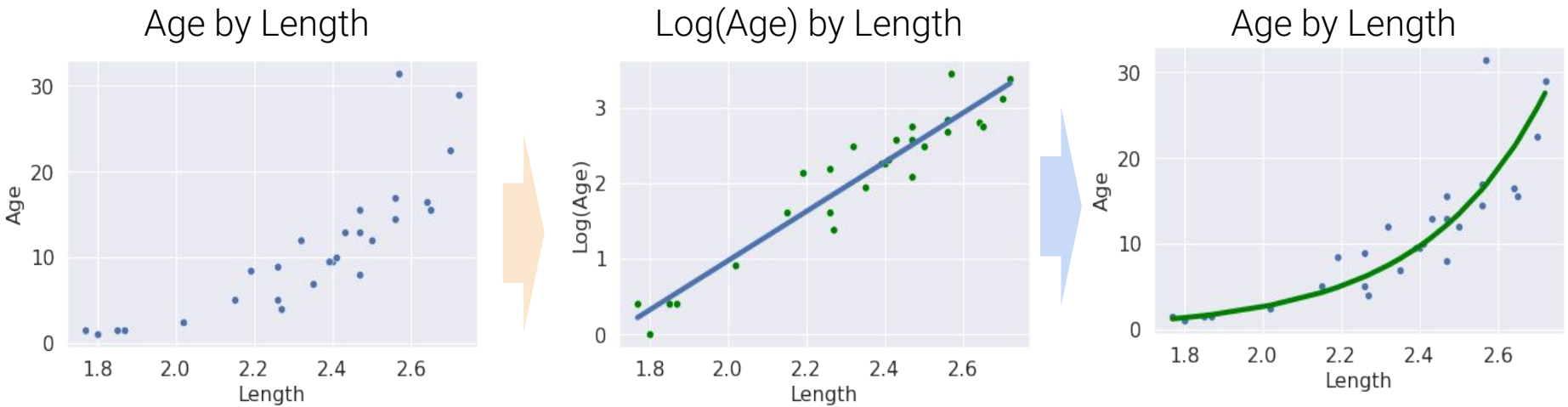
Residual Plot



Fit a Curve using Least Squares Regression

$$z = \log(y)$$

$$y = e^z$$

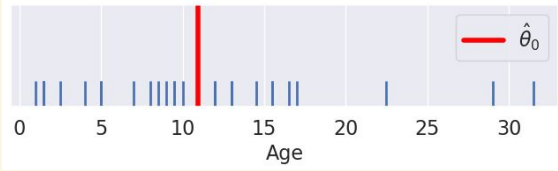


[Fit] Comparing Three Different Models, Both Fit with MSE

Constant Model

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

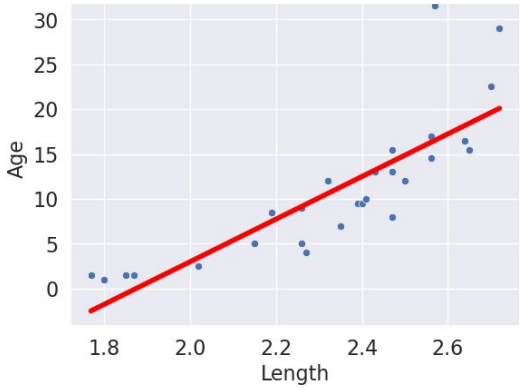
RMSE: **7.72**



Simple Linear Regression

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

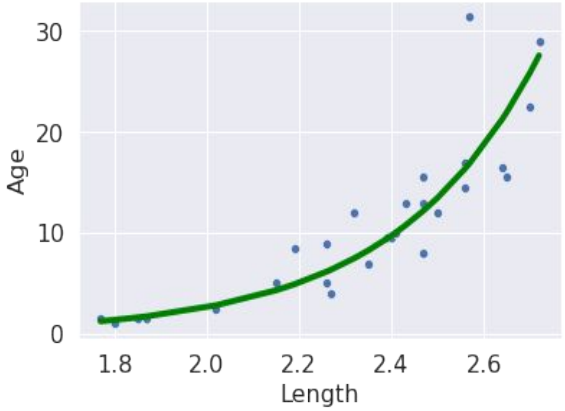
RMSE **4.31**



Log Transformation then Simple Linear Regression:

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

RMSE **3.75**

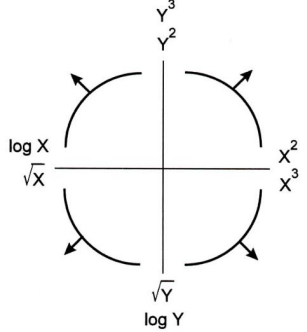


Compare

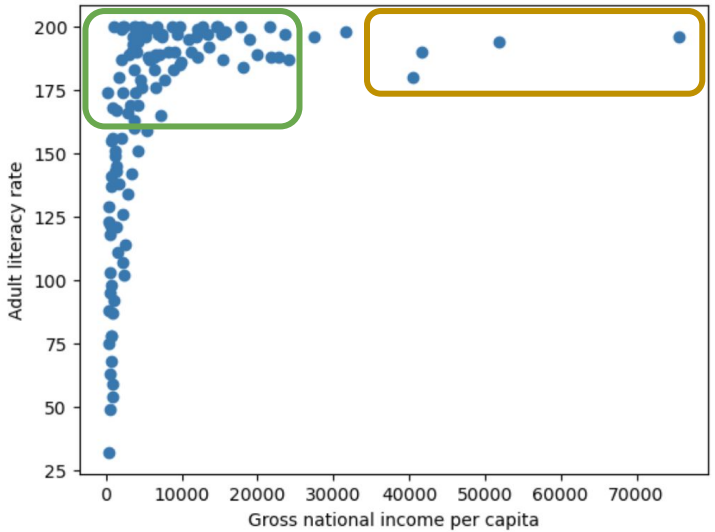
See notebook for code

Appendix More practice with transformations

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.

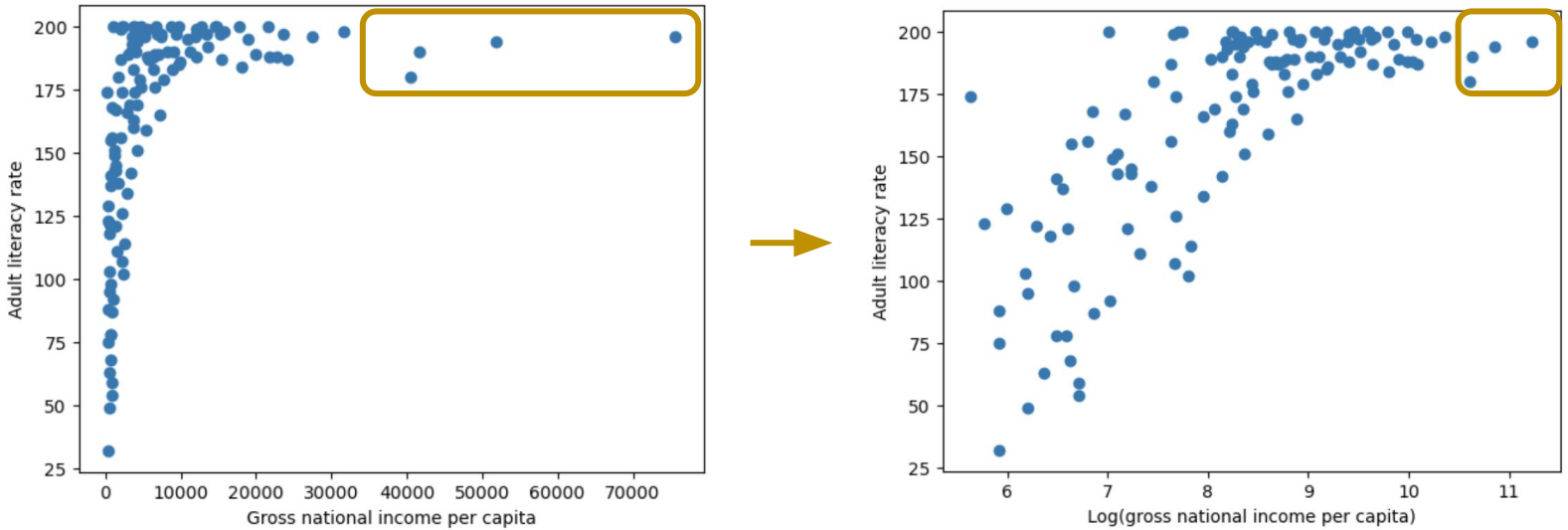
Applying Transformations

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.



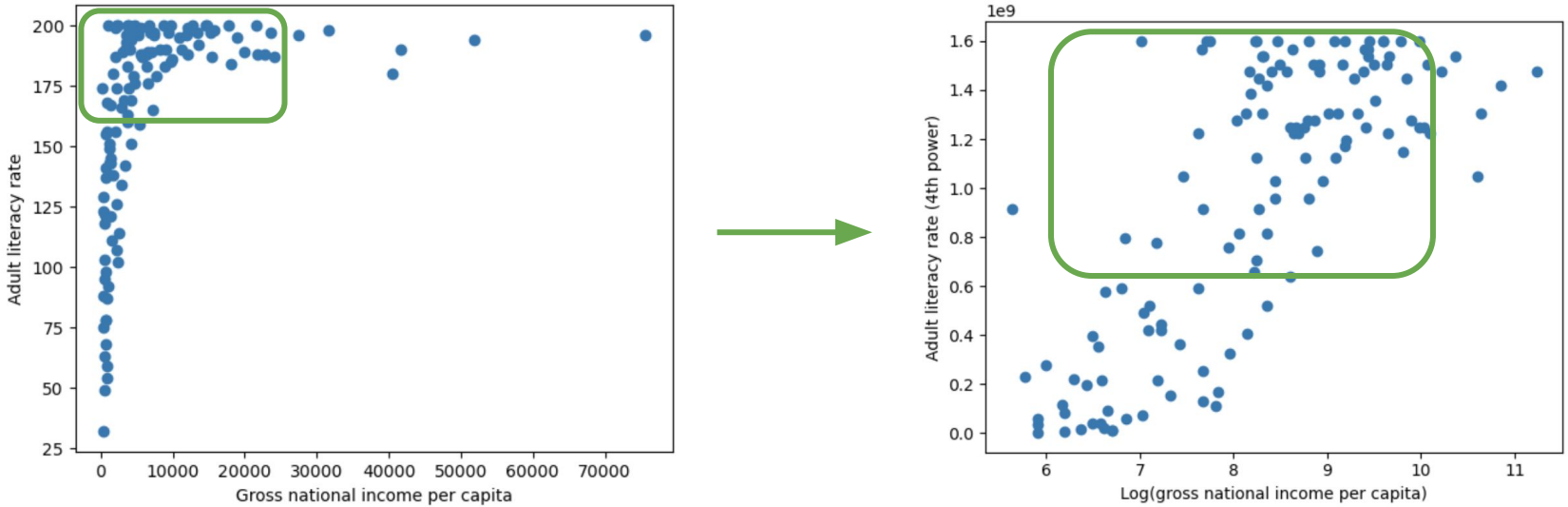
Applying Transformations

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!

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