LECTURE 10

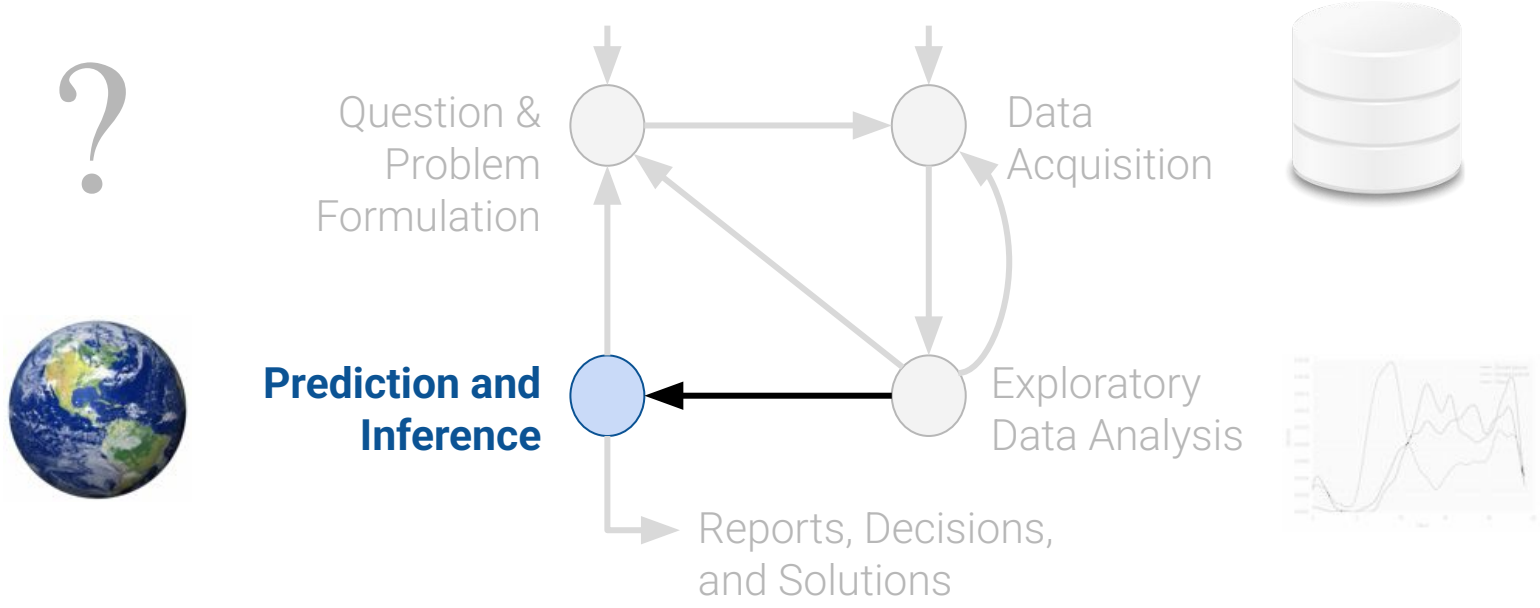
Constant Model, Loss, and Transformations

Adjusting the Modeling Process: different models, loss functions, and data transformations.

Data 100/Data 200, Spring 2022 @ UC Berkeley

Josh Hug and Lisa Yan

Plan for next few lectures: Modeling



(today)

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

Today's Roadmap

Lecture 10, Data 100 Spring 2022

Modeling Process Review

Changing the Model: Constant Model + MSE

Changing the Loss: Constant Model + MAE

Revisiting SLR Evaluation

Transformations to Fit Linear Models

Introducing Notation for Multiple Linear Regression

The Modeling Process (Simple Linear Regression)

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?

Review of the The Modeling Process (Simple Linear Regression)

1. Choose a model

SLR model $\hat{y} = a + bx$

2. Choose a loss function

L2 Loss $L(y, \hat{y}) = (y - \hat{y})^2$

Mean Squared Error (MSE) $R(a, b) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overbrace{(a + bx_i)}^{\hat{y}_i \text{(SLR)}})^2$

3. Fit the model

Minimize average loss with calculus (recording)

$\hat{y} = \hat{a} + \hat{b}x$ $= \bar{y} - \hat{b}\bar{x}$

$\hat{b} = r\frac{\sigma_y}{\sigma_x}$

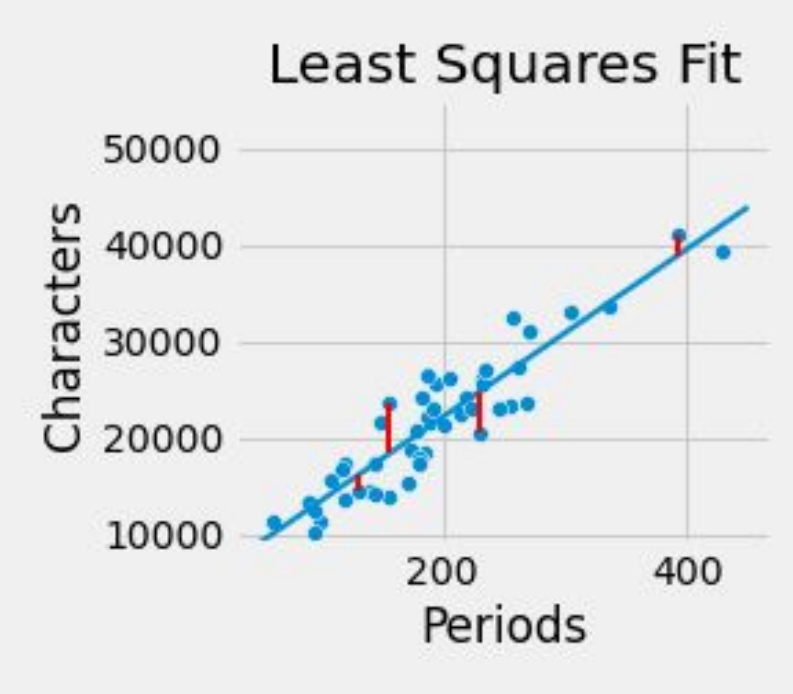
4. Evaluate model performance

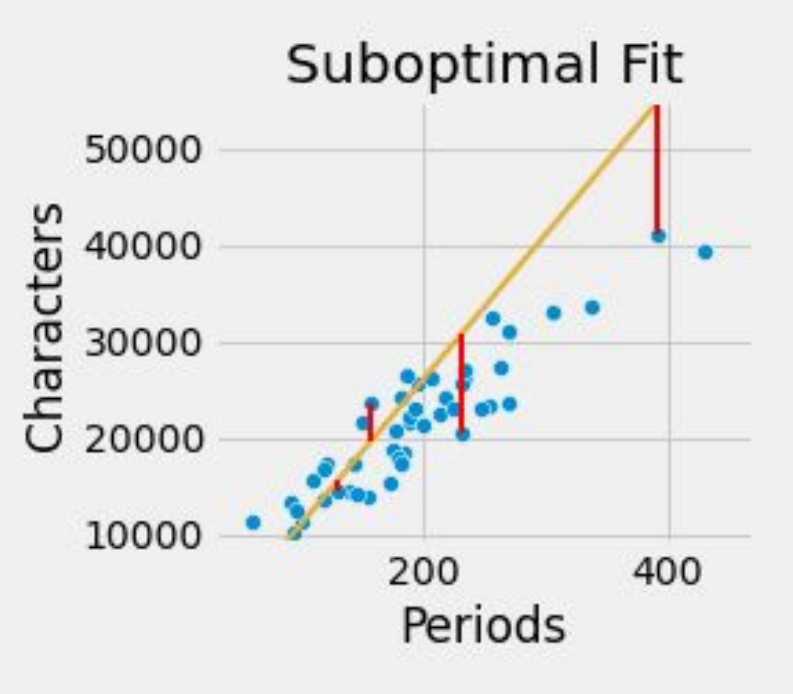
Visualize, Root MSE (to revisit this lecture)

Minimizing MSE is Minimizing Squared Residuals

$$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\underbrace{y_i - \hat{y}_i}_{\text{Residual ("error") in prediction}})^2$$

Lower residuals = better MSE fit!





Terminology: Prediction vs. estimation

These terms are often used somewhat interchangeably, but there is a subtle difference between them.

Estimation is the task of using data to determine model parameters.

Prediction is the task of using a model to predict outputs for unseen data.

We **estimate** parameters by minimizing average loss…

$$\hat{y} = \hat{a} + \hat{b}x$$

…then we **predict** using these estimates.

Least Squares Estimation is when we choose the parameters that minimize MSE.

Changing the Model: Constant Model + MSE

Lecture 10, Data 100 Spring 2022

The Modeling Process: Using a Different Model

1. Choose a model

~~SLR model~~ **Constant Model?** $\hat{y} = ??$

~~$\hat{y} = a + bx$~~

2. Choose a loss function

L2 Loss

Mean Squared Error (MSE)

3. Fit the model

Minimize average loss with calculus

4. Evaluate model performance

Visualize,
Root MSE

The Constant Model

You work at a local bubble tea pop-up store and want to estimate the sales each day.

Here's your data from 5 randomly selected previous days, arbitrarily sorted by number of drinks sold:

$$\{20, 21, 22, 29, 33\}$$

How many drinks will you sell tomorrow?

A. 0
B. 25
C. 22
D. 100
E. Something else

This is a **constant model**.

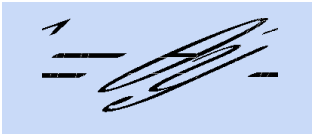
The Constant Model

The **constant model**, also known as a **summary statistic**, summarizes the sample data by always "predicting" the same number—i.e., predicting a constant.

It ignores any relationships between variables:

- For instance, bubble tea sales likely depend on the time of year, the weather, how the customers feel, whether school is in session, etc.
- Ignoring these factors is a **simplifying assumption**.

The constant model is also a parametric, statistical model:



The Constant Model

The **constant model**, also known as a **summary statistic**, summarizes the data by always "predicting" the same number—i.e., predicting a constant.

It ignores any relationships between variables.

- For instance, bubble tea sales likely depend on the time of year, the weather, how the customers feel, whether school is in session, etc.
- Ignoring these factors is a **simplifying assumption**.

The constant model is also a parametric, statistical model:

- Our parameter is 1-dimensional. $\theta \in \mathbb{R}$
- We now have no input into our model; we predict .
- Like before, we can still determine the best $\theta = \hat{\theta}$ that minimizes **average loss** on our data.

The Modeling Process: Using a Different Model

1. Choose a model ✅	~~SLR model~~ $\hat{y} = a + bx$	Constant Model 
2. Choose a loss function	L2 Loss Mean Squared Error (MSE)	**(Let's stick with MSE.)**
3. Fit the model	Minimize average loss with calculus	
4. Evaluate model performance	Visualize, Root MSE	

The Modeling Process: Using a Different Model

1. Choose a model ✅	~~SLR model~~ $\hat{y} = a + bx$	Constant Model 
2. Choose a loss function ✅	L2 Loss Mean Squared Error (MSE)	
3. Fit the model	Minimize average loss with calculus	**How does this step change?**
4. Evaluate model performance	Visualize, Root MSE	

Fit the Model: Rewrite MSE for the Constant Model

Recall that Mean Squared Error (MSE) is average squared loss (L2 loss) over the data $\mathcal{D} = \{y_1, y_2, \ldots, y_n\}$:

$$R(\theta) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{(y_i - \hat{y}_i)^2}_{\text{L2 loss on a single datapoint}}$$

Given the **constant model** $\hat{y} = \theta$:

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

We **fit the model** by finding the optimal $\hat{\theta}$ that minimizes the MSE.

Fit the Model: Three Approaches

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

Approach 1 If you want to prove the general case for any data, you could directly minimize the objective. You showed in HW 1 Q5c that this average loss is minimized by $\hat{\theta} = \textbf{mean}(y) = \bar{y}$.

Approach 2 If you know your data $\mathcal{D} = \{20, 21, 22, 29, 33\}$, you could modify the objective by plugging in values first:

$$R(\theta) = \frac{1}{5}\left((20-\theta)^2 + (21-\theta)^2 + (22-\theta)^2 + (29-\theta)^2 + (33-\theta)^2\right)$$

Approach 3 Algebraic trick.

We review Approach 1 on the next slide.

Approach 2 is left as practice; Approach 3 is in bonus slides.

Fit the Model: Calculus for the General Case

1. Differentiate with respect to θ

$$\frac{d}{d\theta}R(\theta) = \frac{d}{d\theta}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta}(y_i - \theta)^2$$ Derivative of sum is sum of derivatives

$$= \frac{1}{n}\sum_{i=1}^{n}2(y_i - \theta)(-1)$$ Chain rule

$$= \frac{-2}{n}\sum_{i=1}^{n}(y_i - \theta)$$ Simplify constants

2. Set equal to 0.

$$0 = \frac{-2}{n}\sum_{i=1}^{n}(y_i - \theta)$$

3. Solve for $\hat{\theta}$.

Fit the Model: Calculus for the General Case

1. Differentiate with respect to θ.

$$\frac{d}{d\theta}R(\theta) = \frac{d}{d\theta}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta}(y_i - \theta)^2$$ Derivative of sum is sum of derivatives

$$= \frac{1}{n}\sum_{i=1}^{n}2(y_i - \theta)(-1)$$ Chain rule

$$= \frac{-2}{n}\sum_{i=1}^{n}(y_i - \theta)$$ Simplify constants

2. Set equal to 0.

$$0 = \frac{-2}{n}\sum_{i=1}^{n}(y_i - \theta)$$

3. Solve for $\hat{\theta}$.

$$0 = \frac{-2}{n}\sum_{i=1}^{n}(y_i - \theta) = \sum_{i=1}^{n}(y_i - \theta)$$

$$= \sum_{i=1}^{n}y_i - \sum_{i=1}^{n}\theta$$ Separate sums

$$= \left(\sum_{i=1}^{n}y_i\right) - n\theta$$ c + c + … + c = n*c

$$n\theta = \left(\sum_{i=1}^{n}y_i\right)$$

$$\hat{\theta} = \frac{1}{n}\left(\sum_{i=1}^{n}y_i\right) \Rightarrow \boxed{\hat{\theta} = \bar{y}}$$

For more calculus practice, see the SLR derivation video for Lecture 09 (link).

Interpreting $\hat{\theta} = \bar{y}$

This is the optimal parameter for constant model + MSE.

- It holds true regardless of what data sample you have.
- It provides some formal reasoning as to why the mean is such a common summary statistic.

Fun fact:
The minimum MSE is the **sample variance**.

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

Note the difference:

$$R(\hat{\theta}) = \min_{\theta} R(\theta) = \sigma_y^2$$

The **minimum value** of constant + MSE

vs

$$\hat{\theta} = \operatorname*{argmin}_{\theta} R(\theta) = \bar{y}$$

The **arg**ument that **min**imizes constant + MSE

In modeling, we care less about **minimum loss** $R(\hat{\theta})$ and more about the **minimizer** of loss $\hat{\theta}$.

The Modeling Process: Using a Different Model

1. Choose a model ✅ Constant Model

2. Choose a loss function ✅ L2 Loss

Mean Squared Error (MSE)

3. Fit the model ✅ Minimize average loss with calculus

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

$$\hat{\theta} = \mathbf{mean}(y) = \bar{y}$$

4. Evaluate model performance Visualize, Root MSE

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

Suppose we wanted to predict dugong ages.

[image source]

Compare

Constant Model

$$h = \theta$$

Data: Sample of ages.

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

Simple Linear Model

$$\hat{y} = a + bx$$

Data: Sample of (length, age)s.

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

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

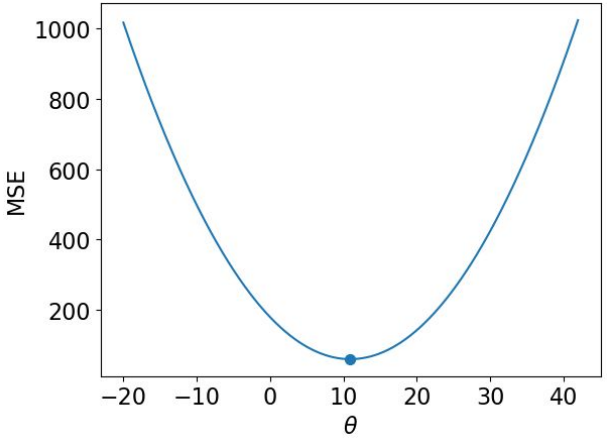
Compare

Constant Model

$\hat{y} = h$

h is **1-D**.

Loss surface is **2-D**.



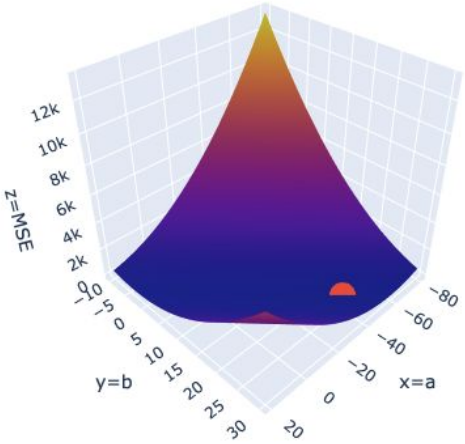
$$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2$$

Simple Linear Model

$$\hat{y} = a + bx$$

$h = (a, b)$ is **2-D**.

Loss surface is **3-D**.



$$R(a, b) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

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

Constant Model

RMSE: **7.72**

Simple Linear Regression

$$\hat{y} = a + bx$$

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)

Compare

See notebook for code

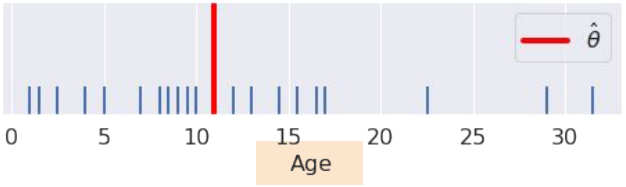
[Fit] Comparing Two Different Models, Both Fit with MSE

Constant Model

$$h = \theta$$

RMSE: **7.72**

Predictions on a **rug plot**.

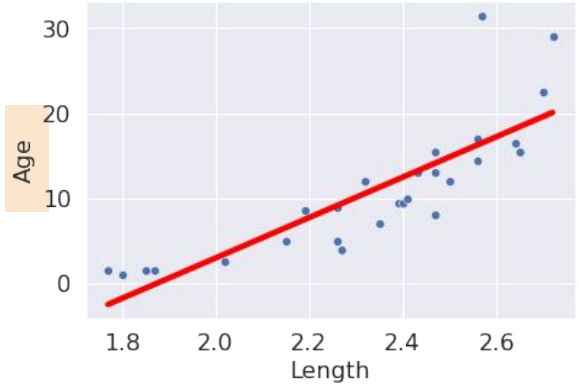


Simple Linear Regression

$$\hat{y} = a + bx$$

RMSE **4.31**

Predictions on a **scatter plot**.



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

Compare

See notebook for code

Changing the Loss: Constant Model + MAE

Lecture 10, Data 100 Spring 2022

The Modeling Process: Using a Different Loss Function

1. Choose a model ✅

Constant Model

2. Choose a loss function ✅

~~L2 Loss~~

~~Mean Squared Error (MSE)~~

Suppose instead we use **L1 loss**.
Average loss then becomes
Mean Absolute Error (MAE).

3. Fit the model

Minimize average loss with calculus

4. Evaluate model performance

Visualize,
Root MSE

The Modeling Process: Using a Different Loss Function

1. Choose a model ✅	Constant Model	
2. Choose a loss function ✅	~~L2 Loss~~ ~~Mean Squared Error (MSE)~~	Suppose instead we use **L1 loss**. Average loss then becomes **Mean Absolute Error (MAE)**.
3. Fit the model	Minimize average loss with calculus	**How does this step change?**
4. Evaluate model performance	Visualize, Root MSE	

Fit the Model: Rewrite MAE for the Constant Model

Recall that Mean **Absolute** Error (MAE) is average **absolute** loss (L1 loss) over the data $\mathcal{D} = \{y_1, y_2, \ldots, y_n\}$:

$$R(\theta) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{|y_i - \hat{y}_i|}_{\text{L1 loss on a single datapoint}}$$

Given the **constant model** $\hat{y} = \theta$:

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

We **fit the model** by finding the optimal $\hat{\theta}$ that minimizes the MAE.

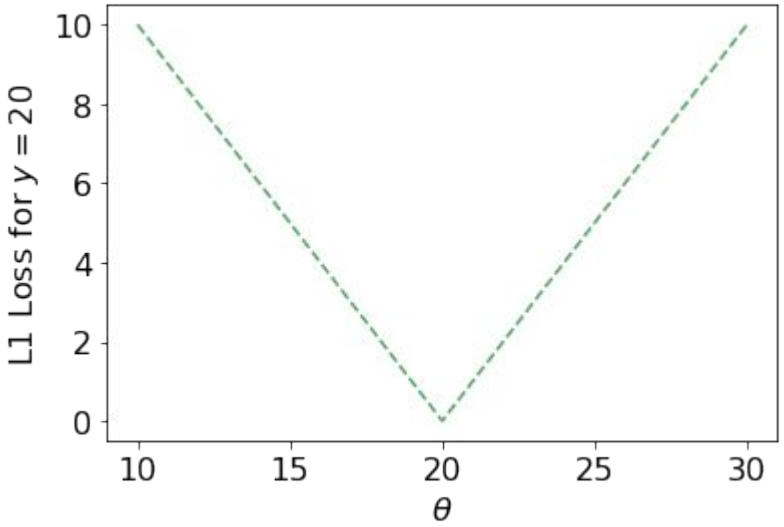
Exploring MAE: A Piecewise function

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

For the bubble tea dataset {20, 21, 22, 29, 33}:

Absolute (L1) Loss on one observation:

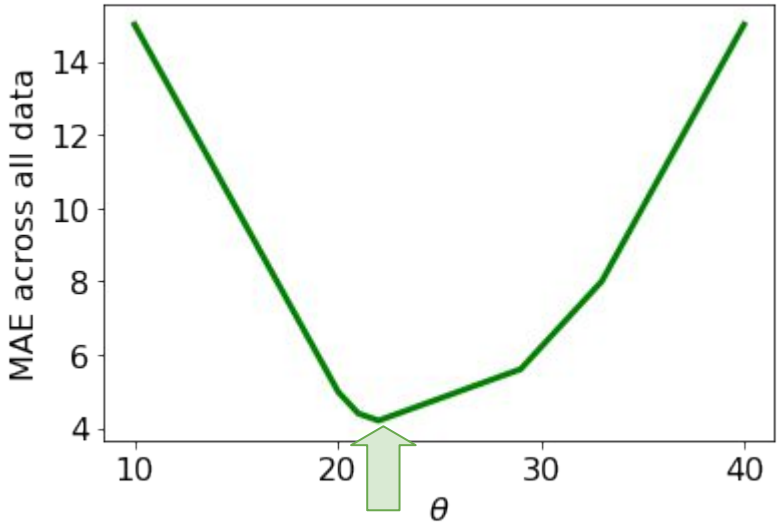
$$L_1(20, \theta) = |20 - \theta|$$



An absolute value curve, centered at theta = 20.

MAE (average absolute loss) across all data:

$$R(\theta) = \frac{1}{5}\left(|20 - \theta| + |21 - \theta| + |22 - \theta| + |29 - \theta| + |33 - \theta|\right)$$



Piecewise linear function… minimized at…theta = 22?

Fit the Model: Calculus

1. Differentiate with respect to θ.

$$\frac{d}{d\theta}R(\theta) = \frac{d}{d\theta}\left(\frac{1}{n}\sum_{i=1}^{n}|y_i - \theta|\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta}|y_i - \theta|$$

⚠️ Absolute value!

The following derivation is beyond what we expect you to generate on your own. But you should understand it.

Fit the Model: Calculus

1. Differentiate with respect to θ.

$$\frac{d}{d\theta}R(\theta) = \frac{d}{d\theta}\left(\frac{1}{n}\sum_{i=1}^{n}|y_i - \theta|\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta}|y_i - \theta|$$

$$|y_i - \theta| = \begin{cases} y_i - \theta & \text{if } \theta \leq y_i \\ \theta - y_i & \text{if } \theta > y_i \end{cases}$$

$$\frac{d}{d\theta}|y_i - \theta| = \begin{cases} -1 & \text{if } \theta < y_i \\ 1 & \text{if } \theta > y_i \end{cases}$$

$$= \frac{1}{n}\left[\sum_{\theta<y_i}(-1) + \sum_{\theta>y_i}(+1)\right]$$

Note: The derivative of the absolute value when the argument is 0 (i.e. when $y_i = \theta$) is technically undefined. We ignore this case in our derivation, since thankfully, it doesn't change our result (proof left to you).

Take some time to process this math. Talk it out!

Fit the Model: Calculus

1. Differentiate with respect to θ.

$$\frac{d}{d\theta}R(\theta) = \frac{d}{d\theta}\left(\frac{1}{n}\sum_{i=1}^{n}|y_i - \theta|\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta}|y_i - \theta|$$

$$|y_i - \theta| = \begin{cases} y_i - \theta & \text{if } \theta \le y_i \\ \theta - y_i & \text{if } \theta > y_i \end{cases}$$

$$\frac{d}{d\theta}|y_i - \theta| = \begin{cases} -1 & \text{if } \theta < y_i \\ 1 & \text{if } \theta > y_i \end{cases}$$

$$= \frac{1}{n}\left[\sum_{\theta<y_i}(-1) + \sum_{\theta>y_i}(+1)\right]$$

Sum up for $i = 1, \ldots, n$:
−1 if observation y_i **>** our prediction θ ;
+1 if observation y_i **<** our prediction θ .

Fit the Model: Calculus

1. Differentiate with respect to θ.

$$\frac{d}{d\theta}R(\theta) = \frac{d}{d\theta}\left(\frac{1}{n}\sum_{i=1}^{n}|y_i - \theta|\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta}|y_i - \theta|$$

$$|y_i - \theta| = \begin{cases} y_i - \theta & \text{if } \theta \leq y_i \\ \theta - y_i & \text{if } \theta > y_i \end{cases}$$

$$\frac{d}{d\theta}|y_i - \theta| = \begin{cases} -1 & \text{if } \theta < y_i \\ 1 & \text{if } \theta > y_i \end{cases}$$

$$= \frac{1}{n}\left[\sum_{\theta<y_i}(-1) + \sum_{\theta>y_i}(+1)\right]$$

2. Set equal to 0.

$$0 = \frac{1}{n}\left[\sum_{\theta<y_i}(-1) + \sum_{\theta>y_i}(+1)\right]$$

3. Solve for $\hat{\theta}$.

$$0 = -\sum_{\theta<y_i} 1 + \sum_{\theta>y_i} 1$$

$$\sum_{\theta<y_i} 1 = \sum_{\theta>y_i} 1$$

Where do we go from here?

Median Minimizes MAE for the Constant Model

The constant model parameter $\theta = \hat{\theta}$ that minimizes MAE must satisfy:

$$\underbrace{\sum_{\theta < y_i} 1}_{\text{\# observations greater than } \hat{\theta}} = \underbrace{\sum_{\theta > y_i} 1}_{\text{\# observations less than } \hat{\theta}}$$

In other words, theta needs to be such that there are **an equal # of points to the left and right**.

This is the definition of the **median!**

$$\hat{\theta} = \text{median}(y)$$

For example, in our bubble tea dataset {20, 21, 22, 29, 33}, the point in **green (22)** is the median.

It is the value in the "middle."

Summary: Loss Optimization, Calculus, and…Critical Points?

First, define the **objective function** as average loss.

- Plug in L1 or L2 loss.
- Plug in model so that resulting expression is a function of w.

Then, find the **minimum** of the objective function:

1. Differentiate with respect to w.
2. Set equal to 0.
3. Solve for $\hat{\theta}$.

Repeat w/partial derivatives if multiple parameters

Recall **critical points** from calculus: $R(\hat{\theta})$ could be a minimum, maximum, or saddle point!

- We should technically also perform the second derivative test, i.e., show $R''(\hat{\theta}) > 0$.
- You will prove on homework that MSE has a property—**convexity**—that guarantees that $R(\hat{\theta})$ is a global minimum.
- The proof of convexity for MAE is beyond this course.

The Modeling Process: Using a Different Loss Function

1. Choose a model ✅

Constant Model

2. Choose a loss function ✅

L1 Loss

Mean Absolute Error (MAE)

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

3. Fit the model ✅

Minimize average loss with calculus

$$\hat{\theta} = \textbf{median}(y)$$

4. Evaluate ~~model performance~~ loss

Visualize, ~~Root MSE~~

[Estimator] Two Constant Models, Fit to Different Losses

MSE (Average Squared Loss)

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

Minimized with **sample mean**:

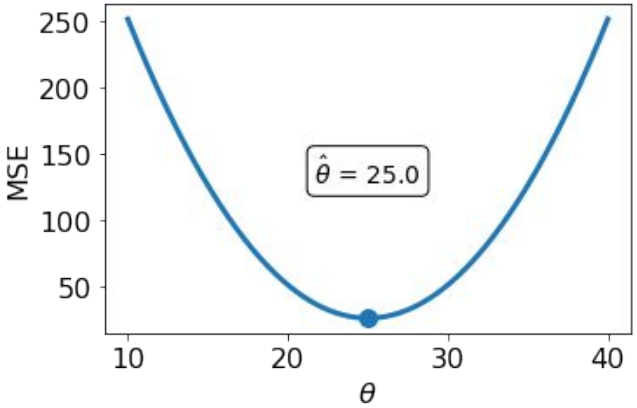
$$\hat{\theta} = \textbf{mean}(y)$$

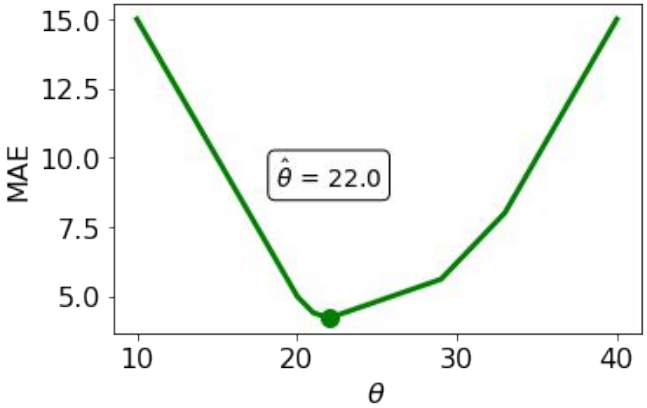
MAE (Average Absolute Loss)

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

Minimized with **sample median**:

$$\hat{\theta} = \textbf{median}(y)$$





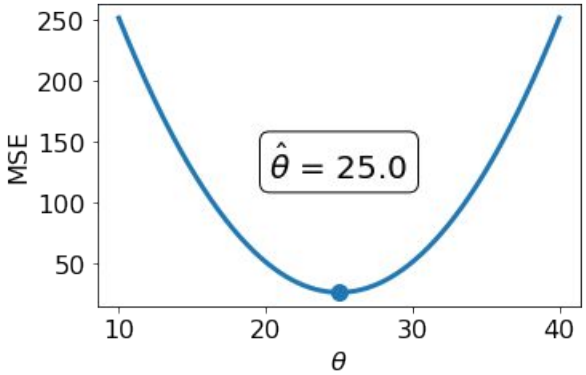
Compare

[Loss] Two Constant Models, Fit to Different Losses

Compare

MSE (Average Squared Loss)

$$\hat{\theta} = \text{mean}(y)$$

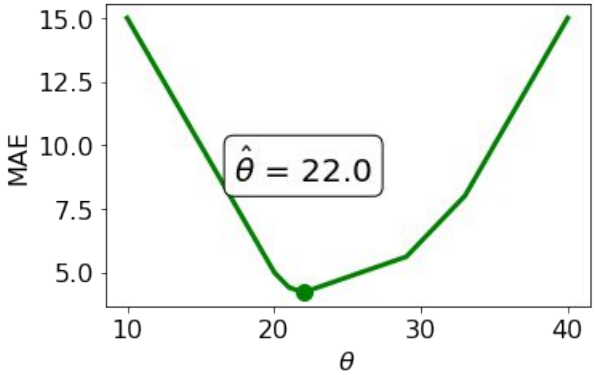


Smooth. Easy to minimize using numerical methods (in a few weeks).

⚠️ **Sensitive** to outliers. Adding 1000 to largest observation → theta = 225

MAE (Average Absolute Loss)

$$\hat{\theta} = \text{median}(y)$$



⚠️ **Piecewise**. at each of the "kinks," it's not differentiable. Harder to minimize.

Robust to outliers. Adding 1000 to largest observation keeps theta = 22.

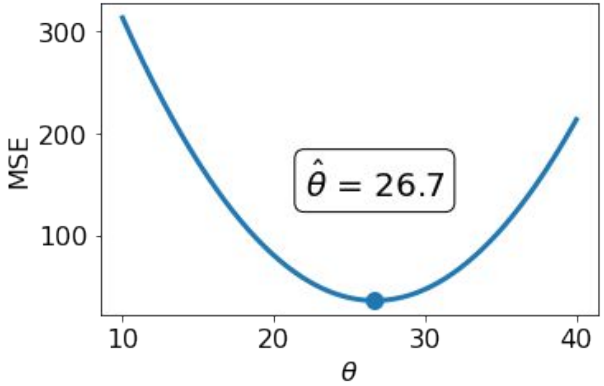
Compare

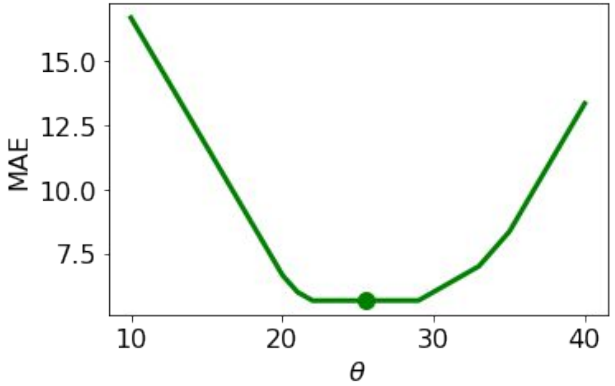
Uniqueness under Different Loss Functions

MSE (Average Squared Loss)

MAE (Average Absolute Loss)

Suppose we add a 6th observation to our bubble tea dataset:
{20, 21, 22, 29, 33, **35**}





Unique $\hat{\theta}$:

$$\hat{\theta} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)$$

⚠️ **Infinitely many** $\hat{\theta}$**s**. Any θ in range (22,29) minimizes MAE.

Updated 2/21

(In practice: With an even # of datapoints, set median to mean of two middle points, e.g., 25.5).

Interlude

Dugong
(marine mammal)

Dewgong
(Gen I Pokemon)

Midterm 1 Announcement

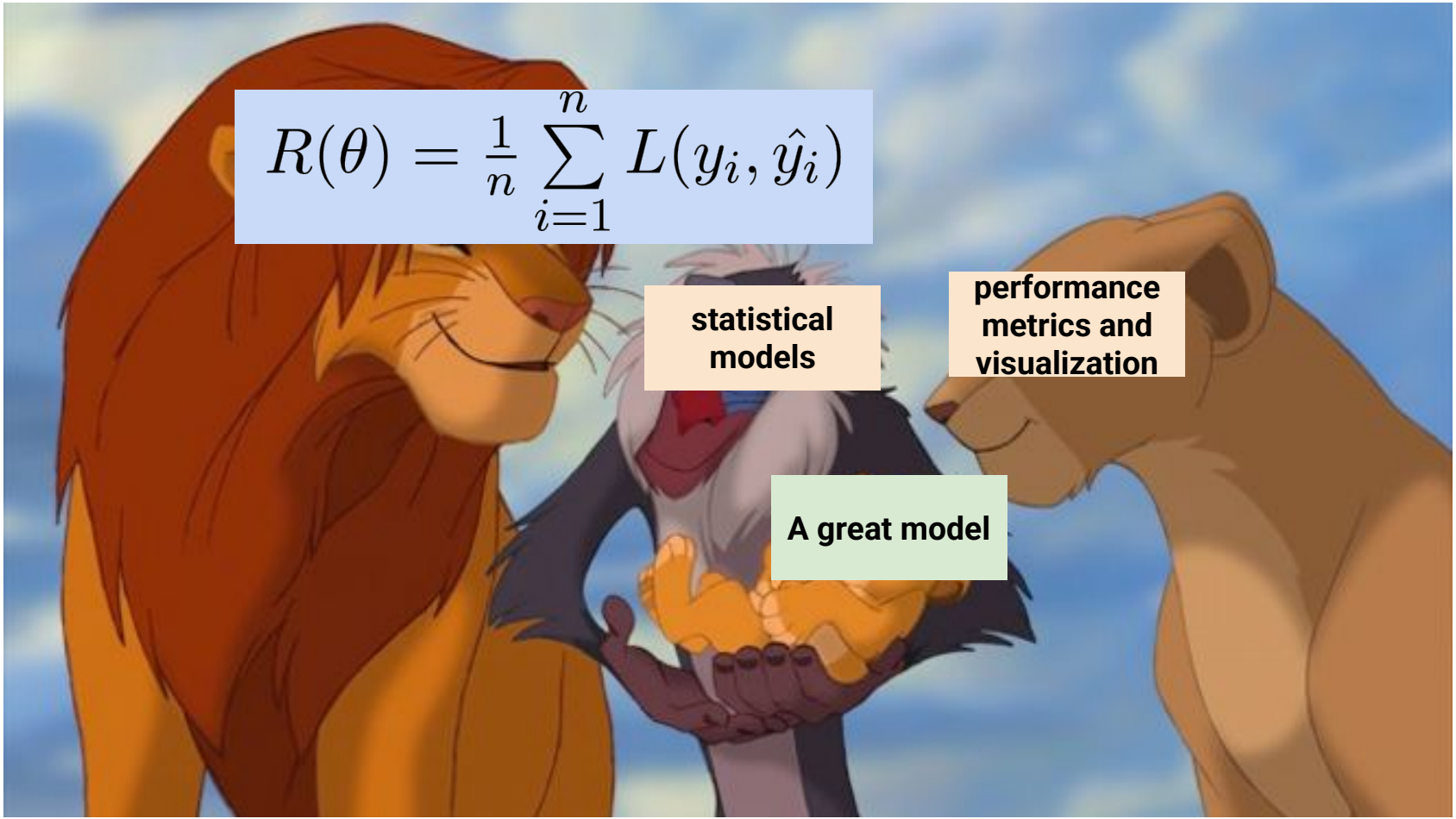
(Josh)

Break (3 min)

Revisiting SLR Evaluation (from Lecture 09)

Lecture 10, Data 100 Spring 2022

Finding the Best Model



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
```

Visualize, then quantify!

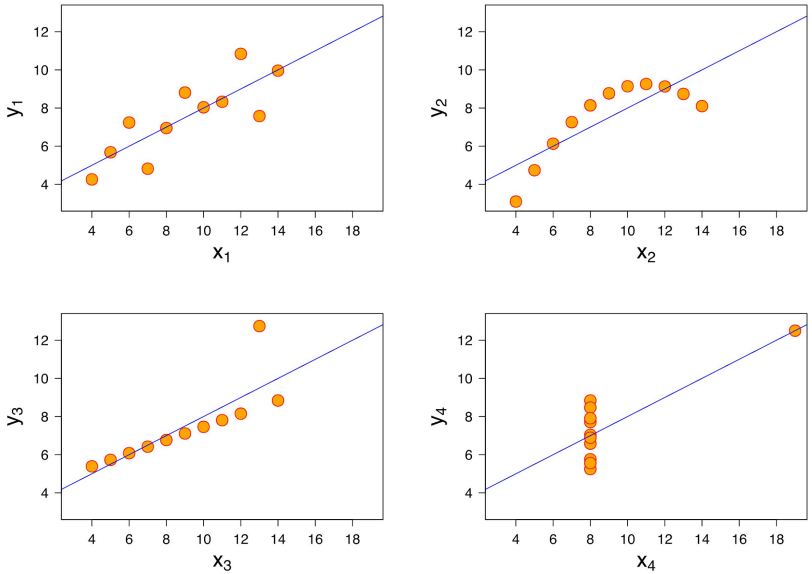
Anscombe's quartet refers to the following four sets of points on the right.

- 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!

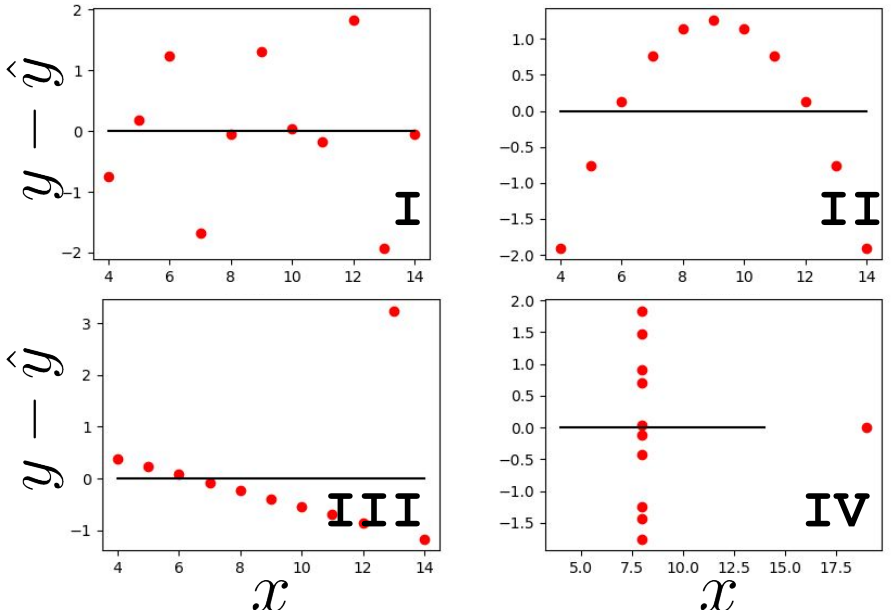
Four mysterious datasets + Least Squares

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
```



From Data 8 (textbook):

> The residual plot of a good regression shows no pattern.

Transformations to Fit Linear Models

Lecture 10, Data 100 Spring 2022

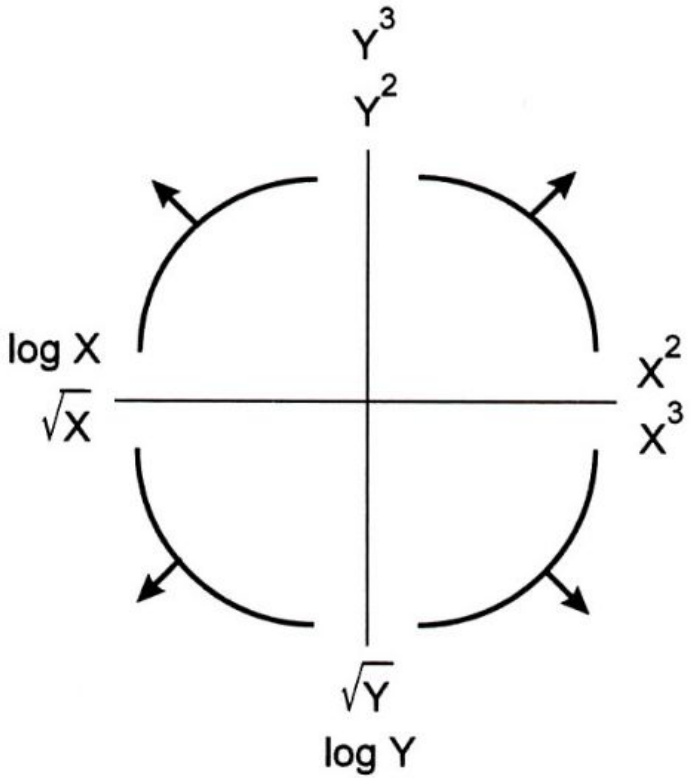
Tukey-Mosteller Bulge Diagram

Other transformations are possible to try to get linearity.

- There are multiple solutions. Some will fit better than others.
- sqrt and log make a value “smaller”. Raising to a value to a power makes it “bigger”.
- Each of these transformations equates to increasing or decreasing the scale of an axis.
- At a later date, we may discuss the Tukey-Mosteller Bulge Diagram (on the right), which provides guidance for linearization.

Today!

And other goals are possible, e.g. makin
appear more symmetric (see lab 4).



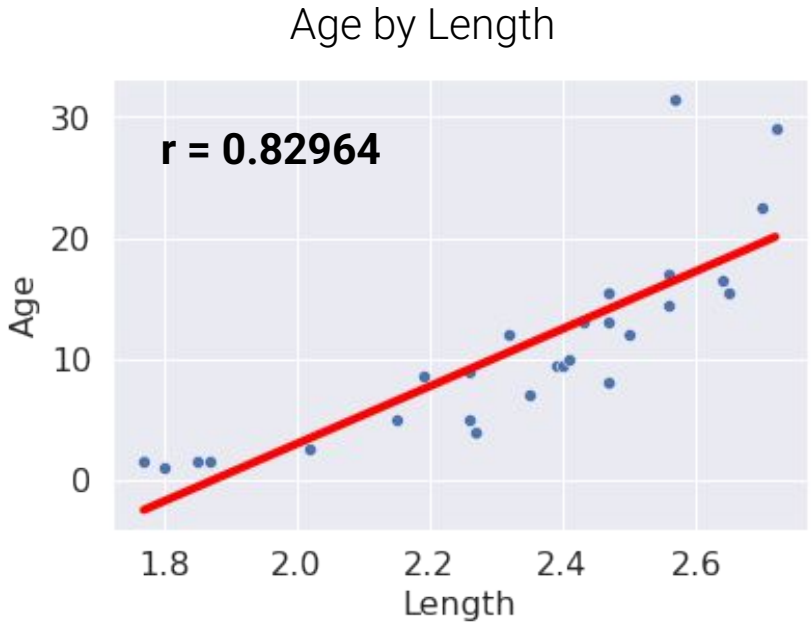
Back to Least Squares Regression with Dugongs

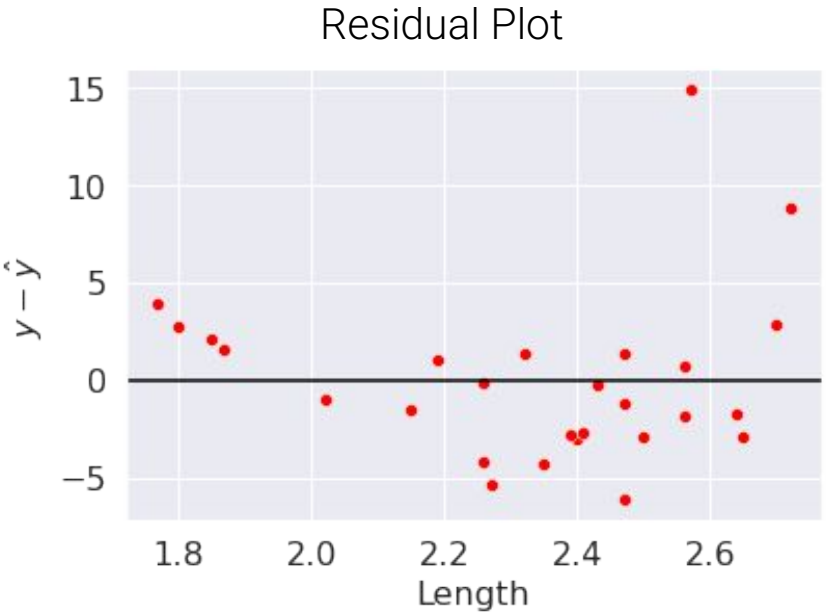
From Data 8 (textbook):

> The residual plot of a good regression shows no pattern.

https://inferentialthinking.com/chapters/15/5/Visual_Diagnostics.html

Back to Least Squares Regression with Dugongs





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

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

A: **Transform the Data**.

Transforming Dugongs

Suppose we do a log(y) transformation (we'll explain why soon).

Notice that the resulting model is still **linear in the parameters** $\theta = (a, b)$:

$$\widehat{\log(y)} = a + bx$$

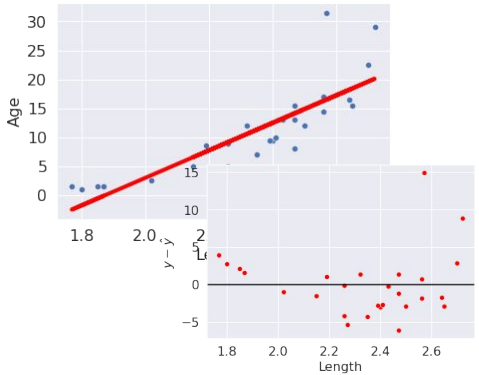
In other words, if we apply the variable transform $z = \log(y)$:

$$\hat{z} = a + bx$$

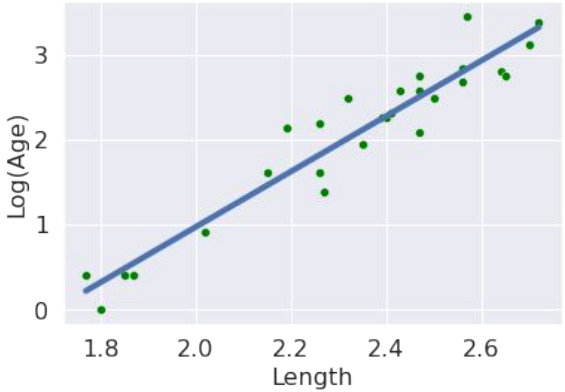
$$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)^2$$

$$\hat{a} = \bar{\mathbf{z}} - \hat{b}\bar{x} \qquad \hat{b} = r\frac{\sigma_{\mathbf{z}}}{\sigma_x}$$

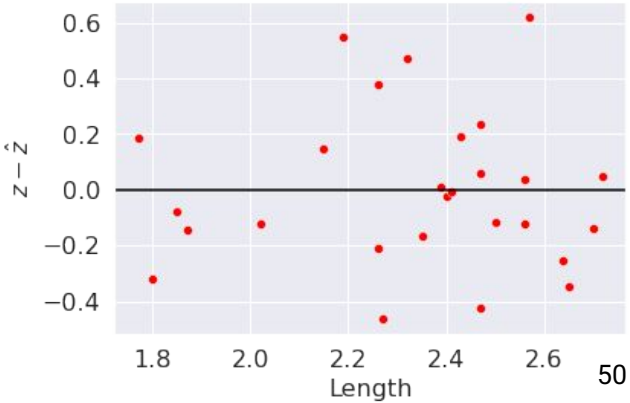
Original (Age by Length)



Log(Age) by Length



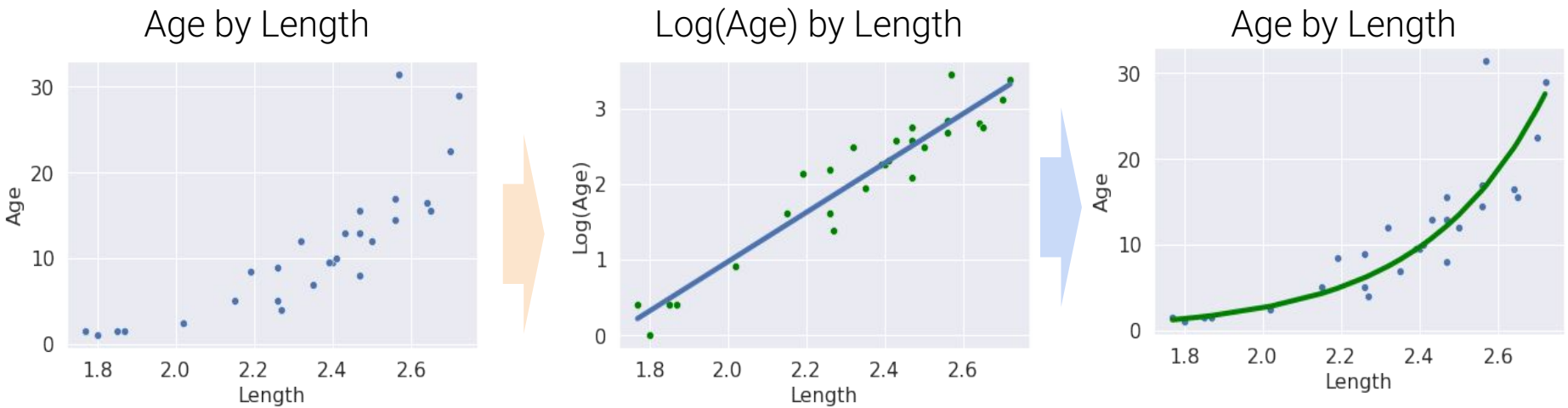
Residual Plot



Fit a Curve using Least Squares Regression

$$z = \log(y)$$

$$y = e^z$$



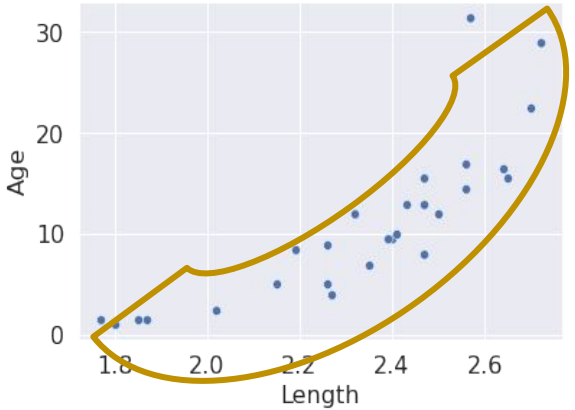
Tukey-Mosteller Bulge Diagram

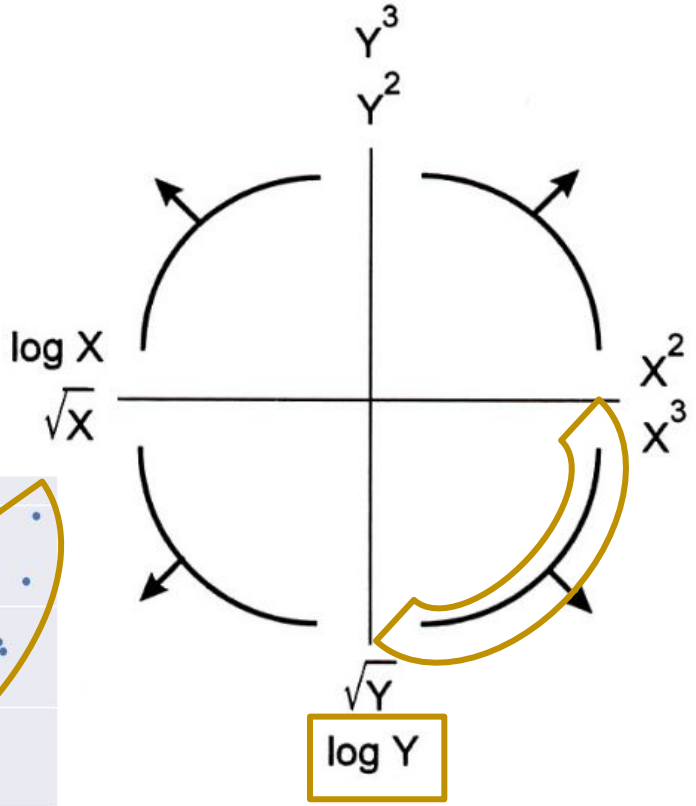
If your data “bulges” in a direction, transform x and/or y in that direction.

- 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.

Age by Length





Introducing Notation for Multiple Linear Regression

Lecture 10, Data 100 Spring 2022

A Note on Terminology

There are several equivalent terms in the context of regression.

Feature(s)

Covariate(s)

Independent variable(s)

Explanatory variable(s)

Predictor(s)

Input(s)

Regressor(s)

Output

Outcome

Response

Dependent variable

Weight(s)

Parameter(s)

Coefficient(s)

Prediction

Predicted response

Estimated value

Estimator(s)

Optimal parameter(s)

Bolded terms are the most common in this course.

Match each column with the appropriate term: $x, y, \hat{y}, \theta, \hat{\theta}$

A Note on Terminology

There are several equivalent terms in the context of regression.

Feature(s)

Covariate(s)

Independent variable(s)

Explanatory variable(s)

Predictor(s)

Input(s)

Regressor(s)

x

Output

Outcome

Response

Dependent variable

y

Weight(s)

Parameter(s)

Coefficient(s)

θ

Prediction

Predicted response

Estimated value

$\hat{y}$

Estimator(s)

Optimal parameter(s)

$\hat{\theta}$

Bolded terms are the most common in this course.

A datapoint (x, y) is also called an **observation**. **(addendum 2/21)**

Multiple Linear Regression

Define the **multiple linear regression** model:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

Parameters are $\theta = (\theta_0, \theta_1, \ldots, \theta_p)$.

Is this linear in θ?

A. no
B. yes
C. maybe

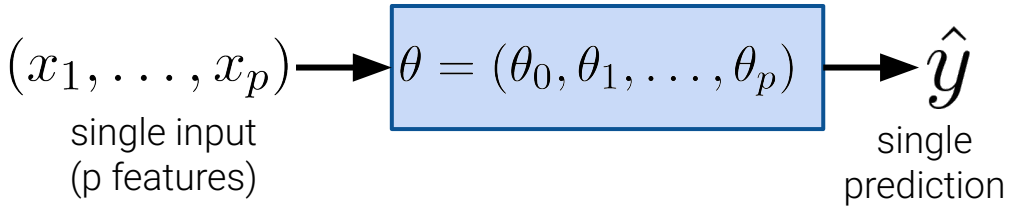
Multiple Linear Regression

Define the **multiple linear regression** model:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

Parameters are $\theta = (\theta_0, \theta_1, \ldots, \theta_p)$.

Yes! This is a **linear combination** of θ_j'S, each scaled by x_j.



Example: Predict dugong ages $\hat{y}$ as a linear model of 2 features:
length x_1 **and** weight x_2.

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

intercept | parameter for length | parameter for weight

Congratulations! You've learned Multiple Linear Regression!

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

Multiple linear regression

In general, the **multiple linear regression** model is of the form

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \ldots + \theta_p x_p = \theta_0 + \sum_{j=1}^{p} \theta_j x_j$$

- We say this model has p **features**, plus an **intercept** term.
- The weight associated with feature x_j is θ_j .

Be careful: x_j here refers to feature j, not data point j.

If we set $x_0 = 1$ for each observation, then we can simplify further:

$$\hat{y} = \sum_{j=0}^{p} \theta_j x_j$$

This is the notation we will use moving forward.

- Think about how you can rewrite this in terms of a vector multiplication!

Multiple regression as a dot product

We previously stated that the multiple regression model was of the form

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \ldots + \theta_p x_p = \theta_0 + \sum_{j=1}^{p} \theta_j x_j$$

This can be restated as a dot product between two vectors.

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} \quad x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

$$\hat{y} = x^T \theta$$

Even though they don't have arrows on top of them, x and θ are still **vectors**!

scalar

Linear Algebra Topics

Going forward, we will assume you have familiarity with the following linear algebra topics:

- [HW01] Dot product
- Vector norm (L2 norm)
- **Span**
- **Orthogonality**
- [HW01] Rank (full column rank)
- [HW01] Linearly independent
- Invertibility

We will briefly cover them as part of next lecture, but it's worth brushing up.

- There will be an Ed post for linear algebra resources posted soon.
- Multiple Linear Regression material is **not** on Midterm 1.
- SLR + Constant Model is midterm content (up through **this lecture**).

A Linear Algebra Preview for Next Time

This is a slide from next lecture. We are not covering it in any detail here; it's included as a preview.

When looking at a **single observation**, our model is

$$\hat{y} = x^T \theta$$

- x is a **vector** of size p + 1.
- $\hat{y}$ is a **scalar**.
- θ is a **vector** of size p + 1.

When looking at **multiple observations**, our model is

$$\hat{\mathbb{Y}} = \mathbb{X}\theta$$

- $\mathbb{X}$ is a **matrix** of size n x (p + 1).
- $\hat{\mathbb{Y}}$ is a **vector** of size n (i.e. $\hat{\mathbb{Y}} \in \mathbb{R}^n$).
- θ is a **vector** of size p + 1.

$$R(\theta) = \frac{1}{n} \underbrace{||\mathbb{Y} - \hat{\mathbb{Y}}||_2^2}_{\text{L2 norm of residual vector}} = \frac{1}{n} ||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

Bonus: Constant Model MSE, Approach 3

MSE minimization using an algebraic trick

It turns out that in this case, there's another rather elegant way of performing the same minimization algebraically, but without using calculus.

- We present this derivation in the next few slides. The lecture video will walk through it in detail.
- In this proof, you will need to use the fact that the **sum of deviations from the mean is 0** (in other words, that $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$). We present that proof here:

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y}) &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \bar{y} \\ &= \sum_{i=1}^{n} y_i - n\bar{y} = \sum_{i=1}^{n} y_i - n \cdot \frac{1}{n}\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i \\ &= 0\end{aligned}$$

For example, this mini-proof shows **1 + 2 + 3 + 4 + 5** is the same as **3 + 3 + 3 + 3 + 3**.

- Our proof will also use the definition of the variance of a sample. As a refresher:

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

Equal to the MSE of the sample mean!

MSE minimization using an algebraic trick

$$
\begin{aligned}
R(\theta) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[(y_i - \bar{y}) + (\bar{y} - \theta)\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[(y_i - \bar{y})^2 + 2(y_i - \bar{y})(\bar{y} - \theta) + (\bar{y} - \theta)^2\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 + 2(\bar{y} - \theta)\sum_{i=1}^{n}(y_i - \bar{y}) + n(\bar{y} - \theta)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 + \frac{2}{n}(\bar{y} - \theta)\cdot 0 + (\bar{y} - \theta)^2 \\
&= \sigma_y^2 + (\bar{y} - \theta)^2
\end{aligned}
$$

variance of sample!

from the previous slide

This proof relies on an algebraic trick. We can write the difference **a - b** as **(a - c) + (c - b)**, where a, b, and c are any numbers.

Using that fact, we can write $y_i - \theta = (y_i - \bar{y}) + (\bar{y} - \theta)$, where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, our sample mean.

Also note: going from line 3 to 4, we distribute the sum to the individual terms. This is a property of sums you should become familiar with!

Minimization using an algebraic trick

In the previous slide, we showed that $R(\theta) = \sigma_y^2 + (\bar{y} - \theta)^2$

- Since variance can't be negative, the first term is greater than or equal to 0.
  - Of note, **the first term doesn't involve θ at all.** Changing our model won't change this value, so for the purposes of determining $\hat{\theta}$, we can ignore it.
- The second term is being squared, and so also must be greater than or equal to 0.
  - This term does involve θ, and so picking the right value of θ will minimize our average loss.
  - We need to pick the θ that sets the second term to 0.
  - This is achieved when $\theta = \bar{y}$. In other words:

$$\hat{\theta} = \bar{y} = \textbf{mean}(y)$$

Looks familiar!