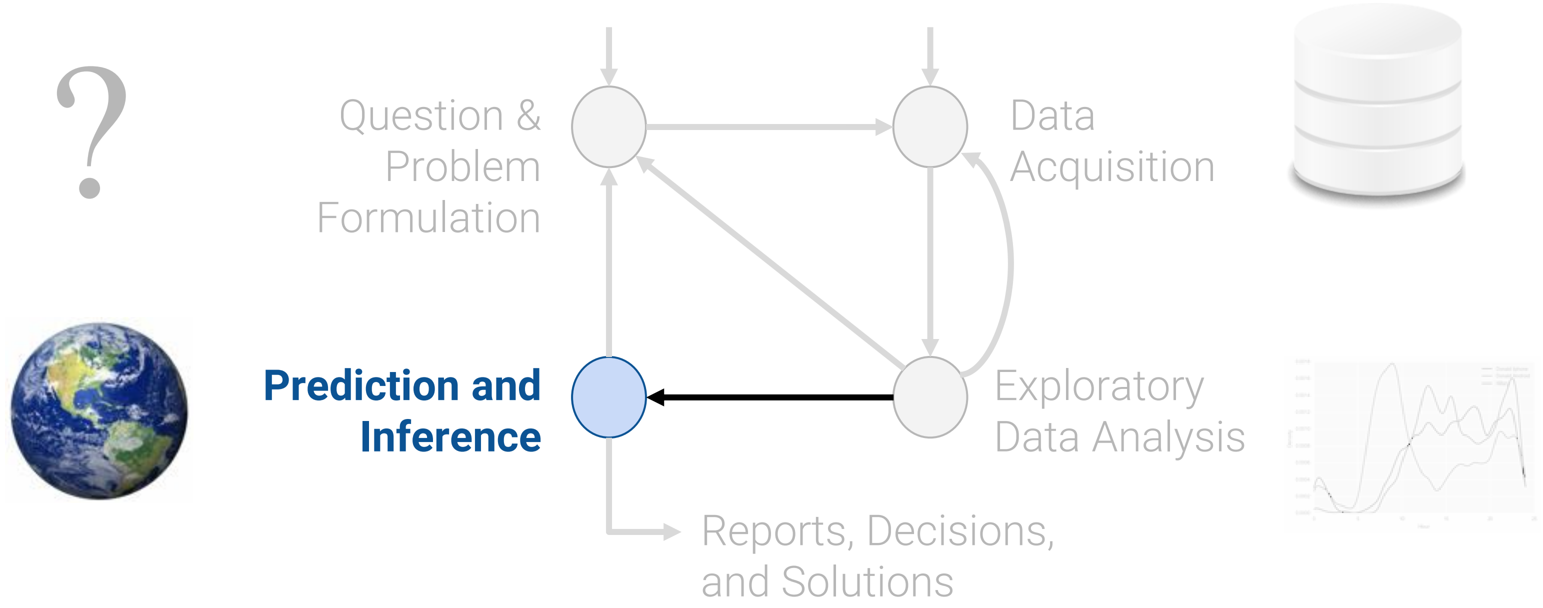
LECTURE 10

Constant Model, Loss, and Transformations

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

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

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

Prediction
Predicted response
Estimated value

$\hat{y}$

Weight(s)
Parameter(s)
Coefficient(s)

θ

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**.

Today's Roadmap

Modeling Process Reiteration

- Evaluating Model the SLR Model
- Iteration 2: Constant Model + MSE
- Iteration 3: Constant Model + MAE

Transformations to Fit Linear Models

Notation for Multiple Linear Regression

Evaluating the Model

Recap from last time...

1. Choose a model

How should we represent the world?

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

SLR model

2. Choose a loss function

How do we quantify prediction error?

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

Squared loss

3. Fit the model

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

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

MSE for SLR

4. Evaluate model performance

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

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

Evaluating 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.

1. **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)

1. **Visualization:**

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

Demo

Four Mysterious Datasets (Anscombe's quartet)

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:

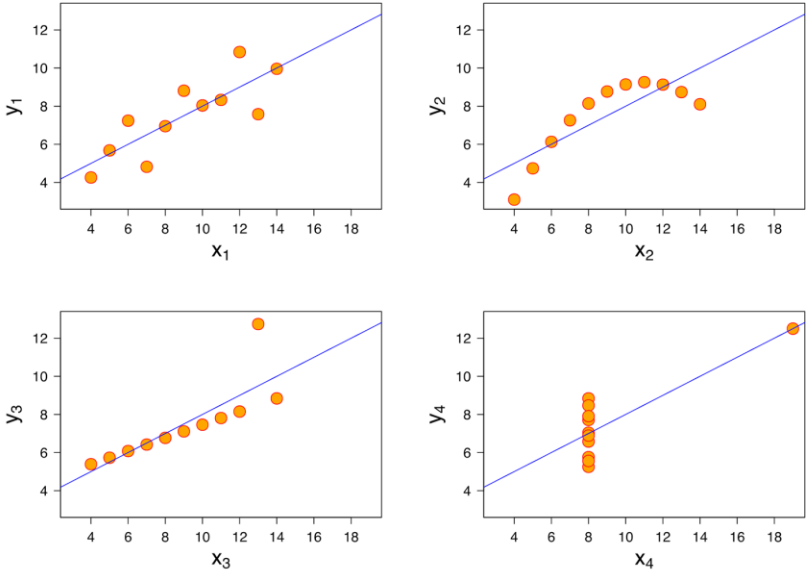
```
x_mean : 9.00, y_mean : 7.50
x_stdev: 3.16, y_stdev: 1.94
r = Correlation(x, y): 0.816
theta_0_hat: 3.00, theta_1_hat: 0.50
RMSE: 1.119
```

Four Mysterious Datasets (Anscombe's quartet)

- **The four dataset** 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** and RMSE.

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

Before modeling, you should always **visualize** your data first!



Demo

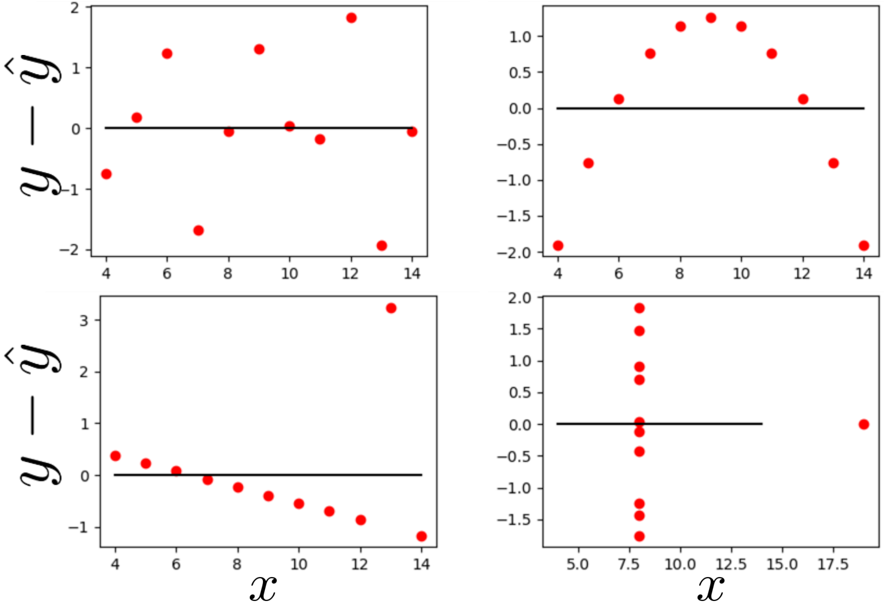
Anscombe's quartet: Residuals

Demo

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**

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



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?

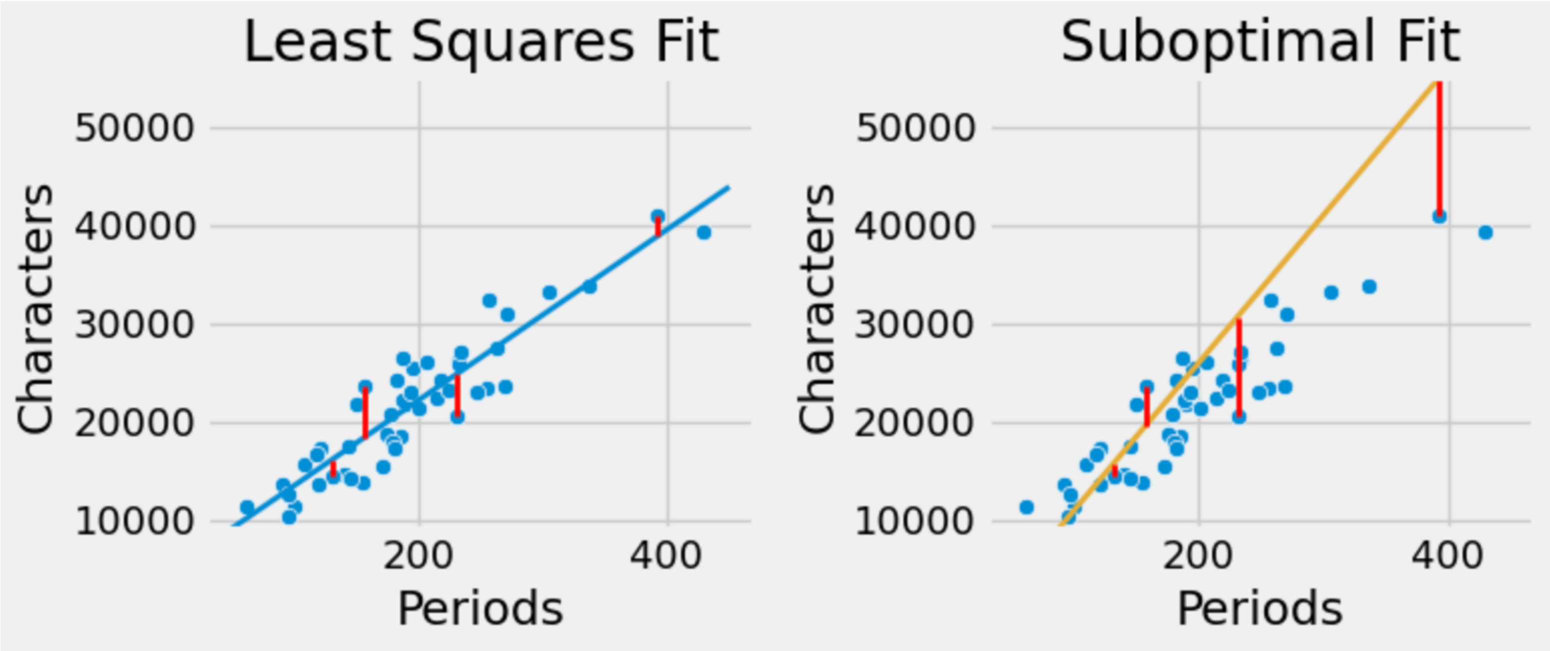
Review of the The Modeling Process (Simple Linear Regression)

1. Choose a model	SLR model	$\hat{y} = \theta_0 + \theta_1 x$
2. Choose a loss function	L2 Loss Mean Squared Error (MSE)	$L(y, \hat{y}) = (y - \hat{y})^2$ $\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overbrace{(\theta_0 + \theta_1 x)}^{\hat{y}_i \text{ (SLR)}})^2$
3. Fit the model	Minimize average loss with calculus	$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x \quad \begin{cases} \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} \\ \hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x} \end{cases}$
4. Evaluate model performance	Visualize, Root MSE	

Minimizing MSE is Minimizing Squared Residuals

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

Lower residuals = better regression 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 calculate model parameters.

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

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

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

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

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

Iteration 2: Constant Model + MSE

Modeling Process Reiteration

- Evaluating Model the SLR Model
- **Iteration 2: Constant Model + MSE**
- Iteration 3: Constant Model + MAE

Transformations to Fit Linear Models

Notation for Multiple Linear Regression

The Modeling Process: Using a Different Model

1. Choose a model

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

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

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

You work at a local boba shop 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

slido

You work at a local boba tea 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}

ⓘ Start presenting to display the poll results on this slide.

The Constant Model

You work at a local boba shop 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 data by always "predicting" the same number—i.e., predicting a constant.

It ignores any relationships between variables:

- For instance, boba 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:

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

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, boba 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:

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

- Our parameter θ_0 is 1-dimensional. $\theta_0 \in \mathbb{R}$
- We now have no input into our model; we predict $\hat{y} = \theta_0$.
- Like before, we can still determine the best θ_0 that minimizes **average loss** on our data.

$$\boxed{\theta_0} \rightarrow \hat{y}$$

The Modeling Process: Using a Different Model

1. Choose a model

~~SLR model~~
$\cancel{\hat{y} = \theta_0 + \theta_1 x}$

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

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~~ $\cancel{\hat{y} = \theta_0 + \theta_1 x}$	Constant Model $\hat{y} = \theta_0$
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 $\mathcal{D} = \{y_1, y_2, \ldots, y_n\}$:

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

Given the **constant model** $\hat{y} = \theta_0$:

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

We **fit the model** by finding the optimal $\hat{\theta}_0$ that minimizes the MSE.

Fit the Model: Three Approaches

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

Approach 1 If you want to prove the general case for any data, you could directly minimize the objective. We can show that average loss is minimized by

$$\hat{\theta}_0 = \text{mean}(y) = \bar{y}$$

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

$$R(\theta) = \frac{1}{5}((20-\theta_0)^2 + (21-\theta_0)^2 + (22-\theta_0)^2 + (29-\theta_0)^2 + (33-\theta_0)^2)$$

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_0:

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

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

Derivative of sum is sum of derivatives

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

Chain rule

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

Simplify constants

2. Set equal to 0.

$$0 = \frac{-2}{n}\sum_{i=1}^{n}\left(y_i - \hat{\theta}_0\right)$$

3. Solve for $\hat{\theta}_0$.

Fit the Model: Calculus for the General Case

1. Differentiate with respect to θ_0:

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

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

Derivative of sum is sum of derivatives

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

Chain rule

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

Simplify constants

2. Set equal to 0.

$$0 = \frac{-2}{n}\sum_{i=1}^{n}\left(y_i - \hat{\theta}_0\right)$$

3. Solve for $\hat{\theta}_0$.

$$0 = \cancel{\frac{-2}{n}}\sum_{i=1}^{n}\left(y_i - \hat{\theta}_0\right) = \sum_{i=1}^{n}\left(y_i - \hat{\theta}_0\right)$$

$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\hat{\theta}_0$$

Separate sums

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

c + c + … + c = nxc

$$n \cdot \hat{\theta}_0 = \sum_{i=1}^{n} y_i$$

$$\hat{\theta}_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right) \Longrightarrow \boxed{\hat{\theta}_0 = \bar{y}}$$

Interpreting $\hat{\theta}_0 = \bar{y}$

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**.

$$R(\hat{\theta}_0) = R(\bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sigma_y^2$$

Note the difference:

$$R(\hat{\theta}_0) = \min_{\theta_0} R(\theta_0) = \sigma_y^2$$

The **minimum value** of constant + MSE

vs

$$\hat{\theta}_0 = \underset{\theta_0}{\operatorname{argmin}}\, R(\theta_0) = \bar{y}$$

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

In modeling, we care less about **minimum loss** $R(\hat{\theta}_0)$ and more about the **minimizer** of loss $\hat{\theta}_0$.

Revisit the Boba Shop Example

You work at a local boba shop 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. $\boxed{25}$
C. 22
D. 100
E. Something else

We will predict the mean of the previous five days' sale:

$(20 + 21 + 22 + 29 + 33)/5 = 25$.

The Modeling Process: Using a Different Model

1. Choose a model	Constant Model	Constant Model $\hat{y} = \theta_0$
2. Choose a loss function	L2 Loss Mean Squared Error (MSE)	$\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_0)^2$
3. Fit the model	Minimize average loss with calculus	$\hat{\theta}_0 = mean(y) = \bar{y}$
4. Evaluate model performance	Visualize, Root MSE	

Demo

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

Suppose we wanted to predict dugong ages.

A Dugong [image source]

Not a Dugong, a Dewgong [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.

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

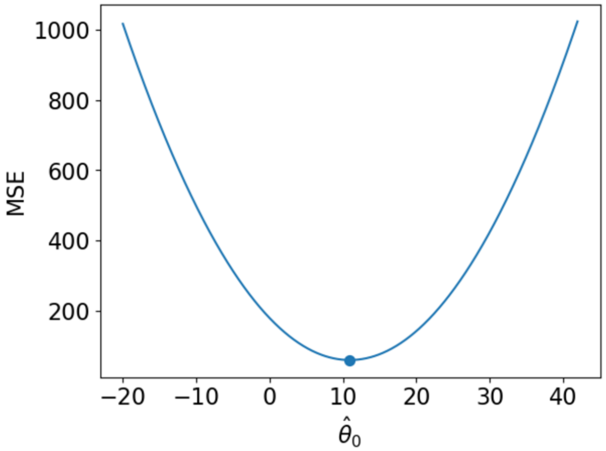
[Loss] Comparing Two Different Models, Both Fit with MSE

Constant Model

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

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

Loss surface is **2-D**.



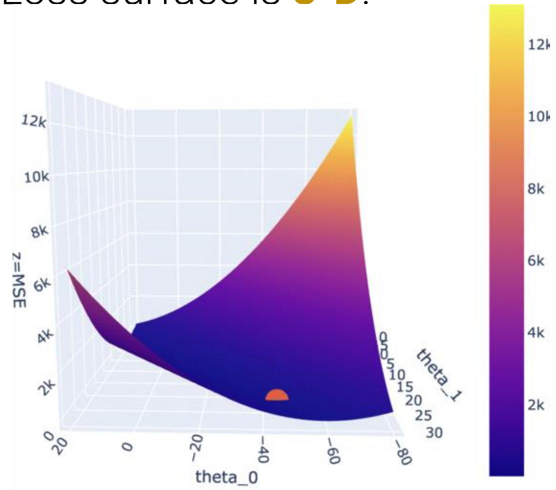
$$\hat{R}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_0)^2$$

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**.



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

Demo

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

Constant Model

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

RMSE: **7.72**

Simple Linear Regression

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

RMSE **4.31**

Interpret the RMSE (Root Mean Square Error):

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

Demo

See notebook for code

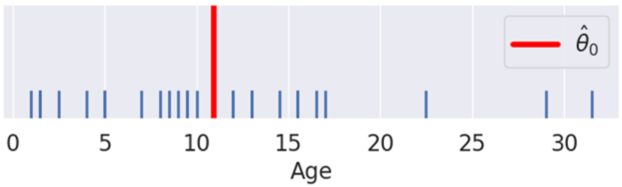
[Fit] Comparing Two Different Models, Both Fit with MSE

Constant Model

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

RMSE: **7.72**

Predictions on a **rug plot**.

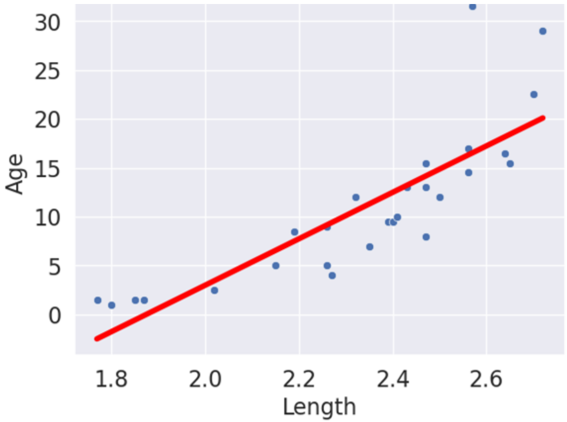


Simple Linear Regression

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

RMSE **4.31**

Predictions on a **scatter plot**.



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

Demo

See notebook for code

Interlude

- Tomorrow’s lecture relies on the geometric interpretation of linear algebra; I recommend watching this 3Blue1Brown video (or better, the entire series) tonight to get a solid understanding of the geometrics of Lin Alg
- Midterm is approaching! More details soon.

Iteration 3: Constant Model + MAE

Modeling Process Reiteration

- Evaluating Model the SLR Model
- Iteration 2: Constant Model + MSE
- **Iteration 3: Constant Model + MAE**

Transformations to Fit Linear Models

Notation for Multiple Linear Regression

The Modeling Process: Using a Different Loss Function

1. Choose a model

Constant Model

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

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	$\hat{y} = \theta_0$
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 $\mathcal{D} = \{y_1, y_2, \ldots, y_n\}$:

$$\hat{R}(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{|y_i - \hat{y}_i|}_{\text{L1 loss on a single datapoint}}$$

Given the **constant model** $\hat{y} = \theta_0$:

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

We **fit the model** by finding the optimal $\hat{\theta}_0$ that minimizes the MAE.

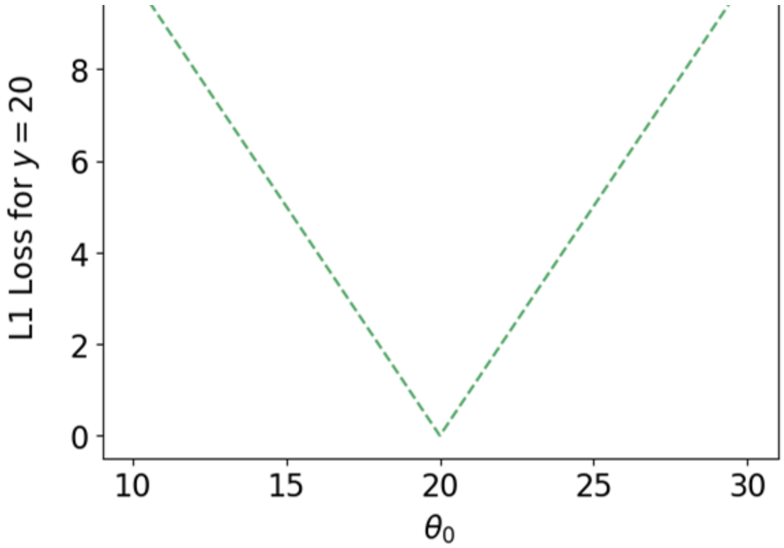
Exploring MAE: A Piecewise function

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

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

Absolute (L1) Loss on one observation:

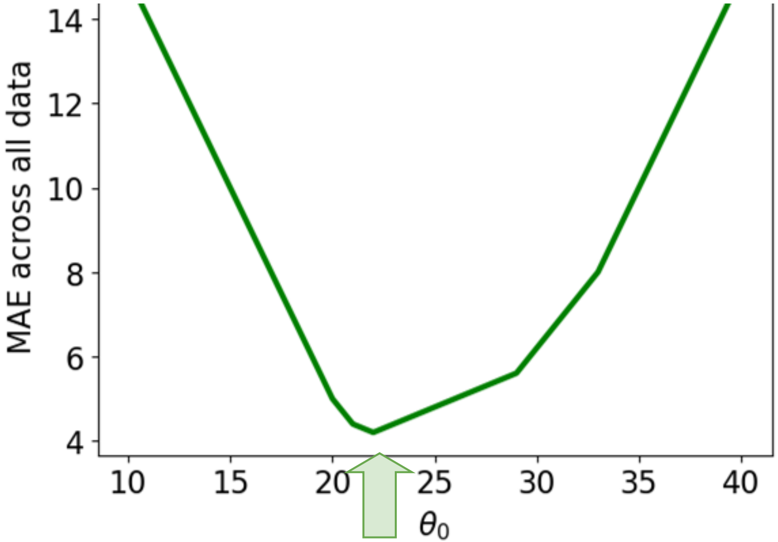
$$L_1(20, \theta_0) = |20 - \theta_0|$$



An absolute value curve, centered at $\hat{\theta}_0$ = 20.

MAE (Mean Absolute Error) across all data:

$$\hat{R}(\theta_0) = \frac{1}{5}(|20-\theta_0| + |21-\theta_0| + |22-\theta_0| + |29-\theta_0| + |33-\theta_0|)$$



Piecewise linear function… minimized at… $\hat{\theta}_0$= 22?

Fit the Model: Calculus

1. Differentiate with respect to $\hat{\theta}_0$.

$$\frac{d}{d\theta_0}R(\theta_0) = \frac{d}{d\theta_0}\frac{1}{n}\sum_{i=1}^{n}|y_i - \theta_0|$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta_0}|y_i - \theta_0|$$

⚠ 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 $\hat{\theta}_0$.

$$\frac{d}{d\theta_0} R(\theta_0) = \frac{d}{d\theta_0} \frac{1}{n} \sum_{i=1}^{n} |y_i - \theta_0|$$

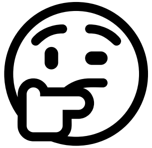
$$= \frac{1}{n} \sum_{i=1}^{n} \frac{d}{d\theta_0} |y_i - \theta_0|$$

$$|y_i - \theta_0| = \begin{cases} y_i - \theta_0 & if\ \theta_0 \leq y_i \\ \theta_0 - y_i & if\ \theta_0 > y_i \end{cases}$$

$$\frac{d}{d\theta_0} |y_i - \theta_0| = \begin{cases} -1 & if\ \theta_0 < y_i \\ 1 & if\ \theta_0 > y_i \end{cases}$$

$$= \frac{1}{n} [\sum_{\theta_0 < y_i} (-1) + \sum_{\theta_0 > y_i} (+1)]$$

Note: The derivative of the absolute value when the argument is 0 (i.e. when $\hat{y} = \theta_0$) 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!

Fit the Model: Calculus

1. Differentiate with respect to $\hat{\theta}_0$.

$$\frac{d}{d\theta_0}R(\theta_0) = \frac{d}{d\theta_0}\frac{1}{n}\sum_{i=1}^{n}|y_i - \theta_0|$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta_0}|y_i - \theta_0|$$

$$|y_i - \theta_0| = \begin{cases} y_i - \theta_0 & if\ \theta_0 \leq y_i \\ \theta_0 - y_i & if\ \theta_0 > y_i \end{cases}$$

$$\frac{d}{d\theta_0}|y_i - \theta_0| = \begin{cases} -1 & if\ \theta_0 < y_i \\ 1 & if\ \theta_0 > y_i \end{cases}$$

$$= \frac{1}{n}\left[\sum_{\theta_0 < y_i}(-1) + \sum_{\theta_0 > y_i}(+1)\right]$$

Sum up for $i = 1, \ldots, n$:

−1 if observation y_i **>** our prediction $\hat{\theta}_0$;

+1 if observation y_i **<** our prediction $\hat{\theta}_0$.

Fit the Model: Calculus

1. Differentiate with respect to $\hat{\theta}_0$.

$$\frac{d}{d\theta_0}R(\theta_0) = \frac{d}{d\theta_0}\frac{1}{n}\sum_{i=1}^{n}|y_i - \theta_0|$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta_0}|y_i - \theta_0|$$

$$|y_i - \theta_0| = \begin{cases} y_i - \theta_0 & if\ \theta_0 \le y_i \\ \theta_0 - y_i & if\ \theta_0 > y_i \end{cases}$$

$$\frac{d}{d\theta_0}|y_i - \theta_0| = \begin{cases} -1 & if\ \theta_0 < y_i \\ 1 & if\ \theta_0 > y_i \end{cases}$$

$$= \frac{1}{n}\left[\sum_{\theta_0 < y_i}(-1) + \sum_{\theta_0 > y_i}(+1)\right]$$

2. Set equal to 0.

$$0 = \frac{1}{n}\sum_{\hat{\theta}_0 < y_i}(-1) + \frac{1}{n}\sum_{\hat{\theta}_0 > y_i}(1)$$

3. Solve for $\hat{\theta}_0$.

$$0 = -\frac{1}{n}\sum_{\hat{\theta}_0 < y_i}(1) + \frac{1}{n}\sum_{\hat{\theta}_0 > y_i}(1)$$

$$\sum_{\hat{\theta}_0 < y_i}(1) = \sum_{\hat{\theta}_0 > y_i}(1)$$

Where do we go from here?

Median Minimizes MAE for the Constant Model

The constant model parameter $\theta = \hat{\theta}_0$ that minimizes MAE must satisfy:

$$\underbrace{\sum_{\hat{\theta}_0 < y_i} (1)}_{\substack{\text{\# observations} \\ \textbf{greater than } \hat{\theta}_0}} = \underbrace{\sum_{\hat{\theta}_0 > y_i} (1)}_{\substack{\text{\# observations} \\ \textbf{less than } \hat{\theta}_0}}$$

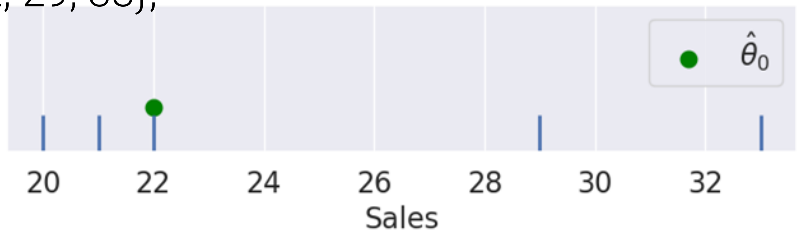
In other words, theta needs to be such that there are **an equal # of points to the left and right**.

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

$$\boxed{\hat{\theta}_0 = median(y)}$$

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 θ.

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

1. Differentiate with respect to θ.
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 $R''(\hat{\theta}) > 0$.
- 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	$\hat{y} = \theta_0$
2. Choose a loss function	L1 Loss Mean Absolute Error (MAE)	
3. Fit the model	Minimize average loss with calculus	$\hat{R}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} \lvert y_i - \theta_0 \rvert$ $\hat{\theta}_0 = median(y)$
4. Evaluate ~~model performance~~ loss	Visualize, ~~Root MSE~~	

Demo

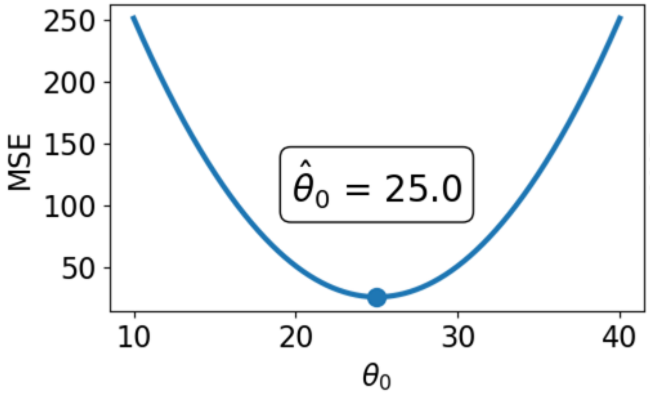
MSE and MAE: Comparing Optimal Parameters

MSE (Mean Squared Loss)

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

Minimized with **sample mean**:

$$\hat{\theta}_0 = \text{mean}(y) = \bar{y}$$

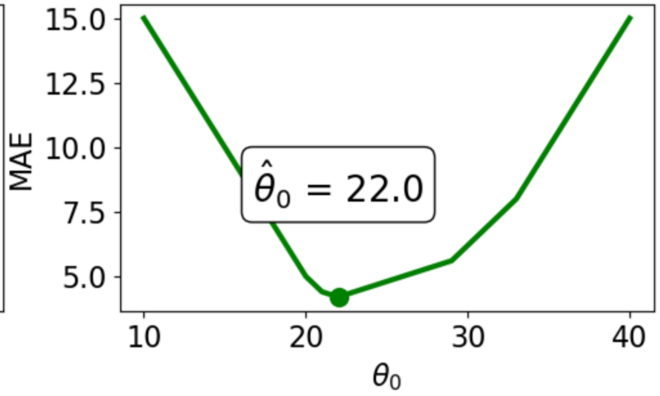


MAE (Mean Absolute Loss)

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

Minimized with **sample median**:

$$\hat{\theta}_0 = \text{median}(y)$$

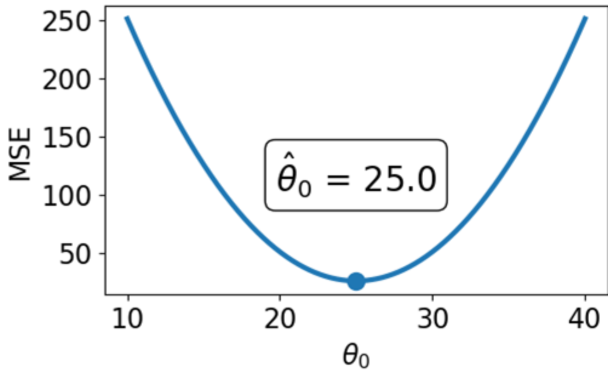


MSE and MAE: Comparing Loss Surfaces

Demo

MSE (Mean Squared Loss)

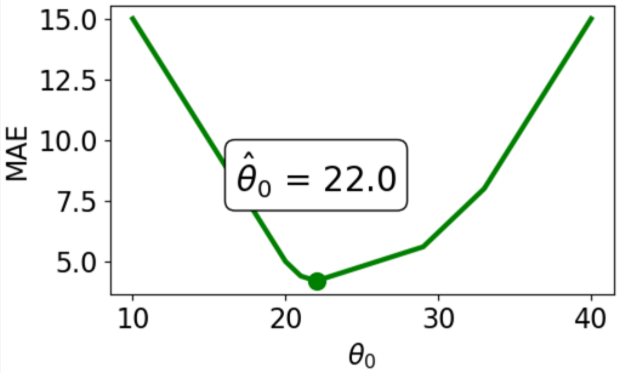
$$\hat{\theta}_0 = mean(y) = \bar{y}$$



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

MAE (Mean Absolute Loss)

$$\hat{\theta}_0 = median(y)$$



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

MSE and MAE: Comparing Sensitivity to Outliers

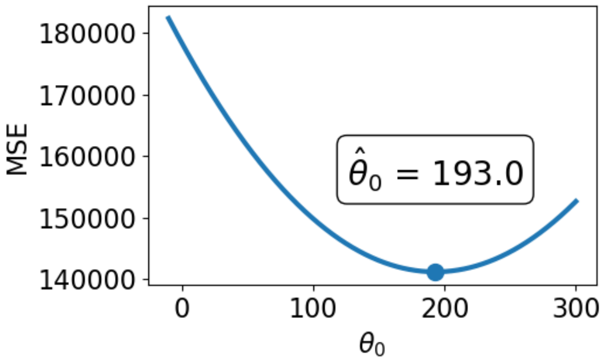
MSE (Mean Squared Loss)
Minimized with **sample mean**:

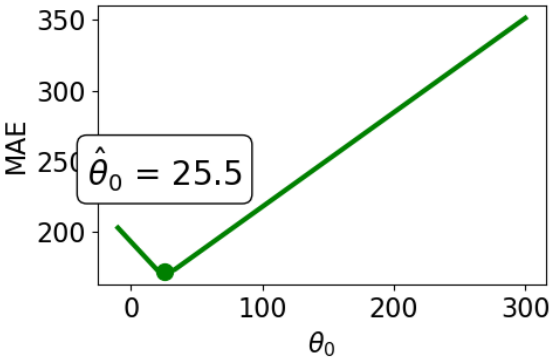
$$\hat{\theta}_0 = \text{mean}(y) = \bar{y}$$

MAE (Mean Absolute Loss)
Minimized with **sample median**:

$$\hat{\theta}_0 = \text{median}(y)$$

data = {20, 21, 22, 29, 33, **1033**}





⚠ **Sensitive** to outliers (since they change mean substantially). Sensitivity also depends on the dataset size.

More robust to outliers.

Demo

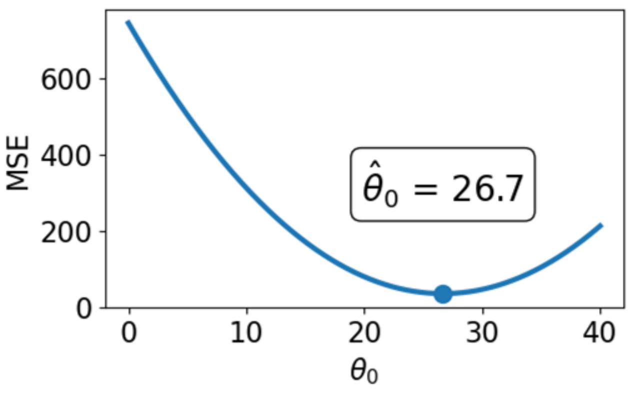
Demo

MSE and MAE: Comparing Uniqueness of Solutions

MSE (Mean Squared Error)

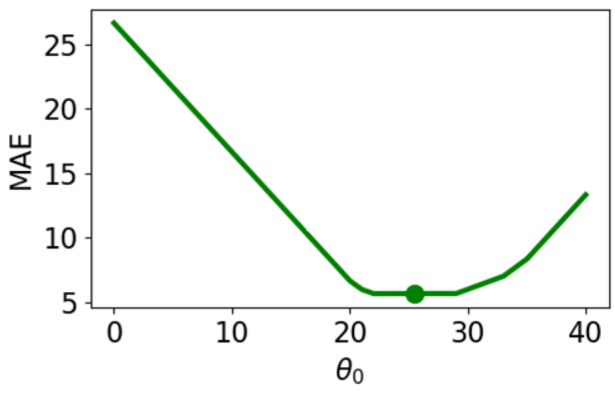
MAE (Mean Absolute Error)

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



Unique $\hat{\theta}_0$:

$$\hat{\theta}_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)$$



⚠ **Infinitely many** $\hat{\theta}_0$ **s**. Any $\hat{\theta}_0$ in range (22, 29) minimizes MAE.

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

slido

The best estimator for a constant model with MAE loss is the ------ of the y values.

ⓘ Start presenting to display the poll results on this slide.

Transformations to Fit Linear Models

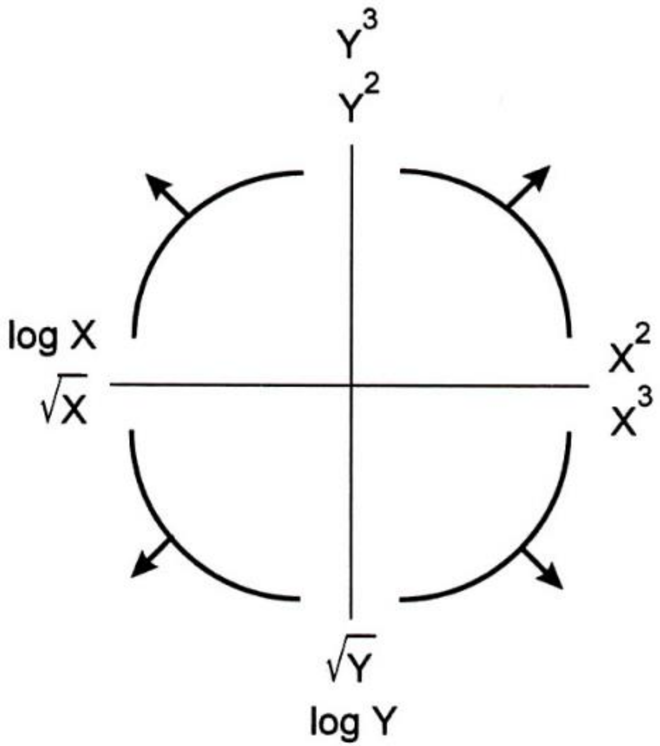
Tukey-Mosteller Bulge Diagram (From Lecture 7)

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

- There are multiple solutions. Some will fit better than others.
- sqrt and log make a value “smaller”.
- Raising a value to a power makes it “bigger”.
- Each of these transformations equates to increasing or decreasing the scale of an axis.

Other goals other than linearity are possible

- E.g. make data appear more symmetric.
- Linearity allows us to fit lines to the transformed data



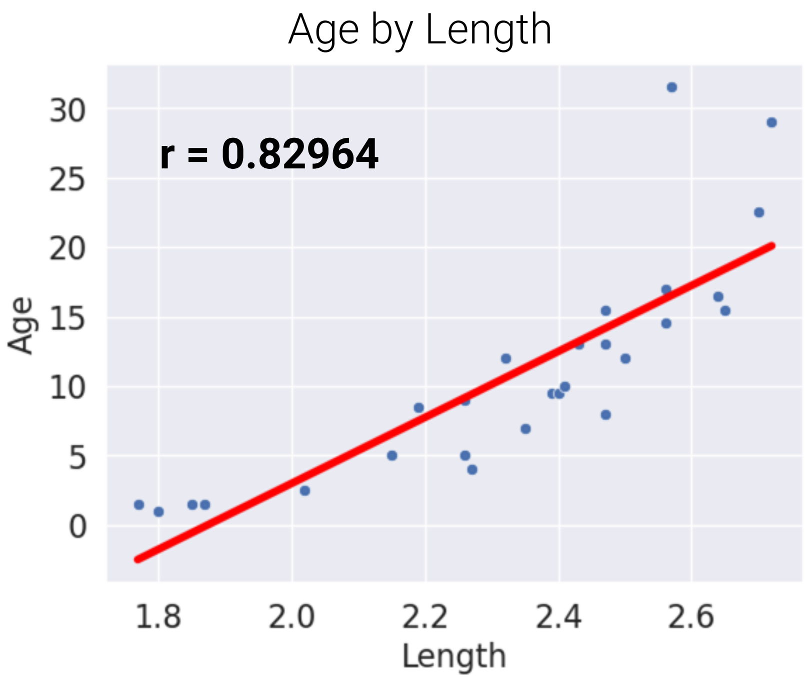
Back to Least Squares Regression with Dugongs

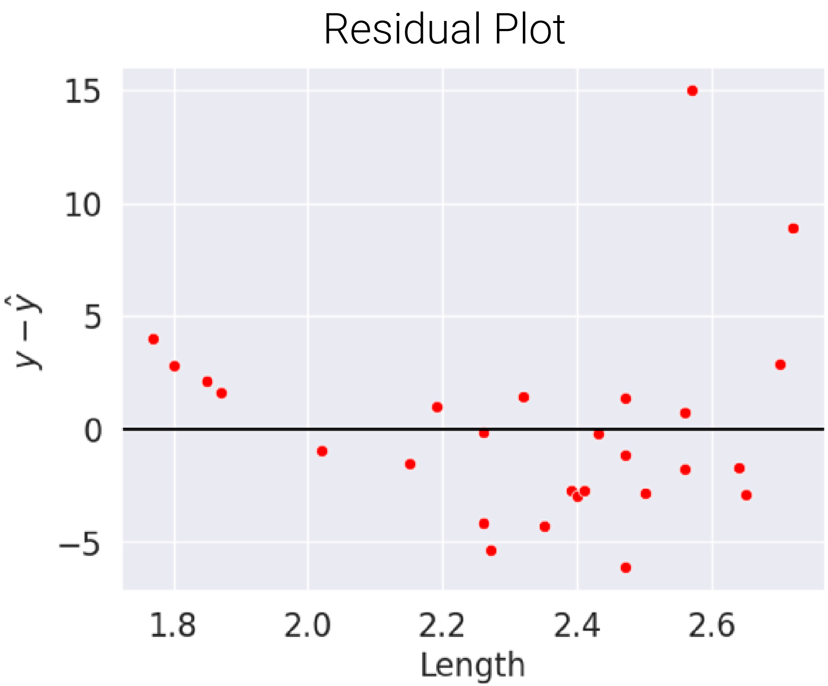
From Data 8 (textbook):

> 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 scatter plot **curves upward**.

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

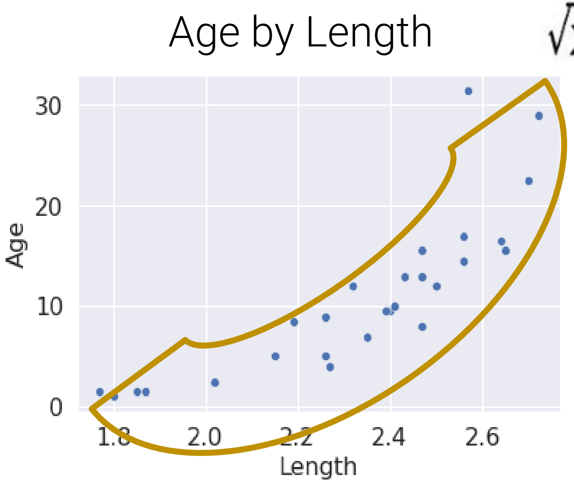
A: **Transform the Data**.

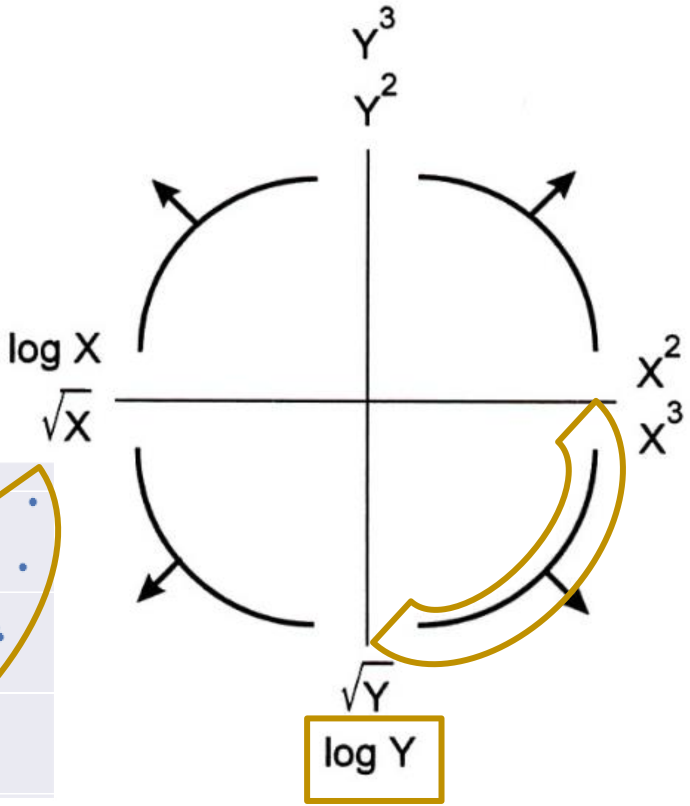
Tukey-Mosteller Bulge Diagram

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

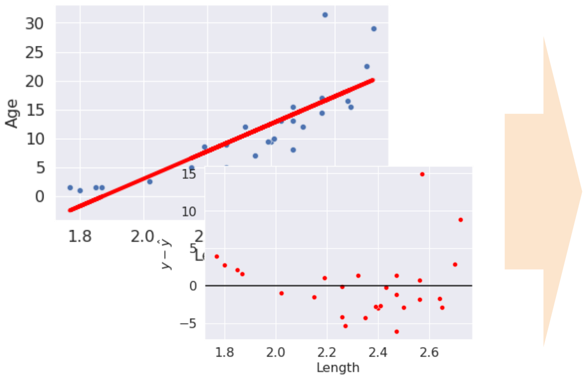
In other words, if we apply the variable transform $z = \log(y)$:

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

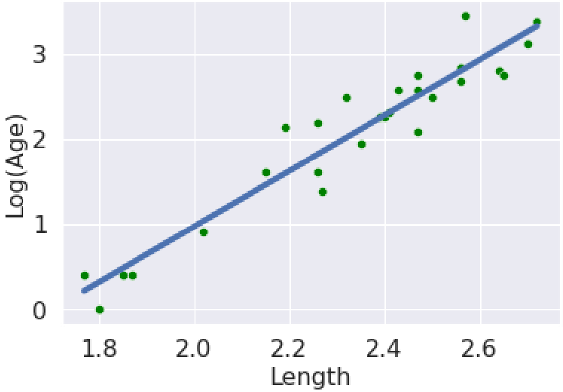
$$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)^2$$

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

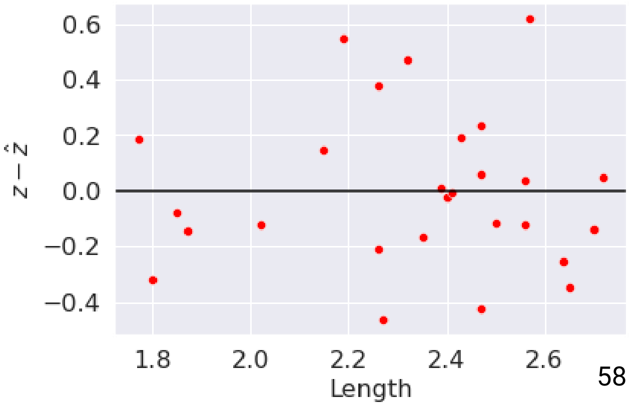
Original (Age by Length)



Log(Age) by Length

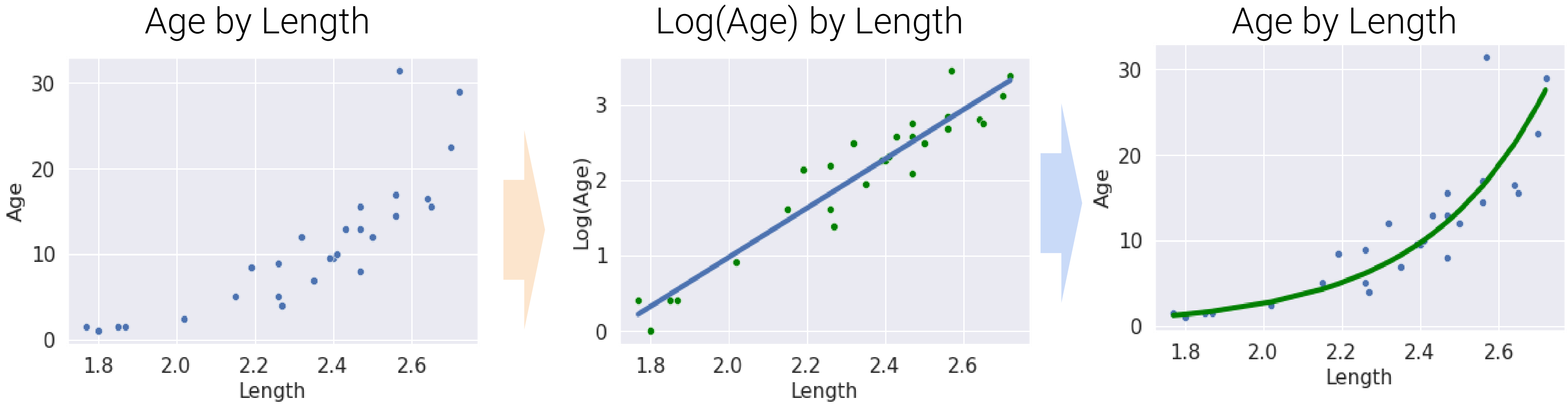


Residual Plot



Fit a Curve using Least Squares Regression

$$z = \log(y) \quad \Rightarrow \quad \hat{y} = e^{\hat{z}} = e^{\theta_0 + \theta_1 x}$$



Notation for Multiple Linear Regression

Modeling Process Reiteration

- Evaluating Model the SLR Model
- Iteration 2: Constant Model + MSE
- Iteration 3: Constant Model + MAE

Transformations to Fit Linear Models

Notation for Multiple Linear Regression

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

Prediction
Predicted response
Estimated value

$\hat{y}$

Weight(s)
Parameter(s)
Coefficient(s)

θ

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**.

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

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.

$$(x_1, \ldots, x_p) \longrightarrow \boxed{\theta = [\theta_0, \theta_1, \ldots, \theta_p]} \longrightarrow \hat{y}$$

single input
(p features)

single
prediction

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!

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

More on Multiple Linear Regression tomorrow

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.
- 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

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

from the previous slide

variance of sample!

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 wher $\theta = \bar{y}$. In other words:

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

Looks familiar!