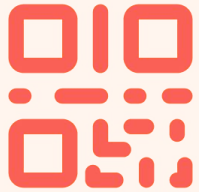


slido



**Join at slido.com
#3280763**

① Start presenting to display the joining instructions on this slide.

LECTURE 11

Constant Model, Loss, and Transformations

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

Data 100/Data 200, Spring 2023 @ UC Berkeley

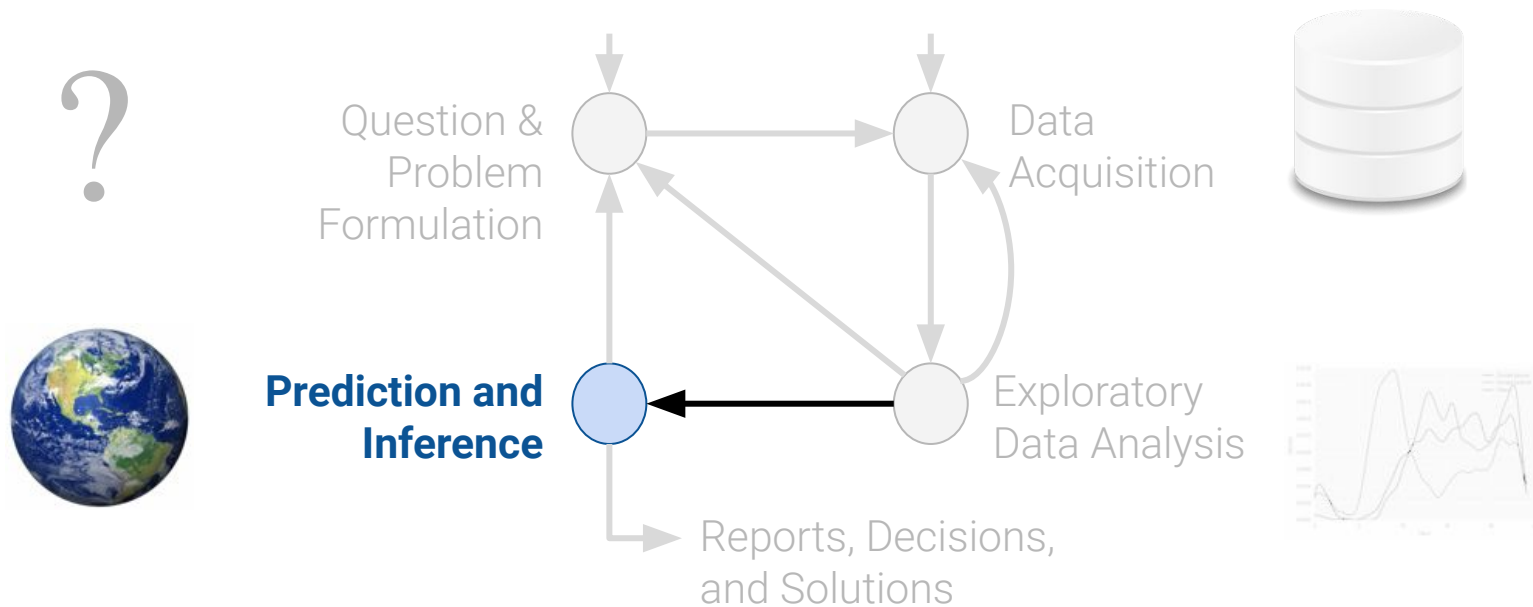
Narges Norouzi and Lisa Yan

Content credit: [Acknowledgments](#)



3280763

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 11, Data 100 Spring 2023

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



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?



3280763

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)

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

3. Fit the model

Minimize average loss with calculus

$$\hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^n (y_i - \underbrace{(\theta_0 + \theta_1 x)}_{\hat{y}_i \text{ (SLR)}})^2$$

4. Evaluate model performance

Visualize, Root MSE

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x \quad \left\{ \begin{array}{l} \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} \\ \hat{\theta}_1 = r \frac{\sigma_y}{\sigma_x} \end{array} \right.$$



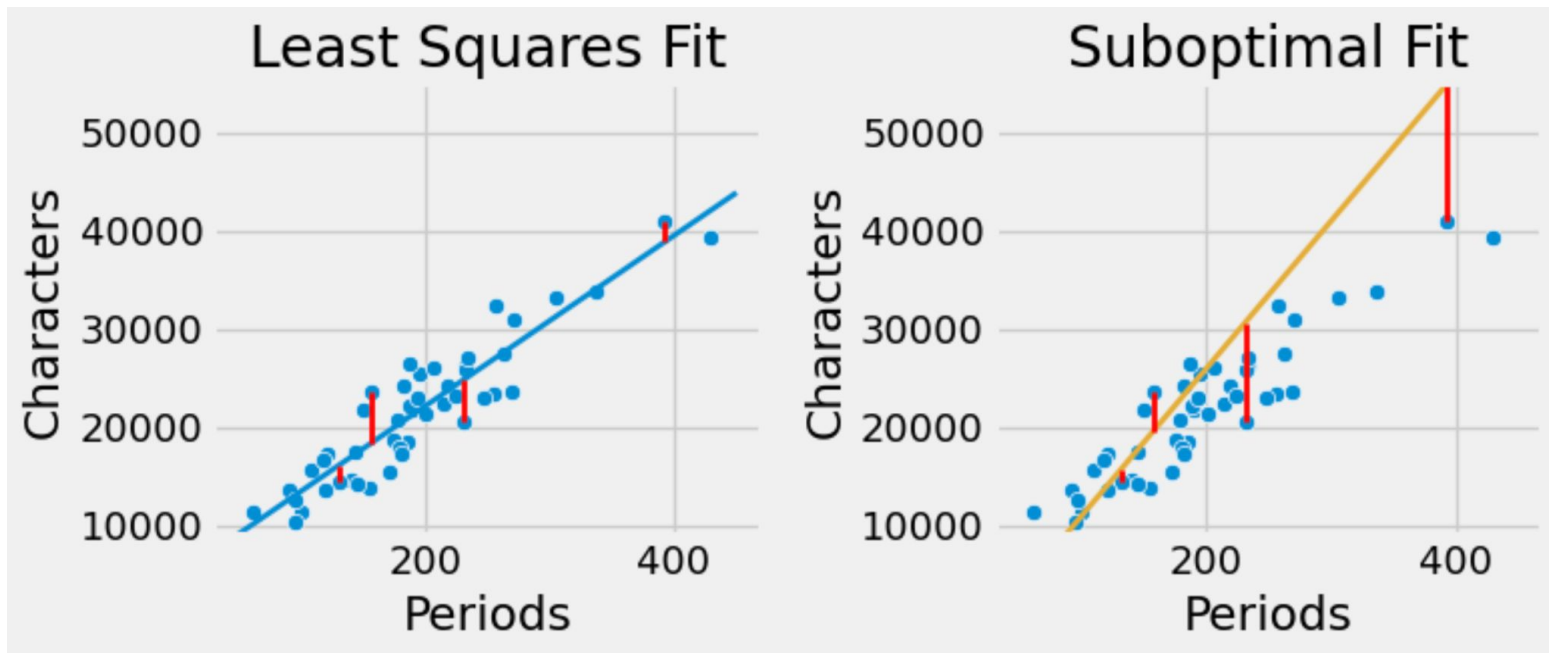
3280763

Minimizing MSE is Minimizing Squared Residuals

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

Residual ("error") in prediction

Lower residuals = better MSE fit!





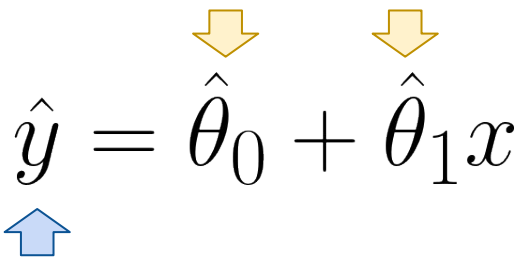
Terminology: Prediction vs. Estimation

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

Estimation is the task of using data to 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.



Changing the Model: Constant Model + MSE

Lecture 11, Data 100 Spring 2023

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



1. Choose a model

~~SLR model~~

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

Constant Model?

$$\hat{y} = ??$$

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 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:

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

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}

① Start presenting to display the poll results on this slide.

The Constant Model



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:

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

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

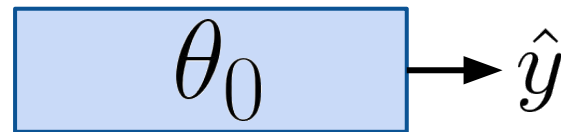
It ignores any relationships between variables.

- 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.





3280763

The Modeling Process: Using a Different Model



1. Choose a model

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

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



3280763

The Modeling Process: Using a Different Model

1. Choose a model



~~SLR model~~

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

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 $\mathcal{D} = \{y_1, y_2, \dots, y_n\}$:

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

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.



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

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.



1. Differentiate with respect to θ_0 :

$$\begin{aligned}\frac{d}{d\theta_0}R(\theta) &= \frac{d}{d\theta_0}\left(\frac{1}{n}\sum_{i=1}^n(y_i - \theta_0)^2\right) \\ &= \frac{1}{n}\sum_{i=1}^n \underbrace{\frac{d}{d\theta_0}(y_i - \theta_0)^2}_{\text{Chain rule}} \quad \text{Derivative of sum is sum of derivatives} \\ &= \frac{1}{n}\sum_{i=1}^n 2(y_i - \theta_0)(-1) \quad \text{Chain rule} \\ &= \frac{-2}{n}\sum_{i=1}^n (y_i - \theta_0) \quad \text{Simplify constants}\end{aligned}$$

2. Set equal to 0.

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

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



Fit the Model: Calculus for the General Case

1. Differentiate with respect to θ_0 :

$$\begin{aligned}\frac{d}{d\theta_0} R(\theta) &= \frac{d}{d\theta_0} \left(\frac{1}{n} \sum_{i=1}^n (y_i - \theta_0)^2 \right) \\ &= \frac{1}{n} \sum_{i=1}^n \frac{d}{d\theta_0} (y_i - \theta_0)^2 && \text{Derivative of sum is sum of derivatives} \\ &= \frac{1}{n} \sum_{i=1}^n 2(y_i - \theta_0)(-1) && \text{Chain rule} \\ &= \frac{-2}{n} \sum_{i=1}^n (y_i - \theta_0) && \text{Simplify constants}\end{aligned}$$

2. Set equal to 0.

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

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

$$\begin{aligned}0 &= \cancel{\frac{-2}{n}} \sum_{i=1}^n (y_i - \theta_0) = \sum_{i=1}^n (y_i - \theta_0) \\ &= \sum_{i=1}^n y_i - \sum_{i=1}^n \theta_0 && \text{Separate sums} \\ &= \left(\sum_{i=1}^n y_i \right) - n \times \theta_0 && c + c + \dots + c = n \times c \\ n \times \theta_0 &= \left(\sum_{i=1}^n y_i \right) \\ \hat{\theta}_0 &= \frac{1}{n} \left(\sum_{i=1}^n y_i \right) \implies \boxed{\hat{\theta}_0 = \bar{y}}\end{aligned}$$



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:

$$R(\hat{\theta}_0) = \min_{\theta_0} R(\theta_0) = \sigma_y^2 \quad \text{vs} \quad \hat{\theta}_0 = \operatorname{argmin}_{\theta_0} R(\theta_0) = \bar{y}$$

The **minimum value** of
constant + MSE

The **argument** that **minimizes**
constant + MSE

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



3280763

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

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

4. Evaluate model performance

Visualize, Root MSE



Suppose we wanted to predict dugong ages.



[\[image source\]](#)

Compare

Constant Model

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

Data: Sample of ages.

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

Simple Linear Regression

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

Data: Sample of (length, age)s.

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



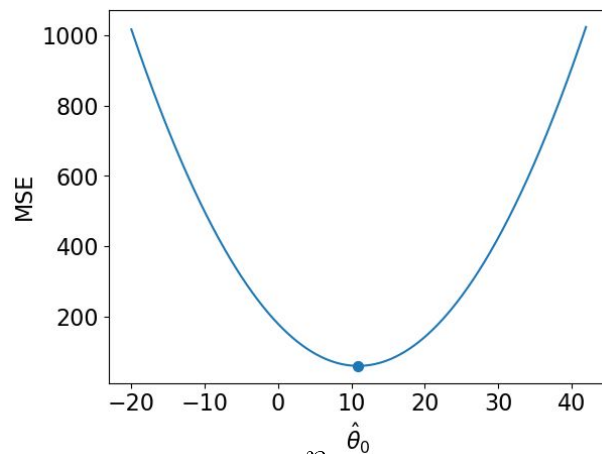
Compare

Constant Model

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

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

Loss surface is **2-D**.



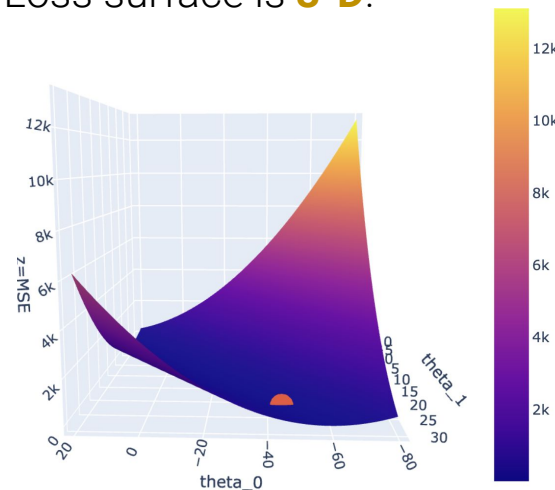
$$\hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^n (y_i - \theta_0)^2$$

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**.



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



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)

Compare

See notebook for code

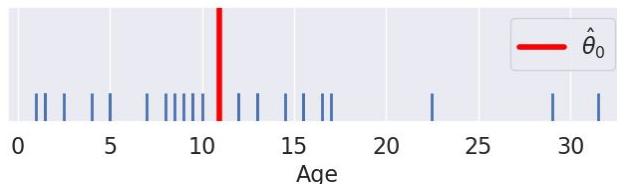


Constant Model

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

RMSE: 7.72

Predictions on a **rug plot**.

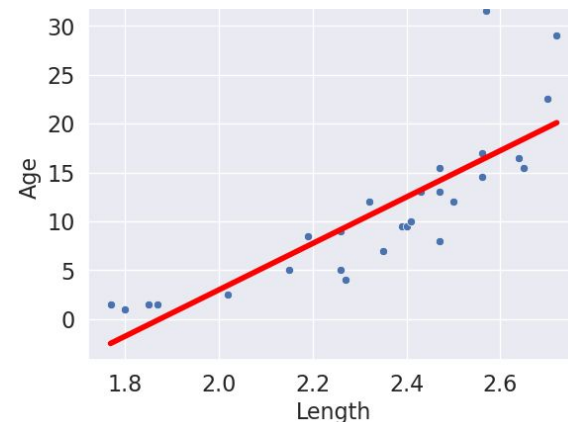


Simple Linear Regression

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

RMSE 4.31

Predictions on a **scatter plot**.



Compare

See notebook for code

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

slido



**The best estimator for a
constant model with MSE loss
is the mean of the y values**

① Start presenting to display the poll results on this slide.



Changing the Loss: Constant Model + MAE

Lecture 11, Data 100 Spring 2023

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



3280763

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



3280763

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



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

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

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.



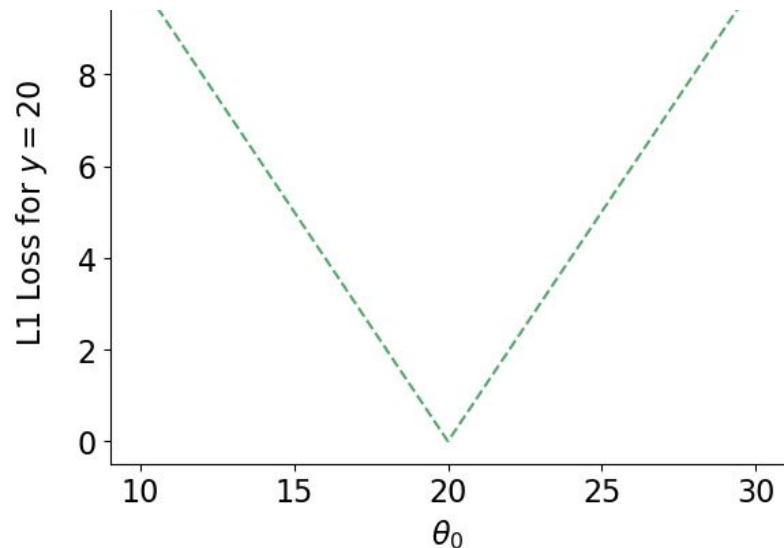
Exploring MAE: A Piecewise function

For the boba tea 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:

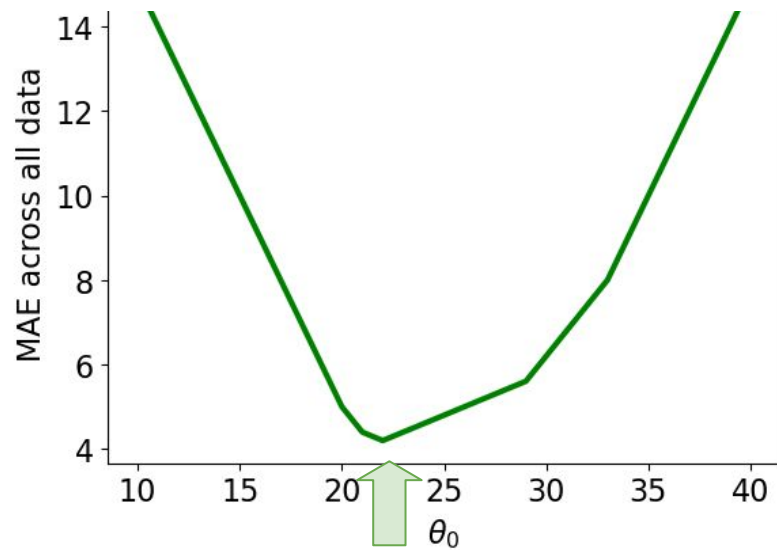
$$L_1(20, \theta_0) = |20 - \theta_0|$$



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... $\hat{\theta}_0 = 22$?



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

$$\begin{aligned}\frac{1}{d\theta_0}R(\theta_0) &= \frac{d}{d\theta_0}\left(\frac{1}{n}\sum_{i=1}^n |y_i - \theta_0|\right) \\ &= \frac{1}{n}\sum_{i=1}^n \frac{d}{d\theta_0}|y_i - \theta_0|\end{aligned}$$



Absolute value!

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



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

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

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

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

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

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

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!



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

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

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

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

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

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

Sum up for $i = 1, \dots, n$:
 -1 if observation y_i > our prediction $\hat{\theta}_0$;
 +1 if observation y_i < our prediction $\hat{\theta}_0$.



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

$$\begin{aligned} \frac{1}{d\theta_0} R(\theta_0) &= \frac{d}{d\theta_0} \left(\frac{1}{n} \sum_{i=1}^n |y_i - \theta_0| \right) \\ &= \frac{1}{n} \sum_{i=1}^n \frac{d}{d\theta_0} |y_i - \theta_0| \end{aligned}$$

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

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

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

2. Set equal to 0.

$$0 = \frac{1}{n} \left[\sum_{\theta_0 < y_i} (-1) + \sum_{\theta_0 > y_i} (+1) \right]$$

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

$$0 = - \sum_{\theta_0 < y_i} 1 + \sum_{\theta_0 > y_i} 1$$

$$\sum_{\theta_0 < y_i} 1 = \sum_{\theta_0 > y_i} 1$$

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:

$$\underbrace{\sum_{\theta_0 < y_i} 1}_{\substack{\text{\# observations} \\ \text{\textbf{greater than}} \hat{\theta}_0}} = \underbrace{\sum_{\theta_0 > y_i} 1}_{\substack{\text{\# observations} \\ \text{\textbf{less than}} \hat{\theta}_0}}$$

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**!

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

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 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 $R''(\hat{\theta}) > 0$.
- 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.



3280763

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

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

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

4. Evaluate model performance loss

Visualize,
~~Root MSE~~



3280763

MSE (Mean Squared Loss)

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

Minimized with **sample mean**:

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

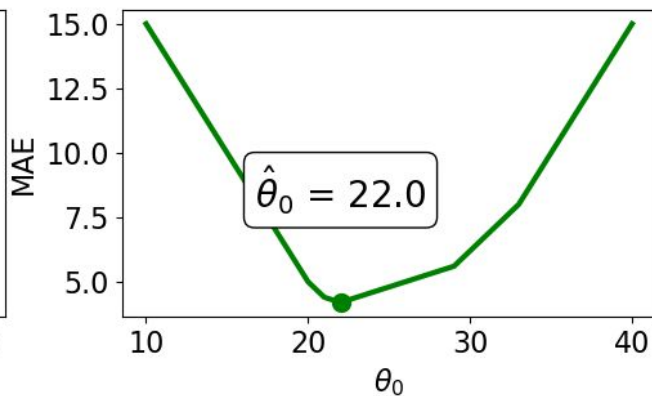
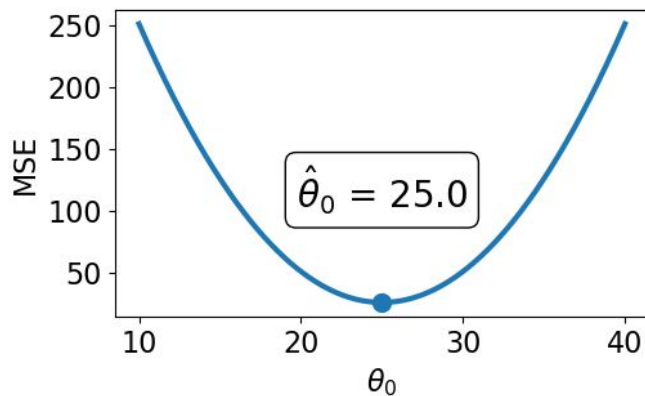
MAE (Mean Absolute Loss)

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

Minimized with **sample median**:

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

Compare

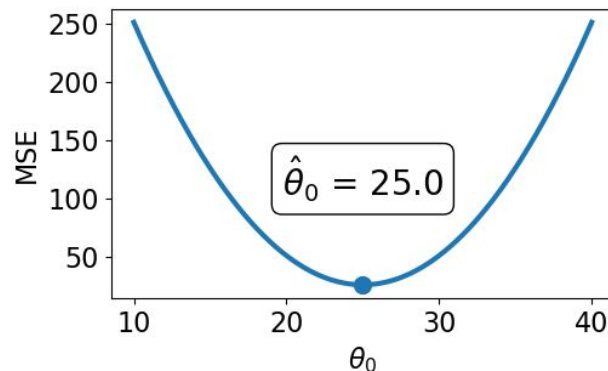




Compare

MSE (Mean Squared Loss)

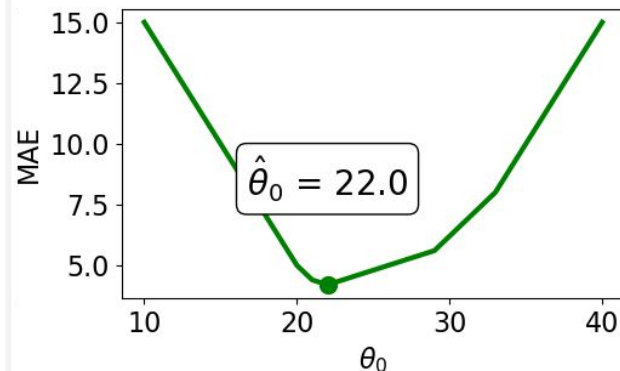
$$\hat{\theta}_0 = \text{mean}(y) = \bar{y}$$



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

MAE (Mean Absolute Loss)

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



! Piecewise. at each of the “kinks,” it’s not differentiable. Harder to minimize.



3280763

MSE (Mean Squared Loss)

Minimized with **sample mean**:

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

! **Sensitive** to outliers (since they change mean substantially).

Sensitivity also depends on the dataset size.

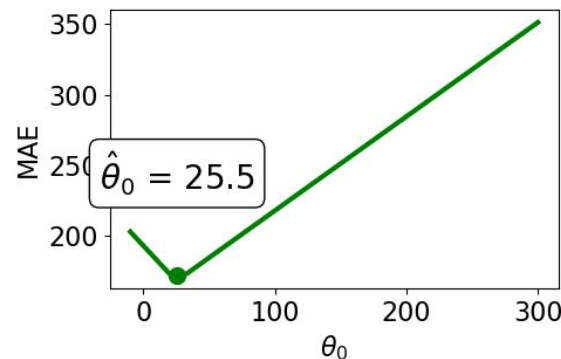
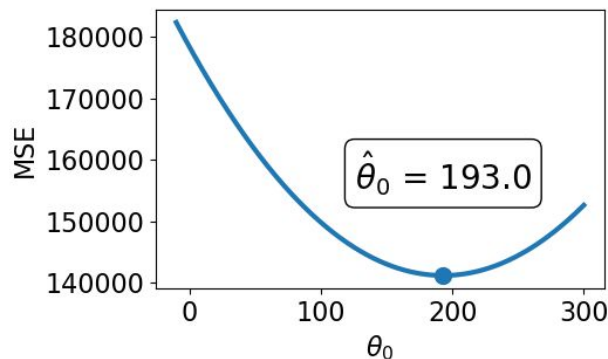
MAE (Mean Absolute Loss)

Minimized with **sample median**:

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

More robust to outliers.

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



Compare

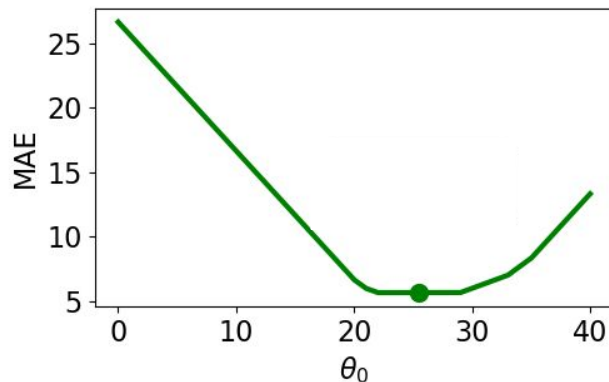
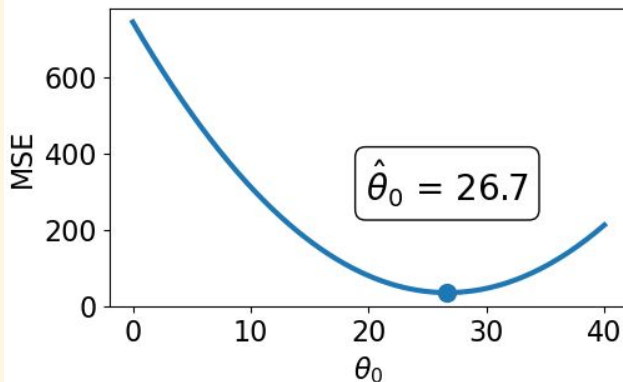


MSE (Mean Squared Error)

MAE (Mean Absolute Error)

Suppose we add a 6th observation to our bubble tea dataset:

{20, 21, 22, 29, 33, **35**}



Compare

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.

① Start presenting to display the poll results on this slide.



Revisiting SLR Evaluation (from Lecture 10)

Lecture 11, Data 100 Spring 2023

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



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



3280763

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**.

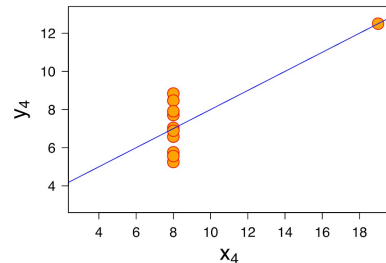
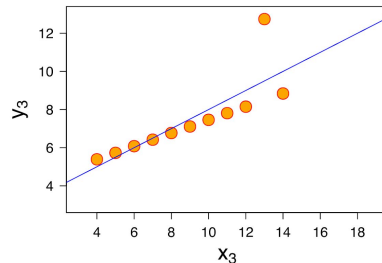
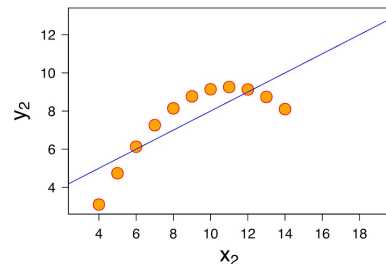
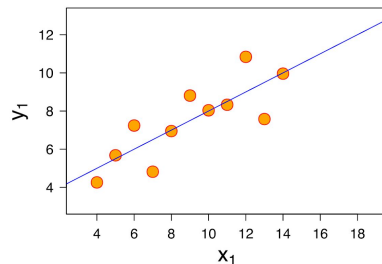
However, only one of these four sets of data makes sense to model using SLR.

Before modeling, you should always visualize your data first!

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

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

$$r = 0.816$$





3280763

Four Mysterious Datasets + Least Squares

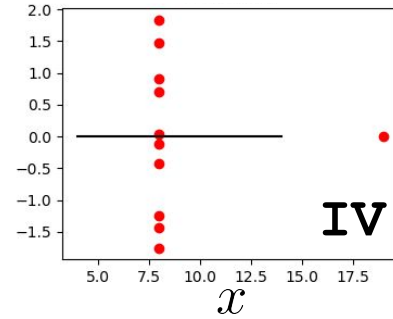
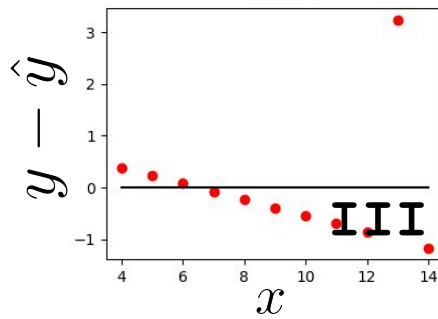
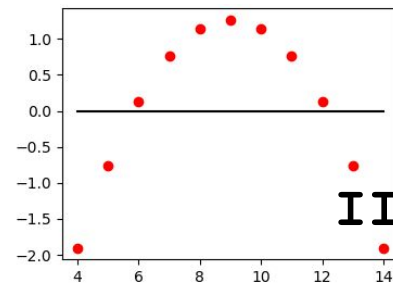
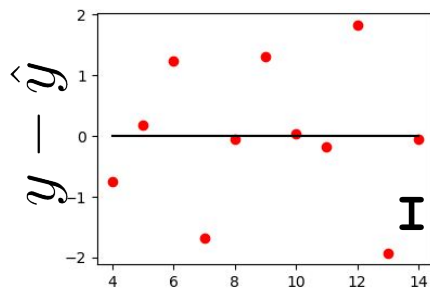
Ideal model evaluation steps, in order:

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

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

$x_{\text{mean}} : 9.00$, $y_{\text{mean}} : 7.50$
 $x_{\text{stdev}} : 3.16$, $y_{\text{stdev}} : 1.94$
 $r = \text{Correlation}(x, y) : 0.816$
 $\hat{a} : 3.00$, $\hat{b} : 0.50$
RMSE: 1.119

3. Residual Visualization



From Data 8 ([textbook](#)):

The residual plot of a good regression shows no pattern.



Transformations to Fit Linear Models

Lecture 11, Data 100 Spring 2023

Modeling Process Review

Changing the Model: Constant Model + MSE

Changing the Loss: Constant Model + MAE

Revisiting SLR Evaluation

Transformations to Fit Linear Models

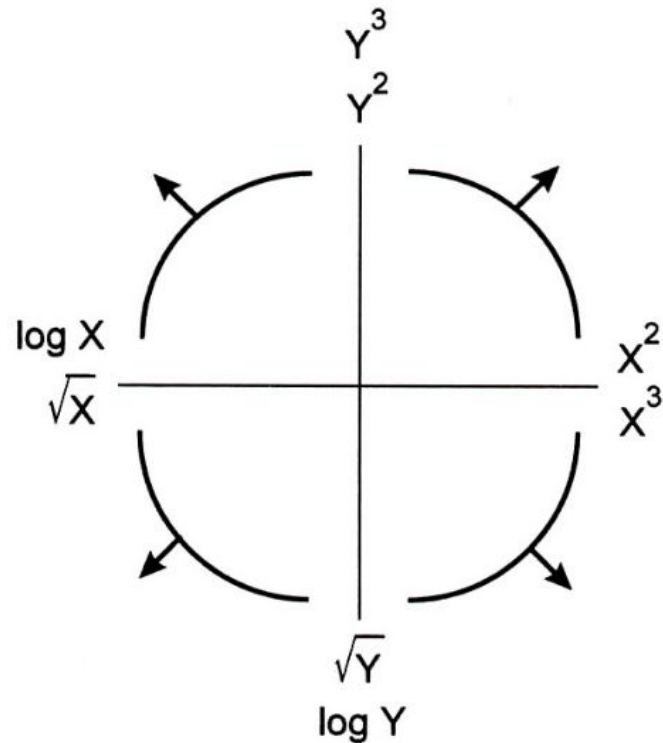
Introducing Notation for Multiple Linear
Regression

The **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 to 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





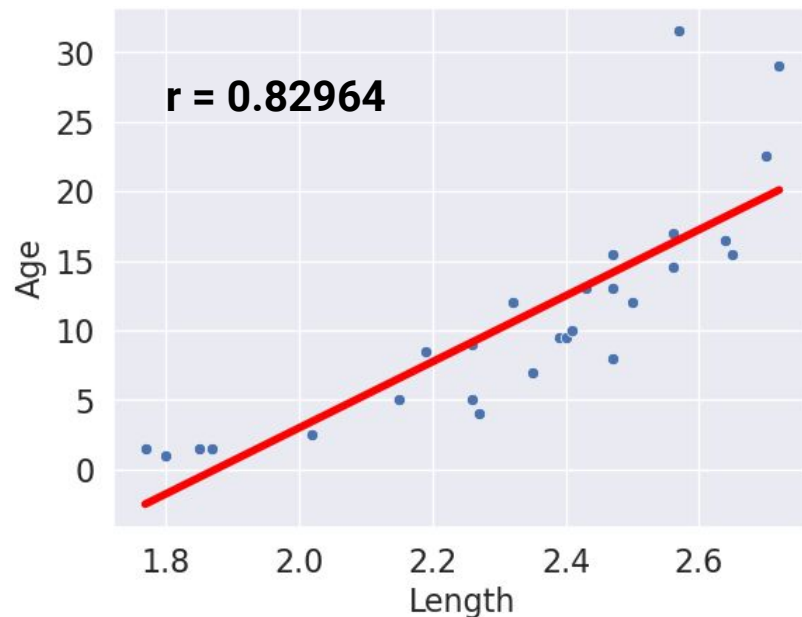
From Data 8 ([textbook](#)):

The residual plot of a good regression shows no pattern.

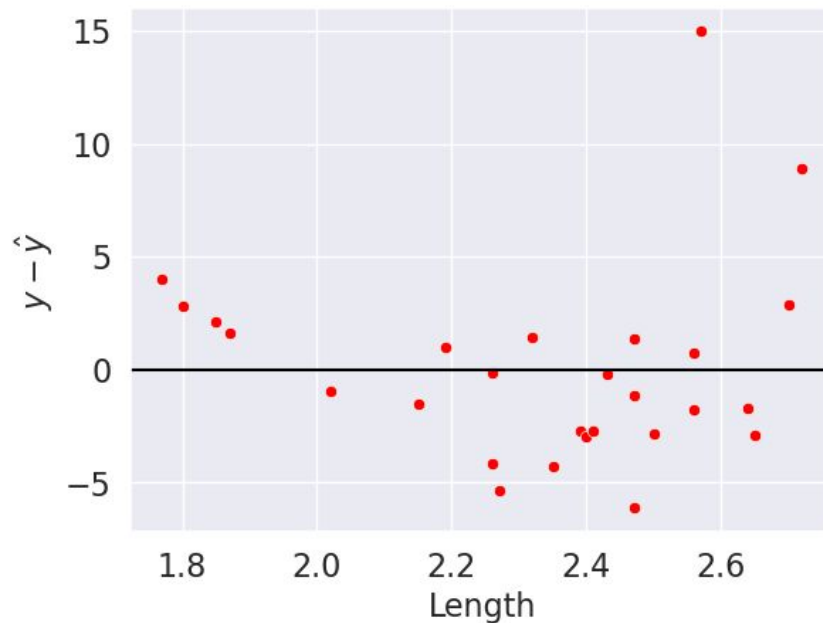
https://inferentialthinking.com/chapters/15/5/Visual_Diagnostics.html



Age by Length



Residual Plot



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.**



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

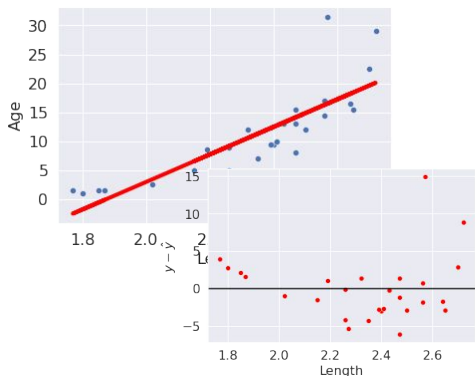
In other words, if we apply the variable transform $z = \log(y)$:

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

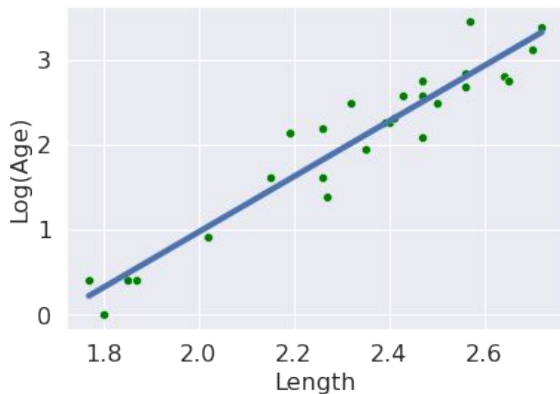
$$R(\theta) = \frac{1}{n} \sum_{i=1}^n (z_i - \hat{z}_i)^2$$

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

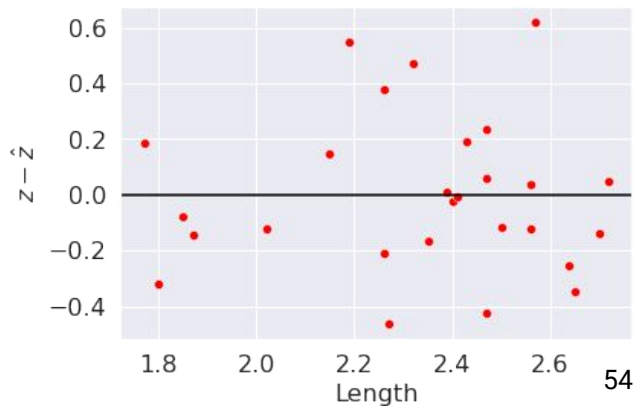
Original (Age by Length)



Log(Age) by Length



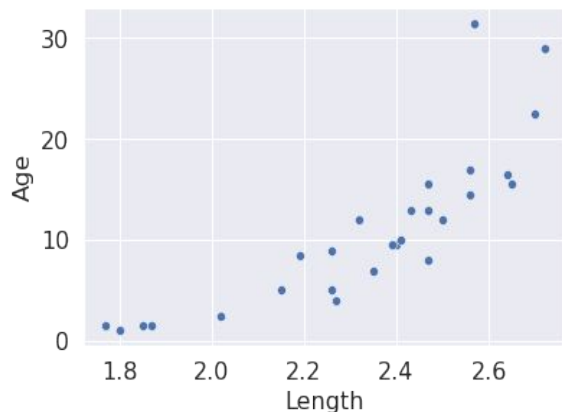
Residual Plot



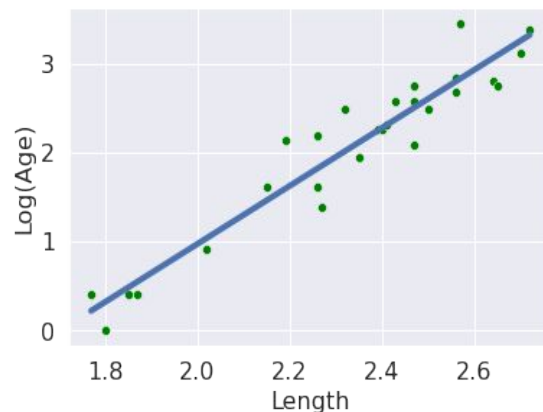
$$z = \log(y)$$

$$y = e^z$$

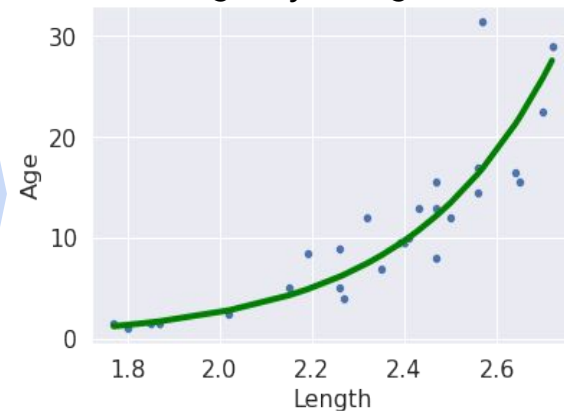
Age by Length



Log(Age) by Length



Age by Length

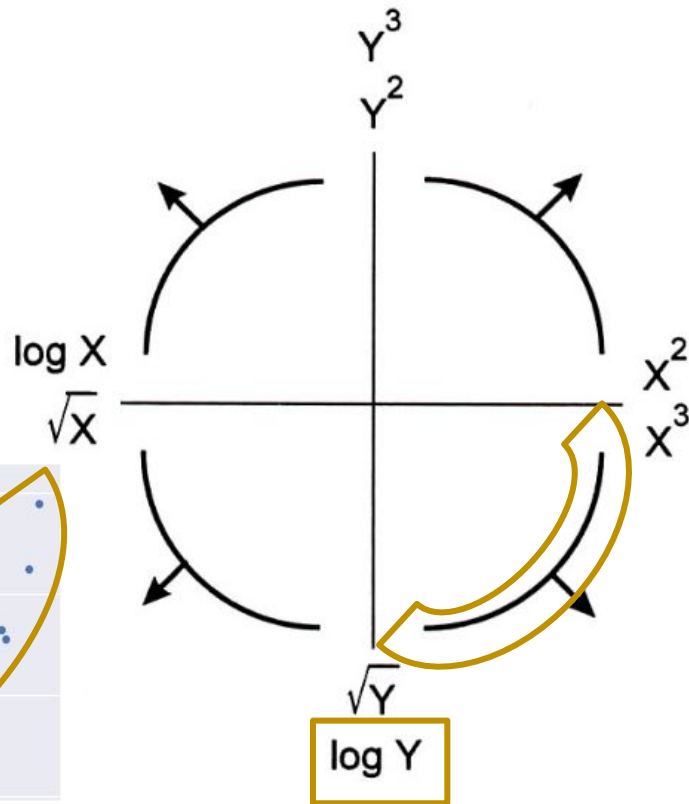
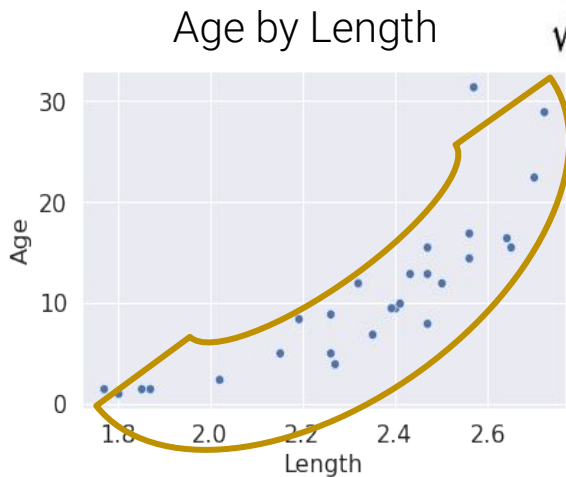


Tukey-Mosteller Bulge Diagram

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.





Introducing Notation for Multiple Linear Regression

Lecture 11, Data 100 Spring 2023

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**



3280763

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**.



Define the **multiple linear regression** model:

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

Parameters are $\theta = [\theta_0, \theta_1, \dots, \theta_p]$

Is this linear in θ ?

- A. no
- B. yes
- C. maybe



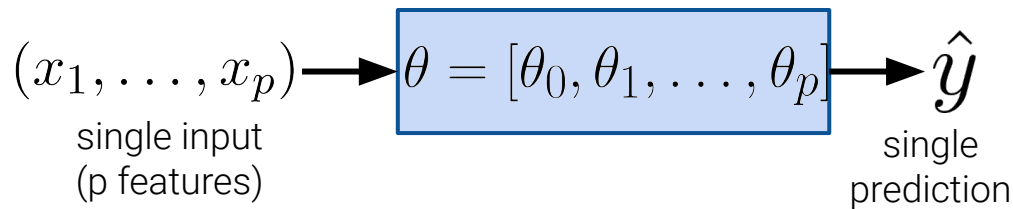
Multiple Linear Regression

Define the **multiple linear regression** model:

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

Parameters are $\theta = [\theta_0, \theta_1, \dots, \theta_p]$

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



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

More on Multiple Linear
Regression on Thursday



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!



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

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!



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:

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

Looks familiar!