2341552

slido

Join at slido.com
#2341552

Click **Present with Slido** or install our Chrome extension to display joining instructions for participants while presenting.

CC BY NC SA

⚠️ Reminder to start the Zoom recording!

2341552

LECTURE 11

Constant Model, Loss, and Transformations

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

Data 100/Data 200, Spring 2025 @ UC Berkeley

Narges Norouzi and Josh Grossman

Content credit: Acknowledgments

CC BY NC SA

2341552

Announcements

Lecture 11, Data 100 Spring 2025

Grad Students:

- Grad project released!
- Project proposal due **Monday, March 3**
- Use Ed thread #666 for group matching

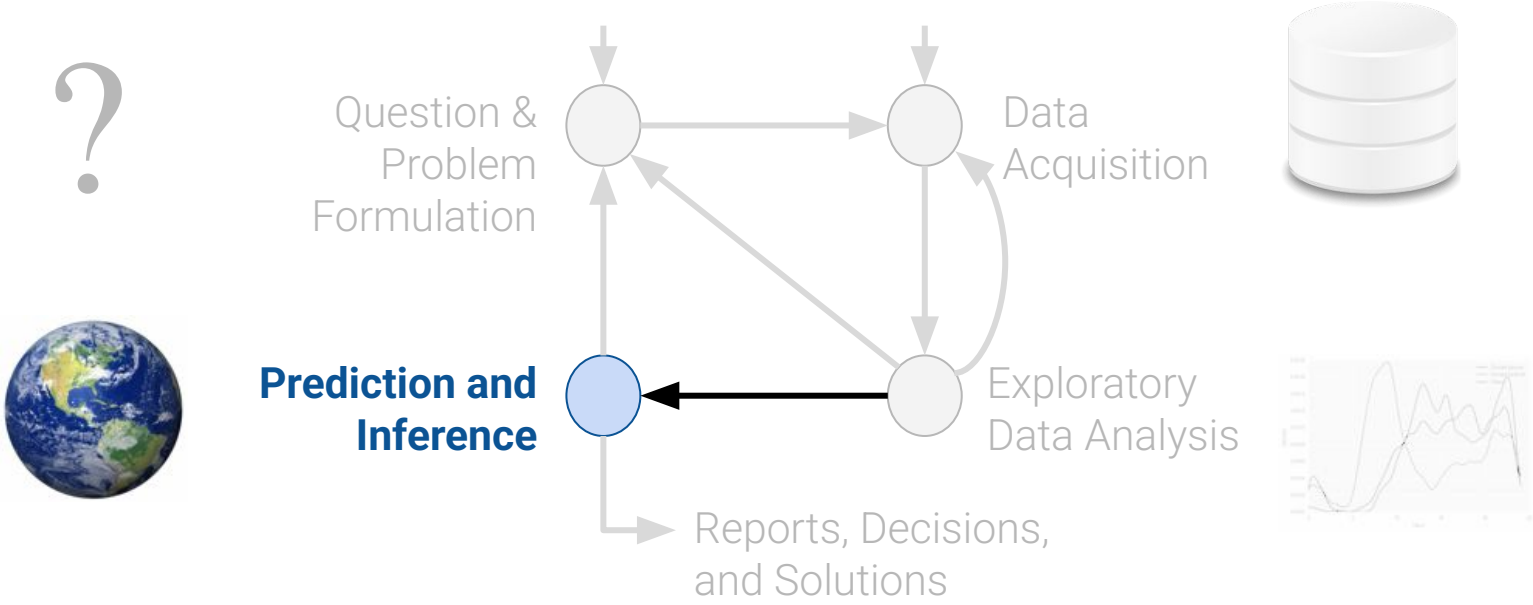
Catch-up hour:

- Every Friday, 3-4pm in VLSB 2032
- Q&A session for **past content**

CC BY NC SA

Plan for Next Few Lectures: Modeling

2341552



(today)

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

CC BY NC SA

2341552

Today's Roadmap

Lecture 11, Data 100 Spring 2025

Modeling Process

- Iteration 1: SLR + MSE (last time!)
- Iteration 2: Constant Model + MSE
- Iteration 3: Constant Model + MAE

Transformations of Linear Models

Notation for Multiple Linear Regression

CC BY NC SA

2341552

Iteration 2: Constant Model + MSE

Lecture 11, Data 100 Spring 2025

Modeling Process

- Iteration 1: SLR + MSE (last time!)
- **Iteration 2: Constant Model + MSE**
- Iteration 3: Constant Model + MAE

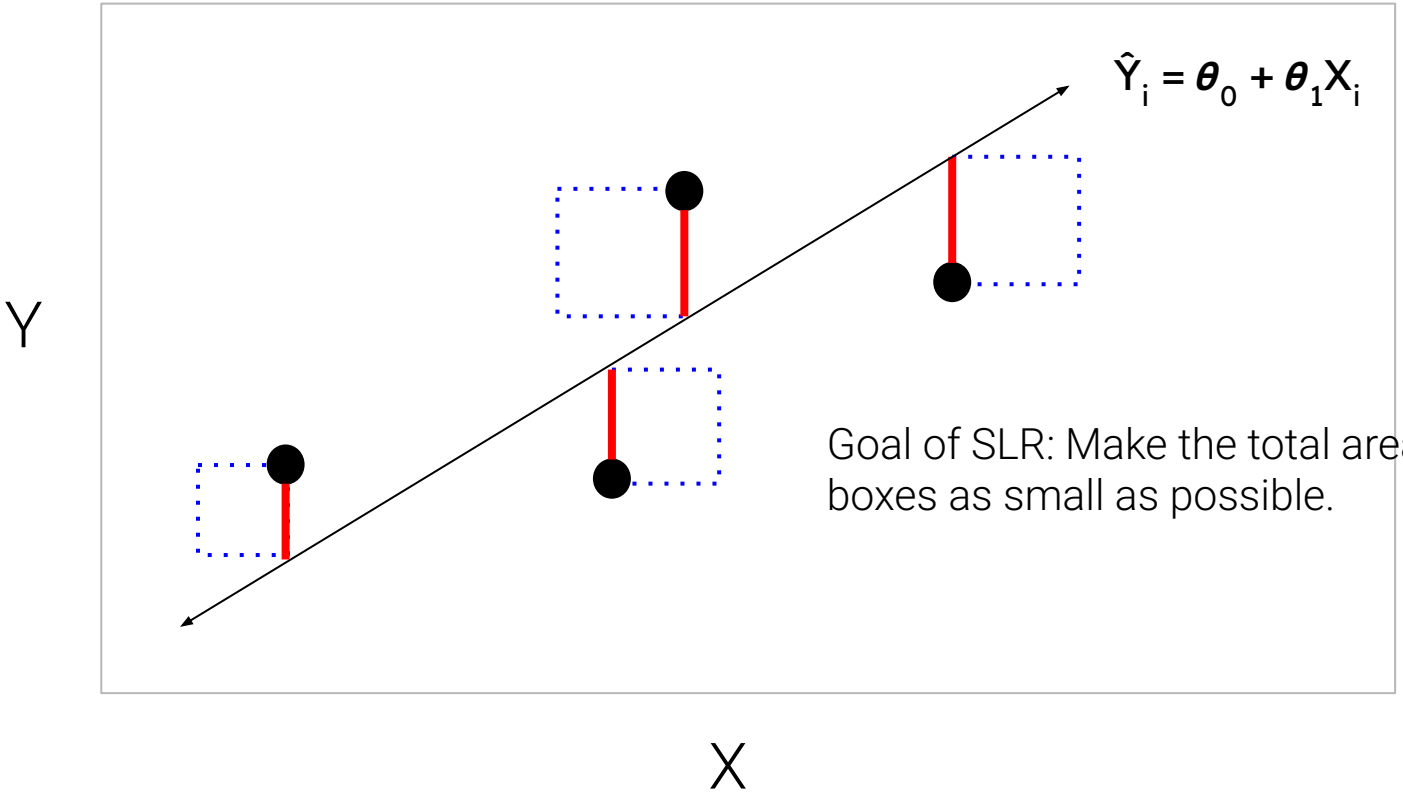
Transformations of Linear Models

Notation for Multiple Linear Regression

CC BY NC SA

Visualizing the sum of squared error for simple linear regression (SLR)

2341552

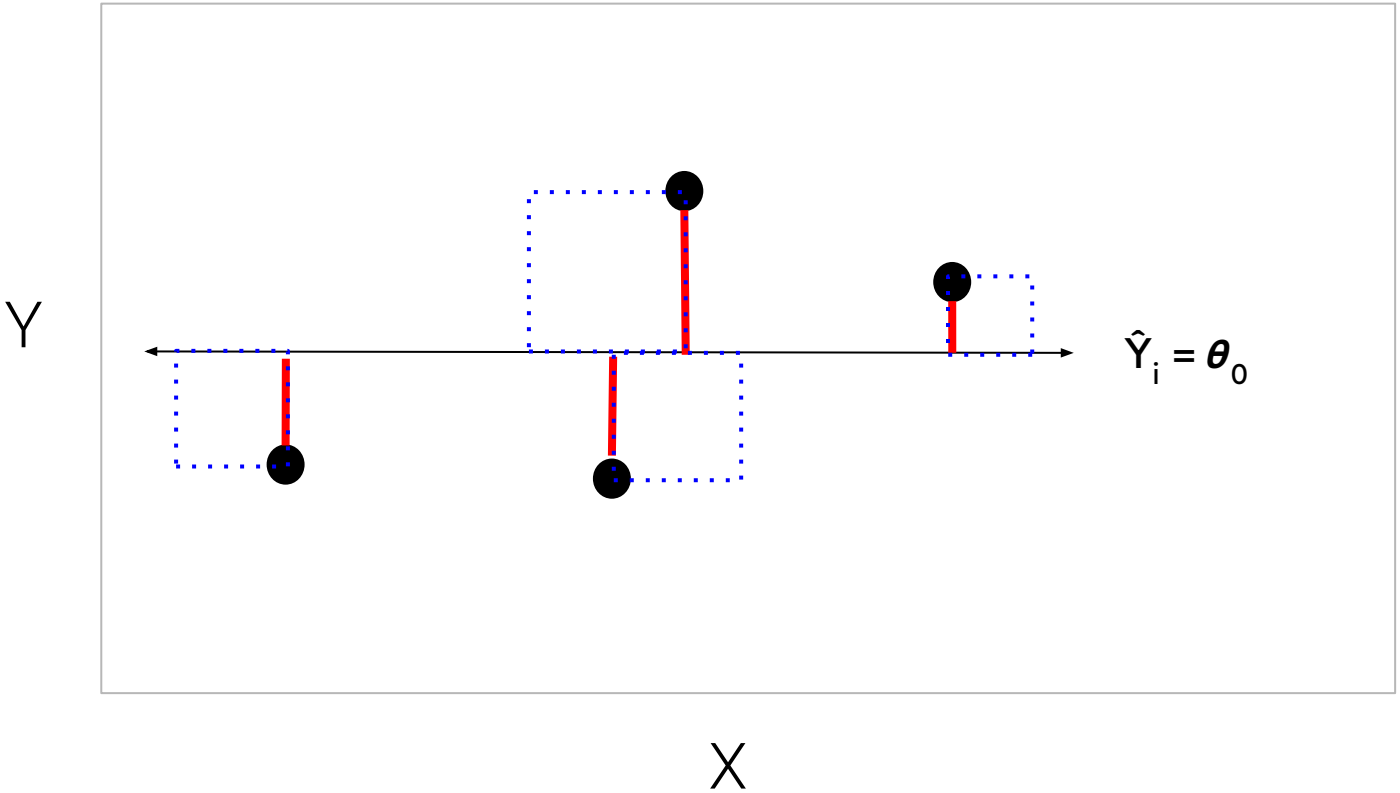


Goal of SLR: Make the total area of the boxes as small as possible.

CC BY NC SA

Visualizing the sum of squared error for a constant model

2341552



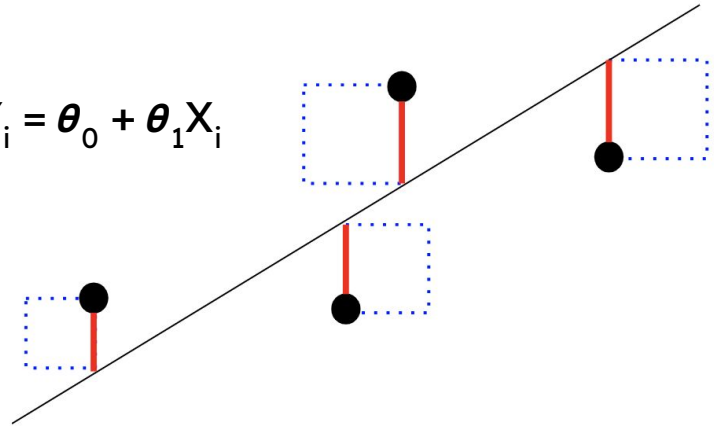
Also known as an intercept-only model.

CC BY NC SA

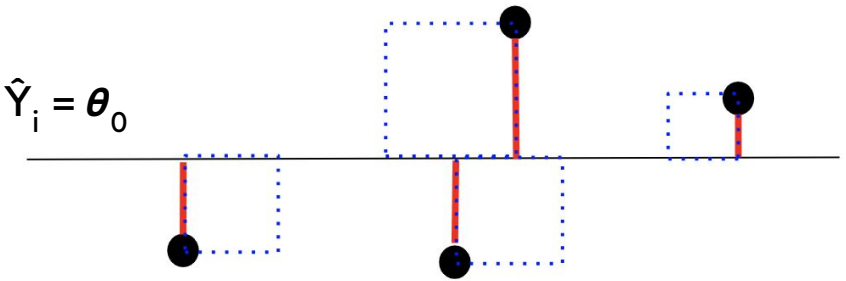
R^2 measures quality of linear model fit relative to constant model

2341552

$\hat{Y}_i = \theta_0 + \theta_1 X_i$



$\hat{Y}_i = \theta_0$



$$R^2 = \frac{\Delta \text{ in area}}{\text{Constant model area}}$$

For SLR, $R^2 = r^2$. More on R^2 in a later lecture!

CC BY NC SA

The Modeling Process: Using a Different Model

2341552

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

4. Evaluate model performance

CC BY NC SA

The Constant Model

2341552

You work at a local boba shop and want to predict the daily number of drinks sold.

Here's your data from 5 randomly selected previous days:

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

You have other info (e.g., the weather and day of week), but you *ignore everything except the historical # of drinks sold*.

Your model always predicts the same number.

This is a **constant model**.

CC BY NC SA

The Constant Model

2341552

The **constant model** **always predicts the same number**—i.e., 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**.

Like SLR, the constant model is a parametric, statistical model:

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

CC BY NC SA

The Constant Model

2341552

Our parameter $\boldsymbol{\theta}_0$ is 1-dimensional. $\theta_0 \in \mathbb{R}$

Unlike SLR, no input needed to generate a prediction; we always predict $\hat{y} = \theta_0$

Like before, we can still determine the best $\boldsymbol{\theta}_0$ that minimizes **average loss** on our data.

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

CC BY NC SA

The Modeling Process: Using a Different Model

2341552

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)

$$\hat{R}(\theta) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{"Average loss" "Objective function" "Mean squared error"}}$$

3. Fit the model

Minimize average loss with calculus

4. Evaluate model performance

Why is there a hat above R? We are *estimating* the average loss with a sample. Estimated values often have a hat. Technically, there is also a "true" loss for the entire population from which our sample is drawn. We would write the expression for this "true" loss without a hat. Don't worry about this for Data 100!

CC BY NC SA

The Modeling Process: Using a Different Model

2341552

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)

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

3. Fit the model

Minimize average loss with calculus

How does this step change?

4. Evaluate model performance

Why is there a hat above R? We are *estimating* the average loss with a sample. Estimated values often have a hat. Technically, there is also a "true" loss for the entire population from which our sample is drawn. We would write the expression for this "true" loss without a hat. Don't worry about this for Data 100!

CC BY NC SA

Fit the Model: Rewrite MSE for the Constant Model

2341552

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

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

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

CC BY NC SA

Fit the Model: Where We Are Headed

2341552

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

With calculus, we can show that the average loss (i.e., our objective function) is minimized by

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

Given a dataset, we can also find the minimizing $\boldsymbol{\theta}_0$ with an iterative algorithm. More soon!

CC BY NC SA

Fit the Model: Calculus for the General Case

2341552

1. Differentiate with respect to θ_0:

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

2. Set equal to 0.

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

CC BY NC SA

Fit the Model: Calculus for the General Case

2341552

1. Differentiate with respect to θ_0:

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

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta_0}(y_i - \theta_0)^2$$ 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 $\hat{\theta}_0$.

CC BY NC SA

Fit the Model: Calculus for the General Case

2341552

1. Differentiate with respect to θ_0:

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

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta_0}(y_i - \theta_0)^2$$ 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 $\hat{\theta}_0$.

CC BY NC SA

Fit the Model: Calculus for the General Case

2341552

1. Differentiate with respect to θ_0:

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

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta_0}(y_i - \theta_0)^2$$ 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 $\hat{\theta}_0$.

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

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

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

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

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

CC BY NC SA

Fit the Model: Calculus for the General Case

2341552

1. Differentiate with respect to θ_0:

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

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta_0}(y_i - \theta_0)^2$$ 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 $\hat{\theta}_0$.

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

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

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

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

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

CC BY NC SA

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

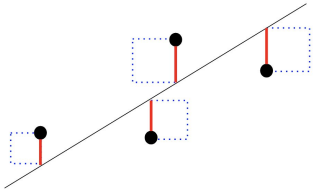
2341552

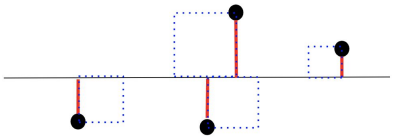
The **mean** of the outcomes achieves the minimum MSE of the constant model.

Minimum MSE is the **sample variance** of y → $$R(\hat{\theta}_0) = R(\bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sigma_y^2$$

R^2 (from earlier) is interpreted as the fraction of **variance in y** "explained" by a linear model.

$$\mathrm{R}^2 = \frac{\Delta \text{ in area}}{\text{Constant model area}} = \frac{\mathrm{MSE}_{\text{constant model}} - \mathrm{MSE}_{\text{linear model}}}{\mathrm{MSE}_{\text{constant model}}}$$





Technical note: The area is the sum of squared errors, not the MSE. Area/n=MSE.
Also, the R in the average loss expression is unrelated to R^2.

Why "explained" is in quotes: There is no causality here!

CC BY NC SA

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

2341552

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

The **minimum** MSE

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

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

.

CC BY NC SA

The Modeling Process: Using a Different Model

2341552

1. Choose a model ✅

Constant Model

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

2. Choose a loss function ✅

L2 Loss

Mean Squared Error (MSE)

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

3. Fit the model ✅

Minimize average loss with calculus

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

4. Evaluate model performance

CC BY NC SA

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

2341552

Modeling the relationship of dugong age (y) and length (x).

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.

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

Demo

lec10.ipynb

CC BY NC SA

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

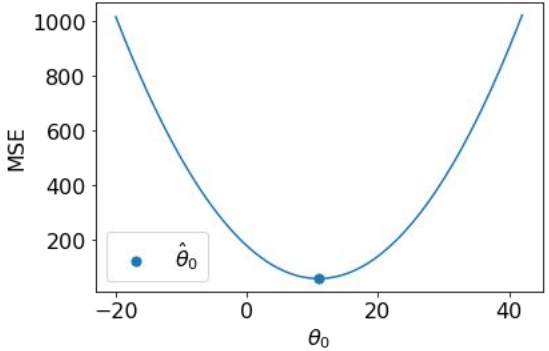
2341552

Constant Model

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

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

Loss surface is **2-D**.



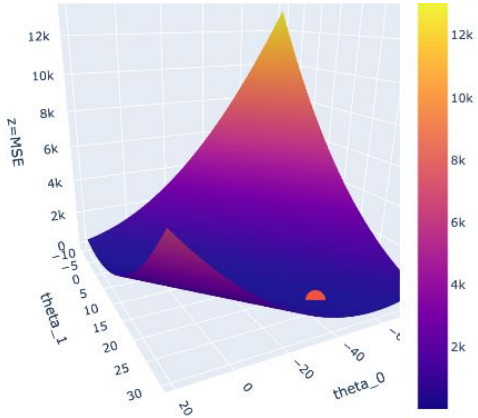
$$\hat{R}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_0)^2$$

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

lec10.ipynb

CC BY NC SA

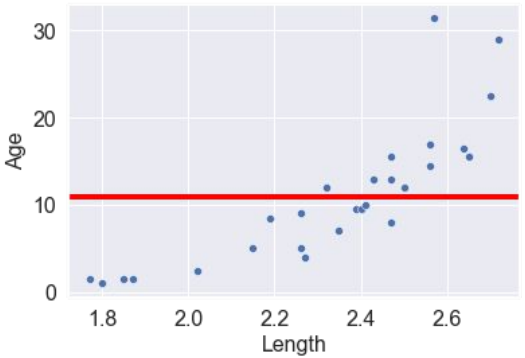
[Fit] Comparing Two Different Models, Both Fit with MSE

2341552

Constant Model

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

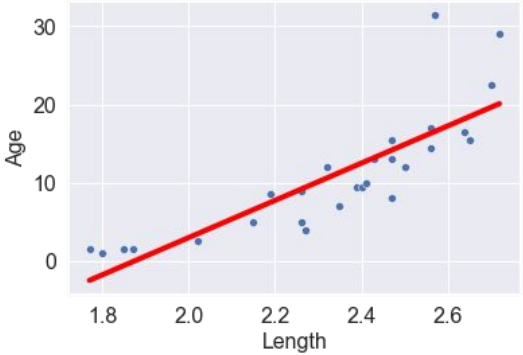
Root MSE (RMSE): **7.72**



Simple Linear Regression

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

Root MSE (RMSE): **4.31**



Interpretation of RMSE:

- Constant model RMSE is **higher** than linear model RMSE
- Constant model has a **poorer fit** than linear model, using RMSE as our metric.

Demo

lec10.ipynb

CC BY NC SA

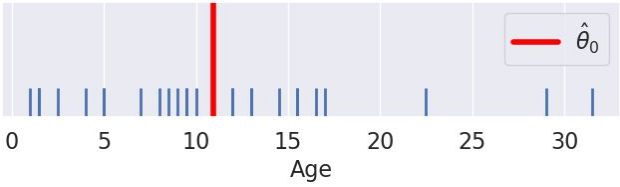
[Fit] Comparing Two Different Models, Both Fit with MSE

2341552

Constant Model

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

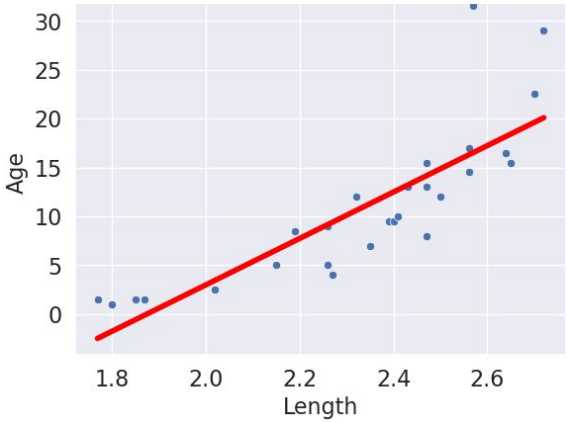
Predictions on a **rug plot**.



Simple Linear Regression

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

Predictions on a **scatter plot**.



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

Demo

lec10.ipynb

CC BY NC SA

2341552

Iteration 3: Constant Model + MAE

Lecture 11, Data 100 Spring 2025

Modeling Process

- Iteration 1: SLR + MSE (last time!)
- Iteration 2: Constant Model + MSE
- **Iteration 3: Constant Model + MAE**

Transformations of Linear Models

Notation for Multiple Linear Regression

CC BY NC SA

The Modeling Process: Using a Different Loss Function

2341552

1. Choose a model

Constant Model

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

2. Choose a loss function

L1 Loss

Mean Absolute Error (MAE)

Let's repeat, but with **Mean Absolute Error (MAE)** as our average loss.

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

3. Fit the model

Minimize average loss with calculus

4. Evaluate model performance

CC BY NC SA

The Modeling Process: Using a Different Loss Function

2341552

1. Choose a model ✅

Constant Model

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

2. Choose a loss function ✅

L1 Loss

Mean Absolute Error (MAE)

Let's repeat, but with **Mean Absolute Error (MAE)** as our average loss.

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

3. Fit the model

Minimize average loss with calculus

How does this step change?

4. Evaluate model performance

CC BY NC SA

Fit the Model: Rewrite MAE for the Constant Model

2341552

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.

CC BY NC SA

Exploring MAE: A Piecewise function

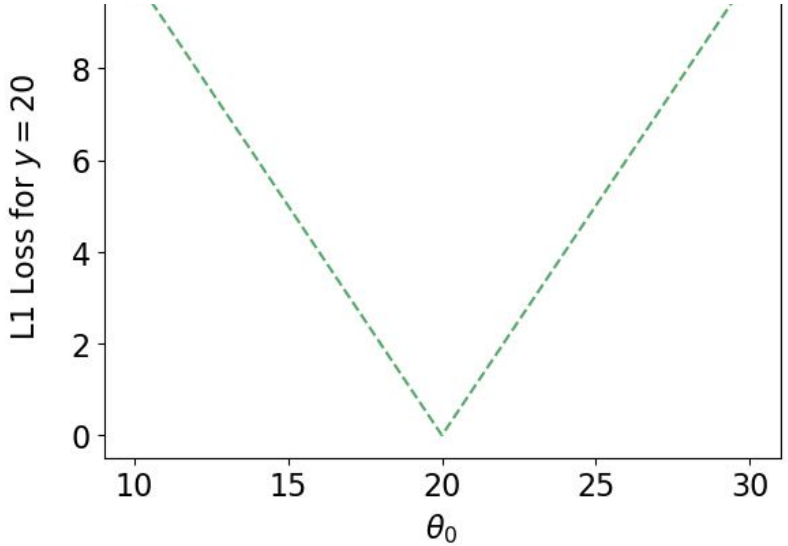
2341552

Boba dataset {20, 21, 22, 29, 33} (Mean=25)

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

Absolute (L1) Loss on one observation:

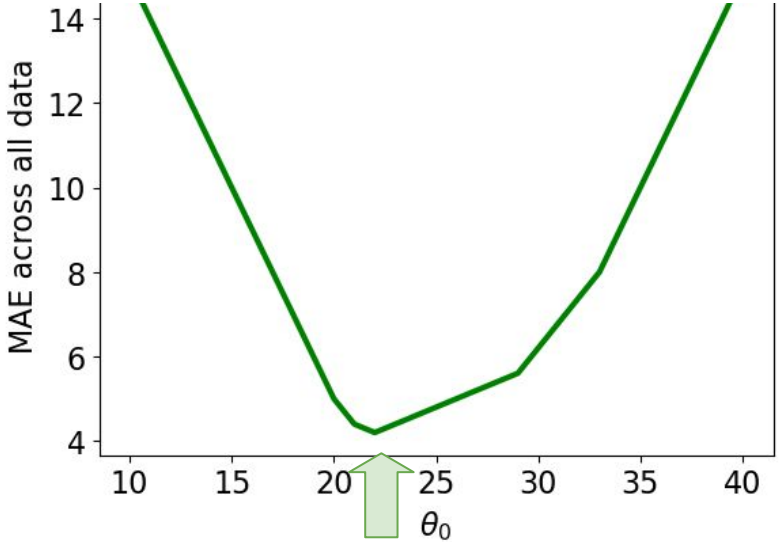
$$L_1(20, \theta_0) = |20 - \theta_0|$$



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?

CC BY NC SA

Fit the Model: Calculus

2341552

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.

CC BY NC SA

Fit the Model: Calculus

2341552

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

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

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

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

CC BY NC SA

Fit the Model: Calculus

2341552

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).

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

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

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

CC BY NC SA

Fit the Model: Calculus

2341552

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

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

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

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

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

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

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

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

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

CC BY NC SA

Fit the Model: Calculus

2341552

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

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

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

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

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

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

2. Set equal to 0.

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

CC BY NC SA

Fit the Model: Calculus

2341552

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

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

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

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

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

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

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 $\hat{\theta}_0$.

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

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

CC BY NC SA

Fit the Model: Calculus

2341552

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

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

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

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

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

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

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 $\hat{\theta}_0$.

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

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

CC BY NC SA

Median Minimizes MAE for the Constant Model

2341552

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

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

$\hat{\theta}_0$ needs to be located where there are **an equal number of y_i's to the left and right**.

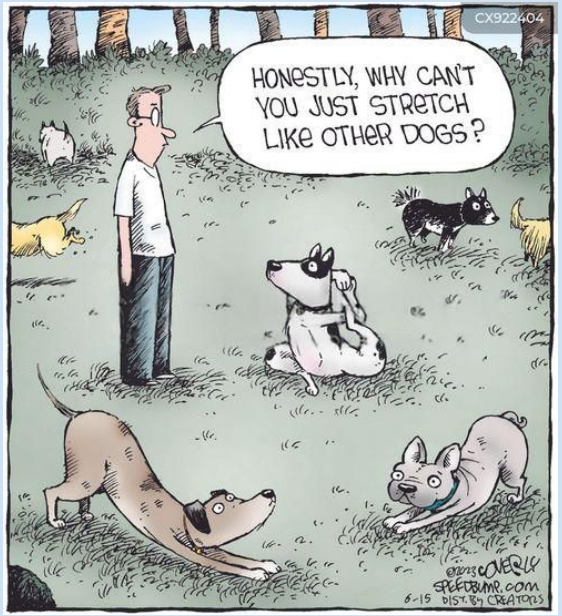
This is the definition of the **median!**

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

For example, in our boba tea dataset {20, 21, 22, 29, 33}, the value **22** is the median.

CC BY NC SA

2341552



2-minute stretch break!

The next lecture relies on the geometric interpretation of linear algebra.

We recommend watching [this 3Blue1Brown video](#) (or better, the entire series) to get a solid understanding of the geometry of linear algebra.

CC BY NC SA

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

2341552

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 $\hat{\theta}$.

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

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

CC BY NC SA

The Modeling Process: Using a Different Loss Function

2341552

1. Choose a model ✅

Constant Model

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

2. Choose a loss function ✅

L1 Loss

Mean Absolute Error (MAE)

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

3. Fit the model ✅

Minimize average loss with calculus

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

4. Evaluate model performance

CC BY NC SA

MSE and MAE: Comparing Optimal Parameters

2341552

MSE (Mean Squared Error)

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

Minimized with **sample mean**:

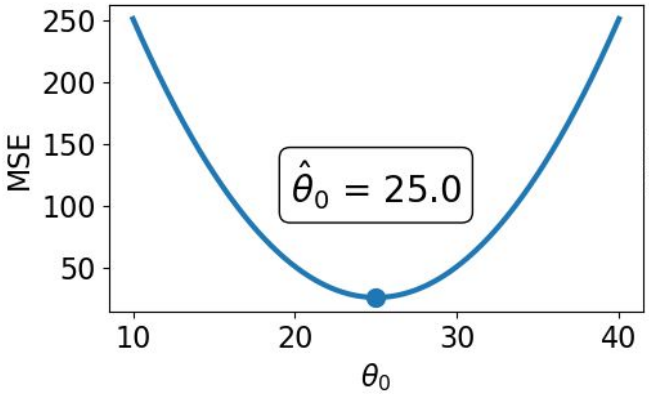
$$\hat{\theta}_0 = \text{mean}(y) = \bar{y}$$

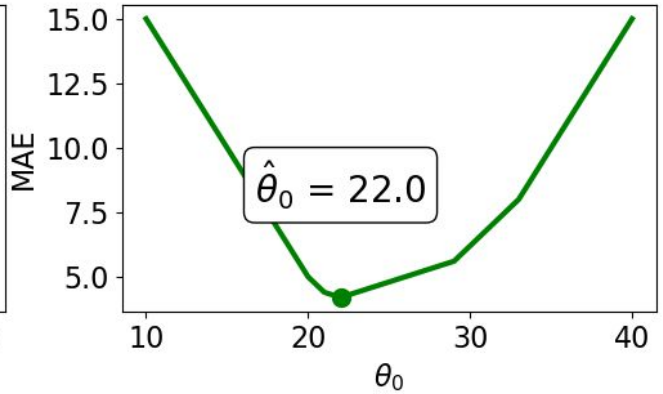
MAE (Mean Absolute Error)

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

Minimized with **sample median**:

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





Demo

lec10.ipynb

CC BY NC SA

MSE and MAE: Finding Minimizing Values

2341552

MSE (Mean Squared Error)

Minimized with **sample mean**:

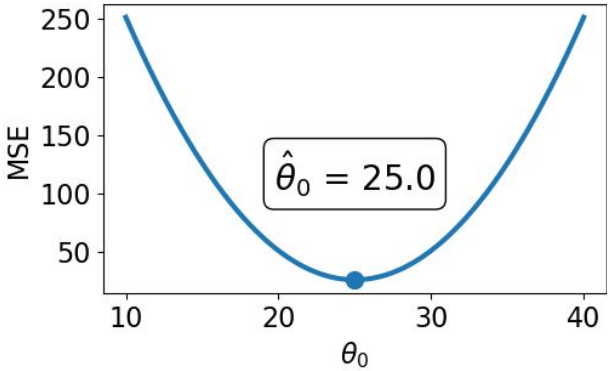
$$\hat{\theta}_0 = \text{mean}(y) = \bar{y}$$

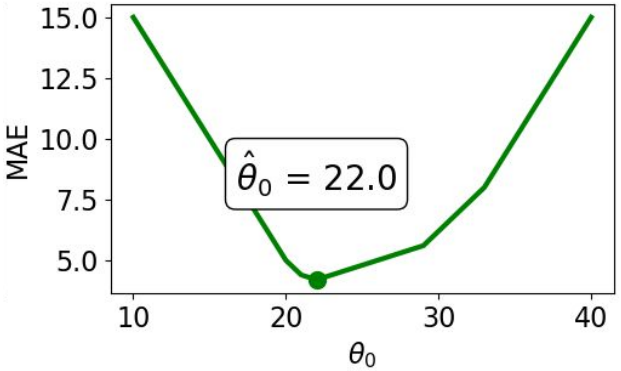
MAE (Mean Absolute Error)

Minimized with **sample median**:

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

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





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

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

Demo

lec10.ipynb

CC BY NC SA

MSE and MAE: Comparing Uniqueness of Solutions

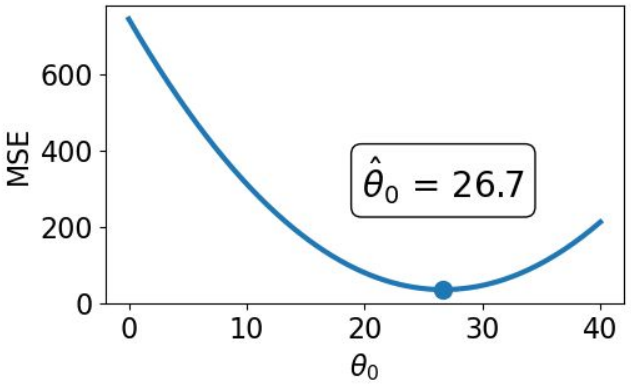
2341552

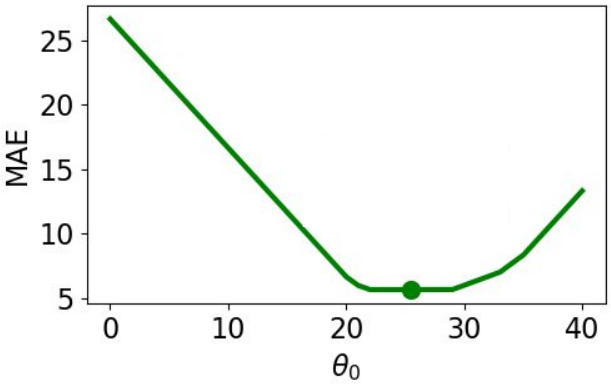
MSE (Mean Squared Error)

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. Equal # of y_i's to left and right.

We typically choose the average of the boundary values (25.5).

Demo

lec10.ipynb

CC BY NC SA

MSE and MAE: Comparing Sensitivity to Outliers

2341552

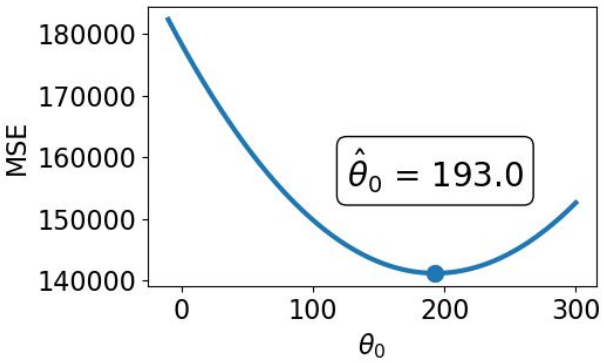
MSE (Mean Squared Error)
Minimized with **sample mean**:

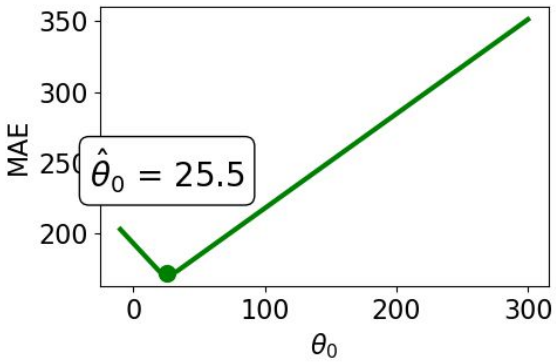
$$\hat{\theta}_0 = \text{mean}(y) = \bar{y}$$

MAE (Mean Absolute Error)
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

lec10.ipynb

CC BY NC SA

2341552

Which of the following loss functions for the constant model are minimized when $\theta_0 = \bar{y}$?

Presenting with animations, GIFs or speaker notes? Enable our Chrome extension

slido

Which loss functions for the constant model are minimized when $\theta_0 = \bar{Y}$?

2341552

✅ MSE: We saw earlier in lecture that $\bar{Y}$ minimizes MSE.

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

✅ SSE (Sum of squared error): The 1/n constant does not impact the minimizing θ_0.

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

❌ MAE: We saw earlier that the median of Y minimizes the MAE.

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

❌ SAE (Sum of absolute error): Like above, the 1/n constant does not impact the minimizing θ_0. So, the median of Y also minimizes the SAE.

$$\text{SAE} = \sum_{i=1}^{n}|y_i - \theta_0|$$

CC BY NC SA

2341552

Transformations of Linear Models

Lecture 11, Data 100 Spring 2025

Modeling Process

- Iteration 1: SLR + MSE (last time!)
- Iteration 2: Constant Model + MSE
- Iteration 3: Constant Model + MAE

Transformations of Linear Models

Notation for Multiple Linear Regression

CC BY NC SA

Back to Least Squares Regression with Dugongs

2341552

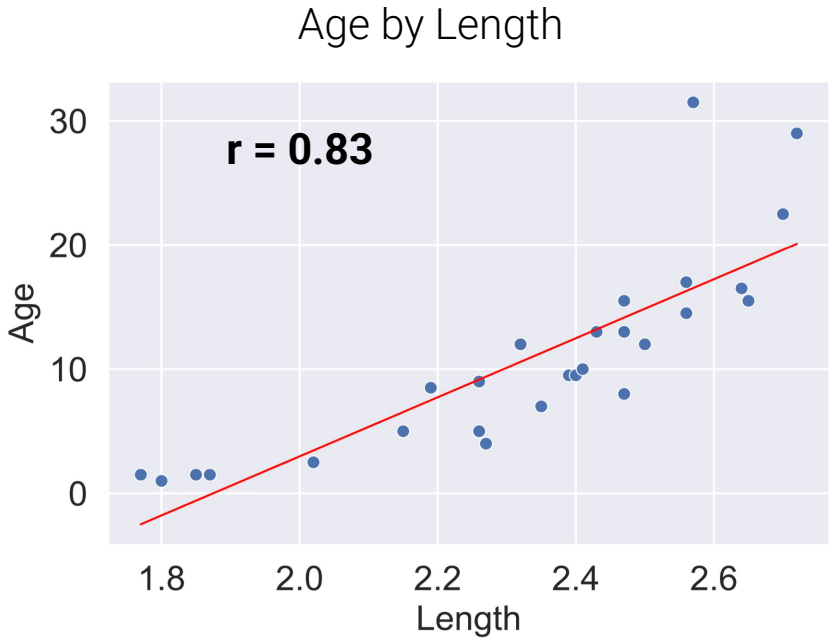
From Data 8 (textbook): The residual plot of a good regression shows **no pattern**.

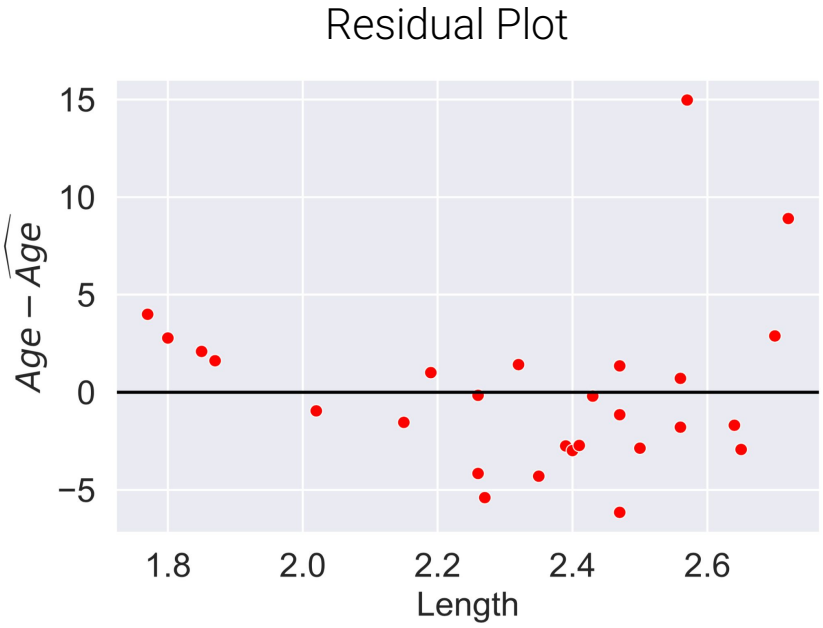
https://inferentialthinking.com/chapters/15/5/Visual_Diagnostics.html

CC BY NC SA

Back to Least Squares Regression with Dugongs

2341552





The scatter plot is **non-linear**, so the SLR **residual plot** shows a pattern indicative of a bad fit.

To improve the fit, we can **linearize** the data with a transformation.

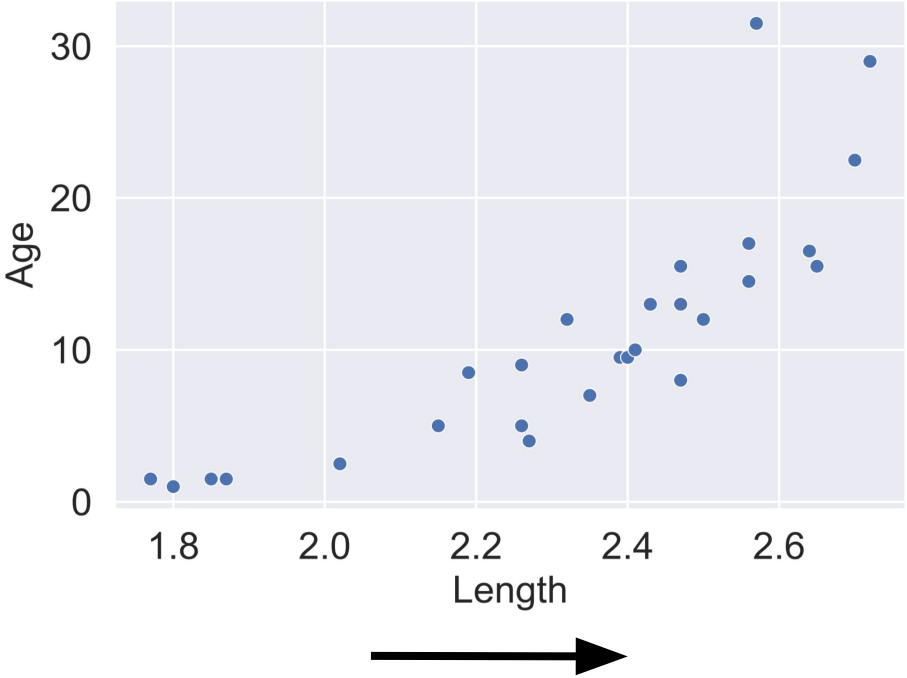
CC BY NC SA

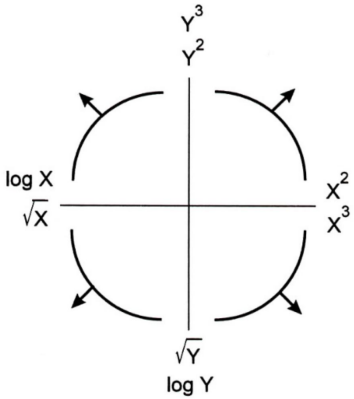
Possible Transformations

2341552

Large age values appear **unconstrained.**

Squish differences with **log(Age)** transform.





Large length values appear **compressed**.

Magnify differences with **(Length)^2** transform.

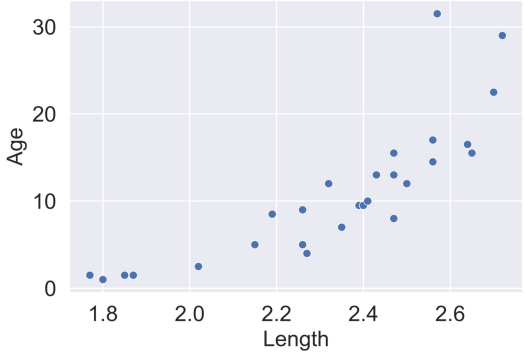
CC BY NC SA

Effect of log and squaring transformations

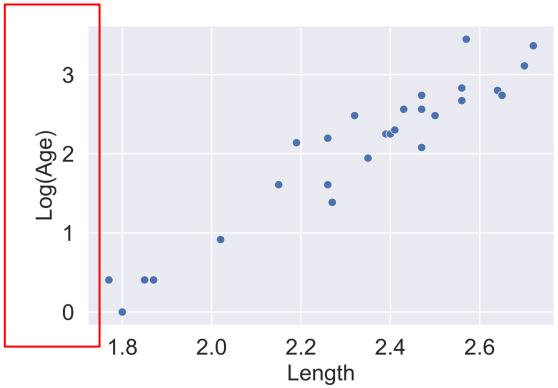
2341552

No transformation

[Raw; natural scale]

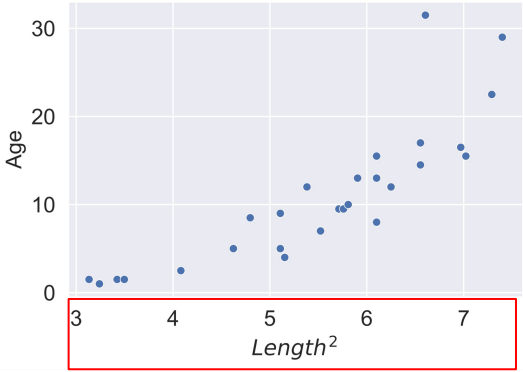


Log transformation



Big improvement in linearity.

Squaring transformation



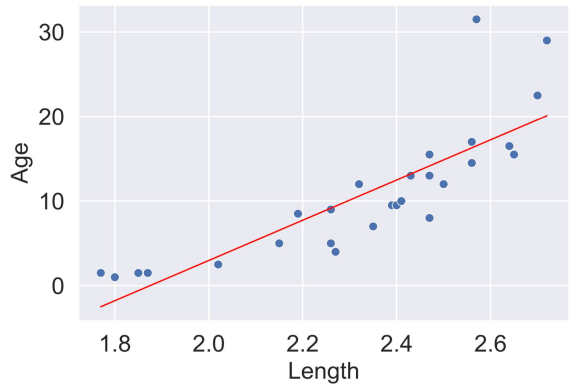
Very subtle improvement in linearity.
Hard to see here—wait a few slides!

CC BY NC SA

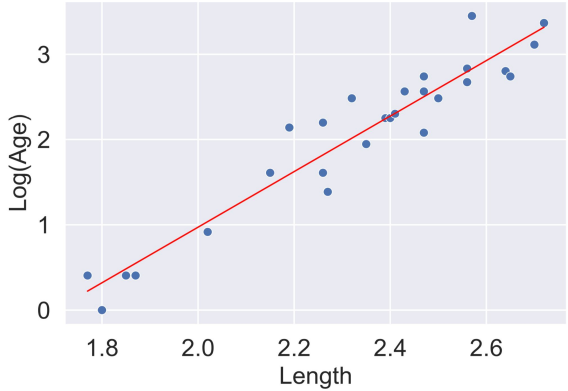
SLR outcomes always linear in the provided feature

2341552

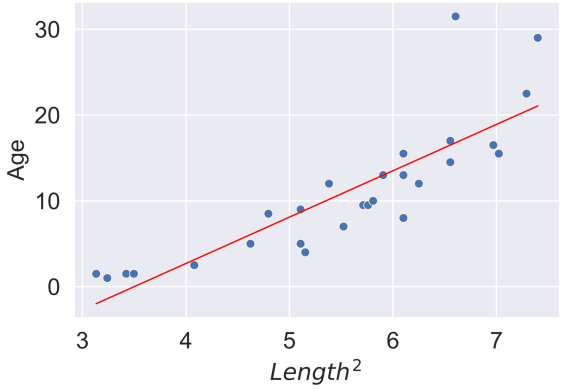
$$\widehat{\text{Age}} = \theta_0 + \theta_1 \text{Length}$$



$$\widehat{\log(\text{Age})} = \theta_0 + \theta_1 \text{Length}$$



$$\widehat{\text{Age}} = \theta_0 + \theta_1 \text{Length}^2$$



SLR doesn't "know" that the output is logged Age or the input is squared Length.

- To SLR, log(Age) and Length^2 are just vectors of values!
- The outcome provided to SLR is always modeled as linear in the provided feature.

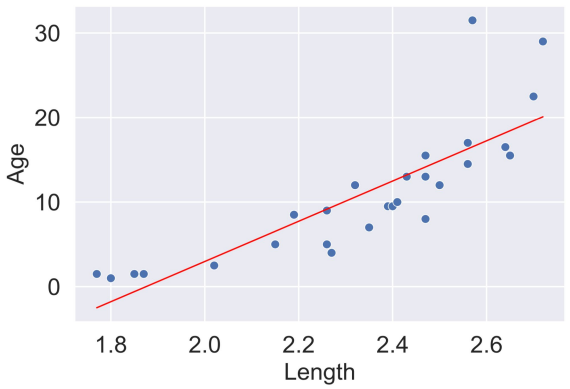
Note: When squaring a feature, we usually include the linear feature for extra flexibility (e.g., $\hat{Y} = \theta_0 + \theta_1 X + \theta_2 X^2$). Next lecture!

CC BY NC SA

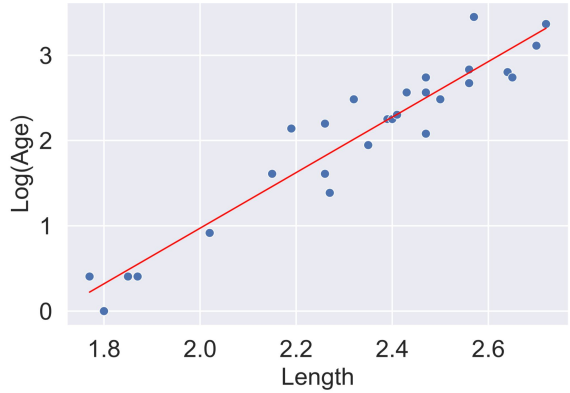
SLR outcomes always linear in the provided feature

2341552

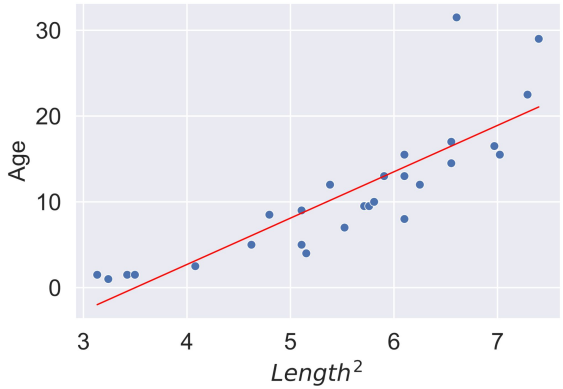
$$\widehat{\text{Age}} = \theta_0 + \theta_1 \text{Length}$$

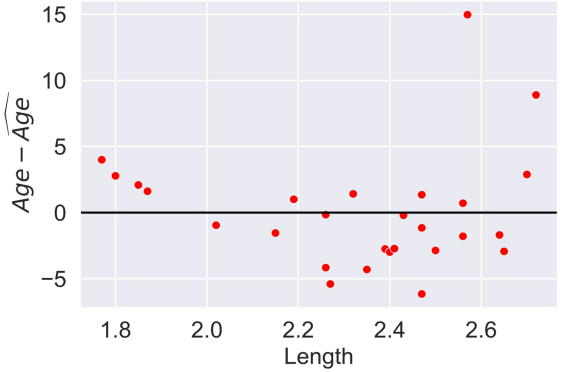


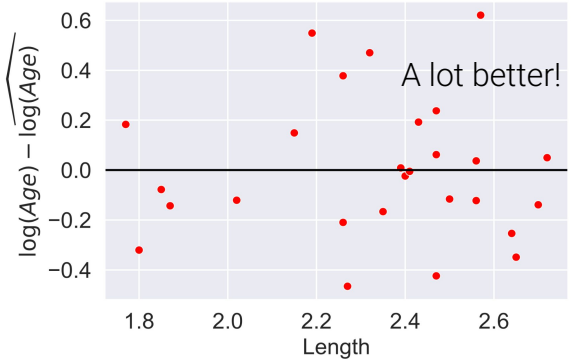
$$\widehat{\log(\text{Age})} = \theta_0 + \theta_1 \text{Length}$$

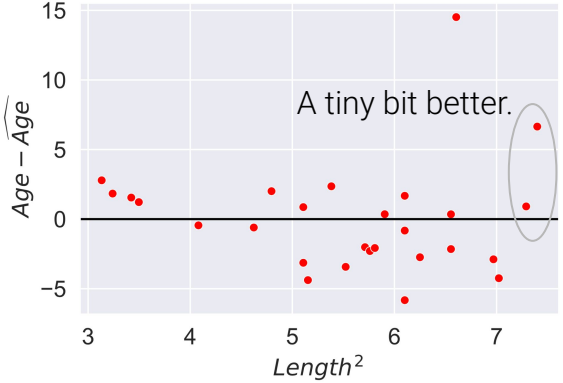


$$\widehat{\text{Age}} = \theta_0 + \theta_1 \text{Length}^2$$









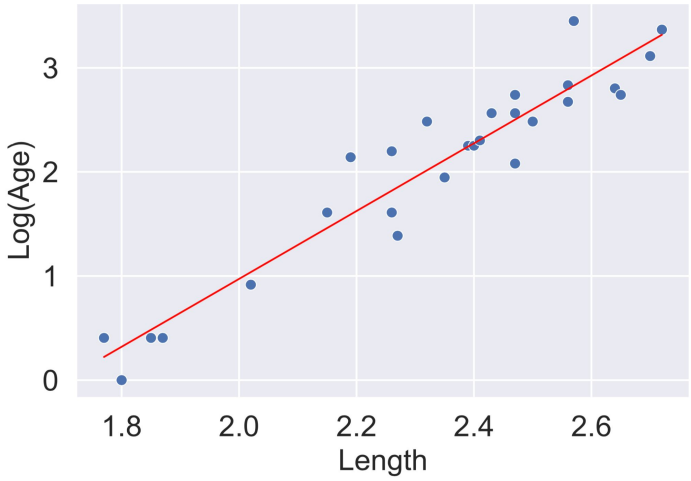
Note: When squaring a feature, we usually include the linear feature for extra flexibility (e.g., $\hat{Y} = \theta_0 + \theta_1 X + \theta_2 X^2$). Next lecture!

CC BY NC SA

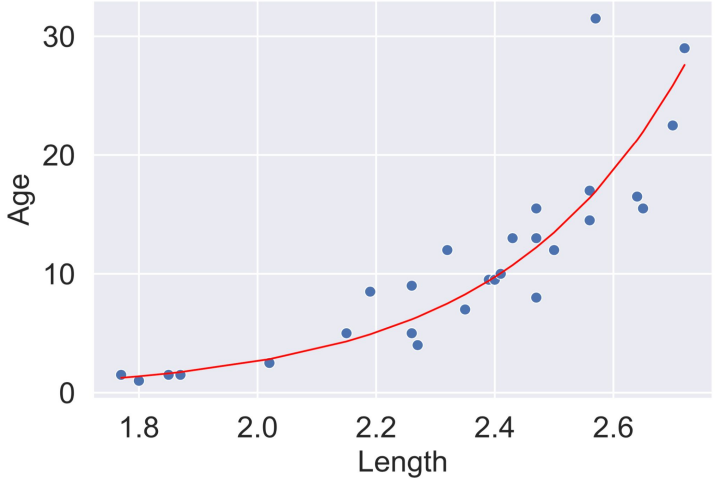
Visualizing log-transformed predictions on natural scale

2341552

$$\widehat{\log(\text{Age})} = \theta_0 + \theta_1 \text{Length}$$



$$\widehat{\text{Age}} = e^{\theta_0 + \theta_1 \text{Length}}$$



Unlike SLR, <u>we</u> know that the y-variable is log(Age).

- To get Age predictions, we exponentiate the log(Age) predictions.

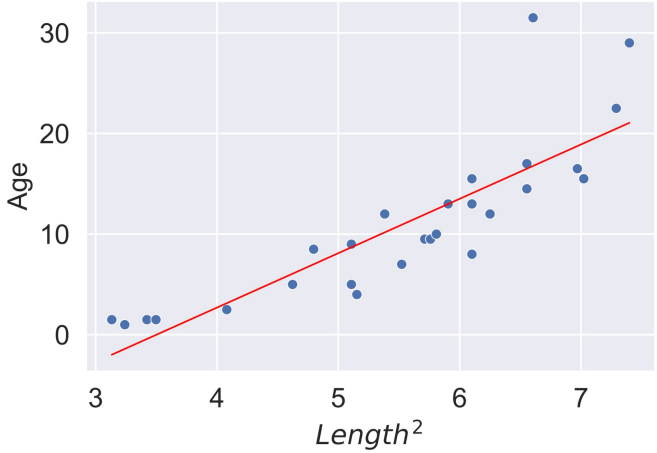
Log(Age) predictions are linear in Length. Age predictions are <u>not</u> linear in Length.

CC BY NC SA

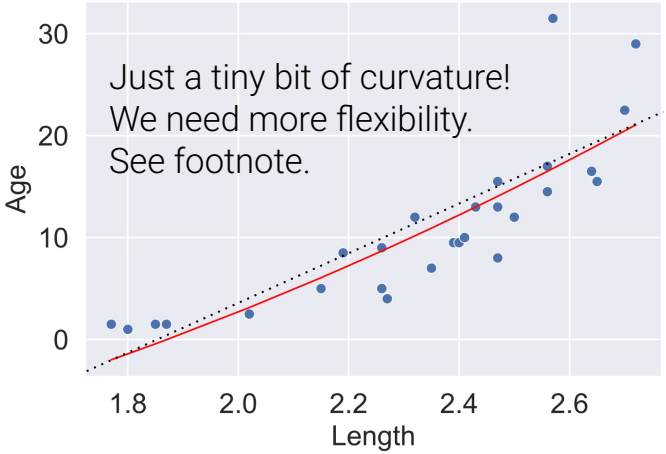
Visualizing squared features on natural scale

2341552

$$\widehat{\text{Age}} = \theta_0 + \theta_1 \text{Length}^2$$



$$\widehat{\text{Age}} = \theta_0 + \theta_1 \text{Length}^2$$



Black line is straight!

Unlike SLR, <u>we</u> know that the x-variable is Length^2!

- Predictions are already on the Age scale, so no need to change the y-axis.
- We can use the non-transformed feature (Length) on the x-axis to see non-linearities.

Age predictions are linear in Length^2, but not linear in Length.

Note: When squaring a feature, we usually include a linear term for extra flexibility (e.g., $\hat{Y} = \theta_0 + \theta_1 X + \theta_2 X^2$). Next lecture!

CC BY NC SA

2341552

Notation for Multiple Linear Regression

Lecture 11, Data 100 Spring 2025

Modeling Process

- Iteration 1: SLR + MSE (last time!)
- Iteration 2: Constant Model + MSE
- Iteration 3: Constant Model + MAE

Transformations of Linear Models

Notation for Multiple Linear Regression

CC BY NC SA

A Note on Terminology

2341552

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)

1

Output

Outcome

Response

Dependent variable

2

Weight(s)

Parameter(s)

Coefficient(s)

4

Prediction

Predicted response

Estimated value

3

Estimator(s)

Optimal parameter(s)

5

Bolded terms are the most common in this course.

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

CC BY NC SA

2341552

Please rank the variables to match them with their appropriate terms (i.e., rank 2 should match "Outcome" / "Response" and rank 3 should match "Prediction"):

Presenting with animations, GIFs or speaker notes? Enable our Chrome extension

slido

A Note on Terminology

2341552

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**.

CC BY NC SA

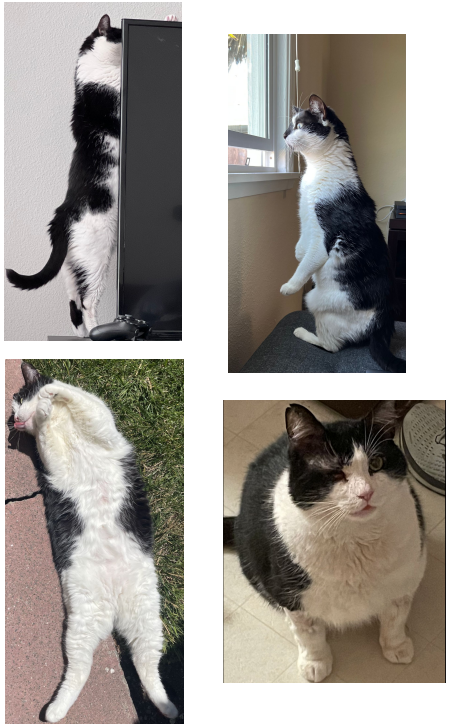
Extending SLR to more than one input

2341552

In SLR, we predict one output with one input.

In real prediction scenarios, we often want to account for more than one input.

Example: Predict cat's 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

CC BY NC SA

Multiple Linear Regression (i.e., ordinary least squares, or OLS)

2341552

Define the **multiple linear regression** model:

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

Parameter vector $\theta = [\theta_0, \theta_1, \ldots, \theta_p]$

CC BY NC SA

Multiple Linear Regression (i.e., ordinary least squares, or OLS)

2341552

Define the **multiple linear regression** model:

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

Parameter vector $\theta = [\theta_0, \theta_1, \ldots, \theta_p]$

This is a **linear combination** of θ_j's, each scaled by x_j

In other words, $\hat{y}$ is linear in θ

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

single input
(p features)

single
prediction

CC BY NC SA

Multiple Linear Regression (i.e., ordinary least squares, or OLS)

2341552

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

More on OLS in the next lecture!

CC BY NC SA

2341552

Lecture 11

Constant Model, Loss, and Transformations

Content credit: Acknowledgments

CC BY NC SA

2341552

Bonus: Constant Model MSE, Algebraic Approach

Lecture 11, Data 100 Spring 2025

MSE Minimization Using an Algebraic Trick

2341552

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!

CC BY NC SA

MSE Minimization Using an Algebraic Trick

2341552

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

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!

CC BY NC SA

Minimization Using an Algebraic Trick

2341552

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!

CC BY NC SA