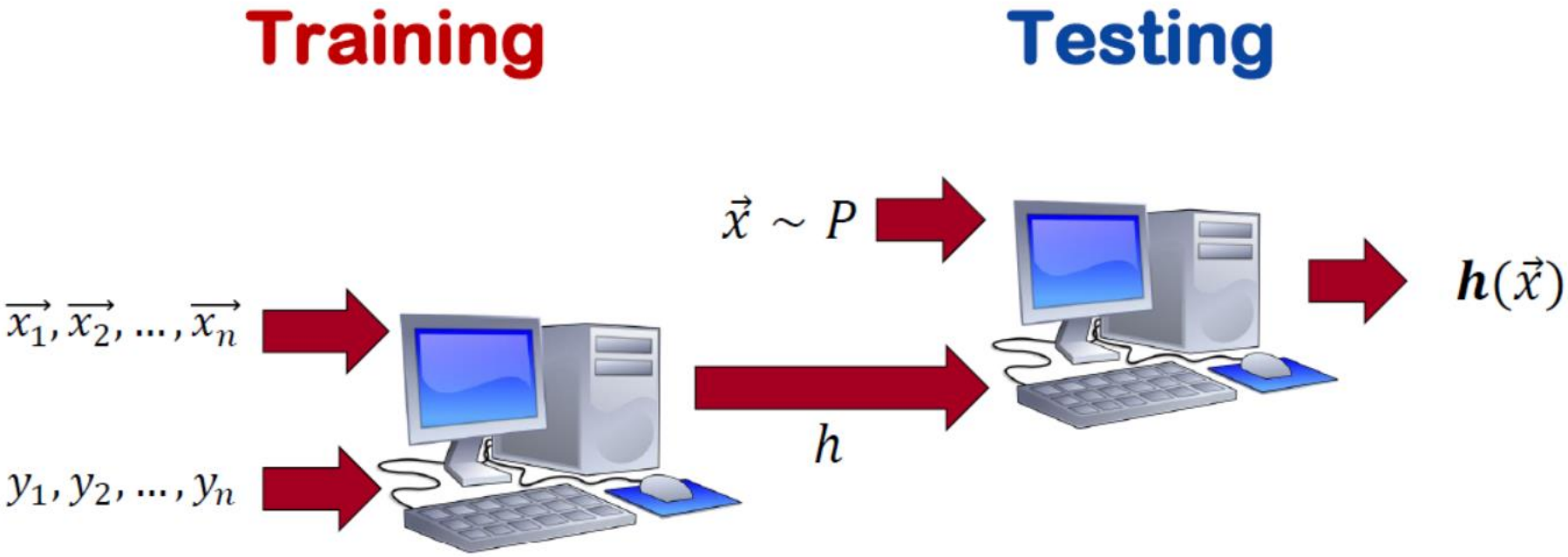
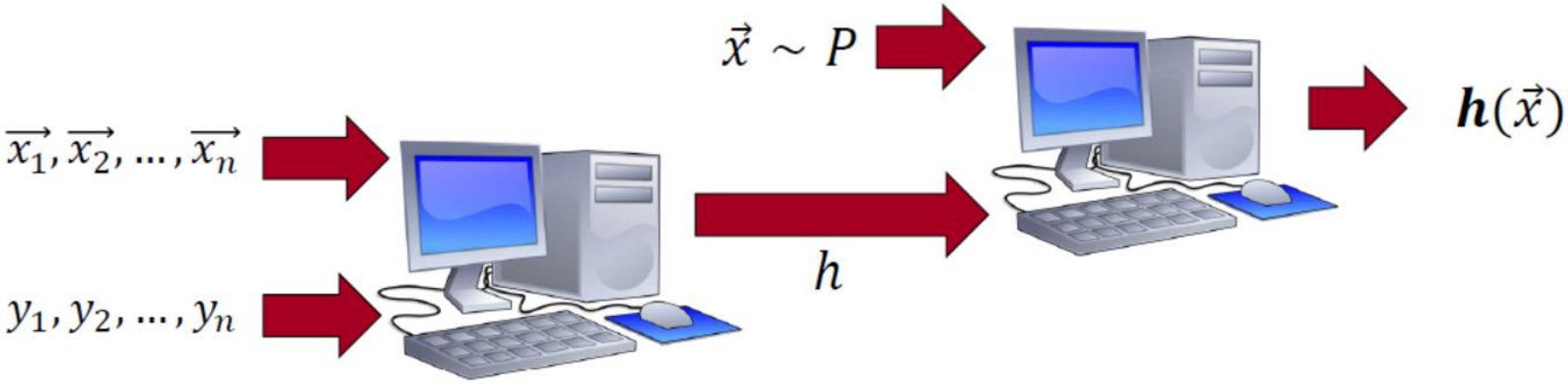
LECTURE 8

What is Machine Learning?


Training
Testing
$\vec{x} \sim P$
$\overrightarrow{x_1}, \overrightarrow{x_2}, \ldots, \overrightarrow{x_n}$
$y_1, y_2, \ldots, y_n$
h
$\boldsymbol{h}(\vec{x})$

Training

Testing

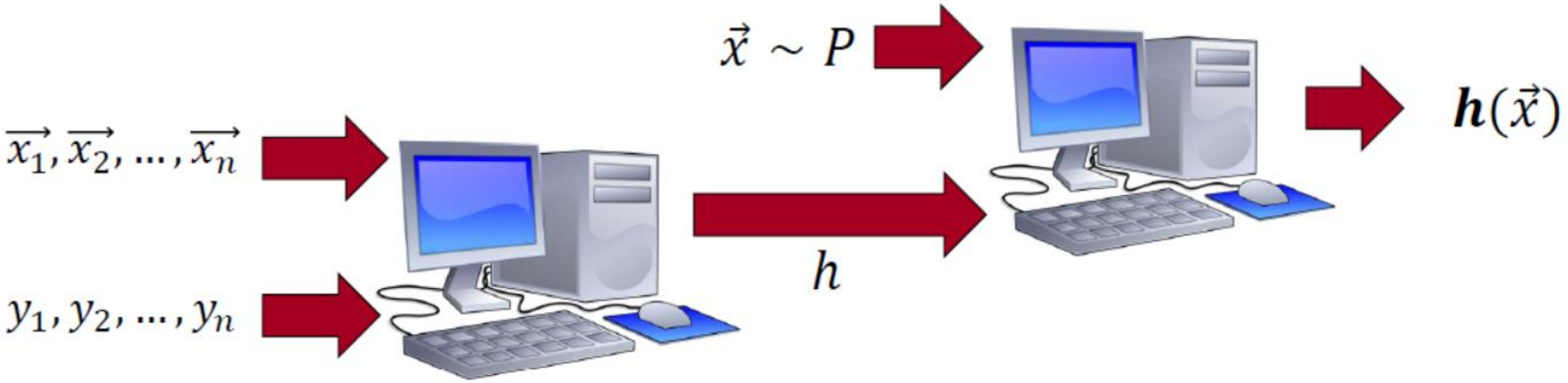


Evaluation

Compare $\boldsymbol{h}(\vec{x})$ and $\boldsymbol{y}$

Training

Testing



Evaluation

Compare $\boldsymbol{h}(\vec{x})$ and $\boldsymbol{y}$

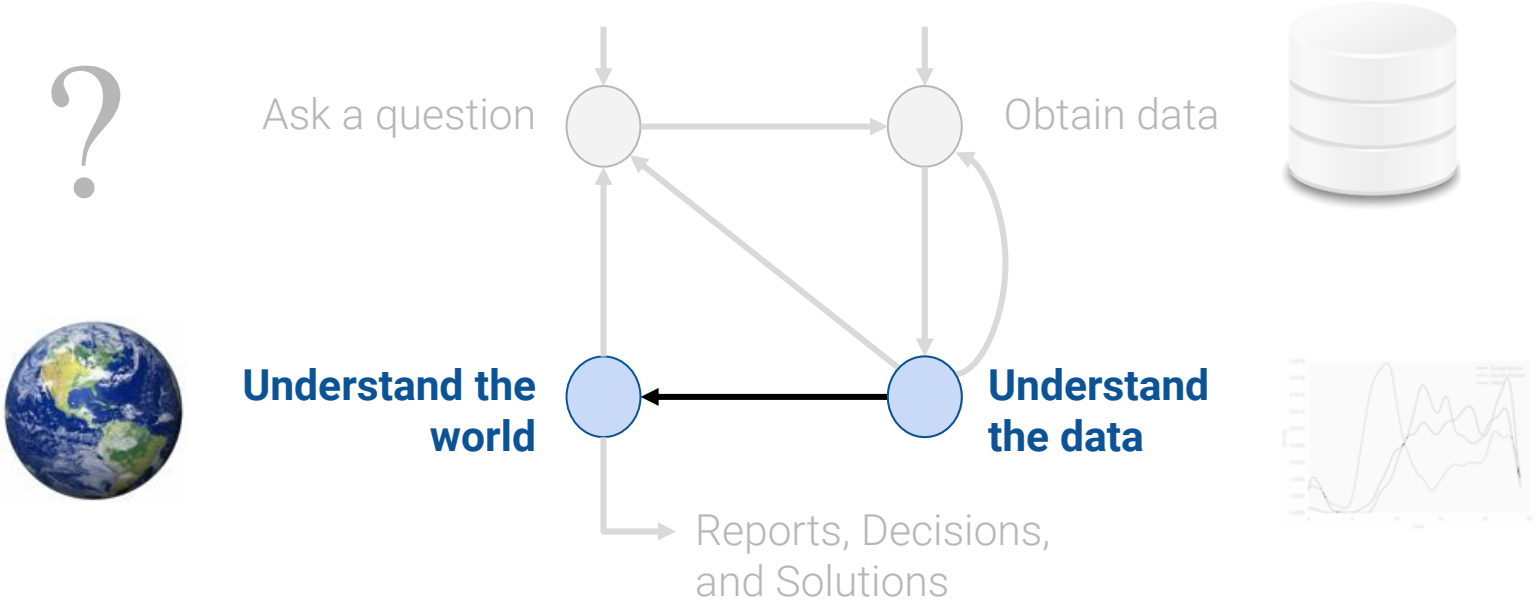
Our hope: $\boldsymbol{h}(\vec{x}) \approx \boldsymbol{y}$

LECTURE 9

Introduction to Modeling, SLR

Understanding the usefulness of models and the simple linear regression model

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

- What is a Model?
  - Regression Line, Correlation
- The Modeling Process
  - Choose a Model
  - Choose a Loss Function
  - Fit the Model
  - Evaluate the Model

What is a Model?

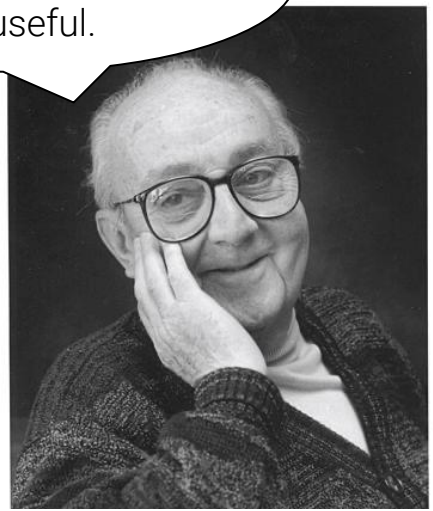
- **What is a Model?**
  - Regression Line, Correlation
- The Modeling Process
  - Choose a Model
  - Choose a Loss Function
  - Fit the Model
  - Evaluate the Model

What is a Model?

A model is an **idealized representation** of a system.

Example:
We model the fall of an object on Earth as subject to a constant acceleration of 9.81 m/s² due to gravity.

- While this describes the behavior of our system, it is merely an approximation.
- It doesn't account for the effects of air resistance, local variations in gravity, etc.
- But in practice, it's accurate enough to be useful!



George Box, Statistician (1919-2013)

Known for	"All models are wrong"
	Response-surface methodology
	EVOP
	q-exponential distribution
	Box–Jenkins method
	Box–Cox transformation

Three Reasons for Building Models

Reason 1:

To explain **complex phenomena** occurring in the world we live in.

- How are the parents' average heights related to the children's average heights?
- How do an object's velocity and acceleration impact how far it travels?

Often times, we care about creating models that are **simple and interpretable**, allowing us to understand what the relationships between our variables are.

Reason 2:

To make **accurate predictions** about unseen data.

- Can we predict if an email is spam or not?
- Can we generate a one-sentence summary of this 10-page long article?

Other times, we care more about making extremely accurate predictions, at the cost of having an uninterpretable model. These are sometimes called **black-box models**, and are common in fields like deep learning.

Reason 3:

To make **causal inferences** about if one thing causes another thing.

- Can we conclude that smoking *causes* lung cancer?
- Does a job training program cause increases in employment and wage?

Much harder question because most statistical tools are designed to infer association not causation

This won't be the focus of this class, but will be if you go on to take more advanced classes (Stat 156, Data 102)

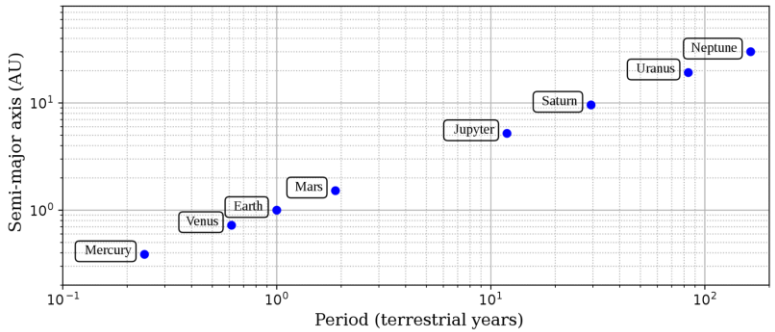
Most of the time, we want to strike a balance between **interpretability** and **accuracy**.

Common Types of Models

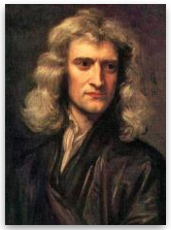
Deterministic physical (mechanistic) models: Laws that govern how the world works.

Kepler's Third Law of Planetary Motion (1619)
The ratio of the square of an object's orbital period with the cube of the semi-major axis of its orbit is the same for all objects orbiting the same primary.

$$T^2 \propto R^3$$

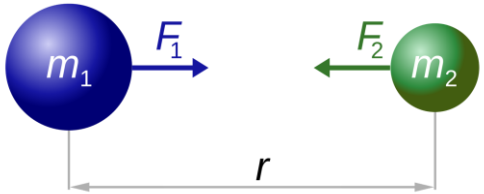


Newton's Laws: motion and gravitation (1687)
Newton's second law of motion models the relationship between the mass of an object and the force required to accelerate it.

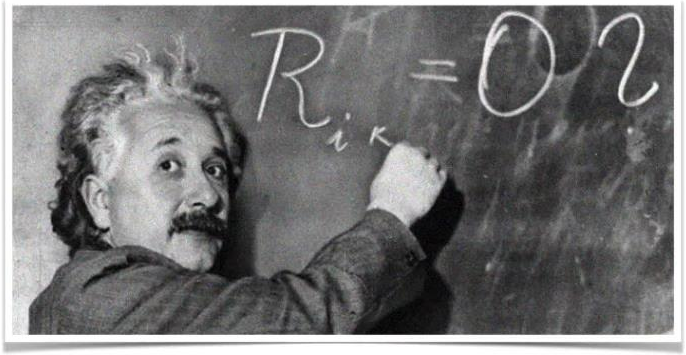


$$\mathbf{F} = m\mathbf{a}$$

$$F = G\frac{m_1 m_2}{r^2}$$



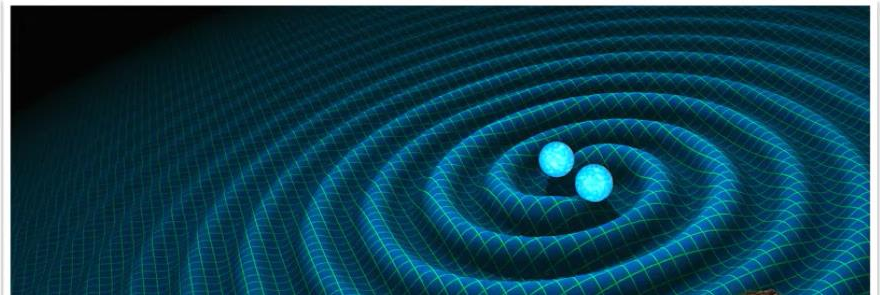
A long time ago in a galaxy far, far away…



$$R_{\mu\nu} - \frac{1}{2}R\,g_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4}T_{\mu\nu}$$

Einstein's Field Equations of General Relativity

Annalen der Physik, 1916



The modeling process

We've implicitly introduced this three-step process, which we will use constantly throughout the rest of the course.

Choose a model

Choose a loss function

Fit the model by minimizing average loss

Regression Line & Correlation

- What is a Model?
- **Regression Line, Correlation**
- The Modeling Process
  - Choose a Model
  - Choose a Loss Function
  - Fit the Model
  - Evaluate the Model

[Review] The Regression Line

From (textbook):

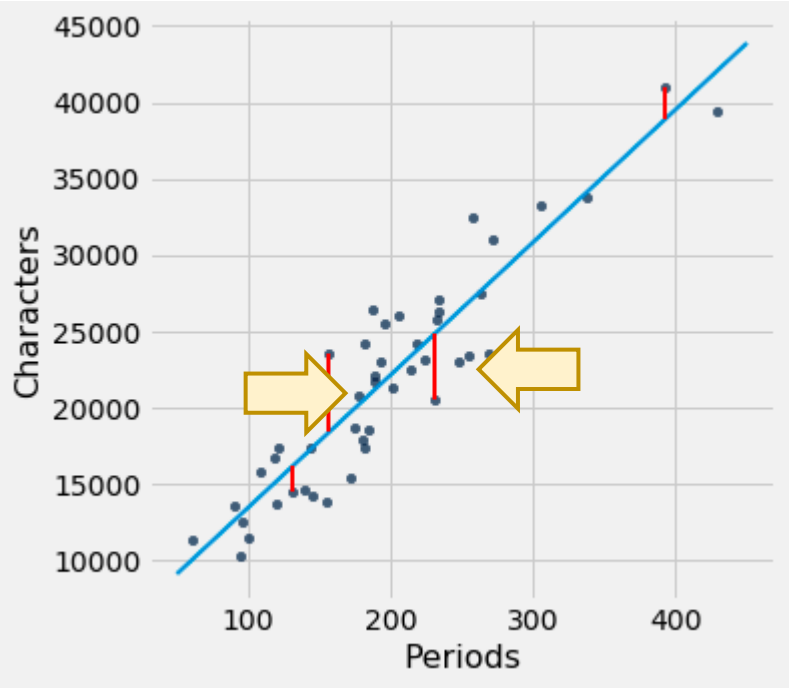
> The **regression line** is the unique straight line that minimizes the **mean squared error** of estimation among all straight lines.

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

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

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

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



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

[Review] Correlation

From (textbook):

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

Define the following:

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

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

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

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

- Correlation measures the strength of a **linear association** between two variables.
- It ranges **between -1 and 1**
  - r = 1 indicates perfect linear association; r = -1 perfect negative association
  - The closer r is to 0, the weaker the linear association is
- It says nothing about **causation** or **non-linear association**
  - Correlation does not imply causation
  - When r = 0, the two variables are **uncorrelated**. However, they could still be related through some non-linear relationship.

The Modeling Process

- What is a Model?
  - Regression Line, Correlation
- **The Modeling Process**
  - Choose a Model
  - Choose a Loss Function
  - Fit the Model
  - Evaluate the Model

Models in Data Science

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

Dataset

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

Observation

(x_i, y_i)

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

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

Prediction

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

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

Models

Some models we will see in the next few lectures:

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

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

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

→ Parametric models

Models in Data Science: Parametric Models

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

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

θ Model parameter(s)

$\hat{\theta}$ Estimated parameter(s), "best" fit to data in some sense

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

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

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

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

Models in Data Science: Parametric Models

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

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

Note: Not all statistical models have parameters!

KDEs, k-Nearest Neighbor classifiers are non-parametric models.

θ Model paramet… …inear model with …meters $\theta = [\theta_0, \theta_1]$

$\hat{\theta}$ Estimated parameter(s), "best" fit to data in some sense } $\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$ The "best" fitting linear model with parameters $\hat{\theta} = [\hat{\theta}_0, \hat{\theta}_1]$

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?

Choose a Model

- What is a Model?
- Data 8 Review
  - Regression Line, Correlation
- **The Modeling Process**
  - **Choose a Model**
  - Choose a Loss Function
  - Fit the Model
  - Evaluate the Model

Simple Linear Regression: Our First Model

Simple Linear Regression Model (SLR)

LA Notation: $\hat{y} = a + bx$ → ML Notation: $\hat{y} = \theta_0 + \theta_1 x$

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

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

x → SLR $\theta = [\theta_0, \theta_1]$ → $\hat{y}$

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

The Modeling Process

1. Choose a model ✓

How should we represent the world?

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

2. Choose a loss function

How do we quantify prediction error?

3. Fit the model

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

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

4. Evaluate model performance

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

Reflect

Loss Functions

- What is a Model?
  - Regression Line, Correlation
- **The Modeling Process**
  - Choose a Model
  - **Choose a Loss Function**
  - Fit the Model
  - Evaluate the Model

The Modeling Process

1. Choose a model ✓

How should we represent the world?

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

SLR model

2. Choose a loss function

How do we quantify prediction error?

3. Fit the model

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

4. Evaluate model performance

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

Loss Functions

We need some metric of how "good" or "bad" our predictions are.

A **loss function** characterizes the **cost**, error, or fit resulting from a particular choice of model or model parameters.

- Loss quantifies how **bad** a prediction is for a **single** observation.
- If our prediction $\hat{y}$ is **close** to the actual value y, we want **low loss**.
- If our prediction $\hat{y}$ is **far** from the actual value y, we want **high loss**.

$$L(y, \hat{y})$$

There are many definitions of loss functions!

The choice of loss function:

- Affects the accuracy and computational cost of estimation.
- Depends on the estimation task:
  - Are outputs quantitative or qualitative?
  - Do we care about outliers?
  - Are all errors equally costly? (e.g., false negative on cancer test)

L2 and L1 Loss

Squared Loss

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

- Widely used.
- Also called "L2 loss".
- Reasonable:
  - $\hat{y} = y$ → good prediction → good fit → no loss
  - $\hat{y}$ far from y → bad prediction → bad fit → *lots of loss*

Absolute Loss

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

- Sounds worse than it is.
- Also called "L1 loss".
- Reasonable:
  - $\hat{y} = y$ → good prediction → good fit → no loss
  - $\hat{y}$ far from y → bad prediction → bad fit → *some loss*

L2 and L1 Loss for SLR

Squared Loss

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

- Widely used.
- Also called "L2 loss".
- Reasonable:
  - $\hat{y} = y$ → good prediction → good fit → no loss
  - $\hat{y}$ far from y → bad prediction → bad fit → *lots of loss*

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

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

Absolute Loss

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

- Sounds worse than it is.
- Also called "L1 loss".
- Reasonable:
  - $\hat{y} = y$ → good prediction → good fit → no loss
  - $\hat{y}$ far from y → bad prediction → bad fit → *some loss*

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

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

Question?

Why don't we use residual error directly and instead we use absolute loss or squared loss?

Residuals as loss function?

Why don't we directly use residual error as the loss function?

$$e = (y - \hat{y})$$

- Doesn't work: big negative residuals shouldn't cancel out big positive residuals!
  - Our predictions can be very off, but we can still get a zero residual

Empirical Risk is Average Loss over Data

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

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

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

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

Function of the parameter θ (holding the data fixed) because θ determines $\hat{y}$.

The average loss on the sample tells us how well the model fits the data (not the population).

But hopefully these are close.

Empirical Risk is Average Loss over Data

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

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

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

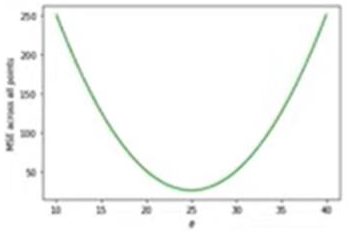
$$\hat{R}(\theta) = \frac{1}{n} \sum_i L(y_i, \hat{y}_i)$$

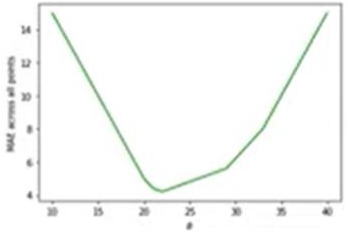
The colloquial term for average loss depends on which loss function we choose.

L2 loss → **Mean Squared Error (MSE)** $\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

L1 loss → **Mean Absolute Error (MAE)** $\frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$

Answer





MSE vs. MAE

What else is different about squared loss (MSE) and absolute loss (MAE)?

Mean squared error (optimal parameter for the constant model is the sample mean)

- Very smooth. Easy to minimize using numerical methods (coming later in the course).
- Very sensitive to outliers, e.g. if we added 1000 to our largest observation, the optimal theta would become 225 instead of 25.

Mean absolute error (optimal parameter for the constant model is the sample median)

- Not as smooth – at each of the "kinks," it's not differentiable. Harder to minimize.
- Robust to outliers! E.g, adding 1000 to our largest observation doesn't change the Median.

The Modeling Process

1. Choose a model ✓

How should we represent the world?

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

SLR model

2. Choose a loss function ✓

How do we quantify prediction error?

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

Squared loss

3. Fit the model

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

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

MSE for SLR

4. Evaluate model performance

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

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

The Modeling Process

1. Choose a model ✓

How should we represent the world?

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

2. Choose a loss function ✓

How do we quantify prediction error?

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

3. Fit the model

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

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

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

4. Evaluate model performance

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

Fit the Model

- What is a Model?
  - Regression Line, Correlation
- **The Modeling Process**
  - Choose a Model
  - Choose a Loss Function
  - **Fit the Model**
  - Evaluate the Model

Minimizing MSE for the SLR Model

Recall: we wanted to pick the **regression line** $\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$

To minimize the (sample) **Mean Squared Error**: $MSE(\theta_0, \theta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1 x_i))^2$

To find the best values, we set derivatives equal to zero to **obtain the optimality conditions:**

$$\frac{\partial}{\partial \theta_0} MSE = 0$$

$$\frac{\partial}{\partial \theta_1} MSE = 0$$

Partial Derivative of MSE with Respect to θ_0, θ_1

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

Derivative of sum is sum of derivatives

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

Chain rule

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

Simplify constants

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

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

Derivative of sum is sum of derivatives

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

Chain rule

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

Simplify constants

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

Estimating Equations

To find the best values, we set derivatives equal to zero to **obtain the optimality conditions:**

$$0 = \frac{\partial}{\partial \theta_0} MSE = -\frac{2}{n} \sum_{i=1}^{n} \left(y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i \right) \iff \frac{1}{n} \sum_i y_i - \hat{y}_i = 0$$

"Equivalent"

$$0 = \frac{\partial}{\partial \theta_1} MSE = -\frac{2}{n} \sum_{i=1}^{n} \left(y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i \right) x_i \iff \frac{1}{n} \sum_i (y_i - \hat{y}_i) x_i = 0$$

Estimating equations

To find the best θ_0, θ_1, we need to solve the **estimating equations** on the right.

From Estimating Equations to Estimators

Goal: Choose θ_0, θ_1 to solve two estimating equations:

$$\frac{1}{n}\sum_i y_i - \hat{y}_i = 0 \quad \boxed{1} \qquad \text{and} \qquad \frac{1}{n}\sum_i (y_i - \hat{y}_i)x_i = 0 \quad \boxed{2}$$

$\boxed{1}$

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i\right) = 0 \iff \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) - \hat{\theta}_0 - \hat{\theta}_1\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = 0$$

$$\iff \bar{y} - \hat{\theta}_0 - \hat{\theta}_1 \bar{x} = 0$$

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

From Estimating Equations to Estimators

Goal: Choose θ_0, θ_1 to solve two estimating equations:

$$\frac{1}{n}\sum_i y_i - \hat{y}_i = 0 \quad \boxed{1} \qquad \text{and} \qquad \frac{1}{n}\sum_i (y_i - \hat{y}_i)x_i = 0 \quad \boxed{2}$$

Now, let's try: $\boxed{2}$ - $\boxed{1}$ * $\bar{x}$

$$\frac{1}{n}\sum_i (y_i - \hat{y}_i)x_i - \frac{1}{n}\sum_i (y_i - \hat{y}_i)\bar{x} = 0 \iff \frac{1}{n}\sum_i (y_i - \hat{y}_i)(x_i - \bar{x}) = 0$$

$$\left(\text{using } \hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_i\right) \Rightarrow \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i\right)(x_i - \bar{x}) = 0$$

$$\left(\text{using } \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}\right) \Rightarrow \frac{1}{n}\sum_i \left(y_i - \bar{y} + \hat{\theta}_1 \bar{x} - \hat{\theta}_1 x_i\right)(x_i - \bar{x}) = 0$$

$$\Rightarrow \frac{1}{n}\sum_i \Big((y_i - \bar{y}) - \hat{\theta}_1 (x_i - \bar{x})\Big)(x_i - \bar{x}) = 0$$

From Estimating Equations to Estimators

$$\Rightarrow \frac{1}{n}\sum_i \left[(y_i - \bar{y})(x_i - \bar{x}) - \hat{\theta}_1 (x_i - \bar{x})^2\right] = 0$$

$$\Rightarrow \frac{1}{n}\sum_i (y_i - \bar{y})(x_i - \bar{x}) = \hat{\theta}_1 \frac{1}{n}\sum_i (x_i - \bar{x})^2$$

Plug in definitions of correlation and SD:

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

Solve for $\hat{\theta}_1$:

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

Reminder

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

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

Estimating Equations

Estimating equations are the equations that the model fit has to solve. They help us:

- Derive the estimates
- Understand what our model is paying attention to

$$\frac{1}{n}\sum_i y_i - \hat{y}_i = 0$$

$$\frac{1}{n}\sum_i (y_i - \hat{y}_i)x_i = 0$$

For SLR:

- The residuals should **average to zero** (otherwise we should fix the intercept!)
- The residuals should be **orthogonal to the predictor variable** (or we should fix the slope!)

Very important for HW 5

Exploring MAE

When we use absolute (or L1) loss, we call the average loss **mean absolute error**. For the constant model, our MAE looks like:

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

Let's again re-visit our toy example of 5 observations, **[20, 21, 22, 29, 33]**.

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

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

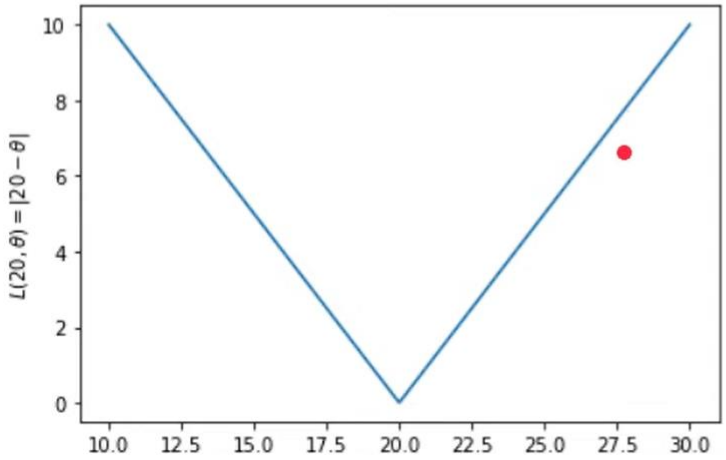
Exploring MAE

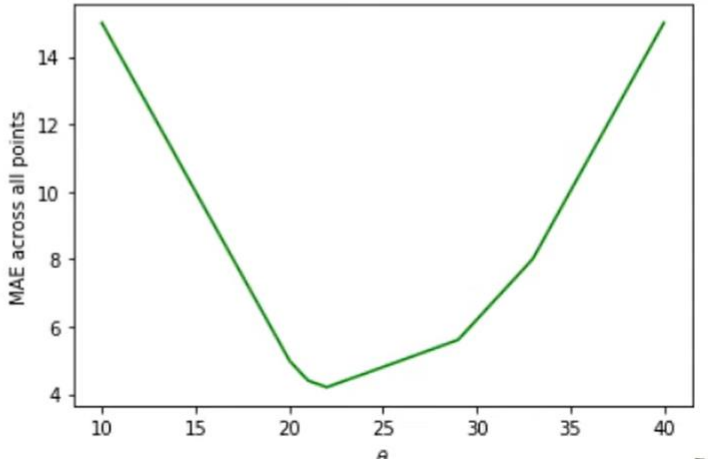
$$L_1(20, \theta) = |20 - \theta|$$

The loss for the first observation (y_1).

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

The average loss across all observations (the MAE).





Answer

$$R(\theta) = \frac{1}{n}\sum_i |y_i - \theta|$$

$$\frac{dR(\theta)}{d\theta} = \frac{1}{n}\sum_i \frac{d|y_i-\theta|}{d\theta}$$

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

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

$|x|$

$$|x| = \begin{cases} x & x \geq 0 \\ -x & x < 0 \end{cases}$$

$$\frac{d|x|}{dx} = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \end{cases}$$

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

What is a Model?

- **What is a Model?**
  - Regression Line, Correlation
- The Modeling Process
  - Choose a Model
  - Choose a Loss Function
  - Fit the Model
  - Evaluate the Model

The Modeling Process

1. Choose a model ✓ — How should we represent the world?

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

2. Choose a loss function ✓ — How do we quantify prediction error?

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

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

4. Evaluate model performance

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

Next Time!

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

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

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

Mean Squared Error (MSE)

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

3. Fit the model

Minimize average loss with calculus

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

4. Evaluate model performance

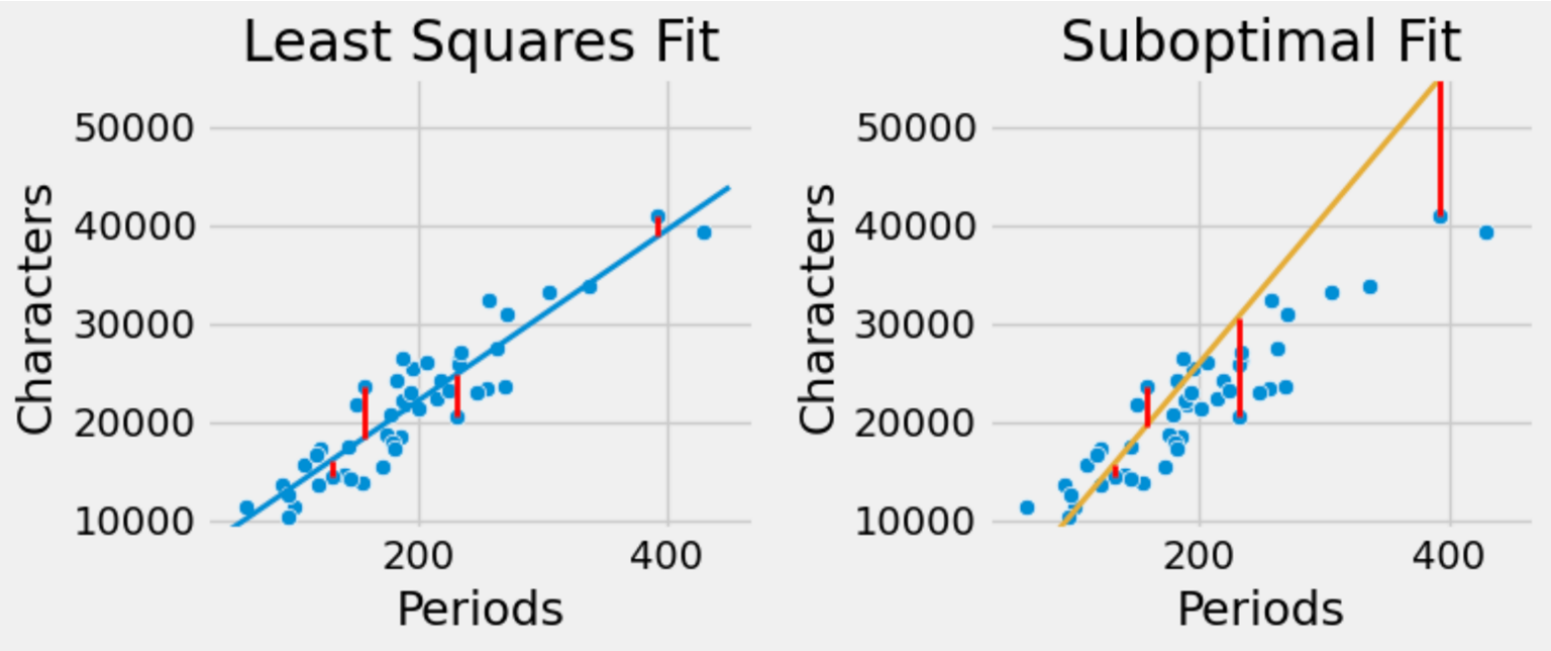
Visualize, Root MSE

Minimizing MSE is Minimizing Squared Residuals

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

Residual ("error") in prediction

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

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

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

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

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.

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	

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

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

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

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

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

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

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

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

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. 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

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

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

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

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

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.

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

Demo

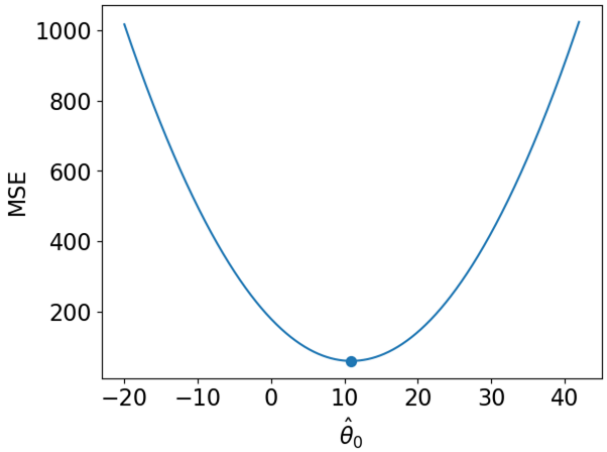
[Loss] Comparing Two Different Models, Both Fit with MSE

Constant Model

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

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

Loss surface is **2-D**.



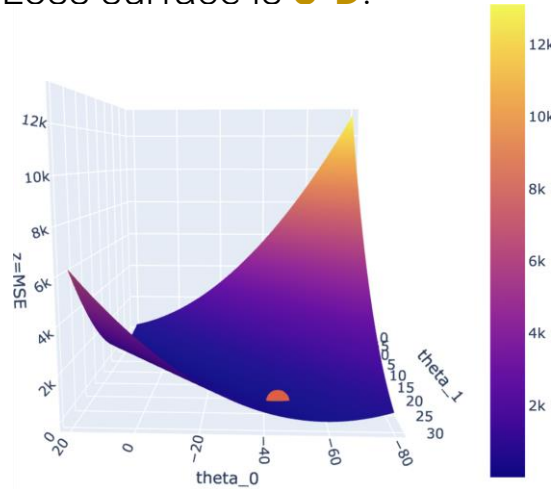
$$\hat{R}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_0)^2$$

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**.



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

[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

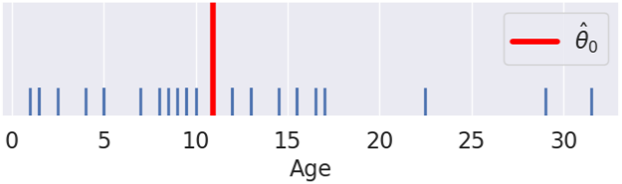
[Fit] Comparing Two Different Models, Both Fit with MSE

Constant Model

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

RMSE: **7.72**

Predictions on a **rug plot**.

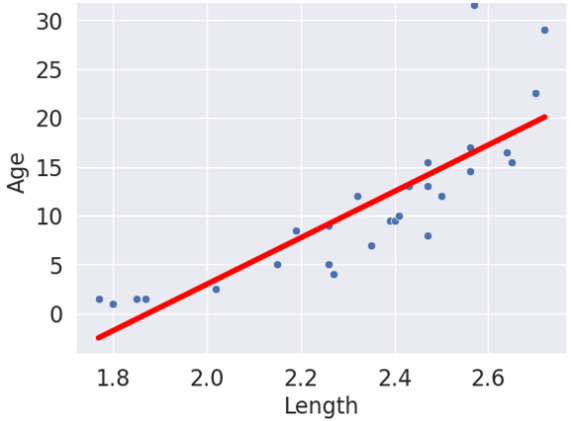


Simple Linear Regression

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

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

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
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

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

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

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 (M**A**E) 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.

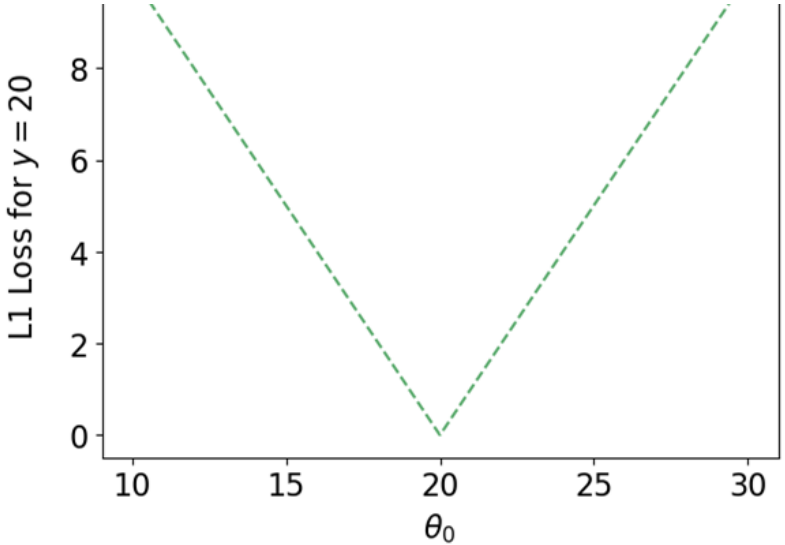
Exploring MAE: A Piecewise function

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:

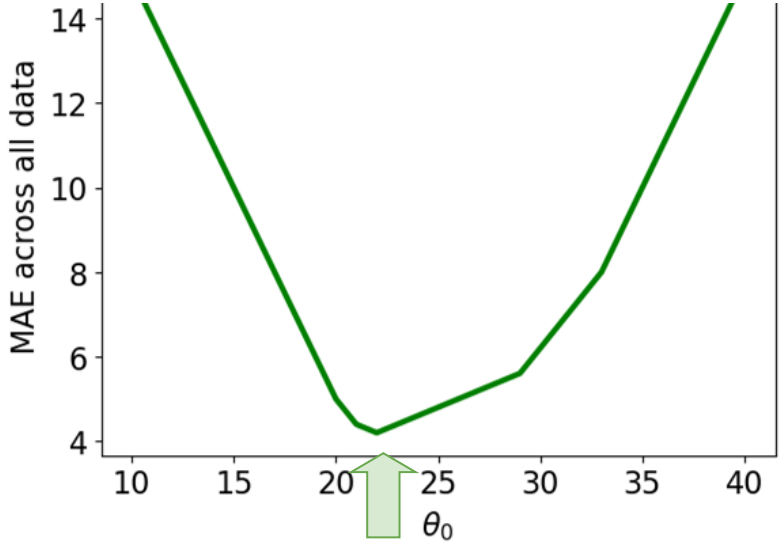
$$L_1(20, \theta_0) = |20 - \theta_0|$$



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 $\hat{\theta}_0$.

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

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

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

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

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

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

Fit the Model: Calculus

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

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

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

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

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

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

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:

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

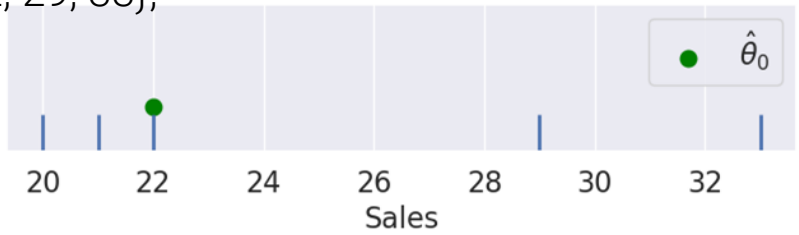
In other words, theta needs to be such that there are **an equal # of points to the left and right**.

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

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

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

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

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

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**:

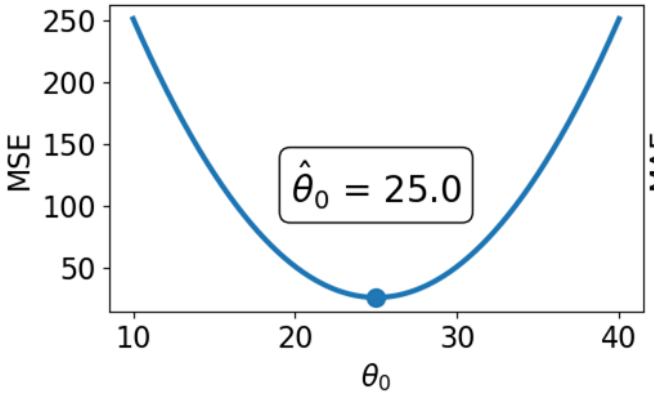
$$\hat{\theta}_0 = \text{mean}(y) = \bar{y}$$

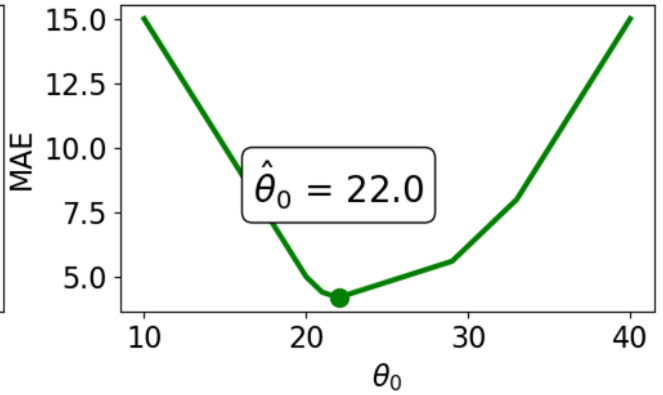
MAE (Mean Absolute Loss)

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

Minimized with **sample median**:

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

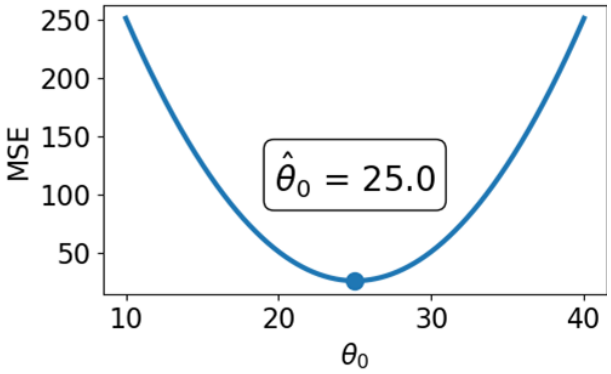




MSE and MAE: Comparing Loss Surfaces

MSE (Mean Squared Loss)

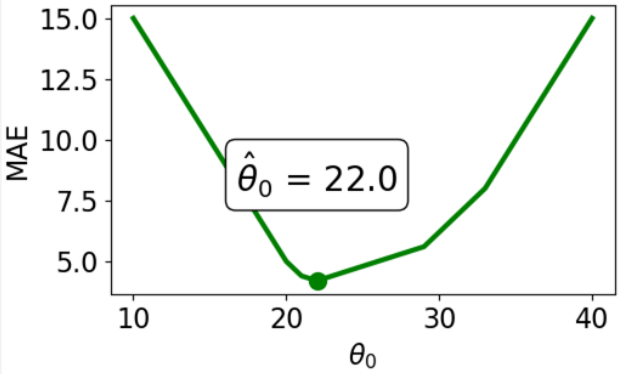
$$\hat{\theta}_0 = mean(y) = \bar{y}$$



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.

Demo

Demo

MSE and MAE: Comparing Sensitivity to Outliers

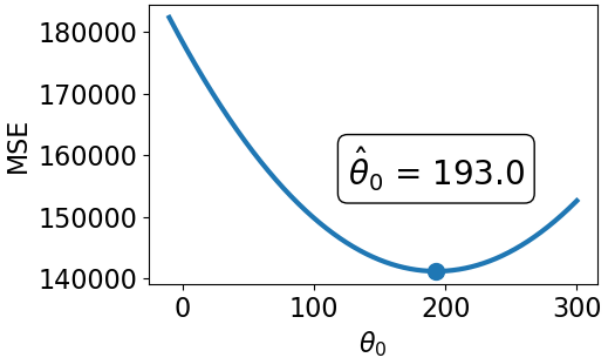
MSE (Mean Squared Loss)
Minimized with **sample mean**:

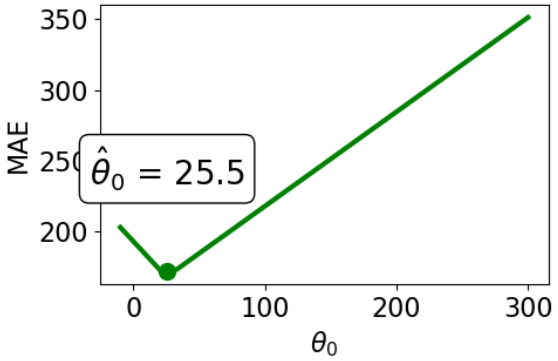
$$\hat{\theta}_0 = \text{mean}(y) = \bar{y}$$

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.

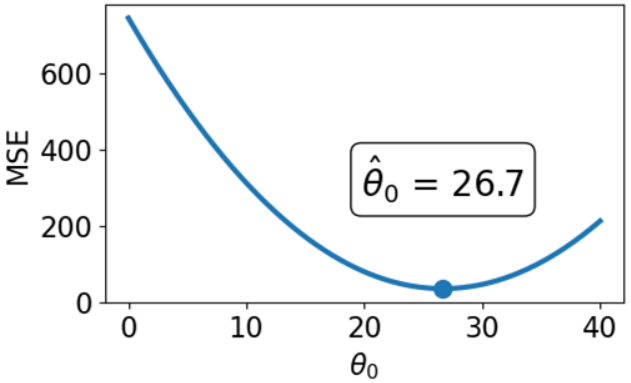
More robust to outliers.

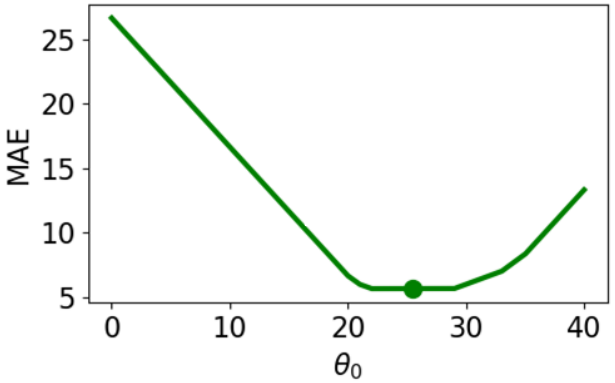
MSE and MAE: Comparing Uniqueness of Solutions

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).

Demo

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 Linear Algorithm.

Transformations to Fit Linear Models

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

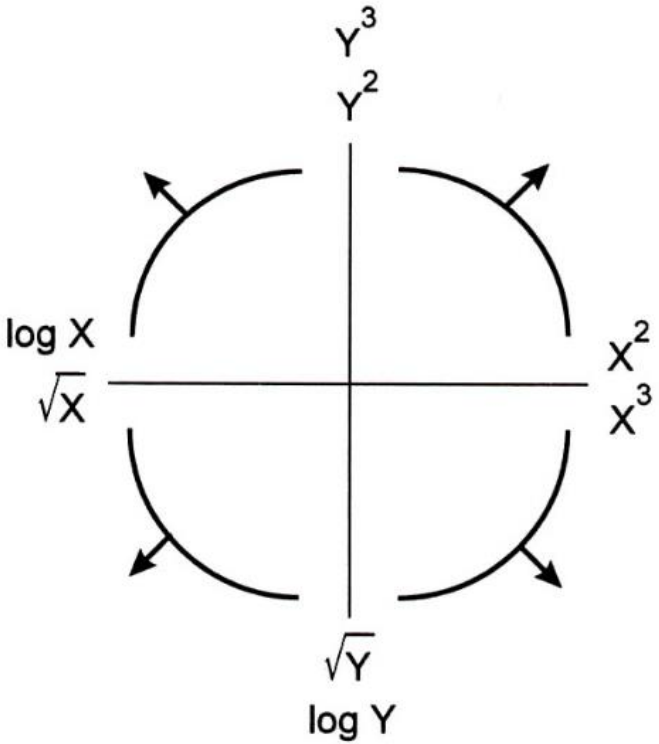
Tukey-Mosteller Bulge Diagram (From Lecture 7)

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



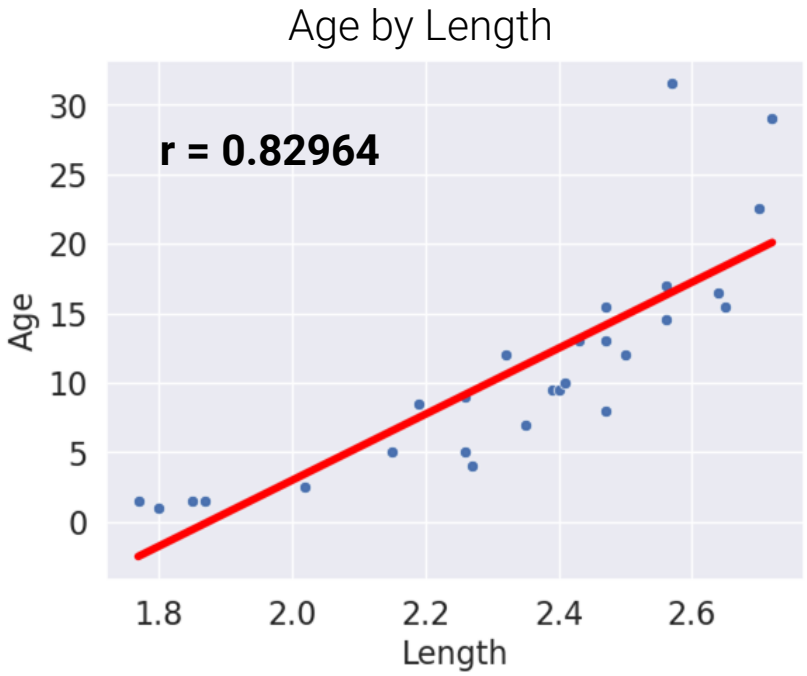
Back to Least Squares Regression with Dugongs

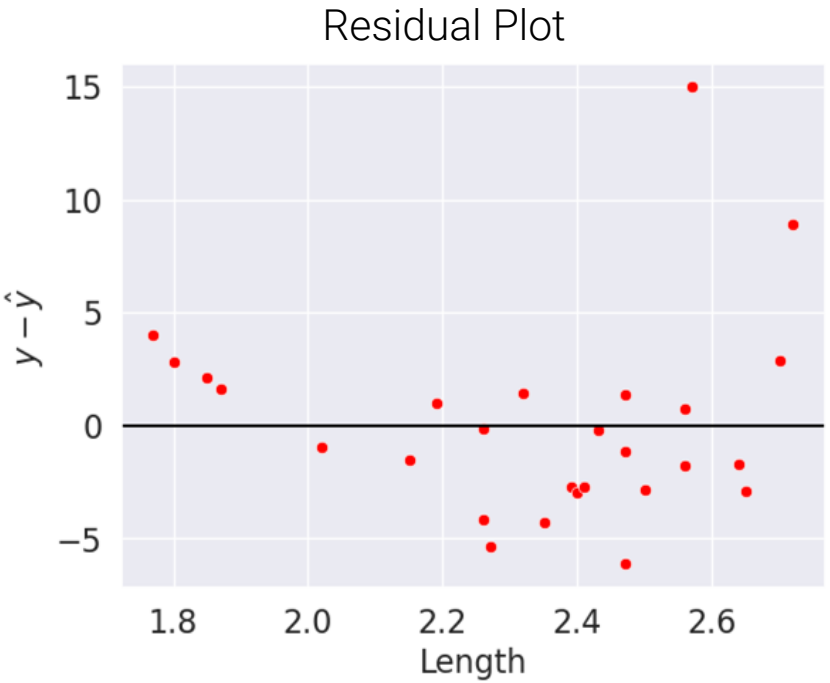
From Data 8 (textbook):

> 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?

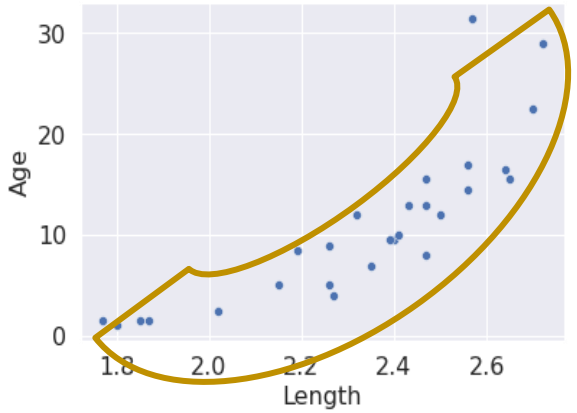
A: **Transform the Data**.

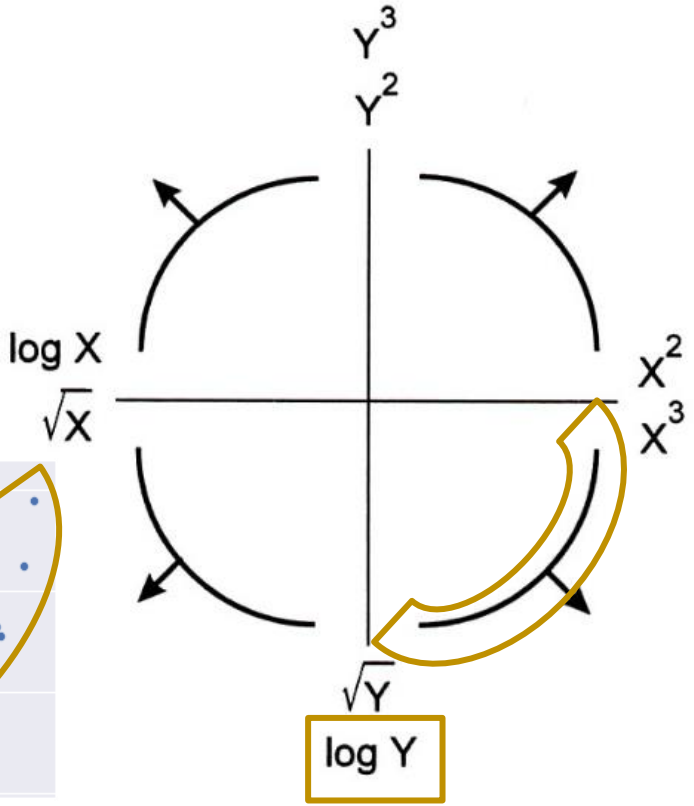
Tukey-Mosteller Bulge Diagram

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

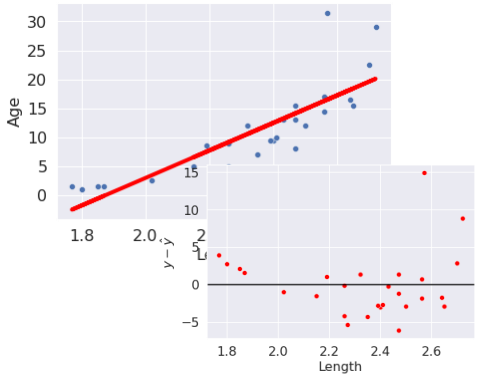
In other words, if we apply the variable transform $z = \log(y)$:

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

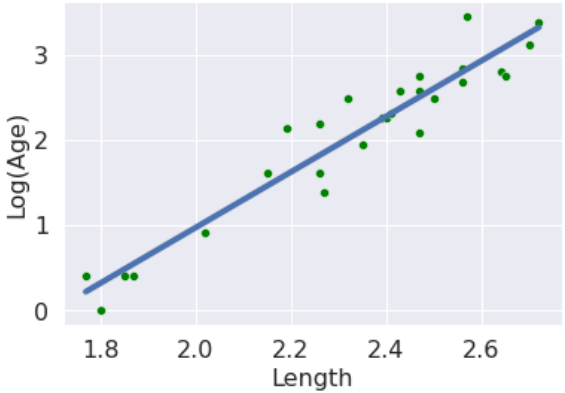
$$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)^2$$

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

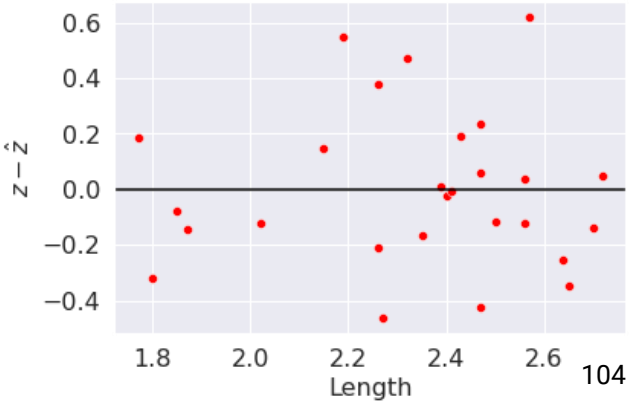
Original (Age by Length)



Log(Age) by Length



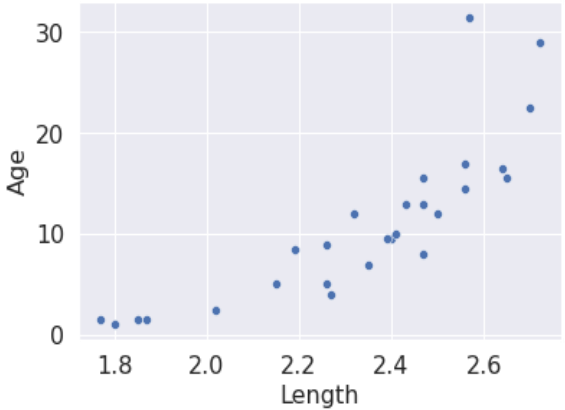
Residual Plot



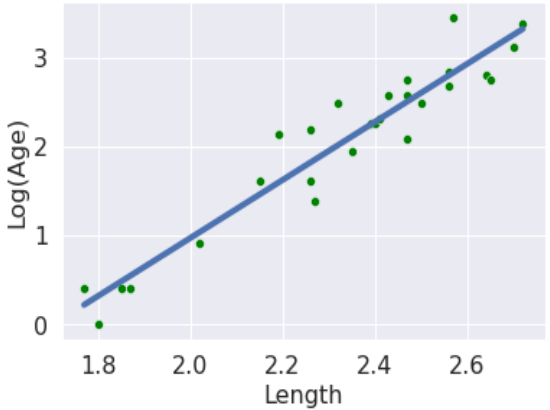
Fit a Curve using Least Squares Regression

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

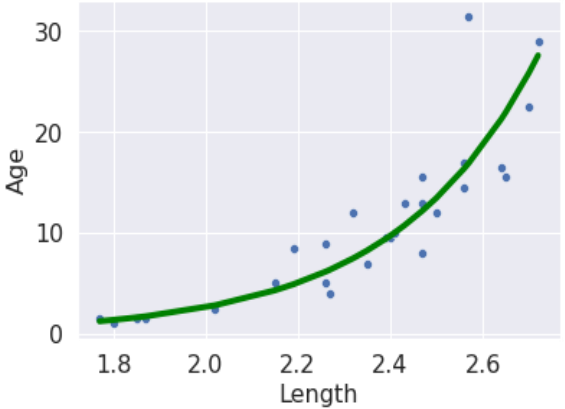
Age by Length



Log(Age) by Length



Age by Length



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

Prediction

Predicted response

Estimated value

Weight(s)

Parameter(s)

Coefficient(s)

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

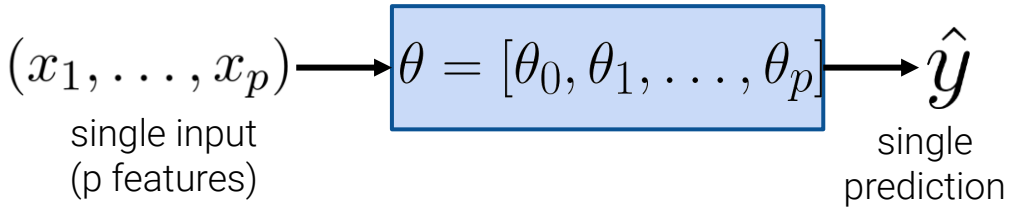
Multiple Linear Regression

Define the **multiple linear regression** model:

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

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

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



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

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

intercept

parameter for length

parameter for weight

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

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

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

variance of sample!

from the previous slide

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

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

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

Minimization using an algebraic trick

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

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

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

Looks familiar!