

LECTURE 26

Introduction to Modeling

Understanding the usefulness of models, and how loss functions help create them.

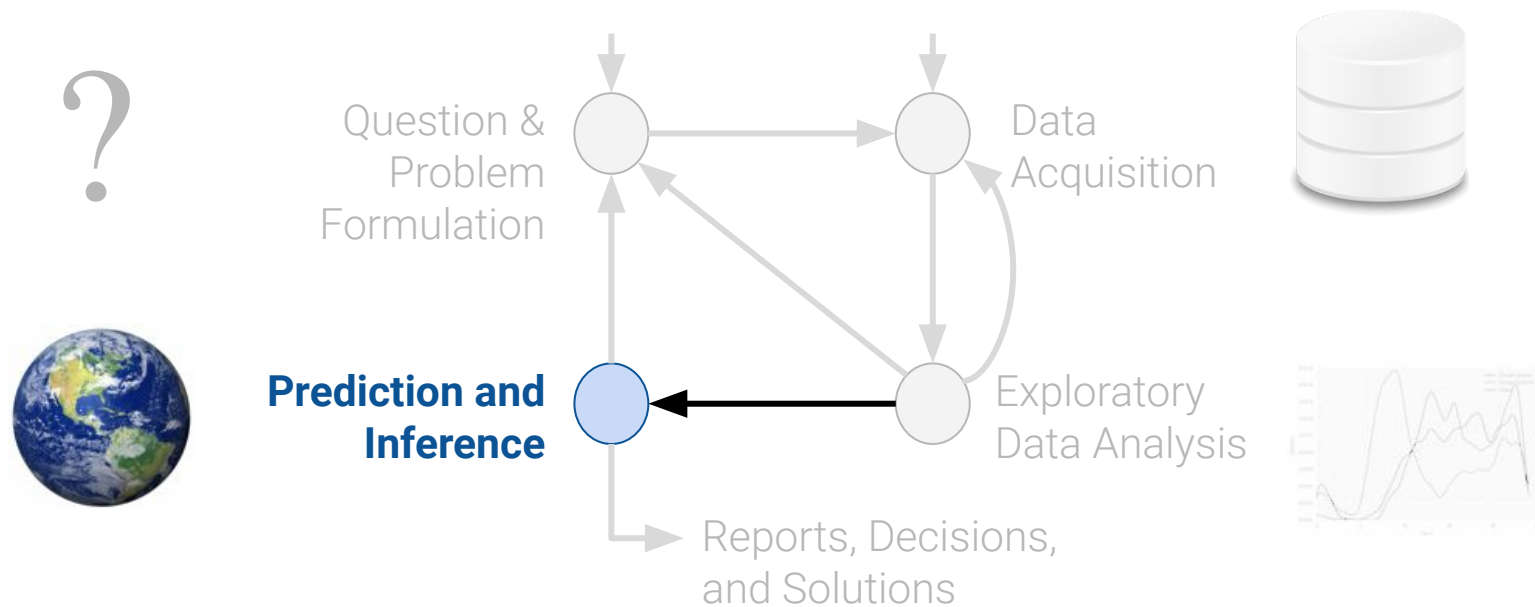
CSCI 3022

Maribeth Oscamou

Content credit: [Acknowledgments](#)

- HW 10 Due Thursday: Corrections made this morning to otter grader for several questions. Please make sure you are using v2
- Lab 10 (Confidence Intervals and Climate Modeling) posted from yesterday
- Quiz 9 Friday:
 - Scope: HW 9,
 - HW 9, nb 9
 - Lesson 21 (A/B Testing, Permutation Tests & Causality)
 - Lesson 22 (Hypothesis Test Errors)
- Exam 2 next Friday
 - Scope: Cumulative. Questions will focus on concepts/topics from
 - Lessons 17-Lessons 25 (i.e. HW 7-10, nb 7-10)

Plan for Rest of Semester: Modeling



(today)

Modeling I:
Different models, loss
functions

Modeling II:
Simple Linear
Regression, linearization

Modeling III:
Multiple Linear
Regression

Today's Roadmap

What Is A Model?

The Modeling Process: Definitions

Loss Functions

Constant Model + MSE

Changing the Loss: Constant Model + MAE

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/s^2 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!

Essentially, all models are wrong, but some are 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

Common Types of Models

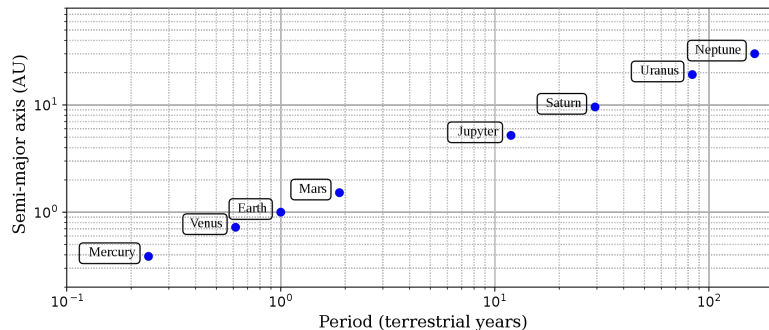
Deterministic physical (mechanistic) models: Laws that govern how the world works.

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.



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



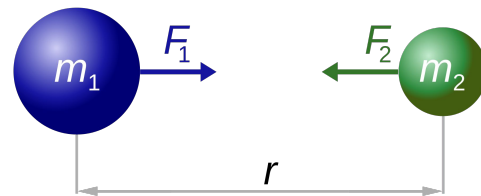
Newton's Laws: motion and gravitation (1687)

Newton's second law of motion models the relationship between the mass of an object and the force required to accelerate it.



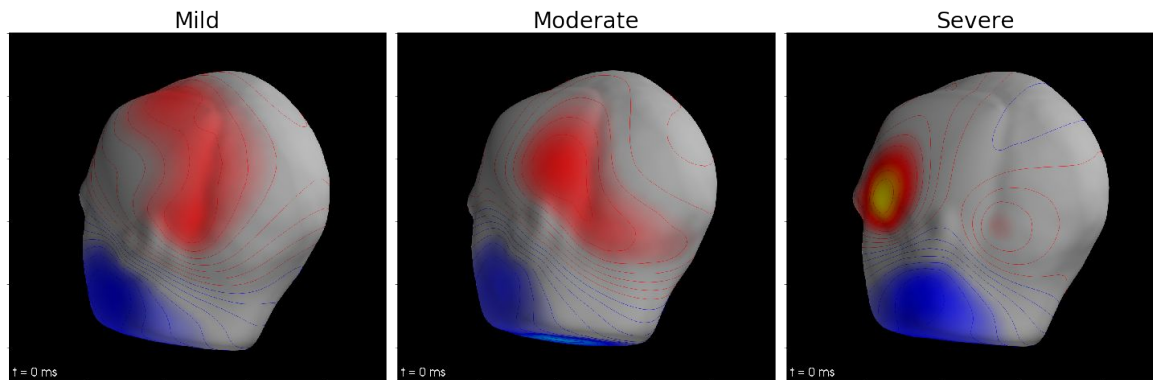
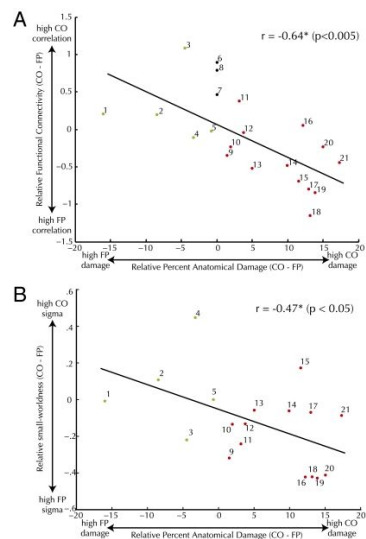
$$\mathbf{F} = m\mathbf{a}$$

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



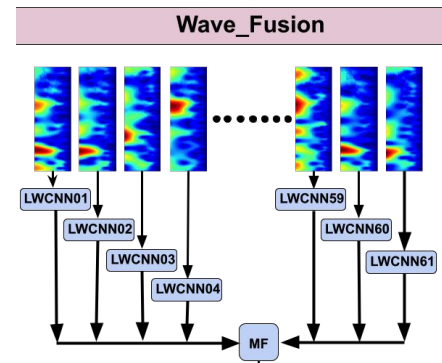
Probabilistic models

- Models of how random processes evolve.
- Often motivated by understanding of an unpredictable system.



<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2900657/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3575518/>



Prediction vs. Inference

Why do we build models?

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

Prediction is the task of using our model to make predictions for the response (output) of unseen data.

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

Inference is the task of using our model to draw conclusions about the underlying true relationship(s) between our features and response.

Prediction vs. Inference

Why do we build models?

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

Prediction is the task of using our model to make predictions for the response (output) of unseen data.

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

Inference is the task of using our model to draw conclusions about the underlying true relationship(s) between our features and response.

Example: Suppose we are interested in studying the relationship between the value of a home and a view of a river, school districts, property size, income level of community, etc.

Prediction: Given the attributes of some house, how much is it worth?

We care more about making accurate predictions, don't care so much about how.

Inference: How much more are houses with river views worth (holding other variables fixed)?

We care more about having model parameters that are interpretable and meaningful.

Modeling Process

What Is A Model?

The Modeling Process: Definitions

Loss Functions

Constant Model + MSE

Changing the Loss: Constant Model + MAE

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?

Example

You work at a local boba tea store and want to estimate the sales each day.

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

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

How many drinks will you sell tomorrow?



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



The Constant Model

You work at a local boba tea store and want to estimate the sales each day.

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

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

How many drinks will you sell tomorrow?



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

This is a **constant model**.

y represents our **true observations** (e.g. the actual observed sales).

y_i represents the i th observation in particular (e.g. y_4 is the sales from the 4th day).
In general, we represent our collected data as y_1, y_2, \dots, y_n .

\hat{y} represents the **predicted observations** given by our model (e.g. the predicted sales).
 \hat{y}_i represents the i th prediction in particular

θ represents the **parameter(s) of our model**. This is what we are trying to **estimate**!
Parameters are what define our model. We make this more clear in the next slide.

$\hat{\theta}$ represents the **fitted, or optimal, parameter(s)** that we solve for. It is our goal to find this!
We want to find $\hat{\theta}$ to make the best possible model.

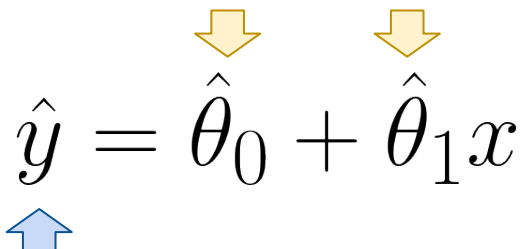
Terminology: Prediction vs. Estimation

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.

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.

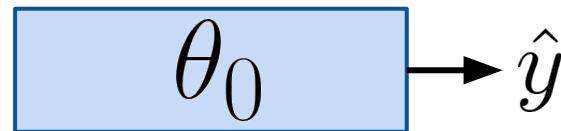
It ignores any relationships between variables.

- 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.



1. Choose a model

Constant Model:

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

2. Choose a loss function

3. Fit the model

4. Evaluate model performance



1. Choose a model

Constant Model

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

2. Choose a loss function

3. Fit the model

4. Evaluate model performance

Loss Functions

What Is A Model?

The Modeling Process: Definitions

Loss Functions

Changing the Model: Constant Model + MSE

Changing the Loss: Constant Model + MAE

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 Loss or Squared Loss

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

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

L1 Loss or Absolute Loss

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

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

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

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

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

The average loss on the sample tells us how well it 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 $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$:

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

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)**

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

Constant Model + Mean Squared Error

What Is A Model?

The Modeling Process: Definitions

Loss Functions

Constant Model + MSE

Changing the Loss: Constant Model + MAE

The Modeling Process:

1. Choose a model



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

2. Choose a loss function



L2 Loss

Mean Squared Error (MSE)

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

3. Fit the model

Minimize
average loss

4. Evaluate model
performance

Fit the Model: Rewrite MSE for the Constant Model

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

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

L2 loss on a
single datapoint

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

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

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

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

Approach 1 If you know your data $\mathcal{D} = \{20, 21, 22, 29, 33\}$ you could modify the objective by plugging in values first:

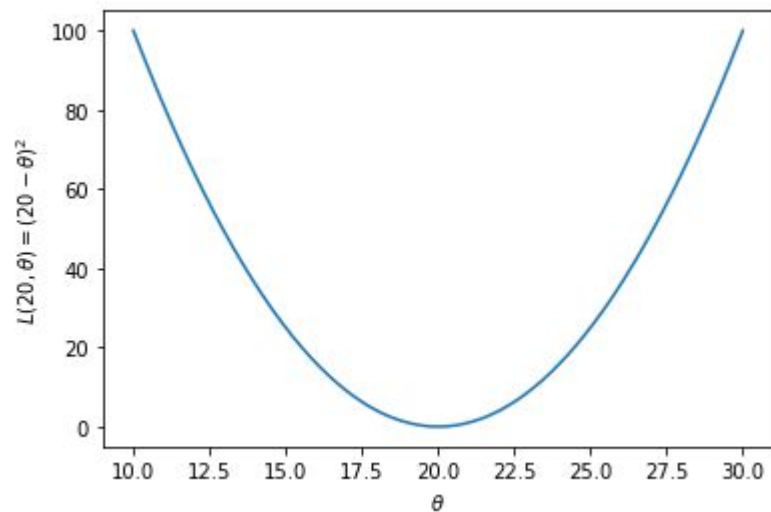
$$R(\theta) = \frac{1}{5} ((20 - \theta_0)^2 + (21 - \theta_0)^2 + (22 - \theta_0)^2 + (29 - \theta_0)^2 + (33 - \theta_0)^2)$$

Approach 2 If you want to prove the general case for any data, you could directly minimize the objective using Calculus. You did this in HW 1!

Exploring MSE: Approach 1

$$L_2(20, \theta) = (20 - \theta)^2$$

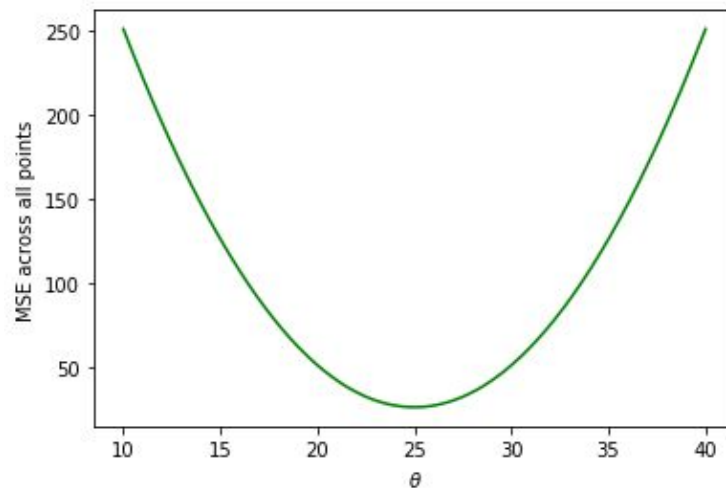
The loss for the first observation (y_1).



A parabola, minimized at $\theta = 20$.

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

The average loss across all observations (the MSE).



Also a parabola! Minimized at $\theta = 25$.

Approach 2: Fit the Model: Calculus for the General Case

1. Differentiate with respect to θ_0 :

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

2. Set equal to 0.

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

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

Approach 2: Fit the Model: Calculus for the General Case (Recap from HW 1)

1. Differentiate with respect to θ_0 :

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

2. Set equal to 0.

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

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

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

Approach 2: Fit the Model: Calculus for the General Case

$$\implies \hat{\theta} = \frac{\sum_{i=1}^n y_i}{n} = \mathbf{mean}(y)$$

We're not done yet! To be thorough, we need to perform the second derivative test, to guarantee that the point we found is truly a **minimum** (rather than a maximum or saddle point). We hope that the second derivative of our objective function is positive, indicating our function is convex opening upwards.

$$\begin{aligned}\frac{d}{d\theta} R(\theta) &= \frac{-2}{n} \sum_{i=1}^n (y_i - \theta) \\ \frac{d^2}{d\theta^2} R(\theta) &= \frac{-2}{n} \sum_{i=1}^n (0 - 1) = \frac{2}{n} \sum_{i=1}^n 1 = 2\end{aligned}$$

Fortunately, it is, so the sample mean truly is the minimizer we were looking for. **We will interpret what this means shortly.**

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

This is the optimal parameter for constant model + MSE.

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

Fun fact:

The minimum MSE is the **sample variance**:

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

Note the difference:

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

The **minimum value** of
constant + MSE

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

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

The Modeling Process:

1. Choose a model



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

2. Choose a loss function



L2 Loss

Mean Squared Error (MSE)

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

3. Fit the model

Minimize
average loss
with calculus

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

4. Evaluate model performance



2771884

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 use the mean of the previous five days' sale as our prediction for tomorrow's sales:

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

Changing the Loss: Constant Model + MAE

What Is A Model?

The Modeling Process: Definitions

Loss Functions

Constant Model + MSE

Changing the Loss: Constant Model + MAE

Fit the Model: Rewrite MAE for the Constant Model

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

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

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

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

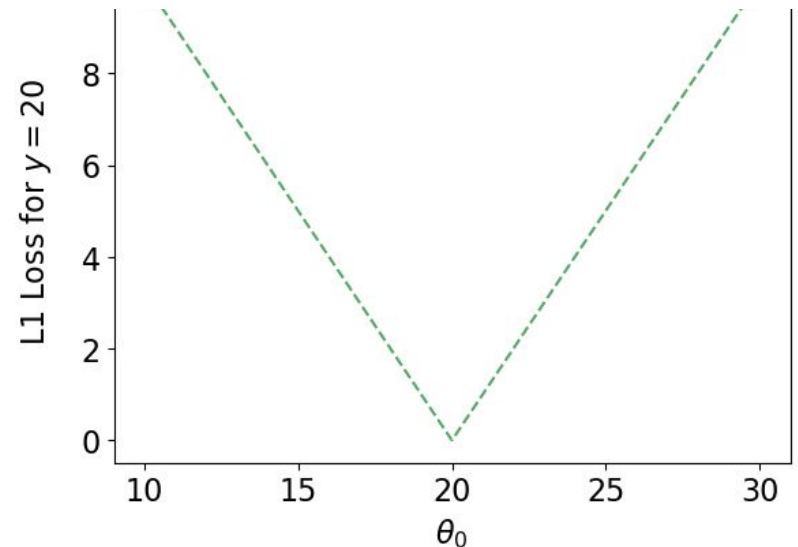
We **fit the model** by finding the optimal $\hat{\theta}_0$ that minimizes the MAE.

Exploring MAE: A Piecewise function

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

Absolute (L1) Loss on one observation:

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

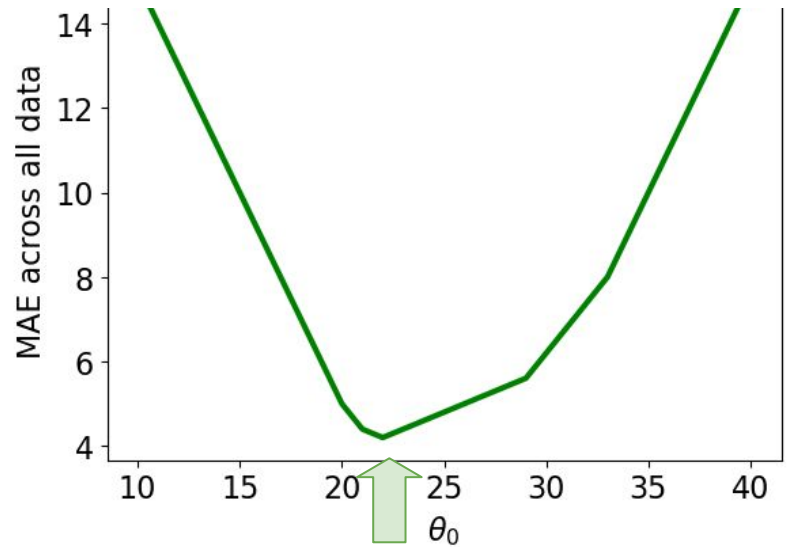


An absolute value curve, centered at $\hat{\theta}_0 = 20$.

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

MAE (Mean Absolute Error) across all data:

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



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

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

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



Absolute value!

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

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

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

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

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

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

Note: The derivative of the absolute value when the argument is 0 (i.e. when $\hat{y} = \theta_0$) is technically undefined. We ignore this case in our derivation, since thankfully, it doesn't change our result (proof left to you).



Take some time to process this math!

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

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

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

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

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

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

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

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

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

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

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

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

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

2. Set equal to 0.

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

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

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

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

Where do we go from here?

Median Minimizes MAE for the Constant Model

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

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

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

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

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

For example, in our boba 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

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

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

4. Evaluate model performance loss

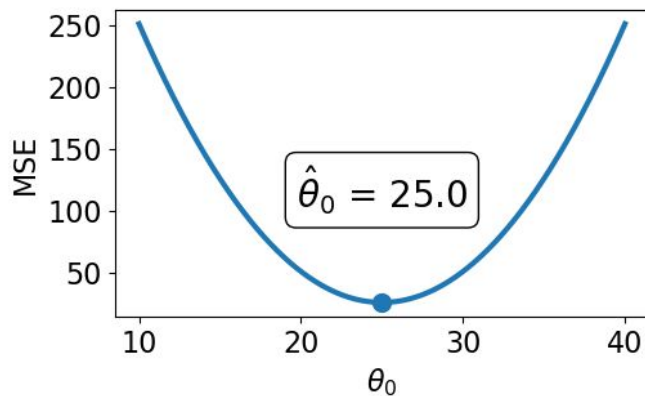
Visualize

MSE (Mean Squared Loss)

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

Minimized with **sample mean**:

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

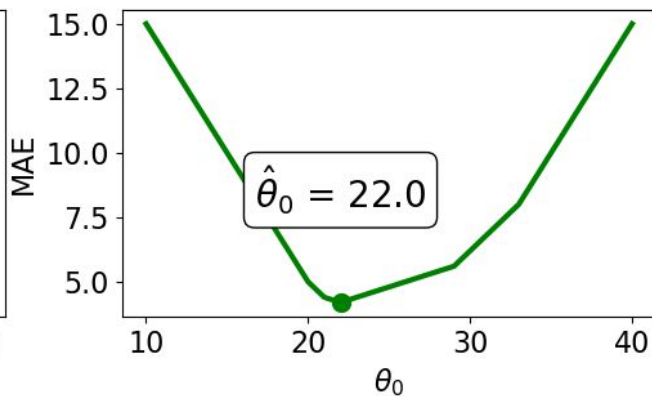


MAE (Mean Absolute Loss)

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

Minimized with **sample median**:

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

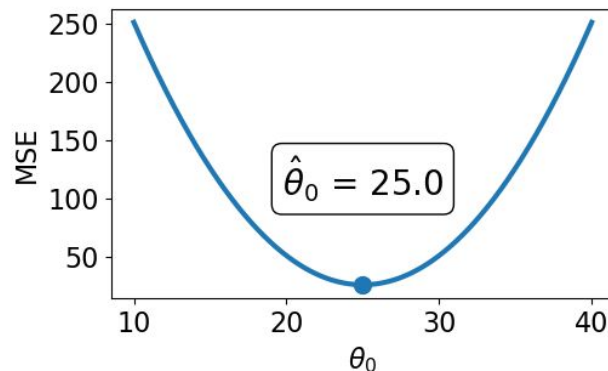


Compare

[Loss] Two Constant Models, Fit to Different Losses

MSE (Mean Squared Loss)

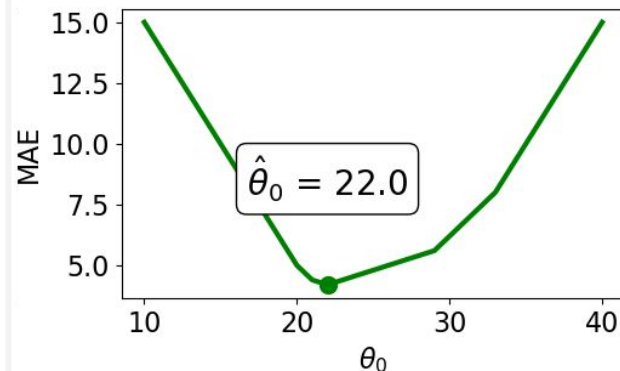
$$\hat{\theta}_0 = \text{mean}(y) = \bar{y}$$



Smooth. Easy to minimize using numerical methods

MAE (Mean Absolute Loss)

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



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

Compare

MSE and MAE: Comparing Sensitivity to Outliers

MSE (Mean Squared Loss)

Minimized with **sample mean**:

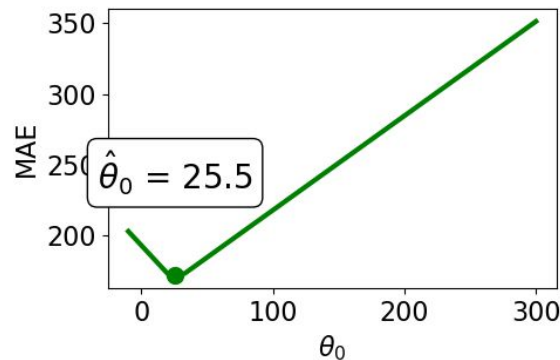
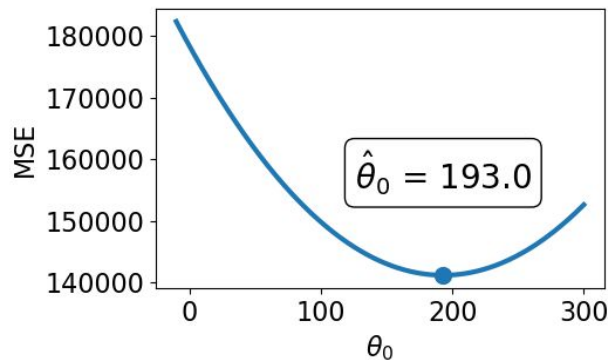
$$\hat{\theta}_0 = \text{mean}(y) = \bar{y}$$

MAE (Mean Absolute Loss)

Minimized with **sample median**:

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

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



Demo

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

More robust to outliers.

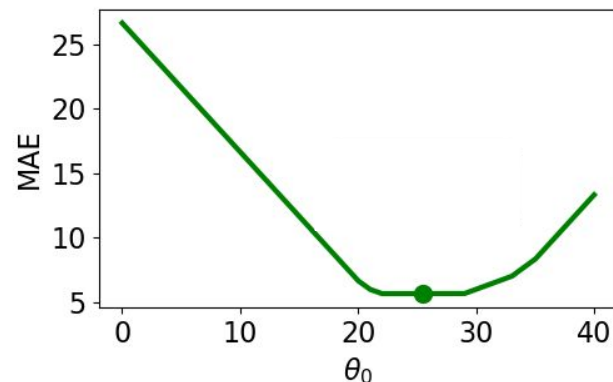
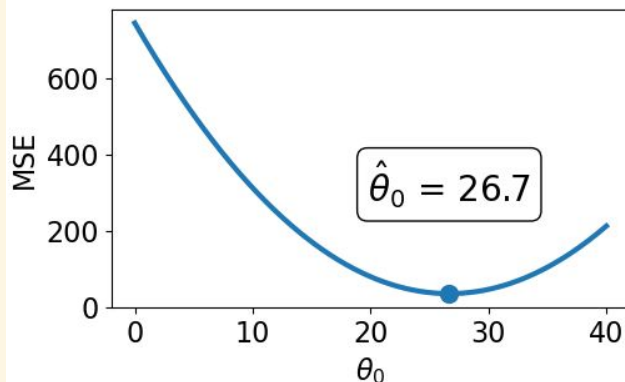
MSE and MAE: Comparing Uniqueness of Solutions

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**}



Demo

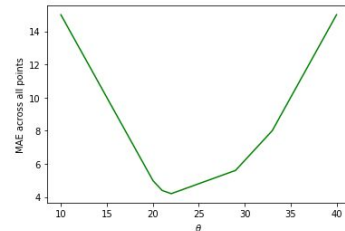
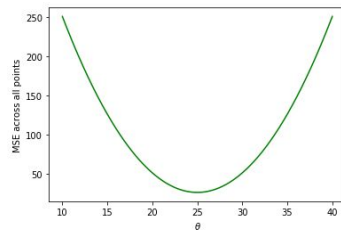
Unique $\hat{\theta}_0$:

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

! 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).

Summary: 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.
- **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.

It’s not clear that one is “better” than the other.

In practice, **we get to choose our loss function!**

Summary

Summary: 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?

In this lecture, we focused exclusively on the **constant model**, which has a single **parameter**.

Parameters define our model. They tell us the relationship between the variables involved in our model. (Not all models have parameters, though!)

In the coming lectures, we will look at more sophisticated models.

Summary: 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?

We introduced two loss functions here: L2 (**squared**) loss and L1 (**absolute**) loss. There also exist others.

Both have their benefits and drawbacks. **We get to choose** which loss function we use, for any modeling task.

Summary: 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 the **optimal parameters** by determining the parameters that **minimize average loss** across our entire dataset. **Different loss functions lead to different optimal parameters.**

This process is called **fitting the model to the data**. We did it by hand here, but in the future we will rely on computerized techniques.

Summary: 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?

Today because we had the simple constant model, we did this visually; in general there are some common performance metrics (we will see in future lectures)

- When we use squared (L2) loss as our loss function, the average loss across our dataset is called **mean squared error**.
 - “Squared loss” and “mean squared error” are not the exact same thing – one is for a single observation, and one is for an entire dataset.
 - But they are closely related.
- A similar relationship holds true between absolute (L1) loss and **mean absolute error**.
- Loss functions and summary statistics you already knew:
 - The **sample mean** is the value of θ that minimizes the **mean squared error**.
 - The **sample median** is the value of θ that minimizes the **mean absolute error**.
- “Average loss” and “empirical risk” mean the same thing for our purposes.
 - So far, our empirical risk was either mean squared error, or mean absolute error.
 - But generally, average loss / empirical risk could be the mean of any loss function across our dataset.

- **Changing the model.**
 - Next, we'll introduce the simple linear regression model.
 - We'll also look at multiple regression (and if time permits, logistic regression).
- **Changing the loss function.**
 - L2 loss (and, hence, mean squared error) will appear a lot.
 - But (if we have time) we'll also introduce new loss functions, like cross-entropy loss.
- **Changing how we fit the model to the data.**
 - We did this largely by hand in this lecture.
 - But shortly, we'll run into combinations of models and loss functions for which the optimal parameters can't be determined by hand - will require numerical approximation techniques.