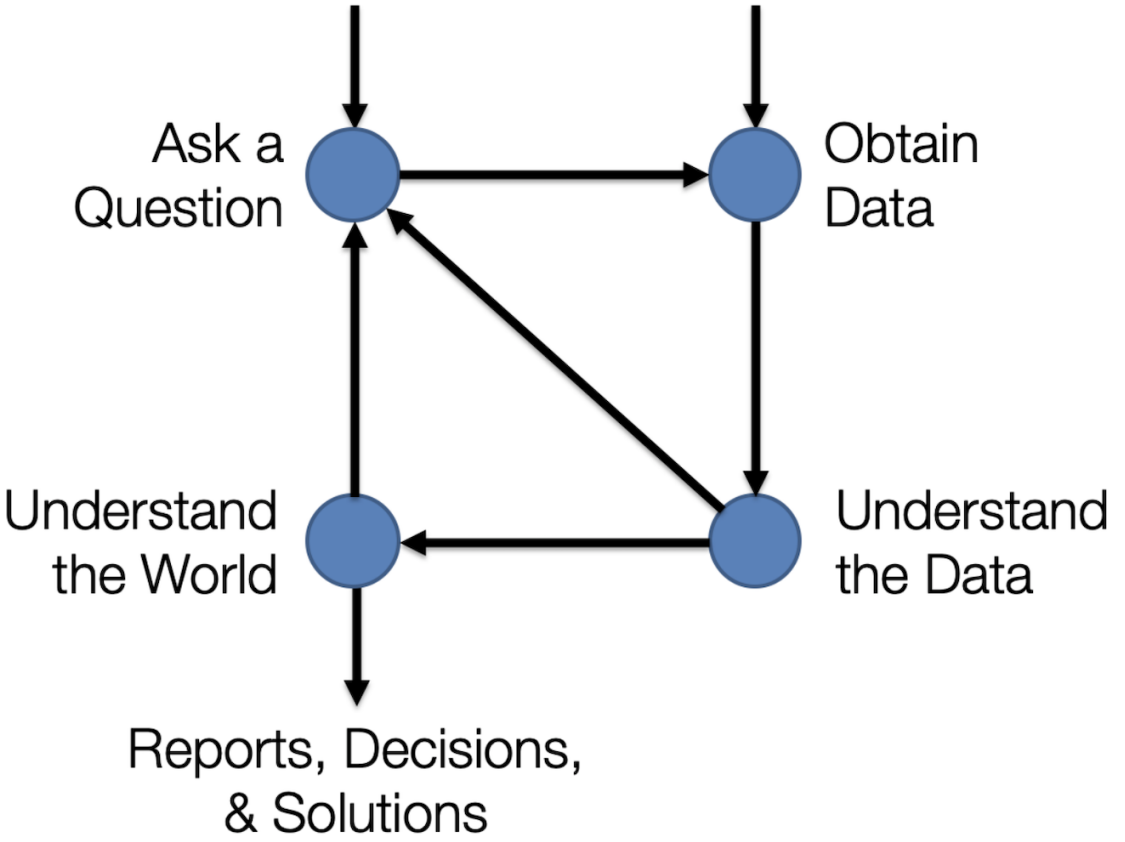
Lecture 11: Modeling

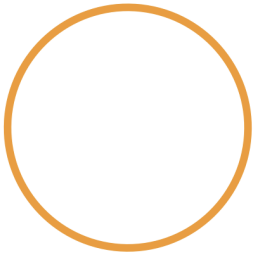
CHOOSING THE MODEL, CHOOSING THE OBJECTIVE, FITTING THE MODEL

Data 100 Spring 2021

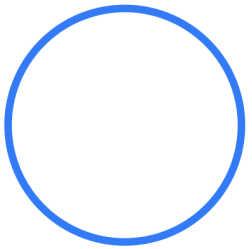
The Data Science Lifecycle



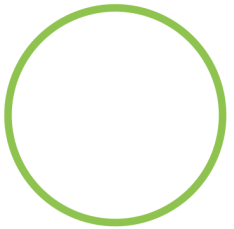
Sampling



Population: the set of all units of interest, size N.



Sampling frame: the set of all possible units that can be drawn into the sample

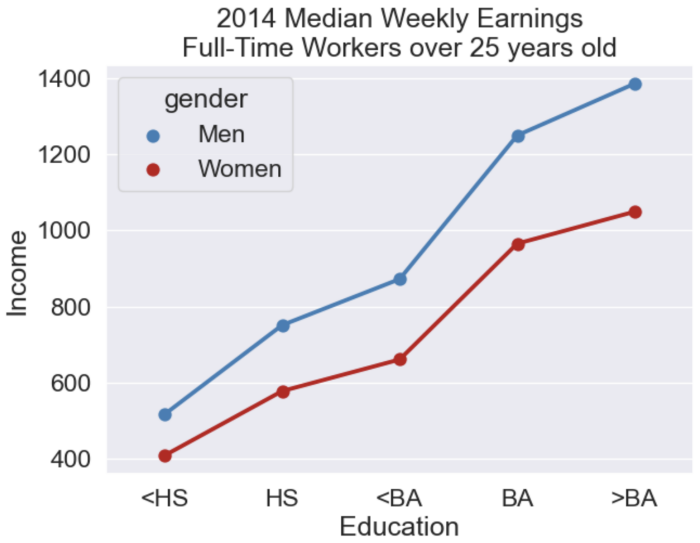


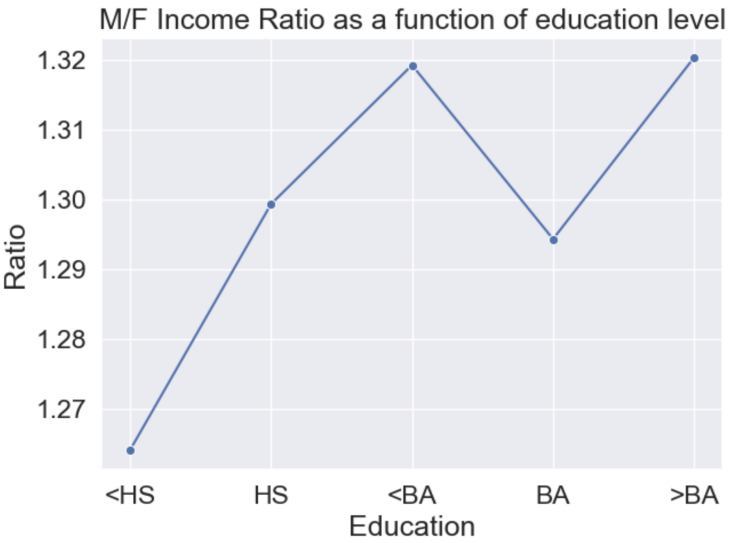
Sample: a subset of the sampling frame, size n.

Data Wrangling

- Filtering rows
- Selecting columns
- Aggregation
- Pivot Tables
- String methods
  - Regular expressions
- Joins

Data Visualization





1. Think about your Data

2. Think about your Model

What is a model?

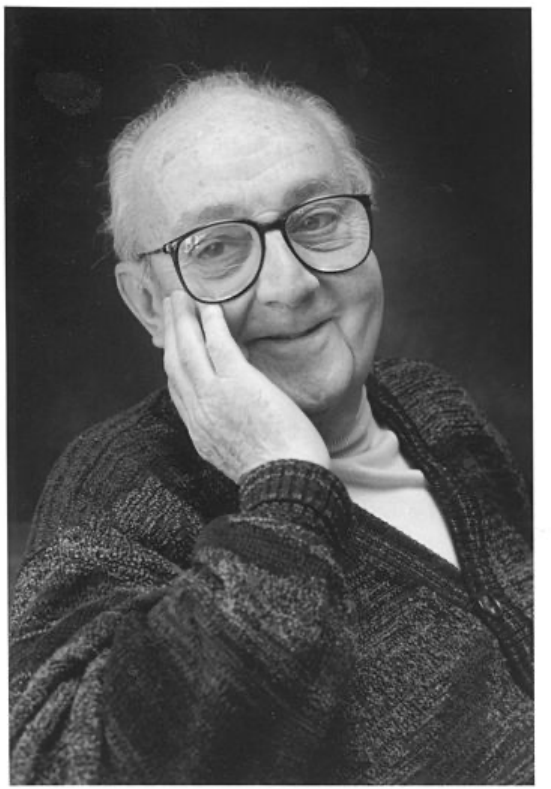
Definition: A model is a useful simplification of reality.

Example: We can model the fall of an object to earth as subject to a constant acceleration due to gravity at 9.81 m/s^2.

The model ignores:

- local variation in gravity
- air resistance
- non-linear dynamics in trajectory

> Essentially, all models are wrong, but some are useful.
>
> George Box (1919 - 2013)



How can a model be *useful*?

DESCRIPTION

To understand the world we live in.

- What factors play a role in the spread of COVID-19?
- How do an object's velocity and acceleration impact how far it travels?

$$d = d_0 + vt + \frac{1}{2}at^2$$

PREDICTION

The predict the value of unseen data.

- Is this email spam?
- Is this shape a pedestrian?

Two classes of models

PHYSICAL MODELS

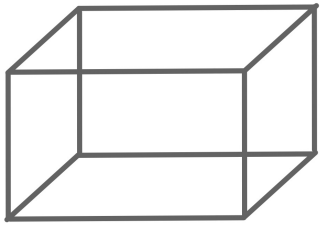
Based upon well-established theories of how the world works.

STATISTICAL MODELS

Based upon observation and data.

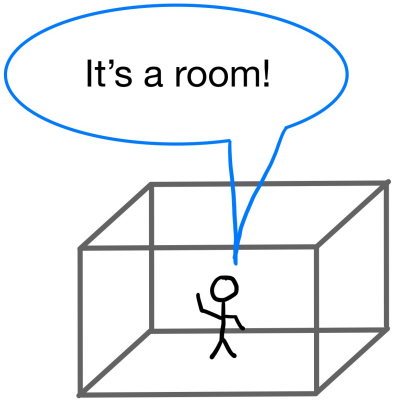
PHYSICAL MODEL

STATISTICAL MODEL



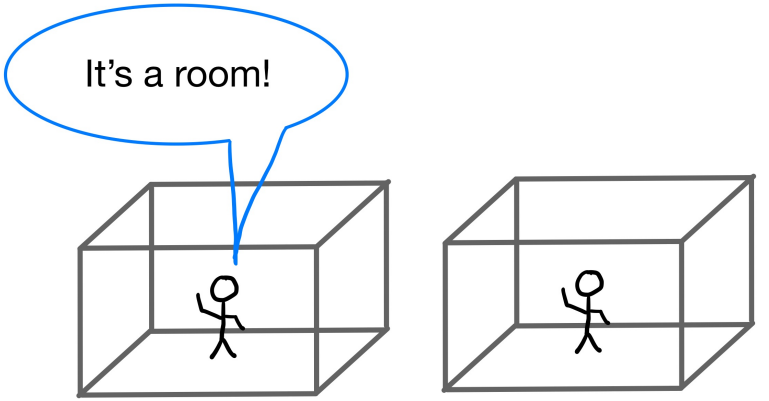
PHYSICAL MODEL

STATISTICAL MODEL



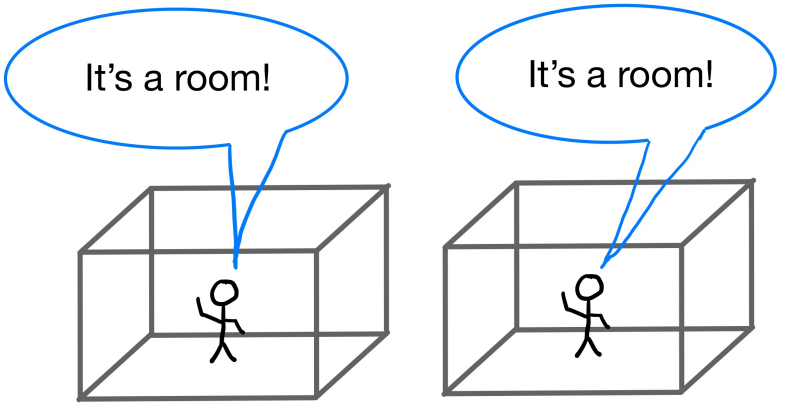
PHYSICAL MODEL

STATISTICAL MODEL



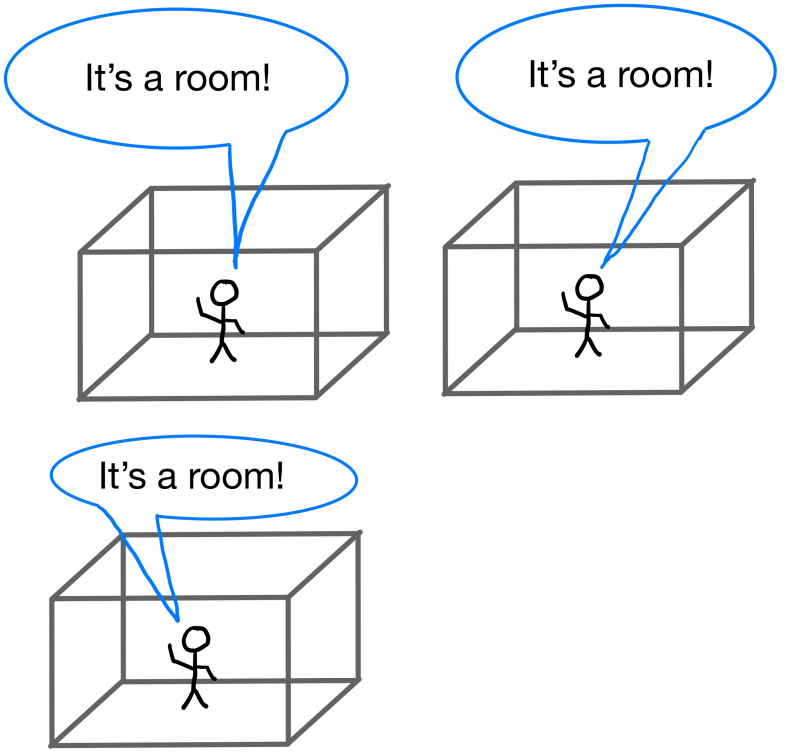
PHYSICAL MODEL

STATISTICAL MODEL



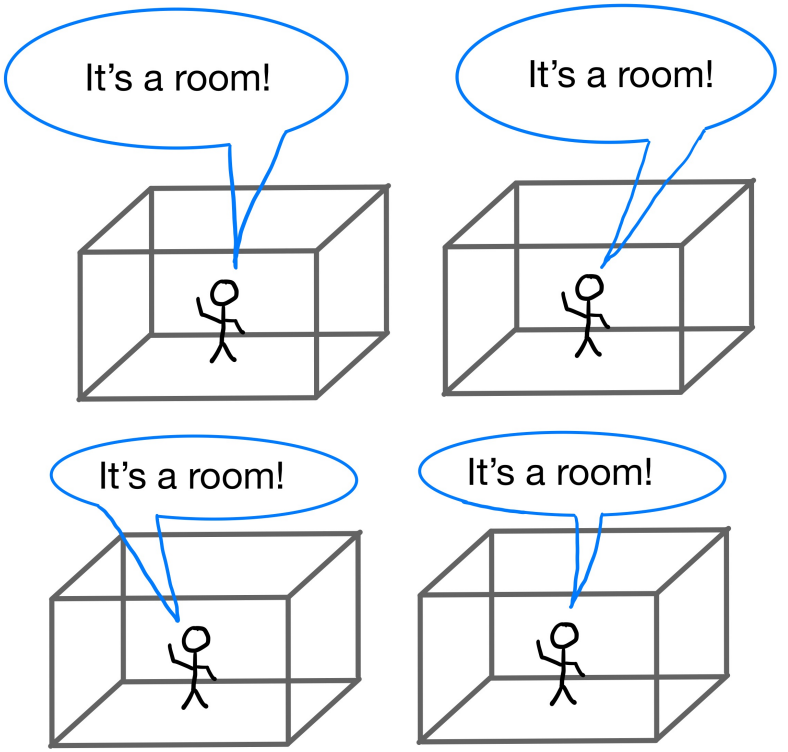
PHYSICAL MODEL

STATISTICAL MODEL

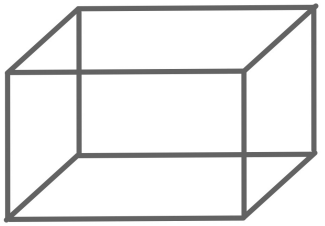


PHYSICAL MODEL

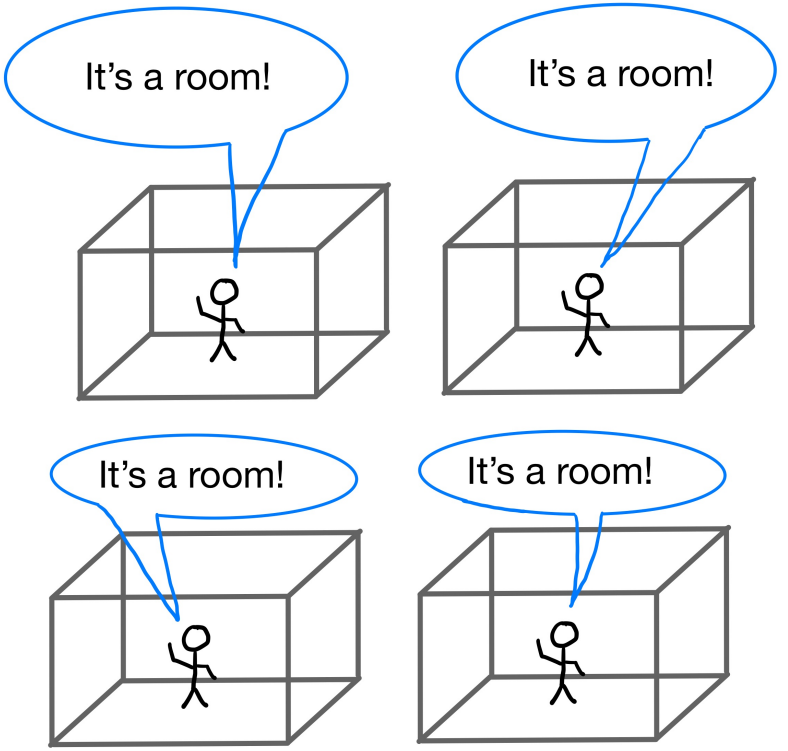
STATISTICAL MODEL



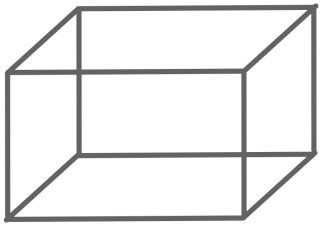
PHYSICAL MODEL



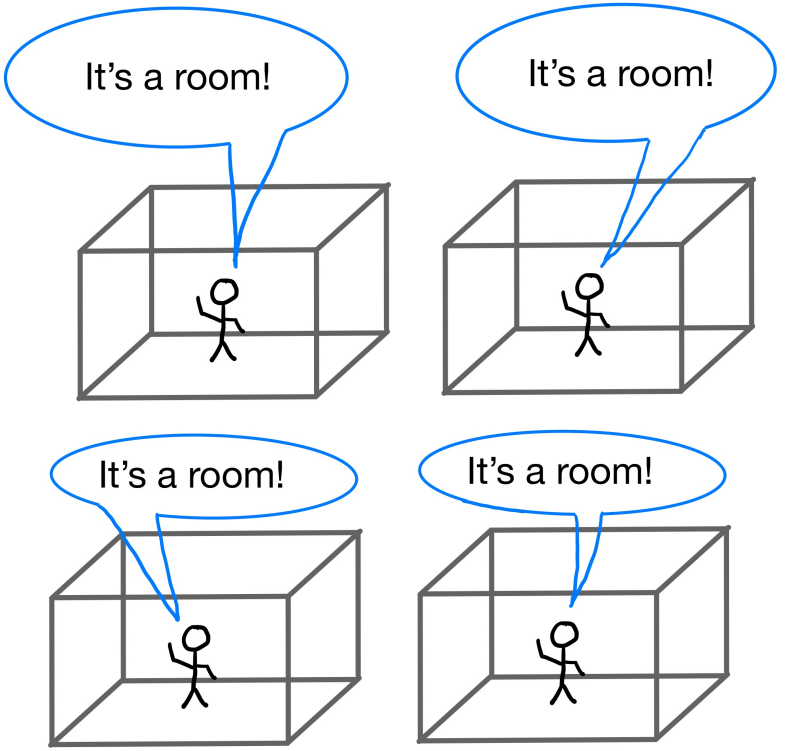
STATISTICAL MODEL



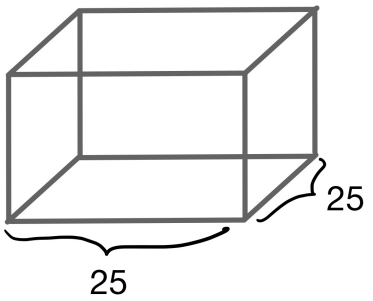
PHYSICAL MODEL



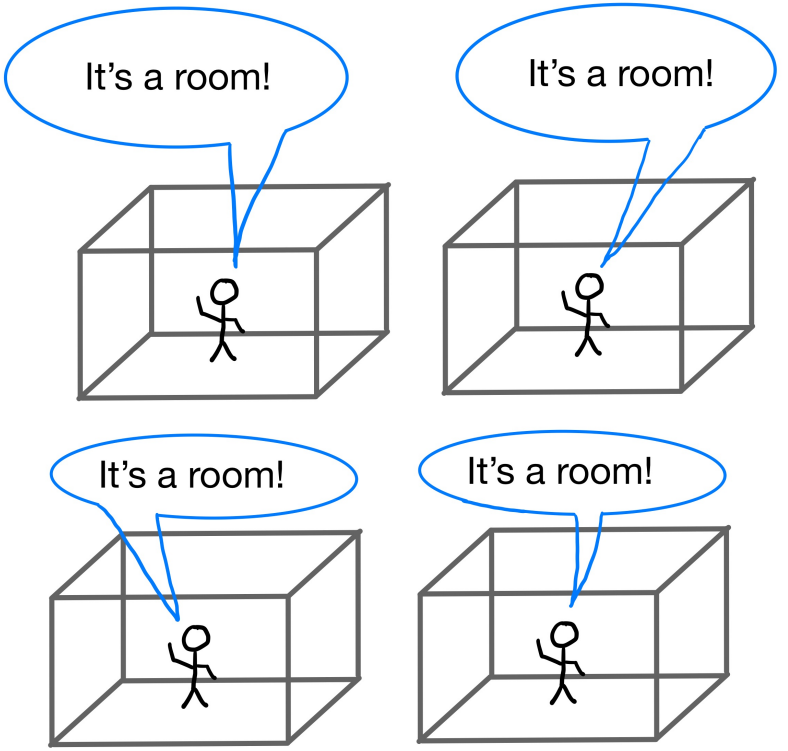
STATISTICAL MODEL



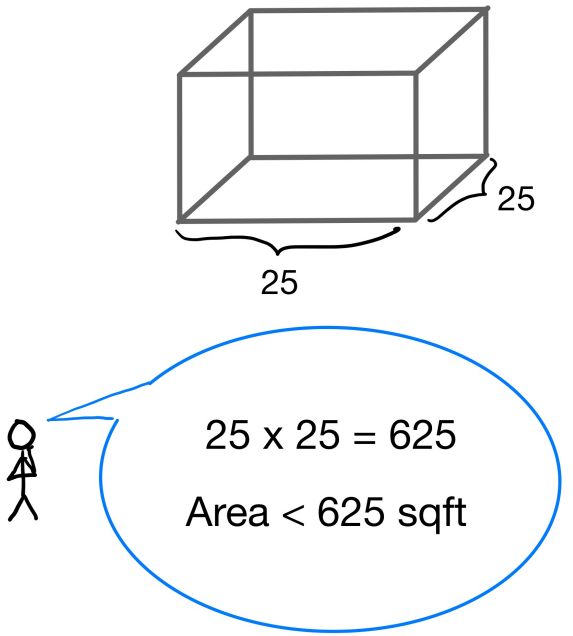
PHYSICAL MODEL



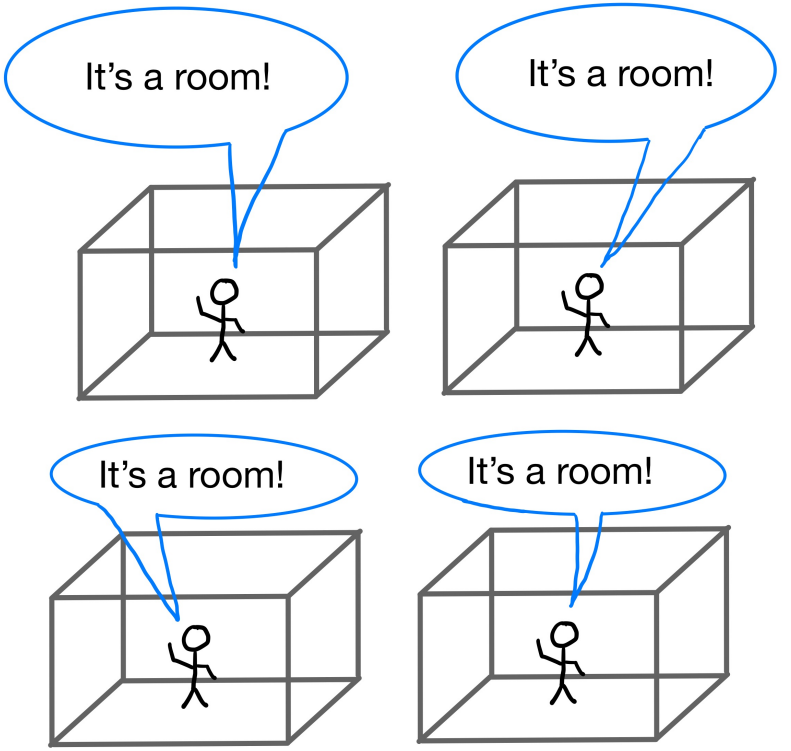
STATISTICAL MODEL



PHYSICAL MODEL

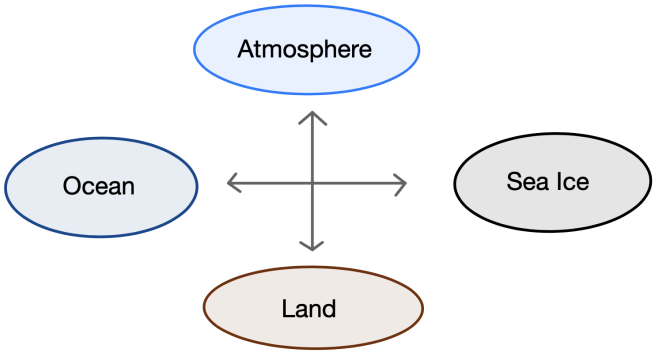


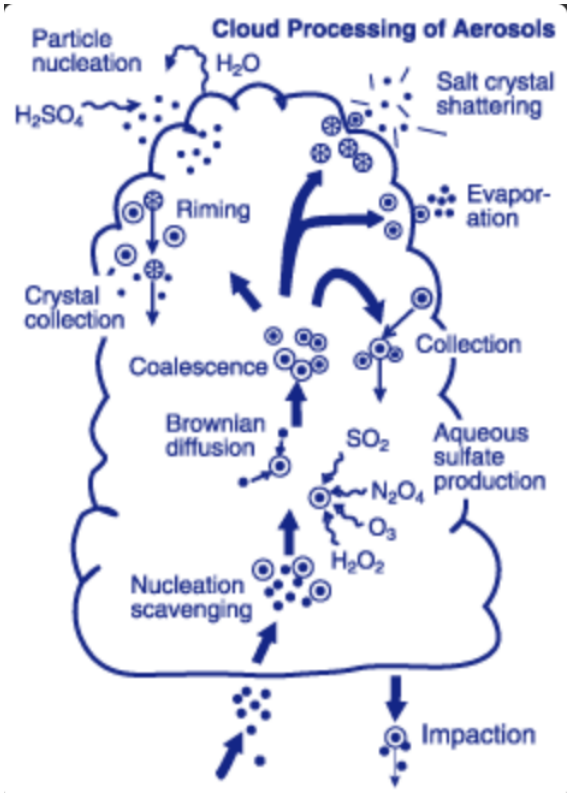
STATISTICAL MODEL



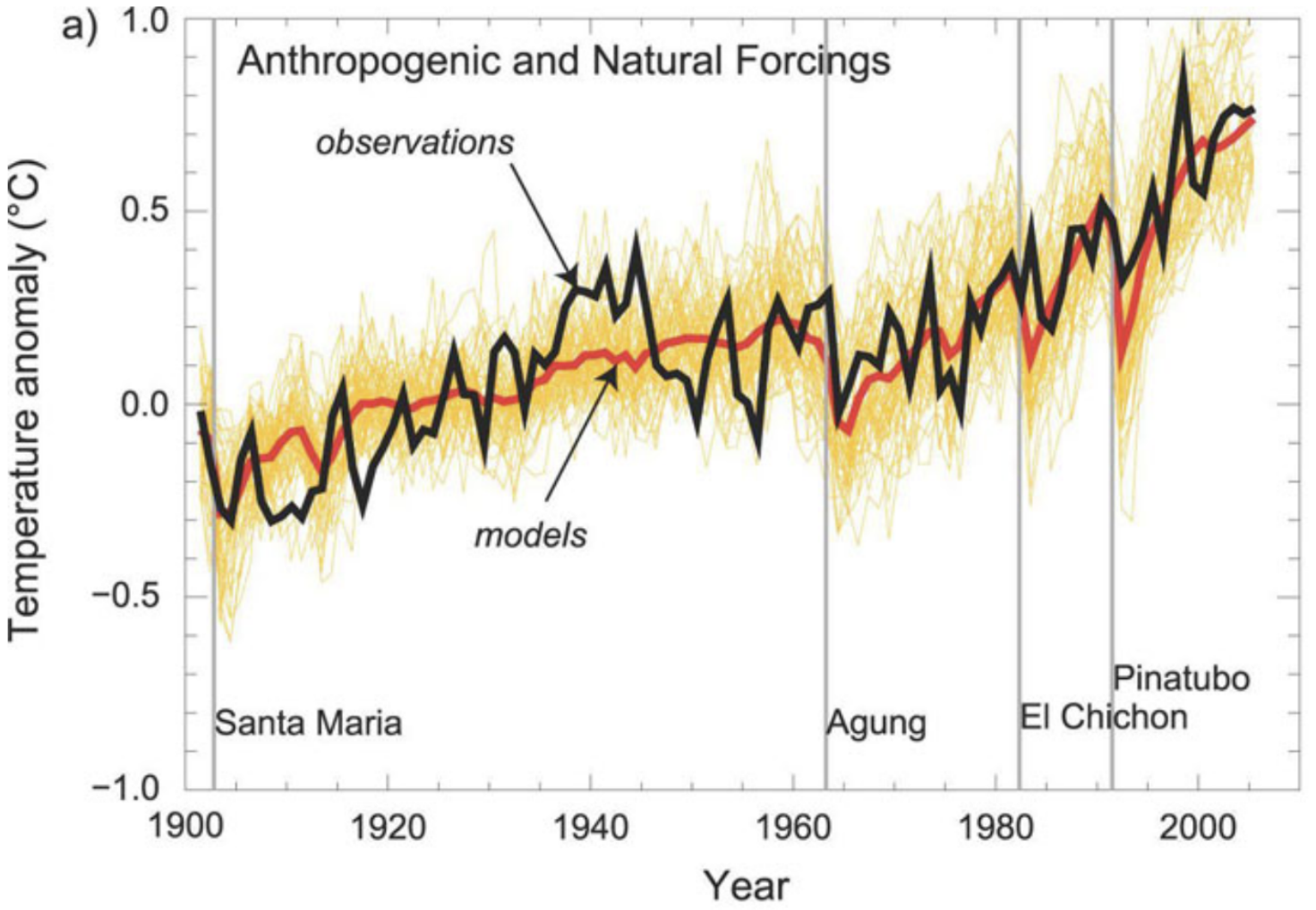
A physical model: GCM

Global Climate Model (GCM)



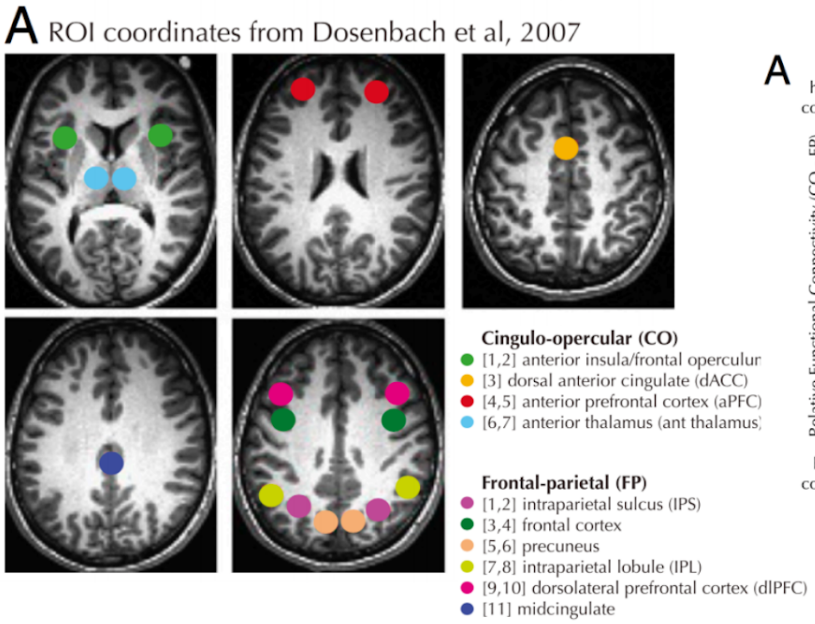


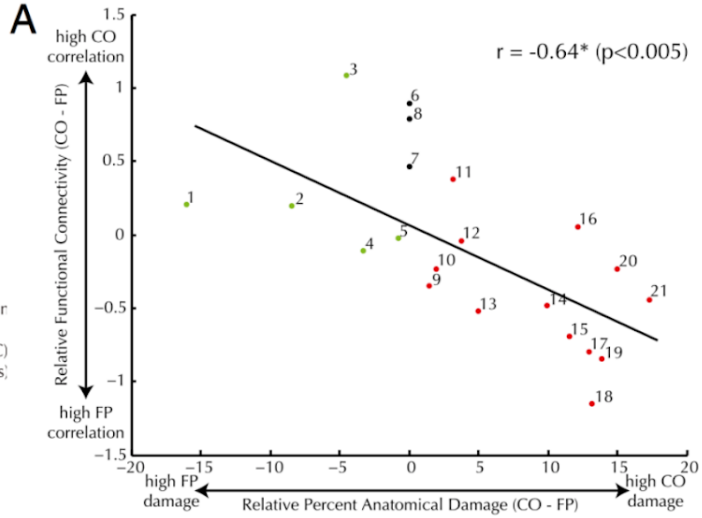
A physical model: GCM



A statistical model: FMRI

Other times, we don't have such a precise understanding of some natural relationship. In such cases, we collect data and use statistical tools to learn more about the relationships between variables.





Choosing the Model

The modeling process: 3 steps

I. CHOOSE A MODEL

II. CHOOSE AN OBJECTIVE FUNCTION

III. FIT THE MODEL BY OPTIMIZING YOUR OBJECTIVE FUNCTION

Choose a Model

What is the simplest model?

> A constant.

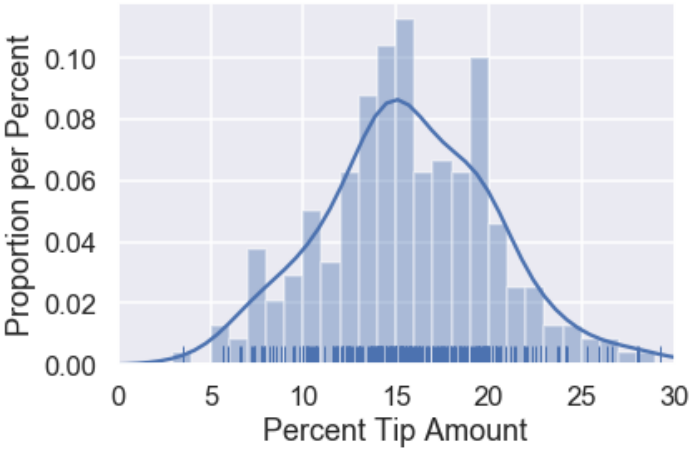
A **constant model** predicts the same number regardless of the circumstances, ignoring all other information.

Example: Tips are 15%

Useful? Description and prediction

Simple? Ignores bill price, time of day, customer info

The Tips Dataset



Tip rate at a restaurant across 244 bills.

Which constant best models these tips?

- 15% seems better than 25%
- Is 15% better than 14%
- We need a more precise formulation of this process.

Notation

y **observations** (data on tip %)

- y_i individual observation
- $y_1, y_2, \ldots, y_n$ data set of size n

$\hat{y}$ **predicted observations** (predicted tip %)

- $\hat{y}_i$ individual prediction

θ **model parameter(s)** ("true" constant tip %)

$\hat{\theta}$ **fitted, or optimal, parameter(s)** (est. constant tip %)

Notation

The constant model can be stated as:

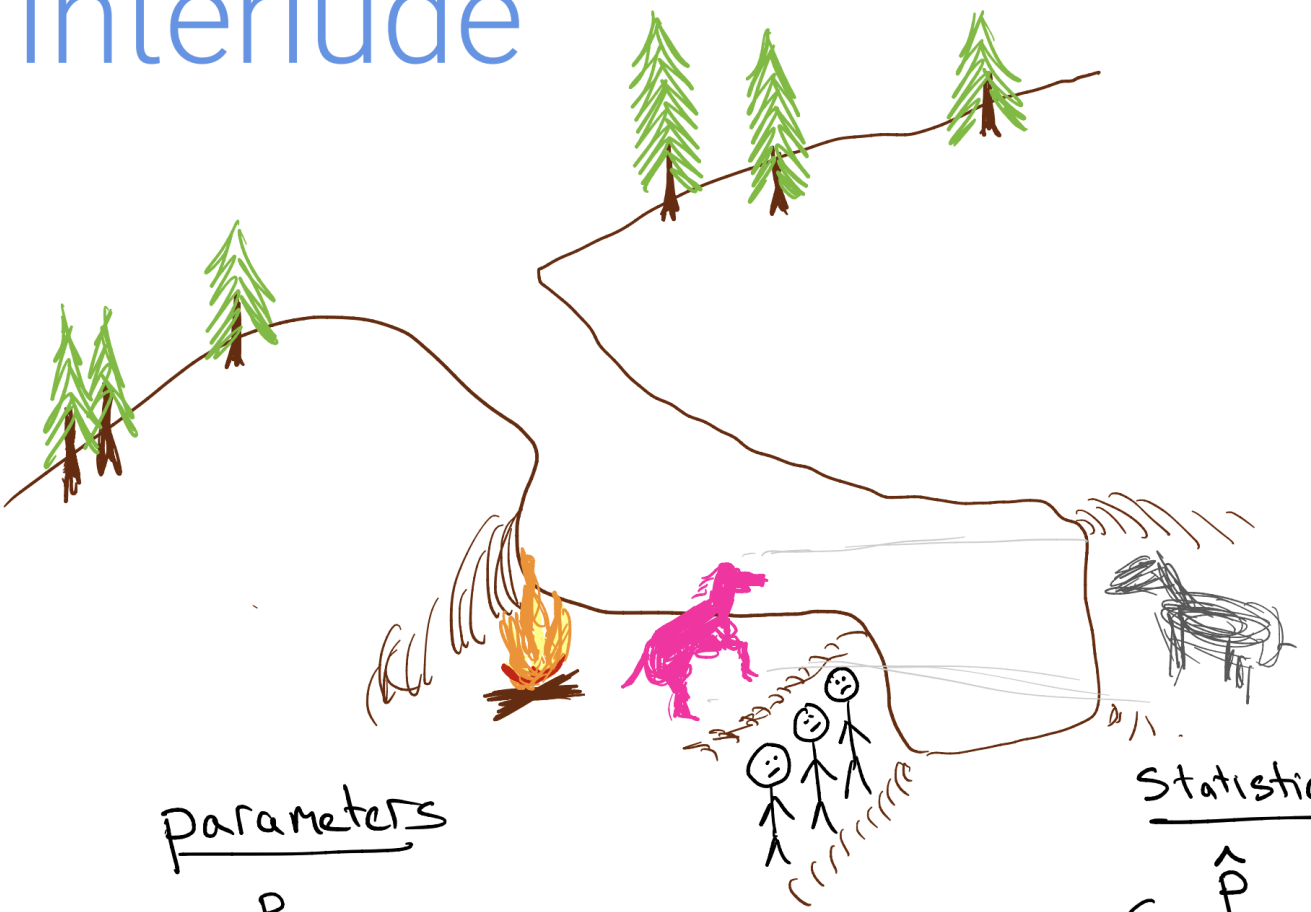
$$\hat{y} = \theta$$

- Parameters define the model
  - Some models are *nonparametric* (e.g. KDEs)
- A constant model ignores any input
- Models can have many parameters

$$\hat{y} = \theta_0 + \theta_1 x \qquad \hat{y} = \frac{1}{1 + \exp(-x^T \theta)}$$

- Goal: find "best" value of the parameter, denoted $\hat{\theta}$

Interlude


parameters
θ
p
μ
σ
β
Statistics
$\hat{\theta}$
$\hat{p}$
$\hat{\mu}$, $\bar{x}$
$\hat{\sigma}$, s
$\hat{\beta}$,

Estimation

Using data to determine model parameters

$$\hat{\theta} = f_1(y, x)$$

Prediction

Using the fitted model parameters to predict outputs for unseen data

$$\hat{y}_i = f_2(\hat{\theta}, x_i)$$

The modeling process: 3 steps

I. CHOOSE A MODEL

- Constant model
- Linear model
- Non-linear model

II. CHOOSE AN OBJECTIVE FUNCTION

- Prediction: Loss function
- Description: e.g. Likelihood function

III. FIT THE MODEL BY OPTIMIZING YOUR OBJECTIVE FUNCTION

- Analytical approach (calculus, algebra)
- Numerical approach (optimization, gradient descent)

Loss functions

The cost of doing business (making predictions)

We need some metric of how “good” or “bad” our predictions are. This is what loss functions provide us with. **Loss functions quantify how bad a prediction is for a single observation**.

- If our prediction is close to the actual value, we want low loss.
- If our prediction is far from the actual value, we want high loss.

A natural choice of loss function is **actual - predicted**, or $y_i - \hat{y}_i$. We call this the **error** for a single prediction.

- But, this treats “negative” predictions and “positive” predictions differently.
  - Predicting 16 when the true value is 15 should be penalized the same as predicting 14.
- This leads to two natural loss functions.

Squared and absolute loss

The most common loss function you'll see is the **squared loss**, also known as L2 loss.

$$L_2(y, \hat{y}) = (y - \hat{y})^2$$

- For a single data point in general, this is $(y_i - \hat{y}_i)^2$.
- For our constant model, since $\hat{y} = \theta$, this is $(y_i - \theta)^2$.

If our prediction is equal to the actual observation, in both cases, our **loss is 0**.

Low loss means a good fit!

Another common loss function is the **absolute loss**, also known as L1 loss.

$$L_1(y, \hat{y}) = |y - \hat{y}|$$

- For our constant model, for a single point, this is $|y_i - \theta|$.

There are benefits and drawbacks to both of the above loss functions. We will examine those shortly. **These are also not the only possible loss functions; we will see more later.**

Loss functions and empirical risk

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 across all points**. Assuming n points:

$$\frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{y}_i)$$

Other names for **average loss** include **empirical risk** and an **objective function**.

The average loss of a model tells us how well it fits the given data. If our model has a low average loss across our dataset, that means it is good at making predictions. As such, we want to **find the parameter(s) that minimize average loss**, in order to make our model as good at making predictions as it can be.

MSE and MAE

If we choose squared loss as our loss function, then average squared loss is typically referred to as **mean squared error (MSE)**, and is of the following form:

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

If we choose absolute loss as our loss function, then average absolute loss is typically referred to as **mean absolute error (MAE)**, and is of the following form:

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

These definitions hold true, regardless of our model. We want to **minimize** these quantities.

Exploring MSE

Average loss is typically written as a function of θ, since θ defines what our model is (and hence what our predictions are). For example, with squared loss and the constant model, our average loss (and hence, the function we want to minimize) is

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

Average loss is also a function of our data. But unlike theta, we can't change our data: it is given to us (i.e. it is fixed).

Mathematically, our goal of finding the optimal $\hat{\theta}$ is stated as:

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

argmin means "the **arg**ument that **min**imizes the following function."

We won't use this notation again in this lecture, but it will come up again in the future.

Exploring MSE

When our model is the constant model, and we choose to use L2 loss, again, our average loss looks like:

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

Let's examine a toy example. Suppose we have 5 observations, **[20, 21, 22, 29, 33]**.

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

The loss for the first observation (y_1).

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

The average loss across all observations (the MSE).

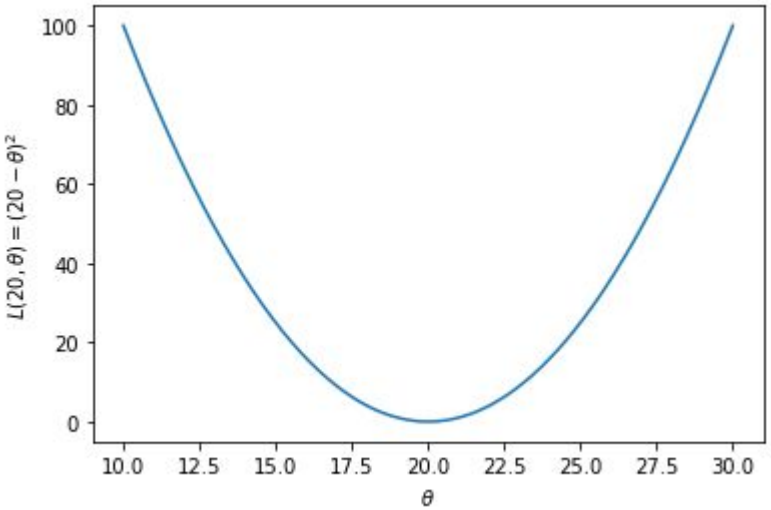
Exploring MSE

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

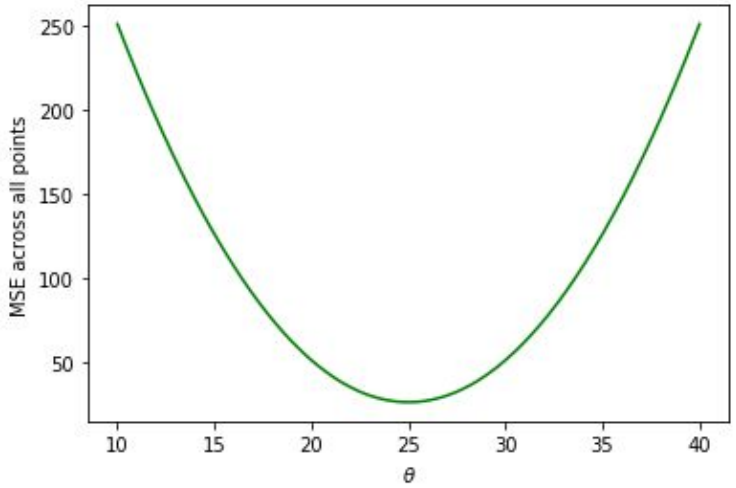
The loss for the first observation (y_1).

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

The average loss across all observations (the MSE).



A parabola, minimized at theta = 20.



Also a parabola! Minimized at theta = 25.

Minimizing mean squared error (MSE)

for the constant model

Minimizing MSE

We saw with the toy example of [20, 21, 22, 29, 33] that the value that minimizes the MSE of the constant model was 25, which was the **mean of our observations**.

We can try other examples if we want to, and we'll end up with the same result. Let's instead pivot to proving this rigorously, using mathematics. There are two ways we'll go about doing this:

- Using calculus.
- Using a neat algebraic trick.

For both derivations, the slides contain the key ideas, but the lecture videos will contain a step-by-step walkthrough.

MSE minimization using calculus

One way to minimize a function is by using calculus: we can take the derivative, set it equal to 0, and solve for the optimizing value.

- The derivative of the sum of several pieces is equal to the sum of the derivative of said pieces.
- The derivative of the loss for a single point is $\frac{d}{d\theta}(y_i - \theta)^2 = 2(y_i - \theta)(-1) = -2(y_i - \theta)$.

Then:

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

since we can pull constants out of sums

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

from above

MSE minimization using calculus

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

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

Setting this term to 0, we have:

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

we can separate sums

$$0 = \sum_{i=1}^{n}(y_i - \theta) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\theta = \sum_{i=1}^{n} y_i - n\theta$$

c + c + ... + c = n * c

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

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

Thus, with squared loss and the constant model, the sample mean minimizes MSE.

MSE minimization using calculus

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

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

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

$$\frac{d^2}{d\theta^2} R(\theta) = \frac{-2}{n} \sum_{i=1}^{n} (0 - 1) = \frac{2}{n} \sum_{i=1}^{n} 1 = 2$$

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

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. The lecture video will walk through it in detail.
- 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 when $\theta = \bar{y}$. In other words:

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

Looks familiar!

Question: What is the value of average loss, when evaluated at $\theta = \hat{\theta}$?

Minimum value of MSE is the sample variance

It's worth noting that when we substitute $\theta = \bar{y}$ back into our average loss, we obtain a familiar result:

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

That is, the **minimum value** that mean squared error can take on (again, for the constant model) **is the sample variance**.

Put another way, the following statement is true whenever $\theta \neq \bar{y}$:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 < \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2$$

Mean minimizes MSE for the constant model

As we determined a variety of ways, for the constant model with squared loss, the mean of the dataset is the optimal model.

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

- This holds true **regardless of the dataset** we use, but it's only true for **this combination of model and loss**.
- If we choose any other constant other than the sample mean, the empirical risk will not be as small as possible, and so our model is "worse" (for this loss).

This is not all that surprising! **It provides some formal reasoning as to why we use means so commonly as summary statistics.** It is the best, in some sense.

Note, we now write $\hat{\theta}$ instead of θ . This is because we are referring to the **optimal parameter**, not just any arbitrary θ .

Minimizing mean absolute error (MAE)

for the constant model

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]**.

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

The loss for the first observation (y_1).

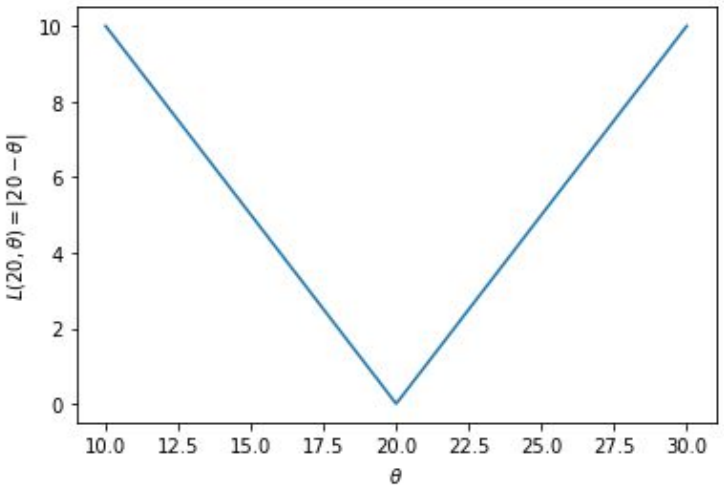
$$R(\theta) = \frac{1}{5}\left(|20 - \theta| + |21 - \theta| + |22 - \theta| + |29 - \theta| + |33 - \theta|\right)$$

The average loss across all observations (the MAE).

Exploring MAE

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

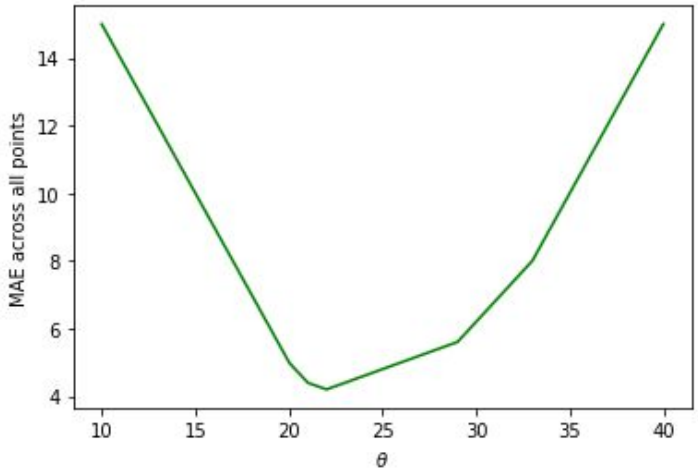
The loss for the first observation (y_1).



An absolute value curve, centered at theta = 20.

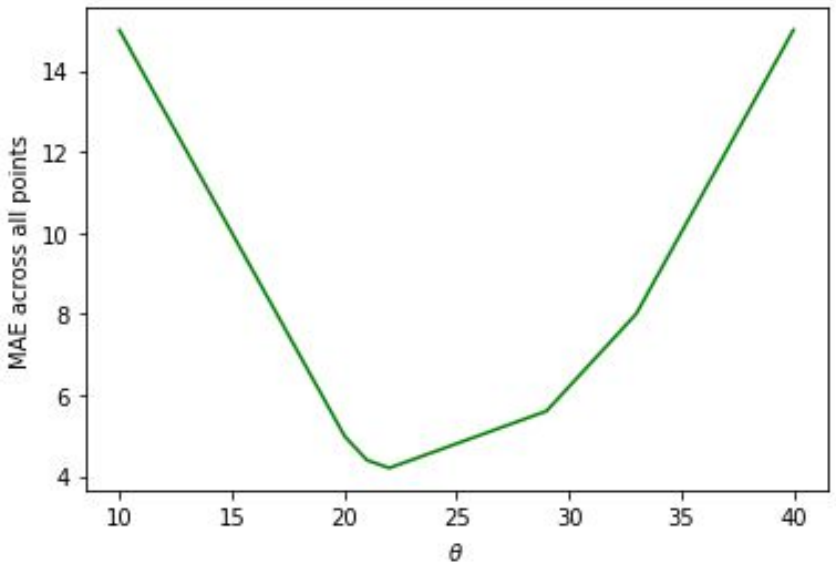
$$R(\theta) = \frac{1}{5}\left(|20 - \theta| + |21 - \theta| + |22 - \theta| + |29 - \theta| + |33 - \theta|\right)$$

The average loss across all observations (the MAE).



Some weird shape.... minimized near theta = 22?

Exploring MAE

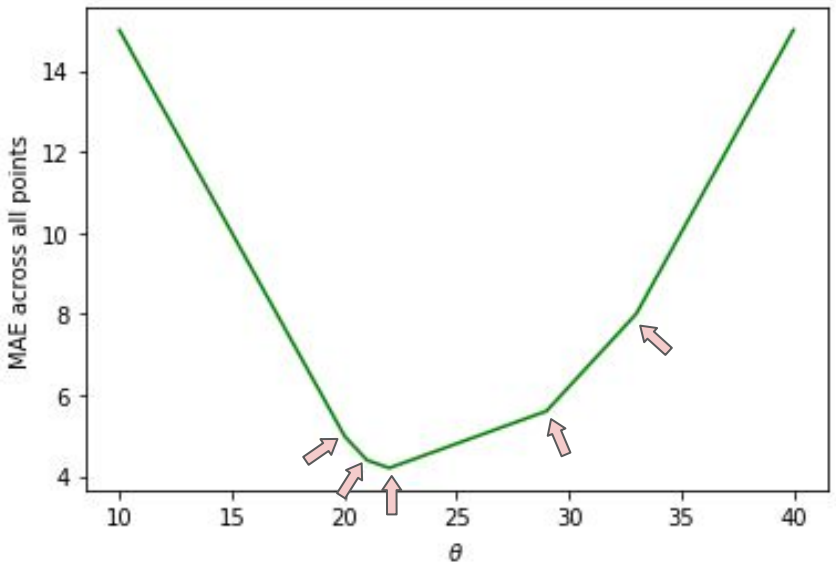


The shape of the MAE with the constant model seems to be jagged. This is because it is the (weighted) sum of several absolute value curves, which results in a **piecewise linear function**.

It also doesn't seem to be immediately clear what the optimal choice of theta should be. It's somewhere in the "middle" of our points, but it's **clearly not 25**, which was the minimizing value for the MSE.

Let's once again resort to calculus!

Exploring MAE



The bends, or "kinks," all appear at our observations! (20, 21, 22, 29, 33)

The shape of the MAE with the constant model seems to be jagged. This is because it is the (weighted) sum of several absolute value curves, which results in a **piecewise linear function**.

It also doesn't seem to be immediately clear what the optimal choice of theta should be. It's somewhere in the "middle" of our points, but it's **clearly not 25**, which was the minimizing value for the MSE.

Let's once again resort to calculus!

MAE minimization using calculus

Once again, we can use calculus to determine the optima $\hat{\theta}$.

The first step is to determine the derivative of our loss function for a single point. Absolute value functions can be written as two piecewise linear functions:

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

The derivative of our loss for a single point, then, is also a piecewise linear function:

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

Note: The derivative of the absolute value when the argument is 0 (i.e. when $y_i = \theta$) is technically undefined. We ignore this case in our derivation, since thankfully, it doesn't change our result.

MAE minimization using calculus

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

From here, we again use the fact that the derivative of a sum is a sum of derivatives:

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

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

Add -1 for each time an observation y_i is greater than our choice of theta.

Add 1 for each time an observation y_i is less than our choice of theta.

MAE minimization using calculus

Setting this derivative equal to 0:

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

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

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

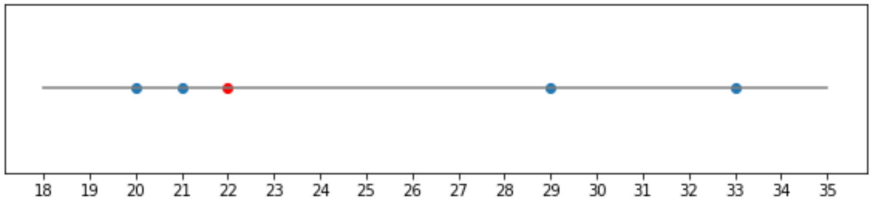
The last line is telling us that in order for our MAE to be minimized, we need to choose a theta such that **the number of observations less than theta** needs to be equal to **the number of observations greater than theta.**

MAE minimization using calculus

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

In order for our MAE to be minimized, we need to choose a theta such that the number of observations less than theta needs to be equal to the number of observations greater than theta. In other words, theta needs to be such that there are **an equal number of points to the left and right.**

This is the definition of the median! For example, in our toy dataset, the point below in red (22) is the median of our observations. It is the value in the "middle."

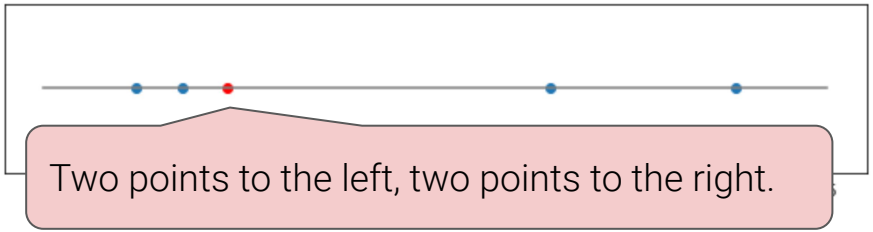


MAE minimization using calculus

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

In order for our MAE to be minimized, we need to choose a theta such that the number of observations less than theta needs to be equal to the number of observations greater than theta. In other words, theta needs to be such that there are **an equal number of points to the left and right.**

This is the definition of the median! For example, in our toy dataset, the point below in red (22) is the median of our observations. It is the value in the "middle."

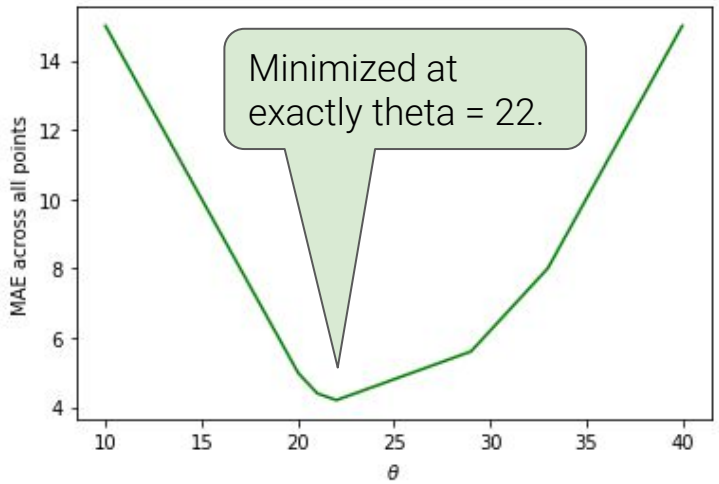


Median minimizes MAE for the constant model

We've now shown that the median minimizes MAE for the constant model.

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

This is consistent with what we saw earlier, when plotting the MAE for our toy dataset:



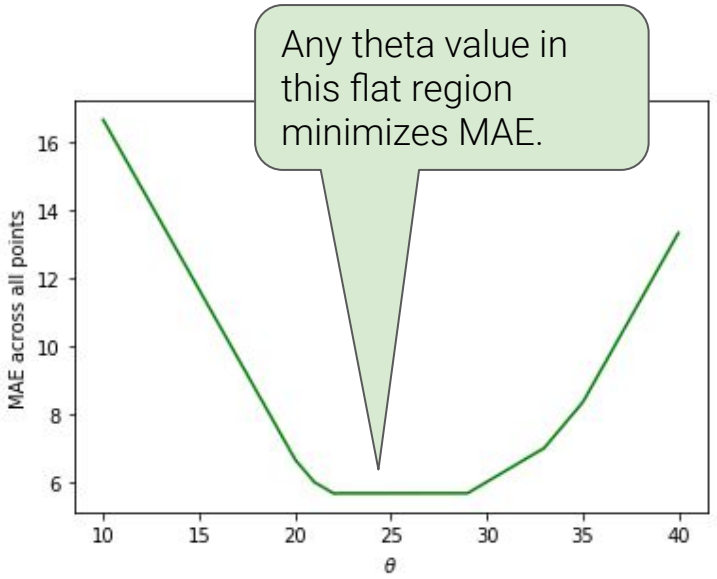
Important note: In general, the mean and median of a dataset are not the same. Therefore, using MSE and MAE give us **different optimal theta values**!

A key takeaway here is that our choice of loss function determines the optimal parameters for our model.

Median minimizes MAE for the constant model

Our toy dataset only had 5 observations. What if it had an even number of observations? Let's say our toy dataset is now **[20, 21, 22, 29, 33, 35]**. The 35 is new.

- **There's no longer a unique solution!**
- Any value in the range [22, 29] minimizes MAE.
- This reflects the fact that there are an even number of observations, and any number in that range has the same number of points to the left and right.
- (When there are an even number of data points, we typically set the median to be the mean of the two middle ones. Here, that'd be 25.5.)

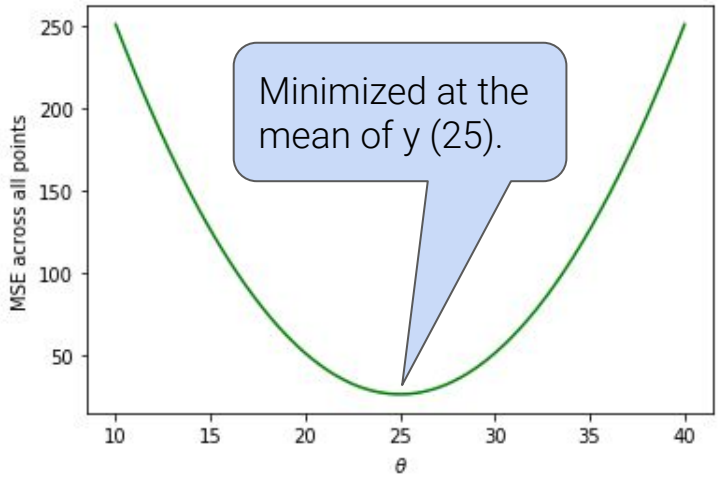


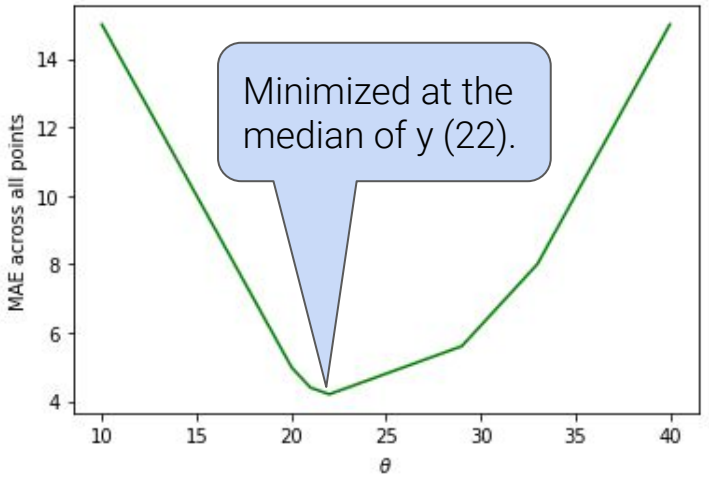
Comparing loss functions

MSE vs. MAE for toy data

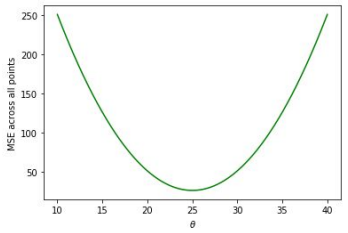
Below, we present the plot of the **loss surface** for our toy dataset, using L2 loss (left) and L1 loss (right).

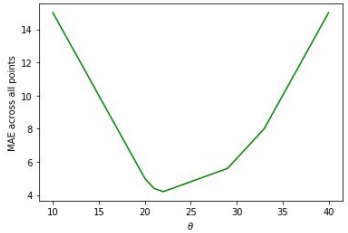
- A **loss surface** is a plot of the loss encountered for each possible value of θ .
- If our model had 2 parameters, this plot would be 3 dimensional.





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.

It's not clear that one is "better" than the other.

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

Summary

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

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

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.

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

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.

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

Lastly, we 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.

Vocabulary review

- 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.

What's next...

- **Changing the model.**
  - Next, we'll introduce the simple linear regression model that you saw in Data 8.
  - We'll also look at multiple regression, logistic regression, decision trees, and random forests, all of which are different types of models.
- **Changing the loss function.**
  - L2 loss (and, hence, mean squared error) will appear a lot.
  - But 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.
  - As such, we'll learn about techniques like gradient descent.