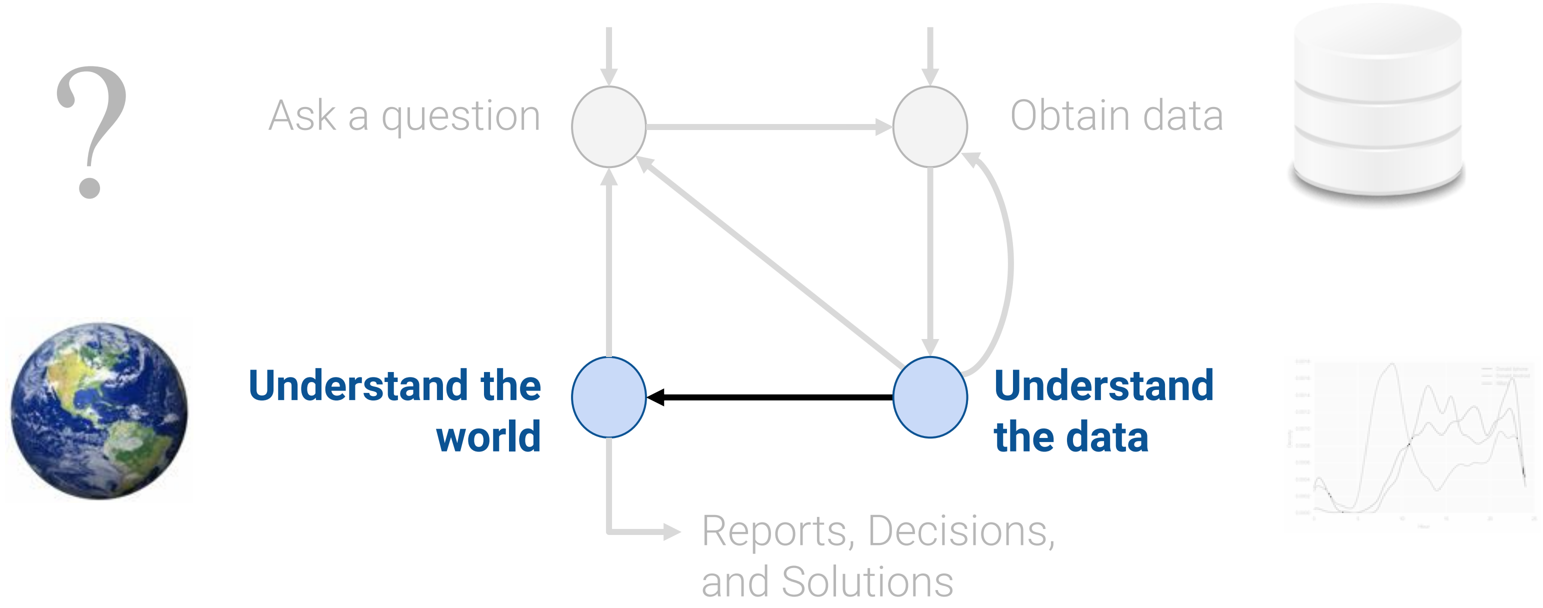
LECTURE 9

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

What is a Model?

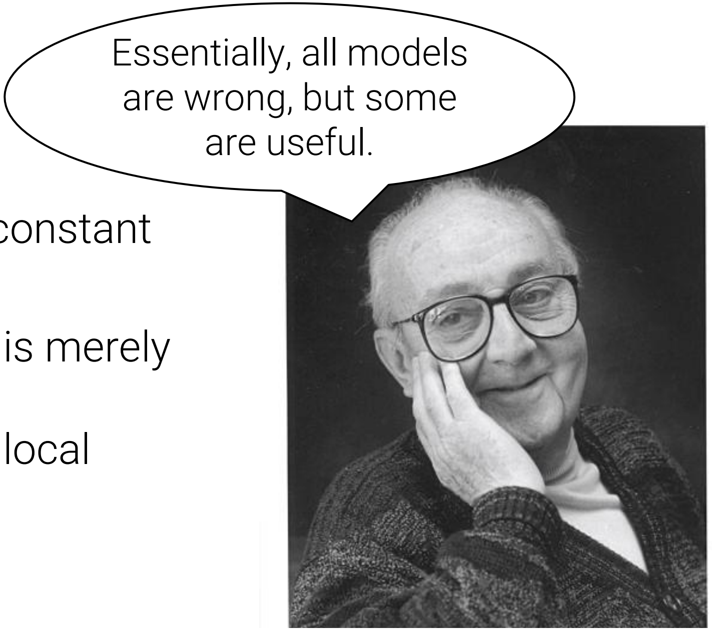
- **What is a Model?**
- Data 8 Review
  - Regression Line, Correlation
- The Modeling Process
  - Choose a Model
  - Choose a Loss Function
  - Fit the Model
  - Evaluate the Model

What is a Model?

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

<u>Example</u>:
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)

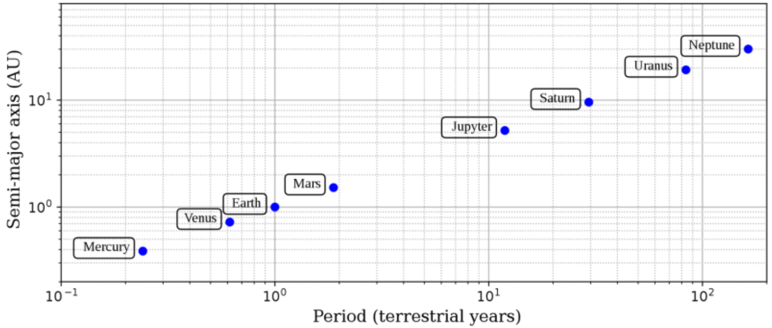
Most of the time, we want to strike a balance between **interpretability** and **accuracy**.

Common Types of Models

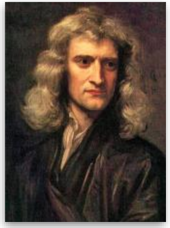
Deterministic physical (mechanistic) models: Laws that govern how the world works.

[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.

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

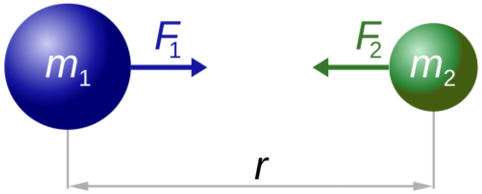


[Newton's Laws: motion and gravitation]() (1687)
Newton's second law of motion models the relationship between the mass of an object and the force required to accelerate it.



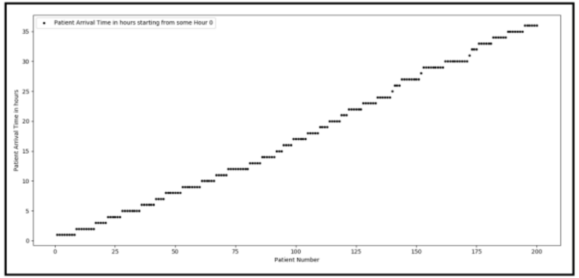
$$\mathbf{F} = m\mathbf{a}$$

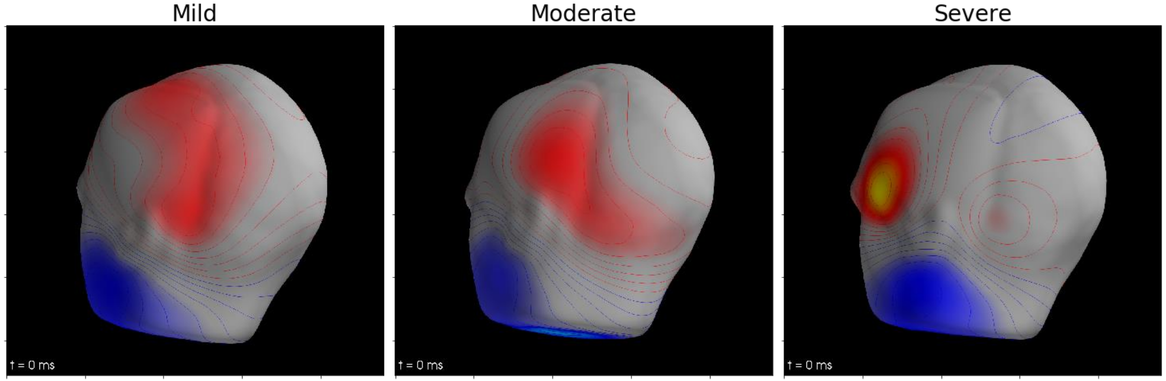
$$F = G\frac{m_1 m_2}{r^2}$$



Probabilistic models

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





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 [illegible]k at more sophisticated mo[illegible]

Regression Line & Correlation

[Review] The Regression Line

From (textbook):

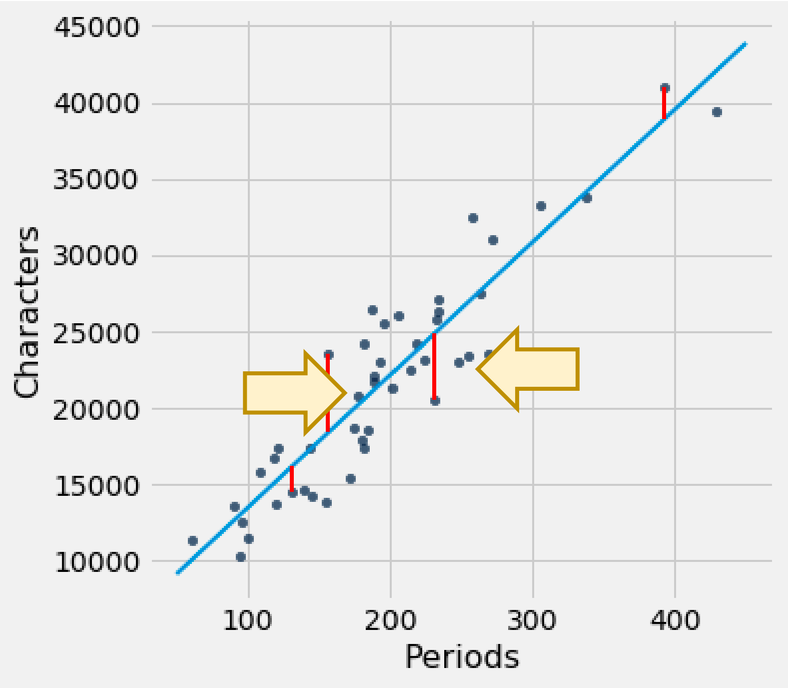
> The **regression line** is the unique straight line that minimizes the **mean squared error** of estimation among all straight lines.

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

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

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

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



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] 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.

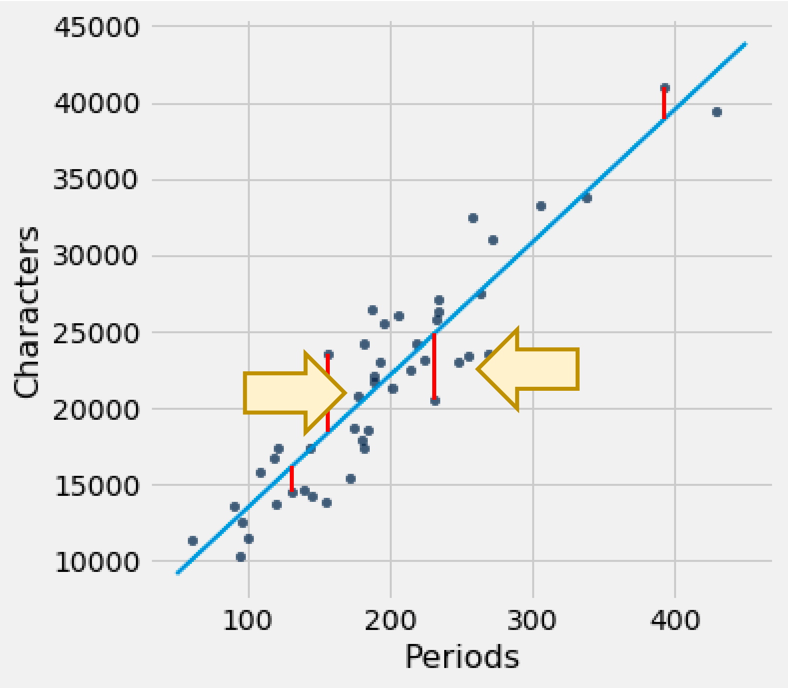
correlation

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

$$\begin{aligned}\text{intercept} = &\ \text{average of } y \\ &- \text{slope} \times \text{average of } x\end{aligned}$$

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

$$\begin{aligned}\text{residual} = &\ \text{observed } y \\ &- \text{regression estimate}\end{aligned}$$



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.

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

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

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

[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.

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

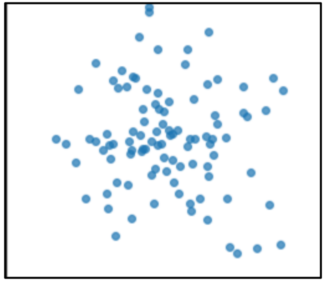
means $\bar{x}, \bar{y}$ standard deviations σ_x, σ_y

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

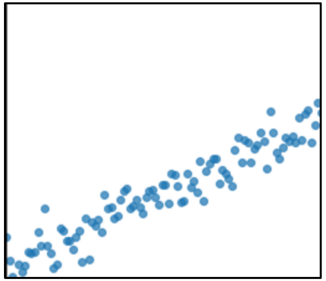
- x_i in standard units: $\frac{x_i - \bar{x}}{\sigma_x}$
- r is also known as Pearson's correlation coefficient.
- Side note: **covariance** is $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = r\sigma_x\sigma_y$

Correlation measures the strength of a **linear association** between two variables.

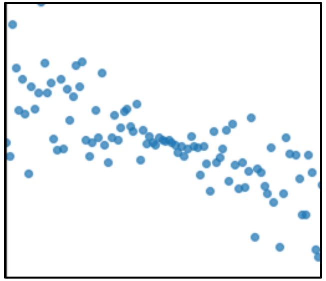
$$|r| \leq 1$$



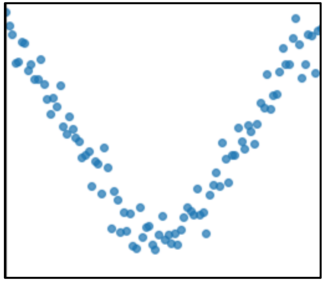
r = −0.121



r = 0.951



r = −0.723



⚠ **r = 0.056**

[Review] The Regression Line

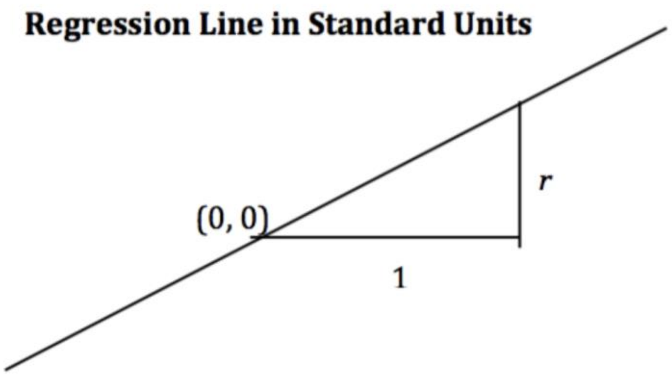
- When the variables x and y are measured in **standard units**, the regression line for predicting y based on x has slope r passes through the origin and the equation will be:

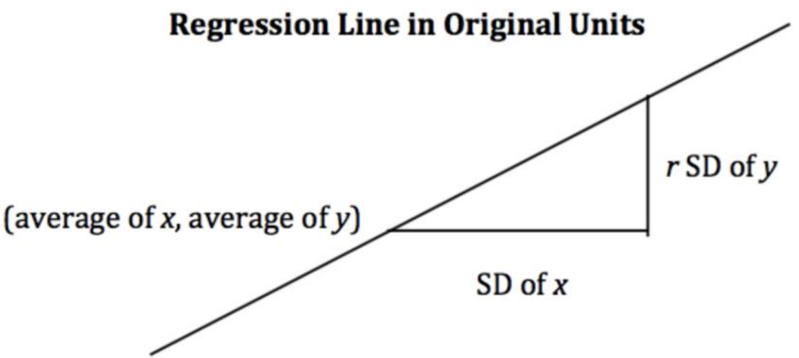
$$\hat{y} = r \times x$$

(both measured in standard units)

- In the original units of the data, this becomes:

$$\frac{\hat{y} - \bar{y}}{\sigma_y} = r \times \frac{\hat{x} - \bar{x}}{\sigma_x}$$





[Review] The Regression Line

$$\frac{\hat{y} - \bar{y}}{\sigma_y} = r \times \frac{x - \bar{x}}{\sigma_x}$$

$$\hat{y} = \sigma_y \times r \times \frac{x - \bar{x}}{\sigma_x} + \bar{y}$$

$$\hat{y} = \boxed{\left(\frac{r\sigma_y}{\sigma_x}\right)} \times x + \boxed{\left(\bar{y} - \frac{r\sigma_y}{\sigma_x}\bar{x}\right)}$$

Recall regression line equation is defined as:

$$\hat{y} = \boxed{\hat{a}} + \boxed{\hat{b}}x$$

Goal: Derive and define everything on this slide!

slope: $r\frac{SD\ of\ y}{SD\ of\ x} = r\frac{\sigma_y}{\sigma_x}$

intercept: $\bar{y} - slope \times \bar{x}$

Error for the i-th data point: $e_i = y_i - \hat{y}_i$

The Modeling Process

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

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

Models

Some models we will see in the next few lectures:

$$\boxed{\begin{aligned} \hat{y}_i &= \theta_0 + \theta_1 x_i \\ \hat{y}_i &= \theta \\ \hat{y}_i &= x_i^\top \theta \end{aligned}}$$

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)

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

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

$\hat{\theta}$ 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 $\hat{\theta} = [\hat{\theta}_0, \hat{\theta}_1]$

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 parameters $\theta = [\theta_0, \theta_1]$

$\hat{\theta}$ 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 Model (SLR)

Data 8 notation: $\hat{y} = a + bx$ → Data 100 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.

$$x \longrightarrow \boxed{\text{SLR} \quad \theta = [\theta_0, \theta_1]} \longrightarrow \hat{y}$$

- 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

Lecture 9, Data 100 Summer 2023

- 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

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

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

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

Click **Present with Slido** or install our [Chrome extension] to activate this poll while presenting.

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

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]**.

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

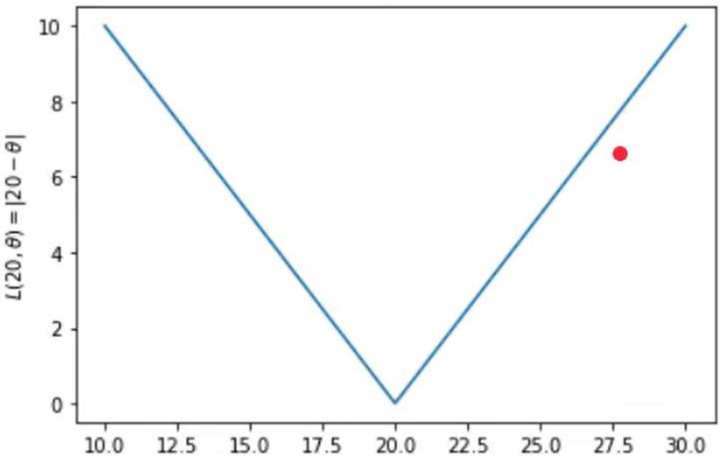
$$R(\theta) = \frac{1}{5}(|20 - \theta| + |21 - \theta| + |22 - \theta| + |29 - \theta| + |33 - \theta|)$$

Question?

Exploring MAE

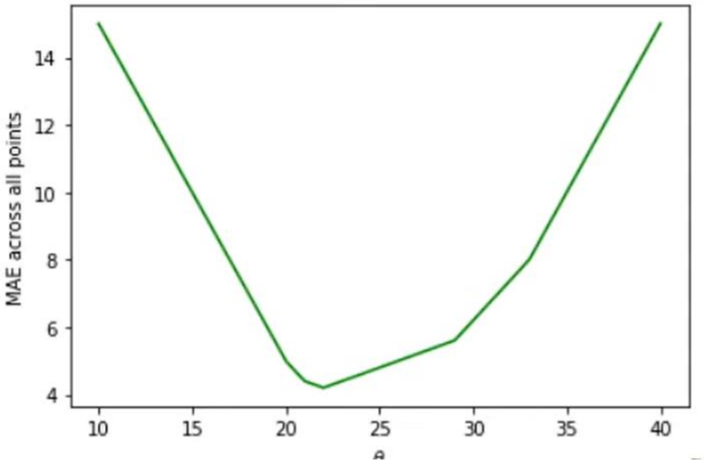
$$L_1(20, \theta) = |20 - \theta|$$

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

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

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

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

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

$|x|$

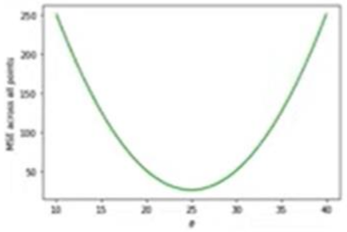
$$|x| = \begin{cases} x & x \geq 0 \\ -x & x < 0 \end{cases}$$

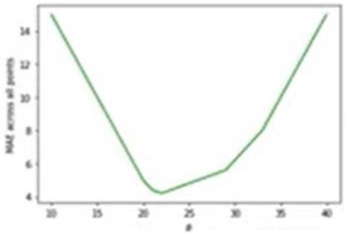
$$\frac{d|x|}{dx} = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \end{cases}$$

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

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.

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

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

MSE for SLR

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?

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?

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

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?

Interlude

Ara Hagopian
@ara_hagopian
"Go hang a salami, I'm a lasagna hog" is a palindrome.
12:18 PM · May 29, 2021
545 Retweets 40 Quotes 4,969 Likes 96 Bookmarks

GO HANG A SALAMI! I'M A LASAGNA HOG!
and Other Palindromes by JON AGEE

Fit 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:

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

and

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

$\boxed{1}$

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i\right) = 0 \Longleftrightarrow \overbrace{\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)}^{\bar{y}} - \hat{\theta}_0 - \hat{\theta}_1 \overbrace{\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)}^{\bar{x}} = 0$$

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

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

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

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

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

Reminder

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

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

Plug in definitions of correlation and SD:

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

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

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

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

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

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

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

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

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

4. Evaluate model performance

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

Next Time!