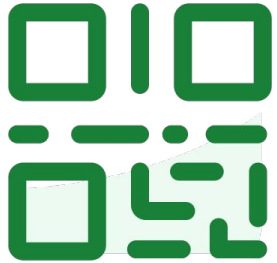




slido



Join at [slido.com](https://slido.com)  
#8041959

① Click **Present with Slido** or install our [Chrome extension](#) to display joining instructions for participants while presenting.



## LECTURE 10

# Introduction to Modeling, SLR

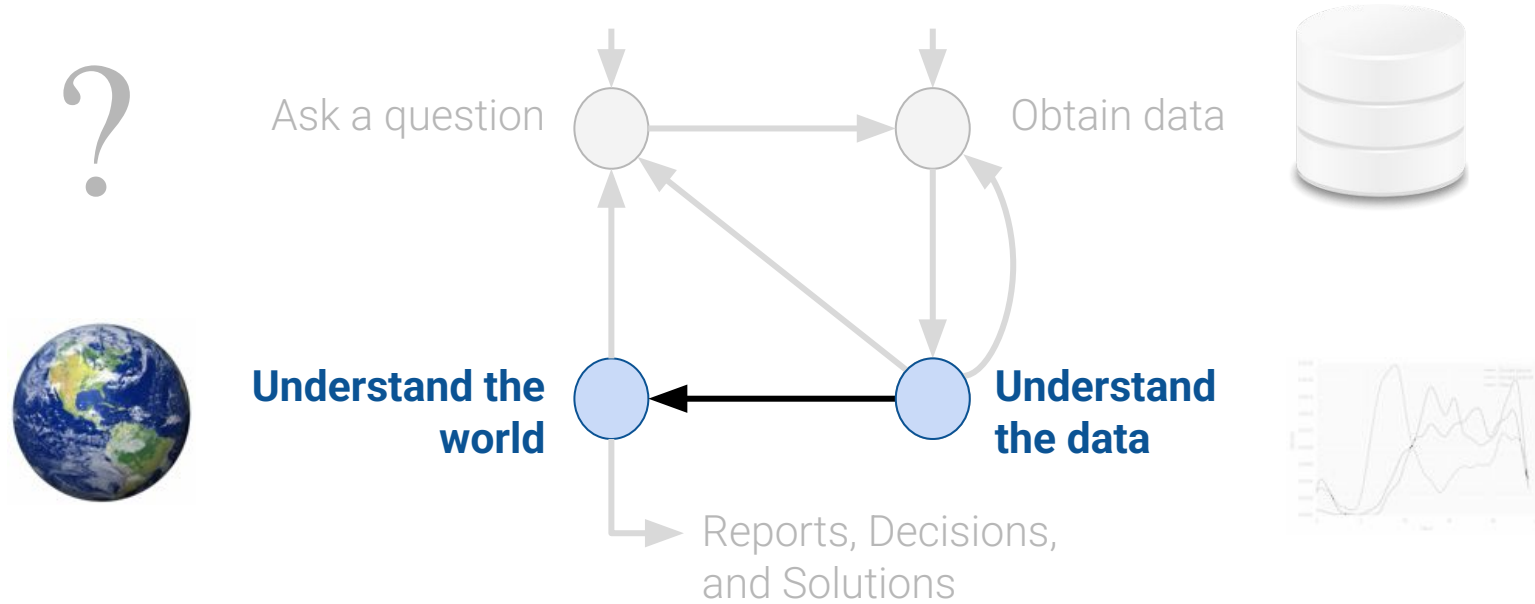
Understanding the usefulness of models and the simple linear regression model

**Data 100/Data 200, Spring 2025 @ UC Berkeley**

Narges Norouzi and Josh Grossman

Content credit: [Acknowledgments](#)

# 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

---

Lecture 10, Data 100 Spring 2025

- 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?

---

Lecture 10, Data 100 Spring 2025

- **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?



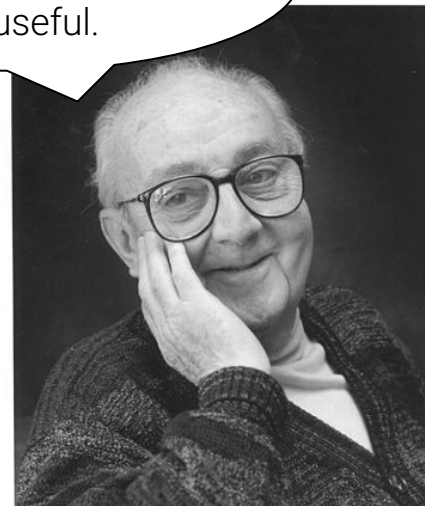
A model is an **idealized representation** of a system.

## Example:

We model the fall of an object on Earth as subject to a constant acceleration of  $9.81 \text{ m/s}^2$  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



8041959

# 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**.



# Data 8 Review: Regression Line & Correlation

---

Lecture 10, Data 100 Spring 2025

- 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





## [Data 8 Review] The Regression Line

From Data 8 ([textbook](#)):

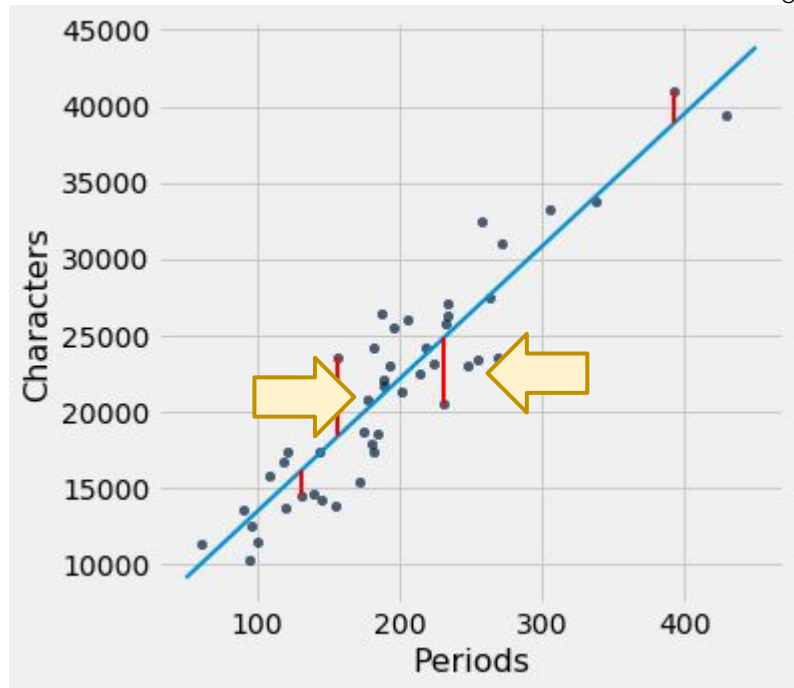
The **regression line** is the unique straight line that minimizes the **mean squared error** of estimation among all straight lines.

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

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

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

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



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



## [Data 8 Review] The Regression Line

From Data 8 (textbook):

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

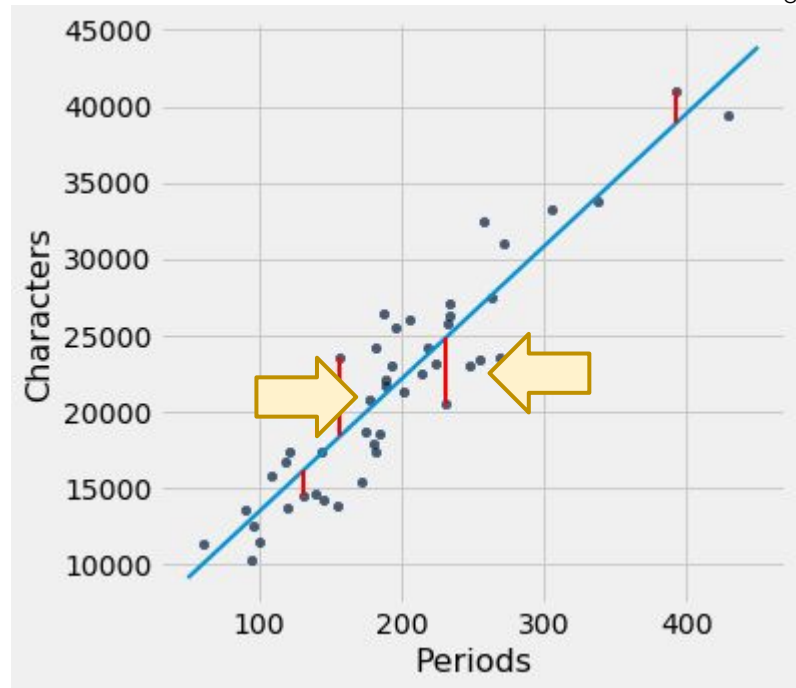
**correlation**

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

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

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

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



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



From Data 8 ([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  $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$

means  $\bar{x}, \bar{y}$  standard deviations  $\sigma_x, \sigma_y$

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



From Data 8 (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  $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$

means  $\bar{x}, \bar{y}$  standard deviations  $\sigma_x, \sigma_y$

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

## [Data 8 Review] Correlation

From Data 8 (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:

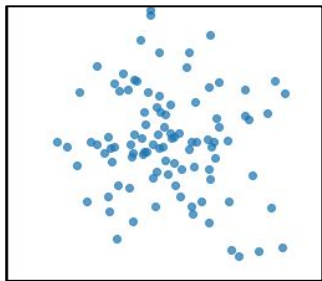
data  $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$

means  $\bar{x}, \bar{y}$  standard deviations  $\sigma_x, \sigma_y$

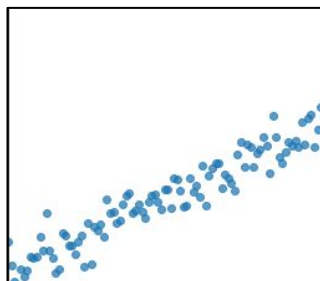
- $x_i$  in standard units:  $\frac{x_i - \bar{x}}{\sigma_x}$
- $r$  is also known as Pearson's correlation coefficient.
- Side note: **covariance** is  $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = r\sigma_x\sigma_y$

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

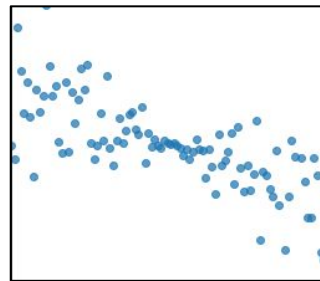
$$|r| \leq 1$$



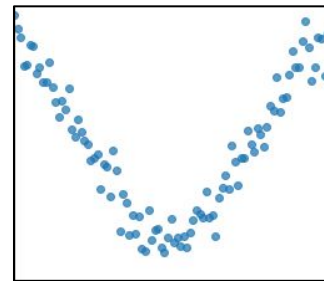
$r = -0.121$



$r = 0.951$



$r = -0.723$



!  $r = 0.056$



- 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:

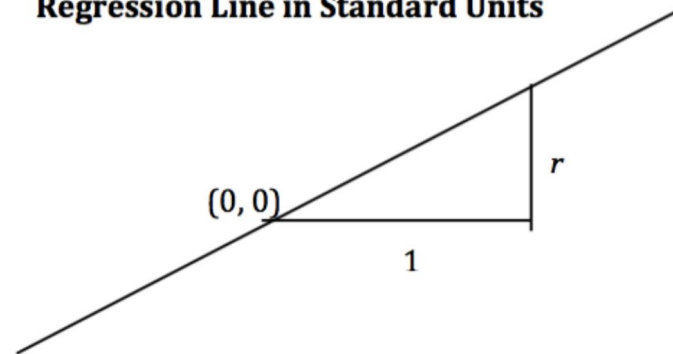
$$\hat{y} = r \times x$$

(both measured in standard units)

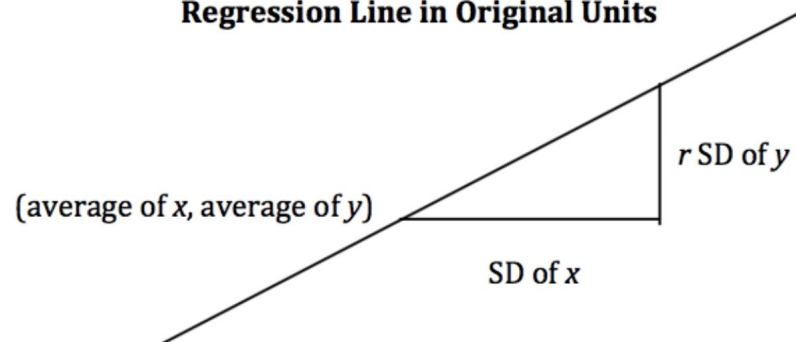
- In the original units of the data, this becomes:

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

**Regression Line in Standard Units**



**Regression Line in Original Units**





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

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:

intercept:

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



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

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

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

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

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

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





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

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

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

Recall regression line equation is defined as:

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

Goal: Derive and define everything on this slide!

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

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

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



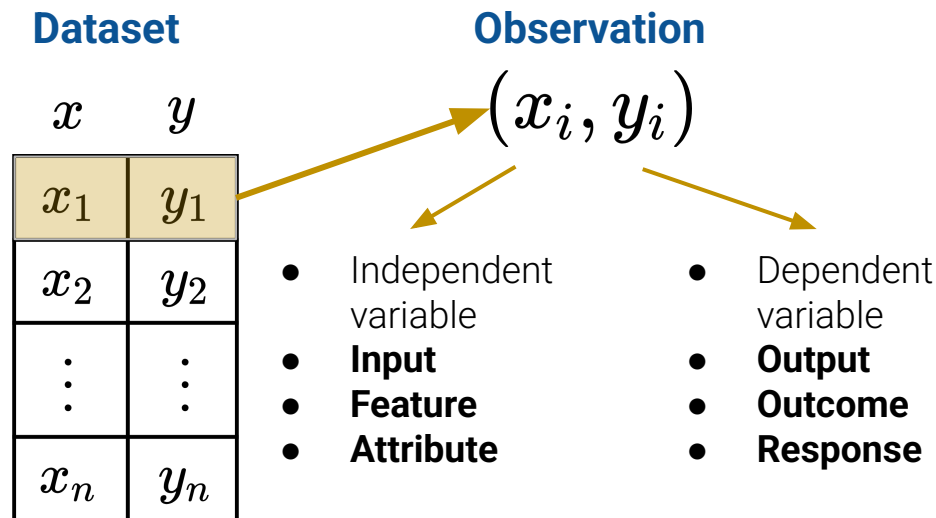
# The Modeling Process

---

Lecture 10, Data 100 Spring 2025

- 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

In Data 100, we'll treat a model as some mathematical rule or function to describe the relationships between variables.



## 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:

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

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

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

**Parametric models**



**Parametric models** are described by a few **parameters** ( $\theta_0, \theta_1$ , 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.


 Model parameter(s)


$$\left. \begin{array}{l} \end{array} \right\} \hat{y} = \theta_0 + \theta_1 x \quad \begin{array}{l} \text{Any linear model with} \\ \text{parameters } \theta = [\theta_0, \theta_1] \end{array}$$



**Parametric models** are described by a few **parameters** ( $\theta_0, \theta_1$ , 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  $\theta$  is written as  $\hat{\theta}$ .
  - The "hat" here is different from the "hat" in  $\hat{y}$ : one means estimate and one means prediction.
- Usually, we pick the parameters that appear "**best**" according to some criterion we choose.

 Model parameter(s)  $\left. \vphantom{\begin{matrix} \text{Model parameter(s)} \\ \text{Estimated parameter(s)} \end{matrix}} \right\} \hat{y} = \theta_0 + \theta_1 x$  Any linear model with parameters  $\theta = [\theta_0, \theta_1]$

 Estimated parameter(s),  
"best" fit to data in some sense  $\left. \vphantom{\begin{matrix} \text{Model parameter(s)} \\ \text{Estimated parameter(s)} \end{matrix}} \right\} \hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$  The "best" fitting linear model with parameters  $\hat{\theta} = [\hat{\theta}_0, \hat{\theta}_1]$



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

- 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  $\theta$  is written as  $\hat{\theta}$ .
  - The "hat" here is different from the "hat" in  $\hat{y}$ : one means estimate and one means prediction.
- Usually, we pick the criterion we choose.

**Note:** Not all statistical models have parameters! KDEs, k-Nearest Neighbor classifiers are non-parametric models.

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

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



## 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?



8041959

# Choose a Model

---

Lecture 10, Data 100 Spring 2025

- 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:

$$\hat{y} = a + bx$$

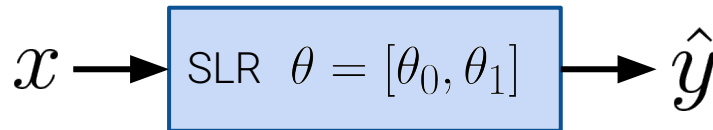


Data 100  
notation:

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

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

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



- 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.



## 1. Choose a model



How should we represent the world?

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



## 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 10, Data 100 Spring 2025

- 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



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?

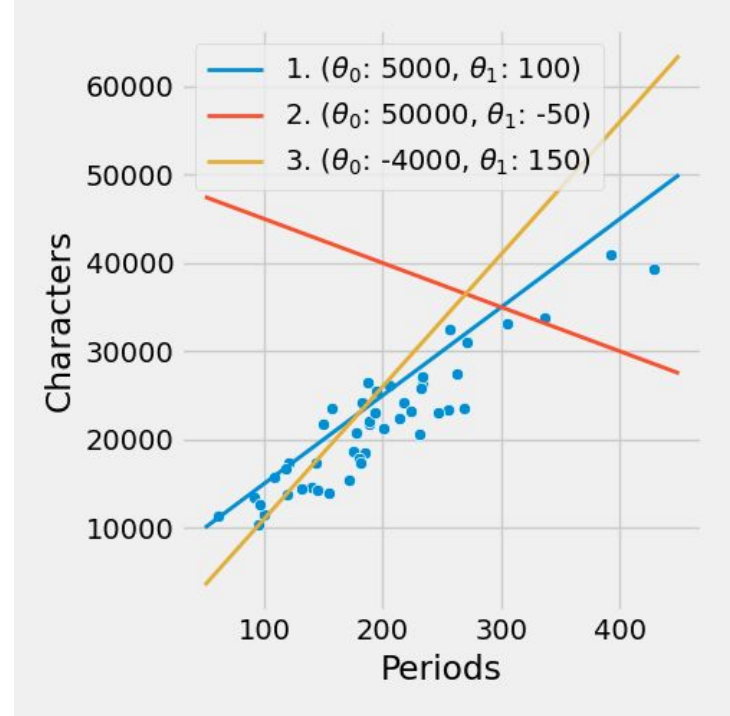
## Which $\hat{y}$ is best?

Based on your interpretation of the data, which are the "optimal parameters" for this linear model?

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

$$\hat{\theta}_0 = ? \quad \hat{\theta}_1 = ?$$

We only had 3 values to choose from to find the optimal parameter. In practice, our parameter domain is all reals, i.e.,  $\theta = [\theta_0, \theta_1] \in \mathbb{R}^2$



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



slido



# Which of these lines matches the data better?

① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.



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)



### 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**

### 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**





### 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

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:
  - $\hat{y} = y \rightarrow$  good prediction  
 $\rightarrow$  good fit  $\rightarrow$  no loss
  - $\hat{y}$  far from  $y \rightarrow$  bad prediction  
 $\rightarrow$  bad fit  $\rightarrow$  some loss

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

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

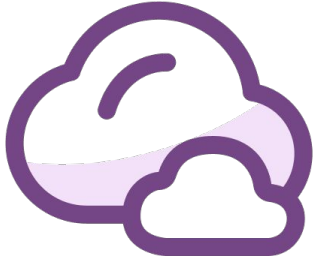
slido



If we want to penalize large residuals more than small residuals, which loss function is more ideal?

① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.

slido



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

① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.



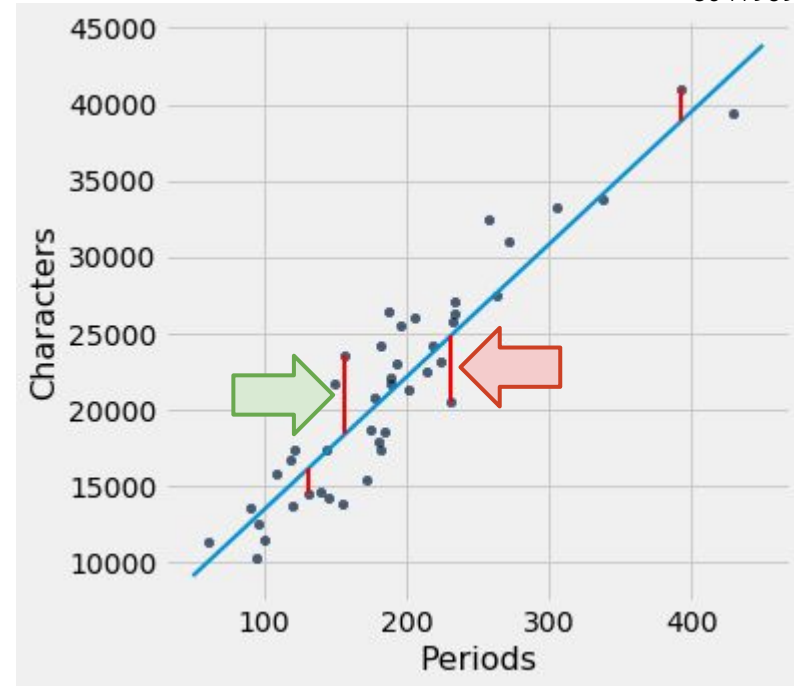
8041959

## 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.



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



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  $\theta$  (holding the data fixed) because  $\theta$  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  $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ :

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

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)**

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



1. Choose a model



How should we represent the world?

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

SLR model

2. Choose a loss function



How do we quantify prediction error?

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

Squared loss

3. Fit the model

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

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

MSE for SLR

4. Evaluate model performance

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

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



# The Modeling Process

1. Choose a model



How should we represent the world?

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

2. Choose a loss function



How do we quantify prediction error?

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

**3. Fit the model**

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

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

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

4. Evaluate model performance

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





# Fit the Model

---

Lecture 10, Data 100 Spring 2025

- 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



## 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:**

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

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



8041959

## Partial Derivative of MSE with Respect to $\theta_0, \theta_1$

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

Derivative of  
sum is sum  
of derivatives

Chain rule

Simplify  
constants

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

Derivative of  
sum is sum  
of derivatives

Chain rule

Simplify  
constants



## Partial Derivative of MSE with Respect to $\theta_0, \theta_1$

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

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

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

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

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

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



## Partial Derivative of MSE with Respect to $\theta_0, \theta_1$

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

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

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

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

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

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



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

$$\begin{aligned} 0 = \frac{\partial}{\partial \theta_0} MSE &= -\frac{2}{n} \sum_{i=1}^n (y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i) \iff \frac{1}{n} \sum_i y_i - \hat{y}_i = 0 \\ 0 = \frac{\partial}{\partial \theta_1} MSE &= -\frac{2}{n} \sum_{i=1}^n (y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i) x_i \iff \frac{1}{n} \sum_i (y_i - \hat{y}_i) x_i = 0 \end{aligned}$$

“Equivalent”

**Estimating equations**

To find the best  $\theta_0, \theta_1$ , we need to solve the **estimating equations** on the right.



**Goal:** Choose  $\theta_0, \theta_1$  to solve two estimating equations:

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

and

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

**1**

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

Separating terms



**Goal:** Choose  $\theta_0, \theta_1$  to solve two estimating equations:

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

and

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

**1**

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i) &= 0 \xLeftrightarrow{\text{Separating terms}} \left( \frac{1}{n} \sum_{i=1}^n y_i \right) - \hat{\theta}_0 - \hat{\theta}_1 \left( \frac{1}{n} \sum_{i=1}^n x_i \right) = 0 \\ &\xLeftrightarrow{\quad} \overbrace{\left( \frac{1}{n} \sum_{i=1}^n y_i \right)}^{\bar{y}} - \hat{\theta}_0 - \hat{\theta}_1 \overbrace{\left( \frac{1}{n} \sum_{i=1}^n x_i \right)}^{\bar{x}} = 0 \\ &\xLeftrightarrow{\quad} \bar{y} - \hat{\theta}_0 - \hat{\theta}_1 \bar{x} = 0 \\ &\xLeftrightarrow{\quad} \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} \end{aligned}$$





**Goal:** Choose  $\theta_0, \theta_1$  to solve two estimating equations:

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

and

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

$\boxed{1}$

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i) &= 0 \xLeftrightarrow{\text{Separating terms}} \left( \overbrace{\frac{1}{n} \sum_{i=1}^n y_i}^{\bar{y}} \right) - \hat{\theta}_0 - \hat{\theta}_1 \left( \overbrace{\frac{1}{n} \sum_{i=1}^n x_i}^{\bar{x}} \right) = 0 \\ &\xLeftrightarrow{\quad} \bar{y} - \hat{\theta}_0 - \hat{\theta}_1 \bar{x} = 0 \\ &\xLeftrightarrow{\quad} \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} \end{aligned}$$

# From Estimating Equations to Estimators

**Goal:** Choose  $\theta_0, \theta_1$  to solve two estimating equations:

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

and

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

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

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

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

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

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



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

**Plug in definitions of correlation and SD:**

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

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

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

### Reminder

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

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



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

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

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

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

For SLR:

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



# Evaluate the Model

---

Lecture 10, Data 100 Spring 2025

- 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**



1. Choose a model



How should we represent the world?

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

2. Choose a loss function



How do we quantify prediction error?

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

3. Fit the model



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

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

**4. Evaluate model performance**

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

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x \quad \left\{ \begin{array}{l} \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} \\ \hat{\theta}_1 = r \frac{\sigma_y}{\sigma_x} \end{array} \right.$$

What are some ways to determine if our model was a good fit to our data?

## 1. Visualize data, compute statistics:

Plot original data.

Compute column means, standard deviation.

If we want to fit a linear model, compute correlation  $r$ .

## 2. Performance metrics:

### Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

- It is the square root of MSE, which is the average loss that we've been minimizing to determine optimal model parameters.
- RMSE is in the **same units** as  $y$ .
- A lower RMSE indicates more "accurate" predictions (lower "average loss" across data)

## 3. Visualization:

Look at a residual plot of  $e_i = y_i - \hat{y}_i$  to visualize the difference between actual and predicted values.



Ideal model evaluation steps, in order:

1. **Visualize original data, Compute Statistics**

2. **Performance Metrics**

For our simple linear least square model, use RMSE (we'll see more metrics later)

3. **Residual Visualization**

4 datasets could have similar aggregate statistics but still be wildly different:

```
x_mean : 9.00, y_mean : 7.50  
x_stdev: 3.16, y_stdev: 1.94  
r = Correlation(x, y): 0.816  
theta_0_hat: 3.00, theta_1_hat: 0.50  
RMSE: 1.119
```

Demo



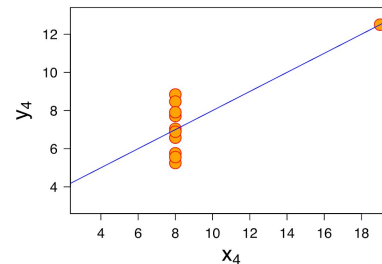
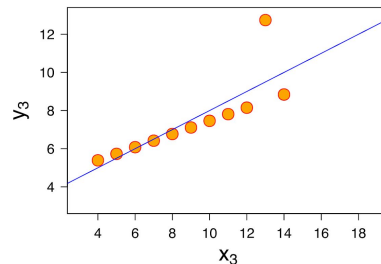
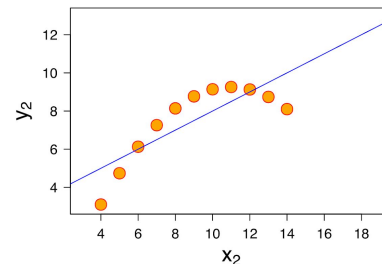
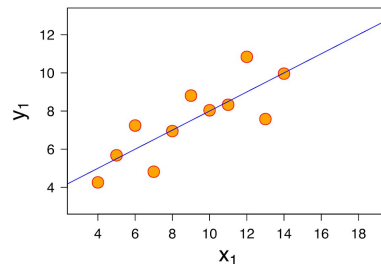
## Four Mysterious Datasets (Anscombe's quartet)



- **The four dataset** each have the same mean of  $x$ , mean of  $y$ , SD of  $x$ , SD of  $y$ , and  $r$  value.
- Since our optimal Least Squares SLR model only depends on those quantities, they all have the **same regression line** and RMSE.

However, only one of these four sets of data makes sense to model using SLR.

Before modeling, you should always **visualize** your data first!



Demo



Ideal model evaluation steps, in order:

1. **Visualize original data, Compute Statistics**
2. **Performance Metrics**

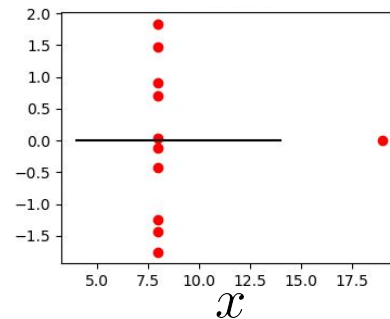
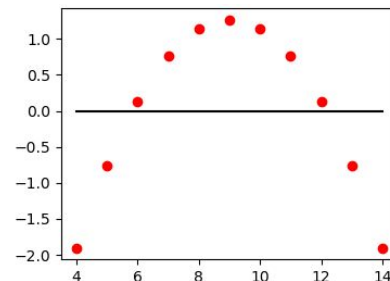
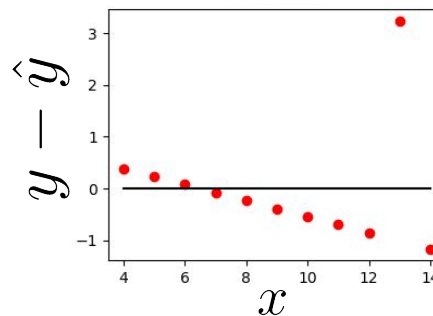
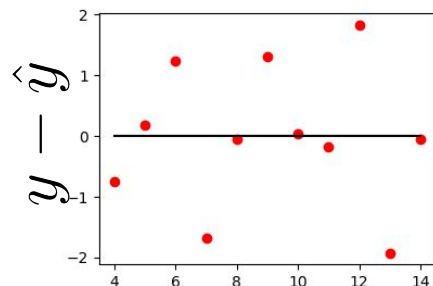
For our simple linear least square model, use RMSE (we'll see more metrics later)

3. **Residual Visualization**

Demo

From Data 8:

The residual plot of a good regression shows **no pattern**.





## Lecture 10

# Introduction to Modeling, SLR

Content credit: [Acknowledgments](#)