

LECTURE 10

# Introduction to Modeling, SLR

Understanding the usefulness of models and the simple linear regression model

# Today's Roadmap

---

## **Review: Regression Line, Correlation**

What is a model?

The Modeling Process: Definitions

Loss Functions

Minimizing Average Loss (Empirical Risk)

Interpreting SLR: Slope, Anscombe's Quartet

Evaluating the Model: RMSE, Residual Plot

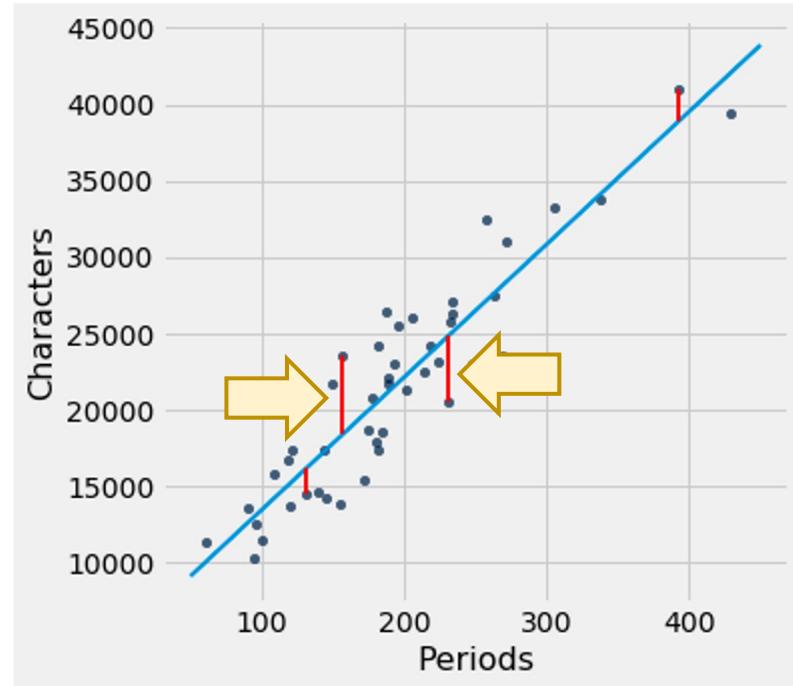
# The Regression Line

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

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

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

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



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

# The Regression Line

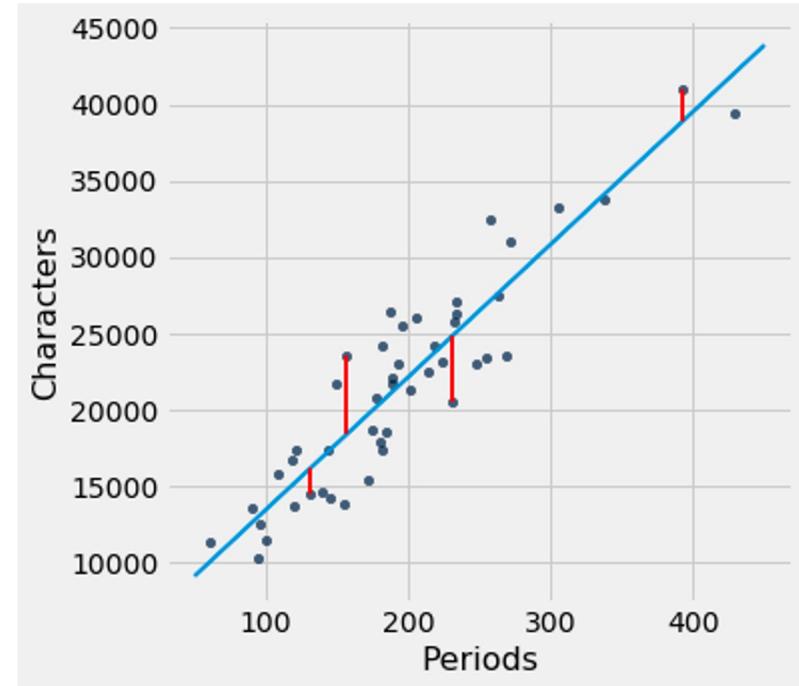
The **regression line** is the unique straight line that minimizes the **mean squared error** of estimation among all straight lines.

## correlation

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

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

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



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

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:

$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$  data  
 $\bar{x}, \bar{y}$  means;  $\sigma_x, \sigma_y$  standard deviations

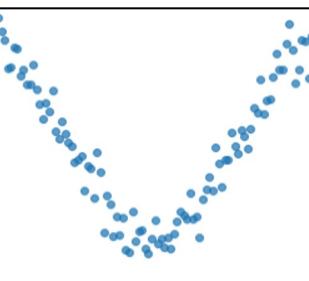
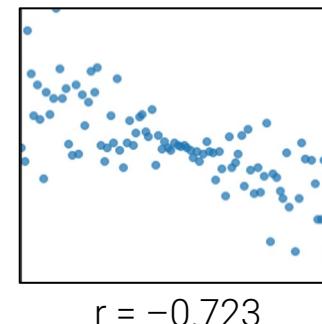
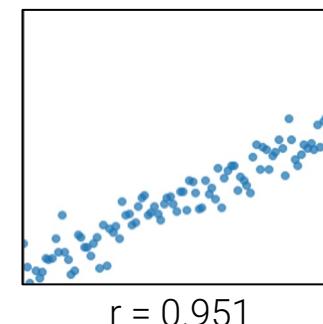
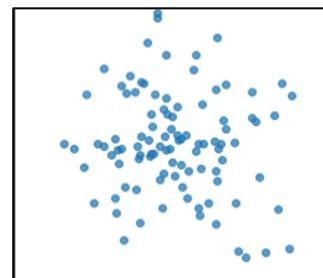
- $x_i$  in standard units:  $\frac{x_i - \bar{x}}{\sigma_x}$
- $r$  is also known as Pearson's correlation coefficient.
- Side note: **covariance** is  $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = r\sigma_x\sigma_y$

# Correlation

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

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

Correlation measures the strength of a **linear association** between two variables.  
 $|r| < 1$



Define the following:

$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$  data  
 $\bar{x}, \bar{y}$  means;  $\sigma_x, \sigma_y$  standard deviations

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

## Expressing the Regression Line Mathematically

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

Define the following:

$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$  data  
 $\bar{x}, \bar{y}$  means;  $\sigma_x, \sigma_y$  standard deviations;  
 $r$  correlation coefficient

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

**regression line**

1. slope  $= r \cdot \frac{\text{SD of } y}{\text{SD of } x}$

2. intercept  $= \text{average of } y - \text{slope} \cdot \text{average of } x$

3. residual  $= \text{observed value} - \text{regression estimate}$

?

Rewrite each expression using math notation.



## Expressing the Regression Line Mathematically

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

Define the following:

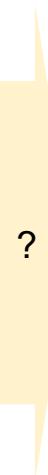
$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$  data  
 $\bar{x}, \bar{y}$  means;  $\sigma_x, \sigma_y$  standard deviations;  
 $r$  correlation coefficient

$$\hat{y} = \hat{a} + \hat{b}x \quad \text{regression line}$$

1. slope  $= r \cdot \frac{\text{SD of } y}{\text{SD of } x}$

2. intercept  $= \text{average of } y - \text{slope} \cdot \text{average of } x$

3. residual  $= \text{observed value} - \text{regression estimate}$



$$\hat{b} = r \frac{\sigma_y}{\sigma_x}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$e_i = y_i - \hat{y}_i$$

Error for the i-th datapoint

## Today's Goal

---

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

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

$$\hat{b} = r \frac{\sigma_y}{\sigma_x}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$e_i = y_i - \hat{y}_i$$

Goal: Derive and define everything on this slide!

# What is a model?

---

Review: Simple Linear Regression and Correlation

## **What is a model?**

The Modeling Process: Definitions

Loss Functions

Minimizing Average Loss on Data

Interpreting SLR: Slope

Evaluating the Model: RMSE, Residual Plot

# What is a model?

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

Example:

We model the fall of an object on Earth as subject to a constant acceleration of 9.81 m/s<sup>2</sup> 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

# Why do we build models?

Reason 1:

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

- What factors play a role in the growth of COVID-19?
- How do an object's velocity and acceleration impact how far it travels?  
(Physics:  $d = d_0 + vt + \frac{1}{2}at^2$  )

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 this 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.

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

# Two common types of models

## Physical (mechanistic) models

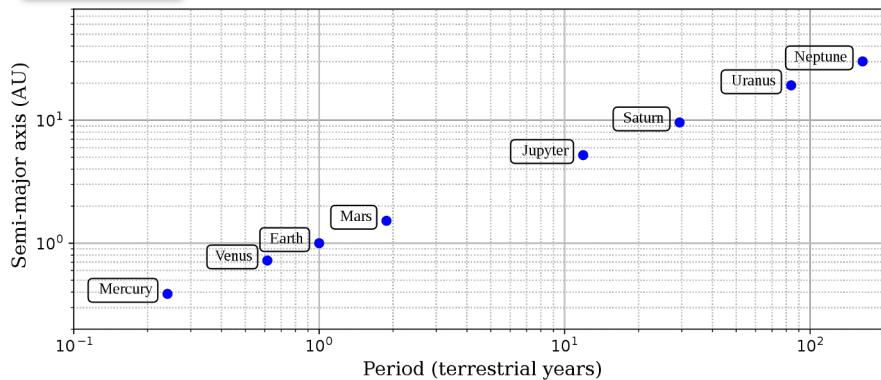
Laws that govern how the world works.

Kepler's Third Law of Planetary Motion (1619)

[\[Wikipedia\]](#)

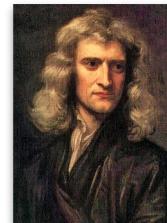


$$T^2 \propto R^3$$



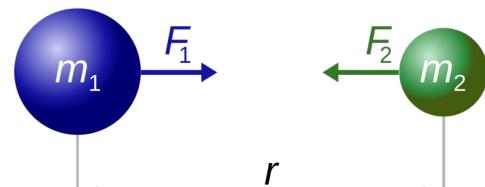
Newton's Laws: motion and gravitation (1687)

[\[Wikipedia\]](#)



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

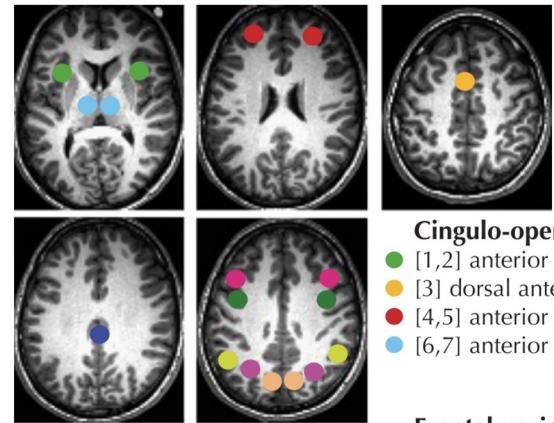
$$F = G \frac{m_1 m_2}{r^2}$$



# Two common types of models

## Statistical models

Relationships between variables found through data and statistical analysis.

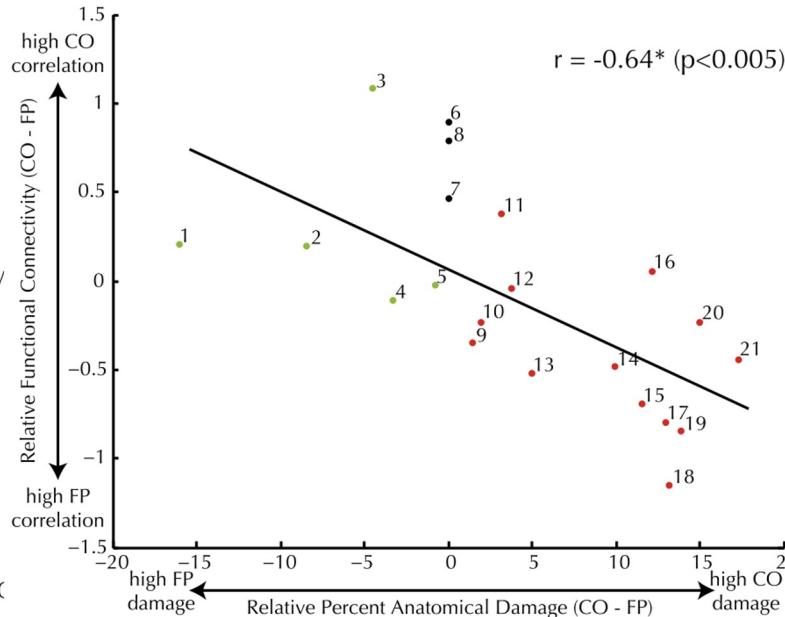


### Cingulo-opercular (CO)

- [1,2] anterior insula/frontal operculum (al/FO)
- [3] dorsal anterior cingulate (dACC)
- [4,5] anterior prefrontal cortex (aPFC)
- [6,7] anterior thalamus (ant thalamus)

### Frontal-parietal (FP)

- [1,2] intraparietal sulcus (IPS)
- [3,4] frontal cortex
- [5,6] precuneus
- [7,8] intraparietal lobule (IPL)
- [9,10] dorsolateral prefrontal cortex (dlPFC)
- [11] midcingulate



Nomura et al.,  
PNAS 2010  
[paper]

# The Modeling Process: Definitions

---

Review: Simple Linear Regression and Correlation

What is a model?

## **The Modeling Process: Definitions**

Loss Functions

Minimizing Average Loss on Data

Interpreting SLR: Slope

Evaluating the Model: RMSE, Residual Plot

### Simple Linear Regression Model (SLR)

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

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

What do we mean by “best”?  
What’s the difference between  
 $a, b$       and     $\hat{a}, \hat{b}$  ?

## Parametric Model Notation

$y$  True outputs

$\hat{y}$  Predicted outputs

$\theta$  Model parameter(s)

$\hat{\theta}$  Optimal parameter(s),  
for some definition of optimal

For data:

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

The i-th datapoint is an **observation**:

- $y_i$  is the i-th **output** (aka dependent variable)
- $x_i$  is the i-th **feature** (aka independent variable)
- $\hat{y}_i$  is the i-th **prediction** (aka estimation).

$$\left. \begin{array}{l} \hat{y} = a + bx \\ \hat{y} = \hat{a} + \hat{b}x \end{array} \right\}$$

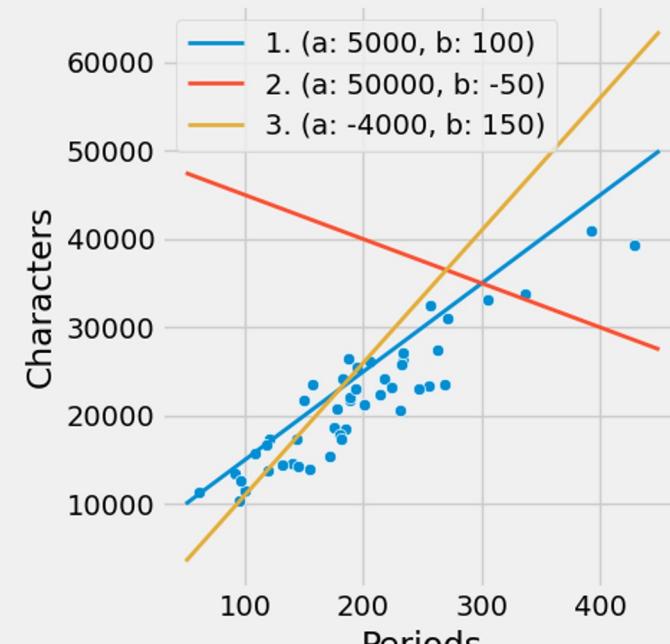
Any linear model with parameters  $\theta = (a, b)$

The “best” linear model with parameters  $\hat{\theta} = (\hat{a}, \hat{b})$

## Which $\theta$ is best?

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

$$\hat{y} = a + bx$$
$$\hat{a} = ? \quad \hat{b} = ?$$



We only had 3 values to choose from to find the optimal parameter. In practice, our parameter domain is all reals, i.e.,  $(a, b) \in \mathbb{R}^2$

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



# Simple Linear Regression: Our First Model

## Simple Linear Regression Model (SLR)

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

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.  $x \rightarrow \text{SLR } \theta = (a, b) \rightarrow \hat{y}$
- $x$  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.
- Other parametric models we'll see soon:  $\hat{y} = \theta$      $\hat{y} = x^T \theta$      $\hat{y} = \frac{1}{1 + \exp(-x^T \vec{\theta})}$
- Note: Not all statistical models have parameters! KDEs are non-parametric models.

# The Modeling Process



## 1. Choose a model

How should we represent the world?

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

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?

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

## 4. Evaluate model performance

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

# Loss Functions

---

Review: Simple Linear Regression and Correlation

What is a model?

The Modeling Process: Definitions

## **Loss Functions**

Minimizing Average Loss on Data

Interpreting SLR: Slope

Evaluating the Model: RMSE, Residual Plot

# The Modeling Process



1. Choose a model

How should we represent the world?

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

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)

### Squared Loss

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

- Widely used
- Also called “L2 loss”
- Reasonable:
  - $\hat{y} = y \rightarrow$  good prediction → good fit → no loss
  - $\hat{y}$  far from  $y \rightarrow$  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:
  - $\hat{y} = y \rightarrow$  good prediction → good fit → no loss
  - $\hat{y}$  far from  $y \rightarrow$  bad prediction → bad fit → *some loss*

## L2 and L1 Loss for SLR

### Squared Loss (L2 Loss)

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

For an SLR model  $\hat{y} = a + bx$  :

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

### Absolute Loss (L1 Loss)

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

For an SLR model  $\hat{y} = a + bx$  :

**1.** What is the SLR L1 Loss?

**2.** Why don't we directly use residual error as the loss function?  $e = (y - \hat{y})$

**3.** Which loss function is better: L1 or L2?



## L2 and L1 Loss for SLR

### Squared Loss (L2 Loss)

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

For an SLR model  $\hat{y} = a + bx$  :

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

### Absolute Loss (L1 Loss)

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

For an SLR model  $\hat{y} = a + bx$  :

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

Why don't we directly use residual error as the loss function?  $e = (y - \hat{y})$

- This unfortunately treats “negative” predictions and “positive” predictions differently.
- Predicting 16 when the true value is 15 should be penalized the same as predicting 14.

Which loss function is better: L1 or L2?

We'll compare tradeoffs next lecture.  
Today we'll focus on Squared (L2) Loss.



## Empirical Risk is Average Loss over Data

---

We care about how bad our model's predictions are for our entire data set, not just for one point. A natural measure, then, is of the **average loss** (aka **empirical risk**) across all points.

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

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

Average loss is a function of the parameter  $\theta$  because **our data do not change**. What defines how well our model works is our choice of  $\theta$ , which determines  $\hat{y}$ .

**The average loss of a model tells us how well it fits the given data.**

We want to **find the parameter(s) that minimize average loss** to best predict the data.

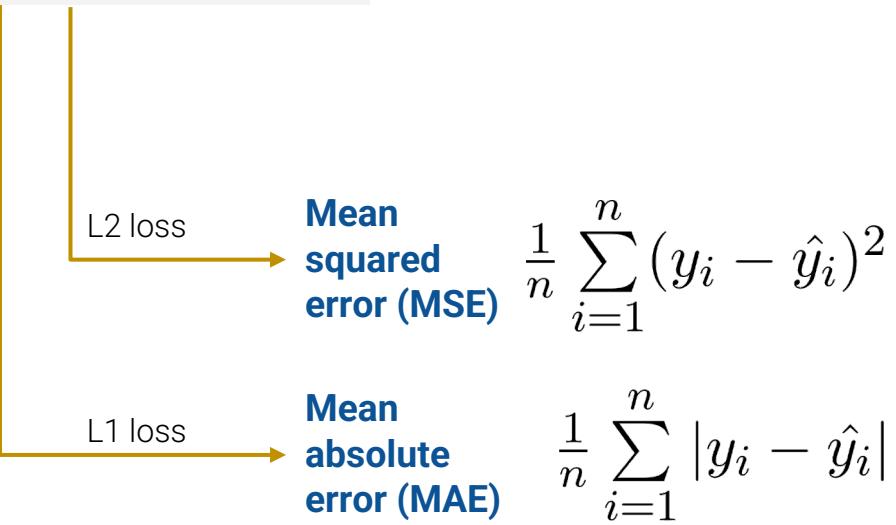
## Empirical Risk is Average Loss over Data

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

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

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



# 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?

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

SLR model

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

Squared loss

$$R(a, b) = \frac{1}{n} \sum_{i=1}^n (y_i - (a + bx_i))^2$$

MSE for SLR

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

# Minimizing Average Loss (Empirical Risk) on Data

---

Review: Simple Linear Regression and Correlation

What is a model?

The Modeling Process: Definitions  
Loss Functions

## **Minimizing Average Loss on Data**

Interpreting SLR: Slope

Evaluating the Model: RMSE, Residual Plot

# 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?

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

SLR model

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

Squared loss

$$R(a, b) = \frac{1}{n} \sum_{i=1}^n (y_i - (a + bx_i))^2$$

We want to find  $\hat{a}, \hat{b}$  that minimize this **objective function**.

## Minimizing MSE for the SLR model

**Objective function:** In optimization theory, the function to minimize.

Find the values of  $a, b$  that minimize the average squared loss (MSE) for the SLR model:

$$R(a, b) = \frac{1}{n} \sum_{i=1}^n (y_i - (a + bx_i))^2$$

Note: The optimal parameters  $a = \hat{a}, b = \hat{b}$  will also minimize this simplified function:

$$\sum_{i=1}^n (y_i - (a + bx_i))^2$$

- This is a quadratic function of two unknowns.
- Remember: The data  $(x_i, y_i)$  are known and—for our purposes—fixed.

Optimize this simplified objective with calculus!

## Step 1 of 2: Fix $b$ and minimize with respect to $a$

1. Rewrite the function:

$$\sum_{i=1}^n (y_i - (a + bx_i))^2 = \sum_{i=1}^n (y_i - bx_i - a)^2$$

2. Differentiate with respect to  $a$ :

$$\begin{aligned}\frac{\partial}{\partial a} \sum_{i=1}^n (y_i - bx_i - a)^2 &= \sum_{i=1}^n \frac{\partial}{\partial a} (y_i - bx_i - a)^2 && \text{Derivative of sum is sum of derivatives} \\ &= \sum_{i=1}^n 2(y_i - bx_i - a)(-1) && \text{Chain rule} \\ &= -2 \sum_{i=1}^n (y_i - bx_i - a) && \text{Simplify constants}\end{aligned}$$

3. Set equal to 0:

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

4. Finally, rearrange and solve for  $\hat{a}$ :

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

$$= \sum_{i=1}^n (y_i - bx_i - a)$$

$$= \sum_{i=1}^n y_i - b \sum_{i=1}^n x_i - na$$

$$na = \sum_{i=1}^n y_i - b \sum_{i=1}^n x_i$$

$$\hat{a} = \frac{1}{n} \sum_{i=1}^n y_i - b \cdot \frac{1}{n} \sum_{i=1}^n x_i$$

$$\hat{a} = \bar{y} - b\bar{x}$$

Pull out scalars

Divide by  $n$

## Step 2 of 2: Plug in $\hat{a}$ and minimize with respect to $b$

Our expression for  $\hat{a}$ :  $\hat{a} = \bar{y} - \hat{b}\bar{x}$

1. Plug in  $\hat{a}$  to our objective function:

$$\begin{aligned} \sum_{i=1}^n (y_i - (a + bx_i))^2 &= \sum_{i=1}^n (y_i - (\bar{y} - b\bar{x} + bx_i))^2 \\ &= \sum_{i=1}^n (y_i - \bar{y} - b(x_i - \bar{x}))^2 \end{aligned}$$

2. Differentiate with respect to  $b$ :

$$\begin{aligned} \sum_{i=1}^n \frac{\partial}{\partial b} (y_i - \bar{y} - b(x_i - \bar{x}))^2 &\quad \text{Derivative of sum is} \\ &\quad \text{sum of derivatives} \\ &= \sum_{i=1}^n 2 \cdot (y_i - \bar{y} - b(x_i - \bar{x})) \cdot (-1) \cdot (x_i - \bar{x}) \\ &\quad \text{Chain rule} \\ &\cdot = -2 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y} - b(x_i - \bar{x})) \quad \text{Simplify} \\ &\quad \text{constants} \end{aligned}$$

3. Set equal to 0:

$$0 = -2 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y} - b(x_i - \bar{x}))$$

4. Finally, rearrange and solve for  $\hat{b}$ :

$$0 = \left[ \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) - b \sum_{i=1}^n (x_i - \bar{x})^2 \right] \quad \text{Distributive prop.}$$

$$= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) - b \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{Separate sums}$$

Add in constants

$$= n\sigma_x\sigma_y \left( \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{\sigma_x} \right) \left( \frac{y_i - \bar{y}}{\sigma_y} \right) \right) - bn \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$= nr\sigma_x\sigma_y - bn\sigma_x^2$$

Definitions of  $r, \sigma_x$

$$b\sigma_x^2 = r\sigma_x\sigma_y$$

$$\hat{b} = \frac{r\sigma_x\sigma_y}{\sigma_x^2} = r \frac{\sigma_y}{\sigma_x}$$

$$\hat{b} = r \frac{\sigma_y}{\sigma_x}$$

# 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?

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

SLR model

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

Squared loss

$$R(a, b) = \frac{1}{n} \sum_{i=1}^n (y_i - (a + bx_i))^2$$

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

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$\hat{b} = r \frac{\sigma_y}{\sigma_x}$$

$$\hat{y} = \hat{a} + \hat{b}x \quad \text{regression line}$$

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

$$\hat{b} = r \frac{\sigma_y}{\sigma_x}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

## Galton demo

# Interpreting SLR: Slope

---

Review: Simple Linear Regression and Correlation

What is a model?

The Modeling Process: Definitions

Loss Functions

Minimizing Average Loss on Data

## **Interpreting SLR: Slope**

Evaluating the Model: RMSE, Residual Plot

# 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?**

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

SLR model

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

Squared loss  
(for today)

$$R(a, b) = \frac{1}{n} \sum_{i=1}^n (y_i - (a + bx_i))^2$$

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

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$\hat{b} = r \frac{\sigma_y}{\sigma_x}$$

## Interpreting the least squares linear regression model

You may sometimes hear the prediction task defined as: “**regressing** y on x.”

Suppose we fit a model that predicts a Chihuahua’s weight (in pounds) given its length (in inches).

$$\hat{y} = \hat{a} + \hat{b}x \quad \hat{a} = \bar{y} - \hat{b}\bar{x} \quad \hat{b} = r \frac{\sigma_y}{\sigma_x}$$

$$\text{predicted weight} = 3 + 2 \cdot \text{length}$$



# Interpreting the least squares linear regression model

You may sometimes hear the prediction task defined as: “**regressing** y on x.”

Suppose we fit a model that predicts a Chihuahua’s weight (in pounds) given its length (in inches).

## Interpreting the slope?

By definition, the slope measures the increase in y (pounds) for a 1 unit increase in x (1 inch).

1. Does this mean that if a chihuahua in the dataset grows 1 inch, we estimate that they will get 2 pounds heavier? What does it actually mean?

$$\hat{y} = \hat{a} + \hat{b}x \quad \hat{a} = \bar{y} - \hat{b}\bar{x} \quad \hat{b} = r \frac{\sigma_y}{\sigma_x}$$

**predicted weight =  $3 + 2 \cdot \text{length}$**



## Predicting on wildly different data?



**Chihuahuas** (left) range from 3-6 pounds, and 9.5-15 inches in length.  
**Great Danes** (right) range from 110-175 pounds, and 35.5-43 inches in length.

2. Should we use this model to predict the weight of Great Danes (a much bigger dog)?



# Interpreting the least squares linear regression model

You may sometimes hear the prediction task defined as: “**regressing** y on x.”

Suppose we fit a model that predicts a Chihuahua’s weight (in pounds) given its length (in inches).

## Interpreting the slope?

By definition, the slope measures the increase in y (pounds) for a 1 unit increase in x (1 inch).

1. Does this mean that if a chihuahua in the dataset grows 1 inch, we estimate that they will get 2 pounds heavier?

No!

$$\hat{y} = \hat{a} + \hat{b}x \quad \hat{a} = \bar{y} - \hat{b}\bar{x} \quad \hat{b} = r \frac{\sigma_y}{\sigma_x}$$

predicted weight =  $3 + 2 \cdot \text{length}$



- The model we created shows **association**, not causation.
- The data we collected is a snapshot of several chihuahuas at one instance of time (**cross-sectional**), not snapshots of chihuahuas over time (**longitudinal**).

Slope interpretation: If two chihuahuas have a 1 inch height difference, their estimated weight difference is 2 lbs.

# Interpreting the least squares linear regression model

You may sometimes hear the prediction task defined as: “**regressing** y on x.”

Suppose we fit a model that predicts a Chihuahua’s weight (in pounds) given its length (in inches).

$$\hat{y} = \hat{a} + \hat{b}x \quad \hat{a} = \bar{y} - \hat{b}\bar{x} \quad \hat{b} = r \frac{\sigma_y}{\sigma_x}$$

**predicted weight** =  $3 + 2 \cdot \text{length}$



## Predicting on wildly different data?



**Chihuahuas** (left) range from 3-6 pounds, and 9.5-15 inches in length.

**Great Danes** (right) range from 110-175 pounds, and 35.5-43 inches in length.

2. Should we use this model to predict the weight of Great Danes (a much bigger dog)? **No!**

- We have no indication that the weight vs. length relationship for Great Danes are the same as Chihuahuas.
- Great Danes’ weights and lengths are well outside of the range of weights and lengths we fit our model on.

If the new data we **test our model on** looks nothing like the data we **fit our model on**, there’s no guarantee that our model will be any good. We will formalize this notion in a few lectures.

# Evaluating the Model: RMSE, Residual Plot

---

Review: Simple Linear Regression and Correlation

What is a model?

The Modeling Process: Definitions

Loss Functions

Minimizing Average Loss on Data

Interpreting SLR: Slope

**Evaluating the Model: RMSE, Residual Plot**

# Evaluating models

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)

- 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)

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

## 3. Visualization

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

It is tempting to only look at step 2. But you  
need to always visualize!!!!

## Demo Slides

# Visualize, then quantify!

**Anscombe's quartet** refers to the following four sets of points on the right.

- They 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**.

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

Before modeling, you should always visualize your data first!

$$\bar{x} = 9, \bar{y} = 7.501$$

$$\sigma_x = 3.162, \sigma_y = 1.937$$

$$r = 0.816$$

