slido

1357201

Join at slido.com #1357201

ⓘ Start presenting to display the joining instructions on this slide.

1357201

LECTURE 10

Introduction to Modeling, SLR

Understanding the usefulness of models and the simple linear regression model

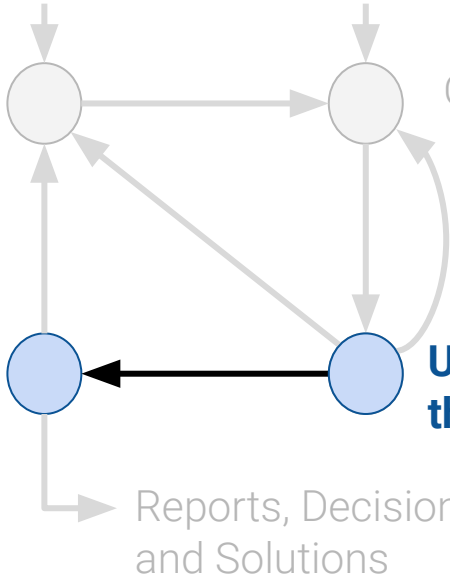
Data 100/Data 200, Spring 2023 @ UC Berkeley

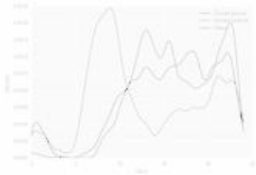
Narges Norouzi and Lisa Yan

Content credit: Narges Norouzi, Will Fithian, Lisa Yan, Suraj Rampure, Ani Adhikari, Deborah Nolan, Joseph Gonzalez

Plan for Next Few Lectures: Predictive modeling

1357201





(today)

Modeling I: Intro to Modeling, Simple Linear Regression

Modeling II: Different models, loss functions, linearization

Modeling III: Multiple Linear Regression

1357201

Today's Roadmap

Lecture 10, Data 100 Spring 2023

What Is A Model?

Data 8 Review: Regression Line, Correlation

The Modeling Process: Definitions

Loss Functions

Minimizing Average Loss (Empirical Risk)

Interpreting SLR: Slope, Anscombe's Quartet

Evaluating the Model: RMSE, Residual Plot

1357201

Last time: Population vs. Sample

- The **population** is a set of individuals (people / bacteria / villages) in the **real world** that we are interested in learning about.
- The **sample** is a (usually smaller) subset of data we can actually collect & analyze.
- If we pick a good sampling scheme, the sample is **representative** of the population (in practice that is often hard to accomplish!).
- If so, we can hope that patterns we find in the data will reflect patterns in the wider world (but **how closely?** Question of **inference** which we will discuss later).

Next few lectures: predictive modeling

- General setting: we observe **pairs** $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
- Assume they are representative of some underlying population of interest.
- **Question (prediction):** If we see x, what is the best prediction for y?
- **Question (association):** if x changes, how much bigger or smaller is y (on avg)?
- **Question (causation):** if we make x one unit larger, how much larger will y get (on avg)?

What Is A Model?

Lecture 10, Data 100 Spring 2023

What Is A Model?

Data 8 Review: Simple Linear Regression and Correlation

The Modeling Process: Definitions

Loss Functions

Minimizing Average Loss on Data

Interpreting SLR: Slope

Evaluating the Model: RMSE, Residual Plot

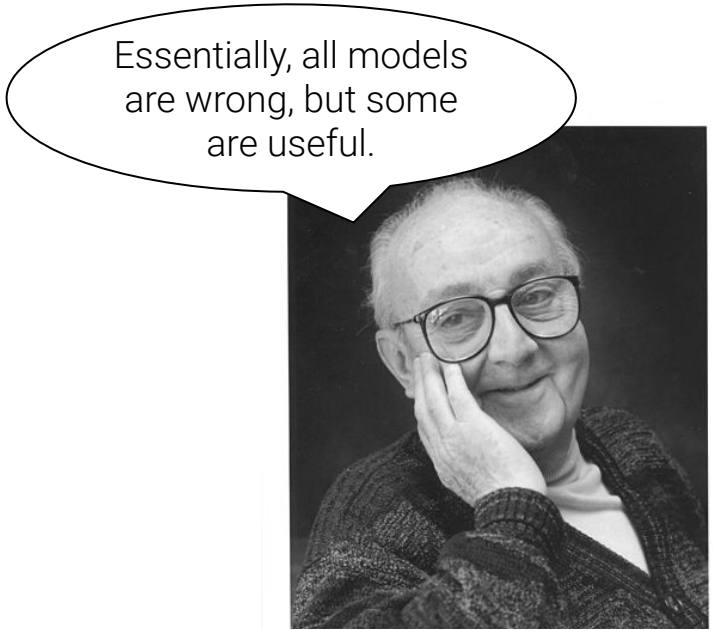
What Is A Model?

1357201

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/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

Two Reasons for Building Models

1357201

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 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.

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

Common Types of Models

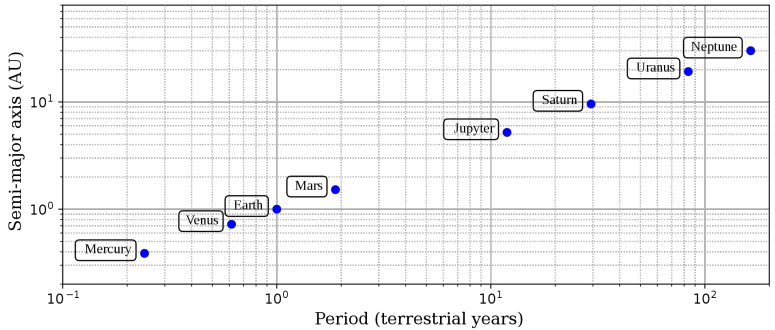
1357201

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

Kepler's Third Law of Planetary Motion (1619)

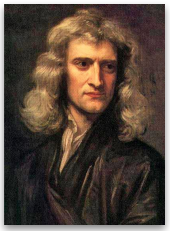
The ratio of the square of an object's orbital period with the cube of the semi-major axis of its orbit is the same for all objects orbiting the same primary.

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



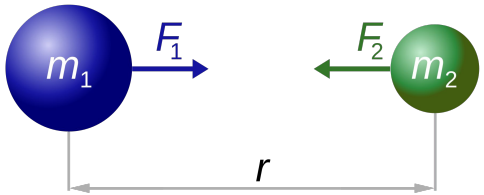
Newton's Laws: motion and gravitation (1687)

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



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

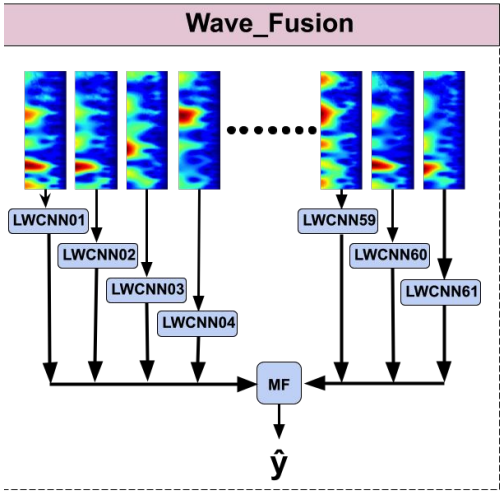
$$F = G\frac{m_1 m_2}{r^2}$$

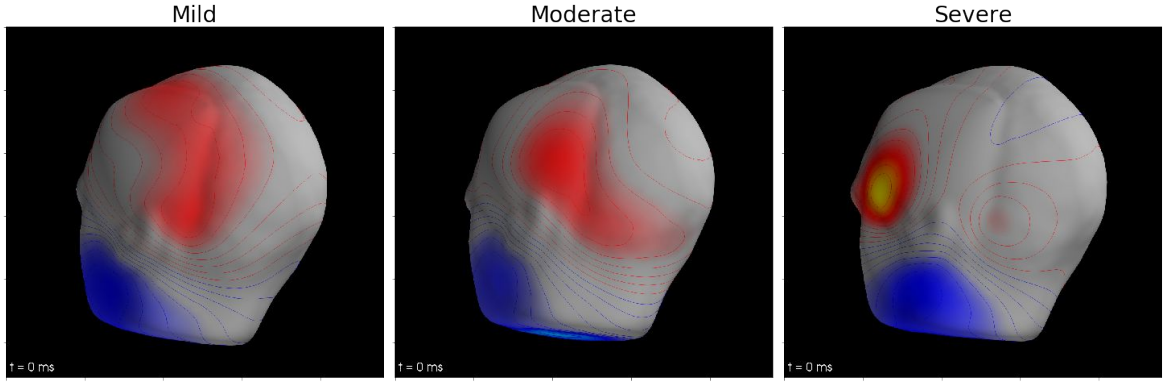


1357201

Probabilistic models

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





1357201

Data 8 Review: Simple Linear Regression & Correlation

Lecture 10, Data 100 Spring 2023

[Data 8 Review] The Regression Line

1357201

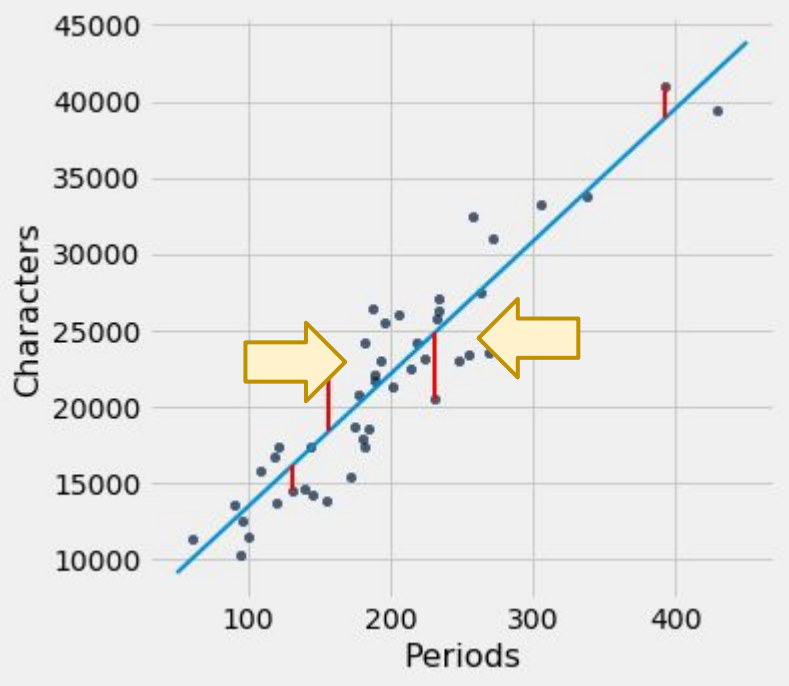
From Data 8 (textbook):

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

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

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

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



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

[Data 8 Review] The Regression Line

1357201

From Data 8 (textbook):

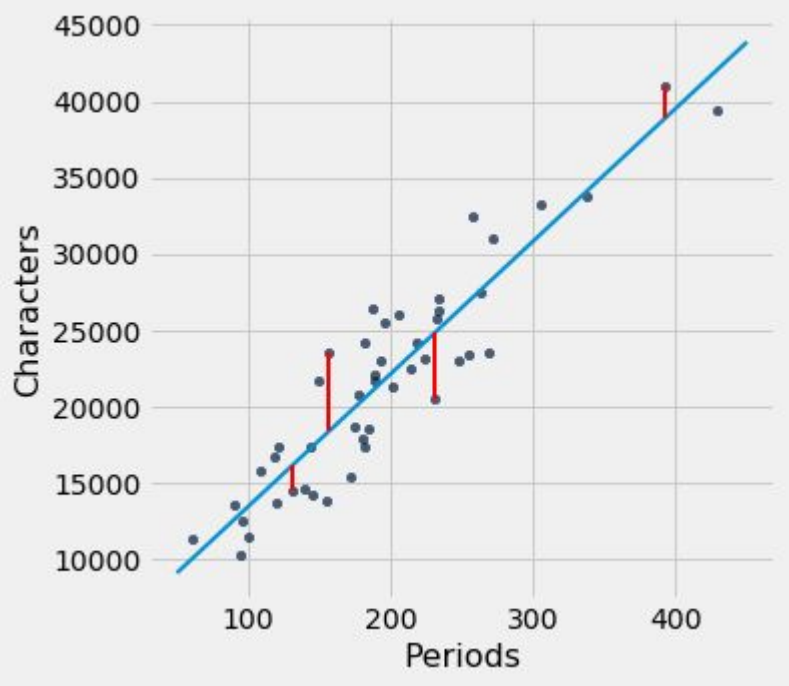
> The **regression line** is the unique straight line that minimizes the **mean squared error** of estimation among all straight lines.

correlation

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

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

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



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

[Data 8 Review] Correlation

1357201

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:

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

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

[Data 8 Review] Correlation

1357201

From Data 8 (textbook):

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

Define the following:

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

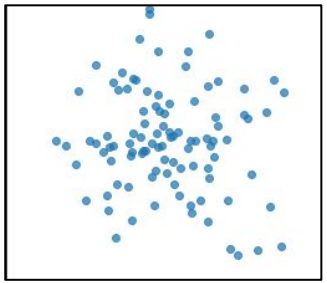
$\bar{x}, \bar{y}$ means; σ_x, σ_y standard deviations

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

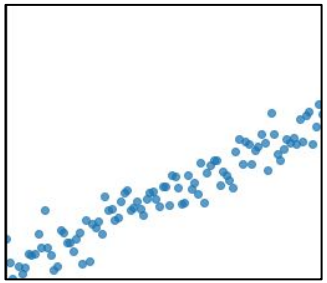
- x_i in standard units: $\frac{x_i - \bar{x}}{\sigma_x}$
- r is also known as Pearson's correlation coefficient.
- Side note: **covariance** is $\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = r \sigma_x \sigma_y$

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

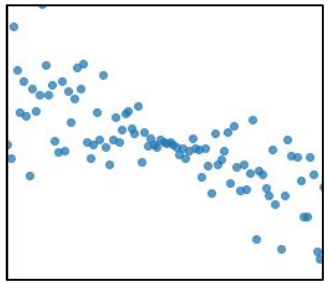
$|r| < 1$



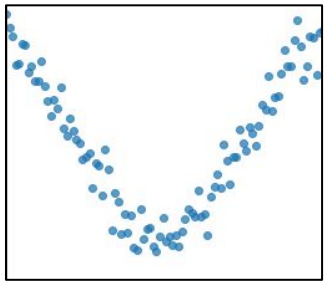
r = −0.121



r = 0.951



r = −0.723



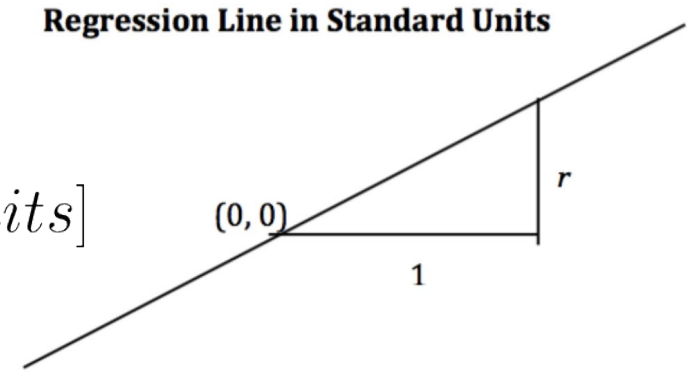
⚠ **r = 0.056**

[Data 8 Review] The Regression Line

1357201

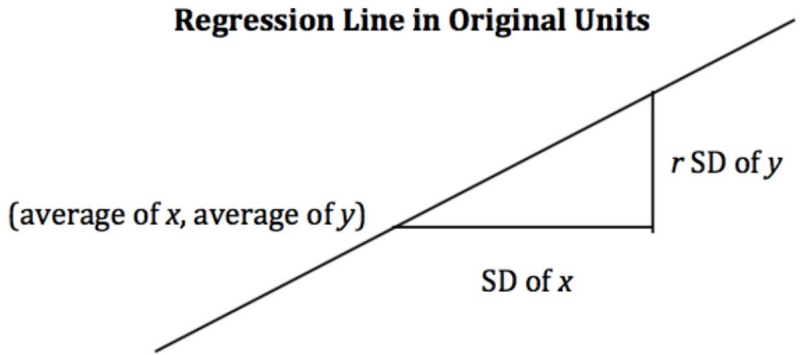
- When the variables x and y are measured in standard units, the regression line for predicting y based on x has slope r passes through the origin and the equation will be:

$$\hat{y} = r \times x \; [both \; measured \; in \; standard \; units]$$



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

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



[Data 8 Review] The Regression Line

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

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

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

Recall regression line equation is defined as:

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

Goal: Derive and define everything on this slide!

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

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

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

1357201

The Modeling Process: Definitions

Lecture 10, Data 100 Spring 2023

1357201

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**: it is described by a few **parameters** (in this case θ_0, θ_1)

- No one tells us the parameters: the data informs us about them.
- The x values are **not** parameters because we directly observe them.
- Sample-based **estimate** of θ_0, θ_1 written as $\hat{\theta}_0, \hat{\theta}_1$.

Usually, we pick the parameters that appear "best" according to some criterion we choose

- Usually standing in as a proxy for fit to new data.

Parametric Model Notation

y True outputs

$\hat{y}$ Predicted outputs

For data:

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

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).

Model parameter(s)

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

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

Estimated parameter(s), “best” fit to data in some sense

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$$

The "best" fitting linear model with parameters $\hat{\theta} = [\hat{\theta}_0, \hat{\theta}_1]$

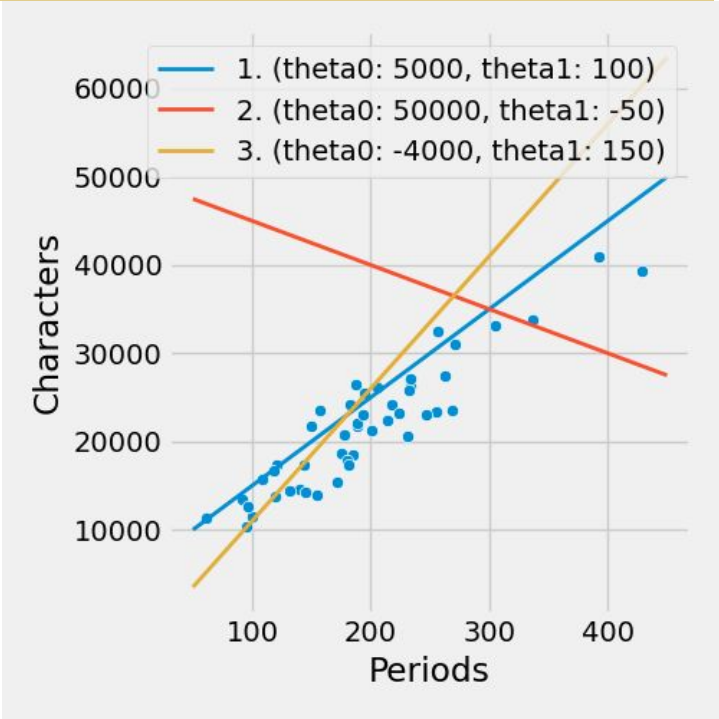
Which is best?

1357201

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

1357201

Which of the lines matches the data better?

ⓘ Start presenting to display the poll results on this slide.

Simple Linear Regression: Our First Model

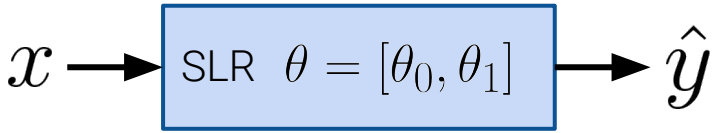
1357201

Simple Linear Regression Model (SLR)

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

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.



- 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

1357201

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 10, Data 100 Spring 2023

The Modeling Process

1357201

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

1357201

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

L2 Loss or 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*

L1 Loss or Absolute Loss

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

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

L2 and L1 Loss for SLR

1357201

Squared Loss (L2 Loss)

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

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

$$L(y, \hat{y}) = (y - (\theta_0 + \theta_1 x))^2$$

Absolute Loss (L1 Loss)

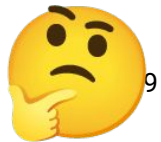
$$L(y, \hat{y}) = |y - \hat{y}|$$

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

1. What is the SLR L1 Loss?

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

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



L2 and L1 Loss for SLR

1357201

Squared Loss (L2 Loss)

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

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

$$L(y, \hat{y}) = (y - (\theta_0 + \theta_1 x))^2$$

Absolute Loss (L1 Loss)

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

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

1 $L(y, \hat{y}) = |y - (\theta_0 + \theta_1 x)|$

L2 and L1 Loss for SLR

1357201

Squared Loss (L2 Loss)

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

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

$$L(y, \hat{y}) = (y - (\theta_0 + \theta_1 x))^2$$

Absolute Loss (L1 Loss)

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

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

1 $L(y, \hat{y}) = |y - (\theta_0 + \theta_1 x)|$

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

slido

1357201

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

ⓘ Start presenting to display the poll results on this slide.

L2 and L1 Loss for SLR

1357201

Squared Loss (L2 Loss)

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

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

$$L(y, \hat{y}) = (y - (\theta_0 + \theta_1 x))^2$$

Absolute Loss (L1 Loss)

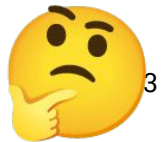
$$L(y, \hat{y}) = |y - \hat{y}|$$

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

1 $L(y, \hat{y}) = |y - (\theta_0 + \theta_1 x)|$

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

- Doesn't work: big negative residuals shouldn't cancel out big positive residuals!



L2 and L1 Loss for SLR

1357201

Squared Loss (L2 Loss)

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

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

$$L(y, \hat{y}) = (y - (\theta_0 + \theta_1 x))^2$$

Absolute Loss (L1 Loss)

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

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

1 $L(y, \hat{y}) = |y - (\theta_0 + \theta_1 x)|$

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

- Doesn't work: big negative residuals shouldn't cancel out big positive residuals!

Which loss function is better: L1 or L2? 3

slido

1357201

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

ⓘ Start presenting to display the poll results on this slide.

L2 and L1 Loss for SLR

1357201

Squared Loss (L2 Loss)

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

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

$$L(y, \hat{y}) = (y - (\theta_0 + \theta_1 x))^2$$

Absolute Loss (L1 Loss)

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

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

1 $L(y, \hat{y}) = |y - (\theta_0 + \theta_1 x)|$

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

- Doesn't work: big negative residuals shouldn't cancel out big positive residuals!

Which loss function is better: L1 or L2? L2 penalizes larger residuals more. 3

Empirical Risk is Average Loss over Data

1357201

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 it fits the data (not the population).

But hopefully these are close.

Empirical Risk is Average Loss over Data

1357201

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)**

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

The Modeling Process

1357201

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

1357201

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?

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

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?

1357201

Minimizing Average Loss on Data

Lecture 10, Data 100 Spring 2023

Minimizing MSE for the SLR Model

1357201

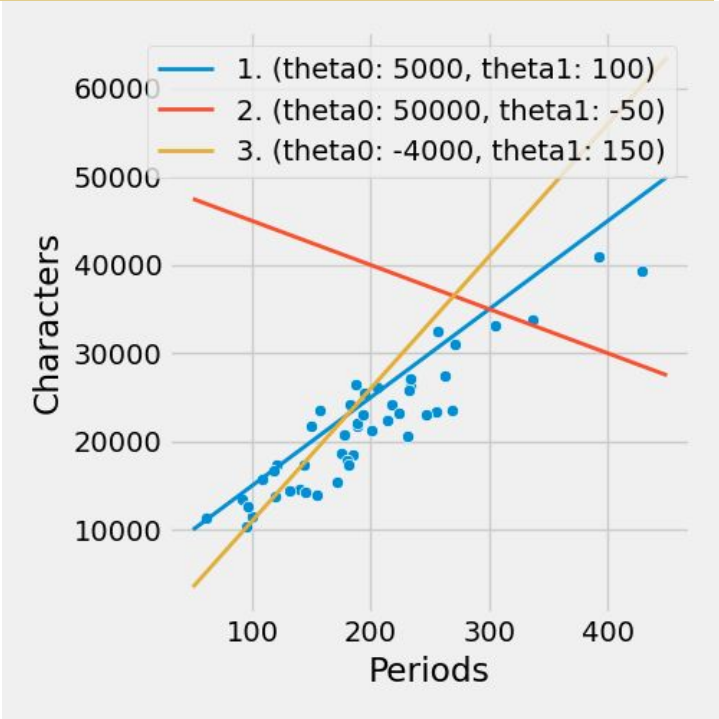
Recall: we wanted to pick the **regression line**

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$$

To minimize the (sample) **Mean Squared Error**:

$$\begin{aligned}\hat{R}(\theta) &= \frac{1}{n}\sum_i L(y_i, \hat{y}_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1 x_i))^2\end{aligned}$$

To find the best values, we **take derivatives** with respect to the choice variables θ_0, θ_1



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

Minimizing MSE for the SLR Model

1357201

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

1357201

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

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

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

Minimizing MSE for the SLR Model

1357201

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:**

$$0 = \frac{\partial}{\partial \theta_0} MSE = -\frac{2}{n}\sum_{i=1}^{n} y_i - \theta_0 - \theta_1 x_i \iff \frac{1}{n}\sum_i y_i - \hat{y}_i = 0$$ **1**

"Equivalent"

$$0 = \frac{\partial}{\partial \theta_1} MSE = -\frac{2}{n}\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i)x_i \iff \frac{1}{n}\sum_i (y_i - \hat{y}_i)x_i = 0$$ **2**

To find the best θ_0, θ_1, we need to solve the **estimating equations** on the right.

From Estimating Equations to Estimators

1357201

Goal: Choose θ_0, θ_1 to solve two estimating equations:

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

1

$$\frac{1}{n}\sum_i (y_i - \theta_0 - \theta_1 x_i) = 0 \iff (\frac{1}{n}\sum_i y_i) - \theta_0 - \theta_1(\frac{1}{n}\sum_i x_i) = 0$$

$$\iff \bar{y} - \theta_0 - \theta_1 \bar{x} = 0$$

$$\iff \theta_0 = \bar{y} - \theta_1 \bar{x}$$

From Estimating Equations to Estimators

1357201

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: $\boxed{2}$ - $\boxed{1}$ * $\bar{x}$

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

$$(using\ \hat{y}_i = \theta_0 + \theta_1 x_i) \Rightarrow \frac{1}{n}\sum_i (y_i - \theta_0 - \theta_1 x_i)(x_i - \bar{x}) = 0$$

$$(using\ \theta_0 = \bar{y} - \theta_1 \bar{x}) \Rightarrow \frac{1}{n}\sum_i (y_i - \bar{y} + \theta_1\bar{x} - \theta_1 x_i)(x_i - \bar{x}) = 0$$

$$\Rightarrow \frac{1}{n}\sum_i (y_i - \bar{y} - \theta_1(x_i - \bar{x}))(x_i - \bar{x}) = 0$$

From Estimating Equations to Estimators

$$\Rightarrow \frac{1}{n}\sum_i (y_i - \bar{y} - \theta_1(x_i - \bar{x}))(x_i - \bar{x}) = 0$$

$$\Rightarrow \frac{1}{n}\sum_i (y_i - \bar{y})(x_i - \bar{x}) = \theta_1 \frac{1}{n}\sum_i (x_i - \bar{x})^2$$

Plug in definitions of correlation and SD:

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

$$r\sigma_x\sigma_y = \theta_1\sigma_x^2$$

Solve for θ_1:

$$\theta_1 = r\frac{\sigma_y}{\sigma_x}$$

Estimating Equations

1357201

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 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

1357201

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?

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

1357201

Interpreting SLR: Slope

Lecture 10, Data 100 Spring 2023

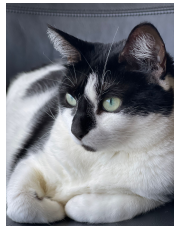
Interpreting the Least Squares Linear Regression Model

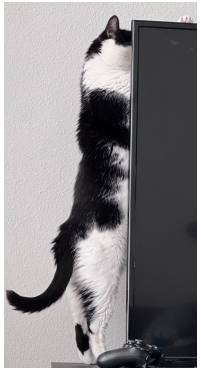
You may sometimes hear the prediction task defined as: "**regressing** y on x."

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

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

```
predicted weight = 2 + 0.5 * length
```





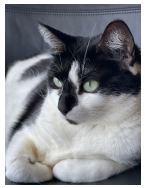
Interpreting the Least Squares Linear Regression Model

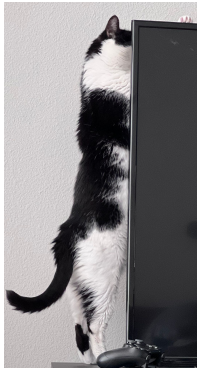
You may sometimes hear the prediction task defined as: "**regressing** y on x."

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

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

```
predicted weight = 2 + 0.5 * length
```





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 cat in the dataset grows 1 inch, we estimate that they will get 0.5 pounds heavier? What does it actually mean?

Interpreting the Least Squares Linear Regression Model

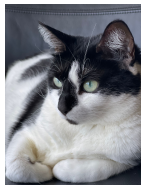
1357201

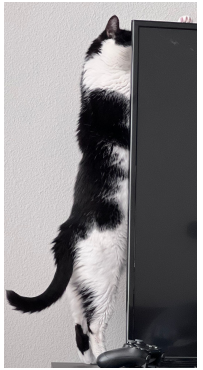
You may sometimes hear the prediction task defined as: "**regressing** y on x."

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

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

```
predicted weight = 2 + 0.5 * length
```





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 cat in the dataset grows 1 inch, we estimate that they will get 0.5 pounds heavier? What does it actually mean? **No!**

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

Slope interpretation: If two cats have a 1 inch height difference, their estimated weight difference is 0.5 lbs.

Interpreting the Least Squares Linear Regression Model

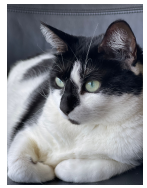
1357201

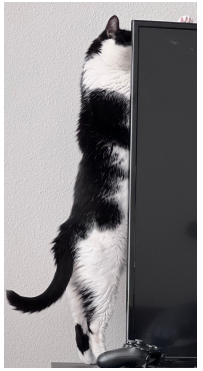
You may sometimes hear the prediction task defined as: "**regressing** y on x."

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

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

```
predicted weight = 2 + 0.5 * length
```





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 cat in the dataset grows 1 inch, we estimate that they will get 0.5 pounds heavier? What does it actually mean? **No!**

Predicting on wildly different data?

Domestic shorthair range from 8-10 pounds, and 13-16 inches in length.
Maine Coon range from 10-25 pounds, and 19-40 inches in length.

2. Should we use this model to predict the weight of all cat breeds? **No!**

Evaluating the Model: RMSE, Residual Plot

Lecture 10, Data 100 Spring 2023

Evaluating Models

1357201

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 $e_i = y_i - \hat{y}_i$ visualize the difference between actual and predicted y values.

Demo Slides

Four Mysterious Datasets + Least Squares

1357201

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!!!!