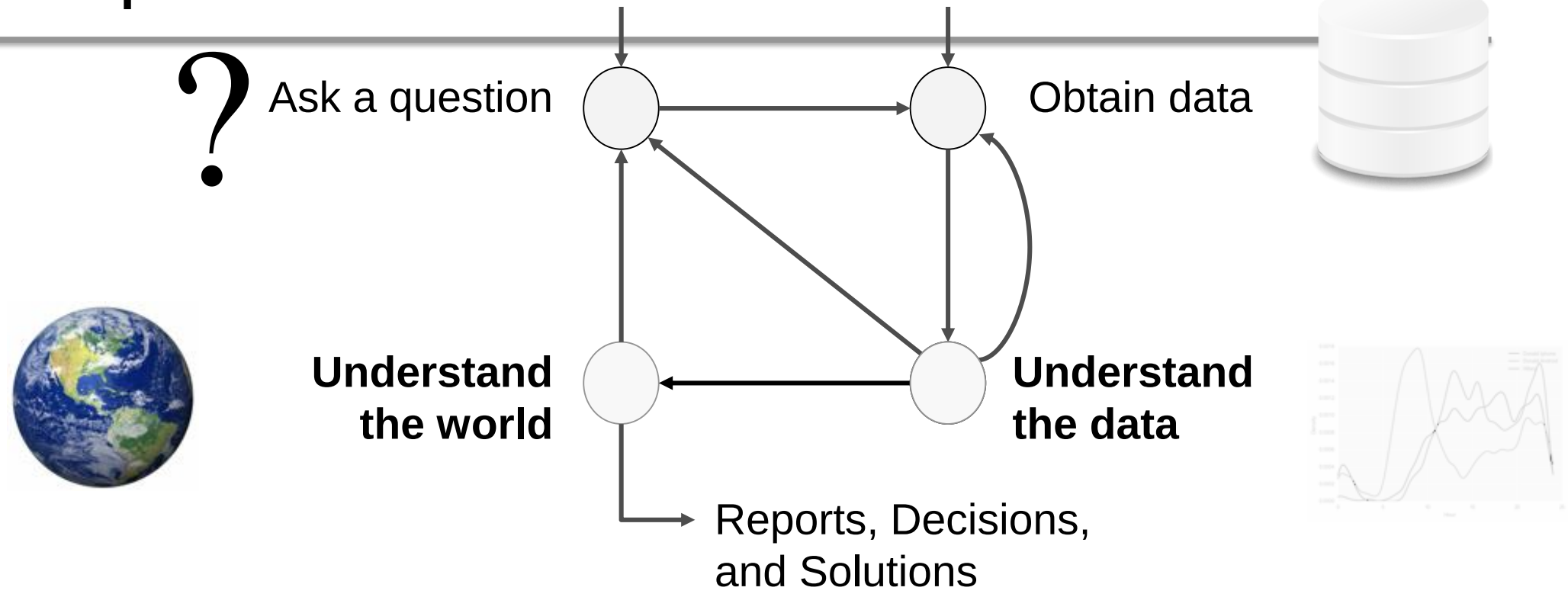
From:
DS100 UC Berkeley

Modeling and Simple Linear Regression

Introduction to Data Science
Spring 1404

Yadollah Yaghoobzadeh

Our plan



(today)

Modeling I:
Intro to Modeling, Simple Linear Regression, Loss functions

Introduction to Modeling

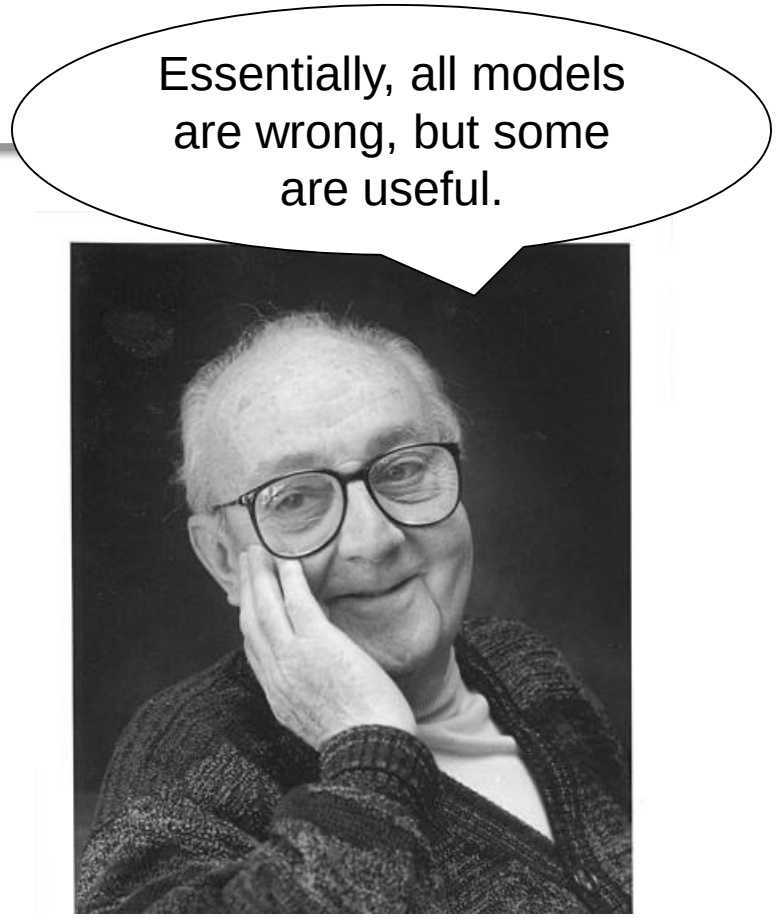
Understanding the usefulness of models

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/s² due to gravity.

- While this describes the behavior of our system, it is merely an approximation.
- It doesn't account for the effects of air resistance, local variations in gravity, etc.
- But in practice, it's accurate enough to be useful!



George Box, Statistician (1919-2013)

Known for	"All models are wrong" Response-surface methodology EVOP q-exponential distribution Box–Jenkins method Box–Cox transformation

Three Reasons for Building Models

Reason 1:

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

- How are the parents' average heights related to the children's average heights?
- How do an object's velocity and acceleration impact how far it travels?

Often times, we care about creating models that are **simple and interpretable**, allowing us to understand what the relationships between our variables are.

Reason 2:

To make **accurate predictions** about unseen data.

- Can we predict if an email is spam or not?
- Can we generate a one-sentence summary of this 10-page long article?

Other times, we care more about making extremely accurate predictions, at the cost of having an uninterpretable model. These are sometimes called **black-box models**, and are common in fields like deep learning.

Reason 3:

To make **causal inferences** about if one thing causes another thing.

- Can we conclude that smoking *causes* lung cancer?
- Does a job training program cause increases in employment and wage?

Much harder question because most statistical tools are designed to infer association not causation

This won't be the focus of this class, but will be if you go on to take more advanced classes.

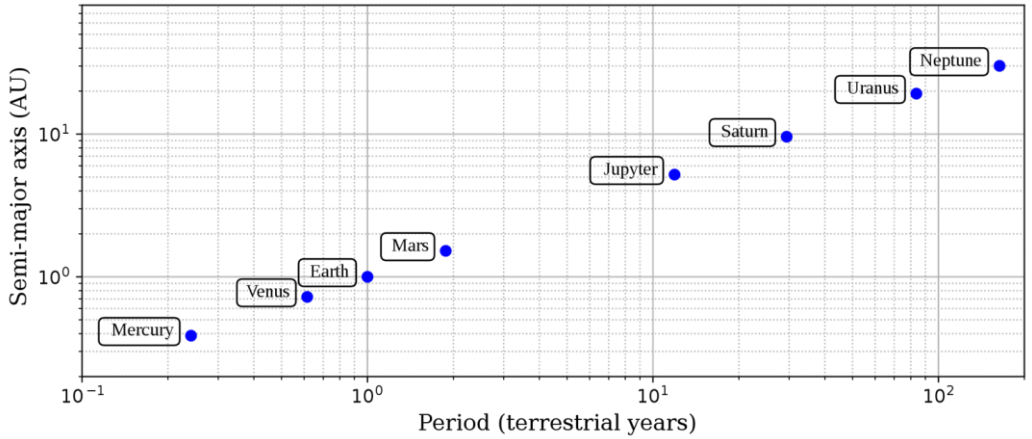
Most of the time, we want to strike a balance between **interpretability** and **accuracy**.

Common Types of Models

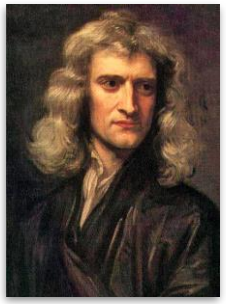
Deterministic physical (mechanistic) models:
Laws that govern how the world works.

Kepler's Third Law of Planetary Motion (1619)
The ratio of the square of an object's orbital period with the cube of the semi-major axis of its orbit is the same for all objects orbiting the same primary.

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

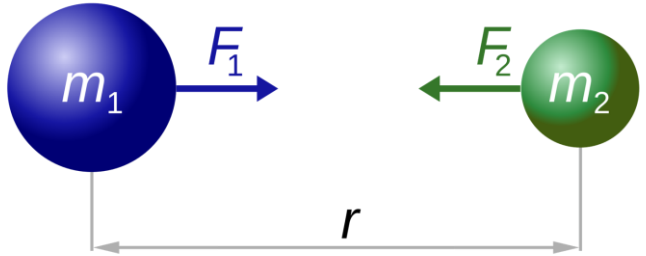


Newton's Laws: motion and gravitation (1687)
Newton's second law of motion models the relationship between the mass of an object and the force required to accelerate it.



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

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



Common Types of Models

Probabilistic models

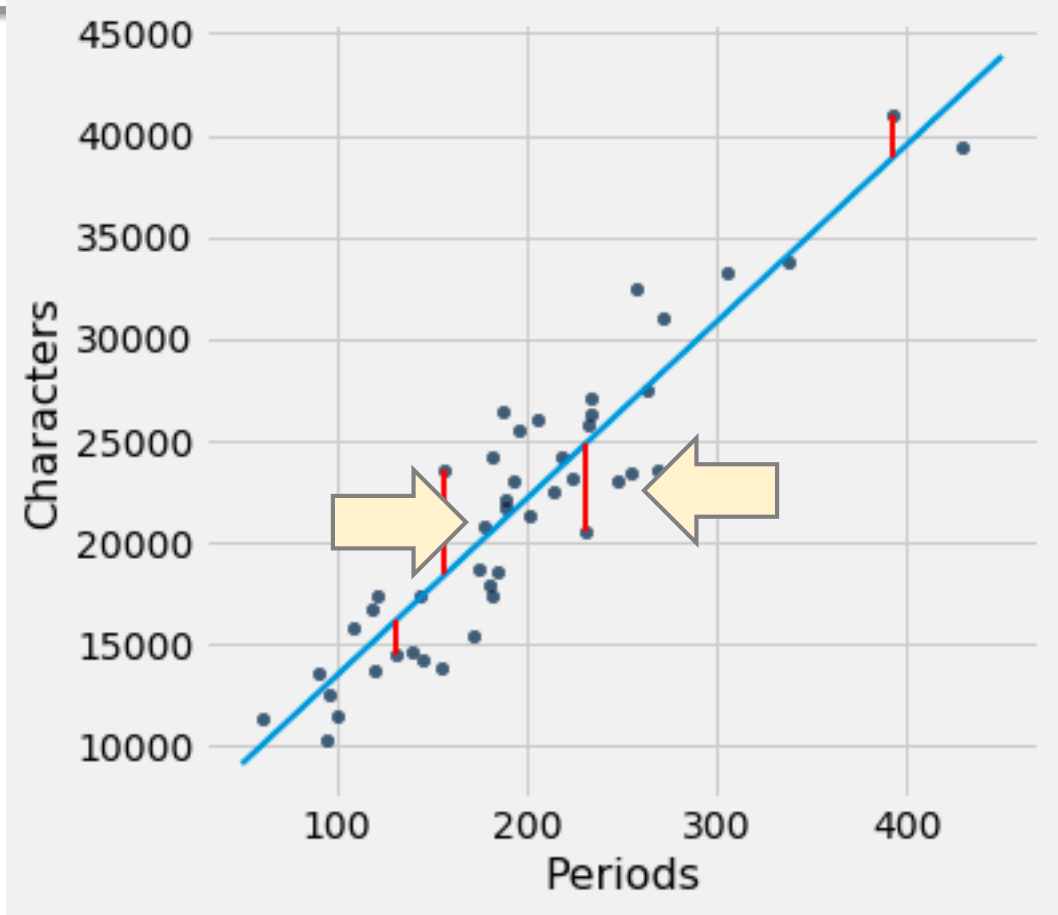
- Models of how random processes evolve.
- Often motivated by understanding of an unpredictable system.

- Examples: Bayesian networks, Markov modelso

Review: Regression Line & Correlation

The Regression Line

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



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

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

Models in DS

In DS, we'll treat a model as some mathematical rule to describe the relationships between variables.

Dataset

x	y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

Observation

(x_i, y_i)

- Independent variable
- **Input**
- **Feature**
- **Attribute**

- Dependent variable
- **Output**
- **Outcome**
- **Response**

Prediction

If we use x to predict y, the predictions are denoted as

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

Models

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

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

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

→ Parametric models

Models in DS: Parametric Models

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 θ is written as $\hat{\theta}$.
- Usually, we pick the parameters that appear "**best**" according to some criterion we choose.

θ Model parameter(s)

$\hat{\theta}$ Estimated parameter(s), "best" fit to data in some sense

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

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

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

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

Models in DS: Parametric Models

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

- No one tells us the parameters: the data informs us about them.
- The x, y values are **not** parameters because we directly observe them.
- Sample-based **estimate** of parameter θ is written as $\hat{\theta}$.
- Usually, we pick the parameters that appear "**best**" according to some criterion we choose

Note: Not all statistical models have parameters!

k-Nearest Neighbor classifiers are non-parametric models.

θ Model paramet… linear model with parameters $\theta = [\theta_0, \theta_1]$

$\hat{\theta}$ Estimated parameter(s), "best" fit to data in some sense $\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$ The "best" fitting linear model with parameters $\hat{\theta} = [\hat{\theta}_0, \hat{\theta}_1]$

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?

The Modeling Process

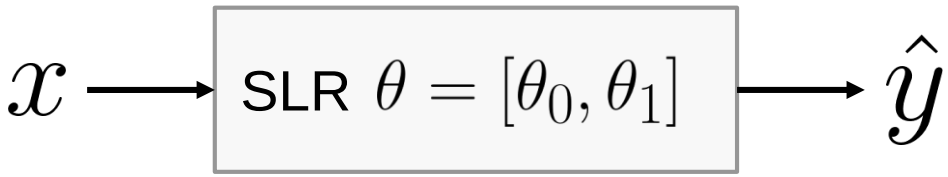
1. Choose a model	**How should we represent the world?**
2. Choose a loss function	How do we quantify prediction error?
3. Fit the model	How do we choose the best parameters of our model given our data?
4. Evaluate model performance	How do we evaluate whether this process gave rise to a good model?

Simple Linear Regression: Our First Model

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

Simple Linear Regression Model (SLR)

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



- 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.

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

The Modeling Process

1. Choose a model

How should we represent the world?

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

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

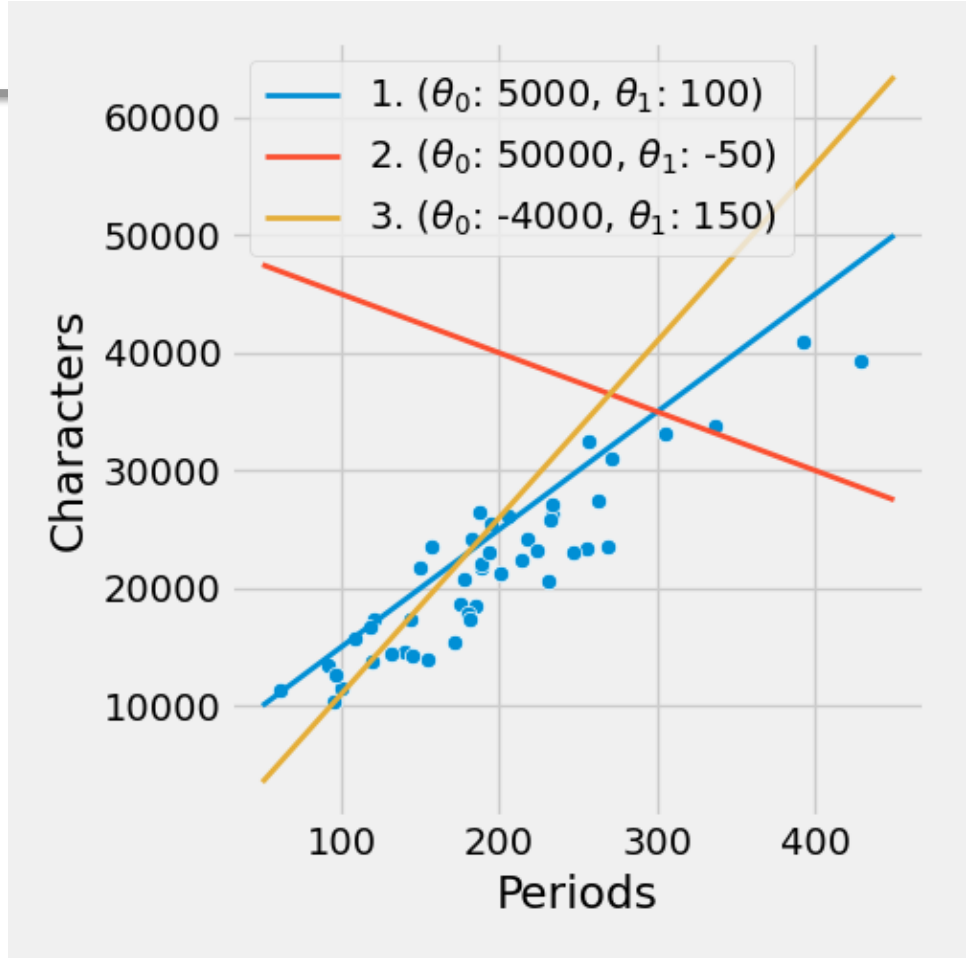
How do we evaluate whether this process gave rise to a good model?

Which θ 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.

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. $L(y, \hat{y})$
- 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**.

- There are many definitions of loss functions!
- The choice of loss function:
  - Affects the accuracy and computational cost of estimation.
  - Depends on the estimation task:
    - Are outputs quantitative or qualitative?
    - Do we care about outliers?
    - Are all errors equally costly? (e.g., false negative on cancer test)

L2 and L1 Loss

Squared Loss

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

- Widely used.
- Also called "L2 loss".
- Reasonable:
  - $\hat{y} = y$ → good prediction → good fit → no loss
  - $\hat{y}$ far from y → bad prediction → bad fit → *lots of loss*

Absolute Loss

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

- 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*

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

Residuals as loss function?

- Why don't we directly use residual error as the loss function?

$$e = (y - \hat{y})$$

- Doesn't work: big negative residuals shouldn't cancel out big positive residuals!
  - Our predictions can be very off, but we can still get a zero residual.

Empirical Risk is Average Loss over Data

We care about how bad our model's predictions are for our entire data set, not just for one point.

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

A natural measure, then, is of the **average loss** (aka **empirical risk**) across all points.

Given data

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

Function of the parameter θ (holding the data fixed) because θ determines $\hat{y}$.

The average loss on the sample tells us how well the model fits the data (not the population).

But hopefully these are close.

Empirical Risk is Average Loss over Data

We care about how bad our model's predictions are for our entire data set, not just for one point. $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

A natural measure, then, is of the **average loss** (aka **empirical risk**) across all points.

Given data

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

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

1. Choose a model

How should we represent the world?

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

2. Choose a loss function

How do we quantify prediction error?

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

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

3. Fit the model

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

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?

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:

$$\frac{\partial}{\partial \theta_0} MSE = 0 \qquad \frac{\partial}{\partial \theta_1} MSE = 0$$

Partial Derivative of MSE with Respect to θ_0, θ_1

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

Derivative of sum is sum of derivatives

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

Chain rule

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

Simplify constants

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

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

Derivative of sum is sum of derivatives

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

Chain rule

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

Simplify constants

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

From Estimating Equations to Estimators

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

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

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

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

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

The Modeling Process

1. Choose a model

How should we represent the world?

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

2. Choose a loss function

How do we quantify prediction error?

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

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

3. Fit the model

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

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?

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

The Modeling Process

1. Choose a model

How should we represent the world?

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

2. Choose a loss function

How do we quantify prediction error?

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

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

3. Fit the model

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

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?

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

Evaluating Models

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

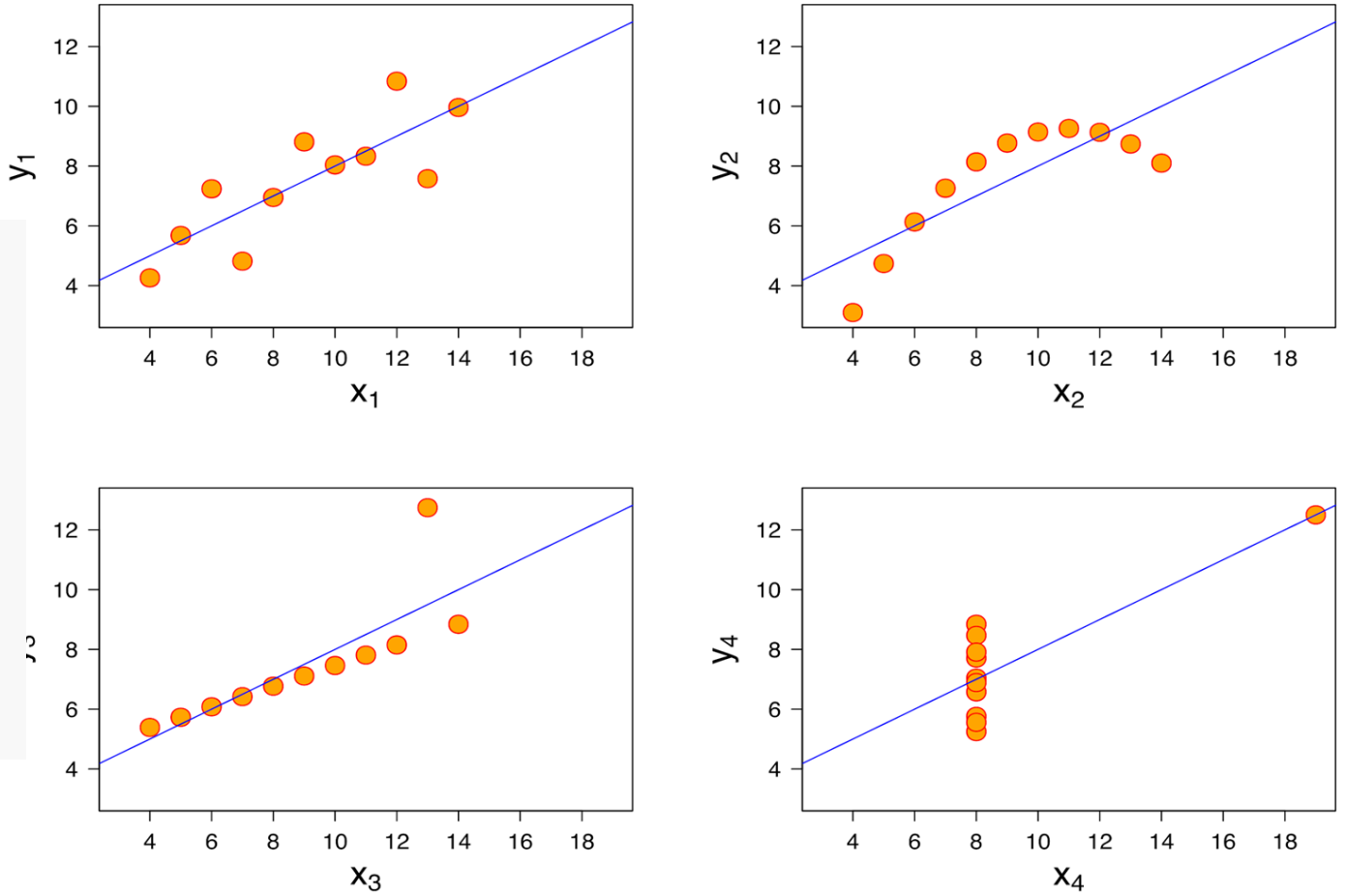
- Visualize data, compute statistics:
  - Plot original data.
    Compute column means, standard deviation.
    If we want to fit a linear model, compute correlation.
- Performance metrics:
  - Root Mean Square Error (RMSE)

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

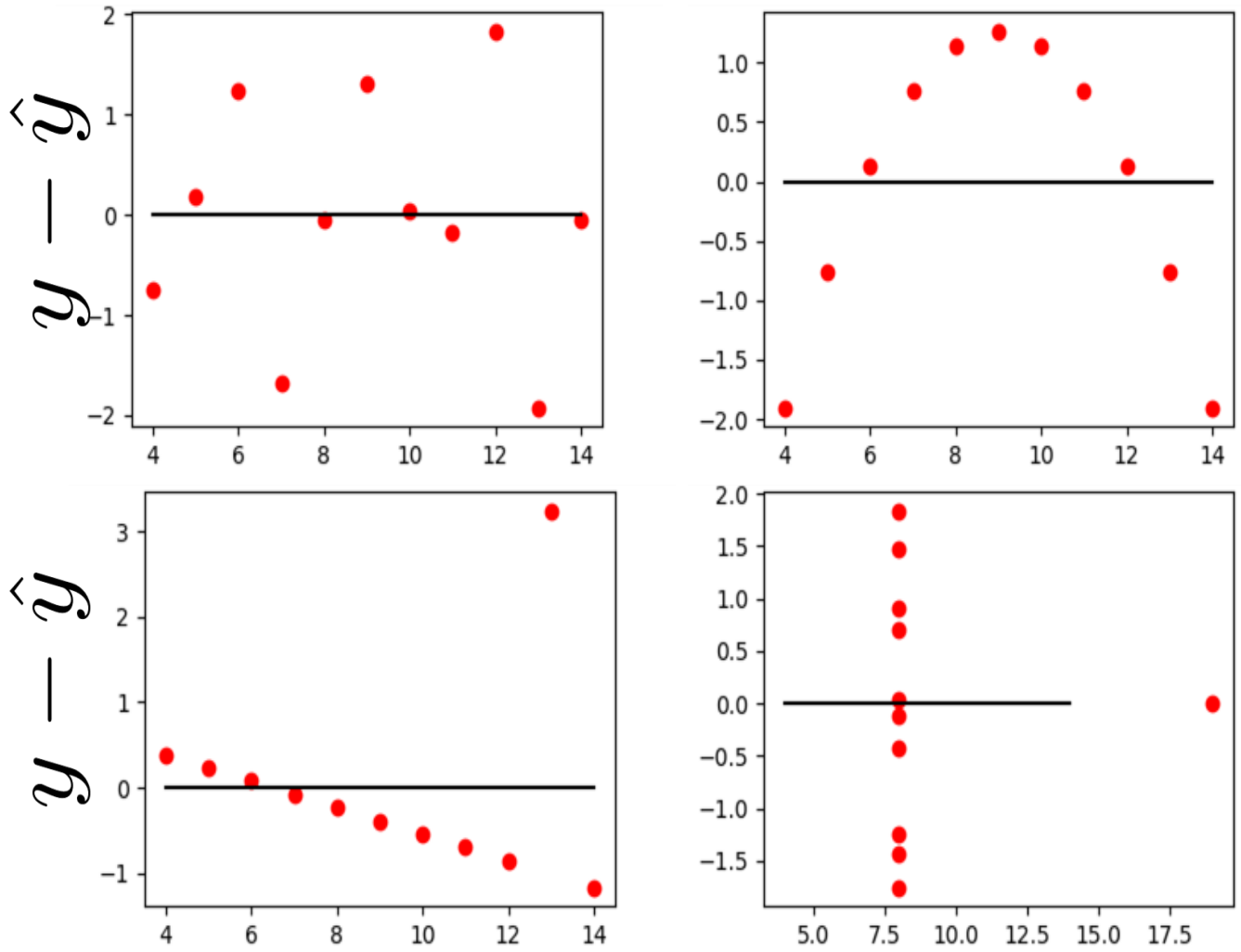
  - It is the square root of MSE, which is the average loss that we've been minimizing to determine optimal model parameters.
  - RMSE is in the same units as y.
  - A lower RMSE indicates more "accurate" predictions (lower "average loss" across data)
- Visualization:
  - Look at a residual plot of $e_i = y_i - \hat{y}_i$ to visualize the difference between actual and predicted values.

Four datasets with the same MSE

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



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



The Modeling Process

1. Choose a model

SLR model

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

2. Choose a loss function

L2 Loss

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

Mean Squared Error (MSE)

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

3. Fit the model

Minimize average loss with calculus

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

4. Evaluate model performance

Visualize, Root MSE

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