LECTURE 12

Introduction to Modeling, Linear Regression

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)

Evaluating the Model

The Regression Line

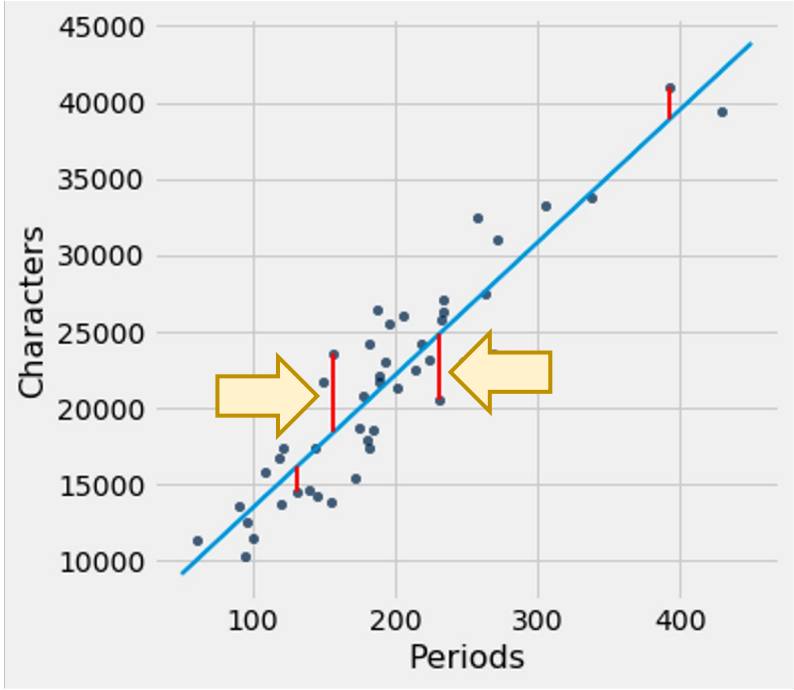
> The **regression line** is the unique straight line that minimizes the **mean squared error** of estimation among all straight lines.

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

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

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

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



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

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).

θ Model parameter(s)

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

$\hat{y} = a + bx$ Any linear model with parameters $\theta = (a, b)$

$\hat{y} = \hat{a} + \hat{b}x$ The "best" linear model with parameters $\hat{\theta} = (\hat{a}, \hat{b})$

The Regression Line

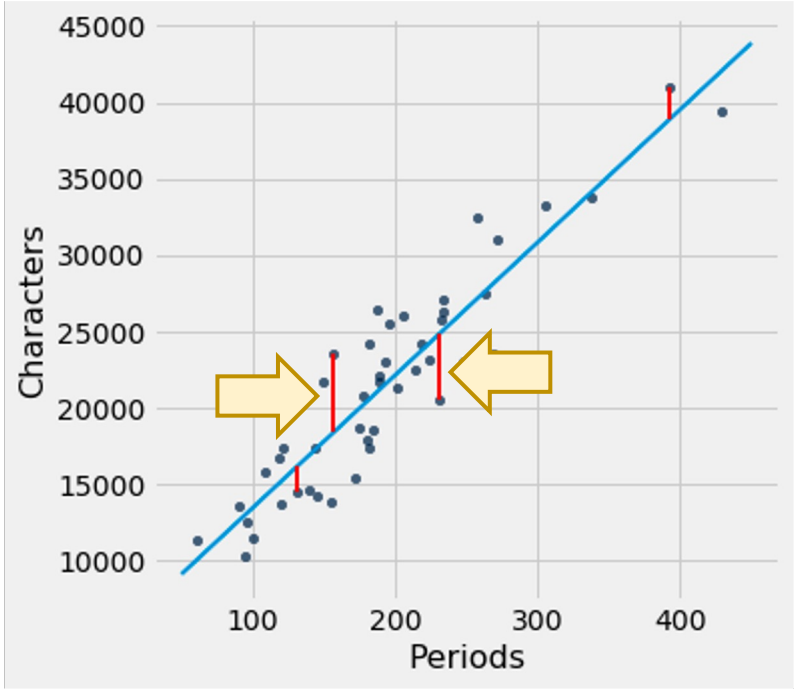
> The **regression line** is the unique straight line that minimizes the **mean squared error** of estimation among all straight lines.

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

slope $\hat{b} = r\dfrac{\sigma_y}{\sigma_x}$ **r: correlation**

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

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



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

Correlation

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

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

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

Correlation

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

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

Define the following:

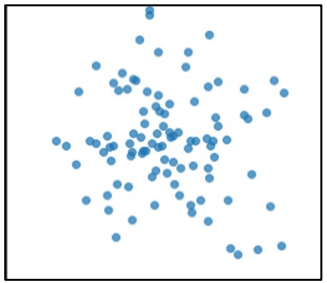
$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ data

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

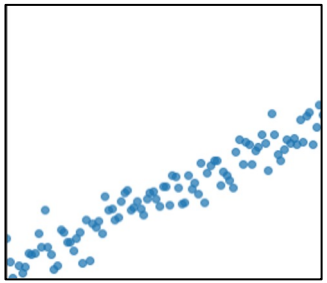
- x_i in standard units: $\frac{x_i - \bar{x}}{\sigma_x}$
- r is also known as Pearson's correlation coefficient.
- Side note: **covariance** is $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = r\sigma_x\sigma_y$

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

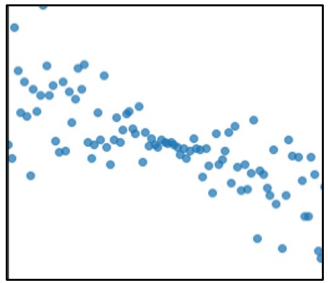
$|r| < 1$



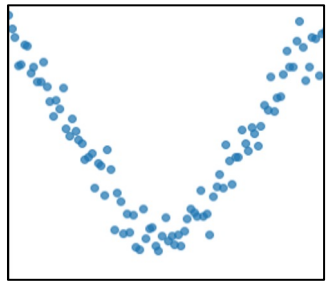
r = −0.121



r = 0.951



r = −0.723



⚠ **r = 0.056**

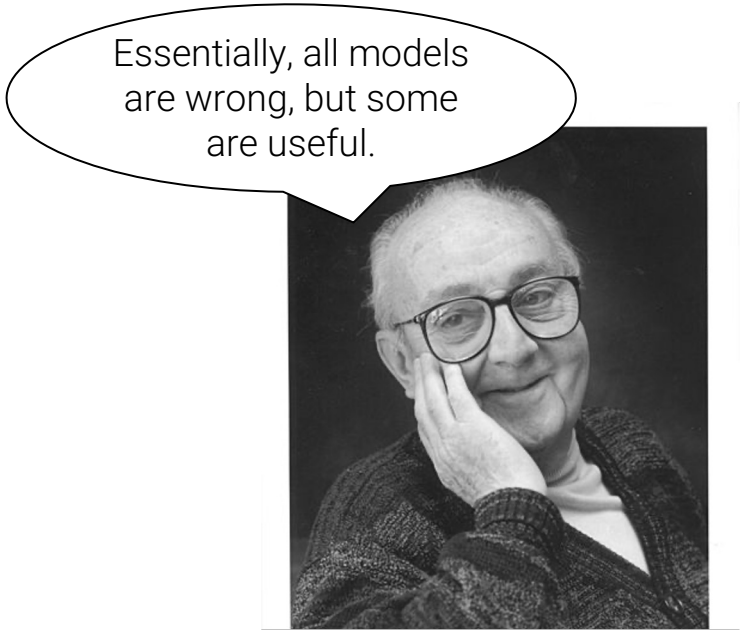
What is a model?

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 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!



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.

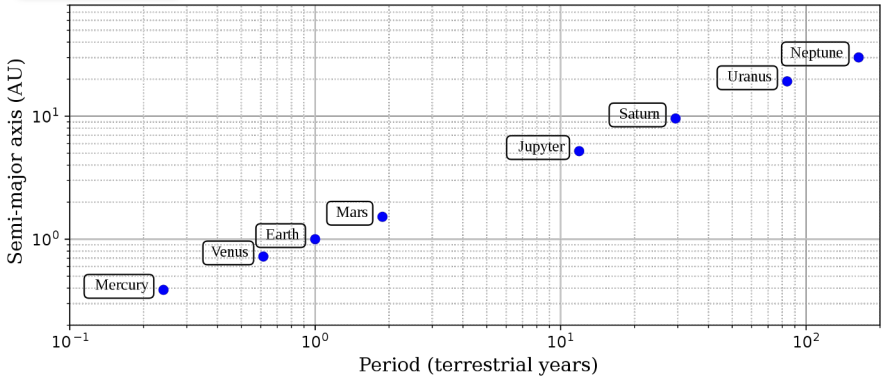
Two common types of models

Physical (mechanistic) models

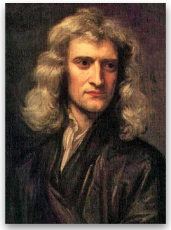
Laws that govern how the world works.

Kepler's Third Law of Planetary Motion (1619)
[Wikipedia]

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

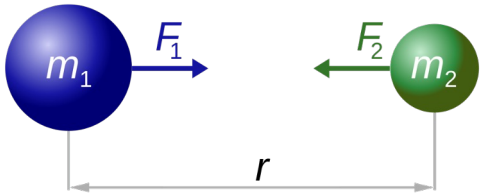


Newton's Laws: motion and gravitation (1687)
[Wikipedia]



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

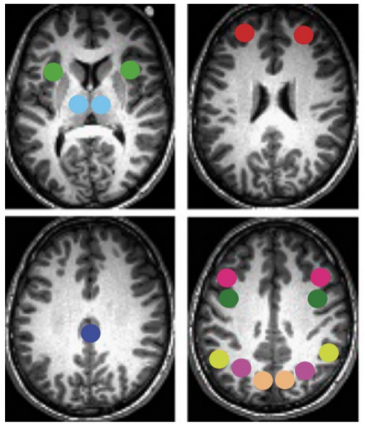
$$F = G\frac{m_1 m_2}{r^2}$$



Two common types of models

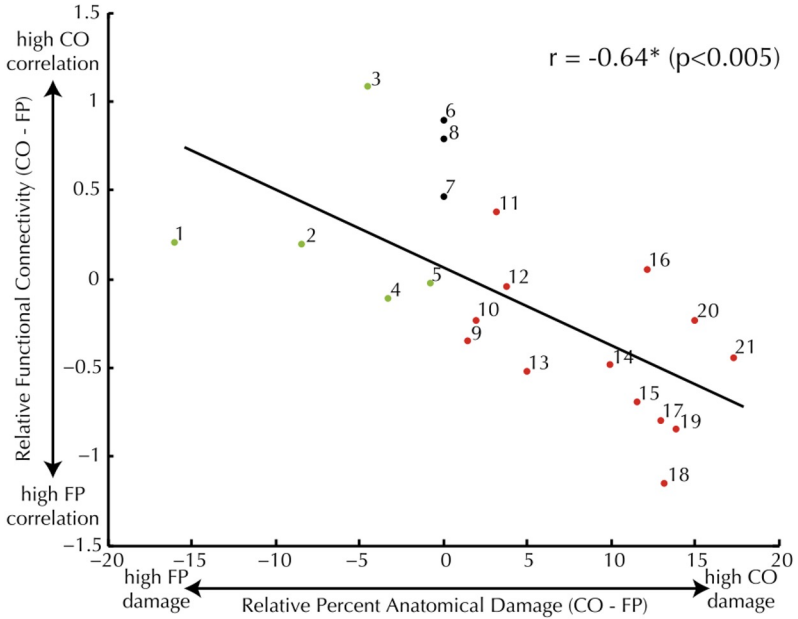
Statistical models

Relationships between variables found through data and statistical analysis.



Cingulo-opercular (CO)
- [1,2] anterior insula/frontal operculum (aI/
- [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

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?

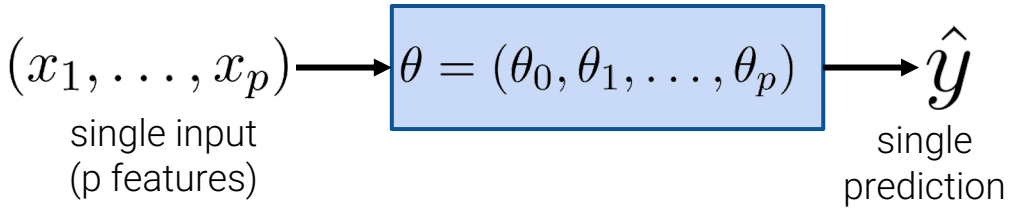
Multiple Linear Regression

Define the **multiple linear regression** model:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

Parameters are $\theta = (\theta_0, \theta_1, \ldots, \theta_p)$.

This is a **linear combination** of θ_j s, each scaled by x_j.



Example: Predict dugong ages $\hat{y}$ as a linear model of 2 features: length x_1 **and** weight x_2 .

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

intercept — parameter for length — parameter for weight

A Linear Algebra Preview

When looking at a **single observation**, our model is

$$\hat{y} = x^T \theta$$

- x is a **vector** of size p + 1.
- $\hat{y}$ is a **scalar**.
- θ is a **vector** of size p + 1.

When looking at **multiple observations**, our model is

$$\hat{\mathbb{Y}} = \mathbb{X}\theta$$

- $\mathbb{X}$ is a **matrix** of size n x (p + 1).
- $\hat{\mathbb{Y}}$ is a **vector** of size n (i.e. $\hat{\mathbb{Y}} \in \mathbb{R}^n$).
- θ is a **vector** of size p + 1.

$$R(\theta) = \frac{1}{n} \underbrace{||\mathbb{Y} - \hat{\mathbb{Y}}||_2^2}_{\text{L2 norm of residual vector}} = \frac{1}{n} ||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

NBA 2018-2019 Dataset

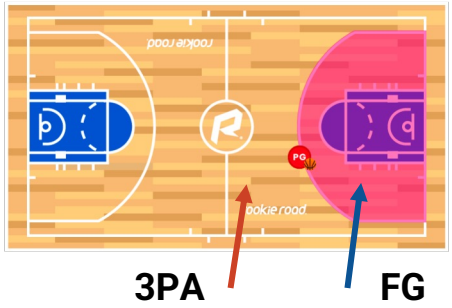
How many points does an athlete score per game?
PTS (average points/game)

To name a few factors:

- **FG**: average # 2 point field goals
- **AST**: average # of assists
- **3PA**: average # 3 point field goals attempted

	FG	AST	3PA	PTS
1	1.8	0.6	4.1	5.3
2	0.4	0.8	1.5	1.7
3	1.1	1.9	2.2	3.2
4	6.0	1.6	0.0	13.9
5	3.4	2.2	0.2	8.9
6	0.6	0.3	1.2	1.7

Rows correspond to individual players.



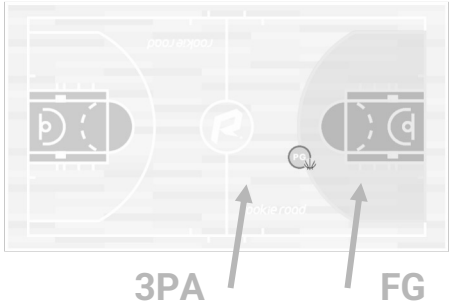
assist: a pass to a teammate that directly leads to a goal

Multiple Linear Regression Model

How many points does an athlete score per game?
PTS (average points/game)

To name a few factors:

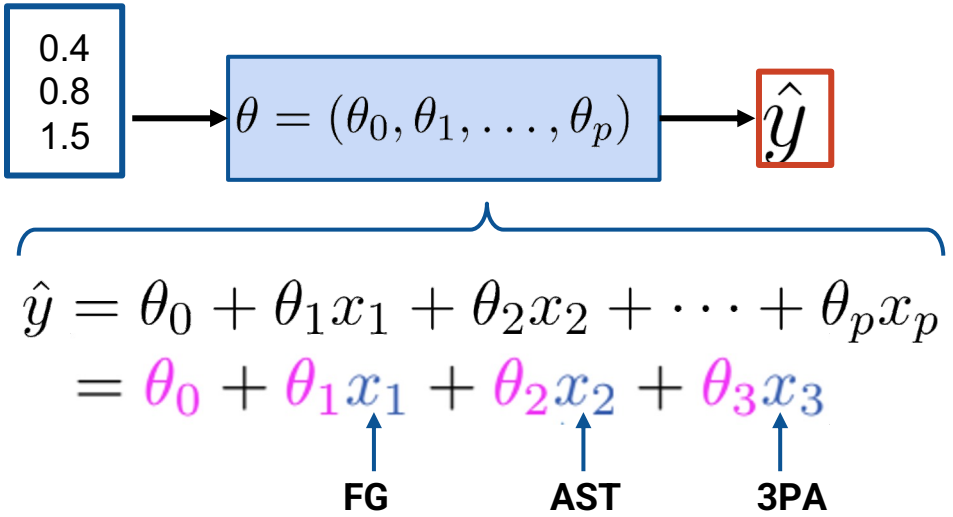
- **FG**: average # 2 point field goals
- **AST**: average # of assists
- **3PA**: average # 3 point field goals attempted



assist: a pass to a teammate that directly leads to a goal

	FG	AST	3PA	PTS
1	1.8	0.6	4.1	5.3
2	0.4	0.8	1.5	1.7
3	1.1	1.9	2.2	3.2
4	6.0	1.6	0.0	13.9
5	3.4	2.2	0.2	8.9
6	0.6	0.3	1.2	1.7

Rows correspond to individual players.



The Modeling Process

1. Choose a model ✅

How should we represent the world?

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

$$\hat{\mathbb{Y}} = \mathbb{X}\theta$$

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?

Vector Notation

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

$$= \theta_0 + \sum_{j=1}^{p} \theta_j x_j$$

$$= x^T \theta \qquad x, \theta \in \mathbb{R}^{(p+1)} : x = \begin{bmatrix} 1 \\ 0.4 \\ 0.8 \\ 1.5 \end{bmatrix} \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0.4 & 0.8 & 1.5 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = \hat{y} \in \mathbb{R}$$

NBA Data

	FG	AST	3PA	PTS
1	1.8	0.6	4.1	5.3
2	0.4	0.8	1.5	1.7
3	1.1	1.9	2.2	3.2
4	6.0	1.6	0.0	13.9
5	3.4	2.2	0.2	8.9
6	0.6	0.3	1.2	1.7

Rows correspond to individual players.

Matrix Notation

Data	FG	AST	3PA	PTS
1	1.8	0.6	4.1	5.3
2	0.4	0.8	1.5	1.7
3	1.1	1.9	2.2	3.2

To make predictions on all n datapoints in our sample:

$$\hat{y}_1 = x_1^T\theta \quad \text{where } x_1^T = \begin{bmatrix} 1 & x_{11} & x_{12} \dots & x_{1p} \end{bmatrix}$$ Datapoint 1

$$\hat{y}_2 = x_2^T\theta \quad \text{where } x_2^T = \begin{bmatrix} 1 & x_{22} & x_{22} & \dots & x_{2p} \end{bmatrix}$$ Datapoint 2

$$\vdots \qquad \vdots \qquad\qquad\qquad \vdots$$

$$\hat{y}_n = x_n^T\theta \quad \text{where } x_n^T = \begin{bmatrix} 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$ Datapoint n

same for all preds $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$

Matrix Notation

Data	FG	AST	3PA	PTS
1	1.8	0.6	4.1	5.3
2	0.4	0.8	1.5	1.7
3	1.1	1.9	2.2	3.2

To make predictions on all n datapoints in our sample:

$$\begin{aligned}
\hat{y}_1 &= \begin{bmatrix} 1 & x_{11} & x_{12} \dots & x_{1p} \end{bmatrix} \theta &&= x_1^T \theta \\
\hat{y}_2 &= \begin{bmatrix} 1 & x_{22} & x_{22} & \dots & x_{2p} \end{bmatrix} \theta &&= x_2^T \theta \\
&\vdots \qquad\qquad \vdots &&\vdots \\
\hat{y}_n &= \begin{bmatrix} 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \theta &&= x_n^T \theta
\end{aligned}$$

same $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$ for all preds

n row vectors, each with dimension **(p+1)**

Expand out each datapoint's (transposed) input

Matrix Notation

To make predictions on all n datapoints in our sample:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \dots & x_{1p} \\ 1 & x_{22} & x_{22} & \dots & x_{2p} \\ & & \vdots & \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \theta$$

n row vectors, each with dimension **(p+1)**

Data	FG	AST	3PA	PTS
1	1.8	0.6	4.1	5.3
2	0.4	0.8	1.5	1.7
3	1.1	1.9	2.2	3.2

same $\theta =$ for all preds $\begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$

Vectorize predictions and parameters to encapsulate all n equations into a single matrix equation.

Matrix Notation

To make predictions on all n datapoints in our sample:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \mathbb{X} \theta$$

Design matrix with dimensions n x (p + 1)

Data	FG	AST	3PA	PTS
1	1.8	0.6	4.1	5.3
2	0.4	0.8	1.5	1.7
3	1.1	1.9	2.2	3.2

same $\theta =$ for all preds $\begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$

The Design Matrix $\mathbb{X}$

We can use linear algebra to represent our predictions of all n datapoints at once.

One step in this process is to stack all of our input features together into a **design matrix**:

$$\mathbb{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & x_{31} & x_{32} & \dots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

What do the **rows** and **columns** of the design matrix represent in terms of the observed data?

Field Goals, Assists, 3-Point Attempts

Bias	FG	AST	3PA	PTS
1	1.8	0.6	4.1	5.3
1	0.4	0.8	1.5	1.7
1	1.1	1.9	2.2	3.2
1	6.0	1.6	0.0	13.9
1	3.4	2.2	0.2	8.9
...	...	...	...	...
1	4.0	0.8	0.0	11.5
1	3.1	0.9	0.0	7.8
1	3.6	1.1	0.0	8.9
1	3.4	0.8	0.0	8.5
1	3.8	1.5	0.0	9.4

Example design matrix
708 rows x (3+1) cols

The Design Matrix $\mathbb{X}$

We can use linear algebra to represent our predictions of all n datapoints at once.

One step in this process is to stack all of our input features together into a **design matrix**:

$$\mathbb{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & x_{31} & x_{32} & \dots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

A **row** corresponds to one **observation**, e.g., all (p+1) features for datapoint 3

A **column** corresponds to a **feature**, e.g. feature 1 for all n data points

Special all-ones feature often called the **bias/intercept**

	Bias	FG	AST	3PA	PTS
	1	1.8	0.6	4.1	5.3
	1	0.4	0.8	1.5	1.7
	1	1.1	1.9	2.2	3.2
	1	6.0	1.6	0.0	13.9
	1	3.4	2.2	0.2	8.9
	...	...	...	...	...
	1	4.0	0.8	0.0	11.5
	1	3.1	0.9	0.0	7.8
	1	3.6	1.1	0.0	8.9
	1	3.4	0.8	0.0	8.5
	1	3.8	1.5	0.0	9.4

Field Goals, Assists, 3-Point Attempts

Example design matrix
708 rows x (3+1) cols

The Multiple Linear Regression Model using Matrix Notation

We can express our linear model on our entire dataset as follows:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & x_{31} & x_{32} & \dots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}$$

$$\hat{\mathbb{Y}} = \mathbb{X}\theta$$

Prediction vector $\mathbb{R}^n$

Design matrix $\mathbb{R}^{n \times (p+1)}$

Parameter vector $\mathbb{R}^{(p+1)}$

Note that our **true output** is also a vector:

$$\mathbb{Y} \in \mathbb{R}^n$$

Loss Functions

The Modeling Process

1. Choose a model ✅

How should we represent the world? $\hat{\mathbb{Y}} = \mathbb{X}\theta$

2. Choose a loss function

How do we quantify prediction error? $R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$

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)

Empirical Risk is Average Loss over Data

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

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

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

Average loss is a function of the parameter θ because **our data do not change**. What defines how well our model works is our choice of θ , 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 $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$:

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

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

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

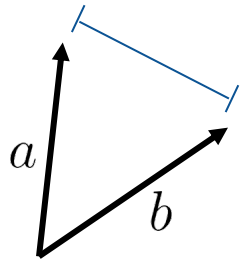
[Linear Algebra] The L2 Norm Is a Measure of Distance

$$||x||_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

The L2 vector norm is a generalization of the Pythagorean theorem into n dimensions.

It can therefore be used as a measure of **distance** between two vectors.

- For n-dimensional vectors a, b , their distance is $||a - b||_2$.



Note: The square of the L2 norm of a vector is the sum of the squares of the vector's elements:

$$||x||_2^2 = \sum_{i=1}^{n} x_i^2$$

Looks like Mean Squared Error!!

Mean Squared Error with L2 Norms

We can rewrite mean squared error as a squared L2 norm:

$$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$= \frac{1}{n}||\mathbb{Y} - \hat{\mathbb{Y}}||_2^2$$

With our linear model $\hat{\mathbb{Y}} = \mathbb{X}\theta$:

$$R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

Ordinary Least Squares

The **least squares estimate** $\hat{\theta}$ is the parameter that **minimizes** the objective function $R(\theta)$:

$$R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

How should we interpret the OLS problem?

A. Minimize the mean squared error for the linear model $\hat{\mathbb{Y}} = \mathbb{X}\theta$

B. Minimize the **distance** between true and predicted values $\mathbb{Y}$ and $\hat{\mathbb{Y}}$

C. Minimize the **length** of the residual vector, $e = \mathbb{Y} - \hat{\mathbb{Y}} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix}$

D. All of the above

E. Something else

Ordinary Least Squares

The **least squares estimate** $\hat{\theta}$ is the parameter that **minimizes** the objective function $R(\theta)$:

$$R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

How should we interpret the OLS problem?

A. Minimize the mean squared error for the linear model $\hat{\mathbb{Y}} = \mathbb{X}\theta$

B. Minimize the **distance** between true and predicted values $\mathbb{Y}$ and $\hat{\mathbb{Y}}$

C. Minimize the **length** of the residual vector, $e = \mathbb{Y} - \hat{\mathbb{Y}} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix}$ Important for today

D. All of the above

E. Something else

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

Evaluating the Model

The Modeling Process

1. Choose a model ✅

Multiple Linear Regression

$$\hat{\mathbb{Y}} = \mathbb{X}\theta$$

2. Choose a loss function ✅

L2 Loss

Mean Squared Error (MSE)

$$R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

3. Fit the model

Minimize average loss with ~~calculus~~ geometry

The calculus derivation requires matrix calculus (out of scope, but here's a link if you're interested).

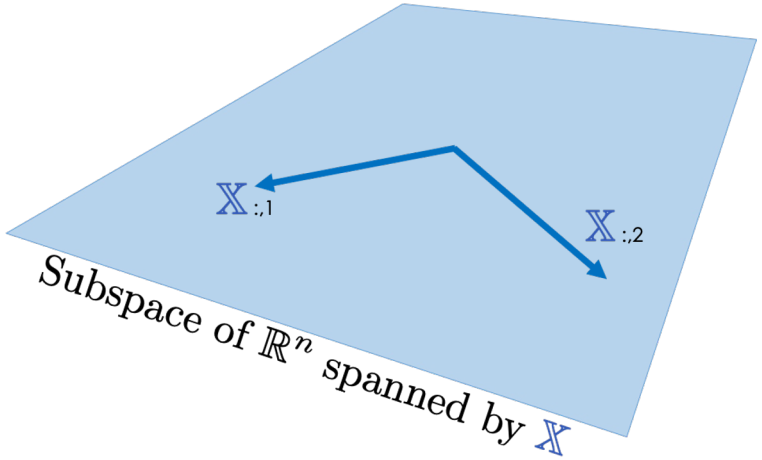
Instead, we will derive $\hat{\theta}$ using a **geometric argument**.

4. Evaluate model performance

[Linear Algebra] Span

The set of all possible linear combinations of the columns of X is called the **span** of the columns of X (denoted $\text{span}(\mathbb{X})$), also called the **column space**.

- Intuitively, this is all of the vectors you can "reach" using the columns of X.
- If each column of X has length n, $\text{span}(\mathbb{X})$ is a subspace of $\mathbb{R}^n$.



A linear combination of columns

$$\hat{\mathbb{Y}} = \mathbb{X}\,\theta$$

So far, we've thought of our model as horizontally stacked predictions per datapoint:

$$\underset{n}{\begin{pmatrix} | \\ \hat{\mathbb{Y}} \\ | \end{pmatrix}}_{1} = \begin{bmatrix} \text{—} x_1^T \text{—} \\ \text{—} x_2^T \text{—} \\ \vdots \\ \text{—} x_n^T \text{—} \end{bmatrix} \underset{p+1}{\begin{pmatrix} | \\ \theta \\ | \end{pmatrix}}_{1} =$$

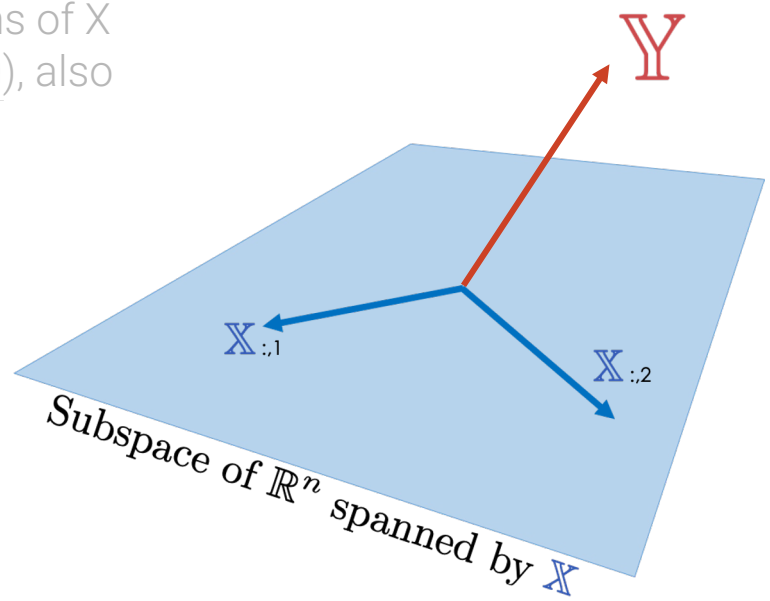
We can also think of $\hat{\mathbb{Y}}$ as a **linear combination of feature vectors**, scaled by **parameters**.

$$\underset{n}{\begin{pmatrix} | \\ \hat{\mathbb{Y}} \\ | \end{pmatrix}}_{1} = \underset{n}{\begin{pmatrix} | & | \\ \mathbb{X}_{:,1} & \mathbb{X}_{:,2} \\ | & | \end{pmatrix}}_{p+1} \underset{p+1}{\begin{pmatrix} | \\ \theta \\ | \end{pmatrix}}_{1} = \theta_1 \begin{pmatrix} | \\ \mathbb{X}_{:,1} \\ | \end{pmatrix} + \theta_2 \begin{pmatrix} | \\ \mathbb{X}_{:,2} \\ | \end{pmatrix}$$

A linear combination of columns

The set of all possible linear combinations of the columns of X is called the **span** of the columns of X (denoted $\text{span}(\mathbb{X})$), also called the **column space**.

- Intuitively, this is all of the vectors you can "reach" using the columns of X.
- If each column of X has length n, $\text{span}(\mathbb{X})$ is a subspace of $\mathbb{R}^n$.

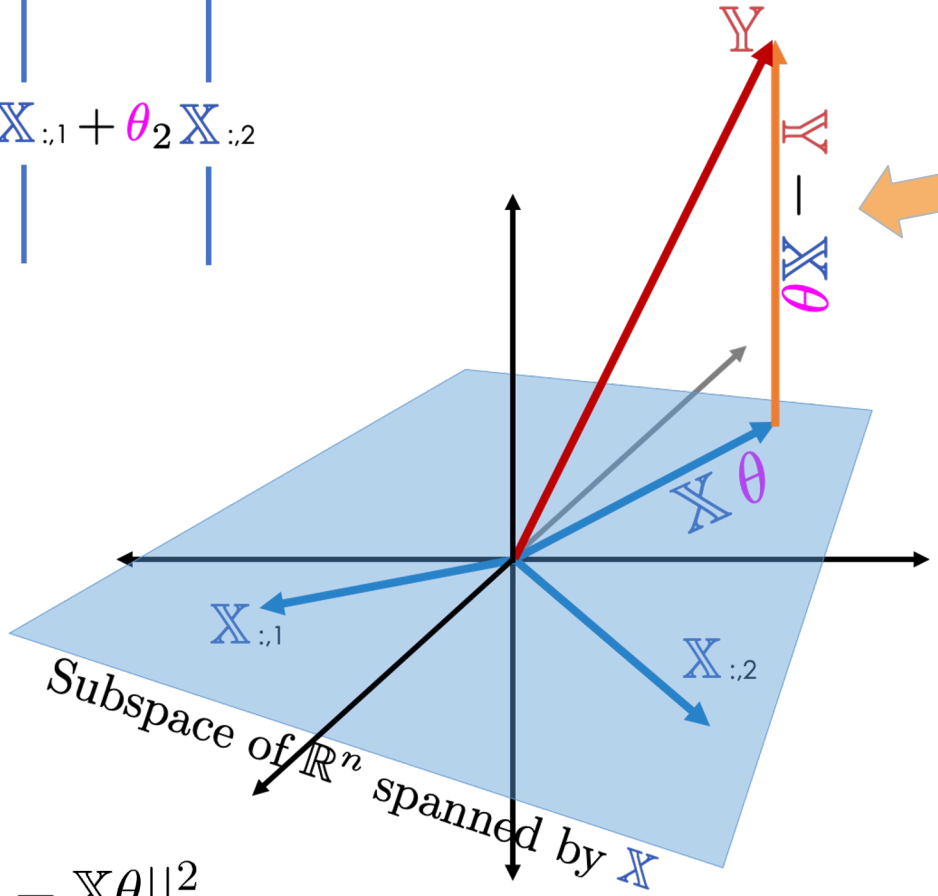


Our prediction $\hat{\mathbb{Y}} = \mathbb{X}\theta$ is a **linear combination** of the columns of $\mathbb{X}$. Therefore $\hat{\mathbb{Y}} \in \text{span}(\mathbb{X})$.

Interpret: Our linear prediction $\hat{\mathbb{Y}}$ will be in $\text{span}(\mathbb{X})$, even if the true values $\mathbb{Y}$ might not be.

Goal: Find the vector in $\text{span}(\mathbb{X})$ that is **closest** to $\mathbb{Y}$.

$$\begin{bmatrix} \\ \hat{\mathbb{Y}} \\ \\ \end{bmatrix}_{n \times 1} = \theta_1 \mathbb{X}_{:,1} + \theta_2 \mathbb{X}_{:,2}$$



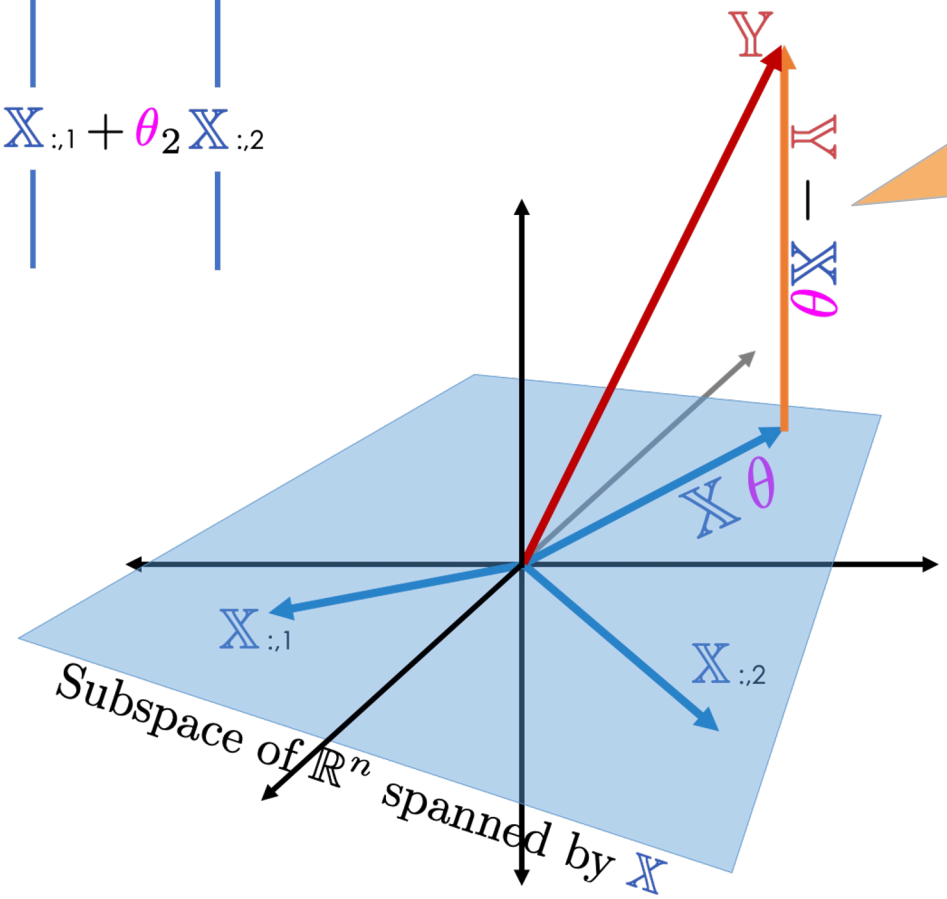
This is the residual vector, $e = \mathbb{Y} - \hat{\mathbb{Y}}$.

Goal:

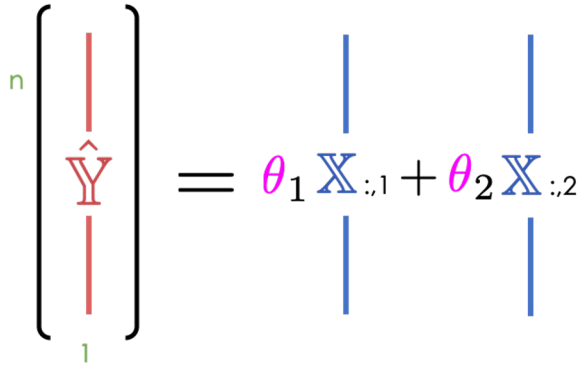
Minimize the $L2$ norm of the residual vector. i.e., get the predictions $\hat{\mathbb{Y}}$ to be "as close" to our true y values as possible.

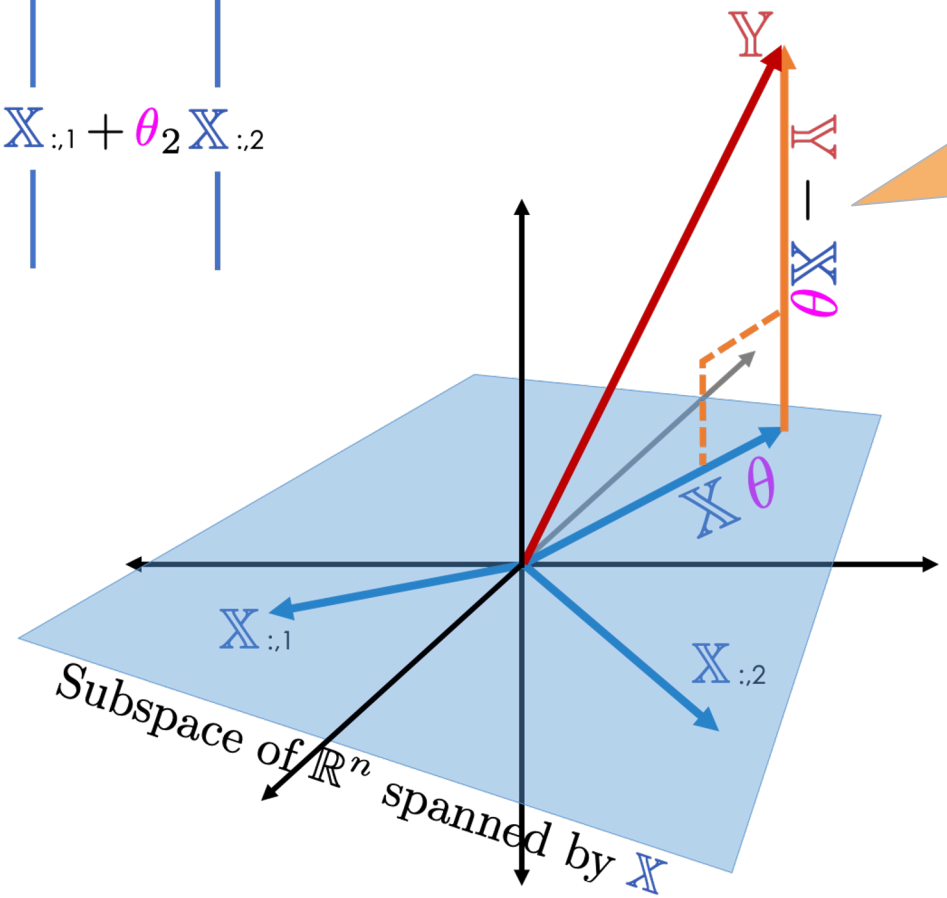
$$R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

$$\underset{n}{\begin{bmatrix} | \\ \hat{\mathbb{Y}} \\ | \end{bmatrix}}_{1} = \theta_1 \mathbb{X}_{:,1} + \theta_2 \mathbb{X}_{:,2}$$



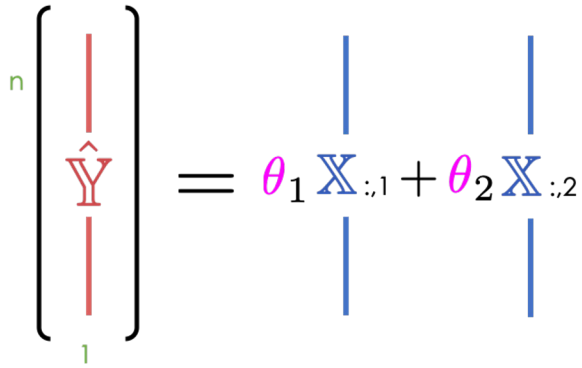
How do we minimize this distance – the norm of the residual vector (squared)?

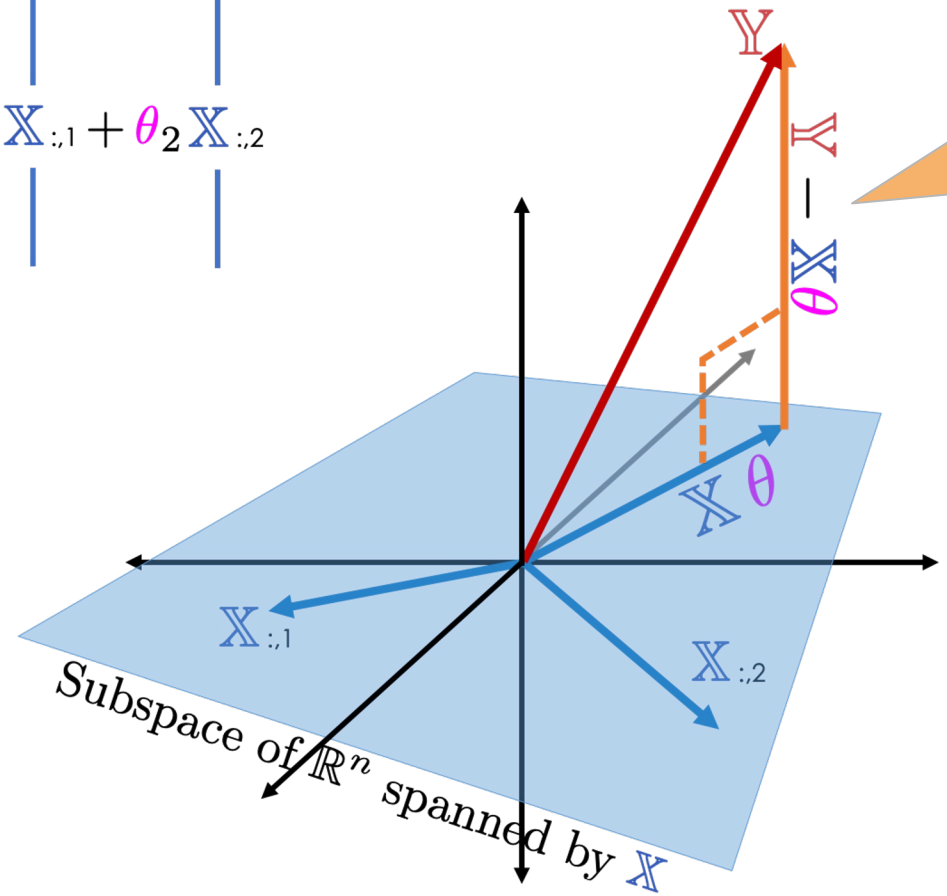




How do we minimize this distance – the norm of the residual vector (squared)?

The vector in span(X) that is closest to Y is the **orthogonal projection** of Y onto span(X).





How do we minimize this distance – the norm of the residual vector (squared)?

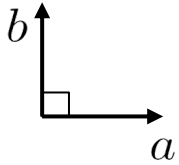
The vector in span(X) that is closest to Y is the **orthogonal projection** of Y onto span(X).

Thus, we should choose the θ that makes the residual vector **orthogonal** to span(X).

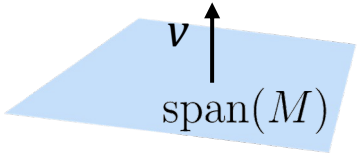
[Linear Algebra] Orthogonality

1. Vector a and Vector b are **orthogonal** if and only if their dot product is 0: $a^T b = 0$

This is a generalization of the notion of two vectors in 2D being perpendicular.



2. A vector v is **orthogonal** to $\text{span}(M)$, the span of the columns of a matrix M, if and only if v is orthogonal to **each column** in M.



Let's express **2** in matrix notation. Let $v \in \mathbb{R}^{n \times 1}$ $M \in \mathbb{R}^{n \times d}$,

$$M = \begin{bmatrix} | & | & & | \\ m_1 & m_2 & \dots & m_d \\ | & | & & | \end{bmatrix}$$

$$\begin{aligned} m_1^T v &= 0 \\ m_2^T v &= 0 \\ &\vdots \\ m_d^T v &= 0 \end{aligned}$$

$$\Rightarrow \begin{bmatrix} m_1^T v \\ m_2^T v \\ \vdots \\ m_d^T v \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \Rightarrow \underbrace{M^T}_{M^T \in \mathbb{R}^{d \times n}} v = \underbrace{\vec{0}}_{\text{zero vector } (d\text{-length vector full of 0s})}$$

v is orthogonal to each column of M, $m_j \in \mathbb{R}^{n \times 1}$

Ordinary Least Squares Proof

The **least squares estimate** $\hat{\theta}$ is the parameter θ that minimizes the objective function $R(\theta)$:

$$R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

Equivalently, this is the $\hat{\theta}$ such that the residual vector $\mathbb{Y} - \mathbb{X}\hat{\theta}$ is orthogonal to $\text{span}(\mathbb{X})$

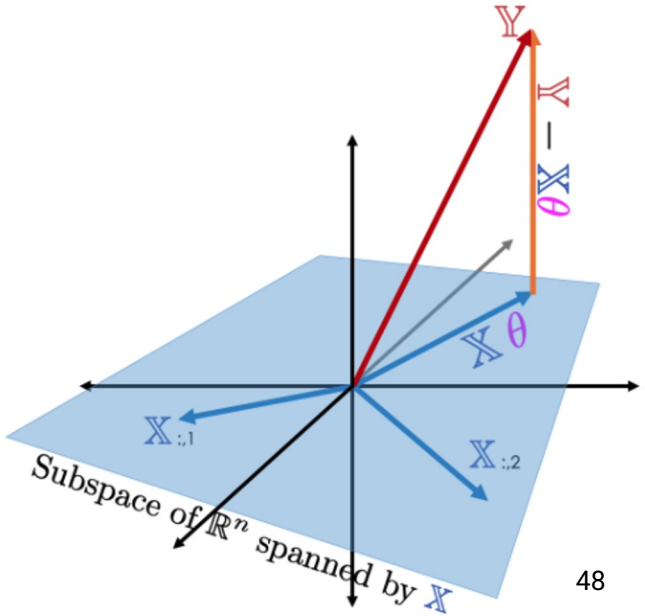
.

Definition of orthogonality $\mathbb{X}^T\left(\mathbb{Y} - \mathbb{X}\hat{\theta}\right) = 0$

Rearrange terms $\mathbb{X}^T\mathbb{Y} - \mathbb{X}^T\mathbb{X}\hat{\theta} = 0$

The **normal equation** $\mathbb{X}^T\mathbb{X}\hat{\theta} = \mathbb{X}^T\mathbb{Y}$

If $\mathbb{X}^T\mathbb{X}$ is invertible $\boxed{\hat{\theta} = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{Y}}$



$$\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

This result is so important that it deserves its own slide.

It is the **least squares estimate** and the solution to the normal equation $\mathbb{X}^T\mathbb{X}\hat{\theta} = \mathbb{X}^T\mathbb{Y}$

.

Least Squares Estimate

1. Choose a model

Multiple Linear Regression

$$\hat{\mathbb{Y}} = \mathbb{X}\theta$$

2. Choose a loss function

L2 Loss

Mean Squared Error (MSE)

$$R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

3. Fit the model ✅

Minimize average loss with ~~calculus~~ geometry

$$\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

4. Evaluate model performance

Evaluating the Model

Least Squares Estimate

1. Choose a model ✅

Multiple Linear Regression

$$\hat{\mathbb{Y}} = \mathbb{X}\theta$$

2. Choose a loss function ✅

L2 Loss

Mean Squared Error (MSE)

$$R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

3. Fit the model ✅

Minimize average loss with ~~calculus~~ geometry

$$\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

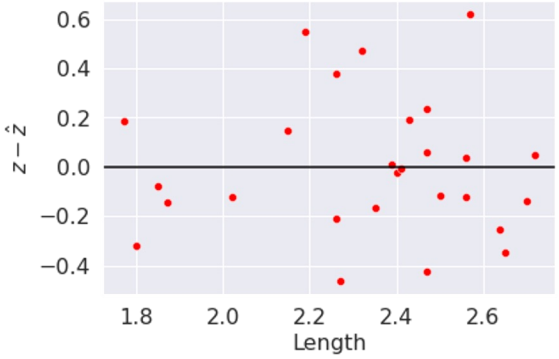
4. Evaluate model performance

Visualize, Multiple R^2

[Visualization] Residual Plots

Simple linear regression

Plot residuals vs
the single feature x.



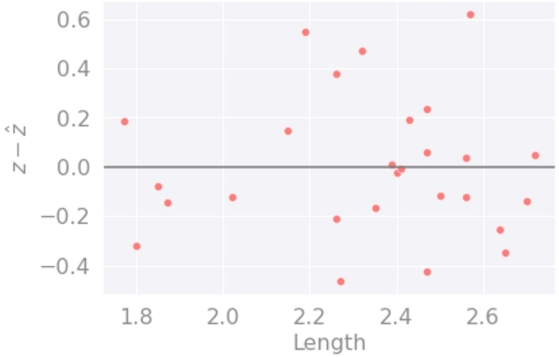
Compare

[Visualization] Residual Plots

Compare

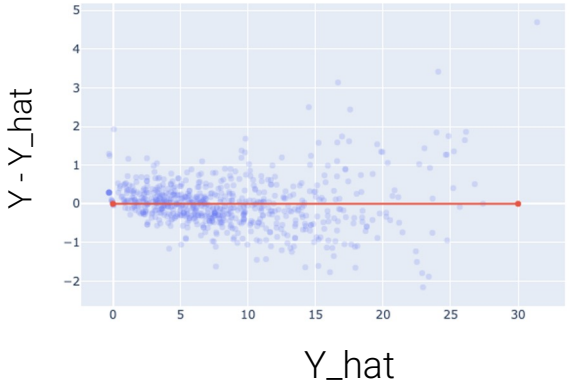
Simple linear regression

Plot residuals vs
the single feature x.



Multiple linear regression

Plot residuals vs
fitted (predicted) values $\hat{y}$.
Check distribution around



Some interpretation

- A good residual plot shows no pattern.
- A good residual plot also has a similar vertical spread throughout the entire plot. Else (heteroscedasticity), the accuracy of the predictions is not reliable.

Compare

[Metrics] Multiple R^2

Simple linear regression

Error
RMSE

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

Linearity
Correlation coefficient, ***r***

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

Multiple linear regression

Error
RMSE

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

Linearity
Multiple R^2, also called the **coefficient of determination**

$$R^2 = \frac{\text{variance of fitted values}}{\text{variance of } y} = \frac{\sigma^2_{\hat{y}}}{\sigma^2_y}$$

Compare

[Metrics] Multiple R^2

We define the **multiple R²** value as the **proportion of variance** or our **fitted values** (predictions) $\hat{y}$ to our true values y.

$$R^2 = \frac{\text{variance of fitted values}}{\text{variance of } y} = \frac{\sigma^2_{\hat{y}}}{\sigma^2_y}$$

Also called the **correlation of determination**.

R^2 ranges from 0 to 1 and is effectively
"the proportion of variance that the **model explains**."

For OLS with an intercept term (e.g. $\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$),
$R^2 = [r(y, \hat{y})]^2$ is equal to the square of correlation between y, $\hat{y}$.

- For SLR, $R^2 = r^2$, the correlation between x, y.

Residual Properties

When using the optimal parameter vector, our residuals $e = \mathbb{Y} - \mathbb{X}\hat{\theta}$ are orthogonal to $\text{span}(\mathbb{X})$.

$$\mathbb{X}^T e = 0$$

Proof For all linear models:

Since our predicted response is in by definition, it is orthogonal to the residuals.

$\hat{\mathbb{Y}}$ $\text{span}(\mathbb{X})$

$$\hat{\mathbb{Y}}^T e = 0$$

For all linear models with an **intercept term**, the **sum of residuals is zero**.

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

$$\sum_{i=1}^{n} e_i = 0$$

(Proof hint) $\mathbb{1}^T e = 0$

$$\mathbb{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & x_{31} & x_{32} & \dots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

Properties when our model has an intercept term

For all linear models with an **intercept term**, the **sum of residuals is zero**.

$$\sum_{i=1}^{n} e_i = 0$$

(previous slide)

- This is the real reason why we don't directly use residuals as loss.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$$

- This is also why positive and negative residuals will cancel out in any residual plot where the (linear) model contains an intercept term, even if the model is terrible.

It follows from the property above that for linear models with intercepts, the average predicted y value is equal to the average true y value.

$$\bar{y} = \overline{\hat{y}}$$

These properties are true when there is an intercept term, and not necessarily when there isn't.