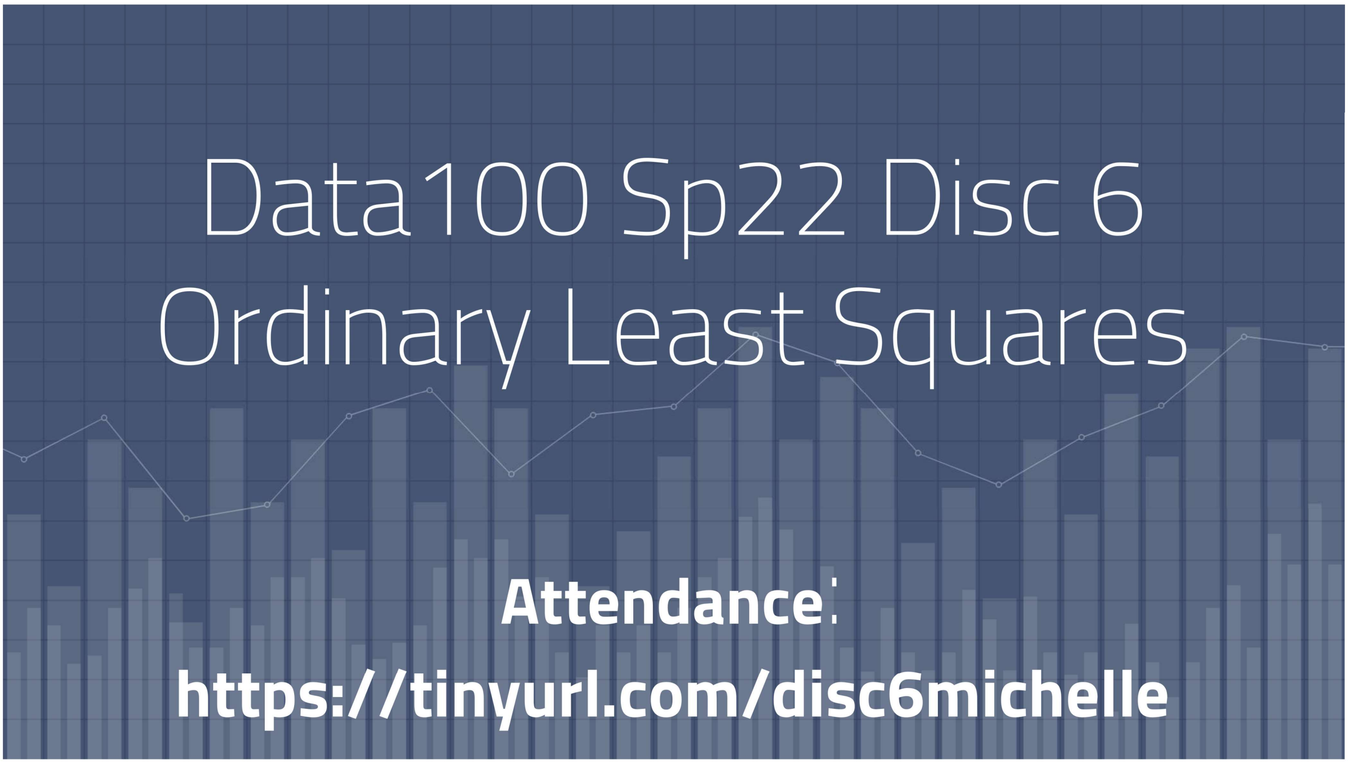
# Data100 Sp22 Disc 6

25 February 2022 08:56

Data100 Sp22 Disc 6



# Announcements

**Due Dates**

-Homework 5 due March 3 (start early)

-Lab 6 due March 1

-Weekly check 6 due Feb 28

**Other**

-Congrats on finishing the midterm!

-Calendly: https://forms.gle/3BJjPjPbMkNHs2ir9

# Modeling Process

1

# Recap: Modeling Process

1. Choose a model
2. Choose a loss function
3. Fit the model
4. Evaluate model performance

# Recap: Modeling Process
# Simple Linear Regression

**1. Choose a model**

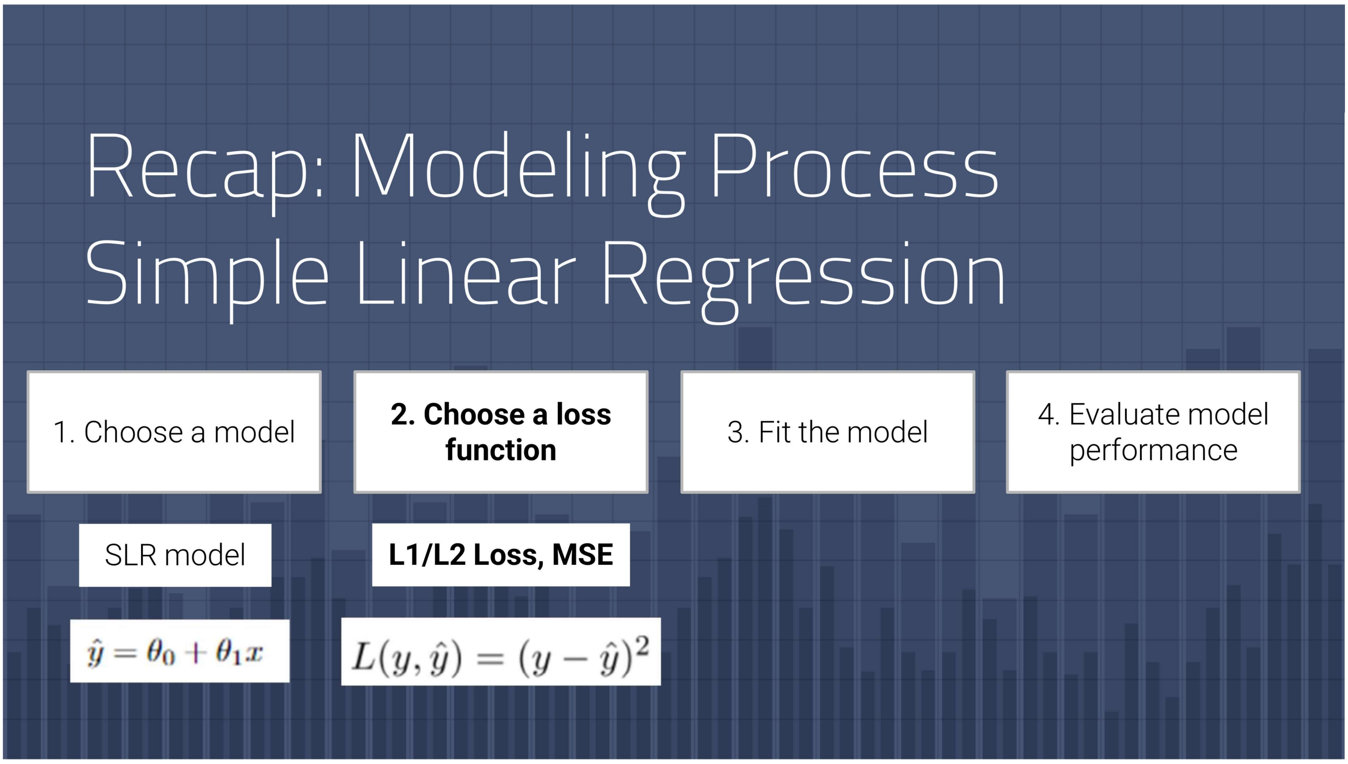
2. Choose a loss function

3. Fit the model

4. Evaluate model performance

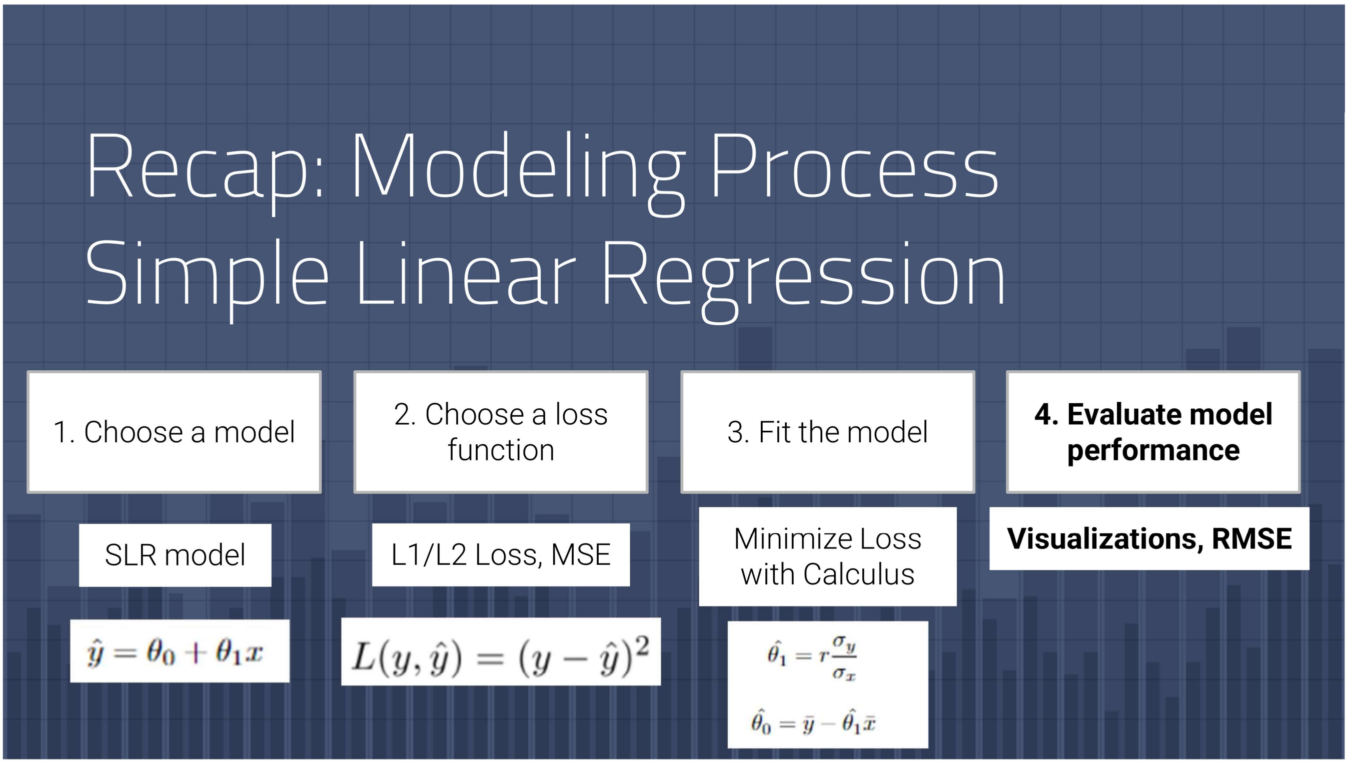
**SLR model**

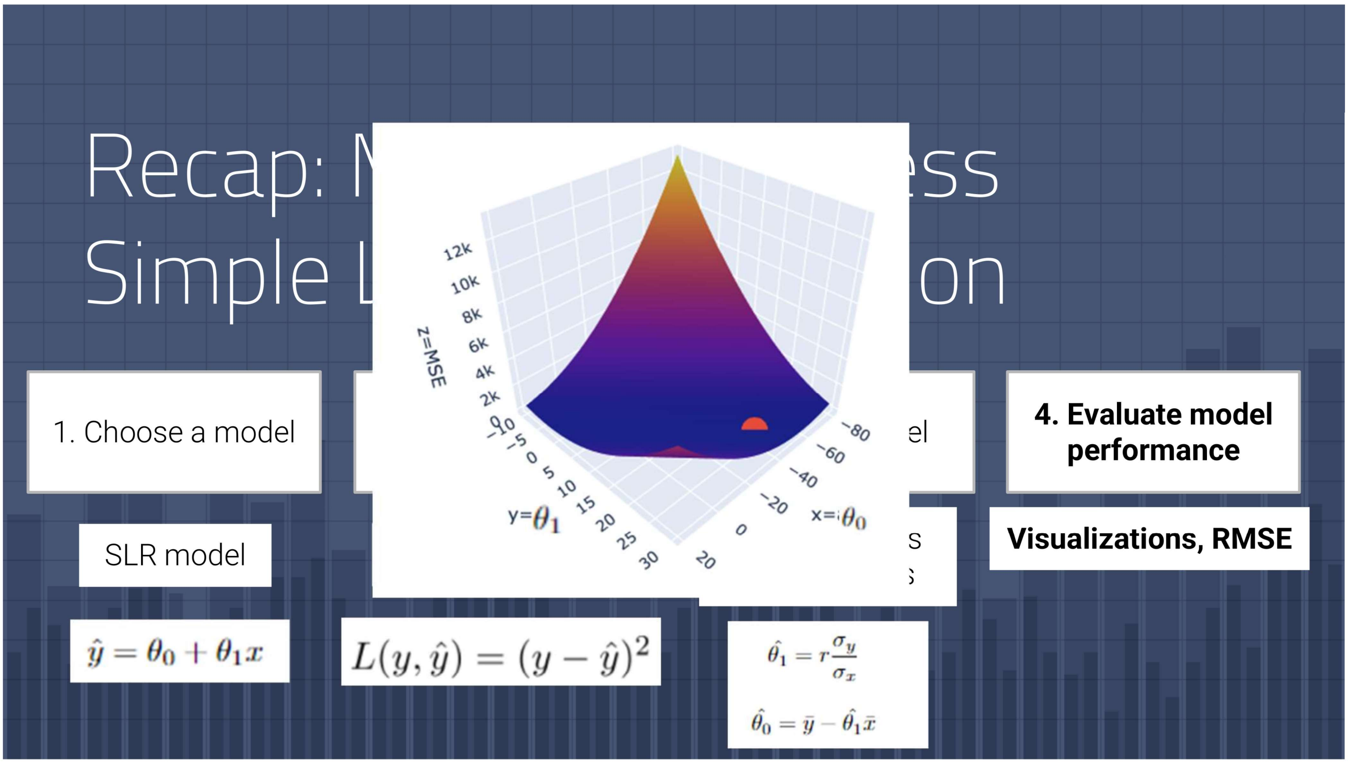
$$\hat{y} = \theta_0 + \theta_1 x$$


Recap: Modeling Process
Simple Linear Regression
1. Choose a model
2. Choose a loss function
3. Fit the model
4. Evaluate model performance
SLR model
L1/L2 Loss, MSE
$\hat{y} = \theta_0 + \theta_1 x$
$L(y, \hat{y}) = (y - \hat{y})^2$

# Recap: Modeling Process Simple Linear Regression

| 1. Choose a model | 2. Choose a loss function | **3. Fit the model** | 4. Evaluate model performance |
|---|---|---|---|
| SLR model | L1/L2 Loss, MSE | **Minimize Loss with Calculus** | |
| $\hat{y} = \theta_0 + \theta_1 x$ | $L(y, \hat{y}) = (y - \hat{y})^2$ | $\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x}$<br>$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$ | |


Recap: Modeling Process
Simple Linear Regression
1. Choose a model
2. Choose a loss function
3. Fit the model
4. Evaluate model performance
SLR model
L1/L2 Loss, MSE
Minimize Loss with Calculus
Visualizations, RMSE
$\hat{y} = \theta_0 + \theta_1 x$
$L(y, \hat{y}) = (y - \hat{y})^2$
$\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x}$
$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$



Recap:
Simple
1. Choose a model
SLR model
$\hat{y} = \theta_0 + \theta_1 x$
$L(y, \hat{y}) = (y - \hat{y})^2$
$\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x}$
$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1\bar{x}$
4. Evaluate model performance
Visualizations, RMSE
z=MSE
y=$\theta_1$
x=$\theta_0$

# Why Multiple Linear Regression?

- **Simple Linear Regression not enough for all use cases**
  - **Often want to predict the value of the response variable based on multiple predictor variables.**

# Why Multiple Linear Regression?

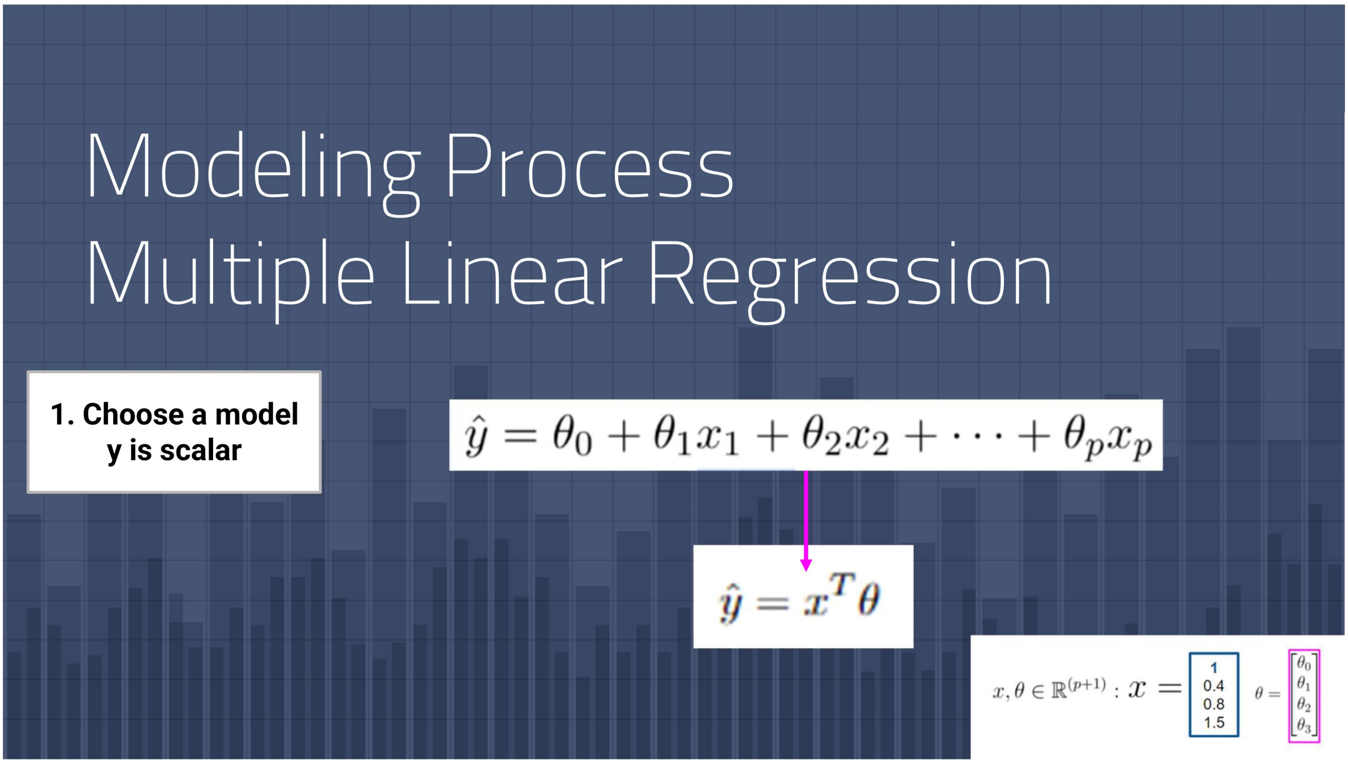
- Simple Linear Regression not enough for all use cases
  - Often want to predict the value of the response variable based on multiple predictor variables.
  - E.g. predict points based on all 3 of
    Field Goals (FG), Assists (AST), and 3 pointers (3PA)
  - SLR - can only predict points based on one out of {FG, AST, 3PA}

| | 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 |

# Modeling Process
# Multiple Linear Regression

**1. Choose a model**
**y is scalar**

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


Modeling Process
Multiple Linear Regression
1. Choose a model
y is scalar
$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$
$\hat{y} = x^T \theta$
$x, \theta \in \mathbb{R}^{(p+1)} : x = \begin{bmatrix} 1 \\ 0.4 \\ 0.8 \\ 1.5 \end{bmatrix}$ $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$

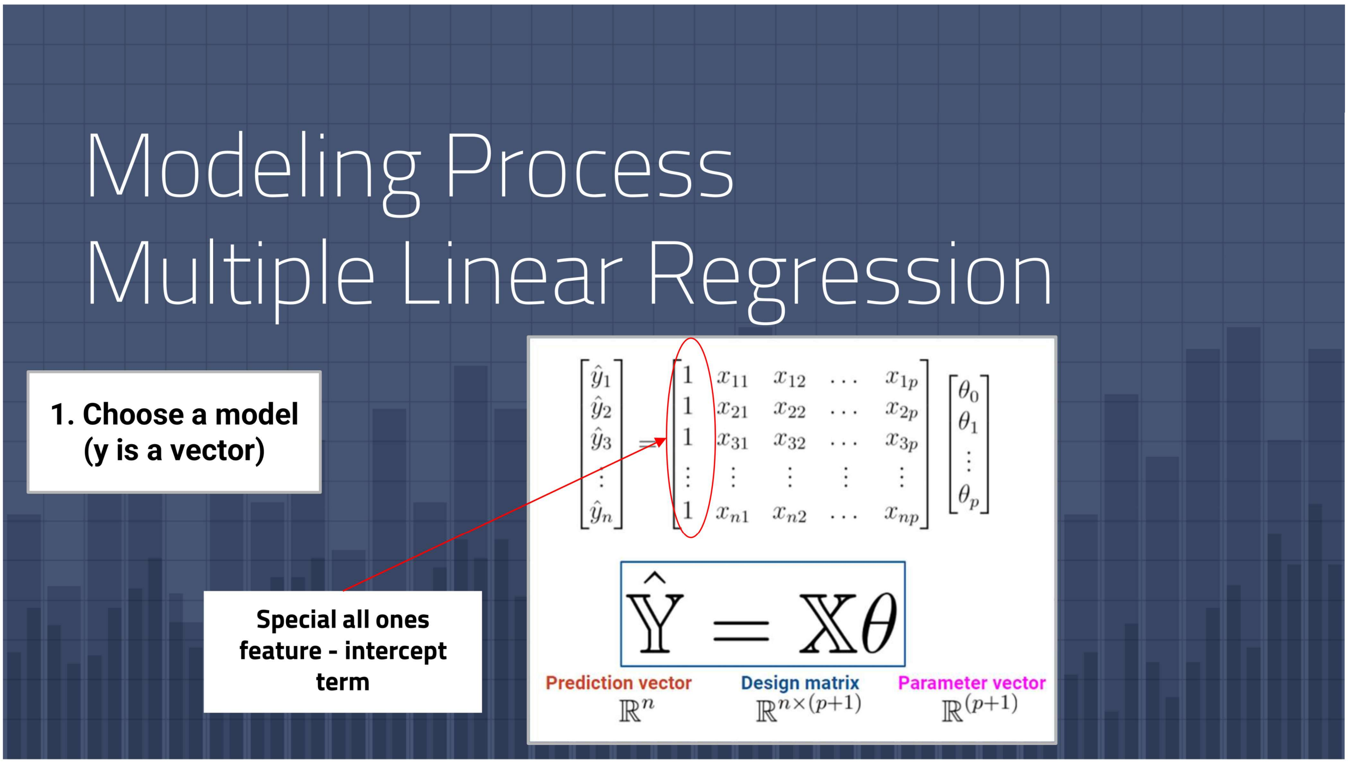
# Modeling Process
# Multiple Linear Regression

**1. Choose a model (y is a vector)**

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

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

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


Modeling Process
Multiple Linear Regression
1. Choose a model
(y is a vector)
Special all ones
feature - intercept
term
$\hat{\mathbb{Y}} = \mathbb{X}\theta$
Prediction vector
$\mathbb{R}^n$
Design matrix
$\mathbb{R}^{n\times(p+1)}$
Parameter vector
$\mathbb{R}^{(p+1)}$

# Modeling Process
# Multiple Linear Regression

**2. Choose a loss function**

L2 Loss

Mean Squared Error (MSE)

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

# Modeling Process
# Multiple Linear Regression

**2. Choose a loss function**

L2 Loss

Mean Squared Error (MSE)

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

For the n-dimensional vector $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}$, the **L2 vector norm** is

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

# Modeling Process
# Multiple Linear Regression

**3. Fit the Model**

**The value of theta that minimizes the MSE loss (R(θ)) is:**

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

**Normal Equation**

# Modeling Process
# Multiple Linear Regression

**3. Fit the Model**

**The value of theta that minimizes the MSE loss (R(θ)) is:**

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

**Least squares estimate**

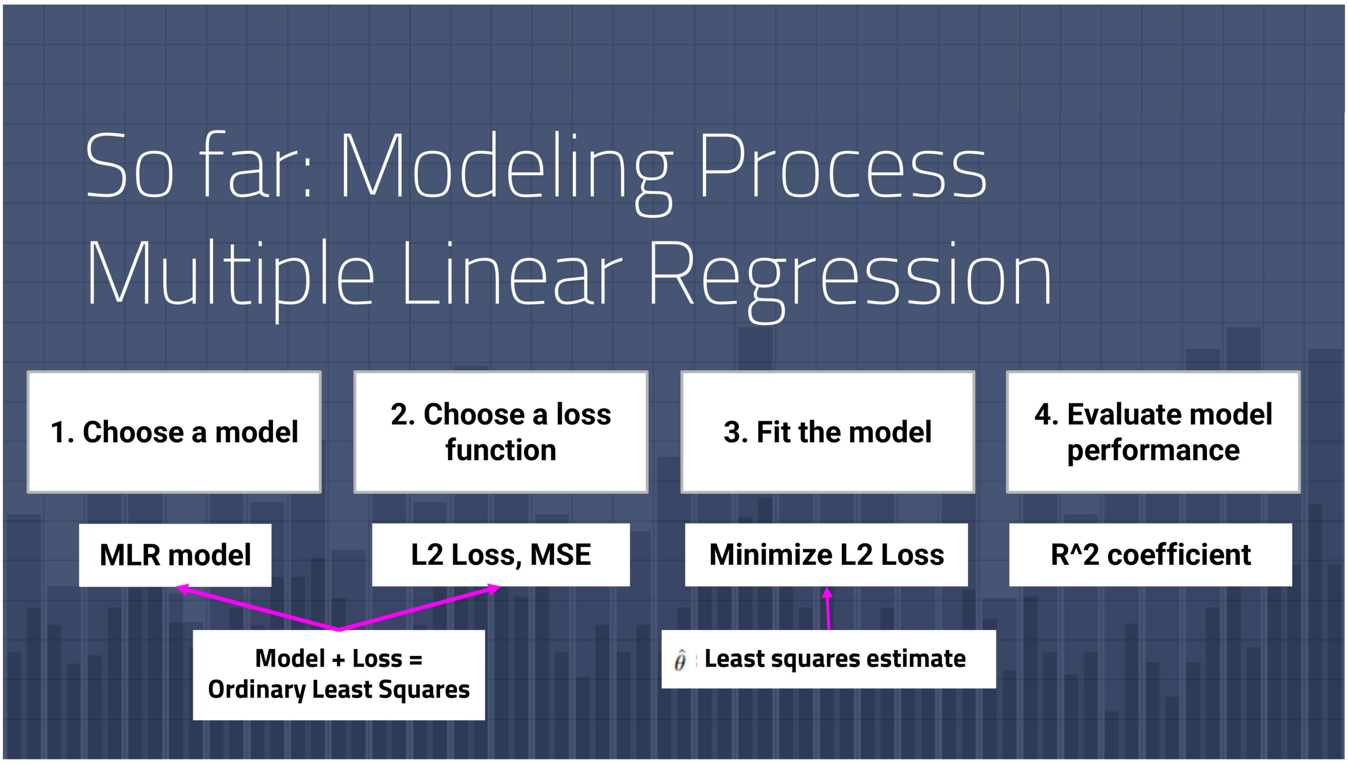
**Normal Equation**

# Modeling Process
# Multiple Linear Regression

**4. Evaluate Model Performance**

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

$$R^2 = \frac{\text{variance of fitted values}}{\text{variance of } y} = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2}$$


So far: Modeling Process
Multiple Linear Regression
1. Choose a model
2. Choose a loss function
3. Fit the model
4. Evaluate model performance
MLR model
L2 Loss, MSE
Minimize L2 Loss
R^2 coefficient
Model + Loss =
Ordinary Least Squares
$\hat{\theta}$ : Least squares estimate

2. Which of the following are true about the optimal solution $\hat{\theta}$ to ordinary least squares (OLS)? Recall that the least squares estimate $\hat{\theta}$ solves the normal equation $(\mathbb{X}^T\mathbb{X})\theta = \mathbb{X}^T\mathbb{Y}$.

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

**Hint: OLS optimizes the MSE**

- A. Using the normal equation, we can derive an optimal solution for simple linear regression with an $L_2$ loss.
- ☐ B. Using the normal equation, we can derive an optimal solution for simple linear regression with an $L_1$ loss.
- C. Using the normal equation, we can derive an optimal solution for a constant model with an $L_2$ loss.
- ☐ D. Using the normal equation, we can derive an optimal solution for a constant model with an $L_1$ loss.
- E. Using the normal equation, we can derive an optimal solution for the model specified option B in question 1 ($\hat{y} = \theta_1 x + \theta_2 \sin(x^2)$).

3. Which of the following conditions are required for the least squares estimate in Question 2?

- A. $\mathbb{X}$ must be full column rank.
- ☐ B. $\mathbb{Y}$ must be full column rank.
- ☐ C. $\mathbb{X}$ must be invertible.
- ☐ D. $\mathbb{X}^{T}$ must be invertible.

# Geometric Intuition

2

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

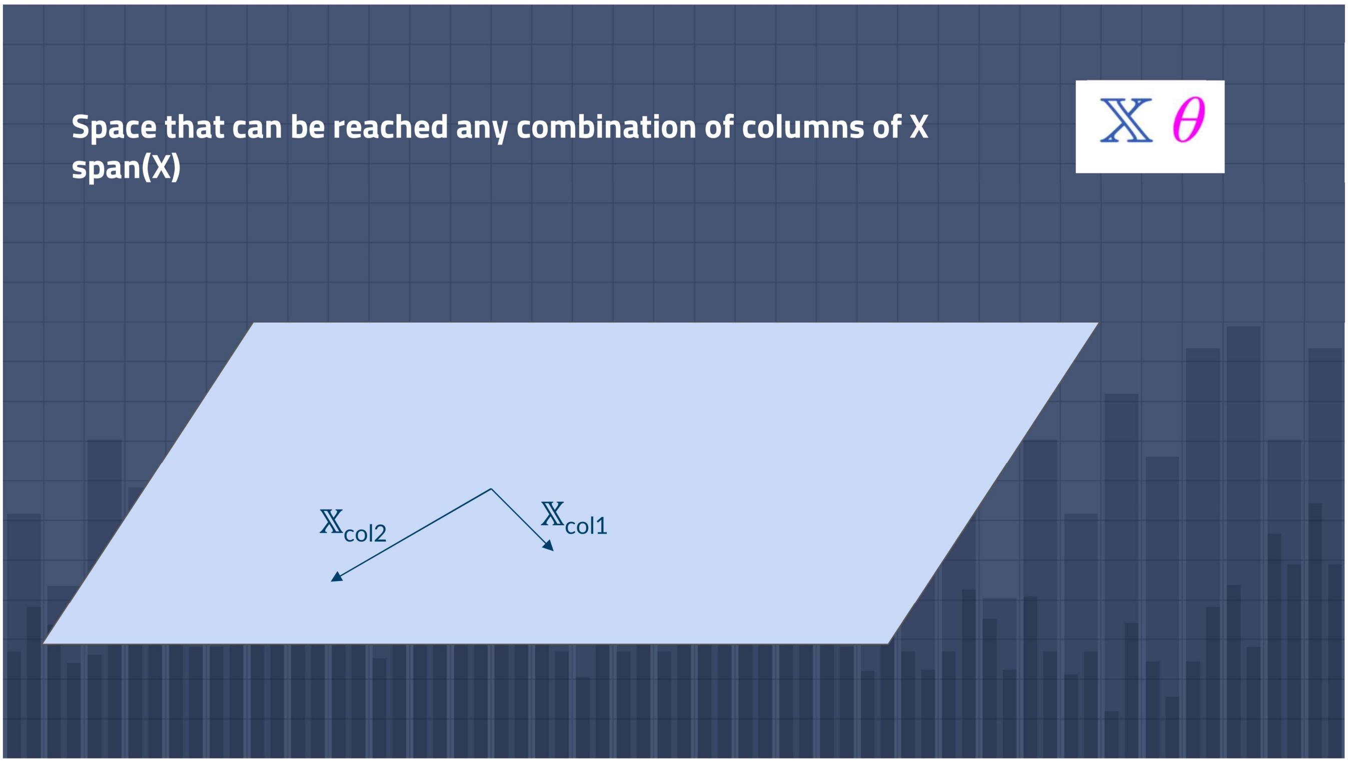
So far, we've thought of our model as horizontally stacked predictions per datapoint:

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

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

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

30



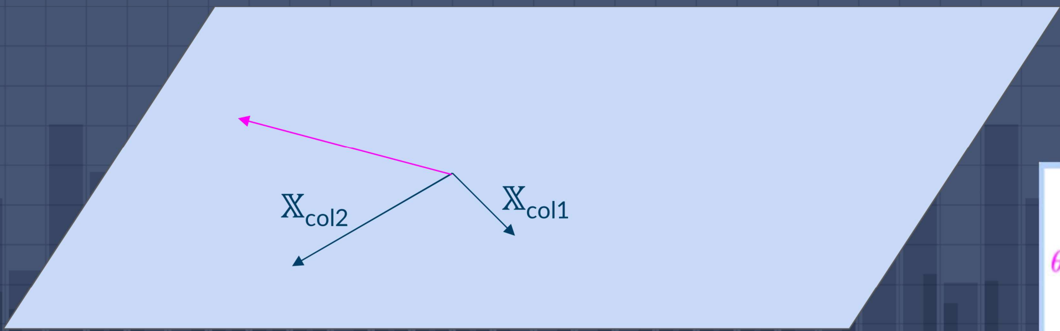
Space that can be reached any combination of columns of X
span(X)
𝕏 θ
𝕏col2
𝕏col1

Space that can be reached by any combination of columns of X

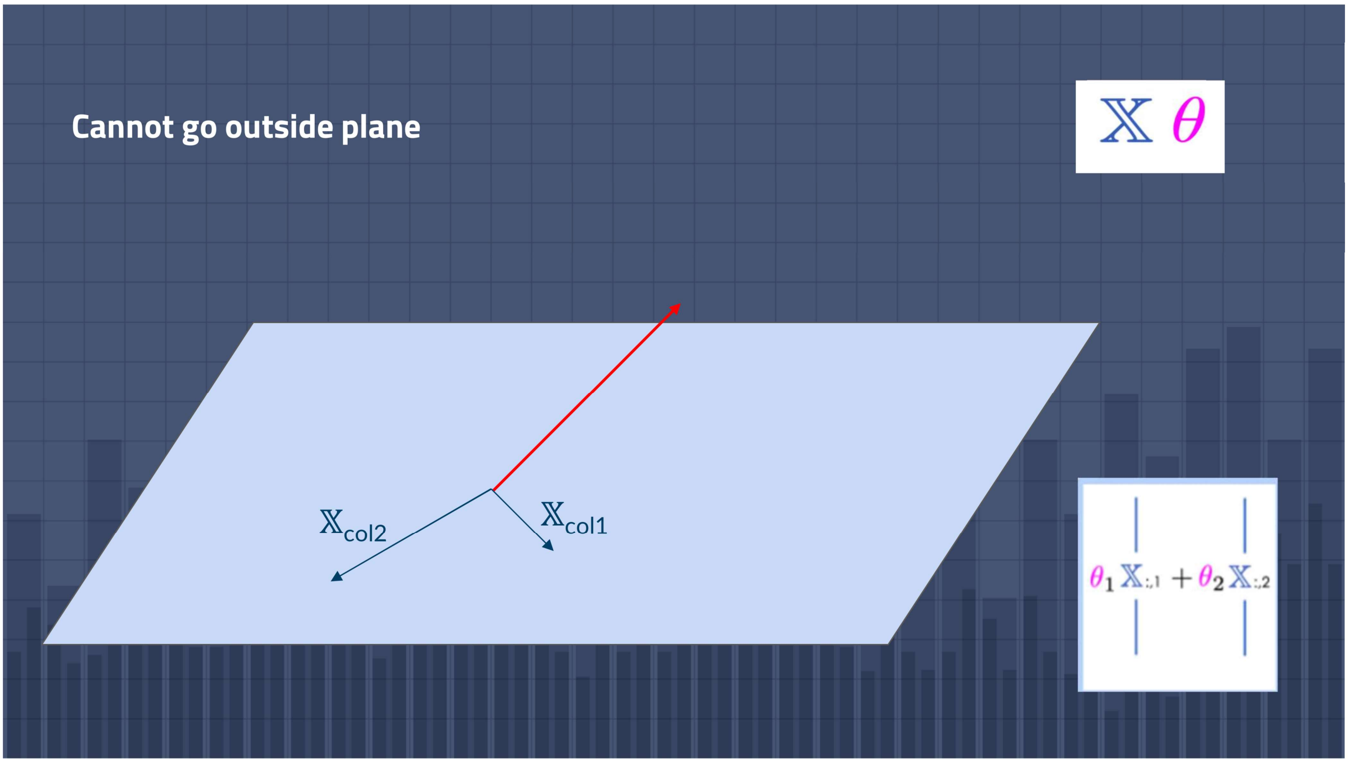
- span(X)

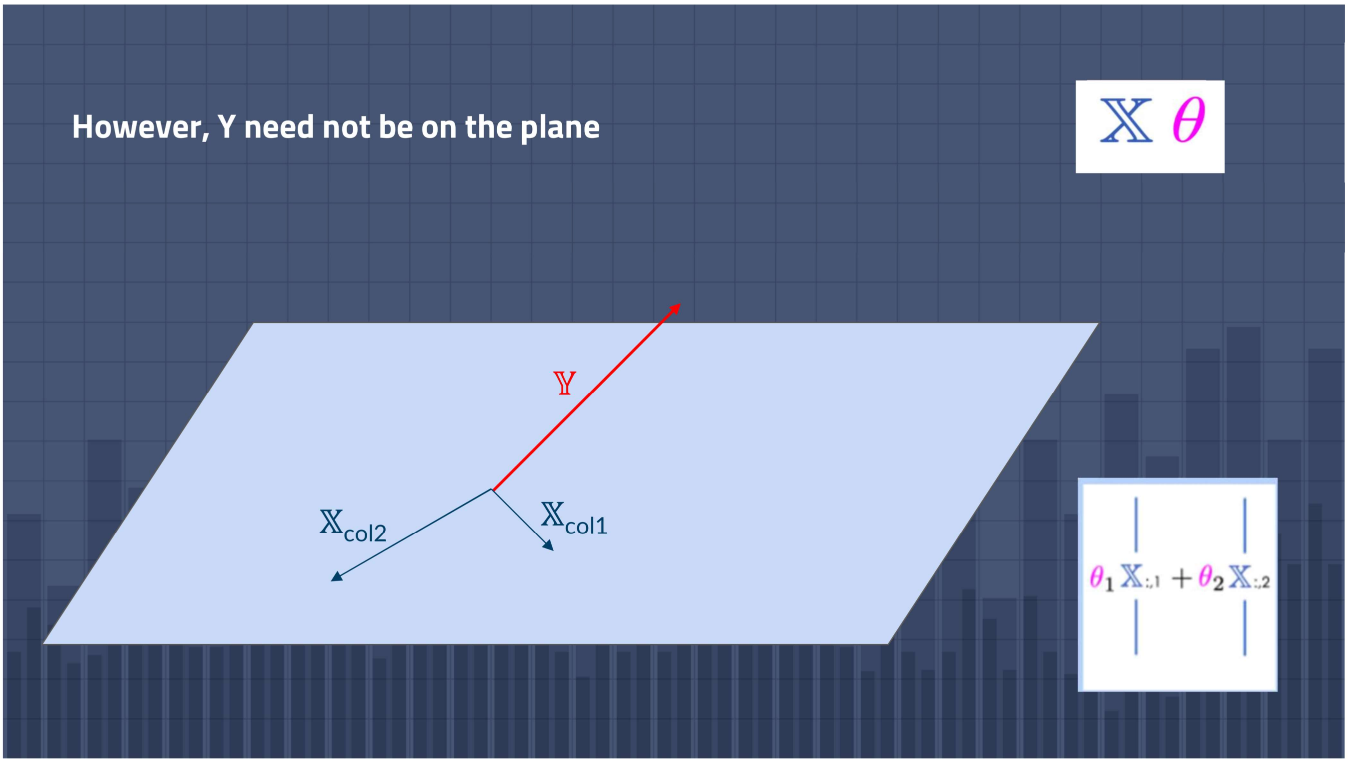
Could be any linear combination (e.g. this could be -2*col1 + 0.7*col2

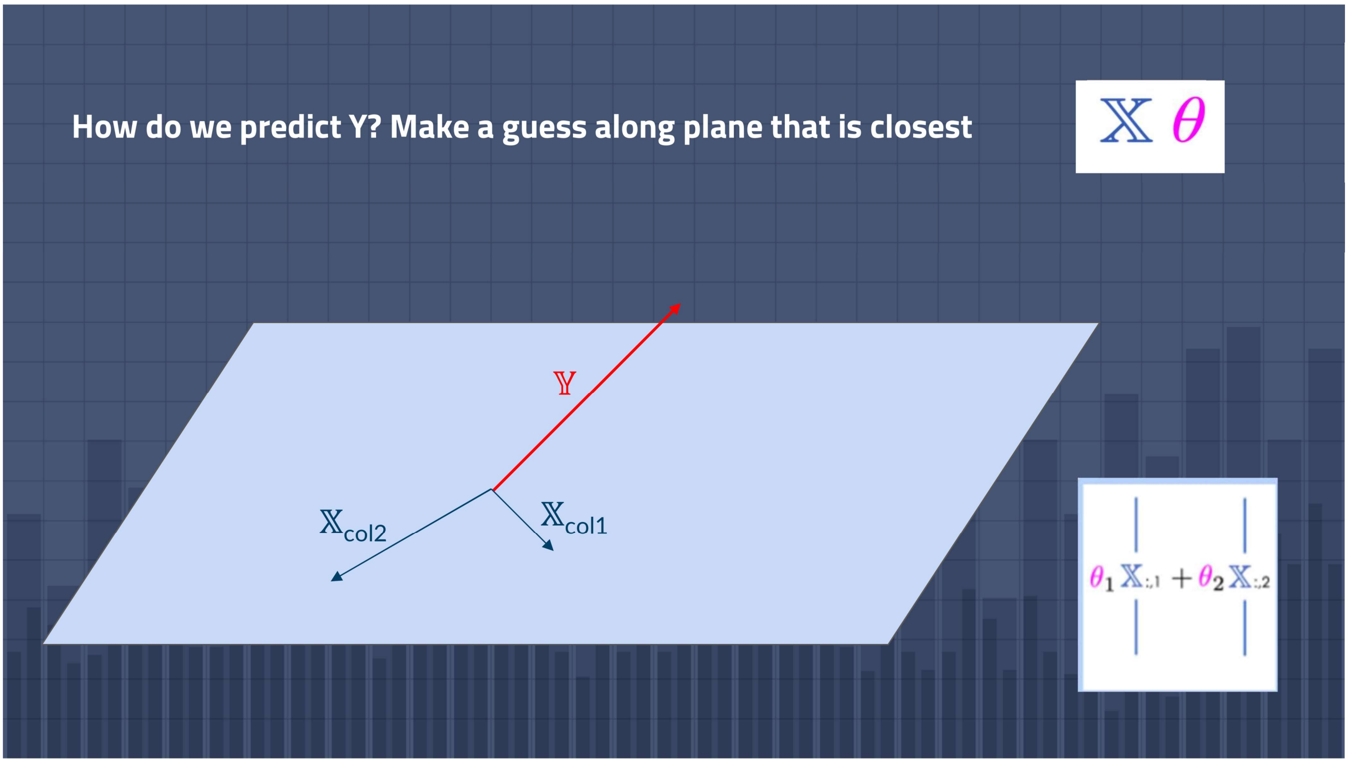
$$\mathbb{X}\,\theta$$

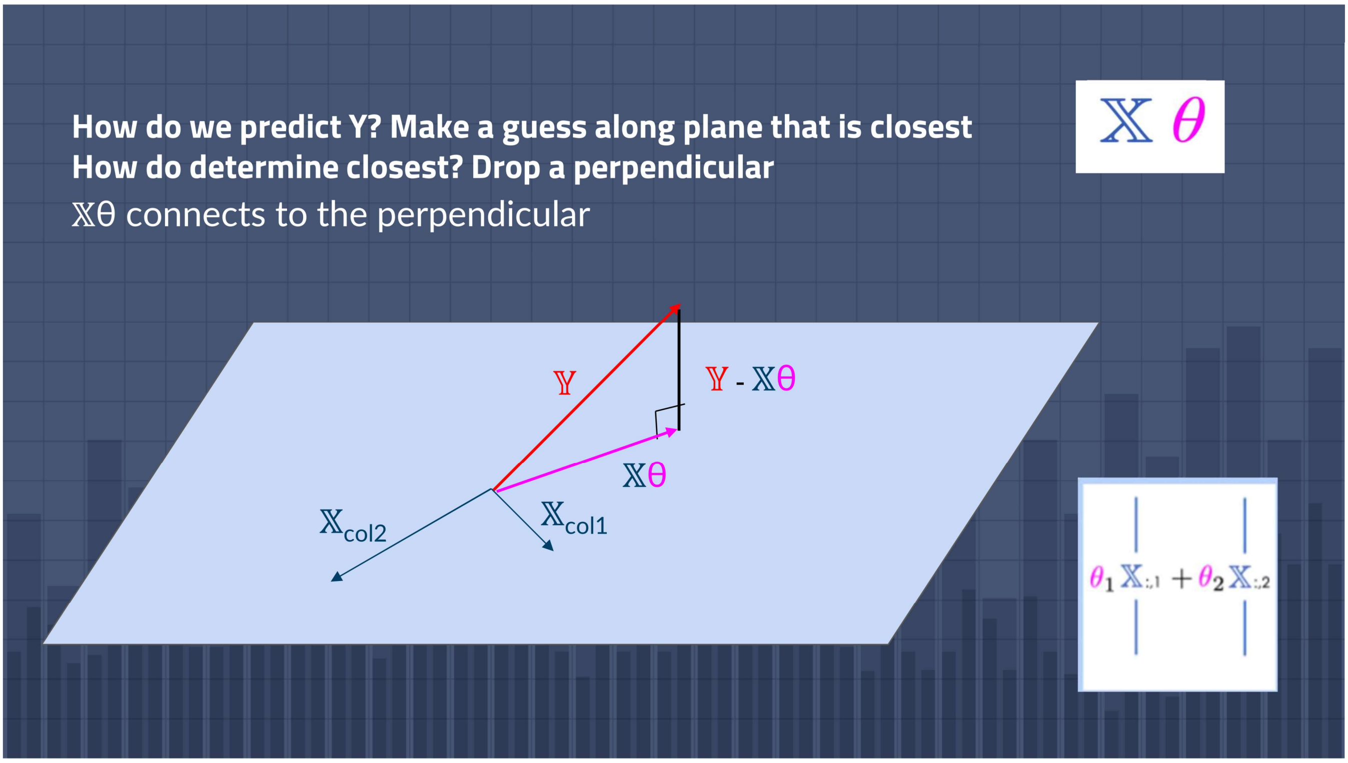


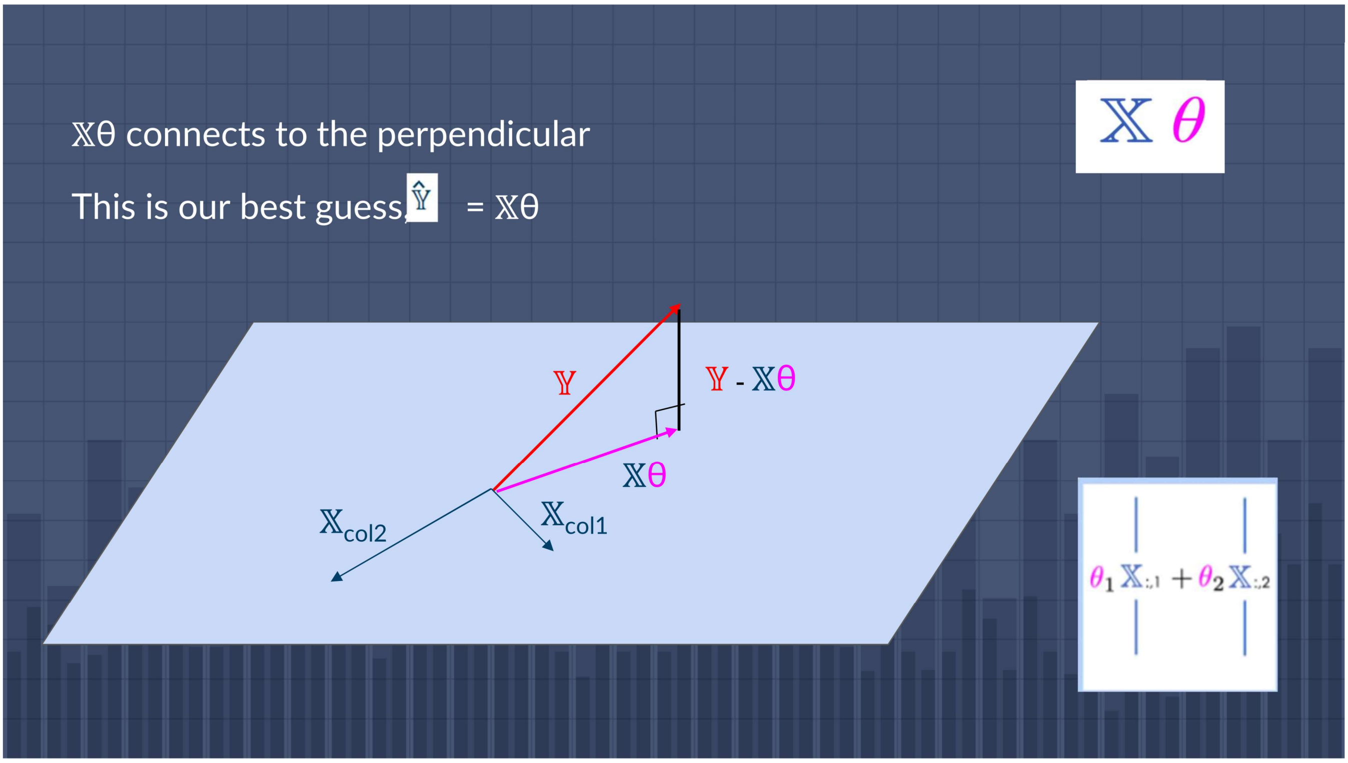

$$\theta_1 \mathbb{X}_{:,1} + \theta_2 \mathbb{X}_{:,2}$$


Cannot go outside plane
$\mathbb{X}\,\theta$
$\mathbb{X}_{col2}$
$\mathbb{X}_{col1}$
$\theta_1 \mathbb{X}_{:,1} + \theta_2 \mathbb{X}_{:,2}$

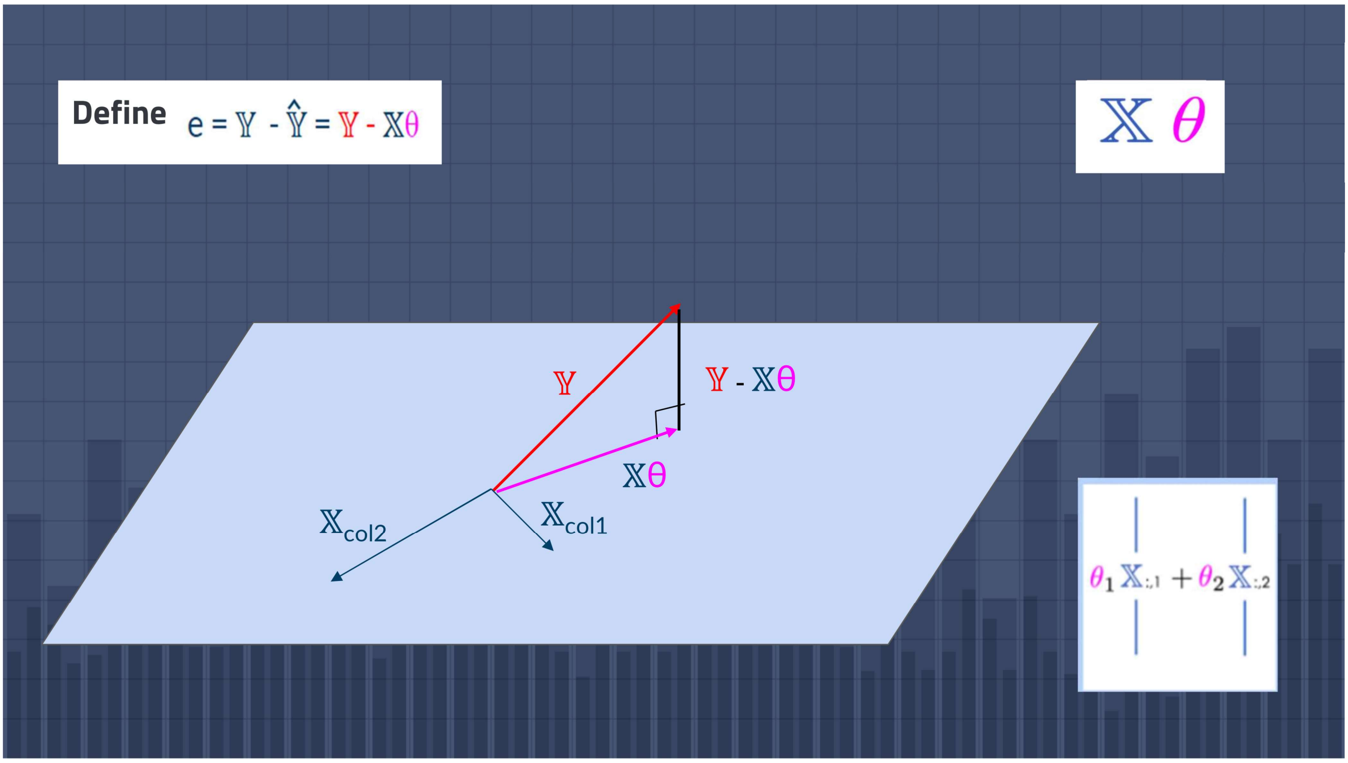

However, Y need not be on the plane
$\mathbb{X}\,\theta$
$\mathbb{Y}$
$\mathbb{X}_{col2}$
$\mathbb{X}_{col1}$
$\theta_1 \mathbb{X}_{:,1} + \theta_2 \mathbb{X}_{:,2}$


How do we predict Y? Make a guess along plane that is closest
$\mathbb{X}\,\theta$
$\mathbb{Y}$
$\mathbb{X}_{col2}$
$\mathbb{X}_{col1}$
$\theta_1 \mathbb{X}_{:,1} + \theta_2 \mathbb{X}_{:,2}$


How do we predict Y? Make a guess along plane that is closest
How do determine closest? Drop a perpendicular
𝕏θ connects to the perpendicular
𝕏 θ
𝕐
𝕐 - 𝕏θ
𝕏θ
𝕏col2
𝕏col1
θ1 𝕏:,1 + θ2 𝕏:,2



𝕏θ connects to the perpendicular
This is our best guess, $\hat{\mathbb{Y}}$ = 𝕏θ
$\mathbb{X}\ \theta$
𝕐
𝕐 - 𝕏θ
𝕏θ
$\mathbb{X}_{col2}$
$\mathbb{X}_{col1}$
$\theta_1 \mathbb{X}_{:,1} + \theta_2 \mathbb{X}_{:,2}$

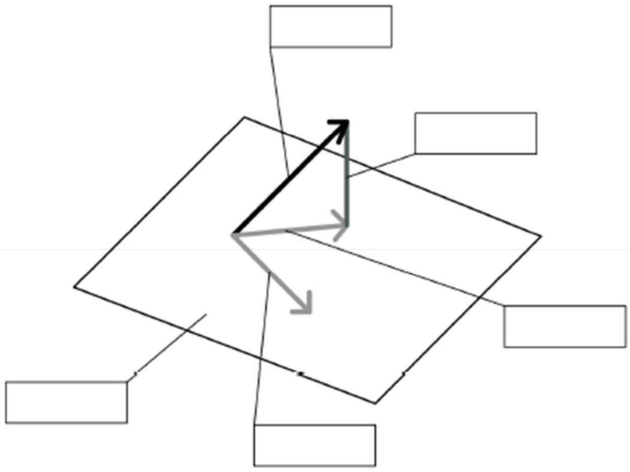


Define e = 𝕐 - 𝕐̂ = 𝕐 - 𝕏θ
𝕏 θ
𝕐
𝕐 - 𝕏θ
𝕏θ
𝕏col2
𝕏col1
θ1 𝕏:,1 + θ2 𝕏:,2

# Some nice properties

- When using θ, residuals (e) are orthogonal to span($\mathbb{X}$) $\mathbb{X}^T e = 0$
- Linear models with an intercept terms WILL HAVE the sum of their residuals to be 0 $\sum_{i=1}^{n} e_i = 0$
- A least squares estimate θ is unique only if $\mathbb{X}$ is full column rank

4. Suppose we have a dataset represented with the design matrix span($\mathbb{X}$) and response vector $\mathbb{Y}$. We use linear regression to solve for this and obtain optimal weights as $\hat{\theta}$. Label the following terms on the geometric interpretation of ordinary least squares:
   - $\mathbb{X}$ (i.e., span($\mathbb{X}$))
   - The response vector $\mathbb{Y}$
   - The residual vector $\mathbb{Y} - \mathbb{X}\hat{\theta}$
   - The prediction vector $\mathbb{X}\hat{\theta}$ (using optimal parameters)
   - A prediction vector $\mathbb{X}\alpha$ (using an arbitrary vector $\alpha$).



(a) What is always true about the residuals in least squares regression? Select all that apply.

☐ A. They are orthogonal to the column space of the design matrix.

☐ B. They represent the errors of the predictions.

☐ C. Their sum is equal to the mean squared error.

☐ D. Their sum is equal to zero.

☐ E. None of the above.

(b) Which are true about the predictions made by OLS? Select all that apply.

☐ A. They are projections of the observations onto the column space of the design matrix.

☐ B. They are linear combinations of the features.

☐ C. They are orthogonal to the residuals.

☐ D. They are orthogonal to the column space of the features.

(c) We fit a simple linear regression to our data $(x_i, y_i), i = 1, 2, 3$, where $x_i$ is the independent variable and $y_i$ is the dependent variable. Our regression line is of the form $\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$. Suppose we plot the relationship between the residuals of the model and the $\hat{y}s$, and find that there is a curve. What does this tell us about our model?

- ☐ A. The relationship between our dependent and independent variables is well represented by a line.
- ☐ B. The accuracy of the regression line varies with the size of the dependent variable.
- ☐ C. The variables need to be transformed, or additional independent variables are needed.

(d) Which are the following is true of the mystery quantity $\vec{v} = (I - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T)\mathbb{Y}$?

- ☐ A. The vector $\vec{v}$ represents the residuals for any linear model.
- ☐ B. If the $\mathbb{X}$ matrix contains the $\vec{1}$ vector, then the sum of the elements in vector $\vec{v}$ is 0 (i.e. $\sum_i v_i = 0$).
- ☐ C. All the column vectors $x_i$ of $\mathbb{X}$ are orthogonal to $\vec{v}$.
- ☐ D. If $\mathbb{X}$ is of shape $n$ by $p$, there are $p$ elements in vector $\vec{v}$.
- ☐ E. For any $\alpha$, $\mathbb{X}\alpha$ is orthogonal to $\vec{v}$.