Lecture 8

Regression and Linear Algebra

DSC 40A, UCSD

Agenda

- Overview: Spans and projections.
- Regression and linear algebra.
- Multiple linear regression.

Question 🤔

Take a moment to pause and reflect...

If you have any questions please post online to our forms/Q&A site.

Course staff will answer them ASAP!

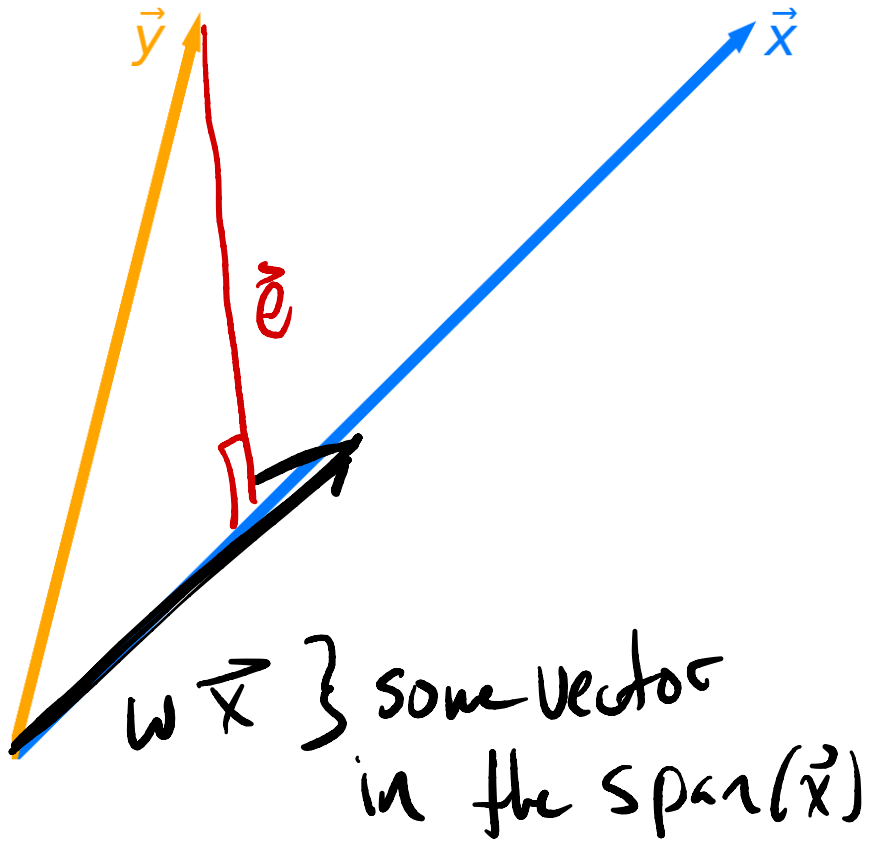
Overview: Spans and projections

Projecting onto the span of a single vector

- **Question**: What vector in $\text{span}(\vec{x})$ is closest to $\vec{y}$?
- The answer is the vector $w\vec{x}$, where the w is chosen to minimize the **length** of the error vector:

$$\|\vec{e}\| = \|\vec{y} - w\vec{x}\|$$

- **Key idea**: To minimize the length of the error vector, choose w so that the error vector is **orthogonal** to $\vec{x}$.



Projecting onto the span of a single vector

- **Question**: What vector in $\text{span}(\vec{x})$ is closest to $\vec{y}$?
- **Answer**: It is the vector $w^* \vec{x}$, where:

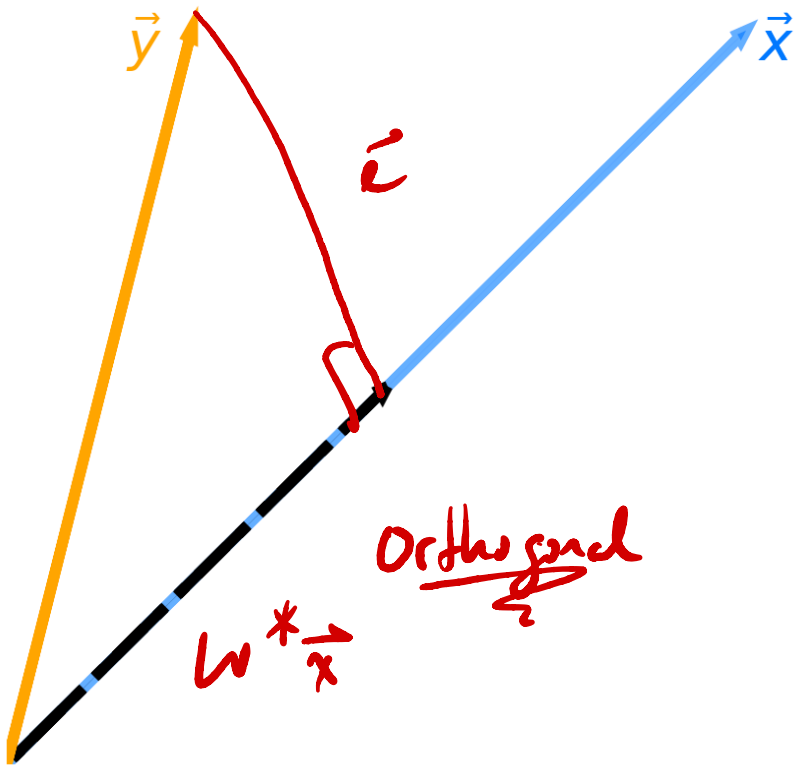
$$w^* = \frac{\vec{x} \cdot \vec{y}}{\vec{x} \cdot \vec{x}}$$

w^* = a scalar

to find w^*

$$\vec{x} \cdot (\vec{y} - w^* \vec{x}) = 0$$

$\vec{e}$

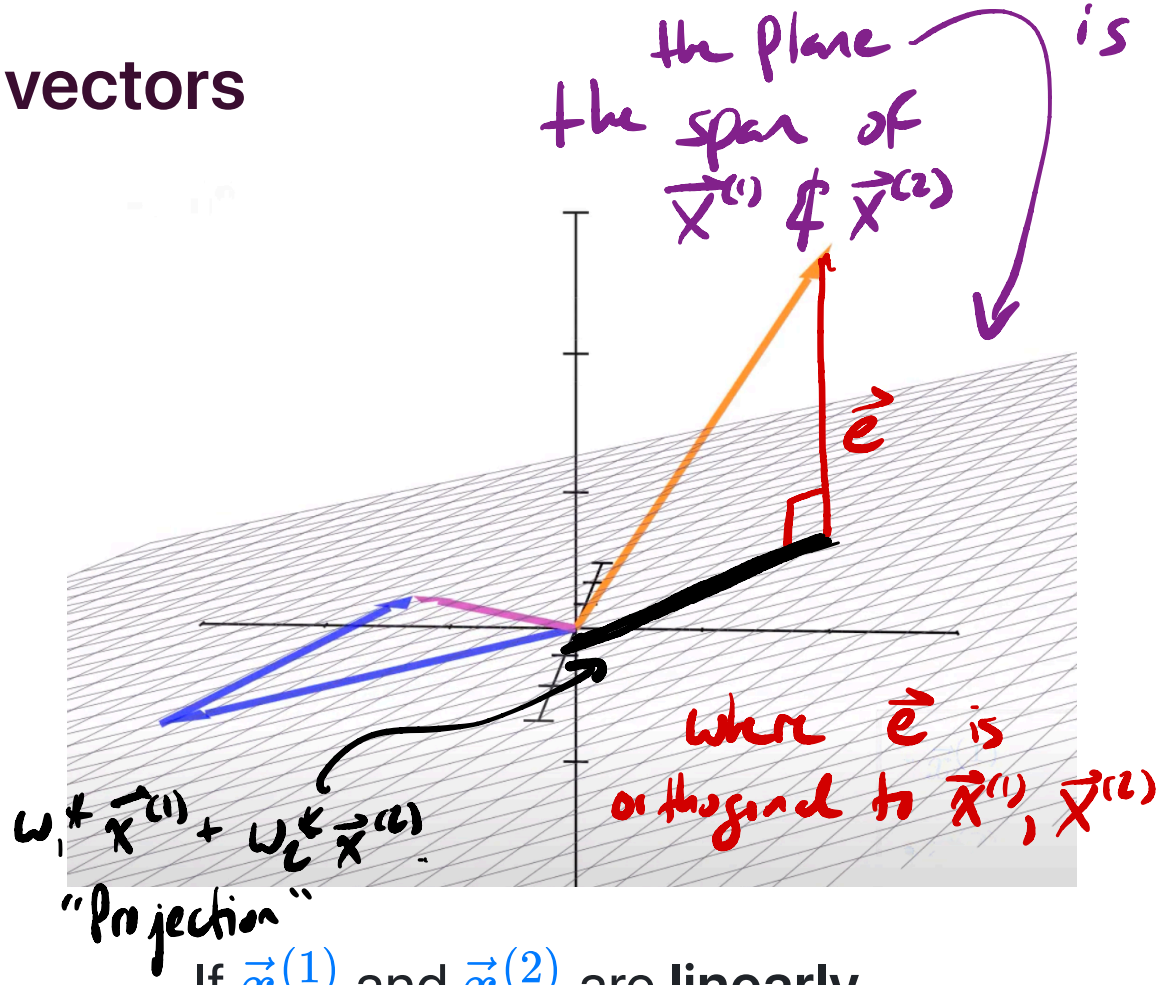


Projecting onto the span of multiple vectors

- **Question**: What vector in $\text{span}(\vec{x}^{(1)}, \vec{x}^{(2)})$ is closest to $\vec{y}$?
- The answer is the vector $w_1\vec{x}^{(1)} + w_2\vec{x}^{(2)}$, where w_1 and w_2 are chosen to minimize the **length** of the error vector:

$$\|\vec{e}\| = \|\vec{y} - w_1\vec{x}^{(1)} - w_2\vec{x}^{(2)}\|$$

- **Key idea**: To minimize the length of the error vector, choose w_1 and w_2 so that the error vector is **orthogonal** to **both** $\vec{x}^{(1)}$ **and** $\vec{x}^{(2)}$.



If $\vec{x}^{(1)}$ and $\vec{x}^{(2)}$ are **linearly independent**, they span a **plane**.

Matrix-vector products create linear combinations of columns!

- **Question**: What vector in $\text{span}(\vec{x}^{(1)}, \vec{x}^{(2)})$ is closest to $\vec{y}$?
- To help, we can create a **matrix**, X, by stacking $\vec{x}^{(1)}$ and $\vec{x}^{(2)}$ next to each other:

$$X = \begin{bmatrix} | & | \\ \vec{x}^{(1)} & \vec{x}^{(2)} \\ | & | \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ 5 & 0 \\ 3 & 4 \end{bmatrix} \quad \vec{y} = \begin{bmatrix} 1 \\ 3 \\ 9 \end{bmatrix}$$

(3x2)

- Then, instead of writing vectors in $\text{span}(\vec{x}^{(1)}, \vec{x}^{(2)})$ as $w_1\vec{x}^{(1)} + w_2\vec{x}^{(2)}$, we can say:

$$X\vec{w} \qquad \text{where } \vec{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

(3x2)(2x1) (2x1)

- **Key idea**: Find $\vec{w}$ such that the error vector, $\vec{e} = \vec{y} - X\vec{w}$, is **orthogonal** to **every column of** X.

$X\vec{w} = w_1\vec{x}^{(1)} + w_2\vec{x}^{(2)}$

Constructing an orthogonal error vector

- **Key idea**: Find $\vec{w} \in \mathbb{R}^d$ such that the error vector, $\vec{e} = \vec{y} - X\vec{w}$, is **orthogonal** to the **columns of** X.
  - Why? Because this will make the error vector as short as possible.
- The $\vec{w}^*$ that accomplishes this satisfies:

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

A vector

$X^T(\vec{y} - X\vec{w}) = 0$

  - Why? Because $X^T\vec{e}$ contains the **dot products** of each column in X with $\vec{e}$. If these are all 0, then $\vec{e}$ is **orthogonal** to **every column of** X!

$$X = \begin{bmatrix} 1 & 1 \\ \vec{x}^{(1)} & \vec{x}^{(2)} \\ 1 & 1 \end{bmatrix}_{(3\times 2)} \Rightarrow X^T\vec{e} = \begin{bmatrix} - & \vec{x}^{(1)T} & - \\ - & \vec{x}^{(2)T} & - \end{bmatrix}\vec{e} = \begin{bmatrix} \vec{x}^{(1)T}\vec{e} \\ \vec{x}^{(2)T}\vec{e} \end{bmatrix}_{(2\times 1)}$$

$\vec{x} \cdot e$

The normal equations

- **Key idea**: Find $\vec{w} \in \mathbb{R}^d$ such that the error vector, $\vec{e} = \vec{y} - X\vec{w}$, is **orthogonal** to the **columns of** X.

- The $\vec{w}^*$ that accomplishes this satisfies:

$$X^T\vec{e} = 0$$
$$X^T(\vec{y} - X\vec{w}^*) = 0$$
$$X^T\vec{y} - X^TX\vec{w}^* = 0$$
$$\implies X^TX\vec{w}^* = X^T\vec{y}$$

- The last statement is referred to as the **normal equations**.

- Assuming X^TX is invertible, this is the vector:

$$\boxed{\vec{w}^* = (X^TX)^{-1}X^T\vec{y}}$$

  - This is a big assumption, because it requires X^TX to be **full rank**.
  - If X^TX is not full rank, then there are infinitely many solutions to the normal equations, $X^TX\vec{w}^* = X^T\vec{y}$.

What does it mean?

- **Original question**: What vector in $\text{span}(\vec{x}^{(1)}, \vec{x}^{(2)})$ is closest to $\vec{y}$?
- **Final answer**: Assuming $X^T X$ is invertible, it is the vector $X\vec{w}^*$, where:

$$\vec{w}^* = (X^T X)^{-1} X^T \vec{y}$$

- Revisiting our example:

$$X = \begin{bmatrix} | & | \\ \vec{x}^{(1)} & \vec{x}^{(2)} \\ | & | \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ 5 & 0 \\ 3 & 4 \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} 1 \\ 3 \\ 9 \end{bmatrix}$$

- Using a computer gives us $\vec{w}^* = (X^T X)^{-1} X^T \vec{y} \approx \begin{bmatrix} 0.7289 \\ 1.6300 \end{bmatrix}$.
- So, the vector in $\text{span}(\vec{x}^{(1)}, \vec{x}^{(2)})$ closest to $\vec{y}$ is $0.7289\vec{x}^{(1)} + 1.6300\vec{x}^{(2)}$.

An optimization problem, solved

- We just used linear algebra to solve an **optimization problem**.
- Specifically, the function we minimized is:

$$\text{error}(\vec{w}) = \|\vec{y} - X\vec{w}\|$$

  - This is a function whose input is a vector, $\vec{w}$, and whose output is a scalar!

- The input, $\vec{w}^*$, to $\text{error}(\vec{w})$ that minimizes it is one that satisfies the **normal equations**:

$$X^T X\vec{w}^* = X^T\vec{y}$$

  If $X^T X$ is invertible, then the unique solution is:

$$\vec{w}^* = (X^T X)^{-1} X^T\vec{y}$$

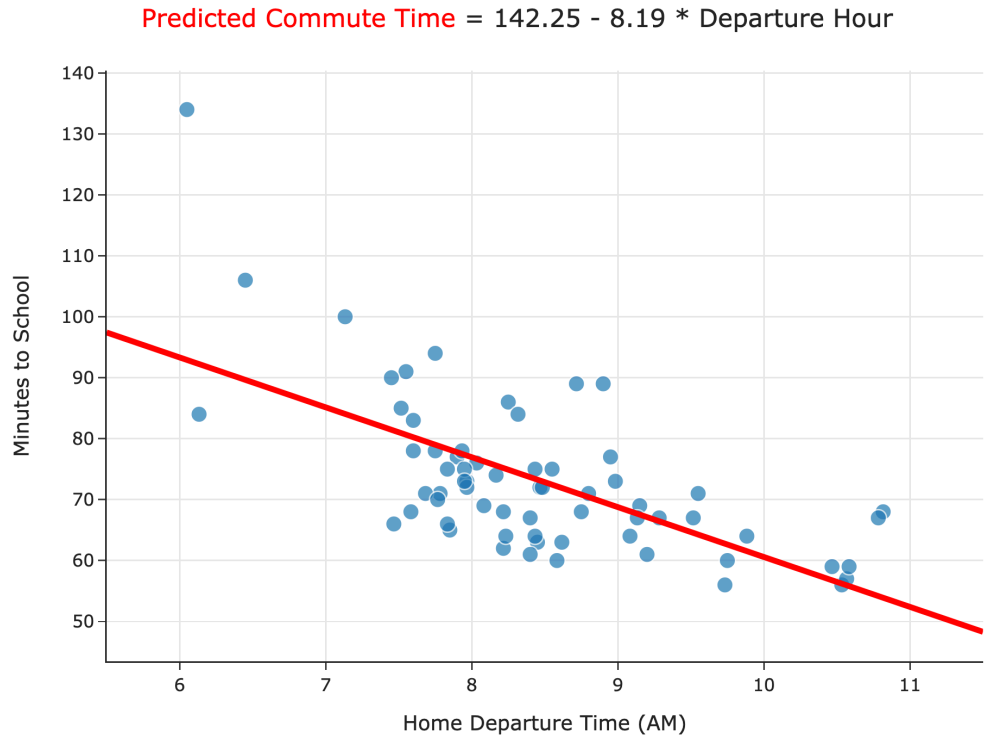
- We're going to use this frequently!

Regression and linear algebra

Wait... why do we need linear algebra?

- Soon, we'll want to make predictions using more than one feature.
  - Example: Predicting commute times using departure hour and temperature.
- Thinking about linear regression in terms of **matrices and vectors** will allow us to find hypothesis functions that:
  - Use multiple features (input variables).
  - Are non-linear in the features, e.g. $H(x) = w_0 + w_1 x + w_2 x^2$.
- Let's see if we can put what we've just learned to use.

Simple linear regression, revisited



Intercept

slope

- **Model**: $H(x) = w_0 + w_1 x.$
- **Loss function**: $(y_i - H(x_i))^2.$
- To find w_0^* and w_1^*, we minimized empirical risk, i.e. average loss:

Best!!!

$$R_{\text{sq}}(H) = \frac{1}{n} \sum_{i=1}^{n} (y_i - H(x_i))^2$$

AVG loss

- **Observation**: $R_{\text{sq}}(w_0, w_1)$ *kind of* looks like the formula for the norm of a vector,

generalized Pythagorean theorem

$$\|\vec{v}\| = \sqrt{v_1^2 + v_2^2 + \ldots + v_n^2}.$$

n = Observations = rows in Dataset

Regression and linear algebra

$\vec{y} = \begin{bmatrix} 25 \\ 30 \\ 51 \\ \vdots \end{bmatrix}_{1 \times n}$

Let's define a few new terms:

- The **observation vector** is the vector $\vec{y} \in \mathbb{R}^n$. This is the vector of observed "actual values".
- The **hypothesis vector** is the vector $\vec{h} \in \mathbb{R}^n$ with components $H(x_i)$. This is the vector of predicted values.
- The **error vector** is the vector $\vec{e} \in \mathbb{R}^n$ with components:

$$e_i = y_i - H(x_i)$$

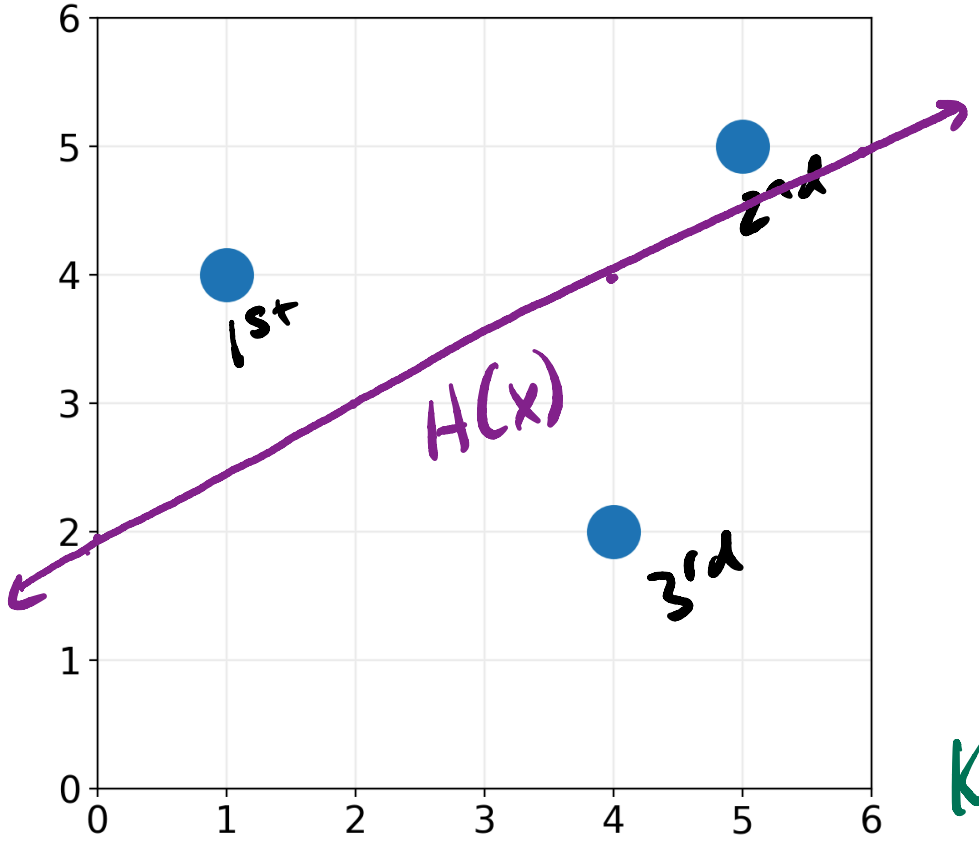
$\vec{h}$

$\vec{h} = \begin{bmatrix} 26 \\ 38 \\ 55 \\ \vdots \end{bmatrix}_{1 \times n}$

in min for our Data

Example

Consider $H(x) = 2 + \frac{1}{2}x$.



$$\vec{y} = \begin{bmatrix} 4 \\ 5 \\ 2 \end{bmatrix} \quad 1\times 3$$

$$\vec{h} = \begin{bmatrix} 2+\frac{1}{2}(1) \\ 2+\frac{1}{2}(5) \\ 2+\frac{1}{2}(4) \end{bmatrix} = \begin{bmatrix} \frac{5}{2} \\ \frac{9}{2} \\ 4 \end{bmatrix} \quad 1\times 3$$

$$\vec{e} = \vec{y} - \vec{h} = \begin{bmatrix} 4 - \frac{5}{2} \\ 5 - \frac{9}{2} \\ 2 - 4 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 0.5 \\ -2 \end{bmatrix}$$

$$R_{\text{sq}}(H) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - H(x_i)\right)^2$$

Key Idea:

$$R_{sq}(H) = \frac{1}{3}\|\vec{e}\|^2 = \frac{1}{3}\left[1.5^2 + 0.5^2 + (-2)^2\right]$$

Regression and linear algebra

Let's define a few new terms:

- The **observation vector** is the vector $\vec{y} \in \mathbb{R}^n$. This is the vector of observed "actual values".
- The **hypothesis vector** is the vector $\vec{h} \in \mathbb{R}^n$ with components $H(x_i)$. This is the vector of predicted values.
- The **error vector** is the vector $\vec{e} \in \mathbb{R}^n$ with components:

$$e_i = y_i - H(x_i)$$

- **Key idea**: We can rewrite the mean squared error of H as:

$$R_{\text{sq}}(H) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - H(x_i)\right)^2 = \frac{1}{n}\|\vec{e}\|^2 = \frac{1}{n}\|\vec{y} - \vec{h}\|^2$$

The hypothesis vector

- The **hypothesis vector** is the vector $\vec{h} \in \mathbb{R}^n$ with components $H(x_i)$. This is the vector of predicted values.
- For the linear hypothesis function $H(x) = w_0 + w_1 x$, the hypothesis vector can be written:

$$\vec{h} = \begin{bmatrix} w_0 + w_1 x_1 \\ w_0 + w_1 x_2 \\ \vdots \\ w_0 + w_1 x_n \end{bmatrix}_{n \times 1} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}_{n \times 2} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}_{2 \times 1}$$

Design Matrix

Parameter Vector $\vec{w}$

$H(x_i) = w_0 + w_1 x_i$

Rewriting the mean squared error

- Define the **design matrix** $X \in \mathbb{R}^{n \times 2}$ as:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

- Define the **parameter vector** $\vec{w} \in \mathbb{R}^2$ to be $\vec{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$. w_0 → y intercept; w_1 → slope

- Then, $\vec{h} = X\vec{w}$, so the mean squared error becomes:

$$R_{\text{sq}}(H) = \frac{1}{n}\|\vec{y} - \vec{h}\|^2 \implies \boxed{R_{\text{sq}}(\vec{w}) = \frac{1}{n}\|\vec{y} - X\vec{w}\|^2}$$

Minimizing mean squared error, again

- To find the optimal model parameters for simple linear regression, w_0^* and w_1^*, we previously minimized:

$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))^2$$

- Now that we've reframed the simple linear regression problem in terms of linear algebra, we can find w_0^* and w_1^* by finding the $\vec{w}^* = \begin{bmatrix} w_0^* & w_1^* \end{bmatrix}^T$ that minimizes:

$$\boxed{R_{\text{sq}}(\vec{w}) = \frac{1}{n}\|\vec{y} - X\vec{w}\|^2}$$

- Do we already know the $\vec{w}^*$ that minimizes $R_{\text{sq}}(\vec{w})$?

Best: intercept slope

An optimization problem we've seen before

- The optimal parameter vector, $\vec{w}^* = \begin{bmatrix} w_0^* & w_1^* \end{bmatrix}^T$, is the one that minimizes:

$$R_{\text{sq}}(\vec{w}) = \frac{1}{n} \|\vec{y} - X\vec{w}\|^2$$

- Previously, we found that $\vec{w}^* = (X^T X)^{-1} X^T \vec{y}$ minimizes the length of the error vector, $\|\vec{e}\| = \|\vec{y} - X\vec{w}\|$
- $R_{\text{sq}}(\vec{w})$ is closely related to $\|\vec{e}\|$:

$$R_{sq}(\vec{w}) = \frac{1}{n} \|\vec{e}\|^2$$

minimize $\|\vec{y} - X\vec{w}\|$ = minimizing $\frac{1}{n} \|\vec{y} - X\vec{w}\|^2$

- The minimizer of $\|\vec{e}\|$ is the same as the minimizer of $R_{\text{sq}}(\vec{w})$!
- **Key idea**: $\vec{w}^* = (X^T X)^{-1} X^T \vec{y}$ also **minimizes** $R_{\text{sq}}(\vec{w})$!

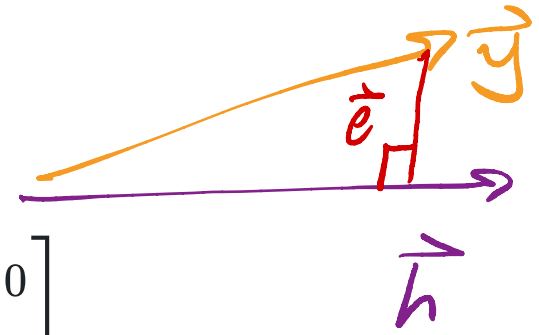
The optimal parameter vector, $\vec{w}^*$

- To find the optimal model parameters for simple linear regression, w_0^* and w_1^*, we previously minimized $R_{\text{sq}}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))^2$.
  - We found, using calculus, that:
    - $\boxed{w_1^* = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r\frac{\sigma_y}{\sigma_x}}$.
    - $\boxed{w_0^* = \bar{y} - w_1^*\bar{x}}$.
- Another way of finding optimal model parameters for simple linear regression is to find the $\vec{w}^*$ that minimizes $R_{\text{sq}}(\vec{w}) = \frac{1}{n}\|\vec{y} - X\vec{w}\|^2$.
  - The minimizer, if X^TX is invertible, is the vector $\boxed{\vec{w}^* = (X^TX)^{-1}X^T\vec{y}}$.
- These formulas are equivalent!

Summary: Regression and linear algebra

- Define the **design matrix** $X \in \mathbb{R}^{n \times 2}$, **observation vector** $\vec{y} \in \mathbb{R}^n$, and parameter vector $\vec{w} \in \mathbb{R}^2$ as:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \vec{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$



- How do we make the hypothesis vector, $\vec{h} = X\vec{w}$, as close to $\vec{y}$ as possible? Use the parameter vector $\vec{w}^*$:

$$\vec{w}^* = (X^T X)^{-1} X^T \vec{y}$$

- We chose $\vec{w}^*$ so that $\vec{h}^* = X\vec{w}^*$ is **the projection of $\vec{y}$ onto the span of the columns of the design matrix, X.**

Multiple linear regression

	departure_hour	day_of_month	minutes
0	10.816667	15	68.0
1	7.750000	16	94.0
2	8.450000	22	63.0
3	7.133333	23	100.0
4	9.150000	30	69.0
...	...	...	...

So far, we've fit **simple** linear regression models, which use only **one** feature (`'departure_hour'`) for making predictions.

Incorporating multiple features

- In the context of the commute times dataset, the simple linear regression model we fit was of the form:

$$\begin{aligned}\text{pred. commute} &= H(\text{departure hour}) \\ &= w_0 + w_1 \cdot \text{departure hour}\end{aligned}$$

- Now, we'll try and fit a multiple linear regression model of the form:

$$\begin{aligned}\text{pred. commute} &= H(\text{departure hour}) \\ &= w_0 + w_1 \cdot \text{departure hour} + w_2 \cdot \text{day of month}\end{aligned}$$

- Linear regression with **multiple** features is called **multiple linear regression**.
- How do we find w_0^*, w_1^*, and w_2^*?

Geometric interpretation

- The hypothesis function:

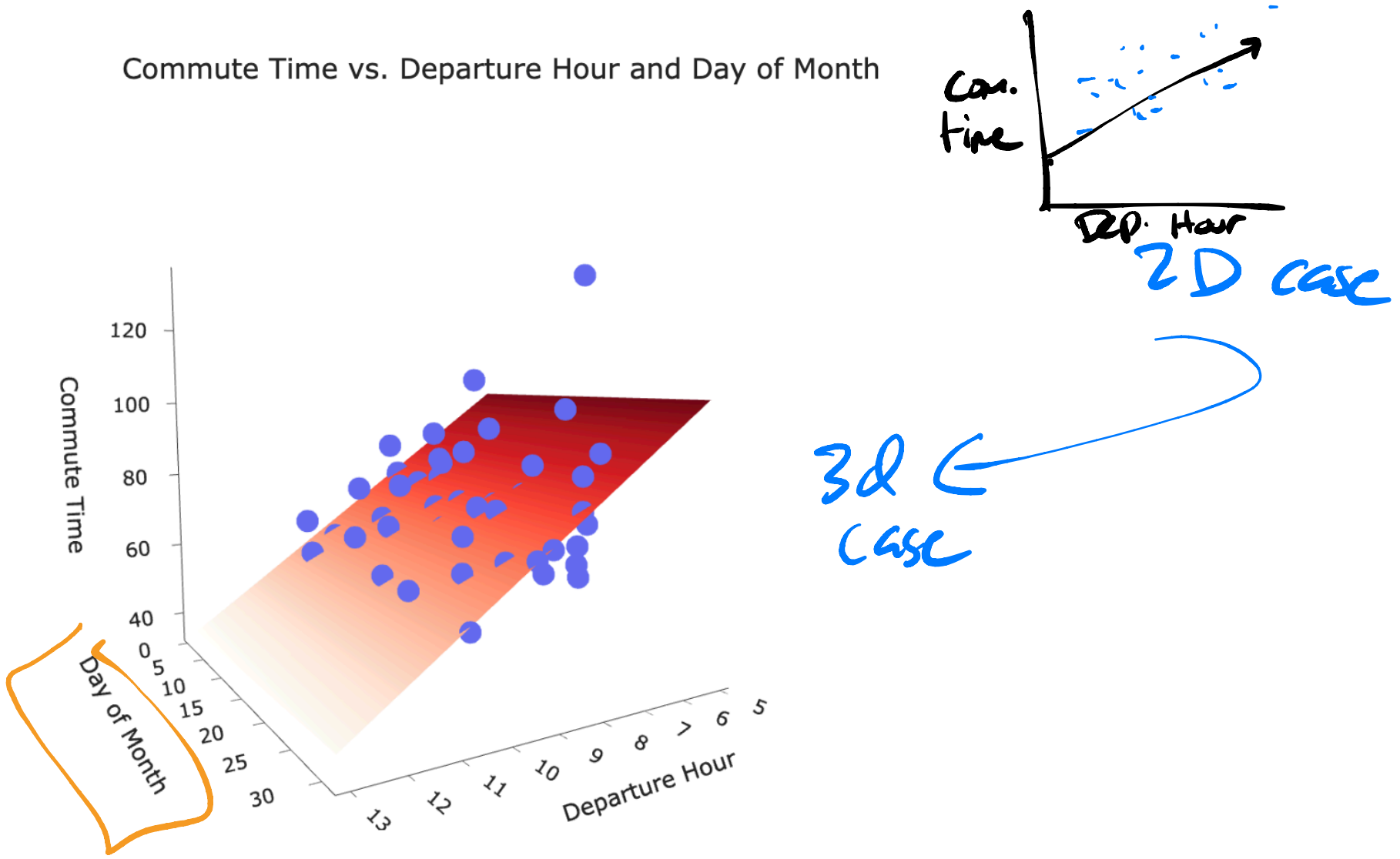
$$H(\text{departure hour}) = w_0 + w_1 \cdot \text{departure hour}$$

  looks like a **line** in 2D.

- **Questions**:
  - How many dimensions do we need to graph the hypothesis function:

$$H(\text{departure hour}) = w_0 + w_1 \cdot \text{departure hour} + w_2 \cdot \text{day of month}$$

  - What is the shape of the hypothesis function?



Our new hypothesis function is a **plane** in 3D!

The setup

- Suppose we have the following dataset.

row	departure_hour	day_of_month	minutes
1	8.45	22	63.0
2	8.90	28	89.0
3	8.72	18	89.0

- We can represent each day with a **feature vector**, $\vec{x}$:

The hypothesis vector

- When our hypothesis function is of the form:

$$H(\text{departure hour}) = w_0 + w_1 \cdot \text{departure hour} + w_2 \cdot \text{day of month}$$

the hypothesis vector $\vec{h} \in \mathbb{R}^n$ can be written as:

$$\vec{h} = \begin{bmatrix} H(\text{departure hour}_1, \text{day}_1) \\ H(\text{departure hour}_2, \text{day}_2) \\ \dots \\ H(\text{departure hour}_n, \text{day}_n) \end{bmatrix} = \begin{bmatrix} 1 & \text{departure hour}_1 & \text{day}_1 \\ 1 & \text{departure hour}_2 & \text{day}_2 \\ \dots & \dots & \dots \\ 1 & \text{departure hour}_n & \text{day}_n \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$$

x_1 x_2 w_0 intercept x_1 x_2 find

Finding the optimal parameters

- To find the optimal parameter vector, $\vec{w}^*$, we can use the **design matrix** $X \in \mathbb{R}^{n \times 3}$ and **observation vector** $\vec{y} \in \mathbb{R}^n$:

$$X = \begin{bmatrix} 1 & \text{departure hour}_1 & \text{day}_1 \\ 1 & \text{departure hour}_2 & \text{day}_2 \\ \dots & \dots & \dots \\ 1 & \text{departure hour}_n & \text{day}_n \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} \text{commute time}_1 \\ \text{commute time}_2 \\ \vdots \\ \text{commute time}_n \end{bmatrix}$$

w_0 w_1 w_2

$w_1 = 25.21$

$w_2 = 0.005$

Param. vector: $\begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$

- Then, all we need to do is solve the **normal equations**:

$$X^T X \vec{w}^* = X^T \vec{y}$$

If $X^T X$ is invertible, we know the solution is:

$$\vec{w}^* = (X^T X)^{-1} X^T \vec{y}$$

Roadmap

- To wrap up today's lecture, we'll find the optimal parameter vector $\vec{w}^*$ for our new two-feature model in code. We'll switch back to our notebook, linked here.
- Next class, we'll present a more general framing of the multiple linear regression model, that uses d features instead of just two.
- We'll also look at how we can **engineer** new features using existing features.
    - e.g. How can we fit a hypothesis function of the form $H(x) = w_0 + w_1 x + w_2 x^2$?