Lecture 8

Regression and Linear Algebra

DSC 40A, Summer 2024

Announcements

- Homework 3 is due **tomorrow**.
    - ~~We moved some office hours around – we now have some on Saturday~~!

- Midterm next week.

Agenda

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

Question 🤔

Answer at q.dsc40a.com

Remember, you can always ask questions at q.dsc40a.com!

If the direct link doesn't work, click the "🤔 Lecture Questions" link in the top right corner of dsc40a.com.

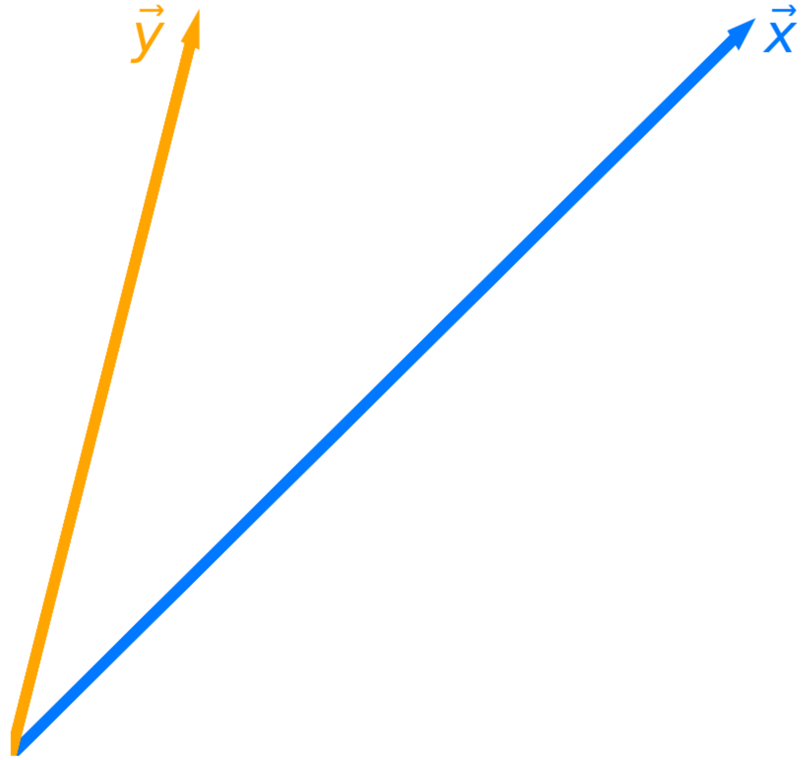
Overview: Spans and projections

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:

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

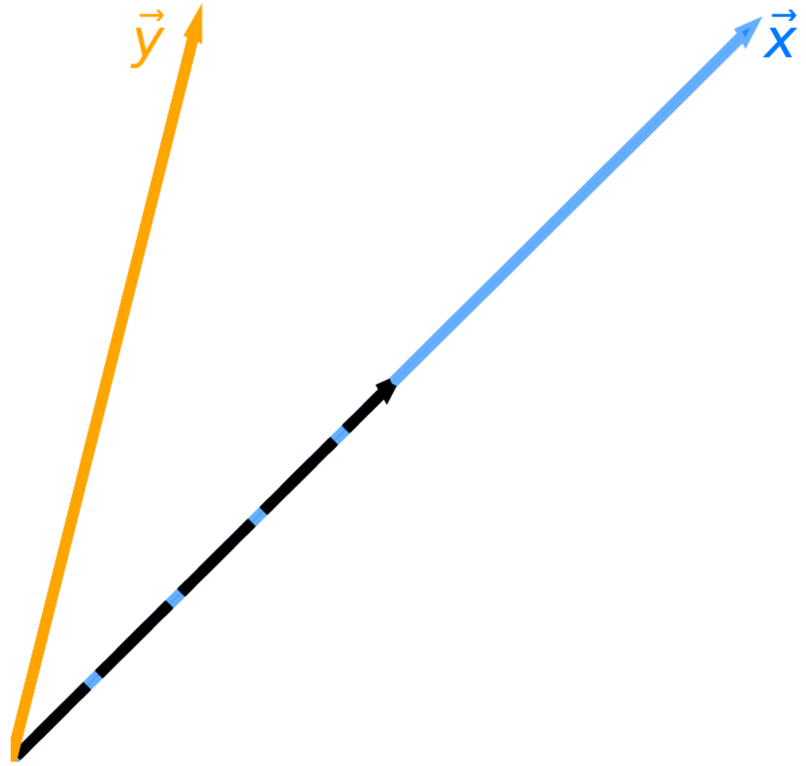
- **Key idea**: To minimize the length of the error vector, choose w so that the error vector is **orthogonal** to $\vec{x}$.



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:

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

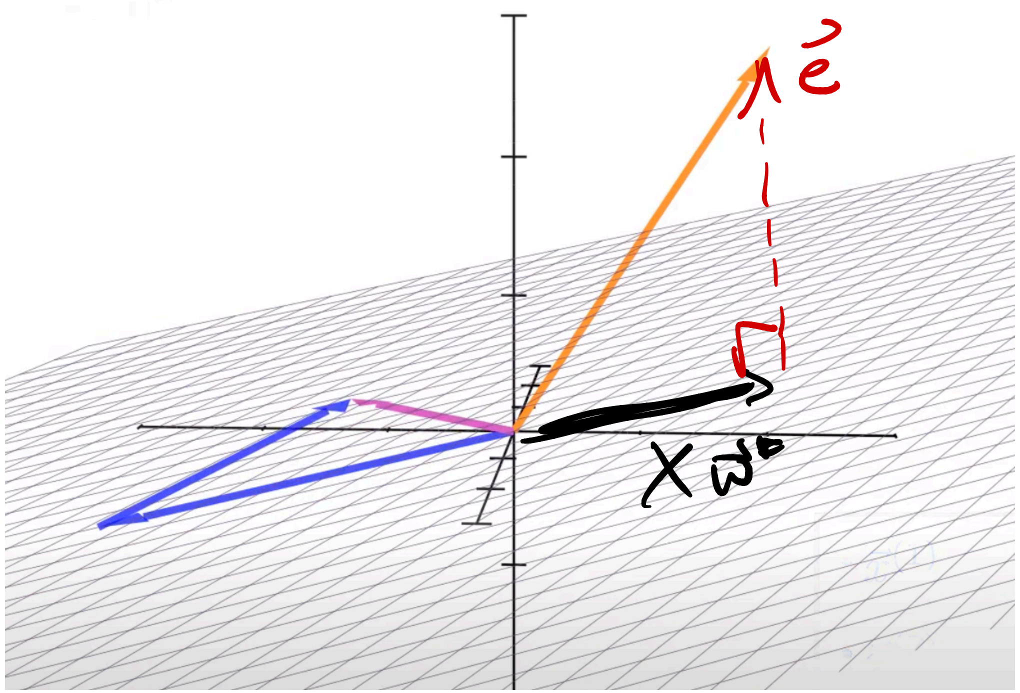


Projecting onto the span of multiple vectors

- **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:

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

- 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:

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

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

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:

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

    - 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**!

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

just dot products

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:

$$\begin{aligned} X^T\vec{e} &= \vec{0} \\ X^T(\vec{y} - X\vec{w}^*) &= \vec{0} \\ X^T\vec{y} - X^TX\vec{w}^* &= \vec{0} \\ \implies \quad X^TX\vec{w}^* &= X^T\vec{y} \end{aligned}$$

- 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**. = invertible
  - 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^TX is invertible, it is the vector $X\vec{w}^*$, where:

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

- 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 $\vec{w}^* = (X^TX)^{-1}X^T\vec{y} \approx \begin{bmatrix} 0.7289 \\ 1.6300 \end{bmatrix}$.
- 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**:

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

  If X^TX is invertible, then the unique solution is:

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

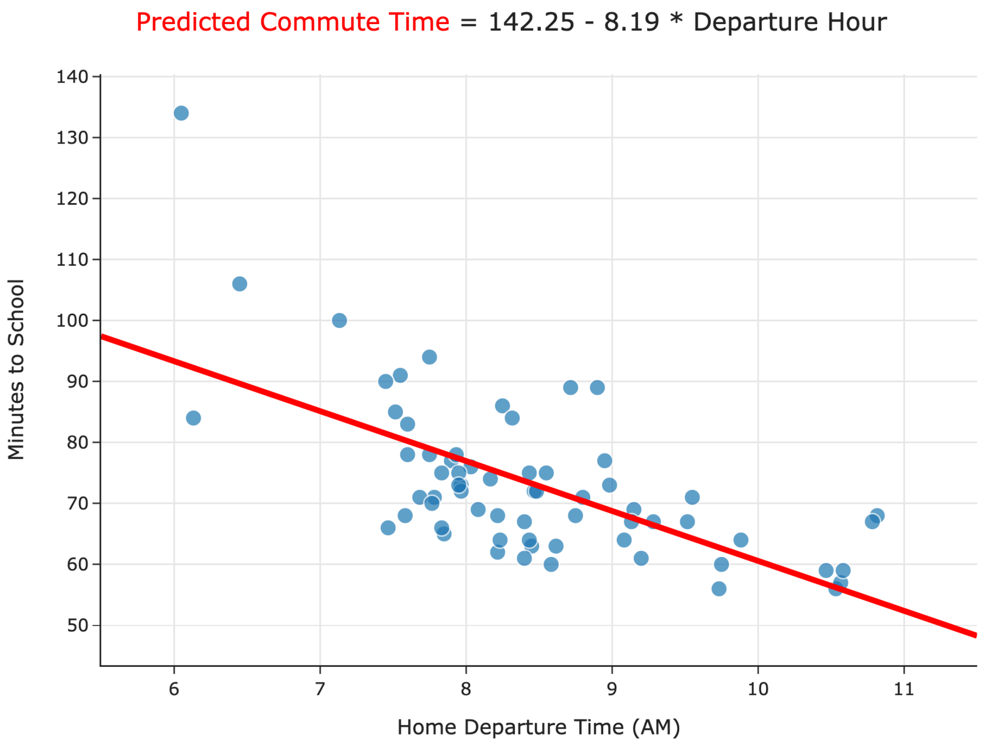
- We're going to use this frequently!

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**: $H(x) = w_0 + w_1 x$.

actual   predicted

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

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

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

$$\underbrace{\|\vec{v}\| = \sqrt{v_1^2 + v_2^2 + \ldots + v_n^2}}_{\text{generalized Pythagorean theorem}}.$$

Regression and linear algebra

Let's define a few new terms:

n rows in my dataset

- 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:

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

$$\vec{e} = \vec{y} - \vec{h}$$

$$\vec{y} = \begin{bmatrix} 42 \text{ min} \\ 74 \text{ min} \\ \vdots \end{bmatrix}_{n \times 1}$$

actual

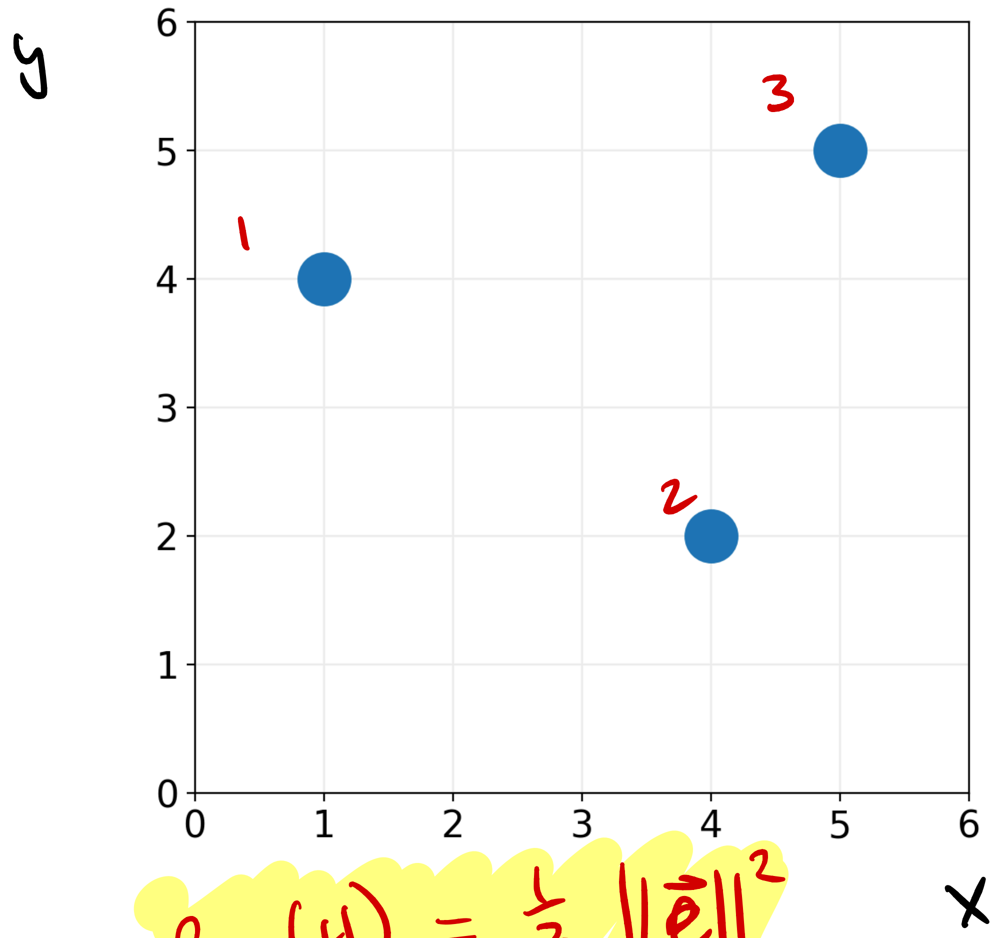
$$\vec{h} = \begin{bmatrix} 52 \text{ min} \\ 71 \text{ min} \\ \vdots \end{bmatrix}_{n \times 1}$$

predicted

Example

not necessarily the optimal line

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



$$R_{sq}(H) = \frac{1}{3}\|\vec{e}\|^2$$

$$\vec{y} = \begin{bmatrix} 4 \\ 2 \\ 5 \end{bmatrix} \qquad \vec{h} = \begin{bmatrix} 2 + \frac{1}{2} \cdot 1 \\ 2 + \frac{1}{2} \cdot 4 \\ 2 + \frac{1}{2} \cdot 5 \end{bmatrix} = \begin{bmatrix} \frac{5}{2} \\ 4 \\ \frac{9}{2} \end{bmatrix}$$

$$\vec{e} = \vec{y} - \vec{h} = \begin{bmatrix} 4 - \frac{5}{2} \\ 2 - 4 \\ 5 - \frac{9}{2} \end{bmatrix} = \begin{bmatrix} \frac{3}{2} \\ -2 \\ \frac{1}{2} \end{bmatrix}$$

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

$$= \frac{1}{3}\left[\left(\frac{3}{2}\right)^2 + (-2)^2 + \left(\frac{1}{2}\right)^2\right]$$

$$\vec{e} = \begin{bmatrix} \frac{3}{2} \\ -2 \\ \frac{1}{2} \end{bmatrix}$$

$$\|\vec{e}\| = \sqrt{\left(\frac{3}{2}\right)^2 + (-2)^2 + \left(\frac{1}{2}\right)^2}$$

pythagorean thm but w/ 3 things

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

$$\frac{1}{3}\|\vec{e}\|^2 = R_{sq}(H)$$

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:

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

$\|\cdot\| \to$ magnitude
length
norm ☆

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

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

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:

intercepts slopes all same

$H(x_i) = w_0 \cdot 1 + w_1 x_i$

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

intercept

slope

x_i's are different!

Still

$H(x_i) = w_0 + w_1 x_i$

Rewriting the mean squared error

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

$$X = \begin{array}{c} \begin{array}{cc} w_0 & w_1 \end{array} \\ \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \end{array}$$

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

- 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:

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

- 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 $R_{\text{sq}}(\vec{w})$?

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:

Empirical risk

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

as a vector norm

- Previously, we found that $\vec{w}^* = (X^TX)^{-1}X^T\vec{y}$ minimizes the length of the error vector, $\|\vec{e}\| = \|\vec{y} - X\vec{w}\|$

Motivation: project $\vec{y}$ onto span $\{\vec{x}_i\}$s

- $R_{\text{sq}}(\vec{w})$ is closely related to $\|\vec{e}\|$:

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

Minimizing $\|\vec{y} - X\vec{w}\|$ is the same as 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**: $\vec{w}^* = (X^TX)^{-1}X^T\vec{y}$ 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:
    - $\boxed{w_1^* = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = r \frac{\sigma_y}{\sigma_x}}$. best slope
    - $\boxed{w_0^* = \bar{y} - w_1^* \bar{x}}$. best intercept
- 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!

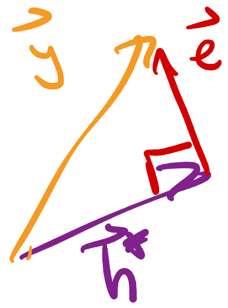
Roadmap

- To give us a break from math, we'll switch to a notebook, linked here, showing that both formulas – that is, (1) the formulas for w_1^* and w_0^* we found using calculus, and (2) the formula for $\vec{w}^*$ we found using linear algebra – give the same results.
- Then, we'll use our new linear algebraic formulation of regression to incorporate **multiple features** in our prediction process.

$H(x_i) = w_0 + w_1 x_i$

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 $\vec{w}^*$:

best predictions

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

- 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:

1 input

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

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

2 inputs

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

- Linear regression with **multiple** features is called **multiple linear regression**.

- How do we find $\underbrace{w_0^*, w_1^*, \text{and } w_2^*}_{}$?

with the Normal Equations!

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:

, day of month)

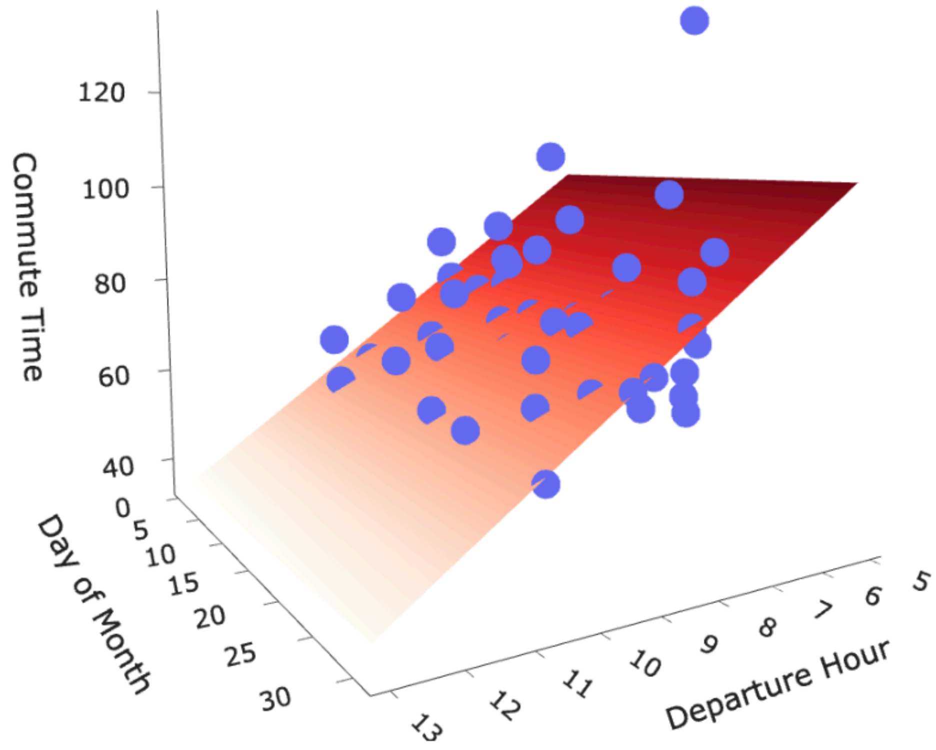
$$H(\text{departure hour}) = w_0 + w_1 \cdot \text{departure hour} + w_2 \cdot \text{day of month}$$

  - What is the shape of the hypothesis function?

$$z = a + bx + cy$$

w_0 w_1 w_2

Commute Time vs. Departure Hour and Day of Month



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

Our goal is to find the **plane** of best fit that pierces through this cloud of points.

The setup

- Suppose we have the following dataset.

$\vec{y}$

	departure_hour	day_of_month	minutes
row			
1	$\vec{x}_1$ 8.45	22	63.0
2	$\vec{x}_2$ 8.90	28	89.0
3	$\vec{x}_3$ 8.72	18	89.0

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

$$\vec{x}_1 = \begin{bmatrix} 8.45 \\ 22 \end{bmatrix} \qquad \vec{x}_2 = \begin{bmatrix} 8.9 \\ 28 \end{bmatrix} \qquad \vec{x}_3 = \begin{bmatrix} 8.72 \\ 18 \end{bmatrix}$$

$$\vec{h} = X \vec{w}$$

hypothesis design param

The hypothesis vector

- When our hypothesis function is of the form:

day)

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

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

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

$w_0 + w_1 \cdot \text{dep.hour}_2 + w_2 \cdot \text{day}_2$

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** $\vec{y} \in \mathbb{R}^n$:

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

- 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.
- On Monday, 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 $H(x) = w_0 + w_1 x + w_2 x^2$?