Lectures 8-10

Linear algebra: Dot products and Projections

DSC 40A, Fall 2024

Announcements

- Homework 2 was released Friday. Remember that using the Overleaf template is required for Homework 2 (and only Homework 2).
- Groupwork 3 is due **tonight**.
- Check out FAQs page and the tutor-created supplemental resources on the course website.

Agenda

- Recap: Simple linear regression and correlation.
- Connections to related models.
- Dot products.
- Spans and projections.

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.

Simple linear regression

- Model: $H(x) = w_0 + w_1 x$.
- Loss function: squared loss, i.e. $L_{\text{sq}}(y_i, H(x_i)) = (y_i - H(x_i))^2$.
- Average loss, i.e. empirical risk:

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

- Optimal model parameters, found by minimizing empirical risk:

$$w_1^* = \frac{\displaystyle\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\displaystyle\sum_{i=1}^{n} (x_i - \bar{x})^2} = r \frac{\sigma_y}{\sigma_x} \qquad\qquad w_0^* = \bar{y} - w_1^* \bar{x}$$

The correlation coefficient

- The correlation coefficient, r, is defined as the **average of the product of x and y, when both are in standard units**.
- Let σ_x be the standard deviation of the x_is, and $\bar{x}$ be the mean of the x_is.
- x_i in standard units is $\frac{x_i - \bar{x}}{\sigma_x}$.
- The correlation coefficient, then, is:

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

Correlation and mean squared error

- **Claim**: Suppose that w_0^* and w_1^* are the optimal intercept and slope for the regression line. Then,

$$R_{\text{sq}}(w_0^*, w_1^*) = \sigma_y^2(1 - r^2)$$

- That is, the mean squared error of the regression line's predictions and the correlation coefficient, r, always satisfy the relationship above.
- Even if it's true, why do we care?
    - In machine learning, we often use both the mean squared error and r^2 to compare the performances of different models.
    - If we can prove the above statement, we can show that **finding models that minimize mean squared error** is equivalent to **finding models that maximize** r^2.

Proof that $R_{\text{sq}}(w_0^*, w_1^*) = \sigma_y^2(1 - r^2)$

Connections to related models

Exercise

Suppose we choose the model $H(x) = w_0$ and squared loss.

What is the optimal model parameter, w_0^*?

Exercise

Suppose we choose the model $H(x) = w_1 x$ and squared loss.

What is the optimal model parameter, w_1^*?

Comparing mean squared errors

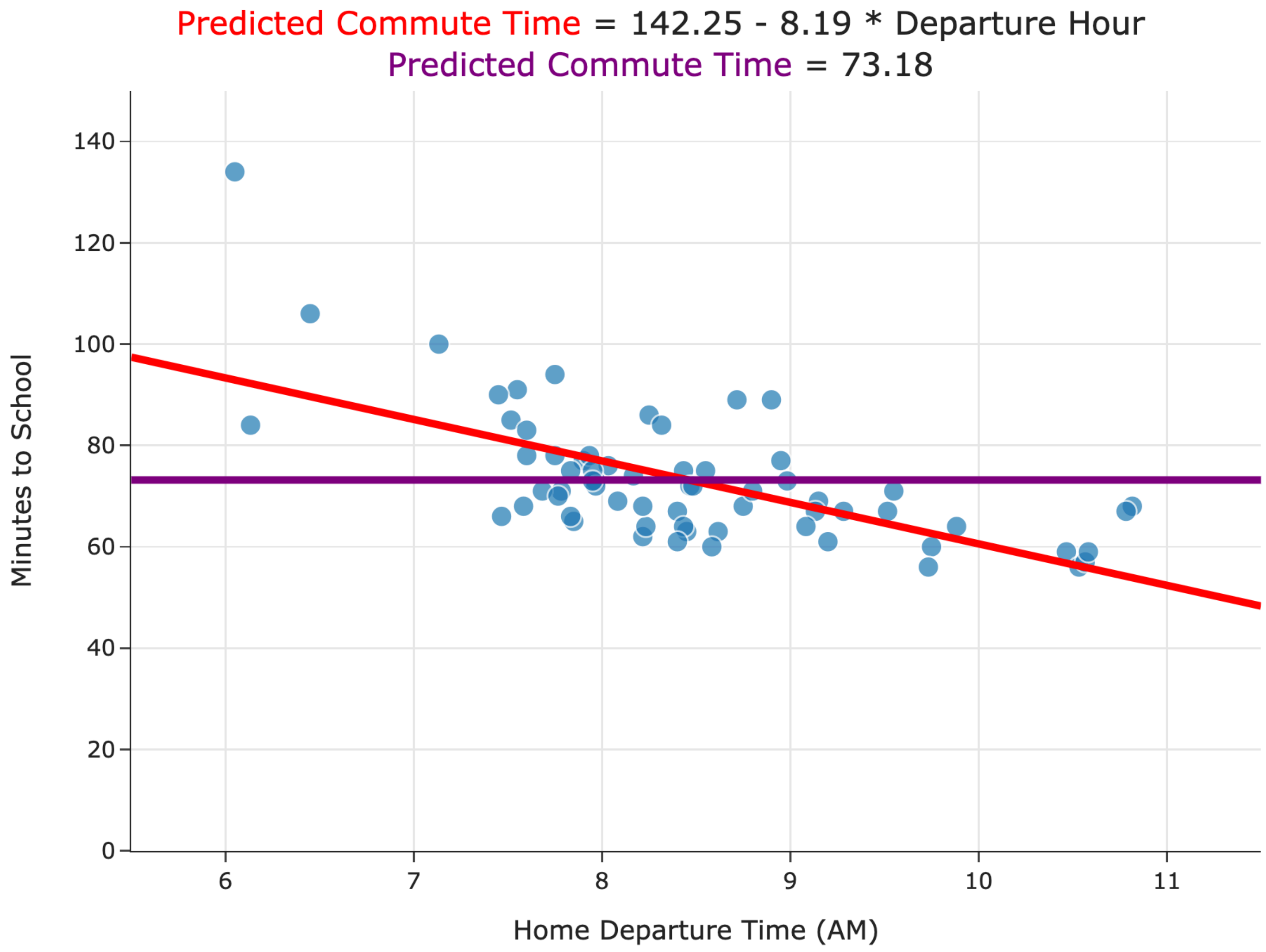
- With both:
    - the constant model, $H(x) = h$, and
    - the simple linear regression model, $H(x) = w_0 + w_1 x$,

  when we chose squared loss, we minimized mean squared error to find optimal parameters:

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

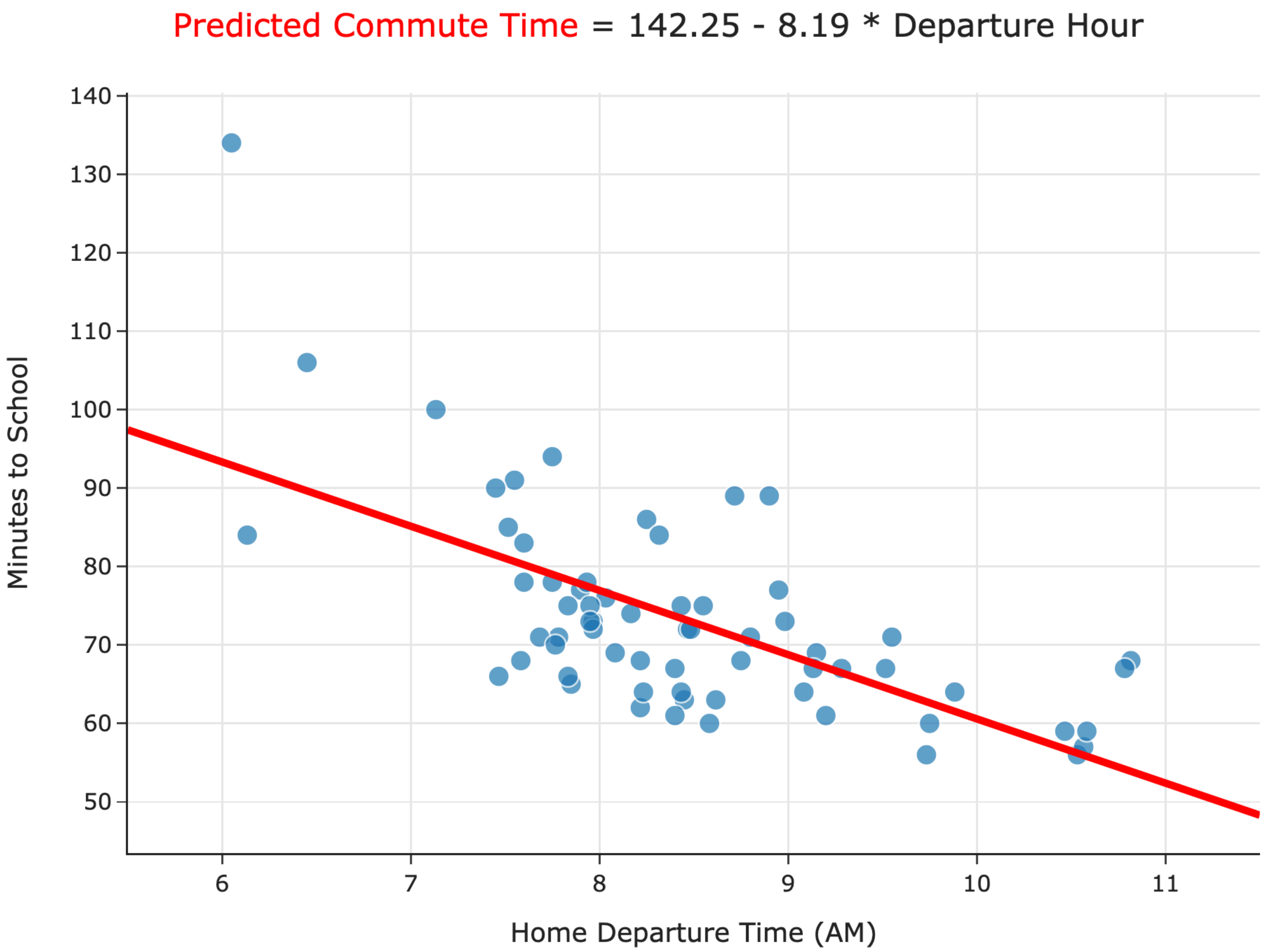
- **Which model minimizes mean squared error more?**

Comparing mean squared errors



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

- The MSE of the best simple linear regression model is ≈ 97
- The MSE of the best constant model is ≈ 167
- The simple linear regression model is a more flexible version of the constant model.



Predicted Commute Time = 142.25 - 8.19 * Departure Hour
Minutes to School
140
130
120
110
100
90
80
70
60
50
6
7
8
9
10
11
Home Departure Time (AM)

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 nonlinear in the features, e.g. $H(x) = w_0 + w_1 x + w_2 x^2$.

Warning ⚠️

- We're **not** going to cover every single detail from your linear algebra course.
- There will be facts that you're expected to remember that we won't explicitly say.
    - For example, if A and B are two matrices, then $AB \neq BA$.
    - This is the kind of fact that we will only mention explicitly if it's directly relevant to what we're studying.
    - But you still need to know it, and it may come up in homework questions.
- We **will** review the topics that you really need to know well.

Dot Products

Vectors

- A **vector** in $\mathbb{R}^n$ is an **ordered collection of n numbers**.
- We use lower-case letters with an arrow on top to represent vectors, and we usually write vectors as **columns**.

$$\vec{v} = \begin{bmatrix} 8 \\ 3 \\ -2 \\ 5 \end{bmatrix}$$

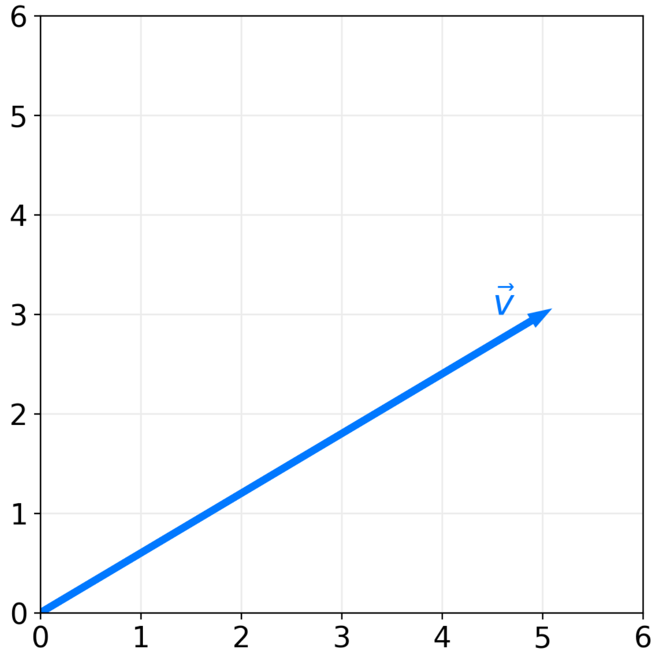
- Another way of writing the above vector is $\vec{v} = [8, 3, -2, 5]^\intercal$.
- Since $\vec{v}$ has four **components**, we say $\vec{v} \in \mathbb{R}^4$.

The geometric interpretation of a vector

- A vector $\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$ is an arrow to the point $(v_1, v_2, \ldots, v_n)$ from the origin.
- The **length**, or L_2 **norm**, of $\vec{v}$ is:

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

- A vector is sometimes described as an object with a **magnitude/length** and **direction**.



Dot product: coordinate definition

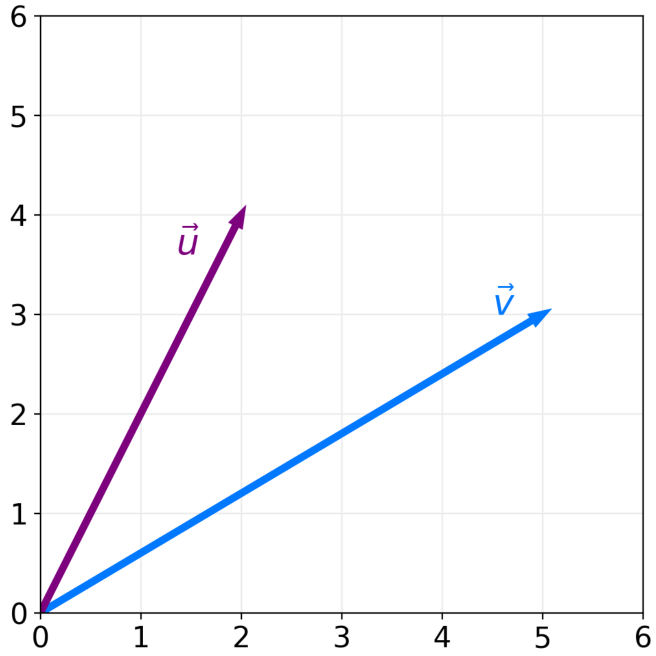
- The **dot product** of two vectors $\vec{u}$ and $\vec{v}$ in $\mathbb{R}^n$ is written as:

$$\vec{u} \cdot \vec{v} = \vec{u}^\intercal \vec{v}$$

- The computational definition of the dot product:

$$\vec{u} \cdot \vec{v} = \sum_{i=1}^{n} u_i v_i = u_1 v_1 + u_2 v_2 + \ldots + u_n v_n$$

- The result is a **scalar**, i.e. a single number.



Dot product: geometric definition

- The computational definition of the dot product:

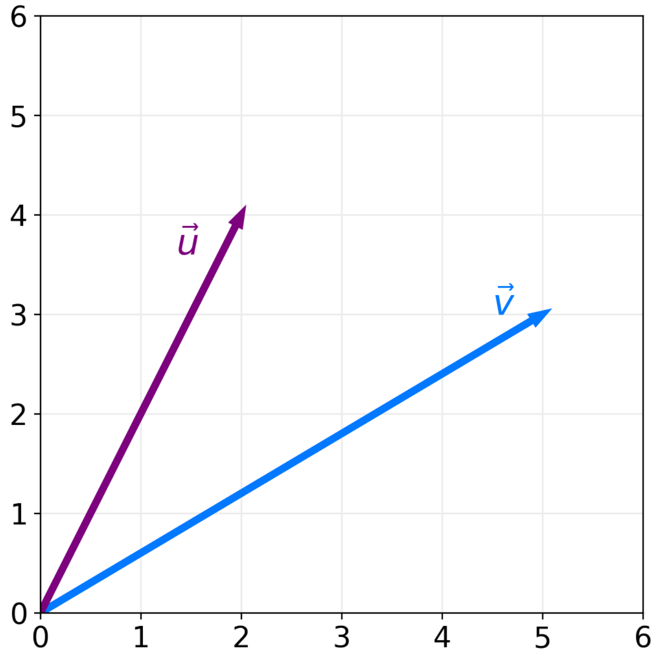
$$\vec{u} \cdot \vec{v} = \sum_{i=1}^{n} u_i v_i = u_1 v_1 + u_2 v_2 + \ldots + u_n v_n$$

- The geometric definition of the dot product:

$$\vec{u} \cdot \vec{v} = \|\vec{u}\| \, \|\vec{v}\| \cos\theta$$

where θ is the angle between $\vec{u}$ and $\vec{v}$.

- The two definitions are equivalent! This equivalence allows us to find the angle θ between two vectors.



Orthogonal vectors

- Recall: $\cos 90^{\mathrm{o}} = 0$.
- Since $\vec{u} \cdot \vec{v} = \|\vec{u}\| \|\vec{v}\| \cos\theta$, if the angle between two vectors is 90^{o}, their dot product is $\|\vec{u}\| \|\vec{v}\| \cos 90^{\mathrm{o}} = 0$.
- If the angle between two vectors is 90^{o}, we say they are perpendicular, or more generally, **orthogonal**.
- **Key idea**:

$$\boxed{\text{two vectors are } \textbf{orthogonal} \iff \vec{u} \cdot \vec{v} = 0}$$

Exercise

Find a non-zero vector in $\mathbb{R}^3$ orthogonal to:

$$\vec{v} = \begin{bmatrix} 2 \\ 5 \\ -8 \end{bmatrix}$$

Spans and projections

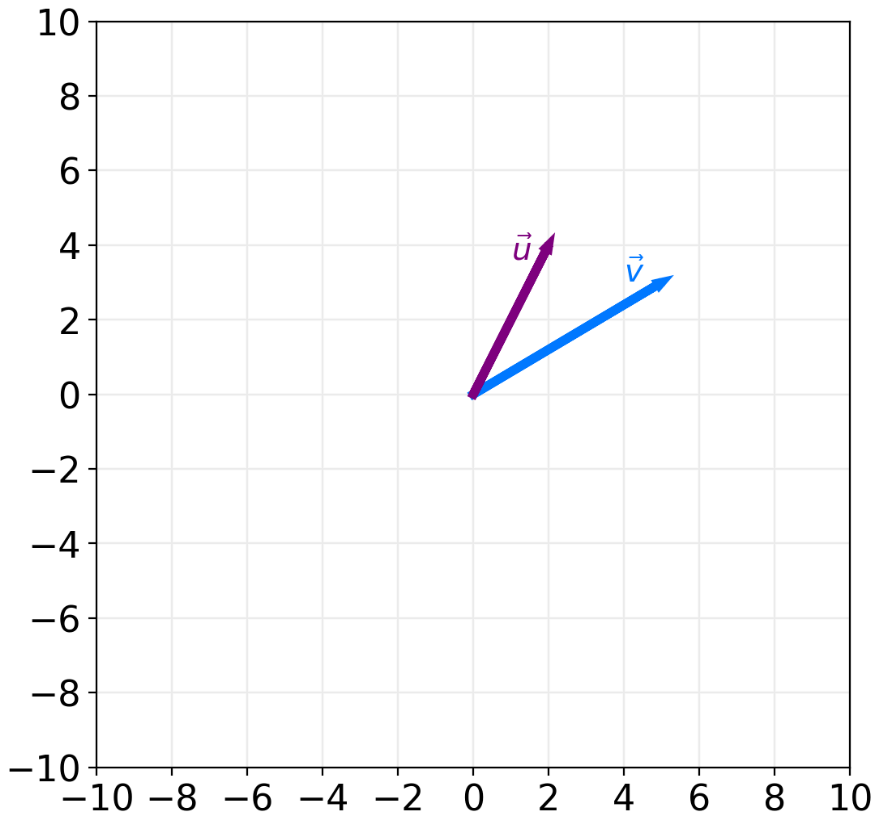
Adding and scaling vectors

- The sum of two vectors $\vec{u}$ and $\vec{v}$ in $\mathbb{R}^n$ is the **element-wise sum** of their components:

$$\vec{u} + \vec{v} = \begin{bmatrix} u_1 + v_1 \\ u_2 + v_2 \\ \vdots \\ u_n + v_n \end{bmatrix}$$

- If c is a scalar, then:

$$c\vec{v} = \begin{bmatrix} cv_1 \\ cv_2 \\ \vdots \\ cv_n \end{bmatrix}$$



Linear combinations

Let $\vec{v}_1$, $\vec{v}_2$, ..., $\vec{v}_d$ all be vectors in $\mathbb{R}^n$.

A **linear combination** of $\vec{v}_1$, $\vec{v}_2$, ..., $\vec{v}_d$ is any vector of the form:

$$a_1\vec{v}_1 + a_2\vec{v}_2 + \ldots + a_d\vec{v}_d$$

where a_1, a_2, ..., a_d are all scalars.

Span

- Let $\vec{v}_1$, $\vec{v}_2$, ..., $\vec{v}_d$ all be vectors in $\mathbb{R}^n$.
- The **span** of $\vec{v}_1$, $\vec{v}_2$, ..., $\vec{v}_d$ is the set of all vectors that can be created using linear combinations of those vectors.
- Formal definition:

$$\text{span}(\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_d) = \{a_1\vec{v}_1 + a_2\vec{v}_2 + \ldots + a_d\vec{v}_d : a_1, a_2, \ldots, a_n \in \mathbb{R}\}$$

Exercise

Let $\vec{v}_1 = \begin{bmatrix} 2 \\ -3 \end{bmatrix}$ and let $\vec{v}_2 = \begin{bmatrix} -1 \\ 4 \end{bmatrix}$. Is $\vec{y} = \begin{bmatrix} 9 \\ 1 \end{bmatrix}$ in $\mathrm{span}(\vec{v}_1, \vec{v}_2)$?

If so, write $\vec{y}$ as a linear combination of $\vec{v}_1$ and $\vec{v}_2$.

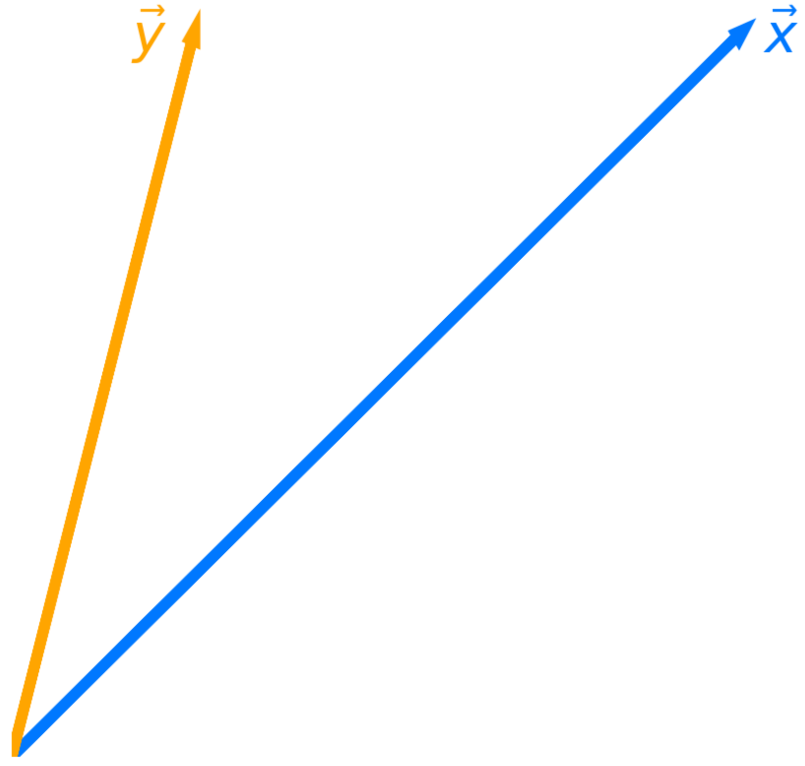
Projecting onto a single vector

- Let $\vec{x}$ and $\vec{y}$ be two vectors in $\mathbb{R}^n$.
- The span of $\vec{x}$ is the set of all vectors of the form:

$$w\vec{x}$$

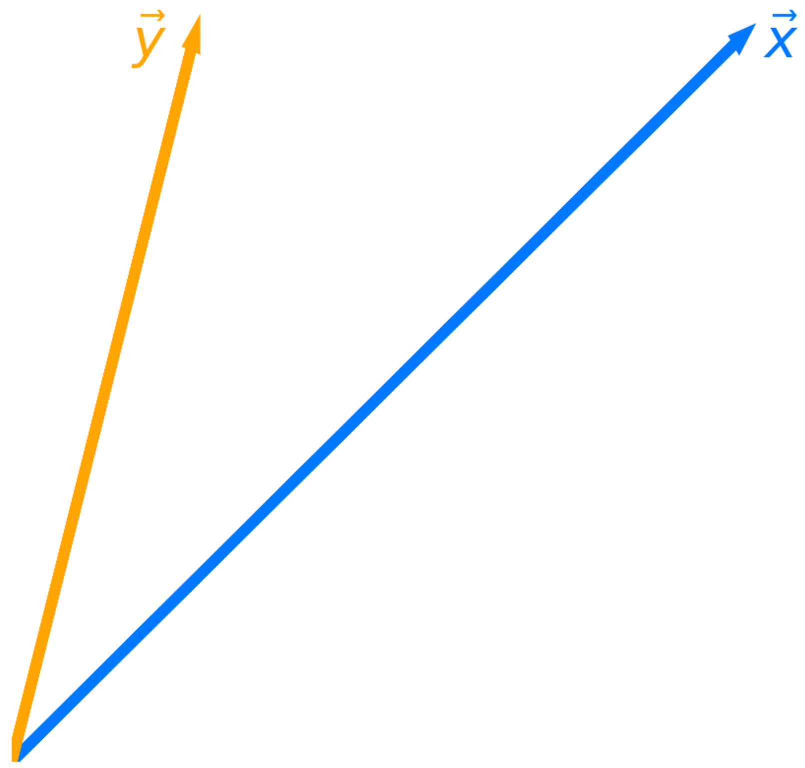
  where $w \in \mathbb{R}$ is a scalar.

- **Question**: What vector in $\text{span}(\vec{x})$ is closest to $\vec{y}$?
- The vector in $\text{span}(\vec{x})$ that is closest to $\vec{y}$ is the __________ **projection of $\vec{y}$ onto $\text{span}(\vec{x})$**.



Projection error

- Let $\vec{e} = \vec{y} - w\vec{x}$ be the **projection error**: that is, the vector that connects $\vec{y}$ to $\text{span}(\vec{x})$.
- **Goal: Find the w that makes $\vec{e}$ as short as possible**.
    - That is, minimize:
    $$\|\vec{e}\|$$
    - Equivalently, minimize:
    $$\|\vec{y} - w\vec{x}\|$$
- **Idea**: To make $\vec{e}$ has short as possible, it should be **orthogonal to $w\vec{x}$**.



Minimizing projection error

- **Goal**: **Find the w that makes $\vec{e} = \vec{y} - w\vec{x}$ as short as possible**.
- **Idea**: To make $\vec{e}$ as short as possible, it should be **orthogonal to** $w\vec{x}$.
- Can we prove that making $\vec{e}$ orthogonal to $w\vec{x}$ minimizes $||\vec{e}||$?

Minimizing projection error

- **Goal: Find the w that makes $\vec{e} = \vec{y} - w\vec{x}$ as short as possible.**
- Now we know that to minimize $\|\vec{e}\|$, $\vec{e}$ must be orthogonal to $w\vec{x}$.
- Given this fact, how can we solve for w?

Orthogonal projection

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

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

- Note that w^* is the solution to a minimization problem, specifically, this one:

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

- We call $w^* \vec{x}$ the **orthogonal projection of** $\vec{y}$ **onto** $\text{span}(\vec{x})$.
    - Think of $w^* \vec{x}$ as the "shadow" of $\vec{y}$.

Exercise

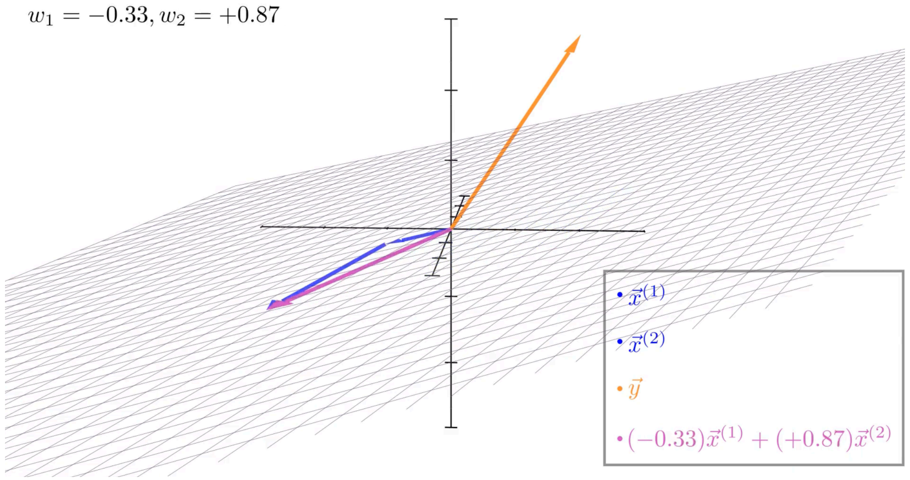
Let $\vec{a} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$ and $\vec{b} = \begin{bmatrix} -1 \\ 9 \end{bmatrix}$.

What is the orthogonal projection of $\vec{a}$ onto $\text{span}(\vec{b})$?

Your answer should be of the form $w^* \vec{b}$, where w^* is a scalar.

Moving to multiple dimensions

- Let's now consider three vectors, $\vec{y}$, $\vec{x}^{(1)}$, and $\vec{x}^{(2)}$, all in $\mathbb{R}^n$.
- **Question**: What vector in $\text{span}(\vec{x}^{(1)}, \vec{x}^{(2)})$ is closest to $\vec{y}$?
    - Vectors in $\text{span}(\vec{x}^{(1)}, \vec{x}^{(2)})$ are of the form $w_1 \vec{x}^{(1)} + w_2 \vec{x}^{(2)}$, where w_1, $w_2 \in \mathbb{R}$ are scalars.
- Before trying to answer, let's watch 🎥 **this animation that Jack, one of our tutors, made**.



Minimizing projection error in multiple dimensions

- **Question**: What vector in $\text{span}(\vec{x}^{(1)}, \vec{x}^{(2)})$ is closest to $\vec{y}$?
  - That is, what vector minimizes $\|\vec{e}\|$, where:

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

- **Answer**: It's the vector such that $w_1\vec{x}^{(1)} + w_2\vec{x}^{(2)}$ is **orthogonal** to $\vec{e}$.
- **Issue**: Solving for w_1 and w_2 in the following equation is difficult:

$$\left(w_1\vec{x}^{(1)} + w_2\vec{x}^{(2)}\right) \cdot \underbrace{\left(\vec{y} - w_1\vec{x}^{(1)} - w_2\vec{x}^{(2)}\right)}_{\vec{e}} = 0$$

Minimizing projection error in multiple dimensions

- It's hard for us to solve for w_1 and w_2 in:

$$\left(w_1\vec{x}^{(1)} + w_2\vec{x}^{(2)}\right) \cdot \underbrace{\left(\vec{y} - w_1\vec{x}^{(1)} - w_2\vec{x}^{(2)}\right)}_{\vec{e}} = 0$$

- **Observation**: All we really need is for $\vec{x}^{(1)}$ and $\vec{x}^{(2)}$ to individually be orthogonal to $\vec{e}$.
    - That is, it's sufficient for $\vec{e}$ to be orthogonal to the spanning vectors themselves.
- If $\vec{x}^{(1)} \cdot \vec{e} = 0$ and $\vec{x}^{(2)} \cdot \vec{e} = 0$, then:

Minimizing projection error in multiple dimensions

- **Question**: What vector in $\text{span}(\vec{x}^{(1)}, \vec{x}^{(2)})$ is closest to $\vec{y}$?
- **Answer**: It's the vector such that $w_1 \vec{x}^{(1)} + w_2 \vec{x}^{(2)}$ is **orthogonal** to $\vec{e} = \vec{y} - w_1 \vec{x}^{(1)} - w_2 \vec{x}^{(2)}$.
- **Equivalently, it's the vector such that $\vec{x}^{(1)}$ and $\vec{x}^{(2)}$ are both orthogonal to $\vec{e}$:**

$$\boxed{\begin{aligned} \vec{x}^{(1)} \cdot \left(\vec{y} - w_1 \vec{x}^{(1)} - w_2 \vec{x}^{(2)}\right) &= 0 \\ \vec{x}^{(2)} \cdot \underbrace{\left(\vec{y} - w_1 \vec{x}^{(1)} - w_2 \vec{x}^{(2)}\right)}_{\vec{e}} &= 0 \end{aligned}}$$

- This is a system of two equations, two unknowns (w_1 and w_2), but it still looks difficult to solve.

Now what?

- We're looking for the scalars w_1 and w_2 that satisfy the following equations:

$$\vec{x}^{(1)} \cdot \left(\vec{y} - w_1 \vec{x}^{(1)} - w_2 \vec{x}^{(2)}\right) = 0$$
$$\vec{x}^{(2)} \cdot \underbrace{\left(\vec{y} - w_1 \vec{x}^{(1)} - w_2 \vec{x}^{(2)}\right)}_{\vec{e}} = 0$$

- In this example, we just have two spanning vectors, $\vec{x}^{(1)}$ and $\vec{x}^{(2)}$.
- If we had any more, this system of equations would get extremely messy, extremely quickly.
- **Idea**: Rewrite the above system of equations **as a single equation, involving matrix-vector products**.

Matrices

Matrices

- An $n \times d$ **matrix** is a table of numbers with n rows and d columns.
- We use upper-case letters to denote matrices.

$$A = \begin{bmatrix} 2 & 5 & 8 \\ -1 & 5 & -3 \end{bmatrix}$$

- Since A has two rows and three columns, we say $A \in \mathbb{R}^{2 \times 3}$.
- **Key idea**: Think of a matrix as **several column vectors, stacked next to each other**.

Matrix addition and scalar multiplication

- We can add two matrices only if they have the same dimensions.
- Addition occurs elementwise:

$$\begin{bmatrix} 2 & 5 & 8 \\ -1 & 5 & -3 \end{bmatrix} + \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 7 & 11 \\ -1 & 6 & -1 \end{bmatrix}$$

- Scalar multiplication occurs elementwise, too:

$$2 \begin{bmatrix} 2 & 5 & 8 \\ -1 & 5 & -3 \end{bmatrix} = \begin{bmatrix} 4 & 10 & 16 \\ -2 & 10 & -6 \end{bmatrix}$$

Matrix-matrix multiplication

- **Key idea**: We can multiply matrices A and B **if and only if**:

$$\boxed{\#\text{ columns in } A = \#\text{ rows in } B}$$

- If A is $n \times d$ and B is $d \times p$, then AB is $n \times p$.
- Example: If A is as defined below, what is A^TA?

$$A = \begin{bmatrix} 2 & 5 & 8 \\ -1 & 5 & -3 \end{bmatrix}$$

Question 🤔

Answer at q.dsc40a.com

Assume A, B, and C are all matrices. Select the **incorrect** statement below.

- A. $A(B+C) = AB + AC$.
- B. $A(BC) = (AB)C$.
- C. $AB = BA$.
- D. $(A+B)^T = A^T + B^T$.
- E. $(AB)^T = B^T A^T$.

Matrix-vector multiplication

- A vector $\vec{v} \in \mathbb{R}^n$ is a matrix with n rows and 1 column.

$$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

- Suppose $A \in \mathbb{R}^{n \times d}$.
  - What must the dimensions of $\vec{v}$ be in order for the product $A\vec{v}$ to be valid?
  - What must the dimensions of $\vec{v}$ be in order for the product $\vec{v}^T A$ to be valid?

One view of matrix-vector multiplication

- One way of thinking about the product $A\vec{v}$ is that it is **the dot product of $\vec{v}$ with every row of A**.
- Example: What is $A\vec{v}$?

$$A = \begin{bmatrix} 2 & 5 & 8 \\ -1 & 5 & -3 \end{bmatrix} \qquad \vec{v} = \begin{bmatrix} 2 \\ -1 \\ -5 \end{bmatrix}$$

Another view of matrix-vector multiplication

- Another way of thinking about the product $A\vec{v}$ is that it is **a linear combination of the columns of A, using the weights in $\vec{v}$.**
- Example: What is $A\vec{v}$?

$$A = \begin{bmatrix} 2 & 5 & 8 \\ -1 & 5 & -3 \end{bmatrix} \qquad \vec{v} = \begin{bmatrix} 2 \\ -1 \\ -5 \end{bmatrix}$$

Matrix-vector products create linear combinations of columns!

- **Key idea**: It'll be very useful to think of the matrix-vector product $A\vec{v}$ as a linear combination of the columns of A, using the weights in $\vec{v}$.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1d} \\ a_{21} & a_{22} & \cdots & a_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nd} \end{bmatrix} \qquad \vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_d \end{bmatrix}$$

$$\downarrow$$

$$\boxed{A\vec{v} = v_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix} + v_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{bmatrix} + \ldots + v_d \begin{bmatrix} a_{1d} \\ a_{2d} \\ \vdots \\ a_{nd} \end{bmatrix}}$$

Spans and projections, revisited

Moving to multiple dimensions

- Let's now consider three vectors, $\vec{y}$, $\vec{x}^{(1)}$, and $\vec{x}^{(2)}$, all in $\mathbb{R}^n$.
- **Question**: What vector in $\text{span}(\vec{x}^{(1)}, \vec{x}^{(2)})$ is closest to $\vec{y}$?
    - That is, what values of w_1 and w_2 minimize $\|\vec{e}\| = \|\vec{y} - w_1\vec{x}^{(1)} - w_2\vec{x}^{(2)}\|$?

Matrix-vector products create linear combinations of columns!

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

- Combining $\vec{x}^{(1)}$ and $\vec{x}^{(2)}$ into a single matrix gives:

$$X = \begin{bmatrix} | & | \\ \vec{x}^{(1)} & \vec{x}^{(2)} \\ | & | \end{bmatrix} = \begin{bmatrix} \_\_\_ & \_\_\_ \\ \_\_\_ & \_\_\_ \\ \_\_\_ & \_\_\_ \end{bmatrix}$$

- Then, if $\vec{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$, linear combinations of $\vec{x}^{(1)}$ and $\vec{x}^{(2)}$ can be written as $X\vec{w}$.
- The **span of the columns of** X, or $\text{span}(X)$, consists of all vectors that can be written in the form $X\vec{w}$.

Minimizing projection error in multiple dimensions

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

- **Goal**: Find the vector $\vec{w} = \begin{bmatrix} w_1 & w_2 \end{bmatrix}^T$ such that $\|\vec{e}\| = \|\vec{y} - X\vec{w}\|$ is minimized.
- As we've seen, $\vec{w}$ must be such that:

$$\vec{x}^{(1)} \cdot \left(\vec{y} - w_1 \vec{x}^{(1)} - w_2 \vec{x}^{(2)} \right) = 0$$
$$\vec{x}^{(2)} \cdot \underbrace{\left(\vec{y} - w_1 \vec{x}^{(1)} - w_2 \vec{x}^{(2)} \right)}_{\vec{e}} = 0$$

- **How can we use our knowledge of matrices to rewrite this system of equations as a single equation?**

Simplifying the system of equations, using matrices

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

$$\vec{x}^{(1)} \cdot \left(\vec{y} - w_1 \vec{x}^{(1)} - w_2 \vec{x}^{(2)} \right) = 0$$

$$\vec{x}^{(2)} \cdot \underbrace{\left(\vec{y} - w_1 \vec{x}^{(1)} - w_2 \vec{x}^{(2)} \right)}_{\vec{e}} = 0$$

Simplifying the system of equations, using matrices

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

1. $w_1\vec{x}^{(1)} + w_2\vec{x}^{(2)}$ can be written as $X\vec{w}$, so $\vec{e} = \vec{y} - X\vec{w}$.
2. **The condition that $\vec{e}$ must be orthogonal to each column of X is equivalent to condition that $X^T\vec{e} = 0$.**

The normal equations

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

- **Goal**: Find the vector $\vec{w} = \begin{bmatrix} w_1 & w_2 \end{bmatrix}^T$ such that $\|\vec{e}\| = \|\vec{y} - X\vec{w}\|$ is minimized.
- We now know that it is the vector $\vec{w}^*$ such that:

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

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

The general solution to the normal equations

$$X \in \mathbb{R}^{n \times d} \qquad \vec{y} \in \mathbb{R}^n$$

- **Goal, in general**: Find the vector $\vec{w} \in \mathbb{R}^d$ such that $\|\vec{e}\| = \|\vec{y} - X\vec{w}\|$ is minimized.
- We now know that it is the vector $\vec{w}^*$ such that:

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

- 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**: 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:

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

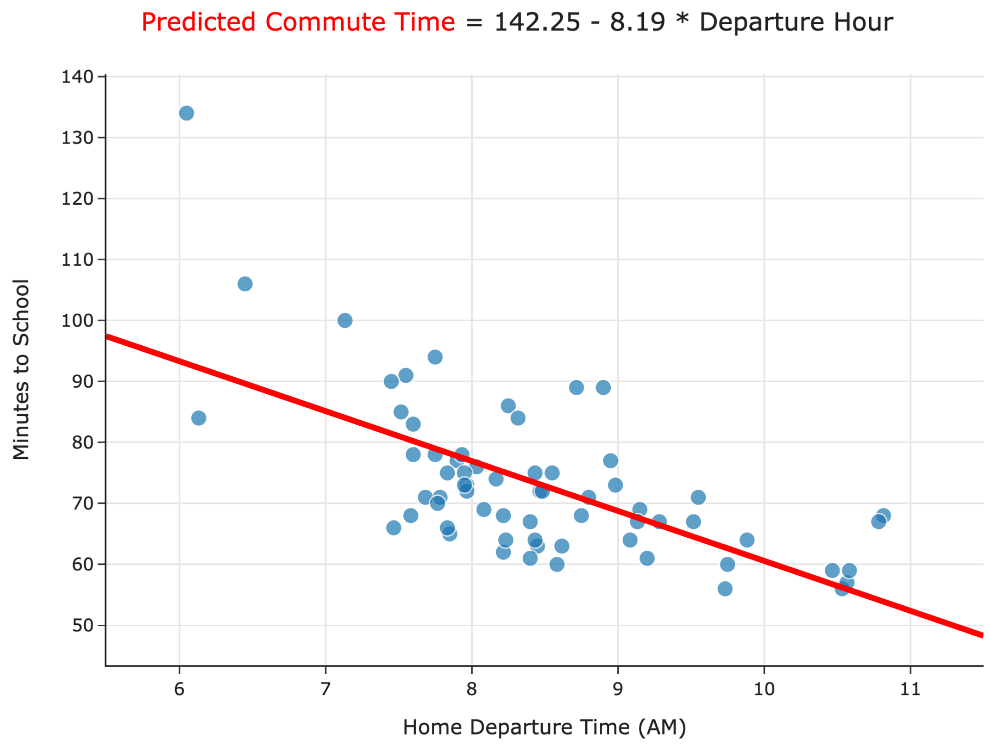
- We're going to use this frequently!

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



- **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:

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

- **Observation**: $R_{\text{sq}}(w_0, w_1)$ *kind of* looks like the formula for the norm of a vector, $\|\vec{v}\| = \sqrt{v_1^2 + v_2^2 + \ldots + v_n^2}$.

Regression and linear algebra

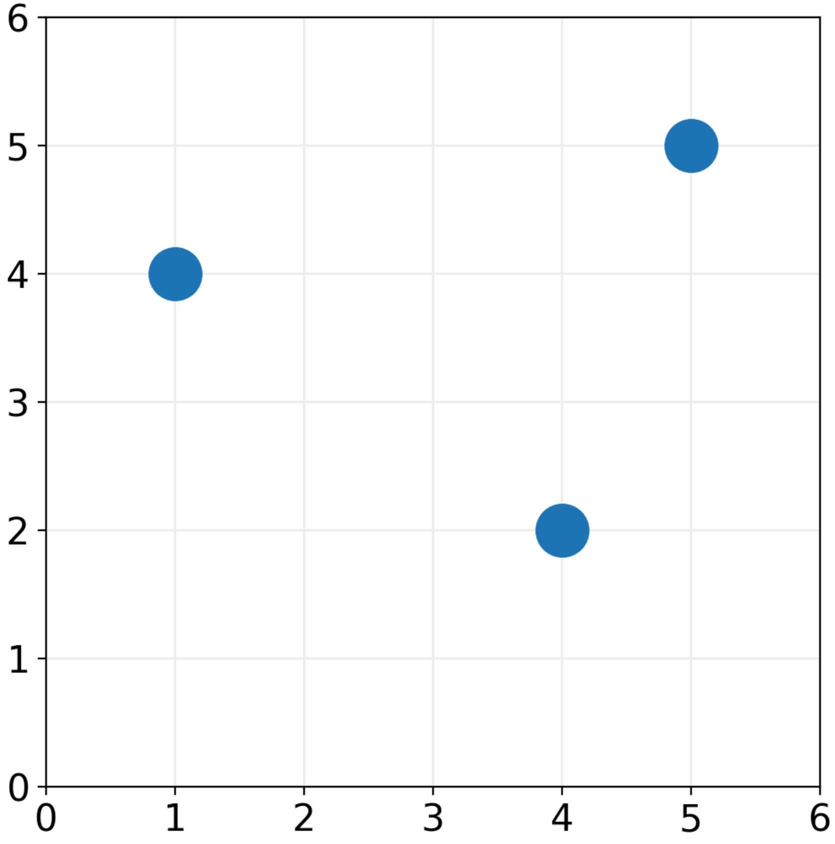
Let's define a few new terms:

- 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:

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

Example

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



$\vec{y} =$

$\vec{h} =$

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

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

$$=$$

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:

$$\vec{h} = \begin{bmatrix} w_0 + w_1 x_1 \\ w_0 + w_1 x_2 \\ \vdots \\ w_0 + w_1 x_n \end{bmatrix} =$$

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 $\vec{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$.

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

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

What's next?

- 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 minimizing:

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

- **We've already solved this problem! Assuming X^TX is invertible, the best $\vec{w}$ is:**

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