

Lecture 14

Regression using Linear Algebra

EECS 398: Practical Data Science, Spring 2025

practicaldsc.org • github.com/practicaldsc/sp25 •  See latest announcements [here on Ed](#)

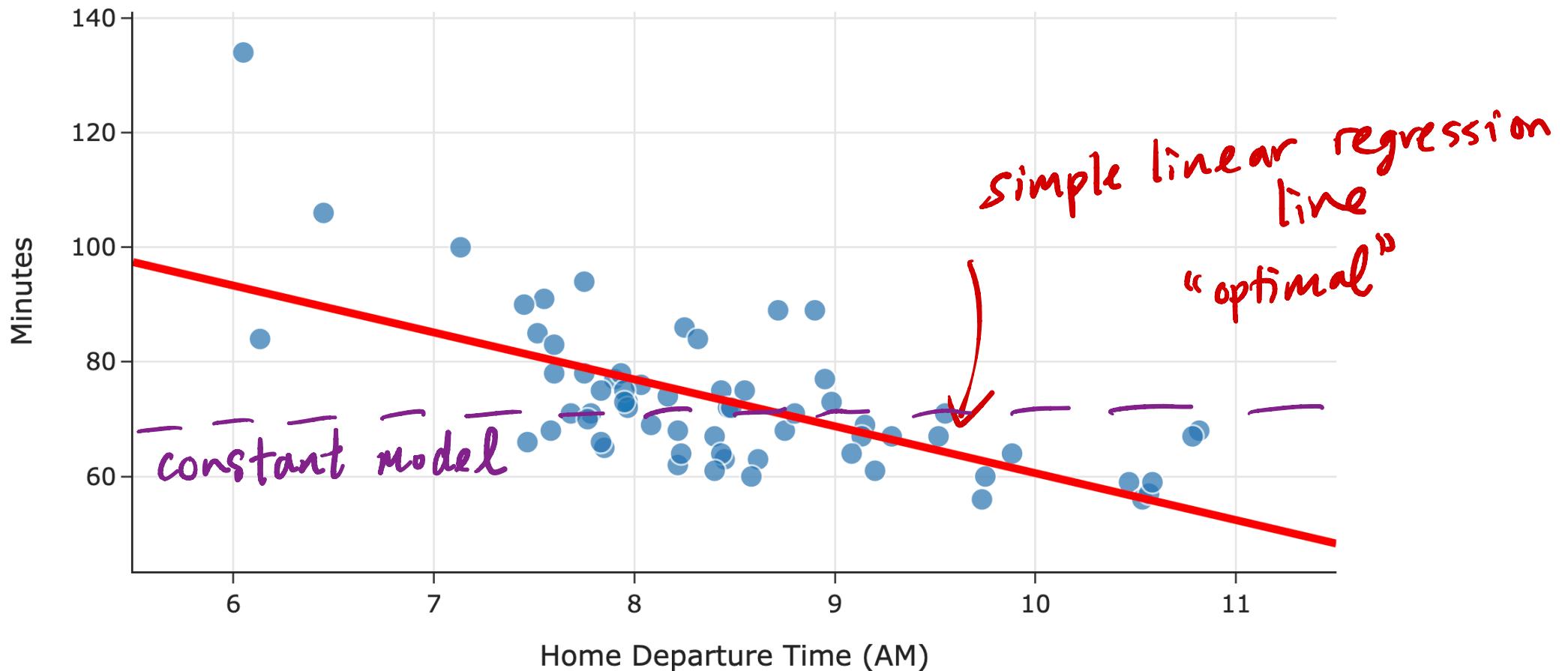
Agenda



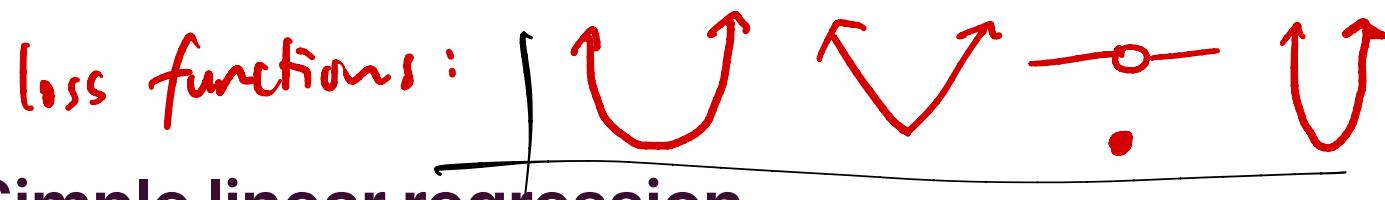
- Recap: Simple linear regression.
- Interpreting the formulas.
- Regression and linear algebra.

Recap: Simple linear regression

$$\text{Predicted Commute Time} = 142.25 - 8.19 * \text{Departure Hour}$$



In Lecture 12, we said that the line in red is the regression line.
But how did we find this line?



Recap: Simple linear regression

- Goal: Use the modeling recipe to find the "best" simple linear hypothesis function.

1. Model: $H(x_i) = w_0 + w_1 x_i$.

2. Loss function: $L_{\text{sq}}(y_i, H(x_i)) = (y_i - H(x_i))^2$.

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

*average of loss function
across entire dataset*

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

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

optimal slope

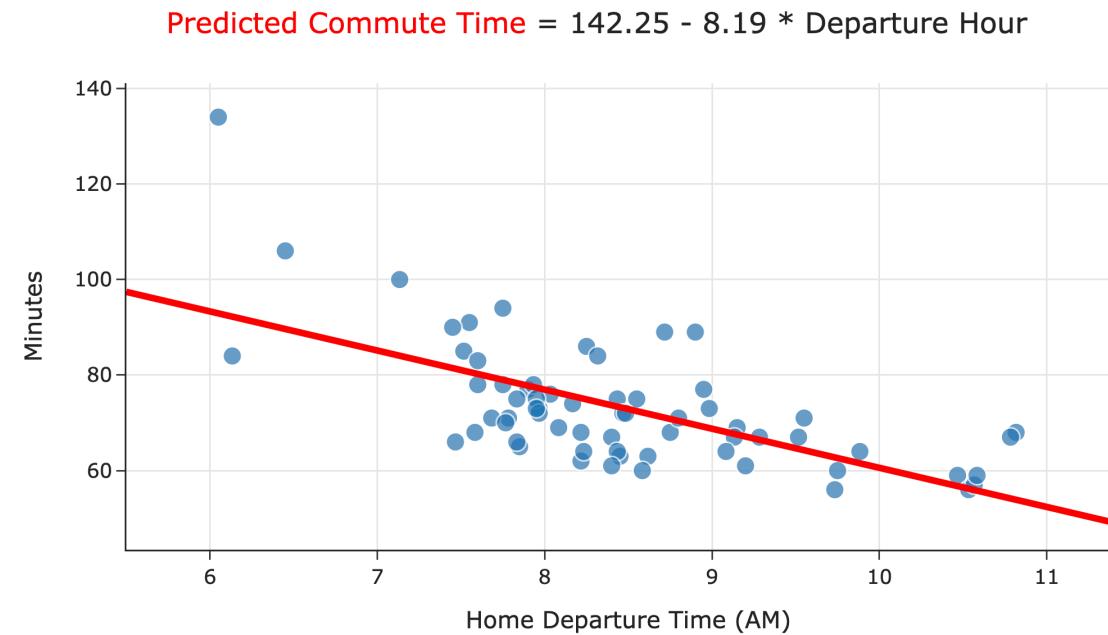
$$w_0^* = \bar{y} - w_1^* \bar{x}$$

optimal intercept

- The resulting line, $H^*(x_i) = w_0^* + w_1^* x_i$, is the unique line that minimizes MSE.

Code demo

- Before we go any further, let's test out our formulas in code.

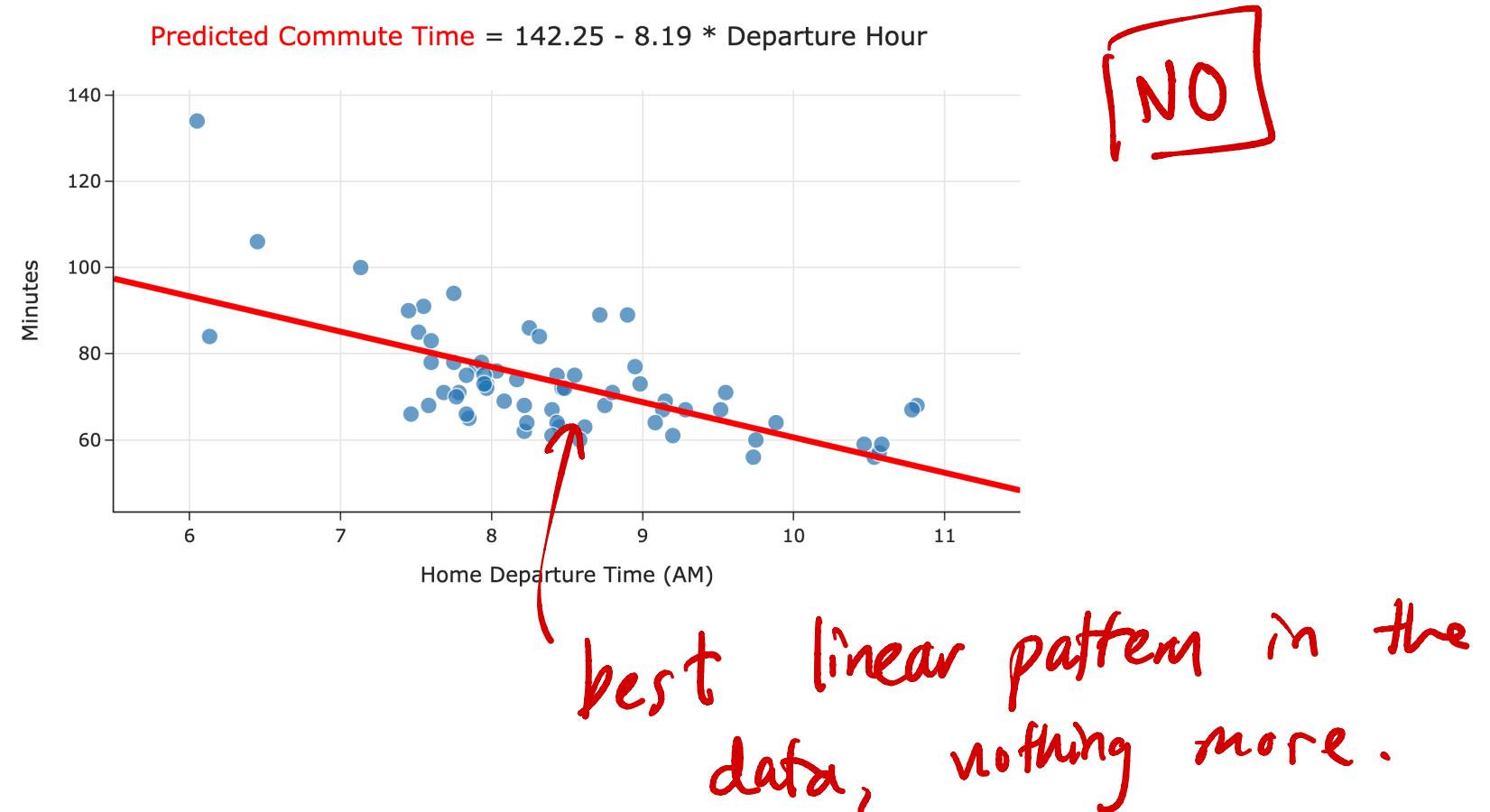


- The supplementary notebook is posted in the usual place on [GitHub](#) and the [course website](#).

Interpreting the formulas

Causality

- Can we conclude that leaving later **causes** you to get to school earlier?



Interpreting the slope

hard to extrapolate
outside of
6-11 AM
(the range of
the data)

- The units of the slope are **units of y per units of x** .
- In our commute times example, in $H^*(x_i) = 142.25 - 8.19x_i$, our predicted commute time decreases by 8.19 minutes per hour.

What if we leave
at 8PM?

$$x_i = 20$$
$$H^*(20) = 142.25 - 170 < 0$$

midnight?

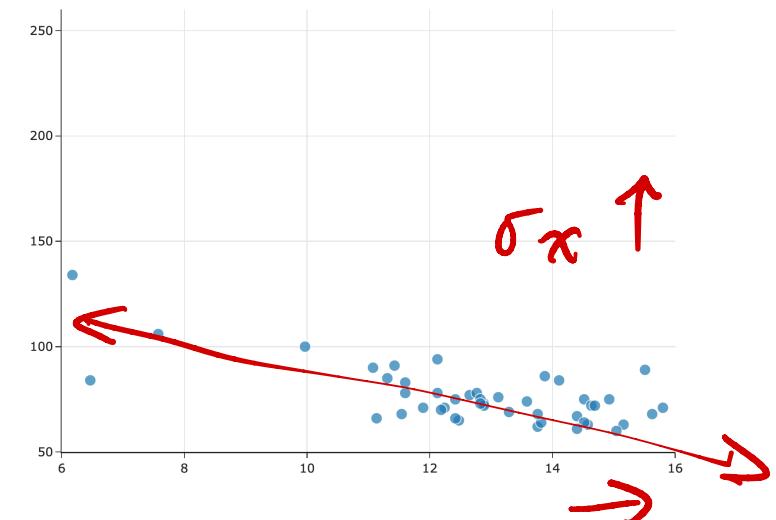
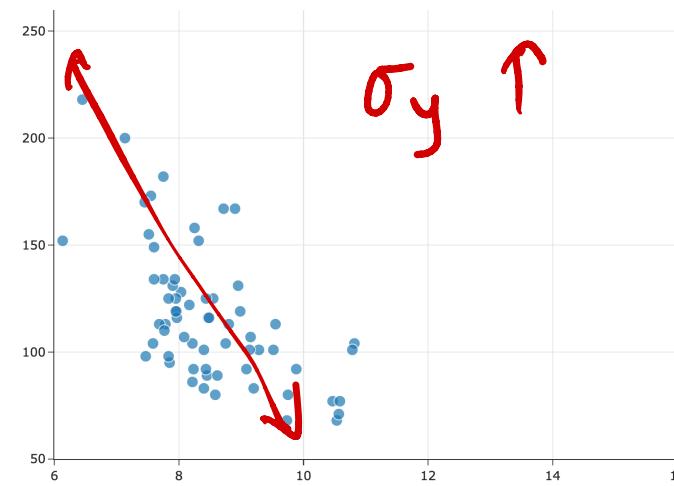
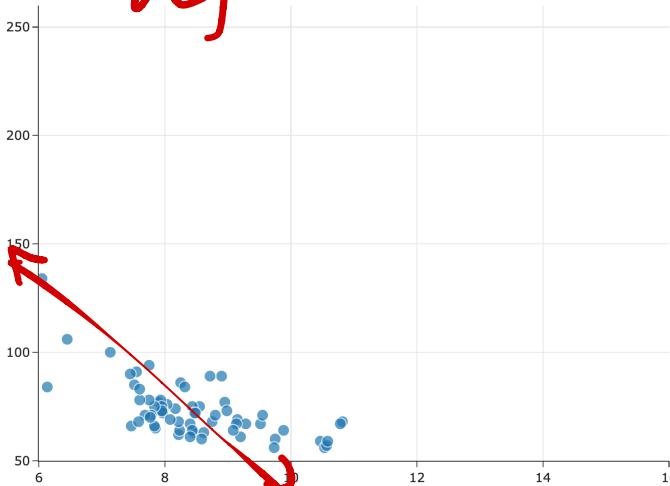
$$x_i = 0$$
$$H^*(0) = 142.25$$

Interpreting the slope

$$w_1^* = r \frac{\sigma_y}{\sigma_x}$$

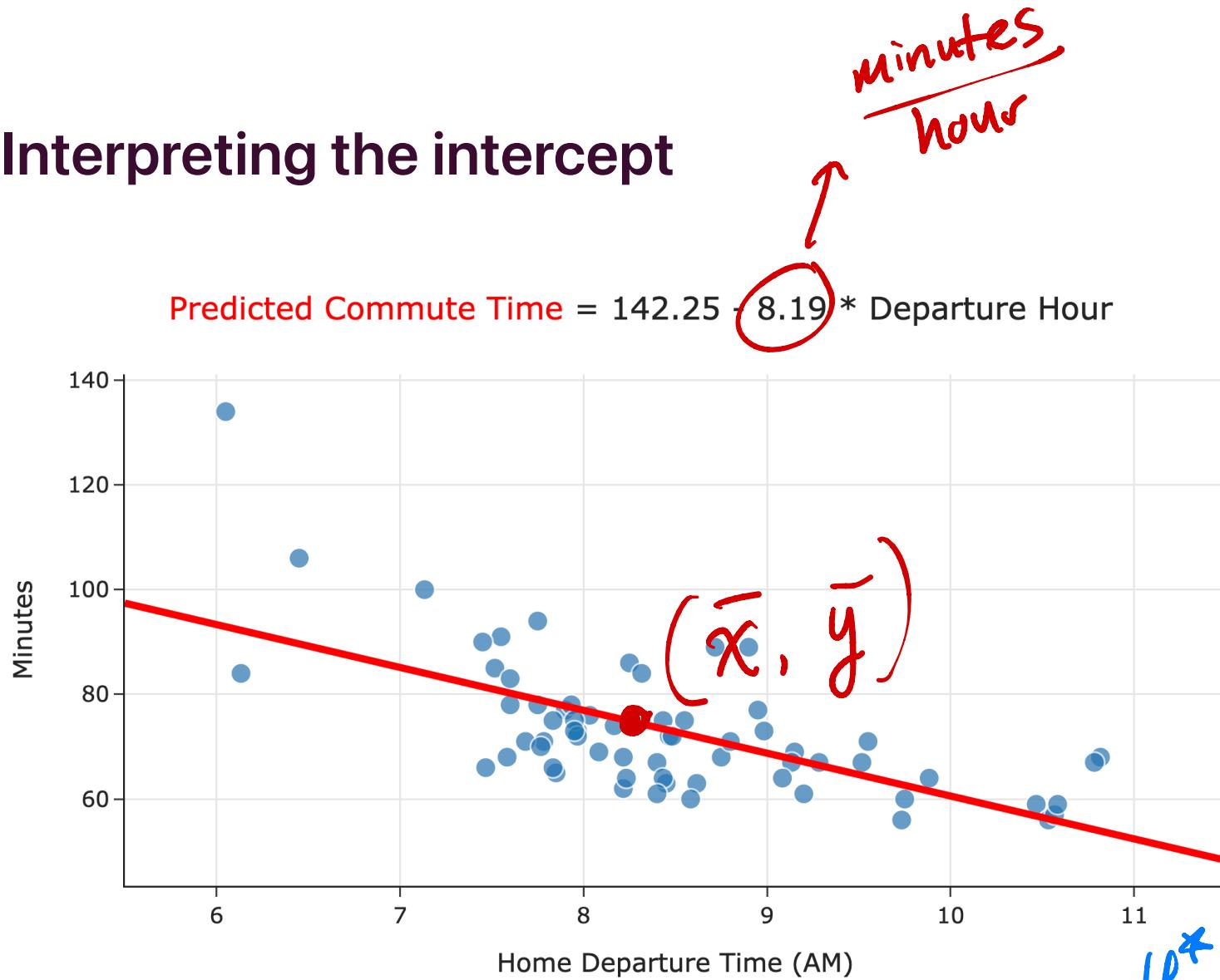
$-1 \leq r \leq 1$

default



- Since $\sigma_x \geq 0$ and $\sigma_y \geq 0$, the slope's sign is r 's sign.
- As the y values get more spread out, σ_y increases, so the slope gets steeper.
- As the x values get more spread out, σ_x increases, so the slope gets shallower.

Interpreting the intercept



$$w_0^* = \bar{y} - w_1^* \bar{x}$$

- What are the units of the intercept?

units of y (minutes)

- What is the value of

$$H^*(\bar{x}) = \bar{y}$$

$$w_0^* = \bar{y} - w_1^* \bar{x}$$

$$H^*(x_i) = \bar{y} - w_1^* \bar{x} + w_1^* x_i$$

if $x_i = \bar{x}$, then

Question 🤔

Answer at practicaldsc.org/q

We fit a regression line to predict commute times given departure hour. Then, we add 75 minutes to all commute times in our dataset. What happens to the resulting regression line?

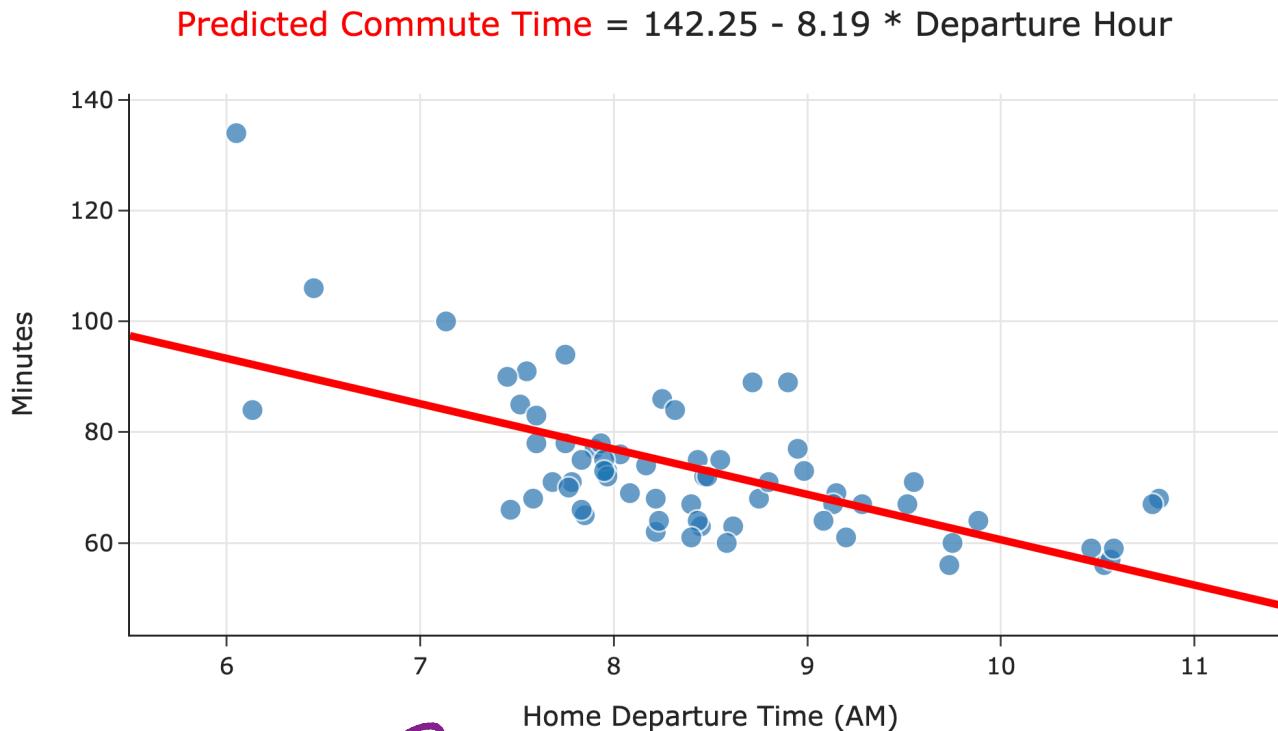
- A. Slope increases, intercept increases.
- B. Slope decreases, intercept increases.
- C. Slope stays the same, intercept increases.
- D. Slope stays the same, intercept stays the same.

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 the day of the month.
- 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. $H(x_i) = w_0 + w_1x_i + w_2x_i^2$.

Simple linear regression, revisited



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

- **Model:** $H(x_i) = w_0 + w_1 x_i$.
- **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,

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

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:

$$\vec{h} = \begin{bmatrix} H(x_1) \\ H(x_2) \\ \vdots \\ H(x_n) \end{bmatrix}$$

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

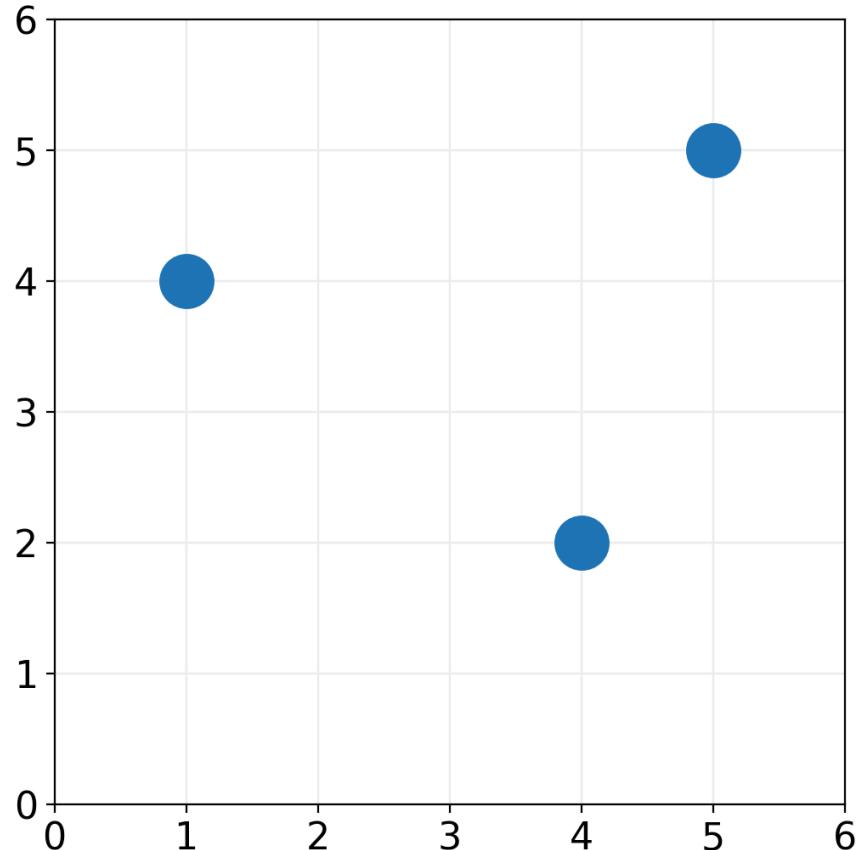
$\vec{y} \in \mathbb{R}^n$:
 \vec{y} is a vector made
of n real numbers

other classes $\hat{\vec{y}}$ \hat{y}

$$e_i = y_i - H(x_i)$$
$$\vec{e} = \vec{y} - \vec{h} = \begin{bmatrix} y_1 - H(x_1) \\ y_2 - H(x_2) \\ \vdots \\ y_n - H(x_n) \end{bmatrix}$$

Example

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



$$\vec{y} = \quad \vec{h} =$$

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

$$R_{\text{sq}}(\textcolor{violet}{H}) = \frac{1}{n} \sum_{i=1}^n (\textcolor{orange}{y}_i - H(x_i))^2 \\ =$$

Regression and linear algebra

Let's define a few new terms:

$$\vec{e} = \begin{bmatrix} y_1 - H(x_1) \\ y_2 - H(x_2) \\ \vdots \\ y_n - H(x_n) \end{bmatrix}$$

- The **observation vector** is the vector $\vec{y} \in \mathbb{R}^n$. This is the vector of observed "actual values". $\frac{1}{n} \|\vec{e}\|^2 = \frac{1}{n} \sum_{i=1}^n (y_i - H(x_i))^2$
- 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:

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

The hypothesis vector

Very important!

the column of all 1s exists for the intercept term!

- 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 $H(x_i) = w_0 + w_1 x_i$, the hypothesis vector can be written:

$$\begin{bmatrix} H(x_1) \\ H(x_2) \\ \vdots \\ H(x_n) \end{bmatrix}$$

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

"design matrix"

"parameter vector"

before: $\frac{1}{n} \sum_{i=1}^n (y_i - (w_0 + w_i x_i))^2$

Rewriting the mean squared error

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

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \vec{h} = \cancel{\vec{X}} \vec{w}$$

- 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:

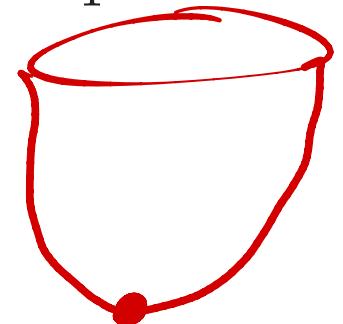
$$R_{\text{sq}}(\vec{H}) = \frac{1}{n} \|\vec{y} - \vec{h}\|^2 \implies R_{\text{sq}}(\vec{w}) = \frac{1}{n} \|\vec{y} - X\vec{w}\|^2$$

Minimizing mean squared error, again

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

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

same thing!



- 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 $\vec{w}^* = \begin{bmatrix} w_0^* \\ w_1^* \end{bmatrix}$ that minimizes:

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

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

Minimizing mean squared error, using projections?

- \mathbf{X} and \vec{y} are fixed: they come from our data.
- Our goal is to pick the \vec{w}^* that minimizes:

mean squared
error!!!

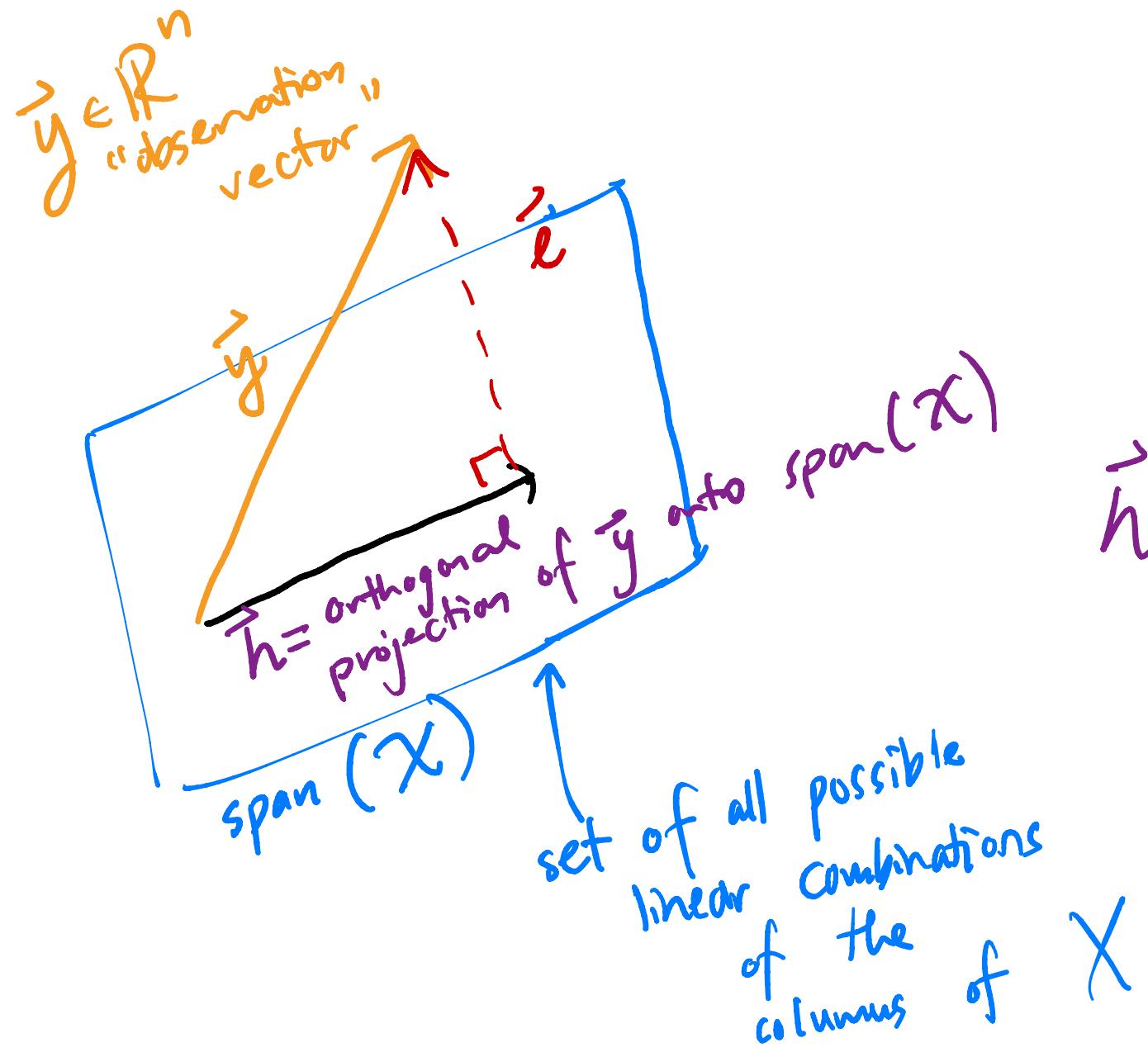
$$R_{\text{sq}}(\vec{w}) = \frac{1}{n} \|\vec{y} - \mathbf{X}\vec{w}\|^2$$

- This is equivalent to picking the \vec{w}^* that minimizes:

$$\|\vec{e}\| = \|\vec{y} - \mathbf{X}\vec{w}\|$$

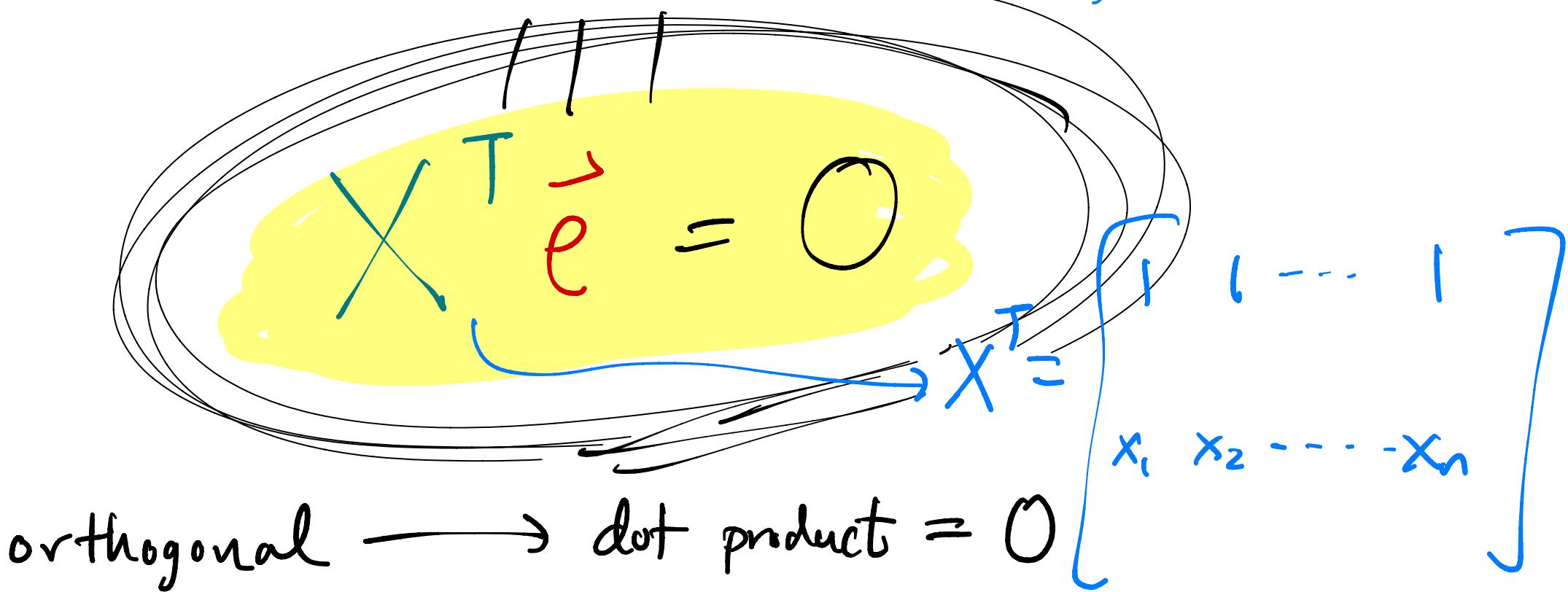
find \vec{w} that
minimizes $\|\vec{e}\|$

- This is equivalent to finding the w_0^* and w_1^* so that $\mathbf{X}\vec{w}^*$ is as "close" to \vec{y} as possible.
- **Solution:** Find the **orthogonal projection** of \vec{y} onto $\text{span}(\mathbf{X})$!
- We already did this in [Linear Algebra Guide 4](#), which you're reviewing in [Homework 6, Question 5!](#)



Million dollar goal:

Pick \vec{w} so that \vec{e} orthogonal
to span(X)



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

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

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

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

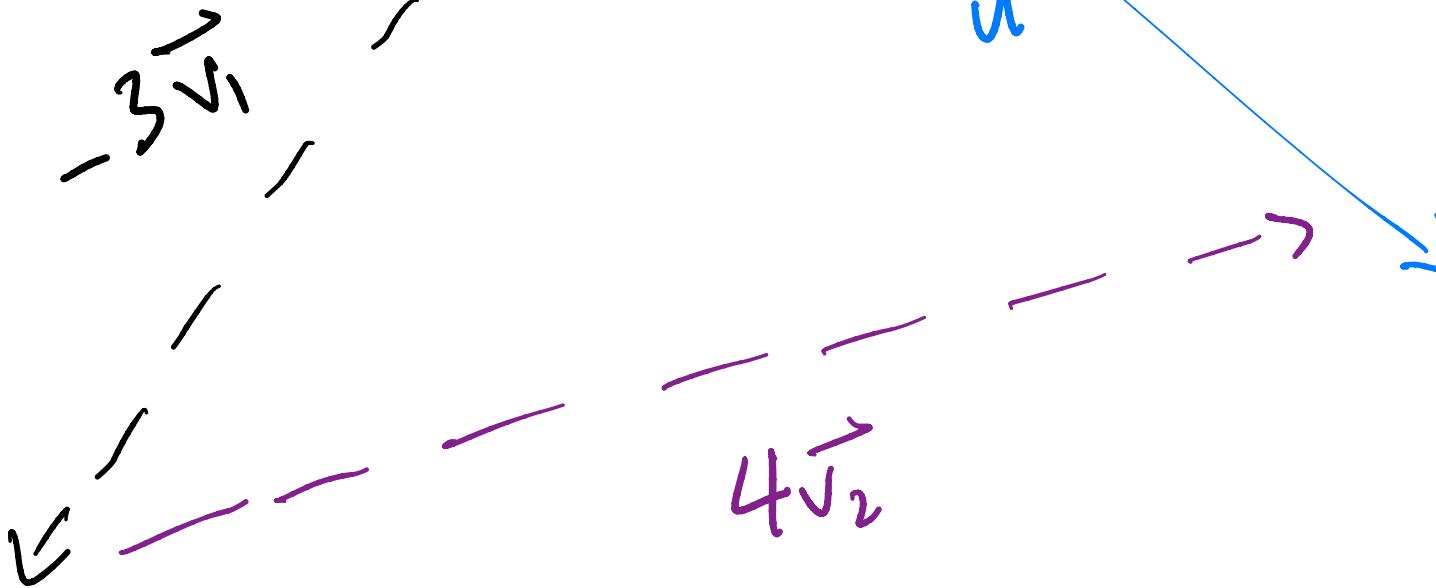
if $(X^T X)^{-1}$ exists, then

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

$$X_{2 \times n}^T X_{n \times 2} = X^T X_{2 \times 2}$$

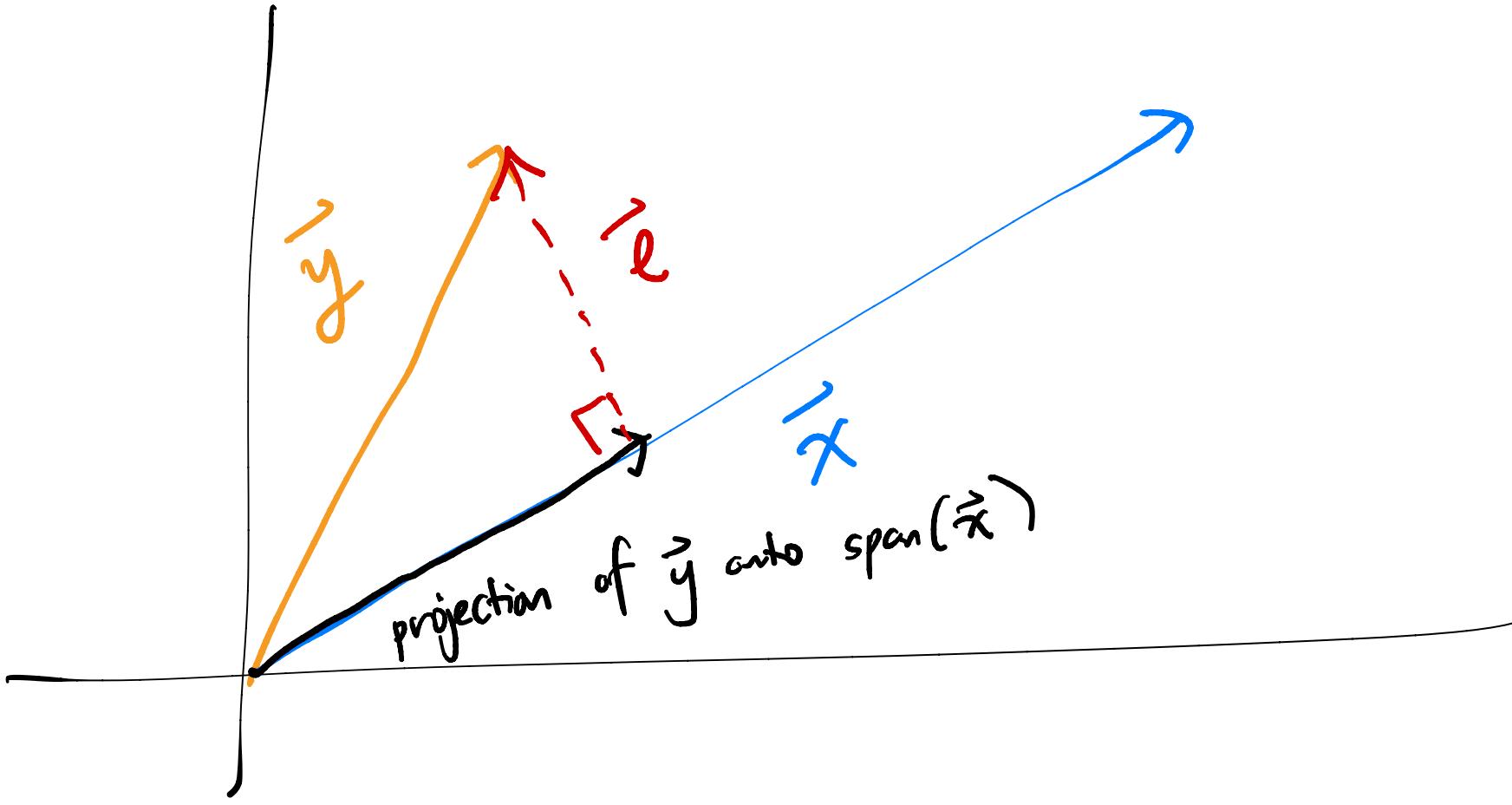
Aside: span

\vec{v}_1 and \vec{v}_2
SPAN
all of \mathbb{R}^2 .



$$\vec{u} = -3\vec{v}_1 + 4\vec{v}_2,$$

so,
 \vec{u} can be expressed
as a
linear combination
of \vec{v}_1, \vec{v}_2 !



An optimization problem we've seen before

- The optimal parameter vector, $\vec{w}^* = [w_0^* \quad w_1^*]^T$, is the one that minimizes:

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

minimizing this

- In the **Linear Algebra Guide** (and your linear algebra class), we showed that the \vec{w}^* that minimizes the length of the error vector, $\|\vec{e}\| = \|\vec{y} - \vec{X}\vec{w}\|$, is the one that satisfies the **normal equations**:

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

is the same as
solving this!

- The minimizer of $\|\vec{e}\|$ is the same as the minimizer of $R_{\text{sq}}(\vec{w})$.

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

- Key idea:** The \vec{w}^* that solves the normal equations also **minimizes** $R_{\text{sq}}(\vec{w})$!

The normal equations

- The normal equations are the system of 2 equations and 2 unknowns defined by:

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

- Why are they called the **normal** equations?
- If $\mathbf{X}^T \mathbf{X}$ is invertible, there is a unique solution to the normal equations:

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

- If $\mathbf{X}^T \mathbf{X}$ is not invertible, then there are infinitely many solutions to the normal equations. We will explore this idea as the semester progresses.

Same as formulas for w_0, w_1, \dots before!

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

- To find the optimal model parameters for simple linear regression, w_0^* and w_1^* , we previously minimized $R_{\text{sq}}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^n (\textcolor{orange}{y}_i - (w_0 + w_1 \textcolor{blue}{x}_i))^2$.

- We found, using calculus, that:

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

- $$w_0^* = \bar{y} - w_1^* \bar{x}.$$

the same!!!

- Another way of finding optimal model parameters for simple linear regression is to find the \vec{w}^* that minimizes $R_{\text{sq}}(\vec{w}) = \frac{1}{n} \|\vec{y} - \mathbf{X}\vec{w}\|^2$.

- The minimizer, if $\mathbf{X}^T \mathbf{X}$ is invertible, is the vector

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

- These formulas are equivalent!

Code demo

- To give us a break from math, we'll switch to a notebook, 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.
 - You'll prove this in Homework 7 😊.
- We'll use the same supplementary notebook as earlier, posted in the usual place on GitHub and the [course website](#).
- Then, in Lecture 15, we'll use our new linear algebraic formulation of regression to incorporate **multiple features** in our prediction process.

Summary: Regression and linear algebra

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

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

- Which \vec{w} makes the hypothesis vector, $\vec{h} = \mathbf{X}\vec{w}$, as close to \vec{y} as possible? Use the solution to the normal equations, \vec{w}^* :

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

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