Lecture 11

Regression and Linear Algebra

DSC 40A, Fall 2024

Announcements

- Homework 3 is due on **Friday, October 25th**.
- Homework 1 scores are available on Gradescope.
  - Regrade requests are due tonight.
- The Midterm Exam is on **Monday, Nov 4th in class**.

Agenda

- Regression and linear algebra.
- Finding the optimal parameter vector
  - by minimizing the projection error (linear algebra).
  - by minimizing empirical risk (multivariate calculus).

Question 🤔

Answer at q.dsc40a.com

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

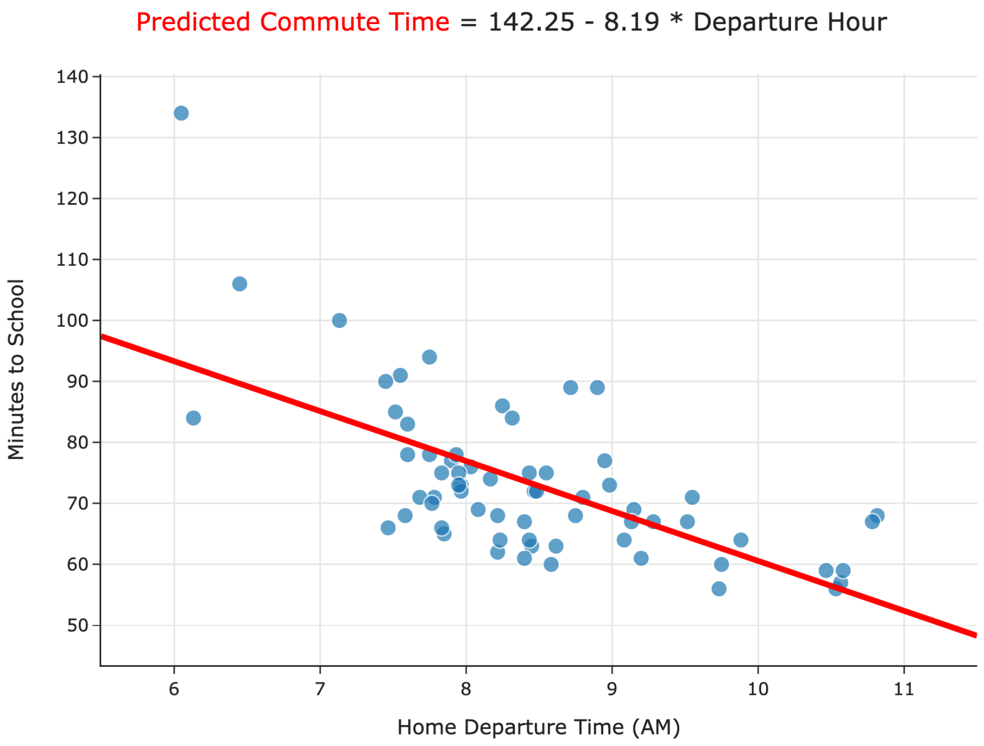
If the direct link doesn't work, click the "🤔 Lecture Questions" link in the top right corner of dsc40a.com.

Regression and linear algebra

Wait... why do we need linear algebra?

- We 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), e.g., $H(x) = w_0 + w_1 x^{(1)} + w_2 x^{(2)}$.
  - 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 learned last week 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:

$$R_{\text{sq}}(H) = \underbrace{\frac{1}{n} \sum_{i=1}^{n}}_{\text{averaging}} \underbrace{\left(\overbrace{y_i}^{\text{observed}} - \overbrace{H(x_i)}^{\text{prediction}}\right)^2}_{\text{loss}}$$

- **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 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:

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

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

This is the vector of signed errors.

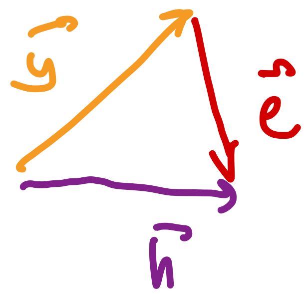
$$\vec{y} = \begin{bmatrix} 35 \text{ minutes} \\ 72 \text{ minutes} \\ 27 \text{ minutes} \\ \vdots \end{bmatrix} \qquad \vec{h} = \begin{bmatrix} H(x_1) \\ H(x_2) \\ \vdots \\ H(x_n) \end{bmatrix} = \begin{bmatrix} 41 \text{ minutes} \\ 70 \text{ minutes} \\ \vdots \end{bmatrix} \qquad \vec{e} = \begin{bmatrix} e_1 = y_1 - H(x_1) \\ e_2 = y_2 - H(x_2) \\ \vdots \end{bmatrix}$$

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 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:

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



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:

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

every element

$h_i = H(x_i) = w_0 + w_1 x_i$

X - design matrix

model parameters

$\vec{1}$ all-ones

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

slope

intercept

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

An optimization problem we've seen before

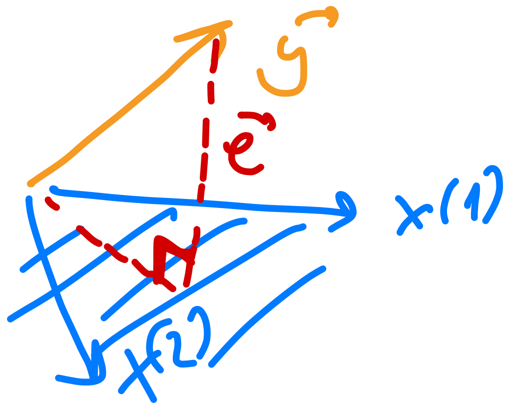
- The optimal parameter vector, $\vec{w}^* = \begin{bmatrix} w_0^* & w_1^* \end{bmatrix}^T$, is the one that minimizes:

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

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

$$\vec{w}^* = \underset{\vec{w}}{\arg\min}\, R_{\text{sq}} = \underset{\vec{w}}{\arg\min}\, \|\vec{e}\|$$

- Last week we found that the vector in the span of the columns of X that is closest to $\vec{y}$ is the vector $X\vec{w}$ such that $\|\vec{e}\| = \|\vec{y} - X\vec{w}\|$ is minimized.



The modeling recipe

1. Choose a model.

$$H(x) = [1 \quad x]^T \vec{w} = w_0 + w_1 x$$

SLR

2. Choose a loss function.

squared loss

$$e^2 = (y - [1 \quad x]^T w)^2$$

3. Minimize average loss to find optimal model parameters.

$$\vec{w}^* = \underset{\vec{w}}{\arg\min}\, R_{\text{sq}}(\vec{w}) = \underset{\vec{w}}{\arg\min} \left\{ \frac{1}{n} \|\vec{y} - X\vec{w}\|^2 \right\} = \underset{\vec{w}}{\arg\min} \left\{ \frac{1}{n} \|\vec{e}\|^2 \right\}$$

$\vec{a} \cdot \vec{b} = 0 \iff \vec{a}, \vec{b}$ are orthogonal

An optimization problem we've seen before

- **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:

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

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

- 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!

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

$A\vec{v}$ – dot product of v with the rows of A

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

$$= \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & & x_n \end{bmatrix} \vec{e} = 0$$

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:

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

- The **normal equations**:

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

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

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

$$\boxed{\vec{w}^* = (X^TX)^{-1}X^T\vec{y}} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

  - This is a big assumption, because it requires X^TX to be **full rank**. all columns are linearly independent
  - If X^TX is not full rank, then there are infinitely many solutions to the normal equations.

$\downarrow$ X is full rank

An optimization problem, solved

- We just used linear algebra to solve an **optimization problem**. (no calculus)
- Specifically, the function we minimized is:

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

- 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:

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

- **Key idea**: $\vec{w}^* = (X^T X)^{-1} X^T \vec{y}$ also **minimizes** $R_{\text{sq}}(\vec{w})$!
- We're going to use this frequently!

Alternative solution

- Our goal is to find the vector $\vec{w}$ that minimize mean squared error:

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

- Strategy: calculus
- Problem: This is a function *of a vector*. What does it even mean to take the derivative of $R_{\text{sq}}(\vec{w})$ with respect to a vector $\vec{w}$?

A function of a vector

- **Solution:** A function *of a vector* is really just a function *of multiple variables*, which are the components of the vector. In other words,

$$R_{\text{sq}}(\vec{w}) = R_{\text{sq}}(w_0, w_1, \dots, w_d) \in \mathbb{R}$$

$$\vec{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} \in \mathbb{R}^d$$

  where $w_0, w_1, \dots, w_d$ are the entries of the vector $\vec{w}$.
  In our case, $\vec{w}$ has just two components, w_0 and w_1. We'll be more general since we eventually want to use prediction rules with even more parameters.

- We know how to deal with derivatives of multivariable functions: the gradient!

The gradient with respect to a vector

- The **gradient of $R_{\text{sq}}(\vec{w})$ with respect to $\vec{w}$** is the vector of partial derivatives:

$$\begin{bmatrix} \frac{\partial R_{sq}}{\partial w_0} \\ \\ \frac{\partial R_{sq}}{\partial w_1} \end{bmatrix} \in \mathbb{R}^2 \quad =$$

$$\nabla_{\vec{w}} R_{\text{sq}}(\vec{w}) = \frac{dR_{\text{sq}}}{d\vec{w}} = \begin{bmatrix} \frac{\partial R_{\text{sq}}}{\partial w_0} \\ \frac{\partial R_{\text{sq}}}{\partial w_1} \\ \vdots \\ \frac{\partial R_{\text{sq}}}{\partial w_d} \end{bmatrix} \in \mathbb{R}^d$$

where $w_0, w_1, \ldots, w_d$ are the entries of the vector $\vec{w}$.

Goal

- We want to minimize the mean squared error: as a function of vector $\vec{w}$

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

- Strategy:

1. Compute the gradient of $R_{\text{sq}}(\vec{w})$.
2. Set it to zero and solve for $\vec{w}$.
    - The result is the optimal parameter vector $\vec{w}^*$.

- Let's start by rewriting the mean squared error in a way that will make it easier to compute its gradient.

Question 🤔

Answer at q.dsc40a.com

Which of the following is equivalent to $R_{\text{sq}}(\vec{w}) = \frac{1}{n}\|\vec{y} - X\vec{w}\|^2$?

A) $\frac{1}{n}(\vec{y} - X\vec{w}) \cdot (X\vec{w} - y)$

B) $\frac{1}{n}\sqrt{(\vec{y} - X\vec{w}) \cdot (y - X\vec{w})}$ $= \frac{1}{n}\|\vec{e}\|$

C) $\frac{1}{n}(\vec{y} - X\vec{w})^T(y - X\vec{w})$

D) $\frac{1}{n}(\vec{y} - X\vec{w})(y - X\vec{w})^T$

$\|$

$\frac{1}{n}\|\vec{e}\|^2 = \frac{1}{n}\vec{e}\cdot\vec{e} = \frac{1}{n}\vec{e}^T\vec{e}$

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

$= \frac{1}{n}(\vec{y} - X\vec{w}) \cdot (y - X\vec{w})$

Rewriting mean squared error

Remider: $\boxed{(AB)^T = B^T A^T}$ $\boxed{A(BC) = (AB)C}$

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

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

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

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

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

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

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

$$= \frac{1}{n}\left(\vec{y}^T\vec{y} - (X^T\vec{y})\cdot\vec{w} - \vec{w}\cdot(X^T\vec{y}) + \vec{w}^T X^T X \vec{w}\right) =$$

$$\boxed{\vec{a}\cdot\vec{b} = \vec{b}\cdot\vec{a}}$$

$$= \frac{1}{n}\left(\vec{y}^T\vec{y} - 2(X^T\vec{y})\cdot\vec{w} + \vec{w}^T X^T X \vec{w}\right)$$

Compute the gradient

$$\frac{dR_{\text{sq}}}{d\vec{w}} = \frac{d}{d\vec{w}}\left(\frac{1}{n}\left(\vec{y}\cdot\vec{y} - 2X^T\vec{y}\cdot\vec{w} + \vec{w}^T X^T X\vec{w}\right)\right)$$

$$= \frac{1}{n}\left(\frac{d}{d\vec{w}}\left(\vec{y}\cdot\vec{y}\right) - \frac{d}{d\vec{w}}\left(2X^T\vec{y}\cdot\vec{w}\right) + \frac{d}{d\vec{w}}\left(\vec{w}^T X^T X\vec{w}\right)\right)$$

Question 🤔

Answer at **q.dsc40a.com**

Which of the following is $\frac{d}{d\vec{w}}(\vec{y} \cdot \vec{y})$?

A. $\vec{y} \cdot \vec{y}$

B. $2\vec{y}$

C. 1

D. 0

$\vec{y}$ doesn't depend on $\vec{w}$

Compute the gradient

$$\begin{aligned}\frac{dR_{\text{sq}}}{d\vec{w}} &= \frac{d}{d\vec{w}}\left(\frac{1}{n}\left(\vec{y}\cdot\vec{y} - 2X^T\vec{y}\cdot\vec{w} + \vec{w}^TX^TX\vec{w}\right)\right) \\ &= \frac{1}{n}\left(\frac{d}{d\vec{w}}\cancel{\left(\vec{y}\cdot\vec{y}\right)} - \frac{d}{d\vec{w}}\left(2X^T\vec{y}\cdot\vec{w}\right) + \frac{d}{d\vec{w}}\left(\vec{w}^TX^TX\vec{w}\right)\right)\end{aligned}$$

0 $2X^T\vec{y}$ $2X^TX\vec{w}$

- $\frac{d}{d\vec{w}}\left(\vec{y}\cdot\vec{y}\right) = 0.$
  - Why? $\vec{y}$ is a constant with respect to $\vec{w}$.
- $\frac{d}{d\vec{w}}\left(\vec{2}X^T\vec{y}\cdot\vec{w}\right) = 2X^Ty.$
  - Why? In groupwork today you will show $\frac{d}{d\vec{x}}\vec{a}\cdot\vec{x} = \vec{a}$.
- $\frac{d}{d\vec{w}}\left(\vec{w}^TX^TX\vec{w}\right) = 2X^TX\vec{w}.$
  - Why? You will prove in homework 4.

Compute the gradient

$$\frac{dR_{\text{sq}}}{d\vec{w}} = \frac{d}{d\vec{w}}\left(\frac{1}{n}\left(\vec{y}\cdot\vec{y} - 2X^T\vec{y}\cdot\vec{w} + \vec{w}^T X^T X\vec{w}\right)\right)$$

$$= \frac{1}{n}\left(\frac{d}{d\vec{w}}\left(\cancel{\vec{y}\cdot\vec{y}}\right) - \frac{d}{d\vec{w}}\left(2X^T\vec{y}\cdot\vec{w}\right) + \frac{d}{d\vec{w}}\left(\vec{w}^T X^T X\vec{w}\right)\right)$$

$$= \frac{1}{n}\left(-2X^T\vec{y} + 2X^T X\vec{w}\right)$$

The normal equations (again)

- To minimize $R_{\text{sq}}(\vec{w})$, set its gradient to zero and solve for $\vec{w}$:

$$\begin{aligned} -2X^T\vec{y} + 2X^TX\vec{w} &= 0 \\ \implies \boxed{X^TX\vec{w} = X^T\vec{y}} \end{aligned}$$

- We have seen this system of equations in matrix form before: the **normal equations**.
- If X^TX is invertible, the solution is

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

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 (Solution 1)

- 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 and minimized the length of the projection error $\|\vec{e}\| = \|\vec{y} - X\vec{w}\|$.

Summary: Regression and linear algebra (Solution 2)

- 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 minimize the mean squared error $R_{\text{sq}}(\vec{w}) = \frac{1}{n} \|\vec{y} - X\vec{w}\|^2$? Using calculus the optimal paramter vector $\vec{w}^*$ is:

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

Roadmap

- 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 $H(x) = w_0 + w_1 x + w_2 x^2$?