# DSC 40A

## Theoretical Foundations of Data Science I

# Last Time

- We used linear algebra to write the mean squared error for a linear prediction rule $H(x) = w_0 + w_1 x$ as

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

where

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

  - $X$ is the **design matrix**.
  - $\vec{w}$ is the **parameter vector**.
  - $\vec{y}$ is the **observation vector**.

## In This Video

We minimize the mean squared error using calculus. The result will soon help us generalize to more exciting regression problems.

## Recommended Reading

Course Notes: Chapter 2, Section 2
Review: Linear Algebra Textbook

## Key Linear Algebra Facts

If $A$ and $B$ are matrices, and $\vec{u}, \vec{v}, \vec{w}, \vec{z}$ are vectors:

- $(A+B)^T = A^T + B^T$
- $(AB)^T = B^T A^T$
- $\vec{u} \cdot \vec{v} = \vec{v} \cdot \vec{u} = \vec{u}^T \vec{v} = \vec{v}^T \vec{u}$
- $\|\vec{u}\|^2 = \vec{u} \cdot \vec{u}$
- $(\vec{u} + \vec{v}) \cdot (\vec{w} + \vec{z}) = \vec{u} \cdot \vec{w} + \vec{u} \cdot \vec{z} + \vec{v} \cdot \vec{w} + \vec{v} \cdot \vec{z}$

## Goal

- We want to minimize the mean squared error:

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

- Strategy: Calculus.

## Goal

- We want to minimize the 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 $\vec{w}$?

## 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,

$$R_{\text{sq}}(\vec{w}) = R_{\text{sq}}(w_0, w_1, \ldots, w_d),$$

  where $w_0, w_1, \ldots, w_d$ are the entries of the vector $\vec{w}$.[1]

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

[1] 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.

## 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:

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

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:

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

- Strategy:
  1. Compute the gradient of $R_{\text{sq}}(\vec{w})$.
  2. Set it to zero and solve for $\vec{w}$.

# Rewrite the Mean Squared Error

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

## Question

Which of the following is equivalent to $R_{\text{sq}}(\vec{w})$ ?

a) $\frac{1}{n}(\vec{y} - X\vec{w}) \cdot (X\vec{w} - y)$
b) $\frac{1}{n}\sqrt{(\vec{y} - X\vec{w}) \cdot (y - X\vec{w})}$
c) $\frac{1}{n}(\vec{y} - X\vec{w})^T(y - X\vec{w})$
d) $\frac{1}{n}(\vec{y} - X\vec{w})(y - X\vec{w})^T$

# Rewrite the Mean Squared Error

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

# Rewrite the Mean Squared Error

$R_{\text{sq}}(\vec{w}) =$

# Compute the Gradient

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

# Compute the Gradient

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

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

## Question

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

a) $\vec{y}\cdot\vec{y}$

b) $2\vec{y}$

c) $1$

d) $0$

## Compute the Gradient

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

## The Normal Equations

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

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

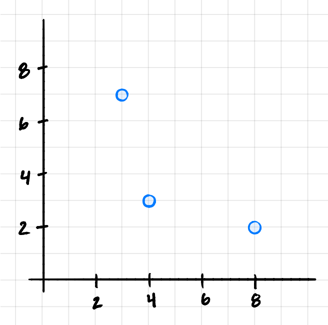
- This is a system of equations in matrix form, called the **normal equations**.

- If inverse exists, solution is[2]

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

[2] Don't actually compute inverse! Use Gaussian elimination or matrix decompositions.

# Example




| $x_i$ | $y_i$ |
|---|---|
| 3 | 7 |
| 4 | 3 |
| 8 | 2 |

# Summary

- We used linear algebra to do simple linear regression in a new way.
- Instead of using our formulas for $w_0$ and $w_1$, we can find these parameters by solving the **normal equations**:

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

- **Next time:** We'll change the form of our prediction rule, and we'll see when the linear algebra still works.