Lecture 9 – Regression in Action and Linear Algebra Review

DSC 40A, Winter 2024

Announcements

- Homework 3 is due **Wed at 11:59pm**.
  - Last HW before the first midterm

- Discusssion session today
  - We modify the scope of discussion session/groupwork so that it aligns with the course better.

- First Midterm exam on Friday next week (Feb. 9th)
  - I will post a practice midterm today

- My OH will be tomorrow 10-12 at HDSI 155

Agenda

- Recap of Lecture 8.
- Connection with correlation.
- Interpretation of formulas.
- Regression demo.
- Linear algebra review.

Recap of Lecture 8

Strategy

$$-\frac{2}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right) = 0 \qquad -\frac{2}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right) x_i = 0$$

1. Solve for w_0 in first equation.
   - The result becomes w_0^*, since it is the “best intercept”.

2. Plug w_0^* into second equation, solve for w_1.
   - The result becomes w_1^*, since it is the “best slope”.

Solve for w_0^*

$$-\frac{2}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i)) = 0$$

$$-\frac{2}{n}\left[\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} w_0 - \sum_{i=1}^{n} w_1 x_i\right] = 0$$

nw_0

Σw_0

$$-\frac{2}{n}\sum y_i + \frac{2}{n} n w_0 + \frac{2}{n}\sum_{i=1}^{n} w_1 x_i = 0$$

$$\rightarrow 2 w_0 = -\frac{2}{n}\sum y_i + \frac{2}{n}\sum w_1 x_i$$

$$w_0 = \frac{1}{n}\sum y_i - \frac{1}{n}\sum w_1 x_i$$

$\bar{y}$

$$w_0 = \bar{y} - \frac{1}{n}\sum w_1 x_i$$

Solve for w_1^*

$$-\frac{2}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))\, x_i = 0$$

$$w_0^* = \bar{y} - \frac{1}{n}\sum w_1 x_i$$

$$-\frac{2}{n}\sum \left(y_i - \left(\left(\bar{y} - \frac{1}{n}\sum w_1 x_i\right) + w_1 x_i\right)\right) x_i = 0$$

$$\sum \left(y_i - \bar{y} + \frac{1}{n}\sum w_1 x_i - w_1 x_i\right)\cdot x_i = 0$$

$$\bar{x}$$

$$\sum \left(y_i - \bar{y} + w_1\left(\frac{1}{n}\sum x_i - x_i\right)\right)\cdot x_i = 0$$

$$\sum \left[(y_i - \bar{y})\cdot x_i + w_1(\bar{x} - x_i)\cdot x_i\right] = 0$$

$$\sum (y_i - \bar{y})\cdot x_i = -\sum w_1(\bar{x} - x_i)\cdot x_i$$

$$w_1 = \frac{\sum (y_i - \bar{y})x_i}{-\sum(\bar{x} - x_i)\cdot x_i} = \frac{\sum (y_i - \bar{y})x_i}{\sum (x_i - \bar{x})x_i}$$

Least squares solutions

- We've found that the values w_0^* and w_1^* that minimize the function $R_{sq}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2$ are

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

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- Let's re-write the slope w_1^* to be a bit more symmetric.

Key fact

The **sum of deviations from the mean** for any dataset is 0.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \qquad \sum_{i=1}^{n}(y_i - \bar{y}) = 0$$

Proof:

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

$$= \sum_{i=1}^{n} x_i - n \cdot \bar{x}$$

$$= \sum_{i=1}^{n} x_i - \cancel{n} \cdot \frac{1}{\cancel{n}} \sum_{i=1}^{n} x_i = 0$$

Equivalent formula for w_1^*

Claim

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

$$= \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})(x_i - \bar{x})} \quad (x_i - \bar{x})^2$$

Proof:

$$\sum (x_i - \bar{x}) = 0, \quad \sum (y_i - \bar{y}) = 0 \Rightarrow \sum (x_i - \bar{x}) \cdot \bar{x} = 0 \cdot \bar{x} = 0, \quad \sum (y_i - \bar{y}) \cdot \bar{x} = 0 \cdot \bar{x} = 0$$

$$w_1^* = \frac{\sum_{i=1}^{n} (y_i - \bar{y})x_i - 0}{\sum_{i=1}^{n} (x_i - \bar{x})x_i - 0} = \frac{\sum (y_i - \bar{y})x_i - \sum (y_i - \bar{y})\bar{x}}{\sum (x_i - \bar{x})x_i - \sum (x_i - \bar{x})\bar{x}}$$

Least squares solutions

- The **least squares solutions** for the slope w_1^* and intercept w_0^* are:

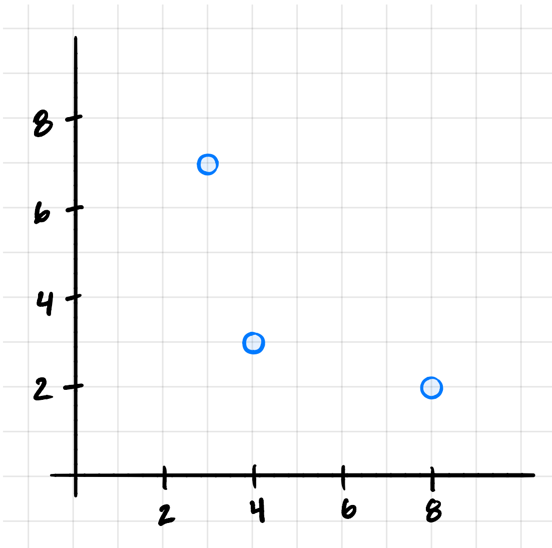
$$w_1^* = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad w_0^* = \bar{y} - w_1\bar{x}$$

  - We also say that w_0^* and w_1^* are **optimal parameters**.

- To make predictions about the future, we use the prediction rule

$$H^*(x) = w_0^* + w_1^* x$$

Example



$\bar{x} = 5$

$\bar{y} = 4$

$$w_1^* = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{-6-6+1}{4+1+9} = -\frac{11}{14}$$

$$w_0^* = \bar{y} - w_1\bar{x} = 4 + \frac{11}{14}\cdot 5 = \frac{56}{14} + \frac{55}{14} = \frac{111}{14}$$

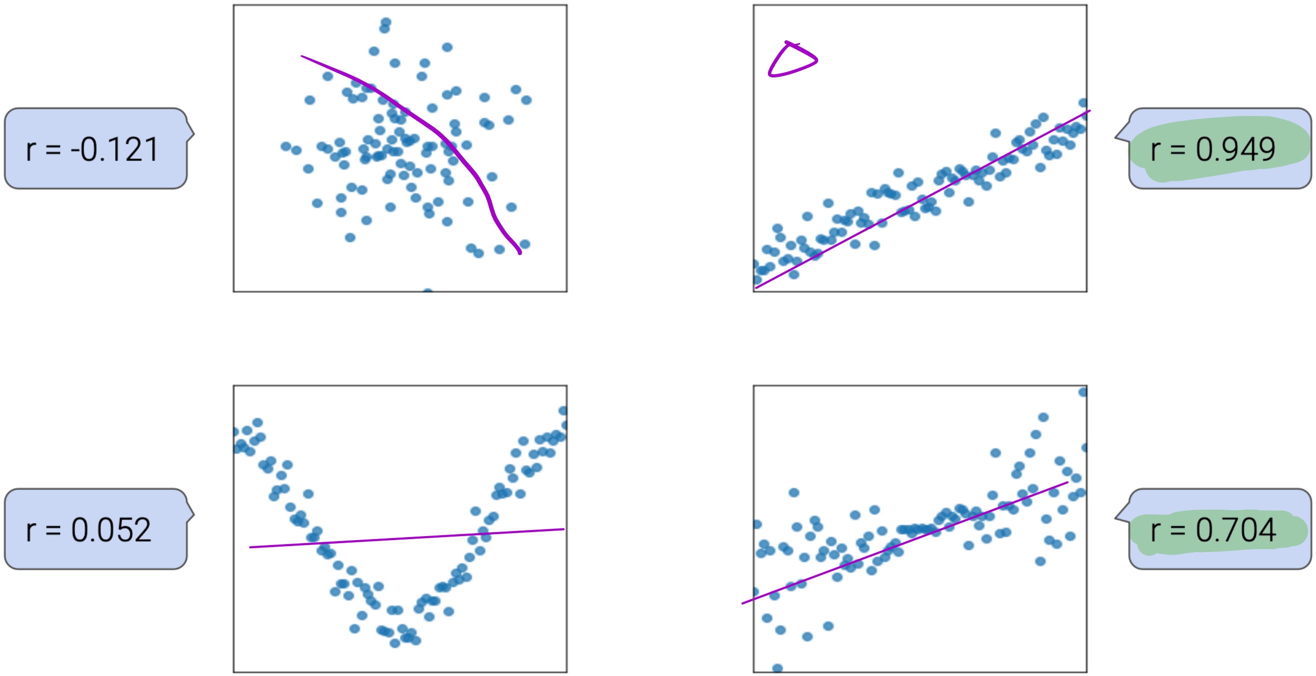
x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
3	7	−2	3	−6	4
4	3	−1	−1	1	1
8	2	3	−2	−6	9

Connection with correlation

Correlation coefficient

- In DSC 10, you were introduced to the idea of correlation.
  - It is a measure of the strength of the **linear association** of two variables, x and y.
  - Intuitively, it measures how tightly clustered a scatter plot is around a straight line.
  - It ranges between -1 and 1.

Patterns in scatter plots



Definition of 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_i's, and $\bar{x}$ be the mean of the x_i's.
  - x_i in standard units is $\dfrac{x_i - \bar{x}}{\sigma_x}$.
  - The correlation coefficient is

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

Another way to express w_1^*

- It turns out that w_1^*, the optimal slope for the linear prediction rule, can be written in terms of r!

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

- It's not surprising that r is related to w_1^*, since r is a measure of linear association.

- Concise way of writing w_0^* and w_1^*:

$$w_1^* = r\frac{\sigma_y}{\sigma_x} \qquad w_0^* = \bar{y} - w_1^*\bar{x}$$

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

$$w^* = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

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

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

$$= \frac{1}{n}\sum \frac{(x_i - \bar{x})\cdot(y_i - \bar{y})}{\sigma_x \cdot \sigma_x}$$

Variance of x

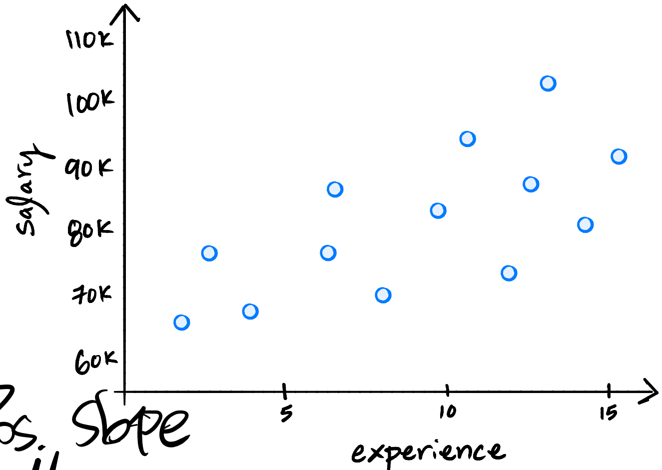
$$= \frac{\frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum (x_i - \bar{x})^2}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Interpretation of formulas

Interpreting the slope

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



>0 → Pos. Corr. → Pos. Slope

<0 → Neg. Corr. → Neg. Slope.

- σ_y and σ_x are always non-negative. As a result, the sign of the slope is determined by the sign of r.
- As the y values get more spread out, σ_y increases and so does the slope.
- As the x values get more spread out, σ_x increases and the slope decreases.

Interpreting the intercept

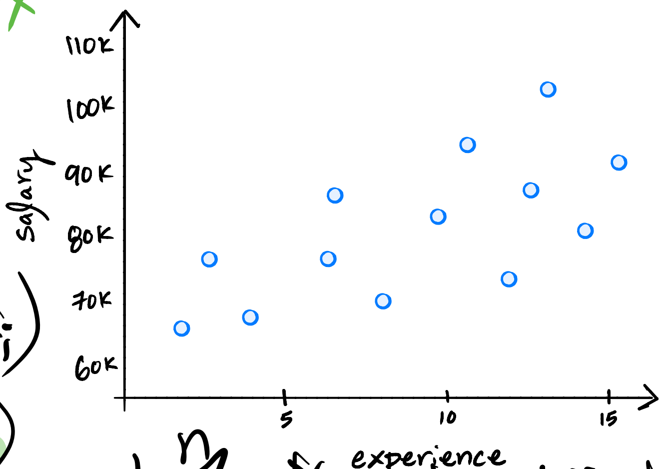
$H^*(x) = w_0^* - w_1^* x$

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

$w_0^* = \frac{1}{n}\sum y_i - w_1^* \cdot \left(\frac{1}{n}\sum x_i\right)$

$= \frac{1}{n}\sum (y_i - w_1^* x_i)$

$w_0^* = \frac{1}{n}\sum_{i=1}^{n} w_0^* = \frac{1}{n} n \cdot w_0^* = w_0^*$



- What is $H^*(\bar{x})$?

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

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

$= \bar{y} - r\frac{\sigma_y}{\sigma_x}(\bar{x} - x)$

$H^*(\bar{x}) = \bar{y} - r\frac{\sigma_y}{\sigma_x}(\bar{x} - \bar{x})$ → 0

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

Discussion Question

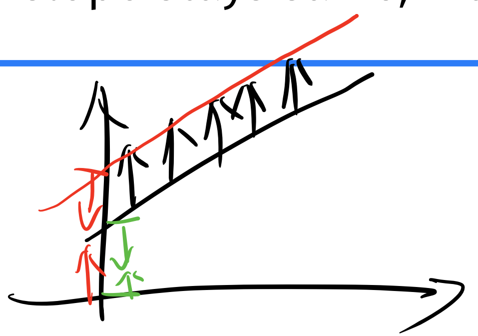
We fit a linear prediction rule for salary given years of experience. Then everyone gets a $5,000 raise. Which of these happens?

a) slope increases, intercept increases

b) slope decreases, intercept increases

c) slope stays same, intercept increases

d) slope stays same, intercept stays same



Regression demo

Let's see regression in action. Follow along here.

Linear algebra review

Wait... why do we need linear algebra?

- Soon, we'll want to make predictions using more than one feature (e.g. predicting salary using years of experience and GPA).
- Thinking about linear regression in terms of **linear algebra** will allow us to find prediction rules that
  - use multiple features.
  - are non-linear.
- Before we dive in, let's review.

Matrices

- An $m \times n$ **matrix** is a table of numbers with m rows and n columns.

- We use upper-case letters for matrices.

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

- A^T denotes the transpose of A:

$$A^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

Matrix addition and scalar multiplication

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

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} + \begin{bmatrix} 7 & 8 & 9 \\ -1 & -2 & -3 \end{bmatrix} = \begin{bmatrix} 8 & 10 & 12 \\ 3 & 3 & 3 \end{bmatrix}$$

- Scalar multiplication occurs elementwise, too:

$$2 \cdot \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 6 \\ 8 & 10 & 12 \end{bmatrix}$$

Matrix-matrix multiplication

- We can multiply two matrices A and B only if

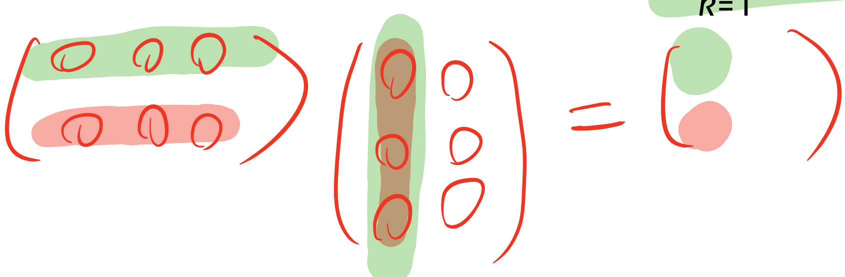
  # columns in A = # rows in B.

- If A is $m \times n$ and B is $n \times p$, the result is $m \times p$.
  - This is **very useful**.

- The ij entry of the product is:

$(2\times3)(3\times2) \rightarrow (2\times2)$

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}$$



Some matrix properties

- Multiplication is Distributive:

$$A(B + C) = AB + AC$$

- Multiplication is Associative:

$$(AB)C = A(BC)$$

- Multiplication is **not commutative**:

$$AB \neq BA$$

A (m×n)
B (n×p)
AB → (m×p)   BA
(n×p) · (m×n)

- Transpose of sum:

$$(A + B)^T = A^T + B^T$$

- Transpose of product:

(m×p)^T → (p×m)

$$(AB)^T = B^T A^T$$

Vectors

- An **vector** in $\mathbb{R}^n$ is an $n \times 1$ matrix.

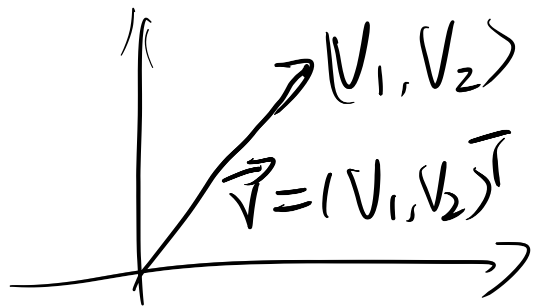
- We use lower-case letters for vectors.

$$\vec{v} = \begin{bmatrix} 2 \\ 1 \\ 5 \\ -3 \end{bmatrix}$$

- Vector addition and scalar multiplication occur elementwise.

Geometric meaning of vectors

- A vector $\vec{v} = (v_1, \ldots, v_n)^T$ is an arrow to the point $(v_1, \ldots, v_n)$ from the origin.



- The **length**, or **norm**, of $\vec{v}$ is $\|\vec{v}\| = \sqrt{v_1^2 + v_2^2 + \ldots + v_n^2}$.

Dot products

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

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

- Definition:

$\vec{u} = (n \times 1) \rightarrow \vec{u}^T = (1 \times n)$ $\vec{u}^T \vec{v}$

$\vec{v} = (n \times 1)$ $= (1 \times n) \cdot (n \times 1)$

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

$= (1 \times 1) \rightarrow$ Scalar

- The result is a **scalar**!

Discussion Question

Which of these is another expression for the length of $\vec{u}$?

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

b) $\sqrt{\vec{u}^2}$

c) $\sqrt{\vec{u} \cdot \vec{u}}$

d) $\vec{u}^2$

Properties of the dot product

- Commutative:

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

- Distributive:

$$\vec{u} \cdot (\vec{v} + \vec{w}) = \vec{u} \cdot \vec{v} + \vec{u} \cdot \vec{w}$$

Matrix-vector multiplication

- Special case of matrix-matrix multiplication.
- Result is always a vector with same number of rows as the matrix.
- One view: a “mixture” of the columns.

$$\begin{bmatrix} 1 & 2 & 1 \\ 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = a_1 \begin{bmatrix} 1 \\ 3 \end{bmatrix} + a_2 \begin{bmatrix} 2 \\ 4 \end{bmatrix} + a_3 \begin{bmatrix} 1 \\ 5 \end{bmatrix}$$

- Another view: a dot product with the rows.

Discussion Question

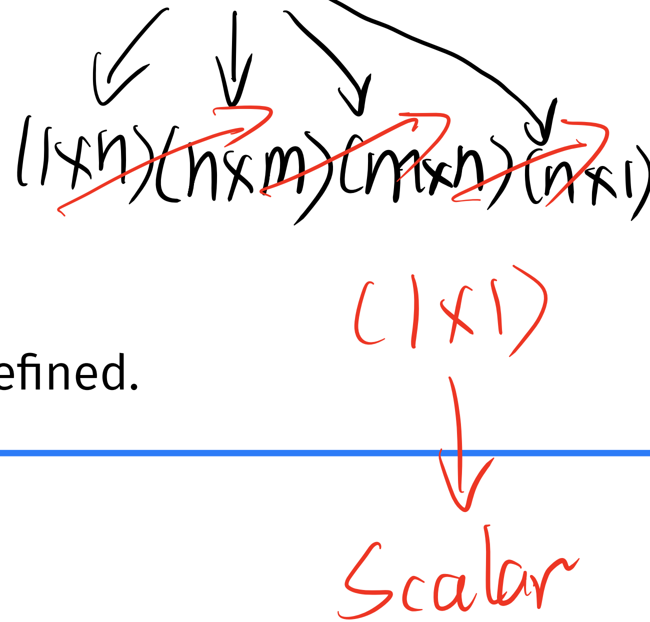
If A is an $m \times n$ matrix and $\vec{v}$ is a vector in $\mathbb{R}^n$, what are the dimensions of the product $\vec{v}^T A^T A \vec{v}$?

a) $m \times n$ (matrix)

b) $n \times 1$ (vector)

c) 1×1 (scalar)

d) The product is undefined.



Summary

Summary, next time

- We can re-write the optimal parameters for the regression line

$$w_1^* = r\frac{\sigma_y}{\sigma_x} \qquad w_0^* = \bar{y} - w_1^*\bar{x}$$

- We can then make predictions using $H^*(x) = w_0^* + w_1^* x$.
- We will need linear algebra in order to generalize regression to work with multiple features.
- **Next time**: Continue linear algebra review. Formulate linear regression in terms of linear algebra.