

Lecture 9 – Regression in Action and Linear Algebra Review



DSC 40A, Spring 2023

Announcements

- ▶ Homework 3 is due **Tuesday at 11:59pm**.
 - ▶ Come to office hours. See dsc40a.com/calendar for the schedule.
 - ▶ It's a pretty long homework. Start early!
- ▶ Solutions to Groupwork 3 and Homework 2 are now available on Campuswire.
 - ▶ Reviewing them will help you on upcoming assignments and exams.

Agenda

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

Recap of Lecture 8

The best **linear** prediction rule

- ▶ Last time, we used multivariable calculus to find the slope w_1^* and intercept w_0^* that minimized the MSE for a linear prediction rule of the form

$$\underline{H(x) = w_0 + w_1 x}$$

int slope

- ▶ In other words, we minimized this function:

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

actual *predicted*

Optimal parameters

- ▶ We found the optimal parameters to be:

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

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

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

- ▶ This line is the **regression line**.

Connection with correlation

Correlation coefficient

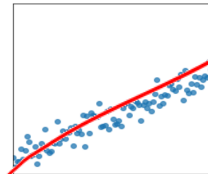
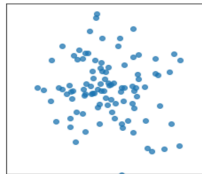
- ▶ In DSC 10, you were introduced to the idea of correlation.
 - ▶ It is a measure of the strength of the **linear association** of two variables, x and y .
 - ▶ Intuitively, it measures how tightly clustered a scatter plot is around a straight line.
 - ▶ It ranges between -1 and 1 .



Patterns in scatter plots

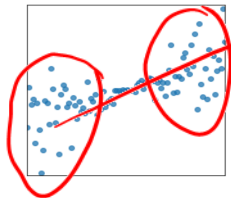
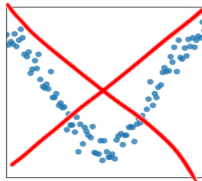


$r = -0.121$



$r = 0.949$

$r = 0.052$



$r = 0.704$

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.

1 SD

- ▶ x_i in standard units is $\frac{x_i - \bar{x}}{\sigma_x}$. ← how many SDs above mean
- ▶ The correlation coefficient is

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

Standard units

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 !

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

← formula from DSC 10

- ▶ 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^* :

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

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

rhs:

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

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

← distributive rule

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

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

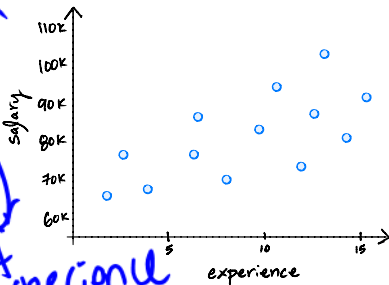
$$\leftarrow (\sigma_x)^2 = \text{var}(x)$$

Interpretation of formulas

Interpreting the slope

units: \$ per year
represents how much
more salary you'd
make for every
year of experience

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



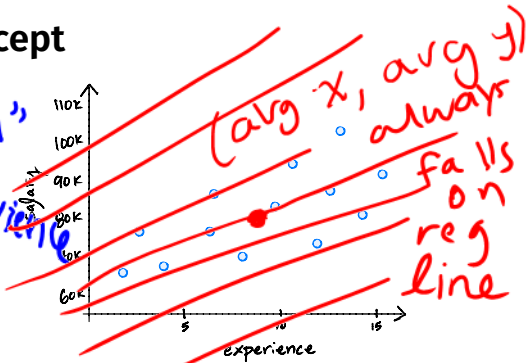
- ▶ σ_y and σ_x are always non-negative. As a result, the sign of the slope is determined by the sign of r .
- ▶ As the y values get more spread out, σ_y increases and so does the slope.
- ▶ As the x values get more spread out, σ_x increases and the slope decreases.

Interpreting the intercept

units: dollars
represents 'new
salary - 0 years

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

grad
of
experience



- ▶ What is $H^*(\bar{x})$?

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

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

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

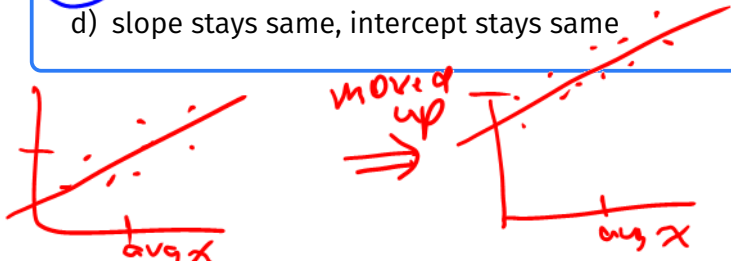
$$= \bar{y} \leftarrow \text{avg } y$$

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?

$$w_0^* = \bar{y} - w_1^* \bar{x} \quad \rightarrow \text{inc by } 5000$$

- 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 gradient descent 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

*rows first
columns*

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

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

- ▶ A^T denotes the transpose of A :

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

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:

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

Some matrix properties

- ▶ Multiplication is Distributive:

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

- ▶ Multiplication is Associative:

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

- ▶ Multiplication is **not commutative**:

$$AB \neq BA$$

- ▶ Transpose of sum:

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

- ▶ Transpose of product:

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

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

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

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:

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

- ▶ 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

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

- ▶ 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.