Lecture 7 – More Simple Linear Regression

DSC 40A, Fall 2021 @ UC San Diego

Suraj Rampure, with help from **many others**

Announcements

- Groupwork 3 is due **tonight at 11:59pm**.
- Homework 3 is due **Monday at 11:59pm**. **No slip days allowed!**
  - Everyone now has 5 slip days, though.
- Midterm exam is on **Thursday, 10/21, from 11AM-12:30PM**. Fully remote.
  - Covers Lectures 1-7.
  - Will receive a PDF on Gradescope and must submit it back within 90 minutes (80 minutes for the exam + 10 minutes for uploading).
  - More details this weekend.
- Midterm review session on **Tuesday, 10/19 from 5-8PM in PCNYH 109**.

Midterm study strategy

- Review the solutions to previous homeworks and groupworks.
  - Homework 2 solutions are now up.
- Identify which concepts are still iffy. Re-watch lecture, post on Campuswire, come to office hours.
- Look at the past exams at `https://dsc40a.com/resources`.
- Study in groups.
- Make a "cheat sheet".
- **Remember:** it's just an exam.

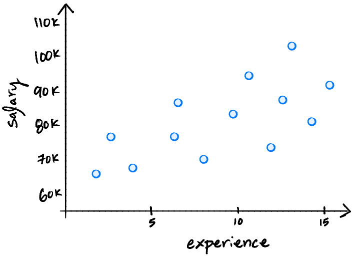
Agenda

- Recap of Lecture 6.
- Correlation.
- Practical demo.
- Linear algebra review.

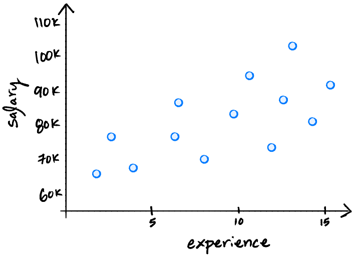
Recap of Lecture 6

Linear prediction rules

- **New:** Instead of predicting the same future value (e.g. salary) h for everyone, we will now use a **prediction rule** $H(x)$ that uses **features**, i.e. information about individuals, to make predictions.
- We decided to use a **linear** prediction rule, which is of the form $H(x) = w_0 + w_1 x$.
  - w_0 and w_1 are called **parameters**.



Before



Now

Finding the best linear prediction rule

- In order to find the best linear prediction rule, we need to pick a loss function and minimize the corresponding empirical risk.
  - We chose squared loss, $\left(y_i - H(x_i)\right)^2$, as our loss function.

- The MSE is a function R_{sq} of a function H.

$$R_{\text{sq}}(H) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - H(x_i)\right)^2$$

- But since H is linear, we know $H(x_i) = w_0 + w_1 x_i$.

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

Finding the best linear prediction rule

- Our goal last lecture was to find the slope w_1^* and intercept w_0^* that minimize the MSE, $R_{\text{sq}}(w_0, w_1)$:

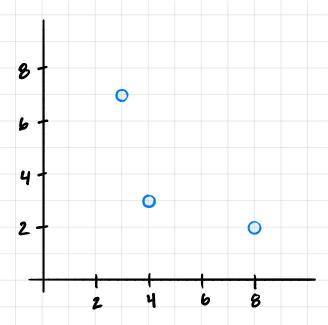
$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2$$

- We did so using multivariable calculus.

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

- To make predictions: $H^*(x) = w_0^* + w_1^*(x)$.

Example



$\bar{x} =$

$\bar{y} =$

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

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

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				
4	3				
8	2				

Terminology

- x: **features**.
- y: **response variable**.
- w_0, w_1: **parameters**.
- w_0^*, w_1^*: **optimal parameters**.
  - Optimal because they minimize mean squared error.
- The process of finding the optimal parameters for a given prediction rule and dataset is called "**fitting to the data**".
- $R_{sq}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - (w_0 + w_1 x_i)\right)^2$: **mean squared error**, **empirical risk**.

Discussion Question

Consider a dataset with just two points, $(2, 5)$ and $(4, 15)$. Suppose we want to fit a linear prediction rule to this dataset by minimizing mean squared error.

What are the values of w_0^* and w_1^* that minimize mean squared error?

a) $w_0^* = 2, w_1^* = 5$

b) $w_0^* = 3, w_1^* = 10$

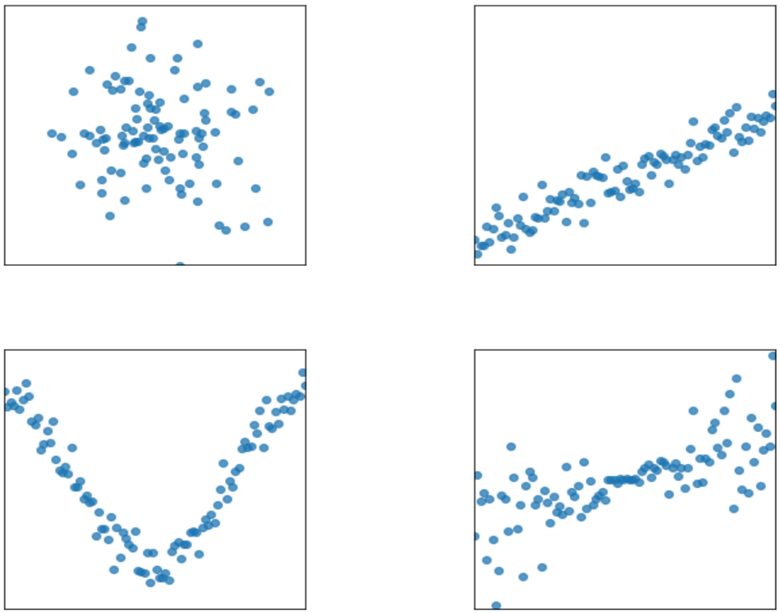
c) $w_0^* = -2, w_1^* = 5$

d) $w_0^* = -5, w_1^* = 5$

e) Impossible to tell

To answer, go to `menti.com` and enter the code 3640 8748.

Correlation

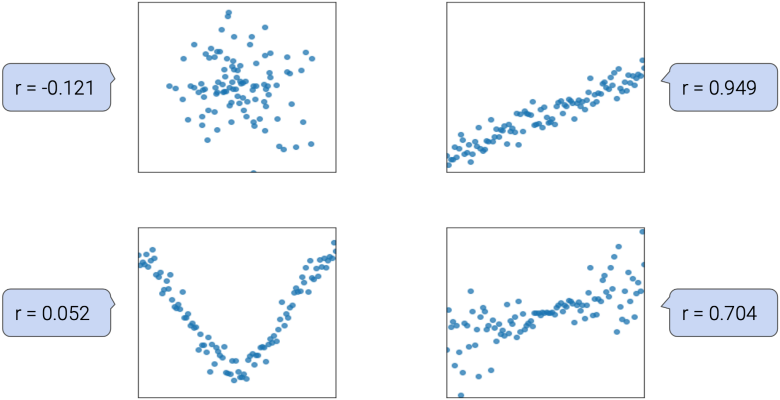


Correlation coefficient

- 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 is a measure of how tightly clustered a scatter plot is around a straight line.
- The correlation coefficient, r, is defined as **the average of the product of x and y, when both are in standard units**.
  - x_i in standard units: $\dfrac{x_i - \bar{x}}{\sigma_x}$.

Properties of the correlation coefficient r

- r has no units.
- It ranges between -1 and 1.
  - $r = 1$ indicates a perfect positive linear association (x and y lie exactly on a straight line that is sloped upwards).
  - $r = -1$ indicates a perfect negative linear association between x and y.
  - The closer r is to 0, the weaker the linear association between x and y is.
  - r says nothing about non-linear association.
- **Correlation != causation.**


r = -0.121
r = 0.949
r = 0.052
r = 0.704

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

Note that

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

so $n\sigma_x^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2$.

$$r\frac{\sigma_y}{\sigma_x}$$

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

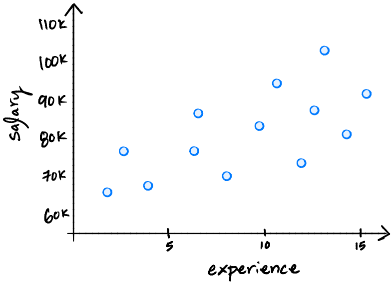
$$= \frac{1}{n\sigma_x\cancel{\sigma_y}}\cdot\frac{\cancel{\sigma_y}}{\sigma_x}\left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]$$

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

Nice!

Interpreting the slope

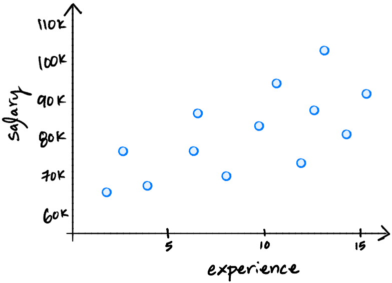
$$w_1^* = r\frac{\sigma_y}{\sigma_x}$$



- 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

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



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

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

To answer, go to `menti.com` **and enter the code 3640 8748.**

Practical demo

Follow along with the demo by clicking the **code** link on the course website next to Lecture 7.

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.
- **No linear algebra on the midterm :)**

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:

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

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 $\vec{v} = (v_1, \ldots, v_n)$ 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:

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

- 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) $\vec{u}^2$

To answer, go to `menti.com` and enter the code 3640 8748.

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.

To answer, go to `menti.com` and enter the code 3640 8748.

Summary

Summary, next time

- The correlation coefficient, r, measures the strength of the linear association between two variables x and y.

- We can re-write the optimal parameters for the linear prediction rule (under squared loss) as

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

- 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**: Formulate linear regression in terms of linear algebra.