Lecture 9 – More Simple Linear Regression

DSC 40A, Fall 2022 @ UC San Diego

Midterm study strategy

- Review the solutions to previous homeworks and groupworks.
- 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".

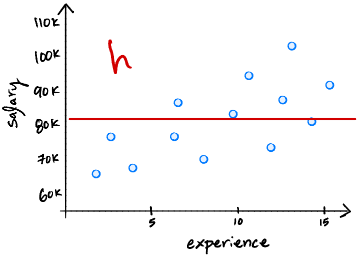
Agenda

- Recap of Lecture 8.
- Correlation.
- Practical demo.

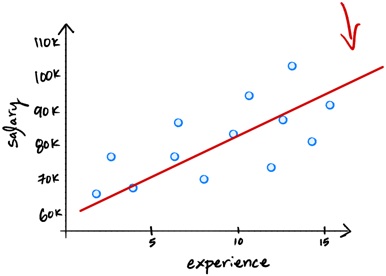
Recap of Lecture 8

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, $(y_i - H(x_i))^2$, as our loss function.

↳ single data point

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

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

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

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

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, $R_{sq}(w_0, w_1)$:

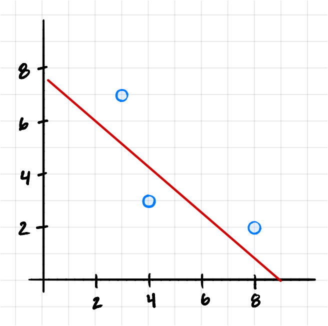
$$R_{sq}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - (w_0 + w_1 x_i)\big)^2$$

- We did so using multivariable calculus.

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

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

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**".

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

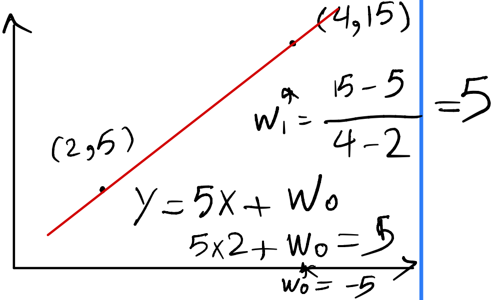
a) $w_0^* = 2, w_1^* = 5$

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

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

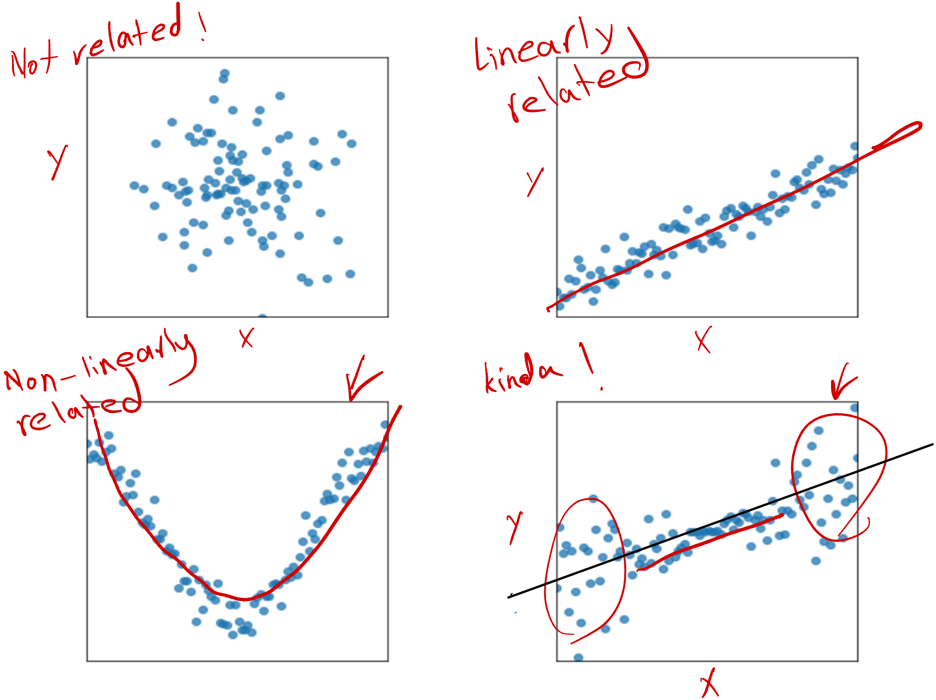
d) $w_0^* = -5, w_1^* = 5$

e) Impossible to tell



To answer, go to `menti.com` **and enter the code 4821 5997.**

Correlation


Not related!
y
x
Linearly related
y
x
Non-linearly related
kinda!
y
x

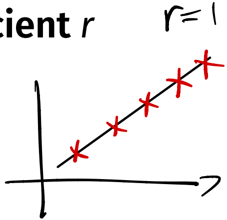
Correlation coefficient

- 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: $\frac{x_i - \bar{x}}{\sigma_x}$.

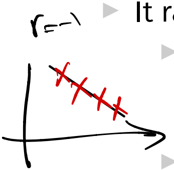
$$r = \frac{1}{n} \sum_{i=1}^{n} x_i y_i \quad \left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

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

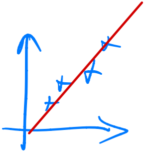
Properties of the correlation coefficient r

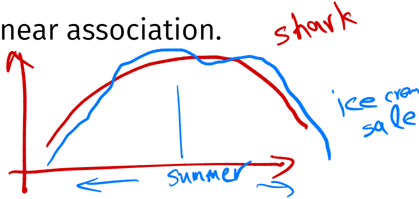


- r has no units.

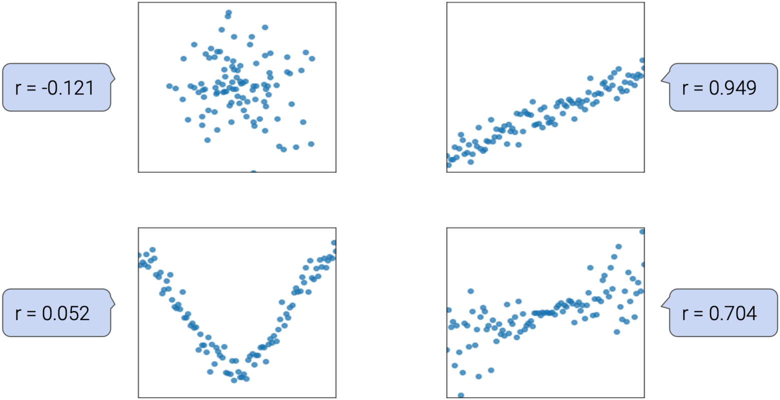


- It ranges between -1 and 1.
  - r = 1 indicates a perfect positive linear association (x and y lie exactly on a straight line that is sloped upwards).
  - r = –1 indicates a perfect negative linear association between x and y.
  - The closer r is to 0, the weaker the linear association between x and 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!

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

no units

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

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

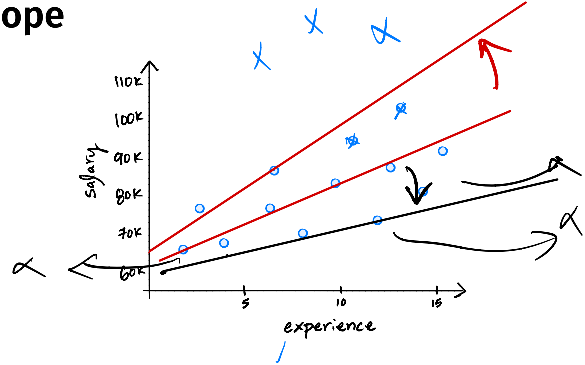
$$= \frac{1}{n\sigma_x\sigma_y}\frac{\sigma_y}{\sigma_x}\left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]$$

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

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

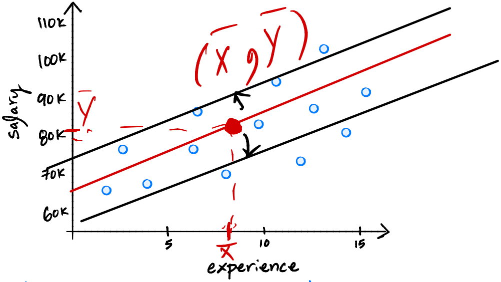
Interpreting the slope

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



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

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

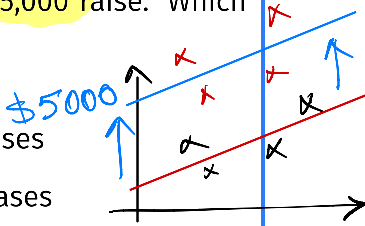
$$H^*(\bar{x}) = w_0^* + w_1^* \bar{x} = (\bar{y} - w_1^* \bar{x}) + w_1^* \bar{x}$$

$$= \bar{y}$$

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

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

To answer, go to `menti.com` **and enter the code 4821 5997.**

Practical demo

Follow along with the demo by clicking the **code** link on the course website next to Lecture 9.

Summary

Summary

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

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