Lecture 8 – More Simple Linear Regression

DSC 40A, Fall 2022 @ UC San Diego

Dr. Truong Son Hy, with help from **many others**

Announcements

- Look at the readings linked on the course website!

- Groupwork Relsease Day: Thursday afternoon
  Groupwork Submission Day: Monday midnight
  Homework Release Day: Friday after lecture
  Homework Submission Day: Friday before lecture

- See `dsc40a.com/calendar` for the Office Hours schedule.

Midterm study strategy

- Review the solutions to previous homeworks and groupworks.
- Re-watch lecture, post on Campuswire, come to office hours.
- Look at the past exams at `https://dsc40a.com/resources`.
- Study in groups.
- **Remember:** it's just an exam.

Agenda

- Recap of simple linear regression.
- Correlation.
- Linear algebra review.

Recap of simple linear regression

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**.

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

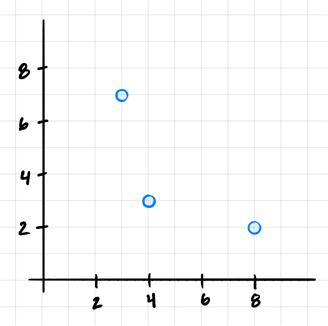
$$R_{sq}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - (w_0 + w_1 x_i)\right)^2$$

- We did so using multivariable calculus.

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

- 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				

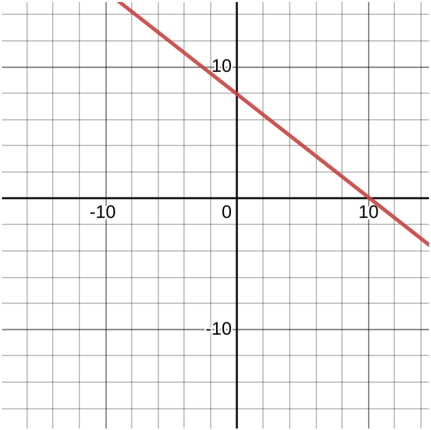
Let's solve it by computer programming!

```python
import numpy as np
def simple_lr(x, y):
    x_bar = np.mean(np.array(x))
    y_bar = np.mean(np.array(y))
    print('Table:')
    num_samples = len(x)
    sum_products = 0
    sum_squares = 0
    for i in range(num_samples):
        x_diff = x[i] - x_bar
        y_diff = y[i] - y_bar
        prod = x_diff * y_diff
        square = x_diff * x_diff
        sum_products += prod
        sum_squares += square
        print(x[i],y[i],x_diff,y_diff,prod,square)
    w1_star = sum_products / sum_squares
    w0_star = y_bar - w1_star * x_bar
    return x_bar, y_bar, w0_star, w1_star
```

```
x = [3, 4, 8]
y = [7, 3, 2]
x_bar, y_bar, w0_star, w1_star = simple_lr(x, y)

print('x_bar =', x_bar)
print('y_bar =', y_bar)
print('w1_star = ', w1_star)
print('w0_star = ', w0_star)
-------------------------------------------------
Table:
3 7 -2.0 3.0 -6.0 4.0
4 3 -1.0 -1.0 1.0 1.0
8 2 3.0 -2.0 -6.0 9.0
x_bar = 5.0
y_bar = 4.0
w1_star =  -0.7857142857142857
w0_star =  7.928571428571429
```

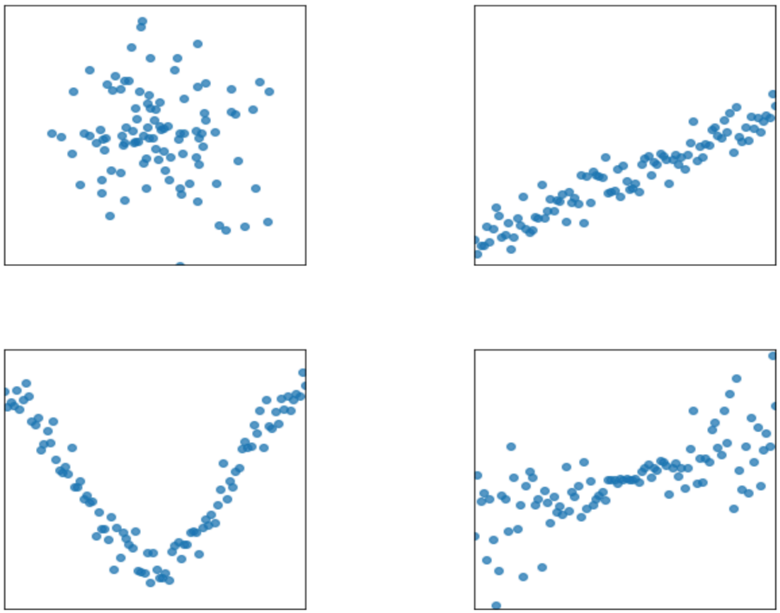
Solution to example



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**.

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

    y_i in standard units: $\dfrac{y_i - \bar{y}}{\sigma_y}$.

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

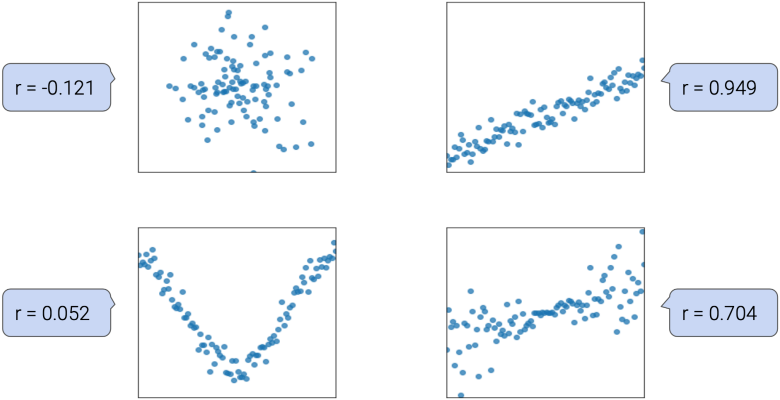
    y_i in standard units: $\frac{y_i - \bar{y}}{\sigma_y}$.

  - Definition of r:

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

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.


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

By definition, we have:

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

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

By definition, we have:

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

Thus:

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

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

By definition, we have:

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

Thus:

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

On the another hand, we also have:

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

That leads to:

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

By definition, we have:

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

Thus:

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

On the another hand, we also have:

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

That leads to:

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

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

By definition, we have:

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

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

By definition, we have:

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

Therefore:

$$w_1^* = \frac{nr\sigma_x\sigma_y}{n\sigma_x^2} =$$

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

By definition, we have:

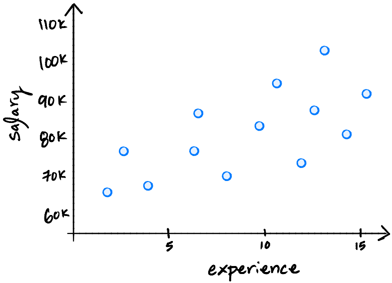
$$w_1^* = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Therefore:

$$w_1^* = \frac{nr\sigma_x\sigma_y}{n\sigma_x^2} = r\frac{\sigma_y}{\sigma_x}$$

Interpreting the slope

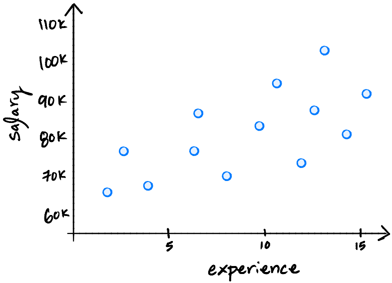
$$w_1^* = r \frac{\sigma_y}{\sigma_x}$$



- 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

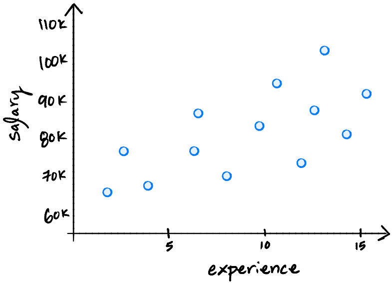
$$w_0^* = \bar{y} - w_1^* \bar{x}$$



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

Interpreting the intercept

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

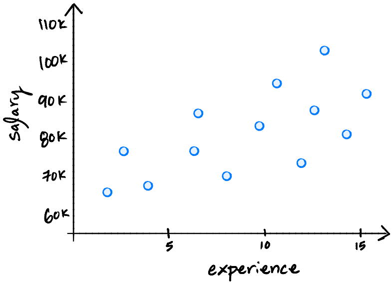


- What is $H^*(\bar{x})$?

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

Interpreting the intercept

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



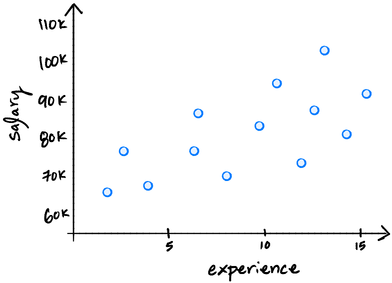
- What is $H^*(\bar{x})$?

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

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

Interpreting the intercept

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



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

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

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

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

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

Answer: C

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

$$\text{\# columns in } A = \text{\# rows in } B.$$

- 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) $\vec{u}^2$

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) $\vec{u}^2$

Answer: C

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.

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.

Answer: C

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.