

Lecture 8 – Simple Linear Regression



DSC 40A, Spring 2023

Announcements

- ▶ Discussion is tonight at 7pm or 8pm in FAH 1101.
 - ▶ Please attend the section you are enrolled in.
 - ▶ Come to work on Groupwork 3, which is due **tonight at 11:59pm.**
 - ▶ It's a pretty long groupwork assignment; it's okay if you don't finish, but review the solutions afterwards because they'll help with Homework 3.
- ▶ Homework 3 is out, due **Tuesday at 11:59pm.**

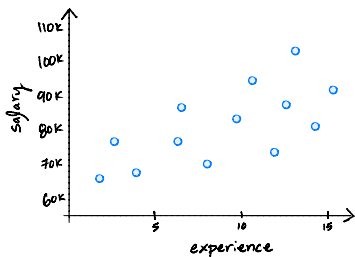
Agenda

- ▶ Recap of Lecture 7.
- ▶ Minimizing mean squared error for the linear prediction rule.
- ▶ Connection with correlation.

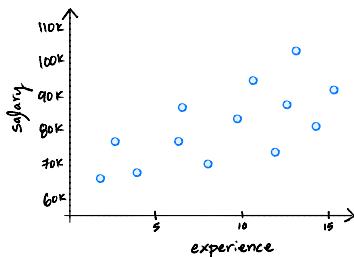
Recap of Lecture 7

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 $H(x) = w_0 + w_1x$.
 - ▶ 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.
- ▶ The MSE is a function R_{sq} of a function H .

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

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

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

Finding the best **linear** prediction rule

- ▶ **Goal:** Find the slope w_1^* and intercept w_0^* that minimize the MSE, $R_{\text{sq}}(w_0, w_1)$:

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

- ▶ **Strategy:** To minimize $R(w_0, w_1)$, compute the gradient (vector of partial derivatives), set it equal to zero, and solve.

Minimizing mean squared error for the linear prediction rule

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

Discussion Question

Choose the expression that equals $\frac{\partial R_{\text{sq}}}{\partial w_0}$.

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

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

$$\frac{\partial R_{\text{sq}}}{\partial w_0} =$$

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

$$\frac{\partial R_{\text{sq}}}{\partial w_1} =$$

Strategy

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

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^*

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

Solve for w_1^*

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

Least squares solutions

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

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

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.

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

Proof:

Equivalent formula for w_1^*

Claim

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

Proof:

Least squares solutions

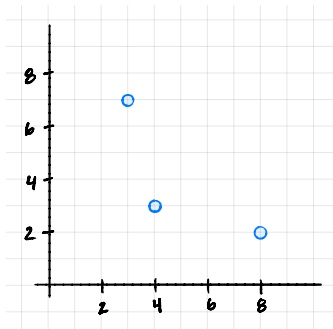
- ▶ The **least squares solutions** for the slope w_1^* and intercept w_0^* are:

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

- ▶ 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



$$\bar{x} =$$

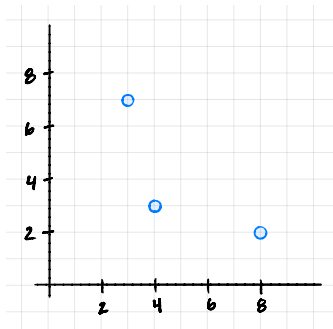
$$\bar{y} =$$

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

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

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				

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**”.
- ▶ $R_{sq}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^n (y_i - (w_0 + w_1 x_i))^2$: **mean squared error, empirical risk**.

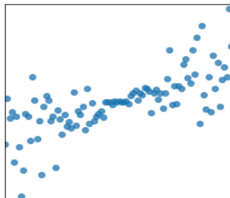
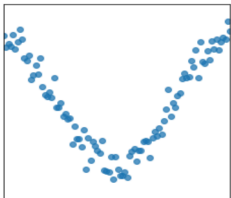
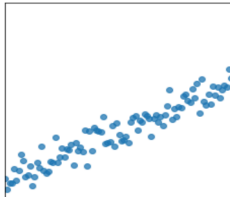
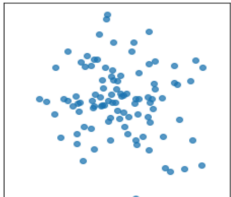
Discussion Question

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

Connection with correlation

Patterns in scatter plots

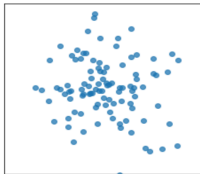


Correlation coefficient

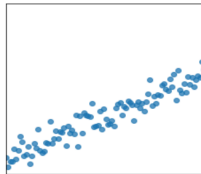
- ▶ In DSC 10, you were introduced to the idea of correlation.
 - ▶ It is a measure of the strength of the **linear association** of two variables, x and y .
 - ▶ Intuitively, it measures how tightly clustered a scatter plot is around a straight line.
 - ▶ It ranges between -1 and 1 .

Patterns in scatter plots

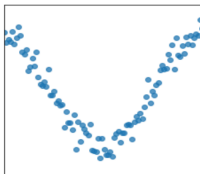
$r = -0.121$



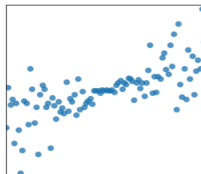
$r = 0.949$



$r = 0.052$



$r = 0.704$



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.

- ▶ x_i in standard units is $\frac{x_i - \bar{x}}{\sigma_x}$.

- ▶ 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 !

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

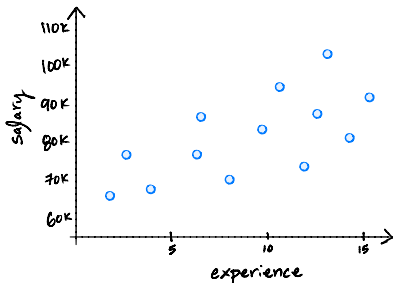
- ▶ It's not surprising that r is related to w_1^* , since r is a measure of linear association.
- ▶ Concise way of writing w_0^* and w_1^* :

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

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

Interpreting the slope

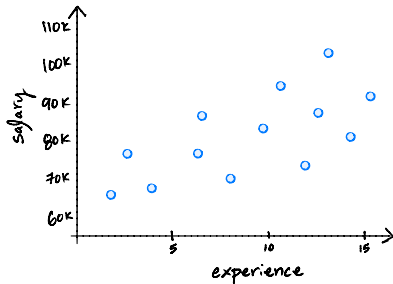
$$W_1^* = r \frac{\sigma_y}{\sigma_x}$$



- ▶ 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 $H^*(\bar{x})$?

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