04/19/2023

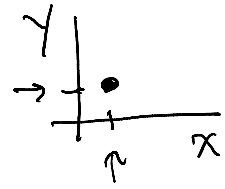
Chapter 10 - Correlation and Linear Regression

Introduction

consider two numerical variables denoted by X and Y.

Ex: i) Height (ft), Weight (lbs)

ii) Time studying (hrs), Exam Score (0 to 100)



Observations (X, Y) are made on subjects or objects under study.

In n = sample size, the sample may be listed as follows:

$$(X_1, Y_1)$$
$$(X_2, Y_2)$$
$$\vdots$$
$$(X_n, Y_n)$$

This is called a bivariate set of data.

Main Questions

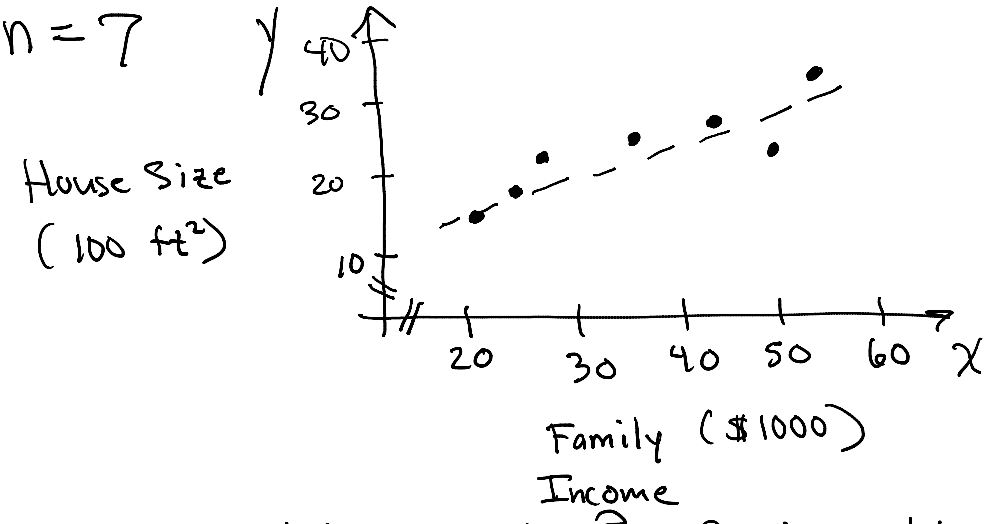
1. How does one describe the relationship between X and Y?
2. What is the "strength" of the relationship between X and Y?
3. Given the value of X, can one predict the value of Y?

Scatterplot

The bivariate data may be visualized using a picture/graph called a **scatterplot**.

Ex. (see new handout)

$n = 7$



Q: Relationship between x and y? Positive relationship

Q: Does a straight line adequately capture the (x, y) relationship? Yes

Q: What is the strength of the relationship between x and y? Strong

Correlation

The strength of the **linear** relationship between x and y can be quantified by determining the sample **correlation coefficient** (r).

r is a measure of how closely the points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ fall on a straight line.

Properties of r (see p. 464)

1. $-1 \leq r \leq 1$

2. r is unitless and doesn't depend on which variable is labeled X and Y.

3. If $r > 0$, then X and Y have a positive relationship

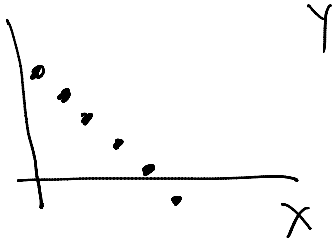
4. If $r < 0$, then X and Y " " negative " .

5. If $r = +1$ or $r = -1$, then X and Y have a perfect positive or negative linear relationship, respectively.

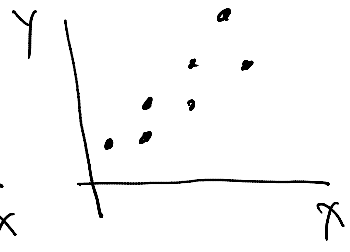
6. If r is close to 0, then X and Y have a very weak linear relationship so X by itself is not useful in predicting Y.

Ex:

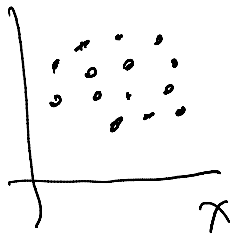
$r = +1$



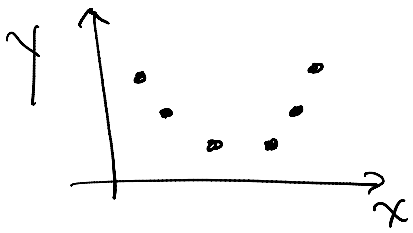
$r = -1$



$r \approx +0.8$ (Brent's guess)



$r \approx 0$



$r = 0$

Def:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}\sqrt{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}}$$

Ex: House size Family Income

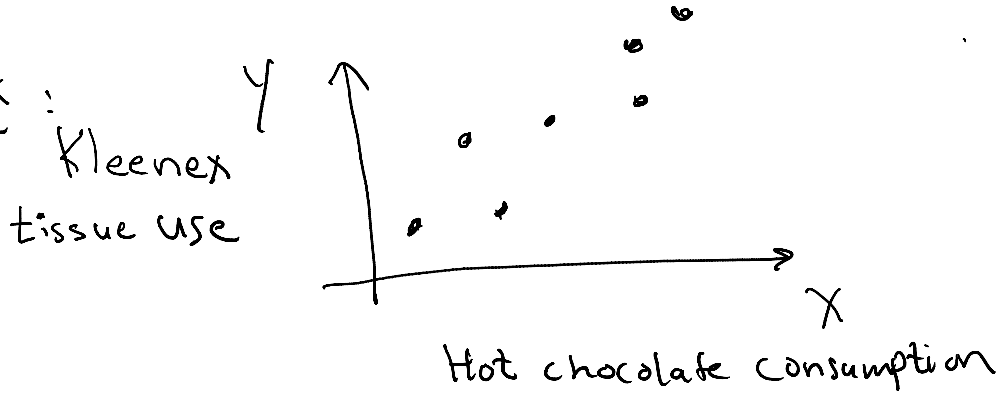
One can show that

$r = 0.82$, which is a strong positive relationship.

NOTE: Consider using statdisk.com or other software to compute r.

Remark: Correlation does not imply causation.

Ex:



There may be other variables that explain the relationship between X and Y.
e.g., cold weather.