Chapter 3 - Describing Bivariate Data

In this chapter, we will introduce:

1. Differences between the association between two variables and causality.
2. Independent and dependent variables.
3. Numeric measure of a bivariate association.
4. Using this numeric measure to estimate the slope and intercept of a line that best fits the association between the two varaibles.

Motivation - Does the knowledge of the values of one variable help you understand the distribution of values of another variable? How do you measure and assess this association?

Tools in this chapter will examine associations or relationships between two variables that are **linear**.

Causality - the value of one variable "causes" or influences the value of the second variable. Changing the value of the first variable may change the value of the second varaible. (example: price and quantity)

Association - knowing the value of first variable allows you to guess at the value of the second variable. Changing the value of the first variable will have no effect on the second variable. (example: height and weight).

Causal relationships are often supported by theory.

Other ways of examining the relationship between two variables

1. Dependent variable - variable of interest - variations in the variable described by exploratory (independent) variable(s). (example: Real GDP)
2. Independent variable(s) - variable used to help describe or explain a dependent variable. (example: Consumption goods, Government expenditures, Investments).

Graphical presentation of bivariate data - scatterplot.

Direction of points and clustering of points.

Numerical Measures of Bivariate data (Interval variables x,y)

Correlation coefficient. Formula:

$$r = \frac{s_{xy}}{s_x s_y}$$

where

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Correlation coefficient (r).

1. Range of values -1 to 1.
2. r = -1 $\rightarrow$ **linear**, negative relationship.
3. r = 1 $\rightarrow$ **linear**, positive relationship.
4. r = 0 $\rightarrow$ no **linear**, relationship.

Correlation coefficient does not describe the slope of the linear relationship.

Regression analysis - describing the best linear relationship between two variables.
The best fitting line is the line that gives the minimum distance between the points of data and the line.

Let's take b to be the slope of the line and a as the intercept value (the value where x=0). The first variable is x and its mean is $\bar{x}$. The second variable is y and its mean is $\bar{y}$. The formulas for b & a are:

$$b = r\left(\frac{s_y}{s_x}\right)$$

$$a = \bar{y} - b\bar{x}$$

Example- question 3.21

Country	GDP per capita (y)	Mental Health Expend/capita (x)
Finland	18,045.00	83.53
The Netherlands	14,449.80	71.38
Denmark	19,842.15	64.79
USA	17,670.15	41.21
Spain	7,563.70	9.17
Japan	19,763.63	54.98

Mental Health expenditures per capita - x variable

$\bar{x} = 54.18$

x	$(x - \bar{x})$	$(x - \bar{x})^2$
83.53	29.35	861.62
71.38	17.20	295.96
64.79	10.61	112.64
41.21	-12.97	168.13
9.17	-45.01	2025.60
54.98	0.80	0.65
Total	0	3464.60

$s_x^2 = 692.92$

$s_x = 26.32$

GDP per capita - y variable

$\bar{y} = 16222.41$

y	$(y - \bar{y})$	$(y - \bar{y})^2$
18045.00	1822.60	3321852.53
14449.80	-1772.61	3142128.49
19842.15	3619.75	13102553.87
17670.15	1447.75	2095965.59
7563.70	-8658.71	74973172.28
19763.63	3541.23	12540274.50
Total	0	109175947.20

$s_y^2 = 21835189.45$

$s_y = 4672.81$

Computing s_{xy}

$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x}) * (y - \bar{y})$
1822.60	29.35	53499.24
-1772.61	17.20	-30494.71
3619.75	10.61	38417.56
1447.75	-12.97	-18772.43
-8658.71	-45.01	389699.45
3541.23	0.80	2844.78
Total		435193.89

$s_{xy} = 87038.78$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{435193.89}{26.32 * 4672.81} = 0.7076$$

$$b = r\left(\frac{s_y}{s_x}\right) = 0.7076 * \frac{4672.81}{26.32} = 125.612$$

$$a = \bar{y} - b\bar{x} = 16222.41 - 125.612 * 54.18 = 9417.18$$