4.1 Scatterplots & Linear Correlation

Two variable statistics are methods used for detecting if there is a relationship between two variables *(e.g. the hotter the day, the more energy is used for air conditioning)*. Once a **cause and effect relationship** is determined, we can then develop mathematical models for these relationships for the purposes of prediction.

Scatter Plots – graphs to determine if there is a relationship blw 2 variables

independent → x-axis , dependent → y-axis

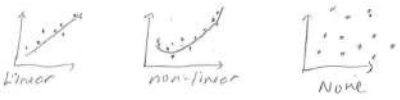
Line of Best Fit – A straight line drawn through data that:

1) passes through as many points as possible
2) Evenly distributed points above / below
3) ignores outliers, wherever possible

Outliers – Data that lies away from the majority. Can affect a regression analysis when data set is small

Correlation – when a change in the independent variable affects the dependent variable

1) Type

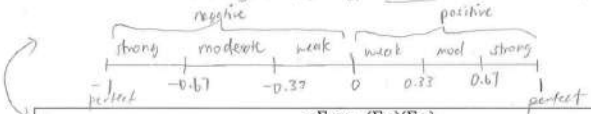


2) Direction

3) Strength

Linear Correlation – when the changes in one variable are proportional to the changes in the other

Correlation Coefficient (*r*) – gives a quantitative measure of the strength of a linear correlation.



$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Example 1 – This table shows data for the full time employees of a small company. Compute the correlation coefficient using the formula above.

Age (years)	*Annual Income ($000)*
33	33
25	31
19	18
44	52
50	56
54	60
38	44
29	35

$n = 8$

Age (x)	*Income (y)*	x^2	y^2	xy
33	33	1089	1089	1089
25	31	625	961	775
19	18	361	324	342
44	52	1936	2704	2288
50	56	2500	3136	2800
54	60	2916	3600	3240
38	44	1444	1936	1672
29	35	841	1225	1015
$\Sigma x = 292$	$\Sigma y = 329$	$\Sigma x^2 = 11,712$	$\Sigma y^2 = 14,975$	$\Sigma xy = 13,221$

50	56	2500	3136	2800
54	60	2916	3600	3240
38	44	1444	1936	1672
29	35	841	1225	1015
$\Sigma x = 292$	$\Sigma y = 329$	$\Sigma x^2 = 11,712$	$\Sigma y^2 = 14,975$	$\Sigma xy = 13,221$

$$r = \frac{(8)(13221) - (292)(329)}{\sqrt{[(8)(11712) - (292)^2][(8)(14975) - (329)^2]}}$$

$$= \frac{9700}{\sqrt{(8432)(11559)}} = 0.98 \leftarrow \text{strong, positive correlation}$$

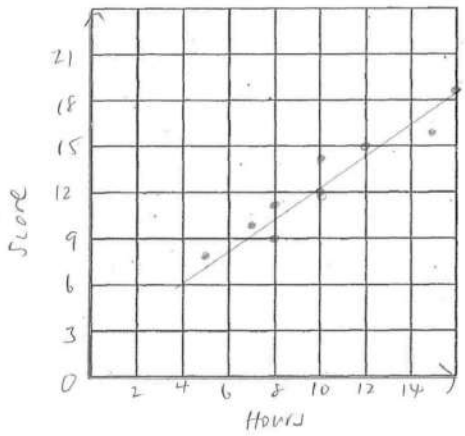
Example 2 – The data below shows scores from two different obedience training methods.

Rogers Method	
Hours x	Score y
10	12
15	16
7	10
12	15
8	9
5	8
8	11
16	19
10	14

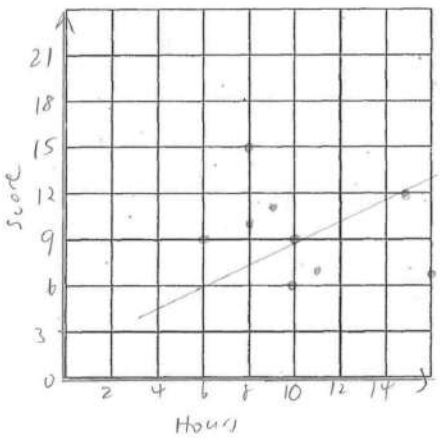
Laing System	
Hours x	Score y
8	10
6	9
15	12
16	7
9	11
11	7
10	9
10	6
8	15

a) Make scatter plots for the data above.

Rogers



Laing



b) Sketch a line of best fit for each graph.

c) What training method do you think is more effective? *Explain*.

The Rogers method is more effective b/c the correlation is stronger.