Unit – I

CORRELATION

Dr. S. DevaArul
Asst. Professor.

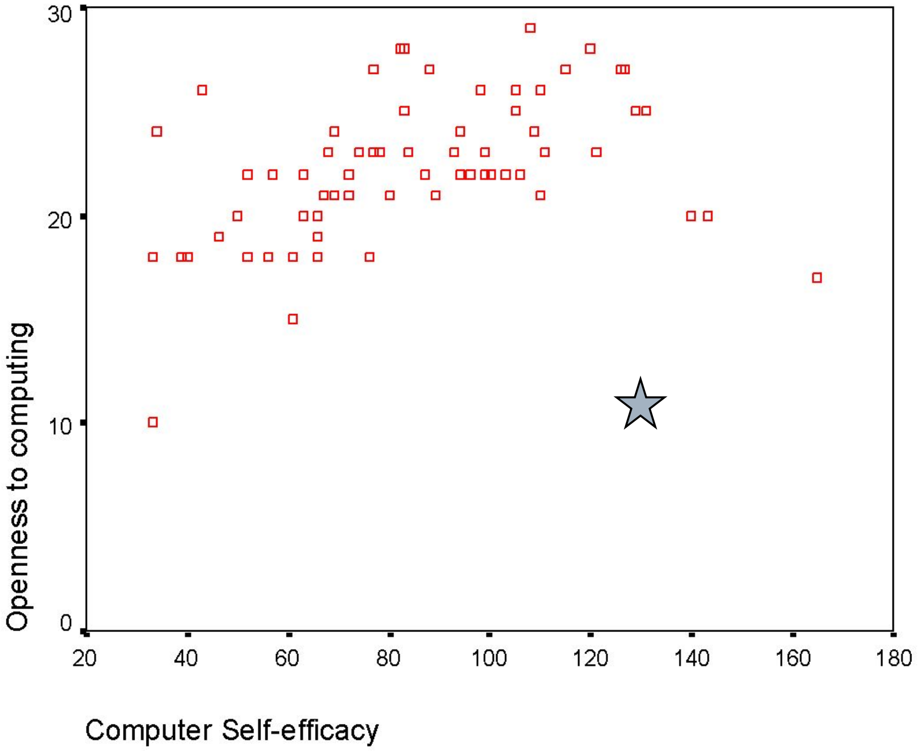
Introduction to Correlation

- The correlation coefficient measures the degree of relation between two variables
- Example, the relationship between daily consumption of fat calories and body weight
- Amount of rainfall and yield
- Price and Demand
- Sometimes both of the variables are treated as “dependent,” meaning that we haven’t ordered them causally.
- Sometimes one of the variables, X, is treated as independent and the other, Y, as dependent.
- The correlation coefficient, Pearson’s *r*, ranges between +1 and -1
- where +1 is a perfect positive relation
- -1 is a perfect negative relation
- A correlation near zero indicates that there is no relationship between scores on the two variables

Scatterplot:

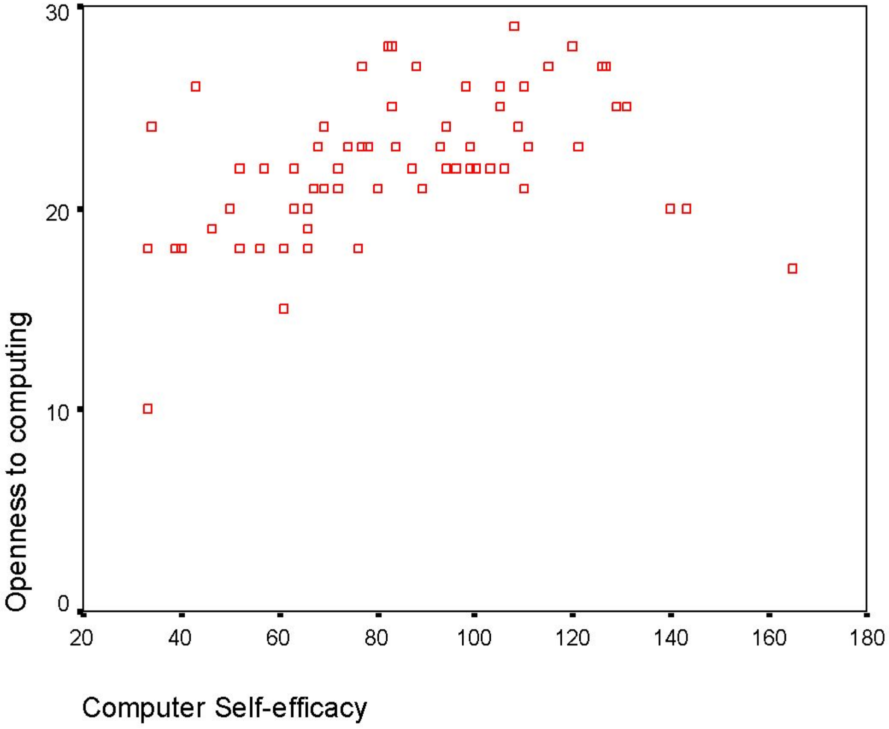
- Visual Representation of the Relationship Measured by the Correlation Coefficient
- The scatterplot is a diagram which plots off cases for which two variables
- For example Price and Demand
- Smoking Habit and Lung disease
- In a scatterplot, one of the variables (usually the independent variable) is plotted along the horizontal or X axis and the other is plotted along the vertical or Y axis
- Each point in a scatterplot corresponds to the scores (X,Y)
- The strength of the linear relationship between X and Y is stronger as the swarm of points in the scatterplot more closely approximates a diagonal “line” across the graph

An Example of a Scatterplot



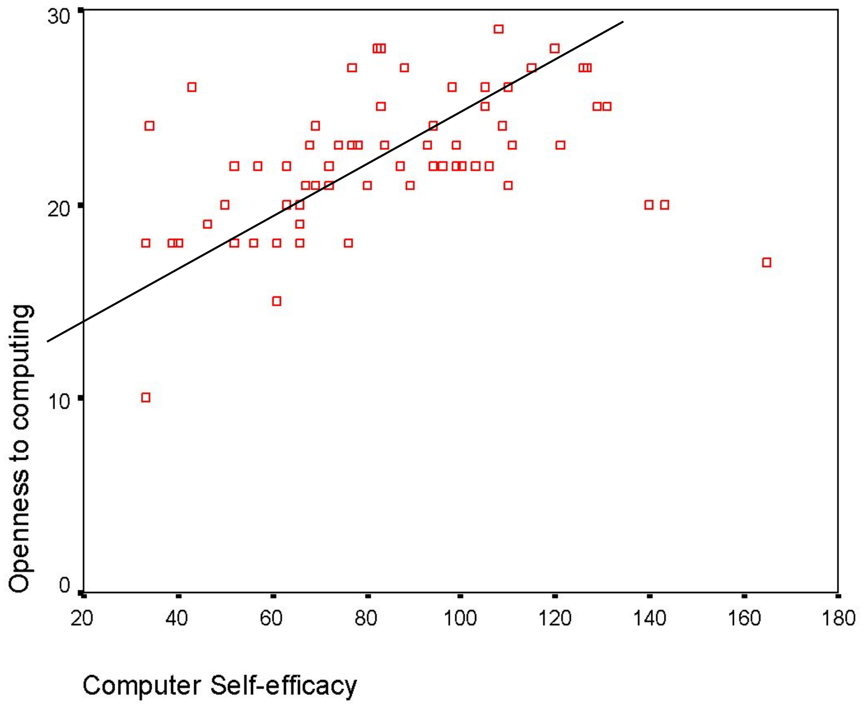
Scatterplot Allows You to Visualize the Relationship between Variables

The purpose of the scatterplot is to visualize the relationship between the two variables represented by the horizontal and vertical axes. Note that although the relationship is not perfect, there is a tendency for higher values of openness to computing to be associated with larger values of computer self-efficacy, suggesting that as openness increases, self-efficacy increases. This indicates that there is a *positive* correlation

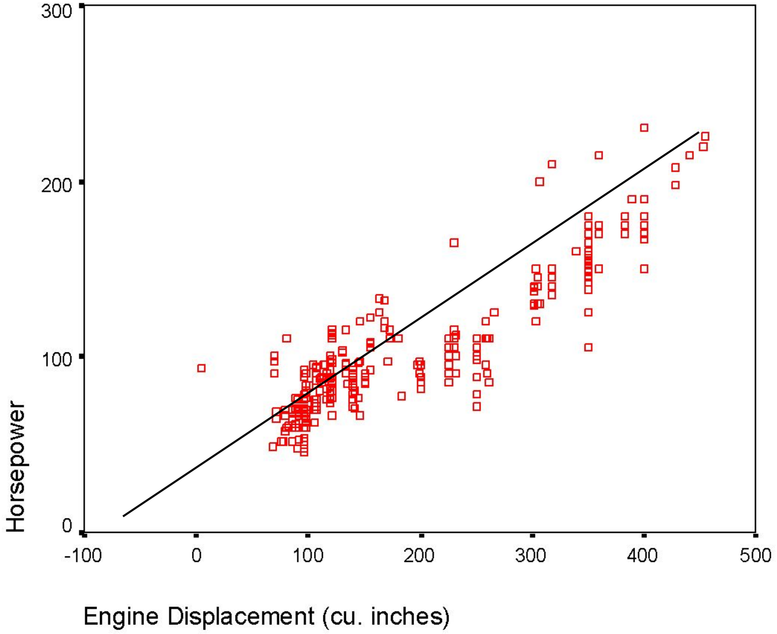


Drawing A Possible Regression Line

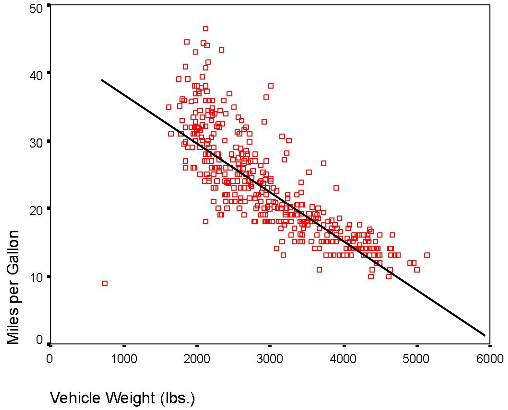
Let’s draw a line through the swarm of points that best “fits” the data set (minimizes the distance between the line and each of the points). This is imposing a *linear* description of the relationship between the two variables, when sometimes you might want to find out if a line that represented a *curvilinear* relationship (in this case an inverted U) was a better fit, but we’ll leave that question for another time. The line that represents this relationship best mathematically is called a “regression line” and the point at which the mathematically best fitting line crosses the *y* axis is called the “intercept”



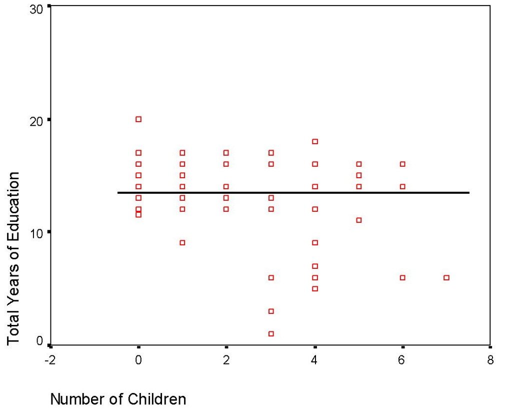
Various Types of Associations



Positive Relationship between X and Y



Strong negative Relationship between X and Y; points tightly clustered around line; nonlinear trend at lower weights



Essentially no relationship between X and Y; points loosely clustered around line

How is the Correlation Coefficient Computed?

- The conceptual formula for the correlation coefficient is a little daunting, but it looks like this:

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{[\Sigma (X - \bar{X})^2]\ [\Sigma (Y - \bar{Y})^2]}}$$

Where X is a person's or case's score on the independent variable, Y is a person's or case's score on the dependent variable, and X-bar and Y-bar are the means of the scores on the independent and dependent variables, respectively. The quantity in the numerator is called the sum of the crossproducts (SP). The quantity in the denominator is the square root of the product of the sum of squares for both variables (SS_x and SS_y)

Computing Formula for Pearson's *r*

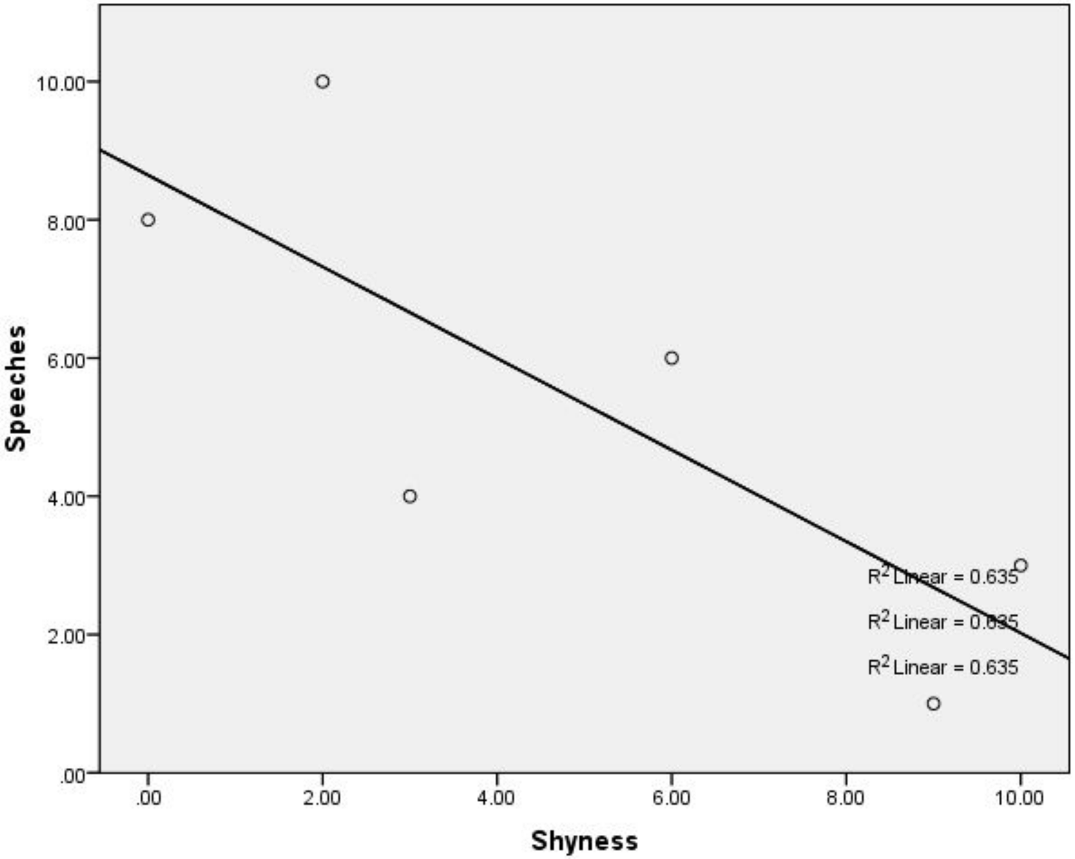
- □ The conceptual formula for Pearson's *r* is rarely used to compute it. You will find a nice illustration here of a computing formula and a brief example

Here is another computing formula

$$r = \frac{N \sum XY - \sum X \sum Y}{\sqrt{[\, N \sum X^2 - (\sum X)^2] \, [N \sum Y^2 - (\sum Y)^2]}}$$

We will do an example using this computing formula next, so let's download the correlation.sav data set

ScatterPlot of Shyness and Speeches



A negative relationship: The more shy you are (the farther you are along the X axis), the fewer speeches you give (the lower you are on the Y axis)

Computational Example of Cor*relation*

$$r = \frac{N\,\Sigma XY - \Sigma X\,\Sigma Y}{\sqrt{[\,N\,\Sigma X^2 - (\Sigma X)^2]\,[N\,\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$\frac{(6 \text{ X } 107) - 30\ (32)}{\sqrt{[6\ (230) - 30^2]\ [6\ (226) - 32^2\ ]}}$$

Shyness **X**	**Speeches** **Y**	XY	X^2	Y^2
0	8	0	0	64
2	10	20	4	100
3	4	12	9	16
6	6	36	36	36
9	1	9	81	1
10	3	30	100	9
30	32	107	230	226

r = -.797 (note crossproducts term in the numerator is negative) and R-square = .635