Association

An association (dependence) exists between two variables if a particular value/category for one variable is more likely to occur with certain values/categories of the other variable.

	OS		
Gen	**Yes**	**No**	**Tot.**
M	600	0	600
F	0	40	40
Tot.	600	40	640

Response and Explanatory variables:

- Response variable: The dependent variable, it is the outcome variable on which we are making comparisons.
- Explanatory variable: The indipendent variable, we it compared with respect to the values/categories on the response variable

Example: Response/Explanatory

- Blood alcohol level/# of beers consumed
- Grade on test/Amount of study time

Gen	OS Yes	OS No	Tot.
M	500	100	600
F	10	30	40
Tot.	510	130	640

Condingency table

Useful for looking at the associations between two categorical variables.

Gen	OS Yes	OS No	Tot.
M	5	1	6
F	1	3	4
Tot.	6	4	10

- It displays two categorical variables
- The rows list the categories of one variable
- The columns list the categories of the other variable
- Entries in the table are frequencies

The original table was:

Code	Gen	OS
1	M	YES
2	F	YES
3	M	NO
4	F	NO
5	M	YES
6	M	YES
7	M	YES
8	F	NO
9	M	YES
10	F	NO

Conditional proportions or percentages

Let's have an example and assume that:

- OS is the response.
- Gen is the explanatory variable.
- We watch the distribution of OS change as Gen changes.

	OS		
Gen	Yes	No	Tot.
M	500	100	600
	0.83	0.17	1
F	10	30	40
	0.25	0.75	1
Tot.	510	130	640
	0.80	0.20	1

Looking at this table:

- 0.83 = proportion of YES under the condition Gen=M (conditional prop.)
- 0.25 = proportion of YES under the condition Gen=F (conditional prop.)
- 0.80 = proportion of YES (marginal proportion)

We can see that men are more likely to say YES.

If there is no association between OS and Gen, then the conditional proportions for the response variable categories (OS) would be the same for each gender, like this:

Gen	OS Yes	OS No	Tot.
M	500 0.83	100 0.17	600 1
F	50 0.83	10 0.17	60 1
Tot.	550 0.83	110 0.17	660 1

In this case the two variables are said to be indipendent.

Another example:

Food Type	Pesticide Status: Present	Not Present	Total
Organic	29	98	127
Conventional	19,485	7,086	26,571
Total	**19,514**	**7,184**	**26,698**

The response is Pesticide Status

Food Type	Pesticide Status: Present	Not Present	Total	*n*
Organic	0.23	0.77	**1.00**	127
Conventional	0.73	0.27	**1.00**	26,571

There is a **dependence**.

And another one:

Example 1

Gen	OS: Yes	No	Tot.
M	500	100	600
	0.83	0.17	1
F	70	30	100
	0.70	0.30	1
Tot.	570	130	700
	0.81	0.19	1

- The categories in correspondence are **YES-M** and **NO-F**.
- This means that **YES** is more likely to "happen" under the "condition" **M** rather than F and **NO** is more likely to "happen" under the "condition" **F** rather than M.

And another one:

Example 2

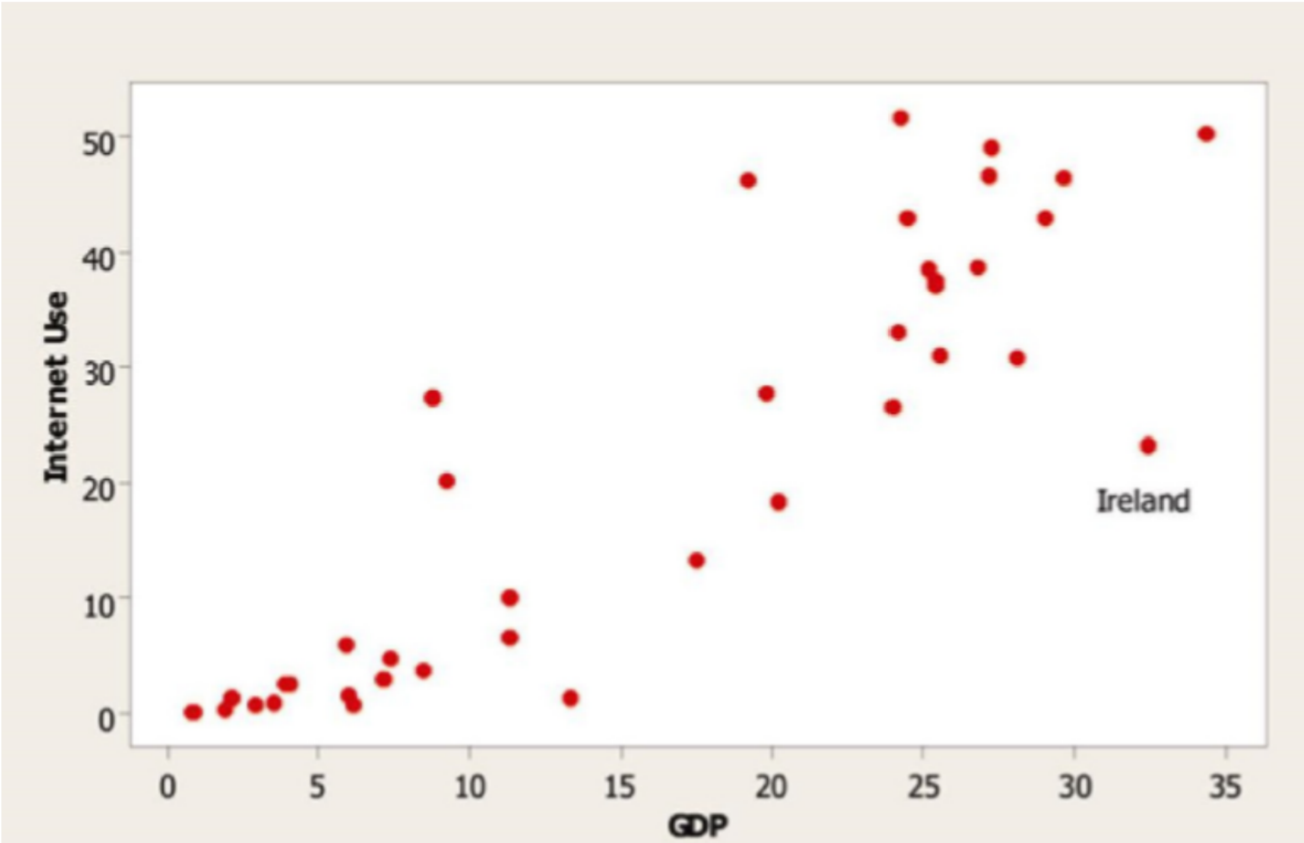
Region	OS Yes	OS No	Tot.
North	540 90%	60 10%	600 100 %
Center	320 80%	80 20%	400 100%
South	210 70%	90 30%	300 100%
Tot.	1070 82%	230 18%	1300 100%

- The categories in correspondence are: **YES-North**, **NO-Center** and **NO-South**.

> **Hint**
>
> In these examples, for each category of the response we find under which category of the explanatory variable its percentage is greater than the corresponding marginal.

Association of quantitative variables, Scatterplot

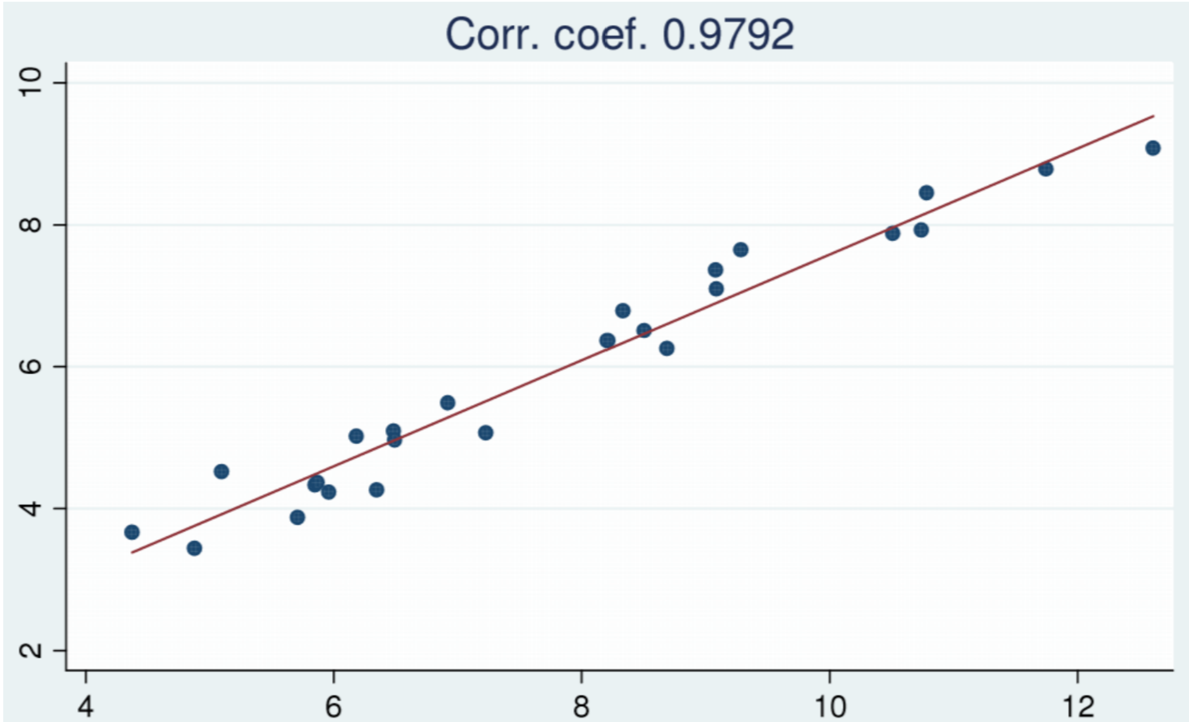
- Horizontal Axis: Explanatory variable, x.
- Vertical Axis: Response variable, y.



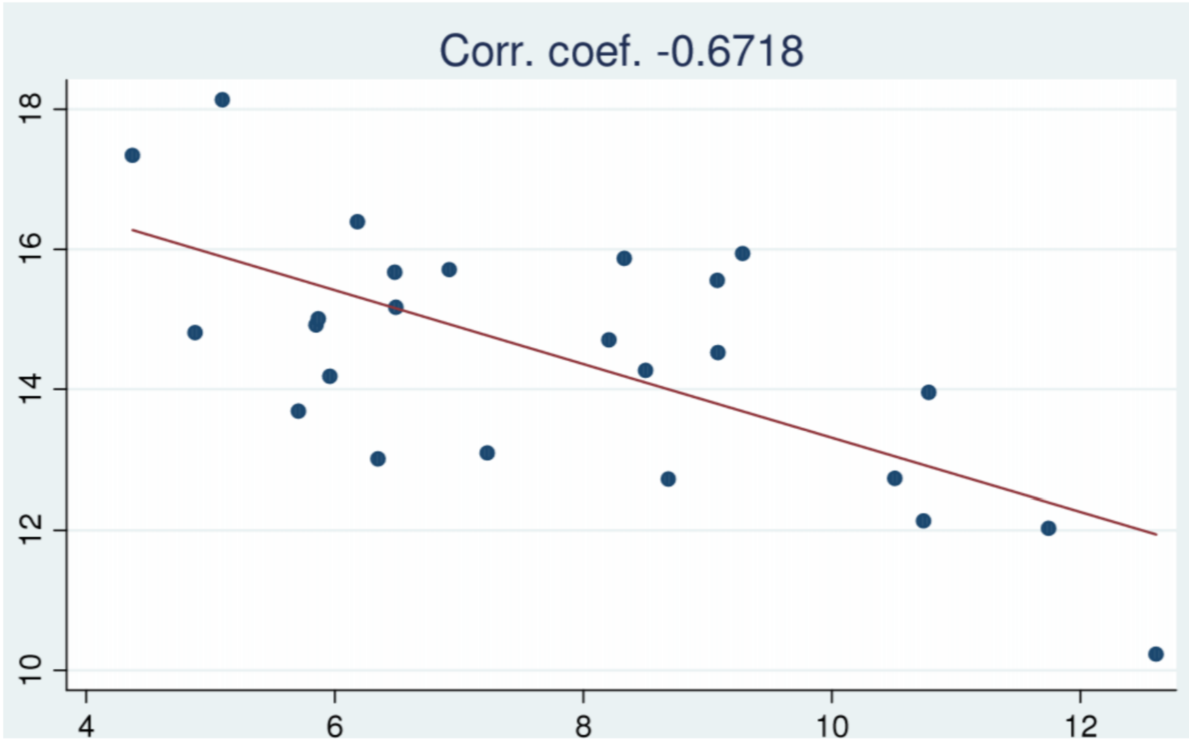
Interpreting the scatterplot:

The variables are:

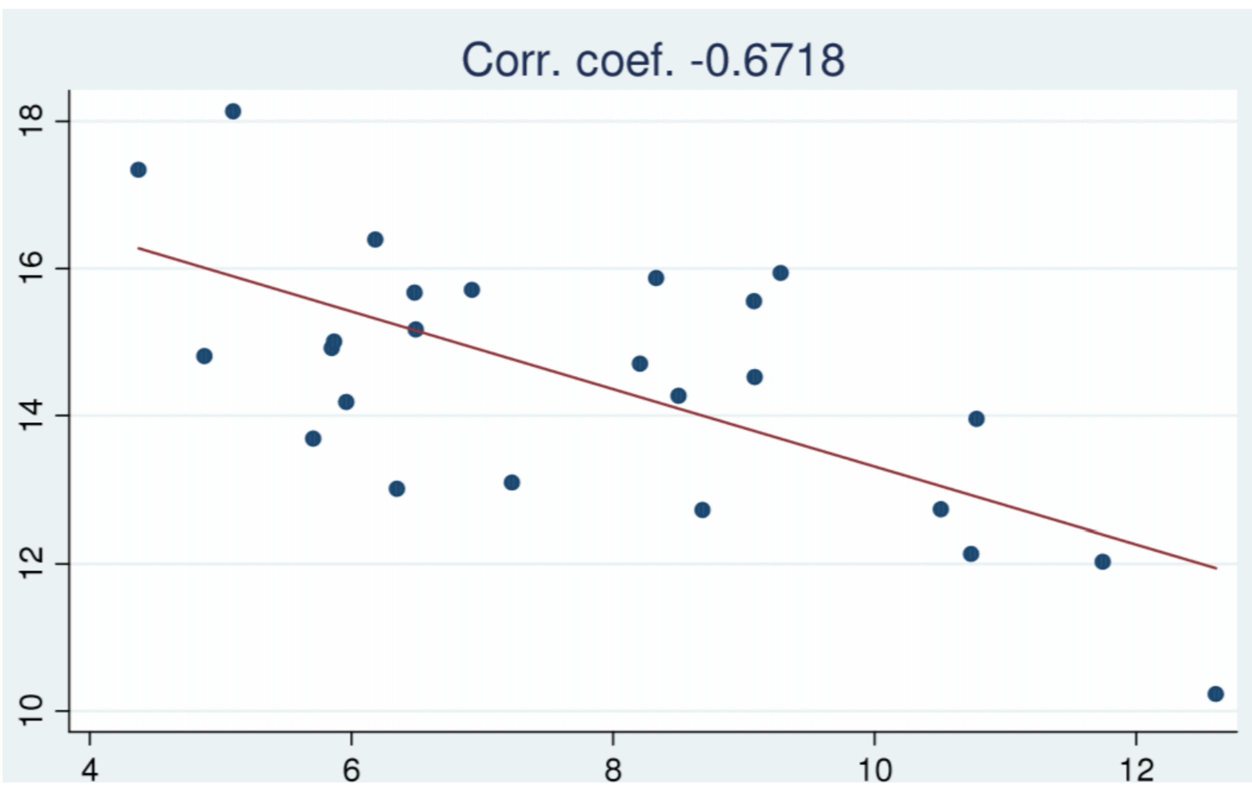
- **Positively associated** when
  - High values of x tend to occur with high values of y
  - Low values of x tend to occur with low values of y



- **Negatively associated** when
  - high values of one variable tend to pair with low values of the other variable
  - High values of x tend to occur with low values of y
  - Low values of x tend to occur with high values of y



- Not associated:



The strength of the association can be measured through the correlation coefficient.