Ch. 10: Two Quantitative Variables

Navigation

By Section

10.1: Two Quantitative Variables

Scatterplots and Correlation

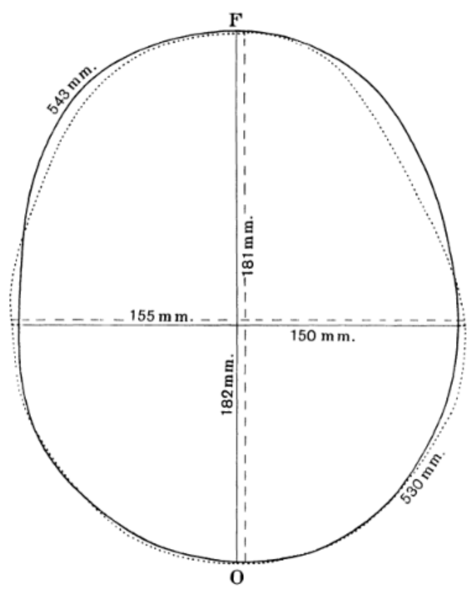
Graphical Summaries of Quantitative Variables

R.J. Gladstone (1905). "A Study of the Relations of the Brain to the Size of the Head", Biometrika, Vol. 4, p 105-123.

Data collected during 237 autopsies at Middlesex Hospital in London, excluding cases "in which the brain showed a distinctly pathological condition which would have obviously affected its weight"

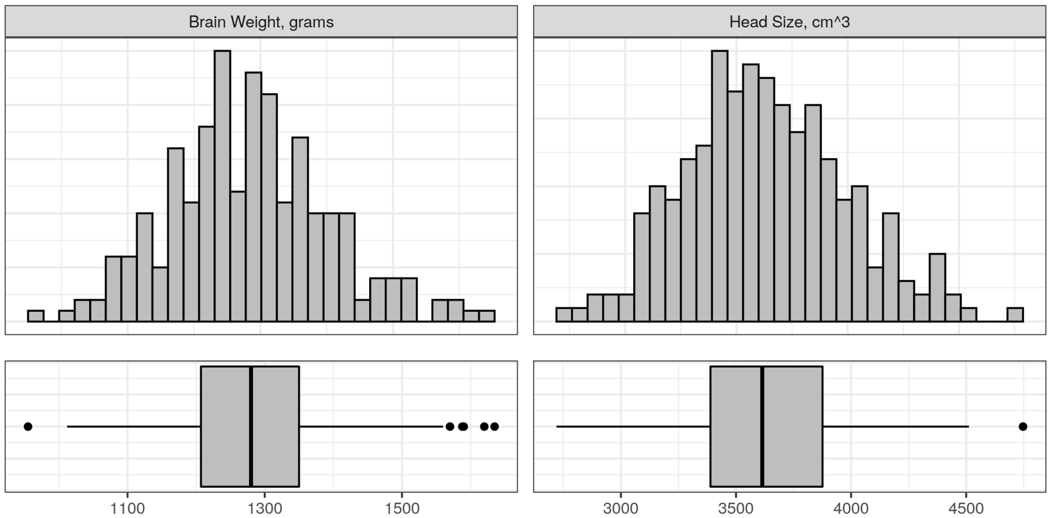
Variables:

- Gender
- Age (20 - 45 or 46+)
- Brain Weight (g)
- Head Size (cubic cm) the smallest rectangular block which could contain the head



Graphical Summaries of Quantitative Variables

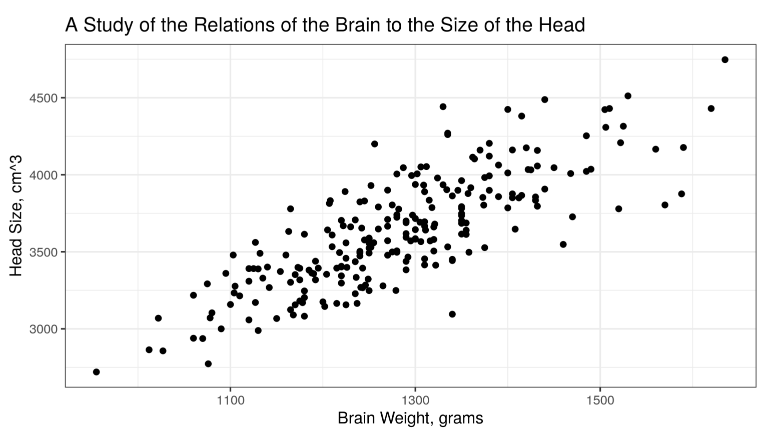
A single quantitative variable can be summarized visually using a histogram or a bar chart:



Graphical Summaries of Quantitative Variables

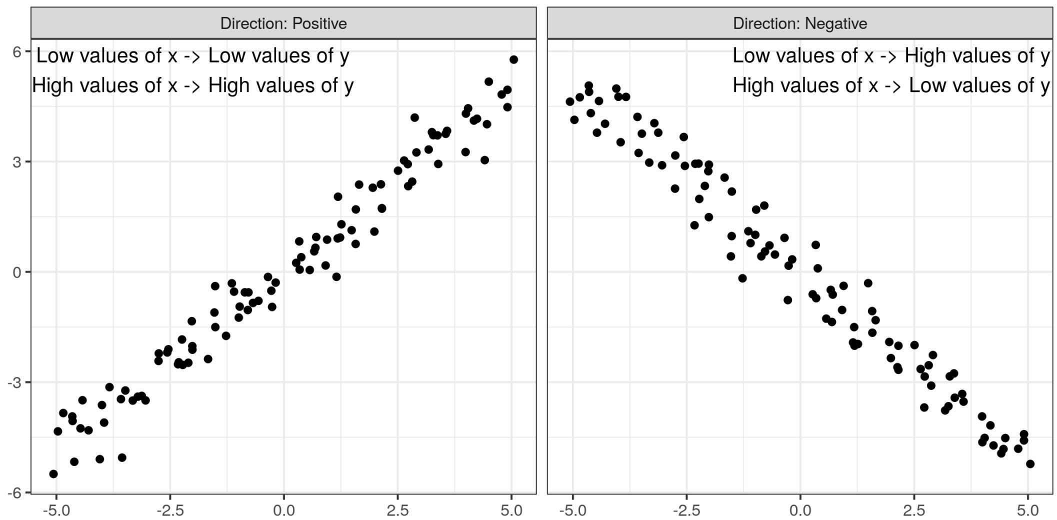
But, summarizing each variable separately doesn't tell us how the two variables might be related.

Is there a relationship between brain weight and head size? How do you know?



A **scatterplot** is a plot with the explanatory variable on the x-axis, and the response variable on the y-axis. Observations are shown as points corresponding to a set of quantitative measurements.

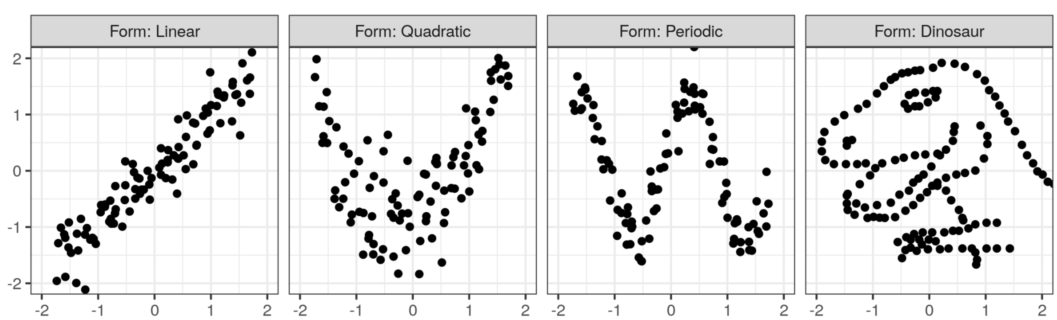
Describing Variable Relationships: Direction



Positive slope: as x increases, y increases too.

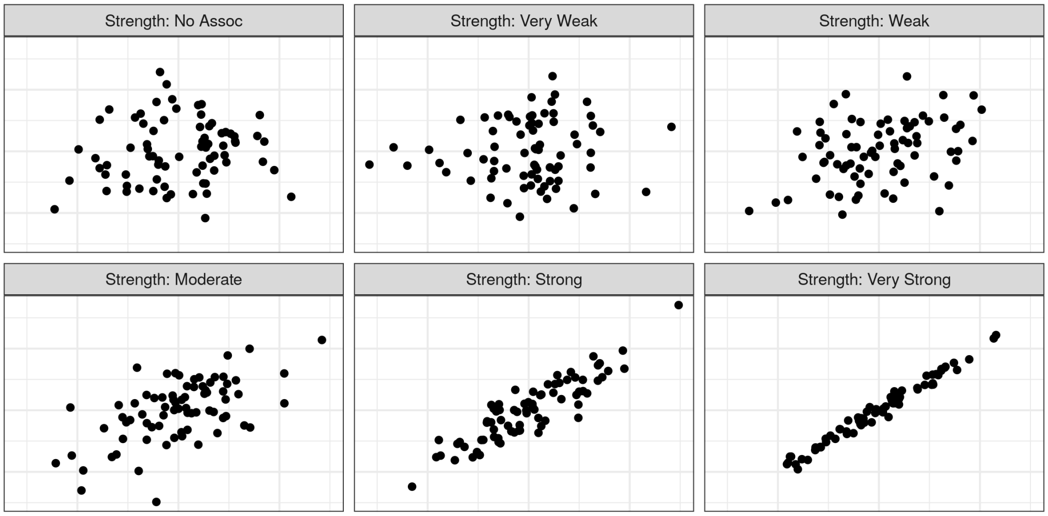
Describing Variable Relationships: Form

The **form** of an association is whether it follows a linear pattern, or some sort of more complicated pattern - periodic, polynomial (quadratic, cubic, etc.)



Describing Variable Relationships: Strength

The **strength** of an association indicates how well the value of one variable can be predicted if you know the value of the other variable.

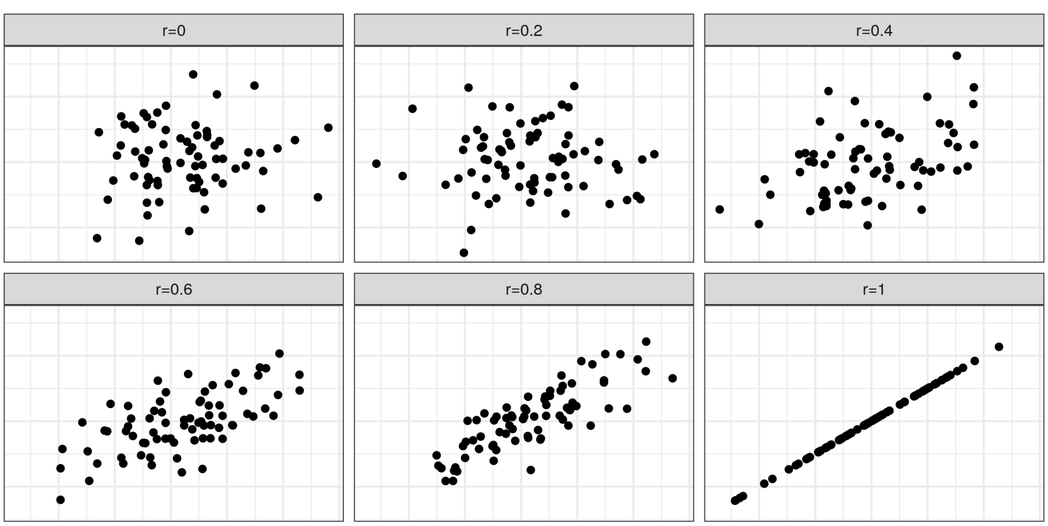


We can describe the strength and direction of a *linear* relationship using the **correlation coefficient**

Correlation Coefficient

The **correlation coefficient**, r, is always between -1 and 1.

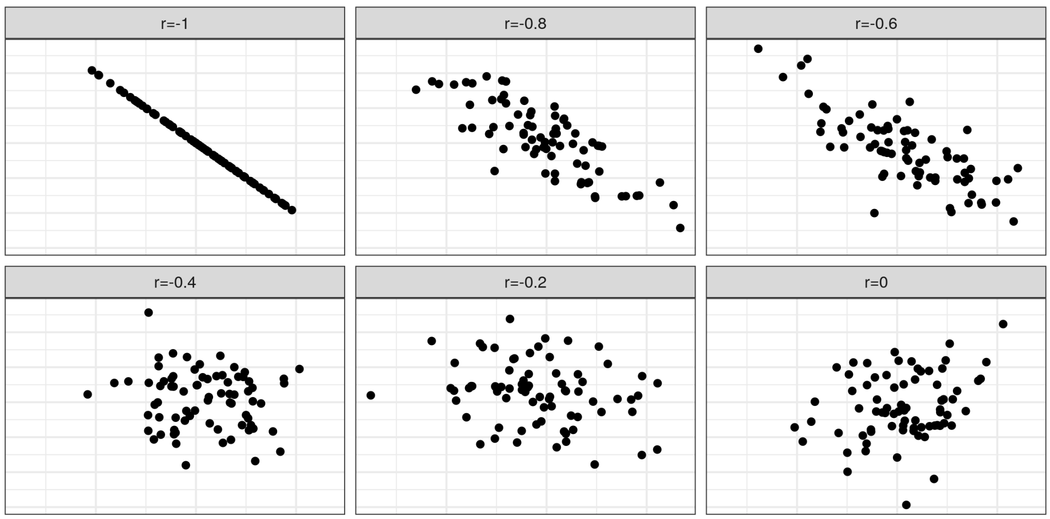
-1	-0.5	0	0.5	1
perfect negative correlation	moderate negative correlation	no correlation	moderate positive correlation	perfect positive correlation



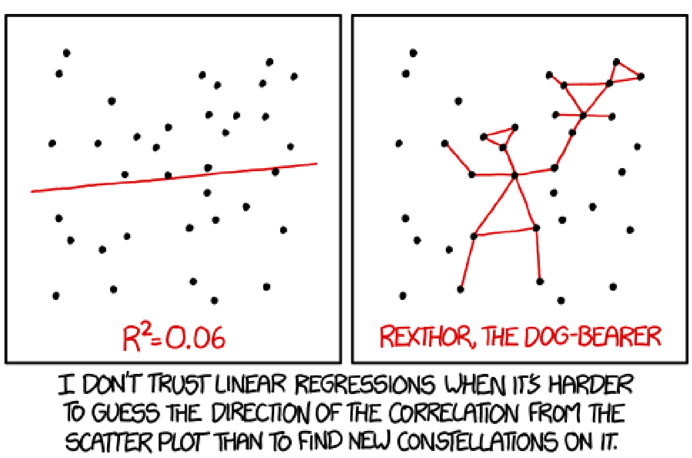
Correlation Coefficient

The **correlation coefficient**, r, is always between -1 and 1.

-1	-0.5	0	0.5	1
perfect negative correlation	moderate negative correlation	no correlation	moderate positive correlation	perfect positive correlation

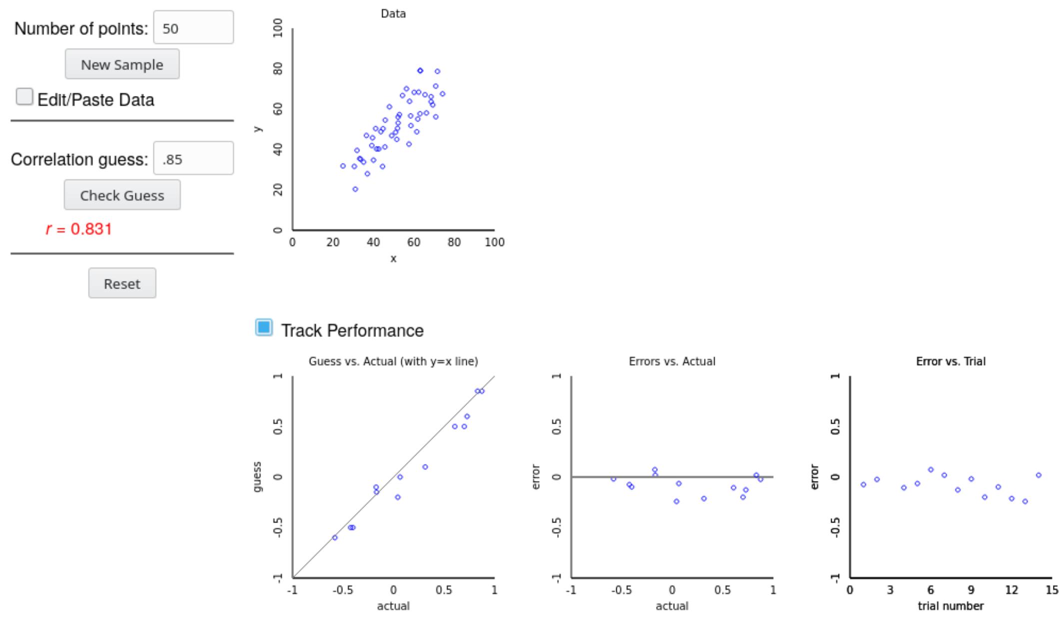


Correlation Coefficient



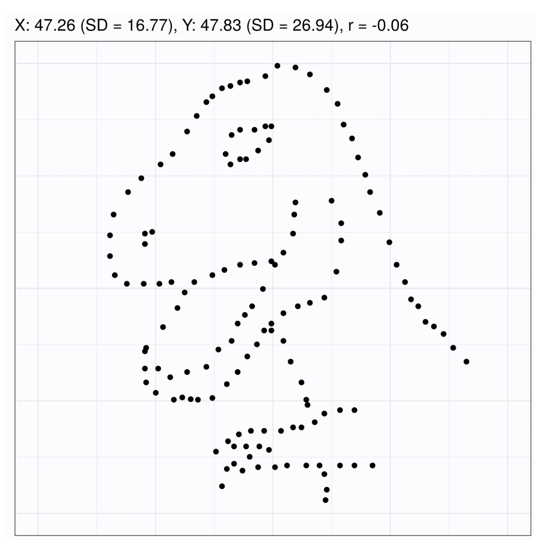
Correlation Coefficient

Get a feel for it by playing the correlation guessing game!



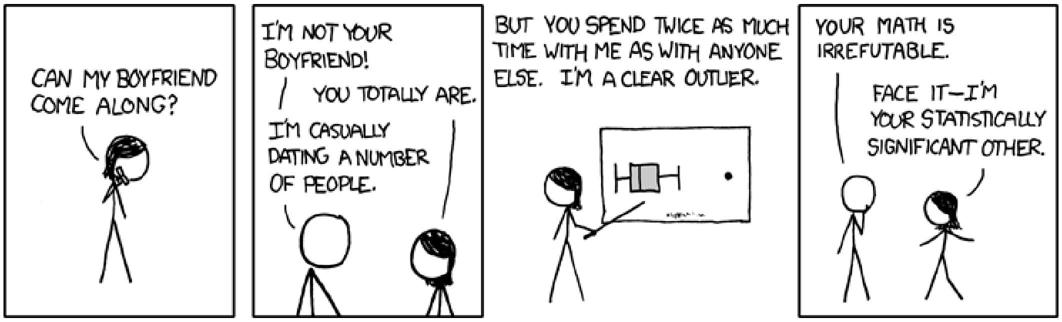
Correlation Coefficient

The correlation coefficient is only useful for showing the strength of linear relationships.



All of these plots have essentially the same correlation coefficient, but in some cases there are very clear associations between x and y

Outliers and Influential Observations

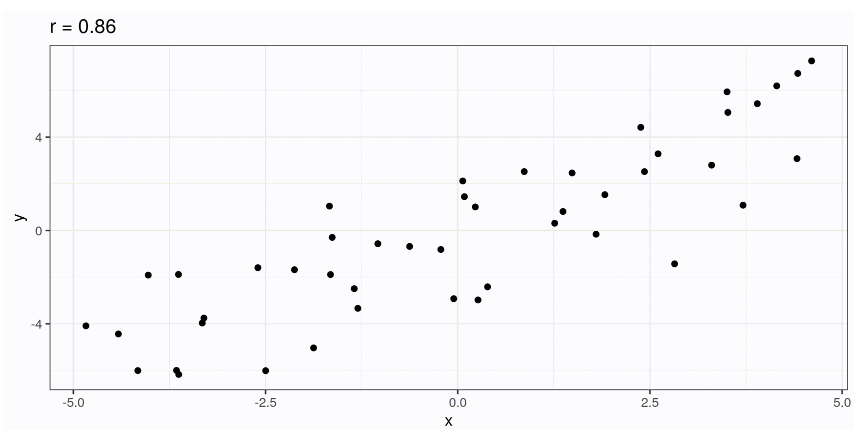


With one variable, outliers are fairly easy to spot

When there are two variables, we don't just have to worry about outliers in one dimension; we also have to worry about **influential observations**

Outliers and Influential Observations

Influential observations are observations which, if included, change our understanding of the relationship between two variables.



Outliers and Influential Observations

Influential observations are observations which, if included, change our understanding of the relationship between two variables.