

Correlation

Ben Babcock
University of Minnesota

The Basic Idea

Suppose that we have two variables that seem to have some sort of relationship and we want an index (number) to tell us how they relate to one another.

If we stick to a linear relationship, the Pearson correlation coefficient (r) is a good index of association.

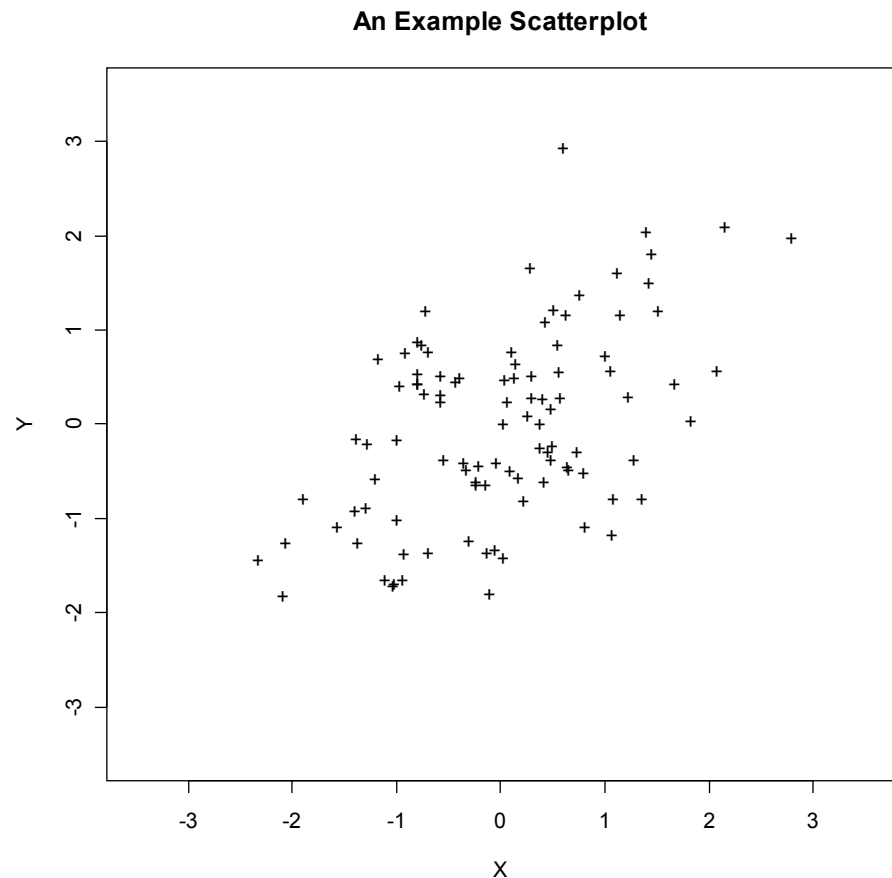
What do I mean by “linear relationship”?

A Linear Relationship: Graphically

Suppose that you have two variables, X and Y . You can see the relationship between the variables in a scatterplot.

Each “+” represents one person's score on a pair of variables.

Each axis, x and y , represents a variable.



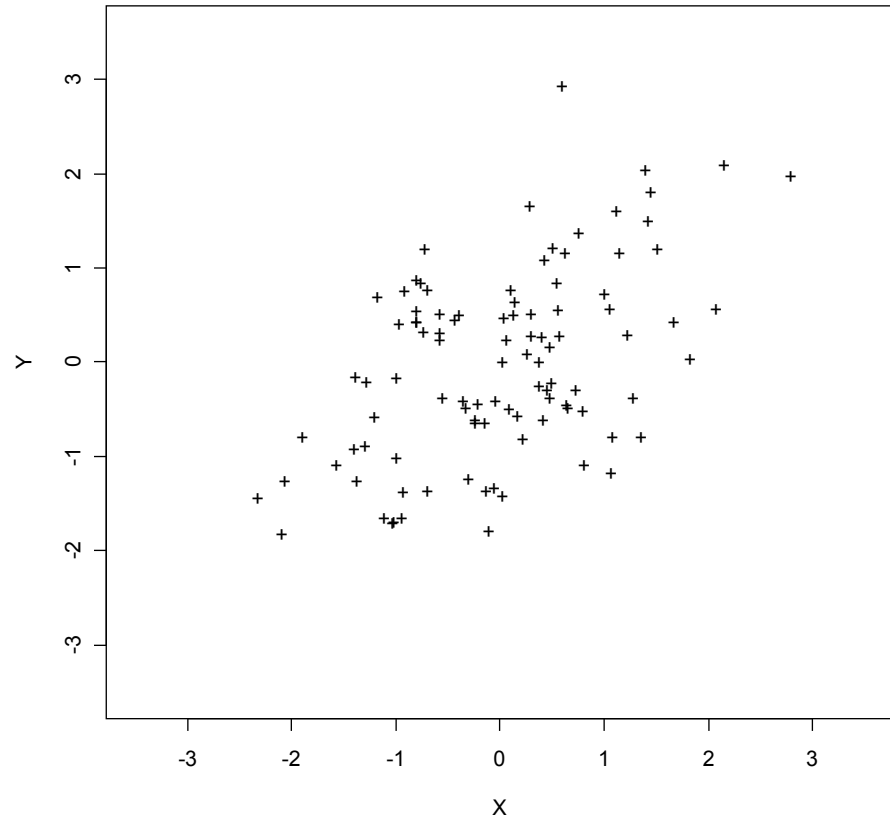
A Linear Relationship: Graphically

Notice that scores on X and Y tend to increase together.

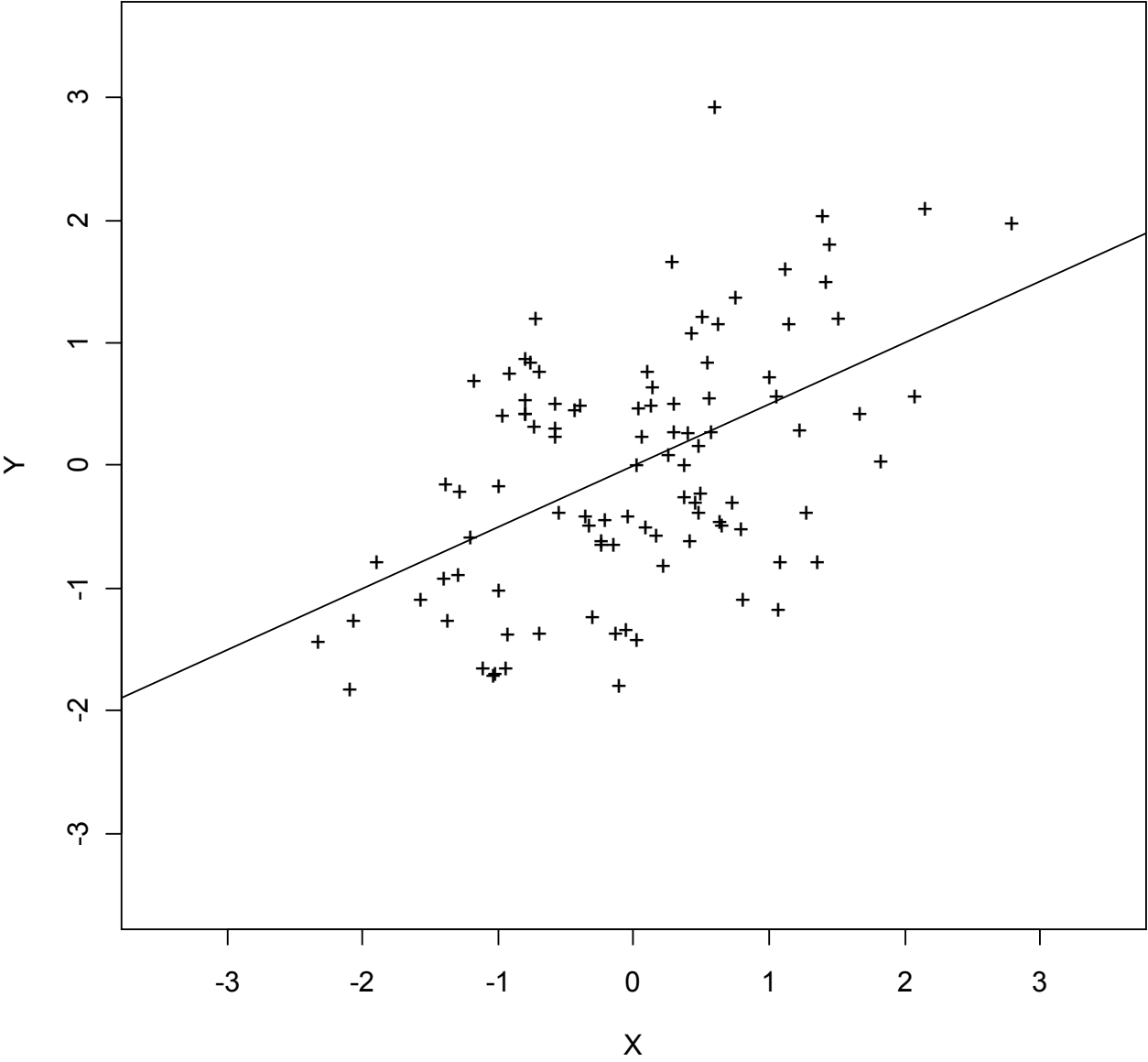
In fact, we could draw a line through the points to describe one type of relationship.

If there is a lot of scatter (variation; go crazy) around the line, the relationship is _____ . If the points cluster tightly around the line, the linear relationship will be _____ .

An Example Scatterplot



An Example Scatterplot



A Measure of Association

Let's think of a possible index to measure the relationship between two variables. This relationship should measure how much one variable CO VARRIES (!) with another variable.

Variance:

$$s_x^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N-1} = \frac{\sum_{i=1}^N (X_i - \bar{X})(X_i - \bar{X})}{N-1}$$

How could we work in another variable into the mix?

Covariance

Let's replace one of the X variables with Y .

This would be the *covariance* (cov_{XY}). H, 181

$$cov_{XY} = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{N - 1}$$

where X and Y are variables and N is sample size.

This is just a spin off of the _____. How does this work?

What is the unit of measure?

What are the bounds (how big or small can it be) of this index?

The Pearson Correlation Coefficient

Covariance units don't make sense. Let's make something that is unit free. This is the Pearson Product Moment Correlation (r_{XY}).

$$r_{XY} = \frac{COV_{XY}}{s_X s_Y}$$

where COV_{XY} is the covariance between variables X and Y and s_X and s_Y are the standard deviations of variables X and Y .

By dividing by the standard deviations of X and Y , we take out the unit of measurement. H, 182. z-scores are also unit free.

We also have a measure of linear relation between two variables that ranges from _____ to _____. Why is _____ the highest that r_{XY} can go?

The Pearson Correlation Coefficient

Values close to indicate a weak relationship. Values close to or indicate a strong relationship. Positive values indicate that Y increases as X increases. Negative values indicate that Y decreases as X increases.

Squaring the Pearson correlation (r^2) gives how much _____ in one variable can be linearly predicted using the other variable.

So if two variables correlate 0.50, ...

So if two variables correlate -0.30, ...

Pearson Correlation: Alternative Equations

Another way to express the correlation is as the average (adjusted) of the product of the z-scores:

$$r_{XY} = \frac{\sum_{i=1}^N X_{zi} Y_{zi}}{N-1}$$

where X_z and Y_z are z-scores for X and Y and N is sample size. I love this equation. Why does this work?

There is also a computational equation in Howell, 182. I hate that equation.

Pearson Correlation: Big and Little

Cohen (year) set out some guidelines for correlation size in the social sciences.

Small: $r =$ _____

Medium: $r =$ _____

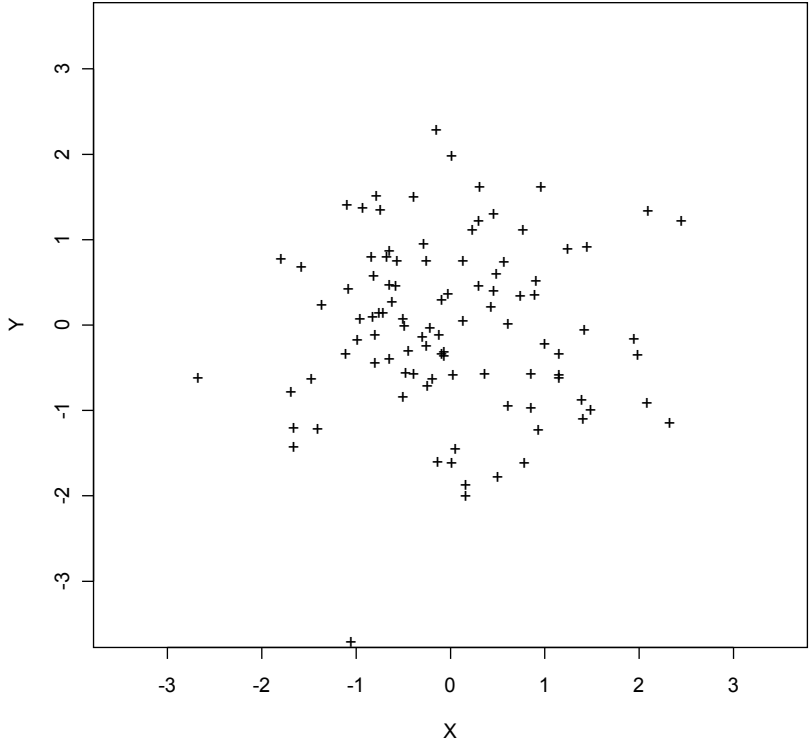
Large: $r =$ _____

However, the “typical” correlation varies widely with your subject area. Examples:

What's big?

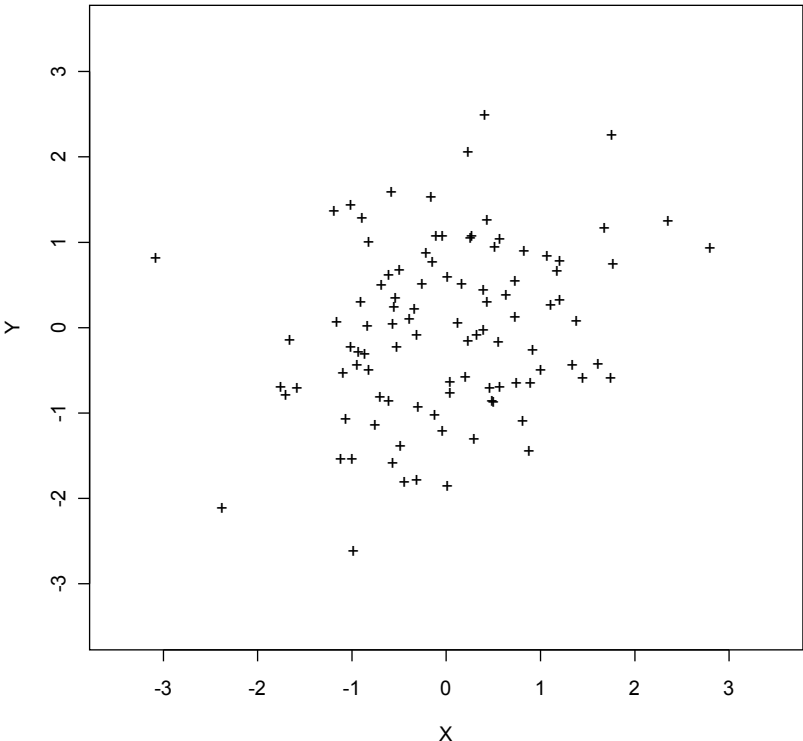
Pearson Correlation: What Does it Look Like?

Plot of 100 Data Points, $r=0.0$



$$r = 0.0$$

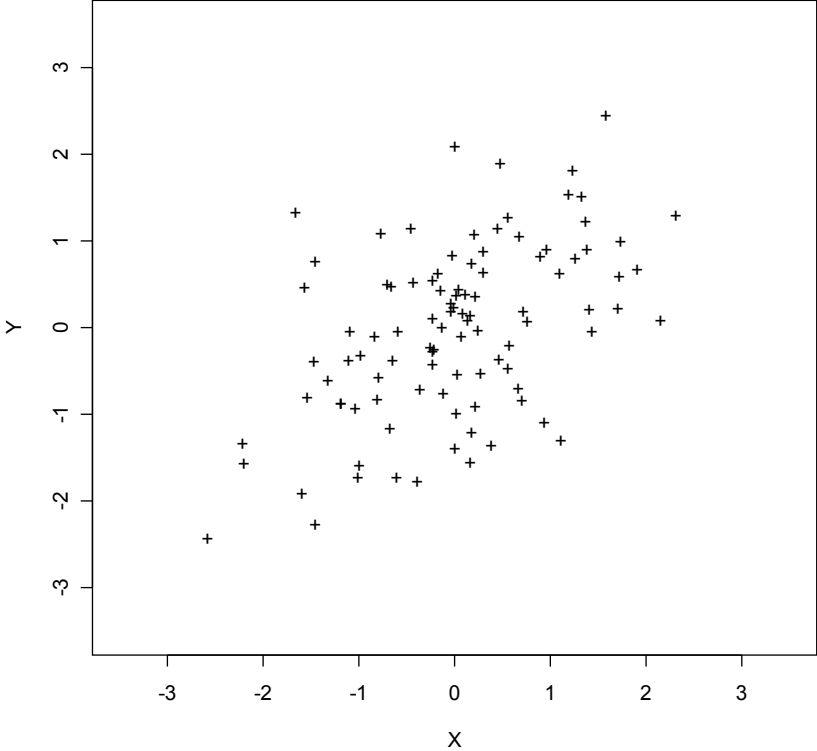
Plot of 100 Data Points, $r=.25$



$$r = .25$$

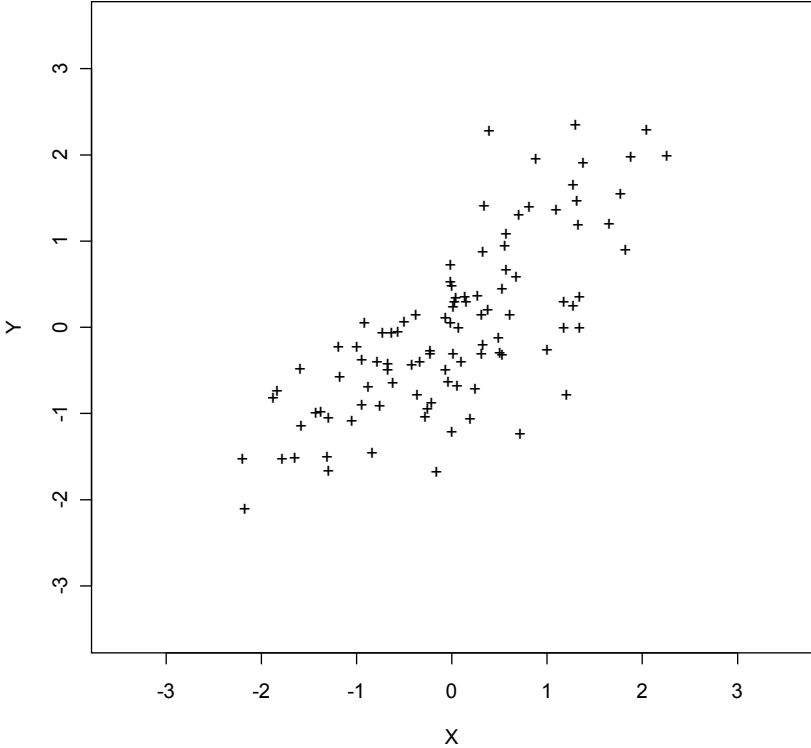
Pearson Correlation: What Does it Look Like?

Plot of 100 Data Points, $r=.50$



$$r = .50$$

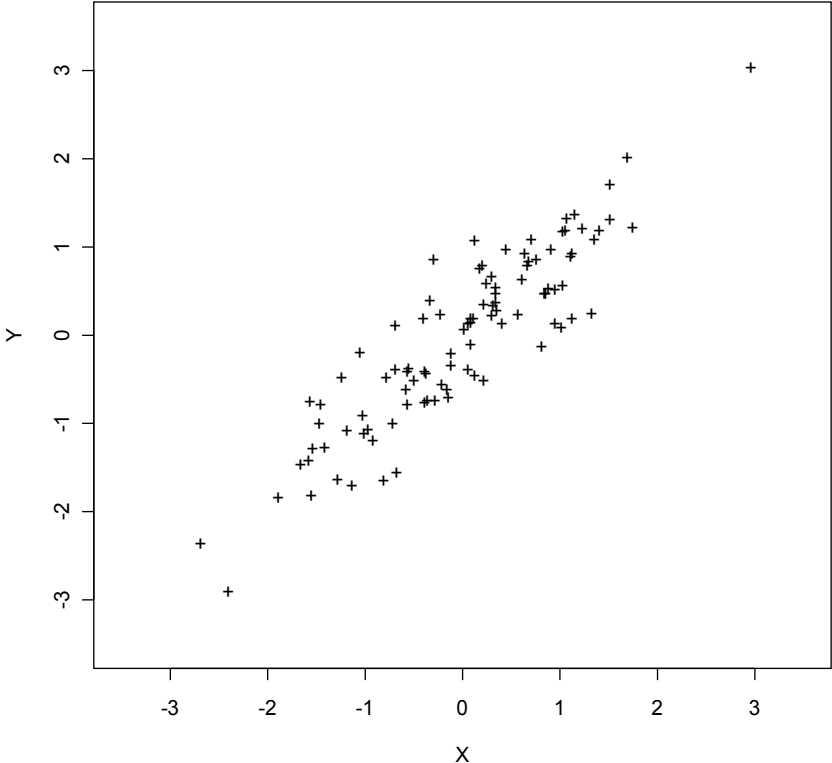
Plot of 100 Data Points, $r=.75$



$$r = .75$$

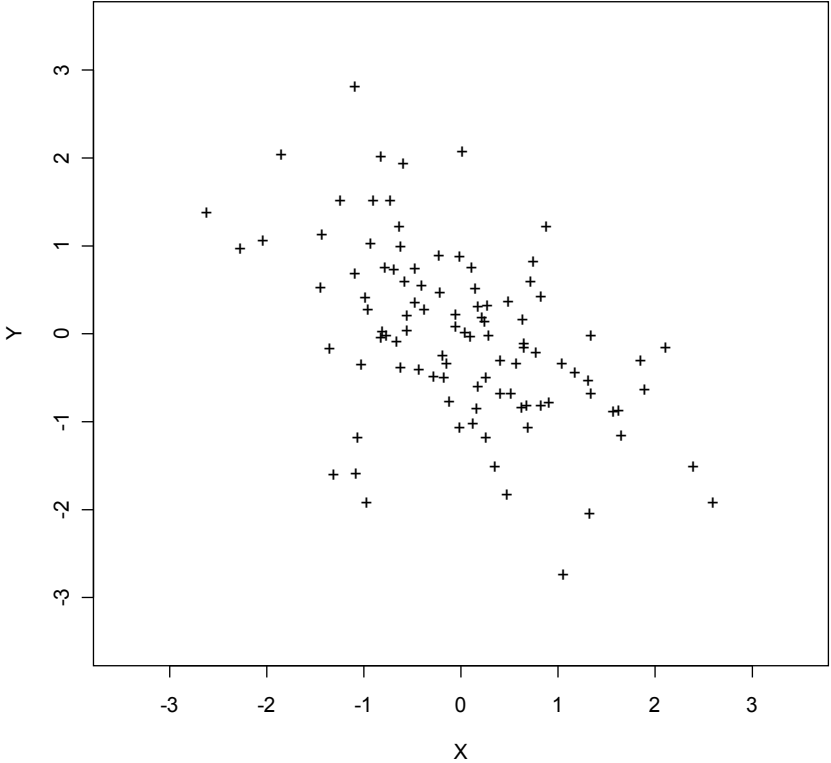
Pearson Correlation: What Does it Look Like?

Plot of 100 Data Points, $r=.90$



$$r = .90$$

Plot of 100 Data Points, $r=-.50$



$$r = -.50$$

Pearson Correlation: Problem Children

Range Restriction

Heterogenous Subsamples

Outliers

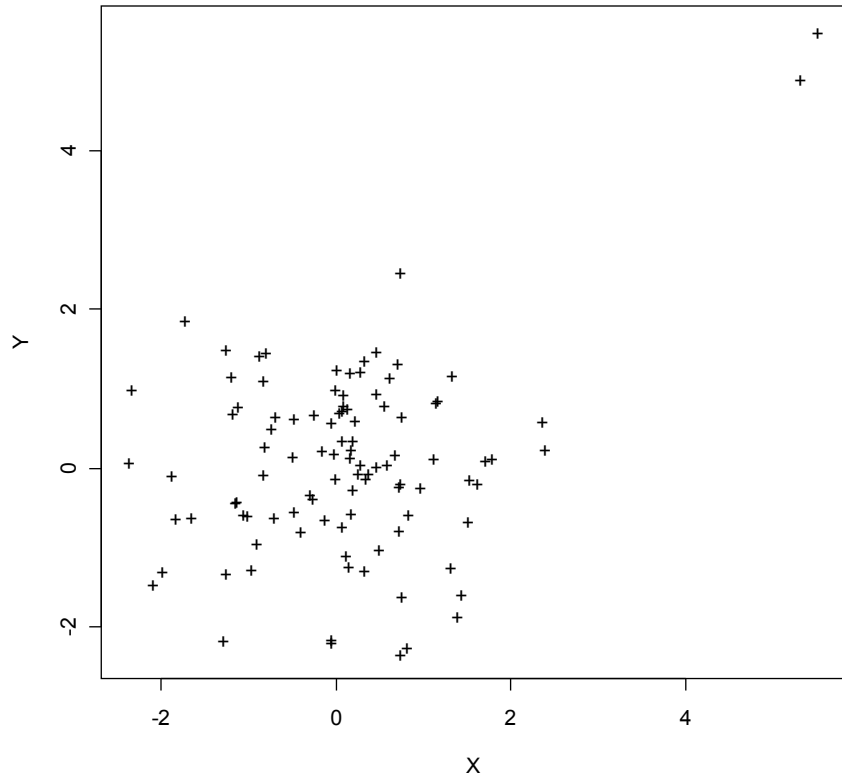
Non-linear Effects

These phenomena can make the correlation coefficient an unrepresentative index of association. How?

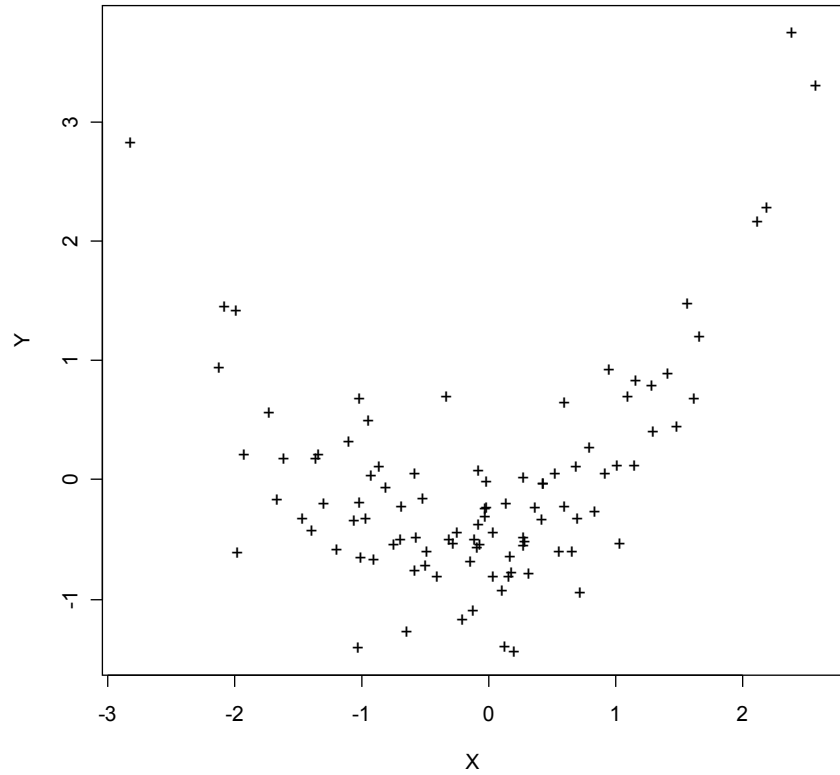
The graphs on the next pages correspond to these problems. There may be more than one problem with any one graph.

Pearson Correlation: Problem Children

Plot 1

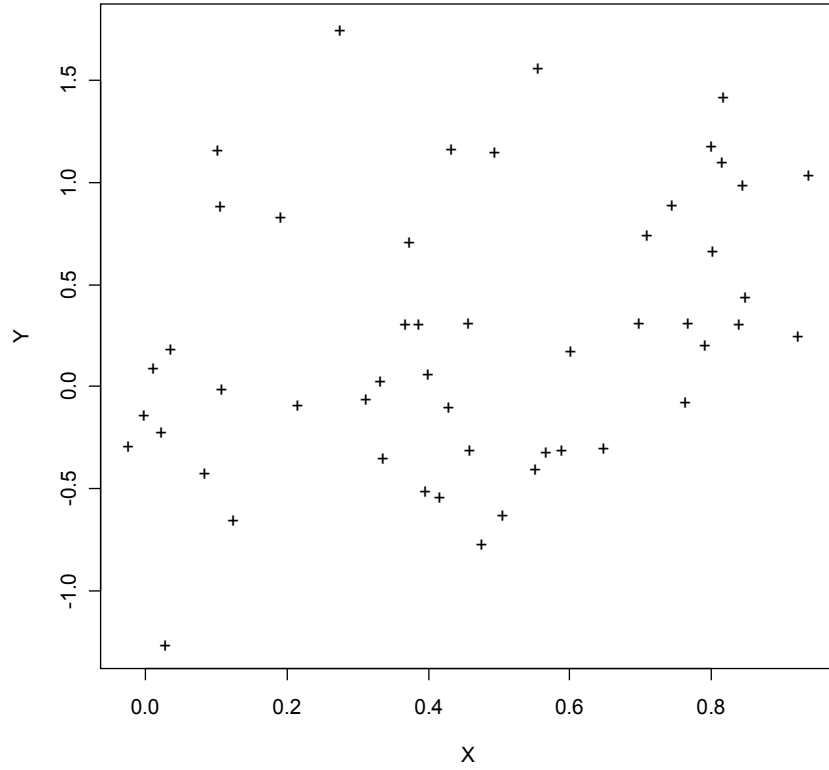


Plot 2

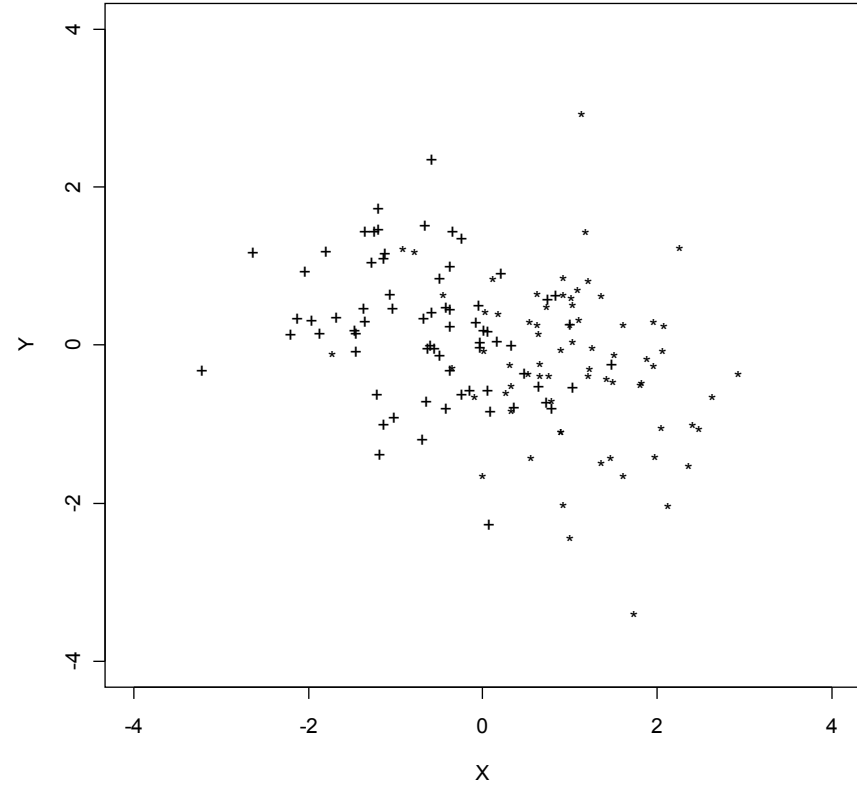


Pearson Correlation: Problem Children

Plot 3

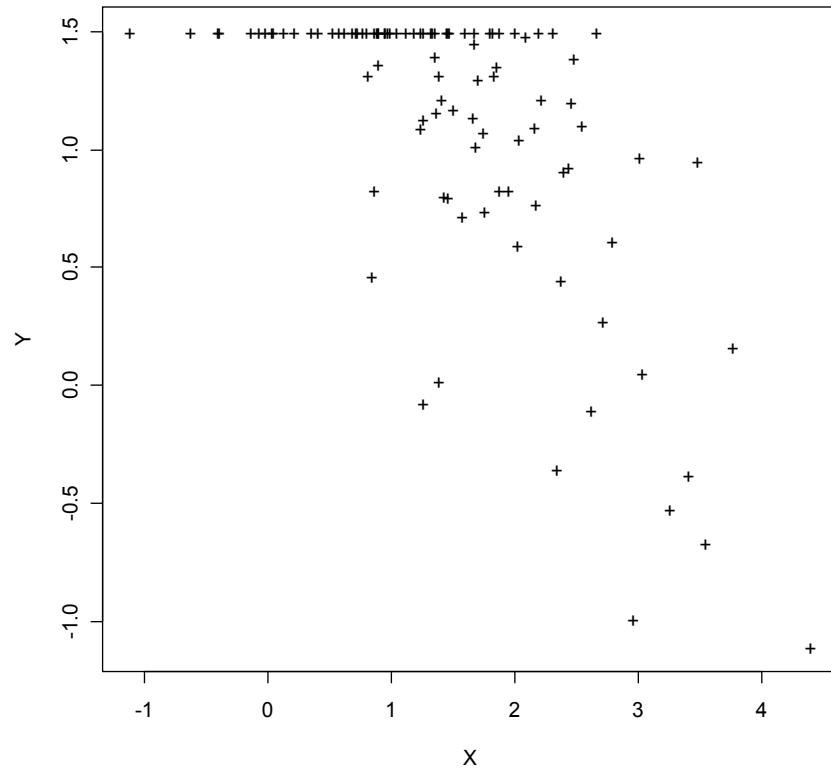


Plot 4

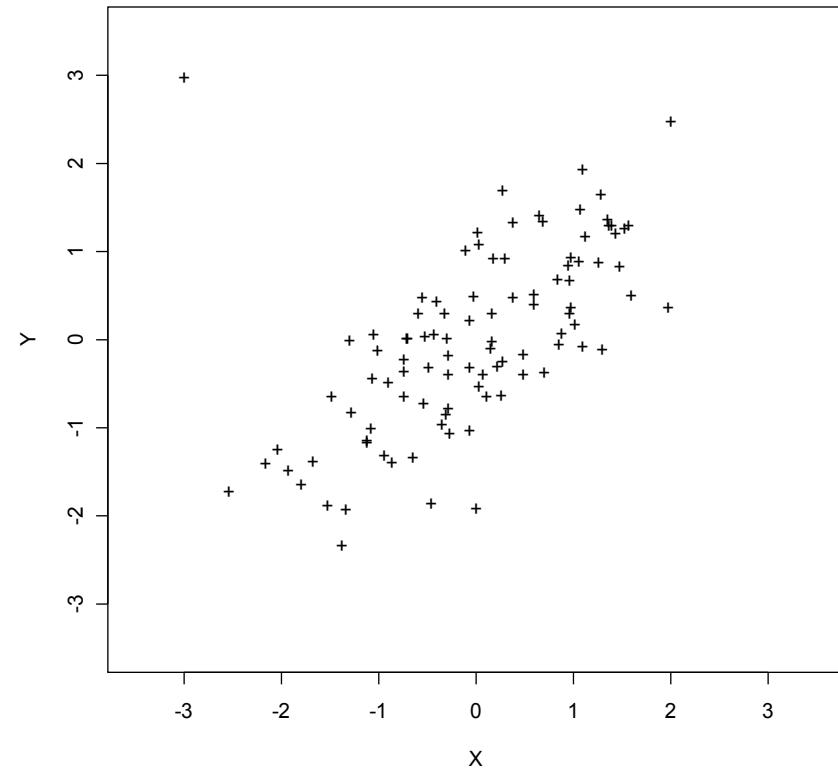


Pearson Correlation: Problem Children

Plot 5



Plot 6



Correlation Does Not Imply Causation

I repeat:

CORRELATION DOES NOT
IMPLY CAUSATION!!!!

Howell, 189-192