Correlation[1]

December 4, 2017

[1] HMS, 2017, v1.1

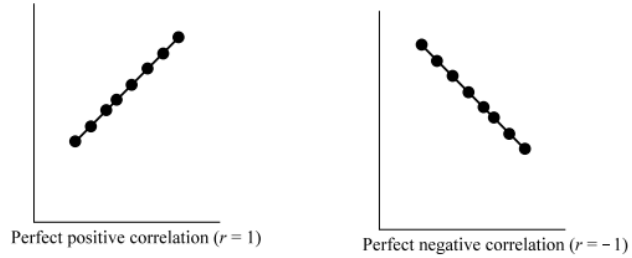
Chapter References

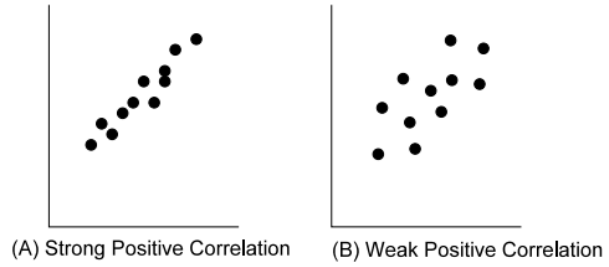
- Diez: Chapter 7
- Navidi, Chapter 7

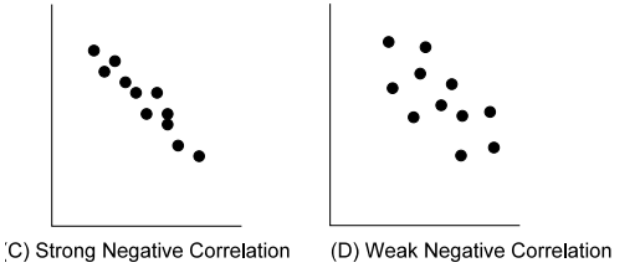
I don't expect you to learn the proofs what will follow.

Correlation

The sample correlation coefficient, r, is used to indicate the strength of the relationship between y and x variables. r ranges from -1 to $+1$. Unlike the residual mean square, it has no units and is therefore comparable between different plots.

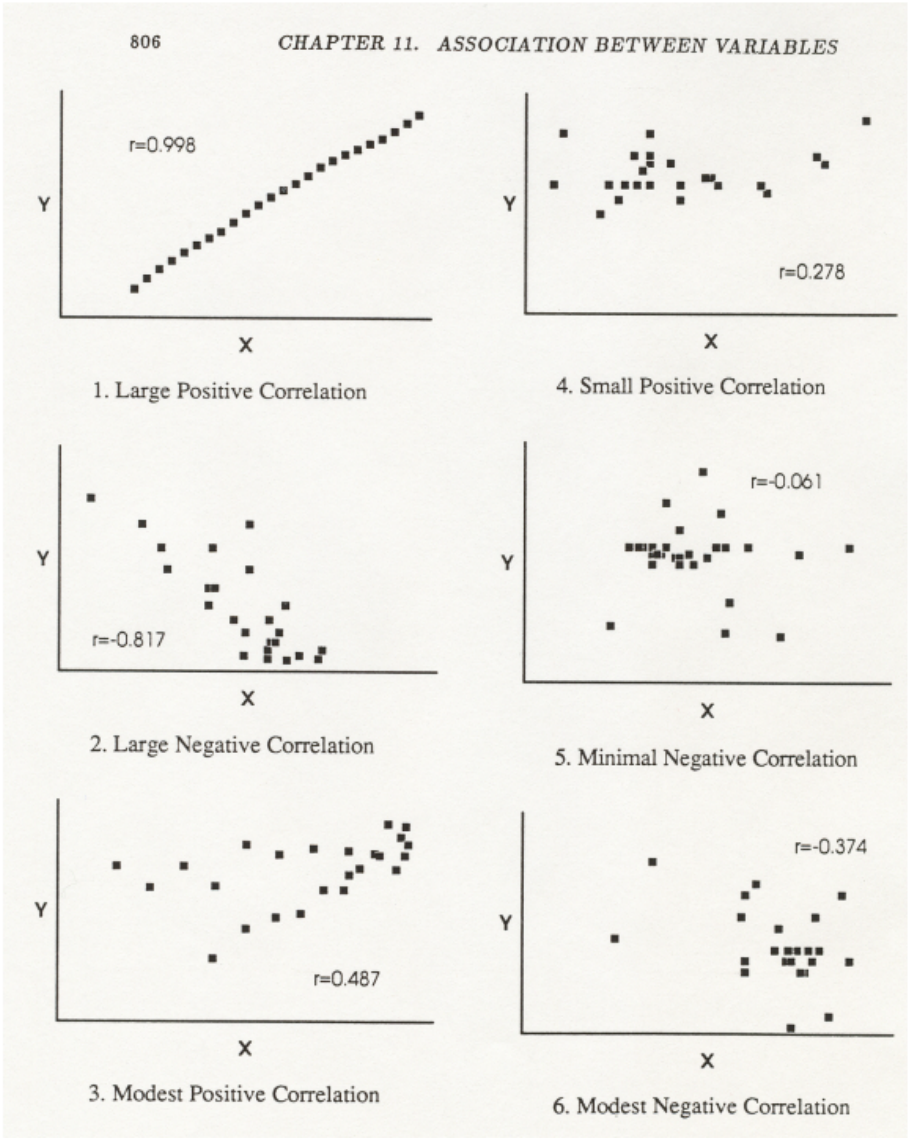






Correlation

Examples of r values:



Correlation

The sample correlation coefficient, r is useful because:

1. It is invariant to linear transformations in x and y, eg $y' = ay + b$.
2. It is independent of units
3. Does not distinguish between dependent and independent variables, r is the same for x vs y and y vs x

Correlation

Computing r, also called the **Pearson product moment correlation coefficient.**

$$SS_X = \sum (x_i - \bar{x})^2 \qquad SS_Y = \sum (y_i - \bar{y})^2$$

$$SS_{XY} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$r = \frac{SS_{XY}}{\sqrt{SS_X \; SS_Y}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}$$

Also

$$r = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \; \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

Connections

Recall:

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_X} \quad \therefore \quad SS_{XY} = \hat{\beta}_1 SS_X$$

Squaring r:

$$r^2 SS_X SS_Y = (SS_{XY})^2$$

$$r^2 SS_X SS_Y = \hat{\beta}_1{}^2 (SS_X)^2$$

$$r^2 SS_Y = \hat{\beta}_1{}^2 SS_X$$

$$r = \frac{\hat{\beta}_1 \sqrt{SS_X}}{\sqrt{SS_Y}}$$

$$r = \frac{\hat{\beta}_1 sd_X}{sd_Y}$$

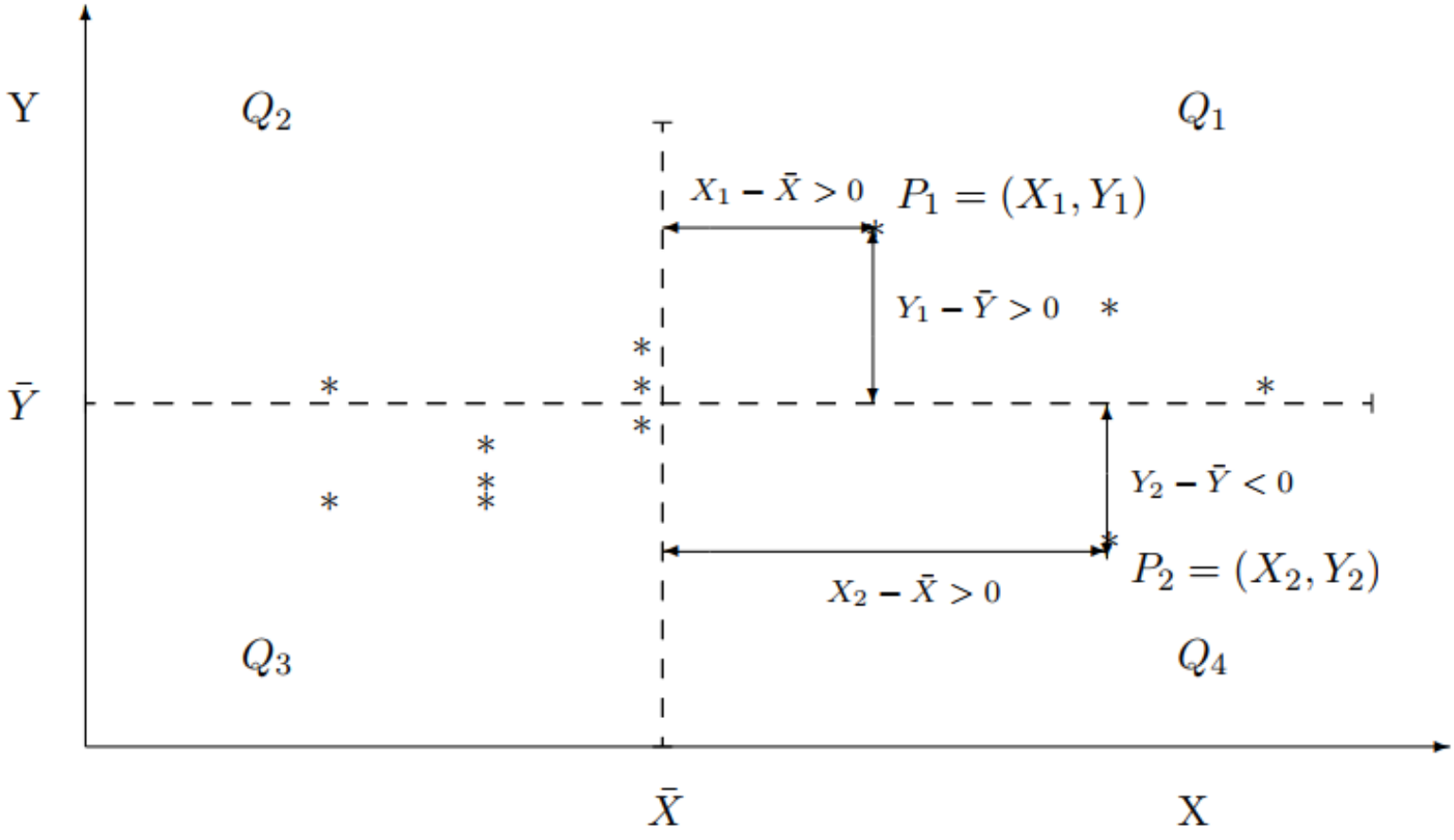
How does it work?

The denominator is a normalization term and ensures that r ranges from -1 to $+1$.

The real work happens in the numerator which is call the covariance of x and y.

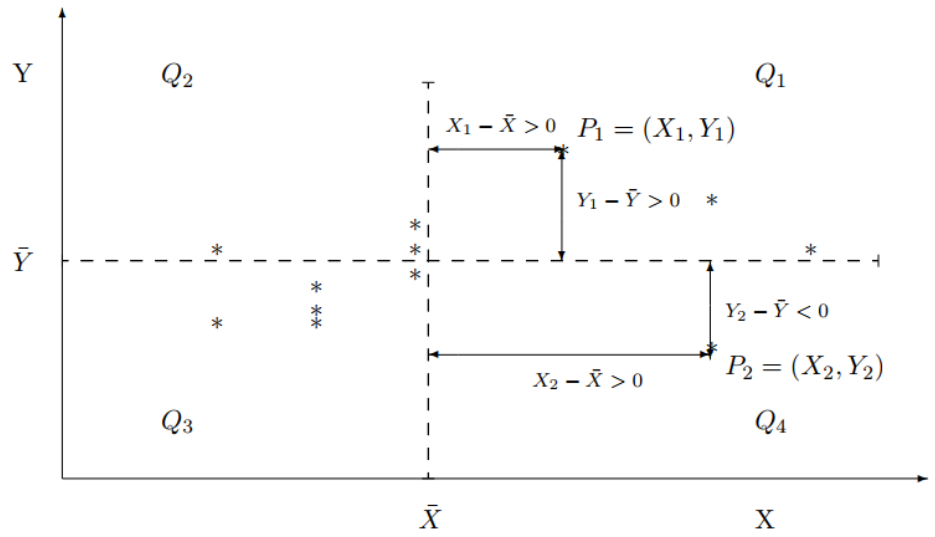
How does it work?

Covariance: $\sum(x_i - \bar{x})(y_i - \bar{y})$



How does it work?

Covariance: $\sum(x_i - \bar{x})(y_i - \bar{y})$



$$
\begin{array}{ll}
Q_1 & (x - \bar{x})(y - \bar{y}) > 0 \\
Q_2 & (x - \bar{x})(y - \bar{y}) < 0 \\
Q_3 & (x - \bar{x})(y - \bar{y}) > 0 \\
Q_4 & (x - \bar{x})(y - \bar{y}) < 0
\end{array}
$$

If there are lots of points in Q_1 and Q_3 the sum of the products will tend to be positive high

If there are lots of points in Q_2 and Q_4 the sum of the products will tend to be very negative

If there are lots of points in all quadrants the sum of the products will tend to cancel out, resulting in a low value.

Correlation

Hypothesis testing

The sample correlation coefficient r is the estimate of the population correlation coefficient (ρ).

The null and alternative hypotheses are:

$$H_o : \rho = 0 \qquad H_1 : \rho \neq 0 \quad \text{Note: Two sided}$$

The hypothesis can be tested with a t statistic:

$$t = \frac{r}{se_r}$$

where se_r is the standard error of the correlation coefficient:

$$se_r = \sqrt{\frac{1 - r^2}{n - 2}} \qquad \text{also } t = \frac{r\sqrt{n-2}}{\sqrt{1 - r^2}}$$

Degrees of freedom $= n - 2$

Correlation

Hypothesis testing: Example (Could the correlation have happened by chance alone?)

Let $r = -0.849$, and $n = 12$ therefore

$$se_r = \sqrt{\frac{1 - (-0.849)^2}{12 - 2}} = 0.167$$

$$t = \frac{-0.849}{0.167} = -5.08$$

With df $= 10$ the p-value in a t table is 0.00048.

We therefore reject the null hypothesis at $\alpha = 0.05$

Correlation

Coefficient of determination: r^2

The coefficient of determination represents the percent of the data that is the closest to the line of best fit.

For example, if $r = 0.922$, then $r^2 = 0.850$, which means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in y remains unexplained.

We will visit this concept again....

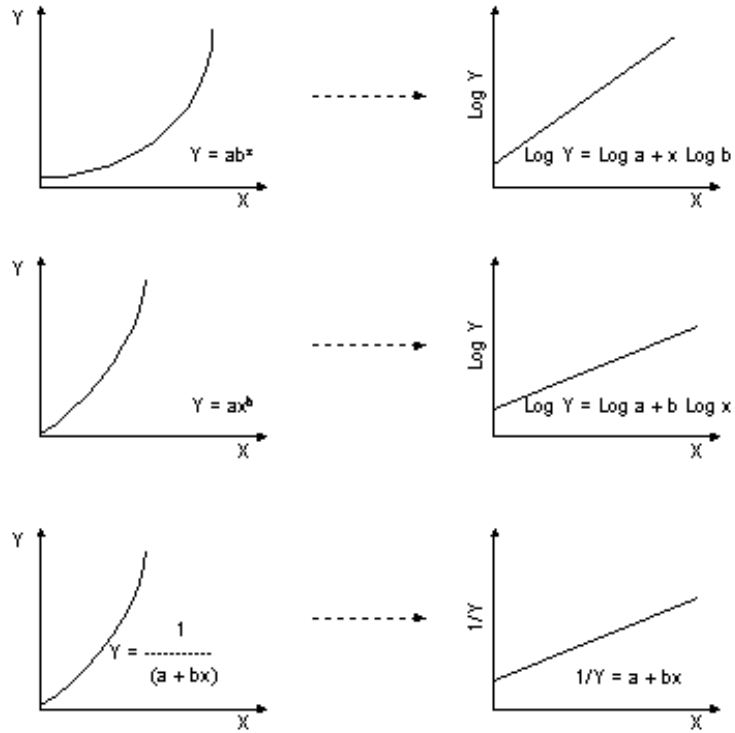
Transformations

Textbooks will often suggest transforming the data in order to make the data behave linearly. The argues goes that once in linear form we can apply the usual regression techniques.

I personally do not recommend this.

Examples of Transformations

Function	Transform	Linear Form
$y = Ae^{\alpha x}$	$y' = \ln(y)$	$y' = \ln(A) + \alpha x$
$y = Ax^{\alpha}$	$y' = \log(t), x' = \log(x)$	$y' = \log(A) + \alpha x'$
$y = 1/(a + bx)$	$y' = 1/y$	$1/y' = a + bx$



Examples of Transformations

The alternative to linear transformations is to use **non-linear regression** to fit the actual nonlinear function.