

Bus 274: Further Statistics For Business

Harald Schmidbauer



Chapter 14:

Correlation



14.1 Introduction

Assumptions and the problem.

In this chapter, we assume that observations (x_i, y_i) , $i = 1, \dots, n$, from a bivariate metric variable (X, Y) are given.

How can we measure the degree of linear dependence between X and Y ?

Whatever the goal of our analysis is, the first step is usually to plot the data.



14.1 Introduction

Example: The expenditure (in euros) of 508 customers on certain groups of goods at a supermarket was recorded.

Recorded were among others: Expenditure on. . .

- bread
- cheese
- dairy products
- fruit
- tea & coffee

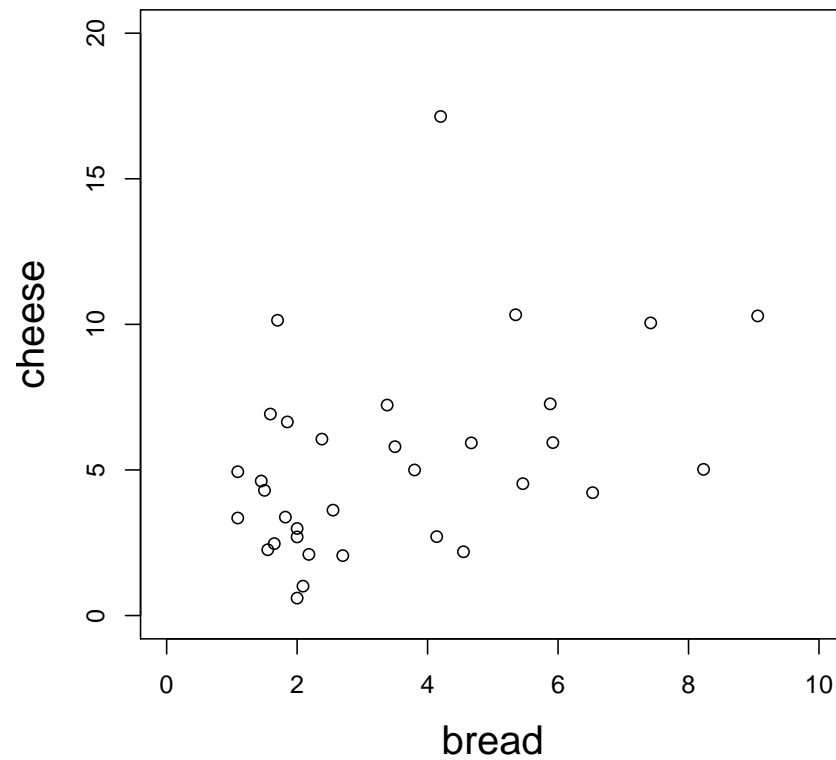
What is the relation between these variables? — Is there any?

Scatterplots will provide us with first insight.



14.1 Introduction

Expenditure on bread and cheese.

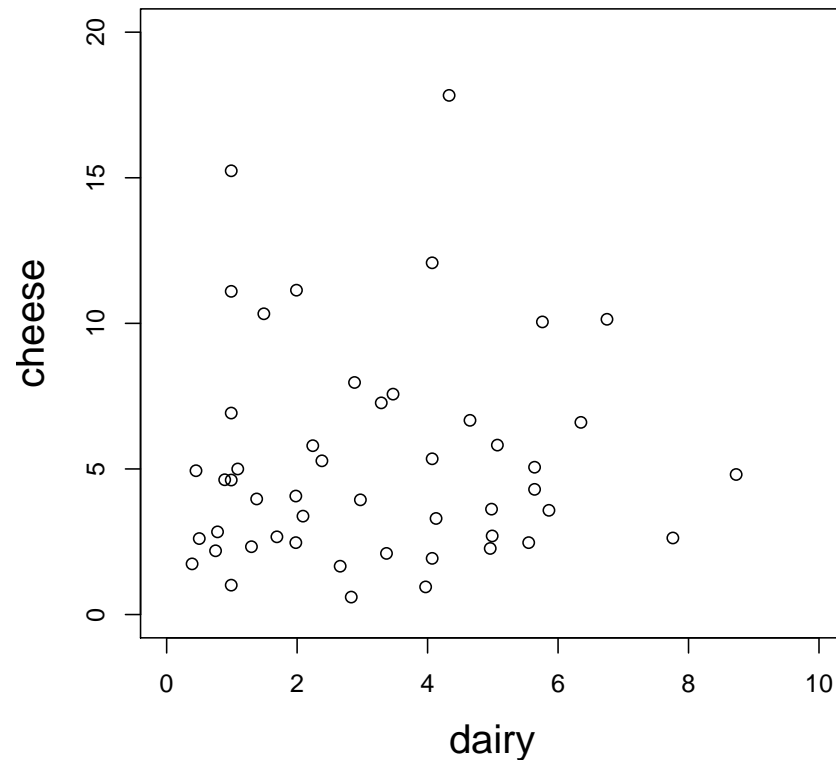


(Shown: only those customers who actually bought both groups.)



14.1 Introduction

Expenditure on dairy products and cheese.

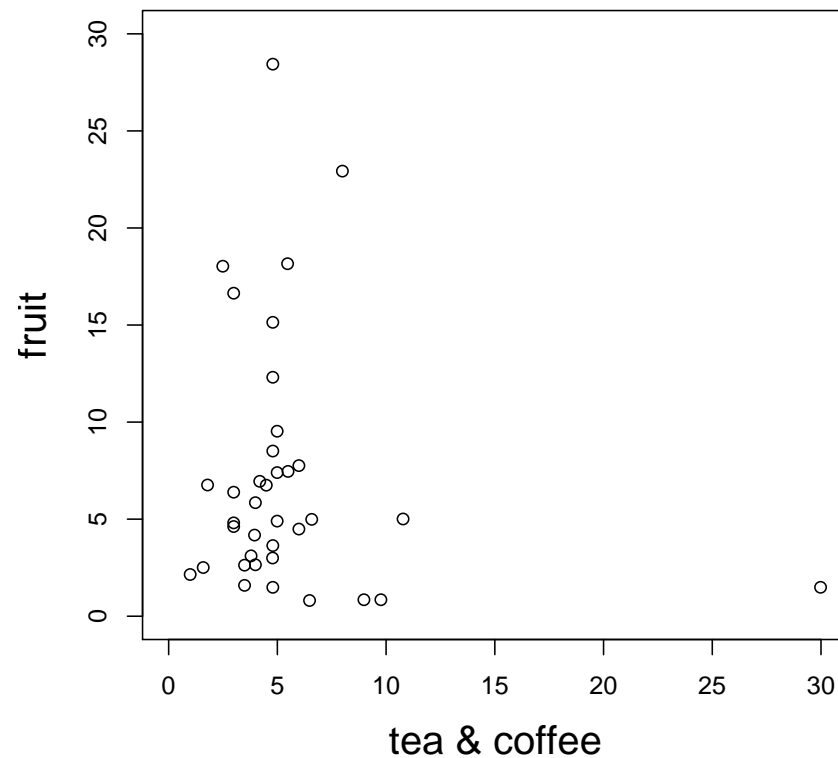


(Shown: only those customers who actually bought both groups.)



14.1 Introduction

Expenditure on tea/coffee and fruit.

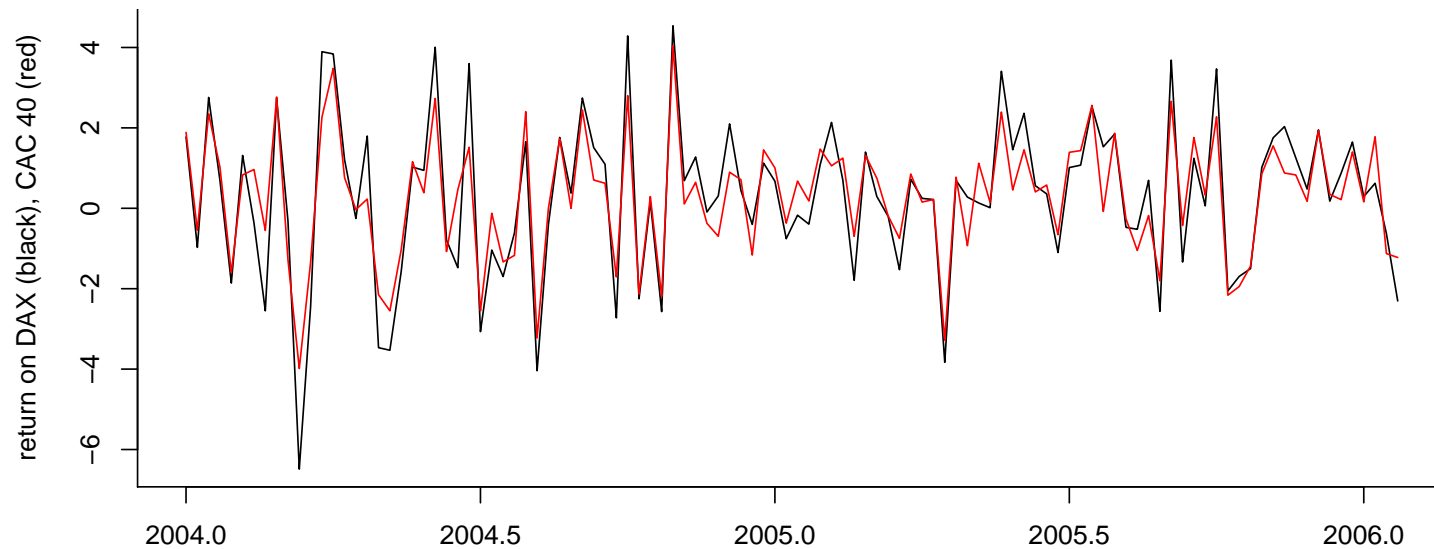


(Shown: only those customers who actually bought both groups.)



14.1 Introduction

Example: Weekly returns on stock indices DAX (gdaxi) and CAC 40 (fchi).



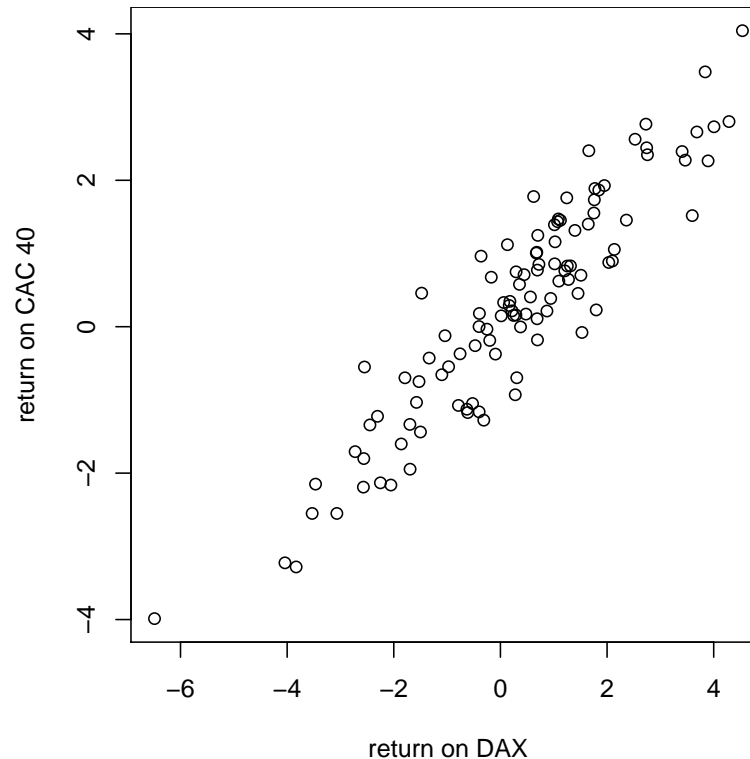
There is obviously a close association between DAX and CAC 40.

But to investigate this, another display is more useful.



14.1 Introduction

Using a scatterplot.

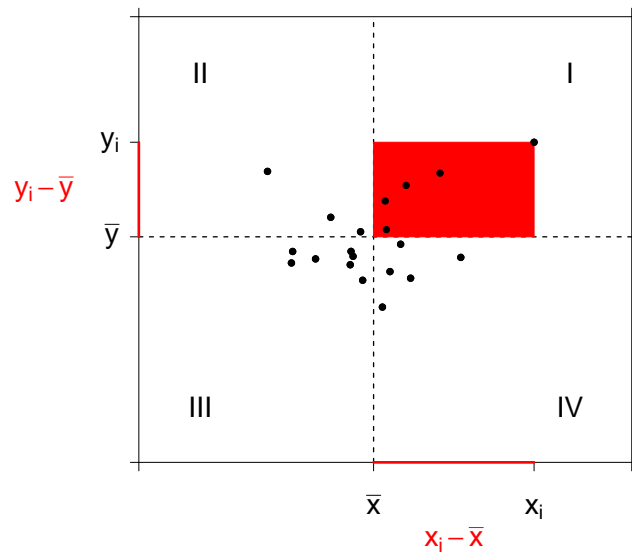


The scatterplot reveals the high correlation between returns on DAX and returns on CAC 40.



14.2 Covariance

Defining the covariance.



$$\text{Area: } (x_i - \bar{x})(y_i - \bar{y})$$

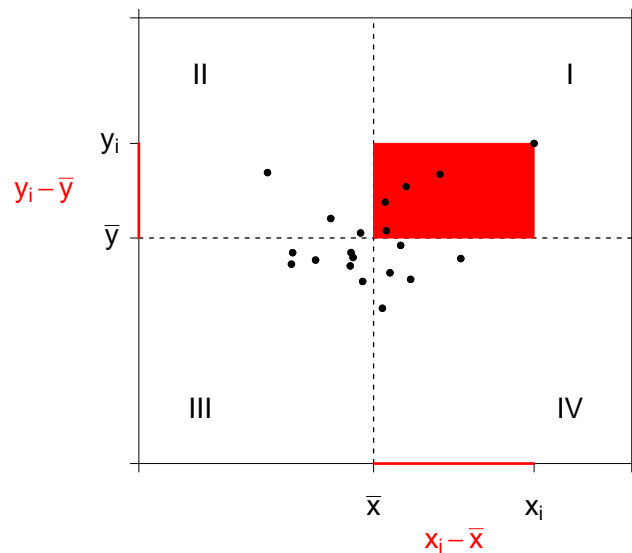
The covariance is defined as the average size of all rectangles:

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$



14.2 Covariance

Interpreting the covariance.



In I and III:

$$(x_i - \bar{x})(y_i - \bar{y}) > 0$$

In II and IV:

$$(x_i - \bar{x})(y_i - \bar{y}) < 0$$

If the points (x_i, y_i) are predominantly in quadrant. . .

. . . I and III: $\text{cov}(X, Y) > 0$

. . . II and IV: $\text{cov}(X, Y) < 0$



14.2 Covariance

Some properties of the covariance.

- The sign of $\text{cov}(X, Y)$ tells us in which direction X and Y are associated.
- The covariance is symmetric: $\text{cov}(X, Y) = \text{cov}(Y, X)$
- It holds that $\text{cov}(aX + b, Y) = a \cdot \text{cov}(X, Y)$;
in particular: The covariance depends on the unit of measurement.

This makes it sometimes difficult to use.

This is why we often prefer to investigate the relationship between two variables using the *correlation*, rather than the covariance.



14.3 Correlation

Definition: The correlation of X and Y is defined as

$$r = \text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}}$$

It has the same sign as the covariance.

Reminder:

$$\text{var}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$



14.3 Correlation

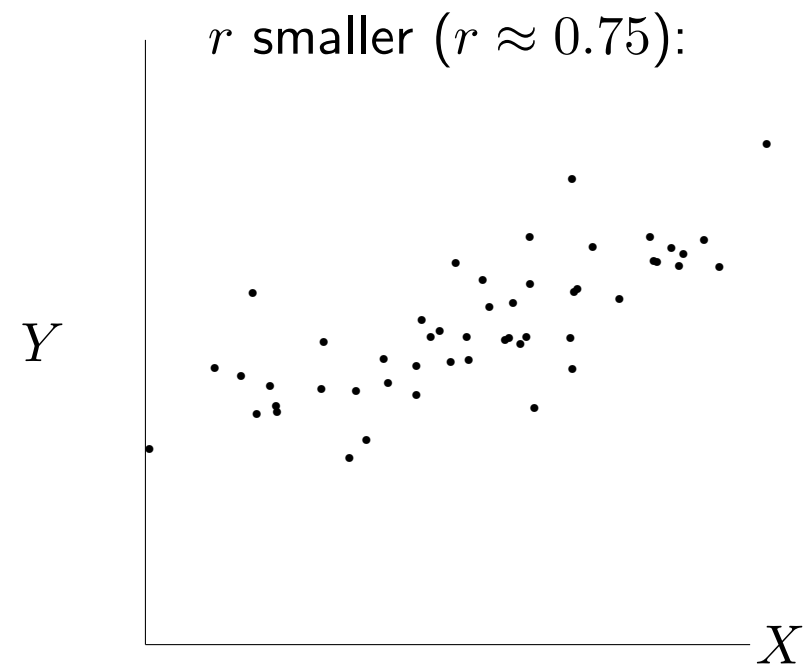
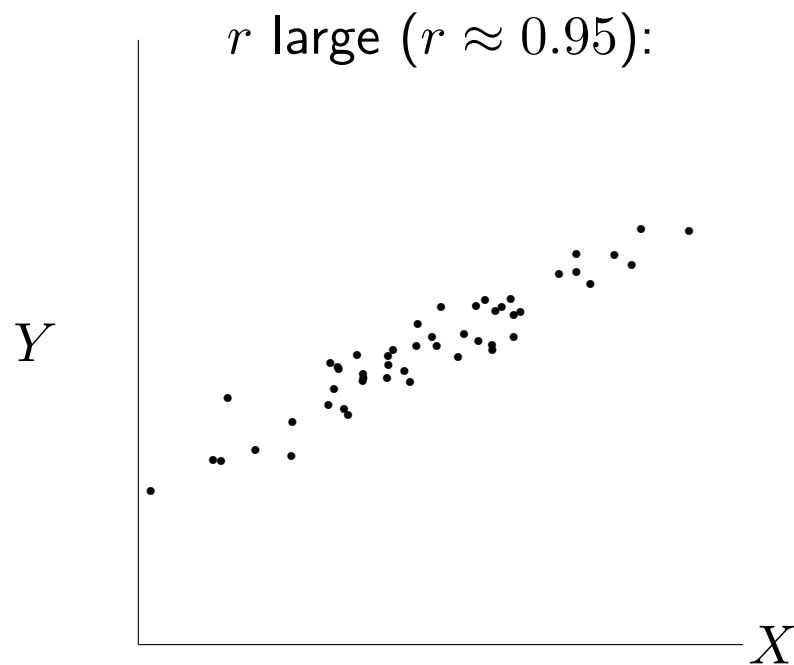
Some properties of the correlation.

- The sign of $\text{cor}(X, Y)$ tells us in which direction X and Y are associated.
- The correlation is *normed*: $-1 \leq \text{cor}(X, Y) \leq +1$.
- It holds that $\text{cor}(X, Y) = \pm 1$ if and only if all points (x_i, y_i) are on a straight line with positive (negative) slope.
- The correlation is symmetric: $\text{cor}(X, Y) = \text{cor}(Y, X)$
- It holds that $\text{cor}(aX + b, Y) = \text{cor}(X, Y)$ ($a > 0$);
in particular: The correlation does not depend on the unit of measurement.



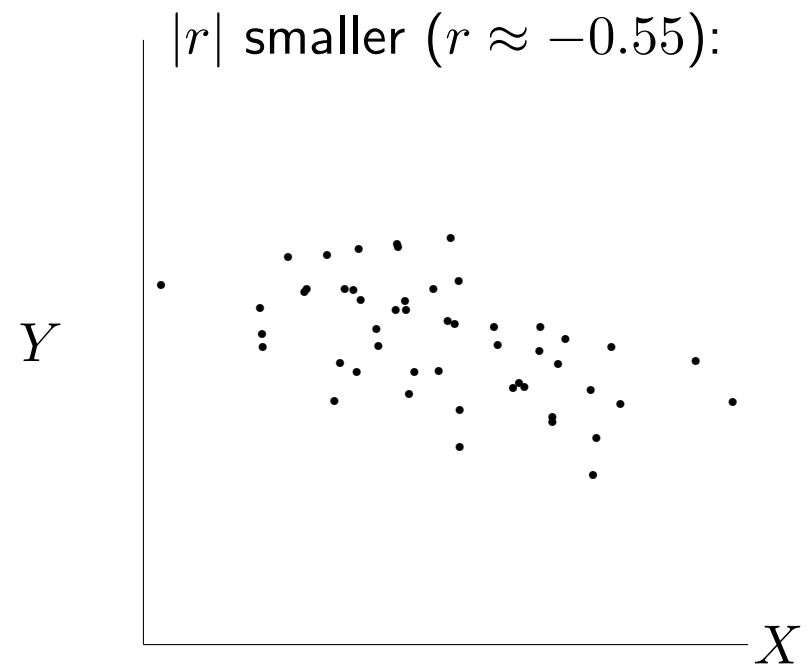
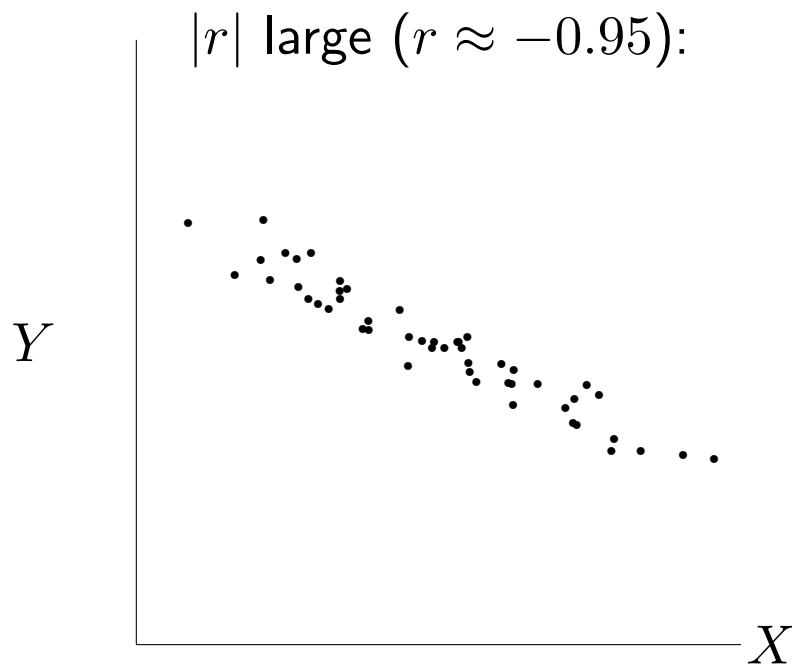
14.3 Correlation

Correlation patterns I: $r > 0$, i.e. the linear relation between X and Y is positive



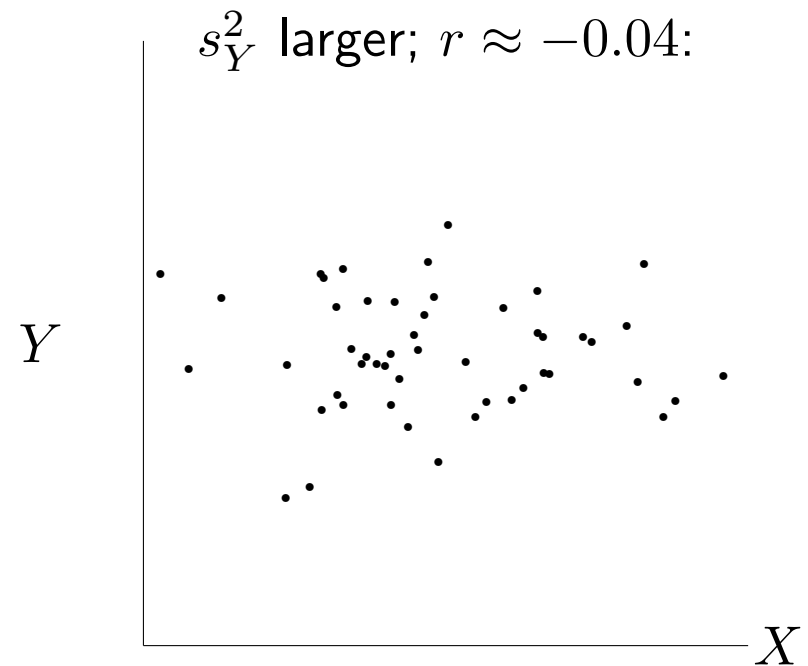
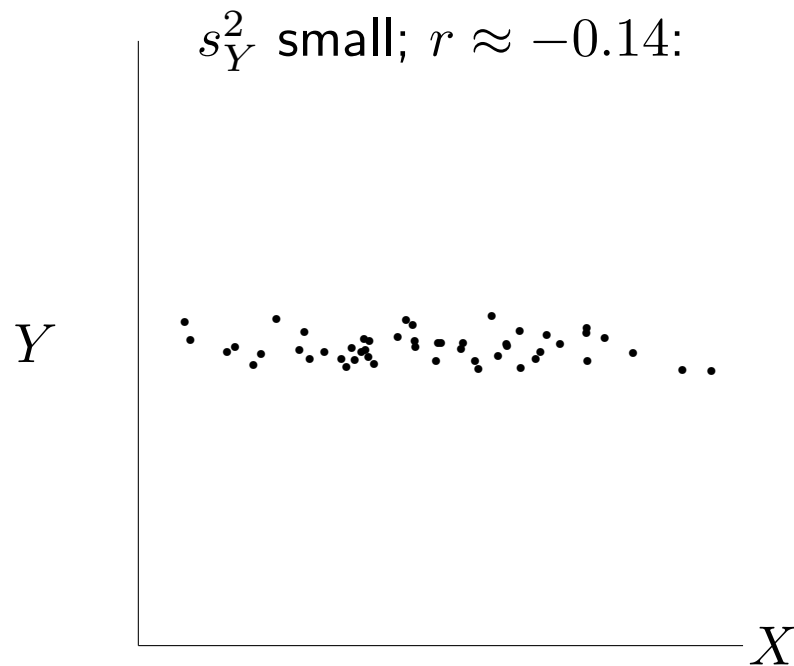
14.3 Correlation

Correlation patterns II: $r < 0$, i.e. the linear relation between X and Y is negative



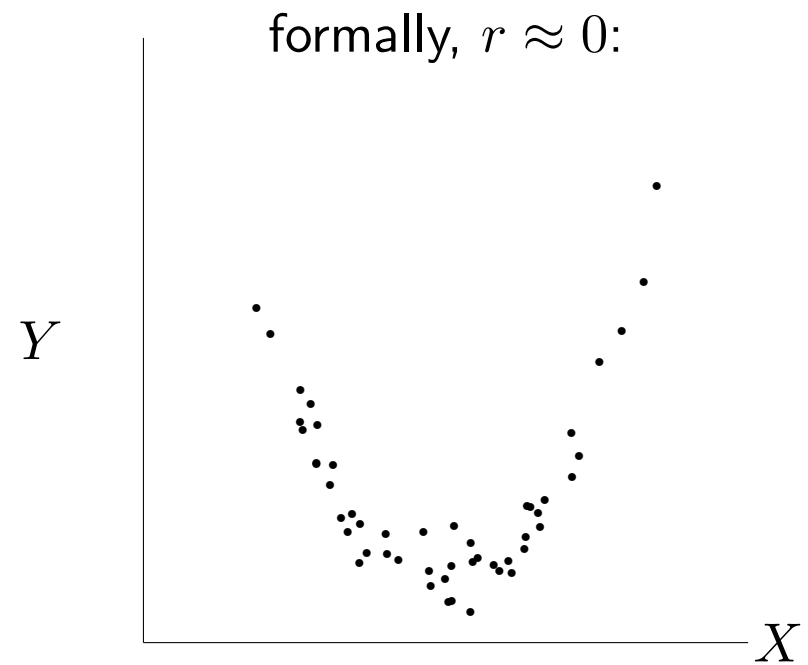
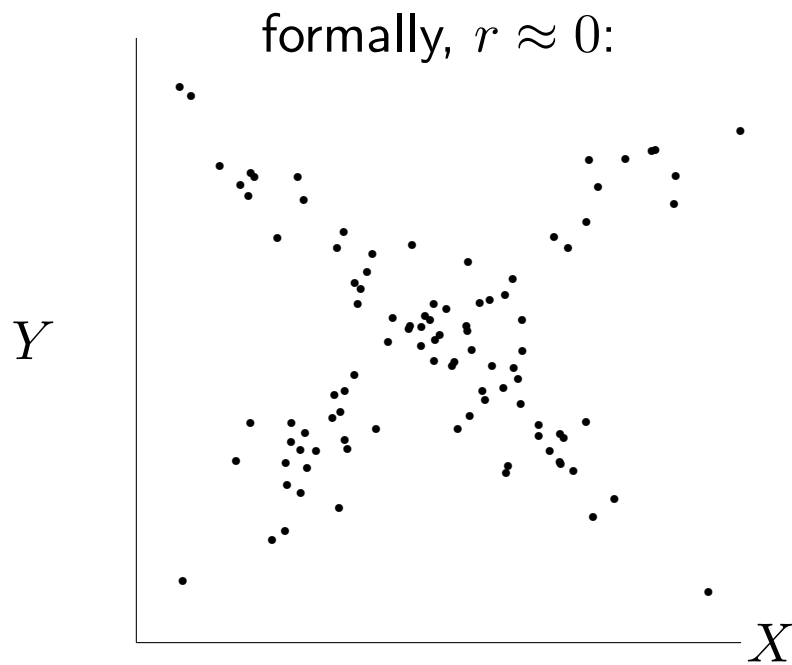
14.3 Correlation

Correlation patterns III: r close to 0, with no apparent relation between X and Y



14.3 Correlation

Correlation patterns IV: r not meaningful because there is a nonlinear relation between X and Y



14.3 Correlation

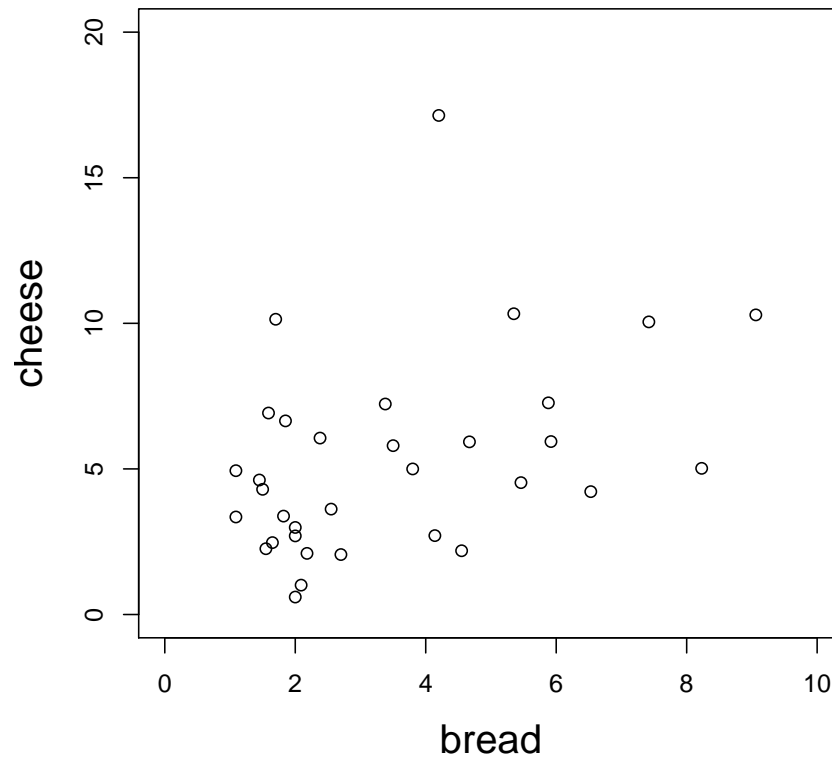
Uncorrelated and independent are not the same.

- Two variables are called uncorrelated if $\text{cor}(X, Y) = 0$.
- The last two figures show that being uncorrelated is a relatively weak property: There can be a strong non-linear relationship between uncorrelated variables.
- Being independent is much stronger: Independent variables have no relation whatsoever.



14.4 Examples

Expenditure on bread and cheese.



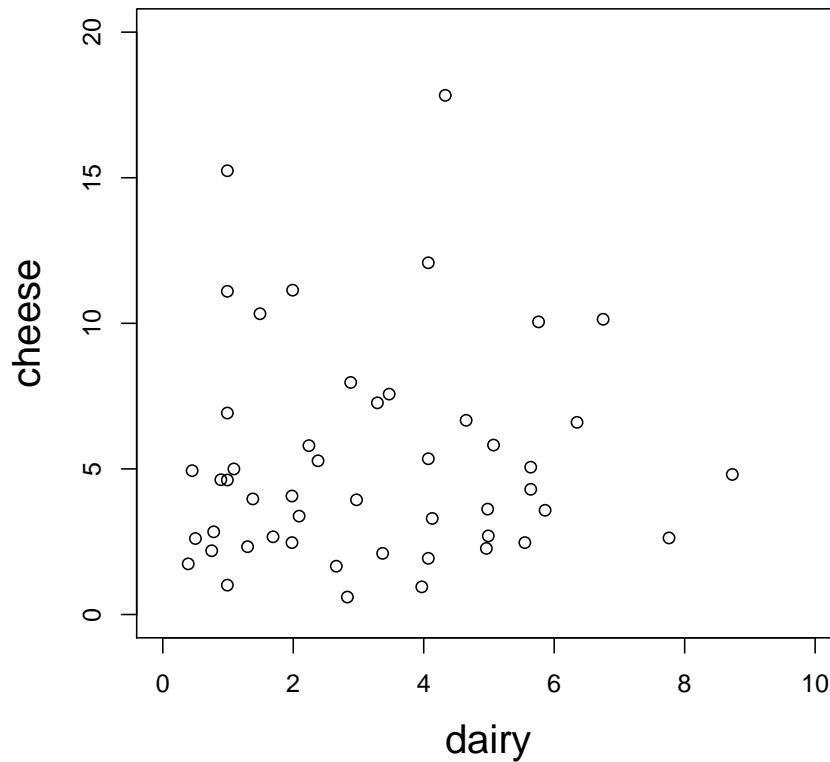
$$r = 0.41$$

Moderate positive correlation.



14.4 Examples

Expenditure on dairy products and cheese.



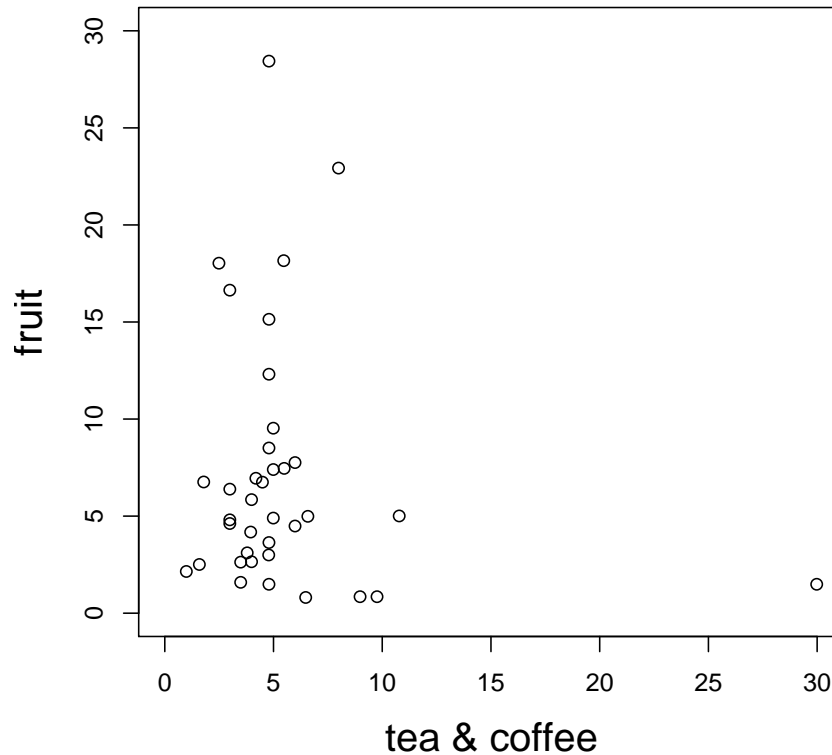
$$r = 0.05$$

Practically uncorrelated.



14.4 Examples

Expenditure on tea/coffee and fruit.



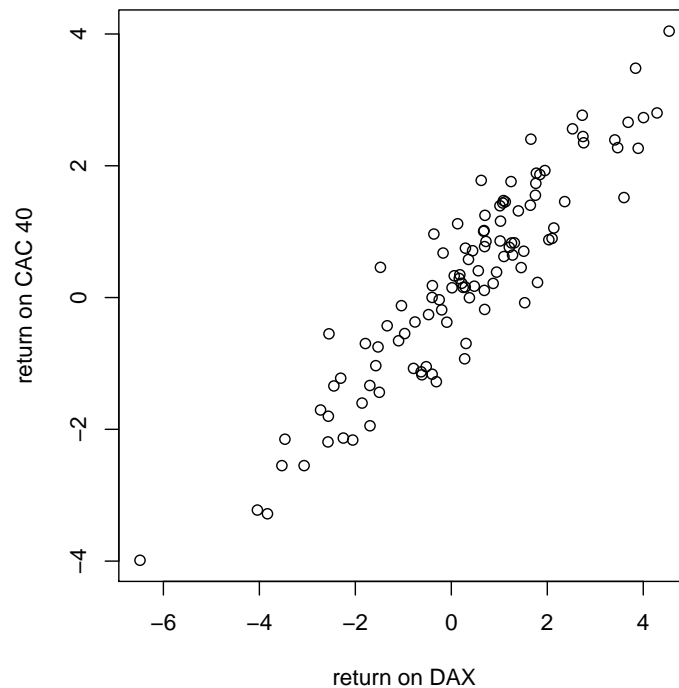
$$r = -0.13$$

The correlation is negative — but this has no meaningful interpretation.



14.4 Examples

Weekly returns on stock indices DAX and CAC 40.



$$r = 0.925$$

Returns on DAX and CAC 40 are highly correlated. Here, the correlation is very useful.



14.4 Examples

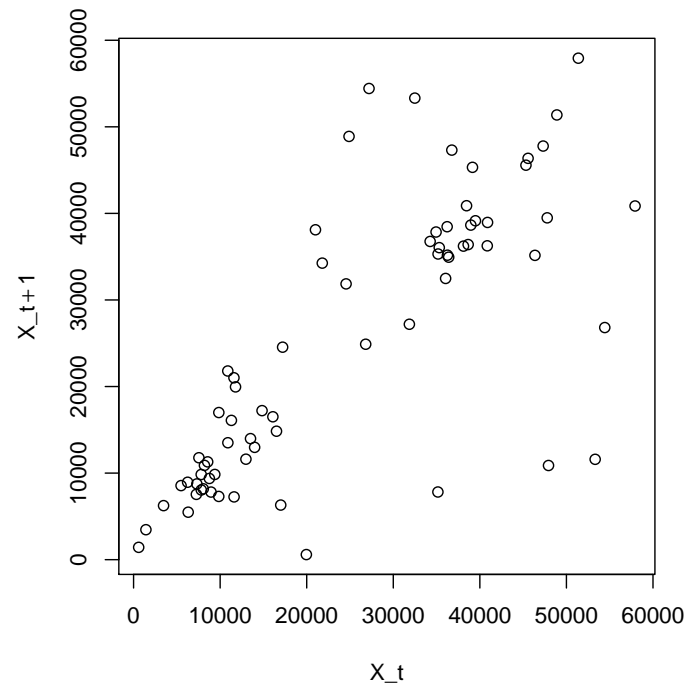
Another application of correlation.

- We have seen the correlation, as applied to two different variables X, Y .
- The concept of correlation can also be applied to a series $(X_t) = X_1, X_2, X_3, \dots$
- $\text{cor}(X_t, X_{t+1})$ is called the autocorrelation (at lag 1) of the series (X_t) .
- Autocorrelation is a very important tool in the analysis of a time series.



14.4 Examples

Example: Monthly car sales in Turkey.



$$r = 0.76$$

This high autocorrelation can be used for forecasting purposes.



14.5 Outlook

A final remark.

- We have used correlation only in the context of descriptive statistics.
- We shall come back to inductive statistics in the next chapter, which deals with a related topic.

