Bus 701: Advanced Statistics

Harald Schmidbauer

Chapter 12:

Correlation

12.1 Introduction

Assumptions and the problem.

In this chapter, we assume that observations (x_i, y_i), $i = 1, \ldots, n$, 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.

12.1 Introduction

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

Recorded were among others: Expenditure for. . .

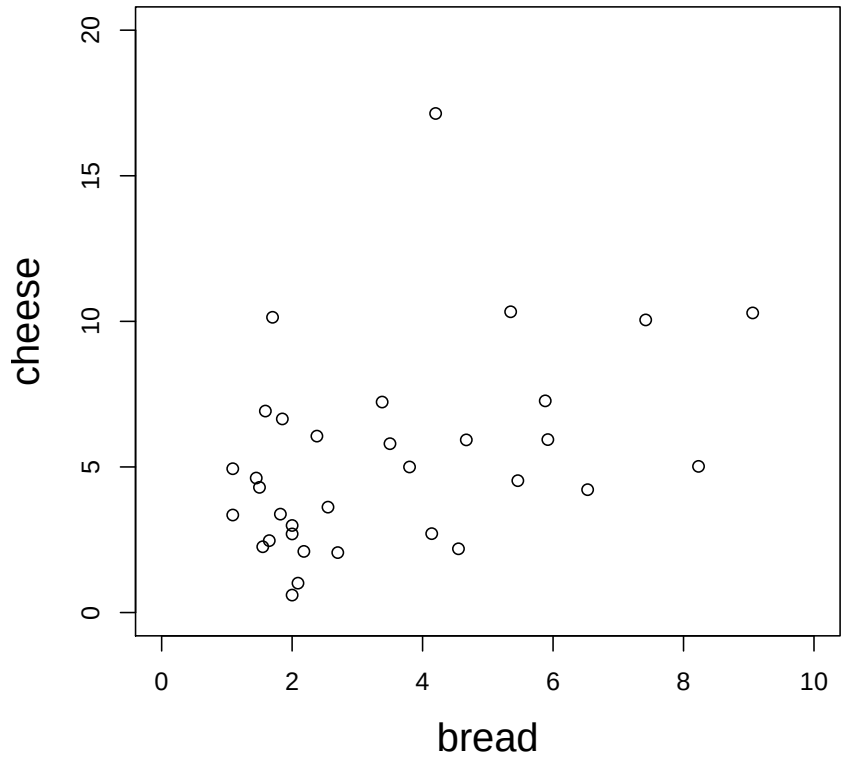
- bread
- cheese
- dairy products
- fruit
- tea & coffee

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

Scatterplots will provide us with first insight.

12.1 Introduction

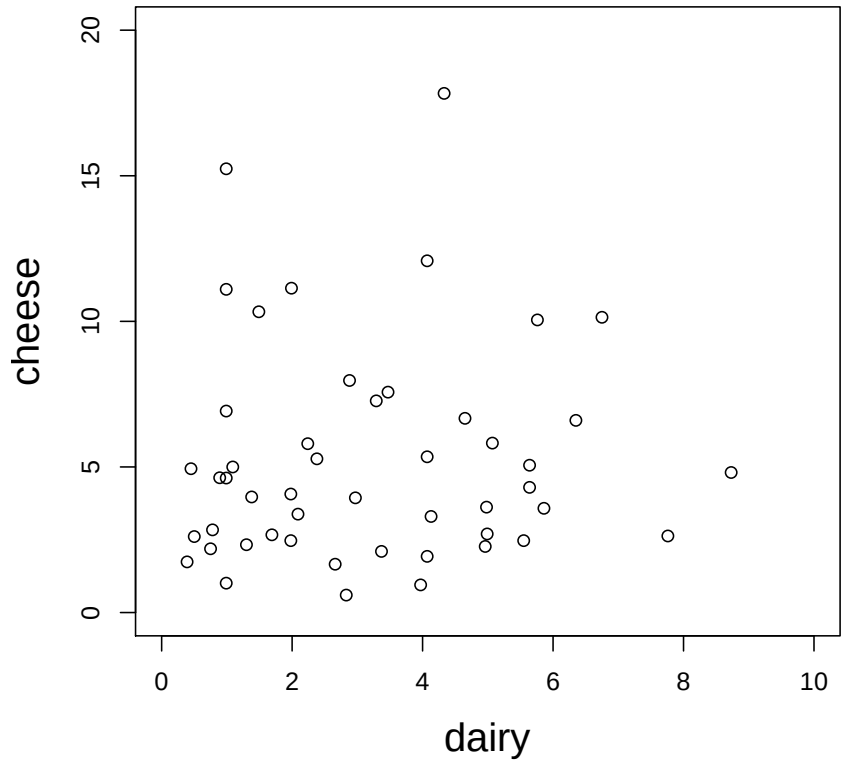
Expenditure for bread and cheese.



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

12.1 Introduction

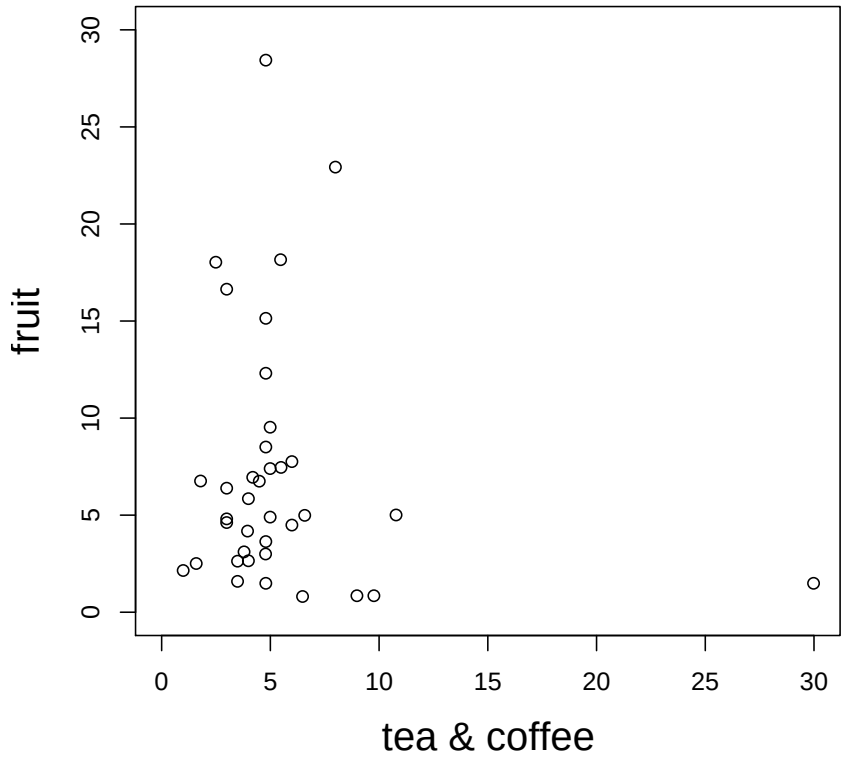
Expenditure for dairy products and cheese.



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

12.1 Introduction

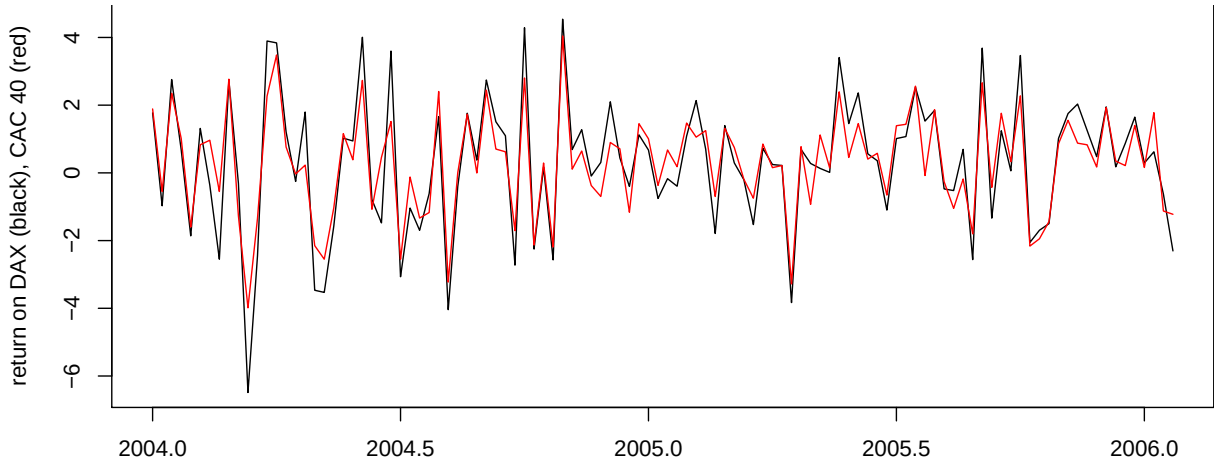
Expenditure for tea/coffee and fruit.



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

12.1 Introduction

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

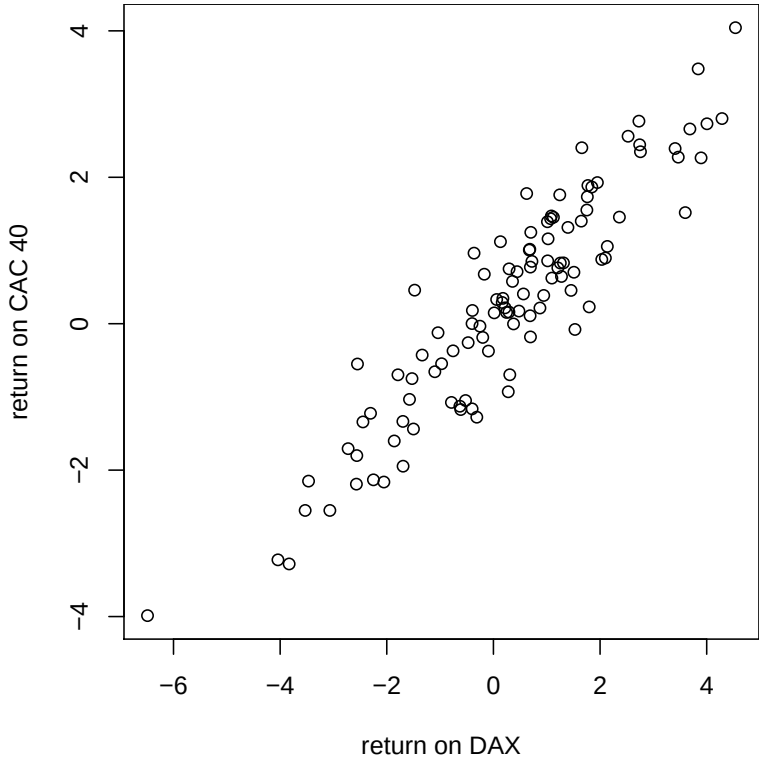


There is obviously a close association between DAX and CAC 40.

But to investigate this, another display is more useful.

12.1 Introduction

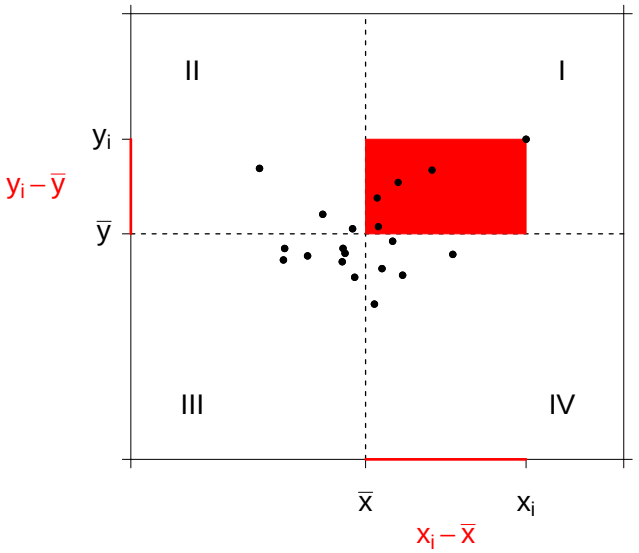
Using a scatterplot.



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

12.2 Covariance

Defining the covariance.



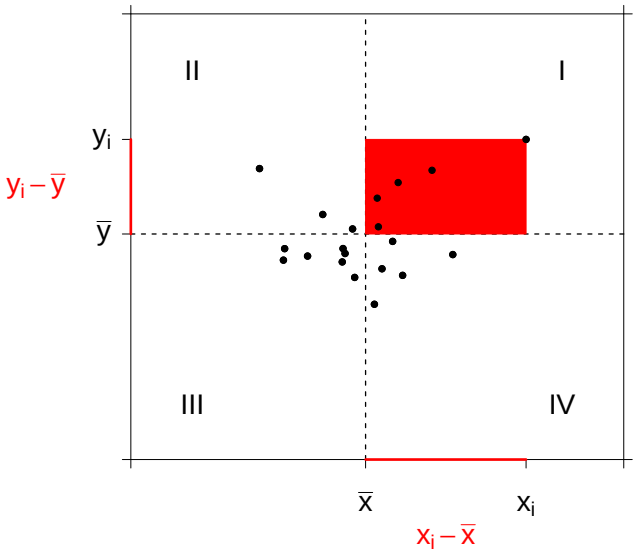
Area: $(x_i - \bar{x})(y_i - \bar{y})$

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

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

12.2 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: $\mathrm{cov}(X, Y) > 0$

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

12.2 Covariance

Some properties of the covariance.

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

12.3 Correlation

Definition: The correlation of X and Y is defined as

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

It has the same sign as the covariance.

Reminder:

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

12.3 Correlation

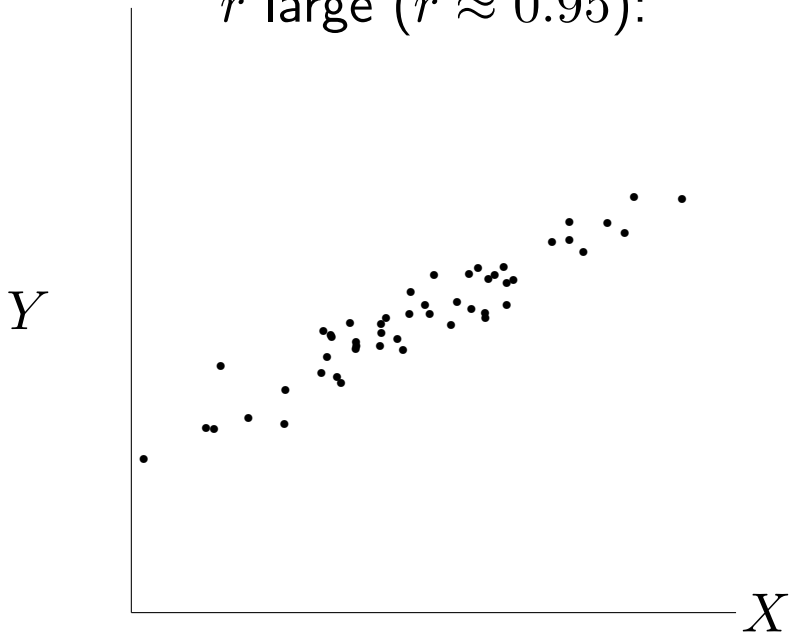
Some properties of the correlation.

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

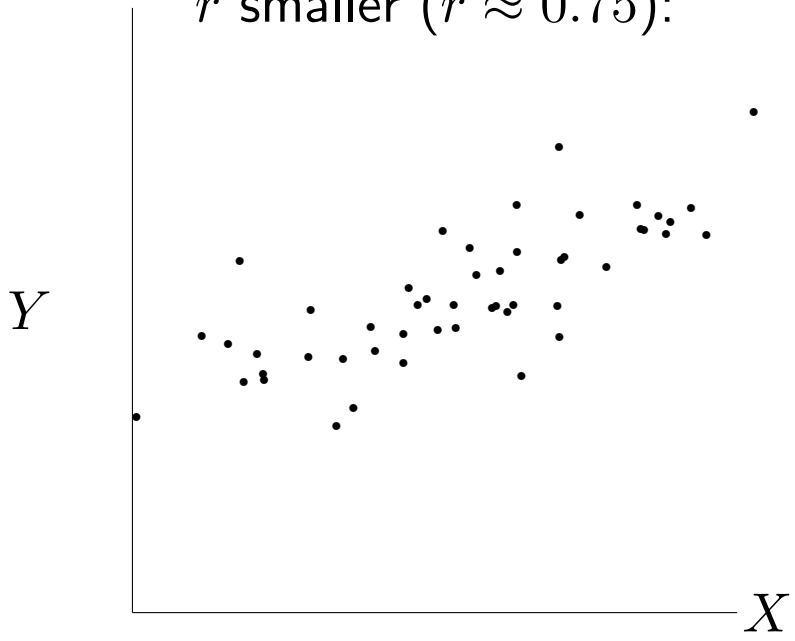
12.3 Correlation

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

r large ($r \approx 0.95$):



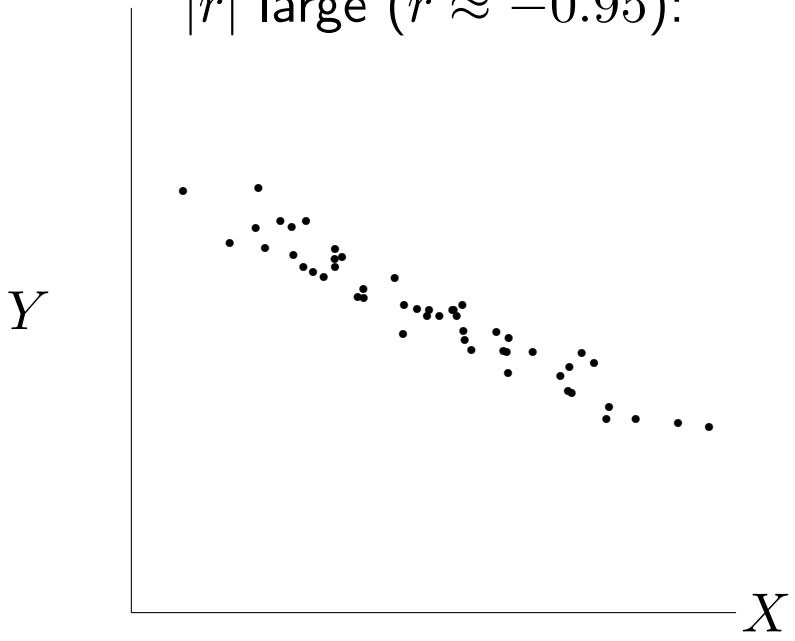
r smaller ($r \approx 0.75$):



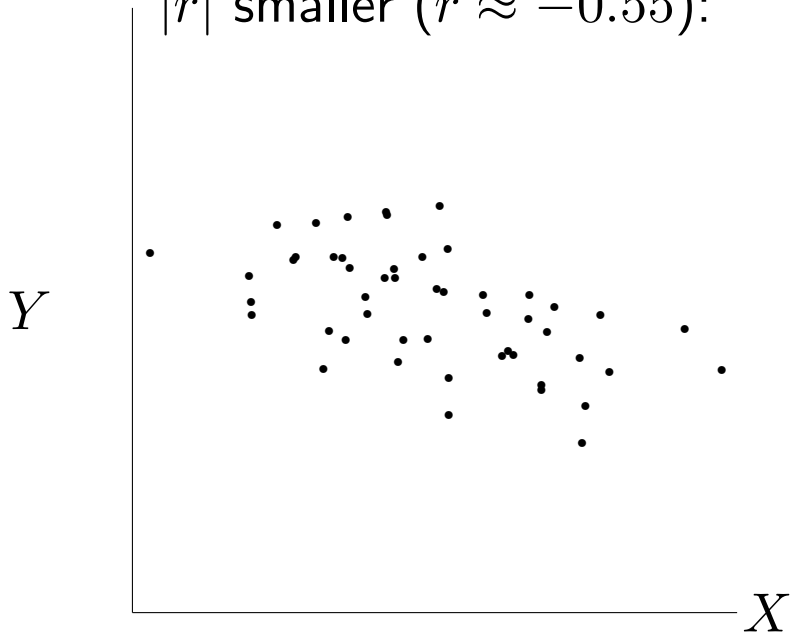
12.3 Correlation

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

$|r|$ large ($r \approx -0.95$):



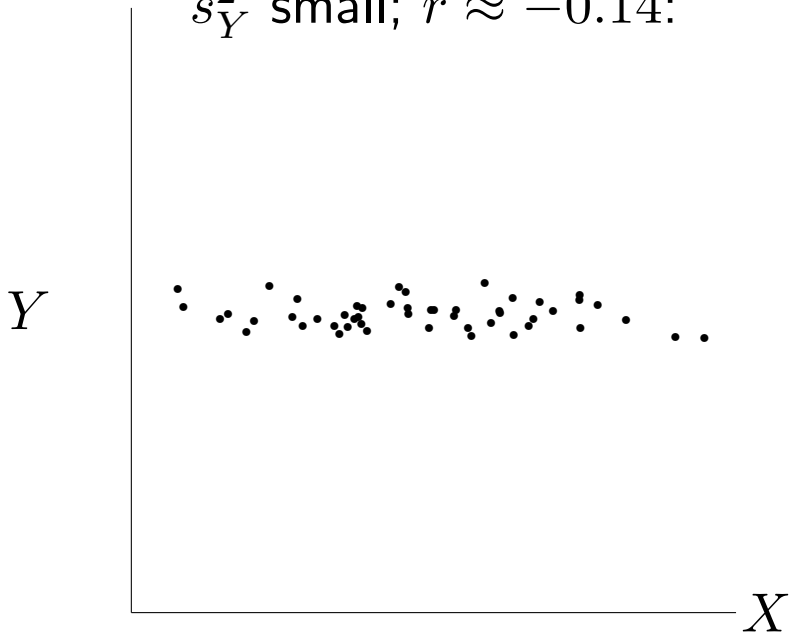
$|r|$ smaller ($r \approx -0.55$):



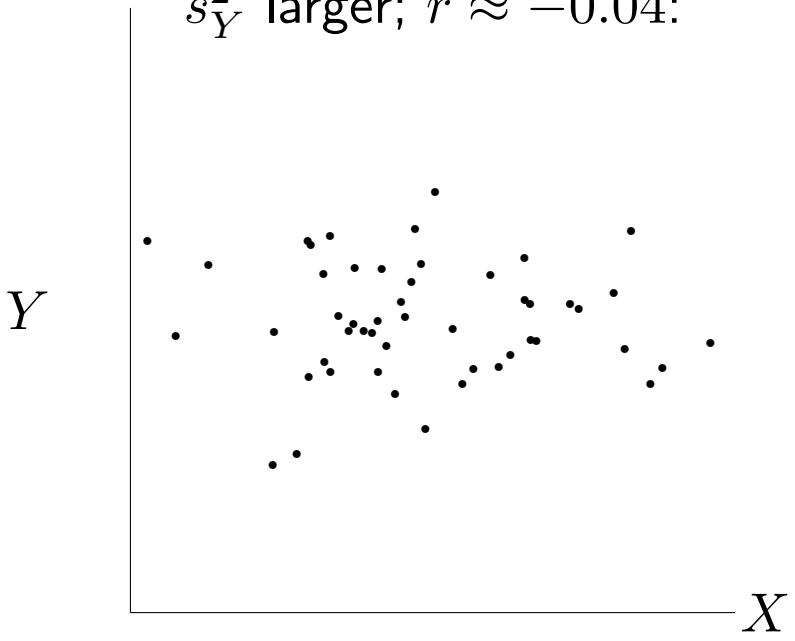
12.3 Correlation

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

s_Y^2 small; $r \approx -0.14$:

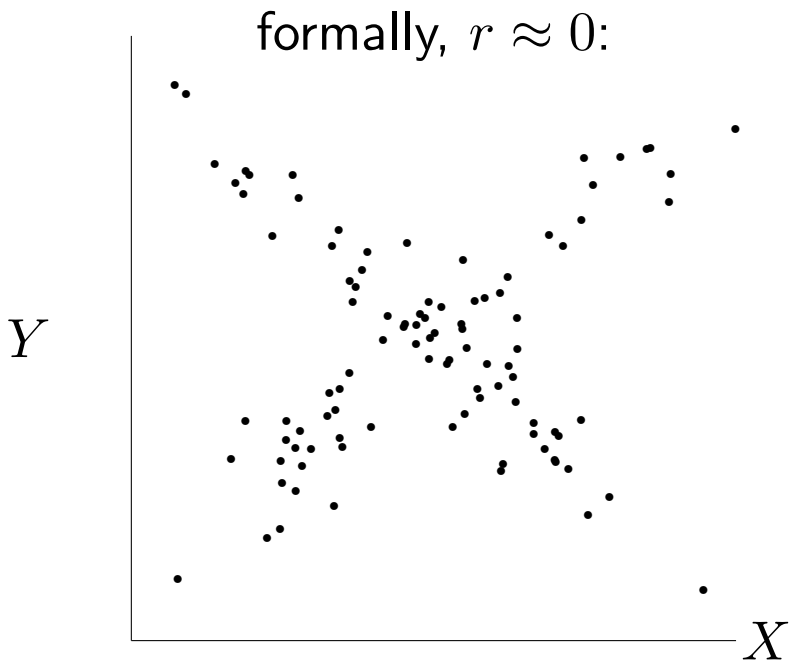


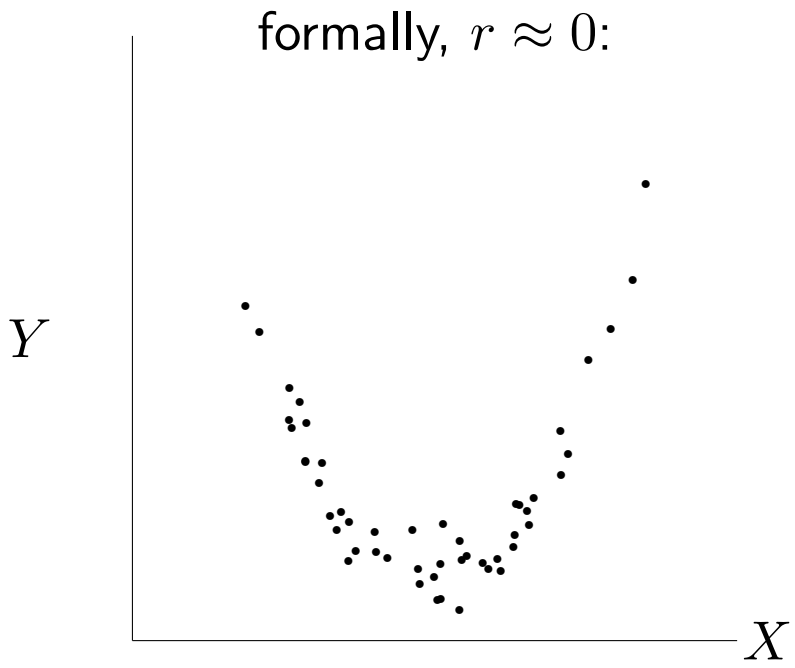
s_Y^2 larger; $r \approx -0.04$:



12.3 Correlation

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





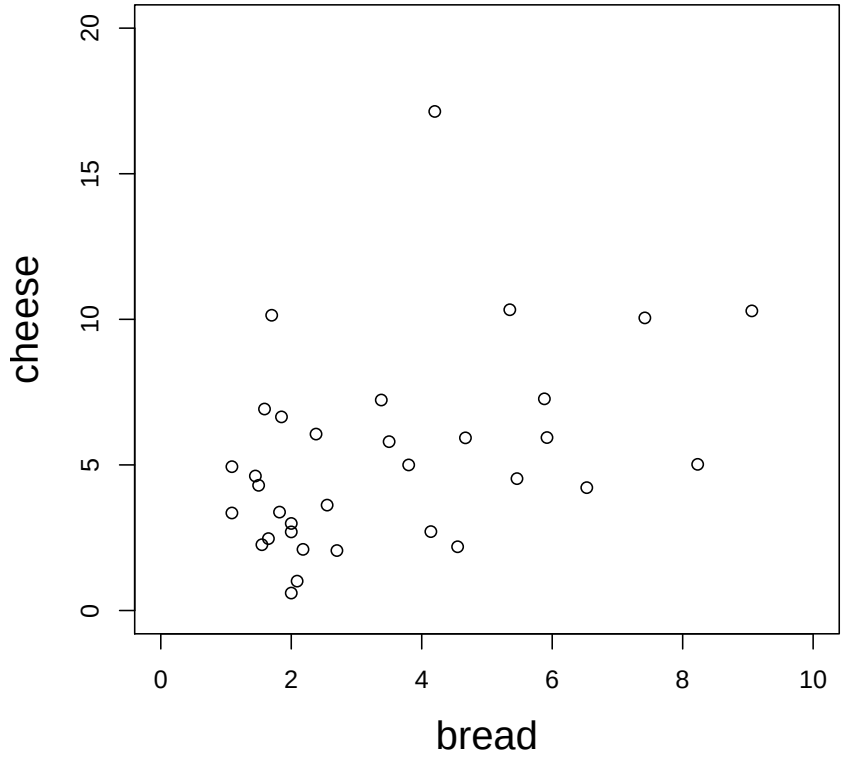
12.3 Correlation

Uncorrelated and independent are not the same.

- Two variables are called uncorrelated if $\mathrm{cor}(X, Y) = 0$.
- 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.

12.4 Examples

Expenditure for bread and cheese.

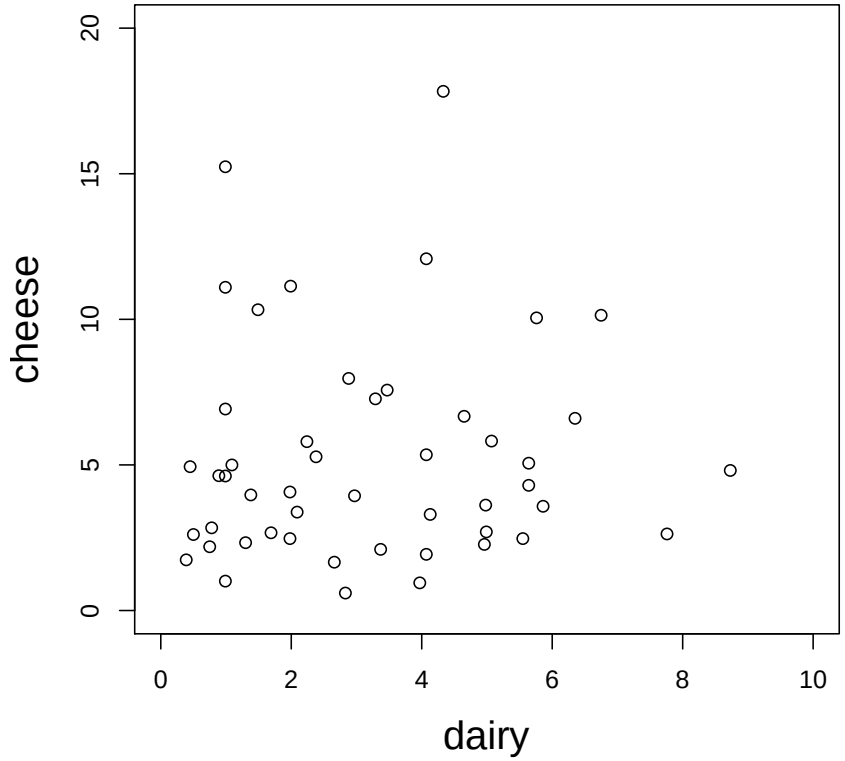


$$r = 0.41$$

Moderate positive correlation.

12.4 Examples

Expenditure for dairy products and cheese.

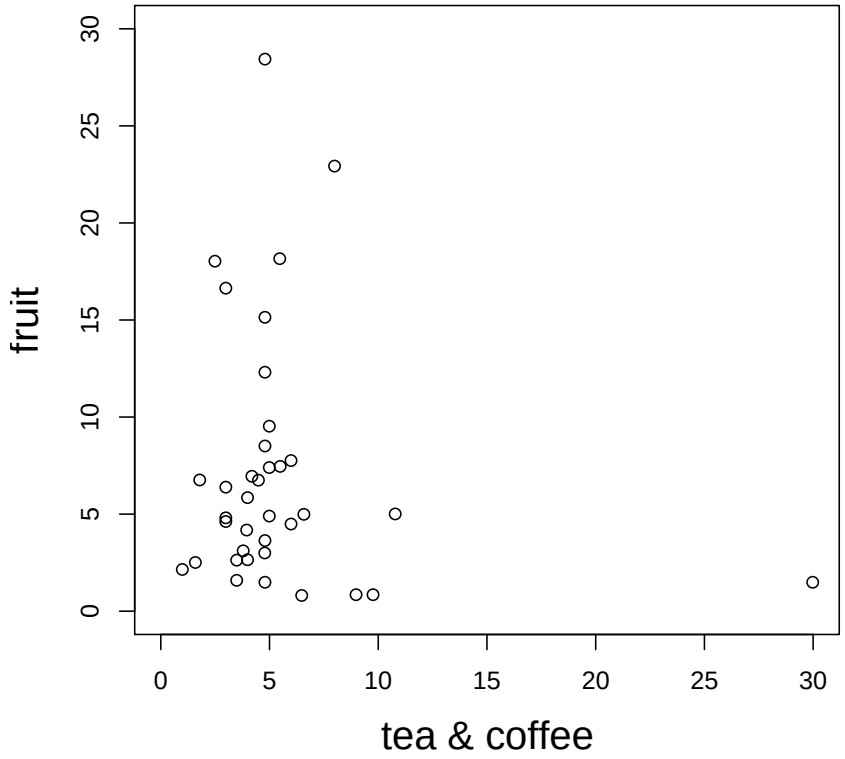


$$r = 0.05$$

Practically uncorrelated.

12.4 Examples

Expenditure for tea/coffee and fruit.

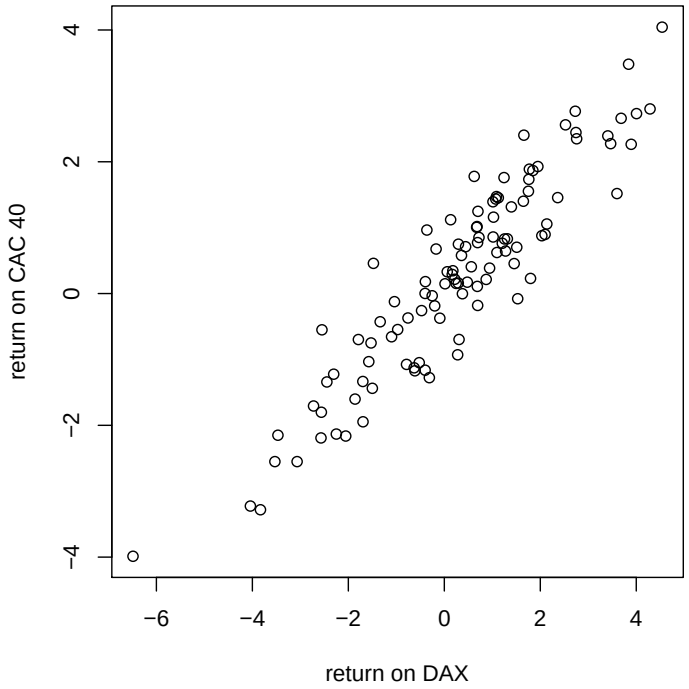


$$r = -0.13$$

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

12.4 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.

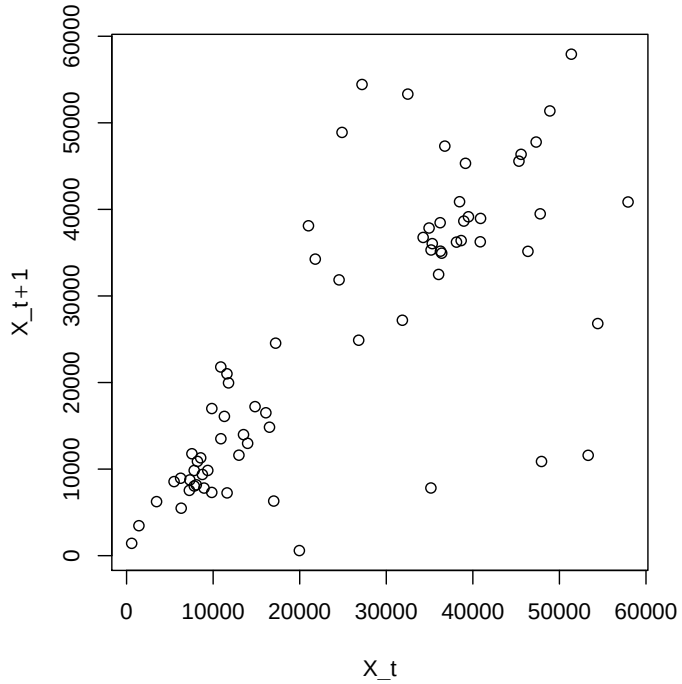
12.4 Examples

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 $(X_t) = X_1, X_2, X_3, \ldots$
- $\mathrm{cor}(X_t, X_{t+1})$ 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.

12.4 Examples

Example: Monthly car sales in Turkey.



$$r = 0.76$$

This high autocorrelation can be used for forecasting purposes.

12.5 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.