3. Measures of Relative Standing, Box Plots and Linear Regression

Percentile

The pth percentile of a set of measurements is a value for which:

- at least $p\%$ of the measurements are less or equal than that value
- at least $(100-p)\%$ of all the measurements are greater or equal than that value

Finding pth percentile

- Step 1. Sort the original data set: $X_1, \dots, X_n \to X_{(1)}, \dots, X_{(n)}$
- Step 2. Compute the value of the *locator*

$$L = \frac{p}{100} \times n$$

Here n is the sample size.

- Step 3. Computing the percentile now depends on the value of the locator.

a) If L is a whole number, then the percentile is given by

$$\frac{X_{(L)} + X_{(L+1)}}{2}.$$

b) If L is not a whole number, then round L *up* to $\lceil L \rceil$, and then the percentile is given by $X_{(\lceil L \rceil)}$.

Commonly used percentiles

- First (lower) decile = 10th percentile
- First (lower) quartile Q_1 = 25th percentile
- Second (middle) quartile Q_2 = 50th percentile
- Third quartile Q_3 = 75th percentile
- Ninth (upper) decile = 90th percentile

Box plots

- ***Interquartile Range:*** $\boldsymbol{IQR = Q_3 - Q_1}$
- ***Inner Fences:*** $[\boldsymbol{Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR}]$

Box plots

Box plot is a pictorial display that provides the main descriptive measures of the measurement set:

- S - The smallest measurement inside the inner fences
- Q_1 - The first quartile
- Q_2 - The second quartile or median
- Q_3 - The third quartile
- L - The largest measurement inside the inner fences

Outliers: boxplot criterion

- An *outlier* is an observation located at a distance of more than $1.5 \times IQR$ from the box, or the one which is outside the inner fences.

- Outliers are marked on the boxplot separately by stars.

Box plot - example

Example 1

Suppose that the return on investment for 21 companies in a certain industry for a certain year is

$$-24.6 \quad -2.6 \quad 2.4 \quad 2.7 \quad 3.8 \quad 5.6 \quad 5.9 \quad 6.7 \quad 7.0 \quad 7.2$$
$$7.5 \quad 8.0 \quad 8.2 \quad 8.5 \quad 8.6 \quad 8.8 \quad 9.0 \quad 9.2 \quad 9.7 \quad 10.0 \quad 20.5$$

Draw a boxplot of these data.

Box plot - example

Solution

With $n = 21$, median is the eleventh score, 7.5. The 25^{th} percentile is 5.6. The 75^{th} percentile is 8.8. Thus, $IQR = 8.8 - 5.6 = 3.2$.

The fences are:

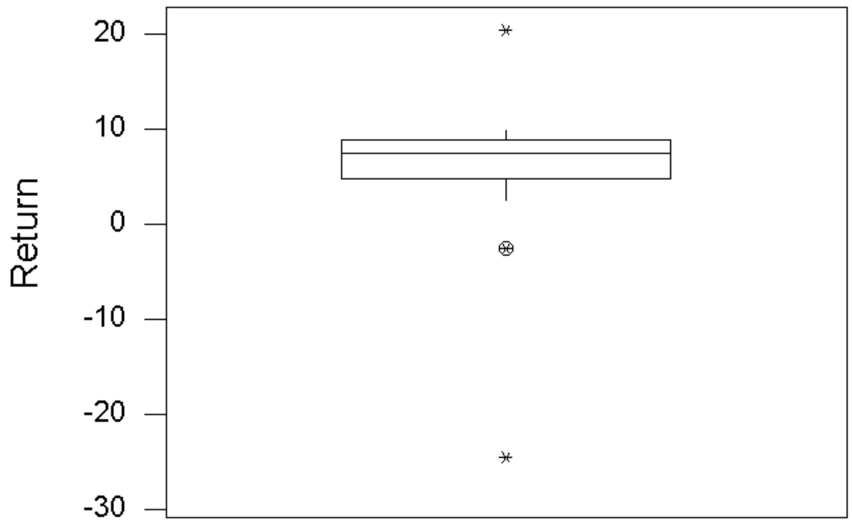
lower inner fence $= 5.6 - 1.5 \times 3.2 = .8$

upper inner fence $= 8.8 + 1.5 \times 3.2 = 13.6$

The fence test identifies three outliers, $-2.6, -24.6$ and 20.5. The smallest and largest non-outliers are 2.4 and 10.

Box plot - example

The box plot is shown below:



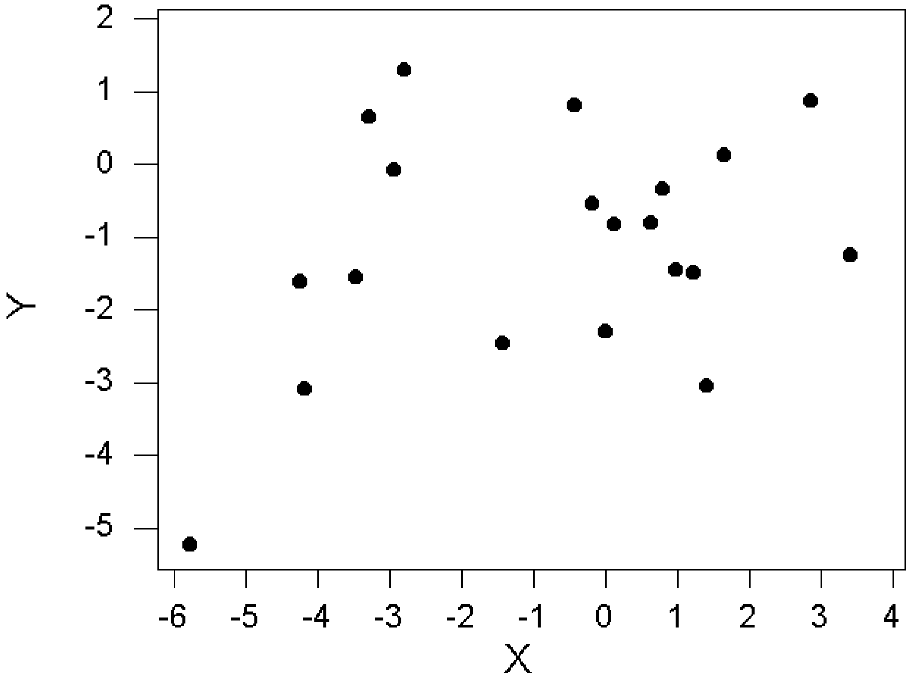
Scatterplot

Often we are interested in the relationships between two numerical variables. *Scatterplot* is a two-dimensional plot of one variable versus the other one.

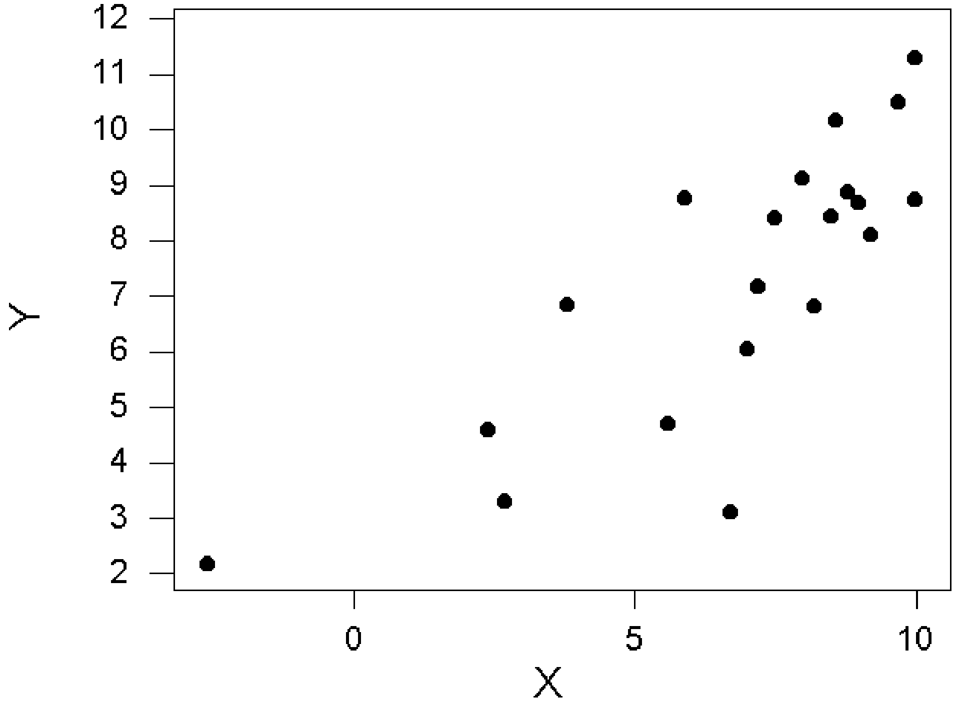
Typical Patterns

- No relationship
- Positive linear relationship
- Negative linear relationship
- Nonlinear (concave, convex) relationship

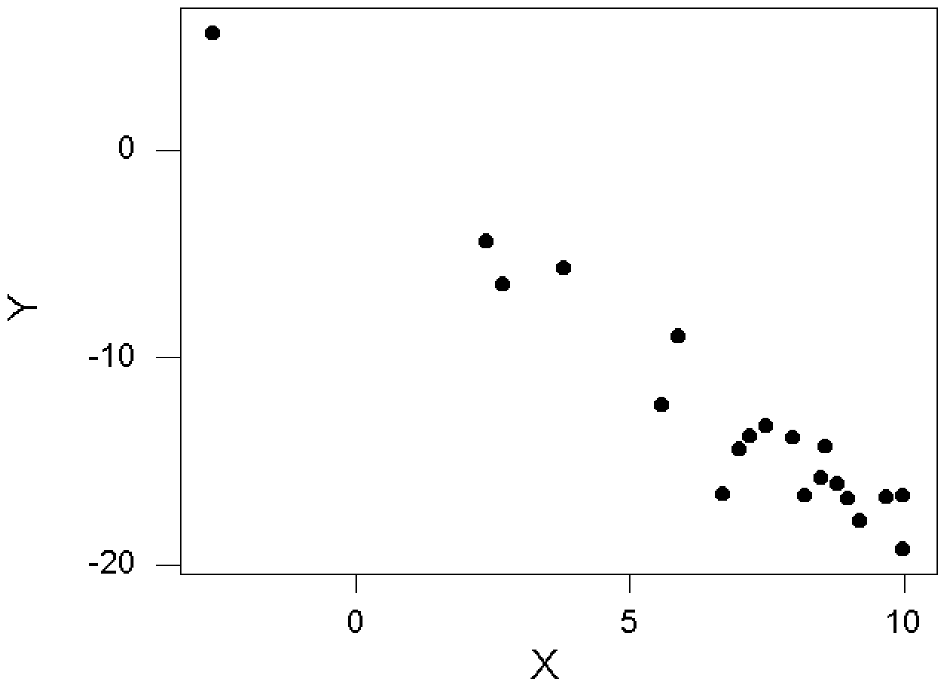
No relationship



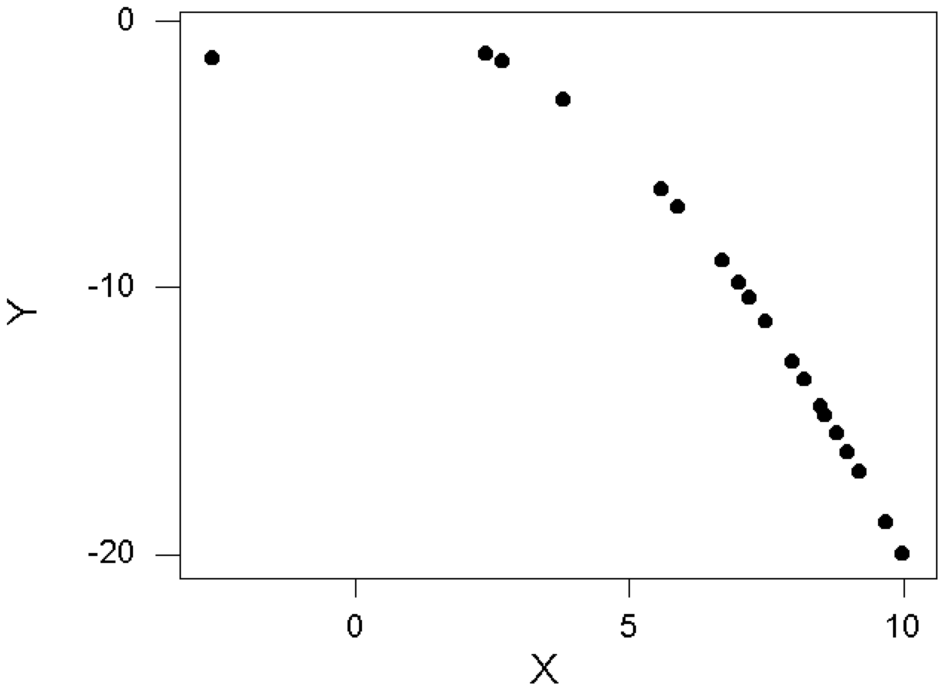
Positive linear relationship



Negative linear relationship



Nonlinear relationship



Correlation coefficient

Correlation coefficient is used for the description of linear relationship between two variables depicted in the scatterplot.

Correlation coefficient

Let us assume that we have two related data sets: $\boldsymbol{X_1, \dots, X_n}$ and $\boldsymbol{Y_1, \dots, Y_n}$.

Then the *sample correlation coefficient* is given by

$$\boldsymbol{r = \frac{1}{n-1}\left(Z_{X_1}Z_{Y_1} + \cdots + Z_{X_n}Z_{Y_n}\right)}$$

where $\boldsymbol{Z_{X_i}}$ and $\boldsymbol{Z_{Y_i}}$ are $\boldsymbol{z}$-scores of $\boldsymbol{X_i}$ and $\boldsymbol{Y_i}$, that is,

$$\boldsymbol{Z_{X_i} = \frac{X_i - \bar{X}}{s_X}}, \qquad \boldsymbol{Z_{Y_i} = \frac{Y_i - \bar{Y}}{s_Y}}$$

Correlation coefficient

One can show mathematically that r is always between -1 and 1.

- If two variables move in the same direction (both increase or both decrease), the correlation coefficient is positive. The closer to 1 the stronger the relationship.
- If two variables move in two opposite directions (one increases when the other one decreases), the correlation is negative. The closer to -1 the stronger the relationship.
- If two variables are unrelated, the correlation will be close to zero.

Regression line

The *regression line* is the straight line $y = mx + b$ that provides the best least-square fit to the data.

With help of calculus one can show that

$$m = \frac{n(X_1Y_1 + \cdots + X_nY_n) - (X_1 + \cdots + X_n)(Y_1 + \cdots + Y_n)}{n\left(X_1^2 + \cdots + X_n^2\right) - (X_1 + \cdots + X_n)^2}$$

and

$$b = \frac{(Y_1 + \cdots + Y_n) - m(X_1 + \cdots + X_n)}{n}$$

Regression line and correlation

The slope of regression line and correlation coefficient are related by the following formula.

$$r = m\frac{s_X}{s_Y}$$