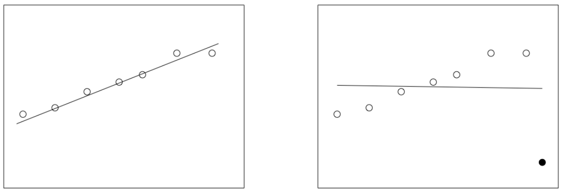
In [7]:
```
# settings↔
```

```
Populating the interactive namespace from numpy and matplotlib
```

2.7. Identification of Outliers

Outlier = Deviations from a model (ideal situation), measurement error

The presence of outliers in data can severely corrupt the results of analysis



- *Left*: Linear regression suggests increasing trend
- *Right*: Outlier causes linear regression to suggest a decreasing trend

2.7.1. Outlier Detection

Question: when do we have to regard a data point as outlier?

- requires comparison with expectation, for instance at hand of other data points, a model, or prior knowledge
  - For example: in the list of body sizes of students in [m]
  $$(1.78, 1.82, 1.68, 2.02, 1.72, 1.60, 172, 1.79, 1.85, 1.59, 1.94, \dots)$$
  - if you find the value 172, it is likely that the actual value should have been 1.72 and only a decimal point was lost, i.e. the value was entered in another unit [cm].
  - such a value can be easily spotted and corrected.
  - what helps is semantic knowledge / domain knowledge
- Often (in lack of such knowledge): expectation of a normal distribution of the data (after elimination of outliers)

z-Test

Let's assume we wish to detect outliers in the feature x_i of a d-dimensional data vector.

Estimate

- the expected value $\hat{\mu}_i$
- and the standard deviation $\hat{\sigma}_i$
- of data in the environment of each data point $\vec{x}$
  - this may require the choice of a suitable neighborhood
- and compute
$$z_i = \frac{|y_i - \hat{\mu}_i|}{\sigma_i}$$
- Data points with $z_i > C$ will be classified as outliers.
- C is an arbitrary threshold (e.g. standard is $C = 3$).
- Instead of using an heuristically chosen value for C:
  - $\rightarrow$ implicit derivation by assumung a small fraction $\alpha \ll 1$ of outliers, e.g. 1%.
- C results from isolation of the condition
$$1 - \int_{-C}^{C} P(s)ds = 2 \int_{C}^{\infty} P(s)ds = \alpha$$
where
$$P(s) = \frac{1}{\sqrt{2\pi}} \exp(-s^2/2)$$
is the normal density function with mean 0 and variance 1.
- as an improvement, when testing data point $\vec{x}_i^{\alpha}$, mean and sigma can be computed without including the checked 'potential outlier'.

[WS2018EOT1107]

- Z-scores can be misleading with small sample size N, because the maximum Z-score has an upper limit of $\frac{(N-1)}{\sqrt{N}}$.
- A remedy is to use the modified Z-score
$$M_i = 0.6745 \frac{x_i - q_{0.5}}{m_{AD}}$$
where $q_{0.5}$ is the median (i.e. the 50% quantile) and $m_{AD} = median(|x_i - q_{0.5}|)$ is the median absolute deviation.
  - source: B. Iglewicz and D. Hoaglin (1993), "Volume 16: How to Detect and Handle Outliers", The ASQC Basic References in Quality Control: Statistical Techniques, E. F. Mykytka, Editor.
- The authors recommend $M_i > 3.5$ as outlier criterion.

Problem:

- many outliers falsify the estimation of $\hat{\mu}_i, \hat{\sigma}_i^2$

⇒ **Possible Improvements:**

- replace means by more robust medians, or
- so-called **Roesner-Test**: as long as the z-test delivers at least an outlier:
  - remove only the most extreme outlier i, i.e.
$$z_i = \max_j z_j$$
from the data set
  - repeat recursively with the remaining data set.
- In case of multivariate data:
  - project data on selected axis and apply the z-test on the 1d-projections.
  - The choice of suitable axis can for instance be done by using Principal Component Analysis (see Sec. 3.3)
  - Example: Feature x: Age, y: number of children.
  - $x = 3$ is certainly not an outlier
  - $y = 2$ is certainly not an outlier
  - but: $(x, y) = (3, 2)$ is totally implausible
  - → univariate tests would not be capable to discover the outliers.

2.7.2 Outlier Management [...]

- What do we do if we detect an outlier?

→ a set of measures / actions (ranging from minimal invasive to drastic)

1. Marking:
   - data points remain in the data set, only creation of an additional mask M with
$$M_{\alpha k} = \begin{cases} 1 & \text{if} \quad x_k^\alpha \text{ valid} \\ 0 & \text{if} \quad x_k^\alpha \text{ invalid} \end{cases}$$
2. Correction:
   - replace the value by a plausible value:
   - → mean value of that feature
   - → e.g. kernel regression (using correct features as independent variables)
3. Removal of the component
4. Removal of the whole data vector