2 Deskriptive Statistik (2D)

- Zwei Messwerte werden beobachet
- Korrelationen zwischen Daten finden

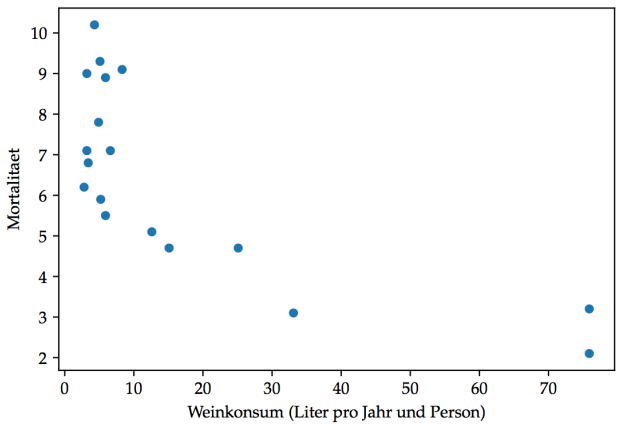
Streudiagramm

- Zwei Messungen als Koordinaten von Punkten dargestellt
- Kausalitäten können anhand Streudiagramme erahnt werden
- Beweist jedoch keinen kausalen Zusammenhang

python

```
data = DataFrame({
    "a": ([2.8, 3.2, 3.2, ..., 75.9]),
    "b": ([6.2, 9.0, 7.1, ..., 2.1])
})

data.plot(kind="scatter", x="a", y="b")
```



Einfache lineare Regression

- lineare Abhängigkeit (z.B. Zusammenhang Seitenzahl x und Buchpreis y → dickere Bücher sind teurer)
- Regressionsgerade: $y = a + bx$

Methode der kleinsten Quadrate

- In Streudiagramm kann man keine Gerade durch alle Punkte ziehen
- **Residuum**: vertikale Differenz zw. Beobachtungspunkt (x_i, y_i) und Gerade (Punkt auf Gerade = $(x_i, a + bx_i)$):

> $$\text{Residuum} = r_i = y_i - (a + bx_i) = y_i - a - bx_i$$

- **Ziel**: Summe der Residuen möglichst klein,

> $$r_1 + r_2 + \ldots + r_n = \sum_i r_i$$

- **Schwäche**: Wenn Hälfte der Punkte weit über und andere Hälfte weit unter der Geraden, ist Summe der Residuen null. Gerade passt aber trotzdem nicht zu Datenpunkten
- **Lösung**: Vorzeichen eliminieren.
  - Option 1: Absolutbetrag verwenden (schwierig abzuleiten)
  - Option 2: Quadrate der Residuen aufsummieren,
  - $r_1^2 + r_2^2 + \ldots + r_n^2 = \sum_i r_i^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$
- python: `b, a = np.polyfit(book["pages"], book["price"], deg=1)`

> **Minimum** (wird aber immer mit Python gemacht):
>
> $$\frac{\delta}{\delta\alpha} \sum_{i=1}^{n} (y_i - (a + bx_i))^2 \doteq 0$$
> $$\frac{\delta}{\delta\beta} \sum_{i=1}^{n} (y_i - (a + bx_i))^2 \doteq 0$$

Die Regressiongerade ist nicht angebracht wenn,

- Die Punkte keiner Gesetzmässigkeit folgen
- Die Punkte einer nichtlinearen Gesetzmässigkeit folgen

Empirische Korrelation

- numerische Zusammenfassung der linearen Abhängigkeit zweier Grössen
- Kennzahl = r oder $\hat{\rho}$
- misst Stärke und Richtung der linearen Abhängigkeit zw. x und y
- **steigende Gerade**: $r = +1$
- **fallende Gerade**: $r = -1$
- **keine Abhängigkeit**: $r = 0$
- python: `data.corr()`

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2) \cdot (\sum_{i=1}^{n}(y_i - \bar{y})^2)}}$$