Einführende Worte zur Regression und Korrelation

Die Regression

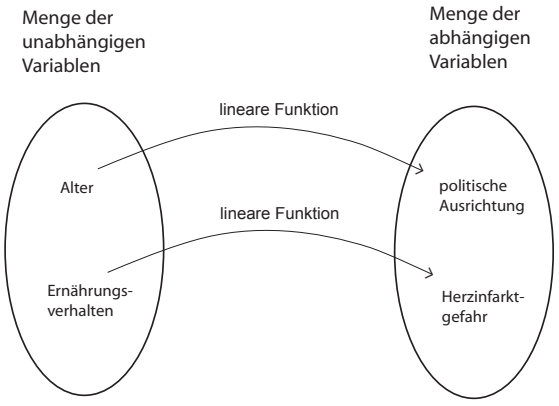
In der Politik, Wirtschaft, Medizin oder einfach im alltäglichen Leben ist man am Erkennen von Zusammenhängen interessiert. 2 Beispiele sind

- Zusammenhang zwischen Alter und politischer Ausrichtung
- Inwieweit prägt das Ernährungsverhalten die Herzinfarktgefahr.

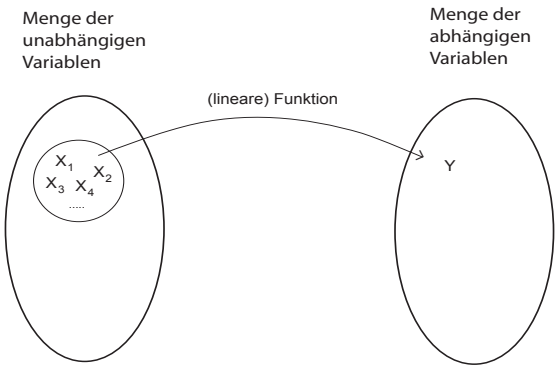
Eine in vielen (aber nicht allen) Situationen anwendbare Methode zur Untersuchung der Zusammenhänge bieten Regressionsmodelle.

Hier unterscheidet man zwischen

- **Einfaches, lineares Regressionsmodell**: Man versucht eine (abhängige) Variable mittels einer linearen Funktion durch eine andere (unabhängige) Variable zu erklären.

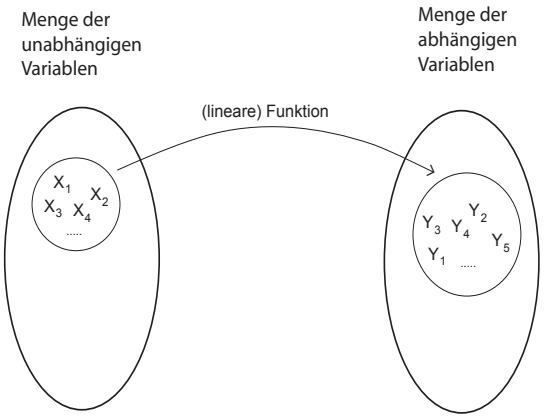


- **Einfaches, nichtlineares Regressionsmodell**: Die "erklärende" Funktion ist nichtlinear.
- **Multiples Regressionsmodell**: Man hat mehr als eine unabhängige Variable, d.h. man untersucht inwieweit eine abhängige Variable Y von mehreren unabhängigen Variablen X_i $i = 1, 2, \ldots, k$ beeinflusst wird.



Auch hier unterscheidet man zwischen einem *linearen* und einem *nichtlinearen* Modell

- **Multivariates Regressionsmodell**: Man hat zusätzlich mehrere abhängige Variablen Y_j, $j = 1, \ldots, l$.



Die Korrelation

Die Korrelation ist ein **Maß** für den linearen Zusammenhang,

- im Falle einer linearen einfachen Regression zwischen der abhängigen Variable (üblicherweise Y genannt) und der unabhängigen Variable (X).

  Die Frage ist also ... wie stark ist der lineare Zusammenhang zweier Variablen?

  Maßzahlen dafür sind primär

  - der *Korrelationskoeffizient* r, $0 \leq r \leq 1$, wobei Werte nahe bei 0 für einen schwachen Zusammenhang sprechen, Werte nahe bei 1 auf einen stark direkt proportionalen Zusammenhang deuten bzw. Werte nahe bei -1 für einen starke umgekehrt proportionale Beziehung sprechen
  - das *Bestimmtheitsmaß* (das Quadrat des Korrelationskoeffizienten), welchen man als den Quotienten

    $$\frac{\text{“durch die Regression erklärte Varianz von } Y\text{”}}{\text{“Gesamtvarianz von } Y\text{”}}$$

    interpretieren kann,

  wobei für das Bestimmheitsmaß r^2 gilt

  - $0 \leq r^2 \leq 1$
  - die Erklärungskraft der Regression ist umso größer, je näher r^2 bei 1 liegt.

- im Falle einer multiplen linearen Regression zwischen der abhängigen Variable Y und den unabhängigen Variablen $X_1, \ldots, X_k$.

  Auch hier existiert ein Maß für die Beurteilung der Güte der Schätzung, das **multiple Bestimmtheitsmaß**, auf welches an dieser Stelle nicht weiter eingegangen wird.