Korrelation und Regression

Wilfried Mann,
Mettmann

Korrelation und Regression beschreiben in der Statistik die Beziehung zwischen zwei (oder mehreren) Variablen; entweder »ungerichtet« (Korrelation) oder »gerichtet« (Regression).

1. Grundsätze

Korrelation und Regression beschreiben die Beziehung zwischen Variablen. Die folgende Darstellung macht die Unterschiede zwischen »ungerichteten« und »gerichteten« Zusammenhängen bei einfachen linearen Beziehungen deutlich.

Der Zusammenhang zwischen Variablen, Merkmalen, Ereignissen, Zuständen oder Funktionen wird mit verschiedenen Zusammenhangs- (Assoziations-) maßen gemessen. Ein bedeutsames Assoziationsmaß ist der dimensionslose Korrelationskoeffizient von Pearson (r) zwischen **zwei** mindestens intervallskalierten Merkmalen. Das Quadrat des Korrelationskoeffizienten stellt das Bestimmtheitsmaß (B oder r^2) dar.

Eine Korrelation (Wechselwirkung) beschreibt jedoch keine Ursache-/Wirkung-Beziehung in die eine und/oder andere Richtung, d.h. aus einem starken Zusammenhang folgt nicht, dass es auch eine eindeutige Ursache-Wirkungs-Beziehung gibt.

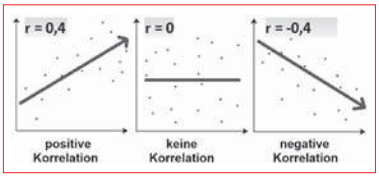
2. Korrelation (ungerichtet)

Da stets der Zusammenhang zwischen **zwei** Variablen linear untersucht wird, wird von einem »bivariaten Zusammenhang« gesprochen.

Diese können wie folgt unterschiedlich ausgeprägt sein:

Gleichsinnige oder positive Korrelation: Hohe (tiefe) Ausprägungen der einen Variablen gehen mit hohen (tiefen) Ausprägungen der zweiten Variablen einher. Zum Beispiel: Je höher der Rohertrag, desto höher ist der Verkehrswert. Je geringer der Rohertrag, desto niedriger ist der Verkehrswert.

Gegenläufige oder negative Korrelation: Hohe Werte der einen Variablen gehen mit tiefen Werten der anderen einher. Zum Beispiel: Je höher der Liegenschaftszinssatz, desto geringer ist der Verkehrswert. Je niedriger der Liegenschaftszinssatz, desto höher ist der Verkehrswert.



Das Korrelationsmaß, der Pearson'sche Korrelationskoeffizient (r), errechnet sich nach der Vorschrift:

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

wobei r alle Werte zwischen -1 und +1 annehmen kann.

Ist r = 0, dann besteht kein Zusammenhang zwischen X und Y.

Ist r = 1, dann besteht ein vollständiger positiver (+) oder negativer (-) Zusammenhang.
Bedingungen sind: stetige Variable (mindestens intervallskaliert), lineare Zusammenhänge, Normalverteilung, keine Ausreißer.

2. Regression (gerichtet)

Die Regression gibt einen Zusammenhang zwischen zwei oder mehr Variablen an. Es wird vorausgesetzt, dass es einen gerichteten linearen Zusammenhang gibt, das heißt, es existieren eine abhängige Variable und mindestens eine unabhängige Variable. Welche Variablen abhängig und welche unabhängig sind, muss aufgrund inhaltlich logischer Überlegungen festgelegt werden, bzw. ergibt sich aus einer konkreten Aufgabenstellung. Folgende Assoziationsmaße können berechnet werden:

Bestimmtheitsmaß (R^2 oder B),
als erklärter Anteil der Varianz einer abhängigen Variablen Y durch ein statistisches Modell (mit X).

Multipler Korrelationskoeffizient (R oder r),
als Wurzel aus dem multiplen Bestimmtheitsmaß.

Korrigiertes R^2 (adj R^2),
berücksichtigt die wirksamen Einflussgrößen stärker und ist immer kleiner als das R^2.

Partielles R^2 (part R^2),
als erklärter Anteil der Varianz einer abhängigen Variablen Y durch eine unabhängige Einflussgröße (X), wobei: Summe part $R^2 = R^2$.

Die Bedingungen zur Ableitung von R^2 entsprechen denen der ungerichteten Korrelation. Im Rahmen einer Regressionsanalyse können quantitative Zusammenhänge mit Hilfe einer Regressionsfunktion (auch statistisches Modell) beschrieben werden. Hierbei sind die Trennung von Signal (Funktion) und Rauschen (Störgröße) sowie die Abschätzung des dabei gemachten Fehlers zu beachten.

3. Anwendung

Korrelationen geben Aufschluss über wechselseitige Beziehungen. Sind diese statistisch »zufälligen« Aussagen aber auch Kausalitäten? Ist also die Wirkung (Reaktion) wirklich eine Folge einer vorausgegangenen Ursache (Aktion)? Dieser Fragen muss sich der Statistiker und auch Sachverständige ehrlich stellen.

Für die Datenanalyse sind Korrelationen zwischen Ziel und Einflussgrößen erwünscht, zwischen den Einflussgrößen aber unerwünscht.

Bei der Datenanalyse von Kaufpreisen (Regressionsanalyse) ist es häufig schwierig, Korrelationen zwischen Einflussgrößen auszuschließen. In Innenstadtlagen werden mehr Büro-/Geschäftshäuser zu finden sein als in städtischen Randbereichen. Somit wirken Effekte aus dem Gebäudetyp auch auf die Lagezuordnung. Korrelationskoeffizienten unter 0,3 können bei großen Kaufpreisstichproben noch als normal gelten. Bei kleinen Stichproben hingegen, sollte dem Sachverstand bei der Beurteilung ein hohes Gewicht zugeordnet werden.