4. Merkmalszusammenhänge

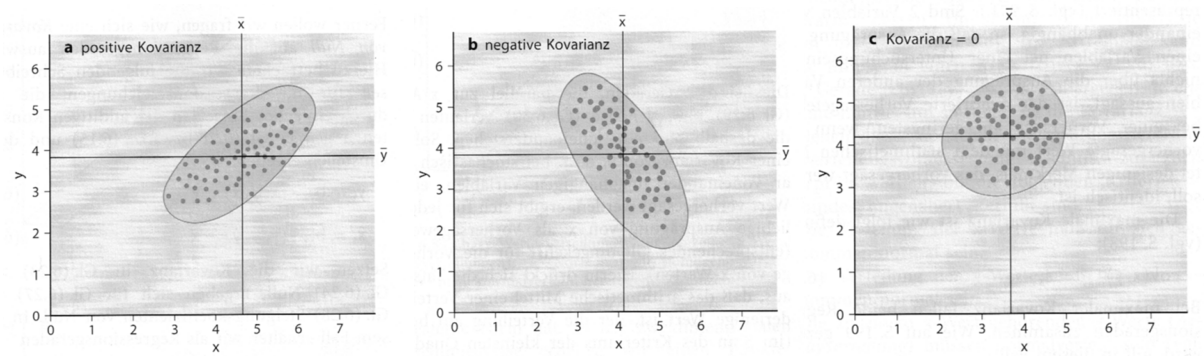
Die meisten Hypothesen über einen empirischen Sachverhalt beinhalten offen oder verdeckt formulierte Annahmen über *Kausalbeziehungen*. Das Aufdecken solcher Kausalzusammenhänge erlaubt uns, über das bloße Beschreiben der phänomenologischen Umwelt hinauszugehen und Erklärungen für empirische Sacherverhalte anzubieten. Die Kenntnis von Zusammenhängen ermöglicht überdies Vorhersagen über künftige Ereignisse.

4.1 Kovarianz und Korrelation

Der Grad des (*nicht-kausalen*) Zusammenhangs zwischen zwei intervallskalierten Variablen lässt sich mathematisch durch die Kovarianz und die auf ihr aufbauende Produkt-Moment-Korrelation beschreiben.

4.1.1 Der Begriff des Zusammenhangs

Ein Zusammenhang kann in zwei „Richtungen“ vorliegen: positiv oder negativ. Wenn, wie im obigen Beispiel, hohe Werte auf der einen Variable hohen Werten auf der anderen entsprechen und niedrige Werte auf der einen Variable niedrigen auf der anderen, so ist der Zusammenhang positiv. Gehen dagegen hohe Werte auf der einen Variable mit niedrigen Werten auf der anderen einher und umgekehrt, so liegt ein negativer Zusammenhang vor.



4.1.2 Die Kovarianz

Die folgende Formel zeigt, dass die Kovarianz im Gegensatz zur Varianz Aussagen über die gemeinsame Variation zweier Merkmale macht:

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

Die Kovarianz ist das durchschnittliche Produkt aller korrespondierenden Abweichungen der Messwerte von den Mittelwerten der beiden Merkmale x und y.

Für jedes Wertepaar wird die Abweichung des x-Werts vom Mittelwert der x-Werte mit der Abweichung des y-Werts vom Mittelwert der y-Werte multipliziert. Die Summe der einzelnen Abweichungsprodukte wird als Kreuzproduktsumme zweier Variablen bezeichnet. Diese Kreuzproduktsumme wird über alle Beobachtungen gemittelt. Allerdings wird analog zur Varianz im Nenner durch n-1 geteilt, um einen erwartungstreuen Schätzer der Populationskovariation zu erhalten.

Eine positive Kovarianz resultiert, wenn die beiden Variablen weitgehend gemeinsam in die gleiche Richtung von ihrem Mittelwert abweichen, d.h. positive Abweichungen der einen Variable werden mit positiven Abweichungen der anderen multipliziert, bzw. negative mit negativen. Der Zusammenhang ist positiv.

Dagegen ergibt sich eine negative Kovarianz, wenn viele entgegengesetzt gerichtete Abweichungen vom jeweiligen Mittelwert auftreten, d.h. eine positive Abweichung auf der einen Variable korrespondiert mit einer negativen Abweichung auf der anderen und umgekehrt. Die Kreuzproduktsumme und somit auch die Kovarianz werden negativ. Die Merkmale weisen einen negativen oder inversen Zusammenhang auf.

Sind die Abweichungen mal gleich, mal entgegengesetzt gerichtet, so heben sich die Abweichungsprodukte gegenseitig auf und es resultiert eine Kovarianz nahe Null. In diesem Fall besteht kein systematischer Zusammenhang zwischen den Variablen x und y. Die Ausprägung des Merkmals x sagt also nichts über die Ausprägung des Merkmals y aus.

Der Betrag der maximalen Kovarianz ist für positive wie auch negative Zusammenhänge identisch. Er ist definiert als das Produkt der beiden Merkmalststreuungen:

$$\left|\text{cov}(\max) = \hat{\sigma}_x \cdot \hat{\sigma}_y\right|$$

Die Kovarianz ist also kein standardisiertes Maß und folglich zur quantitativen Kennzeichnung des Zusammenhangs zweier Merkmale nur bedingt geeignet. Sie kann allerdings in ein standardisiertes Maß überführt werden: den Korrelationskoeffizienten.

4.1.3 Die Produkt-Moment-Korrelation

Die Produkt-Moment-Korrelation nach Pearson ist das gebräuchlichste Maß für die Stärke des Zusammenhangs zweier Variablen. Sie drückt sich aus im Korelationskoeffizienten r. Er stellt die Standardisierung der im vorherigen Abschnitt behandelten Kovarianz dar. Dabei wird die empirisch ermittelte Kovarianz an der maximalen Kovarianz relativiert.

$$r_{xy} = \frac{\text{cov}_{emp}}{\text{cov}_{\max}} = \frac{\text{cov}(x, y)}{\hat{\sigma}_x \cdot \hat{\sigma}_y}$$

Die Formel gibt zu erkennen, dass der Korrelationskoeffizient niemals größer als 1 oder kleiner als -1 werden kann, denn die empirisch gefundene Kovarianz kann die maximal mögliche Kovarianz zwischen den beiden Variablen in ihrem Wert nicht übersteigen. Der Wertebereich der Korrelation ist somit im Gegensatz zu dem der Kovarianz begrenzt zwischen -1 und +1.

Eine Umwandlung der Formel der Korrelation ist sehr aufschlussreich:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{(n-1) \cdot \hat{\sigma}_x \cdot \hat{\sigma}_y} = \frac{1}{n-1} \cdot \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\hat{\sigma}_x} \cdot \frac{y_i - \bar{y}}{\hat{\sigma}_y}\right)$$

Die Quotienten in der Klammer entsprechen der Formel für z-Standardisierung. Die z-Standardisierung übernimmt dabei die Funktion, die unterschiedlichen Streuungen der beiden Verteilungen aus der Kovarianz heraus zu rechnen. Die Korrelation ist also im Grunde genommen nichts anderes als die Kovarianz zweier z-standardisierter Variablen mit den Mittelwerten 0 und der Streuung 1:

$$r = \frac{\sum_{i=1}^{n}(z_{xi} - 0) \cdot (z_{yi} - 0)}{n-1} = \mathrm{cov}(z_{xi}, z_{yi})$$

Der Korrelationskoeffizient r macht den Anschein, als wäre er als Prozentmaß des Zusammenhangs zu verstehen, etwa dergestalt, dass eine Korrelation von 0,8 einen doppelt so hohen Zusammenhang beschreibt wie eine Korrelation von 0,4. Aussagen diesen Typs sind aber mit Korrelationswerten nicht zulässig, da die hierfür erforderliche Äquidistanz nicht gegeben ist.

Exkurs: Korrelation und Kausalität

Wie zu Beginn des Kapitels 4 betont, sagt eine Korrelation noch nichts über zugrunde liegende Ursache-Wirkungs-Beziehungen zwischen den beteiligten Merkmalen aus. Nicht immer ist klar, in welche Richtung die Kausalität verläuft.

Natürlich ist auch eine hohe Korrelation kein Garant dafür, dass überhaupt ein direkter ursächlicher Zusammenhang zwischen den untersuchten Merkmalen besteht. Beide Variablen x und y können von einer dritten gemeinsamen Ursache abhängen. Dieses Phänomen wird als Scheinkorrelation bezeichnet.

4.1.4 Die Fishers Z-Transformation

Liegen zu einem untersuchten Merkmalszusammenhang mehrere Ergebnisse (aus verschiedenen Untersuchungen) in Form von Korrelationen vor, so ist es sinnvoll, einen Mittelwert aus den Ergebnissen zu bilden. Ein Mittelwert aus mehreren Korrelationen ist aber stets mit einem Fehler behaftet, da Korrelationskoeffizienten nicht intervallskaliert sind.

Diesem Problem schafft die Fishers Z-Transformation Abhilfe. Sie ist unter keinen Umständen mit der z-Standardisierung zu verwechseln. Die Aufgabe dieser Transformation ist es, Korrelationen in annähernd intervallskalierte Werte zu überführen, so dass die Bildung des arithmetischen Mittels zulässig ist. Dazu sind drei Schritte notwendig:

1. Transformation der einzelnen Korrelationen in Fishers Z-Werte
2. Bildung des arithmetischen Mittels der Fishers Z-Werte
3. Rücktransformation des arithmetischen Mittels der Fishers Z-Werte in eine Korrelation

Die Berechnungsvorschrift zur Transformation der Korrelationen in Fishers Z-Werte lautet:

$$Z = \frac{1}{2} \cdot \ln\left(\frac{1+r}{1-r}\right)$$

Die Rücktransformation des Mittelwerts der Fishers Z-Werte in eine mittlere Korrelation folgt der Berechnungsvorschrift:

$$\bar{r} = \frac{e^{2\cdot\bar{Z}} - 1}{e^{2\cdot\bar{Z}} + 1}$$

(Anmerkung zur Software: In Excel können die Transformationen mit den Befehlen „Fishers“ und „FisherINV“ durchgeführt werden.)

Fishers Z-Werte sind zwar annähernd intervallskaliert, ihr Wertebereich ist im Gegensatz zu dem der Korrelation aber nicht begrenzt (er geht gegen unendlich). Somit stellen auch sie kein prozentuales Maß für den Zusammenhang zweier Variablen dar. Noch einmal: Korrelationen sind nicht äquidistant, Unterschiede können streng genommen nur als Größer-Kleiner-Relationen interpretiert werden. Fishers Z-Werte dagegen sind nahezu äquidistant, es lassen sich Mittelwerte bilden und die Größe von Abständen interpretieren.

Die Fishers Z-Transformation eignet sich neben der Produkt.Moment-Korrelation auch für zwei weitere Korrelationskoeffizienten, nämlich die punktbiseriale Korrelation und die Rangkorrelation (vgl. Kap. 4.2).

4.1.5 Signifikanz von Korrelationen

Auch die Korrelation lässt sich einem Signifikanztest unterziehen. Dieser verläuft analog zum t-Test mit einem Unterschied: Der Stichprobenkennwert der Testverteilung besteht aus der Korrelation zweier Stichproben, und nicht aus einer Mittelwertsdifferenz. Die Nullhypothese des Signifikanztests für Korrelationen besagt, dass eine empirisch ermittelte Korrelation r zweier Variablen aus einer Grundgesamtheit stammt, in der eine Korrelation ρ („Rho“) von Null besteht.

Der t-Wert aus der empirischen Korrelation r und dem Stichprobenumfang N lässt sich wie folgt berechnen:

$$t_{df} = \frac{r \cdot \sqrt{N-2}}{\sqrt{1-r^2}} \text{ mit df = N - 2}$$

Für den Signifikanztest gilt, dass gegen ein vorher festgelegtes Fehlerniveau α bzw. gegen einen kritischen t-Wert getestet wird. Übertrifft der empirische t-Wert diese Grenzmarke, so ist die Korrelation statistisch signifikant. Die Nullhypothese wird abgelehnt, die Alternativhypothese angenommen.

4.1.6 Konfidenzintervall für eine Korrelation

Die Bestimmung des Konfidenzintervalls für eine Korrelation läuft analog zu der Bestimmung beim Mittelwert ab. Dabei ist zu beachten, dass Korrelationen in ihrer Grundverteilung nicht normalverteilt sind. Sie lassen sich jedoch durch die Fishers Z-Transformation annähernd in eine Normalverteilung überführen. Unter Zuhilfenahme der Formel für die Standardabweichung von Fishers Z-Werten lässt sich so ein symmetrisches Konfidenzintervall um den zugehörigen Fishers Z-Wert der Korrelation bilden. Die ermittelten Grenzwerte können anschließend per Rücktransformation in r-äquivalente Grenzwerte überführt werden. Diese sind dann natürlich nicht mehr symmetrisch um r angeordnet.

4.1.7 Effektstärke

Da eine Korrelation an den Streuungen der beteiligten Variablen standardisiert ist, kann man die Korrelation r als ein Effektstärkenmaß interpretieren. Allerdings empfiehlt es sich aufgrund der fehlenden Äquidistanz zur besseren Vergleichbarkeit Fishers Z-transformierte Korrelationen zu verwenden.

Ein alternatives Effektstärkenmaß ist der so genannte Determinationskoeffizient r^2. Allerdings geht durch die Quadrierung die Information über die Richtung des Zusammenhangs (positiver

vs. negativer Zusammenhang) verloren. Der Determinationskoeffizient steht für den Anteil der Varianz einer Variable, der durch die Varianz der anderen Variable aufgeklärt wird. In diesem Sinne ist r^2 auch als Effektstärkenmaß einer Korrelation zu verstehen. Je mehr Varianz die beiden untersuchten Variablen gemeinsam haben, je stärker sie also kovariieren, desto größer ist der Effekt. Im Gegensatz zu r liefert der Determinationskoeffizient r^2 intervallskalierte Werde und darf als Prozentmaß interpretiert werden.
Liefert eine empirische Untersuchung einen Korrelationskoeffizienten von r = 0,50 zwischen den Variablen A und B, so liegt der Determinationskoeffizient bei $r^2 = 0,25$. Dieser Wert ist so zu interpretieren, dass die Variable B 25% der Varianz von Variable A aufklärt. 75% der Varianz werden durch andere Faktoren verursacht.

4.1.8 Teststärkeanalyse

Wie wir aus dem vorangegangenen Abschnitt bereits wissen, ist r bzw. r^2 bereits als Effektstärke interpretierbar. Die Bestimmung der Teststärke kann entweder mit Hilfe des Nonzentralitätsparameters λ (und den dazugehörigen TPF-Tabellen) erfolgen, oder bequem mit Hilfe von GPower. GPower gibt hier allerdings den Nonzentralitätsparameter δ an. δ ist die Wurzel aus λ. Der Nonzentralitätsparameter berechnet sich dabei wie folgt:

$$\lambda_\alpha = \frac{r^2}{1-r^2} \cdot N$$

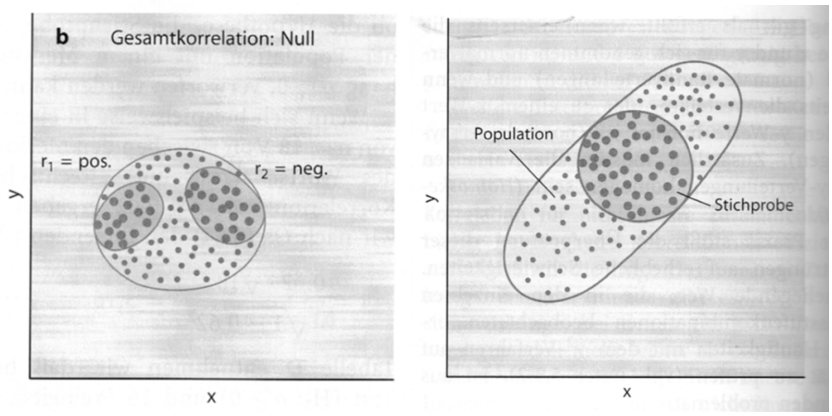
4.1.9 Stichprobenumfangplanung

Um zu ermitteln, wie viele Versuchspersonen nötig sind, um mit einer bestimmten Teststärke einen a priori angenommenen Effekt r bzw. r^2 zu entdecken, wird die obige Formel nach N aufgelöst:

$$N = \frac{\lambda_{\alpha;Teststärke}}{\frac{r^2}{1-r^2}}$$

4.1.10 Stichprobenfehler

Bei der Rekrutierung einer Stichprobe ist zwingend darauf zu achten, dass sie für die interessierende Population repräsentativ ist.



4.1.11 Die Partialkorrelation

Eine Einsatzmöglichkeit der Partialkorrelation ist es, den „versteckten“ Einfluss einer dritten Variablen auf die Merkmale x und y herauszufiltern und somit einen „wahren“ Zusammenhang zwischen den beiden eigentlich interessierenden Variablen bei eventuell vorhandenen Scheinkorrelationen aufzudecken. Allerdings ist dafür natürlich auch die Erfassung dieser Drittvariable für jede Beobachtungseinheit notwenig. In der Sprache der Statistik sagt man auch, dass die Drittvariable z aus x und y heraus partialisiert wird. Daher der Name dieser Korrelationstechnik. Weitere Bezeichnungen lauten bedingte Korrelation oder Korrelation erster Ordnung, im Gegensatz zur oben dargestellten bivariaten Korrelation nullter Ordnung.

$$r_{xy|z} = \frac{r_{xy} - r_{yz} \cdot r_{xz}}{\sqrt{\left(1 - r_{yz}^2\right) \cdot \left(1 - r_{xz}^2\right)}}$$

$r_{xy|z}$: Partialkorrelation der beiden ersten interessierenden Merkmale, diejenige Variable, die auspartialisiert wird, wird mit einem senkrechten Strich, einem Punkt oder einem Komma im Index der Korrelation abgetrennt.

$r_{xy|z}$: Korrelation nullter Ordnung der beiden interessierenden Merkmale

$r_{xy|z}$: Korrelation von x und y mit der Drittvariablen z

Die Signifikanz einer Partialkorrelation kann über den t-Test mit Hilfe folgender Prüfgröße beurteilt werden:

$$t_{df} = r_{xy|z} \cdot \sqrt{\frac{N-2}{1-r_{xy|z}^2}}$$ Die Freiheitsgrade ergeben sich zu df = N-3

Als zweiter Fall ist denkbar, dass die Partialkorrelation größer ist als die Korrelation nullter Ordnung. Dieser Fall tritt ein, wenn die Drittvariable z mit einer der beiden Variablen, sagen wir x, unkorreliert, mit der anderen Variable y dagegen hoch korreliert ist. In diesem Fall ist der Zähler der Formel zur Berechnung der Partialkorrelation gleich der Korrelation r_{xy}. Im Nenner steht jedoch ein Wert kleiner als Eins, so dass die resultierende Partialkorrelation gegenüber r_{xy} erhöht ist. Die Variable z wird in diesem Fall auch als Suppressorvariable bezeichnet, da sie den wahren Zusammenhang zwischen x und y „unterdrückt". Wird z aus y heraus partialisiert, wird y um einen für den Zusammenhang mit x irrelevanten Varianzanteil bereinigt. Dadurch steigt der Anteil gemeinsamer Varianz zwischen x und y an der verbliebenen Varianz von y.

Drittens kann es sein, dass Korrelation und Partialkorrelation sich überhaupt nicht unterscheiden, dann nämlich, wenn die vermeintliche Drittvariable mit beiden interessierenden Variablen x und y unkorreliert ist. In diesem Fall reduziert sich die Formel der Partialkorrelation zur Korrelation zwischen x und y:

$$r_{xy|z} = \frac{r_{xy} - r_{yz} \cdot r_{xz}}{\sqrt{\left(1-r_{yz}^2\right) \cdot \left(1-r_{xz}^2\right)}} = \frac{r_{xy} - 0}{(1-0) \cdot (1-0)} = r_{xy}$$

4.2 Weitere Korrelationstechniken

Der Produkt-Moment-Korrelationskoeffizient wird bei zwei intervallskalierten Variablen berechnet. Nun wissen wir aber, dass Variablenwerte auch andere Skalenebenen repräsentieren können. Für diese verschiedenen Ebenen und deren Kombination bei zwei verschieden skalierten Variablen gibt es weitere Möglichkeiten der Bestimmung ihres Zusammenhangs.

4.2.1 Die punktbiseriale Korrelation

Die punktbiseriale Korrelation ist das geeignete Verfahren, um den Zusammenhang zwischen einem intervallskalierten und einem dichotomen, nominalskalierten Merkmal zu bestimmen. Ein Merkmal ist dann dichotom, wenn es in genau zwei Ausprägungen auftreten kann.
Die Formel lautet:

$$r_{pb} = \frac{\bar{y}_1 - \bar{y}_0}{\hat{\sigma}_y} \cdot \sqrt{\frac{n_0 \cdot n_1}{N^2}}$$

x : dichotome Variable in den Ausprägungen x_0 und x_1 (nicht in der Formel)

y : intervallskalierte Variable

$\bar{y}_o$: Mittelwert der y-Werte in x_0

$\bar{y}_1$: Mittelwert der y-Werte in x_0

n_0 : Stichprobengröße in x_0

n_1 : Stichprobengröße in x_1

N : $n_0 + n_1$ (Anzahl aller Untersuchungseinheiten)

$\hat{\sigma}_y$: geschätzte Populationsstreuung aller y-Werte

Die Formel ist so konzipiert, dass eine positive Korrelation dann resultiert, wenn die y-Werte unter x_0 im Durchschnitt kleiner sind als die y-Werte unter x_1, d.h. die Merkmalsausprägung nimmt von x_0 nach x_1 zu. Entsprechend resultiert eine negative punktbiseriale Korrelation, wenn die durchschnittlichen Merkmalsausprägungen der y-Werte in x_0 über den Ausprägungen in x_1 liegen.

Der zugehörige Signifikanztest erfolgt wie bei der Produkt-Moment-Korrelation über die t-Verteilung mit der Formel:

$$t_{df} = \frac{r_{pb} \cdot \sqrt{N-2}}{\sqrt{1-r_{pb}^2}} \text{ mit df = N-2}$$

Punktbiseriale Korrelation und t-Test

Konzeptuell entsprechen sich punktbiseriale Korrelation und t-Test, mit nur einer Ausnahme: Korrelationen erfassen Zusammenhänge, der t-Test untersucht Mittelwertsunterschiede. Beide Konzepte sind direkt ineinander überführbar.

4.2.2 Die Rangkorrelation

Zur Berechnung der Korrelation zweier ordinalskalierter Merkmale bietet sich die Rangkorrelation nach Spearman an. Der Rangkorrelationskoeffizient r_s stellt eine Analogie zur Produkt-Moment-Korrelation dar, wobei an Stelle intervallskalierter Messwerte die jeweiligen Rangplätze der ordinalskalierten Daten eingesetzt werden. Die Rangkorrelation erfasst, inwieweit zwei Rangreihen systematisch miteinander variieren.

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{N \cdot (N^2 - 1)}$$

d_i stellt die Differenz der Rangplätze einer Untersuchungseinheit i bezüglich der Variablen x und y dar, N die Anzahl der Untersuchungseinheiten.

Für $n \geq 30$ kann der Rangkorrelationskoeffizient durch den folgenden t-Test näherungsweise auf Signifikanz überprüft werden. Wieder hat der Test df = N-2 Freiheitsgrade.

$$t_{df} = \frac{r_s \cdot \sqrt{N-2}}{\sqrt{1-r_s^2}}$$

Sollen eine intervallskalierte und eine ordinale Variable miteinander korreliert werden, so kann ebenfalls die Rangkorrelation verwendet werden. Hierzu ist es nötig, die Verteilung des intervallskalierten Merkmals als lediglich rangskalierte Daten zu betrachten. In diesem Sinne erfährt diese Variable eine Herabstufung des Skalenniveaus. Die Rangkorrelation liefert eine Abschätzung des Zusammenhangs dieser beiden Variablen, wobei in einer Variablen auf vorhandene Information verzichtet wird. Man sollte dieses Vorgehen nur im Notfall verwenden.

	Intervallskala	Rangskala	Nominalskala (dichotom)
Intervallskala	Produkt-Moment-Korrelation	Rangkorrelation	Punktbiseriale Korrelation
Rangskala		Rangkorrelation	Punktbiseriale Korrelation
Nominalskala (dichotom)			Phi-Koeffizient

4.3 Einfache lineare Regression

Der Zusammenhang zweier Variablen lässt sich nach den bisherigen Kenntnissen durch die Korrelation mathematisch beschreiben. In den empirischen Sozialwissenschaften ist es oftmals darüber hinaus von Interesse, auch Vorhersagen über die Ausprägung von Variablen zu machen. Genau das leistet die Regression. Liegen für zwei Merkmale x und y eine Reihe von Wertpaaren vor, so lässt sich aufgrund dieser Daten eine Funktion zur Vorhersage von y aus x bestimmen. Diese Funktion heißt Regressionsgleichung. Dabei ist x die unabhängige Variable, genannt Prädiktor, und y die gesuchte abhängige Variable, das Kriterium. Schon hier wird deutlich, dass bei der Regression zwischen unabhängiger und abhängiger Variable unterschieden wird (also eine kausale Beziehung nahe gelegt wird), wohingegen bei der oben besprochenen Korrelation die beiden Merkmale gewissermaßen gleichberechtigt nebeneinander stehen und allein der Zusammenhang zwischen ihnen von Interesse ist.
Bei „einfachen" Regressionen werden lediglich ein Prädiktor und ein Kriterium verwendet, bei der multiplen Regression werden mehrere Prädiktoren benutzt.

4.3.1 Die Regressionsgerade

Stochastische Zusammenhänge sind unvollkommene Zusammenhänge, die sich graphisch in einer Punktewolke zeigen. Je höher der tatsächliche Zusammenhang ist, desto enger wird die Punktewolke. Bei maximalem Zusammenhang geht die Punktewolke schließlich in eine Gerade über (für jeden x-Wert lässt sich ein y-Wert ablesen). In diesem rein theoretischen Fall liegt ein funktionaler Zusammenhang vor.

4.3.2 Berechnung der Regressionsgleichung

Die Steigung der Regressionsgeraden wird mit b bezeichnet und heißt Regressionsgewicht, die Höhenlage wird mit a bezeichnet. Die Variable y der Funktion wird als Schätzer mit einem Dach ($\hat{y}$) gekennzeichnet, da hypothetische Werte vorhergesagt werden, die nicht unbedingt mit den tatsächlichen Werten übereinstimmen.
Somit lautet die allgemeine Regressionsgleichung:

$$\hat{y} = b \cdot x + a$$

Die Differenz $y_i - \hat{y}_i$ gibt allgemein für jede Versuchsperson an, wie stark ihr wahrer Wert von dem durch die Gerade vorhergesagten Wert abweicht. Die optimale Gerade, die diesen Punkteschwarm am besten wiedergibt, ist diejenige, bei der über alle Versuchspersonen hinweg dieser Vorhersagefehler am Kleinsten ist. Hierzu wird das Kriterium der kleinsten Quadrate genutzt: Die Gerade ist so zu legen, dass die Summe der Quadrate aller

Abweichungen der empirischen y-Werte von den vorhergesagten y-Werten möglichst klein wird. Die Quadrierung hat im Vergleich zu nicht quadrierten Werten den Vorteil, dass sie inhaltlich bedeutsamere Abweichungen stärker berücksichtigt. Zusätzlich fallen Irritationen durch unterschiedliche Vorzeichen weg.

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \min$$

Die beiden „idealen“ Parameter b (Steigung) und a (Höhenlage) der Regressionsgerade lassen sich wie folgt ermitteln:

$$b_{xy} = \frac{\mathrm{cov}(x, y)}{\sigma_x^2} \qquad a_{xy} = \bar{y} - b_{yx} \cdot \bar{x}$$

Die Indizierung der Parameter a und b mit yx besagt, dass die y-Werte aus den x-Werten vorhergesagt werden. Der umgekehrte Fall, nämlich die Vorhersage der x-Werte aus den y-Werten, ist ebenso möglich. Hier sind das Regressionsgewicht b und der y-Achsenabschnitt a entsprechend definiert als:

$$b_{xy} = \frac{\mathrm{cov}(x, y)}{\sigma_y^2} \qquad a_{xy} = \bar{x} - b_{yx} \cdot \bar{y}$$

Die beiden Regressionsgeraden weisen stets einen gemeinsamen Schnittpunkt auf, dessen Koordinaten mit dem Mittelwert der beiden Merkmalsverteilungen identisch sind.

Ein wesentlicher Vorteil der Regressionsanalyse ist die Möglichkeit einer Vorhersage: Ist die Regressionsgleichung zwischen zwei Variablen bekannt, so lässt sich zu einem beliebigen Wert der Prädiktorvariable der zugehörige Kriteriumswert prognostizieren. Dabei ist unbedingt zu beachten, dass es sich hier nicht um eine kausale sondern um eine statistische Prädiktion (um eine Berechnung) handelt! Dass Körpergröße und Körpergewicht zusammen hängen (korrelieren), ist offenkundig. Die errechnete zugehörige Regressionsgerade ermöglicht es nun, zu einem Gewicht von z.B. 85 kg einen Wert für die Körpergröße zu ermitteln, ohne dass es eine Vpn geben muss, die 85 kg schwer gewesen ist. Umgekehrt kann man entsprechend aus einer Größe von 165 cm das Körpergewicht ermitteln, wenn man eine empirisch bestimmte Regressionsgerade zugrunde legt. Das heißt natürlich noch lange nicht, dass die Körpergröße kausal ursächlich für das Gewicht oder das Gewicht kausal ursächlich für die Körpergröße ist.

4.3.3 Wichtige Einsichten und Zusammenhänge

Eine stochastische Unabhängigkeit von Variablen zeigt sich graphisch darin, dass die Geraden senkrecht aufeinander stehen. Mit wachsendem Zusammenhang wird der Betrag der Kovarianz größer. Entsprechend verkleinert sich der Winkel zwischen den Geraden immer mehr. Bei maximaler Kovarianz bzw. perfektem Zusammenhang fallen die beiden Geraden schließlich zusammen.

Regression und z-standardisierte Variablen

Liegen die beiden Merkmale x und y in z-standardisierter Form vor, so haben beiden Verteilungen den Mittelwert 0 und eine Streuung von 1. Es wurde schon zuvor deutlich, dass die Kovarianz zweier z-standardisierter Variablen gleich deren Korrelation ist. Das bedeutet für die Regression, dass die Steigung der Regressionsgerade mit der Korrelation der beiden Variablen identisch ist:

$$b_{xy} = \frac{\mathrm{cov}(Z_x, Z_y)}{1} = r_{xy}$$

4.3.4 Regressionsgewichte

Bleibt die ursprüngliche Maßeinheit erhalten, so wird b als unstandardisiertes Regressionsgewicht bezeichnet.

$$b_{yx} = \frac{Anzahl\_der\_Einheiten\_auf\_y}{pro\_1\_Einheit\_auf\_x}$$

In vielen Fällen ist es jedoch vorteilhaft, die Regressionsgewichte verschiedener Regressionsgleichungen miteinander vergleichen zu können. Um eine einheitliche Metrik für derartige Vergleiche zu erhalten, muss das unstandardisierte Regressionsgewicht von der Originalmetrik der untersuchten Merkmale bereinigt werden, indem b in den Zählereinheiten wie auch in den Nennereinheiten an der Streuung der jeweiligen Merkmale relativiert wird. Das resultierende standardisierte Regressionsgewicht wird mit β (auch „beta-Gewicht“) bezeichnet und errechnet sich wie folgt:

$$\beta_{xy} = \frac{\frac{Anzahl\_Einheiten\_auf\_y}{\sigma_y}}{\frac{1\_Einheit\_auf\_x}{\sigma_x}} = b \cdot \frac{\frac{1}{\sigma_y}}{\frac{1}{\sigma_x}} = b \cdot \frac{\sigma_x}{\sigma_y}$$

Der standardisierte Regressionskoeffizient β ist von den Maßeinheiten der untersuchten Merkmale unabhängig und drückt aus, um wie viele Standardabweichungseinheiten sich y verändert, wenn sich x um eine Standardabweichung vergrößert.
Im Fall der einfachen Regression (d.h. ein Prädiktor, ein Kriterium) ist β außerdem identisch mit der Produkt-Moment-Korrelation zwischen den beiden Merkmalen:

$$b_{xy} = b \cdot \frac{\sigma_x}{\sigma_y} = \frac{\mathrm{cov}(x,y)}{\sigma_x^2} \cdot \frac{\sigma_x}{\sigma_y} = \frac{\mathrm{cov}(x,y)}{\sigma_x \cdot \sigma_y} = r_{xy}$$

4.3.5 Die Güte des Zusammenhangs

Liegt ein unvollständiger Zusammenhang vor, so sind die Vorhersagen der Regressionsgeraden mit einem Fehler behaftet. Das ist der Preis dafür, das eine nicht optimale Messwerteverteilung (Punkteschwarm) in eine exakte Funktion (Gerade) transformiert wird. Dieser Fehler zeigt sich darin, dass die vorhergesagten Werte in der Mehrzahl der Fälle nicht mit den empirischen Daten übereinstimmen. Das Ausmaß dieser Abweichungen ist ein Indikator dafür, wie exakt die Regression in ihrer Vorhersage ist. Dieses Gütemaß der Regressionsvorhersage heißt Standardschätzfehler.
An der Genauigkeit der Vorhersage zeigt sich, inwieweit zwei Merkmale funktional miteinander verknüpft sind. Je erfolgreicher sich die Regressionsgleichung zur Vorhersage eignet, umso größer muss der tatsächliche Zusammenhang zwischen den Merkmalen sein. Zusätzlich zur Kovarianz und Korrelation lässt sich daher der Determinationskoeffizient als Maß für die Güte einer Vorhersage ableiten.
Bei der Regression gibt es für jeden Messwert y_i drei Arten von Abweichungen:
1. Jeder y-Wert weicht von seinem Mittelwert $\bar{y}$ ab. Daraus lässt sich die geschätzte *Populations*- oder *Gesamtvarianz* bestimmen:

$$\hat{\sigma}_y^2 = \frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{N-1}$$

2. Auch die von der Regressionsgeraden vorhergesagten $\hat{y}$-Werte weichen von ihrem Mittelwert $\bar{y}$ ab. Diese Abweichungen ergeben die *Regressionsvarianz*, also diejenige Varianz, die unter den vorhergesagten y-Werten besteht:

$$\hat{\sigma}_{\hat{y}}^2 = \frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2}{N-1}$$

3. Es besteht eine Diskrepanz zwischen den empirischen und den prognostizierten Werten. Diese Abweichungen können als „Fehler“ interpretiert werden, die die Regressionsgerade bei der Vorhersage macht. Zusammen ergeben sie die *Fehler*- oder *Residualvarianz*:

$$\hat{\sigma}^2_{[y/x]} = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N-1}$$

Diese drei Varianzen stehen in einer einfachen mathematischen Beziehung zueinander. Die Gesamtvarianz setzt sich additiv zusammen aus der Regressionsvarianz und der Residualvarianz:

$$\hat{\sigma}^2_y = \hat{\sigma}^2_{\hat{y}} + \hat{\sigma}^2_{[y/x]}$$

Der Standardschätzfehler

Der Standardschätzfehler wird als ein Maß für die Güte einer Regression verwendet. Per definitionem bildet er die Wurzel aus der Residualvarianz:

$$\hat{\sigma}_{[y/x]} = \sqrt{\hat{\sigma}^2_{[y/x]}}$$

Der Standardschätzfehler gibt an, wie stark die empirischen y-Werte durchschnittlich um die von der Regressionsgeraden vorhergesagten Werte streuen. Je kleiner der Standardschätzfehler, umso genauer und zuverlässiger ist die Vorhersage. Er ist kein standardisiertes Maß, seine Größe ist vom gewählten Erhebungsmaß abhängig.

Der Determinationskoeffizient

Ein wesentlich aussagekräftigeres Gütemaß als der Standardschätzfehler ist der Determinationskoeffizient r^2. Er wird durch eine Relativierung der Regressionsvarianz an der Gesamtvarianz gebildet:

$$r^2 = \frac{\hat{\sigma}^2_{\hat{y}}}{\hat{\sigma}^2_y} = \frac{\hat{\sigma}^2_{\hat{y}}}{\hat{\sigma}^2_{\hat{y}} + \hat{\sigma}^2_{[y/x]}}$$

Der Quotient drückt den Anteil der Regressionsvarianz an der Gesamtvarianz aus. Multipliziert mit 100 gibt der Determinationskoeffizient an, wie viel Prozent der gesamten Varianz durch die Regression, also durch die gemeinsame Varianz der x- und y-Werte erklärbar ist. Er ist demnach ein sehr anschauliches Maß für den Zusammenhang zweier Variablen. Ein Determinationskoeffizient von 0,6 bedeutet beispielsweise, dass 60% der Varianz der y-Werte durch die Kenntnis der Prädiktorvariable x aufgeklärt werden können.

Der Determinationskoeffizient trägt nicht ohne Grund die Bezeichnung r^2. Er ist als das Quadrat des Korrelationskoeffizienten r definiert. Somit schließt sich an dieser Stelle der Kreis zwischen Korrelation und Regression. Beide Maße geben Auskunft über den Zusammenhang zweier Merkmale, wobei der Determinationskoeffizient die anschaulichere Größe darstellt.

Im Grunde genommen ist der Determinationskoeffizient nichts anderes als ein Effektstärkenmaß. Er gibt an, welcher Anteil der Variabilität der abhängigen Variable durch die unabhängige Variable aufgeklärt wird.

4.3.6 Vorraussetzungen der linearen Regression

Abschließend sollen die wichtigsten Voraussetzungen für die Durchführung einer Regressionsanalyse angeführt werden.

- Das Kriterium muss intervallskaliert und normalverteilt sein.
- Der Prädiktor kann entweder intervallskaliert und normalverteilt sein, oder dichotom nominalskaliert.
- Die Einzelwerte verschiedener Versuchspersonen müssen voneinander unabhängig zustande gekommen sein.
- Der Zusammenhang der Variable muss theoretisch linear sein.
- Die Streuung der zu einem x-Wert gehörenden y-Werte müssen über den ganzen Wertebereich von x homogen sein (Annahme der Homoskedastizität).