Sozialwissenschaftliche Fakultät

Methodenmodul B.MZS.13: Multivariate Analysemodelle

Skript zu den Lerneinheiten

Teil 1: Drittvariablenkontrolle
in der Tabellenanalyse und im linearen Regressionsmodell

Lerneinheit 1: Bivariate Tabellenanalyse (Wiederholung)

Ziel der bivariaten Tabellenanalyse ist die statistische Untersuchung eines möglichen Zusammenhangs zwischen zwei Variablen.
Dazu werden in der gemeinsamen Verteilung der beiden Variablen die (relativen) Auftretenshäufigkeiten der Ausprägungskombinationen der beiden Variablen betrachtet. Wenn X die Spaltenvariable mit den Ausprägungen $x_1, x_2, \ldots, x_j, \ldots, x_J$ und Y die Zeilenvariablen mit den Ausprägungen $y_1, y_2, \ldots, y_j, \ldots, y_I$ bezeichnet, ergibt sich folgender Tabellenaufbau:

	Spaltenvariable X						
Zeilenvariable Y	x_1	x_2	…	x_j	…	x_J	Σ
y_1							
y_2							
…							
y_i							
…							
y_I							
Σ							

Die inneren Zellen der Tabelle enthalten die gemeinsamen (relativen) Häufigkeiten der Ausprägungskombination (y_i,x_j) der beiden Variablen X und Y, das ist die ***bivariate Verteilung***.
In der rechten Randspalte und der unteren Randzeile werden die Summen der Zeilen bzw. Spalten aufsummiert. Entsprechend enthält die rechte Randspalte die univariate Verteilung der Zeilenvariable Y und die untere Randzeile die univariate Verteilung der Spaltenvariable X.

Symbolik in Tabellen

	Spaltenvariable X						
Zeilenvariable Y	x_1	x_2	…	x_j	…	x_J	Σ
y_1	n_{11}						
y_2							n_{2+1}
…							
y_i							
…							
y_I						p_{IJ}	
Σ				p_{+j}			

Die univariaten Verteilungen werden in der Tabellenanalyse daher auch als ***Randverteilungen*** bezeichnet.
Wenn die Tabelle sich auf Populationsdaten bezieht, werden die absoluten Häufigkeiten in einer beliebigen Tabellenzelle (i,j) durch „N_{ij}“ symbolisiert und die relativen Häufigkeiten durch „π_{ij}“; liegen Stichprobendaten vor, werden die Symbole „n_{ij}“ für die absoluten und „p_{ij}“ für die relativen Häufigkeiten verwendet.

Hinweis:
Da im Englischen der Buchstabe „f“ für „frequencies“ steht, werden die absoluten Häufigkeiten auch oft durch F_{ij} in der Population und f_{ij} in der Stichprobe symbolisiert.

Schätzungen der Populationswerte werden meist durch ein „^“ über dem Symbol gekennzeichnet, also $\hat{N}_{ij}$ für die absoluten und $\hat{\pi}_{ij}$ für die relativen Häufigkeiten.

Die Analyse eines Zusammenhangs in einer Kreuztabelle

In einem Zufallsexperiment sind zwei Ereignisse statistisch unabhängig voneinander, wenn die Wahrscheinlichkeit des gemeinsamen Auftretens gleich dem Produkt des Auftretens der einzelnen Ereignisse ist.
Entsprechend sind zwei Zufallsvariablen X und Y ***statistisch unabhängig*** voneinander, wenn die bivariate Verteilung das Produkt der univariaten Verteilungen ist:

$$\pi_{ij} = \pi_{i+} \cdot \pi_{+j} \text{ für alle i und j.}$$

Statistische Abhängigkeit zwischen zwei Variablen wird dann als Abwesenheit von Unabhängigkeit definiert und liegt somit vor, wenn für mindestens einige Auftretenswahrscheinlichkeiten gilt:

$$\pi_{ij} \neq \pi_{i+} \cdot \pi_{+j} \text{ für alle oder einige i und j.}$$

Stichprobenverteilung der Zellenhäufigkeiten

In einer einfachen Zufallsauswahl sind die absoluten wie die relativen Häufigkeiten konsistente und erwartungstreue sowie asymptotisch normalverteilte Schätzer der entsprechenden Populationshäufigkeiten bzw. der bivariaten Wahrscheinlichkeitsverteilung von zwei Zufallsvariablen.

Bei einer einfachen Zufallsauswahl des Umfangs n mit Zurücklegen bzw. bei einer unbegrenzt großen Population ist die exakte Verteilung jeder Zellenhäufigkeit n_{ij} eine Binomialverteilung:

$$n_{ij} \sim b\left(n_{ij}; n, \pi_{ij}\right) = \binom{n}{n_{ij}} \cdot \pi_{ij}^{n_{ij}} \cdot \left(1-\pi_{ij}\right)^{n-n_{ij}}$$

Stichprobenverteilung der Zellenhäufigkeiten

Bei endlichem Populationsumfang N und einfacher Zufallsauswahl ohne Zurücklegen ist jede Zellenhäufigkeit dagegen hypergeometrisch verteilt:

$$n_{ij} \sim h\left(n_{ij}; n, N, N_{ij}\right) = \binom{N_{ij}}{n_{ij}} \cdot \binom{N-N_{ij}}{n-n_{ij}} / \binom{N}{n}$$

Bei hinreichend großem Stichprobenumfang n kann die Stichprobenverteilung der absoluten Häufigkeit in einer Zelle auch durch eine Poisson-Verteilung berechnet werden:

$$n_{ij} \sim p\left(n_{ij}; \lambda_{ij} = n \cdot \pi_{ij}\right) = \frac{\left(n \cdot \pi_{ij}\right)^{n_{ij}} \cdot e^{-n \cdot \pi_{ij}}}{n_{ij}!}$$

Der Vorteil der Poisson-Verteilung gegenüber der Binomialverteilung bzw. der hypergeometrischen Verteilung ist, dass die einzelnen Zellenhäufigkeiten – gegeben die Verteilungsparameter λ_{ij} – voneinander statistisch unabhängig sind. Da sich die relativen Populationsanteile π_{ij} (bzw. N_{ij}/N) über alle Zellen zu 1.0 aufsummieren müssen, sind die Häufigkeiten bei einer Modellierung über die Binomialverteilugn bzw. die hypergeometrische Verteilung dagegen nicht unabhängig voneinander.

Die Unabhängigkeit der Poissonverteilungen wird genutzt um eine Teststatistik zu finden, die bei einer einfachen Zufallsauswahl die statistische Unabhängigkeit der Spalten- und Zeilenvariable voneinander prüfen kann.

Chiquadrat-Test auf statistische Unabhängigkeit von Zeilen- und Spaltenvariable

Dazu wird die asymptotische Annäherung der Poisson-Verteilung an eine Normalverteilung mit den Parameter $\mu = \lambda$ und $\sigma^2 = \lambda$ genutzt. Die Standardisierung dieser asymptotischen Normalverteilung ergibt für jede Tabellenzelle eine asymptotisch standardnormalverteilte Variable:

$$\frac{n_{ij} - n \cdot \pi_{ij}}{\sqrt{n \cdot \pi_{ij}}} = \frac{n_{ij} - \lambda_{ij}}{\sqrt{\lambda_{ij}}} \underset{n_{ij} \to \infty}{\sim} N(0;1)$$

Da die Summe von quadrierten Standardnormalverteilungen chiquadrat-verteilt ist, folgt für die Aufsummierung aller quadrierten Standardnormalverteilungen:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{n_{ij} - n \cdot \pi_{ij}}{\sqrt{n \cdot \pi_{ij}}}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{ij} - n \cdot \pi_{ij}\right)^2}{n \cdot \pi_{ij}} = \chi^2 \underset{n_{ij} \to \infty}{\sim} \chi^2_{df}$$

Bei statistischer Unabhängigkeit von Zeilen- und Spaltenvariable in der Population gilt dann entsprechend:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{ij} - n \cdot \pi_{i+} \cdot \pi_{+j}\right)^2}{n \cdot \pi_{i+} \cdot \pi_{+j}} = \chi^2 \underset{n_{ij} \to \infty}{\sim} \chi^2_{df}$$

Wenn die λ-Parameter der Poisson-Verteilungen unbekannt sind, müssen sie aus der Stichprobe geschätzt werden.

Chiquadrat-Test auf statistische Unabhängigkeit von Zeilen- und Spaltenvariable

Bei statistischer Unabhängigkeit in der Population werden dazu die relativen Häufigkeiten der Randverteilungen als Schätzungen der entsprechenden Populationsanteile verwendet:

$$\hat{\lambda}_{ij} = n \cdot \hat{\pi}_{ij} \rightarrow \text{ wenn } H_0\text{: } n \cdot \hat{\pi}_{i+} \cdot \hat{\pi}_{+j} = n \cdot p_{i+} \cdot p_{+j} = n \cdot \frac{n_{i+}}{n} \cdot \frac{n_{+j}}{n} = \frac{n_{i+} \cdot n_{+j}}{n}$$

Die Teststatistik berechnet sich dann nach:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{ij} - \frac{n_{+i} \cdot n_{+j}}{n}\right)^2}{\frac{n_{i+} \cdot n_{+j}}{n}} = n \cdot \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(p_{ij} - p_{i+} \cdot p_{+j}\right)^2}{p_{i+} \cdot p_{+j}}$$

Um die Teststatistik anzuwenden, muss die Zahl der Freiheitsgrade bekannt sein. Diese ergibt sich aus der Zahl der quadrierten asymptotischen Standardnormalverteilungen minus der Zahl der für die Schätzung der λ_{ij}-Parameter verwendeten empirischen Informationen:

- Die Zahl der quadrierten asymptotischen Standardnormalverteilungen ist gleich der Zahl der inneren Zellen der Kreuztabelle also I×J.
- Die Zahl der verwendeten empirischen Informationen ist gleich (I–1) + (J–1) +1, da aus der Randverteilung der Zeilenvariable I–1 relative Häufigkeiten p_{i+} zur Schätzung der relativen Populationshäufigkeiten π_{i+} und J–1 relative Häufigkeiten p_{+j} zur Schätzung der relativen Populationshäufigkeiten π_{+j} sowie die Fallzahl n benötigt werden. Dass jeweils I–1 statt I und J–1 statt J Anteile benötigt werden, liegt daran, dass sich die relativen Häufigkeiten einer Verteilung jeweils zu 1.0 summieren.

Chiquadrat-Test auf statistische Unabhängigkeit von Zeilen- und Spaltenvariable

Die Zahl der Freiheitsgrade beträgt daher

$$df = I \times J - (I-1) - (J-1) - 1 = (I-1) \times (J-1)$$

Bei statistischer Unabhängigkeit von Zeilen- und Spaltenvariable in der Population und einfacher Zufallsauswahl ist die Teststatistik somit mit df = (I–1)·(J–1) Freiheitsgraden asymptotisch chiquadrat-verteilt.
Wenn dagegen statistische Abhängigkeit besteht, ist die Teststatistik nichtzentral chiquadrat-verteilt, weil die Erwartungswerte der Poisson-Verteilungen ungleich den unter der Annahme der Unabhängigkeit geschätzten λ_{ij}-Parametern sind. Der Nichtzentralitätsparameter ν (nü) ist eine Funktion der quadrierten Abweichungen der tatsächlich zutreffenden λ_{ij}-Parametern von den bei Unabhängigkeit gültigen Parameten: je stärker die Differenzen sind, desto größer ist ν.
Da der Erwartungswert einer zentralen Chiquadrat-Verteilung gleich der Zahl der Freiheitsgrade ist und der Erwartungswert einer nichtzentralen Chiquadrat-Verteilung gleich der Summe aus Freiheitsgraden und Nichtzentralitätsparameter, ist bei statistischem Zusammenhang in der Population zwischen der Zeilen und der Spaltenvariable mit größeren Werten der Teststatistik zu rechnen als bei statistischer Unabhängigkeit.
Daraus ergibt sich folgende Vorgehensweise zur Prüfung der statistischen Unabhängigkeit zwischen den beiden kreuztabellierten Variablen.

Chiquadrat-Test auf statistische Unabhängigkeit von Zeilen- und Spaltenvariable

Schritt 1: Formulierung von Null- und Alternativhypothese

Die Nullhypothese behauptet statistische Unabhängigkeit, die Alternativhypothese entsprechend eine statistische Beziehung zwischen den beiden Variablen:

H_0: $\pi_{ij} = \pi_{i+} \cdot \pi_{+j}$ für i=1,2,…,I und j=1,2,…,J

vs. H_1: $\pi_{ij} \neq \pi_{i+} \cdot \pi_{+j}$ für einige oder alle i=1,2,…,I und j=1,2,…,J.

Schritt 2: Auswahl von Teststatistik und Testverteilung

Als Teststatistik wird Pearsons Chiquadrat-Statistik herangezogen:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i+} \cdot n_{+j}}{n}\right)^2}{\frac{n_{i+} \cdot n_{+j}}{n}} = n \cdot \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(p_{ij} - p_{i+} \cdot p_{+j}\right)^2}{p_{i+} \cdot p_{+j}}$$

Die Teststatistik ist bei zutreffender Nullhypothese mit df = (I–1)·(J–1) Freiheitsgraden chiquadrat-verteilt. Bei falscher Nullhypothese ist die Teststatistik nichtzentral chiquadrat-verteilt.

Schritt 3: Festlegung von Irrtumswahrscheinlichkeit und Ablehnungsbereich

Da bei falscher Nullhypothese mit großen Werten der Teststatistik zu rechnen ist, wird die Nullhypothese mit der Irrtumswahrscheinlichkeit α abgelehnt, wenn die Teststatistik größer oder gleich dem (1–α)-Quantil der Chiquadrat-Verteilung mit df = (I–1)·(J–1) Freiheitsgraden ist:

$$\chi^2 \geq \chi^2_{1-\alpha;df=(I-1)\cdot(J-1)} \Rightarrow H_1$$

Chiquadrat-Test auf statistische Unabhängigkeit von Zeilen- und Spaltenvariable

Schritt 4: Entscheidung

In Abhängigkeit von dem in der Stichprobe beobachteten Wert wird die Nullhyppthese entsprechend der Regel aus Schritt 3 abgelehnt oder beibehalten.

Schritt 5: Prüfung der Anwendungsvoraussetzungen

Der Test unterstellt, dass die Stichprobendaten aus einer einfachen Zufallsauswahl kommen bzw. alle Stichprobenfälle als unabhängige und identisch verteilte Realisationen einer gemeinsamen bivariaten diskreten Verteilung mit den Ausprägungswahrscheinlichkeiten π_{ij} aufgefasst werden können.

Der Test gilt nur asymptotisch, wobei als Faustregel gilt, dass die Annäherung hinreichend genau ist, wenn in allen Zellen i,j die ***erwarteten Häufigkeiten*** $n_{i+} \cdot n_{+j}/n \geq 5$ ist, oder aber zumindest in 80% aller Zellen und zudem jede Zelle eine erwartete Häufigkeit > 1 aufweist.

Alternative Teststatistik

Anstelle der nach dem Statistiker Pearson benannten Pearson Chiquadrat-Teststatistik kann auch die LR-Teststatistik verwendet werden:

$$L^2 = 2 \cdot \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \cdot \ln\left(\frac{n_{ij}}{n_{i+} \cdot n_{+j} / n} \right) = 2 \cdot \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \cdot \ln\left(\frac{p_{ij}}{p_{i+} \cdot p_{+j}} \right)$$

Die beiden Teststatistiken sind asymptotisch äquivalent, so dass es bis auf die Berechnungsformel bei der Anwendung in den Schritten 1 bis 5 keine Unterschiede zum Vorgehen gegenüber dem Pearsons Chiquadrat-Test gibt.

Stärke eines symmetrischen Zusammenhangs

Neben der Frage, ob zwei Variablen statistisch unabhängig voneinander sind, interessiert im Falle eines statistischen Zusammenhangs auch immer die Frage, wie stark dieser Zusammenhang ist.

Für den Zusammenhang in einer Kreuztabelle sind unterschiedliche Zusammenhangsmaße entwickelt worden. Bei zwei nominalskalierten Variablen wird sehr oft das Zusammenhangsmaß Cramérs V verwendet. Die Idee hinter diesem Maß besteht darin, den in einer Kreuztabelle beobachteten Zusammenhang mit dem maximal möglichen Zusammenhang in Beziehung zu setzen. Ein maximaler oder deterministischer Zusammenhang besteht, wenn in einer Kreuztabelle entweder in jeder Zeile oder in jeder Spalte nur eine einzige Tabellenzelle besetzt ist, also eine Häufigkeit ungleich Null aufweist. Bei quadratischen Tabellen mit gleicher Zahl von Spalten und Zeilen bedeutet dies, dass in jeder Spalte und jeder Zeile genau eine Zelle besetzt ist:

Zeilenvariable Y	Spaltenvariable X						
	x_1	x_2	…	x_j	…	x_J	Σ
y_1	0	0		0		n_{1J}	n_{1J}
y_2	n_{21}	0		0		0	n_{21}
…			…				
y_i	0	n_{i2}		0		0	n_{i2}
…			…		…		…
y_I	0	0		0		0	
Σ	n_{21}	n_{i2}	….			n_{1J}	

Beispiel für einen perfekten Zusammenhang: In jeder Spalte bzw. jeder Zeile darf maximal eine Tabellenzelle eine Häufigkeit ≠0 aufweisen. Dann gibt es entweder für alle Ausprägungen von X oder für alle Ausprägungen von Y genau eine korrepondierende Ausprägung der anderen Variablen.

Stärke eines symmetrischen Zusammenhangs

Es kann gezeigt werden, dass beim Vorliegen eines solchen maximalen Zusammenhangs die Teststatistik Pearsons Chiquadrat in der Stichprobe gerade gleich der Fallzahl mal dem kleineren Wert aus der Spalten- bzw. Zeilenzahl der Tabelle minus Eins ist:

$$\max\left(\chi^2\right) = n \cdot \min\left(I-1, J-1\right)$$

Umgekehrt gilt bei statistischer Unabhängigkeit in einer Kreuztabelle: $p_{ij} = p_{i+} \cdot p_{+j}$. Pearsons Chiquadrat-Statistik ist dann genau Null.

Das symmetrische Zusammenhangsmaß Cramérs V ist nun definiert als positive Quadratwurzel aus Pearsons Chiquadrat-Statistik geteilt durch den maximal möglichen Wert der Statistik:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min\left(I-1, J-1\right)}} = \sqrt{\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(p_{ij} - p_i \cdot p_j\right)^2}{p_i \cdot p_j}}{\min\left(I-1, J-1\right)}}$$

Der Maximalwert 1 von V wird nur bei einem perfekten Zusammenhang erreicht, der Minimalwert 0 nur dann, wenn die beiden Variablen in der Stichprobe statistisch unabhängig voneinander sind.
Aus Erfahrungswerten hat sich folgende Faustregel gebildet: Ist V < 0.05 ist der Zusammenhang sehr klein und eher zu vernachlässigen, bei Werten zwischen 0.05 und 0.1 ist der Zusammenhang als gering, bei Werten über 0.1 bis 0.3 ist er als mäßig, bei Werten >0.3 ist er als groß und bei Werten >0.5 als sehr groß zu bezeichnen.

Interpretation eines symmetrischen Zusammenhangs

Die über ein Zusammenhangsmaß gemessene Stärke des Zusammenhangs gibt einen globalen Eindruck über die Enge der Beziehung zwischen zwei Variablen. Für eine genauere Analyse ist es sinnvoll, die beobachten und bei Unabhängigkeit erwarteten Häufigkeiten in den einzelnen Tabellenzellen zu vergleichen.
Um einen Eindruck darüber zu bekommen, ob eine Differenz groß oder klein ist, werden oft die standardisierten Residuen berechnet:

$$sr_{ij} = \frac{n_{ij} - \frac{n_{i+} \cdot n_{+j}}{n}}{\sqrt{\frac{n_{i+} \cdot n_{+j}}{n}}}$$

Wenn in der Population die relative Häufigkeit π_{ij} einer Ausprägungskombination i,j von Zeilen- und Spaltenvariable gerade gleich dem Produkt der relativen Häufigkeiten $\pi_{i+} \cdot \pi_{+j}$ ist , dann ist das standardisierte Residuum in dieser Zelle asymptotisch standardnormalverteilt. Werte ≤ -2 oder $\geq +2$ weisen dann darauf hin, dass eine Abweichung relativ groß ist.
Bei negativen Werten tritt die Ausprägungskombination dann weniger häufig auf als bei Unabhängigkeit erwartet, bei positiven Werten tritt sie häufiger als erwartet auf. Aus dem Muster der Werte lässt sich so erkennen, welche Ausprägungskombinationen in der bivariaten Verteilung „überzufällig“ oft oder „überzufällig“ selten auftreten.

Anwendungsbeispiel

Als ein Beispiel wird auf der Basis der Daten des Allbus 2006 der Zusammenhang zwischen der Haltung zur Erlaubnis von Schwangerschaftsabbrüchen nach dem Willen der Schwangeren und der Mitgliedschaft in einer Religionsgemeinschaft untersucht.

Es ergibt sich folgende bivariate Häufigkeitstabelle:

Abtreibung erlaubt?	Religionsgemeinsch. nein	ja	Summe
nein	415	1387	1802
ja	689	728	1417
Summe	1104	2115	3219

(Daten: Allbus 2006)

Abtreibung	Religionsg.	erwartet	χ^2-Anteil
nein	nein	618.021	66.603
nein	ja	1183.979	32.813
ja	nein	485.979	84.813
ja	ja	931.102	44.303
SUMME:		3219.081	228.537

Zur Prüfung der Nullhypothese, dass die beiden Variablen in der Population statistisch unabhängig voneinander sind, wird Pearsons Chiquadrat-Statistik berechnet.

Die Berechnung der erwarteten Häufigkeiten nach $n_{i+} \cdot n_{+j}/n$ führt schnell zu Rundungsfehlern. Zur Kontrolle wird die Summe der erwarteten Häufigkeiten berechnet, die gleich der Fallzahl n sein muss. Im Beispiel beträgt der Rundungsfehler 0.081 (=3219.081–3219).
Aus beobachten und erwarteten Häufigkeiten werden dann die Chiquadrat-Anteile für jede Zelle berechnet und aufsummiert. Die Summe ist der Wert der Chiquadrat-Statistik, im Beispiel 228.537.

Anwendungsbeispiel

Abtreibung erlaubt?	Religionsgemeinsch. nein	ja	Summe
nein	415	1387	1802
ja	689	728	1417
Summe	1104	2115	3219

(Daten: Allbus 2006)

Abtreibung	Religionsg.	erwartet	χ^2-Anteil
nein	nein	618.021	66.603
nein	ja	1183.979	32.813
ja	nein	485.979	84.813
ja	ja	931.102	44.303
SUMME:		3219.081	228.537

Bei zwei dichotomen Variablen kann eine alternative Berechnungsformel verwendet werden, die geringere Rundungsfehler ergibt:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(n_{ij} - \frac{n_{i+} \cdot n_{+j}}{n}\right)^2}{\frac{n_{i+} \cdot n_{+j}}{n}} = \frac{n \cdot \left(n_{11} \cdot n_{22} - n_{12} \cdot n_{21}\right)^2}{n_{1+} \cdot n_{2+} \cdot n_{+1} \cdot n_{+2}} = \frac{3219 \cdot \left(689 \cdot 1387 - 728 \cdot 415\right)^2}{1417 \cdot 1802 \cdot 1104 \cdot 2115}$$

$$= 230.589$$

In der Vierfeldertabelle ergeben sich df = (2–1)×(2–1) = 1 Freiheitsgrade. Bei einer Irrtumswahrscheinlichkeit von 5% wird die Nullhypothese, dass kein Zusammenhang besteht, abgelehnt, wenn die Teststatistik das 95%-Quantil der Chiquadratverteilung mit 1 Freiheitgrad erreicht oder übersteigt. Der kritische Quantilwert beträgt 3.84. Da 230.6 sehr viel größer ist, ist die Nullhypothese abzulehnen. Vermutlich besteht in der Population ein Zusammenhang zwischen der Mitgliedschaft in einer Religionsgemeinschaft und der Haltung zu Schwangerschaftsabbrüchen.

Anwendungsbeispiel

Abtreibung erlaubt?	Religionsgemeinsch. nein	ja	Summe
nein	415	1387	1802
ja	689	728	1417
Summe	1104	2115	3219

(Daten: Allbus 2006)

Abtreibung	Religionsg.	erwartet	χ^2-Anteil
nein	nein	618.021	66.603
nein	ja	1183.979	32.813
ja	nein	485.979	84.813
ja	ja	931.102	44.303
SUMME:		3219.081	228.537

Der Allbus basiert nicht auf einer einfachen Zufallsauswahl. Insofern ist eine Anwendungsvoraussetzung nicht gegeben. Es ist zu hoffen, dass das Ergebnis robust ist.
Da die kleinste erwartete Häufigkeit mit einem Wert von 485.979 deutlich größer 5 ist, kann davon ausgegangen werden, dass die asymptotische Annäherung hinreichend genau ist.

Aus dem Chiquadratwert kann Cramérs V berechnet werden. Das Maximum ist in der Vierfeldertabelle gleich der Fallzahl. Damit ergibt sich ein Wert von

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(I-1, J-1)}} = \sqrt{\frac{230.589}{3219}} = 0.268.$$

Es besteht somit ein mittelgroßer Zusammenhang zwischen den beiden Variablen.
In der Vierfeldertabelle ist V (evtl. bis auf das Vorzeichen) gleich dem symmetrischen Zusammenhangsmaß Φ (Phi), das hier ein negatives Vorzeichen aufweist:

$$\Phi = \frac{n_{11} \cdot n_{22} - n_{12} \cdot n_{21}}{\sqrt{n_{1+} \cdot n_{2+} \cdot n_{+1} \cdot n_{+2}}} = -0.268$$

Anwendungsbeispiel

Abtreibung erlaubt?	Religionsgemeinsch. nein	ja	Summe
nein	415	1387	1802
ja	689	728	1417
Summe	1104	2115	3219

(Daten: Allbus 2006)

Abtreibung	Religionsg.	erwartet	std. Resid.
nein	nein	618.021	–8.167
nein	ja	1183.979	5.900
ja	nein	485.979	9.209
ja	ja	931.102	–6.656
SUMME:		3219.081	

Aus der Anordnung der Ausprägungen in der Tabelle und dem negativen Vorzeichen von Φ folgt, dass die Mitgliedschaft in einer Religionsgemeinschaft mit einer regiden Haltung zu Schwangerschaftsabbrüchen einhergeht.
Dies lässt sich auch erkennen, wenn man die standardisierten Residuen berechnet und interpretiert: Es gibt deutlich weniger Fälle, die gleichzeitig gegen die Erlaubnis von Schwangerschaftsabbrüchen sind und keiner Religionsgemeinschaft angehören oder die für ein Erlaubnis Verbot von Abbrüchen sind und einer Religionsgemeinschaft angehören als bei Unabhängigkeit zu erwarten wäre; umgekehrt gibt es deutlich mehr Fälle, die gegen eine Erlaubnis sind und einer Gemeinschaft angehören bzw. für ein Erlaubnis sind und keiner Gemeinschaft angehören.

Stärke eines asymmetrischen Zusammenhangs

Besteht ein Zusammenhang zwischen zwei Variablen, kann zwischen einem symmetrischen oder ungerichteten und einem asymmetrischen oder gerichteten Zusammenhang unterschieden werden. Während bei einem ungerichteten Zusammenhang beide Variable gleich behandelt werden, wird bei einem gerichteten Zusammenhang zwischen abhängiger Variable und unabhängiger Variable unterschieden. Die abhängige Variable heißt auch Kriterium oder Kriteriumsvariable, die unabhängige erklärende Variable, Prädiktor oder Prädiktorvariable.

Die Unterscheidung zwischen abhängiger und unabhängiger Variable in asymmetrischen Beziehungen ist sinnvoll,

- wenn entweder eine kausale Beziehung zwischen den Variablen vermutet oder unterstellt wird, wobei die unabhängige Variable auf die abhängige Variable wirkt,
- oder wenn der Zusammenhang für Vorhersagen genutzt werden soll, wobei dann Realisationen der unabhängigen Vatriable dazu dienen, die Realisationen der abhängigen Variable vorherzusagen.

Ausgangspunkt der Zusammenhangsanalyse bei einem asymmetrischen Zusammenhang ist eine Definition aus der Wahrscheinlichkeitstheorie, dass zwei Ereignisse dann statistisch unabhängig voneinander sind, wenn die bedingte Auftretenswahrscheinlichkeit eines Ereignisses gegeben das andere Ereignis gleich der unbedingten Auftretenswahrscheinlichkeit ist.
Entsprechend wird dann von Abhängigkeit gesprochen, wenn die bedingten Auftretenswahrscheinlichkeiten von den unbedingten abweichen.

Stärke eines asymmetrischen Zusammenhangs

Angewendet auf die asymmetrische Beziehung zwischen zwei Variablen bedeutet das, dass ein asymmetrischer Zusammenhang besteht, wenn sich die bedingten Verteilungen der abhängigen Variablen bei verschiedenen Ausprägungen der erklärenden Variablen unterscheiden.

Hinweis:
Bei statistischer Unabhängigkeit ist die Unterscheidung zwischen symmetrischer und asymmetrischer Beziehung irrelevant, weil die beiden Definitionen statistischer Unabhängigkeit als Produkt der Randwahrscheinlichkeiten (Symmetrie) oder als Gleichheit von bedingter und unbedingter Wahrscheinlichkeit (Gerichtetheit) äquivalent sind.
Wenn dagegen eine Beziehung besteht, ist die Unterscheidung zwischen gerichteter und ungerichteter Beziehung von Bedeutung, weil die Stärke des Zusammenhangs in der Regel mit unterschiedlichen Kriterien gemessen wird.

Da bei einem gerichteten Zusammenhang bedingte Verteilungen betrachtet werden, werden die Häufigkeiten in der Kreuztabelle so umgerechnet, dass sie Unterschiede in den bedingten Verteilungen aufzeigen können:

- Ist die Spaltenvariable erklärende Variable und die Zeilenvariable abhängige Variable, dann enthalten die Elemente in den Spalten die relativen bedingten Häufigkeiten $p_{i(j)}$ oder $p_{i(j)}\%$.
- Ist die Zeilenvariable erklärende Variable und die Spaltenvariable abhängige Variable, dann enthalten die Elemente in den Zeilen die relativen bedingten Häufigkeiten $p_{(i)j}$ oder $p_{(i)j}\%$.

$$p_{i(j)} = \frac{n_{ij}}{n_{+j}} \text{ bzw. } p_{i(j)}\% = 100 \cdot \frac{n_{ij}}{n_{+j}} \text{ und } p_{(i)j} = \frac{n_{ij}}{n_{i+}} \text{ bzw. } p_{(i)j}\% = 100 \cdot \frac{n_{ij}}{n_{i+}}$$

Stärke eines asymmetrischen Zusammenhangs

Zeilenvariable nach Spaltenvariable

Zeilenvariable Y	Spaltenvariable X x_1	…	x_j	…	x_I	$\sum$
y_1	$p_{1(1)}$	…	$p_{1(j)}$	…	$p_{1(J)}$	p_{1+}
…	…	…	…	…	…	…
y_i	$p_{i(1)}$	…	$p_{i(j)}$	…	$p_{i(J)}$	p_{i+}
…	…	…	…	…	…	…
y_I	$p_{I(1)}$	…	$p_{I(j)}$	…	$p_{I(J)}$	p_{I+}
$\sum$	1.0	…	1.0	…	1.0	1.0
	(p_{+1})	…	(p_{+j})	…	(p_{+J})	

Spaltenvariable nach Zeilenvariable

Zeilenvariable Y	Spaltenvariable X x_1	…	x_j	…	x_I	$\sum$	
y_1	$p_{(1)1}$	…	$p_{(1)j}$	…	$p_{(1)J}$	1.0	(p_{1+})
…	…	…	…	…	…	…	…
y_i	$p_{(i)1}$	…	$p_{(i)j}$	…	$p_{(i)J}$	1.0	(p_{i+})
…	…	…	…	…	…	…	…
y_I	$p_{(I)1}$	…	$p_{(I)j}$	…	$p_{(I)J}$	1.0	(p_{I+})
$\sum$	p_{+1}	…	p_{+j}	…	p_{+J}	1.0	

Für die Messung der Stärke einer gerichteten Beziehung werden oft PRE-Maße verwendet, die den Anteil der Reduktion der Vorhersagefehler bei der abhängigen Variable aufgrund der Kenntnis der unabhängigen Variable angeben.
Notwendig ist ein Kennwert für das Ausmaß der Vorhersagefehler. Ein möglicher Kennwert ist ein geeignetes Streuungsmaß: je größer die Streuung, desto schwieriger ist die Vorhersage einer Realisation und desto größer somit die Chance von Vorhersagefehlern.

Bei nominalskalierten Variablen ist die Devianz ein mögliches Streuungsmaß. Für eine Variable Y berechnet sich die Devianz D_Y nach:

$$D_Y = -2 \cdot \sum_{i=1}^{I} n_i \cdot \ln(p_i) = -2 \cdot n \cdot \sum_{i=1}^{I} p_i \cdot \ln(p_i)$$

Stärke eines asymmetrischen Zusammenhangs

Zeilenvariable nach Spaltenvariable

Zeilenvariable Y	Spaltenvariable X x_1	…	x_j	…	x_I	$\sum$
y_1	$p_{1(1)}$	…	$p_{1(j)}$	…	$p_{1(J)}$	p_{1+}
…	…	…	…	…	…	…
y_i	$p_{i(1)}$	…	$p_{i(j)}$	…	$p_{i(J)}$	p_{i+}
…	…	…	…	…	…	…
y_I	$p_{I(1)}$	…	$p_{I(j)}$	…	$p_{I(J)}$	p_{I+}
$\sum$	1.0	…	1.0	…	1.0	1.0
	(p_{+1})	…	(p_{+j})	…	(p_{+J})	

Spaltenvariable nach Zeilenvariable

Zeilenvariable Y	Spaltenvariable X x_1	…	x_j	…	x_I	$\sum$	
y_1	$p_{(1)1}$	…	$p_{(1)j}$	…	$p_{(1)J}$	1.0	(p_{1+})
…	…	…	…	…	…	…	…
y_i	$p_{(i)1}$	…	$p_{(i)j}$	…	$p_{(i)J}$	1.0	(pi_{+})
…	…	…	…	…	…	…	…
y_I	$p_{(I)1}$	…	$p_{(I)j}$	…	$p_{(I)J}$	1.0	(p_{I+})
$\sum$	p_{+1}	…	p_{+j}	…	p_{+J}	1.0	

Ein auf die Devianz bezogenes PRE-Maß erfasst also, um welchen Anteil sich die Devianz der abhängigen Variable verringert, wenn die Werte der unabhängigen Variable bekannt sind und enstprechend die relativen Summen der Devianzen der bedingten Verteilungen anstelle der Devianz der Randverteilung berechnet werden.
Wenn die Zeilenvariable Y abhängige Variable ist, berechnet sich die relative Devianzreduktion, R‘, die auch als Unsicherheitskoeffizient oder MacFaddens Pseudo-R-Quadrat bezeichnet wird, nach:

$$R'_{Y.X} = \frac{-2\sum_{i=1}^{I} n_{i+} \cdot \ln(p_{i+}) - \left(-2\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ij} \cdot \ln(p_{i(j)})\right)}{-2\sum_{i=1}^{I} n_{i+} \cdot \ln(p_{i+})} = 1 - \frac{\sum_{J=1}^{J} p_{+j}\left(\sum_{i=1}^{I} p_{i(j)} \cdot \ln(p_{i(j)})\right)}{\sum_{i=1}^{I} p_{i+} \cdot \ln(p_{i+})}$$

Stärke eines asymmetrischen Zusammenhangs

Zeilenvariable nach Spaltenvariable

Zeilenvar-riable Y	Spaltenvariable X x_1	…	x_j	…	x_J	Σ
y_1	$p_{1(1)}$	…	$p_{1(j)}$	…	$p_{1(J)}$	p_{1+}
…	…	…	…	…	…	…
y_i	$p_{i(1)}$	…	$p_{i(j)}$	…	$p_{i(J)}$	p_{i+}
…	…	…	…	…	…	…
y_I	$p_{I(1)}$	…	$p_{I(j)}$	…	$p_{I(J)}$	p_{I+}
Σ	1.0	…	1.0	…	1.0	1.0
	(p_{+1})	…	(p_{+j})	…	(p_{+J})	

Spaltenvariable nach Zeilenvariable

Zeilenvar-riable Y	Spaltenvariable X x_1	…	x_j	…	x_J	Σ	
y_1	$p_{(1)1}$	…	$p_{(1)j}$	…	$p_{(1)J}$	1.0	(p_{1+})
…	…	…	…	…	…	…	…
y_i	$p_{(i)1}$	…	$p_{(i)j}$	…	$p_{(i)J}$	1.0	(pi_{+})
…	…	…	…	…	…	…	…
y_I	$p_{(I)1}$	…	$p_{(I)j}$	…	$p_{(I)J}$	1.0	(p_{I+})
Σ	p_{+1}	…	p_{+j}	…	p_{+J}	1.0	

Ist die Spaltenvariable X abhängige Variable ist, berechnet sich die relative Devianzreduktion nach:

$$R'_{X.Y} = 1 - \frac{\sum_{i=1}^{I} p_{i+} \left(\sum_{j=1}^{J} p_{(i)j} \cdot \ln\left(p_{(i)j}\right) \right)}{\sum_{j=1}^{J} p_{+j} \cdot \ln\left(p_{+j}\right)}$$

Als PRE-Maß ist die Interpretation einfach: Ein Wert von z.B. R‘ = 0.05 oder 5% besagt, dass sich bei Kenntnis der unabhängigen Variable die Streuung der abhängigen Variable um 5% reduziert wird, oder 5% der Streuung der abhängigen Variable durch die unabhängige Variable erklärt wird, wobei Erklärung nicht unbedingt im Sinne einer kausalen Erklärung gemeint ist.

Stärke eines asymmetrischen Zusammenhangs

Bei der Frage, ab wann ein Zusammenhang als gering, mäßig oder stark zu bezeichnen ist, ist zu beachten, dass die Devianzreduktion in der Regel deutlich kleinere Werte aufweist als etwa Cramérs V.

Um annähernde Vergleichbarkeit herzustellen, muss die positive Quadratwurzel aus der Devianzreduktion gezogen werden. Danach ist ein Zusammenhang praktisch vernachlässigbar, wenn die Devianzreduktion < 0.0025(=0.05^2) ist. Der gerichtete Zusammenhang ist gering, wenn die Devianzreduktion zwischen 0.025 und <0.01 (=0.1^2), mäßig wenn die Reduktion zwischen 0.01 und 0.09 (=0.3^2) liegt, groß ab einem Wert > 0.09 und sehr groß ab einem Wert >0.25 (=0.5^2) .

Hinweis:

Wenn abhängige und unabhängige Variable vertauscht werden, ergeben sich in der Regel unterschiedliche Werte der Devianzreduktion.

Da bei statistischer Unabhängigkeit nicht zwischen symmetrischer und asymmetrischer Beziehung unterschieden werden muss, kann zum Testen der Nullhypothese, dass die Devianzreduktion in der Population Null ist, Pearsons Chiquadrat-Test oder auch der LR-Test verwendet werden. Der LR-Test ist etwas naheliegender, da die Differenz der bedingten von den unbedingten Devianzen stets gleich dem Wert der LR-Statistik ist:

$$L^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \cdot \ln\left(\frac{p_{ij}}{p_{i+} \cdot p_{+j}}\right)$$

$$= -2 \cdot \left(\sum_{i=1}^{I} n_{i+} \cdot \ln\left(p_{i+}\right) - \sum_{j=1}^{J}\sum_{i=1}^{I} n_{ij} \cdot \ln\left(p_{i(j)}\right) \right) = -2 \cdot \left(\sum_{j=1}^{J} n_{+j} \cdot \ln\left(p_{+j}\right) - \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \cdot \ln\left(p_{(i)j}\right) \right)$$

Anwendungsbeispiel

Abtreibung erlaubt?	Religionsgemeinsch. nein	ja	Summe
nein	415	1387	1802
ja	689	728	1417
Summe	1104	2115	3219

(Daten: Allbus 2006)

Abtreibung Erlaubt?	Religionsgemeins. nein	ja	Randvert-teilung
nein	812.086	1170.383	2090.946
ja	649.664	1552.837	2325.379
Devianz	1461.750	2723.220	4416.325

Als Beispiel kann wieder der Zusammenhang zwischen Haltung zu Schwangerschaftsabbrüchen und Mitglied einer Religionsgemeinschaft betrachtet werden. Es liegt nahe die Mitgliedschaft als erklärende Variable zu betrachten, die die Haltung zu Schwangerschaftsabbrüchen beeinflusst, da ein Verbot von Schwangerschaftsabbrüchen von den in den in Deutschland dominierenden christlichen Religionsgemeinschaften gefordert wird.
Ist die Spaltenvariable erklärende Variable, wird zunächst für die Zellen in jeder Spalte der Devianzanteil der Zelle berechnet und aufsummiert, z.B. -2·415·ln(415/1104) = 812.086.

Anschließend kann die Devianzreduktion berechnet werden:

$$R' = 1 - \frac{\sum_{j=1}^{J} D_{Y.X_j}}{D_Y} = 1 - \frac{1461.75 + 2723.22}{4416.325} = 0.052$$

Gemessen über die Devianzreduktion reduziert sich die Streuung der Haltung zu Abtreibungen also um 5.2% , wenn bekannt ist, ob eine Person einer Religionsgemeinschaft angehört.

Anwendungsbeispiel

Abtreibung erlaubt?	Religionsgemeinsch. nein	ja	Summe
nein	415	1387	1802
ja	689	728	1417
Summe	1104	2115	3219

(Daten: Allbus 2006)

Religions-gemeins.	Abtreibung erlaubt? nein	ja	Randvert-teilung
nein	1218.750	993.615	2362.849
ja	726.106	969.690	1776.667
Devianz	1944.856	1963.305	4139.516

Ist die Spaltenvariable erklärende Variable wird zunächst für die Zellen in jeder Spalte der Devianzanteil der Zelle berechnet und aufsummiert, z.B. -2·415·ln(415/1802) = 1218.750.
Anschließend kann die Devianzreduktion berechnet werden:

$$R' = 1 - \frac{\sum_{i=1}^{I} D_{X.Y_i}}{D_X} = 1 - \frac{1944.856 + 1963.305}{4139.516} = 0.056$$

Wird die Mitgliedschaft in einer Religionsgemeinschaft durch die Haltung zu Schwangerschaftsabbrüchen vorhergesagt, beträgt die Devianzreduktion 5.6%, ist also etwas höher als bei der Erklärung der Haltung zur Abtreibung durch Mitgliedschaft in einer Religionsgemeinschaft.

Soll getestet werden, ob in der Population kein Zusammenhang besteht, kann aus den Devianzen leicht die LR-Teststatistik berechnet werden.

$$L^2 = D_X - \sum_{j=1}^{J} D_{X.Y_j} = 4139.516 - 1944.856 - 1963.305 = 231.355$$

Anwendungsbeispiel

Abtreibung erlaubt?	Religionsgemeinsch. nein	ja	Summe
nein	415	1387	1802
ja	689	728	1417
Summe	1104	2115	3219

(Daten: Allbus 2006)

Abtreibung Erlaubt?	Religionsgemeins. nein	ja	Randvert-teilung
nein	812.086	1170.383	2090.946
ja	649.664	1552.837	2325.379
Devianz	1461.750	2723.220	4416.325

Bis auf Rundungsfehler ergibt sich in der bivariaten Kreuztabelle der gleiche Wert auch, wenn die umgekehrte Richtung betrachtet wird oder die LR-Teststatistik für einen symmetrischen Zusammenhang berechnet wird.

$$L^2 = D_Y - \sum_{i=1}^{I} D_{Y.X_i} = 4416.325 - 1461.75 - 2723.22 = 231.355$$

$$L^2 = D_X - \sum_{j=1}^{J} D_{X.Y_j} = 4139.516 - 1944.856 - 1963.305 = 231.355$$

$$L^2 = 2 \cdot \sum^{I} \sum^{J} n_{ij} \cdot \ln\left(n_{ij} / \frac{n_{i+} \cdot n_{+j}}{n} \right) = 231.230$$

Abtreibung	Religionsg.	erwartet	L^2-Anteil
nein	nein	618.021	–330.542
nein	ja	1183.979	439.020
ja	nein	485.979	481.027
ja	ja	931.102	– 358.275
SUMME:		3219.081	231.230

Erwartungsgemäßg weicht der Wert der LR-Statistik mit 231.355 nur geringfügig vom Wert von Pearsons Chiquadrat-Statistik ab, der im Beispiel 231.598 beträgt.

Interpretation der gerichteten Beziehung

Neben der Stärke des Zusammenhangs interessiert bei gerichteten Beziehungen auch stets, wie die Art der Beziehung ist, also wie sich die bedingten Verteilungen der abhängigen Variable bei unterschiedlichen Werten der erklärenden Variablen unterscheiden.
Dazu können die relativen Häufigkeiten oder Prozentwerte einer Ausprägung der abhängigen Variable bei verschiedenen Ausprägungen der unabhängigen Variable miteinander verglichen werden.

Abtreibung erlaubt?	Religionsgemeinsch. nein	ja	Summe
nein	37.6%	56.6%	66.0%
ja	62.4%	34.4%	44.0%
Summe	100.0% (1104)	100.0% (2115)	100.0% (3219)

(Daten: Allbus 2006)

Wird die Haltung zur Abtreibung als abhängig betrachtet, zeigt sich, dass Personen ohne Religionszugehörigkeit mit 37.6% einen um 28 Prozentpunkte niedrigeren Anteil von Gegnern von Abbrüchen haben als Mitglieder von Religionsgemeinschaften mit 56.6%

Abtreibung erlaubt?	Religionsgemeinsch. nein	ja	Summe
nein	23.0%	77.0%	100.0% (1802)
ja	48.6%	51.4%	100.0% (1417)
Summe	34.3%	65.7%	100.0% (3219)

(Daten: Allbus 2006)

Wird die Mitgliedschaft als abhängig betrachtet, zeigt sich, dass Personen, die Abbrüche ablehnen, mit 23.0% einen um 25.6 Prozentpunkte geringeren Anteil von Nichtmitgliedern aufweisen als Personen, die Abbrüche erlauben wollen mit 48.6%.

Interpretation der gerichteten Beziehung

Abtreibung erlaubt?	Religionsgemeinsch. nein	ja	Summe
nein	415	1387	1802
ja	689	728	1417
Summe	1104	2115	3219

(Daten: Allbus 2006)

Additiver Zusammenhang:

$$p_{Y=y_1|X=x_2} + d_{YX} \rightarrow p_{Y=y_1|X=x_1}$$

$$\frac{1387}{2115} + (-0.280) = \frac{415}{1104}$$

Wenn die Veränderungen der Prozentwerte (bzw. Anteile) mittels Prozentsatzdifferenzen (oder Anteilsdifferenzen) verglichen werden, wird die Beziehung als ***additiver Zusammenhang*** modelliert: Es wird betrachtet, welcher Wert bei einem Wechsel von einer Ausprägung der erklärenden Variable zu einer anderen Ausprägung addiert (bzw. bei negativen Werten subtrahiert) werden muss, um vom Ausgangswert der abhängigen Variablen zum neuen Wert der abhängigen Variablen gegeben zu gelangen.
Da bei der Prozentsatzdifferenz der Wert der zweiten Spalte von dem der ersten abgezogen wird, ist die zweite Spalte der Ausgangswert, zu dem ein Wert addiert bzw. subtrahiert werden muss, um zum Wert der ersten Spalte zu gelangen:

$$p_{Y|X=1} - p_{Y|X=x_2} = d_{YX} \rightarrow p_{Y|X=x_2} + d_{YX} = p_{Y|X=x_1}$$

$$\frac{415}{1104} - \frac{1387}{2115} = -0.28 \qquad 0.656 + (-0.28) = 0.376$$

Additive und multiplikative Modellierung eines gerichteten Zusammenhangs

Abtreibung erlaubt?	Religionsgemeinsch. nein	ja	Summe
nein	415	1387	1802
ja	689	728	1417
Summe	1104	2115	3219

(Daten: Allbus 2006)

Additiver Zusammenhang:

$$p_{Y=y_1|X=x_2} + d_{YX} \rightarrow p_{Y=y_1|X=x_1}$$

$$\frac{1387}{2115} + (-0.280) = \frac{415}{1104}$$

Alternativ wird bei kategorialen Variablen eine Beziehung auch oft als ***multiplikativer Zusammenhang*** modelliert: Dann wird betrachtet, mit welchem Wert der Ausgangswert der abhängigen Variable **multipliziert** werden muss, um bei einem Wechsel der Ausprägungen der erklärenden Variable zum geänderten Wert zu gelangen.

Anstelle der Ausgangswerte oder Anteile der abhängigen Variable werden bei der multiplikativen Modellierung in der Regel Ausprägungsverhältnisse betrachtet , die nach der englischen Bezeichnung in der Statistik als ***Odds*** (englisch für „Wette“) bezeichnet werden:

Das Odd „Abtreibung nein zu Abtreibung ja“ beträgt in der Randverteilung im Beispiel 1802 zu 1417 oder 1.272 (=1802/1417). Es gibt 1.272 mal so viele Befragte, die gegen das Erlauben einer Abtreibung sind als es Befragte gibt, die für das Erlauben sind.
Die konditionalen Odds bei Personen, die keiner Religionsgemeinschaft angehören, betragen 415 zu 689 oder 0.602, die derjenigen, die einer Gemeinschaft angehören, betragen 1387 zu 728 oder 1.901.

Multiplikative Modellierung eines gerichteten Zusammenhangs

Abtreibung erlaubt?	Religionsgemeinsch. nein	ja	Summe
nein	415	1387	1802
ja	689	728	1417
Summe	1104	2115	3219

(Daten: Allbus 2006)

Multiplikativer Zusammenhang:

$$\frac{n_{Y=y_1\|X=x_2}}{n_{Y=y_2\|X=x_2}} \times \alpha \rightarrow \frac{n_{Y=y_1\|X=x_1}}{n_{Y=y_2\|X=x_1}}$$

$$1387 / 728 \cdot 0.316 = 415 / 689$$

Als Maß für die Höhe eines multiplikativen Zusammenhangs wird dann das ***Odds-Ratio*** berechnet, das ist der Quotient von zwei Odds der abhängigen Variable bei verschiedenen Ausprägungen der erklärenden Variable.
In der Vierfeldertabelle wird das Odds-Ratio auch als ***Kreuzproduktverhältnis*** bezeichnet und durch den kleinen griechischen Buchstaben ***α*** symbolisiert:

$$\alpha = \frac{n_{11}}{n_{21}} / \frac{n_{12}}{n_{22}} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}} \text{ ; im Beispiel: } \alpha = \frac{415 \cdot 728}{1387 \cdot 689} = 0.316$$

Hinweis:
Das gleiche Symbol α wird auch verwendet, um die Konstante in einer linearen Gleichung oder auch die Irrtumswahrscheinlichkeit bei einem Test zu kennzeichnen. Mann muss daher aufpassen, welche Bedeutung das Symbol jeweils hat.

Das Odds-Ratio gibt dann den Veränderungsfaktor der Odds bei einem Wechsel von der zweiten zur ersten Ausprägung der erklärenden Variable an:

$$n_{11} / n_{21} \; / \; n_{12} / n_{22} = \alpha \rightarrow n_{12} / n_{22} \times \alpha = n_{11} / n_{21}$$

$$415 / 689 \; / \; 1387 / 728 = 0.316 \qquad 1.905 \cdot 0.316 = 0.602$$

Interpretation der gerichteten Beziehung

Abtreibung erlaubt?	Religionsgemeinsch. nein	ja	Summe
nein	415	1387	1802
ja	689	728	1417
Summe	1104	2115	3219

(Daten: Allbus 2006)

$$d_{YX}\% = 100(415/1104 - 1387/2115) = -28.0$$
$$\alpha = 415/689 \; / \; 1387/728 = 0.316$$

Additiver Zusammenhang:

$$\frac{1387}{2115} + (-0.280) = \frac{415}{1104}$$

Multiplikativer Zusammenhang:

$$\frac{1387}{728} \times 0.316 = \frac{415}{689}$$

In der Vierfeldertabelle wird die Prozentsatzdifferenz (im Beispiel –28.0 Prozentpunkte) meist häufiger berechnet als das Odds-Ratio (im Beispiel der Faktor 0.316), da das additive Modell der Differenz von Prozenten oder Anteilen eher der aus dem Alltag gewohnten Sicht auf Veränderungen entspricht.

Allerdings hat das Odds-Ratio in der Vierfeldertabelle den Vorteil, bei Vertauschung von abhängiger und unabhängiger Variable stets den gleichen Wert zu ergeben:

$$\alpha_{YX} = 415/689 \; / \; 1387/728 = 0.316 = \alpha_{XY} = 415/1387 \; / \; 689/728.$$

Die Prozentsatzdifferenzen können sich sich dagegen bei Vertauschung von abhängiger und unabhängiger Variable unterscheiden:

$$d_{YX}\% = 100(415/1104 - 1387/2115) = -28.0 \neq d_{XY}\% = 100(415/1802 - 689/1417) = -25.6.$$

Interpretation der gerichteten Beziehung

Abtreibung erlaubt?	Religionsgemeinsch. nein	ja	Summe
nein	415	1387	1802
ja	689	728	1417
Summe	1104	2115	3219

(Daten: Allbus 2006)

$d_{YX}\% = 100(415/1104 - 1387/2115) = -28.0$

$\alpha = 415/689 \,/\, 1387/728 = 0.316$

Abtreibung erlaubt?	Religionsgemeinsch. nein	ja	Summe
nein	415 ·2	1387 ·2	1802 ·2
ja	689	728	1417
Summe	1519	3502	5021

(Häufigkeiten in 1. Zeile verdoppelt)

$d_{YX}\% = 100(830/1519 - 2774/3502) = -24.6$

$\alpha = 830/689 \,/\, 2774/728 = 0.316$

Werden (z.B. als Folge eines geänderten Stichprobenplans) die Ausprägungshäufigkeiten einer Kategorie der abhängigen Variablen um einen konstanten Faktor verändert, dann hat dies Auswirkungen auf den Wert der Veränderung im additiven Modell, aber nicht im multiplikativen Modell.

Im Beispiel werden die Häufigkeiten in der ersten Zeile (Abtreibung sollte nicht erlaubt sein) verdoppelt. Die Prozentsatzdifferenz sinkt dadurch von –28 Punkten auf –24.6 Punkte. Die Odds-Ratios bleiben dagegen unverändert bei 0.316.

Diesem Vorteil der multiplikativen Sicht steht der Nachteil der ungewohnten Interpretation gegenüber. So bedeutet ein Veränderungsfaktor von 1.0 keine Veränderung (entspricht also im additiven Modell dem Zuwachs 0). Das Wachstum um das Doppelte (Faktor 2.0) ist im multiplikativen Modell genau so groß wie ein Rückgang auf die Hälfte (Faktor 0.5); im additiven Modell kann ein Anstieg und ein Rückgang dagegen direkt verglichen werden.

Lerneinheit 2: Drittvariablenkontrolle bei Kreuztabellen

Mit Hilfe der Drittvariablenkontrolle wird untersucht wird, ob ein bivariater Zusammenhang stabil bleibt oder sich ändert, wenn nicht nur die gemeinsame Verteilung von zwei Variablen, sondern von drei oder auch mehr Variablen gleichzeitig (simultan) betrachtet werden.

Als Beispiel aus der Empirievorlesung wird der Zusammenhang zwischen der Beurteilung von Schwangerschaftsabbrüchen nach dem Telefonbesitz im Haushalt und dem Erhebungsgebiet betrachtet.

Abtreibung, wenn die Frau es will, ...	Telefonanschluss im Haushalt? ja	nein
... sollte verboten sein	54.7%	33.0%
... sollte erlaubt sein	45.3%	67.0%
	(2331)	(782)

(Quelle: ALLBUS 1992)

Bivariat zeigt sich nach den Daten des Allbus 1992, dass Haushalten, die über einen Telefonanschluss verfügten eher für ein Verbot von Schwangerschaftsabbrüchen waren als Personen, die über kein Telefon verfügten.

Abtreibung, wenn die Frau es will, ...	Alte Bundesländer Telefonanschluss? ja	nein	Neue Bundesländer Telefonanschluss? ja	nein
... sollte verboten sein	58.5%	62.8%	28.9%	29.7%
... sollte erlaubt sein	41.5%	37.2%	71.1%	70.3%
	(2026)	(78)	(305)	(704)

(Quelle: ALLBUS 1992)

Allerdings „verschwindet“ der Zusammenhang, wenn zwischen alten und neuen Ländern unterschieden wird.

Drittvariablenkontrolle bei Kreuztabellen

Abtreibung, wenn die Frau es will, ...	Alte Bundesländer Telefonanschluss? ja	nein	Neue Bundesländer Telefonanschluss? ja	nein	Insgesamt Telefonanschluss ja	nein
... sollte verboten sein	58.5%	62.8%	28.9%	29.7%	54.7%	33.0%
... sollte erlaubt sein	41.5%	37.2%	71.1%	70.3%	45.3%	67.0%
	(2026)	(78)	(305)	(704)	(2331)	(782)
(Quelle: ALLBUS 1992)	$d_{YX}\% = -4.3$		$d_{YX}\% = -0.8$		$d_{YX}\% = +21.7$	

Das Beispiel verdeutlicht, dass die bivariate Analyse oft nicht ausreicht, um einen Ausschnitt aus der Wirklichkeit angemessen zu erfassen. Tatsächlich ist es möglich, dass ein zwischen zwei Variablen beobachteter Zusammenhang bei Drittvariablenkontrolle geringer ausfällt, gar nicht vorhanden ist oder bei ordinalen bzw. metrischen Variablen das Vorzeichen ändert. Möglich ist aber auch, dass bei Drittvariablenkontrolle ein Zusammenhang stärker wird oder überhaupt erst sichtbar wird. Schließlich kann es auch vorkommen, dass der Zusammenhang in Abhängigkeit von den Werten der Drittvariablen unterschiedlich ausfällt.

Um zu verstehen, wieso sich bivariate und konditionale Beziehungen unterscheiden können, ist es sinnvoll verschiedene Datenkonstellationen zu betrachten, die sich bei unterschiedlichen kausalen Beziehungen zwischen den Variablen einstellen können.

Drittvariablenkontrolle bei Kreuztabellen

	W														
	w_1					w_2					…	w_K				
	X					X						X				
Y	x_1	x_2	…	x_J	Σ	x_1	x_2	…	x_J	Σ		x_1	x_2	…	x_J	Σ
y_1	n_{111}	n_{121}	…	n_{1J1}	n_{1+1}	n_{112}	n_{122}	…	n_{1J2}	n_{1+2}	…	n_{11K}	n_{12K}	…	n_{1JK}	n_{1+K}
y_2	n_{211}	n_{221}	…	n_{2J1}	n_{2+1}	n_{212}	n_{222}	…	n_{2J2}	n_{2+2}	…	n_{21K}	n_{22K}	…	n_{2JK}	n_{2+K}
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
Y_I	n_{I11}	n_{I21}	…	n_{IJ1}	n_{I+1}	n_{I12}	n_{I22}	…	n_{IJ2}	n_{I+2}	…	n_{I1K}	n_{I2K}	…	n_{IJK}	n_{I+K}
Σ	n_{+11}	n_{+21}	…	n_{+J1}	n_{++1}	n_{+12}	n_{+22}	…	n_{+J2}	n_{++2}	…	n_{+1K}	n_{+2K}	…	n_{+JK}	n_{++K}

Zuvor einige Hinweise zur Notation:
Jede der simultan betrachteten Variablen definiert eine Dimension der Kreuztabelle. Bei drei Variablen ergibt sich so eine dreidimensionale Tabelle. Die Darstellung erfolgt aber meist so, dass mehrere zweidimensionale Kreuztabellen präsentiert werden, die nebeneinander oder untereinander stehen. Die zweidimensionalen Tabellen heißen ***Partialtabellen***, da sie die bivariate Verteilung von zwei Variablen gegeben die Ausprägung(en) der Drittvariable(n) enthalten.

Jede Dimension einer mehrdimensionalen Tabelle wird durch einen eigenen Index identifiziert. Dabei gilt i.A., dass sich der erste Index auf die Zeilenvariable der Partialtabellen (im Beispiel Y) bezieht, der zweite Index auf die Spaltenvariable der Partialtabellen (im Beispiel X) und der dritte und evtl. weitere Indizes auf die Drittvariablen (im Beispiel W) beziehen.
So bezeichnet n_{ijk} die Häufigkeit der Ausprägungskombination $(Y=y_i, X=x_j, W=w_k)$.

Drittvariablenkontrolle bei Kreuztabellen

	W														
	w_1					w_2					…	w_K				
	X					X						X				
Y	x_1	x_2	…	x_J	Σ	x_1	x_2	…	x_J	Σ		x_1	x_2	…	x_J	Σ
y_1	$p_{11(.1)}$	$p_{12(.1)}$	…	$p_{1J(.1)}$	$p_{1+(.1)}$	$p_{11(.2)}$	$p_{12(.2)}$	…	$p_{1J(.2)}$	$p_{1+(.2)}$	…	$p_{11(.K)}$	$p_{12(.K)}$	…	$p_{1J(.K)}$	$p_{1+(.K)}$
y_2	$p_{21(.1)}$	$p_{22(.1)}$	…	$p_{2J(.1)}$	$p_{2+(.1)}$	$p_{21(.2)}$	$p_{22(.2)}$	…	$p_{2J(.2)}$	$p_{2+(.2)}$	…	$p_{21(.K)}$	$p_{22(.K)}$	…	$p_{2J(.K)}$	$p_{2+(.K)}$
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
Y_I	$p_{I1(.1)}$	$p_{I2(.1)}$	…	$p_{IJ(.1)}$	$p_{I+(.1)}$	$p_{I1(.2)}$	$p_{I2(.2)}$	…	$p_{IJ(.2)}$	$p_{I+(.2)}$	…	$p_{I1(.K)}$	$p_{I2(.K)}$	…	$p_{IJ(.K)}$	$p_{I+(.K)}$
Σ	$p_{+1(.1)}$	$p_{+2(.1)}$	…	$p_{+J(.1)}$	$p_{++(.1)}$	$p_{+1(.2)}$	$p_{+2(.2)}$	…	$p_{+J(.2)}$	$p_{++(.2)}$	…	$p_{+1(.K)}$	$p_{+2(.K)}$	…	$p_{+J(.K)}$	$p_{++(.K)}$

In der Regel erfolgt die Berechnung relativer Häufigkeiten innerhalb der Partialtabellen. Zur Kennzeichnung der Drittvariablen werden die Indizes der Drittvariablen in Klammern gesetzt und durch einen Punkt vor dem Index gekennzeichnet.
Die auf eine Partialtabelle bezogene relative Häufigkeit $p_{ij(.k)}$ von Y berechnet sich so nach:

$$p_{ij(.k)} = \frac{n_{ijk}}{n_{++k}}$$

Konditionale relative Häufigkeiten innerhalb der Partialtabellen berechnen sich analog zu bivariaten Tabellen innerhalb der Kreuztabellen. Der Index der unabhängigen Variable ist ebenfalls in Klammern gesetzt allerdings ohne Punkt vor dem Index.
Die durch die Ausprägung $X=x_j$ bedingte relative Häufigkeit von Y in der Partialtabelle $W=w_k$ berechnet sich daher nach:

$$p_{i(j.k)} = \frac{n_{ijk}}{n_{+jk}} = \frac{p_{ij(.k)}}{p_{+j(.k)}}$$

Drittvariablenkontrolle bei Kreuztabellen

Analog ergibt sich die durch $Y=y_i$ bedingte relative Häufigkeit von X in der Partialtabelle $W=w_k$ nach:

$$p_{(i)j(.k)} = \frac{n_{ijk}}{n_{i+k}} = \frac{p_{ij(.k)}}{p_{i+(.k)}}$$

Ähnlich erfolgt die Kennzeichnung von Zusammenhangsmaßen in den Partialtabellen, die als ***konditionale Zusammenhangsmaße*** bezeichnet werden, da sie konditional gegeben die Ausprägungen der Drittvariablen gelten.
Bei Zusammenhangsmaßen werden die Variablen meist explizit genannt: $d_{YX(W=1)}$% ist so die Prozentsatzdifferenz in der Partialtabelle für W=1, wenn Y abhängige und X erklärende Variable ist. Entsprechend bezeichnet $d_{XY(W=1)}$% die Prozentsatzdifferenz, wenn X abhängige und Y unabhängige Variable ist, und $d_{WY(X=1)}$% die Prozentsatzdifferenz mit W als abhängiger, Y als unabhängiger und X als Drittvariable mit der Ausprägung X=1.
Wenn die konditionalen Zusammenhangsmaße bei allen Ausprägungen der Drittvariablen die gleichen Werte aufweisen, spricht man auch von ***partiellen Zusammenhangsmaßen***. Bei partiellen Zusammenhangsmaßen kann die Ausprägung der Kontrollvariablen ausgelassen werden; $d_{YX(W)}$% ist also die ***partielle Prozentsatzdifferenz*** mit Y als abhängiger, X als erklärender und W als Kontrollvariable.

Partialtabellen und Randtabellen

Analog zu den univariaten Randverteilungen in bivariaten Kreuztabellen ergeben sich zwei- oder mehrdimensionale ***Randtabellen*** durch ***Aggregieren***, d.h. Aufsummieren über alle Werte von Drittvariablen. Aus einer dreidimensionalen Kreuztabelle lassen sich so drei zweidimensionale Randtabellen erstellen:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	252	90	63	105	510
Y=0	98	60	87	245	490
∑	350	150	150	350	1000

Aufsummieren über W:

Y	X=1	X=0	∑
Y=1	315	195	510
Y=0	185	305	490
∑	500	500	1000

Aufsummieren über X:

Y	W=1	W=0	∑
Y=1	342	168	510
Y=0	158	332	490
∑	500	500	1000

Aufsummieren über Y:

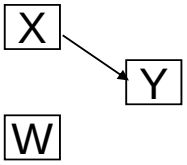
W	X=1	X=0	∑
W=1	350	150	500
W=0	150	350	500
∑	500	500	1000

Kausale Beziehungen zwischen drei Variablen

In den folgenden Beispielen werden die Daten entsprechend einer vorgegebenen Kausalstruktur durch Simulation generiert. Dabei ist Y stets die abhängige Variable und X und W sind unabhängige Variablen, die Y kausal beeinflussen.

1) Nur X wirkt additiv auf Y, X und W sind statistisch unabhängig voneinander

Die Daten sind so generiert, dass X einen additiven Effekt auf Y hat.
Die Drittvariable W ist statistisch unabhängig von X und Y.
Diese Kausalstruktur lässt sich durch ein Pfadmodell darstellen, indem ein gerichteter Pfad von der erklärenden Variable X auf die abhängige Variable Y geht und W weder mit X noch mit Y verbunden ist.



Generierte Daten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	140	110	140	110	500
Y=0	110	140	110	140	500
∑	250	250	250	250	1000

Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.56	.44	.56	.44	.50
Y=0	.44	.56	.44	.56	.50
n	(250)	(250)	(250)	(250)	(1000)

$d_{YX(W=1)}\%=12.0$ $d_{YX(W=0)}\%=12.0$
$\Phi_{XY(W=1)} = 0.12$ $\Phi_{XY(W=0)} = 0.12$

Randtabelle Y nach X:

Y	X=1	X=0	∑
Y=1	.56	.44	.50
Y=0	.44	.56	.50
∑	(500)	(500)	(1000)

$d_{YX} = 12.0$
$\Phi_{XY} = 0.12$

Randtabelle Y nach W:

Y	W=1	W=0	∑
Y=1	.50	.50	.50
Y=0	.50	.50	.50
∑	(500)	(500)	(1000)

$d_{YW} = 0$
$\Phi_{WY} = 0$

Randtabelle W nach X:

W	X=1	X=0	∑
W=1	.50	.50	.50
W=0	.50	.50	.50
∑	(500)	(500)	(1000)

$d_{WX} = 0$
$\Phi_{XW} = 0$

Nur X wirkt additiv auf Y, X und W sind statistisch unabhängig voneinander

Generierte Daten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	140	110	140	110	500
Y=0	110	140	110	140	500
∑	250	250	250	250	1000

Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.56	.44	.56	.44	.50
Y=0	.44	.56	.44	.56	.50
n	(250)	(250)	(250)	(250)	(1000)

$d_{YX(W=1)}\%=12.0$ $d_{YX(W=0)}\%=12.0$
$\Phi_{XY(W=1)} = 0.12$ $\Phi_{XY(W=0)} = 0.12$

Randtabelle Y nach X:

Y	X=1	X=0	∑
Y=1	.56	.44	.50
Y=0	.44	.56	.50
∑	(500)	(500)	(1000)

$d_{YX} = 12.0$
$\Phi_{XY} = 0.12$

Randtabelle Y nach W:

Y	W=1	W=0	∑
Y=1	.50	.50	.50
Y=0	.50	.50	.50
∑	(500)	(500)	(1000)

$d_{YW} = 0$
$\Phi_{WY} = 0$

Randtabelle W nach X:

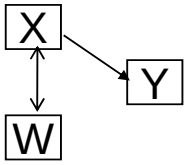
W	X=1	X=0	∑
W=1	.50	.50	.50
W=0	.50	.50	.50
∑	(500)	(500)	(1000)

$d_{WX} = 0$
$\Phi_{XW} = 0$

In dieser Datenkonstellation unterscheidet sich der bivariate Effekt gemessen über die Prozentsatzdifferenz $d_{YX}\%$ nicht von den kondtionalen Zusammenhangsmaßen $d_{YX(W=1)}\%$ und $d_{YX(W=0)}\%$, die daher als partielles Zusammenhangsmaß $d_{YX(W)}\%$ zusammengefasst werden können.
Asymmetrische und symmetrische Zusammenhangsmaße sind nur identisch, wenn bei Vertauschung von abhängiger und erklärender Variable die Werte der asymmetrischen Maße gleich bleiben.

2) Nur X wirkt additiv auf Y, X und W sind nicht statistisch unabhängig

Gegenüber der vorigen Situation unterscheidet sich diese Datenkonstellation dadurch, dass zwar weiterhin nur X auf W additiv wirkt, X aber nicht mehr statistisch unabhängig von W ist.
Der ungerichtete (symmetrische) Zusammenhang zwischen X und W wird im Pfaddiagramm durch eine Verbindungslinie symbolisiert, die an beiden Enden eine Pfeilspitze aufweist.



Generierte Daten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	196	66	84	154	500
Y=0	154	84	66	196	500
∑	350	150	150	350	1000

Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.56	.44	.56	.44	.50
Y=0	.44	.56	.44	.56	.50
n	(350)	(150)	(350)	(150)	(1000)

$d_{YX(W=1)}\%=12.0$ $d_{YX(W=0)}\%=12.0$
$\Phi_{XY(W=1)} = 0.12$ $\Phi_{XY(W=0)} = 0.12$

Randtabelle Y nach X:

Y	X=1	X=0	∑
Y=1	.56	.44	.50
Y=0	.44	.56	.50
∑	(500)	(500)	(1000)

$d_{YX} = 12.0$
$\Phi_{XY} = 0.12$

Randtabelle Y nach W:

Y	W=1	W=0	∑
Y=1	.524	.476	.50
Y=0	.476	.524	.50
∑	(500)	(500)	(1000)

$d_{YW} = 4.8$
$\Phi_{WY} = 0.08$

Randtabelle W nach X:

W	X=1	X=0	∑
W=1	.70	.30	.50
W=0	.70	.70	.50
∑	(500)	(500)	(1000)

$d_{WX} =40.0$
$\Phi_{XW} = 0.40$

Nur X wirkt additiv auf Y, X und W sind nicht statistisch unabhängig

Generierte Daten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	196	66	84	154	500
Y=0	154	84	66	196	500
∑	350	150	150	350	1000

Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.56	.44	.56	.44	.50
Y=0	.44	.56	.44	.56	.50
n	(350)	(150)	(350)	(150)	(1000)

$d_{YX(W=1)}\%=12.0$ $d_{YX(W=0)}\%=12.0$
$\Phi_{XY(W=1)} = 0.12$ $\Phi_{XY(W=0)} = 0.12$

Randtabelle Y nach X:

Y	X=1	X=0	∑
Y=1	.56	.44	.50
Y=0	.44	.56	.50
∑	(500)	(500)	(1000)

$d_{YX} = 12.0$
$\Phi_{XY} = 0.12$

Randtabelle Y nach W:

Y	W=1	W=0	∑
Y=1	.524	.476	.50
Y=0	.476	.524	.50
∑	(500)	(500)	(1000)

$d_{YW} = 4.8$
$\Phi_{WY} = 0.048$

Randtabelle W nach X:

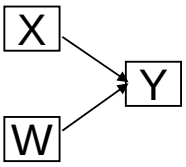
W	X=1	X=0	∑
W=1	.70	.30	.50
W=0	.70	.70	.50
∑	(500)	(500)	(1000)

$d_{WX} =40.0$
$\Phi_{XW} = 0.40$

Die bivariaten und konditionalen Beziehungen zwischen X und Y unterscheiden sich nicht von der Situation einer statistisch unabhängigen Drittvariable.
Allerdings zeigt diese Datenkonstellation bivariat einen Zusammenhang zwischen Y und W, der dadurch hervorgerufen wird, dass X sowohl mit W als auch mit Y zusammenhängt und so einen ***korrelierten Effekt*** zwischen Y und W bewirkt, der nicht kausal interpretiert werden kann und verschwindet, wenn Partialtabellen zwischen Y und W mit X als Drittvariable betrachtet werden.

3) X und W wirken beide additiv auf W, sind aber voneinander unabhängig

Die Daten sind so generiert worden, dass sowohl X als auch W auf Y additiv wirken, die beiden erklärenden Variablen aber statistisch unabhängig voneinander sind.



Generierte Daten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	180	150	105	75	510
Y=0	70	100	145	175	490
∑	250	250	250	250	1000

Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.72	.60	.42	.30	.51
Y=0	.28	.40	.58	.70	.49
n	(250)	(250)	(250)	(250)	(1000)

$d_{YX(W=1)}\%=12.0$ $d_{YX(W=0)}\%=12.0$
$\Phi_{XY(W=1)} = 0.127$ $\Phi_{XY(W=0)} = 0.125$

Randtabelle Y nach X:

Y	X=1	X=0	∑
Y=1	.57	.45	.51
Y=0	.43	.55	.49
∑	(500)	(500)	(1000)

$d_{YX} = 12.0$
$\Phi_{XY} = 0.12$

Randtabelle Y nach W:

Y	W=1	W=0	∑
Y=1	.66	.36	.51
Y=0	.34	.64	.49
∑	(500)	(500)	(1000)

$d_{YW} = 30.0$
$\Phi_{WY} = 0.30$

Randtabelle W nach X:

W	X=1	X=0	∑
W=1	.50	.50	.50
W=0	.50	.50	.50
∑	(500)	(500)	(1000)

$d_{WX} = 0$
$\Phi_{XW} = 0$

X und W wirken beide additiv auf W, sind aber voneinander unabhängig

Generierte Daten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	180	150	105	75	510
Y=0	70	100	145	175	490
∑	250	250	250	250	1000

Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.72	.60	.42	.30	.51
Y=0	.28	.40	.58	.70	.49
n	(250)	(250)	(250)	(250)	(1000)

$d_{YX(W=1)}\%=12.0$ $d_{YX(W=0)}\%=12.0$
$\Phi_{XY(W=1)} = 0.127$ $\Phi_{XY(W=0)} = 0.125$

Randtabelle Y nach X:

Y	X=1	X=0	∑
Y=1	.57	.45	.51
Y=0	.43	.55	.49
∑	(500)	(500)	(1000)

$d_{YX} = 12.0$
$\Phi_{XY} = 0.12$

Randtabelle Y nach W:

Y	W=1	W=0	∑
Y=1	.66	.36	.51
Y=0	.34	.64	.49
∑	(500)	(500)	(1000)

$d_{YW} = 30.0$
$\Phi_{WY} = 0.30$

Randtabelle W nach X:

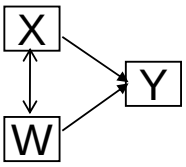
W	X=1	X=0	∑
W=1	.50	.50	.50
W=0	.50	.50	.50
∑	(500)	(500)	(1000)

$d_{WX} = 0$
$\Phi_{XW} = 0$

Wenn X und W statistisch unabhängig voneinander sind und beide auf Y additiv wirken, dann sind die bivariaten Effekte auf die abhängige Variable gleich den konditionalen (bzw. partiellen) Effekten. Dass dies auch für W gilt, wird deutlich, wenn die Rolle von W und X in der Drittvariablenkontrolle vertauscht werden und der konditionale Effekt von W auf Y in den durch X definierten Partialtabellen berechnet wird.
Im Datenbeispiel gilt $d_{YW}\% = d_{YW(X=1)}\% = d_{YW(X=0)}\% = d_{YW(X)}\% = 30.0$.

4) X und W wirken beide additiv auf W, X und W sind nicht unabhängig

Im Unterschied zur vorherigen Situation besteht nun zwischen X und W ein (ungerichteter) statistischer Zusammenhang.
Dies ist eine oft auftretende Konstellation bei Umfragedaten.



Generierte Daten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	252	90	63	105	510
Y=0	98	60	87	245	490
∑	350	150	150	350	1000

Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.72	.60	.42	.30	.51
Y=0	.28	.40	.58	.70	.49
n	(350)	(150)	(150)	(350)	(1000)

$d_{YX(W=1)}\%=12.0$ $d_{YX(W=0)}\%=12.0$
$\Phi_{XY(W=1)} = 0.118$ $\Phi_{XY(W=0)} = 0.116$

Randtabelle Y nach X:

Y	X=1	X=0	∑
Y=1	.63	.39	.51
Y=0	.47	.61	.49
∑	(500)	(500)	(1000)

$d_{YX} = 24.0$
$\Phi_{XY} = 0.24$

Randtabelle Y nach W:

Y	W=1	W=0	∑
Y=1	.684	.336	.51
Y=0	.316	.664	.49
∑	(500)	(500)	(1000)

$d_{YW} = 34.8$
$\Phi_{WY} = 0.348$

Randtabelle W nach X:

W	X=1	X=0	∑
W=1	.70	.30	.50
W=0	.30	.70	.50
∑	(500)	(500)	(1000)

$d_{WX} = 40.0$
$\Phi_{XW} = 0.40$

X und W wirken beide additiv auf W, X und W sind nicht unabhängig

Generierte Daten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	252	90	63	105	510
Y=0	98	60	87	245	490
∑	350	150	150	350	1000

Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.72	.60	.42	.30	.51
Y=0	.28	.40	.58	.70	.49
n	(350)	(150)	(150)	(350)	(1000)

$d_{YX(W=1)}\%=12.0$ $d_{YX(W=0)}\%=12.0$
$\Phi_{XY(W=1)} = 0.118$ $\Phi_{XY(W=0)} = 0.116$

Randtabelle Y nach X:

Y	X=1	X=0	∑
Y=1	.63	.39	.51
Y=0	.47	.61	.49
∑	(500)	(500)	(1000)

$d_{YX} = 24.0$
$\Phi_{XY} = 0.24$

Randtabelle Y nach W:

Y	W=1	W=0	∑
Y=1	.684	.336	.51
Y=0	.316	.664	.49
∑	(500)	(500)	(1000)

$d_{YW} = 34.8$
$\Phi_{WY} = 0.348$

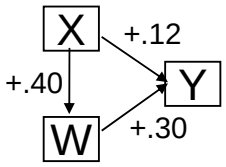
Randtabelle W nach X:

W	X=1	X=0	∑
W=1	.70	.30	.50
W=0	.30	.70	.50
∑	(500)	(500)	(1000)

$d_{WX} = 40.0$
$\Phi_{XW} = 0.40$

Wenn X und W nicht statistisch unabhängig voneinander sind und beide auf Y additiv wirken, dann unterscheiden sich die bivariaten Effekte von den konditionalen (bzw. partiellen) Effekten.
Es wird dann auch davon gesprochen, dass die bivariaten Effekte ***konfundiert*** sind.
Wie Konfundierung entsteht, wird deutlich, wenn man die kausalen Beziehung zwischen den drei Variablen X und W näher betrachtet.

Konfundierung

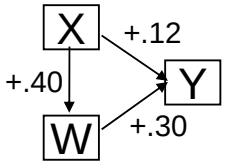


Die Daten wurden so generiert, dass X sowohl auf Y wie auf W wirkt und W zusätzlich auf X.
Um einen vollständigen Eindruck der Datengenerierung zu bekommen, werden die Pfeile mit den Effektstärken (hier gemessen über Anteilsdifferenzen) beschriftet.

Aus dem Pfaddiagramm lässt sich nun folgendes entnehmen:

- Wenn X ansteigt, d.h. bei der in den Tabellen wiedergegebenen Kodierungen von X=0 auf X=1 wechselt, dann bewirkt dies eine Erhöhung des Anteils von Y=1 um +0.12. Dies ist der sogenannte ***direkte Effekt***, der bei Drittvariablenkontrolle mit W als Kontrollvariable sichtbar wird.
- Da X auch auf W wirkt, bewirkt eine Erhöhung von X zusätzlich durch einen direkten Effekt auf W eine Erhöhung des Anteils von W=1 um +0.4.
- Ein Veränderung von W durch einen Wechsel von W=0 auf W=1 bewirkt durch den direkten Effekt von W auf Y einen Anstieg des Anteils von Y=1 um 0.3.
  Durch eine Veränderung in X steigt der Anteil der Fälle von W=1 um 0.4 oder 40% an. Diese zusätzlichen 40% Fälle bewirken nun einen Anstieg des Anteils von Y=1. Da beim Wechsel von X=0 auf X=1 nicht alle, sondern nur 40% von W=0 auf W=1 wechseln, steigt der Anteil von Y=1 nicht um 0.3 sondern nur um 40% von 0.3, also um 0.4×0.3 = 0.12 oder 12%. Dies ist der sogenannte ***indirekte Effekt*** von X auf Y, der über W vermittel wird. **W** wird in diesem Zusammenhang auch als ***Mediatorvariable*** für den Effekt von X auf Y bezeichnet.

Konfundierung



- Insgesamt bewirkt ein Anstieg von X=0 auf X=1 einen sogenannten ***totalen Effekt*** von +0.24 oder 24%, wobei der totale Effekt die Summe aus dem direkten und den indirekten Effekten ist.

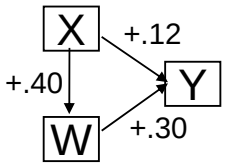
Betrachtet man die bivariate Kreuztabelle von Y und X, so zeigt sich, dass der bivariate Effekt gleich dem totalen Effekt ist.

Eine Ursache der Konfundierung besteht also darin, dass im bivariaten Zusammenhang nicht nur der direkte Effekt berücksichtigt wird, sondern auch mögliche indirekte Effekte. Wenn X und W statistisch unabhängig voneinander sind (wie in den Datenkonstellationen (1) und (3), dann gibt es keinen indirekten Effekt und der bivariate additive Effekt ist gleich dem partiellen additiven Effekt bei Drittvariablenkontrolle.

Auffallend ist weiter, dass auch der bivariate Effekt d_{YW}% der Drittvariable mit 34.8 Prozentpunkten deutlich größer ist als die konditionalen Effekte, die gleich groß sind und damit einen partiellen Effekt von $d_{YW(X)}$% = 30.0 Prozentpunkte aufweisen. Der partielle Effekt von W auf Y bei Drittvariablenkontrolle durch X ist also wieder der direkte Effekt von W auf Y.

- Tatsächlich gibt es eine Gemeinsamkeit von W und Y, die nicht auf den Effekt von W auf Y zurückzuführen ist. Beide Variablen werden nämlich durch X beeinflusst. Diese Gemeinsamkeit bewirkt eine Beziehung zwischen den Variablen und damit eine zusätzliche Quelle für den statistischen bivariaten Zusammenhang. Man spricht hier von einen ***korrelierten Effekt***, wobei das Wort „Effekt“ hier ungünstig ist, da es tatsächlich eher um eine gemeinsame Ursache (X) geht, die sowohl auf W wie auf Y wirkt.

Konfundierung



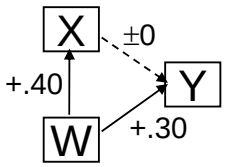
Im Beispiel berechnet sich der korrelierte Effekt als Produkt aus dem direkten Effekt von X auf Y und von X auf W und beträgt somit 0.12×0.4 = 0.048.
Der bivariate Effekt von W auf Y ist die Summe aus den direkten Effekt und dem korrelierten Effekt und beträgt somit 0.30 + 0.048 = 0.348.

Generell gilt für sogenannte ***lineare Modelle***, bei denen nur additive Effekte und Beziehungen bestehen, dass die Höhe des bivariaten Zusammenhangs stets die Summe aus dem direkten Effekt und allen indirekten und korrelierten Effekten ist.

Einen korrelierten Effekt hatten wir auch bereits in der Datenkonstellation 2) beobachtet, bei der W keinen direkten Effekt auf Y hat, aber mit X zusammenhängt. Die Höhe des korrelierten Effekt ist wiederum das Produkt aus dem Effekt von X auf Y und dem Effekt zwischen X und W, im Beispiel 0.12 ×04 = 0.048.
Korrelierte Effekte gibt es auch, wenn die Richtung der Beziehung zwischen zwei Variablen nicht klar ist. Der Unterschied zu indirekten Effekten besteht ausschließlich darin, dass bei indirekten Effekten klar ist, dass eine Wirkung über Mediatorvariablen vermittelt wird.

Wenn direkte und indirekte bzw. korrelierte Effekte in die gleiche Richtung wirken, also wie im Beispiel alle positiv sind oder alle negativ sind, dann sind die partiellen Effekte stets kleiner als die bivariaten Effekte.
Dies ist eine Situation, wie sie in der Sozialforschung **sehr oft** zu beobachten ist.

5) Scheinkausalität



Ein Spezialfall von Konfundierung liegt vor, wenn ein bivariater Effekt besteht, aber die konditionalen Effekte Null sind. Wenn der bivariate Effekt ausschließlich Folge eines korrelierten Effekt ist, spricht man auch von ***Scheinkausalität***. Im Beispiel besteht ein korrelierter Effekt zwischen X und Y in Höhe von 0.4×0.3 = 0.12, der durch die gemeinsame Ursache W bewirkt wird.

Generierte Daten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	210	90	45	105	450
Y=0	140	60	105	245	550
∑	350	150	150	350	1000

Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.60	.60	.30	.30	.45
Y=0	.40	.40	.70	.70	.55
n	(350)	(150)	(150)	(350)	(1000)

$d_{YX(W=1)}\%= 0$ $d_{YX(W=0)}\%= 0$
$\Phi_{XY(W=1)} = 0$ $\Phi_{XY(W=0)} = 0$

Randtabelle Y nach X:

Y	X=1	X=0	∑
Y=1	.51	.39	.45
Y=0	.49	.61	.55
∑	(500)	(500)	(1000)

$d_{YX} = 12.0$
$\Phi_{XY} = 0.121$

Randtabelle Y nach W:

Y	W=1	W=0	∑
Y=1	.60	.30	.45
Y=0	.40	.70	.55
∑	(500)	(500)	(1000)

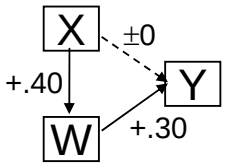
$d_{YW} = 30.0$
$\Phi_{WY} = 0.302$

Randtabelle X nach W:

X	W=1	W=0	∑
X=1	.70	.30	.50
X=0	.30	.70	.50
∑	(500)	(500)	(1000)

$d_{XW} = 40.0$
$\Phi_{XW} = 0.40$

6) Interpretation eines Effektes über eine Mediatorvariable



Die gleiche Datenkonstellation wie die Scheinkausalität ergibt sich, wenn X ausschließlich über eine Kausalkette indirekter Effekte auf Y wirkt. Werden Partialtabellen über die Ausprägungen der Mediatorvariable W gebildet, wird der bivariate Effekt durch die ***Kausalkette*** über W ***interpretiert***. Da es keinen direkten Effekt gibt, sind die konditionalen Effekte wieder Null. Im Beispiel ergibt sich so der bivariate Effekt als 0.4×0.3 = 0.12.

Generierte Daten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	210	90	45	105	450
Y=0	140	60	105	245	550
∑	350	150	150	350	1000

Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.60	.60	.30	.30	.45
Y=0	.40	.40	.70	.70	.55
n	(350)	(150)	(150)	(350)	(1000)

$d_{YX(W=1)}\%= 0$ $d_{YX(W=0)}\%= 0$
$\Phi_{XY(W=1)} = 0$ $\Phi_{XY(W=0)} = 0$

Randtabelle Y nach X:

Y	X=1	X=0	∑
Y=1	.51	.39	.45
Y=0	.49	.61	.55
∑	(500)	(500)	(1000)

$d_{YX} = 12.0$
$\Phi_{XY} = 0.121$

Randtabelle Y nach W:

Y	W=1	W=0	∑
Y=1	.60	.30	.45
Y=0	.40	.70	.55
∑	(500)	(500)	(1000)

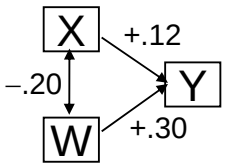
$d_{YW} = 30.0$
$\Phi_{WY} = 0.302$

Randtabelle W nach X:

W	X=1	X=0	∑
W=1	.70	.30	.50
W=0	.30	.70	.50
∑	(500)	(500)	(1000)

$d_{XW} = 40.0$
$\Phi_{XW} = 0.40$

7) Suppression



Wenn die direkten Effekte positiv, die indirekten bzw. korrelierten Effekte dagegen negativ wirken oder umgekehrt, dann spricht man von ***Suppression***, d.h. der bivariate Effekt ist kleiner als die konditionalen bzw. partiellen Effekte. Im Beispiel führt der negative korrelierte Effekt von X auf Y über die ***Suppressorvariable*** W dazu, dass der bivariate Effekt halb so hoch ist wie der direkte Effekt: +.12 –0.2×0.3 = 0.06.

Generierte Daten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	144	180	126	60	510
Y=0	56	120	174	140	490
∑	200	300	300	200	1000

Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.72	.60	.42	.30	.51
Y=0	.28	.40	.58	.70	.49
	(200)	(300)	(300)	(200)	(1000)

$d_{YX(W=1)}\%= 12.0$ $d_{YX(W=0)}\%= 12.0$
$\Phi_{XY(W=1)} = 0.123$ $\Phi_{XY(W=0)} = 0.122$

Randtabelle Y nach X:

Y	X=1	X=0	∑
Y=1	.54	.48	.51
Y=0	.46	.52	.49
∑	(500)	(500)	(1000)

$d_{YX} = 6.0$
$\Phi_{XY} = 0.06$

Randtabelle Y nach W:

Y	W=1	W=0	∑
Y=1	.648	.372	.51
Y=0	.352	.628	.49
∑	(500)	(500)	(1000)

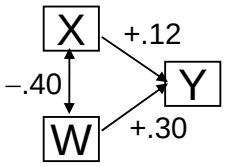
$d_{YW} = 27.6$
$\Phi_{WY} = 0.276$

Randtabelle W nach X:

W	X=1	X=0	∑
W=1	.40	.60	.50
W=0	.60	.40	.50
∑	(500)	(500)	(1000)

$d_{WX} = –20.0$
$\Phi_{XW} = –0.20$

8) Scheinbare Nichtbeziehung



Wenn indirekte oder korrelierte Effekte genau so groß sind wie der direkte Effekt, dann besteht bivariat statistische Unabhängigkeit. Man bezeichnet diese extreme Form der Suppression auch als ***scheinbare Nichtbeziehung***. Im Beispiel hebt der negative korrelierte Effekt von –0.4×0.3 = –0.12 den direkten Effekt von +0.12 exakt auf.

Generierte Daten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	108	210	147	45	510
Y=0	42	140	203	105	490
∑	150	350	350	150	1000

Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.72	.60	.42	.30	.51
Y=0	.28	.40	.58	.70	.49
	(150)	(350)	(350)	(150)	(1000)

$d_{YX(W=1)}$%= 12.0 $d_{YX(W=0)}$%= 12.0
$\Phi_{XY(W=1)}$ = 0.114 $\Phi_{XY(W=0)}$ = 0.113

Randtabelle Y nach X:

Y	X=1	X=0	∑
Y=1	.51	.51	.51
Y=0	.49	.49	.49
∑	(500)	(500)	(1000)

d_{YX} =0
Φ_{XY} = 0

Randtabelle Y nach W:

Y	W=1	W=0	∑
Y=1	.636	.384	.51
Y=0	.364	.616	.49
∑	(500)	(500)	(1000)

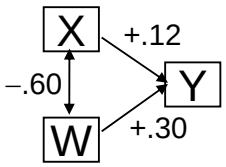
d_{YW} = 25.2
Φ_{WY} = 0.252

Randtabelle W nach X:

W	X=1	X=0	∑
W=1	.30	.70	.50
W=0	.70	.30	.50
∑	(500)	(500)	(1000)

d_{WX} = –40.0
Φ_{XW} = –0.40

9) Verzerrung



Von ***Verzerrung*** spricht man, wenn der bivariate Effekt ein anderes Vorzeichen hat als die konditionalen bzw. partiellen Effekte. Dies tritt auf, wenn die indirekten bzw. korrelierten Effekte größer sind als die direkten Effekte und ein anderes Vorzeichen aufweisen. Im Beispiel ist der indirekte Effekt –0.6×0.3 = –0.18, der direkte Effekt dagegen +0.12. Der bivariate Effekt ist daher 0.12 –0.18 = –0.06.

Generierte Daten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	72	240	168	30	510
Y=0	28	160	232	70	490
∑	100	400	400	100	1000

Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.72	.60	.42	.30	.51
Y=0	.28	.40	.58	.70	.49
	(100)	(400)	(400)	(100)	(1000)

$d_{YX(W=1)}$%= 12.0 $d_{YX(W=0)}$%= 12.0
$\Phi_{XY(W=1)}$ = 0.099 $\Phi_{XY(W=0)}$ = 0.098

Randtabelle Y nach X:

Y	X=1	X=0	∑
Y=1	.48	.54	.51
Y=0	.52	.46	.49
∑	(500)	(500)	(1000)

d_{YX} = –6.0
Φ_{XY} = –0.06

Randtabelle Y nach W:

Y	W=1	W=0	∑
Y=1	.624	.396	.51
Y=0	.376	.604	.49
∑	(500)	(500)	(1000)

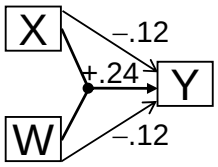
d_{YW} = 22.8
Φ_{WY} = 0.228

Randtabelle W nach X:

W	X=1	X=0	∑
W=1	.20	.80	.50
W=0	.80	.20	.50
∑	(500)	(500)	(1000)

d_{WX} = –60.0
Φ_{XW} = –0.60

10) Interaktionseffekte



Wenn sich die konditionalen Beziehungen in den Partialtabellen unterscheiden, dann besteht eine ***Interaktion*** zwischen erklärenden Variablen bei ihrem Effekt auf die abhängige Variable. Im Beispiel beträgt der konditionale Effekt von X auf Y +0.12, wenn W=1, und –0.12, wenn W=0.

Generierte Daten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	140	110	110	140	500
Y=0	110	140	140	110	500
∑	250	250	250	250	1000

Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.56	.44	.44	.56	.50
Y=0	.56	.44	.56	.44	.50
	(250)	(250)	(250)	(250)	(1000)

$d_{YX(W=1)}\% = 12.0$ $d_{YX(W=0)}\% = -12.0$
$\Phi_{XY(W=1)} = 0.12$ $\Phi_{XY(W=0)} = -0.12$

Randtabelle Y nach X:

Y	X=1	X=0	∑
Y=1	.50	.50	.50
Y=0	.50	.50	.50
∑	(500)	(500)	(1000)

$d_{YX} = 0$
$\Phi_{XY} = 0$

Randtabelle Y nach W:

Y	W=1	W=0	∑
Y=1	.50	.50	.50
Y=0	.50	.50	.50
∑	(500)	(500)	(1000)

$d_{YW} = 0$
$\Phi_{WY} = 0$

Randtabelle W nach X:

W	X=1	X=0	∑
W=1	.50	.50	.50
W=0	.50	.50	.50
∑	(500)	(500)	(1000)

$d_{WX} = 0$
$\Phi_{XW} = 0$

Interaktionseffekte

Generierte Daten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	140	110	110	140	500
Y=0	110	140	140	110	500
∑	250	250	250	250	1000

Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.56	.44	.44	.56	.50
Y=0	.56	.44	.56	.44	.50
	(250)	(250)	(250)	(250)	(1000)

$d_{YX(W=1)}\% = 12.0$ $d_{YX(W=0)}\% = -12.0$
$\Phi_{XY(W=1)} = 0.12$ $\Phi_{XY(W=0)} = -0.12$

Randtabelle Y nach X:

Y	X=1	X=0	∑
Y=1	.50	.50	.50
Y=0	.50	.50	.50
∑	(500)	(500)	(1000)

$d_{YX} = .0$
$\Phi_{XY} = 0$

Randtabelle Y nach W:

Y	W=1	W=0	∑
Y=1	.50	.50	.50
Y=0	.50	.50	.50
∑	(500)	(500)	(1000)

$d_{YW} = 0$
$\Phi_{WY} = 0$

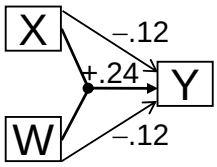
Randtabelle W nach X:

W	X=1	X=0	∑
W=1	.50	.50	.50
W=0	.50	.50	.50
∑	(500)	(500)	(1000)

$d_{WX} = 0$
$\Phi_{XW} = 0$

Beim Vorliegen eines Interaktionseffektes zwischen X und W sind auch die konditionalen Effekte von W verschieden, im Beispiel ist $d_{YW(X=1)}\% = 0.12$ und $d_{YW(X=0)}\% = -0.12$.
Das Beispiel zeigt auch, dass eine (trivariate) statistische Beziehung zwischen den drei Variablen X, Y und W bestehen kann, obwohl bivariat alle Variablen jeweils unabhängig voneinander sind, was hier Folge davon ist, dass sich die positiven und negativen (konditionalen) Effekte bivariat jeweils ausgleichen und zusätzlich die beiden erklärenden Vartiablen unabhängig voneinander sind.

Interaktionseffekte



Im Pfaddiagramm ist der Interaktionseffekt so dargestellt, dass es neben dem direkten Effekt von X bzw. W auf Y noch einen weiteren Pfeil gibt, der zwei Anfänge bei den miteinander interagierenden erklärenden Variablen hat, die wie beim Buchstaben „Y“ zusammenlaufen, bevor sie auf die abhängige Variable weisen. Die beiden direkten Effekte von X und W auf Y in Höhe von jeweils –0.12 werden bisweilen auch als ***Haupteffekte*** bezeichnet, der gemeinsame Effekt ist dann der ***Interaktionseffekt***.

Das Pfadmodell zeigt, dass der gemeinsame Interaktionseffekt als direkter Effekt des Produktes der beiden erklärenden Variablen X und Y betrachtet werden kann:
Wenn W=0 ist, ergibt sich so der negative konditionale Effekt von X auf Y:

$$-0.12\times X + 0.24\times(X\times W) = -0.12\times X + 0.24\times(X\times 0) = -0.12.$$

Wenn W=1, ergibt sich dagegen ein positiver konditionale Effekt:

$$-0.12\times X + 0.24\times(X\times W) = -0.12\times X + 0.24\times(X\times 1) = +0.12.$$

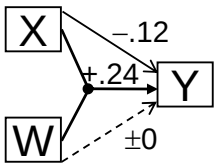
Ganz symmetrisch ergeben sich auch die konditionalen Effekte von W auf Y wenn X=0 als:

$$-0.12\times W + 0.24\times(X\times W) = -0.12\times W + 0.24\times(0\times W) = -0.12,$$

bzw. wenn X=1 als:

$$-0.12\times W + 0.24\times(X\times W) = -0.12\times W + 0.24\times(1\times W) = +0.12.$$

Interaktionseffekte



Bei Vorliegen eines Interaktionseffekt besteht immer eine Symmetrie zwischen erklärender Variable X und Kontrollvariable W: Wenn X bei unterschiedlichen Werten von W unterschiedliche Effekte auf Y aufweist, dann weist auch W unterschiedliche Effekte auf Y bei verschiedenen Werten von X auf.

Wenn im obigen Pfaddiagramm der Haupteffekt von W auf Y Null ist, dann bedeutet das, dass in der Partialtabelle von X=1 kein Zusammenhang zwischen W und Y besteht, bei X=1 dagegen ein positiver Effekt von +0.24:

Bedingte relative Häufigkeiten:

	X=1		X=0		
	W=1	W=0	W=1	W=0	∑
Y=1	.56	.32	.44	.44	.50
Y=0	.56	.58	.66	.66	.50
	(250)	(250)	(250)	(250)	(1000)

$d_{YW(X=1)}\%= 24.0$ $d_{YW(X=0)}\%= 0$
$\Phi_{WY(X=1)} = 0.24$ $\Phi_{WY(X=0)} = 0$

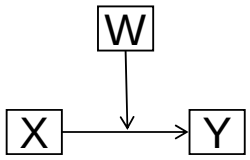
Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.56	.44	.32	.44	.50
Y=0	.56	.44	.68	.66	.50
	(250)	(250)	(250)	(250)	(1000)

$d_{YX(W=1)}\%= 12.0$ $d_{YX(W=0)}\%= -12.0$
$\Phi_{XY(W=1)} = 0.12$ $\Phi_{XY(W=0)} = -0.12$

Aus inhaltlich-theoretischen Gründen wird bisweilen die Symmetrie zwischen erklärender Variable und Drittvariable bei Interaktionseffekten aufgegeben und argumentiert, dass die Drittvariable den Effekt der erklärenden Variable kausal beeinflusst.

Interaktionseffekte



Im Pfaddiagram kann das so ausgedrückt werden, dass die Drittvariable W einen Effekt auf den Pfad von X nach Y hat.
Bei dieser Sichtweise spricht man auch davon, dass W eine ***Moderator-variable*** ist, die den Effekt der erklärenden Variable X auf die abhängige Variable Y moderiert.

Ob eine solche asymmetrische Sichtweise, bei der zwischen erklärender Variable und Moderatorvariable unterschieden wird, oder eine symmetrische Sichtweise mit zwei interaktiv zusammenwirkenden erklärenden Variablen zutrifft, lässt sich anhand der Daten allein nicht entscheiden.
Anders kann die Situation aussehen, wenn ein experimentelles Untersuchungsdesign vorliegt und und X und W unabhängig voneinander in einer bestimmten zeitlichen Abfolge manipuliert werden können. Bei einem reinen Moderatoreffekt sollte der Moderator W bei X=0 Y nicht beeinflussen.

Kausalanalysen

Über bivariate Zusammenhangsanalysen hinausgehende multivariate Analysen von mehr als zwei Variablen werden oft damit begründet, dass nur so die kausalen Zusammenhänge zwischen zwei Variablen sichtbar werden.
In diesem Sinne hat sich so im Ausgangsbeispiel zu dieser Lerneinheit gezeigt, dass der bivariat bestehende Zusammenhang zwischen Telefonbesitz und Einstellung zu Schwangerschaftsabbrüchen bei der dreidimensionalen Betrachtung mit Erhebungsgebiet als Kontrollvariable nicht sichtbar ist, also die bivariate Beziehung offenbar ein Beispiel für Scheinkausalität ist. Mit Hilfe multivariater Analysen ist es so möglich, Scheinkausalitäten aufzudecken und postulierte kausale Zusammenhänge empirisch zu widerlegen.
Man spricht auch von ***Kausalanalysen***, wenn in multivariaten Analysen überprüft wird, ob die beobachteten Beziehungen zwischen Variablen so auftreten, wie es ein Pfadmodell der postulierten Kausalstruktur erwarten lässt.

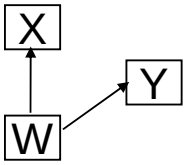
Zeigt sich bei einer solchen Analyse, dass eine empirisch beobachtete Datenkonstellation nicht mit der postulierten Kausalstruktur übereinstimmt, kann daraus geschlossen werden, dass die postulierte Kausalstruktur vermutlich falsch ist. Umgekehrt lassen sich durch Kausalanalysen allerdings zutreffende kausale Vermutungen nicht mit hinreichender Sicherheit bestätigen.

Dies hat mehrere Gründe:

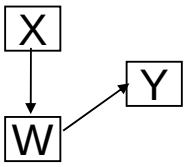
- Zum einen können unterschiedliche Kausalordnungen empirisch die gleiche Form aufweisen. So haben z.B. Scheinkausalität und Interpretation die gleichen empirischen Konsequenzen.

Kausalanalysen

Scheinkausalität:
W beeinfluss
Sowohl X wie Y



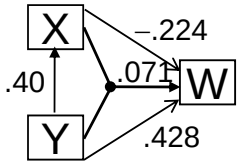
Interpretation:
X beeinflusst
indirekt über W Y

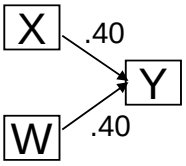


Wenn unterschiedliche kausale Prozesse die gleichen empirischen Konsequenzen haben, wird dies als ***Beobachtungsäquivalenz*** bezeichnet. Bei zwei beobachtungsäquivalenten Modellen lässt sich nicht ausschließlich anhand der Daten entscheiden, welches Modell zutrifft.
Erst mit Hilfe zusätzlicher Informationen lassen sich bisweilen doch Aussagen treffen, welches beobachtungsäquivalente Modell zutrifft. Wenn in einem experimentellen Untersuchungsdesign der Wert von W durch das Treatment festgelegt wird, lässt sich beobachten, ob sich nur Y (wie bei der Interpretation) oder auch X (wie bei Scheinkausalität) verändert. Für aussagekräftigere Kausalprüfungen bedarf es daher neben den Daten zusätzliche Informationen über das Forschungsdesign.

- Dies ist auch notwendig, um gravierende Fehlinterpretationen zu vermeiden, die sich insbesondere dadurch ergeben können, dass anstelle der tatsächlichen kausalen Beziehungen eine ganz andere Kausalstruktur angenommen wird, die dann auch zu einer ganz anderen Dateninterpretation führen kann. Als Beispiel kann eine Datenkonstellation herangezogen werden, bei der zwei statistisch unabhängige Variablen X und W auf Y mit $d_{YX(W)}\% = d_{YW(X)}\% = 40$ erhöhen. Das Beispiel zeigt die Folgen, wenn fälschlicherweise W als abhängige Variable betrachtet wird, die durch X erklärt wird, wobei Y Kontrollvariable ist.

Kausalanalysen





Fehlspezifikation:

	Y=1		Y=0		
	X=1	X=0	X=1	X=0	∑
W=1	216	180	24	180	600
W=0	80	24	80	216	400
∑	296	204	104	396	1000

Korrekte Spezifikation:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	216	180	80	24	500
Y=0	24	180	80	216	500
∑	240	360	160	240	1000

Bedingte relative Häufigkeiten:

	Y=1		Y=0		
	X=1	X=0	X=1	X=0	∑
W=1	.730	.882	.231	.455	.60
W=0	.270	.118	.769	.345	.40
n	(296)	(204)	(104)	(396)	(1000)

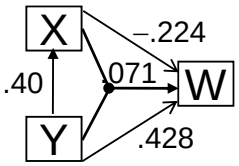
$d_{YX(W=1)}\% = -15.3$ $d_{YX(W=0)}\% = -22.4$
$\Phi_{XY(W=1)} = -0.185$ $\Phi_{XY(W=0)} = -0.185$

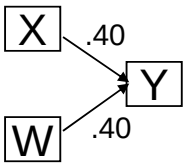
Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.90	.50	.50	.10	.50
Y=0	.10	.50	.50	.90	.50
n	(240)	(360)	(160)	(240)	(1000)

$d_{YX(W=1)}\% = 40.0$ $d_{YX(W=0)}\% = 40.0$
$\Phi_{XY(W=1)} = 0.414$ $\Phi_{XY(W=0)} = 0.447$

Auf der rechten Seite ist die korrekte Kausalstruktur, auf der linken Seite die unzutreffende Kausalstruktur abgebildet. Wird bei der Zusammenhangsanalyse eine unzutreffende Kausalstruktur angenommen, spricht man von einer ***Fehlspezifikation***.





Als Folge der Fehlspezifikation ergibt sich der falsche Eindruck, das X und Y mit einem Interaktionseffekt W beeinflussen. Die empirischen Daten weisen nicht darauf hin, dass die bei der Fehlspezifikation geschätzten Beziehungen Artefakte sind und die Realität (d.h. hier der wahre datengenerierende Prozess) nicht korrekt erfassen.

- Die Datenkonstellation ist auch nicht immer eindeutig. So ist zwar sehr oft zu beobachten, dass bei Drittvariablenkontrolle die Höhe der bivariaten Zusammenhänge abnimmt. Allerdings verschwinden Effekte nur selten vollkommen, wie das im Idealfall der Scheinkausalität oder der Interpretation der Fall sein sollte.
  Dann ist die Datenkonstellationen nicht so klar, dass eindeutig eine der oben dargestellten Konstellationen sichtbar ist. Betrachtet man etwa die Daten des Eingangsbeispiels, lassen die Daten den Schluss zu, dass zwischen Telefonbesitz und Rigidität bei der Bewertung von Schwangerschaftsabbrüchen Scheinkausalität (oder Interpretation durch die Region) vorliegt, weil bei Kontrolle durch die Region der Zusammenhang nicht besteht. Da der verbleibende (sehr geringe) Zusammenhang ein anderes Vorzeichen aufweist, können die Daten auch als Beispiel für eine Verzerrung dienen. Und da sich die konditionalen Zusammenhänge mit Werten von –4.3 und –0.8 unterscheiden, könnte man auch von einem Interaktionseffekt ausgehen.

Abtreibung, wenn die Frau es will, ...	Alte Bundesländer Telefonanschluss? ja	nein	Neue Bundesländer Telefonanschluss? ja	nein	Insgesamt Telefonanschluss ja	nein
... sollte verboten sein	58.5%	62.8%	28.9%	29.7%	54.7%	33.0%
... sollte erlaubt sein	41.5%	37.2%	71.1%	70.3%	45.3%	67.0%
	(2026)	(78)	(305)	(704)	(2331)	(782)
(Quelle: ALLBUS 1992)	d_{YX}% = –4.3		d_{YX}% = –0.8		d_{YX}% = +21.7	

Eine gewisse Hilfestellung bieten hier statistische Tests. Mit ihrer Hilfe können und müssen auch in multivariaten Datenanalysen beobachtete statistische Zusammenhänge abgesichert werden.

So ergeben im Beispiel die Pearsons Chiquadrat-Statistiken, dass in den beiden Partialtabellen keine signifikanten Zusammenhänge bestehen: In den alten Bundesländern beträgt die Teststatistik 0.58, in den neuen Bundesländern 0.07. Bei jeweils df=1 Freiheitsgrad ist der kritische Wert bei einer Irrtumswahrscheinlichkeit von 5% 3.84. Da die Teststatistiken kleiner sind, können die Nullhypothesen, dass es in den Partialtabellen keinen Zusammenhang gibt, nicht ausgeschlossen werden.

Abtreibung erlaubt?	Telefon? ja	nein	Summe	Telefon? ja	nein	Summe
verboten	1185	49	1234	88	209	297
erlaubt	841	29	870	217	495	712
Summe	2026	78	2104	305	704	1009
			χ^2 =0.58			χ^2 =0.07

Kausalanalysen

Die Interpretation von Datenanalysen kommt vor allem nicht ohne theoretischer Überlegungen aus. So ist es im Beispiel des Telefonbesitzes und der Haltung zu Schwangerschaftsabbrüchen zwar offensichtlich, dass der Telefonbesitz nicht die Einstellung zu Abtreibungen kausal beeinflusst. In anderen Situationen mag man aber eher aus einem beobachteten Zusammenhang (einer Korrelation) auf die Existenz einer kausalen Beziehung schließen.

Bei der Beurteilung von Zusammenhängen muss neben der Kontrolle durch geeignete statistische Tests und der Berücksichtigung des Untersuchungsdesigns insbesondere auch die Bedeutung der Variablen bzw. die Operationalisierung der Konstrukte bedacht werden.

So kann im Allbus-Beispiel auch das Erhebungsgebiet keinen kausalen Einfluss auf die Haltung zu Schwangerschaftsabbrüchen haben. Das Erhebungsgebiet kann jedoch als Indikator für Sozialisationsbedingungen in der ehemaligen DDR interpretiert werden. In der DDR galt ein liberaleres Abtreibungsrecht, das in der Bevölkerung akzeptiert war. Zudem war der Einfluss der Kirchen geringer, da der Anteil der Mitglieder geringer war. Es ist daher damit zu rechnen, dass Personen, die in der DDR aufgewachsen sind, eine liberalere Einstellung haben als Personen aus den alten Bundesländern.
Möglicherweise wirkt sich dies auch auf Personen auf, die nach der Auflösung der DDR in die neuen Bundesländer zogen und etwa über Gespräche Einstellungen der bereits zu DDR-Zeiten dort lebenden Personen übernehmen.

Bei solchen theoretischen Überlegungen ist unbedingt zu unterscheiden, was reine ad-hoc-Vermutung ist und was empirisch abgesichert ist bzw. empirisch abzusichern ist. So kann etwa durch aktuellere Daten geprüft werden, ob es auf dem Gebiet der DDR noch heute andere Haltungen zu Schwangerschaftsabbrüchen gibt als in den alten Ländern.

Kausalanalysen

	1992		2006	
Abtreibung, wenn	Erhebungsgebiet			
die Frau es will, ...	West	Ost	West	Ost
... sollte verboten sein	58.7%	29.4%	65.2%	37.8%
... sollte erlaubt sein	41.3%	70.6%	34.8%	62.2%
	(2104)	(1009)	(2148)	(1086)
(Quelle: ALLBUS 1992&2006)	$d_{YX}\% = 29.3$		$d_{YX}\% = 27.4$	

Tatsächlich zeigt der Vergleich der Daten des Allbus 1992 mit denen des Allbus 2006, dass weiterhin eine große Ost-West-Differenz bei der Beurteilung von Schwangerschaftsabbrüchen zu beobachten ist. Die Prozentsatzdifferenz ist nur ganz geringfügig gesunken. Dies spricht dafür, dass sich zumindest bei der Bewertung von Schwangerschaftsabbrüchen bestehende Vorstellungen aus der DDR bis 2006 erhalten habe,
Auffällig ist, dass 2006 in Ost wie West ein etwas höherer Anteil von Personen für ein Verbot sind als 2002. Insgesamt hat also die Ridigität in dieser Frage zugenommen und zwar etwa genau stark bei Personen aus dem Gebiet der ehemaligen DDR wie bei Personen aus den alten Ländern.
Da 2006 alle Befragten über ein Telefon verfügten, kann nicht mehr wie 1992 der Frage nachgegangen werden, ob sich Telefonbesitzer und Personen ohne Telefon in ihrer Einstellung unterscheiden.

Auswahl des Analysemodells

Generierte Daten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	252	90	63	105	510
Y=0	98	60	87	245	490
∑	350	150	150	350	1000

α = 1.714 α = 1.690

Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.72	.60	.42	.30	.51
Y=0	.28	.40	.58	.70	.49
n	(350)	(150)	(150)	(350)	(1000)

$d_{YX(W=1)}$%=12.0 $d_{YX(W=0)}$%=12.0
$\Phi_{XY(W=1)}$ = 0.118 $\Phi_{XY(W=0)}$ = 0.116

Gerade bei Prüfung von Vermutungen über kausale Zusammenhänge kann auch das Analysemodell über das Ergebnis entscheiden.
So können die beispielhaften Datenkonstellationen auch mittel multiplikativer Zusammenhangsmaße analysiert werden. Solange statistische Unabhängigkeit in den (Partial-) Tabellen besteht, unterscheiden sich die Ergebnisse nicht. Unterschiede kann es jedoch geben, wenn Zusammenhänge auftreten.
So zeigt sich bei der Datenkonstellation 4), bei der X und W additiv auf Y wirken und die beiden erklärenden Variablen einen positiven Zusammenhang aufweisen, dass bei Drittvariablenkontrolle die konditionalen Prozentsatzdifferenzen gleich groß sind. Werden stattdessen die multiplikativen Odd-Ratios als Zusammenhangsmaße berechnet, dann ist das Odds-Ratio mit 1.714 in der linken Partialtabelle (W=1) etwas größer als das Odds-Ratio in der rechten Partialtabelle (W=0) mit 1.690.

Auswahl des Analysemodells

Generierte Daten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	252	90	63	105	510
Y=0	98	60	87	245	490
∑	350	150	150	350	1000

α = 1.714 α = 1.690

Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.72	.60	.42	.30	.51
Y=0	.28	.40	.58	.70	.49
n	(350)	(150)	(150)	(350)	(1000)

$d_{YX(W=1)}$%=12.0 $d_{YX(W=0)}$%=12.0
$\Phi_{XY(W=1)}$ = 0.118 $\Phi_{XY(W=0)}$ = 0.116

Im Beispiel ist die Differenz gering, bei anderen Datenlagen kann es aber durchaus vorkommen, dass das multiplikative Modell einen deutlichen Interaktionseffekt aufweist, das additive Modell dagegen nicht.
Häufiger dürfte der umgekehrte Fall auftreten. So zeigt die unten wiedergegebene Datenkonstellation, dass in den beiden Partialtabellen das Odds-Ratio jeweils α=2 ist, sich die Prozentsatzdifferenzen dagegen mit $d_{YX(W=1)}$% = 7.6 und $d_{YX(W=0)}$% = 16.7 deutlich unterscheiden.

Generierte Daten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	200	250	160	120	730
Y=0	20	50	80	120	270
∑	220	300	240	240	1000

α = 2 α = 2

Bedingte relative Häufigkeiten:

	W=1		W=0		
	X=1	X=0	X=1	X=0	∑
Y=1	.571	.833	.667	.500	.73
Y=0	.429	.167	.333	.500	.27
n	(350)	(150)	(150)	(350)	(1000)

$d_{YX(W=1)}$%=7.6 $d_{YX(W=0)}$%=16.7
$\Phi_{XY(W=1)}$ = 0.110 $\Phi_{XY(W=0)}$ = 0.169

Auswahl des Analysemodells

Tatsächlich wurden für das letzte Beispiel die Daten so generiert, dass X auf Y mit einem Odds-Ratio von 2 und W auf Y mit einem Odds-Ratio von 5 wirkt. Nur wenn die Daten additiv mit Prozentsatzdifferenzen analysiert werden, zeigt sich ein Interaktionseffekt.

In der Realität ist der „wahre“ Kausalmechanismus des datengenerierenden Prozesses nicht bekannt. Wenn sich eine Datenkonstellation mit zwei Analysemodellen gleich gut erfassen lässt, wird dann oft das Modell genommen, das „sparsamer“ ist, was in der Statistik bedeutet, dass es durch weniger Parameter beschrieben werden kann.

So ist im letzten Beispiel das multiplikative Modell sparsamer, da es ohne Interaktionseffekt auskommt, während das additive Model einen zusätzlichen Interaktionseffekt benötigt. Umgekehrt war es im anderen Beispiel, wo nur additive Modell ohne Interaktionseffekt auskommt.

Letztlich hängt auch die Auswahl eines Analysmodells neben rein praktischen Gesichtspunkten wie die Verfügbarkeit und leichte Anwendbarkeit von den theoretischen Erwartungen ab, die ein Sozialforscher über den zu untersuchenden Ausschnitt aus der Realität hat.
Wenn sich zeigt, dass die Erwartungen falsch sind, kann es möglicherweise auch sinnvoll sein, auf ein Analysemodell anzuwenden, dass angemessener erscheint.

Lerneinheit 3: Die generelle Logik der Chiquadrat-Tests

	Mitglied in Religionsgemeinschaft?		
Schwangerschaftsabbruch	nein	ja	insgesamt
... sollte erlaubt sein	0.2140 (689)	0.2262 (728)	0.4402 (1417)
... sollte verboten sein	0.1289 (415)	0.4309 (1387)	0.5598 (1802)
insgesamt	0.3429 (1104)	0.6570 (2115)	1.0000 (3219)

Sowohl Pearsons Chiquadrat-Test wie auch der LR-Test prüfen die Nullhypothese, dass die Auftretenswahrscheinlichkeiten von Zeilen- und Spaltenvariable in den Zellen einer bivariaten Kreuztabelle gleich dem Produkt der Auftretenswahrscheinlichkeiten der Randverteilungen sind: H_0: $\pi_{ij} = \pi_{i\bullet} \cdot \pi_{\bullet j}$ vs. H_1: $\pi_{ij} \neq \pi_{i\bullet} \cdot \pi_{\bullet j}$.

Werden Null- und Alternativhypothese verglichen, kann die Nullhypothese auch als eine ***Restriktion*** über die Auftretenswahrscheinlichkeiten in einer Kreuztabelle aufgefasst werden. Die Alternativhypothese beinhaltet dagegen keine Restriktionen über die Wahrscheinlichkeiten. Bei dieser Sichtweise kann man auch davon ausgehen, dass zwei statistische Modelle gegeneinander getestet werden:

- das restriktives Modell der Nullhypothese,
- und das weniger restriktive oder liberale Modell der Alternativhypothese.

Hinweis:
Formal stimmt diese Sicht nicht ganz mit der Hypothesenformulierung nach dem Neyman-Pearsons-Test überein, da das liberale Modell (H_1) ja gerade die Wahrscheinlichkeiten des restriktiven Modells (H_0) ausschließt.

Die generelle Logik der Chiquadrat-Tests

Ähnlich wie bei der Idee der proportionalen Fehlerreduktion bei asymmetrischen Zusammenhangsmaßen kann bei der Unterscheidung zwischen einem liberalen und einem restriktiven Modell geprüft werden, welches Modell besser mit den empirischen Daten einer Stichprobe vereinbar ist.

Tatsächlich sind sowohl Pearsons Chiquadrat-Teststatistik als auch die LR-Teststatistik so aufgebaut, dass Differenzen bzw. Quotienten zwischen den unter der Nullhypothese und der Alternativhypothese berechneten geschätzten Wahrscheinlichkeiten bzw. Populationsanteilen in die Berechnung einfließen.

Unter der Alternativhypothese des liberalen Modells gibt es keine Restriktionen für die beobachteten Häufigkeiten bzw. Anteile. Die geschätzten erwarteten Häufigkeiten e_{ij} bzw. Populationsanteile sind dann gleich den beobachteten Häufigkeiten und Anteilen:

$$e_{ij}\big|H_1 = \hat{\lambda}_{ij}\big|H_1 = n_{ij} \text{ und } \hat{\pi}_{ij}\big|H_1 = p_{ij}$$

Entsprechend gilt dann für die geschätzten erwarteten Häufigkeiten bzw. Anteile:

$$e_{ij}\big|H_0 = \hat{\lambda}_{ij}\big|H_0 = \frac{n_{i+} \cdot n_{+j}}{n} \text{ und } \hat{\pi}_{ij}\big|H_0 = p_{i+} \cdot p_{+j}$$

Die Teststatistiken sind Funktionen dieser geschätzten Häufigkeiten bzw. Anteile.

Die generelle Logik der Chiquadrat-Tests

	Mitglied in Religionsgemeinschaft?		
Schwangerschaftsabbruch	nein	ja	insgesamt
... sollte erlaubt sein	0.2140 (689)	0.2262 (728)	0.4402 (1417)
... sollte verboten sein	0.1289 (415)	0.4309 (1387)	0.5598 (1802)
insgesamt	0.3429 (1104)	0.6570 (2115)	1.0000 (3219)

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i+}\cdot n_{+j}}{n}\right)^2}{\frac{n_{i+}\cdot n_{+j}}{n}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(e_{ij}\middle|H_1 - e_{ij}\middle|H_0\right)^2}{e_{ij}\middle|H_0} = n\cdot\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(\hat{\pi}_{ij}\middle|H_0 - \hat{\pi}_{i}\middle|H_1\right)^2}{\hat{\pi}_{i}\middle|H_1}$$

$$L^2 = 2\cdot\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\cdot\ln\left(\frac{n_{ij}}{\frac{n_{i+}\cdot n_{+j}}{n}}\right) = 2\cdot\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\cdot\ln\left(\frac{e_{ij}\middle|H_1}{e_{ij}\middle|H_1}\right) = 2\cdot\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\cdot\ln\left(\frac{\hat{\pi}_{ij}\middle|H_1}{\hat{\pi}_{ij}\middle|H_1}\right)$$

Wenn die Teststatistiken zu keinem signifikanten Ergebnis führen, kann dies so interpretiert werden, dass sich das restriktive Modell nicht signifikant vom liberalen Modell unterscheidet. Bei einem signifikanten Ergebnis ist das restriktive Modell dagegen signifikant schlechter mit empirischen Daten vereinbar als das liberale Modell.

Die generelle Logik der Chiquadrat-Tests

	Mitglied in Religionsgemeinschaft?		
Schwangerschaftsabbruch	nein	ja	insgesamt
... sollte erlaubt sein	0.2140 (689)	0.2262 (728)	0.4402 (1417)
... sollte verboten sein	0.1289 (415)	0.4309 (1387)	0.5598 (1802)
insgesamt	0.3429 (1104)	0.6570 (2115)	1.0000 (3219)

Wenn das restriktive Modell zutrifft, sind beide Teststatistiken zentral chiquadrat-verteilt, wenn das liberale Modell zutrifft, sind beide Teststatistiken nichtzentral chiquadrat-verteilt, wobei der Wert des Nichtzentralitätsparameters eine Funktion der Unterschiede zwischen den Wahrscheinlichkeiten bzw. Populationsanteilen des restriktiven und des liberalen Modells sind.
Die Freiheitsgrade der Teststatistik können aus den Differenzen der Parameter berechnet werden, die im liberalen bzw. restriktiven Modell geschätzt werden.

Werden die erwarteten Häufigkeiten geschätzt, sind diese im liberalen Modell der H_1 gleich der Anzahl der Tabellenzellen, im Beispiel also 4. Im restriktiven Modell der H_0 berechnen sie sich aus den Randverteilungen, wobei neben der Gesamtfallzahl n nur jeweils eine Häufigkeit zu schätzen ist, da die komplementäre Häufigkeit die Differenz von der Fallzahl ist: $n_{2+} = n - n_{1+}$ und $n_{+2} = n - n_{+1}$. Im Beispiel werden für das restriktive Modell somit 3 Parameter geschätzt. Die Chiquadratverteilung hat daher df= 4–3= 1 Freiheitsgrad.

Die gleiche Zahl ergibt sich bei Berechnung über die zu schätzenden relativen Häufigkeiten. Da die Tabelle 4 Zellen hat und sich Anteile zu 1 summieren, müssen im liberalen Modell 3 Anteile geschätzt werden und im restriktiven Modell 2 Anteile in den beiden Randverteilungen. Die Differenz ist wieder df=1.

Die generelle Logik der Chiquadrat-Tests

	Mitglied in Religionsgemeinschaft?		
Schwangerschaftsabbruch	nein	ja	insgesamt
... sollte erlaubt sein	0.2140 (689)	0.2262 (728)	0.4402 (1417)
... sollte verboten sein	0.1289 (415)	0.4309 (1387)	0.5598 (1802)
insgesamt	0.3429 (1104)	0.6570 (2115)	1.0000 (3219)

Für die Beispieldaten der Vierfeldertabelle berechnen sich die Teststatistiken als:

$$\chi^2 = \frac{\left(689 - \frac{1417 \cdot 1104}{3219}\right)^2}{\frac{1417 \cdot 1104}{3219}} + \frac{\left(728 - \frac{1417 \cdot 2115}{3219}\right)^2}{\frac{1417 \cdot 2115}{3219}} + \frac{\left(415 - \frac{1802 \cdot 1104}{3219}\right)^2}{\frac{1802 \cdot 1104}{3219}}$$

$$+ \frac{\left(1387 - \frac{1802 \cdot 1104}{2115}\right)^2}{\frac{1802 \cdot 1104}{2115}} = 230.6$$

$$L^2 = 2 \cdot \left(689 \cdot \ln\left(\frac{.2140}{.1510}\right) + 728 \cdot \ln\left(\frac{.2262}{.2892}\right) + 415 \cdot \ln\left(\frac{.1289}{.1920}\right) + 1387 \cdot \ln\left(\frac{.4309}{.3679}\right)\right)$$

$$= 231.2$$

Bei einer Irrtumswahrscheinlichkeit von 5% und df=1 Freiheitsgrad sind beide Teststatistiken größer als das 95%-Quantil der Chiquadratverteilung , das 3.84 beträgt.

Die generelle Logik der Chiquadrat-Tests

	Mitglied in Religionsgemeinschaft?		
Schwangerschaftsabbruch	nein	ja	insgesamt
... sollte erlaubt sein	0.2140 (689)	0.2262 (728)	0.4402 (1417)
... sollte verboten sein	0.1289 (415)	0.4309 (1387)	0.5598 (1802)
insgesamt	0.3429 (1104)	0.6570 (2115)	1.0000 (3219)

Diese Logik der beiden Chiquadrattests kann ganz generell für beliebige Hypothesentests angewendet werden, solange die Nullhypothese jeweils einem restriktiven Modell zugeordnet werden kann und die Alternativhypothese einem liberalen Modell ohne Restriktionen.

Als ein weiteres Beispiel kann so z.B. die Hypothese geprüft werden, dass die Populationsanteile in allen vier Tabellenzellen gleich groß und damit gleich 1/4 sind:

$$H_0: \pi_{ij} = 0.25 \; H_1: \pi_{ij} \neq 0.25.$$

Die Teststatistiken berechnen sich bei dieser Nullhypothese folgendermaßen:

$$\chi^2 = 3219 \cdot \left(\frac{(.2140 - .25)^2}{.25} + \frac{(.2262 - .25)^2}{.25} + \frac{(.1289 - .25)^2}{.25} + \frac{(.4309 - .25)^2}{.25} \right) = 634.2$$

$$L^2 = 2 \cdot \left(689 \cdot \ln\left(\frac{.2140}{.25}\right) + 728 \cdot \ln\left(\frac{.2262}{.25}\right) + 415 \cdot \ln\left(\frac{.1289}{.25}\right) + 1387 \cdot \ln\left(\frac{.1387}{.25}\right)\right) = 600.5$$

Die Teststatistiken unterscheiden sich in ihren Absolutwerten, relativ als Quotient sind die Werte aber nicht sehr unterschiedlich: 634.2/600.5 = 1.056

Die generelle Logik der Chiquadrat-Tests

	Mitglied in Religionsgemeinschaft?		
Schwangerschaftsabbruch	nein	ja	insgesamt
... sollte erlaubt sein	0.2140 (689)	0.2262 (728)	0.4402 (1417)
... sollte verboten sein	0.1289 (415)	0.4309 (1387)	0.5598 (1802)
insgesamt	0.3429 (1104)	0.6570 (2115)	1.0000 (3219)

Unter der Nullhypothese wird kein Populationsanteil geschätzt, unter der Alternativhypothese werden wie beim Test auf statistische Unabhängigkeit drei Anteile geschätzt. Also beträgt die Zahl der Freiheitsgrade df = 3–0 = 3. Bei einer Irrtumswahrscheinlichkeit von 5% beträgt der kritische Wert dann 7.815.
Da beide Teststatistiken jeweils sehr viel größer sind, muss die Nullhypothese verworfen werden. Es kann nicht davon ausgegangen werden, dass die Populationsanteile in den vier Zellen gleich sind.

Hierarchische Testung bei LR-Tests

Die Teststatistik des LR-Test gilt für die Aufsummierung über alle Zellen einer beliebig dimensionalen Tabelle. In den folgenden Beispielen wird daher nur ein Index i verwendet, der bei einer bivariaten Kreuztabelle für die Kombination aus Zeilen- und Spaltenindex steht, bei mehrdimensionalen Tabellen für entsprechend mehr Indizes.

Hierarchische Testung bei LR-Tests

Die Alternativhypothese beinhaltet keine Restriktionen über die Zellenhäufigkeiten, so dass die absoluten bzw. relativen Häufigkeiten unter der Alternativhypothese immer gleich den beobachteten absoluten oder relativen Häufigkeiten in den Tabellenzellen sind.
Der LR-Test erlaubt jedoch auch einen Test einer sehr restriktiven Nullhypothese gegen eine weniger restriktive Alternativhypothese.

So ist etwa im obigen Beispiel die Hypothese, dass alle vier relativen Häufigkeiten gleich groß sind, eine restriktivere Annahme als die, dass sie sich aus dem Produkt der Randverteilungen ergeben, wie das die Nullhypothese im Tests auf statistische Unabhängigkeit postuliert.

Der LR-Test erlaubt nun, die stärkere Restriktion gegen die weniger starke zu testen, wenn die stärkere Restriktion als Spezialfall der weniger starken aufzufassen ist.
Das ist hier der Fall, da das Modell der Gleichverteilung als Spezialfall der statistischen Unabhängigkeit formuliert werden kann:

$$H_0: \pi_{ij} = \pi_{i\bullet}\cdot\pi_{\bullet j} \text{ und } \pi_{i\bullet} = \pi_{\bullet j} = 0.5 \text{ vs. } H_1: \pi_{ij} = \pi_{i\bullet}\cdot\pi_{\bullet} \text{ und } \pi_{i\bullet} \neq 0.5 \text{ oder } \pi_{\bullet j} \neq 0.5.$$

In der Statistik spricht man davon, dass zwei statistische Modelle ***hierarchisch*** ineinander geschachtelt *(engl. **nested**)* sind, wenn das restriktivere Modell zusätzliche Restriktionen gegenüber dem weniger restriktiven (Alternativ-) Modell behauptet.
Wenn die Modell mit H_1 und H_0 korrespondieren, erlaubt die LR-Teststatistik die Prüfung des restriktiven Modells unter H_1 gegen das weniger restriktive Alternativmodell unter H_0.

Hierarchische Testung bei LR-Tests

	Mitglied in Religionsgemeinschaft?		
Schwangerschaftsabbruch	nein	ja	insgesamt
... sollte erlaubt sein	0.2140 (689)	0.2262 (728)	0.4402 (1417)
... sollte verboten sein	0.1289 (415)	0.4309 (1387)	0.5598 (1802)
insgesamt	0.3429 (1104)	0.6570 (2115)	1.0000 (3219)

Für die generelle Teststatistik des LR-Tests werden neben den beobachteten Häufigkeiten sowohl die erwarteten Häufigkeiten unter der Nullhypothese wie der Alternativhypothese benötigt:

$$L^2 = 2 \cdot \sum_i n_i \cdot \ln\left(\frac{\hat{\pi}_i|H_1}{\pi_i|H_0}\right) = 2 \cdot \sum_i n_i \cdot \ln\left(\frac{e_i|H_1}{e_i|H_0}\right)$$

Die erwarteten Häufigkeiten unter dem restriktiven H_0-Modell sind im Beispiel stets $e_{ij}|H_0 = 804.75 = 3219 \cdot 0.25$, die unter dem weniger restriktiven H_1-Modell sind die bei statistischer Unabhängigkeit erwarteten Häufigkeiten: $e_{ij}|H_1 = n_{i+} \cdot n_{+j}/n$. Einsetzen der Werte in die allgemeine Formel und Aufsummieren über die 4 Tabellenzellen ergibt den Wert $\chi^2 = 370.5$

Abtreibung	Religionsg.	n_{ij}	$e_{ij}\|H_1$	$e_{ij}\|H_0$	L^2-Anteil
ja	nein	689	485.979	804.75	–693.743
ja	ja	728	931.102	804.75	212.339
nein	nein	415	618.021	804.75	–219.128
nein	ja	1387	1183.979	804.75	1071.054
SUMME:		3219	3219.081 2	3219	370.522

Hierarchische Testung bei LR-Tests

	Mitglied in Religionsgemeinschaft?		
Schwangerschaftsabbruch	nein	ja	insgesamt
... sollte erlaubt sein	0.2140 (689)	0.2262 (728)	0.4402 (1417)
... sollte verboten sein	0.1289 (415)	0.4309 (1387)	0.5598 (1802)
insgesamt	0.3429 (1104)	0.6570 (2115)	1.0000 (3219)

Wenn die Nullhypothese zutrifft, ist die Teststatistik chiquadrat-verteilt, wenn die Alternativhypothese zutrifft, ist die Teststatistik nichtzentral chiquadrat-verteilt. Die Zahl der Freiheitsgrade ergibt sich wieder aus der Differenz der zu schätzenden Parameter. Für die Schätzung der erwarteten Häufigkeiten im weniger restriktiven Modell der Alternativhypothese H_1 müssen zwei Populationsanteile für die Randverteilungen geschätzt werden, für die erwarteten Häufigkeiten im restriktiven Modell der Nullhypothese H_0 gar keine Populationsanteile. Die Zahl der Freiheitsgrade im Test von H_0 vs. H_1 beträgt hier also df = 2– 0 = 2.

Bei einer Irrtumswahrscheinlichkeit von 5% wird die Nullhypothese abgelehnt, wenn die Teststatistik größer ist als das 95%-Quantil der Chiquadrat-Verteilung mit df=2Freiheitsgrad, also größer ist als 5.99. Dies ist hier der Fall. Es ist eher davon auszugehen, dass die beiden Variablen statistisch unabhängig voneinander sind als dass die Populationsanteile in allen vier Tabellenzellen gleich groß sind.

Hierarchische Chiquadrat-Tests haben die Eigenschaft, dass sich die Teststatistiken aufsummieren, was die Berechnung der Teststatistiken erleichtern kann.

Hierarchische Testung bei LR-Tests

	Mitglied in Religionsgemeinschaft?		
Schwangerschaftsabbruch	nein	ja	insgesamt
... sollte erlaubt sein	0.2140 (689)	0.2262 (728)	0.4402 (1417)
... sollte verboten sein	0.1289 (415)	0.4309 (1387)	0.5598 (1802)
insgesamt	0.3429 (1104)	0.6570 (2115)	1.0000 (3219)

So ergab der Test auf Unabhängigkeit der beiden Variablen eine LR-Statistik von 231.2 bei df = 1 Freiheitsgrad. Der Test auf Gleichverteilung in allen vier Zellen ergab eine Teststatistik von 600. 5 bei df = 3 Freiheitsgraden. Die Differenz ist 600.5 – 231.2 = 369.3 bei df = 3–1 = 2 Freiheitsgraden. Bis auf Rundungsfehler ist das genau der Wert der gerade berechnet wurde. Tatsächlich folgt dies aus der allgemeinen Formel für die LR-Statistik:

$$L^2 = 2 \cdot \sum_i n_i \cdot \ln\left(\frac{e_i|H_1}{e_i|H_0}\right) = 2 \cdot \sum_i n_i \cdot \ln\left(\frac{n_i}{e_i|H_0}\right) - 2 \cdot \sum_i n_i \cdot \ln\left(\frac{n_i}{e_i|H_1}\right)$$

Es stellt sich die Frage, warum eine stärker restriktivere (Null-) Hypothese gegen eine weniger starke (Alternativ-) Hypothese getestet werden sollte.
Dies liegt an der Teststärke der beiden Tests:

- Der LR-Test einer strengen gegen eine weniger strenge Hypothese hat eine größere Trennschärfe als der LR-Test der strengen Hypothese gegen die Alternative, dass gar keine Restriktionen formuliert werden.

Hierarchische Testung bei LR-Tests

- Allerdings setzt der hierarchische Test streng genommen voraus, dass zumindest die weniger strenge Hypothese zutrifft. Im Beispiel wurde aber auch die weniger strenge Hypothese abgelehnt.
  Es hat sich jedoch herausgestellt, dass der hierarchische LR-Test i.a. recht robust gegenüber der Annahme der Gültigkeit der Alternativhypothese ist. Der Test führt also auch dann oft zu richtigen Ergebnissen bezüglich der Nullhypothese, wenn die Alternativhypothese nicht zutrifft. Eine falsche Nullhypothese wird also auch dann meistens korrekt abgelehnt, wenn bereits die Alternativhypothese falsch ist.

Die Möglichkeit des hierarchischen Testens hat dazu geführt, dass der LR-Test in der Statistik eine sehr große Bedeutung spielt.
Pearsons Chiquadrat-Statistik weist diese Eigenschaft nicht aus. Zwar kann auch mit Pearsons Chiquadrat-Statistik ein beliebig strenges H_0-Modell getestet werden. Das liberale H_0-Modell besagt aber stets, dass gar keine Restriktionen vorliegen.
Aufgrund der Äquivalenz der beiden Tests bei einer Alternativhypothese ohne jegliche Restriktionen stellt sich allerdings oft das gleiche Ergebnis ein, wenn Differenzen von Pearsons Chiquadrat-Statistiken zur hierarchischen Testung genutzt wird.

Die Annäherung von Pearsons Teststatistik an die Chiquadrat-Verteilung erfolgt schneller als die Annäherung beim LR-Test. Pearsons Chiquadrat-Test ist daher bei Alternativhypothesen, die gar keine Restriktionen postulieren meist vertrauenswürdiger.

Hypothesenprüfung in trivariaten Kreuztabellen

Die Logik des Chiquadrat-Tests kann auch für Tests von Hypothesen in trivariaten Tabellen angewendet werden. Solche Test können insbesondere genutzt werden, um Hypothesen über kausale Zusammenhänge zu prüfen:

- Führt der Test auf Unabhängigkeit aller drei Variablen zu keinem signifikanten Ergebnis, dann ist davon auszugehen, dass es auch keine kausale Beziehung zwischen den drei Variablen gibt. Allerdings ist die Möglichkeit einer scheinbaren Nichtbeziehung nicht ganz ausgeschlossen, die durch indirekte, korrelierte oder Interaktionseffekte mit nicht gemessenen weiteren Variablen hervorgerufen wird.
- Besteht kein signifikanter Zusammenhang mit der Kontrollvariable, dann ist es vermutlich ausgeschlossen, dass die Kontrollvariable einen Interaktionseffekt, einen direkten oder einen indirekten Effekt auf die abhängige Variable hat, so dass eine bivariate Analyse der verbleibenden Variablen hinreichend ist.
- Um Scheinkausalität oder einen indirekten Effekt zu prüfen, muss die gemeinsame Erklärungsgröße oder die intervenierende Variable als Kontrollvariable fungieren. Gibt es dann keine konditionalen Effekte, liegt vermutlich Scheinkausalität bzw. ein indirekter Effekt vor.
- Bei additiven Effekten ergeben sich bei Kontrolle durch die Drittvariable in den Partialtabellen etwa gleich starke Effekte. Dies gilt auch, wenn erklärende Variable und Drittvariable vertauscht werden.
- Dem gegenüber liegt ein Interaktionseffekt vor, wenn sich die konditionalen Effekte in den Partialtabellen stark (und signifikant) unterscheiden.

! Zu beachten ist, dass Fehlspezifikationen bei allen Test zu falschen Schlussfolgerungen führen können.

Test der statistischen Unabhängigkeit unter drei Variablen

Als Beispiel für die Prüfung der Unabhängigkeit von drei Variablen werden aus den Daten des Allbus 2006 die drei Variablen allgemeine Wirtschaftslage (AWL), eigene Wirtschaftslage (EWL) und Erwerbstätigkeit mit einer Irrtumswahrscheinlichkeit betrachtet.

Um nicht zu große Tabellen zu erhalten, sind die ursprünglich jeweils fünf Ausprägungen von AWL und EWL zu den beiden Ausprägungen gut (Code: 1) und nicht gut (2) zusammengefasst. Als erwerbstätig bzw. erwerbssuchend (Code: 1) werden ganztags und halbtags erwerbstätige Personen sowie Arbeitslose aufgefasst, als nicht erwerbstätig (2) die übrigen Befragten, solange diese keine Nebenerwerbstätigkeit ausüben.

Die Durchführung des Tests auf statistische Unabhängigkeit aller drei Variablen erfolgt nach den üblichen vier bzw. fünf Schritten.

Schritt 1: Formulierung von Null- und Alternativhypothese

Die Nullhypothese postuliert, dass alle drei Variablen statistisch unabhängig voneinander sind. Die gemeinsamen Populationsanteile in den Tabellenzellen sind dann gleich den Produkten der entsprechenden Randanteile in der Population:

$$\pi_{ijk} = \pi_{i\bullet\bullet} \cdot \pi_{\bullet j\bullet} \cdot \pi_{\bullet\bullet k} \text{ für alle } i=1,2,...,I,\ j=1,2,...,J \text{ und } k=1,2,...,K$$

Die Alternativhypothese behauptet keine Restriktionen über die gemeinsamen Populationsanteile der inneren Tabellenzellen. Daraus folgt für das zu testende Hypothesenpaar:

$$H_0: \pi_{ijk} = \pi_{i\bullet\bullet} \cdot \pi_{\bullet j\bullet} \cdot \pi_{\bullet\bullet k} \text{ versus } H_1: \pi_{ijk} \neq \pi_{i\bullet\bullet} \cdot \pi_{\bullet j\bullet} \cdot \pi_{\bullet\bullet k}$$

Prüfung der statistischen Unabhängigkeit von drei Variablen

Schritt 2: Auswahl von Teststatistik und Kennwerteverteilung:
Als Teststatistik kommt sowohl Pearsons Chiquadratstatistik als auch die LR-Teststatistik in Frage.

Für die Anwendung des Tests werden Schätzungen der Populationsanteile bzw. der erwarteten absoluten Häufigkeiten unter der Annahme benötigt, dass die Nullhypothese zutrifft. In einfachen Zufallsauswahlen sind die Randanteile $p_{i\bullet\bullet}$, $p_{\bullet j\bullet}$ und $p_{\bullet\bullet k}$ konsistente und erwartungstreue Schätzer der entsprechenden Populationsanteile.
Aus ihnen können dann die bei Gültigkeit der Nullhypothese geschätzten Anteile bzw. erwarteten Häufigkeiten in den Zellen berechnet werden:

$$\hat{\pi}_{ijk}\left|H_0 = \hat{\pi}_i \cdot \hat{\pi}_j \cdot \hat{\pi}_k = p_{i\bullet\bullet} \cdot p_{\bullet j\bullet} \cdot p_{\bullet\bullet k}\right.$$

$$\Rightarrow\ e_{ijk} = n \cdot p_{i\bullet\bullet} \cdot p_{\bullet j\bullet} \cdot p_{\bullet\bullet k} = n \cdot \frac{n_{i\bullet\bullet}}{n} \cdot \frac{n_{\bullet j\bullet}}{n} \cdot \frac{n_{\bullet\bullet k}}{n} = \frac{n_{i\bullet\bullet} \cdot n_{\bullet j\bullet} \cdot n_{\bullet\bullet k}}{n^2}$$

Diese Statistiken können dann in die allgemeine Formeln von Pearsons Chiquadrat-Statistik χ^2 bzw. der LR-Statistik L^2 eingesetzt werden:

$$\chi^2 = n \cdot \sum_i \frac{\left(\hat{\pi}_i\left|H_1 - \hat{\pi}_i\right|H_0\right)^2}{\hat{\pi}_i\left|H_0\right.} = n \cdot \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(p_{ijk} - p_{i\bullet\bullet} \cdot p_{\bullet j\bullet} \cdot p_{\bullet\bullet k}\right)^2}{p_{i\bullet\bullet} \cdot p_{\bullet j\bullet} \cdot p_{\bullet\bullet k}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(n_{ijk} - \frac{n_{i\bullet\bullet} \cdot n_{\bullet j\bullet} \cdot n_{\bullet\bullet k}}{n^2}\right)^2}{\frac{n_{i\bullet\bullet} \cdot n_{\bullet j\bullet} \cdot n_{\bullet\bullet k}}{n^2}}$$

Prüfung der statistischen Unabhängigkeit von drei Variablen X, Y und Z

bzw.:

$$L^2 = 2 \cdot \sum_i n_i \cdot \ln\left(\frac{\hat{\pi}_i\left|H_1\right.}{\hat{\pi}_i\left|H_0\right.}\right) = 2 \cdot \sum_i n_i \cdot \ln\left(\frac{e_i\left|H_1\right.}{e_i\left|H_0\right.}\right) = 2 \cdot \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk} \cdot \ln\left(\frac{n_{ijk}}{\frac{n_{i\bullet\bullet} \cdot n_{\bullet j\bullet} \cdot n_{\bullet\bullet k}}{n^2}}\right)$$

Wenn die Nullhypothese richtig ist, sind die beiden Teststatistiken asymptotisch chiquadrat-verteilt, bei falscher Nullhypothese nichtzentral chiquadrat-verteilt.
Die Zahl der Freiheitsgrade ergibt sich aus der Differenz der aus den Daten zu schätzenden Populationsanteile. Die Tabelle hat insgesamt I×J×K Tabellenzellen. Bei Gültigkeit der Alternativhypothese müssen I·J·K–1 Anteile geschätzt werden, da die Summe aller Zellenanteile 1.0 sein muss. Bei Gültigkeit der Nullhypothese müssen nur (I–1)+(J–1)+(K–1) univariate Randwahrscheinlichkeiten geschätzt werden. Daraus folgt die Zahl der Freiheitsgrade als:

$$df = I \cdot J \cdot K - 1 - (I-1) - (J-1) - (K-1) = I \cdot J \cdot K - I - J - K + 2$$

Schritt 3: Festlegung von Irrtumswahrscheinlichkeiten und kritischen Werten
Die Irrtumswahrscheinlichkeit ist mit 5% vorgegeben.
Da alle drei Variablen dichotomisiert sind, ergeben sich df = 2·2·2–2–2–2+2 = 4 Freiheitsgrade.
Da die nichtzentrale Chiquadrat-Verteilung einen größeren Erwartungswert hat als die zentrale Chiquadrat-Verteilung wird die Nullhypothese bei einer Irrtumswahrscheinlichkeit von 5% abgelehnt, wenn die Teststatistik das 95%-Quantil der Chiquadrat-Verteilung mit df=4 Freiheitsgraden erreicht oder überschreitet.

Prüfung der statistischen Unabhängigkeit von drei Variablen X, Y und Z

	Erwerbstätigkeit (Z)			
	ja (z_1)		nein (z_2)	
Eigene Wirtschaftslage (Y)	Allgemeine Wirtschaftsl.(X)		Allgemeine Wirtschafsl. (X)	
	gut (x_1)	nicht gut (x_2)	gut (x_1)	nicht gut (x_2)
gut (y_1)	64.8% (160)	33.7% (557)	73.1% (152)	40.5% (469)
nicht gut (y_2)	35.2% (87)	66.3% (1094)	26.9% (56)	59.5% (688)

(Quelle: Allbus 2006)

Randverteilungen		
Y	X	Z
1338	455	1898
1925	2808	1365
3263	3263	3263

Aus einer Chiquadrat-Tabelle ist zu entnehmen, dass der kritische Wert, d.h. das 95%-Quantil der Chiquadrat-Verteilung mit df=4 Freieitsgraden 9.488 beträgt.

Schritt 4: Berechnung der Teststatistik und Entscheidung

Für die Berechnung werden die Randverteilung aller drei Variablen benötigt, die sich durch Aufsummieren über die entsprechenden Tabellenzellen ergeben.

Daraus werden dann die die bei gültiger Nullhypothese erwarteten Häufigkeiten berechnet:

$$e_{ijk} = \frac{n_{i\bullet\bullet} \cdot n_{\bullet j\bullet} \cdot n_{\bullet\bullet k}}{n^2} \text{ bei i=1, j=1, k=1, z.B: } e_{111} = \frac{1338 \cdot 455 \cdot 1898}{3263^2} = 108.52$$

Aus den erwarteten und den beobachteten Häufigkeiten werden für jede einzelne Zelle die Chiquadrat-Anteile berechnet. Für die erste Zelle ergeben sich:

$$\chi^2_{111} = \frac{(n_{111} - e_{111})^2}{e_{111}} = \frac{(160 - 108.52)^2}{108.52} = 24.42 \; ; \; L^2_{111} = 2 \cdot n_{111} \cdot \ln\left(\frac{n_{111}}{e_{111}}\right) = 2 \cdot 160 \cdot \ln\left(\frac{160}{108.52}\right) = 124.24$$

Prüfung der statistischen Unabhängigkeit von drei Variablen X, Y und Z

Berechnung der Teststatistiken χ^2 und L^2 bei H_0: $\pi_{ijk} = \pi_{i\bullet\bullet} \cdot \pi_{\bullet j\bullet} \cdot \pi_{\bullet\bullet k}$

erwerbstätig	AWL	EWL	n_{ijk}	e_{ijk}	χ^2-Anteil	L^2-Anteil
ja	gut	gut	160	108.52	24.42	124.24
ja	gut	nicht gut	87	156.14	30.62	-101.76
ja	nicht gut	gut	557	669.75	18.98	-205.35
ja	nicht gut	nicht gut	1094	963.58	17.65	277.75
nein	gut	gut	152	78.05	70.07	202.63
nein	gut	nicht gut	56	112.29	28.22	-77.92
nein	nicht gut	gut	469	481.67	0.33	-25.00
nein	nicht gut	nicht gut	688	692.99	0.04	-9.94
Summe			3263	3262.99	190.33	184.65
				≈3263		

Die Aufsummierung ergibt die Werte der Teststatistiken:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} \frac{\left(n_{ijk} - \frac{n_{i\bullet\bullet} \cdot n_{\bullet j\bullet} \cdot n_{\bullet\bullet k}}{n^2}\right)^2}{\frac{n_{i\bullet\bullet} \cdot n_{\bullet j\bullet} \cdot n_{\bullet\bullet k}}{n^2}} = 190.33 \quad L^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} 2 \cdot n_{ijk} \cdot \ln\left(\frac{n_{ijk}}{\frac{n_{i\bullet\bullet} \cdot n_{\bullet j\bullet} \cdot n_{\bullet\bullet k}}{n^2}}\right) = 184.65$$

Beide Teststatistiken sind deutlich größer als der kritische Wer 9.488. Die Nullhypothese ist daher abzulehnen. Bei einer Irrtumswahrscheinlichkeit von 5% kann davon ausgegangen werden, dass die drei Variablen allgemeine und eigene Wirtschaftslage und Erwerbstätigkeit nicht statistisch unabhängig voneinander sind.

Prüfung der statistischen Unabhängigkeit von drei Variablen X, Y und Z

Berechnung der Teststatistiken χ^2 und L^2 bei H_0: $\pi_{ijk} = \pi_{i\bullet\bullet} \cdot \pi_{\bullet j\bullet} \cdot \pi_{\bullet\bullet k}$

erwerbstätig	AWL	EWL	n_{ijk}	e_{ijk}	χ^2-Anteil	L^2-Anteil
ja	gut	gut	160	108.52	24.42	124.24
ja	gut	nicht gut	87	156.14	30.62	-101.76
ja	nicht gut	gut	557	669.75	18.98	-205.35
ja	nicht gut	nicht gut	1094	963.58	17.65	277.75
nein	gut	gut	152	78.05	70.07	202.63
nein	gut	nicht gut	56	112.29	28.22	-77.92
nein	nicht gut	gut	469	481.67	0.33	-25.00
nein	nicht gut	nicht gut	688	692.99	0.04	-9.94
Summe			3263	3262.99 ≈3263	190.33	184.65

Schritt 5: Kontrolle der Anwendungsvoraussetzungen
Die Chiquadrat-Verteilungen sind nur asymptotisch gültig. Als Faustregel gilt, dass möglichst alle erwarteten Häufigkeiten größer 5 oder aber mindest 80% der Zellen größer 5 und alle Zellen eine erwartete Häufigkeit größer 1 aufweisen.
Wie aus der Spalte e_{ijk} der Tabelle zur Berechnung der Teststatistiken zu entnehmen ist, ist diese Bedingung erfüllt, da die kleinste erwartete Häufigkeit 78.05 > 5 ist.

Prüfung der statistischen Unabhängigkeit einer von zwei anderen Variablen

Der Test auf Unabhängigkeit zeigt, dass vermutlich eine Abhängigkeit zwischen den drei Varablen allgemeine und eigene wirtschaftliche Lage und Erwerbstätigkeit vorliegt. Möglicherweise ist aber die Erwerbstätigkeit von den beiden anderen Variablen unabhängig. Dies kann wiederum mit einem Chiquadrat-Test geprüft werden.

Schritt 1: Formulierung von Null- und Alternativhypothese
Wenn in einer trivariaten Häufigkeitstabelle die Zeilen- und Spaltenvariable von der dritten Variablen statistisch unabhängig sind, muss gelten:

$$\pi_{ijk} = \pi_{ij\bullet} \cdot \pi_{\bullet\bullet k} \text{ für alle } i=1,2,...,I,\ j=1,2,...,J \text{ und } k=1,2,...,K$$

Daraus ergibt sich folgendes Hypothesenpaar:

$$H_0: \pi_{ijk} = \pi_{ij\bullet} \cdot \pi_{\bullet\bullet k} \text{ versus } H_1: \pi_{ijk} \neq \pi_{ij\bullet} \cdot \pi_{\bullet\bullet k}$$

Schritt 2: Auswahl von Teststatistik und Kennwerteverteilung:
Für die Teststatistiken werden wiederum die bei gültiger Nullhypothese erwarteten Häufigkeiten benötigt. Diese berechnen sich wieder über die ensprechenden Anteile in der Randtabelle der allgemeinen und eigenen Wirtschaftslage und der Randverteilung der Erwerbstätigkeit:

$$e_{ijk} = n \cdot p_{ij\bullet} \cdot p_{\bullet\bullet k} = n \cdot \frac{n_{ij\bullet}}{n} \cdot \frac{n_{\bullet\bullet k}}{n} = \frac{n_{ij\bullet} \cdot n_{\bullet\bullet k}}{n}$$

Prüfung der statistischen Unabhängigkeit einer von zwei anderen Variablen

Für die Berechnung kann die Ausgangstabelle so umsortiert werden, dass die Kombinationen der Ausprägungen von allgemeiner und eigener Wirtschaftslage in den Spalten stehen und die Ausprägungen der Erwerbstätigkeit in den Zeilen:

	Erwerbstätigkeit (W)			
	ja (w_1)		nein (w_2)	
Eigene Wirtschaftslage (Y)	Allgemeine Wirtschaftsl.(X)		Allgemeine Wirtschafsl. (X)	
	gut (x_1)	nicht gut (x_2)	gut (x_1)	nicht gut (x_2)
gut (y_1)	64.8% (160)	33.7% (557)	73.1% (152)	40.5% (469)
nicht gut (y_2)	35.2% (87)	66.3% (1094)	26.9% (56)	59.5% (688)

(Quelle: Allbus 2006)

$\Downarrow$

	Kombination von AWL und EWL				
AWL	gut	gut	nicht gut	nicht gut	
EWL	gut	nicht gut	gut	nicht gut	
Erwerbstätigkeit					Summe
ja	160	87	557	1094	1898
nein	152	56	469	688	1365
Summe	312	143	1026	1782	3263

Prüfung der statistischen Unabhängigkeit einer von zwei anderen Variablen

	Kombination von AWL und EWL				
AWL	gut	gut	nicht gut	nicht gut	
EWL	gut	nicht gut	gut	nicht gut	
Erwerbstätigkeit					Summe
ja	160	87	557	1094	1898
nein	152	56	469	688	1365
Summe	312	143	1026	1782	3263

Die Berechnung der erwarteten Häufigkeiten entspricht dann formal der Berechnung bei der Prüfung der Unabhängigkeit in einer bivariaten Kreuztabelle, bei der die Zeilenvariable aus der Kombination aller Ausprägungen der ersten beiden Variablen der trivariaten Tabelle, hier also aus AWL und EWL, gebildet wird. Die bei gültiger Nullhypothese chiquadrat-verteilten Teststatistiken berechnen sich dann nach:

$$\chi^2 = n \cdot \sum_i \frac{\left(p_i - \hat{\pi}_i \middle| H_0\right)^2}{\hat{\pi}_i \middle| H_0} = n \cdot \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(p_{ijk} - p_{ij\bullet} \cdot p_{\bullet\bullet k}\right)^2}{p_{ij} \cdot p_{\bullet\bullet k}} = \sum_{k=1}^{K}\sum_{i,j=1}^{I,J} \frac{\left(n_{k,ij} - \frac{n_{k\bullet} \cdot n_{\bullet,ij}}{n}\right)}{\frac{n_{k\bullet} \cdot n_{\bullet,ij}}{n}}$$

$$L^2 = 2 \cdot \sum_i n_i \cdot \ln\left(\frac{\hat{\pi}_i \middle| H_1}{\hat{\pi}_i \middle| H_0}\right) = 2 \cdot \sum_i n_i \cdot \ln\left(\frac{e_i \middle| H_1}{e_i \middle| H_0}\right) = \sum_{k=1}^{K}\sum_{i,j=1}^{I,J} 2 \cdot n_{k,ij} \cdot \ln\left(\frac{n_{k,ij}}{\frac{n_{k,\bullet} \cdot n_{\bullet,ij}}{n}}\right)$$

Schritt 3: Festlegung von Irrtumswahrscheinlichkeiten und kritischen Werten

Wenn wieder von einer Irrtumswahrscheinlichkeit von 5% ausgegangen wird, wird die Nullhypothese abgelehnt, wenn die Teststatistiken größer oder gleich dem 95%-Quantil einer Chiquadrat-Verteilung sind.

Die Freiheitsgrade ergeben sich aufgrund der formalen Gleichheit mit einem Test auf Unabhängigkeit von zwei Variablen nach df = (I·J–1)·(K–1) Freiheitsgraden. Da hier alle drei Variablen dichotomisiert sind, ergeben sich df = (2·2–1)·(2–1) = 3 Freiheitsgrade.

Der kritische Wert beträgt dann 7.815.

Schritt 4: Berechnung der Teststatistik und Entscheidung

Zur Berechnung der Teststatistik müssen zunächst wieder aus den beobachteten Daten die bei gültiger Nullhypothese erwarteten Häufigkeiten berechnet werden:

Bei Unabhängigkeit erwartete Häufigkeiten (n=3263)				
AWL	gut	gut	nicht gut	nicht gut
EWL	gut	nicht gut	gut	nicht gut
Erwerbstätigkeit	(n=312)	(n=143)	(n=1026)	(n=1782)
ja (n=1898)	181.48	83.18	596.80	1036.54
nein (n=1365)	130.52	59.82	429.20	745.46

Aus den erwarteten und den beobachteten Häufigkeiten werden sodann die Chiquadrat-Anteile in den einzelnen Zellen berechnet. Die Aufsummierung ergibt wieder die Werte der Teststatistiken.

Berechnung der Teststatistiken χ^2 und L^2 bei H_0: $\pi_{ijk} = \pi_{ij\bullet} \cdot \pi_{\bullet\bullet k}$						
erwerbstätig	AWL	EWL	n_{ijk}	e_{ijk}	χ^2-Anteil	L^2-Anteil
ja	gut	gut	160	181.48	2.54	-40.31
ja	gut	nicht gut	87	83.18	0.18	7.81
ja	nicht gut	gut	557	596.80	2.65	-76.88
ja	nicht gut	nicht gut	1094	1036.54	3.19	118.05
nein	gut	gut	152	130.52	0.16	46.32
nein	gut	nicht gut	56	59.82	0.24	-7.39
nein	nicht gut	gut	469	429.20	3.69	83.18
nein	nicht gut	nicht gut	688	745.46	4.43	-110.37
Summe			3263	3263.00	20.88	20.41

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(n_{ijk} - \frac{n_{ij\bullet} \cdot n_{\bullet\bullet k}}{n}\right)^2}{\frac{n_{ij\bullet} \cdot n_{\bullet\bullet k}}{n}} = 20.88 \qquad L^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} 2 \cdot n_{ijk} \cdot \ln\left(\frac{n_{ijk}}{\frac{n_{ij\bullet} \cdot n_{\bullet\bullet k}}{n}}\right) = 20.41$$

Da die Teststatistiken Werte von 20.88 bzw. 20.41 haben und damit größer als der kritische Wert 7.815 sind, ist die Nullhypothese zu verwerfen.

Bei einer Irrtumswahrscheinlichkeit von 5% sind die Erwerbstätigkeit und die bivariate Verteilung der Einschätzung der allgemeinen und der eigenen Wirtschaftslage voneinander statistisch abhängig.

Prüfung der statistischen Unabhängigkeit einer von zwei anderen Variablen

Berechnung der Teststatistiken χ^2 und L^2 bei H_0: $\pi_{ijk} = \pi_{ij\cdot} \cdot \pi_{\cdot\cdot k}$

erwerbstätig	AWL	EWL	n_{ijk}	e_{ijk}	χ^2-Anteil	L^2-Anteil
ja	gut	gut	160	181.48	2.54	-40.31
ja	gut	nicht gut	87	83.18	0.18	7.81
ja	nicht gut	gut	557	596.80	2.65	-76.88
ja	nicht gut	nicht gut	1094	1036.54	3.19	118.05
nein	gut	gut	152	130.52	0.16	46.32
nein	gut	nicht gut	56	59.82	0.24	-7.39
nein	nicht gut	gut	469	429.20	3.69	83.18
nein	nicht gut	nicht gut	688	745.46	4.43	-110.37
Summe			3263	3263.00	20.88	20.41

Die Chiquadrat-Anteile der einzelnen Zellen können genutzt werden, um zu prüfen, ob es in der jeweiligen Zelle eine signifikante Abweichung der beobachteten und der erwarteten Häufigkeiten gibt, da die Zahlen asymptotisch mit df=1 chiquadrat-verteilt sind. Obwohl nur die allerletzte Zelle mit einem Wert von 4.43 größer als der kritische Wert 3.84 sind, zeigt sich, dass insgesamt doch eine deutliche Abhängigkeit besteht.

Schritt 5: Kontrolle der Anwendungsvoraussetzungen

Da die kleinste erwartete Häufigkeite 59.82 > 5 ist, ist die asymptotische Annäherung an die Chiquadratverteilung vermutlich hinreichend genau.

Prüfung der statistischen Unabhängigkeit einer von zwei anderen Variablen

Wenn alle drei Variablen statistisch unabhängig voneinander sind, dann müssen auch zwei Variablen von der dritten statistisch unabhängig sein. Das Modell der Unabhängigkeit aller drei Variablen ist daher ein Spezialfall des Modells der Unabhängigkeit einer Variable von den anderen beiden. Daher kann mit Hilfe des LR-Tests die Unabhängigkeit aller drei Variablen gegen die Alternativhypothese der Unabhängigkeit einer Variablen von den anderen getestet werden:

$$H_0: \pi_{ijk} = \pi_{i++} \cdot \pi_{+j+} \cdot \pi_{++k} \text{ versus } H_1: \pi_{ijk} = \pi_{++k} \cdot \pi_{ij+}.$$

Die Teststatistik ist die Differenz der LR-Statistiken der ersten beiden Tests, die Zahl der Freiheitsgrade ebenfalls gleich der Differenz:

Modell	**L^2**	**df**
M_0: EWL, AWL und Erwerbstätigkeit sind unabhängig	184.65	4
M_1: Erwerbstätigkeit ist unabhängig von AWL u. EWL	20.41	3
Differenz (L^2-Teststatistik)	164.24	1

Die Nullhypothese, dass das strengere Modell M_0 nicht signifikant schlechter mit den Daten zu vereinbaren ist als das weniger strenge Modell M_1 muss bei einer Irrtumswahrscheinlichkeit von 5% abgelehnt werden, da die Teststatistik viel größer ist als der kritische Wert 3.84. der gleich dem 95%-Quantil der Chiquadrat-Verteilung ist.

Prüfung der Beziehung in den Partialtabellen

Innerhalb jeder Partialtabelle lassen sich alle Tests für bivariate Kreuztabellen anwenden.
Die Testergebnisse gelten dann für die Subpopulation, die durch die Ausprägung der Kontrollvariablen definiert ist.
Da die einzelnen Partialtabellen unabhängig voneinander sind, und die Summe von unabhängigen Chiquadrat-Verteilungen wiederum chiquadrat-verteilt ist, können bei Chiquadrat-Tests die Werte der Teststatistiken und der Freiheitsgrade in den Partialtabellen aufsummiert werden und ergeben dann einen Test über alle Partialtabellen. So kann mit Chiquadrat-Tests geprüft werden, ob es bei Kontrolle einer Drittvariablen einen Zusammenhang zwischen zwei Variablen gibt.
Als Beispiel wird der konditionale Zusammenhang zwischen allgemeiner und eigener Wirtschaftslage bei Kontrolle der Erwerbstätigkeit getestet.

Schritt 1: Formulierung von Null- und Alternativhypothese
Das Hypothesenpaar behauptet innerhalb aller Partialtabellen einen statistischen Zusammenhang:

$$H_0: \pi_{ij(k)} = \pi_{i\bullet(k)} \cdot \pi_{\bullet j(k)} \text{ versus } H_1: \pi_{ij(k)} \neq \pi_{i\bullet(k)} \cdot \pi_{\bullet j(k)}$$

Formal wird dazu in jeder Partialtabelle der übliche Test auf statistische Unabhängigkeit durchgeführt und anschließend die Summe der Teststatistiken berechnet.

Prüfung der Beziehung in den Partialtabellen

Schritt 2: Auswahl von Teststatistik und Kennwerteverteilung:
Die Teststatistik ist die Summe aus den Teststatistiken in den Partialtabellen:

$$\chi^2 = n \cdot \sum_i \frac{\left(p_i - \hat{\pi}_i \middle| H_0\right)^2}{\hat{\pi}_i \middle| H_0} = \sum_{k=1}^{K} n_{\bullet\bullet k} \cdot \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(p_{ij(k)} - p_{i\bullet(k)} \cdot p_{\bullet j(k)}\right)^2}{p_{i\bullet(k)} \cdot p_{\bullet j(k)}} = \sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij(k)} - \frac{n_{i\bullet(k)} \cdot n_{\bullet j(k)}}{n_{\bullet\bullet k}}\right)^2}{\frac{n_{i\bullet(k)} \cdot n_{\bullet j(k)}}{n_{\bullet\bullet k}}}$$

$$L^2 = 2 \cdot \sum_i n_i \cdot \ln\left(\frac{e_i \middle| H_1}{e_i \middle| H_0}\right) = \sum_{k=1}^{K} 2 \cdot \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij(k)} \cdot \ln\left(\frac{n_{ij(k)}}{\frac{n_{i\bullet(k)} \cdot n_{\bullet j(k)}}{n_{\bullet\bullet k}}}\right)$$

Wenn die Nullhypothese zutrifft und somit in allen Partialtabelle kein Zusammenhang zwischen Spalten- und Zeilenvariable besteht, dann sind beide Teststatistiken chiquadrat-verteilt.
Gibt es in mindestens einer Partialtabelle einen Zusammenhang, dann sind die Teststatistiken nichtzentral chiquadrat-verteilt.
Die Zahl der Freiheitsgrade ist gleich der Summe der Freiheitsgrade in den Partialtabellen:

$$df = K \cdot (I-1) \cdot (J-1)$$

Prüfung der Beziehung in den Partialtabellen

Schritt 3: Festlegung von Irrtumswahrscheinlichkeiten und kritischen Werten
Wenn von einer Irrtumswahrscheinlichkeit von 5% ausgegangen wird, wird die Nullhypothese abgelehnt, wenn die Teststatistiken größer oder gleich dem 95%-Quantil einer Chiquadrat-Verteilung sind. Da alle Variablen dichotomisiert sind, beträgt die Zahl der Freiheitsgrade hier

$$df = 2 \cdot (2-1) \cdot (2-1) = 2.$$

Aus der Tabelle mit Quantilen der Chiquadrat-Verteilung ist zu entnehmen, dass der kritische Wert 5.99 beträgt.

Schritt 4: Berechnung der Teststatistik und Entscheidung

	Erwerbstätigkeit (W)			
	ja (w_1)		nein (w_2)	
Eigene Wirtschaftslage (Y)	Allgemeine Wirtschaftsl.(X)		Allgemeine Wirtschafsl. (X)	
	gut (x_1)	nicht gut (x_2)	gut (x_1)	nicht gut (x_2)
gut (y_1)	64.8% (160)	33.7% (557)	73.1% (152)	40.5% (469)
nicht gut (y_2)	35.2% (87)	66.3% (1094)	26.9% (56)	59.5% (688)

(Quelle: Allbus 2006)

$$\chi^2 = 1898 \cdot \frac{(160 \cdot 1094 - 557 \cdot 87)^2}{717 \cdot 1181 \cdot 247 \cdot 1651} + 1365 \cdot \frac{(152 \cdot 688 - 469 \cdot 56)^2}{621 \cdot 744 \cdot 208 \cdot 1157} = 88.07 + 75.29 = 163.36$$

Prüfung der Beziehung in den Partialtabellen

	Erwerbstätigkeit (W)			
	ja (w_1)		nein (w_2)	
Eigene Wirtschaftslage (Y)	Allgemeine Wirtschaftsl.(X)		Allgemeine Wirtschafsl. (X)	
	gut (x_1)	nicht gut (x_2)	gut (x_1)	nicht gut (x_2)
gut (y_1)	64.8% (160)	33.7% (557)	73.1% (152)	40.5% (469)
nicht gut (y_2)	35.2% (87)	66.3% (1094)	26.9% (56)	59.5% (688)

(Quelle: Allbus 2006)

$$L^2 = 2 \cdot \left(\begin{array}{l} 160 \cdot \ln\left(\frac{160}{93.31}\right) + 557 \cdot \ln\left(\frac{557}{623.69}\right) + 87 \cdot \ln\left(\frac{87}{153.69}\right) + 1094 \cdot \ln\left(\frac{1094}{1027.31}\right) \\ +152 \cdot \ln\left(\frac{152}{94.63}\right) + 469 \cdot \ln\left(\frac{469}{526.37}\right) + 56 \cdot \ln\left(\frac{56}{113.37}\right) + 688 \cdot \ln\left(\frac{688}{630.63}\right) \end{array} \right)$$
$$= 85.19 + 76.64 = 161.83$$

Da die Teststatistiken mit Werten von 163.36 bzw. 161.83 größer sind als der kritische Wert 5.99, ist die Nullhypothese abzulehnen. Auch bei Kontrolle der Erwerbstätigkeit besteht zwischen der allgemeinen und der eigenen Beurteilung der Wirtschaftslage ein Zusammenhang..

Schritt 5: Kontrolle der Anwendungsvoraussetzungen
Da die kleinste erwartete Häufigkeite 717·247/1898 = 93.31 > 5 ist, ist die asymptotische Annäherung an die Chiquadratverteilung vermutlich hinreichend genau.

Prüfung der Beziehung in den Partialtabellen

Ganz analog kann auch der Zusammenhang zwischen der Erwerbstätigkeit und der eigenen Wirtschaftslage bei Kontrolle der allgemeinen Wirtschaftslage geprüft werden:

	Allgemeine Wirtschaftslage (W)			
	gut (w_1)		nicht gut (w_2)	
Eigene Wirt-	Erwerbstätigkeit (X)		Erwerbstätigkeit (X)	
schaftslage (Y)	ja (x_1)	nein (x_2)	ja (x_1)	nein (x_2)
gut (y_1)	64.8% (160)	73.1% (152)	33.7% (557)	40.5% (469)
nicht gut (y_2)	35.2% (87)	26.9% (56)	66.3% (1094)	59.5% (688)

(Quelle: Allbus 1996)

$$\chi^2 = 3.61 + 13.56 = 17.17 \text{ und } L^2 = 3.63 + 13.51 = 17.14$$

Beide Teststatistiken sind größer als der kritische Wert 5.99, so dass auch hier die Nullhypothese abzulehnen ist, dass in den Partialtabellen kein Zusammenhang besteht.
Da der kleinste Erwartungswert $143 \cdot 208/455 = 65.4 > 5$, sind auch hier die Anwendungsvoraussetzungen erfüllt.

Bei der Betrachtung der Partialtabellen fällt allerdings auf, dass der Zusammenhang zwischen Erwerbstätigkeit und eigener Wirtschaftslage bei der Einschätzung der allgemeinen Wirtschaftslage als gut das Signifikanzniveau von 5% nicht ganz erreicht, da $3.61 < 3.84 = \chi^2_{95\%;df=1}$.
Es scheint daher nicht ausgeschlossen, dass nur bei einer nicht guten Beurteilung der allgemeinen Lage ein Zusammenhang zwischen Erwerbstätigkeit und eigener Wirtschaftslage besteht.

Z-Test auf Interaktion über Vergleich der Prozentsatzdifferenzen

	Allgemeine Wirtschaftslage (W)			
	gut (w_1)		nicht gut (w_2)	
Eigene Wirt-	Erwerbstätigkeit (X)		Erwerbstätigkeit (X)	
schaftslage (Y)	ja (x_1)	nein (x_2)	ja (x_1)	nein (x_2)
gut (y_1)	64.8% (160)	73.1% (152)	33.7% (557)	40.5% (469)
nicht gut (y_2)	35.2% (87)	26.9% (56)	66.3% (1094)	59.5% (688)

(Quelle: Allbus 1996)

Dies würde bedeuten, dass ein Interaktionseffekt zwischen Erwerbtsätigkeit und allgemeiner Wirtschaftslage besteht, da die Erwerbtstätigkeit die eigene Wirtschaftslage nur bei nicht guter Wirtschaftslage beeinflusst.
Auf der anderen Seite beträgt die Prozentsatzdifferenz in der linken Partialtabelle –8.3 Punkte, in der rechten Partialtabelle dagegen nur –6.8 Punkte. Die Insignifikanz der linken Partialtabelle kann also auch Folge der deutlich geringeren Fallzahl sein.
Die Differenz der Prozentsatzdifferenzen von 1.5 Punkten stellt sich notwendigerweise in gleicher Höhe ein, wenn die Erwerbstätigkeit Kontrollvariable ist. Bei Erwerbstätigen beträgt die Prozentsatzdifferenz der allgemeinen Wirtschaftslage 31.1 Punkte, bei nicht Erwerbstätigen mit 32.6 Punkten genau 1.5 Punkte mehr.

Da Prozentsatzdifferenzen asymptotisch normalverteilt sind, kann ein möglicher Interaktionseffekt über einen Z-Test der Differenzen der beiden voneinander unabhängigen Prozentsatzdifferenzen in den Partialtabellen getestet werden.

Z-Test auf Interaktion über Vergleich der Prozentsatzdifferenzen

	Allgemeine Wirtschaftslage (W)			
	gut (w_1)		nicht gut (w_2)	
Eigene Wirtschaftslage (Y)	Erwerbstätigkeit (X)		Erwerbstätigkeit (X)	
	ja (x_1)	nein (x_2)	ja (x_1)	nein (x_2)
gut (y_1)	64.8% (160)	73.1% (152)	33.7% (557)	40.5% (469)
nicht gut (y_2)	35.2% (87)	26.9% (56)	66.3% (1094)	59.5% (688)

(Quelle: Allbus 1996)

Schritt 1: Formulierung von Null- und Alternativhypothese

Bei einem Interaktionseffekt müssen sich die beiden Prozentsatzdifferenzen in der Population unterscheiden. Die Nullhypothese behauptet daher keine Differenz, die Alternativhypthese eine Differenz:

$$H_0: d_{YX(Z=1)}\% - d_{YX(Z=2)}\% = 0 \text{ vs. } H_1: d_{YX(Z=1)}\% - d_{YX(Z=2)}\% \neq 0.$$

Schritt 2: Auswahl von Teststatistik und Kennwerteverteilung:

Prozentsatzdifferenzen bzw. Anteilsdifferenzen sind asymptotisch normalverteilt. Dann ist auch die Differenz von unabhängigen Prozentsatzdifferenzen asymptotisch normalverteilt. Die Teststatistik berechnet sich nach:

$$Z = \frac{\left(d_{YX(1)}\% - d_{YX(1)}\%\right)/100}{\sqrt{\left(\frac{n_{11(1)} \cdot n_{21(1)}}{n^3_{\bullet\bullet(1)}} + \frac{n_{12(1)} \cdot n_{22(1)}}{n^3_{\bullet\bullet(1)}}\right) + \left(\frac{n_{11(2)} \cdot n_{21(2)}}{n^3_{\bullet\bullet(2)}} + \frac{n_{12(2)} \cdot n_{22(2)}}{n^3_{\bullet\bullet(2)}}\right)}}$$

Z-Test auf Interaktion über Vergleich der Prozentsatzdifferenzen

	Allgemeine Wirtschaftslage (W)			
	gut (w_1)		nicht gut (w_2)	
Eigene Wirtschaftslage (Y)	Erwerbstätigkeit (X)		Erwerbstätigkeit (X)	
	ja (x_1)	nein (x_2)	ja (x_1)	nein (x_2)
gut (y_1)	64.8% (160)	73.1% (152)	33.7% (557)	40.5% (469)
nicht gut (y_2)	35.2% (87)	26.9% (56)	66.3% (1094)	59.5% (688)

(Quelle: Allbus 1996)

Schritt 3: Festlegung von Irrtumswahrscheinlichkeiten und kritischen Werten

Wenn von einer Irrtumswahrscheinlichkeit von 5% ausgegangen wird, wird die Nullhypothese im zweiseitigen Test abgelehnt, wenn die Teststatistik kleiner/gleich dem 2.5%-Quantil oder größer/gleich dem 97.5%-Quantil der Standardnormalverteilung ist. Die kritischen Werte sind daher ±1.96.

Schritt 4: Berechnung der Teststatistik und Entscheidung

$$Z = \frac{\left(\frac{160}{247} - \frac{152}{208}\right) - \left(\frac{557}{1651} - \frac{469}{1157}\right)}{\sqrt{\left(\frac{160 \cdot 87}{247^3} + \frac{152 \cdot 56}{208^3}\right) + \left(\frac{557 \cdot 1094}{1651^3} + \frac{469 \cdot 688}{1157^3}\right)}} = -0.319$$

Z-Test auf Interaktion über Vergleich der Prozentsatzdifferenzen

	Allgemeine Wirtschaftslage (W)			
	gut (w_1)		nicht gut (w_2)	
Eigene Wirt-	Erwerbstätigkeit (X)		Erwerbstätigkeit (X)	
schaftslage (Y)	ja (x_1)	nein (x_2)	ja (x_1)	nein (x_2)
gut (y_1)	64.8% (160)	73.1% (152)	33.7% (557)	40.5% (469)
nicht gut (y_2)	35.2% (87)	26.9% (56)	66.3% (1094)	59.5% (688)

(Quelle: Allbus 1996)

Da die Teststatistik größer ist als der untere und kleiner ist als der obere kritische Wert, kann die Nullhypothese bei einer Irrtunswahrscheinlichkeit von 5% nicht verworfen werden.
Es kann also nicht davon ausgegangen werdne, dass ein Interaktionseffekt besteht. Vermutlich sind die Prozentsatzdifferenzen in der Population gleich groß.

Schritt 5: Kontrolle der Anwendungsvoraussetzungen
Zur Prüfung der Anwendungsvoraussetzungen muss für jede Prozentsatzdifferenzdifferenz die hinreichende Annäherung an die Normalverteilung geprüft werden.
Die Fallzahlen sind in allen Spalten größer 60. Darüber hinaus gilt: $247 \cdot 87/160 = 134.3 > 9$, $208 \cdot 56/152 = 76.6 > 9$, $1651 \cdot 557/1094 = 840.6 > 9$ und $1157 \cdot 469/688 = 788.7 > 9$.
Die Anwendungsvoraussetzungen scheinen erfüllt zu sein.

Lerneinheit 4: Bivariate lineare Regression (Wiederholung)

In der Tabellenanalyse wird der Einfluss einer erklärenden Variable X auf eine abhängige Variable Y untersucht, indem die bedingten Verteilungen der abhängigen Variable bei unterschiedlichen Ausprägungen der erklärenden Variable verglichen werden.
Da metrische Variablen in der Regel sehr vielen Ausprägungen haben, ist es praktisch unmöglich, eine große Zahl von bedingten Verteilungen vergleichend zu betrachten.
Statt der bedingten Verteilungen werden daher bei metrischen Variablen nur Kenngrößen dieser bedingten Verteilungen bei unterschiedlichen Werten der erklärenden Variable analysiert.

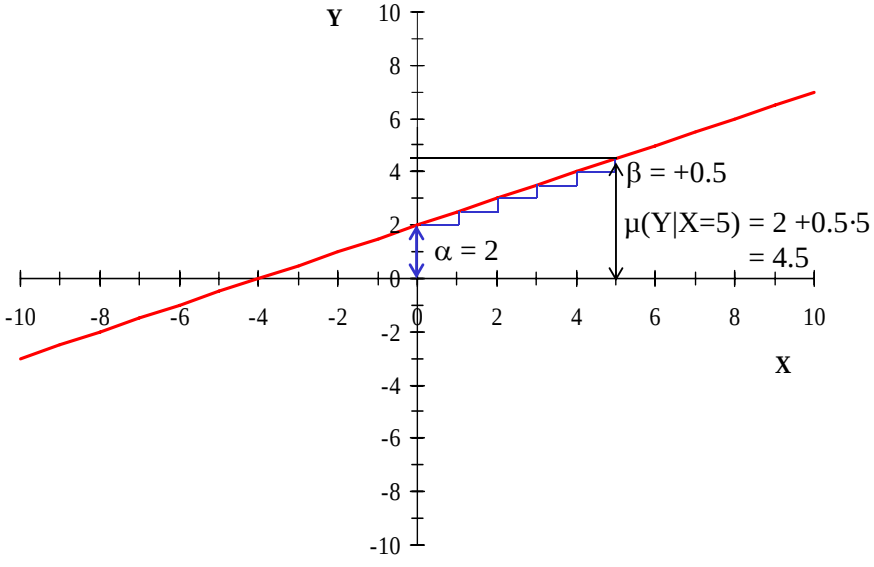
Im ***linearen Regressionsmodell*** wird dabei angenommen, dass sich die ***bedingten Populationsmittelwerte*** der abhängigen Variable durch eine ***lineare Funktion*** der ***Ausprägungen*** der ***erklärenden Variable*** beschreiben lassen.
Wenn $\mu(Y|X = x)$ die bedingten Populationsmittelwerte von Y gegeben die Ausprägung x der erklärenden Variable X symbolisiert, gilt dann:

$$\mu\left(Y\middle|X = x\right) = \alpha + \beta \cdot x$$

Als Beispiel wird der gerichtete Zusammenhang zwischen dem Alter der Partner von insgesamt 185 Befragten aus dem Allbus 2006 betrachtet, die eine Lebenspartnerin bzw. einen Lebenspartner haben, mit dem sie nicht zusammenleben. Da nur das Geschlecht der befragten Person und nicht ihres Partners erfasst wurde, wird in den folgenden Analysen unterstellt, dass es sich um heterosexuelle Partnerschaften handelt.

Bedeutung der Regressionskoeffizienten

Die beiden ***Regressionskoeffizienten*** α (alpha) und β (beta) der linearen Gleichung $\mu(Y|X=x) = \alpha+\beta\cdot x$ (die nicht mit den Symbolen für die Fehlerwahrscheinlichkeiten beim statistischen Testen verwechselt werden dürfen!) bestimmen die Lage der Regressionsgerade bei einer grafischen Darstellung der Beziehung in einem Koordinatensystem.



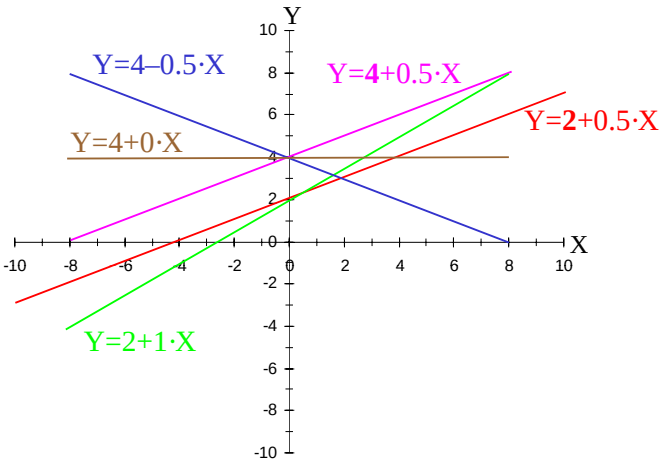
Im Beispiel ist eine Regressionsgerade mit der Regressionskonstante $\alpha = 2$ und $\beta = 0.5$ eingezeichnet:

$$\mu(Y|X) = 2 + 0.5 \cdot X.$$

Wenn X = 0, dann ist der bedingte Mittelwert von Y = 2.
Wenn X um +1 Einheit ansteigt, steigt der bedingte Mittelwert von Y um +0.5 Einheiten an.

- Die ***Regressionskonstante*** (auch als ***Interzept*** bezeichnet) α gibt den Mittelwert von Y an, wenn X = 0 ist.
- Das ***Regressionsgewicht*** β gibt die ***Steigung*** der Geraden an.

Bedeutung der Regressionskoeffizienten



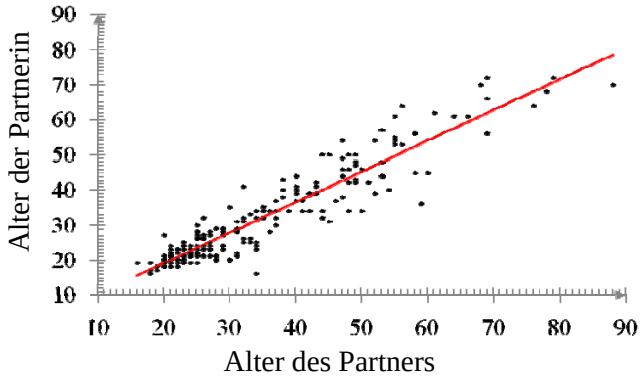
Unterscheiden sich zwei Regressionsgeraden nur beim Wert des Interzepts α, dann verlaufen sie parallel.
Die Gerade mit dem größeren Wert beim Interzept liegt dabei über der Gerade mit dem kleineren Wert beim Interzept.

Je größer der Wert von b, desto steiler ist der Kurvenverlauf.

Bei positiven Werten steigt die Kurve an, bei negativen Werten fällt die Kurve ab, ist b=0, verläuft die Gerade wagerecht.

Hat das Regressionsgewichts einen positiven Wert ($\beta > 0$), weist dies auf eine positive Beziehung hin: je größer der Wert von X, desto größer ist auch der bedingte Mittelwert von Y.
Ist das Regressionsgewicht negativ $\beta < 0$, spricht dies für eine negative Beziehung: je größer der Wert von X, desto kleiner ist der bedingte Mittelwert von Y.
Bei Unabhängigkeit zwischen X und Y gilt notwendigerweise: $\beta = 0$. Alle bedingten Mittelwerte sind dann gleich der Konstante α. Das Umgekehrte gilt nicht: Wenn das Regressionsgewicht null ist, besteht zwar kein linearer bzw. allgemeiner kein tendenziell monotoner Zusammenhang; möglich ist aber ein nichtmonotoner, z.B. wellenförmig verlaufender Zusammenhang.

Bedeutung der Regressionskoeffizienten



Im Beispiel der linearen Regression des Alters des weiblichen Partnerin auf das Alter des männlichen Partners ergibt sich bei den Daten des Allbus 1996 folgende geschätzte Regressionsgleichung:

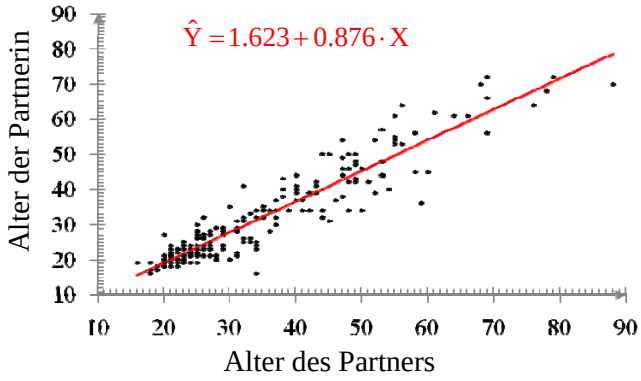
$$\hat{\mu}(Y|X=x) = \hat{Y} = 1.623 + 0.876 \cdot x$$

Mit Hilfe der Regressionskoeffizienten lassen sich für beliebige Werte der erklärenden Variable X die bedingten Populationsmittelwerte der abhängigen Variable Y berechnen.
Da diese Werte zur Vorhersage der Werte der abhängigen Variable benutzt werden können, werden Sie auch als ***Vorhersagewerte*** bezeichnet.

X	$\hat{\mu}(Y\|X) = 1.623 + 0.876 \cdot X$
20	1.623 + 0.876 ·20 = 19.143
21	1.623 + 0.876 ·21 = 20.019
30	1.623 + 0.876 ·30 = 27.903
0	1.623 + 0.876 ·0 = 1.623
–5	1.623 + 0.876·(–5) = –2.757

Wenn im Beispiel der Mann 20 Jahre ist, ist also damit zu rechnen, dass seine Partnerin im Mittel 19.143 Jahre alt ist. Ist der Mann 30, dann ist die Frau im Mittel 27.903 Jahre alt.
Die Berechnung erlaubt allerdings auch ganz unsinnige Vorhersagen: Ist der Mann 0 Jahre alt, ist die Frau 1.623 Jahre alt, ist der Mann –5 Jahre alt, ist die Frau –2.757 Jahre alt.

Bedeutung der Regressionskoeffizienten

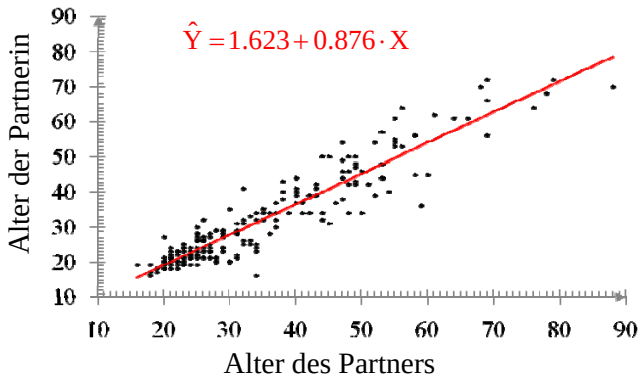


Bei der Interpretation ist also zu beachten, dass die Vorhersagewerte nur bei sinnvollen Werten der erklärenden Variable auch sinnvolle Vorhersagewerte liefern können.

Generell wird bei der Interpretation einer Regressionsgleichung zwischen einer Kausalinterpretation und einer rein prognostischen Interpretation unterschieden.

- **Kausalinterpretation**:
  Der Anstieg des Werte der erklärenden Variable ***bewirkt*** im Mittel eine Veränderung der abhängigen Variable um β Einheiten.
  Im Beispiel würde also bei einem Anstieg des Alters des Mannes (X) um 1 Jahr das Alter der Partnerin (Y) bei unverheiratet zusammenlebenden Paaren im Durchschnitt um 0.853 Jahre ansteigen. Diese kausale Interpretation ist hier offensichtlich nicht sinnvoll!
- **Prognostische Interpretation**:
  Die Vorhersagewerte werden benutzt, um bei einem beliebigen Fall bei Kenntnis des Wertes der erklärenden Variable möglichst gut den Wert der abhängigen Variable ***vorherzusagen***.
  Im Beispiel macht nur die prognostische Interpretation Sinn.

Bedeutung der Regressionskoeffizienten



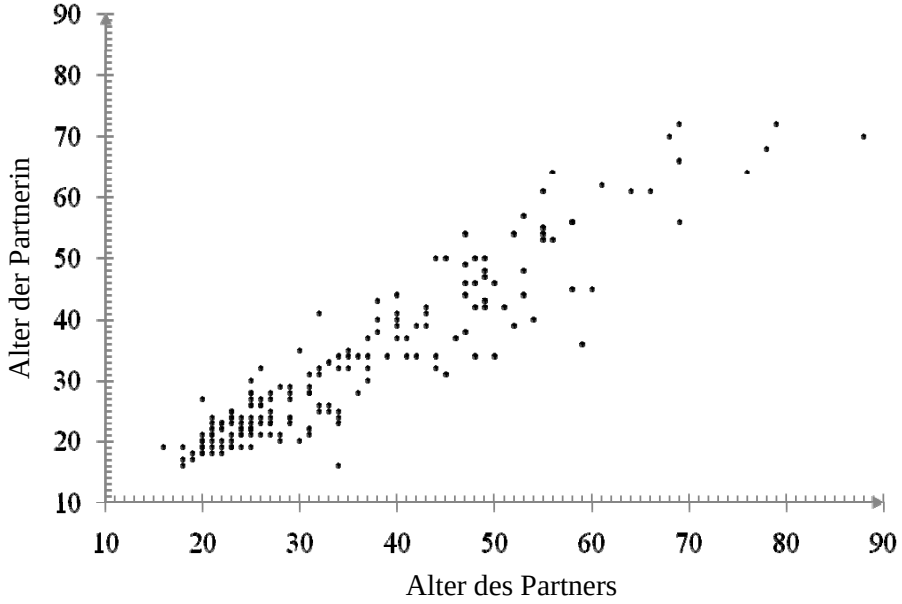
$$\hat{y}_{x=18} = 1.623 + 0.876 \cdot 20 = 19.143$$
$$\hat{y}_{x=45} = 1.623 + 0.876 \cdot 45 = 41.043$$
$$\hat{y}_{x=70} = 1.623 + 0.876 \cdot 70 = 62.943$$

Aus den Vorhersagewerten der Regressionsgleichung ist auch abzulesen, dass mit zunehmenden Alter des Mannes der mittlere Altersabstand zur im Durchschnitt jüngeren Partnerin ansteigt.

Obwohl im Beispiel die Regressionsgleichung selbst nicht kausal interpretiert wird, kann die Vorhersagegleichung doch auch als Hinweis für eine spezifische kausale Interpretation genutzt werden.

So mag es die Norm geben, nach denen in einer Partnerschaft die Frau eher jünger zu sein hat als der Mann. Der zunehmende Altersabstand kann dahingehend gedeutet werden, dass entweder ältere Männer stärker als jüngere Männer deutlich jüngere Frauen bevorzugen bzw. umgekehrt ältere Frauen stärker als jüngere Frauen deutlich ältere Männer, oder dass sich die Norm des älter zu seienden Mannes im Laufe der Zeit abschwächt, so das die Norm heute eher von älteren als von jüngeren Personen beachtet wird. Das Regressionsgewicht ungleich 0 kann aber auch Folge von Messfehlern sein, der zunehmende Abstand dann ein Artefakt hervorgerufen durch die messfehlerbedingte Regression zur Mitte.

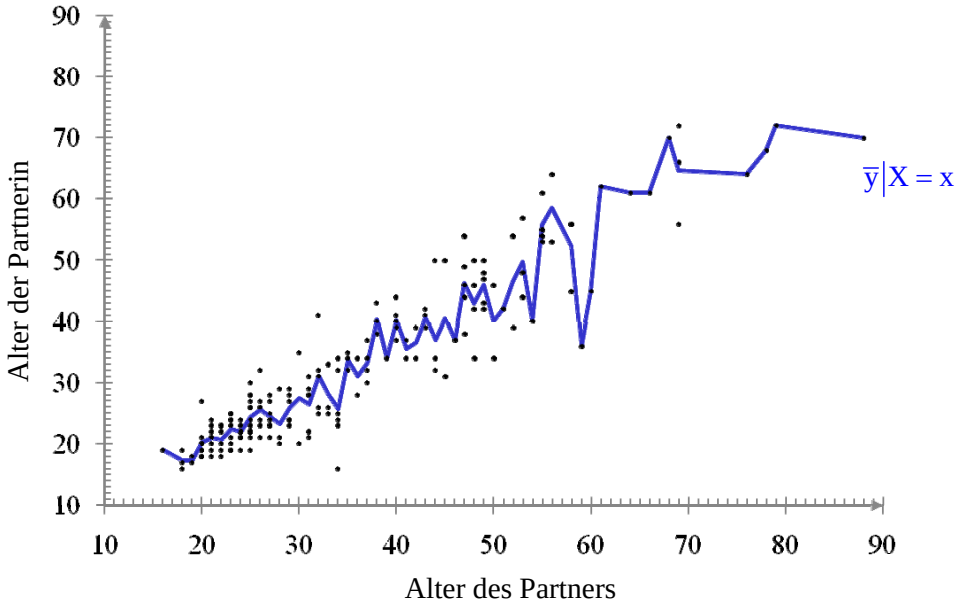
OLS-Schätzung der Regressionskoeffizienten



In empirischen Anwendungen des linearen Regressionsmodells müssen die Werte der Regressionskoeffizienten erst aus Stichprobendaten geschätzt werden. Dazu liegen zunächst nur die Stichprobenrealisierungen vor, die als Punkte in ein Koordinatensystem eingetragen werden können, wobei die Werte der erklärenden Variable entlang der horizontalen X-Achse und die Werte der abhängigen Variable entlang der vertikalen Y-Achse aufgetragen werden.

Das Beispiel zeigt alle 185 Fälle von nicht zusammenlebenden unverheirateten Paaren aus der Stichprobe des Allbus 2006.

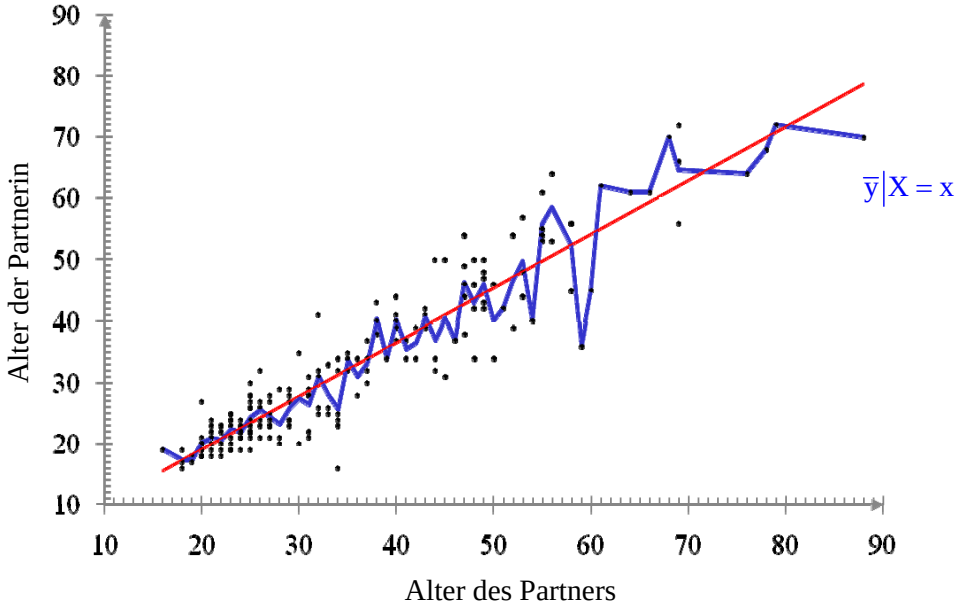
OLS-Schätzung der Regressionskoeffizienten



Werden in der Stichprobe für jede realisierte Ausprägung der erklärenden Variable die bedingten Stichprobenmittelwerte der abhängigen Variable berechnet und in das Koordinatensystem eingetragen, ergibt sich bei einer Verbindung der Mittelwerte durch eine Kurve in der Regel keine gerade Linie, sondern eine unregelmäßig verlaufende Kurve, die jedoch möglicherweise einen Trend erkennen lässt.

Im Beispiel ist so ersichtlich, dass das mittlere Alter der Frau tendenziell mit dem Alter des Mannes ansteigt.

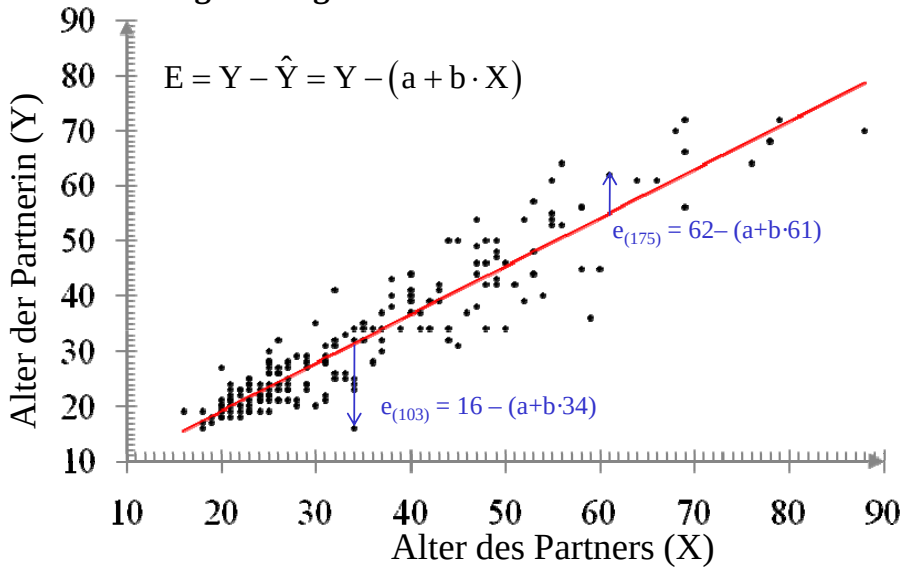
OLS-Schätzung der Regressionskoeffizienten



Für die Schätzung wird angenommen, dass die Abweichungen der bedingten Stichprobenmittelwerte von einer Geraden ausschließlich auf den Zufälligkeiten der Fallauswahl beruhen und sich eine Gerade ergeben würde, lägen alle Fälle der Population vor.

Die unbekannte Regressionsgerade wird dann nach der ***Kleinstquadratmethode*** *(engl: ordinary least squares,* ***OLS****)* so aus den Stichprobendaten geschätzt, dass die Summe der quadrierten Differenzen der Realisierungen der abhängigen Variable von der geschätzten Regressionsgerade minimal ist.

OLS-Schätzung der Regressionskoeffizienten



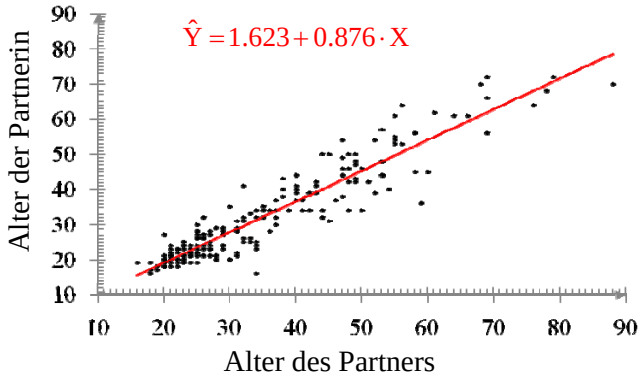
Beim Fall mit der Rangnummer 103 ist der Partner 34 Jahre alt und die Partnerin 16 Jahre.
$\Rightarrow e_{(103)} = 16 - (a+b\cdot 34)$

Die Differenzen $e_i = y_i - (a+b\cdot x_i)$ der Realisierungen y_i der abhängigen Variable von der (geschätzten) Regressionsgerade werden als ***Residuen*** bezeichnet und zur Residualvariable E zusammengefasst.

Da es positive und negative Residuen gibt, werden bei der OLS-Schätzung die geschätzten Regressionskoeffizienten a und b so bestimmt, dass die Summe der quadrierten Stichprobenresiduen minimal ist:

$$Q(a,b) = \sum_{i=1}^{n} \left(y_i - (a + b \cdot x_i)\right)^2 = \sum_{i=1}^{n} e_i^2 \overset{!}{=} \text{minimal}$$

OLS-Schätzung der Regressionskoeffizienten



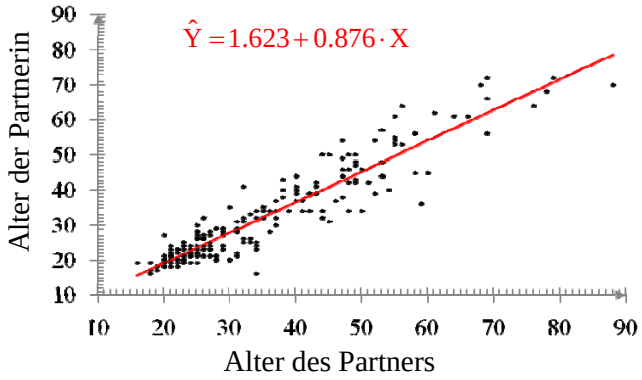
Die Regressionskonstante a ist dann die Differenz von b mal dem Stichprobenmittelwert der erklärenden Variable vom Stichprobenmittelwert der abhängigen Variable.

$$a = \bar{y} - b \cdot \bar{x}$$

Mit Hilfe der Differentialrechnung kann gezeigt werden, dass die ***Minimierungsfunktion*** Q(a,b) genau dann ein Minimum aufweist, wenn b der Quotient aus der Kovariation bzw. Kovarianz zwischen abhängiger und erklärender Variable geteilt durch die Variation bzw. Varianz der erklärenden Variable ist.

$$b = \frac{SP_{YX}}{SS_X} = \frac{\sum_{i=1}^{n}(y_i - \bar{y}) \cdot (x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} y_i \cdot x_i - \frac{\sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{\sum_{i=1}^{n} y_i \cdot x_i - n \cdot \bar{y} \cdot \bar{x}}{\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2} = \frac{s_{YX}}{s_X^2} = \frac{\hat{\sigma}_{YX}}{\hat{\sigma}_X^2}$$

OLS-Schätzung der Regressionskoeffizienten



$$b = \frac{SP_{YX}}{SS_X} = \frac{34087.3243}{38899.9460} = 0.876$$

$$= \frac{s_{YX}}{s_X^2} = \frac{184.2558}{210.2760} = 0.876$$

$$= \frac{\hat{\sigma}_{YX}}{\hat{\sigma}_X^2} = \frac{185.2572}{211.4127} = 0.876$$

$$a = 32.7568 - 0.876 \cdot 35.5405$$

$$\approx 1.623$$

Fall	X	Y	X^2	Y^2	X·Y
1	21	21	441	441	441
...	...	...	...	…	...
185	61	62	3721	3844	3782
∑	6575	6060	272579	232676	249463

$$n = 185 \; ; \; \sum x_i = 6575 \; ; \; \sum y_i = 6060$$

$$\sum x_i^2 = 272579 \; ; \; \sum y_i^2 = 23676$$

$$\sum y_i \cdot x_i = 249463$$

$$\bar{x} = 6575 / 185 = 35.5405 \; ; \; \bar{y} = 6060 / 185 = 32.7568$$

$$SS_X = 272579 - 6575^2 / 185 = 38899.9460$$

$$SS_Y = 232676 - 6060^2 / 185 = 34170.0541$$

$$SP_{YX} = 249463 - 6575 \cdot 6060 / 185 = 34087.3243$$

$$s_X^2 = 42506.196 / 185 = 210.2760$$

$$s_Y^2 = 37311.4378 / 185 = 184.7030$$

$$s_{YX} = 36241.4919 / 185 = 184.2558$$

$$\hat{\sigma}_X^2 = 42506.196 / 184 = 211.4127$$

$$\hat{\sigma}_Y^2 = 37311.4378 / 184 = 185.7068$$

$$\hat{\sigma}_{YX} = 36241.4919 / 184 = 185.2572$$

Eigenschaften der Stichprobenresiduen bei der OLS-Methode

Die Kleinstquadratmethode führt dazu, dass die Stichprobenresiduen bestimmte Eigenschaften haben:

(1) Die Summe der Residuen ist 0:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y_i - a - b \cdot x_i = 0$$

(2) Dann ist auch der Mittelwert der Residuen 0 und die Stichprobenvarianz gleich dem Mittelwert der quadrierten Residuen:

$$\bar{e} = \frac{1}{n} \cdot \sum_{i=1}^{n} e_i = 0 \; ; \; s_E^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} e_i^2$$

Da die Varianz der Stichprobenresiduen gleich dem Mittelwert der quadrierten Residuen ist, minimiert die OLS-Methode also auch diese Residualvarianz in der Stichprobe.

(3) Die Residuen sind mit den Realisierungen der erklärenden Variablen und den Vorhersagewerten unkorreliert:

$$SP_{XE} = \sum_{i=1}^{n} (x_i - \bar{x}) \cdot e_i = \sum_{i=1}^{n} x_i \cdot e_i = 0 \Rightarrow s_{XE} = \frac{SP_{XE}}{n} = 0 \; ; \; r_{XE} = \frac{s_{XE}}{s_X \cdot s_E} = 0$$

$$SP_{\hat{Y}E} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}) \cdot e_i = \sum_{i=1}^{n} y_i \cdot e_i = 0 \Rightarrow s_{\hat{Y}E} = \frac{SP_{\hat{Y}E}}{n} = 0 \; ; \; r_{\hat{Y}E} = \frac{s_{\hat{Y}E}}{s_{\hat{Y}} \cdot s_E} = 0$$

Variationszerlegung

Bei der OLS-Schätzung der Regressionskoeffizienten wird formal die abhängige Variable Y als eine Linearkombination aus der erklärenden Variable X und einer Residualvariable E aufgefasst:

$$Y = \hat{Y} + E = a + b \cdot X + E$$

Aus der Unkorreliertheit der Stichprobenresiduen mit den Vorhersagewerten und den Regeln für Linearkombinationen von unabhängigen Variablen folgt dann, dass die Varianz bzw. Variation der abhängigen Variable gleich der Summe aus der Varianz bzw. Variation der Vorhersagewerte und der Stichprobenresiduen ist:

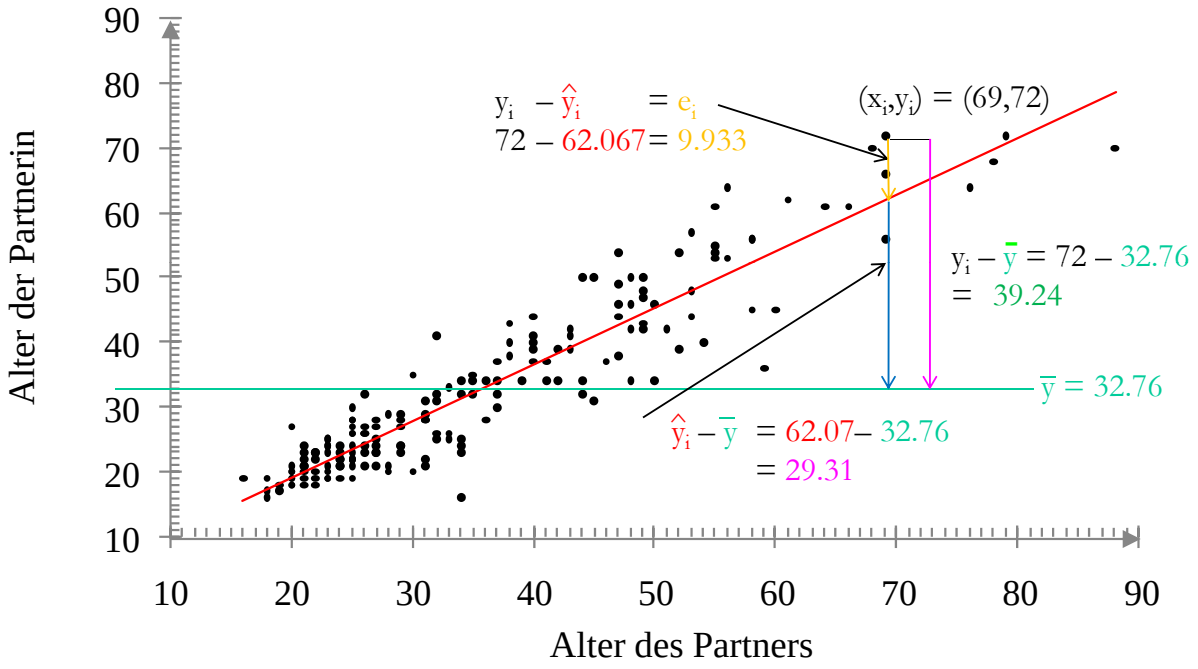
$$SS_Y = SS_{\hat{Y}} + SS_E = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$s_Y^2 = s_{\hat{Y}}^2 + s_E^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \frac{1}{n} \cdot \sum_{i=1}^{n} e_i^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Als PRE-Maß für die Stärke der Beziehung bietet es sich daher an, die Variationsverhältnisse in Beziehung zu setzen. Das so gebildete asymmetrische Zusammenhangsmaß ist der ***Determinationskoeffizient*** R^2*:*

$$R^2 = 1 - \frac{s_E^2}{s_Y^2} = 1 - \frac{SS_E}{SS_Y} = \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad = \quad \frac{s_{\hat{Y}}^2}{s_Y^2} = \frac{SS_{\hat{Y}}}{SS_Y} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Variationszerlegung



In der grafischen Darstellung ist gut erkennbar, wie bei den Fällen die Distanzen der abhängigen Variable zur Regressionsgerade (rot eingezeichnet) meistens geringer sind als zum Mittelwert (grün eingezeichnet) der abhängigen Variable.

Determinationskoeffizient, Korrelation und Regressionsgewichte

In der bivariaten linearen Regression gibt es zwischen dem Determinationskoeffizient, der Produktmomentkorrelation und den beiden Regressionsgewichten der Regressionen von Y auf X und von X auf Y Zusammenhänge:

- Der Determinationskoeffizient ist das Quadrat der Produktmomentkorrelation.

$$R^2 = \frac{SS_{\hat{Y}}}{SS_Y} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(a + b \cdot x_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}((\bar{y} - b \cdot \bar{x}) + b \cdot x_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(b \cdot (x_i - \bar{x}))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= b^2 \cdot \frac{SS_X}{SS_Y} = b^2 \cdot \frac{s_X^2}{s_Y^2} = \frac{(s_{YX})^2}{(s_X^2)^2} \cdot \frac{s_X^2}{s_Y^2} = \frac{(s_{YX})^2}{s_x^2 \cdot s_Y^2} = \left(\frac{s_{YX}}{s_X \cdot s_Y}\right)^2 = (r_{YX})^2$$

- Der Determinationskoeffizient ist das Produkt der Regressionsgewichte der Regressionen von Y auf X und von X auf Y.
  Um die Richtung der Regression zu unterscheiden, werden die Regressionskoeffizienten mit Indizes versehen:

$$Y = a_{Y.X} + b_{Y.X} \cdot X + E_Y \Rightarrow b_{Y.X} = \frac{SP_{YX}}{SS_X} = \frac{s_{YX}}{s_X^2} = \frac{\hat{\sigma}_{YX}}{\hat{\sigma}_X^2} \; ; \; a_{Y.X} = \bar{y} - b_{Y.X} \cdot \bar{x}$$

$$X = a_{X.Y} + b_{X.Y} \cdot Y + E_X \Rightarrow b_{X.Y} = \frac{SP_{XY}}{SS_Y} = \frac{s_{XY}}{s_Y^2} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_Y^2} \; ; \; a_{X.Y} = \bar{x} - b_{X.Y} \cdot \bar{y}$$

Determinationskoeffizient, Korrelation und Regressionsgewichte

Für den Determinationskoeffizient gilt dann:

$$R^2 = \frac{(s_{YX})^2}{s_X^2 \cdot s_Y^2} = \frac{s_{YX}}{s_X^2} \cdot \frac{s_{YX}}{s_Y^2} = b_{Y.X} \cdot b_{X.Y}$$

- Werden abhängige und erklärende Variable über die Z-Transformation standardisiert, dann sind die Regressionskonstanten null und die standardisierten Regressionsgewichte gleich der Produktmomentkorrelation.
  Dies gilt sowohl für die Regression von Y auf X als auch für die Regression von X auf Y.

$$Z_Y = \frac{Y - \bar{y}}{s_Y} \; ; \; Z_X = \frac{X - \bar{x}}{s_X} \; ; \; Z_Y = a_{Z_Y.Z_X} + b_{Z_Y.Z_X} \cdot Z_X + E_{Z_Y} \; ; \; Z_X = a_{Z_X.Z_Y} + b_{Z_X.Z_Y} \cdot Z_Y + E_{Z_X}$$

$$\Rightarrow b_{Z_Y.Z_X} = \frac{s(Z_Y, Z_X)}{s_{Z_X}^2} = \frac{r_{YX}}{1} = r_{YX} \; ; \; a_{Z_Y.Z_X} = \bar{z}_Y - b_{Z_Y.Z_X} \cdot \bar{z}_X = 0 - b_{Z_Y.Z_X} \cdot 0 = 0$$

$$\Rightarrow b_{Z_X.Z_Y} = \frac{s(Z_X, Z_Y)}{s_{Z_Y}^2} = \frac{r_{YX}}{1} = r_{YX} \; ; \; a_{Z_X.Z_Y} = \bar{z}_X - b_{Z_X.Z_Y} \cdot \bar{z}_Y = 0 - b_{Z_X.Z_Y} \cdot 0 = 0$$

Anwendungsbeispiel

Variable (n=185)	Mittelwert	(Ko-) Variationen	Stichproben-(ko)varianzen	Korrelationen
Alter des männl. Partners	35.541	38899.946	210.276	1.000
Alter der Partnerin	32.757	34087.324 34170.054	184.256 184.703	0.935 1.000

Zur Verdeutlichung werden im Beispiel aus dem Allbus 2006 alle Koeffizienten der beiden möglichen Regressionen berechnet.
Als Ausgangsdaten werden neben der Fallzahl die Mittelwerte der Variablen und entweder die Variationen und Kovariationen, die Varianzen und Kovarianzen oder die Korrelationen und Standardabweichungen benötigt.
Bei den Varianzen und Kovarianzen ist darauf zu achten, ob es sich um Stichprobenstatistiken s_Y^2, s_X^2, s_{XY} handelt oder um die geschätzten Populationsparameter $\hat{\sigma}_Y^2, \hat{\sigma}_x^2, \hat{\sigma}_{XY}$.

- Wenn Variationen oder Varianzen vorliegen, werden zunächst die unstandardisierten Regressionskoeffizienten berechnet und daraus anschließend die standardisierten Gewichte.

$$b_{Y.X} = \frac{SP_{XY}}{SS_X} = \frac{34087.324}{38899.946} = 0.876 \qquad b_{X.Y} = \frac{SP_{XY}}{SS_Y} = \frac{34087.324}{34170.054} = 0.998$$

$$b_{Y.X} = \frac{s_{XY}}{s_X^2} = \frac{184.256}{210.276} = 0.876 \qquad b_{X.Y} = \frac{s_{XY}}{s_Y^2} = \frac{184.256}{184.703} = 0.998$$

$$b_{Y.X} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2} = \frac{184.256}{210.276} = 0.876 \qquad b_{X.Y} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_Y^2} = \frac{184.256}{185.707} = 0.998$$

Anwendungsbeispiel

Variable (n=185)	Mittelwert	(Ko-) Variationen		Stichproben-(ko)varianzen		Korrelationen	
Alter des männl. Partners	35.541	38899.946		210.276		1.000	
Alter der Partnerin	32.757	34087.324	34170.054	184.256	184.703	0.935	1.000

Nach dem Regressionsgewicht wird die Regressionskonstante berechnet:

$$a_{Y.X} = \bar{y} - b_{Y.X} \cdot \bar{x} = 32.757 - 0.876 \cdot 35.541 = 1.62$$

$$a_{X.Y} = \bar{x} - b_{X.Y} \cdot \bar{y} = 35.541 - 0.998 \cdot 32.757 = 2.85$$

Da in die Berechnung der Konstante das Regressionsgewicht eingeht, ist die Rechengenauigkeit der Konstante geringer als beim Gewicht. Im Beispiel wurden daher die Konstanten nur mit zwei Nachkommastellen berechnet.

Aus dem unstandardisierten Regressionsgewicht kann dann das standardisierte Regressionsgewicht, das gleichzeitig die Produktmomentkorrelation ist, berechnet werden:

$$b_{Z_Y.Z_X} = b_{Y.X} \cdot \frac{s_X}{s_Y} = 0.876 \cdot \frac{\sqrt{210.276}}{\sqrt{184.703}} = 0.93 \; ; \; b_{Z_Y.Z_X} = b_{Y.X} \cdot \frac{s_X}{s_Y} = 0.998 \cdot \frac{\sqrt{184.703}}{\sqrt{210.276}} = 0.94$$

Auch hier ist die Rechengenauigkeit geringer.
Anstelle der Stichprobenstandardabweichung können auch die geschätzten Populationsstandardabweichungen oder die Wurzeln aus den Variationen verwendet werden.

Anwendungsbeispiel

Variable (n=185)	Mittelwert	Standard-abweichung	Korrelationen	
Alter des männl. Partners	35.541	14.501	1.000	
Alter der Partnerin	32.757	13.591	0.935	1.000

- Wenn als Ausgangsdaten Korrelationen vorliegen, liegt das standardisierte Regressionsgewicht als Korrelation vor, aus dem dann die unstandardisierten Regressionskoeffizienten berechnet werden:

$$b_{Y.X} = b_{Z_Y.Z_x} \cdot \frac{s_Y}{s_X} = 0.935 \cdot \frac{13.591}{14.501} = 0.876 \; ; \; a_{Y.X} = \bar{x} - b_{Y.X} \cdot \bar{y} = 32.757 - 0.876 \cdot 35.541 = 1.62$$

$$b_{X.Y} = b_{Z_X.Z_Y} \cdot \frac{s_X}{s_Y} = 0.935 \cdot \frac{14.501}{13.591} = 0.998 \; ; \; a_{X.Y} = \bar{y} - b_{X.Y} \cdot \bar{x} = 35.541 - 0.998 \cdot 32.757 = 2.85$$

- Nach den Regressionsoeffizienten können entweder zuerst der Determinationskoeffizient und daraus Vorhersage- und Residualvarianz oder erst die Vorhersage- und Residualvarianz und daraus der Determinationskoeffizient berechnet werden.

$$R^2 = r_{YX}^2 = 0.935^2 = 0.874 \quad = \quad b_{Y.X} \cdot b_{X.Y} = 0.876 \cdot 0.998 = 0.874$$

$$\Rightarrow SS_{\hat{Y}} = R^2 \cdot SS_Y = 0.874 \cdot 34087.324 = 29800.775 \; ; \; s_{\hat{Y}}^2 = 0.874 \cdot 184.703 = 161.476$$

$$\Rightarrow SS_{\hat{X}} = R^2 \cdot SS_X = 0.874 \cdot 38899.946 = 34008.200 \; ; \; s_{\hat{X}}^2 = 0.874 \cdot 210.276 = 183.833$$

$$SS_{E_Y} = \left(1 - R^2\right) \cdot SS_Y = 0.126 \cdot 34087.324 = 4286.549 \; = \; SS_Y - SS_{\hat{Y}} = 34087.324 - 29800.775$$

$$SS_{E_X} = \left(1 - R^2\right) \cdot SS_X = 0.126 \cdot 38899.946 = 4891.746 \; = \; SS_X - SS_{\hat{X}} = 38899.946 - 34008.200$$

Anwendungsbeispiel

Variable (n=185)	Mittelwert	(Ko-) Variationen		Stichproben-(ko)varianzen		Korrelationen	
Alter des männl. Partners	35.541	38899.946		210.276		1.000	
Alter der Partnerin	32.757	34087.324	34170.054	184.256	184.703	0.935	1.000

$$SS_{\hat{Y}} = b_{Y.X}^2 \cdot SS_X = 0.876^2 \cdot 38899.946 = 29850.885 \;=\; b_{Y.X} \cdot SP_{YX} = 0.876 \cdot 34087.324$$

$$SS_{\hat{X}} = b_{X.Y}^2 \cdot SS_Y = 0.998^2 \cdot 34170.054 = 34033.510 \;=\; b_{X.Y} \cdot SP_{YX} = 0.998 \cdot 34087.324$$

$$R^2 = \frac{SS_{\hat{Y}}}{SS_Y} = \frac{29850.885}{34170.054} = 0.87 \quad = \quad \frac{SS_{\hat{X}}}{SS_X} = \frac{34033.510}{38899.946} = 0.87$$

Auf der Basis der Berechnungen erfolgt die Interpretation:

Das positive Regressionsgewicht weist auf eine positive Beziehung hin: je höher das Alter des männlichen Partners, desto höher ist im Durchschnitt auch das Alter der Partnerin.

Das standardisierte Regressionsgewicht weist mit einem Wert von 0.935 einen sehr hohen Wert auf. Auch die erklärte Varianz ist mit einem Wert von 0.876 oder 87.6% außergewöhnlich hoch. Es besteht also eine sehr enge Beziehung zwischen den beiden Variablen.

Bei einer Regression des Alters des männlichen Partners auf das Alter der Partnerin ergibt sich im wesentlichen die gleiche Interpretation, da die standardisierten Koeffizienten identisch sind.

Exkurs: Linearkombinationen von Variablen

Aus Statistik I ist bekannt, dass sich die Mittelwerte und Standardabweichungen bzw. Varianzen von Linearkombinationen (linearen Funktion) von statistisch unabhängigen Variablen aus den Gewichten in den linearen Funktionen sowie den Mittelwerten und Varianzen der Ausgangsvariablen berechnen lassen.
Berücksichtigt man zusätzlich die Kovarianzen zwischen den Ausgangsvariablen, lassen sich ganz unabhängig davon, ob die Ausgangsvariablen statistisch unabhängig voneinander sind oder nicht, Mittelwerte und Varianzen von Linearkombinationen berechnen.

Wenn Y eine Linearkombination aus K Variablen X_1, X_2, …, X_K mit der Konstante b_0 und den Gewichten b_1, b_2, …, b_K ist:

$$Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + ... + b_K \cdot X_K = b_0 + \sum_{k=1}^{K} b_k \cdot X_k \;,$$

dann berechnen sich Mittelwert und Varianz von Y nach:

$$\bar{y} = b_0 + b_1 \cdot \bar{x}_1 + b_2 \cdot \bar{x}_2 + ... + b_K \cdot \bar{x}_K = b_0 + \sum_{k=1}^{K} b_k \cdot \bar{x}_k \;,$$

$$s_X^2 = b_1^2 \cdot s_1^2 + b_2^2 \cdot s_2^2 + ... + b_K^2 \cdot s_K^2 + 2 \cdot b_1 \cdot b_2 \cdot s_{12} + ... + 2 \cdot b_1 \cdot b_K \cdot s_{1K} + ... + 2 \cdot b_{K-1} \cdot b_K \cdot s_{K-1,K}$$

$$= \sum_{k=1}^{K} b_k^2 \cdot s_k^2 + 2 \cdot \sum_{k=1}^{K-1} \sum_{r=k+1}^{K} b_k \cdot b_r \cdot s_{kr} = \sum_{k=1}^{K} \sum_{r=1}^{K} b_k \cdot b_r \cdot s_{kr} \;,$$

wobei s^2_k die Varianz von X_k und s_{kr} die Kovarianz zwischen X_k und X_r bezeichnet.

Linearkombinationen von Variablen

$$Y = b_0 + \sum_{k=1}^{K} b_k \cdot X_k$$

$$\Rightarrow \bar{y} = b_0 + \sum_{k=1}^{K} b_k \cdot \bar{x}_k \; ; \; s_Y^2 = \sum_{k=1}^{K}\sum_{r=1}^{K} b_k \cdot b_r \cdot s_k \cdot s_r = \sum_{k=1}^{K} b_k^2 \cdot s_k^2 + 2 \cdot \sum_{k=1}^{K-1}\sum_{r=k+1}^{K} b_k \cdot b_r \cdot s_k \cdot s_r$$

Die gleiche Beziehung gilt analog für die Erwartungswerte und Varianzen von Zufallsvariablen.

Anwendungsbeispiel:
X ist Variable mit Mittelwert 1.2 und Varianz 4, W eine zweite Variable mit Mittelwert 3 und Varianz 9. Die Kovarianz von X und W beträgt 4. Wenn Y die Summe von X und W ist, folgt für Mittelwert und Varianz von Y:

$$Y = 0 + 1 \cdot X + 1 \cdot W$$

$$\Rightarrow \bar{y} = 0 + 1 \cdot \bar{x} + 1 \cdot \bar{w} = 1.2 + 2 = 4.3 \; ; \; s_y^2 = 1^2 \cdot s_X^2 + 1^2 \cdot s_W^2 + 2 \cdot 1 \cdot 1 \cdot s_{XW} = 4 + 9 + 2 \cdot 4 = 21$$

Über die Ausgangsvariablen lässt sich auch die Kovarianz zwischen zwei Linearkombinationen Y und Z berechnen:

$$Y = b_0 + \sum_{k=1}^{K} b_k \cdot X_k \text{ und } Z = c_0 + \sum_{j=1}^{J} c_j \cdot W_k \Rightarrow s_{YZ} = \sum_{k=1}^{K}\sum_{j=1}^{J} b_k \cdot b_j \cdot s_{kj}$$

Wenn die Ausgangsvariablen in Y mit allen Ausgangsvariablen in Z unkorreliert sind , $s_{kj} = 0$ für alle k und j, dann sind auch Y und Z unkorreliert.

Linearkombinationen von Variablen

Da in der linearen Regression die abhängige Variable Y eine Linearkombination der erklärenden Variable X und der Residualvariable E ist, gelten die Regeln für Linearkombinationen auch hier.
Wenn nun als Annahme für das Regressionsmodell gefordert wird, dass die erklärende Variable X und die Residualvariable X unkorreliert und der Mittelwert der Residuen Null sind, folgt für Mittelwert und Varianz von Y und die Kovarianzen zwischen Y und X und zwischen Y und der Residualvariable E:

$$Y = a + b \cdot X + E$$

$$\Rightarrow \bar{y} = a + b \cdot \bar{x} + \bar{w} = a + b \cdot \bar{x} + 0 \Rightarrow a = \bar{y} - b \cdot \bar{x}$$

$$\Rightarrow s_Y^2 = b^2 \cdot s_X^2 + s_E^2 + 2 \cdot b \cdot s_{XE} = b^2 \cdot s_X^2 + s_E^2 + 0 = s_{\hat{Y}}^2 + s_E^2$$

$$\Rightarrow s_{YX} = s\left(a + b \cdot X + E, X\right) = b \cdot s_{XX} + s_{EX} = b \cdot s_X^2 + 0 \Rightarrow b = \frac{s_{XY}}{s_X^2}$$

$$\Rightarrow s_{YE} = s\left(a + b \cdot X + E, e\right) = b \cdot s_{XE} + s_{EE} = b \cdot 0 + s_E^2$$

Aus der Unkorreliertheit von Residuen und erklärender Variable folgt also sowohl die Varianzzerlegung der Varianz der abhängigen Variable in die Summe aus erklärter Varianz und Residualvarinaz und die Gleichung für die Schätzung des Regressionsgewichts.
Aus der zusätzlichen Annahme, dass der Mittelwert der Residuen Null ist, folgt zudem die Schätzgleichung für die Regressionskonstante.

Lerneinheit 5:
Schätzen und Testen im bivariaten Regressionsmodell

In der Regel sollen mit Hilfe der linearen Regression nicht nur Zusammenhänge zwischen Variablen in einer Stichprobe untersucht werden, sondern Zusammenhänge in der Population, aus der die Stichprobe kommt.
Dazu sind in einem Induktionsschluss Stichprobenergebnisse auf die Grundgesamtheit zu verallgemeinern, was ein unvermeidbares Fehlerrisiko beinhaltet.

Wenn die Regressionskurve in der Population tatsächlich eine ***lineare Funktion*** ist, dann sind die bedingten Populationsmittelwerte von Y eine lineare Funktion der Ausprägungen von X:

$$\mu(Y|X) = \alpha + \beta \cdot X$$

Die abhängige Variable kann dann als Summe bzw. Linearkombination aus den bedingten Mittelwerten bzw. der erklärenden Variable und einer unbeobachteten Residualvariable U aufgefasst werden, in der die Populationsresiduen zusammengefasst werden:

$$Y = \mu(Y|X) + U = \alpha + \beta \cdot X + U$$

Für jeden beliebigen einzelnen Fall i der Population gilt entsprechend:

$$y_i = \alpha + \beta \cdot x_i + u_i$$

Es stellt sich dann die Frage, ob überhaupt und wenn ja, unter welchen Voraussetzungen die nach der Kleinstquadrat-Methode berechneten Regressionskoeffizienten a und b geeignete Schätzer der Regressionskoeffizienten α und β in der Population sind.

Schätzen und Testen im bivariaten Regressionsmodell: Voraussetzungen

(1) ***Linearitätsannahme***
Die zentrale Annahme des linearen Regressionsmodell ist die Linearitätsannahme, nach der die Modellgleichung

$$\mu(Y|X) = \alpha + \beta \cdot X$$

tatsächlich für alle Ausprägungen der erklärenden Variable in der Population gilt.

Alternativ wird vorausgesetzt, dass die abhängige Variable als eine lineare Funktion der erklärenden Variable und der Residualvariable dargestellt werden kann:

$$Y = \alpha + \beta \cdot X + U.$$

Formal kann allerdings jede Variable als lineare Funktion einer anderen Variable aufgefasst werden, wenn die Residuen als Differenzen zwischen abhängiger Variable und Vorhersagewerten definiert werden.

Die Linearitätsannahme wird daher auch so interpretiert, dass die (absoluten) Populationsresiduen $u_i = y_i - \alpha - \beta \cdot x_i$ im Mittel kleiner oder gleich den Residuen sind, die sich bei jeder anderen funktionalen Beziehungsform von Y auf X ergeben würden.
Wenn z.B. eine nichtlineare Beziehung der Form $Y = \alpha \cdot \log(\beta \cdot X) + U$ für alle Fälle der Population berechnet werden könnte, dürften die (quadrierten) Residuen dieses multiplikativen Modells im Mittel nicht kleiner sein als die der linearen Gleichung, da anderenfalls das multiplikative und nicht das lineare Modell vorzuziehen wäre.

Schätzen und Testen im bivariaten Regressionsmodell: Voraussetzungen

(2) ***Unkorreliertheitsannahme***

Wenn die Linearitätsannahme für alle bedingten Mittelwerte von Y bei beliebigen Werten von X zutrifft, dann folgt daraus notwendig, dass die erklärende Variable X und die Residualvariable U statistisch unabhängig voneinander sind. Dies liegt daran, dass ein bedingter Mittelwert ein konstanter Wert ist und daher unabhängig von der Variation um diesen Wert sein muss. Diese Eigenschaft wird auch als ***lokale stochastische Unabhängigkeit*** bezeichnet.

Für die ***konsistente und unverzerrte Schätzung*** der Parameter α und β durch die Koeffizienten a und b der OLS-Schätzung ist bereits die Annahme hinreichend, dass die latente Residualvariable U und die beobachtbare erklärende Variable X unkorreliert sind:

$$\rho(X,U) = \rho_{XU} = 0$$

In der Gleichung steht der kleine griechische Buchstabe rho (ρ) als Abkürzung für die Produkt-moment-Korrelation in der Population.

Es gilt dann also:

$$\mu(a) = \mu_a = \alpha \text{ und } \mu(b) = \mu_b = \beta$$

$$\lim_{n\to\infty} P\left(\left|a-\alpha\right| > \varepsilon\right) = 0 \text{ und } \lim_{n\to\infty} P\left(\left|b-\beta\right| > \varepsilon\right) = 0 \text{ für beliebige } \varepsilon>0$$

Schätzen und Testen im bivariaten Regressionsmodell: Voraussetzungen

(3) ***Homoskedastizitätsannahme***

Bei der OLS-Schätzung der Regressionskoeffizienten wird in der Regel angenommen, dass die Streuungen um die Vorhersagewerte in der Population gleich sind. Formal bedeutet dies, dass bei allen Ausprägungen von X die Residualvarianzen in der Population gleich groß sind:

$$\sigma^2\left(U\middle|X=x\right) = \sigma^2_{U|X=x} = \sigma^2_U \text{ für alle Ausprägungen x von X}$$

Die Gleichheit der bedingten Varianzen wird ***Homoskedastizität*** genannt, die Annahme gleicher Varianzen entsprechend Homoskedastizitätsannahme. Sind dagegen die Residual verschieden, spricht von ***Heteroskedastizität*** bzw. heteroskedastischen Residuen.

(4) ***Keine Autokorrelationen***

In einer einfachen Zufallsauswahl werden die Untersuchungseinheiten unabhängig voneinander ausgewählt.
Ergeben sich die Fälle der Stichprobe auf andere Weise, ist dagegen nicht von vornherein ausgeschlossen, dass es Abhängigkeiten zwischen den Fällen gibt. Bei der Anwendung der OLS-Methode wird üblicherweise angenommen, dass die Residuen untereinander unkorreliert sind:

$$\rho\left(u_i, u_j\right) = \rho_{u_i u_j} = 0 \text{ für alle } i \neq j$$

Wenn die Residuen untereinander korreliert sind, spricht man von ***Autokorrelation***. Es wird also angenommen, dass es keine Autokorrelation zwischen den Residuen gibt.

Schätzen und Testen im bivariaten Regressionsmodell: Voraussetzungen

(5) ***Normalverteilte Residuen***

Für die Verwendung von statistischen Tests wird schließlich oft zusätzlich angenommen, dass die Residualvariable U in der Population normalverteilt ist.

Es wurde bereits erwähnt, dass die OLS-Schätzer erwartungstreu und konsistent sind, sowie die Linearitätsannahme und die Unkorreliertheitsannahme erfüllt sind.
Wenn zusätzlich die Homoskedastizitätsannahme zutrifft und keine Autokorrelation vorliegt, dann ist die OLS-Methode auch ***effizient*** in dem Sinne, dass es keine anderen erwartungstreuen linearen Schätzer gibt, die eine geringere Varianz aufweisen.
Nach dem englischen Ausdruck „***best linear unbiased***“ wird davon gesprochen, dass die OLS-Methode die ***BLU-Eigenschaft*** aufweist, bzw. ein BLU-Schätzer ist.
Das „linear“ in „best linear unbiased“ bezieht sich darauf, dass die Schätzfunktionen lineare Funktionen der Realisationen der abhängigen Variable sind. So gilt für das Gewicht b:

$$b = \frac{SP_{YX}}{SS_X} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})\cdot(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} y_i \cdot (x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - \frac{\overbrace{\bar{y}\cdot\sum_{i=1}^{n}(x_i - \bar{x})}^{=0}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \sum_{i=1}^{n} \underbrace{\frac{(x_i - \bar{x})}{\sum_{j=1}^{n}(x_i - \bar{x})^2}}_{c_i} \cdot y_i = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum_{j=1}^{n}(x_i - \bar{x})^2} \cdot (\alpha + \beta \cdot x_i + u_i) = \beta + \sum_{i=1}^{n} \underbrace{\frac{(x_i - \bar{x})}{\sum_{j=1}^{n}(x_i - \bar{x})^2}}_{c_i} \cdot u_i$$

Fixed-X-Annahme

Bei der Darstellung des Schätzers von β als lineare Funktion der abhängigen Variable bzw. der Populationsresiduen wird der Quotient

$$c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

als eine Stichprobenkonstante betrachtet, die keine Funktion von Zufallsvariablen ist. Dies trifft nur dann zu, wenn die Werte x_i der erklärende Variable X als vorgegebene feste Größen bezeichnet werden. Diese sog. ***Fixed-X-Annahme*** ist z.B. in Experimenten erfüllt, bei denen der Versuchsleiter die Ausprägungen der erklärenden Variable, die Treatments, vorgibt.

In Umfragedaten ist es dagegen in der Regel sinnvoller, auch die erklärende Variable X als Zufallsvariable zu interpretieren, deren Realisierungen durch das Zufallsexperiment „einfache Zufallsauswahl“ festgelegt werden.
Die BLU-Eigenschaft gilt auch dann, wenn die vier Annahmen Linearität, Unkorreliertheit von erklärender Variable und Residualvariable, Homoskedastizität und keine Autokorrelation unter den Residuen ***konditional*** für alle Ausprägungen der erklärenden Variable ***X*** erfüllt sind, obwohl die erklärende Variable selbst eine Zufallsvariable ist. Dies ist dann der Fall, wenn die Regressionsfunktion in der Population tatsächlich linear ist und die Residualvarianzen homoskedastisch sind.

Kausalinterpretation des Regressionsmodells

Sehr oft (wenn auch nicht immer) wird die mittels linearer Regression berechnete Beziehung zwischen einer abhängigen Variable Y und einer erklärenden Variable X so ***interpretiert***, dass die erklärende Variable X als kausale Ursache von Y gesehen wird.
Trifft diese ***Kausalinterpretation*** der linearen Regression zu, dann bewirkt die Erhöhung des Wertes der erklärenden Variable X bei einem beliebigen Fall um +1 Einheit bei diesem Fall im Durchschnitt eine Veränderung der abhängigen Variable um β Einheiten.
Zwar ist β ein unbekannter Populationsparameter, doch wird er bei Erfüllung der ersten beiden Modellannahmen konsistent und erwartungstreu aus den Stichprobendaten geschätzt, so dass davon ausgegangen wird, dass die Erhöhung des Wertes der erklärenden Variable X bei einem beliebigen Fall um +1 Einheit zu einer durchschnittlichen Veränderung der abhängigen Variable um etwa b Einheiten führt.

Bei der Kausalinterpretation wird die Residualvariable U ebenfalls kausal interpretiert: U ist dann die Zusammenfassung aller anderen Einflussgrößen, die auch auf die abhängige Variable Y wirken. Nur wenn diese Einflussgrößen und damit U nicht mit der erklärenden Variable X korrelieren, wird der kausale Zusammenhang unverzerrt geschätzt.
Wenn X um plus eine Einheit ansteigt, ist es aber auch dann nicht ausgeschlossen, dass sich zufällig auch die Werte anderer Einflussfaktoren und damit der Wert der Residualvariable U ändert. Nur wenn U bei einer Änderung von X unverändert ist (konstant bleibt), dann steigt auch der Wert von Y um genau β an.

Kennwerteverteilung der Regressionskoeffizienten

Bei Gültigkeit der vier Bedingungen Linearitätsannahme, Unkorreliertheit von erklärender Variable und Residualvariable, Homoskedastizität und keine Autokorrelation folgt nicht nur die BLU-Eigenschaft, sondern in Anwendung einer Verallgemeinerung des zentralen Grenzwertsatzes auch, dass bei einfachen Zufallsauswahlen die Quotienten aus den Differenzen der geschätzten Regressionskoeffizienten und den Populationsparametern geteilt durch die geschätzten Standardfehler ***asymptotisch standardnormalverteilt*** sind:

$$\lim_{n\to\infty} f\left(\frac{a-\alpha}{\hat{\sigma}_a}\right) = N(0,1) = \varphi \text{ und } \lim_{n\to\infty} f\left(\frac{b-\beta}{\hat{\sigma}_b}\right) = N(0,1) = \varphi$$

Wenn zusätzlich gilt, dass die Residualvariable U in der Population normalverteilt ist (Annahme 5), dann sind diese Quotienten unabhängig vom Stichprobenumfang mit df = n–2 Freiheitsgraden exakt ***t-verteilt***:

$$f\left(\frac{a-\alpha}{\hat{\sigma}_a}\right) = t_{df=n-2} \text{ und } f\left(\frac{b-\beta}{\hat{\sigma}_b}\right) = t_{df=n-2}$$

Um die Normalverteilung bzw. die T-Verteilung für die Berechnung von Konfidenzintervallen und die Durchführung statistischer Tests anwenden zu können, müssen die Standardfehler der Regressionskoeffizienten bekannt sein bzw. konsistent geschätzt werden können.

Standardfehler der Regressionskoeffizienten

Wenn die vier Bedingungen der BLU-Eigenschaft erfüllt sind, dann gilt für die Standardschätzfehler der Regressionskoeffizienten unter fixed-X-Annahme bzw. konditional für die beobachteten Realisierungen von X:

$$\sigma(a_{Y.X}) = \sqrt{\frac{\frac{1}{n}\cdot\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sigma_U^2} = \sigma_U\cdot\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{SS_X}} = \frac{\sigma_U}{\sqrt{n}}\cdot\sqrt{1+\frac{\bar{x}^2}{s_X^2}}$$

$$\sigma(b_{Y.X}) = \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sigma_U^2} = \sigma_U\cdot\sqrt{\frac{1}{SS_X}} = \frac{\sigma_U}{\sqrt{n}}\cdot\frac{1}{s_X}$$

Die Formeln gelten für die Regression der abhängigen Variable Y auf die erklärende Variable X. Wird die Regression von X auf Y berechnet, müssen die Gleichungen entsprechend angepasst werden:

$$\sigma(a_{X.Y}) = \frac{\sigma_U}{\sqrt{n}}\cdot\sqrt{1+\frac{\bar{y}^2}{s_Y^2}}\ ;\ \sigma(b_{X.Y}) = \frac{\sigma_U}{\sqrt{n}}\cdot\frac{1}{s_Y}$$

Im folgenden wird davon ausgegangen, dass Y auf X regrediert wird. Bei der umgekehrten Richtung sind alle Formeln entsprechend anzupassen.

Standardfehler der Regressionskoeffizienten

Um die Standardfehler zu schätzen, muss zunächst die Varianz der Populationsresiduen geschätzt werden. Ein konsistenter und erwartungstreuer Schätzer ist:

$$\hat{\sigma}_U^2 = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SS_E}{n-2} = \frac{n}{n-2}\cdot s_E^2$$

Der Quotient n–2 bei der Schätzung der Residualvarianz und der Zahl der Freiheitsgrade bei Anwendung der T-Verteilung ergibt sich daher, dass im Regressionsmodell die *beiden* Regressionskoeffizienten α und β durch a und b geschätzt werden und in die Berechnung der Residuen einfließen.

Die geschätzten Standardfehler der Regressionskoeffizienten sind dann:

$$\hat{\sigma}_a = \sqrt{\frac{\frac{1}{n}\cdot\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\frac{\sum_{i=1}^{n} e_i^2}{n-2}} = \frac{\hat{\sigma}_U}{\sqrt{n}}\cdot\sqrt{1+\frac{\bar{x}^2}{s_X^2}} = \frac{s_E}{\sqrt{n-2}}\cdot\sqrt{1+\frac{\bar{x}^2}{s_X^2}}$$

$$\hat{\sigma}_b = \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\frac{\sum_{i=1}^{n} e_i^2}{n-2}} = \frac{\hat{\sigma}_U}{\sqrt{n}}\cdot\sqrt{\frac{1}{s_X^2}} = \frac{s_E}{\sqrt{n-2}}\cdot\frac{1}{s_X}$$

Standardfehler der Regressionskoeffizienten

Die Schätzung der Regressionskonstante und des Regressionsgewichts sind nicht unabhängig voneinander, da beide Schätzungen auf den gleichen Daten beruhen. Tatsächlich geht in die Schätzgleichung der Regressionskonstante α bereits der Schätzer von β ein:

$$\hat{\alpha} = a = \bar{y} - \hat{\beta} \cdot \bar{x} = \bar{y} - b \cdot \bar{x}$$

Da gezeigt werden kann, dass die Schätzer des Mittelwerts der abhängigen Variable und des Regressionsgewichts statistisch unabhängig voneinander sind, und der Standardfehler des Mittelwerts der abhängigen Variable die geschätzte Residualvarianz geteilt durch die Fallzahl ist, kann die Varianz der Kennwerteverteilung von a nach den generellen Regeln für Varianzen von Linearkombinationen auch als Funktion der Varianz der Kennwerteverteilung von b dargestellt werden, wenn entsprechend der Fixed-X-Annahme die erklärende Variable und damit auch ihr Stichprobenmittelwert nicht als Zufallsvariable, sondern als Konstante aufgefasst wird::

$$\sigma_a^2 = \sigma^2\left(\bar{y} - \bar{x} \cdot b\right) = \sigma_{\bar{y}}^2 + \bar{x}^2 \cdot \sigma_b^2 = \frac{\sigma_U^2}{n} + \bar{x}^2 \cdot \frac{\sigma_U^2}{n \cdot s_X^2} = \left(1 + \frac{\bar{x}^2}{s_X^2}\right) \cdot \frac{\sigma_U^2}{n}$$

Für die Kovarianz der Schätzungen von a und b folgt dann:

$$\sigma_{ab} = \sigma\left(\bar{y} - b \cdot \bar{x}, b\right) = -\bar{x} \cdot \sigma_b^2 = \frac{-\bar{x}}{SS_X} \cdot \sigma_U^2$$

Für den geschätzte Standardfehler wird wiederum σ^2_U durch seinen Schätzer ersetzt.
Aus der negativen Kovarianz folgt, dass bei einer Überschätzung der Regressionskonstante in einer Stichprobe das Regressionsgewicht tendenziell unterschätzt und bei einer Unterschätzung tendenziell überschätzt wird.

Konfidenzintervalle der Regressionskoeffizienten

Da die Kennwerteverteilung der Residuen bekannt ist, können mit Hilfe der geschätzten Standardfehler Konfidenzintervalle für die Regressionskoeffizienten berechnet werden.
Bei hinreichend großen Stichproben ($n \geq 30$) kann die Standardnormalverteilung für die Berechnung asymptotisch gültiger Konfidenzintervalle herangezogen werden.
Sind die Residuen in der Population normalverteilt, dann ergibt die T-Verteilung mit df = n–2 Freiheitsgraden exakte und nicht nur asymptotisch gültige Konfidenzintervalle. Bei einer Irrtumswahrscheinlichkeit α berechnen sich die Intervalle nach:

$$c.i.(\alpha) = a \pm \hat{\sigma}_a \cdot t_{n-2;1-\alpha/2}$$

$$c.i.(\beta) = b \pm \hat{\sigma}_b \cdot t_{n-2;1-\alpha/2}$$

Da die Verwendung der T-Verteilung zu größeren Intervallen führt als die Verwendung der asymptotisch gültigen Standardnormalverteilung, wird die T-Verteilung im Sinne eines vorsichtigen oder konservativen Schätzens meistens auch dann angewendet, wenn die Normalverteilungsannahme für die Populationsresiduen nicht gegeben ist.

Am Beispiel der Daten aus dem Allbus 2006 soll für die Regression des Alters der Partnerinnen auf das Alter der Partner bei unverheirateten Lebensgemeinschaften das 95%-Konfidenzintervalle des Regressionsgewichts b berechnet werden.
Die notwendigen Statistiken sind bereits In Lerneinheit 4 berechnet. Um eine größere Rechengenauigkeit zu erreichen, werden im folgenden die mit dem Statistikprogramm SPSS berechneten Variationen verwendet.

Konfidenzintervalle der Regressionskoeffizienten: Anwendungsbeispiel für b

$$n = 185; \sum_{i=1}^{n} x_i = 6575; \sum_{i=1}^{n} x_i^2 = 272579; \sum_{i=1}^{n} y_i = 6060; \sum_{i=1}^{n} y_i^2 = 232676; \sum_{i=1}^{n} x_i \cdot y_i = 249463;$$

$$a = 1.614 \; ; \; b = 0.876 \; ; \; SS_E = 4301.696 \; ; \; s_E^2 = 23.252$$

Der geschätzte Standardfehler des Regressionsgewichts berechnet sich nach:

$$\hat{\sigma}_b = \sqrt{\frac{1}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)}{n}} \cdot \frac{\sum_{i=1}^{n} e_i^2}{n-2}} = \sqrt{\frac{1}{272579 - \frac{6575^2}{185}} \cdot \frac{4301.696}{185-2}} = 0.025$$

Bei einer Irrtumswahrscheinlichkeit von 5% (=100–95 %) wird das 97.5%-Quantil der T-Verteilung mit df=185–2=183 Freiheitsgraden benötigt, das zwischen 1.98 (df=120) und 1.96 (df=∞) liegt. Im Sinne des vorsichtigen Vorgehens wird der größere Wert verwendet.
Die Grenzen des 95%-Konfidenzintervalls des Regressionsgewichts b betragen dann:

$$\text{c.i.}(\beta) = b \pm \hat{\sigma}_b \cdot t_{n-2;1-\alpha/2} = 0.876 \pm 0.025 \cdot 1.98 = 0.827 \text{ bis } 0.926$$

Mit einer Wahrscheinlichkeit von 95% gehört das Intervall mit den Grenzen 0.827 und 0.926 zu den Intervallen, die den Wert des Regressionsgewichts in der Population enthalten.

Statistische Tests der Regressionskoeffizienten

Tests der Regressionskoeffizienten erfolgen ganz analog zu Tests von Populationsmittelwerten.

Schritt 1: *Formulierung von Null- und Alternativhypothese*
Bei den Regressionsgewichten können folgende Hypothesenpaare geprüft werden:
(1) $H_0: \beta = \beta_0$ versus $H_1: \beta \neq \beta_0$
(2) $H_0: \beta \leq \beta_0$ versus $H_1: \beta > \beta_0$
(3) $H_0: \beta \geq \beta_0$ versus $H_1: \beta < \beta_0$

Schritt 2: *Auswahl der statistischen Prüfgröße und Testverteilung*
Als Teststatistik wird die Differenz des geschätzten Regressionsgewichts minus dem postulierten Wert β_0 durch den Standardfehler geteilt:

$$T = \frac{b - \beta_0}{\hat{\sigma}_b}$$

Bei Zutreffen aller fünf Annahmen der OLS-Methode und einem Populationswert von β_0 ist die Teststatistik mit df = n–2 Freiheitsgraden t-verteilt.
Falls die Normalverteilungsannahme der Residuen nicht erfüllt ist, ist die Teststatistik weiterhin anwendbar. Sie ist dann bei gültiger Nullhypothese (an der Stelle $\beta=\beta_0$) asymptotisch standardnoralverteilt.
Wie bei den Konfidenzintervallen wird die T-Verteilung im Sinne eines vorsichtigen Vorgehens oft auch dann angewendet, wenn die Normalverteilungsannahme nicht gegeben ist.

Statistische Tests der Regressionskoeffizienten

Schritt 3: Festlegung von Irrtumswahrscheinlichkeit und kritischen Werten

- Die Irrtumswahrscheinlichkeit α ist die Wahrscheinlichkeit, mit der die Nullhypothese fälschlichersweise abgelehnt wird, wenn $\beta=\beta_0$.
  In der Regel wird sie auf 5% oder 1% festgesetzt.
- Die kritischen Werte trennen den Annahmebereich des Wertebereichs der Teststatistik vom Ablehnungsbereich.
  Wenn die Nullhypothese falsch ist, ist die Teststatistik T nichtzentral t-verteilt bzw. asymptotisch normalverteilt mit einem Erwartungswert ungleich Null.

Die Nullhypothese wird daher abgelehnt, wenn bei der Prüfung von

(1) $H_0: \beta = \beta_0 \quad T < t_{\alpha/2,\, df=n-2}$ oder $T > t_{1-\alpha/2,\, df=n-2}$ bzw.
(2) $H_0: \beta \le \beta_0 \quad T > t_{1-\alpha,\, df=n-2}$, bzw.
(3) $H_0: \beta \ge \beta_0 \quad T < t_{\alpha,\, df=n-2}$.

Anstelle der Quantile der T-Verteilung können bei großen Fallzahlen auch die entsprechenden Quantile der Standardnormalverteilung herangezogen werden.

Schritt 4: Berechnung der Teststatistik und Entscheidung

Auf der Basis der Stichprobendaten wird der Wert der Teststatistik berechnet mit den kritischen Werten verglichen und die Nullhypothese entsprechend abgelehnt oder beibehalten.

Statistische Tests der Regressionskoeffizienten

Schritt 5: Überprüfung der Anwendungsvoraussetzungen

Zur Überprüfung der Anwendungsvoraussetzungen werden oft ergänzende Datenanalysen durchgeführt, die später vorgestellt werden..

Für die asymptotische Normalverteilung gilt als Faustregel, dass die Stichprobe mindestens 30, besser mindestens 50 Fälle umfassen sollte ($n \ge 30$ bzw. 50).

Tests von Regressionskonstanten

Tests der Regressionskonstanten sind ganz analog.

Die Teststatistik ist hier:

$$T = \frac{a - \alpha_0}{\hat{\sigma}_a}$$

Geprüft werden die Hypothesen

(1) $H_0: \alpha = \alpha_0$ versus $H_1: \alpha \ne \alpha_0$
(2) $H_0: \alpha \le \alpha_0$ versus $H_1: \alpha > \alpha_0$
(3) $H_0: \alpha \ge \alpha_0$ versus $H_1: \alpha < \alpha_0$

Abgelehnt werden die Nullhypothesen, wenn bei

(1) $H_0: \alpha = \alpha_0 \quad T < t_{\alpha/2,\, df=n-2}$ oder $T > t_{1-\alpha/2,\, df=n-2}$ bzw.
(2) $H_0: \alpha \le \alpha_0 \quad T > t_{1-\alpha,\, df=n-2}$, bzw.
(3) $H_0: \alpha \ge \alpha_0 \quad T < t_{\alpha,\, df=n-2}$.

Anstelle der T-Verteilung kann wiederum die Standardnormalverteilung verwendet werden.

Anwendungsbeispiel

Theoretisch sollte bei gleichem durchschnittlichen Alter der beiden Partner die Regressionskonstante gleich Null und das Regressionsgewicht gleich Eins sein. Dies soll mit einer Irrtumswahrscheinlichkeit von jeweils 5% geprüft werden.

Schritt 1: Formulierung der Hypothesenpaare
Im Beispiel können die Forschungshypothesen nicht als Alternativhypothesen formuliert werden. Sie werden daher als Nullhypothese formuliert:

$$H_0: \alpha = 0 \text{ vs. } H_1: \alpha \neq 0 \text{ und } H_0: \beta = 1 \text{ vs. } H_1: \beta \neq 1$$

Schritt 2: Auswahl von Teststatistik und Testverteilung
Auch wenn die Populationsresiduen möglicherweise nicht normalverteilt sind, kann der T-Test mit df=n–2=183 Freiheitsgraden herangezogen werden. Da hier die Forschungshypothese Nullhypothese ist, wird jedoch die Standardnormalverteilung herangezogen.

Schritt 3: Irrtumswahrscheinlichkeit und kritische Werte
Bei der vorgegebenen Irrtumswahrscheinlichkeit von 5% werden die 2.5%- und 97.5%-Quantile der Standardnormalverteilung benötigt, also ±1.96.
Wäre die Forschungshypothese Alternativhypothese würden stattdessen die Quantile der T-Verteilung mit df=183 Freiheitsgraden als kritische Werte verwendet.

Statistische Tests der Regressionskoeffizienten: Anwendungsbeispiel

Schritt 4: Berechnung und Entscheidung
Die Teststatistiken ergeben:

$$H_0: \beta = 1 \text{ vs. } H_1: \beta \neq 1 \Rightarrow Z = \frac{0.876 - 1}{0.025} = -4.96$$

$$H_0: \alpha = 0 \text{ vs. } H_1: \alpha \neq 0 \Rightarrow Z = \frac{1.614}{\sqrt{\frac{272579/185}{272579 - 6575^2/185} \cdot \frac{4301.696}{183}}} = 1.711$$

Da der Z-Wert des Regressionsgewichts mit −3.07 kleiner ist als −1.96, wird die Nullhypothese, dass das Gewicht genau Eins ist, abgelehnt.
Beibehalten wird dagegen die Nullhypothese, dass die Regressionskonstante Null ist, da hier der Z-Wert mit 1.711 < 1.99 ist.
Offenbar gilt nicht nur in der Stichprobe, dass die Altersdifferenz der Partner mit dem Alter systematisch variiert.

Schritt 5: Anwendungsvoraussetzungen
Da die Fallzahl mit n=153 Fällen größer 30 ist, kann davon ausgegangen werden, dass die Regressionskoeffizienten asymptotisch normalverteilt sind, falls die Linearitätsannahme und die Homoskedastiztätsannahme erfüllt sind und keine Autokorrelation vorliegt.

Konfidenzintervalle und Test von Vorhersagen

Konfidenzintervalle und Tests können sich nicht nur auf Regressionskoeffizierten, sondern auch auf die bedingten Mittelwerte (Vorhersagewerte) beziehen.
Da $\hat{y}_j = a + b \cdot x_j$, folgt nach den Regeln für Linearkombinationen von zwei Zufallsvariablen (hier: die Schätzer der Regressionskoeffizienten „a“ und „b“) bei gegebenem (als Konstante betrachteten) Wert x_j:

$$\sigma\left(\hat{\mu}_{Y|X=x_j}\right) = \sigma\left(a + x_j \cdot b\right) = \sigma\left(\bar{y} + \left(x_j - \bar{x}\right) \cdot b\right) = \sqrt{\frac{\sigma_U^2}{n} + \frac{\left(x_j - \bar{x}\right)^2}{s_X^2} \cdot \frac{\sigma_U^2}{n}} = \sqrt{1 + \frac{\left(x_j - \bar{x}\right)^2}{s_X^2}} \cdot \frac{\sigma_U}{\sqrt{n}}$$

Wird die Varianz der Populationsresiduen durch seinen Schätzer ersetzt, ergibt sich der geschätzte Standardfehler für die bedingten Mittelwerte:

$$\hat{\sigma}\left(\hat{\mu}_{Y|X=x_j}\right) = \sqrt{\left(1 + \frac{\left(x_j - \bar{x}\right)^2}{s_X^2}\right) \cdot \frac{\hat{\sigma}_U^2}{n}} = \sqrt{\left(\frac{1}{n} + \frac{\left(x_j - \bar{x}\right)^2}{n \cdot s_X^2}\right) \cdot \hat{\sigma}_U^2} = \hat{\sigma}_U \cdot \sqrt{h_j}$$

Der in der Klammer stehende Faktor $\mathbf{h_j}$ wird in der Literatur als ***Hebelwer***t (engl.: *leverage*) bezeichnet.
Die Standardfehler werden um so größer, je weiter eine Ausprägung x_j vom Mittelwert der erklärenden Variablen entfernt ist, da im Zähler des zweiten Summanden in der Wurzel, also im Hebelwert, der quadrierte Abstand zum Mittelwert vorkommt.

Standardfehler eines bedingten Mittelwerts

Wenn x_j gleich dem Mittelwert der erklärenden Variable ist, dann reduziert sich der Vorhersagewert auf den Mittelwert der abhängigen Variablen. Dieser Wert lässt sich relativ genau schätzen. Der Standardfehler des bedingten Mittelwerts ist dann gleich dem Standardfehler der Residualvariable U geteilt durch die Wurzel aus der Fallzahl.
Je weiter x_j vom Mittelwert entfernt ist, desto stärker muss berücksichtig werden, dass das geschätzte Regressionsgewicht möglicherweise vom tatsächlichen Regressionsgewicht abweicht, weil die mögliche Differenz zwischen der tatsächlichen Regressionsgeraden und der geschätzten Regressionsgeraden größer wird, je stärker man sich den Rändern der Verteilung nähert. Der Standardfehler eines Vorhersagewertes wird daher an den Rändern der Verteilung immer größer.

Soll dass $(1-\alpha)$-Konfidenzintervall eines bedingten Mittelwerts berechnet werden, ergibt es sich somit als:

$$\text{c.i.}\left(\mu_{Y|X=x_j}\right) = \hat{y}_j \pm \hat{\sigma}\left(\hat{\mu}_{Y|X=x_j}\right) \cdot t_{df=n-2;1-\alpha/2}$$

Der geschätzte Standardfehler kann auch verwendet werden, um einen bedingten Mittelwert zu testen. Die Vorgehensweise entspricht dem Test eines Regressionskoeffizienten.

Standardfehler eines bedingten individuellen Vorhersagewertes

Von der Schätzung eines bedingten Mittelwertes zu unterscheiden ist die Schätzung eines individuellen Vorhersagewertes der abhängigen Variable.
Bei der Punktschätzung sind bedingter Mittelwert und individuelle Vorhersage gleich, weil die beste Vorhersage der bedingte Mittelwert ist. Bei der Intervallschätzung oder dem Tests eines individuellen Vorhersagewertes ist dagegen zu berücksichtigen, dass die einzelnen Realisationen um die Regressionskurve streuen.

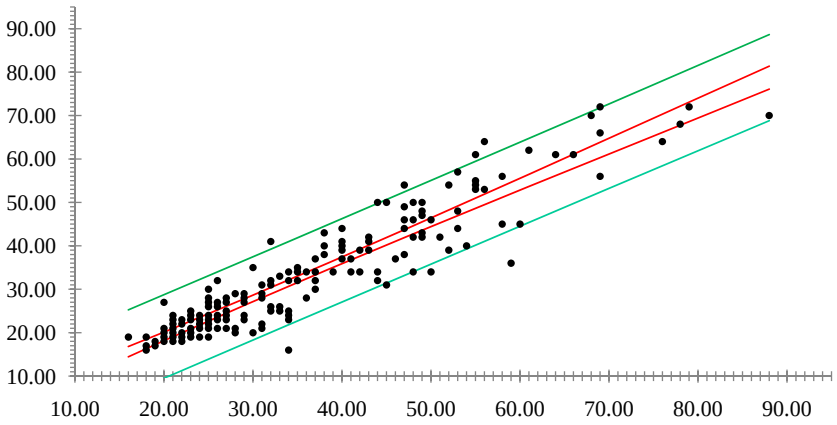
Da bei homoskedastischen Populationsresiduen die Realisationen mit der Varianz σ^2_U um die Regressionsfunktion streuen, ist der quadrierte Standardfehler eines individuellen Vorhersagewertes gleich der Summe der Varianz des bedingten Mittelwertes plus der Residualvarianz. Ein Standardfehler für einen individuellen Vorhersagewert beträgt daher:

$$\hat{\sigma}\left(\hat{Y}_j\right) = \sqrt{\hat{\sigma}^2\left(\hat{\mu}_{Y|X_j}\right) + \hat{\sigma}^2_U} = \sqrt{\left(h_j + 1\right)\cdot\hat{\sigma}^2_U} = \sqrt{\left(\frac{1}{n} + \frac{\left(x_j - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} + 1\right)\cdot\hat{\sigma}^2_U} = \hat{\sigma}_U \cdot \sqrt{1 + h_j}$$

Konfidenzintervall einer individuellen Vorhersage berechnen sich entsprechend nach:

$$c.i.\left(\hat{Y}_j\right) = \hat{y}_j \pm \hat{\sigma}\left(\hat{Y}_j\right)\cdot t_{df=n-2;1-\alpha/2}$$

Konfidentintervall und Test von Vorhersagen



Die Abbildung zeigt die Grenzen der Konfidenzintervalle sowohl der bedingten Mittelwerte wie der individuellen Vorhersagewerte für die 185 Fälle der Allbus-Stichprobe.
Die Grenzen für die einzelnen Realisationen (grüne Kurven) sind deutlich weiter von der Regressionsgerade entfernt als die Grenzen der Intervalle für die Mittelwerte (rote Kurven).
Bei 5% Irrtumswahrscheinlichkeit sollten auch nur etwa 5% der 185 Fälle, also zwischen 9 und 10 Fälle außerhalb der Grenzen der Konfidenzintervalle liegen.
Tatsächlich liegen 10 Fälle auf den Intervallgrenzen oder außerhalb der Intervallgrenzen, was darauf hinweist, dass die Standardfehler recht gut geschätzt werden.

Tests von Hypothesen über Kovarianzen, Korrelationen und Determinationskoeffizienten

Da sich die Fallzahlen bei der Berechnung des Regressionsgewicht herauskürzen, gilt:

$$b = \frac{SP_{XY}}{SS_X} = \frac{\frac{SP_{XY}}{n}}{\frac{SS_X}{n}} = \frac{s_{XY}}{s_X^2} = \frac{s_{XY}}{s_X \cdot s_Y} \cdot \frac{s_Y}{s_X} = r_{XY} \cdot \frac{s_Y}{s_X} = \sqrt{R^2} \cdot \frac{s_Y}{s_X} = \frac{\frac{SP_{XY}}{n-1}}{\frac{SS_X}{n-1}} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2}$$

Wenn das Regressionsgewicht b null ist, ist daher auch die Kovarianz, die Korrelation und ihr Quadrat, der Determinationskoeffizient, null.
Der Test des Regressionsgewichts mit $\beta_0 = 0$ prüft also gleichzeitig auch die Hypothese, dass die Kovarianz, die Korrelation und der Determinationskoeffizient null sind.
Neben dem T-Test bzw. dem asymptotischen Z-Test gibt es im Regressionsmodell alternative Tests mit anderen Testverteilungen, die in späteren Lerneinheiten vorgestellt werden.

Kontrolle der Anwendungsvoraussetzungen

Die Eigenschaften der OLS-Schätzung des linearen Regressionsmodells sind an die oben vorgestellten Anwendungsvoraussetzung bzw. -annahmen gebunden.

Zur Kontrolle der Anwendungsvoraussetzungen gibt unterschiedliche Möglichkeiten. Neben speziellen Tests, die in späteren Lerneinheiten diskutiert werden, werden im Folgenden grafische Inspektionen von Streudiagrammen vorgestellt, um die Linearität der Regressionsfunktion, Homoskedastizität und Normalverteilung der Residuen sowie den Einfluss von Ausreißern in den Daten zu untersuchen.

Kontrolle der Anwendungsvoraussetzungen der Regression

Wenn die Modellannahmen erfüllt sind, sollten die Realisationen der Residualariable U bei allen Ausprägungen der erklärenden Variable X bzw. der Vorhersagewerte

- Erwartungswerte von Null aufweisen (Linearitätsannahme),
- nicht mit den erklärenden Variablen korrelieren (Unkorreliertheit),
- die gleiche Varianz aufweisen (Homoskedastizitätsannahme),
- voneinander unabhängig sein (keine Autokorrelation) und
- möglichst normalverteilt sein (Normalverteilungsannahme).

Da anstelle der Populationskoeffizienten α und β nur deren Schätzungen a und b vorliegen, können anstelle der Realisierungen u_i von U nur die Stichprobenresiduen e_i von E betrachtet werden.

Betrachtet man die Stichprobenresiduen anstelle der Populationsresiduen folgt jedoch aus den Eigenschaften der OLS-Schätzung,

- dass der Mittelwert der Stichprobenresiduen null ist und
- dass die Stichprobenresiduen nicht mit der erklärenden Variablen korrelieren.

Hinzu kommt, dass die Stichprobenresiduen zwangsläufig bei verschiedenen Ausprägungen der erklärenden Variablen unterschiedliche Varianzen aufweisen müssen, selbst wenn die Populationsresiduen homoskedastisch sind. Dies liegt daran, dass die Ausprägungen y_i der abhängigen Variablen Y die Summe der Vorhersagewerte und der Stichprobenresiduen sind:

$$y_i = \hat{y}_i + e_i = \hat{\mu}_{Y|X=x_i} + e_i$$

Kontrolle der Anwendungsvoraussetzungen der Regression

Aufgrund der Unkorreliertheit von Vorhersagewerten und Stichprobenresiduen ist die bedingte Varianz von y_i gegeben x_i die Summe aus der Varianz der Regressionsfunktion an der Stelle x_i und der Varianz des Stichprobenresiduums:

$$\sigma^2\left(y_i \middle| x_i\right) = \sigma^2\left(\hat{\mu}_{Y|X=x_i} + e_i\right) = \sigma^2\left(\hat{\mu}_{Y|X=x_i}\right) + \sigma^2\left(e_i\right) = h_i \cdot \sigma_U^2 + \sigma^2\left(e_i\right)$$

In der Gleichung ist h_i wieder der Hebelwert (leverage) an der Stelle x_i.

Da die bedingte Varianz von Y gegeben $X=x_i$ bei homoskedastischen Residuen aber stets auch gleich der Varianz der Populationsresiduen ist:

$$\sigma^2\left(y_i \middle| x_i\right) = \sigma_U^2,$$

folgt für die Varianz jedes Stichprobenresiduums e_i:

$$\sigma^2(e_i) = \sigma_U^2 - \sigma^2(\mu_{Y|X=x_i}) = \sigma_U^2 - h_i \cdot \sigma_U^2 = \sigma_U^2 \cdot (1 - h_i) = \sigma_U^2 \cdot \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right)$$

Die Varianz eines Stichprobenresiduums e_i ist also um so kleiner, je weiter der zugeordnete Wert x_i vom Mittelwert der erklärenden Variablen entfernt ist.

Kontrolle der Anwendungsvoraussetzungen der Regression

Ersetzt man die Populationsvarianz von U durch den erwartungstreuen Schätzer dieser Varianz und zieht die Wurzel aus der Varianz, ergibt sich der geschätzte Standardfehler eines Residuums e_i:

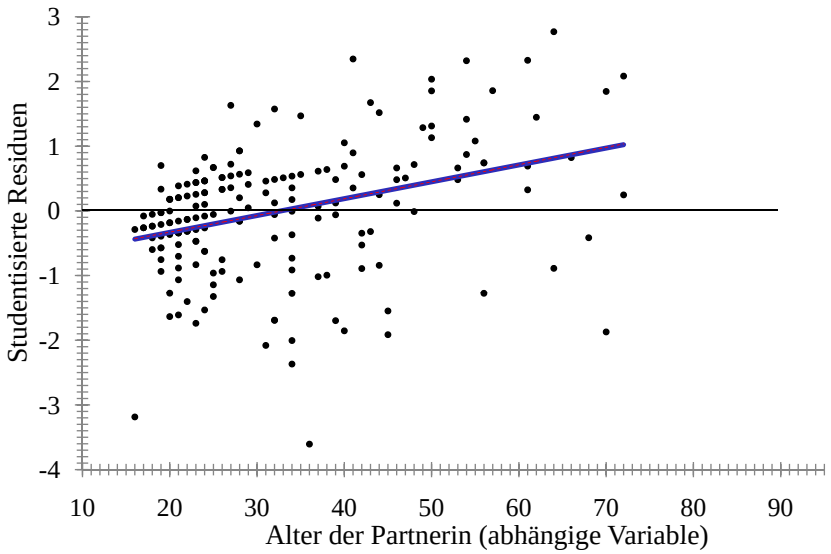
$$\hat{\sigma}(e_i) = \hat{\sigma}_U \cdot \sqrt{1 - h_i} = \hat{\sigma}_U \cdot \sqrt{1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{n \cdot s_X^2}\right)} = \sqrt{\frac{\sum_{j=1}^{n} e_j^2}{n-2} \cdot \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right)}$$

Bei der Kontrolle von Modellannahmen werden daher meist anstelle der Stichprobenresiduen E die ***standardisierten Residuen E^**** betrachtet, die sich ergeben, wenn jedes Residuum e_i durch seinen Standardfehler dividiert wird:

$$e_i^* = \frac{e_i}{\hat{\sigma}(e_i)} = \frac{e_i}{\sqrt{\frac{\sum_{j=1}^{n} e_j^2}{n-2} \cdot \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right)}}$$

Zur Unterscheidung von einer Standardisierung über die Z-Transformation werden die durch ihren Standardfehler geteilten Residuen auch als ***studentisierte Residuen*** bezeichnet.

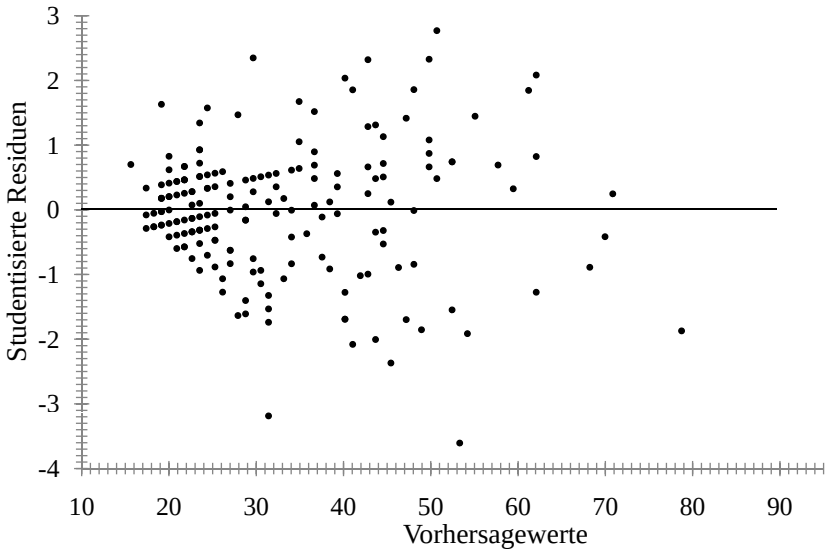
Kontrolle der Linearitätsannahme



Einen ersten Eindruck über die Angemessenheit der Linearitätsannahme gibt ein Streudiagramm der studentisierten Residuen nach den Werten der abhängigen Variable.

Zwischen der abhängiger Variable und standardisierten Residuen besteht notwendigerweise eine positive Korrelation, da die abhängige Variable eine Funktion der unstandardisierten Residuen ist. Auf Nichtlinearität weist ein Streudiagramm hin, bei dem die Punktewolke nicht gleichmäßig ansteigt.
Zur Verdeutlichung sind eine lineare (blau durchgezogen) und eine quadratische (rot gepunktet) Trendlinie in die Grafiken eingefügt. Die Kurven unterscheiden sich praktisch nicht, was auf einen linearen Zusammenhang hinweist.

Kontrolle der Homoskedastizitätsannahme



Das Streudiagramm der studentisierten Residuen gegen die Vorhersagewerte (bzw. die Werte der erklärenden Variablen) sollte bei homoskedastischen Residuen entlang der waagerechten Achse gleichmäßig um den Wert Null streuen.

Dass es im Streudiagramm für die Beispieldaten mehr Fälle bei niedrigen als bei hohen Vorhersagewerten gibt, spricht nicht gegen die Homoskedastizitätsannahme.
Auf der anderen Seite sind große studentisierte Residuen aber vor allem bei mittleren und hohen Vorhersagewerten zu beobachten und kleinere bei niedrigen Vorhersagewerten. Dies könnte ein Hinweis darauf sein, dass bei höherem Alter die Residualvarianzen zunehmen, die individuellen Vorhersagen also ungenauer werden und die Irrtumswahrscheinlichkeiten von Konfidenzintervallen und Tests nicht korrekt sind.

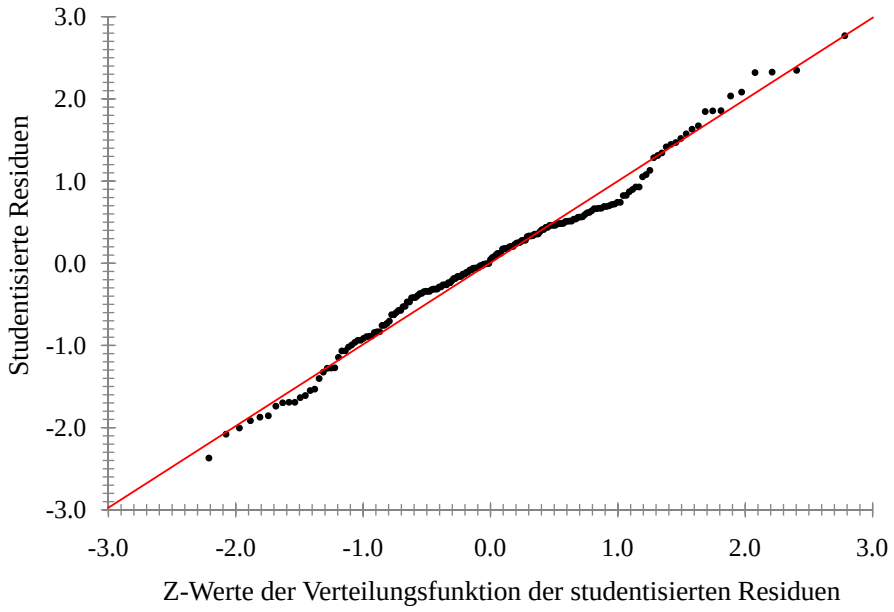
Kontrolle der Normalverteilungsannahme

Obwohl die Normalvereilungsannahme relativ unproblematisch ist, ist es möglich, sie zu überprüfen. Dazu wird oft ein sogenanntes ***Q-Q-Plot*** betrachtet, bei denen die studentisierten Residuen gegen Quantile der Standardnormalverteilung (z-Werte) abgetragen werden, die aus der kumulierten Häufigkeitsverteilung der Residuen berechnet werden.

Die z-Werte berechnen sich nach: $z_i = \Phi^{-1}\left(cp_i - \frac{0.5}{n}\right)$

Im Beispiel der Regression des Alters der Partnerin auf das Alter des Partners ergibt sich für den ersten Fall der der Größe nach sortierten Realisierungen der studentisierten Residuen eine kumulierte relative Häufigkeit von 1/n, bei den Allbus-Daten 1/185. Der z-Wert für dieses Residuum ist dann der Quantilwert der Standardnormalverteilung, der der relativen Häufigkeit von 0.0027 (=1/185 – 0.5/185)entspricht, das ist –2.7822
Der zehntkleinste Wert korrespondiert entsprechend mit dem z-Wert zum relativen Anteil 0.0514 (=10/185 – 0.5/185), was einen z-Wert von –1.6314 entspricht.
Wenn die Residuen normalverteilt sind, sollte das Streudiagramm der studentisierten Residuen gegen die Z-Werte dieser Q-Q-Plot eine Punktewolke zeigen, die relativ eng entlang der 45°-Gerade im Streudiagramm verläuft. Wenn es deutliche Abweichungen gibt, spricht dies gegen die Normalverteilungsannahme

Kontrolle der Normalverteilungsannahme



Die Abbildung zeigt, dass die Punktewolke tatsächlich ziemlich eng entlang der 45°-Achse verläuft, was dafür spricht, dass die Residuen annähernd normalverteilt sind.

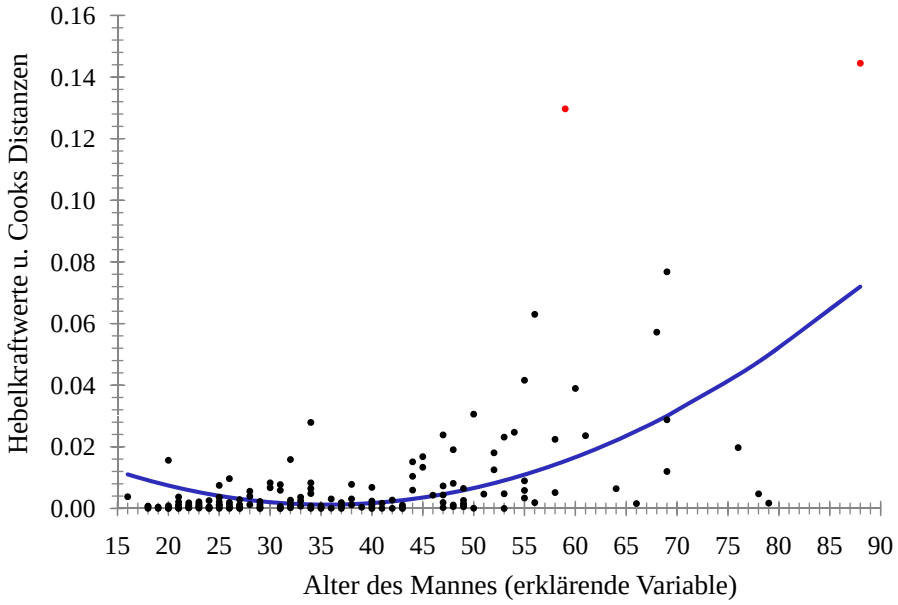
Ausreißer und einflussreiche Fälle

Eine implizite Annahme aller statistischer Analysen ist, dass die Population, aus der die Fälle kommen, homogen ist. Wenn die Stichprobe nämlich Fälle enthält, die von den übrigen Fällen deutlich abweichen, kann es zu Verzerrungen der Ergebnisse kommen.
Generell haben Datenpunkte, die weit vom Schwerpunkt der Punktewolke der abhängigen und unabhängigen Variablen entfernt sind, ein größeres Gewicht bei der Bestimmung der Regressionsgeraden, was daran liegt, das die Regressionsgerade immer durch den Schwerpunkt der Punktewolke verläuft und gleichzeitig die Summe der quadrierten Abweichungen von der Gerade minimiert werden, große Abweichungen also stärker einfließen als kleine Abweichungen.
Sichtbar wird dies an den Hebelwerten h_i, die in die Berechnung der standardisierten Residuen einfließen. Je größer ein Hebelwert ist, desto stärker bestimmt der entsprechende Fall die Lage der Regressionsfunktion.

Neben der Hebelkraft bestimmt auch der Wert des Residuums e_i den Einfluss, den ein Fall i auf die Lage der Regressionsgerade hat. Große Residuen weisen darauf hin, dass ein Fall die Lage relativ stark beeinflusst. Das nach dem Statistiker Cook benannte Maß ***Cooks Distanz D*** erfasst den Einfluss eines Falles, wobei sowohl die Höhe des Residuums als auch der Hebelwert in die Berechnung eingeht:

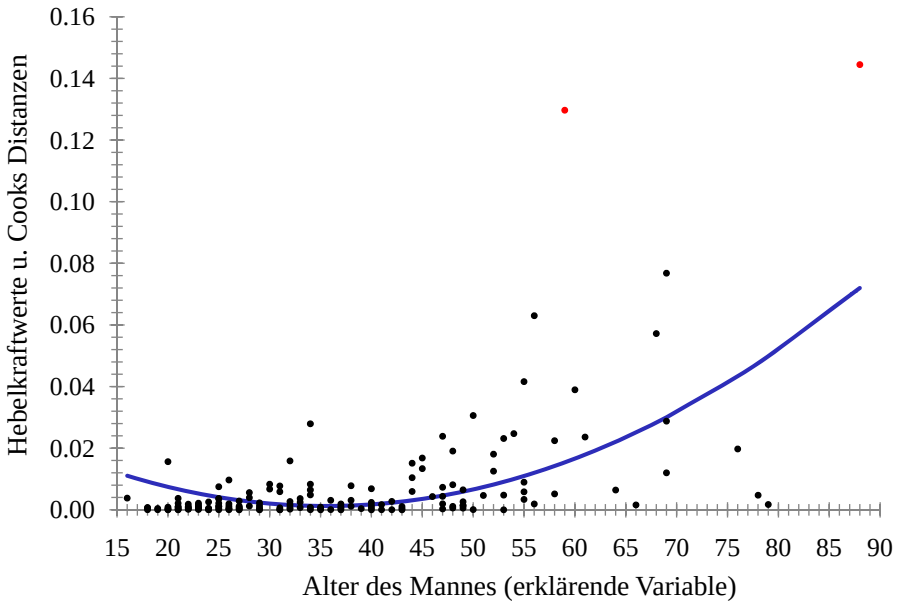
$$D_i = \frac{\left(e_i^*\right)^2}{2} \cdot \frac{h_i}{1-h_i}$$

Ausreißer und einflussreiche Fälle



In der Abbildung sind die Kurve der Hebelwerte h_i und Cooks Distanzen D_i nach den Werten der unabhängigen Variablen X (Alter des Mannes) als Punkte eingezeichnet. Deutlich sichtbar ist der u-förmige Verlauf der Hebelkraftwerte.
Interessanter sind Cooks Distanzen. Während die meisten Werte recht klein sind, gibt, es doch zwei stärker herausragende Fälle, die hier rot markiert sind.

Ausreißer und einflussreiche Fälle



Der größere Wert gehört zu einem 88jährigen Mann, dessen Partnerin 70 Jahre alt ist, der andere zu einem 59jährigen, dessen Partnerin 36 Jahre alt ist.

Werden diese beiden Fälle von der Analyse ausgeschlossen und die Regression für die verbleibenden 183 Fälle berechnet, so ergibt sich als neue Vorhersagegleichung:

$$\hat{Y} = 0.983 + 0.901 \cdot X \text{ anstelle von } \hat{Y} = 1.614 + 0.876 \cdot X$$

Der Ausschluss der beiden Fälle hat also merkliche Auswirkungen auf die Schätzung der Regressionskoeffizienten.

Ausreißer und einflussreiche Fälle

Dieser Effekt allein sollte allerdings nicht als hinreichender Grund genommen werden, die beiden Fälle tatsächlich von der Analyse auszuschließen.
Wenn es sich nämlich nicht um einen Datenfehler handelt, kann der Ausschluss abweichender Fälle zur Missachtung besonders interessanter Informationen führen.
Eine bessere Strategie besteht daher darin, sich solche abweichenden Fälle näher anzusehen.

Konsequenzen von Verletzungen der Modellannahmen

Um zu demonstrieren, welche Auswirkungen es hat, wenn die Anwendungsvoraussetzungen nicht erfüllt sind, können Simulationsstudien durchgeführt werden. Diese führen zu folgenden Ergebnissen:

(1) Die Verletzung der Normalverteilungsannahme hat auf die Schätzung der Regressionskoeffizienten und Standardfehler keine Auswirkungen.

(2) Ist dagegen die Linearitätsannahme verletzt, schätzt die OLS-Regression eine lineare Trendlinie, die im Sinne der kleinsten Quadrate die bestmögliche lineare Annäherung an die tatsächliche nichtlineare Regressionsfunktion ist.
Die Standardfehler der Koeffizienten der Trendlinie können allerdings verzerrt sein, weil selbst bei einer homoskedastischen Residualvarianz der Wahren nichtlinearen Regressionsfunktion die Abweichungen von der Trendlinie heteroskedastisch sind.

Konsequenzen von Verletzungen der Modellannahmen

(3) Besteht in der Population eine lineare Beziehung, bei der die Residuen mit der erklärenden Variable korreliert sind, dann können die Koeffizienten der datengenerierenden linearen Gleichung nur verzerrt geschätzt werden. Der Erwartungswert der Schätzer ist auch hier eine optimale Trendlinie und nicht die wahre lineare Beziehung zwischen den Variablen. Wenn die Korrelation zwischen den Residuen und der erklärenden Variable durch eine lineare Beziehung hervorgerufen wird, ist diese Trendlinie gleichzeitig die zutreffende lineare Regression von Y auf X (allerdings nicht die zutreffende kausale Beziehung zwischen Y und X). Falls die Residuen zudem homoskedastisch sind, werden die Standardfehler der Koeffizienten der Trendlinie unverzerrt geschätzt.

(4) Bei heteroskedastischen Residuen werden die Regressionskoeffizienten unverzerrt geschätzt. Die geschätzte Residualvarianz in der Population und die Standardfehler der Regressionskoeffizienten können jedoch stark verzerrt sein, wodurch auch Konfidenzintervalle und Tests unbrauchbar werden.
Wenn allerdings die Fallzahlen bei allen Ausprägungen der erklärenden Variablen gleich (und größer 1) sind, werden die Standardfehler recht robust geschätzt.

(5) Autokorrelation unter den Residuen hat die gleichen Konsequenzen wie heteroskedastische Residualvarianzen: die Regressionskoeffizienten werden unverzerrt geschätzt, die Standardfehler sind dagegen verzerrt und werden bei positiven Autokorrelationen unterschätzt.

Lerneinheit 6: Tests von Mittelwertdifferenzen

In vielen sozialwissenschaftlichen Fragestellung geht es um den Vergleich von Mittelwerten.

Beispiele sind:

- *Verdienen Frauen weniger als Männer?*
- *Stufen sich die Personen in den neuen Bundesländern im Mittel als stärker links ein als in den alten Bundesländern?*

Statistisch lassen sich solche Fragen auf das Problem zurückführen, ob sich die (Sub-) Populationsmittelwerte μ_1 und μ_2 unterscheiden bzw. ob sich zwei Zufallsvariablen Y_1 und Y_2 in ihren Erwartungswerten μ_1 und μ_2 unterscheiden.
Zur Beantwortung dieser Fragestellung gibt es spezielle Tests, die in der Sozialforschung sehr oft angewendet werden. Diese Tests können aber auch als Spezialfälle der Anwendung des linearen Regressionsmodells aufgefasst werden.

Abhängige und unabhängige Stichproben

Bevor die Logik des Mittelwertvergleichs vorgestellt wird, muss noch eine wichtige Unterscheidung erläutert werden. Statistische Tests über einen möglichen Unterschied zwischen zwei Populationsmittelwerten bzw. Erwartungswerten μ_1 und μ_2 basieren auf Stichprobendaten, die zu zwei Stichprobenmittelwerten $\bar{y}_1$ und $\bar{y}_2$ führen, die Schätzungen der beiden Populationswerte oder Erwartungswerte sind. Die Standardfehler und Teststatistiken unterscheiden sich beim Mittelwertvergleich nämlich danach, ob die Daten aus unabhängigen oder aus abhängigen Stichproben kommen.

Abhängige und unabhängige Stichproben

Unabhängige Stichproben:
Die Fälle, die in die Berechnung des ersten Stichprobenmittelwerts eingehen, bilden eine Variable Y_1, die Fälle die in die Berechnung des zweiten Stichprobenmittelwerts einfließen, bilden eine Variable Y_2. Wenn die Fälle von Y_1 und Y_2 statistisch unabhängig voneinander zufällig ausgewählt sind, liegen ***unabhängigen Stichproben*** vor.
Die Variablen Y_1 und Y_2 beziehen sich dann in der Regel auf die gleiche Eigenschaft, die bei Fällen aus zwei unterschiedlichen Gruppen erfasst wird.

Y_1 könnte z.B. das Einkommen von Frauen bezeichnen und Y_2 das Einkommen von Männern. Die beiden Stichproben sind statistisch unabhängig voneinander, wenn entweder jeweils getrennte Stichproben aus der Population der Männer und der der Frauen gezogen werden, oder wenn - wie in vielen Umfragen üblich - unabhängig voneinander Personen ausgewählt werden und das Geschlecht bei der Auswahl keine Rolle spielt.

Da Populationsmittelwerte von zwei Gruppen (z.B. der Gruppe der Frauen und der Gruppe der Männer) verglichen werden, wird auch von einem ***Gruppenvergleich*** gesprochen.

Abhängige Stichproben:
Wenn sich die Realisationen der beiden Variablen Y_1 und Y_2 auf die gleichen Untersuchungseinheiten beziehen, liegen ***abhängigen Stichproben*** vor.

Ein Beispiel könnte etwa der Vergleich des Einkommens von zusammenlebenden Männern und Frauen sein. Die betrachteten Paare bilden dann jeweils eine Untersuchungseinheit, die zusammen in die Stichprobe aufgenommen werden.

Abhängige und unabhängige Stichproben

Verdeutlichen lässt sich der Unterschied zwischen abhängigen und unabhängigen Stichproben mit Hilfe von Kreuztabellen:

Abhängige Stichproben:

Zeilen-variable Y_1	Spaltenvariable Y_2 Y_2=1	Y_2=2	Y_2=3	∑
Y_1=1	120	100	80	**300**
Y_1=2	100	200	100	**400**
Y_1=3	80	100	120	**300**
∑	**300**	**400**	**300**	1000

Unabhängige Stichproben:

	Gruppierungsvariable (X) X=1	X=2
Y_k=1	**120**	**100**
Y_k=2	**100**	**200**
Y_k=3	**80**	**100**
n	300	400

Bei zwei ***voneinander abhängigen Stichproben*** bzw. Verteilungen (linke Tabelle) werden die beiden Randverteilungen einer Kreuztabelle betrachtet. Dadurch, dass sich die beiden Randverteilungen auf die gleichen Fälle beziehen, kann eine statistische Abhängigkeit bestehen.

Im Beispiel sieht man, dass jeder der 1000 Fälle sowohl eine Ausprägung bei Y_1 wie bei Y_2 hat. Die zu vergleichenden Mittelwerte von basieren daher auf den gleichen Fällen, für die jeweils Realisierungen von 2 Variablen Y_1 und Y_2 vorliegen. Daher sind die beiden Stichproben voneinander abhängig .

Bei zwei ***voneinander unabhängigen Stichproben*** bzw. Verteilungen werden durch die Gruppierungsvariable definierte bedingte Verteilungen analysiert (rechte Tabelle). Jede Verteilung basiert dann auf unterschiedlichen Fällen, wobei auch die Fallzahl in den Gruppen verschieden sein kann. Die Fälle in den beiden Verteilung müssen unabhängig voneinander in die jeweilige Stichprobe aufgenommen sein. Es besteht somit kein statistischer Zusammenhang zwischen der Auswahl eines Falles in der ersten und eines Falles in der zweiten Gruppe.

Unabhängigkeit in abhängigen Stichproben

Unabhängige Stichproben:

	Gruppierungsvariable X=1	X=2
Y_k=1	**120**	**100**
Y_k=2	**100**	**200**
Y_k=3	**80**	**100**
n	300	400

Im Beispiel der unabhängigen Stichproben beträgt die Fallzahl in der ersten Gruppe 300 Fälle und in der zweiten Gruppe 400 Fälle. Kein Fall einer Gruppe ist auch Fall in der anderen Gruppe. Die Stichproben sind unabhängig voneinander.

Hinweis:

Abhängigkeit bei der Stichprobenziehung bedeutet nicht notwendigerweise, dass die beiden Variablen tatsächlich statistisch zusammenhängen. Die gemeinsame Verteilung der beiden Variablen kann so beschaffen sein, dass die Verteilungen voneinander unabhängig sind.

Im folgenden Beispiel sind die Besetzungszahlen (absoluten Häufigkeiten) in den Tabellenzellen so beschaffen, dass statistische Unabhängigkeit besteht.

Obwohl die Stichprobe auf gemeinsamen Fällen beruht, es sich also um eine abhängige Stichprobe handelt, sind Y_1 und Y_2 statistisch unabhängig voneinander und die Kovarianz ist Null.

Zeilen-variable Y_1	Spaltenvariable Y_2 Y_2=1	Y_2=2	Y_2=3	∑
Y_1=1	60	80	60	**200**
Y_1=2	180	240	180	**600**
Y_1=3	60	80	60	**200**
∑	**300**	**400**	**300**	1000

$\bar{y}_1 = \bar{y}_2 = 2$

$s(Y_1,Y_2) = (60\times(1-2)\cdot(1-2) + 80\times(1-2)\cdot(2-2)$
$+ 60\times(1-2)\cdot(3-2) + 180\times(2-2)\cdot(1-2)$
$+ 240\times(2-2)\cdot(2-2) + 180\times(2-2)\cdot(3-2)$
$+ 60\times(3-2)\cdot(1-2) + 80\times(3-2)\cdot(2-2)$
$+ 60\times(3-2)\cdot(3-2)) / 1000 = 0.0$

Mittelwertvergleich bei unabhängigen Stichproben

Für inferenzstatistische Aussagen über Mittelwertvergleiche werden die Standardfehler der Differenzen von Stichprobenmittelwerten benötigt.
Aus dem zentralen Grenzwertsatz folgt, dass die Kennwerteverteilung von Stichprobenmittelwerten bei nahezu beliebiger Verteilung einer Variable in der Population asymptotisch normalverteilt ist, wenn die Fälle, die in die Berechnung der Mittelwerte eingehen, als statistisch unabhängige identisch verteilte Realisierungen von Zufallsvariablen aufgefasst werden können.
Bei einfachen Zufallsauswahlen (mit Zurücklegen) gilt für jeden Mittelwert einer Variable Y_k:

$$\bar{y}_k \underset{n\to\infty}{\sim} N\left(\mu_k; \sigma_k^2 / n_k\right)$$

Dabei steht μ_k für den Erwartungswert bzw. Populationsmittelwert von σ^2_k für die Populationsvarianz von Y_k und n_k für die Fallzahl. Die asymptotische Verteilung gilt unabhängig davon, ob die Populationsvarianz σ^2_k bekannt ist oder aus den Stichprobendaten konsistent geschätzt wird.

Linearkombinationen von (asymptotisch) normalverteilten Variablen sind wiederum (asymptotisch) normalverteilt. Daher ist die Differenz von zwei Stichprobenmittelwerten ebenfalls (asymptotisch normalverteilt). Für die Erwartungswerte und Varianzen der Differenz ergibt sich nach den generellen Regeln für Linearkombinationen (s. L04-22ff.):

$$\mu\left(\bar{y}_1 - \bar{y}_2\right) = \mu\left(0 + 1\cdot\bar{y}_1 + (-1)\cdot\bar{y}_1\right) = \mu\left(\bar{y}_1\right) - \mu\left(\bar{y}_1\right) = \mu_1 - \mu_2$$

$$\sigma^2\left(\bar{y}_1 - \bar{y}_2\right) = \sigma^2\left(0 + 1\cdot\bar{y}_1 + (-1)\cdot\bar{y}_1\right) = \sigma^2\left(\bar{y}_1\right) + \sigma^2\left(\bar{y}_2\right) - 2\cdot\sigma\left(\bar{y}_1, \bar{y}_2\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - 2\cdot 0.$$

Mittelwertvergleich bei unabhängigen Stichproben

Die Varianz der Kennwerteverteilung der Mittelwertdifferenz kann dann konsistent und erwartungstreu aus den Stichprobendaten geschätzt werden:

$$\hat{\sigma}^2\left(\bar{y}_1 - \bar{y}_2\right) = \frac{\sum_{i=1}^{n_1}\left(y_{1,i} - \bar{y}_1\right)}{\left(n_1 - 1\right)\cdot n_1} + \frac{\sum_{i=1}^{n_2}\left(y_{2,i} - \bar{y}_2\right)}{\left(n_2 - 1\right)\cdot n_2} = \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2} = \frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}$$

Daher ist die Teststatistik:

$$Z = \frac{\left(\bar{y}_1 - \bar{y}_2\right) - \mu}{\hat{\sigma}\left(\bar{y}_1 - \bar{y}_2\right)} = \frac{\left(\bar{y}_1 - \bar{y}_2\right) - \mu}{\sqrt{\frac{SS_1}{n_1\cdot\left(n_1 - 1\right)} + \frac{SS_2}{n_2\cdot\left(n_2 - 1\right)}}} = \frac{\left(\bar{y}_1 - \bar{y}_2\right) - \mu}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} = \frac{\left(\bar{y}_1 - \bar{y}_2\right) - \mu}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}}$$

asymptotisch mit einer Varianz von 1 normalverteilt.
Der Erwartungswert von Z ist gleich der Differenz der Konstante μ von der Differenz der Populationsmittelwerte:

$$\mu(Z) = (\mu_1 - \mu_2) - \mu.$$

Wenn die Differenz der Populationsmittelwerte also gleich der in der Teststatistik vorgegebenen Konstante μ ist: $\mu_1 - \mu_2 = \mu$, dann ist Z asymptotisch standardnormalverteilt.

Mittelwertvergleich bei unabhängigen Stichproben

Teststatistik bei gleichen Varianzen in den Gruppen:
Wenn vermutet wird, dass sich die Verteilungen von Y_1 und Y_2 in der Population höchsten nur in ihren Mittelwerten unterscheiden, dann sind die Populationsvarianzen gleich. Statt getrennter Schätzungen der Populationsvarianz bietet es sich dann an, eine gemeinsame Schätzung zu verwenden, da diese einen kleineren Standardfehler aufweist als die getrennten Schätzungen der Varianzen in den Gruppen.

Die gemeinsame (engl: „pooled“) Schätzung der Populationsvarianz beträgt:

$$\hat{\sigma}^2_{pooled} = \frac{\sum_{i=1}^{n}\left(y_{1,i}-\bar{y}_1\right)^2 + \sum_{i=1}^{n}\left(y_{2,i}-\bar{y}_2\right)^2}{n_1+n_2-2} = \frac{n_1\cdot s_1^2 + n_2\cdot s_2^2}{n_1+n_2-2} = \frac{(n_1-1)\cdot\hat{\sigma}_1^2 + (n_2-1)\cdot\hat{\sigma}_2^2}{n_1+n_2-2}$$

Hinweis:
*Diese Schätzung der gleich großen Populationsvarianz um möglicherweise unterschiedliche Mittelwerte sollte nicht mit der Berechnung der **Gesamtvarianz** von Zusammenfassungen verwechselt werden, für die gilt:*

$$s_Y^2 = \frac{\sum_{i=1}^{n_1}\left(y_{1,i}-\bar{y}_1\right)^2 + \sum_{i=1}^{n_2}\left(y_{2,i}-\bar{y}_2\right)^2}{n_1+n_2} + \frac{n_1\cdot\left(\bar{y}_1-\bar{y}\right)^2 + n_2\cdot\left(\bar{y}_2-\bar{y}\right)^2}{n_1+n_2} \quad \text{mit } \bar{y} = \frac{n_1\cdot\bar{y}_1 + n_2\cdot\bar{y}_w}{n_1+n_2}$$

Der erste Summand der Gesamtvarianz stimmt bis auf den Wert –2 im Nenner mit der gemeinsamen Schätzung der Populationsvarianz überein. Der zweite Summand ist der Teil der Gesamtvarianz, der sich durch unterschiedliche Gruppenmittelwerte ergibt.

Mittelwertvergleich bei unabhängigen Stichproben und gleichen Varianzen

Bei angenommener gleicher Populationsvarianz in den Gruppen ändert sich der geschätzte Standardfehler im Nenner der Teststatistik des Mittelwertvergleichs, da die gemeinsame Varianz anstelle der getrennten Varianzschätzungen genutzt wird:

$$Z = \frac{(\bar{y}_1-\bar{y}_2)-\mu}{\hat{\sigma}(\bar{y}_1-\bar{y}_2)} = \frac{(\bar{y}_1-\bar{y}_2)-\mu}{\sqrt{\frac{\hat{\sigma}^2_{pooled}}{n_1}+\frac{\hat{\sigma}^2_{pooled}}{n_2}}} = \frac{(\bar{y}_1-\bar{y}_2)-\mu}{\sqrt{\hat{\sigma}^2_{pooled}\cdot\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{(\bar{y}_1-\bar{y}_2)-\mu}{\sqrt{\frac{n_1\cdot s_1^2+n_2\cdot s_2^2}{n_1+n_2-2}\cdot\left(\frac{1}{n_1-1}+\frac{1}{n_2-1}\right)}}$$

Die Teststatistik ist wieder asymptotisch standardnormalverteilt, wenn die Konstante $\mu = \mu_1 - \mu_2$ ist, ansonsten normalverteilt mit dem Erwartungswert $\mu(Z) = \mu_1 - \mu_2 - \mu$.

Welche Teststatistik sollte angewendet werden?
Wenn bekannt ist, dass die Populationsvarianzen in beiden Gruppen gleich groß sind, sollte die gemeinsame Schätzung der Populationsvarianz verwendet werden. In der Regel ist dies aber nicht bekannt. Simulationen haben nun gezeigt, dass die Schätzung unter der Annahme gleicher Varianzen sehr robust ist, wenn die Varianzen zwar verschieden sind, aber die Fallzahlen in beiden Gruppen gleich groß sind.
Nur wenn die Fallzahlen verschieden sind und sich die Populationsvarianzen unterscheiden, sollte die getrennte Schätzung der Varianzen verwendet werden. Hinweise für unterschiedliche Varianzen geben die Werte s^2_1 und s^2_2. Darüber hinaus gibt es Tests auf Gleichheit von Varianzen, die in späteren Lerneinheiten vorgestellt werden

Mittelwertvergleich bei unabhängigen Stichproben

Standardnormalverteilung oder T-Verteilung als Testverteilung?
Beim Test eines einzelnen Mittelwerts ist die T-Verteilung eine nicht nur asymptotisch, sondern bei jeder Fallzahl exakt zutreffende Verteilung, wenn die betrachtete Variable in der Population normalverteilt ist.
Bei gleicher Varianz in den Gruppen gilt dies auch für die Teststatistik beim Mittelwertvergleich. Anstelle der Standardnormalverteilung kann also auch die T-Verteilung herangezogen werden.
Die Zahl der Freiheitsgrade ist hier:

$$df = n_1 + n_2 - 2,$$

da bezogen auf die Summe der beiden Stichprobenfallzahlen zwei Freiheitsgrade durch die Schätzung der beiden Mittelwerte verloren gehen.

Auch bei ungleichen Varianzen kann die T-Verteilung verwendet werden, wenn die Freiheitsgrade nach folgender Formel berechnet werden:

$$df = \frac{1}{\frac{\left(\frac{\hat{\sigma}_1^2 / n_1}{\hat{\sigma}_1^2 / n_1 + \hat{\sigma}_2^2 / n_2}\right)^2}{(n_1 - 1)} + \frac{\left(\frac{\hat{\sigma}_2^2 / n_1}{\hat{\sigma}_1^2 / n_1 + \hat{\sigma}_2^2 / n_2}\right)^2}{(n_2 - 1)}} = \frac{1}{\frac{\left(\frac{\hat{\sigma}^2(\bar{y}_1)}{\hat{\sigma}^2(\bar{y}_1) + \hat{\sigma}^2(\bar{y}_2)}\right)^2}{(n_1 - 1)} + \frac{\left(\frac{\hat{\sigma}^2(\bar{y}_2)}{\hat{\sigma}^2(\bar{y}_1) + \hat{\sigma}^2(\bar{y}_2)}\right)^2}{(n_2 - 1)}}$$

Die so berechneten Freiheitsgrade sind reelle Zahlen, können also Nachkommastellen haben. Die Kennwerteverteilung ist dann genau genommen eine Verallgemeinerung der T-Verteilung.

Mittelwertvergleich bei unabhängigen Stichproben

Im Sinne eines konservativen Vorgehens bzw. strengen Testens wird die T-Verteilung anstelle der Standardnormalverteilung auch dann angewendet, wenn die Verteilungen in der Population nicht normal sind, da die T-Verteilung zu einem größeren Annahmebereich der Nullhypothese bzw. längeren Konfidenzintervallen führt. Sinnvoll ist dies nur, wenn die Forschungshypothese die Alternativhypothese ist.

Mittelwertgleich als Regressionsmodell:
Der Test der Differenz zweier Mittelwerte aus unabhängigen Stichproben kann auch über ein lineares Regressionsmodell realisiert werden. Dazu wird die Gruppenzugehörigkeit über eine dichotome erklärende Variable X mit den beiden Ausprägungen 0 und 1 gemessen. Wenn ein Fall aus der ersten Gruppe kommt, gilt X=1; wenn er aus der anderen Gruppe kommt, gilt X=0. Anstelle von zwei Variablen Y_1 und Y_2 wird nun nur eine abhängige Variable Y betrachtet, die alle Fälle aus beiden Stichproben umfasst.
Der Zusammenhang mit dem Regressionsmodell wird naheliegender, wenn anstelle der Anordnung der Daten als Kreuztabelle alle Ausprägungskombinationen der zwei Variablen Y und X untereinander geschrieben werden:

	X=1	X=0
Y_k=1	100	100
Y_k=2	100	200
Y_k=3	200	100
n	400	400

$\Rightarrow$

Y	X	n_k	
1	1	100	} Y_1
2	1	100	
3	1	200	
1	0	100	} Y_2
2	0	200	
3	0	100	

Mittelwertvergleich bei unabhängigen Stichproben

Y	X	n_k	
1	1	100	} Y_1
2	1	100	
3	1	200	
1	0	100	
2	0	200	} Y_2
3	0	100	

$$\hat{Y} = a + b \cdot X$$

Wenn X=0: $\hat{y}_0 = a + b \cdot 0 = a = \overline{y}_2$

wenn X=1: $\hat{y}_1 = a + b \cdot 1 = a + b = \overline{y}_1$

Der Mittelwertvergleich lässt sich dann als lineare Regression der abhängigen Variable Y auf die dichotome 0/1-kodierte ***Indikatorvariable*** X modellieren.
Da sich zwei Punkte (Mittelwerte) immer durch eine gerade Linie verbinden lassen, ist die Regressionsfunktion notwendigerweise linear. Daher müssen die Vorhersagewerte immer mit den Mittelwerten in den beiden Teilstichproben exakt übereinstimmen.
Wenn X=0, ist der Vorhersagewert gleich der Regressionskonstante. Die Regressionskonstante schätzt daher den bedingten Populationsmittelwert von Y in der zweiten Gruppe, also μ_2. Wenn X=1, schätzt der Vorhersagewert den Populationsmittelwert in der ersten Gruppe μ_1.
Das Regressionsgewicht b misst dann also genau die Differenz der Gruppenmittelwerte in der Stichprobe:

$$\overline{y}_1 - \overline{y}_2 = (a + b) - a = b$$

Der Test des Regressionsgewichts b prüft daher, ob sich die beiden Gruppenmittelwerte in der Population unterscheiden.

Mittelwertvergleich bei unabhängigen Stichproben

	X=1	X=0
Y_k=1	100	100
Y_k=2	100	200
Y_k=3	200	100
n	400	400

$n_1 = 400$; $\overline{y}_1 = 2.25$; $s_1^2 = 0.6875$

$n_2 = 400$; $\overline{y}_2 = 2.00$; $s_2^2 = 0.5000$

	Y	X	n_k	$n_k \times Y$	$n_k \times Y^2$	$n_k \times X$	$n_k \times X^2$	$n_k \times X \times Y$
Y_1	1	1	100	100	100	100	100	100
	2	1	100	200	400	100	100	200
	3	1	200	600	1800	200	200	600
Y_2	1	0	100	100	100	0	0	0
	2	0	200	400	800	0	0	0
	3	0	100	300	900	0	0	0
				1700	4100	400	400	900

$n = 800$;

$\overline{y} = 2.125$; $s_Y^2 = 0.609375$; $\overline{x} = 0.5$; $s_X^2 = 0.25$; $s_{XY} = 0.0625$

Berechnet man für die fiktiven Beispieldaten das Regressionsgewicht, zeigt sich, dass es tatsächlich die Mittelwertdifferenz misst:

$$b = \frac{s_{XY}}{s_X^2} = \frac{0.0625}{0.250} = 0.25 = 2.25 - 2.00 = \overline{y}_1 - \overline{y}_2$$

Da die Vorhersagewerte gleich den Gruppenmittelwerten sind, ist die Variation der Residuen gleich der Summe der Variationen in den beiden Gruppen:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n_1} (y_{1,i} - (a + b \cdot 1))^2 + \sum_{i=1}^{n_2} (y_{2,i} - (a + b \cdot 0))^2$$

$$= \sum_{i=1}^{n_1} (y_{1,i} - \overline{y}_1)^2 + \sum_{i=1}^{n_2} (y_{2,i} - \overline{y}_2)^2 = n_1 \cdot s_1^2 + n_2 \cdot s_2^2$$

Mittelwertvergleich bei unabhängigen Stichproben

	X=1	X=0
Y_k=1	100	100
Y_k=2	100	200
Y_k=3	200	100
n	400	400

$n_1 = 400$; $\bar{y}_1 = 2.25$; $s_1^2 = 0.6875$

$n_2 = 400$; $\bar{y}_2 = 2.00$; $s_2^2 = 0.5000$

	Y	X	n_k	E	$n_k \times E$	$n_k \times E^2$
Y_1	1	1	100	−1.25	−125	156.25
	2	1	100	−0.25	−25	6.25
	3	1	200	0.75	150	112.50
Y_2	1	0	100	−1.00	−100	100.00
	2	0	200	0.00	0	0.00
	3	0	100	1.00	100	100.00
					0	475.00

$n = 800$;

$\bar{y} = 2.125$; $s_Y^2 = 0.609375$; $\bar{x} = 0.5$; $s_X^2 = 0.25$; $s_{XY} = 0.0625$

$$\sum_{i=1}^{n} e_i^2 = n_1 \cdot s_1^2 + n_2 \cdot s_2^2 = 400 \cdot 0.6875 + 400 \cdot 0.5 = 475$$

Die geschätzte Varianz der Populationsresiduen ist dann gleich der gemeinsamen Populationsvarianz:

$$\hat{\sigma}_U^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{n_1 \cdot s_1^2 + n_2 \cdot s_2^2}{n_1 + n_2 - 2} = \hat{\sigma}_{poled}^2 \; ; \; \frac{475}{800-2} = \frac{400 \cdot 0.6875 + 400 \cdot 0.5}{400 + 400 - 2} = 0.595$$

Bei einer 0/1-kodierten Indikatorvariable X ist die Variation von X eine Funktion der Fallzahlen in den beiden Gruppen:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2 = n_1 - n \cdot \left(\frac{n_1}{n}\right)^2 = n_1 - \frac{n_1^2}{n} = \frac{n_1 \cdot (n - n_1)}{n} = \frac{n_1 \cdot n_2}{n_1 + n_2} = \frac{1}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Mittelwertvergleich bei unabhängigen Stichproben

	X=1	X=0
Y_k=1	100	100
Y_k=2	100	200
Y_k=3	200	100
n	400	400

$n_1 = 400$; $\bar{y}_1 = 2.25$; $s_1^2 = 0.6875$

$n_2 = 400$; $\bar{y}_2 = 2.00$; $s_2^2 = 0.5000$

	Y	X	n_k	$n_k \times Y$	$n_k \times Y^2$	$n_k \times X$	$n_k \times X^2$	$n_k \times X \times Y$
Y_1	1	1	100	100	100	100	100	100
	2	1	100	200	400	100	100	200
	3	1	200	600	1800	200	200	600
Y_2	1	0	100	100	100	0	0	0
	2	0	200	400	800	0	0	0
	3	0	100	300	900	0	0	0
				1700	4100	400	400	900

$n = 800$;

$\bar{y} = 2.125$; $s_Y^2 = 0.609375$; $\bar{x} = 0.5$; $s_X^2 = 0.25$; $s_{XY} = 0.0625$

$$\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2 = n_1 - \frac{n_1^2}{n} = 400 - \frac{400^2}{800} = 200 = 1 / \left(\frac{1}{400} + \frac{1}{400}\right) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}$$

Der geschätzte Standardfehler des Regressionsgewicht weist daher die gleichen Werte auf wie der geschätzte Standardfehler der Differenz zweier Mittelwerte bei gemeinsamer Schätzung der Varianzen:

$$\hat{\sigma}_b = \sqrt{\frac{\frac{1}{n-2} \cdot \sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} = \sqrt{\frac{\sum_{i=1}^{n_1} (y_{1,i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2} (y_{2,i} - \bar{y}_2)^2}{\frac{n-2}{\frac{n_1 \cdot n_2}{n}}}} = \sqrt{\frac{n_1 \cdot s_1^2 + n_2 \cdot s_2^2}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \hat{\sigma}(\bar{y}_1 - \bar{y}_2)$$

Mittelwertvergleich bei unabhängigen Stichproben

	X=1	X=0
Y_k=1	100	100
Y_k=2	100	200
Y_k=3	200	100
n	400	400

Y	X	n_k	E	$n_k \times E$	$n_k \times E^2$
1	1	100	−1.25	−125	156.25
2	1	100	−0.25	−25	6.25
3	1	200	0.75	150	112.50
1	0	100	−1.00	−100	100.00
2	0	200	0.00	0	0.00
3	0	100	1.00	100	100.00
n = 800 ;				0	475.00

$n_1 = 400$; $\bar{y}_1 = 2.25$; $s_1^2 = 0.6875$

$n_2 = 400$; $\bar{y}_2 = 2.00$; $s_2^2 = 0.5000$

$\bar{y} = 2.125$; $s_Y^2 = 0.609375$; $\bar{x} = 0.5$; $s_X^2 = 0.25$; $s_{XY} = 0.0625$

$$\hat{\sigma}_b = \sqrt{\frac{\frac{1}{n-2}\cdot\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} = \sqrt{\frac{475/798}{200}} = 0.055 = \sqrt{\frac{400\cdot 0.6875 + 400\cdot 0.5}{400+400-2}\cdot\left(\frac{1}{400}+\frac{1}{400}\right)} = \hat{\sigma}(\bar{y}_1-\bar{y}_2)$$

Aus der Gleichheit von Regressionsgewicht und Stichprobenmittelwertdifferenz sowie ihrer Standardfehler folgt, dass Tests über das Regressionsgewicht bis auf mögliche Rundungsfehler stets die gleichen Ergebnisse liefern wie die Tests von Mittelwertdifferenzen bei gleichen Varianzen. Die Annahme gleicher Varianzen in den beiden Gruppen entspricht im Regressionsmodell der Homoskedastizitätsannahme. Aus der Robustheit des Mittelwertvergleichs folgt somit auch, dass die Homoskedastizitätsverletzung im Regressionsmodell harmlos ist, wenn die Fallzahlen bei den Ausprägungen von X gleich groß sind.

Mittelwertvergleich bei abhängigen Stichproben

Auch für den Vergleich von Mittelwerten in abhängigen Stichproben gilt, dass die Differenz der beiden Stichprobenmittelwerte eine asymptotisch normalverteilte Kennwerteverteilung aufweist.
Es ist hier allerdings zu berücksichtigen, dass die Populationskovarianz σ_{21} zwischen Y_1 und Y_2 ungleich Null sein kann. Der Standardfehler der Mittelwertdifferenz berechnet sich daher hier nach:

$$\hat{\sigma}(\bar{y}_1-\bar{y}_2) = \sqrt{\frac{\hat{\sigma}_1^2}{n}+\frac{\hat{\sigma}_2^2}{n}-2\cdot\frac{\hat{\sigma}_{21}}{n}} = \sqrt{\frac{s_1^2}{n-1}+\frac{s_2^2}{n-1}-2\cdot\frac{s_{21}}{n-1}}$$

Da es nur eine Stichprobe gibt, bei deren Fällen beide Eigenschaften Y_1 und Y_2 erfasst sind, gibt es im Unterschied zu unabhängigen Stichproben nun eine gemeinsame Fallzahl n.
Analog zur Teststatistik für den Mittelwertvergleich bei unabhängigen Stichproben ergibt sich dann als Teststatistik:

$$Z \text{ bzw. } T = \frac{(\bar{y}_1-\bar{y}_2)-\mu}{\hat{\sigma}(\bar{x}_1-\bar{x}_2)} = \frac{(\bar{y}_1-\bar{y}_2)-\mu}{\sqrt{\frac{s_1^2+s_2^2-2\cdot s_{21}}{n-1}}} = \frac{(\bar{y}_1-\bar{y}_2)-\mu}{\sqrt{\frac{\hat{\sigma}_1^2+\hat{\sigma}_2^2-2\cdot\hat{\sigma}_{21}}{n}}}$$

Wenn in der Population die Mittelwertdifferenz $\mu_1 - \mu_2 = \mu$ ist, dann ist die Teststatistik asymptotisch normalverteilt. Alternativ kann wiederum die T-Verteilung herangezogen werden, die hier df=n–1 Freiheitsgrade hat.

Mittelwertvergleich bei abhängigen Stichproben

Anstelle der Berechnung von Mittelwertdifferenzen ist es bei einer abhängigen Stichprobe oft einfacher, zunächst für jeden Fall die Differenz $D = Y_1 - Y_2$ zu berechnen. Der Populationsmittelwert dieser Differenzvariable muss dann nach den Regeln für Linearkombinationen $\mu_D = \mu_1 - \mu_2$ sein. Die Teststatistik ist daher gleich dem Test eines einfachen Mittelwerts:

$$\text{Z bzw. T} = \frac{\bar{d} - \mu}{\hat{\sigma}(\bar{d})} = \frac{\bar{d} - \mu}{\sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n \cdot (n-1)}}} = \frac{\bar{d} - \mu}{\sqrt{\frac{SS_D}{n \cdot (n-1)}}} = \frac{\bar{d} - \mu}{\hat{\sigma}_D \cdot \sqrt{n}} = \frac{\bar{d} - \mu}{s_D \cdot \sqrt{n-1}}$$

Wenn $\mu_D = \mu$, dann ist die Teststatistik asymptotisch standardnormalverteilt bzw. bei in der Population normalverteilten Ausgangsvariablen Y_1 und Y_2 mit df = n–1 Freiheitsgraden t-verteilt.

Mittelwertgleich als Regressionsmodell:
Auch bei abhängigen Stichproben kann der Mittelwertvergleich mittels linearer Regression erfolgen. Die Differenzvariable D kann nämlich als ein restriktives Regressionsmodell von Y_1 auf Y_2 aufgefasst werden, bei dem die Apriori-Annahme getroffen wird, dass das Regressionsgewicht b=1 ist:

$$D = Y_1 - Y_2 \Rightarrow Y_1 = a + 1 \cdot Y_2 + E$$

Die OLS-Schätzung der Regressionskonstante ist dann gleich der Differenz der beiden Stichprobenmittelwerte.

Mittelwertvergleich bei abhängigen Stichproben

$$a = \bar{y}_1 - b \cdot \bar{y}_2 = \bar{y}_1 - 1 \cdot \bar{y}_2$$

Bei der Berechnung der geschätzten Standardfehler ist zu berücksichtigen, dass nur ein Regressionskoeffizient und nicht zwei geschätzt werden. Die Zahl der Freiheitsgrade ist daher df = n–1 und der erwartungstreue Schätzer der Residualvarianz ergibt sich nach:

$$\hat{\sigma}_U^2 = \frac{\sum_{i=1}^{n} e_i^2}{df} = \frac{\sum_{i=1}^{n}\left(y_{1,i} - \left(a + 1 \cdot y_{2,i}\right)\right)^2}{n-1} = \frac{\sum_{i=1}^{n}\left(y_{i,1} - y_{i,2} - a\right)^2}{n-1} = \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1} = \hat{\sigma}_D^2$$

Da die Regressionskonstante gleich der Differenz der Stichprobenmittelwerte und damit gleich dem Mittelwert der Differenzvariable D ist, folgt zudem, dass der Standardfehler von a im Regressionsmodell mit vorgegebenem Wert b=1 gleich dem Standardfehler vom Stichprobenmittelwert D und damit gleich der geschätzten Residualvarianz geteilt durch die Fallzahl ist:

$$\hat{\sigma}_a = \hat{\sigma}(\bar{d}) = \frac{\hat{\sigma}_D^2}{n} = \frac{\hat{\sigma}_U^2}{n}$$

Wie bei unabhängigen Stichproben ergibt sich also stets das gleiche Ergebnis, unabhängig davon, ob die Differenz der Mittelwerte betrachtet wird oder die Regressionskonstante bei einer Regression von Y_1 auf Y_2, bei der das Regressionsgewicht auf 1 festgesetzt und nicht geschätzt wird. Da es hier zudem nur einen einzigen Vorhersagewert 1 gibt, ist die Homoskedastizitätsannahme (im durch b=1 restringierten Regressionsmodell) irrelevant.

Anwendungen von Mittelwertvergleichen

In inferenzstatistischen Anwendungen von Gruppenvergleichen können Konfidenzintervalle für Mittelwertdifferenzen in der Population oder Tests über Mittelwertdifferenzen durchgeführt werden.
In jedem Fall müssen zunächst Entscheidungen getroffen werden, die die Berechnung der Standardfehler und die Wahl der Kennwerteverteilung festlegen:

(1) Soll die Standardnormalverteilung oder die T-Verteilung als Kennwerteverteilung verwendet werden?
(2) Handelt es sich um abhängige oder um unabhängige Stichproben?
(3) Wenn es sich um unabhängige Stichproben handelt: Soll der Standardfehler unter der Bedingung gleicher oder verschiedener Populationsvarianzen in den Gruppen berechnet werden?

(1) Standardnormalverteilung oder T-Verteilung

Die Standardnormalverteilung ist unabhängig von der Verteilung in der Population anwendbar, setzt aber aufgrund der nur asymptotischen Gültigkeit größere Fallzahlen voraus. Als Faustregel gilt, dass die Annäherung hinreichend genau ist, wenn bei abhängigen Stichproben oder bei unabhängigen Stichproben und gleichen Varianzen die Fallzahl $n \geq 30$ ist und bei unabhängigen Stichproben und verschiedenen Varianzen $n_1 \geq 30$ und $n_2 \geq 30$.
Die T-Verteilung ist auch bei kleineren Fallzahlen anwendbar, setzt jedoch streng genommen Normalverteilung von Y_1 und Y_2 in der Population voraus.

Anwendungen von Mittelwertvergleichen

Da die kritischen Werte bei der T-Verteilung weiter von Null entfernt sind als bei der Standardnormalverteilung, wird die T-Verteilung im Sinne eines konservativen Vorgehens auch dann verwendet, wenn die Normalverteilungsannahmen nicht gegeben ist. Konfidenzintervalle sind dann länger und Nullhypothesen werden nicht so leicht abgelehnt. Sinnvoll ist dies dann, wenn die Forschungshypothese Alternativhypothese ist.
Die Zahl der Freiheitsgrade für die T-Verteilung ist bei unabhängigen Stichproben $df=n_1+n_2-2$ und bei abhängigen Stichproben $df=n-1$.

(2) Abhängige oder unabhängige Stichproben

Abhängige Stichproben liegen immer dann vor, wenn die Messungen der Variablen Y_1 und Y_2 bei den gleichen Erhebungseinheiten (Fällen) durchgeführt wurden.
Unabhängige Stichproben liegen vor, wenn dies nicht der Fall ist und es zudem keinerlei Korrespondenz bei der Auswahl eines Falles der ersten Stichprobe mit der Auswahl eines Falles der zweiten Stichprobe gibt.

(3) Gleiche oder ungleiche Populationsvarianzen

Die Schätzung des Standardfehlers sollte bei unabhängigen Stichproben von gleichen Varianzen ausgehen, wenn die Populationsvarianzen gleich groß zu sein scheinen oder die Fallzahlen in den Gruppen gleich sind. Nur wenn beide Bedingungen nicht erfüllt sind, sollte die Brechnung nach der Formel für ungleiche Varianzen erfolgen.
Als Hinweis kann der Quotient der Varianzen in beiden Gruppen berechnet werden. Wenn der Quotient deutlich von 1 abweicht, ist von ungleichen Varianzen in der Population auszugehen.

Anwendungen von Mittelwertvergleichen

Berechnung von Konfidenzintervallen für Mittelwertvergleiche

Die Berechnung von Konfidenzintervallen für Mittelwertvergleiche erfolgt nach der generellen Formel für Konfidenzintervalle:

$$\text{c.i.}(\mu_1 - \mu_2) = (\bar{y}_1 - \bar{y}_2) \pm \hat{\sigma}(\bar{y}_1 - \bar{y}_2) \cdot Q_{1-\alpha/2}$$

In Abhängigkeit von den Ergebnissen der drei Vorentscheidungen wird für das $(1-\alpha/2)$-Konfidenzintervall entweder der Standardfehler für abhängige Stichproben berechnet, oder bei unabhängigen Stichproben der Standardfehler für gleiche oder aber ungleiche Varianzen. Als Verteilung zur Berechnung des Quantils $Q_{1-\alpha/2}$ wird entweder die Standardnormalverteilung oder die T-Verteilung herangezogen, in der Regel die T-Verteilung. Bei unabhängigen Stichproben ist $df=n_1+n_2-2$ und bei abhängigen Stichproben $df=n-1$.

Hypothesentests über Mittelwertvergleiche

Bei Hypothesentests werden in der Regel folgende Hypothesen getestet:
(a) H_0: $\mu_1 - \mu_2 = \mu$ vs. H_1: $\mu_1 - \mu_2 \neq \mu$ oder
(b) H_0: $\mu_1 - \mu_2 \geq \mu$ vs. H_1: $\mu_1 - \mu_2 < \mu$ oder
(c) H_0: $\mu_1 - \mu_2 \leq \mu$ vs. H_1: $\mu_1 - \mu_2 > \mu$.
Die Teststatistik ist in allen drei Situationen:

$$Z = \frac{\bar{y}_1 - \bar{y}_2 - \mu}{\hat{\sigma}(\bar{y}_1 - \bar{y}_2)}$$

Anwendungen von Mittelwertvergleichen

Der Standardfehler unterscheiden sich in Abhängigkeit von den drei Entscheidungen:

$$\hat{\sigma}(\bar{y}_1 - \bar{\sigma}_2) = \sqrt{\frac{\hat{\sigma}_1^2 + \hat{\sigma}_1^2 - 2 \cdot \hat{\sigma}_{21}}{n}} = \sqrt{\frac{s_1^2 + s_1^2 - 2 \cdot s_{21}}{n-1}}$$

$$\hat{\sigma}(\bar{y}_1 - \bar{\sigma}_2) = \sqrt{\frac{(n_1-1)\cdot\hat{\sigma}_1^2 + (n_2-1)\cdot\hat{\sigma}_2^2}{n_1+n_2-2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{\frac{n_1 \cdot s_1^2 + n_2 \cdot s_2^2}{n_1+n_2-2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\hat{\sigma}(\bar{y}_1 - \bar{\sigma}_2) = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} = \sqrt{\frac{s_1^2}{n_1-1} + \frac{s_2^2}{n_2-1}}$$

Die erste Formel gilt für abhängige Stichproben, die zweite für unabhängige Stichproben und gleiche Varianzen und die letzte für unabhängige Stichproben und verschiedene Varianzen.

Die Nullhypothese wird jeweils abgelehnt, wenn
bei (a) H_0: $\mu_1 - \mu_2 = \mu$ die Teststatistik $Z \leq Q_{\alpha/2}$ oder $Z \geq Q_{1-\alpha/2}$,
bei (b) H_0: $\mu_1 - \mu_2 \geq \mu$ die Teststatistik $Z \leq Q_{\alpha/2}$, bzw.
bei (c) H_0: $\mu_1 - \mu_2 \leq \mu$ die Teststatistik $Z \geq Q_{1-\alpha/2}$.
Die Quantile beziehen sich wie bei Konfidenzintervallen entweder auf die Standardnormalverteilung oder auf die T-Verteilung mit $df=n_1+n_2-2$ bei unabhängigen und $df=n-1$ bei abhängigen Stichproben .

Anwendungsbeispiele

Als Anwendungsbeispiel soll anhand der Daten des Allbus 2006 geprüft werden, ob vollzeitbeschäftigte Männer ein höheres Einkommen haben als Frauen. Aus den Allbus-Daten wurden mit SPSS folgende Ausgangsstatistiken des Nettoeinkommens berechnet:

Gruppe	Männer	Frauen
Fallzahl:	614	330
Mittelwert:	1755.482	1365.664
Standardabweichung:	966.411	637.989

(Quelle: Allbus 2006
Nur ganztags beschäftige Befragte)

Schritt 1: Formulierung von Null- und Altenativhypothese
Wenn das Einkommen der Männer als Y_1 und das der Frauen als Y_2 bezeichnet wird, ergibt sich aus der Forschungshypothese, dass Männer ein höheres Einkommen als Frauen haben, ein einseitiger Test nach oben mit der postulierten Mittelwertdifferenz $\mu > 0$:

$$H_0: \mu_1 - \mu_2 \leq 0 \text{ versus } H_1: \mu_1 - \mu_2 > 0:$$

Schritt 2: Auswahl von Teststatistik und Kennwerteverteilung
Im zweiten Schritt ist zunächst zu prüfen, ob die Daten aus einer abhängigen oder einer unabhängigen Stichprobe kommen und bei unabhängigen Daten, ob die Populationsvarinazen als gleich oder ungleich zu betrachten sind.

Anwendungsbeispiele

Da in der Allbus-Stichprobe nur eine Person in jedem der unabhängig gezogenen Haushalte befragt wurde und keine Paare, ist das Geschlecht der Befragten eine Gruppierungsvariable, so dass es sich um unabhängige Stichproben handelt.

Werden die berechneten Standardabweichungen quadriert, ergeben sich Varianzen von 933950.221 und 409585.920. Sowohl diese Werte wie auch die Fallzahlen sind sehr unterschiedliche, so dass der Standardfehler für ungleiche Varianzen berechnet wird.

Hinweise:

- *Statistikprogramme führen bei einem Mittelwertvergleich bei unabhängigen Stichproben meist zusätzlich einen Test auf gleiche bzw. verschiedene Varianzen durch. Wird die Entscheidung über gleiche bzw. verschiedene Varianzen vom Testergebnis abhängig gemacht, sind die nachfolgenden Signifikanzberechnungen des Mittelwerttests streng genommen nicht mehr gültig, da es sich formal bei letzteren um einen abhängigen Test handelt. Dies gilt immer dann, wenn erst anhand der Stichprobendaten entschieden wird, ob der Test von gleichen oder verschiedenen Varianzen ausgehen soll.*
- *Basiert die Berechnung des Tests wie im Beispiel auf vorgegebenen Fallzahlen, Mittelwerten und Standardabweichungen bzw. Varianzen ist darauf zu achten, ob es sich bei Standardabweichungen und Varianzen um Stichprobenkennwerte oder Schätzungen von Populationsparametern handelt. Statistikprogramme berechnen in der Regel das letztere.*

Anwendungsbeispiele

Die für die Allbus-Daten herangezogene Teststatistik geht von verschiedenen Varianzen aus. Bei den berichteten Standardabweichungen handelt es sich um Schätzungen der Populationswerte. Die Teststatistik berechnet sich daher nach:

$$Z \text{ bzw. } T = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^1}{n_2}}}$$

Zu überlegen ist noch, ob die T- oder die Standardnormalverteilung herangezogen werden soll.

Hinweis:
Statistikprogramme verwenden für den Mittelwertvergleich ausschließlich die T-Verteilung.

Im Beispiel berechnen sich dann die Freiheitsgrade nach:

$$df = \frac{1}{\left(\frac{966.411^2/614}{966.411^2/614 + 637.989^2/330}\right)^2/613 + \left(\frac{637.989^2/330}{966.411^2/614 + 637.989^2/330}\right)^2/329} = 903.4$$

Aufgrund der hohen Zahl von Freiheitsgraden macht es praktisch keinen Unterschied, ob im Beispiel die T-Verteilung oder die Standardnormalverteilung herangezogen wird.

Anwendungsbeispiele

Schritt 3: Festlegung von Irrtumswahrscheinlichkeiten und kritischen Werten
Die kritischen Werte ergeben sich aus der Irrtumswahrscheinlichkeit und der Formulierung von Null- und Alternativhypothese.

Im Beispiel wird eine einseitige Hypothese nach oben geprüft. Bei einer Irrtumswahrscheinlichkeit von 5% wird die Nullhypothese daher abgelehnt, wenn die Teststatistik größer oder gleich dem 95%-Quantil der Standardnormalverteilung bzw. der T-Verteilung mit 903.4 Freiheitsgraden ist. Aus einer Z-Tabelle ist zu entnehmen, dass der kritische Wert 1.645 beträgt. Bei der T-Verteilung ergibt sich der praktisch gleich große Wert 1.646.

Schritt 4: Berechnung der Teststatistik und Entscheidung
Die Berechnung bei den Beispieldaten ergibt:

Gruppe	Männer	Frauen
Fallzahl:	614	330
Mittelwert:	1755.482	1365.664
Standardabweichung:	966.411	637.989

(Quelle: Allbus 2006
Nur ganztags beschäftige Befragte)

$$Z = \frac{1755.482 - 1365.664}{\sqrt{\frac{966.411^2}{614} + \frac{637.989^2}{330}}} = 7.4$$

Da der Wert 7.4 größer ist als der kritische Wert 1.645 bzw. bei einer T-Verteilung als 1.646, ist die Nullhypothese bei einer Irrtumswahrscheinlichkeit von 5% abzulehnen. Es ist daher damit zu rechnen, dass das Nettoeinkommen bei vollzeitbeschäftigten Männern höher ist als bei vollzeitbeschäftigten Frauen.

Die Daten weisen also auf eine Geschlechtsdiskriminierung von Frauen hin. Da nur vollzeitbeschäftigte Personen berücksichtigt wurden, ist das geringere Einkommen der Frauen kein Ergebnis geringerer Arbeitszeit. Nicht berücksichtig sind allerdings Beruf, Beschäftigungsdauer und da es sich um Nettoeinkommen handelt, die Steuerklasse der befragten Personen.

Würde man unterstellen, dass die Populationsvarianzen in den beiden Gruppen gleich groß wären, würde eine andere Teststatistik berechnet:

Gruppe	Männer	Frauen
Fallzahl:	614	330
Mittelwert:	1755.482	1365.664
Standardabweichung:	966.411	637.989

(Quelle: Allbus 2006
Nur ganztags beschäftige Befragte)

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)\cdot \hat{\sigma}_1^2 + (n_2 - 1)\cdot \hat{\sigma}_1^2}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{1755.482 - 1365.664}{\sqrt{\frac{613 \cdot 966.411^2 + 329 \cdot 637.989^2}{942} \cdot \left(\frac{1}{614} + \frac{1}{330}\right)}} = 6.6$$

Die Entscheidung würde sich nicht ändern, Da 6.6 größer ist als der kritische Wert 1.645 des 95%-Quantils der Standardnormalverteilung bzw. des Wertes 1.646 bei einer T-Verteilung mit 942 Freiheitsgraden, ist die Nullhypothese abzulehnen.

Schritt 5: Kontrolle der Anwendungsvoraussetzungen

Der Z-Test der Mittelwertdifferenz basiert auf der Anwendung des zentralen Grenzwertsatzes, nachdem die Summen unabhängiger identisch verteilter Zufallsvariablen asymptotisch normalverteilt sind.

Die Annahme unabhängiger identisch verteilter Zufallsvariablen ist bei einfachen Zufallsauswahlen erfüllt. Bei komplexen Auswahlen ist es allerdings denkbar, dass sich die Populationsmittelwerte zwischen den durch Schichten oder Cluster definierten Subpopulationen unterscheiden. Dies kann dazu führen, dass die für einfache Zufallsauswahlen gültigen Standardfehler nicht korrekt sind.

Die Annäherung an die Normalverteilung hängt von der Verteilung in der Population ab. Als Faustregel gilt, dass ab einer Fallzahl von etwa 30 die Annäherung praktisch immer hinreichend genau ist. Für den Mittelwertvergleich bei unabhängigen Stichproben und verschiedenen Varianzen bedeutet dies, dass vorausgesetzt wird:

$$n_1 \geq 30 \text{ und } n_2 \geq 30.$$

Im Beispiel ist diese Bedingung erfüllt.

Wenn unterstellt werden kann, dass die Populationsvarianzen in den Gruppen gleich groß sind, dann ist die Annäherung leichter zu erfüllen, da hier nur die Fallzahlsumme der beiden Stichproben mindestens 30, besser mindestens 50 Fälle umfassen sollte:

bei gleichen Varianzen: $n_1 + n_2 \geq 30$

Anwendungsbeispiele

In einem weiteren Beispiel soll die Forschungshypothese geprüft werden, dass bei zusammenlebenden Ehepartnern bzw. bei festen Partnern die Frau mehr als zwei Jahre jünger ist als der Mann.

Schritt 1: Formulierung von Null- und Altenativhypothese
Wenn das Alter der Frauen als X_1 und das der Männer als X_2 bezeichnet wird, folgt aus der Forschungshypothese ein einseitiger Test nach unten mit der postulierten Mittelwertdifferenz $\mu=2$:

$$H_0: \mu_1 - \mu_2 \geq -2 \text{ versus } H_0: \mu_1 - \mu_2 < -2:$$

Schritt 2: Auswahl von Teststatistik und Kennwerteverteilung
Da das Alter von Paaren verglichen werden soll, handelt es sich um einen Mittelwertvergleich bei abhängiger Stichprobe.

Die Teststatistik ist dann:

$$Z \text{ bzw. } T = \frac{(\bar{y}_1 - \bar{y}_2) - (-2)}{\hat{\sigma}(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{y}_1 - \bar{y}_2 + 2}{\sqrt{\frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2 \cdot \hat{\sigma}_{21}}{n}}}$$

Da der Allbus eine große Fallzahl aufweist ist es unerheblich, ob die Standardnormalverteilung oder eine T-verteilung als Kennwerteverteilung verwendet wird. Im Allbus 2006 liegen für n=2454 Befragte Altersangaben für den Befragten und dessen Partner bzw. Partnerin vor.

Anwendungsbeispiele

Schritt 3: Festlegung von Irrtumswahrscheinlichkeiten und kritischen Werten
Die kritischen Werte ergeben sich aus der Irrtumswahrscheinlichkeit und der Formulierung von Null- und Alternativhypothese.

Im Beispiel wird eine einseitige Hypothese nach unten geprüft. Bei einer Irrtumswahrscheinlichkeit von 5% wird die Nullhypothese daher abgelehnt, wenn die Teststatistik kleiner oder gleich dem 5%-Quantil der Standardnormalverteilung bzw. der T-Verteilung mit 2553 Freiheitsgraden ist. Aus einer Z-Tabelle ist zu entnehmen, dass der kritische Wert –1.645 beträgt. Bei der T-Verteilung ergibt sich bei 3 Nachkommastellen der gleiche Wert.

Schritt 4: Berechnung der Teststatistik und Entscheidung
Die Berechnung bei den Beispieldaten ergibt:

	Varianzen u. Kovarianzen	
Frau	231.307	
Mann	223.850	237.226
Mittlwerte	48.268	51.018

(Quelle: Allbus 2006, n=2454)

$$Z = \frac{(48.268 - 51.018) - (-2)}{\sqrt{\frac{231.307 + 237.226 - 2 \cdot 223.850}{2454}}} = -8.1$$

Da die Teststatistik –8.1 kleiner ist als der kritische Wert –1.645, ist die Nullhypothese beim einseitigen Test nach unten abzulehnen. Es ist damit zu rechnen, dass Frauen im Mittel mehr als 2 Jahre jünger sind als ihre männlichen Partner.

Anwendungsbeispiele

Alternative Berechnung über Differenzvariable

Bei Vorliegen der Datenmatrix kann auch für jeden Fall der Stichprobe die Differenz aus dem Alter der Frau und der des Mannes berechnet werden. Daten und Teststatistik ergeben dann:

	Altersdifferenz
Fallzahl	2454
Mittelwert	–2.751
Varianz	20.832

(Quelle: Allbus 2006, eigene Berechnung von D)

$$Z = \frac{\bar{d} - \mu}{\sqrt{\frac{\hat{\sigma}_d^2}{n}}} = \frac{-2.751 - (-2)}{\sqrt{\frac{20.832}{2454}}} = -8.2$$

Bis auf Rundungsfehler ergeben sich identische Ergebnisse.
Die Nullhypothese wird wiederum mit einer Irrtumswahrscheinlichkeit von 5% abgelehnt.

Schritt 5: Kontrolle der Anwendungsvoraussetzungen

Aufgrund der hohen Fallzahl kann unabhängig von der Verteilung in der Population davon ausgegangen werden, dass die asymptotische Annäherung an die Normalverteilung gegeben ist.
Da der Allbus allerdings auf einer mehrstufigen Auswahl basiert, muss ohne empirische Prüfung angenommen werden, dass es keine Klumpeneffekte und keine Unterschiede zwischen den alten und neuen Bundesländern gibt.
Darüber hinaus muss unterstellt werden, dass es keine systematischen Ausfälle bei den Altersangaben gibt.

Beziehung zu Test auf Differenzen von Anteilen oder Prozentwerten

Anteile können stets als Mittelwerte von 0/1-kodierten dichotomen Variablen aufgefasst werden. Es liegt daher nahe, die Logik des Mittelwertvergleichs auch auf Anteils- bzw. Prozentsatzdifferenzen anzuwenden. Tatsächlich kann gezeigt werden, dass die bereits in Statistik 1 vorgestellten Tests Spezialfälle der Tests auf Gleichheit zweier Mittelwerte sind.

Zur Verdeutlichung wird wieder das Beispiel der Haltung zu Schwangerschaftsabbrüchen bei Mitgliedern von Religionsgemeinschaften und Konfessionslosen verwendet.

Abtreibung erlaubt? (Y)	Religionsgemeinsch. (X) nein (X=1)	ja (X=0)	Summe
nein (Y=1)	0.376 (415)	0.656 (1387)	0.560 (1802)
ja (Y=0)	0.624 (689)	0.334 (728)	0.440 (1417)
Summe	1.000 (1104)	1.000 (2115)	1.000 (3219)

(Daten: Allbus 2006)

Wie bei der Gruppierungsvariable X (im Beispiel: Mitglied einer Religionsgemeinschaft) ist nun auch die abhängige Variable eine Indikatorvariable, deren Mittelwert und Varianz eine einfache Funktion der Fallzahlen bzw. relativen Häufigkeiten in den beiden Ausprägungen ist:

$$\bar{y}_1 = \frac{1}{n_1} \cdot \sum_{i=1}^{n_1} y_{1,i} = \frac{415}{1104} = 0.376 = p_{1(1)}$$

$$s_1^2 = \frac{1}{n_1} \cdot \sum_{i=1}^{n_1} y_{i,1}^2 - \bar{y}_1^2 = p_{1(1)} - p_{1(1)}^2 = p_{1(1)} \cdot \left(1 - p_{1(1)}\right) = p_{1(1)} \cdot p_{2(1)} = \frac{415 \cdot 689}{1104^2} = 0.2346$$

Beziehung zu Test auf Differenzen von Anteilen oder Prozentwerten

Abtreibung erlaubt? (Y)	Religionsgemeinsch. (X) nein (X=1)	ja (X=0)	Summe
nein (Y=1)	0.376 (415)	0.656 (1387)	0.560 (1802)
ja (Y=0)	0.624 (689)	0.334 (728)	0.440 (1417)
Summe	1.000 (1104)	1.000 (2115)	1.000 (3219)

(Daten: Allbus 2006)

$$\bar{y}_1 = \frac{415}{1104} = 0.376 = p_{1(1)}$$

$$s_1^2 = \frac{415 \cdot 689}{1104^2} = 0.2346$$

$$\bar{y}_2 = \frac{1}{n_2} \cdot \sum_{i=1}^{n_2} y_{2,i} = \frac{1387}{2115} = 0.656 = p_{1(2)}$$

$$s_2^2 = \frac{1}{n_2} \cdot \sum_{i=1}^{n_2} y_{2,1}^2 - \bar{y}_2^2 = p_{1(2)} - p_{1(2)}^2 = p_{1(2)} \cdot \left(1 - p_{1(2)}\right) = p_{1(2)} \cdot p_{2(2)} = \frac{1387 \cdot 728}{2115^2} = 0.2257$$

Eine Besonderheit einer dichotomen Variable ist, dass es nur einen Verteilungsparameter gibt und daher die Varianz der Verteilung eine Funktion des Mittelwertes (bzw. bei Zufallsvariablen des Erwartungswertes) ist. Dies hat die Konsequenz, dass in einfachen Zufallsauswahlen bereits die Stichprobenvarianz ein konsistenter und erwartungstreuer Schätzer der Populationsvarianz ist und somit die Variation durch n und nicht durch n–1 geteilt wird:

$$\hat{\sigma}_1^2 = s_1^2 = p_{1(1)} \cdot \left(1 - p_{1(1)}\right) = p_{1(1)} \cdot p_{2(1)} \; ; \; \hat{\sigma}_2^2 = s_2^2 = p_{1(2)} \cdot \left(1 - p_{1(2)}\right) = p_{1(2)} \cdot p_{2(2)}$$

Weiter folgt auch, dass die beiden Populationsvarianzen σ^2_1 und σ^2_2 nur dann gleich sein können, wenn die Populationsmittelwerte π_1 und π_2 gleich sind.

Beziehung zu Test auf Differenzen von Anteilen oder Prozentwerten

Dann ergibt sich für den geschätzten Standardfehler von Anteilsdifferenzen bei vermutlich unterschiedlichen Populationsanteilen oder Prozentwerten und unabhängigen Stichproben:

$$\hat{\sigma}\left(p_{1(1)} - p_{1(2)}\right) = \sqrt{\frac{p_{1(1)} \cdot p_{2(1)}}{n_{+1}} + \frac{p_{1(2)} \cdot p_{2(2)}}{n_{+2}}} = \sqrt{\frac{a \cdot c}{(a+c)^3} + \frac{b \cdot d}{(b+d)^3}}$$

Bei durch die Nullhypothese postulierten gleichen Populationsanteilen wird dagegen die gemeinsame (pooled) Schätzung der Populationsvarianz verwendet. Hierbei ist zu berücksichtigen, dass für die gemeinsame Schätzung der Varianz aufgrund der Beziehung von Mittelwert und Varianz bei dichotomen Variablen die gemeinsame Schätzung der Varianz über die gemeinsame Schätzung der Populationsanteile $\pi_1 = \pi_2$ in den beiden Gruppen erfolgt:

$$\hat{\sigma}\left(p_{1(1)} - p_{1(2)}\right) = \sqrt{\frac{\left(n_1 \cdot p_{1(1)} + n_2 \cdot p_{1(2)}\right)}{n_1 + n_2} \cdot \frac{\left(n_1 \cdot p_{2(1)} + n_2 \cdot p_{2(2)}\right)}{n_1 + n_2}}{n} \cdot \left(\frac{1}{n_{+1}} + \frac{1}{n_{+2}}\right)$$

$$= \sqrt{\frac{(a+b) \cdot (c+d)}{n^2} \cdot \left(\frac{1}{(a+c)} + \frac{1}{(b+d)}\right)}$$

Setzt man diese Standardfehler in die Teststatistik für Mittelwertvergleiche bei unabhängigen Stichproben ein, ergeben sich die aus Statistik 1 bekannten Teststatistiken für Tests von Prozentsatzdifferenzen bzw. Anteilsdifferenzen.

Test von Anteilsdifferenzen bei abhängigen Stichproben

Auch die Logik des Tests von Mittelwertvergleichen bei abhängigen Stichproben lässt sich auf dichotomen Variablen anwenden. Dabei muss wiederum wieder die Kovarianz zwischen den beiden (dichotomen) Variablen berücksichtigt werden.

Als Beispiel soll die Forschungshypothese geprüft werden, dass der Anteil derjenigen, die einen Schwangerschaftsabbruch befürworten größer ist wenn das Kind vermutlich behindert ist als bei einem Abbruch aus einer finanziellen Notlage.

Schwangerschaftsabbruch ... bei behindertem Kind (Y_1)	... bei finanzieller Notlage sollte (Y_2) ... erlaubt sein	... verboten sein	insgesamt
... solte erlaubt sein	48.4% (a = 1523)	41.3% (b = 1301)	89.7% (2824)
... sollte verboten sein	1.6% (c = 51)	8.7% (d = 374)	10.3% (325)
insgesamt	50.0% (1574)	50.0% (1575)	100.0% (3149)

Wird bei beiden Variablen die Ausprägung „erlaubt“ als 1 und „verboten“ als 0 kodiert, berechnen sich die Stichprobenvarianzen und die Kovarianz nach:

Zelle	Y_1	Y_2	n	$n \cdot Y_1$	$n \cdot Y_2$	$n \cdot Y_1 \cdot Y_2$
a	1	1	1523	1523	1523	1523
b	0	1	1301	0	1301	0
c	1	0	51	51	0	0
d	0	0	374	0	0	0
Summe			3149	1574	2824	1523

$$\bar{y}_1 = p_{1+} = 1574 / 3149 = 0.4998$$
$$\bar{y}_2 = p_{+1} = 2824 / 3149 = 0.8968$$
$$s_1^2 = p_{1+} \cdot (1 - p_{1+}) = 2824 \cdot 325 / 3149^2 = 0.0926$$
$$s_2^2 = p_{+1} \cdot (1 - p_{+1}) = 1574 \cdot 1575 / 3149^2 = 0.2500$$
$$s_{21} = p_{11} - p_{1+} \cdot p_{+1}$$
$$= 1523 / 3149 - 2824 \cdot 1574 / 3149^2 = 0.0354$$

Test von Anteilsdifferenzen bei abhängigen Stichproben

Schwangerschaftsabbruch ... bei behindertem Kind (Y_1)	... bei finanzieller Notlage sollte (Y_2) ... erlaubt sein	... verboten sein	insgesamt
... solte erlaubt sein	48.4% (a = 1523)	41.3% (b = 1301)	89.7% (2824)
... sollte verboten sein	1.6% (c = 51)	8.7% (d = 374)	10.3% (325)
insgesamt	50.0% (1574)	50.0% (1575)	100.0% (3149)

Die Teststatistik ist dann die Anteilsdifferenz minus der nach der Nullhypothese postulierten Differenz geteilt durch den Standardfehler, der sich wie beim Mittelwertvergleich bei abhängigen Stichprobenberechnet, wobei allerdings wieder durch n und nicht durch n–1 geteilt wird:

$$Z = \frac{(p_1 - p_2) - \pi}{\hat{\sigma}(p_1 - p_2)} = \frac{(p_{1+} - p_{+1}) - \pi}{\sqrt{\hat{\sigma}^2(p_{1+}) + \hat{\sigma}^2(p_{+1}) - 2 \cdot \hat{\sigma}(p_{1+}, p_{+1})}}$$

$$= \frac{(p_1 - p_2) - \pi}{\sqrt{\frac{p_{1+} \cdot (1 - p_{1+}) + p_{+1} \cdot (1 - p_{+1}) - 2 \cdot p_{11} \cdot p_{1+} \cdot p_{+1}}{n}}} = \frac{0.897 - 0.500}{\sqrt{\frac{0.0926 + 0.2500 - 2 \cdot 0.0354}{3149}}} = 42.7$$

Bei einer Irrtumswahrscheinlichkeit von 5% ist die Nullhypothese im einseitigen Test nach oben abzulehnen, wenn Z größer /gleich dem 95%-Quantil der Standardnormalverteilung ist. Da 42.7 > 1.645 ist die Nullhypothese abzulehnen. Vermutlich ist auch in der Population die Zustimmung zum Schwangerschaftsabbruch größer, wenn das Kind behindert ist, als wenn eine finanzielle Notlage besteht.

Lerneinheit 7:
Drittvariablenkontrolle im linearen Regressionsmodell

In der Tabellenanalyse werden bei der Drittvariablenkontrolle für alle Ausprägungen der Drittvariablen Partialtabellen gebildet. Obwohl diese Vorgehensweise im Prinzip auch bei der linearen Regression angewendet werden kann, stößt sie schnell an ihre Grenzen, wenn die Drittvariablen metrisch sind und sehr viele Ausprägungen haben. Es gibt dann bei einer Ausprägung der Drittvariable möglicherweise nur einen Fall oder ganz wenige Fälle für die Berechnung der konditionalen Regressionskoeffizienten, was zu sehr großen Standardfehlern führt oder auch die Berechnung unmöglich macht.

Als eine Alternative bietet es sich im linearen Regressionsmodell an, die Modellgleichung einfach um die Kontrollvariable W zum ***trivariaten Regressionsmodell*** zu erweitern:

$$\mu\left(Y \middle| X = x, W = w\right) = \beta_o + \beta_1 \cdot x + \beta_2 \cdot w \text{ bzw. } Y = \mu_{Y|X,W} + U = \beta_o + \beta_1 \cdot X + \beta_2 \cdot W + U$$

Im Unterschied zur bivariaten Regression mit den Koeffizienten α und β werden die Regressionskoeffizienten in der Population einheitlich durch β_k symbolisiert, wobei der Laufindex k bei 0 beginnt und somit β_0 für die Regressionskonstante steht.
In der allgemeinen Darstellung mit vielen Drittvariablen bzw. erklärenden Variablen werden zudem die erklärenden Variablen ebenfalls meist nicht durch unterschiedlichen Buchstaben (z.B. X, W oder Z) sondern durch Indizierung von X (also X_1, X_2, ..., X_K) symbolisiert:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_K \cdot X_K + U.$$

Schätzung der Regressionskoeffizienten

Für die Schätzung des Regressionskoeffizienten β_0, β_1 und β_2 gibt es unterschiedliche Metho-en. Bei der ***Momentenmethode*** werden die Regressionskoeffizienten als Funktion von Mittel-werten, Varianzen und Kovarianzen ausgedrückt und diese Momente über Stichprobendaten geschätzt. Um zu einer eindeutigen Lösung zu kommen, müssen zunächst zwei Annahmen ge-troffen werden, die analog auch beim bivariaten Regressionsmodell vorausgesetzt wurden:
(1) Die erklärenden Variablen X und W korrelieren nicht mit der Residualvariable U.
(2) Der Populationsmittelwert bzw. Erwartungswert von U ist Null.
Dann folgt aus den Regeln für Mittelwerte und Varianzen von Linearkombinationen:

$$\begin{aligned}
& Y = \beta_o + \beta_1 \cdot X + \beta_2 \cdot W + U \\
\Rightarrow\ & \mu_Y = \beta_o + \beta_1 \cdot \mu_X + \beta_2 \cdot \mu_W \\
\Rightarrow\ & \sigma_Y^2 = \beta_1^2 \cdot \sigma_X^2 + \beta_2^2 \cdot \sigma_W^2 + 2 \cdot \beta_1 \cdot \beta_2 \cdot \sigma_{WX} + \sigma_U^2 \\
\Rightarrow\ & \sigma_{YX} = \beta_1 \cdot \sigma_X^2 + \beta_2 \cdot \sigma_{WX} \\
\Rightarrow\ & \sigma_{YW} = \beta_1 \cdot \sigma_{WX} + \beta_W \cdot \sigma_2^2 \\
\Rightarrow\ & \sigma_{YU} = \sigma_U^2
\end{aligned}$$

Aus den beiden Gleichungen für die Kovarianzen zwischen der abhängigen Variable Y und X bzw. W lassen sich dann die Regressionsgewichte β_1 und β_2 als Funktionen der Varianzen und Kovarianzen berechnen.

Schätzung der Regressionskoeffizienten

$$\sigma_{YX} = \beta_1 \cdot \sigma_X^2 + \beta_2 \cdot \sigma_{WX} \text{ und } \sigma_{YW} = \beta_1 \cdot \sigma_{WX} + \beta_2 \cdot \sigma_W^2$$

$$\Rightarrow \beta_1 = \frac{\sigma_W^2 \cdot \sigma_{YX} - \sigma_{WX} \cdot \sigma_{YW}}{\sigma_W^2 \cdot \sigma_X^2 - (\sigma_{WX})^2} \text{ und } \beta_2 = \frac{\sigma_X^2 \cdot \sigma_{YW} - \sigma_{WX} \cdot \sigma_{YX}}{\sigma_X^2 \cdot \sigma_W^2 - (\sigma_{WX})^2}$$

Aus der Gleichung für den Erwartungswert von Y folgt für die Regressionskonstante:

$$\mu_Y = \beta_o + \beta_1 \cdot \mu_X + \beta_2 \cdot \mu_W \Rightarrow \beta_0 = \mu_Y - \beta_1 \cdot \mu_X - \beta_2 \cdot \mu_X$$

Für die Schätzung der Koeffizienten werden die konsistenten und erwartungstreuen Schätzer der Populationsmittelwerte bzw. Erwartungswerte und Varianzen und Kovarianzen genutzt:

$$\hat{\beta}_1 = \frac{\hat{\sigma}_W^2 \cdot \hat{\sigma}_{YX} - \hat{\sigma}_{WX} \cdot \hat{\sigma}_{YW}}{\hat{\sigma}_W^2 \cdot \hat{\sigma}_X^2 - (\hat{\sigma}_{WX})^2} = \frac{s_W^2 \cdot s_{YX} - s_{WX} \cdot s_{YW}}{s_W^2 \cdot s_X^2 - (s_{WX})^2} = \frac{SS_W \cdot SP_{YX} - SP_{WX} \cdot SP_{YW}}{SS_W \cdot SS_X - (SP_{WX})^2}$$

$$\hat{\beta}_2 = \frac{\hat{\sigma}_X^2 \cdot \hat{\sigma}_{YW} - \hat{\sigma}_{WX} \cdot \hat{\sigma}_{YX}}{\hat{\sigma}_X^2 \cdot \hat{\sigma}_W^2 - (\hat{\sigma}_{WX})^2} = \frac{s_X^2 \cdot s_{YW} - s_{WX} \cdot s_{YX}}{s_X^2 \cdot s_W^2 - (s_{WX})^2} = \frac{SS_X \cdot SP_{YW} - SP_{WX} \cdot SP_{YX}}{SS_X \cdot SS_W - (SP_{WX})^2}$$

$$\hat{\beta}_0 = \hat{\mu}_Y - \hat{\beta}_1 \cdot \hat{\mu}_X - \hat{\beta}_2 \cdot \hat{\mu}_W$$

Dass anstelle der geschätzten Populationsvarianzen und Kovarianzen auch die Stichprobenvarianzen und -kovarianzen bzw. die Variationen (SS_k)und Kovariationen (SP_{kj}) eingesetzt werden können, liegt daran, dass sich im Zähler und Nenner die Fallzahlen n bzw. Freiheitsgrade n–1 herauskürzen.

Logik der Drittvariablenkontrolle im linearen Regressionsmodell

Bekannter als die Momentenmethode ist die ***Kleinstquadratschätzung*** (**OLS-Methode** oder **OLS-Schätzung**), die bereits im bivariaten Regressionsmodell vorgestellt wurde. Hier wird zunächst analog zur Gleichung für die Population eine entsprechen Gleichung für die Stichprobe formuliert:

$$Y = \hat{Y} + E = b_0 + b_1 \cdot X + b_2 \cdot W + E$$

Die Regressionskoeffizienten b_0, b_1 und b_2 werden nach der Kleinstquadratmethode wie im bivariaten Modell so bestimmt, dass die Summe der quadrierten Stichprobenresiduen minimal ist:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 \cdot x_i + b_2 \cdot w_i))^2 = \min.$$

Die Kleinstquadratmethode führt zu folgende Berechnungsformeln für die drei Koeffizienten:

$$b_1 = \frac{SS_W \cdot SP_{XY} - SP_{WY} \cdot SP_{XW}}{SS_X \cdot SS_W - (SP_{XW})^2} = \frac{s_W^2 \cdot s_{XY} - s_{WY} \cdot s_{XW}}{s_X^2 \cdot s_W^2 - (s_{XW})^2} = \frac{\hat{\sigma}_W^2 \cdot \hat{\sigma}_{XY} - \hat{\sigma}_{WY} \cdot \hat{\sigma}_{XW}}{\hat{\sigma}_X^2 \cdot \hat{\sigma}_W^2 - (\hat{\sigma}_{XW})^2}$$

$$b_2 = \frac{SS_X \cdot SP_{WY} - SP_{XY} \cdot SP_{XW}}{SS_X \cdot SS_W - (SP_{XW})^2} = \frac{s_X^2 \cdot s_{WY} - s_{XY} \cdot s_{XW}}{s_X^2 \cdot s_W^2 - (s_{XW})^2} = \frac{\hat{\sigma}_X^2 \cdot \hat{\sigma}_{WY} - \hat{\sigma}_{XY} \cdot \hat{\sigma}_{XW}}{\hat{\sigma}_X^2 \cdot \hat{\sigma}_W^2 - (\hat{\sigma}_{XW})^2}$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x} - b_2 \cdot \bar{w}$$

Wie der Vergleich zeigt, sind die OLS-Schätzer mit den Schätzern nach der Momentenmethode identisch.

Trivariate Regression: OLS-Schätzer

W	X	Y	n	X^2	Y^2	X·Y	W^2	W·X	W·Y
0	1	4	1	1	16	4	0	0	0
0	2	3	1	4	9	6	0	0	0
0	3	3	1	9	9	9	0	0	0
0	4	5	1	16	25	20	0	0	0
1	2	1	1	4	1	2	1	2	1
1	3	0	1	9	0	0	1	3	0
1	4	1	1	16	1	4	1	4	1
1	5	3	1	25	9	15	1	5	3
∑ 4	24	20	8	84	70	60	4	14	5

Als Beispiel soll für 8 fiktive Fälle der Zusammenhang zwischen der abhängigen Variablen „Ablehnung von Schwangerschaftsabbrüchen“ (Y) und den beiden erklärenden Variablen „Religiosität“ (X) und „Region“ (W) mit den beiden Ausprägungen alte Bundesländer (W=0) und neue Bundesländer (W=1) untersucht werden.

Durch Aufsummieren werden zunächst die Mittelwerte, Variationen und Kovariationen berechnet:

$$\bar{x} = 24/8 = 3 \; ; \; \bar{y} = 20/8 = 2.5 \; ; \; \bar{w} = 4/8 = 0.5$$

$$s_X^2 = 84/8 - 3^2 = 1.5 \; ; \; s_Y^2 = 70/8 - 2.5^2 = 2.5 \; ; \; s_W^2 = 4/8 - 0.5^2 = 0.25$$

$$s_{YX} = 60/8 - 2.5 \cdot 3 = 0 \; ; \; s_{YW} = 5/8 - 0.5 \cdot 2.5 = -0.625 \; ; \; s_{WX} = 14/8 - 0.5 \cdot 3 = 0.25$$

Trivariate Regression: Interpretation der Regresionskoeffizienten

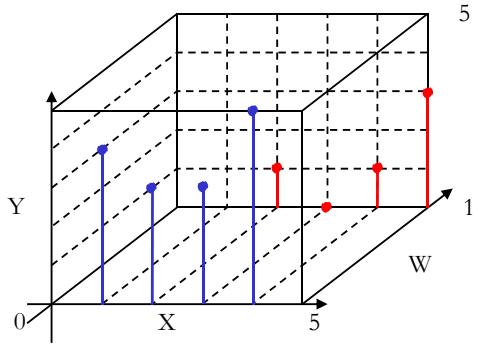
Die Regressionskoeffizienten sind dann:

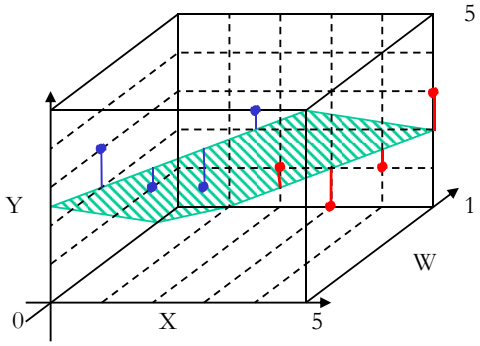
$$b_1 = \frac{s_W^2 \cdot s_{XY} - s_{WY} \cdot s_{XW}}{s_X^2 \cdot s_W^2 - (s_{XW})^2} = \frac{0.25 \cdot 0 - (-0.625) \cdot 0.25}{1.5 \cdot 0.25 - 0.25^2} = 0.5$$

$$b_2 = \frac{s_X^2 \cdot s_{WY} - s_{XY} \cdot s_{XW}}{s_X^2 \cdot s_W^2 - (s_{XW})^2} = \frac{1.5 \cdot (-0.625) - 0 \cdot 0.25}{1.5 \cdot 0.25 - 0.25^2} = -3.0$$

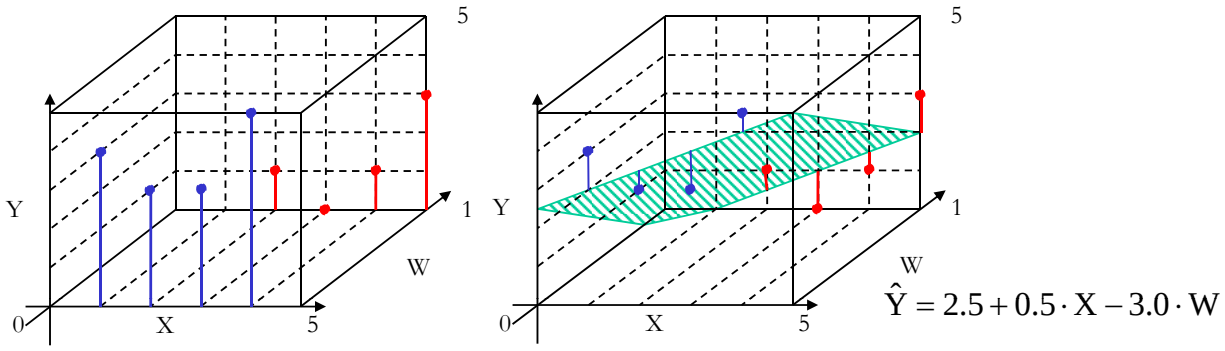
$$b_0 = \bar{y} - b_1 \cdot \bar{x} - b_2 \cdot \bar{w} = 2.5 - 0.5 \cdot 3 - (-3) \cdot 0.5 = 2.5$$

In der grafischen Darstellung ergibt sich eine dreidimensionale Punktewolke, durch die eine zweidimensionale Regressionsebene verläuft. Die drei Regressionskoeffizienten bestimmen die Lage dieser Ebene. Im Beispiel ergibt sich eine nach links unten und nach hinten unten gekippte Ebene.





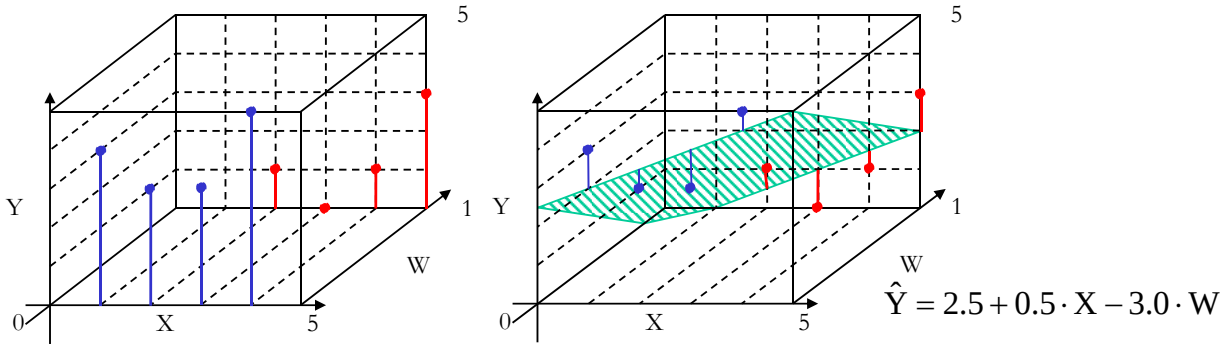
Trivariate Regression: Interpretation der Regresionskoeffizienten



Die ***Regressionskonstante*** $\mathbf{b_0}$ ist der Schnittpunkt dieser Ebene mit der Y-Achse an der Stelle X=0 und W=0. Die Interpretation des Koeffizienten entspricht der Regressionskonstante a in der bivariaten Regression.
Die Konstante gibt im Bespiel den Vorhersagewert an, wenn X=0 und W=0. Für die 8 fiktiven Fälle ergibt sich ein Wert von b_0=2.5.
Da X=0 für keine Religiosität steht, ist bei nicht religiösen Befragten (X=0) aus den alten Ländern (W=0) mit einem Durchschnittswert von 2.5 auf der Skala der Ablehnung von Schwangerschaftsabbrüchen zu rechnen.

Das ***Regressionsgewicht*** $\mathbf{b_1}$ gibt die Richtung und Stärke der Neigung entlang der X-Achse an. Da die Ebene nach links unten bzw. rechts oben gekippt ist, steigen die Vorhersagewerte mit steigenden Werten der erklärenden Variablen X an. Steigt die Religiosität um eine Einheit, ***erhöht*** sich die zu erwartende durchschnittliche Ablehnung von Schwangerschaftsabbrüchen um b_1=0.5 Einheiten.

Trivariate Regression: Interpretation der Regresionskoeffizienten



Ganz analog ist das ***Regressionsgewicht*** $\mathbf{b_2}$ der Drittvariable W zu interpretieren. Die Regres-sionsebene fällt nach hinten, d.h. mit steigenden Werten von W ab. Steigt also W um eine Ein-heit an, ist mit einer durchschnittlichen ***Verringerung*** der durchschnittlichen Ablehnung von Schwangerschaftsabbrüchen um b_2=–3 Einheiten zu rechnen.

Da inhaltlich ein Anstieg um eine Einheit bei der ersten erklärenden Variable X eine Zunahme der Religiosität bedeutet und ein Anstieg um eine Einheit bei der zweiten erklärenden Variable W einen Wechsel von den alten zu den neuen Bundesländern, besagen die beiden Regressions-gewichte, dass mit steigender Religiosität die Ablehnung von Schwangerschaftsabbrüchen zunimmt und dass in den neuen Bundesländern eine geringere Ablehnung von Schwanger-schaftsabbrüchen besteht als in den alten Ländern, die dortigen Befragten also bei dieser Frage liberaler sind als im regideren Westen.

Trivariate Regression: Partielle Effekte

Da alle Vorhersagewerte auf einer Ebene liegen, erfolgt der Anstieg bei Y um b_1=0.5 Einheiten, wenn sich X um eine Einheit erhöht. Die gilt für alle Werten der zweiten erklärenden Variable W.

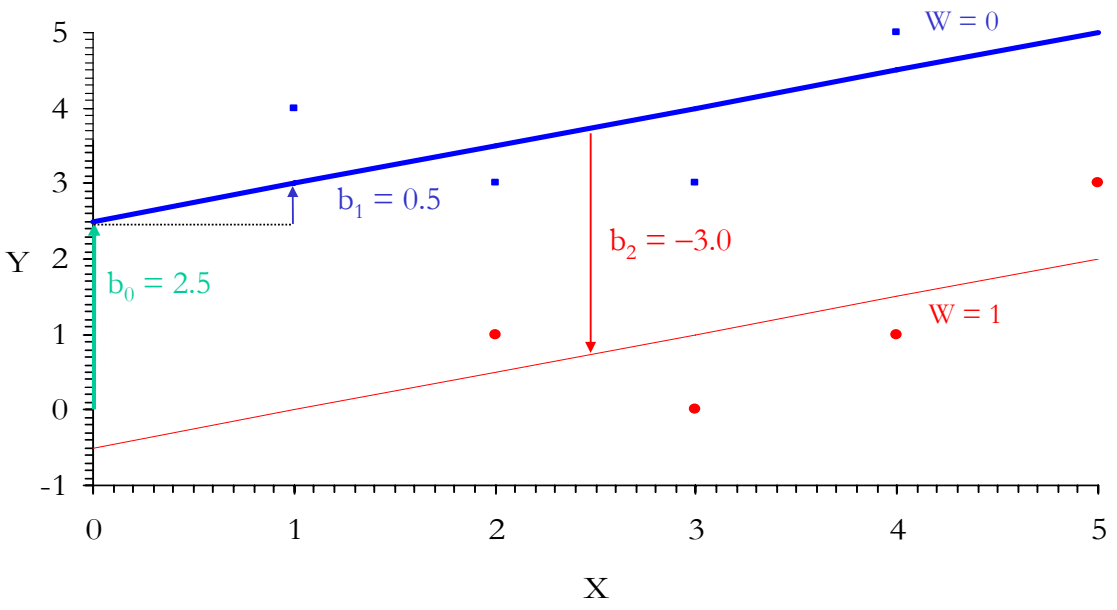
Analog wirkt sich ein Anstieg der zweiten erklärenden Variable W um eine Einheit bei allen Ausprägungen der ersten erklärenden Variable X in gleichem Maße als ein Rückgang der Vor-hersagewerte um b_2=–3 Einheiten aus.

Im Unterschied zur Drittvariablenkontrolle in der Tabellenanalyse, bei der für jede Ausprägung einer Kontrollvariable konditionale Beziehungen in Partialtabellen geschätzt werden, wird in der multiplen Regression daher nicht zwischen erklärenden Variablen und Kontrollvariablen differenziert.

In der trivariaten Regression gibt es entsprechend nicht nur eine, sondern zwei erklärende Variablen, die sich gegenseitig kontrollieren. Jede erklärende Variable ist gleichzeitig Kontrollvariable der anderen erklärenden Variable.

Da das Regressionsgewicht b_1 unabhängig von den Werten der zweiten erklärenden Variablen W gilt und analog das Regressionsgewicht b_2 unabhängig für alle Werte der ersten erklärenden Variablen X gilt, spricht man in der trivariaten (und multiplen) Regression von ***partiellen Effekten*** bzw. ***partiellen Regressionskoeffizienten***.

Trivariate Regression: Partielle Effekte



Die dreidimensionale Punktewolke kann in das zweidimensionale X,Y-Koordinatensystem projeziert werden. Die Abbildung zeigt dann für jede Ausprägung von W parallele Regressions-graden.

Der senkrechte Abstand zwischen den Geraden ist gerade gleich dem Wert des Regressionsgewichts b_2.

Symmetrie zwischen erklärender Variable und Kontrollvariable

Gleiches gilt auch für die bedingte Regression von Y auf W bei verschiedenen Werten von X: Die Regressionsgewichte sind stets gleich, die Regressionskonstanten unterscheiden sich jeweils um den Wert b_1=0.5, wenn X um eine Einheit zunimmt:

Vorhersagewerte: $\hat{Y} = b_0 + b_1 \cdot X + b_2 \cdot W = 2.5 + 0.5 \cdot X - 3.0 \cdot W$

Wenn X=1: $\hat{Y} = b_0 + b_1 \cdot 1 + b_2 \cdot W = 2.5 + 0.5 \cdot 1 - 3.0 \cdot W = 3.0 - 3.0 \cdot W$

Wenn X=2: $\hat{Y} = b_0 + b_1 \cdot 2 + b_2 \cdot W = 2.5 + 0.5 \cdot 2 - 3.0 \cdot W = 3.5 - 3.0 \cdot W$

Wenn X=3: $\hat{Y} = b_0 + b_1 \cdot 3 + b_2 \cdot W = 2.5 + 0.5 \cdot 3 - 3.0 \cdot W = 4.0 - 3.0 \cdot W$

Wenn X=4: $\hat{Y} = b_0 + b_1 \cdot 4 + b_2 \cdot W = 2.5 + 0.5 \cdot 4 - 3.0 \cdot W = 4.5 - 3.0 \cdot W$

Wenn X=5 $\hat{Y} = b_0 + b_1 \cdot 5 + b_2 \cdot W = 2.5 + 0.5 \cdot 5 - 3.0 \cdot W = 5.0 - 3.0 \cdot W$

Generell gilt: Wenn sich X um ΔX ändert, ändert sich der Vorhersagewert um $b_1 \cdot \Delta X$, wenn sich W um ΔW ändert, ändert sich der Vorhersagewert um $b_2 \cdot \Delta W$.

$$\Delta\hat{Y} = b_1 \cdot \Delta X \quad \text{und} \quad \Delta\hat{Y} = b_2 \cdot \Delta W$$

Man bezeichnet diese Eigenschaft der Effekte auch als ***linear-additiv***, weil jede erklärende Variable einen eigenen linearen Effekt aufweist und sich bezogen auf die Vorhersagewerte alle linearen Effekte addieren, also keine Interaktion besteht.

Trivariate Regression als Residuenregression:

Neben der Momentenmethode und der Kleinstquadratmethode gibt es noch eine dritte Methode, um die partiellen Regressionsgewichte zu bestimmen.
Die Idee der Methode der ***Residuenregression*** besteht darin, bei der Drittvariablenkontrolle den (linearen) Effekt bzw. Zusammenhang zwischen der Drittvariable und abhängiger Variable sowie unabhängiger Variable aus den Daten „herauszurechnen".
Dies geschieht dadurch, dass in einem ersten Schritt die Residuen der Regressionen der abhängigen und der unabhängigen Variable in Regressionen auf die Drittvariable berechnet werden und anschießend die bivariate Regressionen zwischen diesen Residuen.

Zunächst wird also die bivariate Regression von Y auf W berechnet:

$$Y = a_Y + b_Y \cdot W + E_{Y.W},$$

wobei die resultierenden Residuen $E_{Y|W}$ als eine zusätzliche Variable in die Datenmatrix aufgenommen wird.

In gleicher Weise wird auch X auf W regrediert:

$$X = a_X + b_X \cdot W + E_{X.W}.$$

Schließlich werden die Residuen von Y auf die Residuen von X regrediert:

$$E_{Y.W} = b_{X.W} \cdot E_{X.W} + E_{YX.W}.$$

Die Regressionskonstante kann bei der Residuenregresssion ausgelassen werden, da die Mittelwerte der Residuen Null sind und daher auch die Konstante Null sein muss.

Trivariate Regression als Residuenregression:

	W	X	Y	n	X^2	Y^2	X·Y	W^2	W·X	W·Y
∑	4	24	20	8	84	70	60	4	14	5

$\bar{x} = 3 ; \bar{y} = 2.5 ; \bar{w} = 0.5 \quad s_X^2 = 1.5 ; s_Y^2 = 2.5 ; s_W^2 = 0.25 \quad s_{YX} = 0 ; s_{YW} = -0.625 ; s_{WX} = 0.25$

Für die acht fiktiven Fälle des Beispiels ergibt die bivariate Regression von Y auf W folgendes Ergebnis:

$$b = s_{YW} / s_W^2 = -0.625 / 0.25 = -2.5; \; a = \bar{y} - b \cdot \bar{w} = 2.5 - (-2.5) \cdot 0.5 = 3.75$$

$$\hat{Y} = 3.75 - 2.5 \cdot W \Rightarrow E_{Y.W} = Y - 3.75 - 2.5 \cdot W$$

Für die Regression von X auf W ergibt sich:

$$b = s_{WX} / s_W^2 = 0.25 / 0.25 = 1.0; \; a = \bar{x} - b \cdot \bar{w} = 3 - 1 \cdot 0.5 = 2.5$$

$$\hat{X} = 2.5 + 1 \cdot W \Rightarrow E_{X.W} = X - 2.5 - 1 \cdot W$$

Für die Varianzen und Kovarianzen von $E_{Y.W}$ und $E_{X.W}$ folgt dann aus den Regeln für Linearkombinationen:

$$s^2(E_{Y.W}) = s_Y^2 + (-2.5)^2 \cdot s_W^2 - 2 \cdot (-2.5) \cdot s_{YW} = 2.5 + (-2.5)^2 \cdot 0.25 - 2 \cdot (-2.5) \cdot (-0.625) = 0.9375$$

$$s^2(E_{X.W}) = s_X^2 + s_W^2 - 2 \cdot s_{WX} = 1.5 + 0.25 - 2 \cdot 0.25 = 1.25$$

$$s(E_{X.W}, E_{YW}) = s_{YX} - s_{YW} - 2.5 \cdot s_{WX} + 2.5 \cdot s_W^2 = 0 - (-0.625) - 2.5 \cdot 0.25 + 2.5 \cdot 0.25 = 0.625$$

Trivariate Regression als Residuenregression:

$\bar{x} = 3 ; \bar{y} = 2.5 ; \bar{w} = 0.5 \quad s_X^2 = 1.5 ; s_Y^2 = 2.5 ; s_W^2 = 0.25 \quad s_{YX} = 0 ; s_{YW} = -0.625 ; s_{WX} = 0.25$

Y	X	W	$E_{Y\|W}$=Y–(3.75–2.5·W)	$E_{X\|W}$=X–(2.5+W)	$E_{X\|W}^2$	$E_{Y\|W} \cdot E_{X\|W}$
4	1	0	4–3.75= 0.25	1–2.5= –1.5	2.25	–0.375
3	2	0	3–3.75= –0.75	2–2.5= –0.5	0.25	0.375
3	3	0	3–3.75= –0.75	3–2.5= 0.5	0.25	–0.375
5	4	0	5–3.75= 1.25	4–2.5= 1.5	2.25	1.875
1	2	1	1–1.25= –0.25	2–3.5= –1.5	2.25	0.375
0	3	1	0–1.25= –1.25	3–3.5= –0.5	0.25	0.625
1	4	1	1–1.25= –0.25	4–3.5= 0.5	0.25	–0.125
3	5	1	3–1.25= 1.75	5–3.5= 1.5	2.25	2.625
Summe:			0	0	10.00	5.000

Die Vorhersagegleichung der Residuenregression der Residuen von Y auf die Residuen von X ist dann:

$$b = s_{Y.W,X.W} / s_{X.W}^2 = 1.25 / 0.625 = 5 / 10 = 0.5$$

Analog kann die Regression der Residuen der Regression von Y auf X auf die Residuen der Regression von W auf X berechnet werden:

$$b = s_{YX} / s_X^2 = 0 / 1.25 = 0; \; a = \bar{y} - b \cdot \bar{x} = 2.5 - 0 \cdot 3 = 2.5 \Rightarrow E_{Y.X} = Y - 2.5$$

$$b = s_{WX} / s_X^2 = 0.25 / 1.5 = 1 / 6; \; a = \bar{w} - b \cdot \bar{x} = 0.5 - 1 / 6 \cdot 3 = 0 \Rightarrow E_{W.X} = W - 1 / 6 \cdot W$$

Trivariate Regression : Auspartialisierung

$\bar{x} = 3$; $\bar{y} = 2.5$; $\bar{w} = 0.5$ $\quad s_X^2 = 1.5$; $s_Y^2 = 2.5$; $s_W^2 = 0.25$ $\quad s_{YX} = 0$; $s_{YW} = -0.625$; $s_{WX} = 0.25$

Y	X	W	$E_{Y\|X} = Y - 2.5$	$E_{W\|X} = W - 0.167 \cdot X$	$E_{W\|X}^2$	$E_{Y\|X} \cdot E_{W\|X}$
4	1	0	4–2.5 = 1.5	0–0.167 = –0.167	0.028	–0.250
3	2	0	3–2.5 = 0.5	0–0.333 = –0.333	0.111	–0.167
3	3	0	3–2.5 = 0.5	0–0.500 = –0.500	0.250	–0.250
5	4	0	5–2.5 = 2.5	0–0.667 = –0.667	0.444	–1.667
1	2	1	1–2.5 = –1.5	1–0.333 = 0.667	0.444	–1.000
0	3	1	0–2.5 = –2.5	1–0.500 = 0.500	0.250	–1.250
1	4	1	1–2.5 = –1.5	1–0.667 = 0.333	0.111	–0.500
3	5	1	3–2.5 = 0.5	1–0.833 = 0.167	0.028	0.083
Summe:			0	0	1.667	–5.000

Die Vorhersagegleichung der Residuenregression der Residuen von Y auf die Residuen von W ist dann:

$$b = s_{Y.X,W.X} / s_{W.X}^2 = \frac{s_{YW} - \frac{1}{6} \cdot s_{YX}}{s_W^2 + \left(\frac{1}{6}\right)^2 \cdot s_X^2 - \frac{2}{6} \cdot s_{WX}} = \frac{-0.625 - \frac{1}{6} \cdot 0}{0.25 + \frac{1.5}{36} - \frac{2}{6} \cdot 0.25} = \frac{-5}{1.667} = -3$$

Vergleicht man die beiden Regressionsgewichte der Residuenregressionen mit den partiellen Regressionsgewichten der trivariaten Regression, so zeigt sich, dass die Werte identisch sind.

Trivariate Regression: Auspartialisierung

In der trivariaten und allgemein auch in der multiplen Regression lassen sich daher die partiellen Regressionsgewichte auch als Koeffizienten von Residuenregressionen interpretieren.

Man spricht in diesem Zusammenhang auch von ***Auspartialisierung***. Dabei werden die (linearen) Effekte der Kontrollvariablen auf die abhängige Variable und auf die jeweils betrachtete erklärende Variablen „herausgerechnet“.

Dies ist, wie gleich bei der konditionalen Regression gezeigt werden wird, nicht identisch mit dem „Konstanthalten“ der Werte der Kontrollvariablen bei der Drittvariablenkontrolle in der Tabellenanalyse.

Konstanthalten und Auspartialisieren führen nur dann zum gleichen Ergebnis, wenn die trivariate Regression von Y auf die beiden erklärenden Variablen X und W tatsächlich linear-additiv sind, die tatsächliche Regressionsfunktion in der Population also der linearen Modellgleichung entspricht.

Gleichwohl hat sich auch in der multiplen Regression der Sprachgebrauch eingebürgert, dass die partiellen Regressionsgewichte den Effekt einer erklärenden Variablen wiedergeben, wenn die Werte der übrigen Prädiktoren „konstant“ gehalten werden, wie das bei der Drittvariablenkontrolle durch konditionale Effekte gegeben einen konstanten Wert der Drittvariablen der Fall ist.

Konditionale bivariate Regression

Bei der ***konditionalen Regression*** werden - analog dem Vorgehen der Analyse bivariater Bezie-hungen in Partialtabellen - für die alten und die neuen Bundesländer jeweils getrennt die Zu-sammenhänge zwischen Ablehnung von Schwangerschaftsabbrüchen und Religiosität berech-net werden:

	W	X	Y	n	X^2	Y^2	X·Y	W^2	W·X	W·Y
				Alte Bundesländer						
∑	0	10	15	4	30	59	39	0	0	0
				Neue Bundesländer						
∑	4	14	5	4	54	11	21	4	14	5

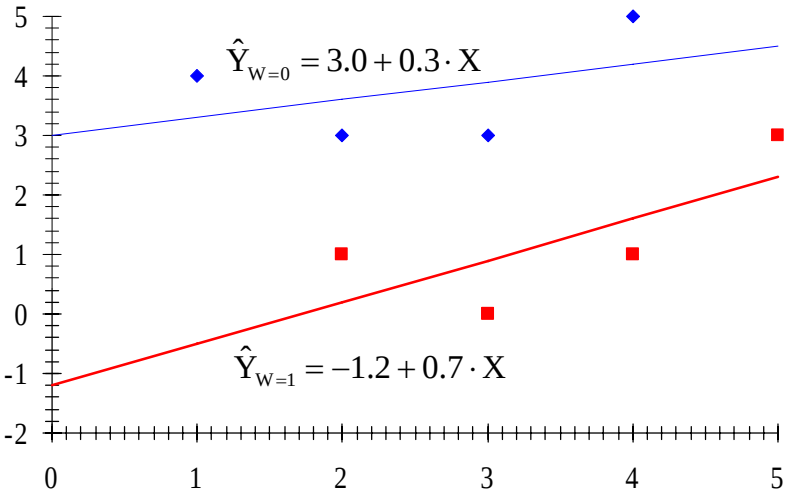
$$b_{W=0} = \frac{39 - \frac{10 \cdot 15}{4}}{30 - \frac{10^2}{4}} = 0.3 \qquad b_{W=1} = \frac{21 - \frac{14 \cdot 5}{4}}{54 - \frac{14^2}{4}} = 0.7$$

$$a_{W=0} = \frac{15}{4} - 0.3 \cdot \frac{10}{4} = 3.0 \qquad a_{W=0} = \frac{5}{4} - 0.7 \cdot \frac{14}{4} = -1.2$$

Die Vorhersagegleichungen für die beiden Ausprägungen der Kontrollvariable betragen:

$$\hat{Y}_{W=0} = a_0 + b_0 \cdot X = 3.0 + 0.3 \cdot X \text{ und } \hat{Y}_{W=1} = a_1 + b_1 \cdot X = -1.2 + 0.7 \cdot X$$

Konditionale Bivariate Regression



Das Regressionsgewicht ist in den alten Bundesländern (W=0) mit einem Wert von b_0=0.3 geringer als in den neuen Bundesländern, wo der Wert b_1=0.7 beträgt. In den neuen Bundes-ländern ist die Beziehung zwischen Ablehnung von Schwangerschaftsabbrüchen und Religio-sität also enger als in den alten Ländern.
Bei den Regressionskonstanten ist umgekehrt der Wert in den neuen Ländern geringer als in den alten Ländern. Die Regressionsgrade startet also in den neuen Ländern von einem tieferem Niveau aus.

Vergleich von partiellen und konditionalen Effekten

Der Unterschied zwischen konditionaler und trivariater Regression wird bei einer Gegenüberstellung der trivariaten Regression mit den beiden konditionalen Regressionsgeraden deutlich.

Dazu werden aus der Vorhersagegleichung des trivariaten Regressionsmodells zwei Regressionsgleichungen für die beiden Erhebungsgebiete berechnet:

Vorhersagewerte: $\hat{Y} = b_0 + b_1 \cdot X + b_2 \cdot W = 2.5 + 0.5 \cdot X - 3.0 \cdot W$

Wenn W=0: $\hat{Y} = b_0 + b_1 \cdot X + b_2 \cdot 0 = 2.5 + 0.5 \cdot X - 3.0 \cdot 0 = 2.5 + 0.5 \cdot X$

Wenn W=1: $\hat{Y} = b_0 + b_1 \cdot X + b_2 \cdot 1 = 2.5 + 0.5 \cdot X - 3.0 \cdot 1 = -0.5 + 0.5 \cdot X$

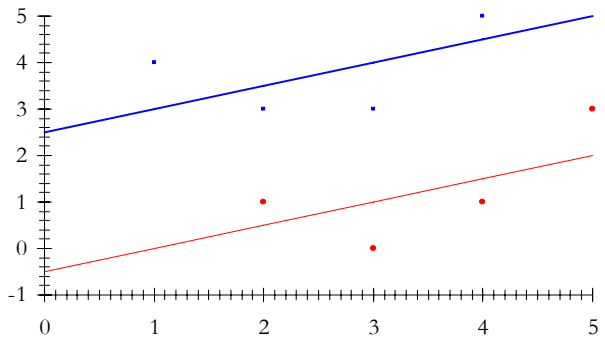
Die beiden Gleichungen unterscheiden sich nur in der Regressionskonstante, die bei der Aus-prägung W=1 um b_2=–3 Einheiten unter der Konstante bei W=1 liegt.

Im konditionalen Regressionsmodell unterscheiden sich dagegen nicht nur die Konstanten, sondern auch die beiden Regressionsgewichte:

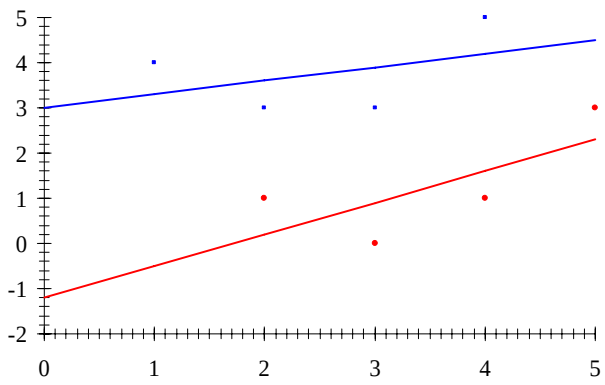
Wenn W=0: $\hat{Y} = a_{W=0} + b_{W=0} \cdot X = 3.0 + 0.3 \cdot X$

Wenn W=1: $\hat{Y} = a_{W=1} + b_{W=1} \cdot X = -1.2 + 0.7 \cdot X$

Vergleich von partiellen und konditionalen Effekten



Linear-additives Modell der trivariaten Regression:
die beiden Regessionskurven verlaufen parallel.



konditionale Regression:
getrennte Berechnung für W=0 und W=1:
Die beiden Regressionskurven verlaufen nicht parallel.

Multiple Regression: Vorteile gegenüber konditionaler Regression

Verglichen mit der Berechnung konditionaler Regressionsmodelle bietet die Drittvariablenkontrolle im trivariaten Regressionsmodell eine Reihe von Vorteilen:

+ Für die Berechnung sind weniger Fälle notwendig.
  Es ist insbesondere nicht notwendig, dass es für alle Ausprägungen der Kontrollvariable jeweils Realisationen zur Berechnung der konditionalen Regressionsgleichungen geben muss. Stattdessen werden alle Regressionsgewichte gemeinsam anhand der vorliegenden Daten geschätzt.
+ Es gibt keine formale Unterscheidung zwischen erklärender Variablen und Kontrollvariablen: Für alle erklärenden Variablen werden Effekte berechnet und jede erklärende Variable ist gleichzeitig Kontrollvariable für alle anderen erklärenden Variablen.
+ Die Interpretation ist einfacher, da für eine erklärende Variable anstelle vieler und sich möglicherweise unterscheidender konditionaler Effekte nur ein einziger partieller Effekt berechnet wird.
+ Das trivariate Regressionsmodell kann sehr leicht durch weitere erklärende Variablen zum Modell der multiplen Regression mit vielen erklärenden Variablen ausgeweitet werden, während die getrennte Berechnung in unterschiedlichen konditionalen Regresionsmodellen nicht nur unübersichtlich wird, sondern auch leicht an mangelnden Fallzahlen scheitern kann.

Multiple Regression: Nachteile gegenüber konditionaler Regression

Die Vorteile dieser Art von Drittvariablenkontrolle hat aber auch ihren Preis:

– es muss angenommen werden, dass die erklärenden Variablen tatsächlich linear-additiv auf die abhängige Variable wirken und somit anstelle der konditionalen Effekte tatsächlich partielle Effekte die Beziehung korrekt beschreiben.
– Die Homoskedastizitätsannahme bei der effizienten Schätzung der Regressionskoeffizienten gilt im Unterschied zur konditionalen Regression nicht nur innerhalb einer Ausprägungskombination der Kontrollvariablen, sondern für alle Ausprägungskombinationen aller erklärender Variablen.

Die zweite Annahme (Homoskedastizität) ist für die BLU-Eigenschaft der Kleinstquadratmethode unverzichtbar. Es gibt aber auch Schätzmethoden für lineare Regressionsmodelle mit heteroskedastischen und/oder autokorrelierten Residuen.

Die Annahme linear-additiver Effekte ist insofern weitgehend unproblematisch, als im Modell der multiplen Regression - wie in späteren Lerneinheiten gezeigt wird - letztlich auch Interaktionseffekte und nichtlineare Beziehungen modelliert werden können.

Empirisches Beispiel für die trivariate Regression:

Im ersten Beispiel wurde mit nur acht Fällen gerechnet, um das Prinzip der multiplen Regression zu verdeutlichen. Im Allbus 2006 gibt es empirische Daten, die als Indikatoren für die Einstellungen zum Schwangerschaftsabbruch und zur Religiosität interpretiert werden können.

So wird in sieben Items danach gefragt, ob ein Schwangerschaftsabbruch erlaubt oder verboten sein sollte, wenn (1) das Kind behindert sein wird, (2) die Schwangere keine weiteren Kinder haben möchte, (3) die Gesundheit der Frau durch die Schwangerschaft gefährdet sei, (4) eine finanzielle Notlage besteht, (5) die Schwangerschaft Folge einer Vergewaltigung ist, (6) die Frau ledig ist und den Vater des Kindes nicht heiraten will oder (7) die Frau einen Abbruch der Schwangerschaft will.

Für die folgende Analyse wurde aus den Antworten auf die sieben Variablen ein Index Y gebildet, der zählt, in wie vielen der 7 Situationen ein Respondent der Ansicht ist, dass ein Schwangerschaftsabbruch verboten sein sollte. Der Minimalwert von null bedeutet also, dass Schwangerschaftsabbrüche in allen Situationen erlaubt sein sollte, der Maximalwert von sieben, dass Schwangerschaftsabbrüche in allen Situationen verboten sein sollte.

Zur Erfassung der Religiosität (X) wurde die Kirchgangshäufigkeit verwendet, wobei die Ausprägungen null bis fünf den Antwortvorgaben „mehrmals in der Woche“ (5), „jede Woche“ (4), „ein- bis dreimal im Monat“ (3), „mehrmals im Jahr“ (2), „seltener“ (1) und „nie“ (0) zugeordnet wurde.

Das Erhebungsgebiet (W) hat wie im Beispiel mit den fiktiven Daten die Ausprägungen 0 für die alten und 1 für die neuen Bundesländer.

Empirisches Beispiel für die trivariate Regression:

n	X	Y	W	X^2	Y^2	W^2	X·Y	W·Y	W·X
$\sum$ 3402	4202	8019	1118	10540	31535	1118	12354	1850	824

Aus den Summen, Quadratsummen und Produktsummen über die gültigen Fälle der Variablen in den Allbus-Daten werden zunächst Mittelwerte, Varianzen und Kovarianzen und anschließend die Regressionskoeffizienten berechnet:

$$\bar{x} = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i = \frac{4202}{3402} = 1.235 \; ; \; s_X^2 = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{10540}{3481} - 1.235^2 = 1.573$$

$$\bar{y} = \frac{1}{n}\cdot\sum_{i=1}^{n} y_i = \frac{8019}{3402} = 2.357 \; ; \; s_Y^2 = \frac{1}{n}\cdot\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = \frac{31535}{3402} - 2.357^2 = 3.713$$

$$\bar{w} = \frac{1}{n}\cdot\sum_{i=1}^{n} w_i = \frac{1118}{3402} = 0.329 \; ; \; s_W^2 = \frac{1}{n}\cdot\sum_{i=1}^{n} w_i^2 - \bar{w}^2 = \frac{1118}{3402} - 0.329^2 = 0.221$$

$$s_{YX} = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i \cdot y_i - \bar{x}\cdot\bar{y} = \frac{12354}{3402} - 1.235\cdot 2.357 = 0.720$$

$$s_{YW} = \frac{1}{n}\cdot\sum_{i=1}^{n} w_i \cdot y_i - \bar{w}\cdot\bar{y} = \frac{1850}{3402} - 0.329\cdot 2.357 = -0.164$$

$$s_{WX} = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i \cdot w_i - \bar{x}\cdot\bar{w} = \frac{824}{3402} - 1.235\cdot 0.329 = -0.231$$

Empirisches Beispiel für die trivariate Regression:

Mittelwerte, Varianzen und Kovarianzen (n=3402):

		Varianz/Kovarianz-Matrix		
Variable	Mittelwert	X	Y	W
Kirchgang (X)	1.235	1.573		
Abtreibung (Y)	2.357	0.720	3.713	
Region (W)	0.329	–0.164	–0.231	0.221

Aus den Beschriftungen der Varianz/Kovarianz-Matrix ist erkennbar, welcher Wert für welche Statistik steht. Die Diagonalelemente enthalten die Varianzen, die darunter liegenden Elemente die Kovarianzen. In der Spalte X und Zeile Y findet sich so z.B. die Kovarianz zwischen X und Y, hier also zwischen Kirchgang und Abtreibung: $s_{YX} = 0.720$.

Die Mittelwerte, Varianzen und Kovarianzen bilden die Ausgangsbasis für die Berechnung der Regressionskoeffizienten:

$$b_1 = \frac{s_W^2 \cdot s_{XY} - s_{WY} \cdot s_{XW}}{s_X^2 \cdot s_W^2 - (s_{XW})^2} = \frac{0.221 \cdot 0.720 - (-.231) \cdot (-.164)}{1.573 \cdot 0.221 - (-.164)^2} = 0.378$$

$$b_2 = \frac{s_X^2 \cdot s_{WY} - s_{XY} \cdot s_{XW}}{s_X^2 \cdot s_W^2 - (s_{XW})^2} = \frac{1.573 \cdot (-.231) - 0.720 \cdot (-.164)}{1.573 \cdot 0.221 - (-.164)^2} = -0.766$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x} - b_2 \cdot \bar{w} = 2.357 - 0.378 \cdot 1.235 - (-.766) \cdot 0.329 = 2.142$$

Hinweis: Die wiedergegebenen Ergebnisse sind mit mehr als 3 Nachkommastellen berechnet.

Empirisches Beispiel für die trivariate Regression:

Mittelwerte, Varianzen und Kovarianzen (n=3402):

		Varianz/Kovarianz-Matrix		
Variable	Mittelwert	X	Y	W
Kirchgang (X)	1.235	1.573		
Abtreibung (Y)	2.357	0.720	3.713	
Region (W)	0.329	–0.164	–0.231	0.221

$$b_0 = 2.142 \text{ , } b_1 = 0.378 \text{ ; } b_2 = -0.766$$

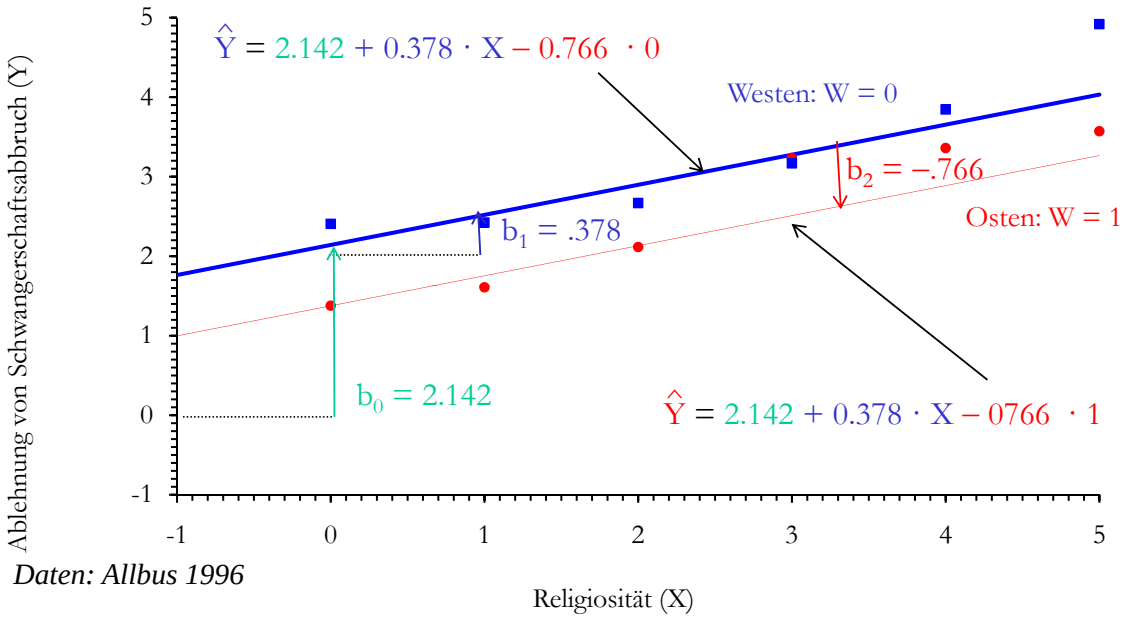
Der positive Wert des Partiellen Regressionsgewichts der Religiosität von $b_1 = 0.378$ weist darauf hin, dass bei Kontrolle des Erhebungsgebiets die Ablehnung von Schwangerschaftsabbrüchen um 0.378 Einheiten zunimmt, wenn die Religiosität um 1 Einheit ansteigt.

Analog bedeutet der negative Wert von $b_2 = -0.766$, dass bei Kontrolle der Religiosität im Osten (W=1) die Ablehnung von Schwangerschaftsabbrüchen um 0.766 Einheiten geringer ist als im Westen (W=0).

Die Regressionskonstante von $b_0 = 2.142$ bedeutet schließlich, dass Befragte aus den alten Bundesländern (W=0), die nie zur Kirche gehen (X=0) im Durchschnitt in 2.142 der sieben Situationen für ein Verbot von Schwangerschaftsabbrüchen sind.

Die Ergebnisse können wie im Eingangsbeispiel in eine Grafik eingetragen werden.

Empirisches Beispiel für die trivariate Regression:



Daten: Allbus 1996

Anstelle der Punktewolke sind die bedingten Stichprobenmittelwerte der Ausprägungskombina-tionen als Punkte eingezeichnet. Jeder Punkt steht also für eine unterschiedliche Anzahl von Befragten.

Vergleich der Effektstärken

In Regressionsmodellen mit mehreren erklärenden Variablen stellt sich die Frage nach der relativen Erklärungskraft der erklärenden Variablen untereinander, also wie die partiellen Zusammenhänge mit der abhängigen Variablen in ihrer Bedeutung und Stärke verglichen werden können.

Als Beispiel soll der Zusammenhang zwischen der Haltung zu Schwangerschaftsabbrüchen, der Religiosität und dem Alter der Befragten betrachtet werden.
Die Haltung zum Schwangerschaftsabbruch (Y) wird wieder über die Anzahl von insgesamt sieben vorgegebenen Situationen erfasst, in denen sich die befragten Personen für ein Verbot von Schwangerschaftsabbrüchen aussprechen.
Die Religiosität (X_1) wird über die Kirchgangshäufigkeit mit den Ausprägungen „mehrmals in der Woche“ (5), „jede Woche“ (4), „ein- bis dreimal im Monat“ (3), „mehrmals im Jahr“ (2), „seltener“ (1) und „nie“ (0) erfasst, das Alter (X_2) wird in Jahren gemessen.
Die geschätzten Populationsmittelwerte, -varianzen und -kovarianzen der n=3394 Fälle zeigt die folgende Tabelle:

Variable	Mittelwerte	Varianzen u. Kovarianzen		
Abtreibung (Y)	2.357	3.711		
Kirchgang (X)	1.234	0.726	1.567	
Alter in Jahren (W)	49.31	1.495	3.734	295.877

(Daten: Allbus 2006)

Vergleich der Effektstärken

Variable	Mittelwerte	Varianzen u. Kovarianzen		
Abtreibung (Y)	2.357	3.711		
Kirchgang (X)	1.234	0.726	1.567	
Alter in Jahren (W)	49.31	1.495	3.734	295.877

(Daten: Allbus 2006)

Die Schätzung der Regressionskoeffizienten ergibt:

$$b_1 = \frac{\hat{\sigma}^2_W \cdot \hat{\sigma}_{XY} - \hat{\sigma}_{WY} \cdot \hat{\sigma}_{XW}}{\hat{\sigma}^2_W \cdot \hat{\sigma}^2_X - \left(\hat{\sigma}_{XW}\right)^2} = \frac{295.877 \cdot 0.726 - 1.495 \cdot 3.734}{295.877 \cdot 1.567 - 3.734^2} = 0.465$$

$$b_2 = \frac{\hat{\sigma}^2_X \cdot \hat{\sigma}_{WY} - \hat{\sigma}_{XY} \cdot \hat{\sigma}_{XW}}{\hat{\sigma}^2_X \cdot \hat{\sigma}^2_W - \left(\hat{\sigma}_{XW}\right)^2} = \frac{1.567 \cdot 1.495 - 3.734 \cdot 0.726}{1.567 \cdot 295.877 - 3.734^2} = -0.0008$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x} - b_2 \cdot \bar{w} = 2.357 - 0.465 \cdot 1.234 - (-0.0008) \cdot 49.314 = 1.823$$

Die Regressionsgewichte besagen:

- *Wenn die Religiosität um eine Einheit ansteigt, dann steigt die Ablehnung von Schwangerschaftsabbrüchen im Durchschnitt um 0.465 Einheiten an,*
- *wenn das Alter um Jahr ansteigt, dann sinkt die Ablehnung von Schwangerschaftsabbrüchen um 0.0008 Einheiten.*

Bedeutet dies, dass die Religosität einen sehr viel größeren Effekt auf die Ablehnung von Schwangerschaftsabbrüchen hat als das Alter?

Vergleich der Effektstärken

Prädiktor	Koeffizient
Konstante (b_0)	1.823
Religiosität (b_1)	0.465
Alter in Jahren (b_2)	–0.001

Prädiktor	Koeffizient
Konstante (b_0)	1.823
Religiosität (b_1)	0.465
Alter (Jahrhdt.) (b_2)	–0.082

Prädiktor		Koeffizient
Konstante	(b^*_0)	0
Z(Relig.)	(b^*_1)	0.302
Z(Alter)	(b^*_2)	–0.007

Diese Frage ist anhand der Daten kaum zu beantworten, da die erklärenden Variablen in nicht vergleichbaren Einheiten gemessen werden.
Auch ändert eine Reskalierung die Koeffizienten. Wird etwa das Alter nicht in Jahren, sondern in Jahrhunderten gemessen, ergibt sich ein Regressionsgewicht von –0.082 (mittlere Tabelle).

Um einen einheitlichen Maßstab zu erhalten, können alle Modellvariablen standardisiert werden:

$$Z_Y = \frac{Y - \bar{y}}{s_Y} \; ; \; Z_X = \frac{X - \bar{x}}{s_X} \; ; \; Z_W = \frac{W - \bar{w}}{s_W}$$

$$Y = b_0 + b_1 \cdot X + b_2 \cdot W + E \Rightarrow Z_Y \cdot s_Y + \bar{y} = b_0 + b_1 \cdot \left(Z_X \cdot s_X + \bar{x}\right) + b_2 \cdot \left(Z_w \cdot s_w + \bar{w}\right)$$

$$Z_Y = \frac{b_0 - \left(\bar{y} - b_1 \cdot \bar{x} - b_2 \cdot \bar{w}\right)}{s_y} + \left(b_1 \cdot \frac{s_X}{s_Y}\right) \cdot Z_X + \left(b_1 \cdot \frac{s_W}{s_Y}\right) \cdot Z_W + \frac{E}{s_Y} = b^*_1 \cdot Z_X + b^*_2 \cdot Z_W + E^*$$

Die Ergebnisse der Regression zwischen den standardisierten Variablen sind in der rechten Tabelle festgehalten.

Vergleich der Effektstärken

Prädiktor	unstandard. Koeffizienten	standardisierte Koeffizienten
Konstante (b_0)	0.823	--
Religiosität (b_1)	0.465	0.302
Alter in Jahren (b_2)	–0.001	–0.007

Da die Mittelwerte standardisierter Variablen null sind, ist auch die Regressioskonstante im ***standardisierten Regressionsmodell*** notwendigerweise null.
Die partiellen Regressionsgewichte werden als ***standardisierte Regressionsgewichte*** oder auch als ***standardisierte Effekte*** bezeichnet. Sie geben an, um wie viel Standardabweichungen die abhängige Variable ansteigt bzw. sinkt, wenn eine erklärende Variable um 1 Standardabweichung ansteigt:

Erhöht sich die Religiosität um 1 Standardabweichung, erhöht sich die Ablehnung von Schwangerschaftsabbrüchen um 0.302 Standardabweichungen;
erhöht sich das Alter um 1 Standardabweichung, verringert sich die Ablehnung von Schwangerschaftsabbrüchen um 0.007 Standardabweichungen.

Bezogen auf den einheitlichen Maßstab „Standardabweichung“ ist daher der Effekt der Religiosität sehr viel höher als der des Alters.

Vergleich der Effektstärken

Prädiktor	unstandard. Koeffizienten	standardisierte Koeffizienten
Konstante (b_0)	0.823	--
Religiosität (b_1)	0.465	0.302
Alter in Jahren (b_2)	–0.001	–0.007

Vor allem in der Umfrageforschung, in der sehr unterschiedliche Antwortskalen Verwendung finden, werden daher neben den unstandardisierten Regressionskoeffizienten zusätzlich (oder auch ausschließlich) die standardisierten Regressionsgewichte berichtet.

Hinweise:

- *Die Interpretation standardisierter Regressionsgewichte schafft zwar einen einheitlichen Maßstab zur Beurteilung der relativen Einflussstärke, doch gehen in diesen Maßstab auch die empirischen Streuungen der Variablen ein.*
  Es ist daher möglich, dass bei zwei erklärende Variablen, die in gleichen Messskalen gemessen werden, die relative Größenordnung der unstandardisierten und der standardisierten Regressionsgewichte umgekehrt ist.
- *In Statistikprogrammen werden die standardisierten Regressiosgewichte sehr oft als Beta-Koeffizienten (β) bezeichnet. Da das griechische β aber auch für die Regressionskoeffizienten in der Population steht, verwende ich stattdessen ein Sternchen (b^*) zur Kennzeichnung standardisierter Koeffizienten.*

Vergleich der Effektstärken

Variable	Mittelwerte	Varianzen u. Kovarianzen			Korrelationen			Koeffizient
Abtreibung (Y)	2.357	3.711			1.000			
Religiosität (X)	1.234	0.726	1.567		0.301	1.000		$b^*_1 = 0.302$
Alter in Jahren (W)	49.314	1.495	3.734	295.877	0.045	0.173	1.000	$b^*_2 = -0.007$

Die standardisierten Regressionsgewichte können auch direkt aus den Produktmoment-Korrelationen berechnet werden, wobei es durch Rundungsfehler zu leichten Abweichungen gegenüber der Berechnung aus den unstandardisierten Koeffizienten kommen kann:

$$b_1^* = \frac{s^2_{Z_W} \cdot s_{Z_X Z_Y} - s_{Z_W Z_Y} \cdot s_{Z_X Z_W}}{s^2_{Z_X} \cdot s^2_{Z_W} - \left(s_{Z_X Z_W}\right)^2} = \frac{1 \cdot r_{XY} - r_{WY} \cdot r_{XW}}{1 \cdot 1 - \left(r_{XW}\right)^2} = \frac{r_{XY} - r_{WY} \cdot r_{XW}}{1 - r_{XW}^2} = \frac{0.301 - 0.045 \cdot 0.173}{1 - 0.173^2} = 0.302$$

$$b_2^* = \frac{s^2_{Z_X} \cdot s_{Z_W Z_Y} - s_{Z_X Z_Y} \cdot s_{Z_X Z_W}}{s^2_{Z_X} \cdot s^2_{Z_W} - \left(s_{Z_X Z_W}\right)^2} = \frac{1 \cdot r_{WY} - r_{XY} \cdot r_{XW}}{1 \cdot 1 - \left(r_{XW}\right)^2} = \frac{r_{WY} - r_{XY} \cdot r_{XW}}{1 - r_{XW}^2} = \frac{0.045 - 0.301 \cdot 0.173}{1 - 0.173^2} = -0.007$$

Da sich standardisierte und unstandardisierte Regressionsgewichten ineinander umrechnen lassen, kann es bei sehr unterschiedlichen Skalierungen der Modellvariablen zur Minimierung von Rundungsfehlern sinnvoll sein, zunächst die standardisierten Koeffizienten zu berechnen und daraus die unstandardisierten Werte.

$$b_k = b_k^* \cdot \frac{s_Y}{s_k} \Leftrightarrow b_k^* = b_k \cdot \frac{s_k}{s_Y}$$

wobei s_k die Standardabweichung der k-ten erklärenden Variablen ist.

Partielle Korrelation

In der bivariaten Regression ist das standardisierte Regressionsgewicht gleich der Korrelation. Die Werte müssen daher zwischen –1 und +1 liegen.
Dies gilt nicht für die partiellen standardisierten Gewichte. Sie können durchaus kleiner –1 bzw. größer +1 sein. Standardisiert sind nämlich nicht die Koeffizienten, sondern die Modellvariablen.
Standardisierte Regessionsgewichte größer Eins können vor allem dann auftreten, wenn die Prädiktoren sehr hoch miteinander korreliert sind.

Korrelationen messen die Stärke symmetrischer Beziehungen, während Regressionsgewichte die asymmetrische Beziehung zwischen abhängigen und unabhängigen Variablen erfassen. Analog zu partiellen Regressionsgewichten lassen sich auch ***partielle Korrelationskoeffizienten*** definieren. Sie messen die Korrelation, also die symmetrische Beziehung zwischen zwei Variablen, nachdem die linearen Einflüsse dritter Variablen auspartialisiert sind.

Die Berechnung kann über die Residuenregression erfolgen. Anstelle des Regressionsgewichts wird dabei die Korrelation der Residuen berechnet. Im Beispiel der trivariaten Regression ist die partielle Korrelation zwischen Kirchgangshäufigkeit und Schwangerschaftsabbruch also gleich der Korrelation zwischen den Residuen der Regressionen der beiden Variablen auf das Alter.

Partielle Korrelation

Die Varianzen der Residuen ergeben sich im Fall der trivariaten Regression zwischen standardisierten Variablen aus der Differenz der quadrierten Korelationen zwischen abhängiger und unabhängiger Variablen. Die Kovarianz zwischen diesen Residuen ist gleich dem Zähler in der Formel zur Berechnung des standardisierten Regressionsgewichts aus Korrelationen. Daher berechnen sich die partiellen Korrelationen bei drei Variablen nach:

$$r_{XY.W} = \frac{r_{XY} - r_{WY} \cdot r_{XW}}{\sqrt{\left(1 - r_{WY}^2\right) \cdot \left(1 - r_{XW}^2\right)}} = \frac{0.301 - 0.045 \cdot 0.173}{\sqrt{\left(1 - 0.045^2\right) \cdot (1 - 0.173^2)}} = 0.298$$

$$r_{WY.X} = \frac{r_{WY} - r_{XY} \cdot r_{XW}}{\sqrt{\left(1 - r_{XY}^2\right) \cdot \left(1 - r_{XW}^2\right)}} = \frac{0.045 - 0.301 \cdot 0.173}{\sqrt{\left(1 - 0.301^2\right) \cdot (1 - 0.173)^2}} = -.0075$$

$$r_{XW.Y} = \frac{r_{XW} - r_{XY} \cdot r_{WY}}{\sqrt{\left(1 - r_{XY}^2\right) \cdot \left(1 - r_{WY}^2\right)}} = \frac{0.173 - 0.301 \cdot 0.045}{\sqrt{\left(1 - .301^2\right) \cdot (1 - 0.045^2)}} = 0.180$$

Da die Zähler der Formeln für partielle Korrelationen und partielle Regressionsgewichte gleich sind, sind auch die Vorzeichen von partiellen unstandardisierten Regressionsgewichten, partiellen standardisierten Regressionsgewichten und partiellen Korrelationen immer identisch. Das bedeutet, dass eine partielle Korrelation genau dann null ist, wenn auch das partielle (standardisierte wie unstandardisierte) Regressionsgewicht null ist.

Eigenschaften von Vorhersagewerten und Residuen

Wie im bivariaten Fall gilt auch in der trivariaten Regression:

(1) der Mittelwert der Vorhersagewerte ist gleich dem Mittelwert der abhängigen Variablen:

$$\bar{\hat{y}} = \frac{1}{n} \cdot \sum_{i=1}^{n} \hat{y}_i = b_0 + b_1 \cdot \frac{\sum_{i=1}^{n} x_i}{n} + b_2 \cdot \frac{\sum_{i=1}^{n} w_i}{n} = b_0 + b_1 \cdot \bar{x} + b_2 \cdot \bar{w} = \bar{y}$$

(2) Die „erklärte“ Variation (und Stichprobenvarianz), d.h. die Variation (Varianz) der Vorhersagewerte ist eine Funktion der Variationen (Varianzen) der erklärenden Variablen:

$$SS_{\hat{Y}} = b_1^2 \cdot SS_X + b_2^2 \cdot SS_W + 2 \cdot b_1 \cdot b_2 \cdot SP_{XW} = b_1 \cdot SP_{XY} + b_2 \cdot SP_{WY}$$

$$s_{\hat{Y}}^2 = b_1^2 \cdot s_X^2 + b_2^2 \cdot s_W^2 + 2 \cdot b_1 \cdot b_2 \cdot s_{XW} = b_1 \cdot s_{XY} + b_2 \cdot s_{WY}$$

(3) Der Mittelwert der Residuen ist Null und die Stichprobenresiduen kovariieren (und korrelieren) nicht mit den Vorhersagewerten und den erklärenden Variablen:

$$s_{\hat{Y}E} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right) \cdot e_i = 0;\ s_{XE} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \bar{x}\right) \cdot e_i = 0;\ s_{WE} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(w_i - \bar{w}\right) \cdot e_i = 0$$

Eigenschaften von Vorhersagewerten und Residuen: Varianzzerlegung

(4) Wie in der bivariaten Regression ist daher die Variation bzw. Stichprobenvarianz der abhängigen Variable gleich der Summe der Variationen bzw. Varianzen der Vorhersagewerte und der Residuen:

$$\begin{aligned} SS_Y &= SS_{\hat{Y}} + SS_E \\ &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 \\ s_Y^2 &= s_{\hat{Y}}^2 + s_E^2 \\ &= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{n} + \frac{\sum_{i=1}^{n} e_i^2}{n} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n} \end{aligned}$$

(5) Aus der Varianzzerlegung folgt als ein PRE-Maß für die Stärke der Zusammenhangs der Anteil der erklärten Varianz (Variation) an der Gesamtvarianz (Gesamtvariation) der abhängigen Variablen, der als ***Determinationskoeffizient*** (R^2) bezeichnet wird:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SS_{\hat{Y}}}{SS_Y} = \frac{s_{\hat{Y}}^2}{s_Y^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SS_E}{SS_Y} = 1 - \frac{s_E^2}{s_Y^2}$$

Eigenschaften von Vorhersagewerten und Residuen: Varianzzerlegung

Die positive Quadratwurzel aus dem Determinationskoeffizienten wird ***multiple Korrelation*** genannt.
Die multiple Korrelation ist gleichzeitig die Produktmomentkorrelation der Vorhersagewerte mit der abhängigen Variablen:

$$R = r_{\hat{Y}Y} = \sqrt{R^2}$$

Für das Beispiel der Erklärung der Haltung zu Schwangerschaftsabbrüchen durch die Religiosität und das Alter aus dem Allbus 1996 ergeben sich folgende Werte:

$$s_{\hat{Y}}^2 = b_1 \cdot s_{XY} + b_2 \cdot s_{WY} = 0.465 \cdot 0.726 + (-.0.00082) \cdot 0.1.495 = 0.337$$

$$s_E^2 = s_Y^2 - s_{\hat{Y}}^2 = 3.710 - 0.337 = 3.373$$

$$R^2 = \frac{s_{\hat{Y}}^2}{s_Y^2} = \frac{0.337}{3.373} = 0.091 = 9.1\%$$

$$R = \sqrt{R^2} = \sqrt{0.091} = 0.301$$

Der Determinationskoeffizient beträgt 9.1%. Die Variation der Religiosität und des Alters der befragten Personen decken zusammen also gut 9% der Unterschiede bei der Ablehnung von Schwangerschaftsabbrüchen. Die multiple Korrelation ist 0.301.

Lerneinheit 8:
Schätzen und Testen im trivariaten Regressionsmodell

Wie in der bivariaten Regression mit einer abhängigen und nur einer erklärenden Variablen soll auch in der trivariaten und allgemeiner in der multiplen Regression mit mehreren erklärenden Variablen die OLS-Schätzung der Regressionskoeffizienten zu möglichst guten Schätzungen der entsprechenden Populationskoeffizienten führen.
Dies ist nur der Fall, wenn die Anwendungsvoraussetzungen erfüllt sind:

1. **Linearitätsannahme**: die abhängige Variable lässt sich als lineare Funktion der Prädiktoren und der Populationsresiduen beschreiben, wobei die Populationsmittelwerte der bedingten Populationsresiduen bei allen Ausprägungskombinationen der erklärenden Variablen null sind.

2. **Unkorreliertheit** von Residualvariable mit allen erklärenden Variablen.

Beide Annahmen sind notwendigerweise erfüllt, wenn die bedingten Mittelwerte der abhängigen Variable in der Population eine lineare Funktion der erklärenden Variablen sind:

$$\mu(Y|X_1, X_2, \ldots, X_K) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_K \cdot X_K$$

Wenn diese beiden Annahmen erfüllt sind, dann sind die nach der OLS-Methode geschätzten Regressionskoeffizienten konsistente und erwartungstreue Schätzer der Koeffizienten in der Population.

3. **Homoskedastizität**: Die bedingten Residualvarianzen sind bei allen Ausprägungen der erklärenden Variablen gleich groß.
4. **keine Autokorrelation**: die Residuen sind nicht autokorreliert.

Wenn die Bedingungen 1. bis 4. erfüllt sind, dann sind die OLS-Schätzer effizient (in der Klasse der linearen Schätzer). Dies ist wieder die BLU-Eigenschaft, nach der es keine anderen erwartungstreuen (linearen) Schätzer mit geringeren Standardfehlern gibt.

Bei der Berechnung von Konfidenzintervallen und statistischen Tests wird zusätzlich i.a. angenommen:

5. **Normalverteilungsannahme**: Die Populationsresiduen sind normalverteilt.

Bei hinreichend großen Fallzahlen ist Annahme (5) nicht so wichtig, weil nach dem zentralen Grenzwertsatz dann die Schätzer unabhängig von der Verteilung der Residuen in der Population asymptotisch normalverteilt sind.
Die Annäherung an die Normalverteilung ist hinreichend genau, wenn die Fallzahl mindestens gleich der Zahl der geschätzten Regressionskoeffizienten plus 30 ($n \geq 30 + K + 1$) oder besser plus 50 ($n \geq 50 + K + 1$) ist.

Schätzung der Residualvarianz in der Population

Da die Standardfehler der Regressionskoeffizienten wie in der bivariaten Regression eine Funktion der unbekannten Varianz der Populationsresiduen sind, wird zunächst eine Schätzung dieser Residualvarianz benötigt.
Dabei gilt wie bei der bivariaten Regression, dass der konsistente und erwartungstreue Schätzer der Residualvarianz gleich der Variation der Stichprobenresiduen geteilt durch die Freiheitsgrade ist, wobei diese die Differenz der Fallzahl minus der Zahl der geschätzten Regressionskoeffizienten ist.

Im trivariaten Regressionsmodell gilt also:

$$\hat{\sigma}_U^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{df} = \frac{\sum_{i=1}^{n} e_i^2}{n-3} = \frac{SS_E}{n-3} = \frac{n}{n-3}\cdot s_E^2$$

Im Beispiel der trivariaten Regression der Ablehnung von Schwangerschaftsabbrüchen (Y) auf die Religiosität (X) und das Alter (W) ergibt sich bei den Daten aus dem Allbus 2006:

$$s_E^2 = s_Y^2 - s_{\hat{Y}}^2 = 3.71 - 0.337 = 3.373$$

Bei n=3394 Fällen berechnet sich dann die geschätzte Varianz der Populationsresiduen als:

$$\hat{\sigma}_U^2 = \frac{3394}{3391}\cdot 3.373 = 3.376$$

Bei größerer Rechengenauigkeit beträgt der Wert 3.377.

Standardfehler der Regressionsgewichte

Der geschätzte Standardfehler des Regressionsgewichts b_1 von X ist in der trivariaten Regression mit den beiden Prädiktoren X und W:

$$\hat{\sigma}(b_1) = \sqrt{\frac{SS_W}{SS_X \cdot SS_W - (SP_{XW})^2}\cdot\frac{SS_E}{n-3}} = \sqrt{\frac{s_W^2}{s_X^2 \cdot s_W^2 - (s_{XW})^2}\cdot\frac{\hat{\sigma}_U^2}{n}} = \sqrt{\frac{\hat{\sigma}_W^2}{\hat{\sigma}_X^2 \cdot \hat{\sigma}_W^2 - (\hat{\sigma}_{XW})^2}\cdot\frac{\hat{\sigma}_U^2}{n-1}}$$

Durch Vertauschen der Rollen von X und W gilt die Formel entsprechend für den Standardfehler des zweiten Regressionsgewichts b_2 von W:

$$\hat{\sigma}(b_2) = \sqrt{\frac{SS_X}{SS_W \cdot SS_X - (SP_{XW})^2}\cdot\frac{SS_E}{n-3}} = \sqrt{\frac{s_X^2}{s_W^2 \cdot s_X^2 - (s_{XW})^2}\cdot\frac{\hat{\sigma}_U^2}{n}} = \sqrt{\frac{\hat{\sigma}_X^2}{\hat{\sigma}_W^2 \cdot \hat{\sigma}_X^2 - (\hat{\sigma}_{XW})^2}\cdot\frac{\hat{\sigma}_U^2}{n-1}}$$

Die geschätzte Kovarianz der Kennwerteverteilungen der beiden Regressionsgewichte beträgt:

$$\hat{\sigma}(b_1, b_2) = \frac{-SP_{XW}}{SS_X \cdot SS_W - (SS_{XW})^2}\cdot\frac{SS_E}{n-3} = \frac{-s_{XW}}{s_X^2 \cdot s_W^2 - (s_{XW})^2}\cdot\frac{\hat{\sigma}_U^2}{n} = \frac{-\hat{\sigma}_{XW}}{\hat{\sigma}_X^2 \cdot \hat{\sigma}_W^2 - (\hat{\sigma}_{XW})^2}\cdot\frac{\hat{\sigma}_U^2}{n-1}$$

Das Vorzeichen der Kovarianz zwischen den Schätzern hängt vom Vorzeichen der Kovarianz bzw. Korrelation der beiden Prädiktoren ab.
Korrelieren die erklärenden Variablen negativ miteinander, dann ist die Kovarianz zwischen den Schätzern positiv, korrelieren die Variablen positiv, dann ist die Kovarianz negativ.

Standardfehler der Regressionsgewichte

Variable	Mittelwerte	Varianzen u. Kovarianzen		
Abtreibung (Y)	2.357	3.711		
Kirchgang (X)	1.234	0.726	1.567	
Alter in Jahren (W)	49.31	1.495	3.734	295.877

(Daten: Allbus 2006)

	unstandard. Koeffizienten	standardisierte Koeffizienten
b_0	0.823	--
b_1	0.465	0.302
b_2	–0.001	–0.007

Für das Allbus-Beispiel berechnen sich die Standardfehler der beiden Regressionsgewichte als:

$$\hat{\sigma}(b_1) = \sqrt{\frac{\hat{\sigma}_W^2}{\hat{\sigma}_X^2 \cdot \hat{\sigma}_W^2 - (\hat{\sigma}_{XW})^2} \cdot \frac{\hat{\sigma}_U^2}{n-1}} = \sqrt{\frac{295.877}{1.567 \cdot 295.877 - 3.734^2} \cdot \frac{3.377}{3393}} = 0.026$$

$$\hat{\sigma}(b_2) = \sqrt{\frac{\hat{\sigma}_X^2}{\hat{\sigma}_X^2 \cdot \hat{\sigma}_W^2 - (\hat{\sigma}_{XW})^2} \cdot \frac{\hat{\sigma}_U^2}{n-1}} = \sqrt{\frac{1.567}{1.567 \cdot 295.877 - 3.734^2} \cdot \frac{3.377}{3393}} = 0.002$$

Wenn alle 5 Anwendungsbedingungen erfüllt sind, kann die T-Verteilung mit df = n – K –1, bei der trivariaten Regression also mit df = n –3 Freiheitsgraden für Konfidenzintervalle und Tests verwendet werden. Ist die Normalverteilungsannahme nicht erfüllt, wird stattdessen die Standardnormalverteilung verwendet.
Im Sinne eines vorsichtigen (konservativen) Vorgehens wird die T-Verteilung i.a. auch dann verwendet, wenn die Populationsresiduen nicht normalverteilt sein, Annahme 5 also nicht zutrifft.

Konfidenzintervalle der Regressionsgewichte

Prädiktor	unstandard. Koeffizienten	Standard-fehler	standardisierte Koeffizienten
Konstante (b_0)	1.823	0.097	--
Religiosität (b_1)	0.465	0.026	0.302
Alter in Jahren (b_2)	–0.001	0.002	–0.007

(Berechnungen mit SPSS)

Für die beiden Regressionsgewichte der trivariaten Regression berechnen sich dann die Grenzen des (1–α)-Konfidenzintervalls nach:

$$\text{c.i.}(\beta_1) = b_1 \pm \hat{\sigma}_{b_1} \cdot t_{1-\alpha/2;df=n-3}$$

$$\text{c.i.}(\beta_2) = b_2 \pm \hat{\sigma}_{b_2} \cdot t_{1-\alpha/2;df=n-3}$$

Beim 95%-Konfidenzintervalle wird das 97.5%-Quantil der T-Verteilung mit 3391 Freiheitsgraden benötigt, das zwischen 1.98 (df=120) und 1.96 (df=∞) liegt. Wenn im Sinne des konservativen Vorgehens der größere Wert verwendet wird, ergeben sich die Intervallgrenzen nach:

$$\text{c.i.}(\beta_1) = 0.465 \pm 0.026 \cdot 1.98 = 0.413 \text{ bis } 0.516$$

$$\text{c.i.}(\beta_2) = -0.001 \pm 0.002 \cdot 1.98 = -0.005 \text{ bis } 0.003$$

Bei einer Irrtumswahrscheinlichkeit von 5% ist also zu vermuten, dass die Ablehnung von Schwangerschaftsabbrüchen in Ost wie West um 0.41 bis 0.52 zunimmt, wenn die Religiosität um eine Einheit ansteigt.
Da der Wert 0 im 95%-Konfidenzintervall des Alters liegt, hat Alter möglicherweise gar keinen Einfluss auf die Haltung zu Schwangerschaftsabbrüchen.

Hypothesentests über Regressionsgewichte

Auch die Hypothesenprüfung erfolgt analog zur bivariaten Regression.
Geprüft werden können die Hypothesenpaare:
(1) H_0: $\beta_k = \beta$ versus H_1: $\beta_k \neq \beta$
(2) H_0: $\beta_k \leq \beta$ versus H_1: $\beta_k > \beta$
(3) H_0: $\beta_k \geq \beta$ versus H_1: $\beta_k < \beta$
Dabei steht β_k für einen unstandardisierten Regressionskoeffizienten der k-ten erklärenden Variablen und β für den Wert, den dieser Koeffizient nach der Nullhypothese aufweist bzw. nicht unter- oder überschreitet.

Als Teststatistik wird die Differenz des geschätzten Regressionsgewichts b_k minus dem postulierten Wert β durch den Standardfehler von b_k geteilt. Bei $\beta_k = \beta$ und Zutreffen der 5 Annahmen der OLS-Methode ist diese Teststatistik t-verteilt:

$$T = \frac{b_k - \beta}{\hat{\sigma}(b_k)}$$

Auch wenn die Normalverteilungsannahme der Residuen nicht erfüllt ist, ist die Teststatistik anwendbar.
Sie ist dann bei gültiger Nullhypothese (an der Stelle $\beta_k=\beta$) asymptotisch standardnormalverteilt.
Wie bei den Konfidenzintervallen wird die T-Verteilung im Sinne eines vorsichtigen Vorgehens auch dann angewendet, wenn die Normalverteilungsannahme nicht gegeben ist. Dies ist nur dann sinnvoll, wenn die Forschungshypothese Alternativhypothese ist.

Hypothesentests über Regressionsgewichte

Wenn die Nullhypothese falsch ist, ist die Teststatistik T nichtzentral t-verteilt bzw. asymptotisch normalverteilt mit einem Erwartungswert ungleich null.
Die Nullhypothese wird daher abgelehnt bei der Prüfung von
(1) H_0: $\beta_k = \beta$ wenn $T < t_{df=n-3;\alpha/2}$ oder $T > t_{df=n-2;1-\alpha/2}$, bzw.
(2) H_0: $\beta_k \leq \beta$ wenn $T > t_{df=n-3;1-\alpha}$, bzw.
(3) H_0: $\beta_k \geq \beta$ wenn $T < t_{df=n-3;\alpha}$.

Prädiktor	unstandard. Koeffizienten	Standard-fehler	standardisierte Koeffizienten
Konstante (b_0)	1.823	0.097	--
Religiosität (b_1)	0.465	0.026	0.302
Alter in Jahren (b_2)	–0.001	0.002	–0.007

(Berechnungen mit SPSS)

Soll für das Anwendungsbeispiel mit einer Irrtumswahrscheinlichkeit von 1% geprüft werden, ob die Religiosität einen Effekt auf die Haltung zu Schwangerschaftsabbrüchen hat, wird das Hypothesenpaar H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$ geprüft. Die Teststatistik ist dann:

$$T = \frac{b_1 - \beta}{\hat{\sigma}(b_1)} = \frac{0.465 - 0}{0.026} = 17.9$$

Der Wert ist bei drei Nachkommastellen ziemlich ungenau. Genauere Werte werden vom Statistikprogramm SPSS mit den geschätzten Parameterwerten ausgedruckt.

Hypothesentests über Regressionsgewichte

Prädiktor	unstandard. Koeffizienten	Standard-fehler	$H_0: \beta_k=0$ T	Signifikanz p	standardisierte Koeffizienten
Konstante (b_0)	1.823	0.097	18.780	≤0.001	--
Religiosität (b_1)	0.465	0.026	18.188	≤0.001	0.302
Alter in Jahren (b_2)	–0.001	0.002	–0.440	0.660	–0.007

(Berechnungen mit SPSS)

Die Nullhypothese ist abzulehnen, wenn die Teststatistik größer ist als das 0.5%-Quantil bzw. kleiner als das 99.5%-Quantil der T-Verteilung mit 3391 Freiheitsgraden. Der kritische Wert liegt zwischen –2.617 (df=120) und 2.576 (df=∞). Die Nullhypothese ist auch beim größeren Wert zu verwerfen. Vermutlich gibt es also auch in der Population einen Zusammenhang zwischen Religiosität und Haltung zu Schwangerschaftsabbrüchen.

In der Regel wird in Statistikprogrammen automatisch für jeden geschätzten Regressionskoeffizienten neben dem Standardfehler auch die Teststatistik und die empirische Signifikanz für den zweiseitigen Test der Nullhypothese berechnet, dass der Regressionskoeffizient in der Population Null ist. Als Teststatistik wird dabei stets die T-Verteilung herangezogen.

Aufgrund größerer Rechengenauigkeit ergeben sich bei Berechnung der Teststatistik per Hand abweichende Werte. Die Schlussfolgerungen sind jedoch die gleichen: Religiosität hat einen signifikanten Effekt, Alter dagegen selbst bei α=10% nicht.

Adjustierter Determinationskoeffizient und F-Test

Bei der Schätzung eines multiplen Regressionsmodells stellt sich die Frage, ob überhaupt ein (monotoner) Zusammenhang zwischen der abhängigen Variablen Y einerseits und den erklärenden Variablen andererseits besteht.
Wenn dies der Fall ist, ist in der Population der Determinationskoeffizient $\rho^2_{Y.X,W}$ größer null.

Der Determinationskoeffizient R^2 in der Stichprobe ist allerdings kein optimaler Schätzers des Determinationskoeffizienten in der Population, da er den Populationswert im Mittel überschätzt. Dies liegt daran, dass bei der OLS-Schätzung selbst zufällige Stichprobengegebenheiten genutzt werden, die Vorhersage der abhängigen Variable zu verbessern. Zudem gilt für R^2 als Quotienten aus zwei positiven Werten notwendigerweise: $R^2 \geq 0$.
Da bei zutreffenden Modellannahmen (1) bis (4) in der Population gilt:

$$\rho^2_{Y.X,W} = \frac{\sigma^2\left(\beta_0+\beta_1\cdot X+\beta_2\cdot W\right)}{\sigma^2_Y} = 1-\frac{\sigma^2_U}{\sigma^2_Y}$$

liegt es nahe, die erwartungstreuen Schätzer der Residualvarianz und der Varianz der abhängigen Variablen als Schätzer zu verwenden. Diese Schätzung wird ***adjustierter Determinationskoeffizient*** genannt.

$$R^2_{adj.} = 1-\frac{\hat{\sigma}^2_U}{\hat{\sigma}^2_Y} = 1-\frac{SS_E/(n-K-1)}{SS_Y/(n-1)} = 1-\frac{n-1}{n-K-1}\cdot\left(1-R^2\right)$$

Adjustierter Determinationskoeffizient

Prädiktor	unstandard. Koeffizienten	Standard-fehler	standardisierte Koeffizienten
Konstante (b_0)	1.823	0.097	--
Religiosität (b_1)	0.465	0.026	0.302
Alter in Jahren (b_2)	–0.001	0.002	–0.007
(Berechnungen mit SPSS)			

Varianzzerlegung: Quelle	Variation	df
Regression	1142.616	2
Residuen	11450.005	3391
Total	12592.621	3393
R^2: 0.091, $R^2_{adj.}$: 0.090, F: 169.197		

Für das Beispiel des Schwangerschaftsabbruchs ergibt sich:

$$R^2_{adj.} = 1 - \frac{11450.005 / 3391}{12592.621 / 3393} = 1 - \frac{3393}{3391} \cdot \left(1 - \frac{1142.616}{12592.621}\right) = 0.090$$

Der adjustierte Determinationskoeffizient ist geringfügig geringer als der einfache Determinationskoeffizient.

Tatsächlich ist der adjustierte Determinationskoeffizient stets kleiner als R^2. Er kann sogar theoretisch (leicht) negative Werte annehmen.

Obwohl der adjustierte Determinationskeffizient ein Quotient aus zwei erwartungstreuen Schätzern ist, ist er selbst nicht erwartungstreu. Tatsächlich gibt es keinen erwartungstreuen Schätzer von $\rho^2_{Y.X,W}$, der unabhängig von Verteilungsannahmen ist.

Die F-Verteilung

Zur Prüfung der Forschungshypothese, dass in der Population der Determinationskoeffizient $\rho^2_{Y.X,W}$ größer Null ist:

$$H_0: \rho^2_{Y.X,W} = 0 \text{ versus } H_1: \rho^2_{Y.X,W} > 0$$

wird eine spezifische Teststatistik herangezogen.

Die Alternativhypothese trifft nur dann zu, wenn mindestens ein Regressionsgewicht ungleich null ist. Alternativ kann das zu testenden Hypothesenpaar daher auch formuliert werden als:

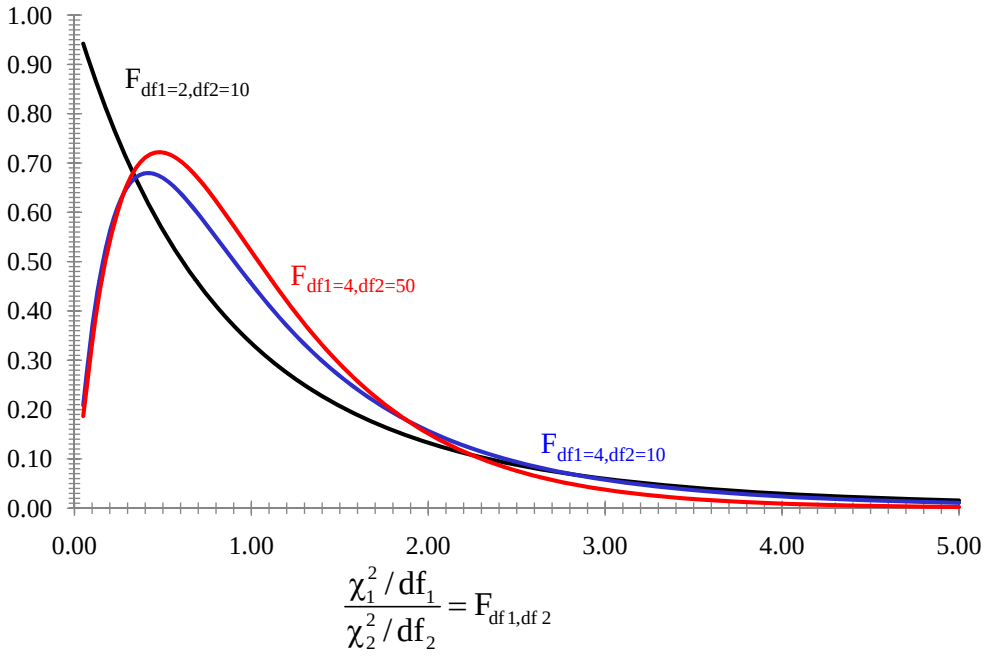
$$H_0: \beta_1 = 0 \text{ und } \beta_2 = 0 \text{ versus } H_1: \beta_1 \neq 0 \text{ oder } \beta_2 \neq 0.$$

Wenn nun (a) die Nullhypothese zutrifft, der Determinationskoeffizient in der Population also Null ist, *und* wenn (b) alle vier Modellannahmen für die BLU-Eigenschaft zutreffen *und* (c) zusätzlich die Populationsresiduen normalverteilt sind, dann sind sowohl die Variation der Residualvarianz wie auch die Variation der Vorhersagewerte geteilt durch die Residualvarianz in der Population chiquadrat-verteilt und statistisch unabhängig voneinander:

$$\frac{\sum_{i=1}^{n} e_i^2}{\sigma_U^2} \sim \chi^2_{df=n-3} \; ; \; \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sigma_U^2} \sim \chi^n_{df=2}$$

Für zwei voneinander unabhängige chiquadrat-verteilten Zufallsvariablen gilt, das der Quotient der beiden Zufallsvariablen jeweils geteilt durch ihre Freiheitsgrade einer ***F-Verteilung*** folgt:

Die F-Verteilung



Die F-Verteilung ist eine rechtsschiefe Verteilungsfamilie, deren Mitglieder durch die beiden Parameter df_1 und df_2 gekennzeichnet sind. Die Abbildung zeigt verschieden F-Verteilungen.

Die F-Verteilung

df_1 = 1				df_1 = 2			df_1 = 3			df_1 = 4		
df_2	90%	95%	99%	90%	95%	99%	90%	95%	99%	90%	95%	99%
1	39.86	161.4	4052.	49.50	199.5	5000.	53.59	215.7	5403.	55.83	224.6	5625.
10	3.285	4.965	10.04	2.924	4.103	7.559	2.728	3.708	6.552	2.605	3.478	5.994
15	3.073	4.543	8.683	2.695	3.682	6.359	2.490	3.287	5.417	2.361	3.056	4.893
30	2.881	4.171	7.562	2.489	3.316	5.390	2.276	2.922	4.510	2.142	2.690	4.018
60	2.791	4.001	7.077	2.393	3.150	4.977	2.177	2.758	4.126	2.041	2.525	3.649
120	2.748	3.920	6.851	2.347	3.072	4.787	2.130	2.680	3.949	1.992	2.447	3.480
∞	2.706	3.842	6.635	2.303	2.996	4.605	2.084	2.605	3.782	1.945	2.372	3.319

In den meisten Statistikbüchern sind Quantile der F-Verteilung aufgeführt.

Aus der obigen Tabelle ist zu entnehmen, dass das 95%-Quantil der F-Verteilung mit df_1=2 und df_2=120 Freiheitsgraden 3.072 ist.

Da eine f-verteilte Variable als Quotient dargestellt werden kann, ist auch ihr Kehrwert f-verteilt:

$$f_{\alpha;df\,1,df\,2} = \frac{1}{f_{1-\alpha;df\,2,df\,1}}$$

Darüber hinaus gibt es weitere Beziehungen zur T-Verteilung und zur χ^2-Verteilung:

$$\left(T_{\alpha/2;df=k}\right)^2 = \left(T_{1-\alpha/2;df=k}\right)^2 = F_{\alpha;df\,1=1,df\,2=k} \;;\; \chi^2_{\alpha;df=k}/k = F_{\alpha;df\,1=k,df\,2=\infty}$$

Die F-Statistik in der linearen Regression

Da die F-Verteilung aus zwei unabhängigen Chiquadrat-Verteilungen konstruiert werden kann, folgt in der trivariaten Regression für die Variation der Vorhersagewerte und der Residuen, wenn der Determinationskoeffizient Null ist und alle fünf Modellannahmen erfüllt sind:

$$F_{df_1,df_2} = \frac{\chi^2_{df_1}/df_1}{\chi^2_{df_2}/df_2} = \frac{\frac{SS_{\hat{Y}}}{\sigma^2_U}/2}{\frac{SS_E}{\sigma^2_U}/(n-3)} = \frac{SS_{\hat{Y}}/2}{SS_E/(n-3)}$$

Durch Umformen ergeben sich alternative Rechenformeln:

$$F = \frac{SS_{\hat{Y}}/2}{SS_E/(n-3)} = \frac{(SS_Y - SS_E)/2}{SS_E/(n-3)} = \frac{R^2/2}{(1-R^2)/(n-3)}$$

Wenn die Nullhypothese falsch ist, also mindestens ein Regressionsgewicht größer null ist, dann ist der Qotient F Teststatistik nichtzentral f-verteilt.
Da die nichtzentrale F-Verteilung bei gleicher Zahl von Freiheitsgraden höhere Werte aufweist als die zentrale F-Verteilung, kann F als Testatistik verwendet werden.
Die Nullhypothese wird bei einer Irrtumswahrscheinlichkeit α abgelehnt, wenn die Teststatistik größer ist als das $(1-\alpha)$-Quantil der F-Verteilung mit $df_1 = 2$ und $df_2 = n-3$ Freiheitsgraden.

Der F-Test

Prädiktor	unstandard. Koeffizienten	Standard-fehler	standardisierte Koeffizienten
Konstante (b_0)	1.823	0.097	--
Religiosität (b_1)	0.465	0.026	0.302
Alter in Jahren (b_2)	–0.001	0.002	–0.007
(Berechnungen mit SPSS)			

Varianzzerlegung: Quelle	Variation	df
Regression	1142.616	2
Residuen	11450.005	3391
Total	12592.621	3393
R^2: 0.091, $R^2_{adj.}$: 0.090, F: 169.197		

In der trivariaten Regression der Haltung zum Schwangerschaftsabbruch auf Religiosität und Alter ergibt sich ein F-Wert von:

$$F = \frac{R^2/2}{(1-R^2)/3391} = \frac{0.091/2}{0.909/3391} = 169.7$$

Die Abweichungen von dem Wert, den das Statistikprogramm SPSS für die Allbus-Daten berechnet hat, ist auf Rundungsfehler zurückzuführen.
Bei einer Irrtumswahrscheinlichkeit von 1% liegt der kritische F-Wert bei $df_1=2$ und $df_2=3473$ zwischen 4.605 und 4.787.
Da die Teststatistik größer ist, ist die Nullhypothese abzulehnen. Bei einer Irrtumswahrscheinlichkeit von 5% ist davon auszugehen, dass der Determinationskoeffizient in der Grundgesamtheit größer Null ist, also mindestens eine erklärende Variable ein Regressionsgewicht ungleich Null aufweist.

F-Test

Anmerkungen:

- Es ist möglich, dass der F-Test zu einem signifikanten Ergebnis kommt, aber gleichwohl alle einzelnen Regressionsgewichte bei gleicher Irrtumswahrscheinlichkeit nicht signifikant von Null verschieden sind.
  Dies kann von Folge von sog. ***Multikollinearität*** sein, die es unmöglich macht, den einzelnen erklärenden Variablen signifikante Beiträge an der Erklärungskraft zuzuordnen, obwohl insgesamt auch in der Population mit einer erklärten Varianz größer Null zu rechnen ist
- Auch der umgekehrte Fall ist möglich: der F-Test besagt, dass alle Regressionskoeffizienten vermutlich in der Population null sind, gleichwohl gibt es einzelne Regressionsgewichte, die bei gleicher Irrtumswahrscheinlichkeit signifikant von Null verschieden sind.
  Dies erscheint paradox, liegt aber daran, dass der Test eines einzelnen Koeffizienten eine größere Trennschärfe hat als der Gesamttest aller Koeffizienten, also die Wahrscheinlichkeit, eine falsche Nullhypothese zu verwerfen, größer ist.
- Der F-Test setzt alle fünf Modellannahmen voraus. Gegenüber der Verletzung der Normalverteilungsannahme ist der Test relativ robust, solange die Homoskedastizitätsannahme erfüllt ist und keine Autokorrelation der Residuen besteht. Bei gleichen Fallzahlen (>1) bei allen Ausprägungskombinationen der erklärenden Variablen ist er auch recht robust bei Verletzung der Homoskedastizitätsannahme.

Multikollinearität

Die Schätzfunktionen sind i.a. miteinander korreliert. So beträgt die Kovarianz zwischen den beiden Regressionsgewichten:

$$\sigma(b_1, b_2) = -\frac{\sigma_U^2 \cdot SP_{XW}}{SS_X \cdot SS_W - (SP_{XW})^2}$$

Bei einer perfekten Korrelation von ±1 zwischen X und W ist die Kovarianz unbestimmt, weil dann die quadrierte Kovariation gleich dem Produkt der Variationen ist, der Nenner also Null wird.
Man bezeichnet diese Situation auch als **perfekte *Multikollinearität*.** In dieser Situation sind auch die Standardfehler der Regressionsgewichte unbestimmt bzw. unendlich groß, weil dann auch dort die Differenzen im Nenner Null sind:

$$\sigma^2(b_1) = \sqrt{\frac{SS_W}{SS_X \cdot SS_W - (SP_{XW})^2} \cdot \sigma_U^2} \; ; \; \sigma^2(b_2) = \sqrt{\frac{SS_X}{SS_X \cdot SS_W - (SP_{XW})^2} \cdot \sigma_U^2}$$

Tatsächlich gibt es dann überhaupt keine eindeutigen Werte für die Regressionskoeffizienten, denn dann weisen auch die Schätzgleichungen einen Nenner von Null auf:

$$b_1 = \frac{SS_W \cdot SP_{XY} - SP_{WY} \cdot SP_{XW}}{SS_X \cdot SS_W - (SP_{XW})^2} \; ; \; b_2 = \frac{SS_X \cdot SP_{WY} - SP_{XY} \cdot SP_{XW}}{SS_W^2 \cdot SS_X^2 - (SP_{XW})^2}$$

Multikollinearität

Die Nichtexistenz einer eindeutigen Schätzung der Regressionsgewichte bei perfekter Kollinearität ist eigentlich nicht verwunderlich, da eine Korrelation von ±1 zwischen den beiden erklärenden Variablen bedeutet, dass beide Prädiktoren dieselben Informationen enthalten und daher eine der beiden Variablen für die Vorhersage der abhängigen Variablen „überflüssig“ ist.

Problematisch ist nicht nur eine perfekte Korrelation zwischen den Prädiktoren, sondern bereits eine sehr hohe Multikollinearität, bei der die geschätzten Regressionsgewichte mit Werten über 0.95 korrelieren. In solcher Situation ist es schwer, den einzelnen Prädiktoren eindeutige direkte Effekte zuzuordnen.

Sichtbar wird dies auch an den Standardfehlern, die bei Multikollinearität stets höher sind als bei unkorrelierten Prädiktoren, da ja im Nenner der Standardfehler das Quadrat der Kovarianz bzw. Kovariation vom Produkt der Varianzen bzw. Variationen der Prädiktoren abgezogen wird.

Als Maß für die Multikollinearität eines Prädiktors wird oft die Differenz von 1.0 des Determinationskoeffizienten bei einer Regression einer erklärenden Variablen auf die übrigen erklärenden Variablen verwendet. Diese Maß wird als ***Toleranz*** bezeichnet.
Wenn ein Prädiktor eine Toleranz von null hat, liegt perfekte Multikollinearität vor. Dann lassen sich die Regressionsgewichte eines Regressionsmodells mit diesem Prädiktor überhaupt nicht eindeutig schätzen.
Toleranzwerte kleiner 0.005 weisen auf eine deutliche Multikollinearität hin.

Multikollinearität

Neben der Toleranz wird oft der Kennwert ***VIF*** *(„variance inflation factor“)* berechnet, der der Kehrwert der Toleranz ist. Er gibt an, um wie viel sich Standardfehler als Folge von Multikollinearität erhöhen.

Bei der trivariaten Regression ist die Toleranz gleich eins minus dem Quadrat der Korrelation zwischen X und W, im Beispiel also $1 – 0.173^2 = 0.970$.
Der Wert von VIF beträgt dann 1/0.970 = 1.031. Die Werte weisen darauf hin, dass keine Multikollinearitätsprobleme bestehen.

Am günstigsten ist es, wenn gar keine Multikollinearität besteht. Dann korrelieren die erklärenden Variablen nicht untereinander.
In experimentellen Untersuchungsdesigns lässt sich Multikollinearität dadurch vermeiden, dass allen Untersuchungs- und Kontrollgruppen, d.h. allen Ausprägungskombinationen aller erklärender Größen (Faktoren) die gleiche Zahl von Fällen zugeordnet wird. Man spricht dann auch von einem ***vollständigen Design***.
In ex-post-facto-Anordnungen lässt sich dies in der Regel nicht realisieren. Bei Multikollinearitätsproblemen ist es dann sinnvoll zu überlegen, warum Prädiktoren eng miteinander zusammenhängen und dies in einer geeigneten Analysestrategie etwa durch Skalenbildung zu berücksichtigen.