

Christian FG Schendera
Regressionsanalyse mit SPSS

Christian FG Schendera

Regressionsanalyse mit SPSS

2. korrigierte und aktualisierte Auflage

DE GRUYTER
OLDENBOURG

ISBN 978-3-11-035985-5
eISBN 978-3-11-037721-7

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.dnb.de> abrufbar.

Library of Congress Cataloging-in-Publication Data

A CIP catalog record for this book has been applied for at the Library of Congress.

© 2014 Oldenbourg Wissenschaftsverlag GmbH
Rosenheimer Straße 143, 81671 München, Deutschland
www.degruyter.com
Ein Unternehmen von De Gruyter

Lektorat: Dr. Stefan Giesen
Herstellung: Tina Bonertz
Titelbild: Norma Cornes/Getty Images/Hemera

Druck und Bindung: CPI books GmbH, Leck

Gedruckt in Deutschland
Dieses Papier ist alterungsbeständig nach DIN/ISO 9706.

Vorwort zur zweiten Auflage

Nach dem Erscheinen von „Regressionsanalyse mit SPSS“ war angesichts der großen Beliebtheit eine Neuauflage schnell absehbar. Immer wieder verzögert durch die erhebliche berufliche Auslastung des Verfassers (die sich, neben der eigentlichen Tätigkeit in der Forschungs- und Unternehmensberatung, u.a. auch im Abfassen von „Clusteranalyse mit SPSS“ (2010) und weiteren Büchern niederschlug), liegt nun eine aktualisierte Auflage von „Regressionsanalyse mit SPSS“ vor. Alle Analysen sind mit SPSS v21 geprüft und wo notwendig, aktualisiert, um mit der Entwicklung von SPSS auf Augenhöhe zu sein (die erste Auflage war noch für v16 geschrieben). An PLS wird eine unkomplizierte Erweiterung von SPSS v21 durch Python-Extensionen veranschaulicht. Ebenfalls wurde sichergestellt, dass die dargestellten Syntax-Beispiele auch in SPSS v21 funktionieren. Kapitel 6, das weitere Regressionsansätze für Menüs bzw. mit Syntax steckbriefartig vorstellt, wurde erweitert. Unter den dort vorgestellten Ansätzen sind nun u.a. die Automatische lineare Modellierung (SPSS Prozedur LINEAR), Probit-Analyse (PROBIT), Weighted Least Square-Methode (syn.: Gewichtungsschätzung, WLS), das Verfahren der zweistufigen kleinsten Quadrate (2SLS), verallgemeinerte lineare Modelle (u.a. für Zähldaten, z.B. Poisson- und Gamma-Regression) (GENLIN), eine quantitative kategoriale Regression (syn.: Optimale Skalierung, CATREG) sowie gemischte Modelle (MIXED, GENLINMIXED), z.B. Zufallskoeffizientenansätze und Syntaxbeispiele für Mehrebenenmodelle. SPSS (als Firma) wurde zwischenzeitlich von IBM aufgekauft. Entsprechend der Branding-Richtlinien heißt SPSS (als Produkt) zwischenzeitlich „IBM SPSS Statistics“. Zwecks Lesbarkeit wird im Text weiterhin kurz und knapp „SPSS“ verwendet; auch, um Nutzer älterer Software-Versionen mit anzusprechen. Die Kommunikation mit der Leserschaft trug zur Qualität dieser Neuauflage mit bei: Stellvertretend für viele Feedbacks seien die engagierten Rückmeldungen und Anregungen genannt von Dipl.-Psych. Jesko Kaltenbaek (Berlin), Mag. Antonia Griesbacher (Graz) und Ewald Gassner (Zürich). Auch über Rückmeldungen zu der hier vorliegenden Neuauflage würde ich mich sehr freuen. Die Emailadresse finden Sie gegen Ende des Buches.

Herrn Dr. Giesen danke ich für das Engagement, von diesem Buch nun auch bei DeGruyter eine zweite Auflage zu veröffentlichen. Nachträglich sei auch den Drs. Adzerßen und Strowitzki von der Heidelberger Brustkrebsstudie gedankt. Meiner Frau, Yun, danke ich für ihre Geduld, Weitsicht und Verständnis. Daten und Syntax zu „Regressionsanalyse mit SPSS“ können von der Webseite des Verfassers unter www.method-consult.ch heruntergeladen werden.

Hergiswil, Januar 2014

Dr. CFG Schendera

Vorwort

Die Regressionsanalyse und ihre Varianten gehört zu den in Wissenschaft und Forschung am häufigsten eingesetzten statistischen Verfahren (z.B. Hsu, 2005; Pötschke & Simonson, 2003; Elmore & Woehlke, 1998, 1996; Goodwin & Goodwin, 1985), wobei dies einer ständigen zu fordernden und zu fördernden Weiterentwicklung von Fachbereichen, Forschung, wie auch der Statistik selbst unterliegt (vgl. z.B. Rigby et al., 2004; Ripoll et al. 1996).

Von einigen Autoren wird die Regressionsanalyse auch zu einem der ältesten statistischen Verfahren gezählt. Stanton (2001) führt die grundlegende statistische Entwicklung der linearen Regression v.a. auf die Arbeiten von Karl Pearson (z.B. 1896) zurück. Howarth (2001) führt z.B. die Regressionsanalyse und verwandte Verfahren sogar bis auf eine Arbeit von Bond im Jahre 1636 zurück (vgl. auch Finney, 1996).

Meilensteine in der Entwicklungsgeschichte der Regressionsanalyse waren grundlegende Rechenansätze und damit im Zusammenhang stehend die Entdeckung des „Kleinste Quadrate“-Ansatzes in der Mitte des 18. Jahrhunderts. Erst die Erfindung von Computern in der Mitte des 20. Jahrhunderts übernahm den mit der Anwendung multivariater regressionsanalytischer Ansätze verbundenen immensen (und fehleranfälligen) Rechenaufwand und beschleunigte dadurch die Anwendung, Weiterverbreitung und Weiterentwicklung u.a. regressionsanalytischer Verfahren. Als weitere (willkürlich ausgewählte) Meilensteine in der Weiterentwicklung regressionsanalytischer Ansätze können z.B. die Ridge-Regression in den 70er Jahren, die robuste Regression in den 80er Jahren sowie die generalisierten bzw. gemischten Ansätze in den 90er Jahren des vergangenen Jahrhunderts gezählt werden. Die Regressionsanalyse als scheinbar traditionelles Verfahren wird auch beim Data Mining jedoch nicht nur gleichrangig neben neueren Ansätzen, wie z.B. neuronalen Netzen, eingesetzt (z.B. SPSS, 2012b, Kap. 10; Rud, 2001; Berry & Linoff, 2000; Graber, 2000), sondern z.T. auch weit häufiger als andere Ansätze (z.B. Rexer et al. 2007; Ayres, 2007). Auch für die bedienungsfreundlichste Data Mining-Anwendung gilt jedoch: Data Mining ersetzt keine Statistik- oder auch Informatikkenntnisse, sondern setzt diese voraus (Schendera, 2007; Khabaza, 2005; Chapman et al., 1999). Die Grundlagen, die dieses Buch für das Anwenden regressionsanalytischer Ansätze in SPSS bereitet, sollten daher auch erste Schritte im Data Mining-Bereich ermöglichen.

Einsteiger in die Statistik könnten überrascht sein, wie mächtig, vielfältig und flexibel Verfahrensfamilien wie die „Regression“ sein können. Für Fortgeschrittene mag es (immer wieder von Neuem) interessant sein, wie mittels regressionsanalytischer Verfahren zahlreiche heterogene Fragestellungen bearbeitet werden können, u.a. einfache und multiple (non)lineare Regressionsanalysen, individuelle Wachstumskurven, Überlebenszeitanalysen, Zeitreihenanalysen usw. Diesem Anwendungsspektrum steht auf der anderen Seite das Problem gegenüber, dass man nicht ohne Weiteres gebetsmühlenartig herunterleiern kann: „Für ein lineares Kausalmodell mit einer intervallskalierten abhängigen Variable verwendet man eine lineare Regression, für ein Kausalmodell mit einer dichotomen abhängigen Variablen z.B. eine logistische Regression usw.“ Oder: „Für eine lineare Regression verwendet man die SPSS Prozedur REGRESSION, für logistische Regressionen nimmt man die SPSS Proze-

duren LOGISTIC oder NOMREG, oder für Überlebensdaten z.B. SURVIVAL, KM oder COXREG usw.“

Im Gegenteil, die Statistik wie auch SPSS sind hochgradig komplex, flexibel und vielseitig. Die Prozedur REGRESSION kann von fortgeschrittenen Anwendern u.a. auf lineare, kollineare Daten oder auch Zeitreihendaten angewandt werden. Die SPSS Prozedur GLM ermöglicht z.B. sowohl die Berechnung einer Varianzanalyse im Rahmen des Allgemeinen Linearen Modells (ALM), aber auch die Durchführung einer Regressionsanalyse mit *mehreren* abhängigen Variablen. Die relativ neue Prozedur GENLIN (seit SPSS Version 15) ermöglicht z.B. als *verallgemeinertes lineares Modell* sowohl die Berechnung einer Gamma-Regression, einer binären logistischen Regression im Messwiederholungsdesign oder auch die einer komplementären Log-Log-Regression für intervallzensierte Überlebensdaten.

Zusätzlich ist die Vielfalt der Statistik deutlich größer als selbst der Funktionsumfang von SPSS. Überlebensdaten können z.B. auch in *ein* erwartetes Risiko (konventionelle Überlebenszeitanalyse, survival analysis), *mehrere* erwartete exklusiv-disjunkte Risiken (competing risk survival analysis) oder auch *wiederkehrende* Risiken (recurrent risk survival analysis) differenziert werden.

Die Wahl des angemessenen regressionsanalytischen Verfahrens hängt somit *nicht* nur von zur Verfügung stehenden SPSS Menüs, Prozeduren oder „Kochrezepten“ ab, sondern konkret von inhaltlichen und methodologischen Aspekten, z.B. von der untersuchenden Fragestellung (Hypothese; es gibt versch. Arten, z.B. Unterschied vs. Zusammenhang), aber auch von festzulegenden Definitionen, wie z.B. Messniveau der Daten, Verteilungen, Transformationen, (Un-)Verbundenheit der Daten, die Modellierung von Haupt- und Interaktionseffekten und vielem anderen mehr ab. Die Auswahl sollte in Absprache mit einem erfahrenen Methodiker bzw. Statistiker erfolgen. Bei speziellen Fragestellungen ist es möglich, dass die erforderlichen statistischen Verfahren nicht in der Standardsoftware implementiert sind. In diesem Falle kann das Verfahren z.B. mit SPSS oder Python selbst programmiert (vgl. u.a. das Makro „Ridge-Regression.sps“) oder auch auf spezielle Analysesoftware ausgewichen werden. Das Vorgehen bei der Modellspezifikation und inferenzstatistischen Hypothesentestung entspricht üblicherweise einer schrittweisen Komplexitätssteigerung (vgl. Schendera, 2007, 401–403).

Dieses Buch führt ein in die grundlegenden Verfahren Korrelation, Regression (linear, multipel, nichtlinear), logistische (binär, multinomial) und ordinale Regression sowie die Überlebenszeitanalyse (Sterbetafel-Methode, Kaplan-Meier-Ansatz sowie Regressionen nach Cox). Weitere Abschnitte behandeln zusätzliche regressionsanalytische Ansätze und Modelle (z.B. die Modellierung individueller Wachstumskurven, PLS-Regression, Ridge-Regression, Gematchte Fall-Kontrollstudie).

Das Ziel dieser Einführung in die Regressionsanalyse mit SPSS ist, eine erste Übersicht und eine tragfähige Grundlage zu schaffen, auf deren Grundlage der dann routinierte Anwender auch im Anwendungsbereich des Data Mining, z.B. mittels Clementine, alleine voranschreiten kann. Viele weitere Regressionsvarianten und -anwendungen (z.B. nichtparametrische Regression, kategoriale Regression, Weibull-Regression, hedonische Regression usw.) können und sollen aus Platzgründen nicht vorgestellt werden (vgl. dazu *Kapitel 6*).

Dieses Buch führt ein in die grundlegenden Ansätze Korrelation (*Kapitel 1*), Regression (linear, multipel, nichtlinear; *Kapitel 2*), logistische und ordinale Regression (*Kapitel 3*) sowie die Überlebenszeitanalyse (Survivalanalyse; *Kapitel 4*). Bei allen Ansätzen werden

Voraussetzungen und häufig begangene Fehler ausführlich erläutert. *Kapitel 5* stellt speziellere Anwendungen der Regression vor (Partial-Regression, Modellierung sog. individueller Wachstumskurven, Ridge-Regression). *Kapitel 6* weist auf weitere Möglichkeiten mit SPSS hin (z.B. Regressionsanalysen für mehrere abhängige Variablen).

Zahlreiche Rechenbeispiele werden von der Fragestellung, der Anforderung der einzelnen Statistiken (per Maus, per Syntax) bis hin zur Interpretation der SPSS Ausgaben systematisch durchgespielt. Es wird auch auf diverse Fehler und Fallstricke eingegangen. Zur Prüfung von Daten vor der Durchführung einer statistischen Analyse wird auf „Datenqualität mit SPSS“ (Schendera, 2007) verwiesen.

Separate Abschnitte stellen die diversen Voraussetzungen für die Durchführung der jeweiligen Analyse sowie Ansätze zu ihrer Überprüfung zusammen. Dieses Buch ist verständlich und anwendungsorientiert geschrieben, ohne jedoch die Komplexität und damit erforderliche Tiefe bei der Erläuterung der Verfahren zu vernachlässigen. Dieses Buch ist für Einsteiger in die Regressionsanalyse, Studierende sowie fortgeschrittene Wissenschaftler und Anwender in den Wirtschafts-, Bio- und Sozialwissenschaften gleichermaßen geeignet.

Das Buch wurde sowohl für die Menüführung, aber auch für SPSS Syntax (derzeit Version 16) entwickelt. Einsteiger in SPSS für Windows sollten dabei wissen, dass SPSS Syntax über Mausclicks automatisch angefordert werden kann bzw. selbst einfach zu programmieren ist (vgl. Schendera, 2007, 2005). Für SPSS Programmierer (aber auch Mauslenker!) wird die Erweiterung von SPSS über Python anhand der Prozedur PLS vorgeführt. Die (u.a.) Ridge-Regression wird anhand des gleichnamigen SPSS-Makros vorgestellt.

Kapitel 1 führt ein in die erfahrungsgemäß unterschätzte Korrelationsanalyse (SPSS Prozedur CORRELATIONS). Gleich zu Beginn beschäftigt sich das Kapitel mit der Interpretation von Zusammenhängen (Kausalität) und stellt mehrere Beispiele von Fehlschlüssen vor, u.a. den oft zitierten Zusammenhang zwischen dem Konsum von gewalttätigen Computerspielen und Gewaltbereitschaft. Jedem Leser, der sich für die Regressionsanalyse interessiert, wird dringend empfohlen, zuvor das Kapitel zur Korrelationsanalyse zu lesen. Anhand der Korrelationsanalyse werden erste, auch für die (lineare) Regressionsanalyse geltende Voraussetzungen wie z.B. Skalenniveau, Homoskedastizität und Kontinuität erläutert. Weitere Abschnitte behandeln die Themen Linearität, Scheinkorrelation und Alphafehler-Kumulation. Auch wird erläutert, warum die bloße Angabe eines Korrelationskoeffizienten im Prinzip Unfug ist. Als spezielle Anwendungen werden u.a. der Vergleich von Korrelationskoeffizienten und die Kanonische Korrelation vorgestellt. Ein abschließender Abschnitt stellt die diversen Voraussetzungen für die Durchführung der Korrelationsanalyse sowie Ansätze zu ihrer Überprüfung zusammen.

Kapitel 2 führt ein in die Regressionsanalyse. Das Kapitel ist schrittweise aufgebaut, um das Grundprinzip beim Durchführen einer Regressionsanalyse transparent zu machen und so den Lesern zu helfen, oft begangene fundamentale Fehler von Anfang an zu vermeiden.

Kapitel 2.1 führt zunächst ein in die einfache, lineare Regressionsanalyse (SPSS Prozedur REGRESSION). Kapitel 2.1 setzt *Kapitel 1* voraus. An einem einfachen Beispiel wird die Überprüfung der Linearität und Identifikation von Ausreißern anhand von Hebelwerten und Residuen erläutert. Auch wird das Überprüfen auf eine möglicherweise vorliegende Autokorrelation erläutert. Im Allgemeinen kann nur eine lineare Funktion mittels einer linearen Re-

gressionsanalyse untersucht werden. Eine *nichtlineare* Funktion mittels einer linearen Regressionsanalyse zu untersuchen mündet in fehlerhaften Ergebnissen.

Kapitel 2.2 erläutert, was getan werden kann, wenn die Daten nicht linear, sondern kurvenförmig verteilt sind. *Kapitel 2.2* setzt *Kapitel 2.1* voraus. Das Kapitel bietet zwei Lösungsmöglichkeiten an: Eine nichtlineare Funktion kann linearisiert und mittels einer linearen Regression analysiert werden. Alternativ kann eine nichtlineare Funktion mittels einer nichtlinearen Regression geschätzt werden (SPSS Prozeduren CNLR und NLR). Die nichtlineare Regression ist das zentrale Thema dieses Kapitels, einschließlich einer nichtlinearen Regression mit zwei Prädiktoren. Darüber hinaus werden Sinn und Grenzen der SPSS Prozedur CURVEFIT für die (non)lineare Kurvenanpassung erläutert. Abschließende Abschnitte stellen die diversen Annahmen der nichtlinearen Regression sowie in einer Übersicht eine Auswahl der bekanntesten Modelle einer nichtlinearen Regression mit einem oder mehr Prädiktoren zusammen.

Kapitel 2.3 führt ein in die multiple lineare Regressionsanalyse (SPSS Prozedur REGRESSION). *Kapitel 2.3* setzt *Kapitel 2.2* voraus. Die Tatsache, dass mehrere anstelle nur einer unabhängigen Variablen im Modell enthalten sind, bedingt Besonderheiten, die v.a. die Verhältnisse der unabhängigen Variablen untereinander betreffen. Dieses Kapitel geht u.a. auf die Themen Modellbildung, Variablenselektion, Multikollinearität und andere Fallstricke ein. Neben dem Identifizieren und Beheben von Multikollinearität wird auch auf das Umgehen mit zeitabhängigen (autoregressiven) Daten eingegangen. Ein abschließender Abschnitt stellt die diversen Voraussetzungen für die Durchführung der (non)linearen Regressionsanalyse sowie Ansätze zu ihrer Überprüfung zusammen.

Kapitel 3 führt ein in die grundlegenden Verfahren der logistischen und ordinalen Regression. Das Kapitel ist nach dem Skalenniveau der abhängigen Variablen aufgebaut. Separate Abschnitte stellen abschließend jeweils die diversen Voraussetzungen der vorgestellten Ansätze zusammen sowie Ansätze zu ihrer Überprüfung.

Die binäre logistische Regression (SPSS Prozedur LOGISTIC REGRESSION, *Kapitel 3.2*) erwartet eine zweistufige abhängige Variable; Ranginformationen in der abhängigen Variablen werden vom Verfahren nicht berücksichtigt. Als grundlegendes Verfahren wird zunächst die binäre logistische Regression vorgestellt und die Gemeinsamkeiten und Unterschiede zu anderen Verfahren erläutert (u.a. Modell, Skalenniveau). Anhand mehrerer Rechenbeispiele werden u.a. die unterschiedlichen Verfahren der Variablenselektion sowie die Interpretation der ausgegebenen Statistiken erläutert. Abschließend wird auf das häufige Auseinanderklaffen von Modellgüte und Vorhersagegenauigkeit eingegangen.

Die ordinale Regression (SPSS Prozedur PLUM, *Kapitel 3.3*) erwartet eine *mindestens* zweistufige (ordinal skalierte) abhängige Variable; Ranginformationen in der abhängigen Variablen werden berücksichtigt. Die Gemeinsamkeiten und Unterschiede mit den anderen Verfahren werden erläutert (u.a. Modell, Skalenniveau). Anhand mehrerer Rechenbeispiele werden u.a. die Interpretation der SPSS Ausgaben für Modelle mit intervallskalierten und kategorial skalierten Prädiktoren erläutert.

Die multinomiale logistische Regression (SPSS Prozedur NOMREG, *Kapitel 3.4*) erwartet ebenfalls eine mindestens zweistufige, jedoch nominal skalierte abhängige Variable; Ranginformationen in der abhängigen Variablen werden vom Verfahren nicht berücksichtigt. Die multinomiale logistische Regression wird analog zu 3.2 behandelt. Zusätzlich wird der Spezialfall der Gematchten Fall-Kontrollstudie (1:1) mit metrischen Prädiktoren vorgestellt.

Kapitel 4 führt ein in die Verfahren der Überlebenszeitanalyse. Die Überlebenszeitanalyse untersucht im Prinzip die Zeit bis zum Eintreten eines definierten Zielereignisses. Dies kann ein erwünschtes Ereignis (z.B. Vertragsverlängerung, Anstellung, Lernerfolg, Heilung usw.) oder auch ein unerwünschtes Ereignis sein (z.B. Kündigung, Defekt, Rückfall, Tod usw.). Aus der unterschiedlichen Bewertung der Zielereignisse rühren auch die ausgesprochen heterogenen Bezeichnungen dieser Verfahrensgruppe, z.B. Survivalanalyse, Lebenszeitanalyse, time to effect bzw. event Analyse etc. Je nach Bewertung der Zielereignisse sind jedoch v.a. die Diagramme unterschiedlich zu interpretieren.

Kapitel 4.1 stellt zunächst das Grundprinzip der Überlebenszeitanalyse sowie beispielhafte Fragestellungen und die Ziele einer Überlebenszeitanalyse vor.

Kapitel 4.2 erläutert die Bestimmung der verschiedenen Überlebensfunktionen (u.a. kumulative $S(t)$, Eins-minus-Überlebensfunktion $(1-S(t))$, Dichtefunktion $f(t)$, logarithmierte Überlebensfunktion $l(t)$ sowie Hazard-Funktion $h(t)$).

Kapitel 4.3 führt in die Zensierung von Daten ein. Bei einer Überlebenszeitanalyse kann es vorkommen, dass bei manchen Fällen das Zielereignis *nicht wie erwartet* eintritt, d.h. das Zielereignis tritt *gar nicht* oder *nicht aus den erwarteten* (definierten) Gründen ein. Um diese Fälle von denjenigen mit den erwarteten Ereignissen abzugrenzen, werden sie anhand sog. Zensierungen markiert. Links-, Rechts- und Intervallzensierung werden vorgestellt, ebenso wie das Interpretieren von Zensierungen im Rahmen eines (non)experimentellen Untersuchungsdesigns.

Kapitel 4.4 erläutert am Beispiel der versicherungsmathematischen Methode (Sterbetafel-Ansatz) und der Kaplan-Meier-Methode, wie diese Verfahren die Überlebensfunktion rechnerisch ermitteln und dabei mit zensierten Fällen umgehen. Beispiele mit SPSS werden ab *Kapitel 4.6* vorgestellt.

Kapitel 4.5 stellt diverse Tests für den Vergleich zwischen Gruppen vor: Log Rang-Test (syn.: Log Rank- bzw. Mantel-Cox-Test), Breslow-Test (syn: modifizierter Wilcoxon-Test, Wilcoxon-Gehan-Test), Tarone-Ware-Test und Likelihood-Ratio-Test. Dieses Kapitel stellt darüber hinaus eine vergleichende Zusammenfassung sowie auch Empfehlungen für die Interpretation dieser Tests zusammen.

In *Kapitel 4.6* werden die versicherungsmathematische Methode (SPSS Prozedur SURVIVAL) und der Kaplan-Meier-Ansatz (SPSS Prozedur KM) mit SPSS gerechnet und interpretiert. Die Regressionen nach Cox werden in einem eigenen Kapitel behandelt. Bei der Sterbetafel-Methode werden Beispiele mit bzw. ohne Faktoren vorgestellt. Beim Kaplan-Meier-Ansatz werden Beispiele mit/ohne Faktoren, mit Schichtvariablen sowie für die Ermittlung von Konfidenzintervallen vorgestellt.

Kapitel 4.7 führt zunächst in die Besonderheiten des Cox-Modells ein (SPSS Prozedur COXREG) und vergleicht diesen Ansatz mit Sterbetafelmethode, Kaplan-Meier und linearer Regression. Anschließend werden mehrere Varianten der Regression nach Cox (zeitunabhängige Kovariaten, zeitabhängige Kovariaten, Interaktionen sowie sog. „Muster“) mit SPSS gerechnet und interpretiert. Separate Abschnitte stellen die Verfahren zur Überprüfung der speziellen Voraussetzungen der Cox-Regression (u.a. Analyse von Zensierungen, Multikollinearität und Proportionalitätsannahme) sowie Möglichkeiten der Bildung von Kontrasten vor („Abweichung“, „Einfach“, „Helmert“ usw.). Weitere Abschnitte stellen abschließend jeweils die diversen Voraussetzungen der vorgestellten Verfahren zusammen sowie Ansätze zu ihrer Überprüfung.

Kapitel 5 stellt anhand von exemplarischen SPSS Analysen weitere Anwendungsmöglichkeiten regressionsanalytischer Ansätze vor.

Kapitel 5.1 stellt die Partial-Regression in zwei Varianten vor. Kapitel 5.1.1 stellt die Partial-Regression mittels partieller kleinster Quadrate (Partial Least Squares, PLS) vor. Die PLS-Regression kann u.a. dann eingesetzt werden, wenn viele Prädiktoren vorliegen, wenn die Einflussvariablen untereinander hoch korrelieren und/oder wenn die Anzahl der Einflussvariablen die Anzahl der Fälle übersteigt. Die PLS-Regression vereinigt in sich Merkmale der Hauptkomponentenanalyse und der multiplen Regression und ermöglicht dadurch, Kausalverhältnisse zwischen einer beliebigen Anzahl (latenter) Variablen mit beliebigem Messniveau als *lineare* Strukturgleichungsmodelle zu modellieren. PLS unterstützt darüber hinaus gemischte Regressions- und Klassifikationsmodelle. Die unabhängigen sowie abhängigen Variablen können intervall- oder kategorialskaliert sein. Der Befehl PLS steht erst ab SPSS Version 16 zur Verfügung. PLS basiert auf einer Python-Erweiterung. In Kapitel 5.1.2 wird eine Variante der Partial-Regression auf der Basis der Korrelationsanalyse mittels der SPSS Prozedur REGRESSION vorgestellt.

Kapitel 5.2 stellt die lineare Modellierung sog. individueller Wachstumskurven anhand des Ansatzes linearer gemischter Modelle (SPSS Prozedur MIXED) vor. Individuelle Wachstumskurven (individual growth modeling) können im vorgestellten Fall in etwa als „Varianzanalysen mit Messwiederholung für Individuen“ umschrieben werden. Bei der „normalen“ linearen Regression würde eine einzige Regressionslinie (wie sie z.B. auch eine Varianzanalyse mit Messwiederholung erzeugen würde) unterschiedlichen individuellen (linearen) Verläufen oft nicht gerecht werden. Eine vorgeschaltete Modellierung mittels des Zufallskoeffizientenmodells ermöglicht jedoch, individuelle Verläufe anhand von Intercept, Steigung und beiden Parametern zugleich zu schätzen. Anhand einer dreistufigen Beispielanalyse wird demonstriert, ob und inwieweit sich Teilnehmer an einem Trainingsprogramm über die Zeit hinweg in ihrer Performanz unterscheiden. Im Beispiel wird konkret geprüft: (a) bewegen sich die Trainingsteilnehmer auf unterschiedlichen (Leistungs-)Niveaus (Intercept), (b) verbessern sie sich unterschiedlich gut bzw. schnell (Steigung), bzw. (c) verbessern sich die Trainingsteilnehmer unterschiedlich unter zusätzlicher Berücksichtigung ihres Leistungsniveaus (beide Parameter)?

Kapitel 5.3 stellt die Ridge-Regression (SPSS Makro „Ridge-Regression.sps“) vor. Die Ridge-Regression ist eine (u.a. visuelle) Möglichkeit, potentiell multikollineare Daten auf eine Analysierbarkeit mittels der multiplen linearen Regressionsanalyse zu überprüfen. Im Gegensatz zu den anderen statistischen Verfahren wird die Ridge-Regression von SPSS nicht über Menü-Führung, sondern ausschließlich in Form eines Makros angeboten. Die Durchführung einer Ridge-Regression ist jedoch unkompliziert. Dieses Kapitel demonstriert u.a. die Visualisierung von Multikollinearität, wie auch die Berechnung einer Ridge-Regression für einen ausgewählten K-Wert.

Kapitel 6 stellt in einer Übersicht weitere Möglichkeiten des Durchführens von Regressionsanalysen mit SPSS vor (z.B. Regressionsanalysen für mehrere abhängige Variablen). Im Gegensatz zu den vorangegangenen Kapiteln werden keine exemplarischen SPSS Analysen durchgerechnet. Diese Zusammenstellung hat keinen Anspruch auf Vollständigkeit. Zur Illustration werden unkommentierte Syntaxbeispiele für exemplarische Analyse zusammengestellt, u.a. weil z.T. die Anforderung nur auf diese Weise möglich bzw. transparent ist.

Zur Beurteilung der SPSS Ausgaben sind Kenntnisse ihrer statistischen Definition und Herleitung unerlässlich. In einem abschließenden Kapitel sind ausgewählte Formeln der wichtigsten behandelten Verfahren zusammengestellt.

Bevor Sie sich an die Analyse machen, stellen Sie bitte *zuvor* sicher, dass Ihre Daten analysereif sind. Prüfen Sie Ihre Daten auf mögliche Fehler (u.a. Vollständigkeit, Einheitlichkeit, Missings, Ausreißer, Doppelte). Vertrauen ist gut; Kontrolle ist besser. Für Kriterien zu Datenqualität und ihre Überprüfung mit SPSS wird der interessierte Leser auf Schendera (2007) verwiesen.

... für das Lächeln im Knurren des chinesischen Drachens ...

Zu Dank verpflichtet bin ich für fachlichen Rat und/oder auch einen Beitrag in Form von Syntax, Daten und/oder auch Dokumentation unter anderem: Prof. Vijay Chatterjee (Mount Sinai Medical School, New York University, USA), Prof. Mark Galliker (Universität Bern, Schweiz), Prof. Jürgen Janssen (Universität Hamburg), Prof. Mitchel Klein (Emory University, Rollins School of Public Health, Atlanta), Prof. Roderick J.A. Little (University of Michigan USA), Prof. Daniel McFadden (University of Berkeley USA), Prof. Rainer Schlittgen (Universität Hamburg), Prof. Stephen G. West (Arizona State University USA), Matthew M. Zack (Centers for Disease Control, Atlanta, Georgia USA).

Mein Dank an SPSS Deutschland geht stellvertretend an Herrn Alexander Bohnenstengel sowie Frau Sabine Wolfrum und Frau Ingrid Abold von der Firma SPSS GmbH Software (München) für die großzügige Bereitstellung der Software und der technischen Dokumentation. Gleichmaßen geht mein Dank an SPSS Schweiz, namentlich an Josef Schmid und Dr. Daniel Schloeth.

Herrn Dr. Schechler vom Oldenbourg Verlag danke ich für das Vertrauen, auch dieses Buch zu veröffentlichen sowie die immer großzügige Unterstützung. Peter Bonata (Köln) schuf die Grundlagen für das Kapitel zur Cox-Regression. Volker Stehle (Eppingen) gestaltete die Druckformatvorlage. Stephan Lindow (Hamburg) entwarf die Grafiken. Markus Schreiner (Heidelberg) stellte freundlicherweise Zufallsdaten spezieller Verteilungen zur Verfügung. Falls in diesem Buch noch irgendwas unklar oder fehlerhaft sein sollte, so liegt die Verantwortung alleine beim Autor.

Bern, Mai 2008

CFG Schendera

Inhalt

Vorworte	V
1 Korrelation	1
1.1 Einführung.....	1
1.2 Erste Voraussetzung: Das Skalenniveau	6
1.3 Weitere Voraussetzungen: Linearität, Homoskedastizität und Kontinuität.....	7
1.4 Exkurs: Grafische Tests auf Linearität	8
1.4.1 Prozedur GRAPH, Scatterplot Option.....	8
1.4.2 SPSS Prozedur CURVEFIT	10
1.5 Statistik und Interpretation des Korrelationskoeffizienten	14
1.5.1 Statistik des Korrelationskoeffizienten.....	14
1.5.2 Interpretation des Korrelationskoeffizienten	15
1.6 Berechnung mit SPSS (Beispiel).....	19
1.7 Fallstricke: Linearität, Scheinkorrelation und Alphafehler-Kumulation	21
1.7.1 Scheinkorrelation und partielle Korrelation	22
1.7.2 Das Problem der Alphafehler-Kumulierung.....	25
1.8 Spezielle Anwendungen	28
1.8.1 Vergleich von Korrelationskoeffizienten	28
1.8.2 Vergleich von Korrelationen auf Gleichheit	29
1.8.3 Kanonische Korrelation.....	30
1.9 Voraussetzungen für die Berechnung des Pearson-Korrelationskoeffizienten.....	32
2 Lineare und nichtlineare Regression	35
2.1 Lineare Regression: Zusammenhang mit Kausalrichtung.....	36
2.1.1 Bivariate lineare Regression: Einführung in die Regressionsanalyse mit REGRESSION	36
2.1.2 Beispiel und Syntax für eine bivariate lineare Regression – Durchgang 1: Überprüfung der Linearität und Identifikation von Ausreißern anhand von Hebelwerten und Residuen.....	43
2.1.3 Output und Erläuterungen	49

2.1.4	Durchgang 2: Der Effekt des Ausschlusses von Ausreißern – Ausgewählter Output	67
2.1.5	Exkurs: Grafik mit eingezeichneter Regressionsgerade (GGRAPH).....	69
2.2	Nichtlineare einfache Regression.....	70
2.2.1	Eine lineare Funktion wird mittels einer linearen Regressionsanalyse untersucht... 71	71
2.2.2	Eine nichtlineare Funktion wird mittels einer linearen Regressionsanalyse untersucht.....	73
2.2.3	Eine nichtlineare Funktion wird linearisiert und mittels einer linearen Regression untersucht.....	74
2.2.4	Eine nichtlineare Funktion wird mittels einer nichtlinearen Regression untersucht: Nichtlineare Regression	76
2.2.5	Etwas Anspruchsvolleres: Eine nichtlineare Regression mit zwei Prädiktoren	91
2.2.6	Die Prozeduren NLR und CNLR für die nichtlineare Regression	93
2.2.7	Annahmen der nichtlinearen Regression.....	98
2.2.8	Übersicht: Modelle für die nichtlineare Regression.....	99
2.3	Multiple lineare Regression: Multikollinearität und andere Fallstricke	102
2.3.1	Besonderheiten der Multiplen Regression	103
2.3.2	Ein erstes Beispiel: Die Interpretation der speziellen Statistiken der multiplen Regression.....	106
2.3.3	Zweites Beispiel: Identifizieren und Beheben von Multikollinearität.....	124
2.4	Voraussetzungen für die Berechnung einer linearen Regression	132
3	Logistische und ordinale Regression	141
3.1	Einführung: Kausalmodell und Messniveau der abhängigen Variable	142
3.2	Binäre logistische Regression	144
3.2.1	Das Verfahren und Vergleich zu anderen Verfahren	144
3.2.2	Beispiel, Maus und Syntax: Schrittweise Methode (BSTEP)	150
3.2.3	Output und Interpretation	155
3.2.4	Beispiel und Syntax: Direkte Methode ENTER.....	164
3.2.5	Output und Interpretation	165
3.2.6	Exkurs: Theoretetest vs. Diagnostik bei der logistischen Regression: Modellgüte (goodness of fit) vs. Prädikative Effizienz (predictive efficiency).....	170
3.2.7	Voraussetzungen für eine binäre logistische Regression	170
3.3	Ordinale Regression.....	178
3.3.1	Das Verfahren und Vergleich mit anderen Verfahren.....	179
3.3.2	Beispiel 1, Maus und Syntax: Intervallskalierte Prädiktoren (WITH-Option).....	182
3.3.3	Output und Interpretation	186
3.3.4	Beispiel 2 und Syntax: Kategoriale Prädiktoren (BY-Option).....	192
3.3.5	Output und Interpretation	193
3.3.6	Voraussetzungen für eine ordinale Regression	201
3.4	Multinomiale logistische Regression	204
3.4.1	Beispiel, Maus und Syntax: Haupteffekt-Modell (dichotome AV).....	204

3.4.2	Output und Interpretation	213
3.4.3	Exkurs: Schrittweise Berechnung eines Modells mit einer dichotomen AV: Vergleich der NOMREG- und LOGISTIC REGRESSION-Ausgaben.....	218
3.4.4	Spezialfall: Gematchte Fall-Kontrollstudie (1:1) mit metrischen Prädiktoren – Beispiel, Syntax, Output und Interpretation	219
3.4.5	Exkurs: LOGISTIC REGRESSION vs. NOMREG (Unterschiede).....	224
3.4.6	Voraussetzungen für eine multinomiale logistische Regression	225
3.5	Vergleich der vorgestellten Regressionsansätze.....	231
4	Survivalanalysen	233
4.1	Einführung in die Überlebenszeitanalyse	234
4.2	Das Grundprinzip der Survivalanalyse.....	237
4.2.1	Die Überlebensfunktion $S(t)$	237
4.2.2	Die Bestimmung der Überlebensfunktion $S(t)$	238
4.2.3	Weitere Funktionen	240
4.3	Zensierte Daten	243
4.3.1	Unerwartete Ereignisse oder nicht eintretende Zielereignisse.....	243
4.3.2	Drei Gründe, zensierte Daten anders als nichtzensierte Daten zu behandeln.....	244
4.3.3	Umgehen mit ausfallenden Daten bzw. Zensierungen (drei Ansätze).....	245
4.4	Verfahren zur Schätzung der Überlebenszeit $S(t)$	246
4.4.1	Versicherungsmathematische Methode bzw. Sterbetafel-Methode.....	247
4.4.2	Schätzung von $S(t)$ mit der Kaplan-Meier-Methode	248
4.4.3	Beispiele ohne und mit Zensierungen (Ansatz: Kaplan-Meier)	249
4.5	Tests für den Vergleich mehrerer Gruppen	253
4.6	Überlebenszeitanalyse mit SPSS.....	256
4.6.1	Beispiel: Kaplan-Meier-Verfahren ohne Faktor	256
4.6.2	Beispiel: Kaplan-Meier-Verfahren mit Faktor	263
4.6.3	Vergleiche mittels mit Faktor- und Schichtvariablen (Kaplan-Meier).....	269
4.6.4	Konfidenzintervalle für Kaplan-Meier Analysen	275
4.6.5	Beispiel einer Sterbetafel-Berechnung ohne Faktor	277
4.6.6	Beispiel einer Sterbetafel-Berechnung mit Faktor	281
4.6.7	Erste Voraussetzungen für die Berechnung einer Survivalanalyse	285
4.7	Regression nach Cox	288
4.7.1	Einführung und Hintergrund des Cox-Modells	288
4.7.2	Cox-Regression mit einer metrischen Kovariaten.....	294
4.7.3	Cox-Regression mit einer dichotomen Kovariaten ($k=2$).....	304
4.7.4	Cox-Regression mit einer kategorialen Kovariaten ($k>2$).....	308
4.7.5	Cox-Regressionen für Interaktionen.....	312
4.7.6	Überprüfung der Voraussetzungen einer Cox-Regression	328
4.7.7	Cox-Regression mit zeitabhängigen, metrischen Kovariaten.....	336
4.7.8	Spezielle Voraussetzungen der Cox-Regression	343
4.7.9	Anhang: Kontraste.....	347

5	Weitere Anwendungsbeispiele der Regressionsanalyse	353
5.1	Partial-Regression	354
5.1.1	Berechnung mit der Prozedur PLS (Python Extension).....	356
5.1.2	Berechnung mit der Prozedur REGRESSION	366
5.2	Individuelle Wachstumskurven (individual growth modeling).....	369
5.2.1	Ansatz 1: Modell zufälliger Intercepts	372
5.2.2	Ansatz 2: Modell zufälliger Steigungen.....	377
5.2.3	Ansatz 3: Modell zufälliger Intercepts und zufälliger Steigungen	379
5.3	Ridge-Regression (SPSS Makro)	383
5.3.1	Visualisierung von Multikollinearität mittels Ridge-Trace.....	383
5.3.2	Berechnung einer Ridge-Regression	387
5.3.3	Das SPSS Makro „Ridge-Regression“	389
6	Weitere Ansätze und Modelle (Ausblick)	397
6.1	Weitere Regressionsansätze über SPSS Menüs	397
6.2	Weitere Regressionsvarianten über Syntax	407
7	Anhang: Formeln	409
8	Literatur	425
9	Ihre Meinung zu diesem Buch	433
10	Autor	435
	Syntaxverzeichnis	437
	Sachverzeichnis	443

1 Korrelation

Kapitel 1 führt ein in die erfahrungsgemäß unterschätzte Korrelationsanalyse (SPSS Prozedur CORRELATIONS). Gleich zu Beginn beschäftigt sich das Kapitel mit der Interpretation von Zusammenhängen (Kausalität) und stellt mehrere Beispiele von Fehlschlüssen vor, u.a. den oft zitierten Zusammenhang zwischen dem Konsum von gewalttätigen Computerspielen und Gewaltbereitschaft (vgl. 1.1). Anhand der Korrelationsanalyse werden erste, auch für die (lineare) Regressionsanalyse geltende Voraussetzungen wie z.B. Skalenniveau, Homoskedastizität und Kontinuität erläutert (vgl. 1.2, 1.3). Weitere Abschnitte behandeln die Themen Linearität, Scheinkorrelation und Alphafehler-Kumulation. Auch wird erläutert, warum die bloße Angabe eines Korrelationskoeffizienten im Prinzip kaum weiterhilft (vgl. 1.7). Die Statistik und Interpretation des Korrelationskoeffizienten wird ebenso vorgestellt wie grafische Tests auf (u.a.) Linearität (vgl. 1.4, 1.5). Anschließend wird der Korrelationskoeffizient mit SPSS berechnet und interpretiert (vgl. 1.6). Als spezielle Anwendungen werden u.a. der Vergleich von Korrelationskoeffizienten und die Kanonische Korrelation vorgestellt (vgl. 1.8). Ein abschließender Abschnitt stellt die diversen Voraussetzungen für die Durchführung der Korrelationsanalyse sowie Ansätze zu ihrer Überprüfung zusammen (vgl. 1.9). Jedem Leser, der sich für die Regressionsanalyse interessiert, wird dringend empfohlen, zuvor dieses Kapitel zur Korrelationsanalyse zu lesen.

1.1 Einführung

Zusammenhang ohne Kausalrichtung Korrelation ist nicht gleich Kausation

Häufig besteht das Ziel einer wissenschaftlichen Untersuchung in der Analyse des Zusammenhangs zwischen zwei Variablen. Folgende Fragestellungen sind Anwendungsbeispiele für Korrelationsanalysen:

- Hängen Schwangerschaftsdauer und Geburtsgewicht von Neugeborenen miteinander zusammen?
- Hängen Gewicht (z.B. in Gramm) und Nährstoffgehalt (z.B. in Joule) von Lebensmitteln zusammen?
- Hängen Motorleistung und Kraftstoffverbrauch eines Autos miteinander zusammen?

Bei der Zusammenhangsanalyse gilt es vorab einen wichtigen Grundsatz zu beachten: Korrelation ist nicht gleich Kausation! „Correlation is no proof of causation“ (Pedhazur, 1982, 579). Eine Korrelationsanalyse erlaubt nur die Aussage darüber, ob und in welchem *Ausmaß* zwei Variablen zusammenhängen, aber nicht über die *Art* ihres Zusammenhangs, also welche der beiden Variablen (wenn überhaupt!) die Ursache, welche die Wirkung repräsentiert. Umgekehrt gilt jedoch: Liegt keine bivariate Korrelation vor, liegt auch keine bivariate Kausation vor.

Wenn z.B. ein statistisch bedeutsamer Zusammenhang zwischen zwei Variablen A und B beobachtet wird, sind im Prinzip vier Kausalinterpretationen möglich (vgl. Pedhazur, 1982, 110ff., 578ff.):

- A beeinflusst B kausal
- B beeinflusst A kausal
- A und B werden von einer dritten oder mehr Variablen kausal beeinflusst
- A und B beeinflussen sich gegenseitig (kausal).

Der Korrelationskoeffizient erlaubt keinen Aufschluss darüber, welche Kausalinterpretation richtig ist. Eine Korrelation zwischen zwei Variablen ist eine notwendige, aber keine hinreichende Bedingung für Kausalzusammenhänge.

Welches der Kausalmodelle das plausibelste ist (es sind im Prinzip viele weitere denkbar), wird nicht durch die Korrelation zwischen A *und* B, sondern allein durch eine sachnahe Theorie bestimmt. Nur Logik und plausible Schlussfolgerungen sind eine tragfähige Grundlage der Interpretation von Korrelationen.

Auch in Veröffentlichungen wird Korrelation häufig mit Kausation verwechselt; dies jedoch sollte, wenn schon nicht vom Autor, dann doch vom Leser kritisch auseinander gehalten werden:

Beispiel 1: Gale et al. (2006) berichten einen Zusammenhang zwischen einem hohen IQ in der Kindheit und einer höheren Wahrscheinlichkeit, im Erwachsenenalter Vegetarier zu sein. Hier wird Korrelation mit Monokausalität verwechselt, denn: Werden Menschen Vegetarier, weil sie intelligent sind? Oder werden Menschen einfach nur deshalb Vegetarier, weil sie Wert auf eine gesunde Ernährung legen, was ja nicht notwendigerweise mit einem hohen IQ gleichzusetzen ist?

Beispiel 2: Eine Publikation aus der Krebsforschung argumentiert ähnlich. Die Gabe von Hormonen während der Hormontherapie begünstigt eher den Brustkrebs bei Frauen, als ihn zu bekämpfen. Laut Peter Ravdin (2006) gehe nun jedoch z.B. die Zahl der Brustkrebsdiagnosen zurück, weil immer mehr Frauen in den USA die Hormontherapie absetzen würden. Zahlreiche Frauen brachen die Hormontherapie ab, mit dem Effekt, dass damit einhergehend auch die Zahl der Brustkrebsfälle innerhalb weniger Monate ungewöhnlich stark gefallen sei. Tatsächlich liegt trotz aller sachnaher Plausibilität eine Gleichsetzung von Korrelation mit Kausation vor: Auch wenn Hormontherapie und erhöhtes Brustkrebsrisiko tatsächlich miteinander zusammenhängen, ist nicht automatisch der Umkehrschluss zulässig, dass weniger Hormontherapie (z.B. durch eine veränderte Verordnungspraxis) automatisch ein verringertes

tes Brustkrebsrisiko nach sich zieht. Die Gabe von Hormonen ist nicht der alleinige karzino- gene Faktor.

Beispiel 3: Cha et al. (2001) veröffentlichten z.B. im angesehenen Journal of Reproductive Medicine einen scheinbar empirisch nachgewiesenen Zusammenhang zwischen Gebeten und der Wahrscheinlichkeit, schwanger zu werden. Demnach würden unfruchtbare Frauen doppelt so häufig schwanger, wenn für sie eine Gebetsgruppe betete, als Frauen, für die nicht gebetet wurde. Die Studie wurde eine Zeit lang als „Miracle Study“ bezeichnet, unter anderem aus dem Grund, weil die Frauen gar nicht wussten, dass für sie gebetet wurde und weil die Gebete in einigen Tausend Kilometern Entfernung stattfanden. Um es bei diesem Beispiel kurz zu machen: Die Veröffentlichung von Cha et al. (2001) zum Zusammenhang zwischen Gebeten und der Wahrscheinlichkeit, schwanger zu werden, war schlichtweg Betrug. Einer der Verfasser wurde wegen mehrfachen Betrugs schließlich rechtskräftig verurteilt. Es ist nicht einmal bewiesen, dass diese Studie jemals durchgeführt wurde. Replikationen konnten den behaupteten Zusammenhang zwischen Beten und Schwangerschaft nicht bestätigen. Die unkritische Öffentlichkeit hindert dies jedoch nicht, sich durch das Zitieren dieser Studie als Beleg für den Effekt des sog. „faith healing“ zu blamieren.

Beispiel 4: Im Forschungsbereich der Medienrezeption werden beim oft behaupteten Zusammenhang zwischen dem Konsum gewalttätiger Computerspiele und der Gewaltbereitschaft der Spieler Ursache und Zusammenhänge oft auf mehreren Ebenen verwechselt. Tatsächlich ist diese „Argumentation“ ein schönes Beispiel für die wissenschaftlich nicht korrekte Interpretation von Zusammenhängen, besonders in Gestalt von Korrelationskoeffizienten:

Ebene 1: Reduktion einer zweiseitigen Korrelation auf eine einseitige Kausalrichtung:

Der behauptete (und nur implizit zweiseitig gemeinte) Zusammenhang des Konsums gewalttätiger Computerspielen und Gewaltbereitschaft wird auf den Konsum gewalttätiger Computerspiele als einseitige *Ursache* von Gewaltbereitschaft reduziert. Die zweite potentielle Wirkrichtung, dass Gewaltbereitschaft eine mögliche Ursache des Konsums gewalttätiger Computerspiele sei, wird inkorrekterweise ausgeklammert; keinesfalls stützt dies die Plausibilität der These der ersten, beibehaltenen Wirkrichtung des Konsums von gewalttätigen Computerspielen als *alleinige* Ursache von Gewaltbereitschaft.

Ebene 2: Reduktion eines komplexen Beziehungsgeflechts auf eine einzige Ursache (Monokausalität):

Implizit ist in der behaupteten alleinigen Wirkrichtung von gewalttätigen Computerspielen auf Gewaltbereitschaft die These der Monokausalität des Konsums von gewalttätigen Computerspielen als *einzigem verursachendem Faktor* formuliert. In den Sozialwissenschaften ist eine solche Auffassung gleichermaßen anachronistisch wie simplifizierend. Die Tatsache, dass ein Zusammenhang, z.B. eine Korrelation, nur zwei Größen in ein Modell zu bringen vermag, besagt nichts darüber aus, ob dieses Modell der Komplexität empirischer Realität angemessen ist.

Ebene 3: Ablenkung von der Komplexität empirischer Realität durch realitätsfern simple Thesen:

Die Komplexität der empirischen Realität sozialwissenschaftlicher Forschungsgegenstände ist nur ausnahmsweise durch monokausal-monotone Variableneinflüsse beschreibbar; stattdessen ist hier üblicherweise von komplex-dynamischen Faktorengeflechten auszugehen. In der Realität des Medienkonsums ist angesichts des Verhältnisses der verkauften bzw. gespielten Spiele und der Anzahl von z.B. Amokläufern eher von folgenden Konstellationen auszugehen: Fast jeder Amokläufer hat gewalttätige Computerspiele gespielt, jedoch läuft nicht jeder Computerspieler Amok. Genau betrachtet lässt diese Beobachtung mind. zwei Fehler im eingangs behaupteten Zusammenhang zwischen dem Konsum gewalttätiger Computerspielen und der Gewaltbereitschaft der Spieler erkennen.

Zum einen liegt ein Ziehungsfehler vor: Es wird von Befunden an einer verhältnismäßig kleinen Extremgruppe, z.B. Amokläufern, als Teil („pars“) auf die Grundgesamtheit aller Spieler („toto“) geschlossen. Aufgrund des Ziehungsfehlers ist die Extremgruppe jedoch nicht repräsentativ für die Gesamtgruppe. Zusätzlich liegt hier der Missbrauch des pars pro toto-Prinzips vor.

Zum anderen liegt ein Denkfehler vor: Über all der Argumentation über die Ursachen bei Amokläufern (die auch gewalttätige Computerspiele konsumieren, jedoch eindeutig zahlenmäßig in der Minderheit sind), gewalttätig zu werden, wird in der Dominanz des Diskurses im Wesentlichen übergangen, warum die überwältigende *Mehrheit* von Spielern gewalttätiger Spiele *nicht* Amok läuft. Die Schlussfolgerung ist eindeutig: (a) Weil die Wirkrichtung womöglich genau umgekehrt sein kann: Die psychosoziale Befindlichkeit der Spieler entscheidet nicht nur, ob man gewalttätige Computerspiele konsumiert, sondern auch, ob man (eigenes) gewalttätiges Verhalten zulässt. (b) Weil damit die These des monokausalen Effekts gewalttätiger Computerspiele gegenstandslos ist. Computerspiele zu verbieten ändert nichts an der vorgebrachten pseudo-wissenschaftlichen Argumentation: Korrelation ist nicht gleich Kausation.

Für die direkte Verbindung zwischen Computer-Spielen (z.B. World of Warcraft) und gewalttätigem Verhalten findet die aktuelle medienpsychologische Forschung keinen Beleg. Im Gegenteil werden Publikationsbias, unerwartete Effekte (z.B. entspanntes Verhalten nach dem Spielen gewalttätiger Computer-Spiele), wie auch vermittelnde Faktoren wie z.B. Alter, Geschlecht, Persönlichkeits- und Sozialstrukturen berichtet (z.B. Kutner & Olson, 2008; Barnett, 2008; Ferguson, 2007). Kausalitätsmodelle können auch mittels komplexerer statistischer Verfahren, z.B. der multiplen Regression „konstruiert“ werden. Solche Kausalitätskonstruktionen sind in dieser Verkleidung oftmals schwieriger zu durchschauen. Solche Kausalitätskonstruktionen münden in pseudo-wissenschaftliche „Erfindungen“ in der Form überzogener Schlussfolgerungen, die von den zugrundeliegenden Daten bei genauerer Analyse gar nicht unterstützt werden. Beispiele für diese „Erfindungen“ sind Behauptungen wie z.B., dass die Todesstrafe oder auch das Tragen von Schusswaffen die Kriminalitätsrate senke (vgl. z.B. Goertzel, 2002).

Die Korrelation zwischen zwei Variablen ist nicht gleichzusetzen mit dem Zusammenhang zwischen zwei Konstrukten; nur angedeutet werden sollen Aspekte wie z.B. operationale Definition oder Stichprobenabhängigkeit. Eine Korrelationsanalyse kann auch als statistische Modellierung einer niedrigkomplexen Operationalisierung wenig erklärungshaltiger, einfa-

cher Theorien verstanden werden. Ein statistisch signifikantes Ergebnis einer Korrelationsanalyse schließt die Gültigkeit anderer, konkurrierender Modelle nicht aus. Die Interpretation einer Korrelationsanalyse als monokausal begeht somit gleich mehrere Fehler gleichzeitig:

- Gleichsetzung einer Variablen mit einem Konstrukt
- Reduktion einer Korrelation auf eine Kausalrichtung
- Reduktion eines komplexen Beziehungsgeflechts auf eine einzige Ursache (Monokausalität)

Der „Beweis“ seitens einer Korrelationsanalyse, dass z.B. ein behaupteter Zusammenhang zwischen zwei Variablen tatsächlich auch statistische Signifikanz erzielt, sollte daher mit einer gewissen Zurückhaltung betrachtet werden (auch, was die Verabsolutierung der Signifikanz angeht, z.B. Witte, 1980; Schendera, 2007).

Soll ein Kausalmodell überprüft werden, also ob eine Variable systematisch die Variation einer anderen Variable verursacht, dann kann u.a. anstelle der Korrelation das Verfahren der Regression gewählt werden. Für komplexere Modelle sind u.a. die Partialkorrelation (5.3) bzw. -regression, ggf. auch die Pfadanalyse (5.1) in Frage kommende Verfahren.

In der Statistik wurden viele Messverfahren entwickelt, um das Ausmaß des Zusammenhangs zwischen zwei Variablen zu messen. Die Terminologie für Zusammenhangsmaße („Korrelation“; „Assoziation“) ist in der Literatur leider nicht immer einheitlich (siehe auch Lorenz, 1992, 58ff.). Wird z.B. die Stärke eines (linearen) Zusammenhangs zwischen Reihen von Wertepaaren untersucht, spricht man bei intervall- bzw. ordinalskalierten Daten von Korrelation (Maß-, Rang-). Bei Kreuztabellen, Vierfeldertafeln bzw. Kontingenztafeln wird von Assoziation oder manchmal auch von Kontingenz gesprochen. Bei ordinalskalierten Variablen wird die Stärke eines Zusammenhangs nach Spearman als Korrelation bezeichnet, nach Kendall oder Somer dagegen als Kontingenz oder Assoziation. Bortz (1993, Kap. 6.3) spricht sogar bei dichotomen Variablen von Korrelation. Die Entscheidung für das Verfahren der Wahl hängt letztlich von der Anzahl der Variablenausprägungen, ihrer Verteilung und dem Skalierungsniveau ab (vgl. auch das Kapitel zur Tabellenanalyse). Wichtig ist u.U. auch, ob ein Kausalmodell vorliegt (z.B. ‚X verursacht Y‘), der Ursache der Beziehung, der Anzahl der Variablen und anderes mehr. Alle Verfahren folgen aber einem generellen Prinzip. Es erfolgt ein Vergleich zwischen den Variablen bezüglich dem empirisch beobachteten und dem theoretisch möglichen maximalen Zusammenhang. Mit anderen Worten: Es wird verglichen, was den Variablen gemeinsam ist, mit dem, was den Variablen hätte gemeinsam sein können, wenn deren Beziehung perfekt gewesen wäre.

Der Pearson-Korrelationskoeffizient (syn.: Produkt-Moment-Korrelation, Bravais-Pearson-Korrelation) beschreibt die Stärke (syn.: Enge) des Zusammenhangs zweier metrisch skalierten, linear verbundener Variablen (Messwertreihen) unabhängig von ihrer Einheit.

1.2 Erste Voraussetzung: Das Skalenniveau

Die Art der Skalierung entscheidet darüber, in welcher Form der (lineare) Zusammenhang zwischen den Variablen nachgewiesen werden kann. Sind die Daten intervallskaliert, kann als Maß die Korrelation nach Pearson bestimmt werden. Stetig ordinalskalierte Variablen werden mittels den Verfahren nach Kendall oder Spearman untersucht; diese Verfahren (z.B. Spearman's Rangkorrelation) können auch dann eingesetzt werden, wenn zwei intervallskalierte Variablen nicht linear, sondern nur monoton zusammenhängen und diese Information ausreichend ist.

Liegen diskret skalierte Ordinalvariablen vor, könnten z.B. die Maße Gamma, Somer usw. eingesetzt werden (zu Details wird auf die Tabellenanalyse verwiesen, vgl. z.B. Schendera, 2004, Kap. 12).

Bei zwei nominalskalierten („qualitativen“) Variablen kann der Zusammenhang z.B. durch einen Kontingenzkoeffizienten beschrieben werden. Wichtig ist also, dass das gewählte Verfahren für eine Analyse auf Zusammenhang dem Skalenniveau der untersuchten Variablen entspricht. Bei Variablenpaaren mit unterschiedlichem Skalenniveau sollte immer auf das Skalenniveau der niedriger skalierten Variable ausgewichen werden.

Übersicht: Eine erste Übersicht zu Maßen der Assoziation und Korrelation

	intervall	ordinal	nominal
intervall	Pearson's r		Eta ² (R ²)
ordinal		Pearson's r (falls gleiche Abstände), Spearman's R Gamma, Kendall's tau-b, Stuart's tau-c, Somer's D	
nominal			Phi-Koeffizient Cramer's V Kontingenzkoeffizient Lambda Unsicherheitskoeffizient Kontingenzkoeffizient Cochran's Q

Die Tabelle enthält mit Ausnahme von Eta² sogenannte symmetrische Maße. Bei der Berechnung einer Korrelation bzw. Assoziation ist dabei unerheblich, welche die unabhängige Variable und welche die abhängige Variable ist. Eta² kann z.B. in dem Fall angewendet werden, wenn die unabhängige Variable intervallskaliert und die abhängige Variable nominalskaliert ist. Wichtig ist, dass die anschließend gewählte Statistik der festgestellten Form

(Funktion) der Messwertverteilung entspricht. Cohen et al. (2003³, 60–63) raten bei der Korrelation zweier Verhältnismaße zur Vorsicht bzw. ganz davon ab.

Im Folgenden wird die Korrelationsanalyse nach Pearson vorgestellt; dieses Verfahren setzt zwei Variablen auf mind. Intervallskalenniveau voraus, aber wie bereits angedeutet auch, dass der Zusammenhang zwischen den beiden untersuchten Variablen *linear* ist.

1.3 Weitere Voraussetzungen: Linearität, Homoskedastizität und Kontinuität

Eine genauer zu betrachtende Voraussetzung der Korrelation nach Pearson ist, dass die beiden untersuchten Variablen (bzw. genauer: ihre Messwertpaare) grafisch einen linearen Zusammenhang zeigen. Messwertpaare ordnen sich als „in etwa“ linear an, z.B. in einem Streudiagramm. Zeigt ein Streudiagramm eine lineare Anordnung der Daten, kann ein Maß für einen linearen Zusammenhang gewählt werden.

Der maximal mögliche positive Zusammenhang ($r=1,0$) wird dann erreicht, wenn beide Variablen exakt dieselbe Verteilungsform aufweisen (jedoch nicht notwendigerweise eine Normalverteilung). Der maximal mögliche negative Zusammenhang ($r=-1,0$) wird dann erreicht, wenn beide Variablen genau spiegelverkehrte Verteilungen besitzen. Für die Beschreibung eines bivariaten Zusammenhangs mittels eines Pearson-Korrelationskoeffizienten müssen die Variablen nicht notwendigerweise normalverteilt sein. Je weiter beide Variablen gegenläufig verteilt sind, umso niedriger wird der Zusammenhang (Cohen et al., 2003³, 53).

Linearität ist genau betrachtet, eine Kombination dreier, auch grafisch (z.B. eines Streudiagramms) überprüfbarer Merkmale:

- Lineare Anordnung der Messwertpaare: Eine kurvilineare Messwertverteilung kann durch den Pearson-Korrelationskoeffizienten nicht angemessen beschrieben werden.
- Streuung um die Gerade: Je enger die Streuung, umso höher ist der Koeffizient. Je diffuser, wolkenartiger die Punkte angeordnet sind, desto niedriger ist der Korrelationskoeffizient.
- Ausschluss von Ausreißern: Ausreißer sind nicht vorhanden. Sowohl Ausreißer in Richtung der Funktion, wie z.B. auch orthogonal zu ihr, verzerren den Koeffizienten (vgl. Schendera, 2007).

Ein hoher Korrelationskoeffizient setzt somit Linearität voraus, um gültig zu sein und kann daher kein Linearitätstest sein, sondern muss Linearität voraussetzen. Ein niedriger Korrelationskoeffizient kann deshalb entstehen, wenn mind. eines der drei Merkmale nicht vorliegt, z.B. lineare Funktion, minimale Streuung und/oder Ausreißer. Weil ein niedriger Korrelationskoeffizient auch deshalb entstehen kann, weil ein einzelner Ausreißer (bei einer ansonsten perfekten linearen Korrelation) massiv den Schätzvorgang verzerrt, kann auch hier der Korrelationskoeffizient nicht als Linearitätstest bezeichnet werden.

Zeigt ein Streudiagramm einen eher kurvenartigen Verlauf der Daten, so kann kein Maß für einen linearen Zusammenhang gewählt werden, sondern stattdessen ein Maß für einen nicht-linearen Zusammenhang. Anders formuliert: Das Standardmaß der Statistik für das Ausmaß eines Zusammenhangs, der Korrelationskoeffizient nach Pearson, setzt eine lineare Anordnung der Daten voraus. Das Vorliegen der sog. „Linearität“ kann unkompliziert anhand eines bivariaten Streudiagramms überprüft werden und ist somit ein sog. „grafischer Test auf Linearität“.

1.4 Exkurs: Grafische Tests auf Linearität

1.4.1 Prozedur GRAPH, Scatterplot Option

Streudiagramme bilden Messwertpaare in einem Koordinatensystem ab („Punkteschwarm“). Streudiagramme sind geeignet, um die Beziehung zwischen mindestens zwei intervallskalierten, metrischen Variablen grafisch abzubilden. Oft werden in Streudiagramme (Regressions-)Geraden eingebettet, die den (kurvi-)linearen Zusammenhang beider Variablen repräsentieren (siehe z.B. weiter unten CURVEFIT).

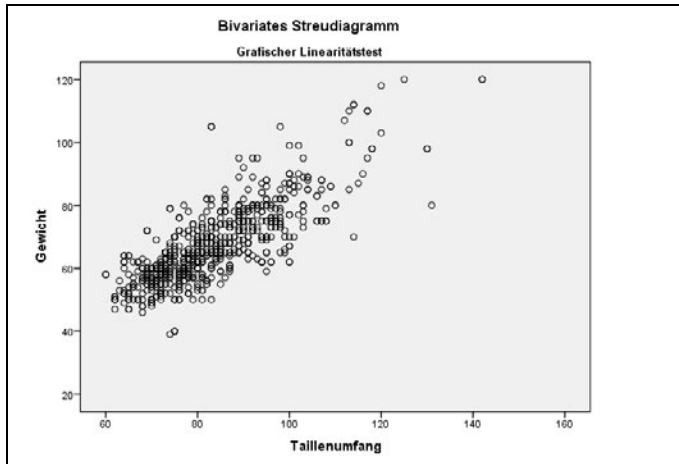
Beschreibung

Das folgende Streudiagramm stellt zwei Variablen auf zwei Skalenachsen dar. Eine Variable legt dabei die horizontale, die andere die vertikale Achse fest. Jeder Wert der Variable auf der X-Achse (hier: Taillenumfang) ist mit dem dazugehörigen Wert der auf der Y-Achse abgetragenen Variable (hier: Gewicht) in das Koordinatensystem eingetragen. Anhand der so entstandenen Punktwolke kann man nun den Zusammenhang zwischen beiden Variablen darstellen. In diesem Beispiel besteht ein positiver linienförmiger Zusammenhang zwischen beiden Variablen (sog. grafischer ‚Linearitätstest‘). Da die Messwertpaare durchgehend in annähernd parallelem Abstand um eine gedachte Korrelationslinie streuen, ist gleichzeitig Homoskedastizität gegeben (Gleichheit der Streuungen). Da die Linie keine Lücken aufweist, z.B. keine unterbrochene Verteilung wie sie z.B. bei der Analyse von Extremgruppen zutage treten könnte, ist auch die Kontinuität der Linie gegeben.

Voreinstellungen

Pfad: Bearbeiten → Optionen → Reiter „Viewer“

Stellen Sie sicher, dass unter „Ausgabeelemente anzeigen“ die Option „Befehle im Log anzeigen“ aktiviert ist. In früheren SPSS Versionen hiess der „Viewer“ noch „Text-Viewer“.



Pfad: Diagramme → Veraltete Dialogfelder [nicht bei älteren SPSS Versionen] → Streu-/Punkt-Diagramm → Einfach → Definieren.

Klicken Sie TAILLE in die X-Achse und GEWICHT in die Y-Achse. Beachten Sie, dass SPSS automatisch die Vorgehensweise „Listenweiser Fallausschluss“ für den Umgang mit fehlenden Werten einstellt (vgl. „Optionen“). Legen Sie über „Titel“ Titel und Untertitel fest. Fordern Sie die Grafik mit „OK“ an.

GRAPH

```
/SCATTERPLOT(BIVAR)=
    taille WITH gewicht
/MISSING=LISTWISE
/TITLE= "Bivariates Streudiagramm"
/SUBTITLE= "Grafischer Linearitätstest".
```

Erläuterung:

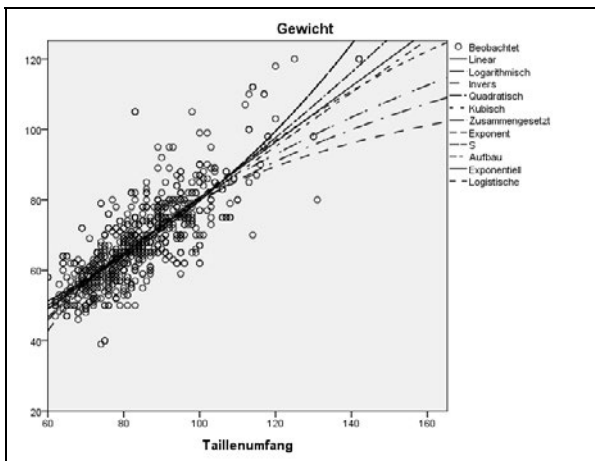
Der Befehl GRAPH fordert ein Diagramm an. Mit /SCATTERPLOT (BIVAR) wird der Typ, ein bivariates Streudiagramm, festgelegt. Nach dem Gleichheitszeichen folgen die Namen der beiden Variablen, deren Wertepaare eingetragen werden sollen. Die zuerst genannte Variable (hier: TAILLE, Taillenumfang) wird auf der x-Achse abgetragen. GEWICHT wird auf der y-Achse abgetragen. Nach MISSING= wird festgelegt, wie mit möglicherweise fehlenden Werten umgegangen werden soll. Die gewählte Option LISTWISE führt den Fallabschluss listenweise durch (über VARIABLEWISE kann alternativ erreicht werden, dass die Fälle Variable für Variable ausgeschlossen werden). Mittels TITLE, SUBTITLE und ggf. FOOTNOTE wird der Text für die Überschrift, den Untertitel und die Fußnote angegeben.

1.4.2 SPSS Prozedur CURVEFIT

Ein linearer Zusammenhang gehört zu einer der Grundvoraussetzungen für einen hohen Korrelationskoeffizienten bzw. für die Berechnung einer Pearson-Regression. Mit der Prozedur CURVEFIT bietet SPSS eine weitere Möglichkeit an, zwei Variablen daraufhin zu überprüfen, ob zwischen ihnen ein linearer Zusammenhang besteht oder ob eine andere Funktion diesen Zusammenhang nicht besser erklären könnte.

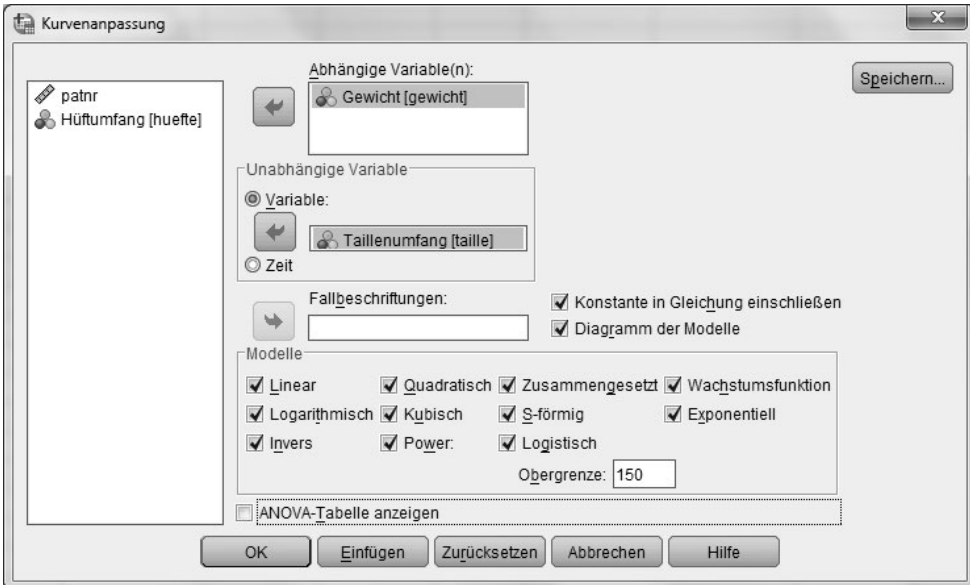
Beschreibung

Der Leistungsumfang der Prozedur CURVEFIT geht über die einfache Anforderung des bivariaten Streudiagramms hinaus. CURVEFIT überprüft nicht nur auf einen möglichen linearen Zusammenhang, sondern auf zehn weitere Zusammenhangsmodelle (u.a. Exponent, exponentiell, invers, kubisch, logarithmisch, quadratisch, S, Wachstum, zusammengesetzt). CURVEFIT trägt in die Anordnung der empirisch vorliegenden Messwertpaare („Beobachtet“) die jeweils ermittelte Funktion als Linie ab (werden zahlreiche Funktionen auf einmal eingezeichnet, kann dies u.U. etwas unübersichtlich werden). CURVEFIT ermittelt für jede Funktion zusätzlich statistische Parameter, wie z.B. R^2 usw. Der Vergleich der verschiedenen Kurvenmodelle erfolgt daher nicht nur per Augenschein anhand der Linien, sondern auch auf der Basis von statistischen Parametern und erlaubt darüber hinaus relativ unkompliziert zu entscheiden, welche Funktion den Zusammenhang zwischen den beiden untersuchten Variablen möglicherweise besser als das lineare Modell erklärt.



Pfad: Analysieren → Regression ... → Kurvenanpassung.

Legen Sie GEWICHT als abhängige Variable und TAILLE als unabhängige Variable fest. Wählen Sie unter den angebotenen Modellen (s.u.) die gewünschten Kurvenfunktionen. Aktivieren Sie „Konstante in Gleichung einschließen“ und „Diagramm der Modelle“. Fordern Sie die Kurvenanpassung mit „OK“ an.



```
TSET NEWVAR=NONE .
CURVEFIT
/VARIABLES=gewicht WITH taille
/CONSTANT
/MODEL=
  LINEAR LOGARITHMIC INVERSE
  QUADRATIC CUBIC COMPOUND
  POWER S GROWTH EXPONENTIAL
  LGSTIC
/UPPERBOUND=150
/PLOT FIT .
```

Erläuterung der Syntax

CURVEFIT fordert das Verfahren zur Kurvenanpassung an. CURVEFIT gibt standardmäßig ein Kurvenanpassungsdiagramm und eine zusammenfassende Tabelle der Regressionsstatistiken aus, darin u.a. Kurvenfunktion bzw. -methode, R^2 , Freiheitsgrade, F-Wert, Signifikanzniveau, ggf. Obergrenze (Upper bound), Konstante (b_0) und Regressionskoeffizient/en (b_1 , b_2 , b_3). Das Konfidenzintervall ist bei 95% voreingestellt. CURVEFIT schließt Missings standardmäßig listenweise aus.

Nach VARIABLES folgen die Namen der beiden Variablen, für die Kurvenfunktionen an die Daten angepasst werden sollen. Die zuerst genannte Variable (hier: TAILLE, Taillenumfang) wird als unabhängige Variable in die Modellbildung einbezogen; es kann nur eine unabhängige Variable angegeben werden. Die anschließend genannte Variable (hier: GEWICHT) bildet die abhängige Variable; es können mehrere abhängige Variablen angegeben werden. Wird TAILLE zur AV und GEWICHT zur UV, sind konsequenterweise andere Funktionen das Ergebnis (siehe dazu die Erläuterung der einfachen linearen Regression); das R^2 ist je-

doch jeweils dasselbe. Es kann nur eine VARIABLES-Anweisung angegeben werden. /CONSTANT bestimmt, ob die Regressionsgleichung eine Konstante enthalten soll (oder nicht: NOCONSTANT). Nach /MODEL= können insgesamt bis zu elf verschiedene Regressionsmodelle auf einmal angegeben werden (alternativ über die Option: ALL). CURVEFIT erzeugt dabei für jede AV und Modellkurve vier neue Variablen. Bei großen Datenmengen sollten nicht mehr als die jeweils notwendigen Modellkurven angefordert werden. Die verschiedenen Regressionsmodelle werden weiter unten im Detail vorgestellt. Wird ein logistisches Regressionsmodell (LGSTIC) angefordert, muss mittels /UPPERBOUND separat ein oberer Grenzwert angegeben werden, der positiv und auch größer als der größte Wert in allen angegebenen abhängigen Variablen ist; für die vorliegenden Daten wurde der Wert 150 angegeben. Dieser Grenzwert wird für das logistische Regressionsmodell im Output angegeben. Mittels PLOT=FIT (voreingestellt) wird das Kurvenanpassungsdiagramm angefordert, PLOT=NONE unterdrückt die Ausgabe des Kurvenanpassungsdiagramms.

Das CURVEFIT vorausgehende TSET NEWVAR= legt Voreinstellungen für den Umgang mit abhängigen Daten fest, z.B. Zeitreihen und Sequenzvariablen. Bei NONE werden keine neuen Variablen abgespeichert. Bei CURRENT (voreingestellt) und ALL wird dagegen abgespeichert, wobei CURRENT bereits vorhandene Variablen ersetzt, ALL dagegen nicht. Das vorgestellte Beispiel ist nicht für Zeitreihendaten geeignet. Die Ausgabe ist gekürzt.

Modellzusammenfassung und Parameterschätzer

Abhängige Variable: Gewicht

Gleichung	Modellzusammenfassung					Parameterschätzer			
	R-Quadrat	F	Freiheitsgrade 1	Freiheitsgrade 2	Sig.	Konstante	b1	b2	b3
Linear	,645	1247,205	1	686	,000	1,532	,791		
Logarithmisch	,632	1176,897	1	686	,000	-232,420	67,965		
Invers	,607	1058,998	1	686	,000	136,157	-5604,780		
Quadratisch	,646	625,993	2	685	,000	14,652	,490	,002	
Kubisch	,646	625,993	2	685	,000	14,652	,490	,002	,000
Zusammengesetzt	,633	1183,187	1	686	,000	26,405	1,011		
Potenzfunktion	,633	1184,755	1	686	,000	,958	,961		
S-förmig	,620	1120,686	1	686	,000	5,176	-79,995		
Aufbaufunktion	,633	1183,187	1	686	,000	3,274	,011		
Exponentiell	,633	1183,187	1	686	,000	26,405	,011		
Logistisch	,643	1236,631	1	686	,000	,051	,978		

Die unabhängige Variable ist Taillenumfang.

Erläuterung der Ausgabe

Das Kurvenanpassungsdiagramm wurde bereits weiter oben abgebildet. In die Anordnung der empirisch vorliegenden Messwertpaare („Beobachtet“) wurde für jede angeforderte Funktion eine Linie in den Punkteschwarm eingezeichnet (linear, logarithmisch usw.). Beschränkt man den Vergleich der Linien auf den Bereich der Datengrundlage (SPSS führt die eingezeichneten Linien über den zugrunde liegenden Datenbereich hinaus, was eigentlich nicht zulässig ist), kann man bereits grafisch anhand des Kurvenanpassungsdiagramms feststellen, dass einige der ermittelten Funktionen zum annähernd gleichen Ergebnis führen; wobei die-

ses „gleich“ nicht mit einem „gleich gut“ zu verwechseln ist. Für eine begründetere Wahl zwischen den Funktionen können die Regressionsstatistiken in der Tabelle befragt werden. Die Tabelle „Modellzusammenfassung und Parameterschätzer“ listet Gleichungen und Kennwerte für das untersuchte Modell auf; die AV („Gewicht“) bzw. UV („Tailenumfang“) sind dabei jeweils über bzw. unter der Tabelle angegeben. Unter „Gleichung“, „Modellzusammenfassung“ und „Parameterschätzer“ sind die Gleichungen der angepassten Kurvenfunktionen und die dazugehörigen R^2 [„R-Quadrat“], F-Wert [„F“], Freiheitsgrade [„1“ bzw. „2“], Signifikanzniveau [„Sig.“] und die Konstante (b0) und ab „b1“ je nach Modell die diversen Regressionskoeffizienten aufgelistet. Im Gegensatz zu früheren SPSS Versionen werden angegebene Obergrenzen, wie auch (nicht) eingehaltene Toleranzen in einer anderen Tabelle ausgegeben. Das Toleranzkriterium für QUA und CUB kann über TSET eingestellt werden.

Voraussetzung für die Wahl einer Kurvenfunktion anhand von Parametern ist zunächst, dass anhand des Kurvenanpassungsdiagramms davon ausgegangen werden kann, dass die jeweils eingezeichnete Linie eine angemessene Schätzung der beobachteten Messwertpaare darstellt. Die besten Parameter sind sinnfrei, wenn die ermittelte Kurve gar nicht die empirische Verteilung widerspiegelt. Das nächste Kriterium ist das der Signifikanz. Es werden zunächst nur die Kurvenfunktionen weiter betrachtet, deren F-Test Signifikanz erreicht (im Beispiel wären dies alle Modelle). Das nächste Kriterium ist das R^2 . Es werden von den signifikanten Modellen nur die Kurvenfunktionen mit dem höchsten R^2 weiter betrachtet (im Beispiel wären dies LIN, QUA und CUB). Das nächste Kriterium ist die Einfachheit der Modellgleichung. Dem Ausmaß der größeren Varianzaufklärung steht also eine komplexere Modellgleichung gegenüber, deren Komplexität jedoch nicht immer durch eine marginal bessere Varianzaufklärung gerechtfertigt ist. Wenn sich z.B. das R^2 des linearen und des kubischen Regressionsmodells nur um 0.001 unterscheidet, der Preis dafür jedoch zwei zusätzliche Variablen in der Gleichung der quadratischen Funktion sind (siehe Übersicht), dann sollte die einfachere Gleichung, also in diesem Falle die lineare Regressionsfunktion vorgezogen werden. Der Vorteil ist dabei nicht nur der, dass mit linearen Korrelations- bzw. Regressionsmodellen ohne substantiellen Informationsverlust weitergerechnet werden kann, sondern dass die Interpretation der Ergebnisse bzw. Modelle vor dem Hintergrund linearer Modelle um einiges einfacher ist, als wenn dies anhand quadratischer Modelle gemacht werden müsste. Dennoch sollte man sich abschließend rückversichern, ob die grafisch und statistisch gefundene Kurvenfunktion tatsächlich zur inhaltlichen Beschreibung der miteinander zusammenhängenden Konstrukte geeignet ist. Für diese abschließende Betrachtung ist es oft aufschlussreich auszuprobieren, welche inhaltlichen Konsequenzen es nach sich ziehen könnte, wenn die gefundene Funktion über den Bereich der vorliegenden Daten verlängert würde oder ob die Gesamtfunktion je nach Datenbereich auch in einzelne, womöglich unterschiedliche Funktionen zerlegt werden könnte.

1.5 Statistik und Interpretation des Korrelationskoeffizienten

Wie wird nun ein Korrelationskoeffizient nach Pearson berechnet und interpretiert?

1.5.1 Statistik des Korrelationskoeffizienten

Für die Berechnung des Pearson-Korrelationskoeffizienten zirkulieren mehrere Formel- bzw. Notationsvarianten.

$$r = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}} \text{ bzw. } r = \frac{\text{cov}(x, y)}{s_x \cdot s_y} \text{ bzw. } r = \frac{\sum_i^n ((x_i - \bar{x}) \cdot (y_i - \bar{y}))}{n \cdot s_x \cdot s_y}$$

$$\text{bzw. } r = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{((n \sum x_i^2) - (\sum x_i)^2)((n \sum y_i^2) - (\sum y_i)^2)}}$$

In der ersten Formel sind z.B. s_x^2 und s_y^2 die Varianzen der beiden Variablen x und y . s_{xy} ist die Kovarianz von x und y ; demnach ist der Pearson-Korrelationskoeffizienten r definiert als die Kovarianz von x und y relativiert an der Quadratwurzel des Produktes der Varianzen der beiden Variablen x und y . Der Koeffizient wird dann maximal, wenn die beiden Variablen gleich variieren (s.u.). Unter Varianz wird die Streuung von mind. intervallskalierten Variablen verstanden. Die beiden anderen Formeln sind Varianten der ersten; die dritte Schreibweise deutet die Ähnlichkeit zur Ermittlung der einfachen linearen Regression an. Als Kovarianz ist die gemeinsame Streuung zweier mind. intervallskalierten Variablen definiert. Bei standardisierten Variablen ist die Kovarianz identisch mit der Korrelation.

Für die Berechnung des Korrelationskoeffizienten r als *deskriptives Maß* sind keinerlei Annahmen über die Verteilungen der beiden Variablen erforderlich. Die Berechnung von r für einen *Hypothesen-* bzw. *Signifikanztest* setzt jedoch eine bivariate Normalverteilung voraus (z.B. Pedhazur, 1982², 40; vgl. Schendera, 2007, wie mit abgeschnittenen, sog. trunkeierten Verteilungen umzugehen ist).

Die Interpretation des Korrelationskoeffizienten setzt einen *linearen* Zusammenhang zwischen den beiden Variablen voraus (erkennbar z.B. an einem Streudiagramm). Bilden die Datenpunkte eine perfekte Linie, wird der Maximalbetrag von 1.0 erreicht. r ist den Daten dann am besten angemessen, weil es auf der minimalen Summe quadrierter Abstände zwischen den beobachteten und den geschätzten Werten basiert. Je größer die (systematische) Streuung der Datenpunkte um die (gedachte) Gerade, desto größer ist die Fehlervarianz und somit niedriger wird r . Je kurvilinearere eine bivariate Streuung ist, desto weniger ist der

lineare Korrelationskoeffizient als ein Index für die Streuung der vorliegenden Daten geeignet. Stattdessen sind passende Kurvenfunktionen zu wählen (z.B. kubisch, quadratisch usw.). Was sagt ein Korrelationskoeffizient aus? r basiert auf standardisierten (z-transformierten) Werten. r ist eine reine Zahl und unabhängig von den Messeinheiten der beiden miteinander korrelierten Variablen. Der absolute r -Wert drückt somit den Grad des linearen Zusammenhangs zwischen zwei Variablen in z-Wert-Einheiten aus. Je höher r ist, desto besser kann die eine der beiden Variablen durch die jeweils vorhergesagt werden. Die Rangkorrelation, der Phi-Koeffizient und u.a. das punktbiseriale r sind lediglich Berechnungsäquivalente auf der Basis der Formel von r (vgl. Cohen et al. 2003³, Kap.2).

Das Quadrat des Korrelationskoeffizienten (r^2) drückt den Anteil der gemeinsamen Varianz von x und y bzw. den Varianzanteil der linearen Assoziation zwischen den beiden Variablen aus („Überlappung“). r^2 wird auch als Determinationskoeffizient bezeichnet.

$1 - r^2$ drückt die nichtgemeinsame Varianz aus bzw. den Anteil der *nicht* linearen Assoziation zwischen den beiden Variablen („Nichtüberlappung“). $1 - r^2$ kann auch als Vorhersagefehler (Varianzschätzfehler) der jeweils einen Variablen durch die andere interpretiert werden.

1.5.2 Interpretation des Korrelationskoeffizienten

Mit einer Korrelationsanalyse wird üblicherweise untersucht, ob zwischen zwei Variablen ein Zusammenhang besteht. Der Korrelationskoeffizient nach Pearson (auch: Maßkorrelationskoeffizient, oder auch einfach nur: Korrelationskoeffizient) ist ein Maß für den linearen Zusammenhang zwischen zwei kontinuierlich skalierten Variablen. Wie eingangs gesehen, gibt es neben dem Pearson-Korrelationskoeffizient noch weitere Zusammenhangsmaße. Der Korrelationskoeffizient kann Werte zwischen -1 und 1 annehmen. Das Vorzeichen des Koeffizienten gibt die Richtung des Zusammenhangs und sein Absolutwert die Stärke an. Eine positive Korrelation bedeutet, dass, sobald sich Werte der einen Variable erhöhen, sich auch Werte der anderen Variable erhöhen (grafisch: von links unten nach rechts oben ansteigende Gerade: positives Vorzeichen). Eine negative Korrelation bedeutet, dass die Werte der einen Variable abnehmen, während die Werte der anderen Variable ansteigen (grafisch: von links oben nach rechts unten fallende Gerade: negatives Vorzeichen). Ein niedriger Korrelationskoeffizient drückt aus, dass kein linearer Zusammenhang zwischen den beiden untersuchten Variablen vorliegt. Werte um ± 1 drücken einen perfekten linearen Zusammenhang aus. Bei einer Korrelation um Null besteht kein linearer Zusammenhang zwischen beiden Variablen.

Unter der Voraussetzung, dass intervallskalierte und zuverlässig gemessene Variablen vorliegen, werden niedrige Korrelationen durch die Messwertverteilung verursacht, nicht durch das Messniveau selbst. Die Korrelation zweier Variablen ist nicht gleichzusetzen mit dem Zusammenhang zweier Konstrukte; es besteht auch die Möglichkeit, dass mind. eines der Konstrukte unzuverlässig gemessen wurde (vgl. Cohen et al., 2003³, 53–55).

Der Signifikanztest zum Korrelationskoeffizienten wird auch von der Größe der zugrunde liegenden Stichprobe mit beeinflusst. Signifikanz bedeutet hier nicht immer auch Relevanz. Wichtiger als die Signifikanz des Hypothesentests ist die Größe des Koeffizienten.

Die Bewertung des Koeffizienten hängt dabei stark von der jeweiligen Fragestellung ab. Oft wird folgende Interpretation für Korrelationsbeträge zwischen 0 und 1 verwendet:

Tabelle: Interpretationshilfe für Korrelationskoeffizienten r

r	Interpretation
$< 0,2$	Sehr geringe Korrelation
$< 0,5$	Geringe Korrelation
$< 0,7$	Mittlere Korrelation
$< 0,9$	Hohe Korrelation
$\geq 0,9$	Sehr hohe Korrelation

Solche „Interpretationshilfen“ sind im Prinzip eine Beeinträchtigung der einander ergänzenden Information von Korrelationskoeffizient (Präzision) und Streudiagramm (Differenziertheit).

Nur im oberen Korrelationsbereich ist eine lineare Funktion dann als hohe Korrelation zu interpretieren, wenn z.B. der Range der Funktion auch aus pragmatischer Sicht ausgesprochen hoch ist, was sich sowohl grafisch in einer linearen Anordnung, wie auch numerisch in einem hohen Koeffizienten ausdrückt. Nicht jeder quantitativ hohe Korrelationskoeffizient kann z.B. als qualitativ hohe Korrelation interpretiert werden.

- Beispiel 1: Die Höhe eines Zusammenhangs ist abhängig von den jew. Verteilungsformen der Variablen, jedoch nicht von ihrer jeweiligen Lage auf der x- und y-Achse. Be trägt z.B. bei einer bivariaten Korrelation der Koeffizient 0,9, so würde dieser Wert z.B. durch die Multiplikation oder Division aller Werte einer (oder auch beider) Variablen durch eine Konstante, z.B. 10, nicht verändert werden. Ein gleich hoher numerischer Koeffizient kann demnach inhaltlich völlig andere Zusammenhänge ausdrücken.
- Beispiel 2: Dasselbe gilt für eine Einschränkung eines Ranges i.S.v. Messwertbereichen (Streuung) der jeweiligen Variablen: Ist z.B. der Range der Daten (z.B. 1 bis 100) auf einen kleinen Ausschnitt eingeschränkt (z.B. 1 bis 10), kann zwar immer noch eine lineare Anordnung bzw. Funktion die Folge sein, jedoch nicht mehr ohne Weiteres als „hoch“ verallgemeinernd i.S.v. repräsentativ für den theoretisch möglichen Range beschrieben werden, auch falls dies ein hoher Koeffizient suggerieren sollte.

Gleich hohe Korrelationskoeffizienten können daher je nach Fragestellung oder Datenlage unterschiedlich bedeutsam sein, wie auch unterschiedlich hohe Korrelationskoeffizienten inhaltlich durchaus dasselbe repräsentieren können.

Vor allem für Koeffizienten im mittleren bis niedrigen Korrelationsbereich ergibt sich zusätzlich das Problem (v.a. bei großen Datenmengen), dass Streuungsmuster nicht mehr ohne Weiteres grafisch interpretiert werden können (z.B. aufgrund übereinander liegender Messwertpaare). Ähnlich erscheinende Punktwolken werden oft durch verschiedene Koeffizienten beschrieben.

Kann jedoch ein präzises Maß, wie z.B. ein Korrelationskoeffizient, nicht einmal mehr anhand der Differenziertheit eines Streudiagramms genauer interpretiert werden, dann umso weniger eine Beschreibung qua Adjektivierung.

Von der Verwendung von wertenden Termini wie z.B. „hoch“ oder „gering“ ist daher ausdrücklich abzuraten; stattdessen sind die Werte der Koeffizienten mit dem empirischen wie auch theoretisch möglichen Range der jeweils dazugehörigen Variablen anzugeben.

Wichtig zu wissen ist, dass die Höhe von Korrelationskoeffizienten nicht nur durch Untersuchungsbedingungen (Versuchsanlage), sondern auch durch Merkmale der Stichprobe mit beeinflusst wird (Zufallscharakter, Merkmalsvariabilität/Repräsentativität und Größe).

Die Höhe des Korrelationskoeffizienten wird auch durch die jeweilige Variabilität, also des Ranges der beiden Variablen mitbestimmt. Ist der Range mind. eine der beiden Variablen beeinträchtigt, wird der Korrelationskoeffizient künstlich gesenkt. An verschiedenen Stichproben ermittelte Zusammenhänge ergeben verschiedene Korrelationseffizienten für dieselbe Grundgesamtheit. Der Zufallscharakter repräsentativer Ziehungen bedingt eine spezifische Stichprobenvariation und damit stichprobenspezifische Korrelationen; bei nichtrepräsentativen Ziehungen fällt dieser Effekt weitaus stärker aus:

Der Range wird bei nichtrepräsentativen Ziehungen oft durch den ausschnittartigen Charakter der Ziehung der Stichprobe beeinträchtigt. Werden z.B. Intelligenzfaktoren nicht an einer repräsentativen, sondern nur an einer studentischen Stichprobe untersucht, ist davon auszugehen, dass nur ein oberer, aber kein repräsentativer Ausschnitt aus dem Gesamtrange der Intelligenzfaktoren in die Analyse eingeht.

Eine präzise Schätzung eines Korrelationskoeffizienten für eine Grundgesamtheit ist nur mittels einer hinreichend großen Stichprobe möglich; nur anhand hinreichend großer Stichproben sind Vergleiche verschiedener Korrelationskoeffizienten möglich (unterstellt, sie stammen aus derselben Grundgesamtheit). Diehl & Kohr (1999, 237–239) geben z.B. eine Tabelle für das Ausmaß der Abhängigkeit der Variation von Korrelationskoeffizienten von der Stichprobengröße an.

Tabelle: Abhängigkeit der Variation des Korrelationskoeffizienten

<i>Stichprobengröße</i>	<i>90% Grenzen des Koeffizienten</i>
5	-0.82 bis + 0.82
10	-0.55 bis + 0.55
20	-0.38 bis + 0.38
30	-0.31 bis + 0.31
50	-0.24 bis + 0.24
100	-0.17 bis + 0.17
300	-0.10 bis + 0.10
1000	-0.05 bis + 0.05

Die Tabelle basiert auf einer angenommenen Korrelation von 0,00 und demonstriert, dass mit zunehmender Stichprobengröße die Variation der Koeffizienten abnimmt. Bei $n=10$ liegen 90% der Koeffizienten im Bereich zwischen ± 0.55 . Bei $n=100$ liegen 90% der Koeffizienten im Bereich zwischen ± 0.17 . Bei $n=1000$ liegen 90% der Koeffizienten im Bereich zwischen ± 0.05 .

Von Anfang an sollte vielleicht auch darauf hingewiesen werden, dass aus einem *Korrelationskoeffizient um Null* nicht auf eine Unkorreliertheit der untersuchten Merkmale geschlossen werden kann. Je weiter die Messwertpaare von der Korrelationslinie abweichen, desto mehr unterschätzt der Korrelationskoeffizient die Stärke der bivariat-linearen Beziehung. Die untersuchten Merkmale können durchaus nicht-linear miteinander zusammenhängen, z.B. in Gestalt von U-förmigen Kurven. Ein Korrelationskoeffizient sollte also erst dann berechnet bzw. interpretiert werden, nachdem man sich mittels eines Streudiagramms ein genaues Bild von der konkreten Verteilung der Messwerte verschafft hat.

Bei der Zusammenhangsanalyse gilt es immer den wichtigen Grundsatz zu beachten: Korrelation ist nicht gleich Kausation! Eine Korrelationsanalyse erlaubt nur die Aussage darüber, ob und in welchem *Ausmaß* zwei Variablen zusammenhängen, aber nicht über die *Art* ihres Zusammenhangs, also welche der beiden Variablen die Ursache, welche die Wirkung repräsentiert. Soll ein Kausalmodell überprüft werden, also ob eine Variable systematisch die Variation einer anderen Variable verursacht, dann ist statt der Korrelation das Verfahren der Regression zu wählen. Umgekehrt gilt jedoch: Liegt keine bivariate Korrelation vor, liegt auch keine bivariate Kausation vor.

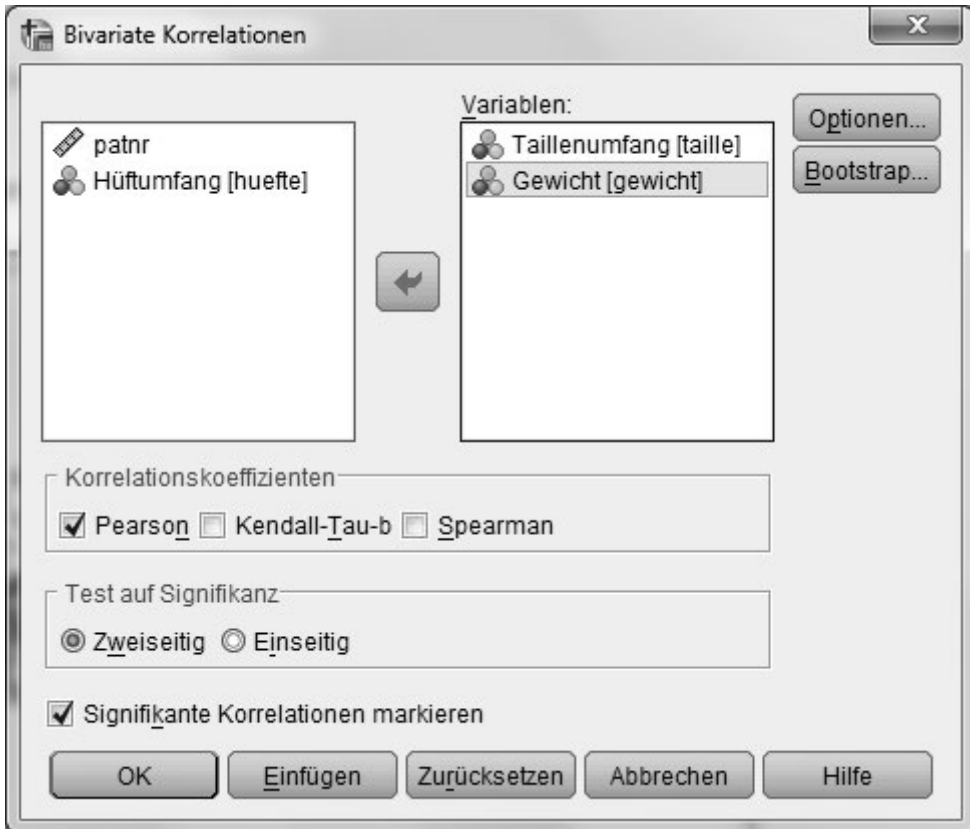
Multivariate Varianten der einfachen (bivariaten) Korrelation sind z.B. die partielle Korrelation, die multiple Korrelation und die kanonische Korrelation. Eine partielle Korrelation (auch: Partialkorrelation) ist z.B. die Korrelation, die zwischen zwei Variablen verbleibt, nachdem die Korrelation rechnerisch entfernt wurde, die durch den Effekt anderer, korrelierender Variablen verursacht worden war.

1.6 Berechnung mit SPSS (Beispiel)

Im Folgenden wird für den ungerichteten Zusammenhang zwischen Taillenumfang und Gewicht das Ausmaß der Korrelation (zweiseitig) nach Pearson berechnet.

Maussteuerung

Pfad: Analysieren → Korrelation → Bivariat...



Markieren und Klicken Sie die beiden Variablen TAILLE und GEWICHT in das Auswahlfenster. Aktivieren Sie als Korrelationskoeffizienten „Pearson“ und unter Test auf Signifikanz „zweiseitig“. Aktivieren Sie „Signifikante Korrelationen markieren“. Wählen Sie unter „Optionen“ „Paarweiser Fallausschluss“ für den Umgang mit fehlenden Werten. Starten Sie die Berechnung mit „OK“.

Syntax:

```

CORRELATIONS
/VARIABLES=taille gewicht
/PRINT=TWOTAIL NOSIG
/MISSING=PAIRWISE .

```

Der Befehl CORRELATIONS fordert die Berechnung einer Pearson Produkt-Moment-Korrelation an. Hinter /VARIABLES= werden die Namen der Variablen (hier: TAILLE GEWICHT) angegeben, deren Zusammenhang berechnet werden soll; es können auch mehr als zwei Variablen angegeben werden. Die Anweisung /PRINT= legt Testrichtung und Markierung signifikanter Korrelationskoeffizienten fest. Über ONETAIL/TWOTAIL kann eine einseitige oder zweiseitige Testrichtung angegeben werden. TWOTAIL (voreingestellt) für eine zweiseitige Testrichtung ist v.a. dann geeignet, wenn die Richtung des Zusammenhangs vorher nicht bekannt ist. Der Zusammenhang kann sowohl positiv, also auch negativ sein. ONETAIL ist dann geeignet, wenn die Richtung des Zusammenhangs vorher bekannt ist bzw. nur eine bestimmte Richtung erwartet wird. Der Zusammenhang darf entweder nur positiv oder nur negativ werden. Über NOSIG können signifikante Korrelationen mit Sternchen hervorgehoben werden. Signifikante Werte bei einem Alpha von 0.05 erhalten ein Sternchen, signifikante Werte bei einem Alpha von 0.01 erhalten zwei Sternchen (vgl. Schendera, 2007, Kap. 19, zur Interpretation von Asterisken). Verwirrend ist, dass SIG für Nichtkennzeichen steht, vermutlich wäre es andersherum leichter nachzuvollziehen. Über MISSING wird der Umgang mit fehlenden Werten festgelegt. PAARWISE schließt Missings paarweise aus, LISTWISE dagegen listenweise. Bei nur zwei Variablen führt dies zum selben Ergebnis; bei mehr als zwei Variablen können drastisch verschiedene N in den jew. Analysen die Folge sein, v.a. beim listenweisen Ausschluss von Missings. Zusätzlich können über INCLUDE anwenderdefinierte Missings in die Analyse einbezogen bzw. über EXCLUDE anwender- und systemdefinierte Missings aus der Analyse ausgeschlossen werden.

Die Syntax erlaubt über die Maussteuerung hinaus u.a. folgende Möglichkeiten: Mit dem Schlüsselwort WITH im Unterbefehl VARIABLES können die Korrelationen zwischen allen Variablen einer ersten Liste und allen Variablen einer zweiten Liste berechnet werden.

```

/VARIABLES=VAR1 VAR2 VAR3 with VAR4 VAR5 VAR6

```

Es werden dabei nur die angegebenen listenweisen Korrelationen ermittelt, z.B. VAR1 mit VAR4 usw., allerdings nicht die Korrelation zwischen VAR1 und VAR2. Die WITH-Option ist sehr geeignet, die manchmal unüberschaubaren Korrelationstabellen auf die interessierenden Variablenpaare zu beschränken.

Mit der Anweisung /MATRIX kann z.B. eine Matrix für Pearson-Korrelationen in einen Datensatz geschrieben werden.

```

CORRELATIONS
/VARIABLES= VAR1 VAR2 VAR3 VAR4 VAR5 VAR6
/MATRIX OUT(*).

```

Im Beispiel werden z.B. für die Variablen VAR1 bis VAR6 die Mittelwerte, Standardabweichungen und N sowie die Korrelationskoeffizienten in einen temporären Datensatz geschrieben. Diese Matrizen können mit mehreren Verfahren weiter verarbeitet werden, u.a. Regression (SPSS Prozedur REGRESSION), Faktoren- und Clusteranalyse (FACTOR, CLUSTER), vgl. auch Schendera (2010).

Output:

Korrelationen

		Taillenumfang	Gewicht
Taillenumfang	Korrelation nach Pearson	1	,803**
	Signifikanz (2-seitig)		,000
	N	711	688
Gewicht	Korrelation nach Pearson	,803**	1
	Signifikanz (2-seitig)	,000	
	N	688	719

** . Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Die Tabelle „Korrelationen“ enthält die Pearson-Korrelationskoeffizienten, Signifikanzen und die Anzahl der Fälle ohne Missings (z.B. N=688 bei der paarweisen Korrelation). In konventionellen Korrelationsmatrizen befindet sich auf der Diagonalen immer der Wert 1, weil jede Variable eine perfekte lineare Korrelation mit sich selbst hat. In konventionellen Korrelationsmatrizen spiegeln sich somit auch die Koeffizienten an dieser Diagonalen. Werden jedoch Korrelationsanalysen mittels der WITH-Option angefordert, sind die Koeffizienten weder gespiegelt, noch enthält die Tabelle eine 1er-Diagonale. Im Beispiel erreicht die Korrelation zwischen Taillenumfang und Gewicht die Signifikanz von $p=0,000$ und einen Korrelationskoeffizienten von 0.803. Da kein negatives Vorzeichen angegeben ist, korrelieren diese beiden Variablen signifikant, positiv und hoch miteinander.

1.7 Fallstricke: Linearität, Scheinkorrelation und Alphafehler-Kumulation

Auch wenn die Korrelation nach Pearson rechnerisch relativ unkompliziert erscheint, werden bei ihrer Interpretation nicht selten Fehler begangen. Die häufigsten Fehler betreffen Linearität, Scheinkorrelation und Alphafehler-Kumulation durch das fishing for significance. Ein nicht seltener Fehler ist die Ableitung der Linearität aus der Signifikanz. Anhand einer Signifikanz kann nicht abgeleitet werden, dass ein linearer Zusammenhang vorliegt i.S.e. Folge der Signifikanz. Im Gegenteil: die Linearität ist die Voraussetzung, überhaupt eine Korrelationsanalyse durchführen zu dürfen. Um es an dieser Stelle nochmals zu wiederholen:

Wenn zwei Variablen miteinander in starker Beziehung stehen, der Zusammenhang *aber nicht* linear ist, ist der Korrelationskoeffizient nach Pearson keine geeignete Statistik zum Messen dieses Zusammenhangs. Etwas knapp formuliert: Die Linearität ist Voraussetzung, nicht Folge der Signifikanz.

Ebenso ist zu beachten, dass eine statistisch hohe Korrelation im rechnerischen Sinne zwar richtig sein mag, aber oft als rein statistischer Zusammenhang nicht ohne Weiteres in die Realität empirischer Zusammenhänge zurück übertragen werden kann. Oftmals kommt eine bivariat-signifikante Korrelation nur dadurch zustande, weil der Einfluss einer Drittvariablen nicht berücksichtigt wurde.

Wenn z.B. eine hohe Korrelation zwischen Lebensalter und dem Preis einer Tageszeitung ermittelt wurde, obwohl zwischen den beiden Variablen eindeutig kein direkter Zusammenhang besteht, dann deshalb, weil eine Drittvariable, die Teuerungsrate, nicht berücksichtigt wurde, die diesen Zusammenhang erst künstlich schuf.

Bei bivariaten Korrelationen besteht nicht selten die Gefahr, dass ein künstlicher Zusammenhang mit einem empirisch gegebenen Zusammenhang verwechselt wird. Wie der Einfluss der Störvariablen kontrolliert werden kann, wird weiter unten an der Partialkorrelation gezeigt.

Eine weitere Gefahr liegt im oft und gerne praktizierten „fishing for significance“. Aus einer planlos angelegten Korrelationsmatrix werden dabei unkritisch die „besten“, also die signifikantesten i.S. einer Hypothesenbestätigung ausgewählt, ohne sich der Gefahr wegen der Alphafehler-Kumulation zufällig entstandener Korrelationen und daher damit einhergehend dem Risiko des Aufsitzens auf Artefakte bewusst zu sein.

1.7.1 Scheinkorrelation und partielle Korrelation

Oft erhält man bei Zusammenhangsberechnungen Werte, die auf hohe bis sehr hohe Korrelationen schließen lassen, z.B. zwischen Güte einer Wohngegend und den jeweils auftretenden Todesfällen durch Krebs, einem Zusammenhang zwischen Alter und BMI bei Heranwachsenden, oder auch einem Zusammenhang der Anzahl der Störche und der jährlichen Geburtenrate.

Eine nachgewiesene Korrelation wird durch eine fehlende Erklärung noch nicht hinfällig. Ihre Abgrenzung von sog. nonsense correlations, z.B. dem Zusammenhang zwischen der Anzahl der Störche und der jährlichen Geburtenzahl, gelingt nur durch sorgfältige inhaltliche und statistische Überprüfung. Das Storch-Beispiel hält einer inhaltlichen Prüfung erst gar nicht stand; eine statistische Überprüfung erübrigt sich daher. Anders verhält es sich bei den beiden anderen Beispielen.

Beim Zusammenhang zwischen der Güte einer Wohngegend (operationalisiert über m²-Preise) und den jeweils auftretenden Todesfällen durch Krebs wurde die ungleiche Verteilung des Alters der Menschen auf die Wohngegenden übersehen. Die teuren Wohngegenden konnten sich nur jüngere Doppelverdienerpaare leisten, die älteren Menschen dagegen nur die einfacheren Wohngegenden. Dass die Menschen in diesen Gegenden eher an Krebs starben, hatte mit ihrem Alter zu tun, aber nicht mit der Wohngegend. Der Zusammenhang zwischen Wohngegend und der Krebssterblichkeit ist also in Wirklichkeit eine (anfangs unentdeckte) Scheinkorrelation.

Mit dem Verfahren der partiellen Korrelation kann aus einem Zusammenhang zwischen zwei Variablen der mögliche Einfluss einer dritten Variablen herausgerechnet werden. Bei der Partialkorrelation wird aus den beiden Variablen QMPREIS und KREBSTOD der Variationsanteil herausgerechnet, der aus der Variation in ALTER linear vorhersagbar ist; anschließend werden die Residuen von QMPREIS und KREBSTOD miteinander korreliert (bei der allgemeineren Semi-Partialkorrelation wird nur aus KREBSTOD der Variationsanteil von ALTER herauspartialisiert; anschließend werden die Residuen von KREBSTOD mit QMPREIS korreliert).

```
CORRELATIONS
/VARIABLES=
  QMPREIS WITH KREBSTOD
/PRINT=TWOTAIL NOSIG
/MISSING=LISTWISE .
```

		Krebssterblichkeit
qm-Preise	Korrelation nach Pearson	,796**
	Signifikanz (2-seitig)	,000
	N	344

** . Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

```
PARTIAL CORR
/VARIABLES=
  QMPREIS KREBSTOD BY ALTER
/SIGNIFICANCE=TWOTAIL
/MISSING=LISTWISE .
```

Kontrollvariablen			qm-Preise	Krebssterblichkeit
Alter	qm-Preise	Korrelation	1,000	,351
		Signifikanz (zweiseitig)	.	,000
		Freiheitsgrade	0	341
Krebssterblichkeit	Krebssterblichkeit	Korrelation	,351	1,000
		Signifikanz (zweiseitig)	,000	.
		Freiheitsgrade	341	0

Der Abschnitt **CORRELATIONS** ermittelt den Zusammenhang zwischen Wohngegend (Variable QMPREISE) und Krebssterblichkeit (Variable KREBSTOD); der Zusammenhang ist signifikant, positiv und hoch ($r=0,796$, $p=0,000$). Es wird im Beispiel der Einfachheit halber davon ausgegangen, dass die weiteren Bedingungen für die Durchführung einer Korrelationsanalyse erfüllt sind.

Der Abschnitt **PARTIAL CORR** rechnet aus dem Zusammenhang zwischen QMPREISE und KREBSTOD den möglichen Einfluss der Variablen ALTER heraus. Das Ergebnis ist eindeutig: Der Zusammenhang zwischen Wohngegend und Krebssterblichkeit ist zwar noch signifikant, jedoch nur noch gering ($r=0,351$, $p=0,000$). Die Signifikanz wird z.T. durch die Größe der Stichprobe ($N=344$) mit beeinflusst. PARTIAL CORR basiert auf dem Pearson-

Korrelationskoeffizienten und hat insofern alle Bedingungen für die Durchführung einer Korrelationsanalyse zu erfüllen. Im Beispiel wird davon ausgegangen, dass die Voraussetzungen aller Variablen erfüllt sind (wichtig ist auch auf die annähernd gleiche Anzahl von Fällen bzw. Missings zu achten).

Der globale Zusammenhang zwischen Alter und BMI bei Heranwachsenden ist unscheinbar, aber nicht minder aufschlussreich. Das folgende Beispiel behandelt nicht eine intervallskalierte, sondern eine kategorial skalierte Störvariable. Effekte kategorial skalierten Störvariablen können über eine partielle Korrelation nicht kontrolliert werden, aber über eine sog. „Blockbildung“ zunächst unkompliziert miteinander verglichen werden.

CORRELATIONS

```
/VARIABLES=
  alter WITH bmi
/PRINT=TWOTAIL NOSIG
/MISSING=LISTWISE .
```

Korrelationen

		BMI
Alter	Korrelation nach Pearson	,231**
	Signifikanz (2-seitig)	,005
	N	144

** . Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

```
SORT CASES BY gschlcht .
SPLIT FILE
  LAYERED BY gschlcht .
```

CORRELATIONS

```
/VARIABLES=age WITH bmi
/PRINT=TWOTAIL NOSIG
/MISSING=LISTWISE .
```

Korrelationen

Geschlecht		BMI	
männlich	Alter	Korrelation nach Pearson	-,028
		Signifikanz (2-seitig)	,812
		N	77
weiblich	Alter	Korrelation nach Pearson	,502**
		Signifikanz (2-seitig)	,000
		N	67

** . Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

```
SPLIT FILE
  OFF .
```

Werden die Daten nach dem Geschlecht (Variable GSCHLCHT) der Personen gruppiert, zeigt sich ein völlig gegenläufiger Effekt. Während der Zusammenhang zwischen ALTER und BMI für die Heranwachsenden insgesamt (N=144) zwar signifikant, aber gering (0,231) war, fallen für Jungen und Mädchen separat ermittelte Korrelationen völlig verschieden aus. Für Jungen (N=67) wird der Zusammenhang zwischen ALTER und BMI erst gar nicht signifikant; für Mädchen (N=77) erreicht der Zusammenhang zwischen ALTER und BMI nicht nur Signifikanz ($p=0,000$), sondern auch ein beträchtliches Ausmaß (0,502). Das erzielte Ergebnis ist also ziemlich komplex:

Die eingangs ermittelte Korrelation war wahrscheinlich durch die Größe der Stichprobe insgesamt (N=144) mit verursacht worden. Darüber hinaus täuschte das Unspezifische dieses ersten Ergebnisses über einen möglicherweise differentiellen Effekt der Variable Geschlecht hinweg, der sich dann in geschlechtsspezifischen Zusammenhängen zwischen ALTER und

BMI niederschlug. Wichtig ist auch hier die annähernd gleiche Anzahl von Fällen bzw. Missings in den miteinander verglichenen Gruppen. Wie verschiedene Korrelationskoeffizienten direkt miteinander verglichen werden können, wird in einem späteren Abschnitt vorgestellt. Bei Kategorialdaten ist die Berechnung partieller Korrelationen durch das Einbeziehen einer Kontrollvariable beim Erstellen von Kreuztabellen (z.B. CROSSTABS) möglich.

Eine Scheinkorrelation entsteht also statistisch dadurch, dass eine dritte Variable zwischen den beiden geprüften Variablen einen künstlichen Zusammenhang herstellt, der dadurch Gefahr läuft, falsch interpretiert zu werden, wenn die Interpretation den Einfluss der Drittvariablen nicht einbezieht (vgl. Cohen et al., 2003³, 75–79; Litz, 2000, 77–91; Pedhazur, 1982, 110–111).

Wichtig ist, dass bei der Herausparsialisierung von Störeffekten durch Partialkorrelationen nur Variablen gewählt werden, die tats. sachlogisch in Frage kommen bzw. überhaupt vorliegen. Wurden die Variablen nicht erhoben, kann ihr Effekt auch nicht ausgeschaltet werden.

Bei der Berechnung der Partialkorrelation sollte man sich über die Theorie, wie auch die zulässigen Kausalmodelle im Klaren sein, die mit diesem Verfahren untersucht werden können. Nach Pedhazur (1982, 110–111) sind z.B. die beiden folgenden Modellstrukturen zwei von mehreren Alternativen:

$X \rightarrow Y \rightarrow Z$ z.B. QMPREISE \rightarrow ALTER \rightarrow KREBSTOD	$X \leftarrow Y \rightarrow Z$ z.B. QMPREISE \leftarrow ALTER \rightarrow KREBSTOD
--	--

Im ersten Modell beeinflusst Y die Effekte von X auf Z (Y ist eine Mediatorvariable); im zweiten Modell ist Y eine gemeinsame Ursache von X und Z und verursacht eine Scheinkorrelation zwischen X und Z.

Die Kausalkorrelation ist dann nicht geeignet, wenn sie Wechselwirkungen oder indirekte Effekte (Moderatorvariablen) untersuchen soll, z.B. wenn mehr Relationen als Prädiktoren vorhanden sind, z.B. wenn X Z sowohl direkt, wie auch indirekt über Y beeinflusst, oder wenn X und Y korrelierte Ursachen von Z sind. Für solche und andere Modelle bietet die Pfadanalyse angemessenere Analysetechniken.

1.7.2 Das Problem der Alphafehler-Kumulierung

Auch wenn die Korrelationsanalyse von manchen Autoren im Kontext der deskriptiven Statistik vorgestellt wird, liegt ihr beim Schluss von einer Stichprobe auf eine Grundgesamtheit dennoch ein Hypothesentest zugrunde. Eine Hypothese wird formuliert, geprüft und anschließend angenommen oder abgelehnt. Zu beachten ist hier, dass eine Korrelation auf zwei Variablen basiert. Die rechnergestützte Datenanalyse (unabhängig davon, ob dies nun per Maus oder Syntax erfolgt) bietet jedoch die Möglichkeit an, Korrelationen von weit mehr als nur zwei Variablen anzufordern und sich aus der Menge der signifikanten Korrelationen die eindrucksvollsten herauszufischen. Wo liegt das Problem? Das Problem dieses sprichwörtli-

chen „fishing for significance“ ist vielfältig und fängt mit dem zugrunde liegenden „running for significance“ an, als ob Zusammenhänge wichtiger wären als ausbleibende Zusammenhänge; im Gegenteil, beide Befundlagen sind formal gesehen gleichwertig. Diese Signifikanz-Fixiertheit führte jedoch zu einem massiven „publishing error“ ausschließlich signifikanter Ergebnisse. Arbeiten, die keinen Unterschied aufzeigen können (vielleicht, weil auch keiner existiert), werden wegen des ‚running for significance‘ und der Alphafehler-Kumulation gar nicht erst veröffentlicht (z.B. Witte, 1980, 51–59; Bredenkamp, 1972, 53) und führen somit auch zu einer verzerrten Repräsentation von (z.B. klinischer) Wirklichkeit (z.B. Turner et al., 2008; Hackbarth, 2008). Eine konkrete Gegenmaßnahme bestand u.a. in der Herausgabe von Zeitschriften, die sich das Ziel gesetzt haben, ausschließlich nicht signifikante Ergebnisse zu veröffentlichen, z.B. das Journal of Negative Results in Biomedicine. Das nächste Problem ist, dass diese Analysepraxis im Prinzip nicht mehr dem klassischen Hypothesentest entspricht; es werden ja gar keine Hypothesen formuliert, sondern gleich alle ausgewählten Zusammenhänge auf einmal analysiert, auf Signifikanzen abgesucht und unreflektiert veröffentlicht (ob überhaupt überprüft wird, ob alle Voraussetzungen für das Anfordern eines Korrelationskoeffizienten erfüllt waren, ist bei solch einer Vorgehensweise nicht sehr wahrscheinlich). Dieses allzu typische „fishing for significance“, die Annahme, je mehr Korrelationsanalysen durchgeführt werden, umso größer sei die Chance, Signifikanzen zu finden, zieht zu keinem Zeitpunkt in Betracht, dass weil so viele Korrelationsanalysen durchgeführt wurden (dies gilt im Prinzip für alle wiederholt angewandten multivariaten Verfahren), aufgrund ihrer großen Zahl auch einige Signifikanzen zufälliger Natur sein könnten. Als Daumenregel gilt, dass bei 100 nacheinander durchgeführten Tests 5 Signifikanzen zufällig zustande gekommen sein können. Das Unwissenschaftliche an dem „fishing for significance“ ist daher, dies nicht nur nicht zu wissen, sondern auch die zufällig zustande gekommenen nicht von den hypothesengeleitet ermittelten (Nicht-)Signifikanzen unterscheiden zu können.

Man stelle sich einen Dartspieler vor, der 100-mal mit verbundenen Augen auf eine Dartscheibe wirft. Jeder Wurf entspräche einer durchgeführten Korrelationsanalyse. Jeder Treffer der Scheibe entspräche einer Signifikanz. Solche Treffer würden aber noch lange nicht beweisen, dass der Betreffende Dart spielen könne, noch dass die Treffer ernst zu nehmen sind, eben weil sie nicht kontrolliert bzw. nicht theoriegeleitet zustande gekommen sind. Und jetzt stelle man sich vor, dass dieser Dartspieler seine 5 mit verbundenen Augen erzielten Treffer veröffentlicht, ohne anzugeben, wie viele Tests (Würfe) er vorgenommen hat, bis er zu seinen Resultaten gelangt ist. Je mehr Analysen (Würfe) durchgeführt werden, umso übertriebener und unglaubwürdiger (weil zufälliger) sind die angegebenen Signifikanzen (Treffer). Ein solches Vorgehen ist in der Wissenschaft, wie auch beim Dartspiel (was durchaus auch eine Wissenschaft für sich sein kann) mehr als zweifelhaft.

Vom klassischen Schema des Hypothesentests her ist bekannt, dass die Wahrscheinlichkeit, die Null-Hypothese korrekt anzunehmen $(1 - 0,05) = 0,95$ beträgt. 0,05 bezeichnet dabei den Alphafehler, die Wahrscheinlichkeit, eine Signifikanz anzunehmen, obwohl in Wirklichkeit die Nullhypothese wahr ist. Bei zwei (unabhängigen) Tests ist diese Wahrscheinlichkeit schon etwas niedriger: $0,95^2 = 0,90$, bei drei Tests ist man schon bei $0,95^3 = 0,86$ angekommen, bei vier bei $0,95^4 = 0,81$ und bei fünf $0,95^5 = 0,77$ usw. Konträr erhöht sich der Alphafehler deutlich von anfangs 0,05, über 0,10, 0,19 hin zu 0,23. Dieses Phänomen, allgemein als

Alphafehler-Kumulierung bekannt, bezieht sich darauf, dass mit zunehmender Testzahl die Wahrscheinlichkeit rapide ansteigt, zufällig zustande gekommene Signifikanzen fälschlicherweise als korrekt anzunehmen.

Beispiel:

Werden nur fünf Variablen miteinander korreliert, entstehen zehn verschiedene Hypothesentests; wem die Formel $(N/2) * (N - 1)$ [im Beispiel entspricht N der Anzahl 5] zu abstrakt ist, kann die Anzahl der Felder links bzw. rechts von der Diagonalen in der Abbildung unten gegenzählen. Die Wahrscheinlichkeit, eine H_0 korrekt anzunehmen, beträgt nur noch 0,60; die Wahrscheinlichkeit, eine H_0 fälschlicherweise anzunehmen dagegen bereits 0,40. Von einem puren Raten ist man an dieser Stelle nicht mehr weit entfernt. Ein weiteres Problem ist: Die seitens SPSS ermittelten Signifikanzen sind nicht gegen die Alphafehler-Kumulation korrigiert. Diese Korrektur muss der Anwender selbst vornehmen.

Der Alphafehler-Kumulation kann jedoch relativ unkompliziert mittels der sog. Bonferonni-Korrektur gegengesteuert werden. Bei der Bonferonni-Korrektur wird einfach der jeweils ermittelte Signifikanzwert mit der Anzahl der Hypothesentests (hier: 10) multipliziert. Liegen die multiplizierten Signifikanzen immer noch unter dem Ausgangs-Alpha (hier: 0,05), dann können sie als signifikant gelten.

Korrelationen

		kör	ene	psy	soz	ess
kör	Korrelation nach Pearson	1	,613**	,517**	,454**	,210**
	Signifikanz (2-seitig)		,000	,000	,000	,003
	N	191	191	191	191	191
ene	Korrelation nach Pearson	,613**	1	,572**	,620**	,144*
	Signifikanz (2-seitig)	,000		,000	,000	,046
	N	191	191	191	191	191
psy	Korrelation nach Pearson	,517**	,572**	1	,418**	,092
	Signifikanz (2-seitig)	,000	,000		,000	,208
	N	191	191	191	191	191
soz	Korrelation nach Pearson	,454**	,620**	,418**	1	,197**
	Signifikanz (2-seitig)	,000	,000	,000		,006
	N	191	191	191	191	191
ess	Korrelation nach Pearson	,210**	,144*	,092	,197**	1
	Signifikanz (2-seitig)	,003	,046	,208	,006	
	N	191	191	191	191	191

** Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

* Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

In der o.a. Abbildung beträgt der Zusammenhang zwischen ESS und ENE $p=0,046$ und ist somit (eventuell zufällig) signifikant. Nach der Korrektur um die Anzahl der durchgeführten Hypothesentests beträgt das $p=0,46$ und ist damit eindeutig nicht mehr signifikant. Der Zusammenhang zwischen ESS und KÖR bleibt nach der Bonferonni-Korrektur signifikant (unkorrigiert: 0,003; korrigiert: 0,03).

1.8 Spezielle Anwendungen

Unter den speziellen Anwendungen werden der direkte Vergleich von Korrelationskoeffizienten, die Überprüfung von Korrelationen auf Gleichheit und die Berechnung der Kanonischen Korrelation vorgestellt.

1.8.1 Vergleich von Korrelationskoeffizienten

Der Vergleich von Korrelationskoeffizienten ist mit SPSS auf mehrere Arten möglich. Die erste Variante erlaubt den Korrelationskoeffizienten einer Stichprobe (CSAMPLE) bei Angabe ihrer Größe (NSAMPLE) gegen den Koeffizienten der Grundgesamtheit (POPCOEFF) zu testen. Die beiden einzigen Annahmen dieses Ansatzes sind, dass der Korrelationskoeffizient der Grundgesamtheit ungleich Null ist und dass die zugrundeliegenden z-Werte in etwa standardnormalverteilt sind. Die Nullhypothese lautet: Der Korrelationskoeffizient der Stichprobe und der Korrelationskoeffizient der Grundgesamtheit sind gleich.

Als Daten werden jeweils die Koeffizienten aus dem Beispiel zum unterschiedlichen Zusammenhang zwischen BMI und ALTER für Jungen und Mädchen verwendet. Der Zusammenhang zwischen ALTER und BMI erreichte für Jungen und Mädchen (N=144) zusammen signifikante 0,231; dieser Wert wird im Beispiel als der Korrelationskoeffizient für die Grundgesamtheit einbezogen. Die Korrelationen für Jungen und Mädchen fielen jedoch jeweils völlig verschieden aus. Für Jungen (N=67) wurde der Zusammenhang zwischen ALTER und BMI (-0,028) erst gar nicht signifikant. Für Mädchen (N=77) erreicht derselbe Zusammenhang signifikante 0,502.

Im nachfolgenden COMPUTE-Beispiel wird der Korrelationskoeffizient der Mädchen (0,502, N=77) gegen den Koeffizienten der Heranwachsenden insgesamt (0,231, N=144) getestet, in einem weiteren Schritt wird zusätzlich der Korrelationskoeffizient für die Jungen auf dieselbe Weise getestet. Es werden jeweils gerundete Werte in die Analyse einbezogen. Der durchgeführte Test ist zweiseitig; der Koeffizient der Grundgesamtheit darf jedoch nicht Null sein. Im COMPUTE-Beispiel können auch negative Koeffizienten angegeben werden. Der POPCOEFF-Wert darf jedoch nicht gleich Null sein.

```
data list free
/ CSAMPLE NSAMPLE POPCOEFF.
begin data
  0,50 77 0,23
-0,03 67 0,23
end data.
compute #ZSAMPLE = .5* (ln ((1 + CSAMPLE) / (1 - CSAMPLE))).
compute #ZPOP = .5* (ln ((1 + POPCOEFF) / (1 - POPCOEFF))).
compute Z = (#ZSAMPLE-#ZPOP)/(1/(sqrt(NSAMPLE-3))).
compute PWERT = 2*(1-cdfnorm(abs(Z))).
exe.
format NSAMPLE (F3.0) CSAMPLE POPCOEFF PWERT (F8.3).
list.
```

Nach DATA LIST werden als CSAMPLE der Koeffizient der jeweiligen Stichprobe (Mädchen, Jungen), nach NSAMPLE die jeweilige Größe der Stichprobe und als POPCOEFF der Koeffizient der Grundgesamtheit eingelesen.

Ausgabe:

CSAMPLE	NSAMPLE	POPCOEFF	Z	PWERT
.500	77	.230	2.71	.007
-.030	67	.230	-2.11	.035

Number of cases read: 2 Number of cases listed: 2

Die ermittelten Signifikanz (PWERT) liegen jeweils unter einem Alpha von 0,05. Der Korrelationskoeffizient der Mädchen unterscheidet sich statistisch signifikant vom Korrelationskoeffizienten der Grundgesamtheit ($p=0,007$); dasselbe gilt für den Korrelationskoeffizienten der Jungen ($p=0,035$).

Die SPSS Command Syntax Reference enthält auch eine Makrovariante für diesen Test; dieses Makro erlaubt jedoch nicht, negative Koeffizienten für die Stichprobe anzugeben.

Der Vergleich von Korrelationskoeffizienten ist nur unter Berücksichtigung des jew. zugrunde liegenden Ranges bzw. der jew. Lage der Verteilung auf x- bzw. y-Achse sinnvoll (vgl. 1.5.2).

1.8.2 Vergleich von Korrelationen auf Gleichheit

Liegen Daten in gruppierter Form vor, z.B. ALTER und BMI nach Geschlecht, so können verschiedene Gruppen auf die Nullhypothese gleicher bivariater Korrelationen überprüft werden, also z.B., ob die Korrelationen zwischen BMI und ALTER für Jungen und Mädchen gleich sind. Der Lösungsansatz basiert auf der Überprüfung der Gleichheit der Steigungen (Regressionshomogenität) auf der Basis z-standardisierter Variablen, wobei die Regressionssteigungen den Korrelationen entsprechen. Diese Variablen werden anschließend in eine Kovarianzanalyse einbezogen, wobei die Interaktion zwischen der Gruppierungsvariable (hier: GSCHLCHT) und einer der beiden z-standardisierten Variablen (welche, ist unerheblich) die Nullhypothese der Gleichheit der Steigungen (Korrelationen) über verschiedene Gruppen hinweg testet.

```
DESCRIPTIVES
  VARIABLES=age bmi
  /SAVE
  /STATISTICS=MEAN STDDEV MIN MAX .
```

```
UNIANOVA
  zBMI BY gschlcht WITH zage
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
```

```
/CRITERIA = ALPHA(.05)
/DESIGN = gschlcht*zage .
```

Beim Ermitteln z-standardisierter Variablen mittels DESCRIPTIVES ist darauf zu achten, dass die Anzahl der Fälle bzw. Missings denen der durchgeführten Korrelationen bzw. Kovarianzanalyse entsprechen.

Zwischensubjektfaktoren

	Wertelabel	N
Geschlecht	1,00	männlich 77
	2,00	weiblich 67

Tests der Zwischensubjekteffekte

Abhängige Variable: Z-Wert: BMI

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Signifikanz
Korrigiertes Modell	19,953 ^a	2	9,977	9,671	,000
Konstanter Term	17,070	1	17,070	16,548	,000
geschlcht * Zage	19,953	2	9,977	9,671	,000
Fehler	145,451	141	1,032		
Gesamt	180,591	144			
Korrigierte Gesamtvariation	165,405	143			

a. R-Quadrat = ,121 (korrigiertes R-Quadrat = ,108)

Diese Interaktion zwischen der Gruppierungsvariable (hier: GSCHLCHT) und der z-standardisierten Variablen ZAGE (z-Standardisierung von AGE) ist signifikant. Die Nullhypothese der Gleichheit der Steigungen (Korrelationen) wird zurückgewiesen. Die Steigungen (Korrelationen) für Jungen und Mädchen sind verschieden. Auch hier gilt jedoch, dass der Vergleich von Korrelationen ohne Berücksichtigung des jew. zugrunde liegenden Ranges bzw. der Lage einer Verteilung auf x- bzw. y-Achse nur eingeschränkt sinnvoll ist (vgl. 1.5.2).

1.8.3 Kanonische Korrelation

Mittels einer kanonischen Korrelation (syn.: set correlation) kann der lineare Zusammenhang zwischen zwei Gruppen (Sätzen, Sets) von Variablen ermittelt werden, z.B. einem Satz von Prädiktoren und einem Satz von Kriterien. Der Einsatz dieses Verfahrens ist also v.a. dann sinnvoll, wenn der Zusammenhang zweier Merkmale untersucht werden soll, die sich jeweils aus mehreren Variablen zusammensetzen. Das größte Problem der kanonischen Korrelation ist die Interpretierbarkeit der ermittelten kanonischen Korrelationskoeffizienten. Maximale lineare Korreliertheit bedingt nicht selten minimale Interpretierbarkeit.

Kurzdarstellung des Verfahrens der kanonischen Korrelation:

Beim Verfahren der kanonischen Korrelation wird in jeder der beiden Gruppen zunächst eine Linearkombination der Variablen (sog. kanonische Variable) mit der Zieleigenschaft ermittelt, dass die Korrelation zwischen den beiden kanonischen Variablen K_1 und K_2 maximal ist. Die Korrelation zwischen K_1 und K_2 ist die erste kanonische Korrelation. Nach der Extraktion von K_1 und K_2 verbleibt für jedes Set eine Restvarianz, für die ebenfalls Linearkombinationen ermittelt werden, die maximal miteinander korrelieren (die ermittelten kanonischen Korrelationen werden sukzessive immer kleiner). Dieser Vorgang wird solange wiederholt, bis die Gesamtvarianz in einem der beiden Sets ausgeschöpft ist. Die Anzahl der kanonischen Korrelationen ist i.A. kleinergleich der Anzahl an Variablen im kleineren der beiden Sets. Die kanonische Korrelation ist immer größergleich wie die größte der einzelnen multiplen Korrelationen. Die kanonischen Variablen innerhalb eines Sets korrelieren miteinander zu Null (Orthogonalität).

Beispiel:

```
GET FILE "C:\...\IHREDATEN.SAV".
INCLUDE "C:\...\SPSS\Canonical correlation.sps".
CANCORR
      SET1=ENE ESS
      / SET2=KÖR PSY SOZ.
```

Das Berechnen der kanonischen Korrelation(en) ist unkompliziert. Über GET FILE wird zunächst der Datensatz angefordert, in dem sich die Variablen befinden, die in die kanonische Korrelation einbezogen werden sollen. Über INCLUDE wird das SPSS Makro „Canonical correlation“ in die Analyse einbezogen. Am Makro muss im Prinzip nichts angepasst werden; es genügt die Einbindung über INCLUDE. Das Makro befindet sich meist im Unterverzeichnis „.../SPSS“. Nach CANCORR werden nur noch die Variablen für die beiden Sets, SET1= und SET2=, angegeben (im Beispiel ENE und ESS für das erste Set und KÖR, PSY und SOZ für das zweite Set), für die die kanonischen Variablen K_1 bis K_n bzw. dazugehörigen kanonischen Korrelationen ermittelt werden sollen.

Ausgabe (Ausschnitt):

Correlations Between Set-1 and Set-2

	kör	psy	soz
ene	,6132	,5715	,6201
ess	,2104	,0916	,1974

Canonical Correlations

1	,761
2	,115

Test that remaining correlations are zero:

	Wilk's	Chi-SQ	DF	Sig.
1	,415	164,596	6,000	,000
2	,987	2,489	2,000	,288

Unter „Correlations Between Set-1 and Set-2“ werden zunächst die multiplen Korrelationskoeffizienten ausgegeben. Unter „Canonical Correlations“ werden die kanonischen Korrelationskoeffizienten zwischen den beiden Sets SET1 und SET2 ausgegeben, diese betragen 0,761 bzw. 0,115. Unter „Test that remaining correlations are zero“ wird nach jeder ermittelten kanonischen Korrelation die Nullhypothese geprüft, dass die nächste kanonische Korrelation gleich Null ist.

Nach der ersten kanonischen Korrelation wird die Nullhypothese zurückgewiesen ($p=0,000$); tatsächlich wird noch eine zweite Korrelation mit 0,115 ermittelt. Nach der zweiten kanonischen Korrelation wird die Nullhypothese jedoch beibehalten ($p=0,288$); es wird keine weitere Korrelation ermittelt.

Zu den wichtigsten Voraussetzungen des Verfahrens gehört, dass die Korrelations- bzw. Varianz-Kovarianzmatrizen auf linearen Zusammenhängen zwischen den einzelnen Variablenpaaren basieren. Darüber hinaus sollten die Variablen u.a. intervallskaliert, multivariat normalverteilt und frei von Ausreißern sein. Bei dichotomen Prädiktorvariablen müssen die Kriteriumsvariablen in allen durch die Prädiktorvariablen definierten Subpopulationen normalverteilt sein. Liegen keine linearen Zusammenhänge vor, kann auf eine nichtlineare, kanonische Korrelationsanalyse mittels der SPSS Prozedur OVERALS ausgewichen werden (Menüpunkt: „Optimale Skalierung ...“).

1.9 Voraussetzungen für die Berechnung des Pearson-Korrelationskoeffizienten

Für die Berechnung einer Pearson-Korrelation sind zahlreiche Voraussetzungen bzw. Besonderheiten zu beachten, die im Folgenden zusammengestellt wurden (vgl. Cohen et al., 2003³):

1. Eine Korrelation erlaubt kein Kausalverhältnis zwischen zwei Variablen zu prüfen (vgl. Regression). Liegt jedoch keine bivariate Korrelation vor, dann besteht auch keine bivariate Kausation.
2. Nonsenskorrelationen, z.B. zwischen der Anzahl nistender Störche und der Anzahl neugeborener Kinder, sind von vornherein sachlogisch auszuschließen.
3. Die paarweisen Messungen x_i und y_i müssen zum selben Objekt gehören. In anderen Worten: Die untersuchten Merkmale werden dem gleichen Element einer Stichprobe entnommen.
4. Für den benötigten Stichprobenumfang liegen mehrere Daumenregeln vor; diese empfehlen mind. $N=50$ (Green, 1991) bzw. $N=30$ (Borg & Gall, 1989). Liegen schiefe Verteilungen oder messfehlerbehaftete Daten vor, werden weit mehr Fälle benötigt. Weniger Messwertpaare (z.B. $N=6$) sind dann zulässig, wenn die Daten annähernd perfekt linear angeordnet sind und eine Korrelation für einen idealen Zusammenhang berechnet werden soll.

5. Die Variablen sind jeweils intervallskaliert. Treppenartige Muster in einem Streudiagramm sind ein Hinweis darauf, dass mind. eine der beiden Variablen nicht intervallskaliert ist.
6. Messfehler (Zuverlässigkeit, Bias): Beide Variablen sollten hoch zuverlässig, also idealerweise ohne Fehler gemessen worden sein. Messfehler verringern die Korrelation. Ein bivariater Bias überhöht die Korrelation (z.B. Selektionsmechanismen, die Division beider Variablen durch denselben Zähler oder Nenner, oder bei der Korrelation von einzelnen Items mit zusammengesetzten Maßen, sog. Teil-Ganzes-Korrelationen).
7. Wenn von einer Stichprobe auf eine Grundgesamtheit geschlossen werden soll, z.B. mittels eines Signifikanztests, müssen die Daten beider Variablen jeweils einer Normalverteilung folgen. Liegt keine sog. zweidimensionale Normalverteilung vor, kann ein nicht-lineares Zusammenhangsmaß oder eines auf einem niedrigeren Skalenniveau berechnet werden.
8. Zwischen beiden Variablen besteht ein linearer Zusammenhang. Der Pearson-Korrelationskoeffizient ist nicht zur Berechnung des Ausmaßes eines Zusammenhangs geeignet, wenn der Zusammenhang zwischen beiden Zufallsvariablen nichtlinear oder kurvilinear ist. Wenn die Messwertpaare in annähernd parallelem (möglichst dichtem) Abstand um die Korrelationslinie streuen, ist Homoskedastizität gegeben (Konstanz der Streuungen).
9. Läge dagegen Heteroskedastizität vor, würde der als Durchschnittswert ermittelte Standardschätzfehler weite Streuungen unterschätzen und dicht anliegende Streuungen überschätzen. Insofern wäre der Durchschnittswert nicht mehr für die Abbildung des linearen Zusammenhangs brauchbar. In diesem Fall bestünde die Möglichkeit, die Gesamtstreuung durch Bildung von Untergruppen zu unterteilen und die Korrelationskoeffizienten für die einzelnen Bereiche separat zu ermitteln. Das ähnliche anmutende Phänomen unterbrochener Linien erfordert ein anderes Vorgehen.
10. Die Korrelationslinie sollte Kontinuität aufweisen. Weist eine Linie (eine oder auch mehrere) Lücken auf, so ist dies als Hinweis auf eine nicht repräsentative Stichprobe zu interpretieren. Es sollte die Stichprobe über weitere Ziehungen so weit aufgefüllt werden, dass die Lücken geschlossen sind. Im anderen Falle wäre ein doppeltes Problem die Folge: Verzerrte Schätzungen für eine nichtrepräsentative Stichprobe.
11. Ausreißer liegen nicht vor. Ausreißer, die auf Störungseinflüssen basieren, sind zu entfernen, da sie zu irreführenden Korrelationen führen (v.a. bei kleinen Stichproben). Ausreißer in Richtung der Funktion überhöhen den Korrelationskoeffizienten; Ausreißer (z.B.) orthogonal zur Funktion führen zu einer Unterschätzung des Korrelationskoeffizienten. Lücken (s.o.) zu Ausreißern, die valide sind, aber auf einer nichtrepräsentativen Stichprobe basieren, sind über zusätzliche Ziehungen aufzufüllen.
12. Die Höhe des Korrelationskoeffizienten wird durch die jeweilige Variabilität der beiden Variablen mitbestimmt. Ist der Range mind. einer der beiden Variablen beeinträchtigt, wird der Korrelationskoeffizient künstlich gesenkt (Cohen et al., 2003³, 57). Dieses Problem kann nur über eine repräsentative Stichprobe behoben werden. Eine Korrelation sollte grafisch weder über den vorliegenden Messwertbereich hinaus geführt, noch darüber hinaus interpretiert werden.
13. Die Signifikanz wird auch von der Größe der zugrunde liegenden Stichprobe mit beeinflusst. Signifikanz bedeutet nicht immer auch Relevanz.