Skript

Computergestützte Statistik II

Prof. Dr. Bernd Rönz

Humboldt-Universität zu Berlin
Wirtschaftswissenschaftliche Fakultät
Institut für Statistik und Ökonometrie

2000

Inhaltsverzeichnis

1. Vorwort **1**

2. Überprüfung von Zusammenhängen **5**

2.1. Explorative Zusammenhangsanalyse 8

2.2. Beziehung zwischen nominal- bzw. ordinalskalierten Daten 29

2.2.1. Die Kontingenztabelle 29

2.2.2. Tests auf Unabhängigkeit zweier Variablen 42

2.2.2.1. Chi-Quadrat-Unabhängigkeitstest nach Pearson 42

2.2.2.2. Likelihood-Quotienten-Test (likelihood-ratio-test) 46

2.2.2.3. Linear-by-Linear Association 46

2.2.2.4. Fisher's exakter Test 46

2.2.2.5. Beispiele 47

2.2.3. Zusammenhangsmaße 51

2.2.3.1. Korrelationen 51

2.2.3.2. Assoziationsmaße für nominalskalierte Variablen 53

2.2.3.3. Zusammenhangsmaße für ordinalskalierte Daten 67

2.2.3.4. Zusammenhangsmaß für eine intervallskalierte Variable mit einer nominalskalierten Variablen (Eta-Koeffizient) 74

2.2.3.5. Kappa-Koeffizient 77

2.2.3.6. Relatives Risiko 80

2.2.3.7. McNemar - Test 86

2.2.3.8. Cochran's und Mantel-Haenszel Test 94

3. Regressionsanalyse **103**

3.1. Lineare Regression 103

3.1.1. Modell der linearen Regression 103

3.1.2. Schätzung des linearen Regressionsmodells und Hypothesenprüfungen . . 105

3.1.3. Zur Modelldiagnose . . . 112
3.2. Die Optionen unter SPSS . . . 117
3.3. Beispiel . . . 129
3.4. Kurvenanpassung . . . 151

4. Reliabilitäts- und Homogenitätsanalyse von Konstrukten 161
4.1. Reliabilitätsanalyse . . . 162
4.1.1. Statistiken und Tests der Reliabilitätsanalyse . . . 165
4.1.2. Modelle der Reliabilitätsanalyse . . . 180
4.2. Homogenitätsanalyse . . . 186
4.2.1. Exkurs zur Faktoranalyse . . . 186
4.2.2. Die Faktoranalyse unter SPSS . . . 195
4.2.3. Die Verwendung der Faktoranalyse zur Homogenitätsprüfung . . . 201

Anhang A: Beispiel für Lowess 209

Anhang B: Testentscheidung unter Verwendung statistischer Software 217

Anhang C: Informationen zum Chi - Quadrat - Unabhängigkeitstest 223

Anhang D: Informationen zum Likelihood - Ratio - Test 229

Anhang E: Informationen zu Fisher's exaktem Test 233

Anhang F: t-Verteilung 239

Anhang G: Konkordanz, Diskordanz und Ties in Kontingenztabellen 241

Anhang H: Eigenwerte 251

Literaturverzeichnis 253

Stichwortverzeichnis 257

Abbildungsverzeichnis

2.1. Menü „Graphs“ . . . 9
2.2. Dialogfeld „Scatterplot“ . . . 9
2.3. Dialogfeld „Simple Scatterplot“ . . . 9
2.4. Dialogfeld „Scatterplot Options“ . . . 10
2.5. Dialogfeld „Scatterplot Options: Fit Line“ . . . 11
2.6. Beispiel eines einfachen Scatterplots mit linearer Kurvenanpassung . . . 12
2.7. Dialogfeld „Scatterplot Options: Fit Line“: Lowess . . . 12
2.8. Einfacher Scatterplot mit Lowess Kurvenanpassung, % of points to fit: 35 . . . 14
2.9. Einfacher Scatterplot mit Lowess Kurvenanpassung, % of points to fit: 65 . . . 14
2.10. Einfacher Scatterplot mit kubischer Kurvenanpassung . . . 15
2.11. Point Selection Ikon . . . 15
2.12. Einfacher Scatterplot mit Punkt-Identifizierung . . . 16
2.13. Einfacher Scatterplot mit linearer Kurvenanpassung und Gruppenvariable . . . 17
2.14. Dialogfeld „Scatterplot-Matrix“ . . . 17
2.15. Beispiel einer Scatterplot-Matrix mit linearer Kurvenanpassung . . . 18
2.16. Dialogfeld „3-D Scatterplot“ . . . 19
2.17. Beispiel eines 3-D Scatterplots . . . 19
2.18. Dialogfeld „3-D Scatterplot: Options“ . . . 20
2.19. Beispiel eines 3-D Scatterplots mit Floor-Projektion . . . 20
2.20. Menüleiste des Chart Editors . . . 21
2.21. Dialogfeld „3-D Rotation“ . . . 21
2.22. Menüleiste des Spinmode . . . 21
2.23. Beispiel eines rotierten 3-D Scatterplots . . . 22
2.24. Beispiel eines 3-D Spektral-Plots . . . 23
2.25. Beispiel eines 3-D Space-Plots . . . 23
2.26. Beispiel eines 3-D Surfaceplots-Plots mittels Distance Weighted Least Squares . . . 23
2.27. Dialogfeld „Overlay Scatterplot“ . . . 24

2.28. Beispiel eines Overlay Scatterplots . . . 25
2.29. Dialogfeld „Bar Charts“ . . . 26
2.30. Dialogfeld „Define Clustered Bar: Summaries for Groups of Cases“ . . . 26
2.31. Clustered Bar Chart der Variablen Schichteinstufung und Schule . . . 27
2.32. Clustered Bar Chart der Variablen Schule und Schichteinstufung . . . 28
2.33. 3-D Balkendiagramm der Variablen Schichteinstufung und Schule . . . 28
2.34. Menü „Analyze“ und Menü „Descriptive Statistics“ . . . 32
2.35. Dialogfeld „Crosstabs“ . . . 33
2.36. Dialogfeld „Crosstabs: Table Format“ . . . 34
2.37. Dialogfeld „Crosstabs: Cell Display“ . . . 34
2.38. Dichtefunktion der Chi-Quadrat-Verteilung und Entscheidungsbereiche des Chi-Quadrat-Unabhängigkeitstests . . . 45
2.39. Dialogfeld „Crosstabs: Statistics“ . . . 47
2.40. Dialogfeld „Means“ . . . 76
2.41. Dialogfeld „Means: Options“ . . . 76
2.42. Dialogfeld „Two-Related-Samples Tests“ . . . 91
2.43. Dialogfeld „Exact Tests“ . . . 91

3.1. Lineare Regressionsfunktion der Grundgesamtheit . . . 105
3.2. Entscheidungsbereiche des Durbin-Watson-Tests . . . 113
3.3. Dialogfeld „Linear Regression“ . . . 117
3.4. Dialogfeld „Linear Regression: Options“ . . . 120
3.5. Dialogfeld „Linear Regression: Set Rule“ . . . 122
3.6. Dialogfeld „Linear Regression: Statistics“ . . . 122
3.7. Dialogfeld „Linear Regression: Plots“ . . . 125
3.8. Dialogfeld „Linear Regression: Save“ . . . 127
3.9. Dialogfeld „Bivariate Correlations“ . . . 130
3.10. Dialogfeld „Bivariate Correlations: Options“ . . . 130
3.11. Dialogfeld „Partial Correlations“ . . . 131
3.12. Histogramm der standardisierten Residuen . . . 145
3.13. Normal P-P-Plot der standardisierten Residuen . . . 145
3.14. Dialogfeld „Line Charts“ . . . 147
3.15. Dialogfeld „Define Simple line: Summaries for Groups of Cases“ . . . 147
3.16. Line-Plot der standardisierten Residuen gegen die Monate . . . 148
3.17. Line-Plot der standardisierten Residuen gegen den Pro-Kopf-Verbrauch von Eiscreme . . . 148
3.18. Line-Plot der standardisierten Residuen gegen die Temperatur . . . 148

3.19. Line-Plot der Temperatur gegen die Monate . . . 149
3.20. Boxplot der standardisierten Residuen . . . 149
3.21. Dialogfeld „Curve Estimation“ . . . 152
3.22. Dialogfeld „Curve Estimation: Save“ . . . 152
3.23. Scatterplot für Pro-Kopf-Verbrauch an Eiscreme und mittlere Temperatur . . . 155
3.24. Lineplot für Pro-Kopf-Verbrauch an Eiscreme und mittlere Temperatur mit kubischem Modell . . . 157
3.25. Line Plot für Telefon und Zeit mit kubischem Modell und Vorhersagewerten . . . 159

4.1. Dialogfeld „Reliability Analysis“ . . . 163
4.2. Dialogfeld „Reliability Analysis: Statistics“ . . . 165
4.3. Dialogfeld „Factor Analysis“ . . . 195
4.4. Dialogfeld „Factor Analysis: Descriptives“ . . . 196
4.5. Dialogfeld „Factor Analysis: Extraction“ . . . 198
4.6. Dialogfeld „Factor Analysis: Rotation“ . . . 199
4.7. Dialogfeld „Factor Analysis: Factor Scores“ . . . 200
4.8. Dialogfeld „Factor Analysis: Options“ . . . 200
4.9. Faktor-Screeplot . . . 204
4.10. Faktor-Plot der rotierten Faktorlösung . . . 207
A.1. Bereich für die Vorhersage von Y an der Stelle x_1 . . . 210
A.2. Nachbarschaftsgewichte im Bereich um x_1 . . . 211
A.3. Beobachtete und geschätzte Y-Werte im Bereich um x_1 . . . 212
A.4. Bereich für die Vorhersage von Y an der Stelle x_5 . . . 212
A.5. Nachbarschaftsgewichte im Bereich um x_5 . . . 213
A.6. Beobachtete und geschätzte Y-Werte im Bereich um x_5 . . . 213
A.7. Beobachtete und geschätzte Y-Werte für alle x_i . . . 214
A.8. Residuen-Plot . . . 215
A.9. Robustheitsgewichte $G(x_i)$ . . . 215
A.10. Funktion der Robustheitsgewichte . . . 215
B.1. Signifikanzniveau α und Entscheidungsbereiche beim rechtsseitigen Test . . . 218
B.2. Überschreitungswahrscheinlichkeit $P = P(V > v|\vartheta_0)$ bei Gültigkeit der H_0 . . . 219
B.3. Signifikanzniveau $\alpha = P(V > c|\vartheta_0)$ und Überschreitungswahrscheinlichkeit $P = P(V > v|\vartheta_0)$ bei Gültigkeit der NullhypotheseH_0 für einen rechtsseitigen Test . . . 220
B.4. Signifikanzniveau $\alpha = P(V > c|\vartheta_0)$ und Überschreitungswahrscheinlichkeit $P = P(V > v|\vartheta_0)$ bei Gültigkeit der Nullhypothese H_0 für einen rechtsseitigen Test . . . 221

Verzeichnis der verwendeten SPSS-Dateien

Datei	Verwendung in	Seite
10items.sav	Kapitel 4	164
allbus.sav	Beispiel 2.6, 2.9	30, 83
autofreier_sonntag.sav	Beispiel 2.13	91
europa.sav	Beispiel 2.1, 2.3, 2.5	11, 18, 24
hke.sav	Beispiel 2.12	94
icecream.sav	Abschnitt 3.3	129
kappa.sav	Beispiel 2.10	78
lowess.sav	Anhang A	209
mieten.sav	Beispiel 2.2	16
percept_91_96.sav	Beispiel 2.14	96
rauchen.sav	Beispiel 2.11	81
schwierigkeiten.sav	Beispiel 2.7	49
studium.sav	Beispiel 2.8, Anhang G	69
telefon.sav	Beispiel 3.2	157
verkehr.sav	Beispiel 2.4	19

1. Vorwort

Statistische Datenanalyse in Wissenschaft, Wirtschaft, Verwaltung und Politik ist heutzutage ohne Unterstützung durch Computer kaum durchführbar. Rechenaufwendige Analysen können in kürzester Zeit bewältigt werden; andererseits sind eine Reihe von statistischen Methoden erst durch schnellere Computer mit ausreichender Speicherkapazität praktisch durchführbar geworden. Für die Bewältigung solcher Datenanalysen stehen eine Vielzahl von statistischen Software-Paketen zur Verfügung. Die Anwendung dieser statistischen Datenanalysesysteme erfordert jedoch ein umfangreiches Wissen in Statistik, um entsprechend der fachlichen Problemstellung die angemessenen statistischen Verfahren unter Berücksichtigung ihrer Voraussetzungen auszuwählen und aus den Ergebnissen, die der Computer als Output liefert, die richtigen Schlüsse zu ziehen und Fehlinterpretationen zu vermeiden. Computergestützte Statistik (Computer-Statistik, Computational Statistics) wird deshalb als Vermittlung von Kenntnissen über die Aufbereitung und Analyse statistischer Daten verstanden, zu deren Durchführung der Computer und ein statistisches Datenanalysesystem (Software-Paket) als Hilfsmittel eingesetzt wird.

Für die Lehrveranstaltung „Comptergestützte Statistik“ wurde SPSS for Windows Release 9.0.0 (im weiteren kurz SPSS genannt, http://www.spss.com) ausgewählt, nicht weil es den anderen Statistik-Softwarepaketen wesentlich überlegen ist, sondern weil es u.E. in vielen Wirtschaftsbereichen praktisch eingesetzt wird. Zum anderen sind spezielle Kenntnisse der Programmsyntax für die grundlegende Handhabung von SPSS nicht erforderlich, da es im wesentlichen menü- und dialoggesteuert ist und viele Aufgaben durch Anklicken mit der Maus ausgewählt und abgearbeitet werden können. Grundlagen der Handhabung von SPSS werden im weiteren vorausgesetzt.

In anderen Statistik-Lehrveranstaltungen im Hauptstudium wird auch andere Statistik-Software herangezogen, wie z.B. XploRe (http://www.xplore-stat.de/index js.html) und S-Plus (http://www.splus.mathsoft.com), so dass keine Einseitigkeit in der Handhabung von statistischen Datenanalysesystemen auftreten kann.

Man sollte sich jedoch vor dem Gedanken hüten, dass man nunmehr ein Statistik-Softwarepaket hat, es mit Daten füttert, statistische Prozeduren mechanisch abarbeitet und anschließend etwas Vernünftiges herauskommt.

Zum einen ist das Wissen um den Inhalt der verwendeten Daten und die diesen Daten zugrundeliegende Problemstellung eine wesentliche Voraussetzung für eine sachgerechte statistische Auswertung. Solange man zur Beantwortung einer wissenschaftlichen oder praktischen Fragestellung selbst die dafür benötigten Daten erhebt (Primärerhebung), ist dies im allgemeinen gegeben. Werden jedoch Sekundärdaten verwendet, so entfallen entscheidende Etappen der statistischen Arbeit:

- die Planungs- und Definitionsphase mit der Formulierung der Problem- und Zielstellung der Untersuchung und ihrer theoretischen Begründung, mit der Festlegung der Grundgesamtheit, der statistischen Einheiten, der zu erhebenden Variablen (einschließlich ihrer Adäquation), der Art und des Umfangs der Erhebung,

- die Erhebungsphase mit der Erarbeitung des Erhebungsinstrumentes (z.B. eines Fragebogens), der Festlegung des Erhebungsplans und der Durchführung der eigentlichen Erhebung,

- die Aufbereitungsphase, worunter hier nur die technische Aufbereitung (wie Codierung der Daten, Übertragung auf computerlesbare Medien, Datenprüfung und -korrektur) verstanden wird.

Alle Festlegungen dieser Arbeitsetappen können mit der eigenen Zielstellung der statistischen Untersuchung differieren. Auch bei einer statistischen Sekundäranalyse sollte man sich deshalb soweit wie möglich einen Überblick über diese Etappen verschaffen, um Fehlverwendungen der Daten vorzubeugen und eine sachgerechte statistische Auswertung zu gewährleisten.

Eine weitere Fehlerquelle für unrealistische oder gar unsinnige statistische Ergebnisse ist eine nicht adäquate Auswahl der anzuwendenden statistischen Verfahren und Modelle, die oftmals in der Unkenntnis der Voraussetzungen dieser Methoden begründet liegt und besonders im Zusammenhang mit der gegebenen Menü- und Dialogsteuerung der Software durch unbedarfte Handhabung hervorgerufen wird. Es ist immer noch der Nutzer der Software, der entscheidet, welches Verfahren auf welche Daten angewendet werden soll. Die statistische Software liefert auch bei Nichteinhaltung der Voraussetzung (fast immer) irgendwelche Ergebnisse. Zum Beispiel werden auch für eine nominalskalierte Variable, deren Ausprägungen in Form von Schlüsselzahlen vorliegen, (arithmetischer) Mittelwert und Streuung berechnet.

Vor Beginn der statistischen Datenanalyse (Analysephase) und somit der Methodenauswahl ist deshalb, ausgehend von der fachlichen Problemstellung , das statistische Ziel der Untersuchung zu definieren. Dies beinhaltet zwei grundsätzliche Aspekte:

- zum einen die Entscheidung über die Anzahl der gleichzeitig zu untersuchenden Variablen, d.h. ob

  - eine **univariate Analyse** (Einbeziehung nur einer Variablen)
  - eine **bivariate Analyse** (Einbeziehung zweier Variablen) oder
  - eine **multivariate Analyse** (Einbeziehung von mehr als zwei Variablen)

  durchzuführen ist,

- zum anderen die Festlegung der statistischen Herangehensweise, d.h. ob

  - eine Beschreibung der Untersuchungsgesamtheit anhand ausgewählter Variablen erfolgen soll. Hierbei kommen vor allem Methoden der **deskriptiven Statistik** zur Anwendung.
  - die Analyse der Generierung von statistisch überprüfbaren Hypothesen dienen soll. Im Vordergrund stehen dabei Verfahren der **explorativen Datenanalyse**.
  - aus Forschungshypothesen abgeleitete statistische Hypothesen getestet werden sollen. Dies erfordert Methoden und Modelle der **induktiven Statistik**.

  Diese drei Herangehensweisen sind nicht im Sinne „entweder ... oder“ zu verstehen, sondern sie werden vielmehr Schritte einer stufenweisen Analyse sein.

Diese Aspekte entscheiden über die Auswahl der statistischen Methode(n), wobei in jeder Phase der statistischen Datenanalyse stets erneut zu prüfen ist, ob die Voraussetzungen ihrer Anwendung auch gegeben sind.

Die Lehrveranstaltung will diese konzeptionelle Vorgehensweise computergestützter Statistik vermitteln. Sie ist kein Kurs zur Vermittlung der ausgewählten Software. Das eingesetzte Software-Paket soll möglichst im Hintergrund bleiben, d.h., es soll wirklich als ein technisches Hilfsmittel begriffen werden, das jederzeit gegen ein anderes ausgetauscht werden kann.

Im Mittelpunkt der Lehrveranstaltung „Computergestützte Statistik II“ stehen folgende Problemkreise:

- *Überprüfung von Zusammenhängen zwischen Variablen*
  Die wohl am häufigsten gestellte Frage bei sozio-ökonomischen Untersuchungen ist die nach möglichen Zusamenhängen bzw. Abhängigkeiten von Variablen. Den Beitrag, den die Statistik zur Beantwortung dieser Frage leisten kann, besteht in der Überprüfung, ob sich aufgrund von Beobachtungen der Variablen an einer Vielzahl von statistischen Einheiten ein solcher Zusammenhang quantitativ nachweisen läßt. Von den Werkzeugen der

explorativen Datenanalyse wird dafür vor allem der Scatterplot, die Scatterplot-Matrix und der 3D-Scatterplot verwendet. Die Auswertung von Kontingenztabellen und die Berechnung von Assoziations-, Kontingenz- und Korrelationskoeffizienten sind Möglichkeiten zur Beurteilung und Messung der Stärke von Zusammenhängen auf verschiedenen Skalenniveaus. Für die konfirmatorische Prüfung von Zusammenhangshypothesen stehen verschiedene Tests zur Verfügung, wie z.B. der Chi-Quadrat-Unabhängigkeitstest nach Pearson, der Likelihood-Ratio-Test, der Mantel-Haenszel-Test, Fisher's exakter Test und der t-Test.

- *Feststellung der Abhängigkeit von Variablen (Regressionsanalyse)*
  Die aus dem Grundstudium bekannte lineare Regressionsanalyse, mit der die Abhängigkeit von Variablen quantifiziert werden kann, wird bezüglich der Schätzung und Hypothesenprüfung unter Verwendung von SPSS behandelt. Einen breiten Raum nehmen vor allem die konfirmatorischen Auswertungen und die Modelldiagnostik ein, d.h. eine eingehende Prüfung der Voraussetzungen des Regressionsmodells und eine detaillierte Analyse der Residuen.

- *Reliabilitäts- und Homogenitätsanalyse von Konstrukten*
  Es gibt bei vielen praktischen Untersuchungen bzw. Forschungsproblemen Merkmale, die nicht oder nur sehr schwierig beobachtet werden können, d.h. einer Messung bzw. Beurteilung nicht zugänglich sind. Solche Merkmale werden theoretisch als Konstrukte bzw. auch als latente Variablen bezeichnet. Diese Konstrukte können nur über die Beobachtung einer Vielzahl von Quellvariablen erfaßt werden, die verschiedene Aspekte dieses theoretischen Konstrukts beinhalten und die anschließend in geeigneter Weise zu einer neuen Variablen zusammenzufassen sind. Bevor mittels solchermaßen erstellter synthetischer Variablen Hypothesen z.B. über Unterschiede zwischen Objekt- oder Personengruppen oder über Zusammenhänge geprüft bzw. Theorien aufgebaut oder verifiziert werden können, muß eine Einschätzung ihrer Konstruktion vorgenommen werden. Mittels der Statistik können dabei Aussagen über ihre Reliabilität (Zuverlässigkeit) und Homogenität getroffen werden. Für die Reliabilitätsanalyse stehen u.a. der Reliabilitätskoeffizient Alpha von Cronbach und das Split-Half Modell zur Verfügung. Zur Homogenitätsprüfung wird auf die Faktorenanalyse zurückgegriffen.

Diese Problemkreise werden an Beispielen ausführlich diskutiert und für SPSS demonstriert.

2. Überprüfung von Zusammenhängen zwischen Merkmalen

Die wohl am häufigsten gestellte Frage bei sozio-ökonomischen Untersuchungen ist die nach möglichen Zusammenhängen bzw. Abhängigkeiten von Variablen. Den Beitrag, den die Statistik zur Beantwortung dieser Frage leisten kann, besteht in der Überprüfung, ob sich aufgrund von Beobachtungen der Variablen an einer Vielzahl von statistischen Einheiten ein solcher Zusammenhang *quantitativ* nachweisen läßt. Im Ergebnis einer solchen Analyse wird somit die Aussage stehen: *Statistisch* läßt sich ein Zusammenhang bzw. eine Abhängigkeit zwischen den Variablen nachweisen bzw. nicht nachweisen, wobei insbesondere die Stärke und die Form der Beziehung quantifiziert werden sowie die Richtung des Einflusses ermittelt wird. Dies impliziert sofort, dass mittels statistischer Methoden keine Kausalitätsuntersuchung durchgeführt, sondern festgestellt wird, ob die untersuchten Variablen in einem bestimmten Ausmaß eine gemeinsame Variation aufweisen bzw. ob die gemeinsame Häufigkeitsverteilung der Variablen eine Beziehung erkennen läßt. Eine statistisch nachgewiesene Beziehung kann auch bei Variablen auftreten, für die sich keine sachlogische Erklärung geben läßt (Nonsense-Beziehung). Jede Untersuchung von Abhängigkeiten und Zusammenhängen sollte deshalb fachwissenschaftlich fundiert sein, um von vornherein sachlogisch unsinnige Analysen zu vermeiden.

Bereits in der „Computergestützten Statistik I“ (Rönz, B. (2001), Abschnitt 6.2.2.3) wurde mit der einfachen Varianzanalyse eine statistische Methode zur Untersuchung der Beziehung von Variablen vorgestellt. Mit dem dort durchgeführten Test zur Prüfung auf signifikante Unterschiede in den Mittelwerten mehrerer Grundgesamtheiten wurde die Frage beantwortet, ob die Unterschiede in den Mittelwerten der zu untersuchenden Variablen (Zielgröße) auf die Wirkung einer anderen Variablen (Faktor) mit verschiedenen Ausprägungen (Faktorstufen), nach denen sich die Grundgesamtheiten unterscheiden, zurückzuführen ist. Die einfache Varianzanalyse setzt jedoch eine metrisch skalierte abhängige Variable bei beliebiger Skalierung der beeinflussenden Variablen sowie normalverteilte Grundgesamtheiten mit gleicher Varianz voraus.

Dies zeigt deutlich, dass die statistischen Methoden zur Prüfung von Zusammenhängen und Abhängigkeiten bestimmte Voraussetzungen an das Skalenniveau der zu analysierenden Variablen stellen. Umgekehrt heißt das: Sind die zu untersuchenden Variablen mit ihren Skalenniveaus gegeben, dann ist ein solches statistisches Verfahren auszuwählen, das diesen Skalenniveaus genügt.

Bei metrisch skalierten Variablen (vor allem bei stetigen Variablen, aber auch bei diskreten Variablen mit sehr vielen möglichen Variablenwerten) tritt es realtiv selten auf, dass exakt derselbe Variablenwert wiederholt beobachtet wird. Ein beobachtetes Wertetupel $(x_{i1}, x_{i2}, \dots)$ der Variablen $X_1, X_2, \dots$ tritt in der Stichprobe im allgemeinen mit einer kleinen absoluten Häufigkeit auf, in der Regel sogar mit $\mathrm{h}(x_{i1}, x_{i2}, \dots) = 1$. Das Augenmerk bei der Untersuchung von Zusammenhängen metrisch skalierter Variablen wird deshalb auf die gemeinsamen Variation der Variablen gerichtet sein; also in dem zunächst einmal sehr allgemeinen Sinne, ob mit großen bzw. kleinen Werten der einen Variablen große bzw. kleine Werte der anderen Variablen auftreten.

Die Untersuchung von Zusammenhängen zwischen nominalskalierten und/oder ordinalskalierten Variablen kann nicht über die Analyse der gemeinsamen Variation in den beobachteten Ausprägungen erfolgen, da für solcherart skalierte Variablen eine Variation entweder nicht meßbar ist (nominalskalierte Variablen) bzw. Abstände nicht sinnvoll interpretiert werden können (ordinalskalierte Variablen). Letzteres trifft auch für metrisch skalierte Variablen zu, deren Variablenwerte klassiert wurden. Die Klassenbildung beinhaltet eine Transformation von der metrischen Skala in die Ordinalskala, da jeder konkrete Variablenwert durch die Klassennummer ersetzt wird. Damit ist im allgemeinen auch eine Kategorisierung verbunden, z.B. von den konkreten Einkommenswerten der einzelnen Personen in die Kategorien niedriges, mittleres oder hohes Einkommen. Abstände zwischen Klassennummern sind deshalb nicht sinnvoll interpretierbar, denn ein Variablenwert z.B. der vierten Klasse muß nicht doppelt so groß sein wie ein Variablenwert der zweiten Klasse.

Bei den zuletzt genannten Arten von Variablen wird dieselbe Variablenausprägung relativ häufig beobachtet (z.B. die Ausprägungen männlich bzw. weiblich der Variablen Geschlecht). Ein beobachtetes Wertetupel $(x_{i1}, x_{i2}, \dots)$ der Variablen $X_1, X_2, \dots$ tritt deshalb in der Stichprobe im allgemeinen mit einer absoluten Häufigkeit größer als Eins auf. Bei der statistischen Analyse von Zusammenhängen dieser Variablen wird man sich somit auf die gemeinsame Häufigkeitsverteilung konzentrieren, indem die beobachteten Häufigkeiten mit denen bei Unabhängigkeit zu erwartenden Häufigkeiten verglichen werden.

An dieser Stelle sei angemerkt, dass das Skalenniveau statistischer Variablen nicht identisch ist mit dem Typ von Variablen unter SPSS. Nominalskalierte Variablen, deren Ausprägungen mittels Schlüsselnummern kodiert wurden, und ordinalskalierte Variablen, deren Ausprägungen

Rangzahlen zugeordnet wurden, werden unter SPSS als numerische Variablen behandelt, auf die dann fälschlicherweise statistische Verfahren angewandt werden könnten, die nur für metrisch skalierte Variablen zulässig sind.
Im weiteren wird davon ausgegangen, dass aufgrund einer fachwissenschaftlichen Problemstellung eine Zufallsstichprobe vom Umfang n aus einer gegebenen Grundgesamtheit gezogen wurde, wobei gleichzeitig mehrere Variablen erfaßt wurden, d.h., es liegen Beobachtungen von mehreren Variablen an n statistischen Einheiten (unter SPSS als Fälle bezeichnet) vor.
Letztendliches Untersuchungsziel ist die Beantwortung der Frage:
Läßt sich für die Grundgesamtheit, aus der die Stichprobe gezogen wurde, ein Zusammenhang zwischen den Variablen bzw. eine Abhängigkeit einer Variablen von anderen Variablen statistisch nachweisen. Dies führt auf einen statistischen Test von entsprechend formulierten Hypothesen hinaus.
Die zur Testdurchführung auszuwählenden Verfahren hängen neben dem oben erwähnten Skalenniveau der Variablen auch von der Anzahl der gleichzeitig einzubeziehenden Variablen ab.
Vor der Auswahl eines solchen Verfahrens wird man sich jedoch im allgemeinen zuerst einmal die Gegebenheiten in der Stichprobe anschauen, was mittels explorativer statistischer Werkzeuge und/oder deskriptiver statistischer Methoden erfolgen kann.

2.1. Explorative Zusammenhangsanalyse

Der Scatterplot (Streuungsdiagramm), die Scatterplot-Matrix (Draftsman-Display) und der 3D-Scatterplot sind explorative Werkzeuge[1] zur visuellen Veranschaulichung von Beziehungen zwischen Variablen, die jedoch nur sinnvoll bei metrisch skalierten Variablen einsetzbar sind.

Beim Scatterplot handelt es sich um die graphische Darstellung der Beobachtungswerte zweier metrisch skalierter Variablen in einem kartesischen Koordinatensystem. Jedes Paar (x_i, y_i) von Beobachtungswerten $(i = 1, \ldots, n)$ erscheint als Punkt in der Variablenebene.

Die Scatterplotmatrix ist ein graphisches Verfahren zur Veranschaulichung von paarweisen Zusammenhängen zwischen mehr als zwei Variablen. Liegen für p metrisch skalierte Variablen $X_1, \ldots, X_p$ $(p > 2)$ Beobachtungen an n statistischen Einheiten vor, so werden für jeweils zwei Variable X_j und X_k $(j \neq k, j, k = 1, \ldots, p)$ die Beobachtungswerte in einem Scatterplot dargestellt, womit sich insgesamt $p \cdot (p-1)$ Plots ergeben. Diese werden in Form einer Matrix angeordnet, wobei die Hauptdiagonale leere Flächen enthält, da die Variable X_j nicht gegen sich selbst abgetragen wird. Die Scatterplots unterhalb der Diagonale sind die Spiegelbilder der Scatterplots oberhalb der Diagonale.

Beim 3D-Scatterplot kann eine augenscheinliche Prüfung auf Beziehungen zwischen drei Variablen X, Y und Z durch Abtragung der Beobachtungen (x_i, y_i, z_i), $i = 1 \ldots, n$, als Punkte im dreidimensionalen Raum vorgenommen werden.

Diese Plots können unter SPSS über

■ Graphs

■ Scatter...

(siehe Abb. 2.1) aufgerufen werden.

Im sich öffnenden Dialogfeld „Scatterplot“ (Abb. 2.2) wird die gewünschte Plotart durch Anklicken gewählt und die Schaltfläche „Define“ (Definieren) betätigt.

[1] Diese explorativen Werkzeuge wurden bereits in der „Computergestützten Statistik I“ (siehe Rönz, B. (2001), Abschnitt 4.3) behandelt und dienten der Identifizierung von potentiellen Ausreißern. Da an dieser Stelle die Zielstellung der Untersuchung eine andere ist, sollen sie wiederholend behandelt werden, zumal einige neue Aspekte hinzukommen

Abbildung 2.1.: Menü „Graphs“

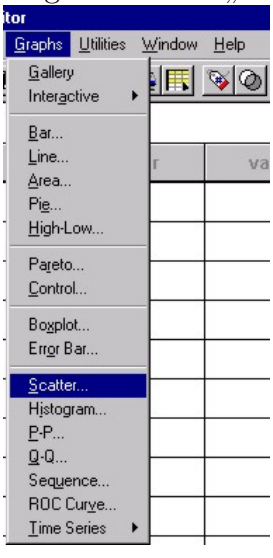
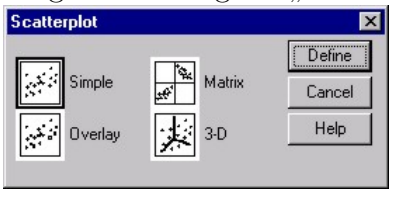
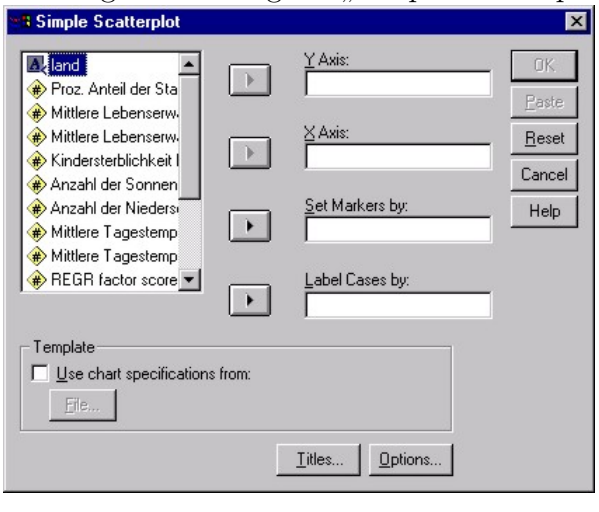


Abbildung 2.2.: Dialogfeld „Scatterplot“



Wurde ein **einfacher Scatterplot** (Simple) gewählt, öffnet sich das nachstehende Dialogfeld.

Abbildung 2.3.: Dialogfeld „Simple Scatterplot“



In diesem Dialogfeld werden die beiden Variablen, für die ein Plot erstellt werden soll, in die Felder „Y Axis:“ bzw. „X Axis:“ gebracht, wobei der Nutzer entscheidet, welche Variable auf welcher Achse erscheinen soll.

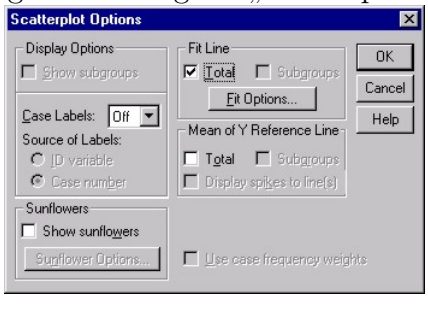
Eine Fallbeschriftung (Label Cases by:) ist möglich, sollte jedoch nur bei einem Plot mit wenigen Punkten verwendet werden, da sonst die Anschaulichkeit erheblich leidet. Desweiteren kann eine Gruppenvariable (Set Markers by:) ausgewählt werden. Die einzelnen Datenpunkte im Scatterplot erscheinen dann gemäß der Ausprägungen dieser Variablen in unterschiedlichen Farben oder mit unterschiedlichen Markierungssymbolen. Wegen der Übersichtlichkeit empfiehlt sich nur eine Gruppenvariable mit zwei oder drei Ausprägungen.

Nach Klick auf „OK“ erscheint im Output-Fenster (SPSS for Windows Viewer) der Scatterplot. Die Grafik kann editiert werden, indem mittels eines Doppelklicks auf die Grafik diese in das Grafikfenster (SPSS for Windows Chart Editor) gebracht wird. Dort stehen über die Menüs „Chart“ und „Format“ bzw. über Icons vielfältige Möglichkeiten für Veränderungen bzw. Ergänzungen des Plots zur Verfügung. Von diesen Möglichkeiten soll im Kontext der hier diskutierten Problemstellung nur eine gezeigt werden. Über

- Graphs
  - Options...

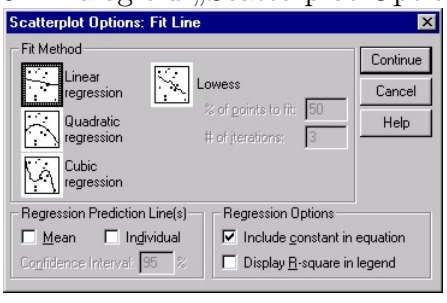
kann in dem Dialogfeld „Scatterplot Options“ (Abb. 2.4) eine Kurvenanpassung (Fit Line) ausgewählt werden. Eine solche Kurvenanpassung ist hilfreich, um Tendenzen in der gemeinsamen Variation der beiden Variablen sichtbar zu machen.

Abbildung 2.4.: Dialogfeld „Scatterplot Options“



Dabei kann die Kurvenanpassung für alle Daten (Total) und/oder für Untergruppen (Subgroups), wenn eine Gruppenvariable ausgewählt wurde, erfolgen. Nach der Betätigung der Schaltfläche „Fit Options...“ kann in dem Dialogfeld „Scatterplot Options: Fit Line“ (Abb. 2.5) unter vier Anpassungslinien gewählt werden.

Abbildung 2.5.: Dialogfeld „Scatterplot Options: Fit Line“



Bei linearer (linear), quadratischer (quadratic) bzw. kubischer (cubic) Regression wird jeweils diejenige Regressionskurve des gewählten Funktionstyps nach der Methode der kleinsten Quadrate bestimmt, die sich den Datenpunkten am besten anpaßt (siehe auch Abschnitt 2.3).

Bei diesen Kurvenanpassungen kann

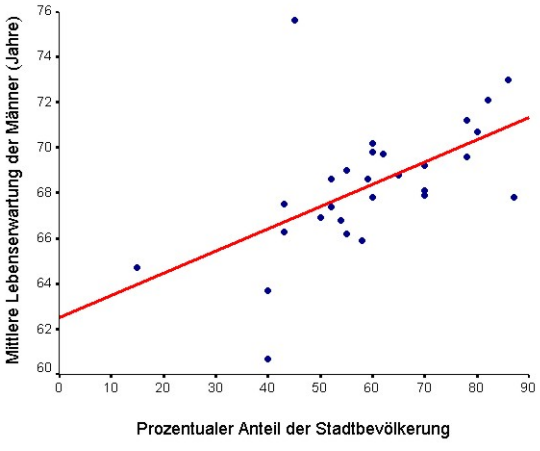
- in dem Feld „Regression Options“ angegeben werden, ob die anzupassende Regressionsfunktion unter Einschluß oder Ausschluß einer Konstanten geschätzt werden soll (Include constant in equation). Desweiteren hat man die Option der Ausgabe des Bestimmtheitsmaßes R^2 (Display R-square in legend).

- in dem Feld „Regression Prediction Line(s)“ die Ausgabe von Konfidenzintervallen für die vorhergesagten Werte (bei „Mean“ anklicken) und/oder für die beobachteten Werte der abhängigen Variablen (bei „Individual“ anklicken) angefordert werden, wobei das Konfidenzniveau frei gewählt werden kann (ein Wert zwischen 10 und 99,9%).

• Beispiel 2.1

Es soll überprüft werden, ob ein Zusammenhang zwischen mittlerer Lebenserwartung der Männer und dem prozentualen Anteil der Stadtbevölkerung besteht. Dazu stehen Daten aus 28 europäischen Ländern zur Verfügung, die der Datei *europa.sav* auf der bei Bühl, A., Zöfel, P. (1994) beiliegenden Diskette entnommen wurden. In diesem Fall wird der prozentuale Anteil der Stadtbevölkerung für die Abszisse und die mittlere Lebenserwartung für die Ordinate ausgewählt. Der resultierende Scatterplot, in dem eine lineare Kurvenanpassung erfolgte, ist in der Abb. 2.6 enthalten. Es ist eine positive Abhängigkeit der mittleren Lebenserwartung der Männer vom prozentualen Anteil der Stadtbevölkerung zu erkennen, da mit steigenden Werten des prozentualen Anteils der Stadtbevölkerung tendenziell auch die mittlere Lebenserwartung der Männer ansteigt. Diese Abhängigkeit läßt sich mittels einer linearen Funktion relativ gut approximieren.

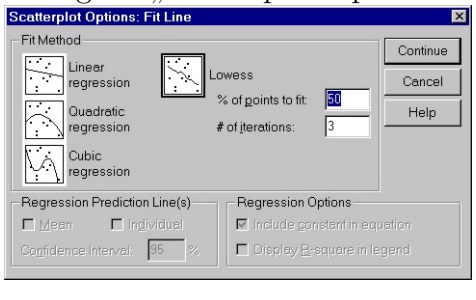
Abbildung 2.6.: Beispiel eines einfachen Scatterplots mit linearer Kurvenanpassung



Eine weitere Möglichkeit der Kurvenanpassung ist Lowess (locally weighted regression scatter plot smoothing), wobei kein spezieller Funktionstyp vorgegeben wird. Dieses Verfahren (siehe Chambers, J.M., Cleveland, W.S. (1985), S. 167 ff.) soll kurz skizziert werden. Ein ausführliches Beispiel ist im Anhang A enthalten.

Mit Lowess erfolgt eine lokal gewichtete Regressionsglättung (smoothing) mittels einer iterativen, gewichteten Methode der kleinsten Quadrate. Es wird eine nach der Größe der x_i-Werte geordnete Reihe der Beobachtungsdaten unterstellt. Für die Vorhersage des Wertes der abhängigen Variablen Y an der Stelle x_i $(i = 1, \ldots, n)$ wird nur ein bestimmter Prozentsatz der gegebenen Punkte einbezogen, wobei der Punkt (x_i, y_i) im Zentrum des ausgewählten Bereichs liegen soll. Dazu muß der Prozentsatz der einzubeziehenden Punkte (% of points to fit; Voreinstellung: 50) vorgegeben werden.

Abbildung 2.7.: Dialogfeld „Scatterplot Options: Fit Line“: Lowess



Je höher der Prozentsatz der Punkte gewählt wird, umso glatter wird die letztendlich angepaßte Kurve. Im allgemeinen sollte ein Wert zwischen 33 und 67% gewählt werden. Aufgrund des vorgegebenen Anteils f wird die Anzahl der einzubeziehenden Punkte ermittelt: $K = f \cdot n$, gerundet zur ganzen Zahl. Die Grenzen des betrachteten X-Bereiches werden nun so gelegt, dass der Abstand von x_i zur Bereichsgrenze gleich dem Abstand von x_i zum K-ten nächsten

Nachbarpunkt ist, wobei x_i als ein Nachbar zu sich selbst gezählt wird. Für jedes x_i sind somit die Bereiche verschieden groß, jedoch so, dass x_i im Zentrum liegt und K Punkte im Bereich liegen.

Der Vorhersagewert $\hat{y}_i$ für die abhängige Variable an der Stelle x_i wird berechnet, indem die einbezogenen Werte x_k der erklärenden Variablen $(k = 1, \ldots, K)$ derart gewichtet werden, dass die nahe bei dem ausgewählten x_i liegenden Werte eine größere Gewichtung als die weiter entfernt liegenden Werte erhalten. Als Funktion zur Vergabe dieser „Nachbarschafts"-Gewichte wird

$$W(x_k) = \left(1 - \left|\frac{x_i - x_k}{d_i}\right|^3\right)^3 \tag{2.1}$$

mit d_i als Abstand von x_i zum K-ten nächsten Nachbarpunkt verwendet. Diese Gewichtsfunktion ist symmetrisch, erreicht den größten Wert an der Stelle x_i und den Wert Null jeweils an den Bereichsgrenzen.

Unter Verwendung der einbezogenen Punkte (x_k, y_k), k = 1,. . . ,K, wird nunmehr eine mit $w(x_k)$ gewichtete lineare Regressionsfunktion geschätzt, die

$$\sum_k w(x_k)(y_k - b_0 - b_1 x_k)^2 \tag{2.2}$$

minimiert. Die Regressionsfunktion $\hat{y}_k = b_0 + b_1 x_k$ zeigt die mittlere lineare Abhängigkeit der Variablen Y von der Varibalen X in diesem ausgewählten Bereich von X-Werten. Der Regreßwert $\hat{y}_i$ wird als der (in der 1. Stufe) geglättete Wert der Variablen Y an der Stelle x_i verwendet. Diese Prozedur wird für alle $i = 1, \ldots, n$ durchgeführt. Man erhält im Ergebnis für jedes i einen geglätteten y-Wert.

Wie leicht einzusehen ist, werden die Regressionsfunktionen in den ausgewählten Bereichen stark durch potentielle Ausreißer beeinflußt. Um gegen Ausreißer robustere geglättete y-Werte zu erhalten, werden in der 2. Stufe die Residuen $\hat{u}_i = y_i - \hat{y}_i$ $(i = 1, \ldots, n)$, der Median aus den Residuenwerten m = median $|\hat{u}_i|$ sowie das Verhältnis $e_i = \hat{u}_i/6m$ berechnet. Wenn die Residuen näherungsweise normalverteilt sind, wird mit m approximativ $2\sigma/3$ und somit mit 6m etwa 4σ geschätzt, wobei σ die Standardabweichung der Grundgesamtheit ist. Für die Residuen werden anschließend „Robustheits"-Gewichte nach

$$G(x_i) = \begin{cases} (1 - e_i^2)^2 & \text{für } |e| < 1 \\ 0 & \text{sonst} \end{cases} \tag{2.3}$$

ermittelt. Je kleiner ein Residuum im Vergleich zu $6m$ ist, desto größer ist sein „Robustheits"-Gewicht; für Residuen, die größer als oder gleich $6m$ sind, ist das „Robustheits"-Gewicht Null. Diese „Robustheits"-Gewichte, multipliziert mit den „Nachbarschafts"-Gewichten, werden für

eine erneute Anpassung einer Regressionslinie innerhalb der einzelnen Bereiche verwendet. Dadurch erhält man eine Reihe neuer geglätteter Werte $\hat{y}_i$.

Der Schritt der Bestimmung von Robustheitsgewichten kann mehrmals wiederholt werden. Je höher die vorgegebene Anzahl der Iterationen (# of iterations, siehe Abb. 2.7) ist, desto genauer ist die Anpassung (wobei jedoch die weitere Erhöhung einer bereits hohen Iterationsanzahl kaum eine sichtbare Verbesserung bringt).

• Beispiel 2.1 (Fortsetzung):

Im Scatterplot der Variablen prozentualer Anteil der Stadtbevölkerung und mittlere Lebenserwartung der Männer erfolgt nun eine Kurvenanpassung nach Lowess. Dabei wird als Prozentsatz der einzubeziehenden Punkte zunächst 35% und danach 65% gewählt. In beiden Fällen wird die Voreinstellung der Anzahl der Iterationen von 3 beibehalten.

Abbildung 2.8.: Einfacher Scatterplot mit Lowess Kurvenanpassung, % of points to fit: 35

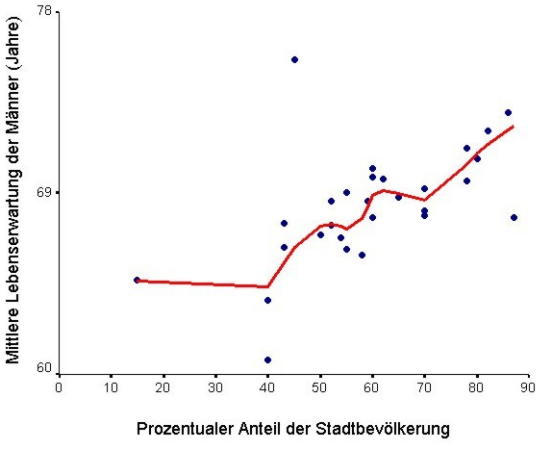
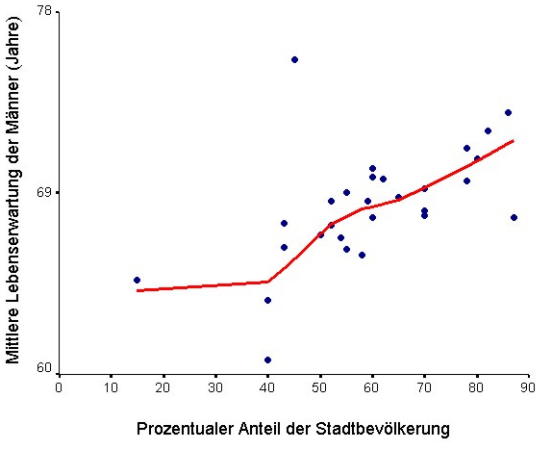


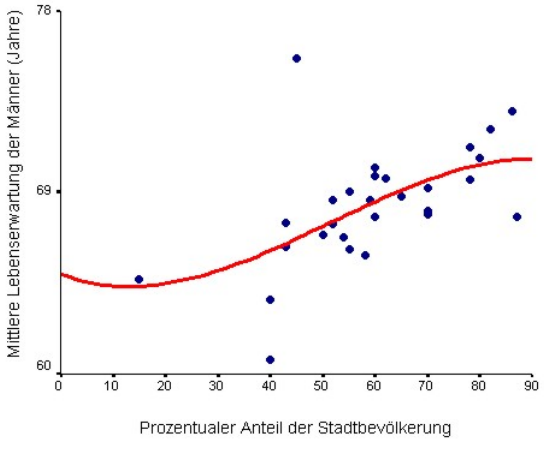
Abbildung 2.9.: Einfacher Scatterplot mit Lowess Kurvenanpassung, % of points to fit: 65



Durch die Erhöhung des Prozentsatzes der einbezogenen Punkte wird die angepaßte Kurve we-

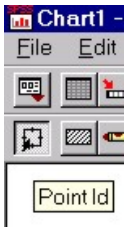
sentlich glatter. Aber selbst bei dem realtiv hohen Prozentsatz von 65% ist der Einfluß zweier „extremer“ Beobachtungspunkte auf die Kurvenanpassung deutlich erkennbar. In diesem Fall führt eine kubische Regressionsfunktion zu einer geringfügigen Verbesserung im Vergleich zur linearen Anpassung (vgl. Abb. 2.6).

Abbildung 2.10.: Einfacher Scatterplot mit kubischer Kurvenanpassung



Dieses Beispiel zeigt, dass es oftmals von Interesse ist, spezielle Punkte in einem Scatterplot zu identifizieren. SPSS ermöglicht diese Identifizierung im Chart Editor durch das Point Selection Ikon.

Abbildung 2.11.: Point Selection Ikon



Ein Klick auf dieses Ikon aktiviert den Point Selection Mode, wodurch der Cursor sein Aussehen zu einem „Fadenkreuz“ verändert. Mit diesem Cursor geht man auf den interessierenden Punkt im Scatterplot und klickt ihn an. Die Identifizierung kann in verschiedener Weise erfolgen:

- Wurde bei der Erstellung des Scatterplots keine Variable für „Label Cases by“ (siehe Abb. 2.3) angegeben, dann erscheint die zu diesem Punkt gehörige Fallnummer.

- Wurde eine Variable für „Label Cases by“ spezifiziert, erfolgt die Identifizierung mit der zu diesem Punkt gehörenden Ausprägung der Variablen.

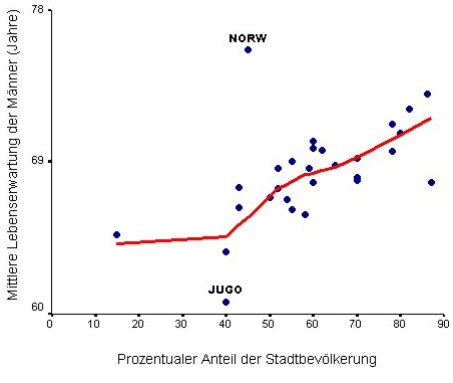
Wenn der Datenfile, aus dem heraus der Scatterplot erstellt wurde, noch geöffnet ist, wird der Fall markiert und kann durch einen Wechsel zum SPSS Data Editor inspiziert werden.

Ein nochmaliger Klick auf den Punkt entfernt die Punkt-Identifikation und ein nochmaliger Klick auf das Point Selection Ikon deaktiviert den Selection Mode.

• Beispiel 2.1 (Fortsetzung):

Bei der Erstellung des Scatterplots der Variablen prozentualer Anteil der Stadtbevölkerung und mittlere Lebenserwartung der Männer wird die Variable Land für „Label Cases by“ ausgewählt. Die Punkt-Identifizierung zeigt an, dass die beiden „extremen Punkte“ zu den Ländern Norwegen und Jugoslawien gehören.

Abbildung 2.12.: Einfacher Scatterplot mit Punkt-Identifizierung



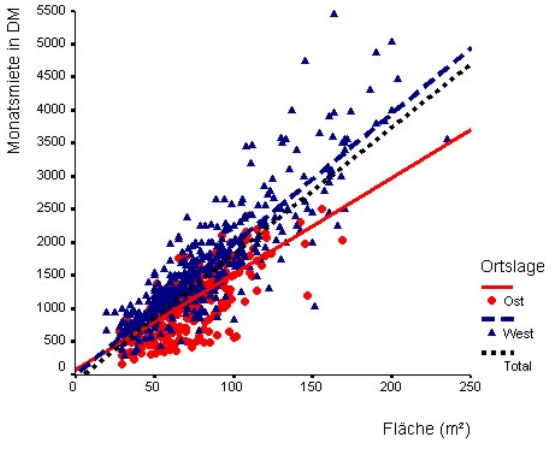
• Beispiel 2.2:

Dieses Beispiel dient dem Zweck, die Verwendung einer Gruppenvariablen bei der Erstellung eines Scatterplots zu demonstrieren.

Es steht zweifelsfrei fest, dass es einen Zusammenhang zwischen der Fläche (in m^2) und der Höhe der Monatsmiete (in DM) gibt, der jedoch nicht funktional ist, da eine Reihe anderer Faktoren ebenfalls die Monatsmiete beeinflussen. Es ist also sinnvoll, die Beziehung zwischen diesen beiden Variablen über einen Scatterplot visuell zu veranschaulichen, bevor ein statistisches Modell zur Schätzung der Abhängigket formuliert wird. Als Datenbasis für dieses Beispiel dient die Datei mieten.sav[2], in der diese Variablen für 815 Berliner Mietwohnungen enthalten sind. Als Gruppenvariable für „Set Markers by“ wird die Variable Ortslage mit den Ausprägungen Ost und West verwendet. Es wird eine lineare Kurvenanpassung gewählt, die sowohl für die Gruppen (Subgroups) als auch für alle Punkte (Total) vorgenommen wird.

[2]Die Datei mieten.sav wurde von Herrn Prof. Dr. P. P. Eckstein, Fachhochschule für Technik und Wirtschaft Berlin, im Internet zur Verfügung gestellt und ist beschrieben in: Eckstein, P. (1997), S. 43 f. (http://www.f3.fhtw-berlin.de/Professoren/Eckstein/download.html).

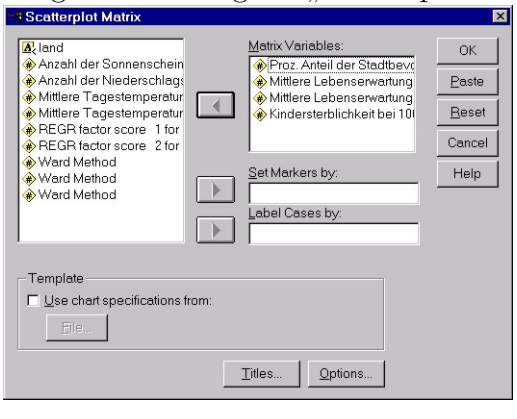
Abbildung 2.13.: Einfacher Scatterplot mit linearer Kurvenanpassung und Gruppenvariable



Anhand des Scatterplots ist zu erkennen, dass im Mittel der Zuwachs der Monatsmiete pro Erhöhung der Fläche um einen m^2 im Westteil größer ist als im Ostteil. Darüber hinaus erscheint die Anpassung der linearen Regressionsfunktion für den Ostteil schlechter zu sein als für den Westteil, da die Verteilung der Werte der Monatsmiete über einen kleineren Bereich von Werten der Fläche erfolgt (nicht sehr lang gestreckte Punktwolke für die Daten des Ostteils).

Wurde eine **Scatterplot-Matrix** gewählt, öffnet sich das nachstehende Dialogfeld.

Abbildung 2.14.: Dialogfeld „Scatterplot-Matrix“

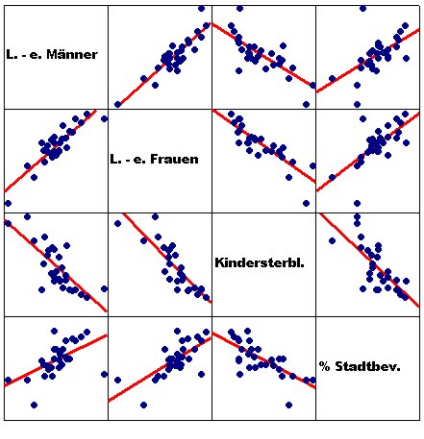


In diesem Dialogfeld werden alle Variablen, für die der paarweise Zusammenhang in jeweils einem Scatterplot überprüft werden soll, in das Feld „Matrix-Variables:“ gebracht. Allerdings sollten nicht zu viele Variablen ausgewählt werden, weil sonst jedes Scatterplot zu klein und der mögliche Zusammenhang nicht mehr erkennbar wird. Eine Fallbeschriftung ist wiederum möglich, sollte jedoch wegen der Übersichtlichkeit weggelassen werden. Ebenso kann wieder eine

Gruppenvariable ausgewählt werden. Eine Kurvenanpassung ist wie beim Scatterplot möglich, jedoch kann für alle in der Matrix enthaltenen Plots nur derselbe Typ der Anpassung gewählt werden. Gleichfalls können wieder einzelne Punkte identifiziert werden. Das ist insofern interessant, um festzustellen, ob ein „extremer“ Beobachtungspunkt für den einen Zusammenhang auch ein „extremer“ Beobachtungspunkt für den Zusammenhang zweier anderer Variablen ist.

• Beispiel 2.3:
Es soll überprüft werden, ob ein Zusammenhang zwischen mittlerer Lebenserwartung der Männer bzw. mittlerer Lebenserwartung der Frauen bzw. Kindersterblichkeit je 1000 Geburten und dem prozentualen Anteil der Stadtbevölkerung besteht. Dazu wird wieder die Datei europa.sav des Beispiels 2.1 verwendet. Diese vier Variablen werden paarweise in einer Scatterplot-Matrix veranschaulicht.

Abbildung 2.15.: Beispiel einer Scatterplot-Matrix mit linearer Kurvenanpassung



Der bereits im Beispiel 2.1 angegebene Scatterplot der beiden Variablen mittlere Lebenserwartung der Männer und prozentualer Anteil der Stadtbevölkerung (Abb. 2.6) ist im Feld (1,4) der Scatterplot-Matrix enthalten. Das Feld (4,1) beinhaltet ebenfalls diesen Zusammenhang, jedoch wurde die mittlere Lebenserwartung der Männer auf der Abszisse und der prozentuale Anteil der Stadtbevölkerung auf der Ordinate abgetragen. Zubeachten ist deshalb, dass

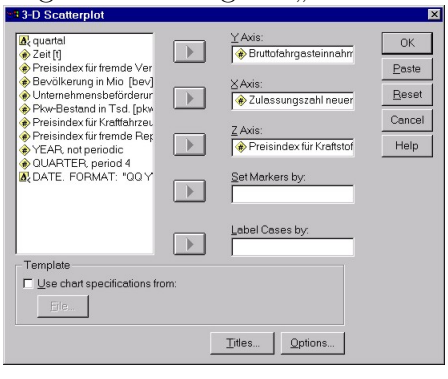
- die lineare Anpassung im Feld (1,4) die mittlere Lebenserwartung der Männer (Y) in Abhängigkeit vom prozentualen Anteil der Stadtbevölkerung (X) beinhaltet,

- die lineare Anpassung im Feld (4,1) den prozentualen Anteil der Stadtbevölkerung (Y) in Abhängigkeit von der mittleren Lebenserwartung der Männer (X) zeigt (was sachlogisch keinen Sinn macht).

Wurde der **3D-Scatterplot** ausgewählt, so sind weitere Festlegungen der Gestaltung des Plots in dem in der Abb. 2.16 enthaltenen Dialogfeld vorzunehmen.

Es sind drei Variablen aus der linken Quelliste auszuwählen und den Achsen zuzuordnen. Wie vorher sind Fallbeschriftung bzw. Gruppierung nach einer weiteren Variablen möglich. Ebenso kann in dem erzeugten 3-D Scatterplot eine Punkt-Identifizierung vorgenommen werden.

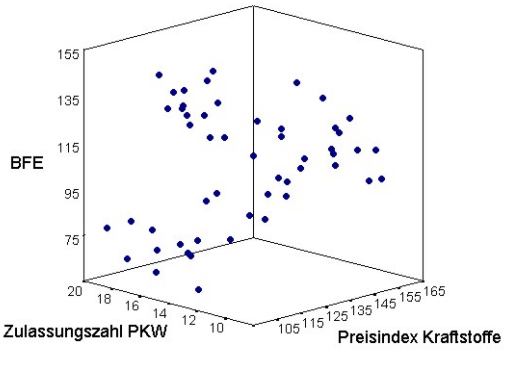
Abbildung 2.16.: Dialogfeld „3-D Scatterplot“



• Beispiel 2.4:

Es soll explorativ untersucht werden, ob ein Zusammmenhang zwischen den Bruttofahrgasteinnahmen (BFE, in Mio. DM) eines regionalen Verkehrsunternehmens, der Zulassungszahl neuer PKW (in 1000) in dieser Region und dem Preisindex für Kraftstoffe existiert. Dazu wurden Beobachtungen dieser drei Variablen über 52 Quartale erhoben, die in der Datei verkehr.sav enthalten sind. Den resultierenden Scatterplot zeigt Abb. 2.17.

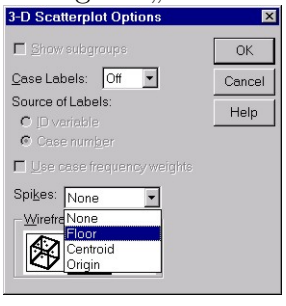
Abbildung 2.17.: Beispiel eines 3-D Scatterplots



Da es bei der Betrachtung eines 3-D Scatterplots oft schwierig ist, sich die Lage der Punkte im dreidimensionalen Raum vorzustellen, gibt es die Möglichkeit von Projektionen. Dazu muß

sich der 3-D Scatterplot im Chart Editor befinden. Durch Anwahl von „Chart“ in der Menü-Leiste des Chart Editors öffnet sich ein Pull-Down-Menü, in dem „Options . . . “ auszuwählen ist. Es kann auch das Ikon „Chart options“ angewählt werden. Es öffnet sich das in Abb. 2.18 enthaltene Dialogfeld.

Abbildung 2.18.: Dialogfeld „3-D Scatterplot: Options“

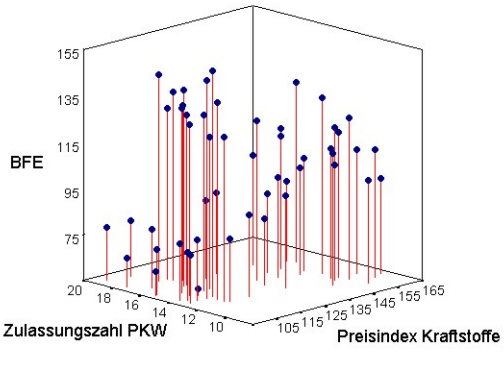


Bei „Spikes:“ kann unter drei verschiedenen Projektionsmöglichkeiten gewählt werden:

- Floor (Parallel): Es erfolgt eine Projektion auf die X-Z-Ebene.
- Centroid (Zentroid): Es erfolgt eine Projektion zum Mittelpunkt der Punktwolke.
- Origin (Ursprung): Es erfolgt eine Projektion zum Koordinatenursprung.

Die folgende Abbildung zeigt den 3-D Scatterplot der Abbildung 2.17 mit der Projektion Floor für die Variablen des Beispiels 2.4.

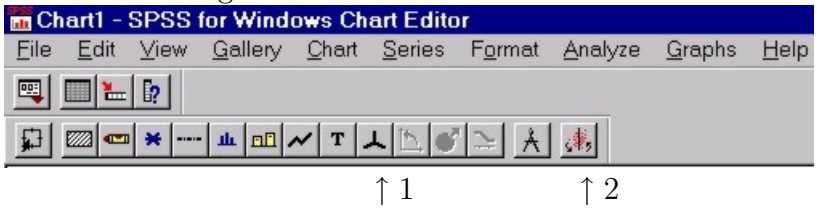
Abbildung 2.19.: Beispiel eines 3-D Scatterplots mit Floor-Projektion



Die 3-D Darstellung bietet weiterhin die Möglichkeit der Rotation der Punktwolke über verschiedene Achsen. Auf diese Weise ist es möglich, die Punktwolke aus einem anderen Blickwinkel zu betrachten, wobei möglicherweise Zusammenhänge deutlicher sichtbar werden bzw. aus der

Punktwolke herausfallende Punkte besser erkennbar sind. Dazu muß sich der 3-D Scatterplot ebenfalls im Chart Editor befinden.

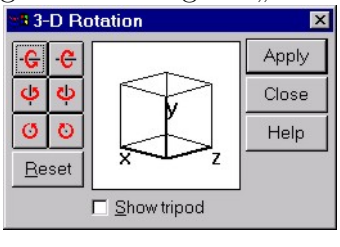
Abbildung 2.20.: Menüleiste des Chart Editors



Für die Rotation bieten sich zwei Möglichkeiten:

1. Eine Möglichkeit der Rotation wird über das Ikon „3-D Rotation“ (siehe mit 1 gekennzeichnetes Ikon in Abb. 2.20) bzw. durch Anwahl von „Format“ in der Menü-Leiste des Chart Editors und anschließender Auswahl von „3-D Rotation“ in dem Pull-Down-Menü angeboten. Es öffnet sich das Dialogfeld der Abb. 2.21.

   Abbildung 2.21.: Dialogfeld „3-D Rotation“

   Über die Schaltflächen auf der linken Seite kann man bei gedrückter Maustaste das Koordinatensystem in die gewünschte(n) Richtung(en) drehen. Nach Betätigung von „Apply“ (Zuweisen) wird die Rotation ausgeführt. Der Nachteil dieser Rotationsmöglichkeit ist, dass der gedrehte 3-D Scatterplot erst nach dem Zuweisen sichtbar wird. Über die Schaltfläche „Reset“ kann die Rotation wieder rückgängig gemacht werden.

2. Die andere Möglichkeit der Rotation ist über das Ikon „Set/exit spin mode“ (siehe mit 2 gekennzeichnetes Ikon in Abb. 2.20) bzw. durch Anwahl von „Format“ in der Menü-Leiste des Chart Editors und anschließender Auswahl von „Spin Mode“ in dem Pull-Down-Menü gegeben, wodurch Ikons in der Menü-Leiste erscheinen, die das Drehen des 3-D Scatterplots in verschiedenen Richtungen erlauben.

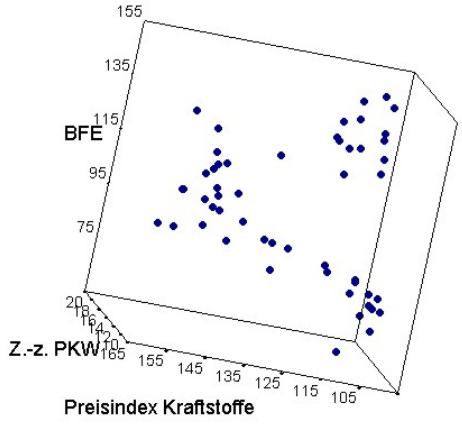
   Abbildung 2.22.: Menüleiste des Spinmode

Wählt man eine dieser Schaltflächen an und hält die linke Maustaste gedrückt, so wird die Drehung sofort ausgeführt, wodurch eine unmittelbare Verfolgung der Veränderung der Punktwolke wesentlich erleichtert wird. Bei nochmaliger Betätigung des Ikon „Set/exit spin mode“ verbleibt der 3-D Scatterplot in der letzten Rotationsposition. Wählt man das Ikon „Reset“, so wird der 3-D Scatterplot in die Ausgangsposition gebracht, wobei der Spin Mode weiter aktiv bleibt. Wählt man das Ikon „Cancel“, wird der 3-D Scatterplot in die Ausgangsposition zurückgesetzt und der Spin Mode deaktiviert.

Abb. 2.23 zeigt den 3-D Scatterplot der Abb. 2.17 in rotierter Form.

Abbildung 2.23.: Beispiel eines rotierten 3-D Scatterplots



Weitere Einblicke in eine 3-D Punktwolke bieten 3-D Spektral-Plots, 3-D Space-Plots sowie 3-D Surface-Plots (Oberflächen-Plots) mit wählbaren Anpassungsfunktionen, die jedoch nicht unter SPSS verfügbar sind. Eine Statistik-Software, die derartige Grafiken liefert, ist u.a. STATISTICA von StatSoft, Inc., Release 6.0 (http://www.statsoft.com). Die Abbildungen 2.24 bis 2.26 zeigen Beispiele der genannten Plots.

Abbildung 2.24.: Beispiel eines 3-D Spektral-Plots

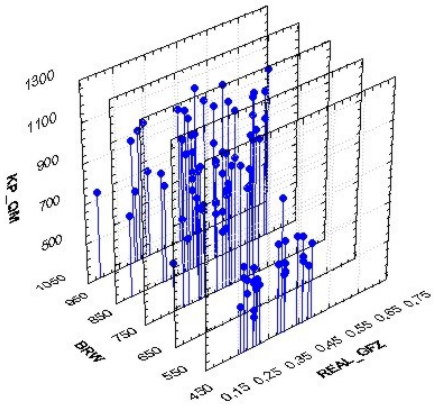


Abbildung 2.25.: Beispiel eines 3-D Space-Plots

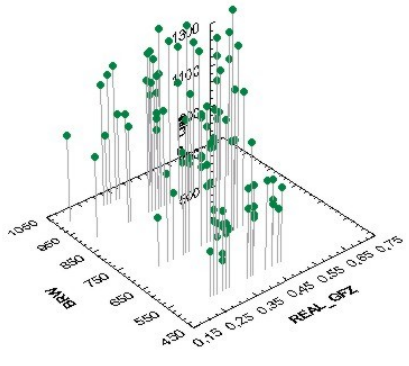
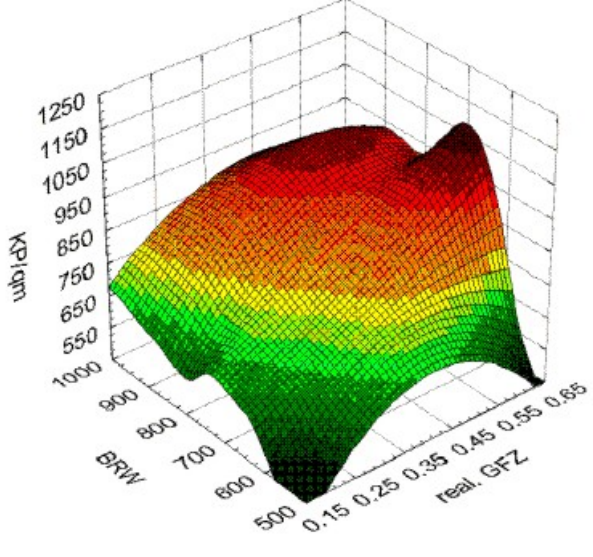
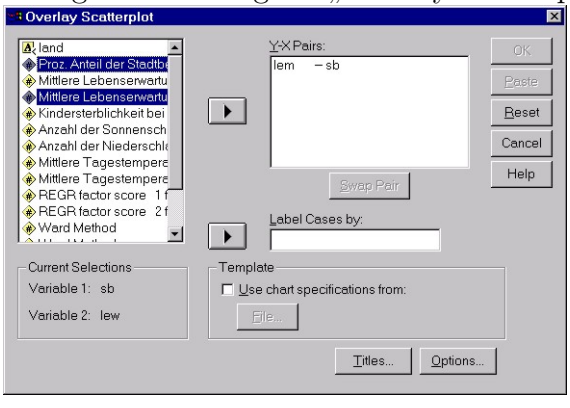


Abbildung 2.26.: Beispiel eines 3-D Surfaceplots-Plots mittels Distance Weighted Least Squares



Eine vierte unter SPSS angebotene Art von Scatterplots (siehe Abb. 2.2) ist der **Overlay Scatterplot** (überlagerter Scatterplot). Dieser ermöglicht, mehrere bivariate Zusammenhänge in einer Grafik darzustellen. Die Vereinbarungen zum Aufbau des Plots werden in dem Dialogfeld „Overlay Scatterplot“ getroffen.

Abbildung 2.27.: Dialogfeld „Overlay Scatterplot“



Es müssen mindestens zwei Paare von Variablen spezifiziert werden. Zur Auswahl eines Paares werden die jeweiligen Variablen nacheinander angeklickt. In dem Feld „Current Selections“ (Auswahl) erscheint dann bei Variable 1 und Variable 2 der Variablenname. Anschließend ist die Schaltfläche ▷ vor dem Feld „Y-X Pairs:“ zu betätigen. Eine bereits ausgewählte Variable kann nochmals in einem anderen Variablenpaar erscheinen. Weist ein ausgewähltes Variablenpaar in dem Feld „Y-X Pairs:“ nicht die gewünschte Reihenfolge auf, wird es durch Anklicken mit der Maus markiert und anschließend die Schaltfläche „Swap Pair“ (Paar vertauschen) betätigt. Eine Kurvenanpassung in der oben bschriebenen Weise kann für jedes Variablenpaar in dem erzeugten Scatterplot vorgenommen werden, jedoch nur für alle Variablenpaare mit dem gleichen Typ der Anpassung. Ebenso ist wiederum eine Identifizierung von Punkten möglich.

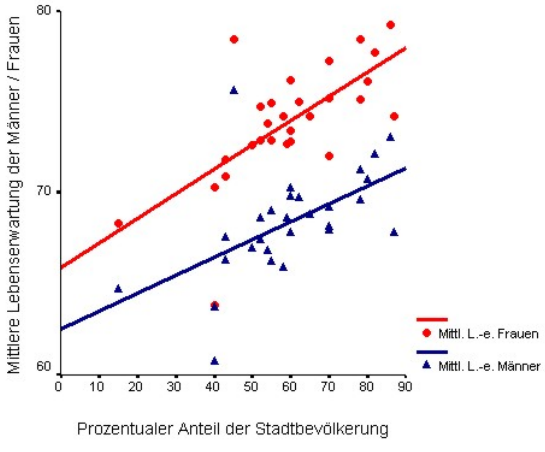
Ein Overlay Scatterplot ist im allgemeinen nur sinnvoll, wenn Paare von Variablen mit gleichen Maßeinheiten ausgewählt werden, um eine gleiche Maßeinteilung der Achsen zu garantieren.

• Beispiel 2.5:

Ausgehend von der im Beispiel 2.1 verwendeten Datei europa.sav soll festgestellt werden, ob zwischen der mittleren Lebenserwartung der Männer bzw. der mittleren Lebenserwartung der Frauen einerseits und dem prozentualen Anteil der Stadtbevölkerung andererseits ein Zusammenhang besteht und inwieweit er sich möglicherweise zwischen Männern und Frauen unterscheidet. In diesem Fall wird der prozentuale Anteil der Stadtbevölkerung für die Abszisse und die mittlere Lebenserwartung der Männer bzw. der Frauen für die Ordinate ausgewählt. Der resultierende Overlay Scatterplot, in dem eine lineare Kurvenanpassung erfolgte, ist in

der Abb. 2.28 enthalten. Sowohl für die Männer als auch für die Frauen ist eine relativ stark ausgeprägte positive Abhängigkeit der mittleren Lebenserwartung vom prozentualen Anteil der Stadtbevölkerung zu erkennen. Diese Abhängigkeit läßt sich mittels einer linearen Funktion gut anpassen, die sich jedoch bezüglich Niveau und Anstieg zwischen beiden Geschlechtern unterscheidet.

Abbildung 2.28.: Beispiel eines Overlay Scatterplots

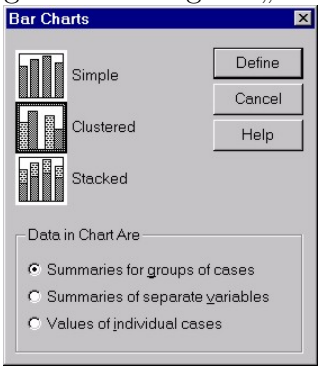


Wie bereits erwähnt, steht bei der Exploration eines Zusammenhangs zwischen zwei **nominalskalierten, ordinalskalierten bzw. metrisch skalierten diskreten Variablen mit wenigen Ausprägungen** die gemeinsame Häufigkeitsverteilung im Blickfeld der Betrachtung. SPSS bietet kein 3-D Balkendiagramm an, in dem die X-Achse die eine Variable, die Y-Achse die andere Variable und die Z-Achse die Häufigkeiten repräsentiert. Die einzige Möglichkeit besteht darin, ein Clustered Bar Chart (gruppiertes Balkendiagramm) zu erzeugen, in dem für jede Ausprägung der einen Variablen eine Gruppe von Balken entsprechend den Ausprägungen der zweiten „Gruppierungs-“ Variablen erzeugt wird. Das Clustered Bar Chart erhält man über

- Graphs
  - Bar...

(siehe Abb. 2.1). In dem Dialogfeld „Bar Charts“ ist „Clustered“ zu wählen, die Voreinstellung „Summaries for groups of cases“ (Kategorien einer Variablen) zu belassen und die Schaltfläche „Define“ zu betätigen.

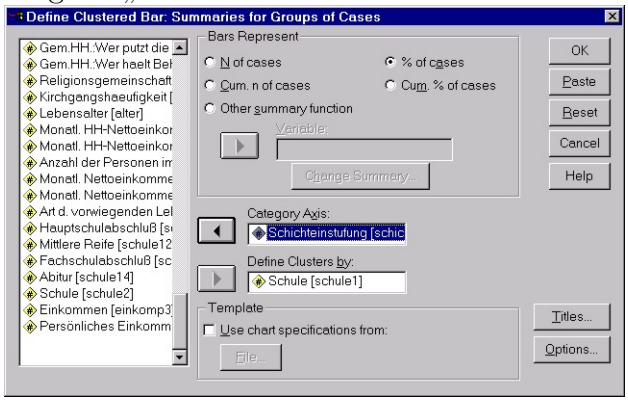
Abbildung 2.29.: Dialogfeld „Bar Charts“



Im Dialogfeld „Define Clustered Bar: Summaries for Groups of Cases“ (Gruppierte Balken: Auswertung über Kategorien einer Variablen, siehe Abb. 2.30) wird diejenige Variable, für deren Ausprägungen die Gruppen von Balken erzeugt werden sollen, aus der linken Quellliste in das Feld „Define Clusters by:“ gebracht. Im Feld „Bars Represent“ (Bedeutung der Balken) kann eine Entscheidung darüber getroffen werden, ob die Balken die absolute Häufigkeit (N of cases, d.h. die Anzahl der Fälle) oder relative Häufigkeit (% of cases) repräsentieren sollen.

Achtung: Bei der Interpretation des Clustered Bar Charts ist darauf zu achten, dass nicht die gemeinsame Häufigkeitsverteilung der beiden Variablen dargestellt wird, sondern die bedingten Verteilungen, und zwar die bedingte Verteilung der Variable im Feld „Category Axis:“, gegeben die Ausprägungen der Variablen im Feld „Define Clusters by:“.

Abbildung 2.30.: Dialogfeld „Define Clustered Bar: Summaries for Groups of Cases“



• Beispiel 2.6:

Bei der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) wurden u.a. folgende Fragen gestellt:

- Es wird viel über die verschiedenen Bevölkerungsschichten gesprochen. Welcher Schicht

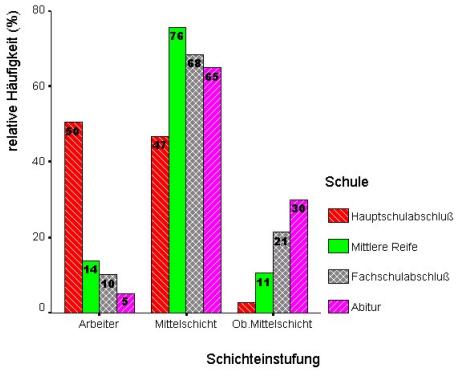
rechnen Sie sich selbst eher zu?

- Welchen allgemeinbildenden Schulabschluß haben Sie?

Für dieses Beispiel wurden nur die Fälle mit den Ausprägungen Arbeiter, Mittelschicht und obere Mittelschicht für die Variable Schichteinstufung (schicht1) und mit den Ausprägungen Hauptschulabschluß, mittlere Reife, Fachschulabschluß und Abitur für die Variable Schule (schule1) ausgewählt. Die Beobachtungsdaten dieser Variablen enthält die Datei allbus.sav[3]. Es interessiert die Frage, ob zwischen den Variablen Schule und Schichteinstufung eine Beziehung besteht. Beide Variablen sind nominalskaliert, so dass nur über ein gruppiertes Balkendiagramm eine explorative Beurteilung erfolgen kann, wobei als Bedeutung der Balken die relative Häufigkeit (% of cases) gewählt wird. Die Balken können mit den Häufigkeiten beschriftet werden, wenn sich das Balkendiagramm im SPSS Chart Editor befindet, dort „Format“ in der Menüleiste und in dem sich öffnenden Pull-Down-Menü „Bar Label Style“ gewählt wird.

Wird die Schichteinstufung als Kategorienvariable (Category Axis:) und die Variable Schule als Gruppierungsvariable (Define Clusters by:) gewählt, so resultiert das nachstehende gruppierte Balkendiagramm.

Abbildung 2.31.: Clustered Bar Chart der Variablen Schichteinstufung und Schule



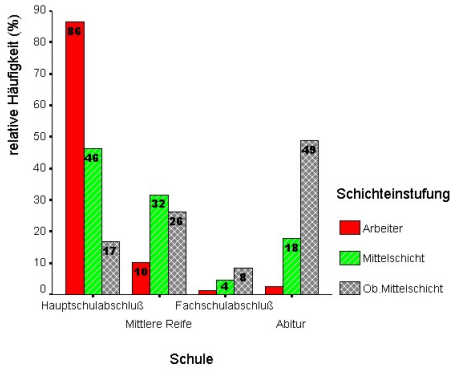
Dieses gruppierte Balkendiagramm zeigt die bedingte Verteilung von Schichteinstufung, gegeben eine Ausprägung der Variablen Schule. Zum Beispiel geben die von links oben nach rechts unten markierten (roten) Balken die bedingte Verteilung der Schichteinstufung an, gegeben die Ausprägung Hauptschulabschluß der Variablen Schule. Die Häufigkeiten dieser Balken müssen sich zu 100% addieren (bis auf Rundungsfehler, da nur ganzzahlige Werte auf den Balken angegeben werden).

Wird die Variable Schule als Kategorienvariable (Category Axis:) und die Schichteinstufung als Gruppierungsvariable (Define Clusters by:) gewählt, so resultiert das gruppierte Balkendia-

[3]Die Datei allbus.sav wurde entnommen aus: Wittenberg, R. (1991), jedoch durch umkodierte Variablen ergänzt.

gramm der Abb. 2.32.

Abbildung 2.32.: Clustered Bar Chart der Variablen Schule und Schichteinstufung

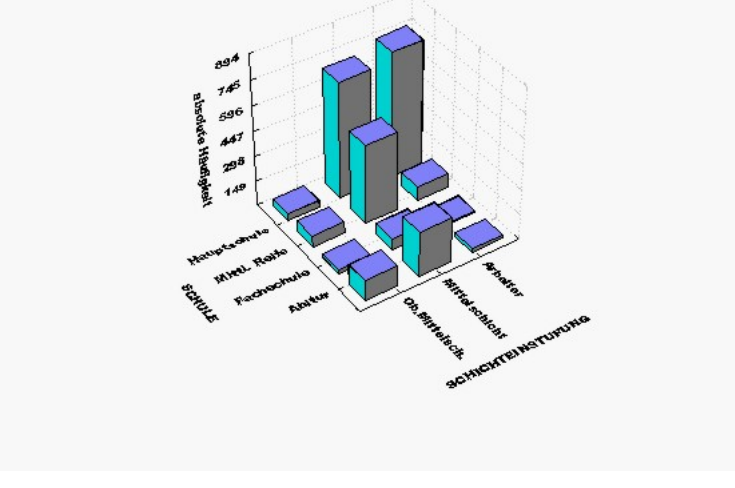


Dieses gruppierte Balkendiagramm zeigt die bedingte Verteilung der Variablen Schule, gegeben eine Ausprägung der Variablen Schichteinstufung. Zum Beispiel enthalten die kreuzmakierten Balken die bedingte Verteilung der Variablen Schule, gegeben die Ausprägung obere Mittelschicht der Variablen Schichteinstufung. Die Häufigkeiten dieser Blaken müssen sich zu 100% addieren.

Jedes gruppierte Balkendiagramm zeigt deutliche Unterschiede in den bedingten Verteilungen, so dass offensichtlich keine Unabhängigkeit zwischen den Variablen Schichteinstufung und Schule besteht.

Ein Statistik-Softwarepaket, das tatsächlich die gemeinsame Häufigkeitsverteilung in einem 3-D Balkendiagramm realisiert, ist z.B. STATISTICA. Die folgende Abbildung zeigt für das Beispiel 2.6 das 3-D Balkendiagramm, in dem die absoluten Häufigkeiten der gemeinsamen Verteilung von Schichteinstufung und Schule auf der Z-Achse abgetragen sind.

Abbildung 2.33.: 3-D Balkendiagramm der Variablen Schichteinstufung und Schule



2.2. Beziehung zwischen nominal- bzw. ordinalskalierten Daten

2.2.1. Die Kontingenztabelle

Wie bereits eingangs zu diesem Kapitel erwähnt, wird bei der Untersuchung von Zusammenhängen nominalskalierter, ordinalskalierter Variablen sowie metrisch klassierter Variablen die gemeinsame Häufigkeitsverteilung herangezogen. Eine geeignete Darstellungsform für die gemeinsame Häufigkeitsverteilung zweier Variablen ist die bivariate Kontingenztabelle (auch als Kreuztabelle bezeichnet).

Es sei allgemein angenommen, dass eine Variable X mit J Variablenausprägungen (j = 1,...,J) und eine Variable Y mit K Variablenausprägungen (k = 1,...,K) gegeben sind und diese Variablen an n statistischen Einheiten (i = 1,...,n) beobachtet wurden. Es gibt somit $J \cdot K$ mögliche Paare von Variablenausprägungen (x_j, y_k). Den allgemeinen Aufbau einer bivariaten Kontingenztabelle zeigt die Tabelle 2.1.

Tabelle 2.1.: Zweidimensionale Kontingenztabelle

Variable X	Variable Y					Randverteilung X
	y_1	$\cdots$	y_k	$\cdots$	y_K	
x_1	h_{11}	$\cdots$	h_{1k}	$\cdots$	h_{1K}	h_{1+}
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_j	h_{j1}	$\cdots$	h_{jk}	$\cdots$	h_{jK}	h_{j+}
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_J	h_{J1}	$\cdots$	h_{Jk}	$\cdots$	h_{JK}	h_{J+}
Randverteilung Y	h_{+1}	$\cdots$	h_{+k}	$\cdots$	h_{+K}	$h_{++} = n$

Die Zeilen der Tabelle entsprechen den Variablenausprägungen der einen Variablen (in der Tabelle 2.1 die Variable X) und die Spalten den Variablenausprägungen der zweiten Variablen (in der Tabelle 2.1 die Variable Y). Innerhalb der Tabelle gibt es $J \cdot K$ Zellen. Diese Zellen nehmen im allgemeinen die absoluten oder relativen Häufigkeiten des Auftretens eines Paares von Variablenausprägungen (x_j, y_k) auf. Es bezeichnen im weiteren für $j = 1, \ldots, J$ und $k = 1, \ldots, K$:

- $$h_{jk} = h(x_j, y_k) = h[(X = x_j) \cap (Y = y_k)] \tag{2.4}$$

  die absolute Zellhäufigkeit als die Anzahl der Stichprobenelemente mit der Variablenausprägung x_j der Variablen X und der Variablenausprägung y_k der Variablen Y;

- $$f_{jk} = f(x_j, y_k) = f[(X = x_j) \cap (Y = y_k)] = h_{jk}/n \tag{2.5}$$

  die relative Zellhäufigkeit als der Anteil der Elemente in der Stichprobe mit der Variablenausprägung x_j der Variablen X und der Variablenausprägung y_k der Variablen Y.

Die letzte Spalte der Kontingenztabelle gibt die eindimensionale Randverteilung der Variablen X an, d.h.,

$$\begin{aligned} h_{j+} &= h(X = x_j) = \sum_{k=1}^{K} h_{jk} \\ f_{j+} &= f(X = x_j) = \sum_{k=1}^{K} f_{jk} = \frac{h_{j+}}{n} \end{aligned} \qquad j = 1, \ldots, J \tag{2.6}$$

beinhaltet die Anzahl bzw. den Anteil der Elemente in der Stichprobe mit der Variablenausprägung x_j der Variaben X.

Die letzte Zeile der Kontingenztabelle beinhaltet die eindimensionale Randverteilung der Variablen Y, d.h.,

$$\begin{aligned} h_{+k} &= h(Y = y_k) = \sum_{j=1}^{J} h_{jk} \\ f_{+k} &= f(Y = y_k) = \sum_{j=1}^{J} f_{jk} = \frac{h_{+k}}{n} \end{aligned} \qquad k = 1, \ldots, K \tag{2.7}$$

symbolisiert die Anzahl bzw. den Anteil der Elemente in der Stichprobe mit der Variablenausprägung y_k der Variablen Y.

In der rechten unteren Ecke der Kontingenztabelle steht die Gesamtzahl der beobachteten statistischen Einheiten (Stichprobenumfang):

$$h_{++} = \sum_{j=1}^{J} \sum_{k=1}^{K} h_{jk} = \sum_{j=1}^{J} h_{j+} = \sum_{k=1}^{K} h_{+k} = n \tag{2.8}$$

bzw. $f_{++} = 1$ (bzw. 100%), wenn die Zellen der Kontingenztabelle die relativen (prozentualen) Häufigkeiten enthalten.

Der Inhalt der Zellen einer Kontingenztabelle kann durch weitere, für die statistische Auswertung wichtige Informationen ergänzt bzw. durch sie ersetzt werden:

- erwartete Zellhäufigkeiten

  Es ist diejenige absolute Häufigkeit in jeder Zelle, die unter der Annahme der stochastischen Unabhängigkeit der beiden Variablen zu erwarten ist.

  Sind die Variablen X und Y stochastisch unabhängig, dann gilt nach dem Multiplikationssatz für unabhängige Ereignisse für die Grundgesamtheit:

$$\begin{aligned} P(\{X = x_j\} \cap \{Y = y_k\}) &= P(X = x_j) \cdot P(Y = y_k) \\ &= p_{j+} \cdot p_{+k} = p_{jk} \qquad \text{(für alle j und k).} \end{aligned} \tag{2.9}$$

Die erwarteten absoluten Häufigkeiten (e_{jk}) in der Grundgesamtheit ergeben sich zu:

$$e_{jk} = n \cdot p_{j+} \cdot p_{+k} = n \cdot p_{jk}. \tag{2.10}$$

Da die wirklichen Randwahrscheinlichkeiten p_{j+} und p_{+k} unbekannt sind, werden sie mittels der beobachteten Randhäufigkeiten f_{j+} (2.6) und f_{+k} (2.7) geschätzt. Für die geschätzten erwarteten absoluten Häufigkeiten folgt dann:

$$\hat{e}_{jk} = n \cdot f_{j+} \cdot f_{+k} = \frac{h_{j+}h_{+k}}{n}. \tag{2.11}$$

- bedingte Häufigkeitsverteilungen

  - die bedingte relative Häufigkeit der Variablen Y für eine gegebene Ausprägung x_j der Variablen X:

    $$f(y_k|x_j) = \frac{f(\{X = x_j\} \cap \{Y = y_k\})}{f(X = x_j)} = \frac{h_{jk}}{h_{j+}}, \quad k = 1, \ldots, K \tag{2.12}$$

    mit

    $$\sum_{k=1}^{K} f(y_k|x_j) = 1; \tag{2.13}$$

  - die bedingte relative Häufigkeit der Variablen X für eine gegebene Ausprägung y_k der Variablen Y:

    $$f(x_j|y_k) = \frac{f(\{X = x_j\} \cap \{Y = y_k\})}{f(Y = y_k)} = \frac{h_{jk}}{h_{+k}}, \quad j = 1, \ldots, J \tag{2.14}$$

    mit

    $$\sum_{j=1}^{J} f(x_j|y_k) = 1. \tag{2.15}$$

    Die bedingten Häufigkeitsverteilungen sind eindimensionale Verteilungen.

- Residuen

  Ein Residuum beinhaltet die Abweichungen zwischen der beobachteten und der unter Unabhängigkeit zu erwartenden Häufigkeit einer Zelle. Dabei können drei Arten von Residuen unterschieden werden:

- unstandardisiertes Residuum

  Es ist die Abweichung zwischen der beobachteten absoluten Häufigkeit und der bei Unabhängigkeit zu erwartenden absoluten Häufigkeit einer Zelle:

$$r_{jk} = h_{jk} - \hat{e}_{jk}. \tag{2.16}$$

- standardisiertes Residuum

  Das standardisierte Residuum einer Zelle ergibt sich als unstandardisiertes Residuum dividiert durch die Quadratwurzel aus der erwarteten absoluten Häufigkeit:

$$rs_{jk} = \frac{h_{jk} - \hat{e}_{jk}}{\sqrt{\hat{e}_{jk}}} = \frac{r_{jk}}{\sqrt{\hat{e}_{jk}}}. \tag{2.17}$$

  Diese standardisierten Residuen erweisen sich bei der Auswertung des weiter unten dargestellten Chi-Quadrat-Tests als nützlich.

- korrigiertes standardisiertes Residuum

  Das korrigierte standardisierte (adjusted) Residuum einer Zelle beinhaltet das unstandardisierte Residuum dividiert durch eine Schätzung des Standardfehlers der erwarteten Zellhäufigkeit. Die Schätzung des Standardfehlers erfolgt dabei unter Berücksichtigung der Zeilenhäufigkeit h_{j+}, der Spaltenhäufigkeit h_{k+} sowie der Gesamtzahl der Beobachtungen n.

$$ra_{jk} = \frac{h_{jk} - \hat{e}_{jk}}{\sqrt{\hat{e}_{jk} \cdot \left(1 - \frac{h_{j+}}{n}\right)\left(1 - \frac{h_{+k}}{n}\right)}} = \frac{r_{jk}}{\sqrt{\hat{e}_{jk} \cdot \left(1 - \frac{h_{j+}}{n}\right)\left(1 - \frac{h_{+k}}{n}\right)}}. \tag{2.18}$$

Unter SPSS kann eine Kreuztabelle über

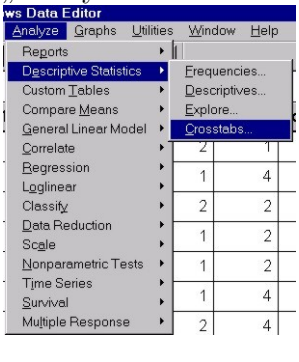
■ Analyze

  ■ Descriptive Statistics

    ■ Crosstabs ...

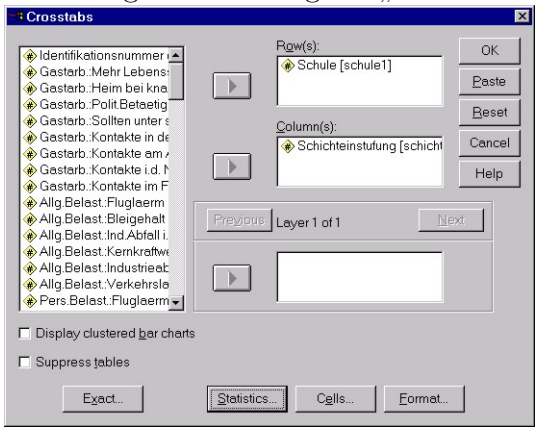
angefordert werden.

Abbildung 2.34.: Menü „Analyze“ und Menü „Descriptive Statistics“



Die weiteren Entscheidungen sind in dem Dialogfeld „Crosstabs“ zu treffen.

Abbildung 2.35.: Dialogfeld „Crosstabs“



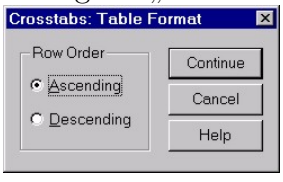
Sowohl in das Feld „Row(s):“ (Zeilen) als auch in das Feld „Column(s):“ (Spalten) ist mindestens eine Variable aus der linken Quelliste einzutragen. Wurde mehr als eine Variable für Zeilen und/oder Spalten ausgewählt, so wird für jede Kombination von zwei Variablen eine Kreuztabelle erstellt. Stehen z.B. drei Variablen in dem Feld „Row(s):“ und zwei Variablen in dem Feld „Column(s):“, so werden $3 \cdot 2 = 6$ Kreuztabellen erzeugt.

Desweiteren können sogenannte Kontrollvariablen in das Feld „Layer 1 of 1“ (Ebene 1 von 1) eingetragen werden. Es werden dann Kreuztabellen der angegebenen Zeilen- und Spaltenvariablen für jede Variablenausprägung (Kategorie) dieser Kontrollvariablen in der ersten Ebene ausgegeben. Wurden Kontrollvariablen auch für eine weitere Ebene (z.B. Layer 2 of 2) ausgewählt, so entstehen Kreuztabellen für jede Kombination von Kategorien der Variablen der ersten Ebene und der Kategorien der Variablen der zweiten Ebene, usw. Durch diese Möglichkeit können existierende Unterschiede bezüglich des Zusammenhanges zweier Variablen nach einer oder mehreren anderen Variablen kontrolliert werden.

Über die Schaltfläche „Format . . . “ kann in dem sich öffnenden Dialogfeld „Crosstabs: Table Format“ die Sortierung der Zeilenvariablen beeinflußt werden (siehe Abb. 2.36):

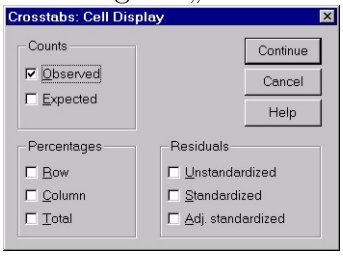
- Ascending (aufsteigend)
- Descending (absteigend).

Abbildung 2.36.: Dialogfeld „Crosstabs: Table Format“



Der Inhalt der Zellen der Kreuztabelle kann ebenfalls variiert werden. Nach Betätigung der Schaltfläche „Cells . . . “ erscheint nachstehendes Dialogfeld.

Abbildung 2.37.: Dialogfeld „Crosstabs: Cell Display“



Als Inhalt der Zellen kann gewählt werden:

- Counts (absolute Werte)

  - Observed (beobachtete)
    Die voreingestellte Anzeige für Kreuztabellen beinhaltet nur die Ausgabe der beobachteten absoluten Häufigkeiten h_{jk} (gemäß Formel 2.4) in jeder Zelle und der beobachteten absoluten Häufigkeiten der Randverteilungen h_{j+} und h_{+k} (gemäß Formel 2.6 und 2.7)
  - Expected (erwartete)
    Es wird die unter der Annahme der stochastischen Unabhängigkeit der beiden Variablen erwartete absolute Häufigkeit in jeder Zelle (gemäß Formel 2.11) ausgegeben.

- Percentages (Prozentwerte)

  - Row (zeilenweise)
    Hierbei handelt es sich um die bedingte relative Häufigkeit der Variablen Y für eine gegebene Ausprägung x_j der Variablen X (gemäß Formel 2.12), angegeben in Prozent.
  - Column (spaltenweise)
    Dies ist die bedingte relative Häufigkeit der Variablen X für eine gegebene Ausprägung y_k der Variablen Y (gemäß Formel 2.14), angegeben in Prozent.

- Total (gesamt)
  Es ist die relative Häufigkeit (gemäß Formel 2.5) des Auftretens eines Paares von Variablenausprägungen (x_j, y_k), angegeben in Prozent.

Alle Prozentangaben werden auf eine Dezimalstelle gerundet angegeben.

- Residuals (Residuen)

  - Unstandardized (unstandardisiert)
    Es ist die Abweichung zwischen der beobachteten absoluten Häufigkeit und der bei Unabhängigkeit zu erwartenden absoluten Häufigkeit einer Zelle (gemäß Formel 2.16).
  - Standardized (standardisiert), gemäß Formel 2.17
  - Adj. Standardized (korrigiert standardisiert), gemäß Formel 2.18

  Alle Residuen werden nur auf eine Dezimalstelle genau ausgegeben.

• Beispiel 2.6: (Fortsetzung):
Für die Variablen Schichteinstufung (als Spaltenvariable) und Schule (als Zeilenvariable) wird die Kontingenztabelle erstellt, wobei zu Demonstrationszwecken alle Möglichkeiten der inhaltlichen Gestaltung der Zellen ausgewählt werden. Den Output im SPSS Viewer enthält Tabelle 2.2.

Insgesamt gibt es n = 2619 gültige Fälle, d.h. befragte Personen, die bei beiden Variablen gültige Ausprägungen ausweisen. Davon haben 1484 Personen oder 56,7 % einen Hauptschulabschluß, 627 Personen oder 23,9 % die Mittlere Reife, 98 Personen oder 3,7 % einen Fachschulabschluß und 410 Personen oder 15,7 % das Abitur. Hierbei handelt es sich um die Randverteilung der Variablen Schule (gemäß Formel 2.6)). Die Randverteilung der Variablen Schichteinstufung (gemäß Formel 2.7) ist in der letzten Zeile gegeben: 866 Personen oder 33,1 % stuften sich als Arbeiter, 1501 Personen oder 57,3 % zur Mittelschicht gehörig und 252 Personen oder 9,6 % zur oberen Mittelschicht gehörig ein.

Tabelle 2.2.: Kontingenztabelle für Schichteinstufung und Schule

Schule * Schichteinstufung Crosstabulation

			SE			
			Arbeiter	MS	Ob.MS	Total
Schule	Haupt-	Count	748	694	42	1484
	schul-	Expected Count	490,7	850,5	142,8	1484,0
	abschluß	% within Schule	50,4%	46,8%	2,8%	100%
		% within Schichteinstufung	86,4%	46,2%	16,7%	56,7%
		% of Total	28,6%	26,5%	1,6%	56,7%
		Residual	257,3	-156,5	-100,8	
		Std. Residual	11,6	-5,4	-8,4	
		Adjusted Residual	21,6	-12,5	-13,5	
	Mittlere	Count	87	474	66	627
	Reife	Expected Count	207,3	359,3	60,3	627,0
		% within Schule	13,9%	75,6%	10,5%	100,0%
		% within Schichteinstufung	10,0%	31,6%	26,2%	23,9%
		% of Total	3,3%	18,1%	2,5%	23,9%
		Residual	-120,3	114,7	5,7	
		Std. Residual	-8,4	6,0	,7	
		Adjusted Residual	-11,7	10,6	,9	
	Fach-	Count	10	67	21	98
	schul-	Expected Count	32,4	56,2	9,4	98,0
	abschluß	% within Schule	10,2%	68,4%	21,4%	100%
		% within Schichteinstufung	1,2%	4,5%	8,3%	3,7%
		% of Total	,4%	2,6%	,8%	3,7%
		Residual	-22,4	10,8	11,6	
		Std. Residual	-3,9	1,4	3,8	
		Adjusted Residual	-4,9	2,3	4,0	
	Abitur	Count	21	266	123	410
		Expected Count	135,6	235,0	39,5	410,0
		% within Schule	5,1%	64,9%	30,0%	100,0%
		% within Schichteinstufung	2,4%	17,7%	48,8%	15,7%
		% of Total	,8%	10,2%	4,7%	15,7%
		Residual	-114,6	31,0	83,5	
		Std. Residual	-9,8	2,0	13,3	
		Adjusted Residual	-13,1	3,4	15,2	
Total		Count	866	1501	252	2619
		Expected Count	866,0	1501,0	252,0	2619,0
		% within Schule	33,1%	57,3%	9,6%	100,0%
		% within Schichteinstufung	100,0%	100,0%	100,0%	100,0%
		% of Total	33,1%	57,3%	9,6%	100,0%

Abkürzungen: SE - Schichteinstufung, MS - Mittelschicht, Ob.MS - obere Mittelschicht, HSA - Hauptschulabschluß, MR - Mittlere Reife, FSA - Fachschulabschluß

Die Angaben in den Zellen, z.B. in der ersten Zelle, sind wie folgt zu interpretieren:

- Zellzeile Count:
  748 Personen (h_{11}) haben einen Hauptschulabschluß **und** stuften sich als Arbeiter ein. (Hinweis: Die gemeinsame Häufigkeitsverteilung von Schichteinstufung und Schule ist in der Abb. 2.33 dargestellt.)

- Zellzeile: % of Total
  28,6% ($f_{11} = 748/2619 \cdot 100\%$, Formel (2.5)) haben einen Hauptschulabschluß **und** stuften sich als Arbeiter ein.

- Zellzeile: % within Schule
  50,4% der Personen mit einem Hauptschulabschluß stuften sich als Arbeiter ein. Hierbei handelt es sich um die bedingte relative Häufigkeit für die Ausprägung Arbeiter der Variablen Schichteinstufung (y_1 von Y) für die gegebene Ausprägung Hauptschulabschluß der Variablen Schule (x_1 von X) gemäß Formel (2.12): $f(y_1|x_1) = 748/1484 \cdot 100\%$.
  (Hinweis: Diese bedingte Häufigkeitsverteilung ist in der Abb. 2.31 dargestellt.)

- Zellzeile: % within Schichteinstufung
  86,4% der Personen, die sich als Arbeiter einstuften, haben einen Hauptschulabschluß. Hierbei handelt es sich um die bedingte relative Häufigkeit für die Ausprägung Hauptschulabschluß der Variablen Schule (x_1 von X) für die gegebene Ausprägung Arbeiter der Variablen Schichteinstufung (y_1 von Y) gemäß Formel (2.14): $f(x_1|y_1) = 748/866 \cdot 100\%$.
  (Hinweis: Diese bedingte Häufigkeitsverteilung ist in der Abb. 2.32 dargestellt.)

- Zellzeile: Expected Count
  Bei Unabhängigkeit der Variablen Schule und Schichteinstufung wären 490,7 Personen mit Hauptschulabschluß und Selbsteinstufung als Arbeiter zu erwarten gewesen, gemäß Formel (2.11): $\hat{e}_{11} = 1484 \cdot 866/2619$.

- Zellzeile: Residual
  Da $h_{11} = 748$ und $\hat{e}_{11} = 490,7$ sind, besteht eine Abweichung von $r_{11} = 257,3$ (Formel (2.16)).

- Zellzeile: Std. Residual
  Das standardisierte Residuum rs_{11} ergibt sich gemäß Formel (2.17) zu:

$$rs_{11} = \frac{257,3}{\sqrt{490,7}} = 11,615.$$

- Zellzeile: Adjusted Residual
  Das korrigiert standardisierte Residuum ra_{11} ergibt sich gemäß Formel (2.18) zu:

$$ra_{11} = \frac{257,3}{\sqrt{490,7 \cdot \left(1 - \frac{1484}{2619}\right)\left(1 - \frac{866}{2619}\right)}} = 21,566.$$

Analog sind die anderen Zellen zu interpretieren.

Eine Einschätzung, ob die beiden in der Kontingenztabelle enthaltenen Variablen unabhängig sind, kann in zweierlei Weise erfolgen:

1. Zum einen kann anhand der unstandardisierten Residuen überprüft werden, ob die beobachteten gleich den erwarteten Häufigkeiten in allen Zellen sind. Die Abweichungen r_{jk} (Residual, 6. Zahlenwert in jeder Zelle) sind zum Teil erheblich.

2. Zum anderen müssen bei Unabhängigkeit die bedingten Verteilungen einer Variablen untereinander und mit der jeweiligen Randverteilung übereinstimmen:
   $f(y_k|x_j) = f(y_k)$ für alle k
   $f(x_j|y_k) = f(x_j)$ für alle j.
   Greift man die dafür relevanten Informationen aus den obigen Kontingenztabellen heraus, so erhält man:

   a) für die bedingte Verteilung der Schichteinstufung (Y) für gegebene Ausprägungen von Schule (X)

Tabelle 2.3.: Bedingte Verteilung von Schichteinstufung (Y) für gegebene Ausprägungen von Schule (X)

Schule * Schichteinstufung

% within Schule

		SE			
		Arbeiter	Mittelschicht	Ob. Mittelschicht	Total
Schule	Hauptschulabschluß	50,4%	46,8%	2,8%	100,0%
	Mittlere Reife	13,9%	75,6%	10,5%	100,0%
	Fachschulabschluß	10,2%	68,4%	21,4%	100,0%
	Abitur	5,1%	64,9%	30,0%	100,0%
Total		33,1%	57,3%	9,6%	100,0%

b) für die bedingte Verteilung von Schule (X) für gegebene Ausprägungen der Schichteinstufung (Y)

Tabelle 2.4.: Bedingte Verteilung von Schule (x) für gegebene Ausprägungen von Schichteinstufung (Y)

Schule * Schichteinstufung

% within Schichteinstufung

		SE			
		Arbeiter	Mittelschicht	Ob. Mittelschicht	Total
Schule	Hauptschulabschluß	86,4%	46,2%	16,7%	56,7%
	Mittlere Reife	10,0%	31,6%	26,2%	23,9%
	Fachschulabschluß	1,2%	4,5%	8,3%	3,7%
	Abitur	2,4%	17,7%	48,8%	15,7%
Total		100,0%	100,0%	100,0%	100,0%

Deutliche Unterschiede zwischen den bedingten Verteilungen und der jeweiligen Randverteilung werden in den Tabellen 2.3 und 2.4 sichtbar. Es liegt die Vermutung nahe, dass keine Unabhängigkeit zwischen den beiden Variablen Schichteinstufung und Schule besteht.

Ein wichtiges Argument bei dieser Auswertung besteht darin, dass das Geschlecht der befragten Personen einen Einfluß auf die Beziehung zwischen Schichteinstufung und Schule haben könnte. Um diesen Einfluß der Variablen Geschlecht bei der Überprüfung des Zusammenhangs zwischen Schichteinstufung und Schule zu berücksichtigen, kann die Variable Geschlecht als Kontrollvariable in das Feld „Layer 1 of 1“ (siehe Abb. 2.35) gebracht werden. Tabelle 2.5 zeigt den Output im SPSS Viewer, für den Count, Expected und unstandardized Residuals angefordert wurden, Tabelle 2.6 den Output mit lediglich Row Percentages und Tabelle 2.7 den Output mit nur Column Percentages.

In jeder der durch eine Ausprägung der Variablen Geschlecht gegebenen Kontingenztabelle von Schule und Schichteinstufung sind die unstandardisierten Residuen weiterhin recht groß, und die bedingten Häufigkeitsverteilungen weisen untereinander und zur jeweiligen Randverteilung erhebliche Unterschiede auf. Sowohl für die Männer als auch für die Frauen kann die Vermutung beibehalten werden, dass zwischen Schichteinstufung und Schule eine Beziehung besteht.

Tabelle 2.5.: Kontingenztabelle für Schichteinstufung und Schule, kontrolliert für die Variable Geschlecht

Schule * Schichteinstufung * Geschlechtszugehoerigkeit Crosstabulation

Geschlecht				SE			Total
				Arbeiter	MS	Ob. MS	
Mann	Schule	Haupt-	Count	353	276	15	644
		schul-	Expected Count	228,9	351,6	63,6	644,0
		abschluß	Residual	124,1	-75,6	-48,6	
		Mittlere	Count	36	195	25	256
		Reife	Expected Count	91,0	139,8	25,3	256,0
			Residual	-55,0	55,2	-,3	
		Fach-	Count	9	34	13	56
		schul-	Expected Count	19,9	30,6	5,5	56,0
		abschluß	Residual	-10,9	3,4	7,5	
		Abitur	Count	16	131	62	209
			Expected Count	74,3	114,1	20,6	209,0
			Residual	-58,3	16,9	41,4	
	Total		Count	414	636	115	1165
			Expected Count	414,0	636,0	115,0	1165,0
Frau	Schule	Haupt-	Count	395	418	27	840
		schul-	Expected Count	261,1	499,7	79,1	840,0
		abschluß	Residual	133,9	-81,7	-52,1	
		Mittlere	Count	51	279	41	371
		Reife	Expected Count	115,3	220,7	35,0	371,0
			Residual	-64,3	58,3	6,0	
		Fach-	Count	1	33	8	42
		schul-	Expected Count	13,1	25,0	4,0	42,0
		abschluß	Residual	-12,1	8,0	4,0	
		Abitur	Count	5	135	61	201
			Expected Count	62,5	119,6	18,9	201,0
			Residual	-57,5	15,4	42,1	
	Total		Count	452	865	137	1454
			Expected Count	452,0	865,0	137,0	1454,0

Tabelle 2.6.: Bedingte Verteilung von Schichteinstufung (Y) für gegebene Ausprägungen von Schule (X), kontrolliert für die Variable Geschlecht

Schule * Schichteinstufung * Geschlechtszugehoerigkeit Crosstabulation

% within Schule

Geschlecht			SE			Total
			Arbeiter	Mittelschicht	Ob. Mittelschicht	
Mann	Schule	HSA	54,8%	42,9%	2,3%	100,0%
		MR	14,1%	76,2%	9,8%	100,0%
		FSA	16,1%	60,7%	23,3%	100,0%
		Abitur	7,7%	62,7%	29,7%	100,0%
	Total		35,5%	54,6%	9,9%	100,0%
Frau	Schule	HSA	47,0%	49,8%	3,2%	100,0%
		MR	13,7%	75,2%	11,1%	100,0%
		FSA	2,4%	78,6%	19,0%	100,0%
		Abitur	2,5%	67,2%	30,3%	100,0%
	Total		31,1%	59,5%	9,4%	100,0%

Tabelle 2.7.: Bedingte Verteilung von Schule (X) für gegebene Ausprägungen von Schichteinstufung (Y), kontrolliert für die Variable Geschlecht

Schule * Schichteinstufung * Geschlechtszugehoerigkeit Crosstabulation

% within Schichteinstufung

Geschlecht			SE			Total
			Arbeiter	Mittelschicht	Ob. Mittelschicht	
Mann	Schule	HSA	85,3%	43,4%	13,0%	55,3%
		MR	8,7%	30,7%	21,7%	22,0%
		FSA	2,2%	5,3%	11,3%	4,8%
		Abitur	3,9%	20,6%	53,9%	17,9%
	Total		100,0%	100,0%	100,0%	100,0%
Frau	Schule	HSA	87,4%	48,3%	19,7%	57,8%
		MR	11,3%	32,3%	29,9%	25,5%
		FSA	,2%	3,8%	5,8%	2,9%
		Abitur	1,1%	15,6%	44,5%	13,8%
	Total		100,0%	100,0%	100,0%	100,0%

Nur anhand dieser Residuen bzw. bedingten Häufigkeitsverteilungen einzuschätzen, ob tatsächlich ein Zusammenhang zwischen Schichteinstufung und Schule gegeben ist, dürfte schwer fallen. Da hier die Ergebnisse einer Stichprobe vorliegen, werden selbst dann, wenn Unabhängigkeit der beiden Variablen in der Grundgesamtheit gegeben ist, immer gewisse Abweichungen zwischen beobachteten und erwarteten Häufigkeiten bzw. zwischen den bedingten Verteilungen zu beobachten sein.

Wie groß dürfen aber die Abweichungen sein, um noch auf Unabhängigkeit zu erkennen, bzw. wann handelt es sich um signifikante Abweichungen? Um hier einer subjektiven Einschätzung vorzubeugen, ist ein statistischer Test auf Unabhängigkeit anzuwenden. Wenn ein Zusammenhang anzunehmen ist, wie das im obigen Beispiel von Schichteinstufung und Schule der Fall ist, interessiert weiterhin die Stärke dieses Zusammenhanges sowie eine konkrete Modellschätzung und Modellanpassung. Letzteres führt zu den verallgemeinerten linearen Modellen, die hier nicht diskutiert werden sollen; es sei auf die gleichnamige Lehrveranstaltung verwiesen: http://ise.wiwi.hu-berlin.de/statistik/glm_d.html

Erläuterungen zur generellen Testbeantwortung bei Verwendung statistischer Software sind im Anhang B enthalten.

2.2.2. Tests auf Unabhängigkeit zweier Variablen

Das zu prüfende Hypothesenpaar lautet:

H_0: Die Zufallsvariablen X und Y sind stochastisch unabhängig, d.h. $p_{jk} = p_{j+} \cdot p_{+k}$ für alle Paare (j,k).

H_1: Die Zufallsvariablen X und Y sind nicht stochastisch unabhängig, d.h. $p_{jk} \neq p_{j+} \cdot p_{+k}$ für mindestens ein Paar (j,k).

Das Signifikanzniveau α und der Stichprobenumfang n sind vor der Testdurchführung festzulegen.

2.2.2.1. Chi-Quadrat-Unabhängigkeitstest nach Pearson

Ein allgemein angewandter Test zur Überprüfung der Nullhypothese ist der Chi-Quadrat-Unabhängigkeitstest nach Pearson[4].

[4]Siehe u.a. Rönz, B., Strohe, H.G. (1994), S. 67 ff.; Bamberg, G., Baur, F. (1991), S. 202 ff.; Schlittgen, R. (1990), S.384 ff.; Büning, H., Trenkler, G. (1978), S. 238 ff.; Schwarze, J. (1990), S. 249 ff.; Bosch, K. (1992), S.384 ff.; Backhaus, K., Erichson, B., Plinke, W., Weiber, R. (1994), S. 164 ff.; Hartung, J., Elpelt, B., Klösener, K.-H. (1993), S. 413 ff.

Bei diesem Test wird geprüft, ob zwei Zufallsvariablen stochastisch unabhängig sind. Der Chi-Quadrat-Unabhängigkeitstest gehört zu den nichtparametrischen Tests. An das Skalenniveau der Zufallsvariablen werden keine Voraussetzungen gestellt; er ist jedoch ungeeignet für stetige Variablen, die nicht klassiert wurden.

Der Chi-Quadrat-Unabhängigkeitstest setzt voraus, dass

- die Beobachtungen unabhängig sind (d.h. eine einfache Zufallsstichprobe gezogen wurde),
- die Aufteilung der Variablenausprägungen vollständig ist, d.h., jede statistische Einheit (Fall) gehört zu einem Paar von Variablenausprägungen (x_j, y_k).

Der Test basiert auf dem Vergleich der in der Stichprobe beobachteten und der bei Gültigkeit der Nullhypothese H_0 erwarteten absoluten Zellhäufigkeiten. Für die konkrete Stichprobe sind die gemeinsamen absoluten Häufigkeiten h_{jk} (j = 1, ..., J; k = 1, ..., K) in den Zellen der zweidimensionalen Häufigkeitstabelle gegeben. Da diese absoluten Häufigkeiten h_{jk} Ergebnis eines Zufallsexperimentes sind, können sie von Stichprobe zu Stichprobe unterschiedliche Werte annehmen, d.h. sie sind Realisationen von Zufallsvariablen H_{jk}. Wenn die Nullhypothese gilt, ergeben sich die erwarteten absoluten Häufigkeiten $\hat{e}_{jk}$ gemäß Formel (2.11). Der Vergleich baut somit auf den Differenzen $H_{jk} - \hat{e}_{jk}$ auf. Eine summarische Größe, die die Abweichungen von der Nullhypothese bewertet, ist die Teststatistik

$$V = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(H_{jk} - \hat{e}_{jk})^2}{\hat{e}_{jk}}. \tag{2.19}$$

Es ist ersichtlich, dass diese Teststatistik die quadrierten standardisierten Residuen nach Formel (2.17) enthält, d.h. rs^2_{jk} ist der Beitrag der Zelle (j,k) mit der Ausprägung x_j von X und der Ausprägung y_k von Y zu dieser Teststatistik.

Durch die Gewichtung der quadrierten Abweichungen (Residuen) mit $1/\hat{e}_{jk}$ wird der relativen Größe der Abweichungen Rechnung getragen: Zum Beispiel ist für $h_{jk} = 8$ und $\hat{e}_{jk} = 12$ bzw. $h_{jk} = 96$ und $\hat{e}_{jk} = 100$ die quadrierte Abweichung mit 16 gleich, im ersteren Fall ist die Abweichung jedoch relativ größer.

Aus der Definition der Teststatistik folgt unmittelbar, dass die Nullhypothese für „zu große“ Werte von V abgelehnt wird.

Unter H_0 ist die Teststatistik V approximativ χ^2-verteilt mit $DF = (J-1)(K-1)$ Freiheitsgraden (degrees of freedom). Je größer der Stichprobenumfang n ist, desto besser ist die Approximation. Für den Test gelten die Approximationsbedingungen:

- Die erwartete Häufigkeit $\hat{e}_{jk}$ jeder Zelle muß größer als 1 sein.
- Höchstens 20% der Zellen dürfen erwartete Häufigkeiten $\hat{e}_{jk}$ kleiner als 5 aufweisen.

Andernfalls ist die Approximation der Verteilung der Teststatistik V an die Chi-Quadrat-Verteilung nicht hinreichend. Sind diese Bedingungen nicht erfüllt, müssen vor der Anwendung des Tests benachbarte Ausprägungen bzw. Klassen zusammengefaßt werden. J und K sind dann die Anzahlen der verbliebenen Ausprägungen bzw. Klassen nach einer eventuell notwendigen Zusammenfassung.

Die Entscheidungsbereiche der Nullhypothese können erst nach Vorliegen der Stichprobe festgelegt werden, da

1. die erwarteten gemeinsamen absoluten Häufigkeiten $\hat{e}_{jk}$ aufgrund der Stichprobe zu schätzen sind,

2. erst dann die Approximationsbedingungen überprüft werden können und ersichtlich ist, ob Ausprägungen zusammenzufassen sind,

3. erst danach die Anzahl der Freiheitsgrade feststeht und der kritische Wert aufgesucht werden kann.

Der kritische Wert $c = \chi^2_{1-\alpha;DF}$ wird für $P(V \leq c) = 1 - \alpha$ und die Anzahl der Freiheitsgrade DF aus der Tabelle der Verteilungsfunktion der Chi-Quadrat-Verteilung (siehe Anhang C) entnommen. Die Entscheidungsbereiche für den Test sind:

- Ablehungsbereich der H_0:

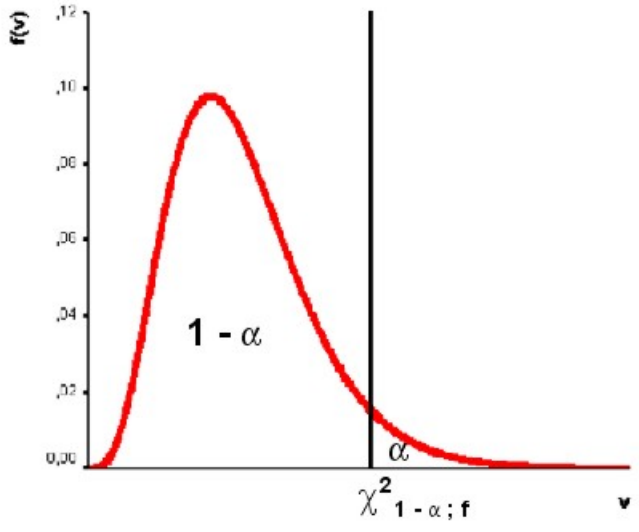
  $$\{v|v > \chi^2_{1-\alpha;DF}\}$$

- Nichtablehnungsbereich der H_0:

  $$\{v|v \leq \chi^2_{1-\alpha;DF}\}$$

Die Wahrscheinlichkeit, dass die Teststatistik V eine Realisation aus dem Nichtablehnungsbereich der H_0 annimmt, ist $P(V \leq \chi^2_{1-\alpha;DF}|H_0) = 1 - \alpha$. Die Wahrscheinlichkeit, dass die Teststatistik V eine Realisation aus dem Ablehnungsbereich der H_0 annimmt, entspricht dem vorgegebenen Signifikanzniveau $\alpha = P(V > \chi^2_{1-\alpha;DF}|H_0)$.

Abbildung 2.38.: Dichtefunktion der Chi-Quadrat-Verteilung und Entscheidungsbereiche des Chi-Quadrat-Unabhängigkeitstests



Nichtablehnungsbereich der H_0 | Ablehnungsbereich der H_0

Die Nullhypothese wird abgelehnt, falls $v > \chi^2_{1-\alpha;DF}$ ist.

Weitere Informationen zum Chi-Quadrat-Unabhängigkeitstest sind im Anhang C enthalten.

Im Fall einer 2x2 Kontingenztabelle wird oftmals eine von Yates vorgeschlagene Stetigkeitskorrektur (continuity correction) verwendet, die eine bessere Approximation an die Chi-Quadrat-Verteilung ergibt:

$$\chi^2_c = \frac{n(|h_{11}h_{22} - h_{12}h_{21}| - 0,5n)^2}{h_{1+}h_{2+}h_{+1}h_{+2}}. \tag{2.20}$$

Die Anzahl der Freiheitsgrade in einer 2x2 Kontingenztabelle ist DF = 1.

Weitere in SPSS enthaltene Tests für den Zusammenhang zweier Variablen einer Kontingenztabelle sind der Likelihood-Quotienten-Test, der Mantel-Haenszel-Test und Fisher's exakter Test.

2.2.2.2. Likelihood-Quotienten-Test (likelihood-ratio-test)

Für diesen Test[5], der auf der Maximum-Likelihood-Theorie basiert, ist die Teststatisitk definiert als:

$$\lambda = -2 \sum_{j=1}^{J} \sum_{k=1}^{K} h_{jk} \ln \left(\frac{\hat{e}_{jk}}{h_{jk}} \right), \tag{2.21}$$

worin ln der natürliche Logarithmus ist. Hintergrundinformationen zu diesem Test enthält der Anhang D.

Die Teststatistik ist approximativ χ^2-verteilt mit $DF = (J-1)(K-1)$ Freiheitsgraden. Der Likelihood-Quotienten-Test wird vor allem dann angewandt, wenn es um die Einschätzung der Güte der Anpassung eines Modells des Zusammenhangs der beiden Variablen im Sinne verallgemeinerter linearer Modelle geht (siehe die gleichnamige Veranstaltung, http://ise.wiwi.hu-berlin.de/statistik/glm_d.html).
Für große Stichproben unterscheiden sich Pearson's Chi-Quadrat-Test und Likelihood-Quotienten-Test im Ergebnis kaum.

2.2.2.3. Linear-by-Linear Association

Die Linear-by-Linear Association, auch als Mantel-Haenszel-Statistik bezeichnet, ergibt sich als

$$Q = (n-1) \cdot r^2, \tag{2.22}$$

worin r der Bravais-Pearson-Korrelationskoeffizient ist. Sie ist approximativ χ^2-verteilt mit DF = 1 Freiheitsgrad. Dieser Test sollte nur bei metrisch skalierten Variablen angewandt werden. Unter SPSS wird er jedoch immer ausgegeben, wenn der Chi-Quadrat-Test angefordert wird und beide Variablen quantitativ sind. Dabei ist zu berücksichtigen, dass auch eine nominalskalierte Variable mit zahlenmäßig kodierten Ausprägungen unter SPSS als quantitative Variable behandelt wird.

2.2.2.4. Fisher's exakter Test

Dieser Test[6] wird auf eine 2×2 Kontingenztabelle angewandt, wenn

- der Stichprobenumfang klein ($n \leq 30$) ist,
- wenigstens eine erwartete Zellhäufigkeit kleiner als 5 ist und

[5] Siehe u.a. Rönz, B., Strohe, H.G. (1994), S. 219; Büning, H., Trenkler, G. (1978), S.50; Berry, D.A., Lindgren, B.W. (1990), S. 509 ff., 579 ff., 608 f.; Hartung, J., Elpelt, B., Klösener, K.-H. (1993), S.435 ff.

[6] Siehe u.a. Büning, H., Trenkler, G. (1978), S. 246 ff.; Bosch, K. (1992), S.388 ff.; Hartung, J., Elpelt, B., Klösener, K.H. (1993), S. 414 ff.; Lienert, G.A. (1973), S. 169 ff.

- die 2x2 Kontingenztabelle eine starke Asymmetrie aufweist.

Als Teststatistik V wird die Zellhäufigkeit h_{11} verwendet. Als Überschreitungswahrscheinlichkeit P (Significance) wird die Wahrscheinlichkeit dafür berechnet, bei Gültigkeit der Nullhypothese die beobachtete 2×2 Kontingenztabelle und alle noch extremeren und damit unwahrscheinlicheren Kontingenztabellen (bei festen Randhäufigkeiten) zu erhalten. Anhang E enthält ausführliche Informationen zu diesem Test.

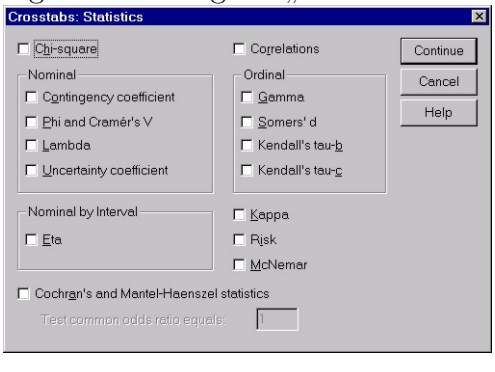
Die Testentscheidung wird im Vergleich der berechneten Überschreitungswahrscheinlichkeit P und dem vorgegebenen Signifikanzniveau α getroffen. Ist $P < \alpha$, so wird die Nullhypothese abgelehnt.

Der Name dieses Tests ergibt sich daraus, dass unter der Nullhypothese die Verteilung der Teststatistik genau bekannt ist (als eine hypergeometrische Verteilung), was beim Chi-Quadrat-Test nicht der Fall ist.

2.2.2.5. Beispiele

Unter SPSS wird die Hypothesenprüfung durchgeführt, indem im Dialogfeld „Crosstabs“ (siehe Abb. 2.35) die Schaltfläche „Statistics“ betätigt und in dem sich öffnenden Dialogfeld „Crosstabs: Statistics“ auf Chi-square entschieden wird.

Abbildung 2.39.: Dialogfeld „Crosstabs: Statistics“



Der Output für die Hypothesenprüfung enthält stets Pearson Chi-Quadrat, Likelihood Ratio und Linear-by-Linear Association. Bei Erfüllung der o.g. Bedingungen werden weiterhin ausgegeben: Continuity Correction bzw. Fisher's Exact Test. Außerdem wird die Anzahl und der Prozentsatz der Zellen mit einer erwarteten Häufigkeit kleiner als 5 sowie das Minimum der erwarteten Häufigkeiten (minimum expected count) ausgegeben.

Für die Tests (außer Fisher's Exact Test) enthält der Output den berechneten Wert der Teststatistik (Value), die Anzahl der Freiheitsgrade (df) und die Überschreitungswahrscheinlichkeit des berechneten Testwertes (Asymp. Sig.) Für Fisher's Exact Test werden nur die zwei-

und einseitigen Überschreitungswahrscheinlichkeiten (Exact Sig.) ausgegeben.

• Beispiel 2.6 (Fortsetzung):
Für die Variablen Schichteinstufung (als Spaltenvariable) und Schule (als Zeilenvariable) soll auf dem 5%-Niveau die Nullhypothese geprüft werden, dass die beiden Variablen stochastisch unabhängig sind. Die sich aufgrund der Stichprobe vom Umfang n = 2619 ergebende Kontingenztabelle wurde bereits in der Tabelle 2.2 gezeigt. Den zusätzlichen Output im SPSS Viewer für den Test enthält Tabelle 2.8.

Tabelle 2.8.: Testergebnis für die Kontingenztabelle für Schichteinstufung und Schule

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	$651,451^a$	6	,000
Likelihood Ratio	670,017	6	,000
Linear-by-Linear Association	538,841	1	,000
N of Valid Cases	2619		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 9,43.

Die bereits bei der Diskussion der gemeinsamen Häufigkeitsverteilung bzw. der bedingten Verteilungen von Schichteinstufung und Schule vermutete Assoziation beider Variablen wird durch den Test auf einem Signifikanzniveau von $\alpha = 0,05$ bestätigt, da wegen $Sig. < \alpha$ die Nullhypothese abgelehnt wird. Zu dem gleichen Testergebnis gelangt man, wenn aus der Tabelle der Chi-Quadrat-Verteilung (Anhang C) der kritische Wert für $\alpha = 0,05$ und df = 6 aufgesucht wird: $\chi^2_{0,95;6} = 12,592$. Wegen $v = 651,451 > \chi^2_{0,95;6} = 12,592$ wird die Nullhypothese auf Unabhängigkeit der Variablen Schichteinstufung und Schule abgelehnt.

Wie ersichtlich, liefern Pearson Chi-Quadrat und Likelihood Ratio aufgrund des großen Stichprobenumfanges Werte der Teststatistik, die sich nur wenig unterscheiden. Man beachte, dass Linear-by-Linear Association nicht ausgewertet werden darf, da es sich bei den beiden Variablen nicht um metrisch skalierte Variablen handelt.

Ein wichtiges Argument bei der Auswertung der Kontingenztabelle für Schichteinstufung und Schule besteht darin, dass das Geschlecht der befragten Person einen Einfluß auf die Beziehung zwischen Schichteinstufung und Schule haben könnte. Um den Einfluß des Geschlechts bei der Überprüfung des Zusammenhangs von Schichteinstufung und Schule auszuschließen, wurde die Variable Geschlecht als Kontrollvariable in das Feld „Layer 1 of 1“ gebracht. Tabelle

2.5 enthielt die resultierenden Kontingenztabellen. Getrennt für die Geschlechter soll wiederum auf dem 5%-Niveau die Nullhypothese geprüft werden, dass die beiden Variablen stochastisch unabhängig sind. Den zusätzlichen Output im SPSS Viewer für die Tests enthält Tabelle 2.9.

Tabelle 2.9.: Testergebnis für die Kontingenztabelle für Schichteinstufung und Schule, kontrolliert für die Variable Geschlecht

Chi-Square Tests

Geschlechtszugehörigkeit		Value	df	Asymp. Sig. (2-sided)
Mann	Pearson Chi-Square	$323,389^a$	6	,000
	Likelihood Ratio	330,361	6	,000
	Linear-by-Linear Association	259,534	1	,000
	N of Valid Cases	1165		
Frau	Pearson Chi-Square	$334,789^b$	6	,000
	Likelihood Ratio	354,893	6	,000
	Linear-by-Linear Association	284,593	1	,000
	N of Valid Cases	1454		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 9,43.

b. 1 cells (8,3%) have expected count less than 5. The minimum expected count is 3,96.

Auch für die Geschlechter getrennt wird die Nullhypothes auf Unabhängigkeit der Variablen Schichteinstufung und Schule auf einem Signifikanzniveau von $\alpha = 0,05$ abgelehnt. Die Auswertung des Chi-Quadrat-Tests bei den Frauen ist noch zulässig, da weniger als 20% der erwarteten Zellhäufigkeiten kleiner als 5 sind.

• Beispiel 2.7:

Für 35 zufällig ausgewählte Schüler wird erfaßt, ob sie aus Familien mit ernsthaften Familienschwierigkeiten kommen (Variable Familienschwierigkeiten mit 1 - ja und 2 - nein) und ob sie größere Lernschwierigkeiten haben (Variable Lernschwierigkeiten mit 1 - ja und 2 - nein). Auf einem Signifikanzniveau von $\alpha = 0,05$ soll geprüft werden, ob eine Assoziation zwischen den beiden Variablen besteht. Die Stichprobe (Datei: schwierigkeiten.sav) liefert nachstehende Kontingenztabelle (siehe Tabelle 2.10). Der Output im SPSS Viewer für die Tests ist in Tabelle 2.11 enthalten.

Tabelle 2.10.: Beobachtete 2x2 Kontingenztabelle für Familienschwierigkeiten und Lernschwierigkeiten

		Lernschwierigkeiten		Total
		ja	nein	
Familienschwierigkeiten	ja	2	13	15
	nein	10	10	20
Total		12	23	35

Tabelle 2.11.: Testergebnis für die Kontingenztabelle für Familienschwierigkeiten und Lernschwierigkeiten

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	$5,115^b$	1	,024		
Continuity Correction[a]	3,617	1	,057		
Likelihood Ratio	5,498	1	,019		
Fisher' Exact Test				0,034	0,026
Linear-by-Linear Association	4,969	1	,026		
N of Valid Cases	35				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 5,14.

Da hier eine 2x2 Kontingenztabelle zugrunde liegt, wird auch der Chi-Quadrat-Test mit Stetigkeitskorrektur (Continuity Correction) und Fisher's exakter Test durchgeführt. Anhand dieser 2x2 Kontingenztabelle soll die Berechnung des Wertes der Teststatistik des Chi-Quadrat-Tests ohne und mit Stetigkeitskorrektur demonstriert werden.
Pearson Chi-Square:

$$\chi^2 = \frac{\left(2-\frac{15\cdot 12}{35}\right)^2}{\frac{15\cdot 12}{35}} + \frac{\left(13-\frac{15\cdot 23}{35}\right)^2}{\frac{15\cdot 23}{35}} + \frac{\left(10-\frac{20\cdot 12}{35}\right)^2}{\frac{20\cdot 12}{35}} + \frac{\left(10-\frac{20\cdot 23}{35}\right)^2}{\frac{20\cdot 23}{35}}$$

$$= 1,9206 + 1,0021 + 1,4405 + 0,7516 = 5,1148$$

Continuity Correction:

$$\chi^2 = \frac{35(|2\cdot 10 - 13\cdot 10| - 0,5\cdot 35)^2}{15\cdot 20\cdot 12\cdot 23} = \frac{35(110-17,5)^2}{82800}$$

$$= \frac{299468,75}{82800} = 3,6168$$

Diese Werte der Teststatistik findet man im Output in der Spalte Value.

Es zeigt sich, dass auf dem 5%-Niveau mit Pearson Chi-Square die Nullhypothese abgelehnt wird, dagegen unter Verwendung der Continuity Correction der Test anzeigt, dass aufgrund der Stichprobe keine Veranlassung besteht, die Nullhypothese zu verwerfen.

Da im Falle einer 2x2 Kontingenztabelle mit Yates Stetigkeitskorrektur bei Gültigkeit der H_0 eine bessere Anpassung der Teststatistik an die Chi-Quadrat-Verteilung erreicht wird, sollte in diesem Fall die Continuity Correction bevorzugt werden. Jedoch ist zu berücksichtigen, dass es sich trotz der Stetigkeitskorrektur weiterhin um einen asymptotischen χ^2-Test handelt, d.h., die Teststatistik unter H_0 nur approximativ χ^2-verteilt ist.

Da in diesem Beispiel die Stichprobe von nur mäßigem Umfang (n = 35) ist und darüber hinaus die 2x2 Kontingenztabelle eine starke Asymmetrie aufweist, muß der asymptotische χ^2-Test durch einen exakten Test ersetzt werden. Unter diesen Gegebenheiten ist Fisher's exakter Test heranzuziehen, der bessere Güteeigenschaften aufweist. Mit ihm wird die Nullhypothese auf dem 5%-Niveau abgelehnt.

Wurde mittels der Hypothesenprüfung ein statistisch gesicherter Zusammenhang festgestellt, interessiert im weiteren die Stärke dieses Zusammenhanges. Auch dafür stellt SPSS eine Reihe von Maßzahlen zur Verfügung, für deren Berechnung man sich ebenfalls in dem Dialogfeld „Crosstabs: Statistics“ (siehe Abb. 2.39) entscheiden kann. Dabei ist jedoch sehr auf das Skalenniveau der Variablen zu achten, denn die Software berechnet im allgemeinen auch die nicht zulässigen Koeffizienten.

2.2.3. Zusammenhangsmaße

2.2.3.1. Korrelationen

Hinter correlations verbergen sich der Bravais-Pearson-Korrelationskoeffizient und der Spearman'sche Rangkorrelationskoeffizient.[7]

Der **Bravais-Pearson-Korrelationskoeffizient** (auch als Produkt-Moment-Korrelationskoeffizient bezeichnet) ist ein Maß für die lineare Beziehung zwischen zwei metrisch skalierten Variablen:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}, \tag{2.23}$$

[7] Die Korrelation wird in jedem guten Statistik-Lehrbuch behandelt. Siehe u.a. auch Rönz, B., Strohe, H.G. (1994), S. 200 ff., 303; Rönz, B., Förster, E. (1992), S. 106 ff., 308 ff.

worin die Kovarianz s_{xy} und die beiden Standardabweichungen s_x und s_y im Fall der Kontingenztabelle wie folgt berechnet werden:

$$s_{xy} = \sum_{j=1}^{J}\sum_{k=1}^{K} x_j y_k h_{jk} - \frac{1}{n}\left(\sum_{j=1}^{J} x_j h_{j+}\right)\left(\sum_{k=1}^{K} y_k h_{+k}\right) \tag{2.24}$$

$$s_x = \sqrt{\sum_{j=1}^{J} x_j^2 h_{j+} - \frac{1}{n}\left(\sum_{j=1}^{J} x_j h_{j+}\right)^2} \tag{2.25}$$

$$s_y = \sqrt{\sum_{k=1}^{K} y_k^2 h_{+k} - \frac{1}{n}\left(\sum_{k=1}^{K} y_k h_{+k}\right)^2}. \tag{2.26}$$

Durch die Norminierung der Kovarianz auf die beiden Standardabweichungen wird erreicht, dass stets $-1 \leq r_{xy} \leq +1$ gilt und der Korrelationskoeffizient eine dimensionslose Zahl ist, die bis auf das Vorzeichen invariant gegenüber linearen Skalentransformationen ist. Der Koeffizient nimmt die Werte $r_{xy} = +1$ bzw. $r_{xy} = -1$ an, wenn eine perfekte lineare Beziehung zwischen den beiden Variablen existiert. $r_{xy} = 0$ beinhaltet, dass kein linearer Zusammenhang vorliegt.

Die statistische Prüfung des Korrelationskoeffizienten gegen Null ($\rho_{xy} = 0$: kein Zusammenhang zwischen den beiden Variablen in der Grundgesamtheit) kann entweder zweiseitig mit dem Hypothesenpaar
$H_0 : \rho_{xy} = 0 \qquad H_1 : \rho_{xy} \neq 0$
oder einseitig mit den Hypothesen
a) $H_0 : \rho_{xy} \leq 0 \qquad H_1 : \rho_{xy} > 0 \qquad$ oder $\qquad$ *b)* $H_0 : \rho_{xy} \geq 0 \qquad H_1 : \rho_{xy} < 0$
vorgenommen werden.

Sind die beiden Zufallsvariablen X und Y (zumindest approximativ) normalverteilt, dann folgt die Teststatistik

$$T = \frac{R_{xy}\sqrt{n-2}}{\sqrt{1-R_{xy}^2}} \tag{2.27}$$

unter H_0 (zumindest approximativ) einer t-Verteilung mit $df = n - 2$ Freiheitsgraden. Der sich aufgrund einer konkreten Stichprobe ergebende Prüfwert t wird mit dem $(1 - \alpha/2)$ - Quantil bzw. mit dem $(1-\alpha)$ - Quantil der t-Verteilung (siehe Anhang F) für ein vorgegebenes Signifikanzniveau α verglichen.

Ist beim zweiseitigen Test $|t| > |t_{df;1-\alpha/2|}$ bzw. beim einseitigen Test für a) $t > t_{df;1-\alpha}$ oder für b) $t < t_{df;\alpha}$ so wird H_0 abgelehnt.

Bei genügend großem Stichprobenumfang n ($n > 30$) können näherungsweise die kritischen Werte aus der Standardnormalverteilung $N(0;1)$ entnommen und für die Testentscheidung verwendet werden.

Unter SPSS wird stets der zweiseitige Test ausgeführt.

Da Kontingenztabellen nicht für metrische Variablen (stetige bzw. diskrete Variablen mit sehr vielen möglichen Variablenwerten) geeignet sind, sollte dieser Korrelationskoeffizient nicht im Zusammenhang mit der Auswertung von Kontingenztabellen verwendet werden.

Der **Spearman'sche Rangkorrelationskoeffizient** ist ein Maß für den linearen Zusammenhang zweier ordinalskalierter Variablen. Dafür werden Rangwerte w_j für x_j und Rangwerte v_k für y_k verwendet, die wie folgt ermittelt werden:

$$w_j = \sum_{p<j} h_{p+} + \frac{h_{j+}+1}{2}$$
$$v_k = \sum_{q<k} h_{+q} + \frac{h_{+k}+1}{2}. \tag{2.28}$$

Diese Rangwerte werden in den Formeln (2.24) bis (2.26) für x_j und y_k substituiert und anschließend in der Formel des Bravais-Pearson-Korrelationskoeffizienten nach (2.23) verwendet. Dabei wird unterstellt, dass die Abstände zwischen den vergebenen Rangwerten als gleich groß (äquidistant) angesehen und somit die Rangwerte metrisch interpretiert und behandelt werden können.

Die Vorgehensweise des Spearman'schen Rangkorrelationskoeffizienten erweist sich zumindest dann als vorteilhaft, wenn die Verteilung der Ausgangsvariablen X und Y deutlich von der Normalverteilung abweicht und damit der mit dem Bravais-Pearson-Korrelationskoeffizient verbundene Unabhängigkeitstest nicht angewandt werden kann.

2.2.3.2. Assoziationsmaße für nominalskalierte Variablen

Die in dem Feld Nominal des Dialogfeldes „Crosstabs: Statistics" (siehe Abb. 2.39) wählbaren Koeffizienten sind Maße[8] für die Stärke der Beziehung (Assoziation) zwischen zwei nominalskalierten Variablen. Mit ihnen können jedoch keine Aussagen über Richtung und Art der Beziehung getroffen werden, was in den Eigenschaften der Nominalskala begründet liegt. Es werden zwei Typen von Maßen unterschieden: die auf der χ^2-Statistik gemäß (2.19) basierenden Maße und sogenannte PRE-Maße.

Der **Kontingenzkoeffizient C** nach Pearson (Contingency coefficient) ist ein Maß für die Stärke des Zusammenhanges zwischen zwei nominalskalierten Variablen und basiert auf der χ^2-Statistik:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \tag{2.29}$$

[8] Siehe u.a. Hartung, J., Elpelt, B., Klösener, K.-H. (1993), S. 439 ff.; Bortz, J., Lienert, G.A., Boehnke, K. (1990), S. 325 ff.

worin χ^2 entsprechend (2.19) berechnet wird und n der Stichprobenumfang ist. Es gilt $0 \leq C < 1$. Bei C = 0 liegt statistische Unabhängigkeit vor. C hat den Nachteil, von der Form und Größe der Tabelle, d.h. von der Anzahl der Zeilen und Spalten, beeinflußt zu sein. Er erreicht den Wert 1 praktisch nie, auch wenn ein völliger Zusammenhang zwischen den beiden Variablen besteht, da n als Stichprobenumfang immer größer Null ist. Er sollte deswegen nur zum Vergleich von Tabellen gleicher Größe und Form verwendet werden.

Phi (Φ) ist geeignet, die Stärke des Zusammenhanges zwischen zwei nominalskalierten, dichotomen Variablen einer 2×2 Kontingenztabelle zu messen, und basiert ebenfalls auf der χ^2-Statistik:

$$\Phi = \sqrt{\frac{\chi^2}{n}} \tag{2.30}$$

Es gilt $0 \leq \Phi \leq 1$, d.h., für $\Phi = 0$ ist statistische Unabhängigkeit, bei $\Phi = 1$ ein totaler Zusammenhang zwischen den betrachteten dichotomen Variablen gegeben. Phi ist wegen seiner Eigenschaft, bei größeren Tabellen unter Umständen Werte > 1 anzunehmen, nur für Vierfeldertabellen geeignet und ist bei größeren Tabellen durch Cramer's V zu ersetzen.

Cramer's V ist ein Maß für die Stärke des Zusammenhanges zwischen zwei Variablen mit nominalen Skalenniveau. Der Koeffizient basiert ebenfalls auf der χ^2-Statistik:

$$V = \sqrt{\frac{\chi^2}{n \cdot (c-1)}} \tag{2.31}$$

mit c als Minimum der Zeilen- bzw. Spaltenanzahl: $c = min(J; K)$. Es gilt $0 \leq V \leq 1$. Cramer's V hängt nicht von der Form und Größe einer Kontingenztabelle ab und ist als eine Erweiterung des Phi-Wertes von 2x2 Tabellen auf größere Tabellen anzusehen. Wenn entweder die Anzahl der Zeilen oder die Anzahl der Spalten gleich 2 ist, stimmen V und Φ überein.

• Beispiel 2.6 (Fortsetzung:
Da mit Pearsons Chi-Quadrat-Test bereits festgestellt wurde, dass eine Assoziation zwischen den Variablen subjektive Schichteinstufung und Schule besteht, soll nunmehr deren Stärke gemessen werden. Dafür eignen sich der Kontingenzkoeffizient und Cramer's V. Wie bereits vorher gezeigt (siehe Tabelle 2.8), ist für diese beiden Variablen $\chi^2 = 651,451$. Somit ergibt sich:

$$C = \sqrt{\frac{651,451}{651,451 + 2619}} = 0,4463$$

$$V = \sqrt{\frac{651,451}{2619(3-1)}} = 0,3527.$$

Will man dieses Ergebnis unter SPSS erhalten, so muß man sich im Dialogfeld „Crosstabs: Statistics“ (siehe Abb. 2.39) für diese beiden Koeffizienten entscheiden. Wie zu ersehen ist, werden PHI und Cramer's V stets zusammen ausgegeben, obwohl es wegen der Anwendbarkeit von PHI keinen Sinn macht. Als Output erhält man:

Tabelle 2.12.: Assoziationsmaße für Schichteinstufung und Schule

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,499	,000
	Cramer's V	,353	,000
	Contingency Coefficient	,446	,000
N of Valid Cases		2619	

a. Not assuming the null hypothesis

b. Using the asymptotic standard error assuming the null hypothesis.

Cramer's V ist aufgrund der Nachteile des Kontingenzkoeffizienten für die Interpretation vorzuziehen. Es besteht somit eine mittelstarke Assoziation zwischen beiden Variablen, die auf dem 5%-Niveau statistisch gesichert ist. Die Signifikanz dieses Maßes wird approximativ mittels Pearsons Chi-Quadrat-Statistik geprüft und zeigt hier das bereits bekannte Resultat, dass eine gesicherte Beziehung zum 5% Signifikanzniveau vorliegt.

Nachfolgend sind noch die Assoziationsmaße für den Fall angegeben, dass für die Variable Geschlecht kontrolliert wird (die zugehörigen Kontingenztabellen enthält Tabelle 2.5 und die Tests Tabelle 2.9).

Tabelle 2.13.: Assoziationsmaße für Schichteinstufung und Schule, kontrolliert für die Variable Geschlecht

Symmetric Measures

			Value	Approx. Sig.
Mann	Nominal by Nominal	Phi	,527	,000
		Cramer's V	,373	,000
		Contingency Coefficient	,446	,000
	N of Valid Cases		1165	
Frau	Nominal by Nominal	Phi	,480	,000
		Cramer's V	,339	,000
		Contingency Coefficient	,433	,000
	N of Valid Cases		1454	

a. Not assuming the null hypothesis

b. Using the asymptotic standard error assuming the null hypothesis.

Die Stärke der Assoziation von Schichteinstufung und Schule ist bei beiden Geschlechtern so

gut wie gleich.

PRE-Maße[9]

Allgemeine Betrachtungen zu den PRE-Maßen

PRE-Maße bringen nicht direkt die Stärke des Zusammenhangs zwischen zwei Variablen zum Ausdruck. Sie beinhalten die proportionale Reduktion des Fehlers bei der Vorhersage (Proportional Reduction in Error), d.h., sie geben an, um wieviel der Fehler bei der Vorhersage der Werte der einen Variablen vermindert werden kann, wenn auch die Verteilung der anderen Variablen bekannt ist. Die PRE-Maße beruhen deshalb auf der Ermittlung zweier Fehler, die bei der Vorhersage begangen werden können: ein erster Fehler, der sich bei der Vorhersage nur aufgrund der Kenntnis der Verteilung der einen Variablen ergibt, und ein zweiter Fehler, der sich bei der Vorhersage aufgrund der Kenntnis auch der Verteilung der anderen Variablen und somit der gemeinsamen Verteilung ergibt.

Die PRE-Maße nehmen Werte zwischen 0 bei keiner Fehlerreduktion durch Kenntnis der anderen Variablen, und 1 bei 100%iger Sicherheit der Vorhersage der einen Variablen unter Einbeziehung der Kenntnis der anderen Variablen an.

Man unterscheidet bei einigen dieser PRE-Maße zwischen symmetrischen und asymmetrischen Maßen.

Symmetrische Maße: Bei den symmetrischen Maßen bleibt offen, welches die abhängige und welches die unabhängige Variable ist.

Asymmetrische Maße: Eine der beiden Variablen wird als abhängige Variable betrachtet. Als abhängige Variable wird im allgemeinen diejenige Variable angesehen, deren Vorhersage von besonderem Interesse ist.

Der Vorteil der PRE-Maße besteht darin, dass sie alle gleich interpretierbar sind, obwohl sie jeweils ein anderes Skalenniveau der involvierten Variablen voraussetzen. Die konkrete Berechnungsweise eines PRE-Maßes hängt jedoch von dem Skalenniveau der zu untersuchenden Variablen ab.

LAMDA, Goodman und Kruskal's Tau und der Unsicherheitskoeffizient werden für nominalskalierte Variablen, GAMMA, Sommers d, Kendalls τ_b und τ_c für ordinalskalierte und der Determinationskoeffizient R^2 für metrische Variablen als PRE-Maße verwendet. ETA wird eingesetzt, wenn die unabhängige Variable nominal- oder ordinalskaliert und die abhängige Variable metrisch skaliert vorliegt.

PRE-Maße für nominalskalierte Variablen

Es kann durchaus auftreten, dass die Koeffizienten C, Phi bzw. Cramer's V einen starken Zusam-

[9] siehe u.a. Wittenberg, R. (1991), S. 155 ff.; Kähler, W.-M. (1994), S. 163 ff.

menhang zwischen zwei nominalskalierten Variablen anzeigen, die PRE-Maße jedoch einen Wert nahe Null aufweisen. Dies ist auf die völlig unterschiedliche Interpretation von χ^2-bezogenen und PRE-Maßen zurückzuführen.

LAMBDA (Goodman and Kruskal's lambda)

Lambda ist ein PRE-Maß für nominalskalierte Variablen. Für Lambda werden zwei asymmetrische Werte und ein symmetrischer Wert berechnet. Dazu muß die zweidimensionale Kontingenztabelle bekannt sein, von deren allgemeinen Darstellung in Tabelle 2.1 ausgegangen wird. Es bezeichnen im weiteren:
h_{jm} - die maximale Zellhäufigkeit in der Zeile j $(j = 1, \ldots, J)$,
h_{mk} - die maximale Zellhäufigkeit in der Spalte k $(k = 1, \ldots, K)$,
$h_{j+,m}$ - die maximale Randhäufigkeit der Variablen X (d.h. der Zeilen),
$f_{j+,m}$ - die maximale relative Randhäufigkeit der Variablen X,
$h_{+k,m}$ - die maximale Randhäufigkeit der Variablen Y (d.h. der Spalten),
$f_{+k,m}$ - die maximale relative Randhäufigkeit der Variablen Y,
wobei m für maximal steht.
a) Y als abhängige Variable (asymmetrisches Lambda-Maß)
Es wird zunächst Y als die abhängige Variable und X als unabhängige Variable betrachtet, d.h., es soll eine Ausprägung der Variablen Y vorhergesagt werden. Wird diese Vorhersage nur auf der Basis der Verteilung der abhängigen Variablen Y erstellt, so ist die beste Schätzung diejenige Kategorie mit der maximalen Randhäufigkeit der Variablen Y (Modal-Kategorie), da die Wahrscheinlichkeit für einen Vorhersagefehler dafür am geringsten ist. Die Wahrscheinlichkeit für einen Vorhersagefehler ist:

$$P_1 = 1 - f_{+k,m}.$$

Das erste Fehlermaß ist entsprechend definiert als

$$E_1 = n - h_{+k,m},$$

d.h., es wird die Anzahl der Fälle ermittelt, die eine von der modalen Y-Kategorie abweichende Ausprägung bei Y aufweisen.

Für die Ermittlung des 2. Fehlers wird auch die Kenntnis der Verteilung der unabhängigen Variablen X und damit der gemeinsamen Häufigkeitsverteilung beider Variablen vorausgesetzt und für die Vorhersage berücksichtigt. In diesem Fall wird für jede Kategorie von X (d.h. für jede Zeile) diejenige Kategorie von Y vorhergesagt, die die größte (modale) Zellhäufigkeit aufweist. Die Wahrscheinlichkeit für einen Fehler bei dieser Vorhersage ist
P_2 = Summe der relativen Zellhäufigkeiten für alle Zellen, die keine Zeilen-Modi sind.

Das zweite Fehlermaß ergibt sich somit zu

$$E_2 = \sum_{j=1}^{J}(h_{j+} - h_{jm}). \tag{2.32}$$

E_2 ist immer kleiner gleich E_1. Bezieht man die Differenz $E_1 - E_2$ auf E_1, so erhält man den Wert des asymmetrischen Lambda-Maßes $\lambda_{Y|X}$ zur Vorhersage von Y bei Kenntnis der Kategorien von X:

$$\lambda_{Y|X} = \frac{E_1 - E_2}{E_1} = \frac{P_1 - P_2}{P_1}. \tag{2.33}$$

Setzt man die vereinbarten Symbole ein, so folgt

$$\lambda_{Y|X} = \frac{\sum_{j=1}^{J} h_{jm} - h_{+k;m}}{n - h_{+k;m}}. \tag{2.34}$$

Der Wert von LAMBDA gibt an, um wieviel sich der Fehler bei der Vorhersage der abhängigen Variablen reduzieren läßt, wenn man die Verteilung der unabhängigen Variablen kennt. Lambda kann Werte im Bereich $0 \leq \lambda \leq 1$ annehmen. Der Wert 0 wird angenommen, wenn die zusätzliche Information über die unabhängige Variable nicht zur Reduktion des Vorhersagefehlers beiträgt. In diesem Sinne hat die unabhängige Variable keinen Einfluß auf die abhängige Variable. Ein Lambda-Wert von 1 bedeutet, dass mit der Kenntnis der unabhängigen Variablen die Kategorien der abhängigen Variablen exakt vorhergesagt werden können.

b) X als abhängige Variable (asymmetrisches Lambda-Maß)

Mit den gleichen Überlegungen erhält man das zweite asymmetrische Lambda-Maß, wenn X die abhängige und Y die unabhängige Variable ist, d.h., wenn eine Ausprägung der Variablen X vorhergesagt werden soll. Die entsprechenden Formeln sind:

$$P_1 = 1 - f_{j+;m};$$
$$E_1 = n - h_{j+;m};$$

P_2 = Summe der relativen Zellhäufigkeiten für alle Zellen, die keine Spalten-Modi sind;

$$E_2 = \sum_{k=1}^{K}(h_{+k} - h_{mk}); \tag{2.35}$$

$\lambda_{X|Y}$ wird analog zu (2.33) berechnet oder nach Einsetzen von E_1 und E_2 als

$$\lambda_{X|Y} = \frac{\sum_{k=1}^{K} h_{mk} - h_{j+;m}}{n - h_{j+;m}}. \tag{2.36}$$

c) symmetrisches Lambda-Maß

Ein symmetrisches Lambda-Maß wird ermittelt, wenn nicht festgelegt ist, welche Variable als unabhängige und welche als abhängige zu betrachten ist. Dabei wird gleichzeitig für die Zeilen- und für die Spaltenvariable eine Ausprägung vorhergesagt. Die beiden Fehlermaße ergeben sich somit zu:

$$E_1 = (n - h_{+k;m}) + (n - h_{j+;m}) \tag{2.37}$$

$$E_2 = \sum_{j=1}^{J}(h_{j+} - h_{jm}) + \sum_{k=1}^{K}(h_{+k} - h_{mk}). \tag{2.38}$$

Bezeichnet man mit λ das symmetrische Maß, so ergibt es sich entweder analog zu (2.33) oder unter Verwendung der Definitionen von E_1 und E_2 als

$$\lambda = \frac{\sum_{j=1}^{J} h_{jm} + \sum_{k=1}^{K} h_{mk} - h_{+k;m} - h_{j+;m}}{2n - h_{+k;m} - h_{j+;m}}. \tag{2.39}$$

• Beispiel 2.6 (Fortsetzung):

a) Y als abhängige Variable (asymmetrisches Lambda-Maß)

Es soll die Frage beantwortet werden, ob die Variable Schule (X) im Sinne der proportionalen Fehlerreduktion einen Einfluß auf die Variable Schichteinstufung (Y) ausübt. Die dazugehörige Kontingenztabelle (Tabelle 2.2) mit den benötigten Informationen sei hier nochmals angegeben.

Tabelle 2.2: Kontingenztabelle für Schichteinstufung (SE) und Schule

Schule * Schichteinstufung Crosstabulation

			SE			
			Arbeiter	MS	Ob.MS	Total
Schule	HSA	Count	748	694	42	1484
		% of Total	28,6%	26,5%	1,6%	56,7%
	Mittlere Reife	Count	87	474	66	627
		% of Total	3,3%	18,1%	2,5%	23,9%
	FSA	Count	10	67	21	98
		% of Total	,4%	2,6 %	,8%	3,7%
	Abitur	Count	21	266	123	410
		% of Total	,8%	10,2%	4,7%	15,7%
Total		Count	866	1501	252	2619
		% of Total	33,1%	57,3%	9,6%	100,0%

Ohne Kenntnis der Variablen Schule würde die Kategorie Mittelschicht der Variablen Schichteinstufung vorhergesagt, da diese Kategorie mit $h_{+k;m} = 1501$ die größte Randhäufigkeit aufweist. Daraus folgt

$$E_1 = 2619 - 1501 = 1118$$

und als Wahrscheinlichkeit für diesen ersten Fehler

$$P_1 = E_1/n = 0,4269.$$

Bei Kenntnis der Variablen Schule und der gemeinsamen Verteilung mit der Variablen Schichteinstufung wird „Arbeiter“ für diejenigen mit Hauptschulabschluß vorhergesagt, da diese Kategorie der Variablen Schichteinstufung in der Zeile Hauptschulabschluß (HSA) die größte Zellhäufigkeit aufweist. Für diejenigen mit Mittlerer Reife ist die Vorhersage „Mittelschicht“, da in der Zeile Mittlere Reife diese Kategorie der Variablen Schichteinstufung die größte Zellhäufigkeit hat. „Mittelschicht“ wird ebenfalls für diejenigen mit Fachschulabschluß (FSA) und auch diejenigen mit Abitur vorhergesagt, da diese Kategorie die modale Zellhäufigkeit in der jeweiligen Zeile aufweist. Das Fehlermaß E_2 beträgt

$$E_2 = (1484 - 748) + (627 - 474) + (98 - 67) + (410 - 266) = 1064$$

und somit die Wahrscheinlichkeit für einen Vorhersagefehler bei Kenntnis der Variablen Schule

$$P_2 = E_2/n = 0,4063.$$

Entsprechend (2.33) folgt:

$$\lambda_{Y|X} = (1118 - 1064)/1118 = 0,0483.$$

Wenn zur Vorhersage der Schichteinstufung auch die Verteilung von Schule bekannt ist, kann eine Fehlerreduktion von 4,83% erreicht werden. In diesem Sinne der Fehlerreduktion hat die Schule einen gewissen Einfluß auf die Variable Schichteinstufung.

b) X als abhängige Variable (asymmetrisches Lambda-Maß)

Es soll die Frage beantwortet werden, ob die Variable Schichteinstufung (Y) im Sinne der proportionalen Fehlerreduktion einen Einfluß auf die Variable Schule (X) ausübt. Ohne Kenntnis der Variablen Schichteinstufung würde die Kategorie Hauptschulabschluß der Variablen Schule vorhergesagt, da diese Kategorie mit $h_{j+;m} = 1484$ die größte Randhäufigkeit aufweist. Daraus folgt

$$E_1 = 2619 - 1484 = 1135$$

und als Wahrscheinlichkeit für diesen ersten Fehler

$$P_1 = E_1/n = 0,4334.$$

Bei zusätzlicher Kenntnis der Variablen Schichteinstufung und der gemeinsamen Verteilung mit der Variablen Schule wird Hauptschulabschluß für diejenigen vorhergesagt, die sich als Arbeiter einstuften, da in der Spalte Arbeiter diese Kategorie der Variablen Schule die größte Zellhäufigkeit aufweist. Hauptschulabschluß wird ebenfalls für die Schichteinstufung Mittelschicht vorhergesagt, da in der Spalte Mittelschicht diese Kategorie der Variablen Schule die größte Zellhäufigkeit hat. Für die Kategorie Obere Mittelschicht der Variablen Schichteinstufung wird Abitur vorhergesagt. Das Fehlermaß E_2 beträgt

$$E_2 = (866 - 748) + (1501 - 694) + (252 - 123) = 1054$$

und somit die Wahrscheinlichkeit für einen Vorhersagefehler bei zusätzlicher Kenntnis der Variablen Schichteinstufung

$$P_2 = E_2/n = 0,4024.$$

Entsprechend (2.33) folgt:

$$\lambda_{X|Y} = (1135 - 1054)/1135 = 0,0714$$

Wenn zur Vorhersage einer Kategorie der Variablen Schule auch die Verteilung von Schichteinstufung bekannt ist, kann eine Fehlerreduktion von 7,14% erreicht werden. In diesem Sinne der Fehlerreduktion hat die Variable Schichteinstufung einen gewissen Einfluß auf die Variable Schule.

c) symmetrisches Lambda-Maß

Wird bei den Variablen Schichteinstufung und Schule keine als abhängige bzw. unabhängige Variable spezifiziert, so folgt:

$$\begin{aligned} E_1 &= (2619 - 1501) + (2619 - 1484) = 2253 \\ E_2 &= 1064 + 1054 = 2118 \\ \lambda &= (2253 - 2118)/2253 = 0,0599. \end{aligned}$$

Bei Kenntnis der gemeinsamen Verteilung von Schichteinstufung und Schule kann eine Fehlerreduktion bei der Vorhersage von 5,99% erreicht werden. Die Variablen sind im Sinne der Fehlerreduktion voneinander abhängig.

Diese Ergebnisse zeigt der SPSS-Output in folgender Weise (obere Teil der Tabelle 2.14), wenn im Dialogfeld „Crosstabs: Statistics“ (siehe Abb. 2.39) Lambda gewählt wurde.

Tabelle 2.14.: Assoziationsmaße für Schichteinstufung und Schule

Directional Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	,060	,017	3,375	,001
		Schule Dependent	,071	,011	6,354	,000
		Schichteinstufung Dependent	,048	,033	1,423	,155
	Goodman and Kruskal tau	Schule Dependent	,130	,008		,000[c]
		Schichteinstufung Dependent	,118	,009		,000[c]

a. Not assuming the null hypothesis

b. Using the asymptotic standard error assuming the null hypothesis

c. Based on chi-square approximation

Neben dem Wert (Value) für Lambda wird ein asymptotischer Standardfehler unter der Alternativhypothese ausgegeben. Mittels eines approximativen t-Tests (unter Verwendung eines asymptotischen Standardfehlers unter der Nullhypothese) kann die Nullhypothese $H_0 : \lambda = 0$ überprüft werden. Dieser Test zeigt an, dass die Fehlerreduktion von 0,0599 im Falle der symmetrischen Betrachtung und die Fehlerreduktion von 0,0714 bei der Vorhersage der Variablen Schule unter Kenntnis der Verteilung von Schichteinstufung zum 5%-Niveau signifikant sind, während bei der Vorhersage der Schichteinstufung unter Kenntnis der Verteilung der Variablen Schule keine signifikante Fehlerreduktion erreicht werden kann.

Goodman und Kruskal's Tau

Hierbei handelt es sich ebenfalls um ein PRE-Maß, bei dem wiederum zwei Fehlermaße zugrunde gelegt werden. Für Goodman and Kruskal's Tau werden nur symmetrische Werte berechnet.

a) Y als abhängige Variable

Es wird Y als die abhängige Variable und X als die unabhängige Variable betrachtet. Wird eine Vorhersage nur aufgrund der Kenntnis der Verteilung der abhängigen Variablen Y erstellt, so begeht man einen ersten Fehler, der sich auf Basis der gesamten Randverteilung der abhängigen (anstatt des Modus bei Lambda) ergibt.

Die Anzahl der korrekt vorhergesagten Fälle für die k-te Kategorie von Y ergibt sich unter Verwendung der relativen Randhäufigkeiten f_{+k} von Y als: $f_{+k} \cdot h_{+k}$ $(k = 1, \ldots, K)$. Damit

resultieren insgesamt korrekt vorhergesagte Fälle für die abhängige Variable Y:

$$K_1 = \sum_{k=1}^{K} f_{+k} \cdot h_{+k} = \frac{\sum_{k=1}^{K} h_{+k}^2}{n}. \tag{2.40}$$

Die Anzahl der inkorrekt vorhergesagten Fälle für die abhängige Variable Y und damit die Größe des ersten Fehlers ist die Differenz zum Stichprobenumfang:

$$E_1 = n - \frac{\sum_{k=1}^{K} h_{+k}^2}{n} = \frac{n^2 - \sum_{k=1}^{K} h_{+k}^2}{n}. \tag{2.41}$$

Können zusätzliche Informationen über die unabhängige Variable X für die Vorhersage von Y einbezogen werden, dann basiert die Vorhersage auf den bedingten relativen Zellhäufigkeiten $f(y_k|x_j)$ (gemäß (2.12)). Gegeben die j-te Kategorie der Variablen X würden $f(y_k|x_j) \cdot h_{jk}$ Fälle für die k-te Kategorie von Y korrekt vorhergesagt. Die Gesamtzahl der korrekt vorhergesagten Fälle ist in diesem Fall:

$$K_2 = \sum_{j=1}^{J} \sum_{k=1}^{K} f(y_k|x_j) \cdot h_{jk} = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{h_{jk}^2}{h_{j+}}. \tag{2.42}$$

Die Gesamtzahl der inkorrekt vorhergesagten Fälle für die abhängige Variable Y und damit die Größe des zweiten Fehlers ist die Differenz zum Stichprobenumfang:

$$E_2 = n - \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{h_{jk}^2}{h_{j+}}. \tag{2.43}$$

Goodman und Kruskal's Tau wird berechnet, indem die Differenz der beiden Fehler relativ zur Größe des ersten Fehlers ermittelt wird:

$$\tau_{Y|X} = \frac{E_1 - E_2}{E_1} \tag{2.44}$$

bzw. nach Einsetzen von E_1 und E_2

$$\tau_{Y|X} = \frac{n \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{h_{jk}^2}{h_{j+}} - \sum_{k=1}^{K} h_{+k}^2}{n^2 - \sum_{k=1}^{K} h_{+k}^2}. \tag{2.45}$$

b) X als abhängige Variable

Es wird X als die abhängige Variable und Y als die unabhängige Variable betrachtet. Es resultieren analoge Formeln.

Anzahl der korrekt vorhergesagten Fälle für die abhängige Variable X nur aufgrund der Kenntnis der Verteilung von X:

$$K_1 = \sum_{j=1}^{J} f_{j+} \cdot h_{j+} = \frac{\sum_{j=1}^{J} h_{j+}^2}{n}, \tag{2.46}$$

Anzahl der inkorrekt vorhergesagten Fälle für die abhängige Variable X nur aufgrund der Kenntnis der Verteilung von X und damit die Größe des ersten Fehlers:

$$E_1 = n - \frac{\sum_{j=1}^{J} h_{j+}^2}{n} = \frac{n^2 - \sum_{j=1}^{J} h_{j+}^2}{n}, \tag{2.47}$$

Anzahl der korrekt vorhergesagten Fälle von X bei Kenntnis der Verteilung von Y:

$$K_2 = \sum_{j=1}^{J} \sum_{k=1}^{K} f(x_j|y_k) \cdot h_{jk} = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{h_{jk}^2}{h_{+k}}, \tag{2.48}$$

Anzahl der inkorrekt vorhergesagten Fälle von X bei Kenntnis der Verteilung von Y, d.h. die Größe des zweiten Fehlers:

$$E_2 = n - \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{h_{jk}^2}{h_{+k}}, \tag{2.49}$$

Goodman und Kruskal's Tau:

$$\tau_{X|Y} = \frac{E_1 - E_2}{E_1} = \frac{n \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{h_{jk}^2}{h_{+k}} - \sum_{j=1}^{J} h_{j+}^2}{n^2 - \sum_{j=1}^{J} h_{j+}^2}. \tag{2.50}$$

• Beispiel 2.6 (Fortsetzung):

Die Schichteinstufung (Y) sei die abhängige und Schule (X) die unabhängige Variable. Für diesen Fall ist

$$\begin{aligned}
K_1 &= (866^2 + 1501^2 + 252^2)/2619 = 1170,852 \\
E_1 &= (2619 - 1170,852 = 1448,148 \\
K_2 &= (748^2 + 694^2 + 42^2)/1484 + (87^2 + 474^2 + 66^2)/627 + (10^2 + 67^2 + 21^2)/98 \\
&\quad + (21^2 + 266^2 + 123^2)/410 \\
&= 702,7655 + 377,3541 + 51,3265 + 210,5512 = 1341,9973 \\
E_2 &= 2619 - 1341,9973 = 1277,003 \\
\tau_{Y|X} &= (1448,148 - 1277,003)/1448,148 = 0,1182.
\end{aligned}$$

Der Prozentsatz der falschen Vorhersage bei der Variablen Schichteinstufung kann durch Kenntnis der unabhängigen Variablen Schule um 11,82% reduziert werden. Analog ergibt sich für den Fall, dass Schule (X) die abhängige und Schichteinstufung (Y) die unabhängige Variable ist:

$$\tau_{X|Y} = 0,13.$$

Der Prozentsatz falscher Vorhersagen bei der Variablen Schule kann durch Kenntnis der unabhängigen Variablen Schichteinstufung um 13% reduziert werden.

Goodman und Kruskal's Tau wird im SPSS-Output zusammen mit Lambda ausgegeben (siehe Tabelle 2.14). Neben dem Wert für Tau wird ein asymptotischer Standardfehler unter der Alternativhypothese H_1 angegeben. Ein Test der Nullhypothese, dass $\tau = 0$ ist, basiert auf der Teststatistik $(n-1)(K-1)\tau_{Y|X}$ bzw. $(n-1)(J-1)\tau_{X|Y}$, die unter H_0 approximativ einer Chi-Quadrat-Verteilung mit $(J-1)(K-1)$ Freiheitsgraden folgt.

Unsicherheitskoeffizient

Ein weiteres PRE-Maß ist der Unsicherheitskoeffizient, der auf der Entropie beruht und für den ein symmetrischer Wert und zwei asymmetrische Werte ausgegeben werden. Die Entropie eines Versuches[10] als Maßzahl der Unbestimmtheit eines Versuches ist wie folgt definiert:

$$H = -\sum_{i=1}^{N} P(A_i) \ln P(A_i), \tag{2.51}$$

worin $P(A_i)$ die Wahrscheinlichkeit des Auftretens des Ereignisses A_i ist. Jede Information, die man über das Auftreten der Ereignisse A_i erlangt, kann als Beseitigung von Unbestimmtheit interpretiert werden, was zur Verringerung des Betrages an Entropie führt. Da im Kontext der Kontingenztabelle die Wahrscheinlichkeiten unbekannt sind, werden sie durch die beobachteten relativen Häufigkeiten ersetzt. Somit erhält man für die Variable X

$$U_X = -\sum_{j=1}^{J} f_{j+} \ln f_{j+}, \tag{2.52}$$

für die Variable Y

$$U_Y = -\sum_{k=1}^{K} f_{+k} \ln f_{+k} \tag{2.53}$$

und für beide Variablen X und Y

$$U_{XY} = -\sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk} \ln f_{jk} \tag{2.54}$$

[10] Vgl. Rönz, B., Strohe, H.G. (1994), S. 106 f.

Der Unsicherheitskoeffizient $U_{Y|X}$, der die proportionale Reduktion in der Unsicherheit der abhängigen Variablen Y durch Kenntnis der unabhängigen Variablen X beinhaltet, ist definiert als

$$U_{Y|X} = \frac{U_X + U_Y + U_{XY}}{U_Y}. \tag{2.55}$$

Entsprechend ist der Unsicherheitskoeffizient $U_{X|Y}$, der die proportionale Reduktion in der Unsicherheit der abhängigen Variablen X durch Kenntnis der unabhängigen Variablen Y beinhaltet, gegeben mit

$$U_{X|Y} = \frac{U_X + U_Y + U_{XY}}{U_X}. \tag{2.56}$$

Der symmetrische Unsicherheitskoeffizient lautet

$$U = 2\left(\frac{U_X + U_Y - U_{XY}}{U_Y + U_X}\right). \tag{2.57}$$

Der Unsicherheitskoeffizient nimmt Werte im Bereich $0 \leq U \leq 1$ an. Je mehr U gegen Eins geht, umso mehr Informationen können über das Auftreten von Kategorien der einen Variablen durch die Kenntnis der Verteilung der anderen Variablen gewonnen werden.

Unter SPSS erhält man den Unsicherheitskoeffizienten, wenn im Dialogfeld „Crosstabs: Statistics“ (siehe Abb. 2.39) Uncertainty coefficient gewählt wurde.

• Beispiel 2.6 (Fortsetzung):
Für die beiden Variablen Schichteinstufung und Schule ergeben sich die in Tabelle 2.15 enthaltenen Unsicherheitskoeffizienten.

Tabelle 2.15.: Assoziationsmaße für Schichteinstufung und Schule

Directional Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Uncertainty coefficient	Symmetric	,129	,009	14,288	,000[c]
		Schule Dependent	,119	,008	14,288	,000[c]
		Schichteinstufung Dependent	,141	,009	14,288	,000[c]

a. Not assuming the null hypothesis

b. Using the asymptotic standard error assuming the null hypothesis

c. Likelihood ration chi-square probability

Auch für den Unsicherheitskoeffizienten wird neben seinem Wert (Value) ein asymptotischer

Standardfehler unter der Alternativhypothese ausgegeben. Mittels eines approximativen t-Tests (unter Verwendung eines asymptotischen Standardfehlrs unter der Nullhypothese) kann die Nullhypothese überprüft werden, dass der Unsicherheitskoeffizient Null ist. Dieser Test zeigt an, dass durch Kenntnis der anderen Variablen eine signifikante Reduktion in der Unsicherheit der vorhergesagten Variablen erreicht werden kann.

2.2.3.3. Zusammenhangsmaße für ordinalskalierte Daten

Die in dem Feld Ordinal des Dialogfeldes „Crosstabs: Statistics“ (siehe Abb. 2.39) wählbaren Koeffizienten sind Maße für die Stärke und Richtung der Beziehung zwischen zwei ordinalskalierten Variablen.

Zunächst sei angemerkt, dass der Zusammenhang ordinalskalierter Variablen auch mit den beschriebenen Maßen für nominalskalierte Variablen gemessen werden könnten. Da diese Maße jedoch die Ordnungsrelation nicht in Betracht ziehen, gehen in der Stichprobe enthaltene Informationen verloren. Es gibt weitere Zusammenhangsmaße[11], die speziell die Rangordnung der Daten berücksichtigen und damit neben einer Aussage über die Stärke des Zusammenhanges auch eine solche über die Richtung der Beziehung erlauben. Ein Maß für den linearen Zusammenhang zwischen zwei ordinalskalierten Variablen wurde bereits mit dem Spearman'schen Rangkorrelationskoeffizienten vorgestellt. Die nunmehr einzuführenden Maße basieren auf dem Vergleich der Ordnungsrelation für alle möglichen Paare von beobachteten Variablenwerten.

Gegeben seien n statistische Einheiten (Fälle), für die bezüglich zweier ordinalskalierter Variablen X und Y die Rangzahlen gegeben sind.

Als konkordant (oder gleichsinnig)[12] bezeichnet man jene Paare von statistischen Einheiten, die eine gleiche Ordnungsrelation in den Rangzahlen von X und Y aufweisen, d.h., wenn für ein Paar $(i;h)$ von statistischen Einheiten mit $i \neq h$ $(i,h = 1,\ldots,n)$ gilt:

$$\{X_i < X_h; Y_i < Y_h\} \text{ bzw. } \{X_i > X_h; Y_i > Y_h\}.$$

Die Anzahl konkordanter Paare sei mit C symbolisiert.
Als diskordant (oder gegensinning)[13] bezeichnet man jene Paare von statistischen Einheiten, die eine entgegengesetzte Ordnungsrelation in den Rangzahlen von X und Y aufweisen, d.h, wenn für ein Paar $(i;h)$ von statistischen Einheiten mit $i \neq h$ gilt:

$$\{X_i < X_h; Y_i > Y_h\} \text{ bzw. } \{X_i > X_h; Y_i < Y_h\}.$$

Die Anzahl diskordanter Paare sei mit D symbolisiert.
Als Ties (Bindungen, Verknüpfungen) bezeichnet man in der Statistik den Tatbestand, dass bei

[11] Siehe u.a. Litz, H.P. (1997), S. 137 ff.; Wittenberg, R. (1991), S. 155 ff.; Kähler, W.-M. (1994), S. 169 ff.
[12] Vgl. Rönz, B., Strohe, H.G. (1994), S.190
[13] Vgl. Rönz, B., Strohe, H.G. (1994), S. 85 f.

mindestens ordinalskalierten Variablen mehrere statistische Einheiten (Fälle) die gleichen Ausprägungen aufweisen und somit keine reinen Rangreihen gebildet werden können. Es gibt dann mehr Fälle als Ausprägungen der ordinalskalierten Variablen. Ties können in drei verschiedenen Varianten auftreten:

- Ties in der Variablen X, jedoch nicht in der Varibalen Y: $\{X_i = X_h; Y_i < Y_h\}$.
- Ties in der Variablen Y, jedoch nicht in der Variablen X: $\{X_i < X_h; Y_i = Y_h\}$.
- Ties in X und Y: $\{X_i = X_h; Y_i = Y_h\}$.

Die entsprechenden Anzahlen seien mit T_X, T_Y und T_{XY} symbolisiert.
Die Anzahl aller Paare ergibt sich als Anzahl der Kombinationen ohne Wiederholung zur 2. Klasse:

$$\binom{n}{2} = \frac{n!}{2!(n-2)!} = \frac{n(n-1)}{2} = C + D + T_X + T_Y + T_{XY} \tag{2.58}$$

Beispiel: Zwei Gutachter X und Y prüfen 4 Bauprojekte nach bestimmten Kriterien und bringen sie in eine Rangfolge:

Bauprojekt	BP_1	BP_2	BP_3	BP_4
Gutachter X	3	4	2	1
Gutachter Y	3	1	4	2

Folgende Paare von Bauprojekten weisen Konkordanz in den Rangdaten auf:
$\{(X_1 = 3) > (X_4 = 1); (Y_1 = 3) > (Y_4 = 2)\}$ und $\{(X_3 = 2) > (X_4 = 1); (Y_3 = 4) > (Y_4 = 2)\}$.
Folgende Paare von Bauprojekten weisen Diskordanz in den Rangdaten auf:
$\{(X_1 = 3) < (X_2 = 4); (Y_1 = 3) > (Y_2 = 1)\}; \{(X_1 = 3) > (X_3 = 2); (Y_1 = 3) < (Y_3 = 4)\};$
$\{(X_2 = 4) > (X_3 = 2); (Y_2 = 1) < (Y_3 = 4)\}; \{(X_2 = 4) > (X_4 = 1); (Y_2 = 1) < (Y_4 = 2)\}$.
Ties sind hier nicht aufgetreten, so dass $C = 2$, $D = 4$, $T_X = T_Y = T_{XY} = 0$ sowie $n(n-1)/2 = 4 \cdot 3/2 = 6$ sind.

Überwiegen die konkordanten Paare, so spricht man von einer positiven Assoziation, denn mit steigenden (fallenden) Rängen der einen Variablen steigen (fallen) in der Tendenz auch die Ränge der anderen Variablen. Überwiegen die diskordanten Paare, liegt eine negative Assoziation vor, denn mit steigenden Rängen der einen Variablen fallen in der Tendenz die Ränge der anderen Variablen. Ist die Anzahl der konkordanten und der diskordanten Paare gleich, ist keine Assoziation zwischen beiden Variablen gegeben.

Gamma (Goodman and Kruskal's gamma)
Goodman und Kruskal's Gamma basiert auf der Anzahl der konkordanten und diskordanten

Paare, wobei verbundene Paare ignoriert werden und keine Unterscheidung zwischen abhängiger und unabhängiger Variable vorgenommen wird. Gamma ist somit ein symmetrisches Maß für das Ausmaß und die Richtung der Beziehung zwischen ordinalskalierten Variablen. Gamma ist definiert als

$$\gamma = \frac{C - D}{C + D} \tag{2.59}$$

Es gilt $-1 \leq \gamma \leq +1$. Gamma kann auch als PRE-Maß interpretiert werden: Der absolute Wert von Gamma gibt an, um wieviel der Fehler bei der Vorhersage der einen Variablen unter Kenntnis der Verteilung der anderen Variablen vermindert werden kann. Zu berücksichtigen ist, dass bei Vorliegen von Ties (die im allgemeinen in einer Kontingenztabelle auftreten) der Wert von Gamma erhöht ausgewiesen wird.

• Beispiel 2.8:

Bei einer Befragung von Studenten wurde u.a. die soziale Lage und die Beurteilung des Studiums in die Untersuchung einbezogen. Die resultierenden Ergebnisse sind in der Datei studium.sav[14] enthalten. Die ursprüngliche Bewertung des Studiums mit 1 (sehr gut), 2 (gut), 3 (befriedigend), 4 (ausreichend), 5 (mangelhaft), 6 (ungenügend) wurde in folgender Weise verändert: 1 (sehr gut, gut), 2 (befriedigend), 3 (schlecht). Diese Veränderung ist in der Variablen stud enthalten. Ebenso wurde eine Umkodierung der Bewertung der sozialen Lage von 1 (sehr gut), 2 (gut), 3 (befriedigend), 4 (ausreichend), 5 (mangelhaft), 6 (ungenügend) in 1 (sehr gut, gut), 2 (befriedigend), 3 (schlecht) vorgenommen und in der Variablen lage abgespeichert. Der Grund für diese Umkodierung war die Voraussetzung, in der Kontingenztabelle erwartete Häufigkeiten von 5 und größer in jeder Zelle zu erhalten. Beide Variablen weisen somit ordinales Skalenniveau auf. Tabelle 2.16 gibt die Kontingenztabelle und Tabelle 2.17 die Testergebnisse wieder.

Tabelle 2.16.: Kontingenztabelle für die Beurteilung des Studiums und soziale Lage

Beurteilung des Studiums * Soziale Lage Crosstabulation

Count

		Soziale Lage			
		sehr gut, gut	befriedigend	schlecht	Total
Beurteilung des Studiums	sehr gut, gut	17	4	3	24
	befriedigend	23	16	19	58
	schlecht	2	5	18	25
Total		42	25	40	107

Wie Pearson's Chi-Quadrat-Test zeigt, besteht auf dem 5%-Signifikanzniveau ein wesentlicher

[14] Die Datei wurde der Diskette entnommen, die Bühl, A., Zöfel, P. (1994) beiliegt.

Zusammenhang zwischen beiden Variablen.

Tabelle 2.17.: Testergebnis für Beurteilung des Studiums und soziale Lage

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	25,666[a]	4	,000
Likelihood Ratio	27,241	4	,000
Linear-by-Linear Association	23,788	1	,000
N of Valid Cases	107		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 5,61.

Aus der Kontingenztabelle ergeben sich die Anzahl der konkordanten und diskordanten Paare wie folgt (zur genaueren Demonstration siehe Anhang G):
$C = 17(16 + 19 + 5 + 18) + 4(19 + 18) + 23(5 + 18) + 16(18) = 1951$
$D = 4(23 + 2) + 3(23 + 16 + 2 + 5) + 16(2) + 19(2 + 5) = 403$
Daraus ergibt sich für Gamma:
$\gamma = (1951 - 403)/(1951 + 403) = 0,6576.$
Wählt man im Dialogfeld „Crosstabs: Statistics“ unter Ordinal den Koeffizienten Gamma aus, so erhält man die Angaben in Tabelle 2.18.

Tabelle 2.18.: Gamma für Beurteilung des Studiums und soziale Lage

Symmetric Measures

	Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal Gamma	,658	,093	5,915	,000
N of Valid Cases	107			

a. Not assuming the null hypothesis

b. Using the asymptotic standard error assuming the null hypothesis

Gamma zeigt eine mittelstarke positive Assoziation zwischen der Beurteilung des Studiums und der sozialen Lage an. Wenn bekannt ist, dass für zwei Fälle (Studenten) bei der Variablen soziale Lage eine konkordante Ordnungsrelation besteht, so wird man für dieses Paar auch für die Variable Beurteilung des Studiums die gleiche Ordnungsrelation vorhersagen. Durch die Anwendung dieser Vorhersageregel auf alle nicht gebundenen Paare von Studenten kann der Vorhersagefehler um 65,8% reduziert werden, im Gegensatz zu einer Vorhersage, bei der die

Verteilung der Variablen soziale Lage nicht einbezogen wird.

Somers' D

Somers' D ist ein Maß für die Stärke und die Richtung der Beziehung zwischen ordinalskalierten Variablen, dass sowohl symmetrisch als auch asymmetrisch berechnet wird. Ein weiterer Unterschied zu Gamma besteht darin, dass Somers' D das Auftreten von Ties berücksichtigt (zur Ermittlung von Ties in Kontingenztabellen siehe Anhang G). Folgende Maße sind definiert:
a) wenn Y die abhängige und X die unabhängige Variable ist (asymmetrisches Maß):

$$D_{Y|X} = \frac{C-D}{C+D+T_Y} = \frac{2(C-D)}{n^2 - \sum_{j=1}^{J} h_{j+}^2} \tag{2.60}$$

b) wenn X die abhängige und Y die unabhängige Variable ist (asymmetrisches Maß):

$$D_{X|Y} = \frac{C-D}{C+D+T_X} = \frac{2(C-D)}{n^2 - \sum_{k=1}^{K} h_{+k}^2} \tag{2.61}$$

c) wenn keine der Variablen zur abhängigen bzw. unabhängigen Variablen erklärt wurde (symmetrisches Maß):

$$D = \frac{C-D}{C+D+\frac{1}{2}(T_Y+T_X)} = \frac{2(C-D)}{\frac{1}{2}\left(n^2 - \sum_{j=1}^{J} h_{j+}^2 + n^2 - \sum_{k=1}^{K} h_{+k}^2\right)}. \tag{2.62}$$

Somers' D nimmt Werte im Bereich $-1 \leq D \leq +1$ an.
Bei asymmetrischen Kontingenztabellen kann dieser Koeffizient niemals den Maximalwert ± 1 erreichen. Es wird deshalb empfohlen, ihn nur auf symmetrische Tabellen anzuwenden.

• Beispiel 2.8 (Fortsetzung):
Entsprechend obiger Kontingenztabelle ist X die Bewertung des Studiums und Y die soziale Lage. $C = 1951$ und $D = 403$ ist aus der Anwendung von Gamma bekannt. T_Y und T_X ergeben sich zu:

$$\begin{aligned} T_Y &= 17(23+2) + 23(2) + 4(16+5) + 16(5) + 3(19+18) + 19(18) = 1088 \\ T_X &= 17(4+3) + 4(3) + 23(16+19) + 16(19) + 2(5+18) + 5(18) = 1376 \end{aligned}$$

Damit folgt für (2.60) bis (2.62):

$$\begin{aligned} D_{Y|X} &= (1951-403)/(1951+403+1088) = 0,44974 \\ D_{X|Y} &= (1951-403)/(1951+403+1376) = 0,41501 \\ D &= (1951-403)/(1951+403+1088/2+1376/2) = 0,43168. \end{aligned}$$

Wählt man im Dialogfeld „Crosstabs: Statistics“ unter Ordinal den Koeffizienten Somers' D aus, so erhält man die Angaben in Tabelle 2.19.

Tabelle 2.19.: Somers' D für Beurteilung des Studiums und soziale Lage

Directional Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Somers' d	Symmetric	,432	,070	5,915	,000
		Beurteilung des Studiums Dependent	,415	,071	5,915	,000
		Soziale Lage Dependent	,450	,070	5,915	,000

a. Not assuming the null hypothesis

b. Using the asymptotic standard error assuming the null hypothesis.

Interpretiert man z.B. den Wert $D_{X|Y}$, so zeigt sich unter den Paaren, die in der Variablen Y (Soziale Lage als unabhängige Variable) nicht gebunden sind, eine positive Assoziation, da die Anzahl der konkordanten Paare größer ist als die Anzahl der diskordanten Paare, d.h., Studenten mit einer sehr guten/guten (schlechten) sozialen Lage bewerten in der Tendenz das Studium als sehr gut/gut (schlecht).

Somers' D liefert einen kleineren Wert als Gamma, da im Nenner neben C und D auch die Anzahl der Ties steht. Während Ties in der unabhängigen Variablen weder für noch gegen einen Zusammenhang sprechen, sprechen Ties in der abhängigen Variablen gegen die Hypothese einer Abhängigkeit, da für unterschiedliche Werte der unabhängigen Variablen gleiche Werte der abhängigen Variablen auftreten.

Kendall's Tau-Werte

Sie sind symmetrische Maße für die Messung der Assoziation und Richtung des Zusammenhangs zwischen zwei ordinalskalierten Variablen. Durch die unterschiedliche Berücksichtigung von Ties sowie der Form der Kontingenztabelle gibt es drei Tau-Werte. Für die Tau-Werte gilt allgemein: $-1 \leq \tau \leq +1$.

Kendall's Tau$_a$:

$$\tau_a = \frac{C - D}{\frac{n(n-1)}{2}} \tag{2.63}$$

Dieser Koeffizient sollte nur für Kontingenztabellen verwendet werden, in denen keine Ties vorkommen. Dieser Koeffizient ist nicht unter SPSS verfügbar, da es praktisch keine Tabellen ohne Ties gibt.

Kendall's Tau$_b$:

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_X)(C + D + T_Y)}} \tag{2.64}$$

Dieser Koeffizient berücksichtigt Ties in der Variablen X und in der Variablen Y separat, jedoch keine Ties in beiden Variablen gleichzeitig. τ_b sollte nur auf quadratische Kontingenztabellen angewandt werden, da sonst ± 1 nicht erreicht werden kann.

Kendall's Tau$_c$:

$$\tau_c = \frac{2m(C - D)}{(m - 1)n^2}, \tag{2.65}$$

worin $m = min\{J, K\}$ ist. Dieser Koeffizient kann auf beliebig große und asymmetrische Kontingenztabellen angewandt werden.

• Beispiel 2.8 (Fortsetzung):

Es war $C = 1951$, $D = 403$, $T_Y = 1088$ und $T_X = 1376$. Damit folgt für (2.64) und (2.65):

$$\begin{aligned} \tau_b &= \frac{1951 - 403}{\sqrt{(1951 + 403 + 1376)(1951 + 403 + 1088)}} = 0,43203 \\ \tau_c &= \frac{2 \cdot 3(1351 - 403)}{107^2(3 - 1)} = 0,40562 \end{aligned}$$

Wählt man im Dialogfeld „Crosstabs: Statistics“ unter Ordinal die Koeffizienten Kendall's tau-b und Kendall's tau-c aus, so erhält man den SPSS-Output in Tabelle 2.20.

Tabelle 2.20.: Kendall's Tau für Beurteilung des Studiums und soziale Lage

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	,432	,070	5,915	,000
	Kendall's tau-c	,406	,069	5,915	,000
N of Valid Cases		107			

a. Not assuming the null hypothesis

b. Using the asymptotic standard error assuming the null hypothesis

2.2.3.4. Zusammenhangsmaß für eine intervallskalierte Variable mit einer nominalskalierten Variablen (Eta-Koeffizient)

Im Dialogfeld „Crosstabs: Statistics“ (siehe Abb. 2.39) ist unter Nominal by Intervall der Koeffizient Eta wählbar. Dieses PRE-Maß ist geeignet, die Stärke des Zusammenhanges zwischen zwei Variablen zu messen, wenn die abhängige Variable Y metrisch skaliert und die unabhängige Variable X nominal- oder ordinalskaliert ist. Ist die unabhängige Variable ebenfalls metrisch skaliert, wird man eher den Bravais-Pearson-Korrelationskoeffizient verwenden. Eta ist ein asymmetrisches Maß und setzt keine linearen Beziehungen zwischen den Variablen voraus. Es kann als derjenige Anteil der Gesamtvarianz der abhängigen Variablen interpretiert werden, der durch die Kenntnis der Werte der unabhängigen Variablen erklärt werden kann. Dies kann in der folgenden Weise verdeutlicht werden.

Wird nur unter Kenntnis der Verteilung der metrischen Variablen Y ein Variablenwert vorhergesagt, so verwendet man das arithmetische Mittel $\bar{y}$. Der Vorhersagefehler ist gleich der Summe der quadratischen Abweichungen der beobachteten Y-Werte y_k vom arithmetischen Mittel (Zähler der Varianz von Y):

$$E_1 = \sum_{k=1}^{K}(y_k - \bar{y})^2 h_{+k} = \sum_{k=1}^{K} y_k^2 h_{+k} - \frac{1}{n}\left(\sum_{k=1}^{K} y_k h_{+k}\right)^2 \tag{2.66}$$

mit

$$\bar{y} = \frac{\sum_{k=1}^{K} y_k h_{+k}}{n}. \tag{2.67}$$

Wird nunmehr auch die Kenntnis der Variablen X und somit der gemeinsamen Verteilung von X und Y vorausgesetzt, so wird als Vorhersagewert für ein statistisches Element mit dem Variablenwert x_j das arithmetische Mittel $\bar{y}_j$ der Y-Werte dieser Gruppe j verwendet. Der Vorhersagefehler bei gegebenem Variablenwert x_j ist die Summe der quadratischen Abweichungen der beobachteten Y-Werte y_{jk} vom arithmetischen Mittel $\bar{y}_j$ (Zähler der Varianz von Y innerhalb der Gruppe j):

$$\sum_{k=1}^{K}(y_{jk} - \bar{y}_j)^2 h_{jk} \tag{2.68}$$

mit

$$\bar{y}_j = \frac{\sum_{k=1}^{K} y_{jk} h_{jk}}{h_{j+}}. \tag{2.69}$$

Der Vorhersagefehler bei Einbeziehung aller Variablenwerte x_j ist folglich die Summe der quadratischen Abweichungen innerhalb aller Gruppen:

$$E_2 = \sum_{j=1}^{J} \left[\sum_{k=1}^{K} (y_{jk} - \bar{y}_j)^2 h_{jk} \right] = \sum_{j=1}^{J} \sum_{k=1}^{K} y_{jk}^2 h_{jk} - \sum_{j=1}^{J} \frac{1}{h_{j+}} \left(\sum_{k=1}^{K} y_{jk} h_{jk} \right)^2 \tag{2.70}$$

Im Sinne eines PRE-Maßes ist dann:

$$\eta^2 = \frac{E_1 - E_2}{E_1} \tag{2.71}$$

bzw.

$$\eta = \sqrt{\frac{E_1 - E_2}{E_1}}. \tag{2.72}$$

Im Zähler von (2.71) steht mit $E_1 - E_2$ die Differenz der Summe der quadratischen Abweichungen insgesamt (total sum of squares) und der Summe der quadratischen Abweichungen innerhalb der Gruppen (sum of squares within groups), was identisch ist mit der Summe der quadratischen Abweichungen zwischen den Gruppen (sum of squares between groups):

$$\sum_{j=1}^{J} (\bar{y}_j - \bar{y})^2 h_{j+}. \tag{2.73}$$

Diese Summe der quadratischen Abweichungen zwischen den Gruppen ist jedoch auf das Wirken der Variablen X zurückzuführen.

Da η^2 nach (2.71) ein Anteilswert ist, ist der Wertebereich von η mit $0 \leq \eta \leq 1$ gegeben.

• Beispiel 2.9:

Aus der Datei allbus.sav werden die Variablen monatliches persönliches Einkommen (einkomp1 = Y) und Geschlecht (sex = X) ausgewählt, um die Frage zu beantworten, ob eine Beziehung zwischen diesen beiden Variablen existiert. Da das Einkommen metrisch skaliert und das Geschlecht nominalskaliert ist, wird η als Maß zur Messung der Stärke der Beziehung verwendet.

Zur Verdeutlichung der obigen Herleitung von Eta werden E_1 und E_2 berechnet. Die Gesamtvarianz von Y, die Varianz innerhalb der Gruppen und die Varianz zwischen den Gruppen erhält man über:

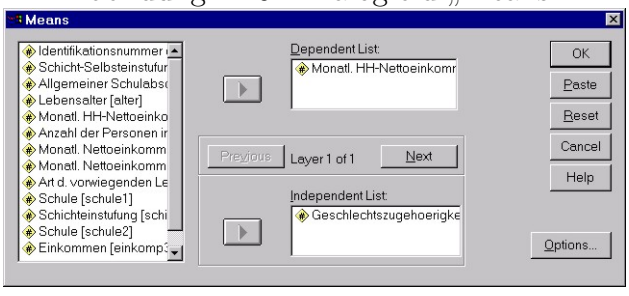
■ Analyze

■ Compare Means

■ Means. . . ,

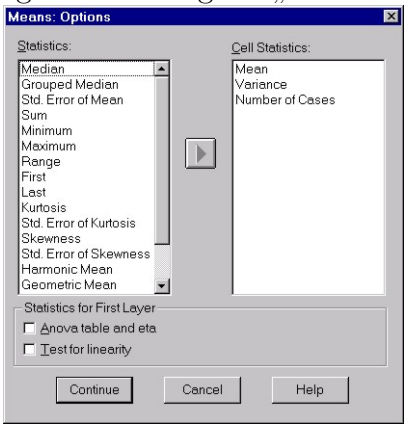
indem man in dem Dialogfeld „Means“ die Variable einkomp1 in das Feld „Dependent List:“ und die Variable sex in das Feld „Independent List:“ bringt.

Abbildung 2.40.: Dialogfeld „Means“



Über die Schaltfläche „Options. . . “ gelangt man in das Dialogfeld „Means: Options“, in dem Mittelwert, Varianz und die Anzahl der Fälle ausgewählt werden können.

Abbildung 2.41.: Dialogfeld „Means: Options“



Wie ersichtlich kann auch in diesem Dialogfeld Eta angefordert werden, indem „Anova table and eta“ angekreuzt wird.
Den SPSS-Output enthält Tabelle 2.21.

Tabelle 2.21.: Mittelwerte, Varianzen und die Anzahl der Fälle für monatliches persönliches Nettoeinkommen, gruppiert nach dem Geschlecht

Report

Monatliches Nettoeinkommen in DM

Geschlecht	Mean	Variance	N
Mann	2287,24	1358691,30	451
Frau	1190,69	505612,93	265
Total	1881,40	1322520,30	716

Daraus lassen sich die Summen der Abweichungsquadrate und damit die Fehler E_1 und E_2 ermitteln:

$$E_1 = s_y^2(n_G - 1) = 1.322.520,3 \cdot 715 = 945.602.014,5$$

$$
\begin{aligned}
E_2 &= s_y^2(M)(n_M - 1) + s_y^2(F)(n_F - 1) = 1.358.691,3 \cdot 450 + 505.612,93 \cdot 264 \\
&= 611.411.085 + 133.481.813,5 = 744.892.898,5 \\
\eta^2 &= (945.602.014,5 - 744.892.898,5)/945.602.014,5 = 0,2122 \\
\eta &= 0,4607
\end{aligned}
$$

Der SPSS-Output für Eta aus dem Dialogfeld „Means: Options“ (Abb. 2.41) heraus ist wie folgt:

Tabelle 2.22.: Eta für monatliches persönliches Nettoeinkommen und Geschlecht

Measures of Association

	Eta	Eta squared
Monatliches Nettoeinkommen * Geschlecht	,461	,212

Im SPSS-Output aus dem Dialogfeld „Crosstabs: Statistics“ (Abb. 2.39) heraus werden beide assymetrischen Versionen ausgegeben, obwohl der Version mit der nominalskalierten Variablen als abhängige Variable nicht sinnvoll ist, wie das auch für das Beispiel der Fall ist. Da das Geschlecht jedoch mit 1 für Mann und 2 für Frau kodiert wurde, wird diese Variable unter SPSS als eine numerische Variable behandelt.

Tabelle 2.23.: Eta für monatliches persönliches Nettoeinkommen und Geschlecht

Directional Measures

			Eta
Nominal by Intervall	Eta	Monatliches Nettoeinkommen Dependent	,461
		Geschlecht Dependent	,674

Der Vorhersagefehler für die Variable monatliches persönliches Nettoeinkommen kann somit bei Kenntnis der Ausprägung der Variablen Geschlecht um rund 21% (= η^2) verringert werden. Es besteht in diesem Sinne eine schwach ausgeprägte Beziehung zwischen diesen Variablen.

2.2.3.5. Kappa-Koeffizient

Cohens Kappa-Koeffizient[15] wird zur Messung des Grades der Übereinstimmung von Beurteilungen der gleichen Objekte bzw. Tatbestände durch zwei Personen verwendet. Da beide Personen nach den gleichen Kriterien beurteilen, stehen in den Zeilen und Spalten der Kontingenztabelle die gleichen Ausprägungen. Daraus folgt, dass die Kontingenztabelle quadratisch sein muß.

[15] Siehe u.a. Bortz, J., Lienert, G.A., Boehnke, K. (1990), S. 450 ff., 458 ff.

Der Kappa-Koeffizient basiert zum einen auf dem beobachteten Anteil der übereinstimmenden Beurteilungen, der mit f_b symbolisiert wird. Die übereinstimmenden Beurteilungen stehen in der Hauptdiagonale der Kontingenztabelle, so dass sich der Anteil f_b als

$$f_b = \sum_{j=1}^{J} f_{jj} = \frac{\sum_{j=1}^{J} h_{jj}}{n} \tag{2.74}$$

ergibt.

Da jedoch übereinstimmende Beurteilungen auch auftreten können, wenn die Personen nach dem Zufallsprinzip bewerten würden, muß um diese Zufallsübereinstimmung korrigiert werden. Dazu wird der Anteil der übereinstimmenden Beurteilungen (symbolisiert mit f_z) ermittelt, der bei zufälliger Bewertung erwartet würde, d.h. aufgrund des Produkts der Randhäufigkeiten:

$$f_z = \frac{\sum_{j=1}^{J} h_{j+}h_{+j}}{n^2}. \tag{2.75}$$

$1 - f_z$ ist entsprechend der theoretisch mögliche Anteil der über den Zufall hinausgehenden übereinstimmenden Bewertungen. Die Differenz $f_b - f_z$ beinhaltet dagegen den tatsächlich beobachteten Anteil der über den Zufall hinausgehenden übereinstimmenden Bewertungen. Diese Differenz wird zum Zwecke der Normierung durch $1 - f_z$ dividiert. Das Ergebnis ist der Kappa-Koeffizient:

$$\begin{aligned} \kappa &= \frac{f_b - f_z}{1 - f_z} \\ &= \frac{n\sum_{j=1}^{J} h_{jj} - \sum_{j=1}^{J} h_{j+}h_{+j}}{n^2 - \sum_{j=1}^{J} h_{j+}h_{+j}}. \end{aligned} \tag{2.76}$$

Der Wertebereich des Kappa-Koeffizienten ist $-1 \leq \kappa \leq 1$.

• Beispiel 2.10:
Ein einfaches Beispiel soll den Kappa-Koeffizienten verdeutlichen. Zur Evaluierung der Lehrveranstaltungen werden zwei Studenten, die die gleichen Lehrveranstaltungen während des Studiums besucht haben, gebeten, diese zu beurteilen. Als Bewertungskriterien werden anspruchsvoll (1), mittelmäßig (2) und langweilig (3) vorgegeben. Der Grad der Übereinstimmung in ihren Beurteilungen wird mit dem Kappa-Koeffizienten gemessen. Es wird angenommen, dass die beiden Studenten $n = 50$ gleiche Lehrveranstaltungen besuchten.

Ausgehend von den in der Datei kappa.sav enthaltenen Daten ergibt sich nachstehender SPSS-Output, wobei auch die Kontingenztabelle angefordert wird.

Tabelle 2.24.: Kontingenztabelle für die Bewertung der Lehrveranstaltungen durch zwei Studenten und Kappa-Koeffizient

Student 1 * Student 2 Crosstabulation

			Student 2			
			anspruchsvoll	mittelmäßig	langweilig	Total
Student 1	anspruchsvoll	Count	6	4	5	15
		% of Total	12,0%	8,0%	10,0%	30,0%
	mittelmäßig	Count	8	12	4	24
		% of Total	16,0%	24,0%	8,0%	48,0%
	langweilig	Count	1	3	7	11
		% of Total	2,0%	6,0%	14,0%	22,0%
Total		Count	15	19	16	50
		% of Total	30,0%	38,0%	32,0%	100,0%

Symmetric Measures

	Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig
Measure of Agreement Kappa	,239	,106	2,417	,016
N valid of Cases	50			

a. Not assuming the null hypothesis
b. Using the asymptotic standard error assuming the null hypothesis.

Der Anteil der beobachteten Übereinstimmung in der Beurteilung der Lehrveranstaltungen ist

$$f_b = 0,12 + 0,24 + 0,14 = 0,5,$$

d.h., in 50% aller beurteilten Lehrveranstaltungen stimmen die beiden Studenten überein. Der Anteil übereinstimmender Beurteilungen, der bei zufälliger Bewertung erwartet würde, beträgt:

$$f_z = (15 \cdot 15 + 24 \cdot 19 + 11 \cdot 16)/50^2 = 857/2500 = 0,3428.$$

Bei gegebenen Randhäufigkeiten sind 34,28% Zufallsübereinstimmungen zu erwarten. Der theoretisch mögliche Anteil der über den Zufall hinausgehenden übereinstimmenden Beurteilungen beträgt 65,72%. Der tatsächlich beobachtete Anteil der über den Zufall hinausgehenden übereinstimmenden Beurteilungen beläuft sich auf 15,72%. Für den Kappa-Koeffizienten folgt:

$$\kappa = (0,5 - 0,3428)/(1 - 0,3428) = 0,2392.$$

Es besteht somit eine mäßige Übereinstimmung in den Beurteilungen der Lehrveranstaltungen durch diese beiden Studenten.

Der Kappa-Koeffizient kann bei Gültigkeit der Nullhypothese $H_0 : \kappa = 0$ mit einem approximativen T-Test geprüft werden, der für das Beispiel Signifikanz zum 5%-Niveau anzeigt.

2.2.3.6. Relatives Risiko

Das relative Risiko[16] ist ein Maß für die Stärke des Zusammenhanges zwischen einer Faktorvariablen und dem Eintreten eines Ereignisses. Es wird im weiteren vorausgesetzt, dass die Faktorvariable dichotom ist, d.h., die Faktorausprägung vorhanden ist oder nicht. Ebenso ist die Variable, die das Ereignis aufnimmt, dichotom, d.h., das Ereignis ist eingetreten bzw. nicht eingetreten. Es resultiert somit eine 2x2 Kontingenztabelle.

Bei der Berechnung des relativen Risikos muß zwischen zwei Arten der Untersuchung unterschieden werden: der Kohorten- oder prospektiven Studie[17] und der Fall-Kontroll- bzw. retrospektiven Sudie.

Kohorten- oder prospektive Studie

Bei der Kohorten-Studie werden zwei Gruppen gebildet, wobei bei der einen Gruppe die Faktorausprägung vorhanden, bei der anderen Gruppe nicht vorhanden ist. Beide Gruppen sind jedoch zu Beginn der Studie gleich bezüglich des Nichtvorhandenseins des interessierenden Ereignisses. Diese Gruppen werden über einen bestimmten Zeitraum beobachtet. Am Ende des Untersuchungszeitraumes wird in jeder Gruppe festgestellt, wie oft das Ereignis eingetreten ist. Je Gruppe kann dann die Auftretensrate (Inzidenzrate) ermittelt werden. Geht man davon aus, dass in der 2×2-Kontingenztabelle das Vorhandensein der Faktorausprägung in der ersten Zeile und das Auftreten der Ereignisses in der ersten Spalte steht, so resultiert nachstehende Tabelle.

Tabelle 2.25.: Kontingenztabelle für Faktorvariable und Ereignis-Variable

Faktorausprägung	Ereignis		Randverteilung
	eingetreten	nicht eingetreten	
vorhanden	h_{11}	h_{12}	$h_{11} + h_{12}$
nicht vorhanden	h_{21}	h_{22}	$h_{21} + h_{22}$
Randverteilung	$h_{11} + h_{21}$	$h_{12} + h_{22}$	n

Die Inzidenzraten je Gruppe werden wie folgt berechnet:

[16] Siehe u.a. Bortz, J., Lienert, G.A., Boehnke, K. (1990), S. 341 f.; Fahrmeir, L., Künstler, R., Pigeot, I., Tutz, G. (1997), S. 119 ff.

[17] Siehe u.a. Rönz, B., Strohe, H.G. (1994), S. 183

I(Faktorausprägung vorhanden) $= h_{11}/(h_{11} + h_{12})$
I(Faktorausprägung nicht vorhanden) $= h_{21}/(h_{21} + h_{22})$.
Das relative Risiko ist das Verhältnis der Inzidenzrate der Gruppe, bei der die Faktorausprägung vorhanden ist, zur Inzidenzrate der Gruppe ohne die Faktorausprägung:

$$r_1 = \frac{I(Faktor)}{I(Nicht - Faktor)} = \frac{h_{11}(h_{21} + h_{22})}{h_{21}(h_{11} + h_{12})}. \tag{2.77}$$

Fall-Kontroll- oder retrospektive Studie

Bei einer Fall-Kontroll-Studie (case-control-study) wird eine Gruppe ausgewählt, bei der das Ereignis eingetreten ist (Fall-Gruppe, cases), und eine Kontroll-Gruppe, bei der Ereignis nicht eingetreten ist. Nunmehr wird festgestellt, wie oft in jeder Gruppe die Faktorausprägung vorhanden ist und wie oft nicht. Bei dieser Art der Untersuchung kann kein relatives Risiko im obigen Sinne bestimmt werden, da das Ereignis bereits eingetreten ist. Statt dessen wird das odds-ratio[18] berechnet.

Es wird im weiteren analog zur Tabelle 2.25 davon ausgegangen, dass die Fall-Gruppe (Cases) in der ersten Spalte und die Kontroll-Gruppe (Control) in der zweiten Spalte, die vorhandene Faktorausprägung in der ersten Zeile und die nicht vorhandene Faktorausprägung in der zweiten Zeile stehen.

Die relativen Chancen (odds) für das Eintreten des Ereignisses in der Gruppe, in der die Faktorausprägung vorhanden ist, ergeben sich zu
odds (Faktorausprägung vorhanden) $= h_{11}/h_{12}$
und die odds für das Eintreten des Ereignisses in der Gruppe, in der die Faktorausprägung nicht vorhanden ist, zu
odds (Faktorausprägung nicht vorhanden) $= h_{21}/h_{22}$.
Der Quotient

$$r_2 = \frac{odds(\text{Faktorausprägung vorhanden})}{odds(\text{Faktorausprägung nicht vorhanden})} = \frac{h_{11}h_{22}}{h_{12}h_{21}} \tag{2.78}$$

wird als odds ratio (Verhältnis der relativen Chancen bezeichnet).

• Beispiel 2.11:

In der Datei rauchen.sav sind die Daten für Raucher (Faktorausprägung vorhanden) und Nichtraucher (Faktorausprägung nicht vorhanden) sowie Auftreten von Lungenkrebs (interessierendes Ereignis) enthalten. Es soll untersucht werden, ob zwischen Rauchen und dem Auftreten von Lungenkrebs eine Beziehung existiert.

a) Kohorten-Studie

Wenn man davon ausgeht, dass eine Kohorten-Studie durchgeführt wird, so wird eine Gruppe

[18] Siehe u.a. Kleinbaum, D. G. (1994)

von Rauchern und eine Gruppe von Nichtrauchern ausgewählt. Alle Personen weisen zu Beginn der Studie keine Auffälligkeiten hinsichtlich von Lungenkrebs auf (Nichtvorhandensein des Ereignisses). Nach einer bestimmten Anzahl von Jahren wird festgehalten, wie oft in jeder Gruppe Lungenkrebs (Eintreten des Ereignisses) aufgetreten ist. Die zugehörige Kontingenztabelle und das relative Risiko enthält der nachstehende SPSS-Output.

Tabelle 2.26.: Kontingenztabelle für Rauchen und Lungenkrebs sowie relatives Risiko

Rauchen * Lungenkrebs Crosstabulation

Count

		Lungenkrebs		
		ja	nein	Total
Rauchen	ja	10	2	12
	nein	8	30	38
Total		18	32	50

Risk Estimate

		95% Confidence Intervall	
	Value	Lower	Upper
Odds Ratio for Rauchen (ja/nein)	18,750	3,402	103,335
For cohort Lungenkrebs = ja	3,958	2,034	7,702
For cohort Lungenkrebs = nein	,211	,059	,756
N of Valid Cases	50		

Nach (2.77) läßt sich das relative Risiko leicht aus der Kontingenztabelle berechnen. Die Inzidenzraten sind wie folgt:

$$\begin{aligned} I(Raucher) &= h_{11}/(h_{11}+h_{12}) = 10/12 = 0,8333 \\ I(Nicht-Raucher) &= h_{21}/(h_{21}+h_{22}) = 8/38 = 0,2105 \end{aligned}$$

Das relative Risiko für das Auftreten von Lungenkrebs bei den Rauchern beträgt 83,33%, während es für Nichtraucher nur 21,05% ausmacht.
Als relatives Risiko für das Auftreten von Lungenkrebs (im Verhältnis Raucher zu Nichtraucher) resultiert:

$$r_1 = 10 \cdot 38/8 \cdot 12 = 0,8333/0,2105 = 3,9583.$$

Bei den untersuchten Personen ist für die Raucher das Risiko, Lungenkrebs zu bekommen, circa viermal größer als für die Nichtraucher.

Dieses Ergebnis findet man im SPSS-Output in der Zeile For cohort Lungenkrebs = ja. Dabei gilt die Vereinbarung, dass das interessierende Ereignis in der 1. Spalte der Kontingenztabelle steht und die Inzidenzrate der 1. Zeile auf die Inzidenzrate der 2. Zeile bezogen wird.

Wenn das interessierende Ereignis jedoch das Nichteintreten von Lungenkrebs (2. Spalte der Kontingenztabelle) gewesen wäre, würden sich die Inzidenzraten zu

$$\begin{aligned} I(Raucher) &= h_{12}/(h_{11}+h_{12}) = 2/12 = 0,1667 \\ I(Nicht-Raucher) &= h_{22}/(h_{21}+h_{22}) = 30/38 = 0,7895 \end{aligned}$$

ergeben. Als relatives Risiko für das Nichtauftreten von Lungenkrebs (im Verhältnis Raucher zu Nichtraucher) resultiert:

$$\begin{aligned} r_1 &= [h_{12}/(h_{11}+h_{12})]/[h_{22}/(h_{21}+h_{22})] = h_{12}(h_{21}+h_{22})/h_{22}(h_{11}+h_{12}) \\ &= 2 \cdot 38/30 \cdot 12 = 0,1667/0,7895 = 0,2111 \end{aligned}$$

Dieses Ergebnis findet man im SPSS-Output in der Zeile For cohort Lungenkrebs = nein. Dabei gilt die Vereinbarung, dass das interessierende Ereignis in der 2. Spalte der Kontingenztabelle steht und die Inzidenzrate der 1. Zeile auf die Inzidenzrate der 2. Zeile bezogen wird.

Mit dem ebenfalls ausgegebenen 95%-Konfidenzintervall kann die Nullhypothese „Die beiden Inzidenzraten sind gleich; d.h. $r_1 = 1$“ geprüft werden. Überdeckt das Konfidenzintervall den Wert 1, so besteht keine Veranlassung, die Nullhypothese zu verwerfen. Liegt der Wert 1 nicht im Konfidenzintervall, so wird die Nullhypothese auf dem 5% Signifikanzniveau verworfen. Für das Beispiel wird auf Basis der Stichprobe die Nullhypothese verworfen, so dass von einer signifikanten Assoziation zwischen Rauchen und Lungenkrebs ausgegangen werden kann.

b) Fall-Kontroll-Studie

Wenn man davon ausgeht, dass eine Fall-Kontroll-Studie durchgeführt wird, so wird eine Gruppe von Personen, die an Lungenkrebs erkrankt sind (Fall-Grupe, cases), ausgewählt und eine Gruppe von nicht an Lungenkrebs erkrankten Personen (Kontroll-Gruppe, control), die aber bezüglich der Zusammensetzung hinsichtlich Geschlecht, Alter usw. mit den Personen der Fall-Gruppe vergleichbar sind. Nach dieser Gruppenbildung wird festgestellt, wieviele Personen in jeder Gruppe rauchen (Faktorausprägung vorhanden) und wie viele Personen nicht rauchen (Faktorausprägung nicht vorhanden). Es sei angenommen, dass sich im Ergebnis der Datenerhebung die gleiche Kontingenztabelle wie in Tabelle 2.26 ergeben habe.
Da das interessierende Ereignis (Lungenkrebs) bereits eingetreten ist, wird das odds-ratio nach (2.78) berechnet. Die relativen Chancen (odds) für das Eintreten von Lungenkrebs (im Verhältnis zum Nichteintreten von Lungenkrebs) in der Gruppe der Raucher ergeben sich zu

$$odds\ (Raucher) = h_{11}/h_{12} = 10/2 = 5$$

und die odds für das Eintreten von Lungenkrebs (im Verhältnis zum Nichteintreten von Lungenkrebs) in der Grppe der Nichtraucher zu

$$odds\ (Nichtraucher) = h_{21}/h_{22} = 8/30 = 0,2667$$

Als odds ratio resultiert:

$$r_2 = (h_{11}/h_{12})/(h_{21}/h_{22}) = h_{11} \cdot h_{22}/h_{12} \cdot h_{21} = 10 \cdot 30/8 \cdot 2 = 5/0,2667 = 18,75.$$

Lungenkrebs tritt in der Stichprobe rund 18mal häufiger bei Rauchern als bei Nichtrauchern auf.

Dieses Ergebnis findet man im SPSS-Output in der Zeile Odds Ratio for Rauchen (ja/nein). Dabei gilt die Vereinbarung, dass das interessierende Ereignis in der 1. Spalte der Kontingenztabelle steht und die odds der 1. Zeile auf die odds der 2. Zeile bezogen werden.

Mit dem ebenfalls ausgegebenen 95%-Konfidenzintervall kann die Nullhypothese „Die beiden odds sind gleich; d.h. $r_2 = 1$“ geprüft werden. Überdeckt das Konfidenzintervall den Wert 1, so besteht keine Veranlassung, die Nullhypothese zu verwerfen. Liegt der Wert 1 nicht im Konfidenzintervall, so wird die Nullhypothese auf dem 5% Signifikanzniveau verworfen. Für das Beispiel wird auf der Basis der Stichprobe die Nullhypothese auf dem 5%-Niveau verworfen.

Wie gezeigt ist aufgrund der beobachteten Daten und der Betrachtung der Kontingenztabelle nicht ersichtlich, ob es sich um eine Kohorten- oder um eine Fall-Kontroll-Studie handelt. Im SPSS-Output sind deshalb sowohl das relative Risiko als auch das odds ratio enthalten. Der Nutzer muß denjenigen Wert auswählen, der für seine Untersuchung und die Art seiner Studie zutrifft.

Wenn bei derartigen Untersuchungen eine signifikante Beziehung festgestellt wurde, interessiert oftmals auch das Ausmaß des Einflusses. Da die abhängige Variable Y (auch als Response-Variable bezeichnet) eine dichotome oder binäre Variable ist, kann offensichtlich nicht die bekannte Regressionsanalyse angewandt werden. Die Modellierung einer binären Response-Variablen in Abhängigkeit von Variablen beliebigen Skalenniveaus führt zu einem Logit-Modell (logistische Regression) oder zu einem Probit-Modell. Diese gehören zu den verallgemeinerten linearen Modellen, die in der gleichnamigen Lehrveranstaltung behandelt werden.

• Beispiel 2.12:
Von $n = 100$ zufällig ausgewählten Personen wurde erfaßt, ob eine Herzkranzgefäßerkrankung (HKE) vorliegt (1) oder nicht (2). Zusätzlich wurde das Alter jeder Person registriert.[19] Die Altersangaben wurden in folgender Weise dichotomisiert: 1 = „55 Jahre und älter“, 2 = „bis 54

[19] Die Ausgangsdaten wurden aus Hosmer, D.W., Lemeshow, S. (1989), S. 2 ff. entnommen.

Jahre“. Die Frage lautet: Gibt es eine Beziehung zwischen Auftreten von HKE und dem Alter? Mit diesen Daten, die in der Datei hke.sav enthalten sind, liegt eine Fall-Kontroll-Studie vor. Neben der Kontingenztabelle wurde im Dialogfeld „Crosstabs: Statistics“ (Abb. 2.39) auf die Ausgabe von Risk entschieden.

Tabelle 2.27.: Kontingenztabelle für Alter und Herkranzgefäßerkrankung sowie odds ratio

Alter * Herzkranzgefäßerkrankung Crosstabulation

Count

		Herzkranzgefäßerkrankung		
		ja	nein	Total
Alter	55-	21	6	27
	-54	22	51	73
Total		43	57	100

Risk Estimate

		95% Confidence Intervall	
	Value	Lower	Upper
Odds Ratio for Alter (55-/-54)	8,114	2,880	22,861
For cohort Herzkranzgefäßerkrankung = ja	2,581	1,723	3,863
For cohort Herzkranzgefäßerkrankung = nein	,318	,155	,655
N of Valid Cases	100		

Die odds für die Gruppe der älteren und die Gruppe der jüngeren Personen sind:

$$o(55-) = h_{11}/h_{12} = 21/6 = 3,5$$
$$o(-54) = h_{21}/h_{22} = 22/51 = 0,4314.$$

Die relative Chance für HKE (im Verhältnis zum Nichtauftreten von HKE) beträgt in der Stichprobe bei den älteren Personen (55 Jahre und älter) 3,5 und bei den „jüngeren“ Personen weniger als die Hälfte.
Als odd ratio folgt gemäß (2.78):

$$r_2 = 21 \cdot 51/6 \cdot 22 = 3,5/0,4314 = 8,114.$$

HKE tritt also in der Stichprobe rund achtmal häufiger bei Personen 55 Jahre und älter auf als unter den „jüngeren“ Personen. Dieses Ergebnis steht in der Output-Zeile Ods Ratio for Alter (55-/-54). Die Prüfung der Nullhypothese, dass die odds gleich sind, d.h. $r_2 = 1$, kann auf dem 5%-Niveau abgelehnt werden.

2.2.3.7. McNemar - Test

Dieser Option im Dialogfeld „Crosstabs: Statistics“ (siehe Abb. 2.39) unterliegt ebenfalls eine spezielle Problemstellung. Der McNemar-Test[20] ist ein nichtparametrischer Test für zwei abhängige Stichproben, der die Hypothese prüft, dass die Kombinationen der verschiedenen Ausprägungen einer dichotomen Response-Variablen gleich wahrscheinlich sind.

Abhängige Stichproben[21] (auch als verbundene, korrelierte, gepaarte Stichproben bezeichnet) ergeben sich immer dann, wenn zwischen den Elementen zweier (oder mehrerer) Stichproben eine gegenseitige Beeinflussung ihres Zufallsverhaltens bzw. eine gewisse Informationsbeziehung besteht. Das ist z.B. der Fall, wenn

- die Elemente ein und derselben Stichprobe mehrmals auf ein Merkmal hin beobachtet werden. Die Wiederholung der Beobachtung kann z.B. nach einem zeitlichen Abstand oder unter veränderten Bedingungen vorgenommen werden.
  Beispiele: Beobachtung einer biologischen Reaktion an Patienten vor, während und nach einer ärztlichen Behandlung; Veränderung der Einstellung vergleichbarer Individuen unter verschiedenen Bedingungen (beispielsweise Medieneinwirkung); Einschätzung ein und derselben Leistung durch zwei Prüfer; Veränderung der Wahrnehmung durch mehrmalige Betrachtung.

- nach einem bestimmten Kriterium Parallelstichproben gebildet werden, deren Elemente zufallsbedingt einer der Stichproben zugeordnet werden (parallelisierte Stichproben, matched samples).
  Beispiel: Es werden Paare von Probanden gebildet, die hinsichtlich festgelegter Merkmale (z.B. Geschlecht, Alter) aufeinander abgestimmt und somit möglichst homogen sind. Die Zuordnung der Partner zu den beiden Stichproben (Versuchsgruppen) erfolgt zufällig. Jede Stichprobe von Probanden wird dann einer bestimmten Behandlungsmethode unterzogen, beispielsweise die Probanden der einen Gruppe (control subjects) mit einer Standardmethode und die Probanden der anderen Gruppe (experimental subjects) mit einer neuen Methode.

Der McNemar-Test wird bei zwei abhängigen Stichproben mit dichotomer Response-Variable angewandt, d.h., im Ergebnis der Stichprobenerhebung liegen Beobachtungspaare vor, die einander zugeordnet sind und nur dieselben dichotomen Werte (z.B. ja/nein, Reaktion/keine Reaktion, positiv/negativ, dafür/dagegen), codiert beispielsweise mit 1 und 2, annehmen können.

[20] Vgl. u.a. Bortz, J., Lienert, G.A., Boehnke, K. (1990), S. 160 ff.; Daniel, W.W. (1990), S. 162 ff.; Hartung, J., Elpelt, B., Klösener, K.-H. (1993), S. 423; Lindgren, B. W. (1993), S. 381 ff.; Lienert, G.A. (1973), S.191 ff.; Sachs, L. (1992), S.467 ff.; Clauß, G., Finze, F.-R., Partzsch, L. (1994), S. 240 ff.

[21] Vgl. u.a. Bortz, J. (1993), S. 135

Zur Vereinfachung der weiteren Darlegungen wird davon ausgegangen, dass Elemente ein und derselben Stichprobe zweimal beobachtet werden. Die zugehörigen Häufigkeiten lassen sich in einer 2×2 Kontingenztabelle (Vierfeldertafel) darstellen.

Tabelle 2.28.: Kontingenztabelle für abhängige Stichproben

1. Untersuchung	2. Untersuchung		Randverteilung
	Reaktion 1	Reaktion 2	von X
Reaktion 1	h_{11}	h_{12}	h_{1+}
Reaktion 2	h_{21}	h_{22}	h_{2+}
Randverteilung von Y	h_{+1}	h_{+2}	n

Darin sind :
n - Stichprobenumfang
$h_{1+} = h_{11} + h_{12}$: Anzahl der Elemente, die bei der 1. Untersuchung Reaktion 1 aufweisen,
$h_{2+} = h_{21} + h_{22}$: Anzahl der Elemente, die bei der 1. Untersuchung Reaktion 2 aufweisen,
$h_{+1} = h_{11} + h_{21}$: Anzahl der Elemente, die bei der 2. Untersuchung Reaktion 1 aufweisen,
$h_{+2} = h_{12} + h_{22}$: Anzahl der Elemente, die bei der 2. Untersuchung Reaktion 2 aufweisen,
h_{11}: Anzahl der Elemente, die bei beiden Untersuchungen Reaktion 1 aufweisen,
h_{12}: Anzahl der Elemente, die bei der 1. Untersuchung Reaktion 1 und bei der 2. Untersuchung Reaktion 2 aufweisen,
h_{21}: Anzahl der Elemente, die bei der 1. Untersuchung Reaktion 2 und bei der 2. Untersuchung Reaktion 1 aufweisen,
h_{22}: Anzahl der Elemente, die bei beiden Untersuchungen Reaktion 2 aufweisen.
Es soll nunmehr geprüft werden, ob die Wahrscheinlichkeiten bei beiden Untersuchungen gleich sind.

Wenn analog zu den beobachteten Häufigkeiten die (unbekannten) Wahrscheinlichkeiten mit p symbolisiert werden, so können die Hypothesen wie folgt formuliert werden:

- wenn die Reaktion 1 diejenige ist, auf die sich das Interesse bei der Untersuchung richtet:
  $H_0 : p_{1+} = p_{+1} \qquad H_1 : p_{1+} \neq p_{+1}$,
- wenn die Reaktion 2 diejenige ist, auf die sich das Interesse bei der Untersuchung richtet:
  $H_0 : p_{2+} = p_{+2} \qquad H_1 : p_{2+} \neq p_{+2}$.

Wegen $p_{1+} = p_{11} + p_{12}$ und $p_{+1} = p_{11} + p_{21}$ bzw. $p_{2+} = p_{21} + p_{22}$ und $p_{+2} = p_{12} + p_{22}$ ergibt sich in beiden Fällen die äquivalente Hypothesenformulierung:
$H_0 : p_{12} = p_{21} \qquad H_1 : p_{12} \neq p_{21}$.
Es werden nur noch die Zellen der Kontingenztabelle betrachtet, bei denen eine Veränderung

der Reaktion eingetreten ist; Zellen der Kontingenztabelle mit gleichen Reaktionen bei beiden Untersuchungen liefern keine Information zur Problemstellung und können weggelassen werden.

Hinter der Nullhypothese steht die Annahme, dass die Veränderungen in den Reaktionen rein zufällig sind und die Zellhäufigkeiten h_{12} und h_{21} nur zufällige Stichprobenschwankungen aufweisen. Dies impliziert, dass die Hälfte der Veränderungen von Reaktion 1 nach Reaktion 2 (enthalten in Zelle (1,2) der Kontingenztabelle) und die andere Hälfte der Veränderungen von Reaktion 2 nach Reaktion 1 (enthalten in Zelle (2,1) der Kontingenztabelle) erfolgte. Die bei Gültigkeit der Nullhypothese erwarteten absoluten Zellhäufigkeiten der Zellen (1,2) und (2,1) sind somit:

$$\hat{e}_{12} = \hat{e}_{21} = (h_{12} + h_{21})/2.$$

Abseits der Hauptdiagonalen wird eine symmetrische Häufigkeitsverteilung erwartet, weshalb der McNemar-Test auch als Vierfeldersymmetrietest bezeichnet wird. Große Abweichungen der beobachteten Zellhäufigkeiten h_{12} und h_{21} von den erwarteten Häufigkeiten sprechen gegen die Nullhypothese.

Zur Prüfung von H_0 wird die bekannte Chi-Quadrat-Teststatistik verwendet, jedoch nur noch unter Einbeziehung der Zellen (1,2) und (2,1):

$$\begin{aligned} V &= \frac{(h_{12} - \hat{e}_{12})^2}{\hat{e}_{12}} + \frac{(h_{21} - \hat{e}_{21})^2}{\hat{e}_{21}} \\ &= \frac{\left(h_{12} - \frac{h_{12}+h_{21}}{2}\right)^2}{\frac{h_{12}+h_{21}}{2}} + \frac{\left(h_{21} - \frac{h_{12}+h_{21}}{2}\right)^2}{\frac{h_{12}+h_{21}}{2}} = \frac{(h_{12} - h_{21})^2}{h_{12} + h_{21}}. \end{aligned} \tag{2.79}$$

Da die Teststatistik V stetig verteilt ist, die Häufigkeiten dagegen diskret sind, wird für den Fall $(h_{12} + h_{21}) < 30$ im allgemeinen eine Stetigkeitskorrektur vorgenommen:

$$V_{korr.} = \frac{(|h_{12} - h_{21}| - 1)^2}{h_{12} + h_{21}}. \tag{2.80}$$

Bei Gültigkeit der Nullhypothese ist die Teststatistik V bzw. $V_{korr.}$ approximativ chi-quadratverteilt mit einem Freiheitsgrad.

Wenn die Nullhypothese gilt, wird die Teststatistik kleine Werte annehmen, so dass H_0 für „zu große“ Werte von V abgelehnt wird. Den kritischen Wert $c = \chi^2_{1-\alpha;1}$ entnimmt man für $P(V \leq c) = 1 - \alpha$ und die Anzahl der Freiheitsgrade df = 1 aus der Tabelle der Verteilungsfunktion der Chi - Quadrat - Verteilung (siehe Anhang C). Die Entscheidungsbereiche für den Test sind:

- Ablehnungsbereich der H_0:
  $\{v | v > \chi^2_{1-\alpha;1}\}$

- Nichtablehnungsbereich der H_0:
  $\{v|v \leq \chi^2_{1-\alpha;1}\}$.

Die Wahrscheinlichkeit, dass die Teststatistik V eine Realisation aus dem Ablehnungsbereich der H_0 annimmt, entspricht dem vorgegeben Signifikanzniveau

$$\alpha = P(V > \chi^2_{1-\alpha;1}|H_0).$$

Die Teststatistik (2.79) von McNemar läßt sich auch wie folgt begründen:
Für die relative Häufigkeit der Reaktion 1 bei Untersuchung 1 ergibt sich

$$f_{1+} = h_{1+}/n = (h_{11} + h_{12})/n$$

und für die relative Häufigkeit der Reaktion 1 bei Untersuchung 2

$$f_{+1} = h_{+1}/n = (h_{11} + h_{21})/n.$$

Die Differenz dieser relativen Häufigkeit ist

$$f_{1+} - f_{+1} = (h_{11} + h_{12})/n - (h_{11} + h_{21})/n = (h_{12} - h_{21})/n.$$

Die geschätzte Varianz dieser Differenz ergibt sich zu

$$\hat{\sigma}^2(f_{1+} - f_{+1}) = (h_{12} + h_{21})/n^2.$$

Die Zufallsvariable Z (als Differenz der relativen Häufigkeiten bezogen auf den Standardfehler)

$$Z = \frac{f_{1+} - f_{+1}}{\hat{\sigma}(f_{1+} - f_{+1})} = \frac{h_{12} - h_{21}}{\sqrt{h_{12} + h_{21}}} \tag{2.81}$$

ist bei Gültigkeit der Nullhypothese $H_0 : p_{1+} = p_{+1}$ (bzw. $H_0 : p_{1+} - p_{+1} = 0$) und bei genügend großem Stichprobenumfang standardnormalverteilt. Das Quadrat der standardnormalverteilten Zufallsvariable Z ist χ^2-verteilt mit 1 Freiheitsgrad: $Z^2 \sim \chi^2(1)$. Mit Z^2 resultiert jedoch die Teststatistik (2.79).
Einseitige Hypothesenformulierung:
Falls über die Richtung der zu erwartenden Reaktionsveränderungen schon vor der Stichprobenerhebung eine begründete Aussage getroffen werden kann, läßt sich auch ein entsprechender einseitiger Test durchführen, also entweder

$$H_0 : p_{12} \leq p_{21} \qquad H_1 : p_{12} > p_{21} \; (rechtsseitiger\ Test)$$

oder

$$H_0 : p_{12} \geq p_{21} \qquad H_1 : p_{12} < p_{21} \; (linksseitiger\ Test).$$

Die Voraussetzungen des McNemar-Tests sind:

- 2 abhängige Stichproben,
- Beobachtung einer dichotomen Response-Variablen,
- reine Zufallsauswahl der Stichprobenelemente sowie bei parallelisierten Stichproben zufällige Zuordnung der Stichprobenelemente zu den beiden Stichproben. Daraus folgt:
  - Wenn die Elemente ein und derselben Stichprobe zweimal beobachtet werden, dann sind die Stichprobenelemente wechselseitig unabhängig, die beiden Beobachtungen an demselben Element sind jedoch abhängig.
  - Wenn parallelisierte Stichproben verwendet werden, so sind die Stichprobenpaare unabhängig, aber die Beobachtungen eines gegebenen Paares sind abhängig.
- eindeutige und vollständige Zuordnung der Stichprobenelemente zu den Zellen der 2×2 Kontingenztabelle für abhängige Stichproben (Tabelle 2.28),
- die erwarteten Zellhäufigkeiten e_{12} und e_{21} sind größer als 5.

Ist die letztgenannte Voraussetzung nicht erfüllt, so wird für die Zellhäufigkeit h_{12} oder h_{21} ein Binomialtest[22] durchgeführt, da unter H_0 bei gegebener Anzahl $h = h_{12} + h_{21}$ (Anzahl der Wechsel der Reaktionen) die Zellhäufigkeit h_{12} bzw h_{21} B(h;0,5)-verteilt ist.

Unter SPSS wird beim Aufruf des McNemar-Tests aus dem Dialogfeld „Crosstabs: Statistics“ (siehe Abb. 2.39) heraus stets die Binomialverteilung zur Bestimmung des (zweiseitigen) exakten Signifikanzniveaus verwendet.

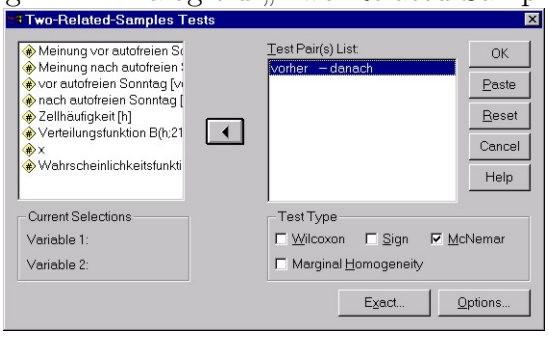
Eine weitere Möglichkeit für den Aufruf des McNemar-Tests ist über

■ Analyze
  ■ Nonparametric Tests
    ■ 2 related Samples...

gegeben.

[22] Vgl. u.a. Bortz, J., Lienert, G.A., Boehnke, K. (1990), S.88 ff.; Clauß, G., Finze, F.-R., Partzsch, L. (1994), S. 192 f.; Rönz, B. (2001),

Abbildung 2.42.: Dialogfeld „Two-Related-Samples Tests“

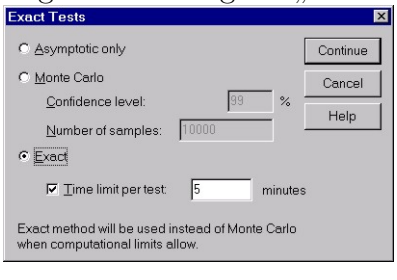


Hier wird

- die Chi-Quadrat-Teststatistik nach (2.80) verwendet, wenn $h_{12} + h_{21} > 25$ ist,
- die Binomialverteilung verwendet, wenn $h_{12} + h_{21} \leq 25$ gilt.

Allerdings kann man sich auch beide ausgeben lassen, wenn man über die Schaltfläche „Exact...“ in das Dialogfeld „Exact Tests“ (Abb. 2.43) geht und auf „Exact“ entscheidet, wobei neben Exact Sig. (2-tailed) auch Exact Sig. (1-tailed) und Point Probability (Wahrscheinlichkeit, genau die beobachtete Zellhäufigkeit h_{12} zu erhalten) im Output erscheint.

Abbildung 2.43.: Dialogfeld „Exact Tests“



• Beispiel 2.13:

40 zufällig ausgewählten Personen wird die Frage gestellt, ob sie für oder gegen autofreie Sonntage sind. Nach einem autofreien Sonntag werden dieselben Personen erneut nach ihrer Meinung (dafür - 1 oder dagegen - 2) befragt. Für jede Person der Stichprobe erhält man ein Beobachtungspaar: Eine Meinungsäußerung vor dem autofreien Sonntag und eine Meinungsäußerung nach dem autofreien Sonntag, wobei beide die gleichen möglichen Ausprägungen aufweisen. Aufgrund der Stichprobenziehung habe sich die nachstehende 2×2 Kontingenztabelle (Vierfeldertafel) ergeben.

Tabelle 2.29.: Kontingenztabelle für Meinung zum autofreien Sonntag

Meinung vor autofreien Sonntag * Meinung nach autofreien Sonntag Crosstabulation

Count

		Meinung nach autofreien Sonntag		
		dafür	dagegen	Total
Meinung vor autofreien Sonntag	dafür	8	5	13
	dagegen	16	11	27
Total		24	16	40

Es interessiert die Frage, ob das Erleben eines autofreien Sonntags eine signifikante Veränderung in der Auffassung verursacht hat. Die 8 bzw. 11 Befragten, die eine gleiche Meinung vor und nach dem autofreien Sonntag haben, sagen nichts über die möglichen Veränderungen in der Auffassung aus. Die wesentlichen Informationen zur Beantwortung der Frage entnimmt man den Zellen (1,2) und (2,1). Auf dem 5% Niveau wird deshalb die Nullhypothese $H_0 : p_{12} = p_{21}$ gegen die Alternativhypothese $H_1 : p_{12} \neq p_{21}$ geprüft. Bei Gültigkeit der H_0 dürfte sich der Anteil derjenigen, deren Meinung von „dafür“ nach „dagegen“ wechselte, nur zufallsbedingt von dem Anteil derjenigen unterscheiden, die erst „dagegen“ und dann „dafür“ sind, und die Häufigkeiten h_{12} und h_{21} sollten sich wie 1:1 verhalten. Dies läßt sich auch mittels der Binomialverteilung prüfen. Unter H_0 und bei gegebener Anzahl $h = h_{12} + h_{21}$ (Anzahl der Wechsel der Meinung) ist die Zufallsvariable H_{12} (Häufigkeit der Zelle (1,2)) binomialverteilt mit den Parametern $h = 21$ und $p = 0,5$. Die Wahrscheinlichkeit, dass diese Zellhäufigkeit höchstens den Wert 5 annimmt, ergibt sich zu

$$P(H_{12} \leq 5 | h = 21, p = 0,5) = \sum_{h_{12}=0}^{5} \binom{21}{h_{12}} \cdot 0,5^{h_{12}} \cdot 0,5^{21-h_{12}} = 0,0133. \tag{2.82}$$

Da die Binomialverteilung für $p = 0,5$ symmetrisch ist, folgt unmittelbar:

$$P(H_{12} \geq 16 | h = 21, p = 0,5) = 0,0133.$$

Für die zweiseitige Hypothesenprüfung resultiert somit ein exaktes Signifikanzniveau von 0,0266. Dies ist im SPSS-Output (gerundet auf drei Dezimalstellen) enthalten.

Tabelle 2.30.: McNemar-Test zum Beispiel: Meinung zum autofreien Sonntag

Chi-Square Tests

	Value	Exact Sig. (2-sided)
McNemar Test		,027[a]
N of Valid Cases	40	

a. Binomial distribution used.[23]

Da Sig. $< \alpha = 0,05$ ist, wird die Nullhypothese abgelehnt. Das Erleben eines autofreien Sonntags hat eine signifikante Veränderung in der Auffassung der Befragten bewirkt. Wenn vor der Testdurchführung die Überzeugung besteht, dass die Praktizierung von autofreien Sonntagen von seinen Vorteilen überzeugt, hätte auch ein einseitiger Test durchgeführt werden können:

$$H_0 : p_{12} \geq p_{21} \qquad H_1 : p_{12} < p_{21}.$$

Exact Sig. in Tabelle 2.30 ist dann zu halbieren und mit dem vorgegebenen Signifikanzniveau zu vergleichen.

Wenn der McNemar-Test über das Dialogfeld „Two-Related-Samples Tests“ (siehe Abb. 2.42) angefordert wird, so erhält man für das Beispiel das gleiche Ergebnis, da $h_{12}+h_{21} \leq 25$ ist und deshalb die Binomialverteilung verwendet wird. Entscheidet man sich dabei im Dialogfeld „Exact Tests“ (siehe Abb. 2.43) für „Exact“, so erhält man neben Exact Sig. (2-tailed) auch Exact Sig. (1-tailed) und Point Probability (die Punktwahrscheinlichkeit $\mathrm{P}(H_{12} = 5)$).

Angenommen, es wurde eine andere Stichprobe vom Umfang $n = 50$ gezogen und es habe sich die Kontingenztabelle der Tabelle 2.31 ergeben.

Tabelle 2.31.: Kontingenztabelle für Meinung zum autofreien Sonntag ($n = 50$)

Meinung vor autofreien Sonntag & Meinung nach autofreien Sonntag

Meinung vor autofreien Sonntag	Meinung nach autofreien Sonntag	
	1	2
1	10	7
2	20	13

In diesem Fall wird die Chi-Quadrat-Teststatistik (2.80) mit Stetigkeitskorrektur verwendet.

[23] Die Fußnote *a*. „Binomial distribution used“ im SPSS-Output mag kontradiktorisch zur Überschrift „Chi-Square Tests“ erscheinen. Offensichtlich handelt es sich aber um eine generelle Überschrift für die aus dem Dialogfeld „Crosstabs: Statistics“ erzeugten Test-Outputs.

Tabelle 2.32.: McNemar-Test über das Dialogfeld „Two-Related-Samples Tests“ für Meinung zum autofreien Sonntag ($n = 50$)

Test Statistics[b]

	Meinung vor autofreien Sonntag & Meinung nach autofreien Sonntag
N	50
Chi-Square[a]	5,333
Asymp. Sig.	,021

a. Continuity Corrected.

b. McNemar Test

Bei Entscheidung für „Exact“ im Dialogfeld „Exact Tests“ (siehe Abb. 2.43) erhält man den nachfolgenden Output.

Tabelle 2.33.: McNemar-Test über das Dialogfeld „Two-Related-Samples Tests“ für Meinung zum autofreien Sonntag ($n = 50$) mit Entscheidung für „Exact“

Test Statistics[b]

	Meinung vor autofreien Sonntag & Meinung nach autofreien Sonntag
N	50
Chi-Square[a]	5,333
Asymp. Sig.	,021
Exact Sig. (2-tailed)	,019
Exact Sig. (1-tailed)	,010
Point Probability	,007

a. Continuity Corrected.

b. McNemar Test

2.2.3.8. Cochran's und Mantel-Haenszel Test

Cochran's und Mantel-Haenszel Test[24] sind Tests auf bedingte Unabhängigkeit zwischen zwei dichotomen Variablen, wobei um die Einflüsse einer nominalen Kontrollvariable Z bereinigt wird. Dabei ist die eine der beiden Variablen eine Faktor- bzw. Gruppierungsvariable und die andere Variable eine Response-Variable.

[24]Siehe u.a. Agresti, A. (1990), S. 230 ff.; Agresti A. (1996), S. 60 ff.

Im Ergebnis der Stichprobenziehung resultiert eine dreidimensionale Kontingenztabelle, die die Häufigkeiten für die Kombinationen der Ausprägungen der drei Variablen enthält. Angenommen die Variable Z weist C Ausprägungen auf, dann erhält man eine 2x2xC Tabelle mit den Häufigkeiten h_{jkc} $(j = 1, 2; k = 1, 2; c = 1, \ldots, C)$.

Das Hauptaugenmerk der Analyse ist auf die Assoziation zwischen X und Y gerichtet. Da die Variable Z die Assoziation zwischen X und Y beeinflussen kann, muss um diesen Einfluß bereinigt (kontrolliert) werden. Dies geschieht, indem jede Schicht c der Kontingenztabelle als eine partielle 2x2 Tabelle angesehen wird, die die X-Y-Assoziation für fest vorgegebene Ausprägungen von Z enthält. In diesem Sinne spricht man von bedingter Assoziation von X und Y. Diese bedingte X-Y-Assoziation kann durchaus verschieden von der Assoziation zwischen X und Y in der 2x2 Randtabelle (marginale Kontingenztabelle) sein. Die 2x2 Randtabelle für die Beziehung zwischen X und Y erhält man durch die Summation der Häufigkeiten gleicher Zellen der partiellen Tabellen:

$$\begin{aligned} h_{11} &= \sum_{c=1}^{C} h_{11c}, \qquad h_{12} = \sum_{c=1}^{C} h_{12c} \\ h_{21} &= \sum_{c=1}^{C} h_{21c}, \qquad h_{22} = \sum_{c=1}^{C} h_{22c}. \end{aligned} \tag{2.83}$$

Damit ist in der 2x2 Randtabelle keine Information bezüglich der Variablen Z mehr enthalten.

Die zu prüfende Nullhypothese lautet: Die Variablen X und Y sind bedingt unabhängig, gegeben die Variable Z, d.h., X und Y sind in jeder partiellen Tabelle unabhängig. Eine äquivalente Formulierung der Nullhypothese ist, dass die bedingten odds ratio $(or_{XY(c)})$ zwischen X und Y

$$or_{XY(c)} = \frac{h_{11c}h_{22c}}{h_{12c}h_{21c}} \tag{2.84}$$

in jeder partiellen Tabelle gleich 1 sind.

Mantel und Haenszel haben einen Test vorgeschlagen, der auf folgenden Überlegungen basiert. In einer partiellen 2x2 Kontingenztabelle mit gegebenen festen Randhäufigkeiten h_{1+c}, h_{2+c}, h_{+1c} und h_{+2c} kann die gemeinsame Häufigkeitsverteilung auf die Angabe der Zellhäufigkeit h_{11c} beschränkt werden, da damit auch die drei anderen Zellhäufigkeiten bestimmt sein, und die Stichprobenverteilung dieser Zellhäufigkeit ist hypergeometrisch (vgl. Anhang E). Bei Gültigkeit der Nullhypothese ergeben sich Mittelwert (erwartete absolute Zellhäufigkeit) und Varianz der Zellhäufigkeit H_{11c} zu:

$$\hat{e}_{11c} = \frac{h_{1+c}h_{+1c}}{h_{++c}} \tag{2.85}$$

$$s^2(h_{11c}) = \frac{h_{1+c}h_{2+c}h_{+1c}h_{+2c}}{h_{++c}^2(h_{++c} - 1)}. \tag{2.86}$$

Für fest vorgegebene Randverteilungen aller partiellen Tabellen sind die Zellhäufigkeiten aus verschiedenen partiellen Tabellen unabhängig, so dass die Zufallsvariable $\sum_c H_{11c}$ den Mittelwert $\sum_c \hat{e}_{11c}$ und die Varianz $\sum_c s^2(h_{11c})$ hat. Als Teststatistik, die die Informationen aller partiellen Tabellen zusammenfaßt, ergibt sich, wobei MH für Mantel-Haenszel steht:

$$MH = \frac{\left(\left|\sum_{c=1}^{C} h_{11c} - \sum_{c=1}^{C} \hat{e}_{11c}\right| - 0,5\right)^2}{\sum_{c=1}^{C} s^2(h_{11c})}. \tag{2.87}$$

Unter der Nullhypothese ist diese Teststatistik approximativ chi-quadrat-verteilt mit einem Freiheitsgrad.

Eine ähnliche Teststatistik stammt von Cochran, wobei jedoch auf die Stetigkeitskorrektur verzichtet wird und nur die Randhäufigkeiten der Gruppierungsvariablen in jeder partiellen Tabelle als fest vorgegeben behandelt werden, so dass die Stichprobenverteilung der Zellhäufigkeit H_{11c} eine Binomialverteilung ist und sich die Varianz zu

$$s^2(h_{11c}) = \frac{h_{1+c}h_{2+c}h_{+1c}h_{+2c}}{h_{++c}^3} \tag{2.88}$$

ergibt.

Die Teststatistiken von Mantel-Haenszel und Cochran nehmen große Werte an, wenn die Differenz $h_{11c} - \hat{e}_{11c}$ in allen partiellen Tabellen entweder große positive oder große negative Werte annimmt. Große Werte der Teststatistik sprechen jedoch gegen die Nullhypothese. Die Tests sollten nicht verwendet werden, wenn sich die Assoziation von X und Y zwischen den partiellen Tabellen drastisch verändert, vor allem die Vorzeichen wechseln.

Unter der letztgenannten Voraussetzung unterbreiten Mantel und Haenszel auch einen Vorschlag für die Schätzung der Stärke der Assoziation zwischen X und Y:

$$or_{XY}(MH) = \frac{\sum_{c=1}^{C} \frac{h_{11c}h_{22c}}{h_{++c}}}{\sum_{c=1}^{C} \frac{h_{12c}h_{21c}}{h_{++c}}}, \tag{2.89}$$

was als gemeinsames odds ratio (common odds ratio) bezeichnet wird.

• Beispiel 2.14:

Bei der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS[25]) wurde u.a.

[25] Der ALLBUS ist in den Jahren 1980-86 von der Deutschen Forschungsgemeinschaft (DFG), ab 1988 von Bund und Ländern über die GESIS (Gesellschaft sozialwissenschaftlicher Infrastruktureinrichtungen) finanziert worden. Er wird vom ZUMA (Zentrum für Umfragen, Methoden und Analysen e.V., Mannheim) und vom Zentralarchiv für Empirische Sozialforschung (Köln) in Zusammenarbeit mit den Mitgliedern des

folgende Frage gestellt: Wie beurteilen Sie ganz allgemein die heutige wirtschaftliche Lage in Deutschland? Diese Response-Variable (Y) wurde in der Weise dichotomisiert, dass positiv - 1 und sonst - 2 sind.

Darüber hinaus wurden aus dem ALLBUS die Gruppierungsvariable (X) Erhebungsgebiet mit 1 - alte Bundesländer und 2 - neue Bundesländer und die Kontrollvariable (Z) Familienstand des Befragten mit 1 - verheiratet zusammenlebend, 2 - verheiratet getrennt lebend, geschieden, verwitwet und 3 - ledig entnommen. Die Beobachtunsdaten dieser Variablen sind in der Datei percept_91_96.sav für die Jahre 1991 und 1996 enthalten.

Es interessiert die Frage, ob zwischen den Variablen X und Y, d.h. zwischen der Perzeption der Wirtschaftslage und dem Erhebungsgebiet, eine Assoziation besteht, wobei für die Variable Familienstand kontrolliert wird. Geprüft wird somit die Nullhypothese: Die Variablen Perzeption der Wirtschaftlage und Erhebungsgebiet sind bedingt unabhängig. Der Test der H_0 erfolgt mittels des Mantel-Haenszel- und des Cochran-Tests auf einem Signifikanzniveau von $\alpha = 0,05$.

Für 1991 resultiert nachstehende $2\times2\times3$ Kontingenztabelle, in der auch die erwarteten absoluten Zellhäufigkeiten nach (2.85) enthalten sind. Aufgrund der Anzahl der Ausprägungen der Kontrollvariablen Familienstand ergeben sich C = 3 partielle Tabellen.

ALLBUS-Ausschusses realisiert. Die vorgenannten Institutionen tragen keine Verantwortung für die Verwendung der Daten in diesem Skript. Alle inhaltlichen Ausführungen zum ALLBUS beziehen sich auf: „Allgemeine Bevölkerungsumfrage der Sozialwissenschaften, ALLBUS 1980-94“, Codebuch, ZA-Nr. 1795, Zentralarchiv für Empirische Sozialforschung an der Universität Köln, Zentrum für Umfragen, Methoden und Analysen Mannheim.

Tabelle 2.34.: Kontingenztabelle für Perzeption der Wirtschaftslage, Erhebungsgebiet und Familienstand (1991)

Erhebungsgebiet 1991*Wirtschaftslage in Deutschland 1991*Familienstand 1991 Crosstabulation

Familienstand				Wirtschaftslage in Dtl. positiv	Wirtschaftslage in Dtl. sonst	Total
verheiratet	Erhebungsgebiet	Alte Bundesl.	Count	595	272	867
			Expected Count	522,6	344,4	867,0
		Neue Bundesl.	Count	552	484	1036
			Expected Count	624,4	411,6	1036,0
	Total		Count	1147	756	1903
			Expected Count	1147,0	756,0	1903,0
getr., gesch., verw.	Erhebungsgebiet	Alte Bundesl.	Count	169	84	253
			Expected Count	138,5	114,5	253,0
		Neue Bundesl.	Count	113	149	262
			Expected Count	143,5	118,5	262,0
	Total		Count	282	233	515
			Expected Count	282,0	233,0	515,0
ledig	Erhebungsgebiet	Alte Bundesl.	Count	209	135	344
			Expected Count	199,4	144,6	344,0
		Neue Bundesl.	Count	108	95	203
			Expected Count	117,6	85,4	203,0
	Total		Count	317	230	547
			Expected Count	317,0	230,0	547,0

Die marginale 2x2 Kontingenztabelle für X und Y enthält die Tabelle 2.35, deren Zellhäufigkeiten sich gemäß (2.83) ergeben.

Tabelle 2.35.: Marginale Kontingenztabelle für Perzeption der Wirtschaftslage und Erhebungsgebiet (1991)

Erhebungsgebiet 1991*Wirtschaftlage in Deutschland 1991 Crosstabulation

			Wirtschaftslage in Deutschland positiv	Wirtschaftslage in Deutschland sonst	Total
Erhebungsgebiet 1991	Alte Bundesländer	Count	973	491	1464
		Expected Count	862,1	601,9	1464,0
	Neue Bundesländer	Count	773	728	1501
		Expected Count	883,9	617,1	1501,0
Total		Count	1746	1219	1965
		Expected Count	1746,0	1219,0	2965,0

Da die Nullhypothese impliziert, dass die bedingten odds ratio ($or_{XY(c)}$) zwischen X und Y in jeder partiellen Tabelle gleich 1 sind, werden die beobachteten bedingten odds ratio für jede Ausprägung des Familienstandes in der Tabelle 2.36 angegeben. Die zur Ermittlung des Testwertes der Mantel-Haenszel- und Cochran-Statistik notwendigen Varianzen der Zellhäufigkeiten H_{11c} sind ebenfalls in der Tabelle 2.36 enthalten. Unter Verwendung der Zellhäufigkeiten aus Tabelle 2.34 lassen sich die odds ratio und Varianzen leicht nachrechnen.

Tabelle 2.36.: Odds ratio und Varianzen der Zellhäufigkeiten H_{11c} (1991)

Familienstand	$or_{XY(c)}$	$s^2(h_{11c})$ nach Mantel-Haenszel	$s^2(h_{11c})$ nach Cochran
c = 1 (verheiratet)	1,9180	113,0774424	113,0180218
c = 2 (getr., gesch., verw.)	2,6529	31,9484111	31,8863754
c = 3 (ledig)	1,3618	31,1655132	31,1085379

Für den Mantel-Haenszel-Test bzw. Cochran-Test erhält man:

$$MH = \frac{(|973 - 860,4615| - 0,5)^2}{176,1914} = 71,244$$
$$CO = \frac{(973 - 860,4614)^2}{176,0129} = 71,954.$$

Für das vorgegebene Signifikanzniveau $\alpha = 0,05$ und die Anzahl der Freiheitsgrade $df = 1$ findet man in der Tabelle der Verteilungsfunktion der Chi-Quadrat-Verteilung (Anhang C) den kritischen Wert $\chi^2_{1;0,95} = 3,841$. Da die Testwerte größer sind als der kritische Wert, wird die Nullhypothese auf bedingte Unabhängigkeit zwischen Perzeption der Wirtschaftslage und Erhebungsgebiet bei beiden Tests auf dem 5%-Niveau abgelehnt. Der Familienstand hat im Jahre 1991 einen signifikanten Einfluß auf die Assoziation zwischen Perzeption der Wirtschaftslage und Erhebungsgebiet.

Aufgrund des Testergebnisses ist u.E. die Angabe eines common odds ratio nach Mantel-Haenszel nicht sinnvoll, soll jedoch für Demonstrationszwecke hier erfolgen, da eine Schätzung des common ratio im SPSS-Output stets enthalten ist:

$$or_{XY}(MH) = \frac{\frac{595 \cdot 484}{1903} + \frac{169 \cdot 149}{515} + \frac{209 \cdot 95}{847}}{\frac{552 \cdot 272}{1903} + \frac{113 \cdot 84}{515} + \frac{108 \cdot 135}{847}} = 1,90768.$$

Diese Ergebnisse sind im nachstehenden SPSS-Output enthalten.

Tabelle 2.37.: Testergebnisse und Schätzung des common odds ratio (1991)

Test for Homogeneity of the Odds Ratio

Statistics		Chi-Squared	df	Asymp. Sig. (2-sided)
Conditional	Cochran's	71,954	1	,000
Independence	Mantel-Haenszel	71,244	1	,000
Homogeneity	Breslow-Day	6,831	2	,033
	Tarone's	6,831	2	,033

Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected ist 0.

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			1,908
ln(Estimate)			,646
Std. Error of ln(Estimate)			,077
Asymp. Sig. (2-sided)			,000
Asymp. 95% Confidence	Common Odds Ratio	Lower Bound	1,642
Intervall		Upper Bound	2,217
	ln(Common Odds Ratio)	Lower Bound	,496
		Upper Bound	,796

The Mantel Heanszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1,000 assumption. So is the natural log of the estimate.

Die Breslow-Day-Statistic und die Tarone-Statistik sind Tests auf Homogenität der odds ratio ($H_0 : or_{XY(1)} = \ldots = or_{XY(C)}$) unter der Voraussetzung, dass die erwarteten Zellhäufigkeiten aus partiellen Tabellen mit den beobachteten Randhäufigkeiten, jedoch mit einem nach Mantel-Haenszel bestimmten common odds ratio $or_{XY(MH)}$ geschätzt wurden. Diese Teststatistiken sind approximativ chi-quadrat-verteilt mit C-1 Freiheitsgraden. Auf ihre Darstellung soll hier nicht näher eingegangen werden.

Für 1996 soll aus Platzgründen auf die Angabe der 2x2x3 Kontingenztabelle sowie die ausführliche Darstellung verzichtet und nur auf den SPSS-Output Bezug genommen werden.

Tabelle 2.38.: Testergebnisse und Schätzung des common odds ratio (1996)

Test for Homogeneity of the Odds Ratio

Statistics		Chi-Squared	df	Asymp. Sig. (2-sided)
Conditional	Cochran's	2,282	1	,131
Independence	Mantel-Haenszel	2,114	1	,146
Homogeneity	Breslow-Day	1,618	2	,445
	Tarone's	1,618	2	,445

Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected ist 0.

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			1,188
ln(Estimate)			,172
Std. Error of ln(Estimate)			,114
Asymp. Sig. (2-sided)			,132
Asymp. 95% Confidence	Common Odds Ratio	Lower Bound	,949
Intervall		Upper Bound	1,487
	ln(Common Odds Ratio)	Lower Bound	-,052
		Upper Bound	,396

The Mantel Heanszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1,000 assumption. So is the natural log of the estimate.

Da das ausgegebene Sig. größer als das vorgegebene Signifikanzniveau $\alpha = 0,05$ ist, kann die Nullhypothese auf bedingte Unabhängigkeit zwischen Perzeption der Wirtschaftslage und dem Erhebungsgebiet auf dem 5%-Niveau nicht abgelehnt werden. Der Familienstand hat im Jahre 1996 keinen signifikanten Einfluß auf die Assoziation zwischen Perzeption der Wirtschaftslage und dem Erhebungsgebiet. Hier macht die Verwendung eines common odds ratio dann auch Sinn.

3. Regressionsanalyse

Die Regressionsanalyse gehört ebenfalls zum Problemkreis der Überprüfung von Zusammenhängen und Abhängigkeiten. Aufgrund ihrer besonderen Stellung bei sozioökonomischen Untersuchungen und des Umfangs der notwendigen Erläuterungen wird sie in einem separaten Kapitel behandelt.

Wie bereits eingangs zum Kapitel 2 erwähnt, ist bei der Untersuchung von Abhängigkeiten zwischen metrisch skalierten Variablen das Interesse auf die gemeinsame Variation der Variablen gerichtet. Die Regressionsanalyse[26] ist im Verbund mit der Korrelationsanalyse dafür eine geeignete statistische Analysemethode. Einen ersten (explorativen) Anhaltspunkt über die gemeinsame Variation und die Form der Beziehung zwischen den Variablen liefern die im Abschnitt 2.1 behandelten Scatterplots.

3.1. Lineare Regression

3.1.1. Modell der linearen Regression

Im weiteren wird davon ausgegangen, dass eine abhängige (zu erklärende, endogene) Variable Y und m unabhängige (erklärende, exogene) Variablen X_k mit $k = 1, \ldots, m$ gegeben sind, und unterstellt, dass sich die Abhängigkeit der Variablen Y von den erklärenden X-Variablen in Form einer linearen Funktion approximieren läßt.

Das wahre Regressionsmodell in der Grundgesamtheit läßt sich wie folgt notieren:

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \ldots + \beta_m x_{im} + u_i, \quad i = 1, \ldots, N. \tag{3.1}$$

Darin sind:
y_i - die Werte der abhängigen Variablen Y_i, $i = 1, \ldots, N$ (N-Umfang der Grundgesamtheit),
x_{i0} - die Werte einer Scheinvariablen (dummy-Variablen) X_0 für den Achsenabschnitt im mehrdimensionalen Koordinatensystem mit $x_{i0} = 1$ für alle i,

[26] Die Regressions- und die Korrelationsanalyse sind in ihren Grundzügen in jedem guten Statistik-Lehrbuch enthalten. Siehe weiterhin Rönz, B., Förster, E. (1992); Judge, G.G. et al. (1988).

x_{ik} - die Werte der erklärenden Variablen X_k, $k = 1, \ldots, m$,
β_0 - die Regressionskonstante,
β_k - der Regressionskoeffizient bei der Variablen X_k, $k = 1, \ldots, m$,
u_i - die Werte der Restgröße U_i.
Die Abhängigkeit der Variablen Y von den X-Variablen manifestiert sich in den Koeffizienten β_k $(k = 1, \ldots, m)$, die jedoch unbekannt sind. Zur Schätzung von (3.1) und zur Prüfung eines der Hypothesenpaare

a) $H_0 : \beta_k = 0 \qquad H_1 : \beta_k \neq 0;$

b) $H_0 : \beta_k \leq 0 \qquad H_1 : \beta_k > 0;$

c) $H_0 : \beta_k \geq 0 \qquad H_1 : \beta_k < 0 \qquad k = 1, \ldots, m$

wird der Grundgesamtheit eine Stichprobe vom Umfang n $(n > m + 1)$ entnommen. Das Stichprobenregressionsmodell lautet:

$$y_i = b_0 x_{i0} + b_1 x_{i1} + \ldots + b_m x_{im} + \hat{u}_i, \quad i = 1, \ldots, n, \tag{3.2}$$

worin b_k $(k = 0, 1, \ldots, m)$ die geschätzten Regressionsparameter und $\hat{u}_i$ $(i = 1, \ldots, n)$ die geschätzten Werte (Residuen) der Restvariablen U_i sind:
Die mit dem Regressionsmodell verbundenen Annahmen sind:

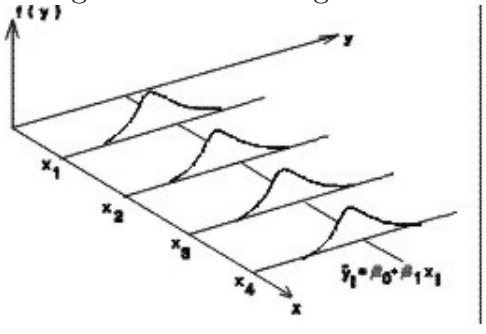
1. Die Werte der X-Variablen sind feste (nichtzufällige) Größen.

2. Die erklärenden X-Variablen umfassen alle wesentlichen Einflußgrößen auf Y, so dass die Restgrößen U_i nur zufallsbedingte Einflüsse beinhalten.

3. Zwischen den Werten der X-Variablen treten keine funktionalen linearen Abhängigkeiten auf (Abwesenheit von extremer Multikollinearität).

4. Die Regressionsfunktion ist linear in den Parametern; bezüglich der erklärenden Variablen braucht dies nicht der Fall zu sein. Diese Linearität bedeutet, dass die wahren Regressionsparameter β_k über alle gegebenen Wertetupel der X-Variablen konstante Werte haben, d.h., die (bedingten) Mittelwerte $\mu_{Y|X}$ liegen auf einer Hyperebene, die die wahre Regressionsebene ist. Für den Fall nur einer erklärenden X-Variablen wird das in der Abb. 3.1 veranschaulicht.

Abbildung 3.1.: Lineare Regressionsfunktion der Grundgesamtheit



5. Die Zufallsvariablen U_i $(i = 1, \ldots, n)$ haben jeweils den Erwartungswert $E(U_i) = 0$ und die Varianz $Var(U_i) = \sigma_i^2 = \sigma_u^2$, d.h. sie ist bei allen n statistischen Einheiten gleich und konstant.

6. Die n Zufallsvariablen U_i sind nicht miteinander korreliert (wahrscheinlichkeitstheoretisch unabhängig voneinander): $COV(U_i U_j) = \sigma_{ij} = 0$ für alle $i, j = 1, \ldots, n$ und $i \neq j$.

7. Die Zufallsvariablen U_i sind normalverteilt. Diese Annahme ist nicht für die Schätzung der unbekannten Parameter notwendig, jedoch für die nachfolgenden Hypothesenprüfungen. Diese Annahme impliziert: Für jedes feste Wertetupel der X-Variablen ist Y ebenfalls eine Zufallsvariable, die normalverteilt ist mit dem Mittelwert $\mu_{y|x}$ und einer konstanten Varianz σ_y^2.

(3.2) zusammen mit diesen Annahmen wird als lineares Regressionsmodell bezeichnet. Ein Teil der Annahmen läßt sich vor der Schätzung, einige Annahmen jedoch erst nach der Schätzung beurteilen.

3.1.2. Schätzung des linearen Regressionsmodells und Hypothesenprüfungen

Die Schätzung der Regressionsparameter erfolgt nach der Methode der kleinsten Quadrate (MkQ), für deren Anwendung keine Annahme über die Verteilung der Störvariablen notwendig ist. Zur Vereinfachung der Darstellung wird im weiteren die Matrixnotation verwendet, so dass sich (3.2) wie folgt schreiben läßt:

$$\mathbf{y} = \mathbf{Xb} + \hat{\mathbf{u}} \tag{3.3}$$

Darin sind:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}; \quad \mathbf{X} = \begin{pmatrix} x_{10} & \cdots & x_{1k} & \cdots & x_{1m} \\ \vdots & \vdots & \vdots & \vdots & \\ x_{i0} & \cdots & x_{ik} & \cdots & x_{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n0} & \cdots & x_{nk} & \cdots & x_{nm} \end{pmatrix}; \quad \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \\ \vdots \\ b_m \end{pmatrix}; \quad \hat{\mathbf{u}} = \begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_i \\ \vdots \\ \hat{u}_n \end{pmatrix}. \tag{3.4}$$

Die Minimierungsforderung der MkQ lautet:

$$Q(\mathbf{b}) = \hat{\mathbf{u}}^T\hat{\mathbf{u}} = (\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{y}^T\mathbf{y} - 2\mathbf{b}^T\mathbf{X}^T\mathbf{y} + \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b} \rightarrow \min. \tag{3.5}$$

Differenzierbarkeit von $Q(\mathbf{b})$ vorausgesetzt, ergeben sich die ersten partiellen Ableitungen $\partial Q(\mathbf{b})$ nach den Parametern $b_0, b_1, \ldots, b_m$ zu

$$\partial Q(\mathbf{b})/\partial \mathbf{b} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{b}.$$

Nullsetzten der ersten partiellen Ableitungen führt zu $m + 1$ Normalgleichungen, die nach den Parametern zu lösen sind:

$$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y}. \tag{3.6}$$

Darin sind, wobei der Laufindex der Summen in den Matrizen $i = 1$ bis n ist:

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & \sum x_{i1} & \cdots & \sum x_{ik} & \cdots & \sum x_{im} \\ \sum x_{i1} & \sum x_{i1}^2 & \cdots & \sum x_{i1}x_{ik} & \cdots & \sum x_{i1}x_{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum x_{ik} & \sum x_{ik}x_{i1} & \cdots & \sum x_{ik}^2 & \cdots & \sum x_{ik}x_{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum x_{im} & \sum x_{im}x_{i1} & \cdots & \sum x_{im}x_{ik} & \cdots & \sum x_{im}^2 \end{pmatrix}; \mathbf{X}^T\mathbf{y} = \begin{pmatrix} \sum y_i \\ \sum x_{i1}y_i \\ \vdots \\ \sum x_{ik}y_i \\ \vdots \\ \sum x_{im}y_i \end{pmatrix}$$

Unter der Voraussetzung, dass die Matrix $\mathbf{X}^T\mathbf{X}$ regulär oder, was gleichwertig ist, der Rang $\mathbf{X} = m + 1$ und somit $\det(\mathbf{X}^T\mathbf{X}) \neq 0$ ist, erhält man als Lösung des Normalgleichungssystems die gesuchten Parameterschätzungen:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{3.7}$$

Mit

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} \tag{3.8}$$

ist der Vektor der geschätzten Regreßwerte (geschätzte Mittelwerte von Y bei gegebenen Werten $x_{i1}, \ldots, x_{im}$ der erklärenden X-Variablen) und mit

$$\mathbf{y} - \hat{\mathbf{y}} = \hat{\mathbf{u}} \tag{3.9}$$

der Vektor der Residuen gegeben.

Zur Güte der Anpassung der Regressionsfunktion

Zur Einschätzung der Güte der Anpassung der Regressionsfunktion an die beobachteten Daten wird u.a. das Bestimmtheitsmaß (R Square, coefficient of determination) verwendet. Die Berechnung des Bestimmheitsmaßes beruht auf der Aufspaltung der Varianz der abhängigen Variablen Y:

$$\begin{aligned}\sum_{i=1}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2=\sum_{i=1}^{n}[(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y}]^2\\ &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2.\end{aligned}$$

Der zweite Term auf der rechten Seite kann wie folgt geschrieben werden:

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) &= \sum_{i=1}^{n}\hat{u}_i(\hat{y}_i-\bar{y})=\sum_{i=1}^{n}\hat{u}_i\hat{y}_i-\bar{y}\sum_{i=1}^{n}\hat{u}_i\\ &= \sum_{i=1}^{n}\hat{u}_i(b_0+b_1x_{i1}+\ldots+b_mx_{im})-\bar{y}\sum_{i=1}^{n}\hat{u}_i\\ &= b_0\sum_{i=1}^{n}\hat{u}_i+b_1\sum_{i=1}^{n}\hat{u}_ix_{i1}+\ldots+b_m\sum_{i=1}^{n}\hat{u}_ix_{im}-\bar{y}\sum_{i=1}^{n}\hat{u}_i=0.\end{aligned}$$

Dieser Term ist gleich Null, da wegen Annahme 5 $\sum_i \hat{u}_i = 0$ ist und aufgrund von Annahmen 1 und 5 die X-Variablen mit den Störvariablen unkorreliert sind. Für die Varianzaufspaltung resultiert somit:

$$\sum_{i=1}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y})^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2. \tag{3.10}$$

Die einzelnen Quadratsummen (sum of squares) haben folgende Bedeutung:

- die Totalquadratsumme (sum of squares total; gesamte Variation in Y)

$$SS-Total=\sum_{i=1}^{n}(y_i-\bar{y})^2,$$

- die Fehler- oder Restquadratsumme (sum of squares residual; die durch das lineare Regressionsmodell nicht erklärte Variation in Y)

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

- die Modellquadratsumme (sum of squares regression; die durch das lineare Regressionsmodell erklärte Variation von Y)

$$SS-Regression=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2.$$

Wie leicht zu sehen ist, gilt $SS - Total = SS - Regression + SS - Residual$.
Das Bestimmheitsmaß ist als der Anteil der durch die Regressionsfunktion erklärten Variation in Y an der Gesamtvariation in Y definiert:

$$R^2 = SS - Regression/SS - Total$$

bzw.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{3.11}$$

Das Bestimmtheitsmaß kann nur Werte im Bereich 0 bis 1 annehmen: $0 \leq R^2 \leq 1$, wobei $R^2 = 0$ bedeutet, dass keine lineare (!) Beziehung existiert.
Um die Hypothese zu testen, dass keine lineare Beziehung zwischen Y und den X-Variablen besteht, wird das Bestimmheitsmaß gegen Null geprüft:

$$H_0 : R^2 = 0 \text{ gegen } H_1 : R^2 > 0.$$

H_0 ist identisch mit der Prüfung aller Regressionskoeffizienten gegen Null: $\beta_k = 0$ für alle k.
Die Testfunktion, wobei MS mean square (mittlere Quadratsumme) bedeutet,

$$F = \frac{MS - Regression}{MS - Residual} = \frac{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{m}}{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - m - 1}} = \frac{R^2(n - m - 1)}{m(1 - R^2)} \tag{3.12}$$

folgt unter H_0 einer F-Verteilung mit $f_1 = m$ und $f_2 = n-m-1$ Freiheitsgraden. Der Ablehungsbereich der H_0 ist $\{F|F > F_{f1;f2;1-\alpha}\}$, wobei $F_{f1;f2;1-\alpha}$ das $(1-\alpha)$-Quantil der F-Verteilung mit f_1 und f_2 Freiheitsgraden ist. Wenn die unter SPSS ausgegebene Überschreitungswahrscheinlichkeit des F-Wertes kleiner ist als das vorgegebene α, so ist die Nullhypothese abzulehnen. Die Ablehnung der Nullhypothese bedeutet, dass die in dem linearen Regressionsmodell enthaltenen X-Variablen in ihrer Gesamtheit einen wesentlichen Teil der Variation von Y erklären. Es beinhaltet jedoch nicht, dass jede einzelne X-Variable einen wesentlichen Beitrag dazu leistet, dazu sind die einzelnen Regressionsparameter gegen Null zu prüfen.

Neben R^2 wird ein korrigiertes Bestimmtheitsmaß R_a^2 (adjusted R-square) zur Einschätzung der Güte der Anpassung herangezogen, weil R^2 allein durch Hinzunahme weiterer Variablen vergrößert werden kann, ohne dass dadurch die Anpassung in sachgerechter Weise verbessert

würde. Dem trägt R_a^2 dadurch Rechnung, dass die Anzahl der Freiheitsgrade berücksichtigt wird, die sich in analoger Weise wie die Totalquadratsumme aufspalten läßt:

$$n - 1 = (n - m - 1) + m.$$

Damit folgt für das korrigierte Bestimmtheitsmaß R_a^2:

$$\begin{aligned} R_a^2 &= 1 - \frac{\frac{SS - Residual}{n - m - 1}}{\frac{SS - Total}{n - 1}} = 1 - (1 - R^2)\frac{n - 1}{n - m - 1} \\ &= R^2 - \frac{m(1 - R^2)}{n - m - 1}. \end{aligned} \tag{3.13}$$

Das korrigierte Bestimmtheitsmaß R_a^2 kann bei Aufnahme weiterer Variablen auch kleiner werden.

Zu den geschätzten Regressionsparametern

Die geschätzten Regressionsparameter variieren von Stichprobe zu Stichprobe, d.h., sie sind Zufallsvariablen. Um konfirmatorische Aussagen über sie treffen zu können, muß die Verteilung dieser Schätzfunktion[27] bekannt sein. Gilt die eingangs angegebene Normalverteilungsannahme der U_i, so sind sie normalverteilt mit dem Erwartungswert β_k. Die Varianz-Kovarianz-Matrix der Regressionsparameter $\sum(b)$ ist wie folgt definiert:

$$\mathbf{\Sigma(b)} = E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})^T]. \tag{3.14}$$

In der Hauptdiagonalen dieser Varianz-Kovarianz-Matrix stehen die Varianzen der Regressionsparameter $\sigma^2(b_k) = E[(b_k - \beta_k)^2]$ und abseits der Diagonalen die Kovarianzen $\sigma^2(b_k b_j) = E[(b_k - \beta_k)(b_j - \beta_j)]$.

Unter Verwendung des wahren Regressionsmodells $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ geht (3.7) über in

$$\begin{aligned} \mathbf{b} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ \mathbf{b} &= \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{u}, \end{aligned}$$

so dass die Abweichung der geschätzten von den wahren Regressionsparametern

$$\mathbf{b} - \boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{u}$$

ist. Dises Ergebnis wird in (3.14) eingesetzt:

$$\mathbf{\Sigma(b)} = E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{u}\mathbf{u}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\mathbf{u}\mathbf{u}^T)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}.$$

[27] Die Schätzfunktionen als Zufallsvariablen werden ebenfalls mit kleinen Buchstaben geschrieben. In den weiteren Ausführungen geht aus dem Kontext hervor, ob es sich um die Zufallsvariable bzw. den Schätzwert handelt.

$E(\mathbf{uu}^T)$ ist die Varianz-Kovarianz-Matrix der Störvariablen U_i. Aufgrund der Annahmen 5 und 6 des linearen Regressionsmodells ($Var(U_i) = \sigma_i^2 = \sigma_u^2$ und $Cov(U_iU_j) = \sigma_{ij} = 0$ für alle $i, j = 1, \ldots, n,\ i \neq j$) folgt $E(\mathbf{uu}^T) = \sigma_u^2\mathbf{I}$ mit $\mathbf{I}$ als Einheitsmatrix. Einsetzen in die Varianz-Kovarianz-Matrix der Regressionsparameter führt zu

$$\mathbf{\Sigma}(\mathbf{b}) = \sigma_u^2(\mathbf{X}^T\mathbf{X})^{-1}. \tag{3.15}$$

Da die Varianz der Störvariablen σ_u^2 unbekannt ist, muss sie aus der Stichprobe geschätzt werden:

$$S_u^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n - m - 1}. \tag{3.16}$$

Mit s_u^2 als Schätzwert erhält man für die geschätzte Varianz-Kovarianz-Matrix der Regressionsparameter:

$$\mathbf{S}(\mathbf{b}) = s_u^2(\mathbf{X}^T\mathbf{X})^{-1}. \tag{3.17}$$

Die Wurzel aus den Diagonalelementen dieser Matrix sind die Standardabweichungen (Standardfehler, standard error) $s(b_k)$ der geschätzten Regressionsparameter, $k = 0, 1, \ldots, m$. Sie beinhalten die mittlere Streuung der möglichen Schätzwerte des jeweiligen Regressionsparameters um seinen wahren (aber unbekannten) Wert in der Grundgesamtheit.

In einem multiplen Regressionsmodell können standardisierte Regressionskoeffizienten b_k^* ($k = 1, \ldots, m$) berechnet werden, um die relative Bedeutung jeder X-Variablen einzuschätzen. Sie ergeben sich als Schätzungen eines linearen Regressionsmodells mit den standardisierten Variablen

$$\frac{y_i - \bar{y}}{s_y}, \qquad \frac{x_{ik} - \bar{x}_k}{s_k}$$

Man kann sie jedoch auch mittels der geschätzten Regressionskoeffizienten b_k durch die Multiplikation des Regressionskoeffizienten mit dem Quotienten aus der Standardabweichung der jeweiligen erklärenden Variablen und der Standardabweichung der abhängigen Variablen berechnen:

$$b_k^* = b_k\frac{s_k}{s_y} \tag{3.18}$$

Sie sind dimensionslose Koeffizienten und eignen sich daher für den Vergleich. Sie hängen jedoch von den tatsächlich im Regressionsmodell enthaltenen X-Variablen ab und sind durch deren Korrelation beeinflußt. Im Fall eines linearen Regressionsmodell mit nur einer erklärenden X-Variablen ist der standardisierte Regressionskoeffizient b_1^* identisch mit dem Korrelationskoeffizienten r_{XY}.

Zur Prüfung von Hypothesen über die Regressionskoeffizienten gegen Null wird die Testfunktion

$$T_k = \frac{b_k}{s(b_k)} \quad k = 0, 1, \ldots, m \tag{3.19}$$

verwendet. Soll jedoch gegen einen anderen hypothetischen Wert β_k^0 geprüft werden, so ist

$$T_k = \frac{b_k - \beta_k^0}{s(b_k)} \quad k = 0, 1, \ldots, m \tag{3.20}$$

zu verwenden. Die Testfunktion (3.19) bzw. (3.20) folgt unter H_0 einer t-Verteilung mit $f = n - m - 1$ Freiheitsgraden unter der Voraussetzung der Gültigkeit der Annahme 7 (Normalverteilung der Störvariablen U_i). Als Ablehnungsbereich der H_0 ergibt sich

a) beim zweiseitigen Test ($H_0 : \beta_k = 0$ gegen $H_1 : \beta_k \neq 0$) : $\{t | t < -t_{f;1-\alpha/2}$ oder $t > t_{f;1-\alpha/2}\}$,

b) beim rechtsseitigen Test ($H_0 : \beta_k \leq 0$ gegen $H_1 : \beta_k > 0$) : $\{t | t > t_{f;1-\alpha}\}$,

c) beim linksseitigen Test ($H_0 : \beta_k \geq 0$ gegen $H_1 : \beta_k < 0$) : $\{t | t < -t_{f;1-\alpha}\}$,

wobei $t_{f,1-\alpha/2}$ bzw. $t_{f,1-\alpha}$ Quantile der t-Verteilung mit f Freiheitsgraden sind.

Der Test der Regressionskoeffizienten gegen Null wird unter SPSS ausgeführt. Wenn die ausgegebene Überschreitungswahrscheinlichkeit des t-Wertes kleiner ist als das vorgegebene α, so ist H_0 abzulehnen. Zum vorgegebenen Konfidenzniveau $1 - \alpha$

$$P[b_k - t_{n-m-1;1-\frac{\alpha}{2}} \cdot s(b_k) \leq \beta_k \leq b_k + t_{n-m-1;1-\frac{\alpha}{2}} \cdot s(b_k)] = 1 - \alpha \tag{3.21}$$

lassen sich Konfidenzintervalle für die Regressionskoeffizienten bestimmen:

$$[b_k - t_{n-m-1;1-\frac{\alpha}{2}} \cdot s(b_k);\ b_k + t_{n-m-1;1-\frac{\alpha}{2}} \cdot s(b_k)] \tag{3.22}$$

Zu den Regreßwerten

Die Regreßwerte (vorhergesagten Werte) erhält man nach der Schätzung der Regressionsparameter und der Residuen über $\hat{\mathbf{y}} = \mathbf{Xb}$ (3.8) oder $\hat{\mathbf{y}} = \mathbf{y} - \hat{\mathbf{u}}$. Sie sind Zufallsvariablen und streuen zufallsbedingt um die wahren Regreßwerte $\tilde{y}_i$. Mit $\hat{\mathbf{y}} = \mathbf{Xb}$ und $\tilde{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}$ lassen sich die Abweichungen der geschätzten Regreßwerte von den wahren Regreßwerten wie folgt schreiben:

$$\hat{\mathbf{y}} - \tilde{\mathbf{y}} = \mathbf{X}(\mathbf{b} - \boldsymbol{\beta})$$

Die Varianzen und Kovarianzen der Regreßwerte sind in ihrer Varianz-Kovarianz-Matrix wie folgt gegeben:

$$\begin{aligned}
\boldsymbol{\Sigma}(\hat{\mathbf{y}}) &= E[(\hat{\mathbf{y}} - \tilde{\mathbf{y}})(\hat{\mathbf{y}} - \tilde{\mathbf{y}})^T] = E[\mathbf{X}(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})^T\mathbf{X}^T] \\
&= \mathbf{X}E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})^T]\mathbf{X}^T \\
&= \mathbf{X}\boldsymbol{\Sigma}(\mathbf{b})\mathbf{X}^T \\
&= \sigma_u^2\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T.
\end{aligned}$$

Daraus folgt mit der geschätzten Varianz der Residuen

$$\mathbf{S}(\hat{\mathbf{y}}) = s_u^2 \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T. \tag{3.23}$$

Als Standardfehler der Regreßwerte bezeichnet man die Wurzel aus $s^2(\hat{y}_i)$, die in der Hauptdiagonale von $\mathbf{S}(\hat{\mathbf{y}})$ stehen.
Zum vorgegebenen Konfidenzniveau $1-\alpha$

$$P[\hat{y}_i - t_{n-m-1;1-\frac{\alpha}{2}} \cdot s(\hat{y}_i) \leq \ \tilde{y}_i \ \leq \hat{y}_i + t_{n-m-1;1-\frac{\alpha}{2}} \cdot s(\hat{y}_i)] = 1-\alpha \tag{3.24}$$

erhält man Konfidenzintervalle für die Regreßwerte:

$$[\hat{y}_i - t_{n-m-1;1-\frac{\alpha}{2}} \cdot s(\hat{y}_i); \hat{y}_i + t_{n-m-1;1-\frac{\alpha}{2}} \cdot s(\hat{y}_i)]. \tag{3.25}$$

Zur Vorhersage von individuellen Werten der Variablen Y
Bei der Vorhersage von individuellen Werten der Variablen Y sind Fehlerquellen sowohl die geschätzten Regressionskoeffizienten, die nicht mit den wahren Regressionsparametern übereinstimmen, als auch die Störvariable U_i. Für die Varianz der vorhergesagten Werte y_p erhält man dementsprechend

$$s^2(y_p) = s^2(\hat{y}_i) + s_u^2 = s_u^2[1 + \mathbf{x}_p^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_p], \tag{3.26}$$

worin $\mathbf{x}_p$ der Vektor der Werte der X-Variablen für die Vorhersage an der Stelle p ist. Der Standardfehler (auch als Prognosefehler bezeichnet) ist die Wurzel daraus.
Zum vorgegebenen Konfidenzniveau $1-\alpha$

$$P[\hat{y}_p - t_{n-m-1;1-\frac{\alpha}{2}} \cdot s(y_p) \leq \ y_p \ \leq \hat{y}_p + t_{n-m-1;1-\frac{\alpha}{2}} \cdot s(y_p)] = 1-\alpha \tag{3.27}$$

erhält man Konfidenzintervalle für individuelle Werte von Y:

$$[\hat{y}_p - t_{n-m-1;1-\frac{\alpha}{2}} \cdot s(y_p); \hat{y}_p + t_{n-m-1;1-\frac{\alpha}{2}} \cdot s(y_p)]. \tag{3.28}$$

3.1.3. Zur Modelldiagnose

Ein wesentlicher Teil der Modellüberprüfung bezieht sich auf die Analyse der Residuen, d.h., kann erst nach der Schätzung der Regressionsfunktion durchgeführt werden. Diese Modelldiagnostik kann durch eine (vor allem graphische) Überprüfung der Residuen, durch weitere Maßzahlen sowie durch geeignete Tests erfolgen.

Die Residuen, definiert gemäß (3.9), sollten bei einem korrekt spezifizierten Modell gemäß den Annahmen 5 bis 7 zumindest approximativ normalverteilt mit $E(\hat{\mathbf{u}}) = \mathbf{0}$ und geschätzter Varianz-Kovarianz-Matrix $E(\mathbf{u}\mathbf{u}^T) = s_u^2\mathbf{I}$ sein, wobei $\mathbf{I}$ die Einheitsmatrix ist. Explorative Werkzeuge zur Entdeckung von gravierenden Abweichungen der Verteilung der Residuen von der Normalverteilung sind u.a. Wahrscheinlichkeitsplots (Q-Q-Plots oder P-P-Plots)

der Residuen und Scatterplots der Residuen gegen Y oder gegen jeweils eine X-Variable. Ein Test auf Normalverteilung der Residuen ist z.B. der Kolmogorov-Smirnov-Test[28]. Bei Verletzung dieser Annahme des Regressionsmodells können geeignete Transformationen der Variablen Y oftmals eine bessere Approximation an die Normalverteilung herbeiführen.

Ist eine bestimmte Ordnung der Fälle gegeben, wie z.B. bei Zeitreihen, so kann in den Residuen Autokorrelation auftreten, die es zu überprüfen gilt. Dazu wird im allgemeinen der Durbin-Watson-Test verwendet. Der Test unterstellt einen autoregressiven Prozeß erster Ordnung für die Störvariable U_i:

$$u_i = \varrho\, u_{i-1} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.29}$$

worin n der Stichprobenumfang und ϱ der Autokorrelationskoeffizient sind. Für den Autokorrelationskoeffizienten gilt $-1 \leq \varrho \leq +1$.
Es wird die Nullhypothese $H_0 : \varrho = 0$ gegen die Alternativhypothese $H_1 : \varrho \neq 0$ geprüft. Die Testvariable lautet:

$$d = \frac{\sum_{i=2}^{n} (\hat{u}_i - \hat{u}_{i-1})^2}{\sum_{i=1}^{n} \hat{u}_i^2}. \tag{3.30}$$

Für großes n gilt: $d \approx 2(1 - \varrho)$. d nimmt somit bei Abwesenheit von Autokorrelation den Wert 2, bei vollständiger positiver Autokorrelation den Wert 0 und bei vollständig negativer Autokorrelation den Wert 4 an. Die Entscheidungsbereiche für den Test zeigt folgende Abbildung.

Abbildung 3.2.: Entscheidungsbereiche des Durbin-Watson-Tests

Ablehnung von H_0, Annahme von H_1:		Annahme von H_0:		Ablehnung von H_0, Annahme von H_1:
Positive Autokorrelation 1. Ordnung der Residuen	?	keine Autokorrelation der Residuen	?	Negative Autokorrelation 1. Ordnung der Residuen

0 $\quad d_u \quad d_o \quad 4 - d_o \quad 4 - d_u \quad$ 4

Die Werte d_u und d_o sind von der vorgegebenen Irrtumswahrscheinlichkeit α, dem Stichprobenumfang n und der Anzahl m der erklärenden X-Variablen abhängig. Sie liegen für ausgewählte

[28] Vgl. u.a. Rönz, B. (2001)

Werte von α, n und m in Tabellen vor. Keine Testentscheidung ist für $d_u \leq d \leq d_o$ und $4 - d_o \leq d \leq 4 - d_u$ möglich.

Eine weitere Diagnosemaßzahl ist der Leverage (Hebelwert, Einfluß), der auf der sogenannten Projektions- oder Hat-Matrix basiert.[29] Die Regreßwerte ergaben sich gemäß Formel (3.8) als $\hat{\mathbf{y}} = \mathbf{Xb}$. Ersetzt man darin den Vektor $\mathbf{b}$ durch Formel (3.7), d.h. $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, so erhält man

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{3.31}$$

Die darin enthaltene Matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ wird als Hat-Matrix bezeichnet, da sie angibt, wie die beobachteten y-Werte in die „y-hats“ (Regreßwerte) überführt werden.
Das i-te Element der Diagonale

$$h_i = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i \qquad (i = 1, \ldots, n) \tag{3.32}$$

dieser Matrix $\mathbf{H}$ gibt den Einfluß (leverage) der i-ten Beobachtungen der erklärenden Variablen $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{im})$ auf die geschätzten Parameter und die Reduktion der Varianz der geschätzten Parameter an. Diese Werte werden deshalb als leverage bezeichnet. Es gilt: $0 \leq h_i \leq 1$.
Je weiter ein Punkt $\mathbf{x}_i$ von der Mehrheit der Beobachtungspunkte entfernt liegt, um so größer ist sein Einfluß und um so größer ist h_i. Wenn alle Beobachtungen der X-Variablen relativ nahe an ihren jeweiligen Mittelwerten liegen, wird kaum eine präzise Schätzung des Effektes der Veränderung der erklärenden X-Variablen auf die abhängige Variable Y möglich sein. Umgekehrt tragen Beobachtungen der X-Variablen, die weit entfernt von ihren jeweiligen Mittelwerten liegen, viel dazu bei. Beobachtungspunkte $\mathbf{x}_i$ mit großen h_i, z.B. größer als das Zwei- oder Dreifache des Mittelwertes m/n, sollten dahingehend überprüft werden, ob sie auf Erfassungsfehler zurückzuführen sind, denn wenn sie einen großen Einfluß ausüben, sollten sie korrekt sein. Wenn es mehrere Beobachtungspunkte $\mathbf{x}_i$ mit hohem h_i gibt, dann kann dies auch ein Anhaltspunkt dafür sein, daß das lineare Regressionsmodell, zumindest für einen Teil der Beobachtungen, nicht adäquat ist.

Da die Residuen $\hat{u}_i$ von der Maßeinheit der Variablen Y abhängen, sagen ihre absoluten Werte nicht sehr viel über das Auftreten von großen Fehlern aus. Um Ausreißer (outliers) zu entdecken, führt man eine Standardisierung der Residuen durch. Neben den Residuen (3.9) wurde eine Reihe weiterer Residuen definiert, z.B.

- die standardisierten Residuen

$$r_i = \hat{u}_i / s_u, \tag{3.33}$$

  wobei s_u die Standardabweichung der Residuen als Wurzel aus (3.16) ist.

[29] Für eine detaillierte Diskussion siehe u.a. Judge, G.G. et. al. (1988), S. 892 ff.

- die studentisierten Residuen

$$r_{i;(s)} = \frac{\hat{u}_i}{s_u\sqrt{1-h_i}}. \tag{3.34}$$

Im Nenner von (3.34) ist neben dem Standardfehler der Residuen auch der Leverage h_i enthalten, d.h., im Nenner steht die Standardabweichung des jeweiligen Residuums, die von Beobachtungspunkt zu Beobachtungspunkt in Abhängigkeit der Distanz die Beobachtungswerte $\mathbf{x}_i$ der erklärenden X-Variablen von ihren Mittelwerten verschieden sein kann.

- die ausgeschlossenen Residuen (deleted residuals)

$$r_{i;(d)} = \frac{\hat{u}_i}{\sqrt{1-h_i}}. \tag{3.35}$$

Diese Residuen ergeben sich, wenn man die Schätzung der Regressionsfunktion ohne den i-ten Beobachtungspunkt (d.h. durch Streichung der i-ten Zeile von $\mathbf{X}$ und des i-ten Elements von $\mathbf{y}$) durchführt. Man erhält bei Ausschluß der i-ten Beobachtung eine Schätzung $\mathbf{b}(i)$ der Regressionsparameter und $s_u^2(i)$ der Varianz der Residuen. Anschließend werden für alle Beobachtungspunkte (einschließlich des i-ten Beobachtungspunktes) die Regreßwerte $\hat{y}_i = \mathbf{x}_i^T\mathbf{b}(i)$ und die Residuen $y_i - \mathbf{x}_i^T\mathbf{b}(i)$ ermittelt. Diese sogenannten ausgeschlossenen Residuen lassen sich jedoch nach (3.35) ermitteln.

- die studentisierten ausgeschlossenen Residuen

$$r_{i;(sd)} = \frac{\hat{u}_i}{s_u(i)\sqrt{1-h_i}} \tag{3.36}$$

mit

$$s_u^2(i) = \frac{(n-m-1)s_u^2}{n-m-2} - \frac{\hat{u}_i^2}{(n-m-2)(1-h_i)}. \tag{3.37}$$

Wenn die Annahme der Normalverteilung zutrifft, dann sind die Residuen t-verteilt mit $n-m-2$ Freiheitsgraden. Wenn $r_{i;(sd)}$ auftreten, die größer als z.B. 2 sind, so werden sie als Ausreißer betrachtet und können ein Indiz für nicht modellierte Aspekte sein, die auf diese Beobachtung zutreffen. Zuviele Ausreißer weisen wiederum auf Abweichungen von der Normalverteilung hin.

Diese Residuen sollten gegen die Regreßwerte $\hat{y}_i$, gegen die Werte jeder erklärenden Variablen und/oder gegen die Werte bisher nicht einbezogener X-Variablen geplottet werden. Zeigen sich

dabei deutliche Muster, so deutet dies auf eine Fehlspezifikation des Modells bzw. „ungewöhnliche“ Beobachtungen hin.

Obwohl die Differenz zwischen den studentisierten Residuen und den ausgeschlossenen Residuen den Einfluß der i-ten Beobachtung angibt, reflektiert sie nicht die Veränderungen in den anderen Residuen, wenn die i-te Beobachtung ausgeschlossen wird. Cook's Distanz erfaßt die Veränderungen, die bei Ausschluß der i-ten Beobachtung bei allen Residuen auftritt:

$$C(i) = \frac{\sum_{k=1}^{n}(\hat{y}_k(i) - \hat{y}_k)^2}{(n-m)s_u^2}. \tag{3.38}$$

Weiter Analysemöglichkeiten bieten die DFBETAS und DFFITS.

- DFBETA(i) gibt die Differenz zwischen dem Schätzer $\mathbf{b}$ bei Einbeziehung aller Beobachtungen und dem Schätzer $\mathbf{b}(i)$ bei Ausschluß der i-ten Beobachtung an:

$$DFBETA(i) = \mathbf{b} - \mathbf{b}(i) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\hat{u}_i/(1-h_i). \tag{3.39}$$

- Da die Regressionsparameter von den Maßeinheiten der Variablen abhängen, ergibt sich die Notwendigkeit einer angemessenen Skalierung, die sich jedoch aus diesem Grund auf den einzelnen Parameter bezieht. Es ergeben sich die SDBETAS. Es bezeichnen: $b_j - b_j(i)$ das j-te Element von (3.39), a_{jj} das j-te Diagonalelement von $(\mathbf{X}^T\mathbf{X})^{-1}$ und c_{ji} das (j,i)-te Element von $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. $b_j - b_j(i)$ wird maßstabsunabhängig gemacht, indem durch die Standardabweichung von b_j, die Wurzel des j-ten Elementes $s_u^2(i)(\mathbf{X}^T\mathbf{X})^{-1}$, vgl. (3.17), dividiert wird:

$$SDBETA_{ji} = \frac{b_j - b_j(i)}{s_u^2(i)\sqrt{a_{jj}}} = \frac{c_{ji}r_{i;(sd)}}{\sqrt{a_{jj}(1-h_i)}}. \tag{3.40}$$

  Große studentisierte ausgeschlossene Residuen und/oder h_i-Werte rufen große $SDBETA_{ji}$ hervor. Als Kriterium für groß, was eine weitere Untersuchung erforderlich macht, gilt $|SDBETA_{ji}| > 2/\sqrt{n}$. Die Division durch $\sqrt{n}$ erfolgt, um dem geringeren Einfluß einzelner Beobachtungen bei größeren Stichprobenumfängen Rechnung zu tragen.

- Analog werden die theoretischen oder erwarteten Werte $\hat{y}_i$ mittels der DFFITS und SDFITS analysiert, um festzustellen, welche Auswirkung die Herausnahme der i-ten Beobachtung hat. Es ist

$$DFFIT_i = \hat{\mathbf{y}}_\mathbf{i} - \hat{\mathbf{y}}_\mathbf{i}(i) = \mathbf{x}_i^T[\mathbf{b} - \mathbf{b}(i)] = h_i\hat{u}_i/(1-h_i) \tag{3.41}$$

Die Skalierung erfolgt, indem (3.41) durch die Standardabweichung von $\hat{y}_i$, d.h. $s_u(i)h_i^{1/2}$, dividiert wird:

$$SDFIT_i = \frac{DFFIT_i}{s_u(i)\sqrt{h_i}} = \frac{\sqrt{h_i}\,\hat{u}_i}{s_u(i)(1-h_i)} = r_{i;(sd)}\sqrt{\frac{h_i}{1-h_i}}. \tag{3.42}$$

Kriterium für große SDFITS ist $|SDFIT_i| > 2(m/n)^{1/2}$.

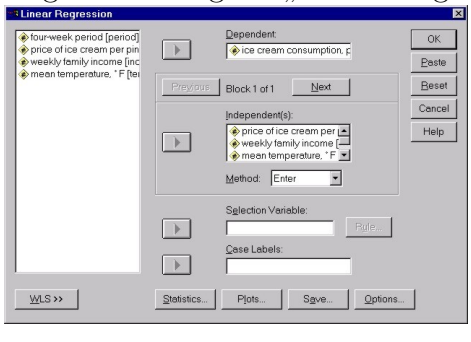
3.2. Die Optionen unter SPSS

Nachfolgend wird die lineare Regressionsanalyse unter SPSS mit den damit verbundenen Entscheidungen dargestellt. Der Aufruf erfolgt über

- Analyze
  - Regression
    - Linear ...

In dem sich öffnenden Dialogfeld

Abbildung 3.3.: Dialogfeld „Linear Regression“



wird als erstes die abhängige Variable aus der linken Quellliste in das Feld „Dependent:“ gebracht. Bezüglich der erklärenden X-Variablen gibt es zwei Möglichkeiten ihrer Aufnahme in die lineare Regressionsfunktion:

- Alle entsprechenden X-Variablen bringt man nacheinander aus der linken Quellliste in das Feld „Independent(s):“. Sie werden dann als ein einziger Block betrachtet.

- Die in die Regressionsfunktion aufzunehmenden X-Variablen können in verschiedene Blöcke unterteilt werden, z.B. entsprechend ihres fachwissenschaftlichen Inhalts oder nach anderen Kriterien. Es erfolgt dann die Auswahl der Variablen des ersten Blockes in das Feld „Independent(s):“. Danach wird die Schaltfläche „Next“ betätigt, wodurch die davor stehende Ausschrift zu „Block 2 of 2“ wechselt. Nunmehr werden die X-Variablen

des 2. Blocks aus der Quellliste in das Feld „Independent(s):“ gebracht. Dies wird solange fortgesetzt, bis alle Blöcke mit ihren X-Variablen vereinbart sind. Es besteht jederzeit die Möglichkeit, durch die Betätigung der Schaltfläche „Previous“ in die vorherigen Blöcke zurückzukehren, um dort eventuell notwendige Korrekturen vorzunehmen. Für die Schätzung der Regressionsfunktion beginnt SPSS mit den Variablen des ersten Blocks entsprechend der ausgewählten Methode. Mit der sich zuletzt ergebenden Regressionsfunktion dieses Blocks beginnt die Hinzufügung der Variablen des zweiten Blocks entsprechend der in diesem Block ausgewählten Methode usw.

Wurden die X-Variablen ausgewählt, muss noch eine Entscheidung über die Methode ihrer Aufnahme in die Regressionsfunktion getroffen werden. Dabei kann je Block durchaus eine andere Methode gewählt werden. Es stehen folgende Selektionsmethoden zur Verfügung:

- Enter (Einschluß):
  Alle X-Variablen eines Blocks werden in einem Schritt in die Regressionsfunktion aufgenommen. Bei mehreren Blöcken von X-Variablen werden die Variablen eines Blocks mit der Methode Einschluß zu der sich zu letzt ergebenden Regressionsfunktion des vorherigen Blocks als Gruppe hinzugefügt.

- Forward (vorwärts):
  Als erste X-Variable für die Aufnahme in die Regressionsfunktion wird diejenige mit der größten (absoluten) Korrelation mit der Y-Variablen ausgewählt. Es wird dann ein F-Test zur Prüfung des Regressionskoeffizienten dieser X-Variable auf Null durchgeführt. Wird dabei das Einschlusskriterium (siehe weiter unten bei der Behandlung der Optionen) erfüllt, verbleibt diese X-Variable in der Regressionsfunktion. Der nächste Aufnahmekandidat ist diejenige X-Variable mit der größten (absoluten) partiellen Korrelation mit der Y-Variablen bei Kontrolle für die bereits in der Regressionsfunktion enthaltenen X-Variablen. Dies ist äquivalent mit der Auswahl derjenigen X-Variablen mit dem größten F-Wert. Ist das Einschlusskriterium erfüllt, wird die X-Variable aufgenommen. Dann erfolgt die Auswahl der nächsten X-Variablen. Die Prozedur bricht ab, wenn keine X-Variable mehr aufgenommen werden kann.

- Backward (rückwärts):
  Diese Selektionsmethode beginnt mit einer Regressionsfunktion, die alle ausgewählten X-Variablen enthält. Es wird dann diejenige X-Variable mit der kleinsten (absoluten) partiellen Korrelation mit der Y-Variablen gesucht (entspricht derjenigen X-Variablen mit dem kleinsten F-Wert). Bei Erfüllung des Ausschlusskriteriums (siehe weiter unter bei der Behandlung der Optionen) wird die X-Variable aus der Regressionsfunktion heraus-

genommen. Anschließend wird in gleicher Weise der nächste Ausschlusskandidat gesucht. Die Prozedur bricht ab, wenn keine X-Variable mehr ausgeschlossen werden kann.

- Stepwise (schrittweise):
  Diese Selektionsmethode beinhaltet einen ständigen Wechsel zwischen Forward- und Backward-Selektion. Die erste X-Variable wird entsprechend der Vorgehensweise von Forward ausgewählt. Wurde sie in die Regressionsfunktion aufgenommen, schließt sich die Backward-Selektion mit ihren Regeln an. Dann wird nach der Forward-Selektion die nächste X-Variable für die Aufnahme ausgewählt. Anschließend werden alle X-Variablen in der Regressionsfunktion nach dem Kriterium der Backward-Selektion geprüft. Dabei kann es vorkommen, dass eine bereits aufgenommene X-Variable wieder aus der Regressionsfunktion entfernt wird, da sie durch die Hinzunahme weiterer X-Variablen nicht mehr das Kriterium des Verbleibs in der Regressionsfunktion erfüllt. Die Prozedur bricht ab, wenn keine X-Variable mehr aufgenommen oder ausgeschlossen werden kann.

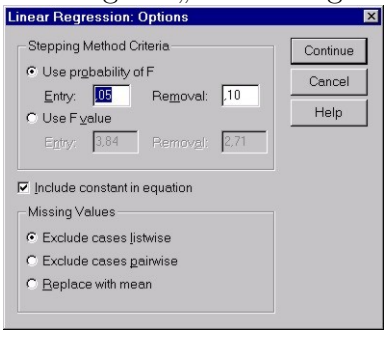
- Remove (Ausschluß):
  Alle X-Variablen eines Blocks werden in einem Schritt aus der Regressionsfunktion herausgenommen. Damit ist klar, dass diese Methode nicht im 1. Block, sondern nur in nachfolgenden Blöcken verwendet werden kann.

Die Verwendung verschiedener Selektionsmethoden auf die gleichen X-Variablen kann durchaus zu einer unterschiedlichen Auswahl von X-Variablen führen. Keine dieser Selektionsmethoden kann als die „beste“ angesehen werden. Allerdings ist Stepwise die am häufigsten verwendete Selektionsmethode.

Bei Verwendung der Selektionsmethoden Stepwise, Forward bzw. Backward müssen noch die Kriterien für den Ein- und Ausschluß von X-Variablen festgelegt werden. Dies geschieht in dem über die Schaltfläche „Options...“ zu erreichenden Dialogfeld (siehe Abb. 3.4).

Zur Festlegung der Kriterien für den Ein- und Ausschluß von X-Variablen gibt es zwei Möglichkeiten.

Abbildung 3.4.: Dialogfeld „Linear Regression: Options“



- Es wird das Signifikanznivau des F-Tests (Use probability of F) vereinbart. Voreingestellt sind für die Aufnahme von X-Variablen $\alpha = 0,05$, als PIN (probability of F-to-enter) im SPSS-Output bezeichnet, und für den Ausschluß von X-Variablen $\alpha = 0,10$, als POUT (probability of F-to-remove) im SPSS-Output bezeichnet. Diese Werte können verändert werden. Der Aufnahmewert (Entry) muss dabei immer kleiner sein als der Ausschlusswert (Removal). Eine X-Variable wird nur in die Regressionsfunktion aufgenommen, wenn die mit dem F-Test verbundene Wahrscheinlichkeit kleiner oder gleich dem ausgewählten Signifikanzniveau ist. Eine X-Variable wird aus der Regressionsfunktion herausgenommen, wenn die mit dem F-Test verbundene Wahrscheinlichkeit größer als das ausgewählte Signifikanzniveau ist.

- Es wird der Wert der Testgröße F vereinbart (Use F value). Voreingestellt sind für die Aufnahme von X-Variablen F = 3,84, als FIN (F-to-enter) im SPSS-Output bezeichnet, und für den Ausschluß von X-Variablen F = 2,71, als FOUT (F-to-remove) im SPSS-Output bezeichnet. Der Aufnahmewert muß dabei immer größer sein als der Ausschlusswert. Eine X-Variable wird in die Regressionsfunktion aufgenommen, wenn der Wert von F größer oder gleich dem gewählten Aufnahmewert ist. Eine X-Variable wird aus der Regressionsfunktion herausgenommen, wenn der Wert von F unter dem gewählten Ausschlusswert liegt.

  Zu beachten ist, dass zu einem festgelegten F-Wert in den einzelnen Schritten der Selektionsprozedur unterschiedliche Signifikanzniveaus gehören können, da sich durch die Aufnahme bzw. Herausnahme einer X-Variablen die Anzahl der Freiheitsgrade f_1 und f_2 des F-Tests verändern.

Eine weitere Vereinbarung für die Spezifikation der Regressionsfunktion, die im Dialogfeld „Linear Regression: Options“ zu treffen ist, betrifft die Aufnahme oder den Ausschluss einer Konstanten. Voreingestellt ist eine Regressionsfunktion mit Konstante (siehe Abb. 3.4). Zur Unterdrückung der Regressionskonstanten muss das Kreuz vor „Include constant in equation“ durch

Anklicken mit der Maus entfernt werden. Es wird dann eine Regressionsfunktion geschätzt, die durch den Ursprung des m-dimensionalen Koordinatensystems führt. Allerdings sind in diesem Fall das Bestimmheitsmaß R^2 und einige andere Resultate der Schätzung nicht mehr wie bei einer Regressionsfunktion mit Konstante zu interpretieren.

Über das Dialogfeld „Linear Regression: Options“ kann auch die Behandlung von Missing-Values gesteuert werden:

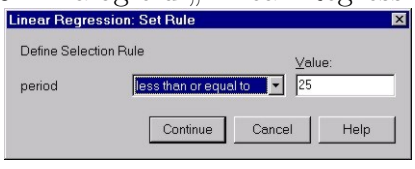
- Exclude cases listwise (fallweiser Ausschluss):
  Es werden nur Fälle mit gültigen Werten für alle Variablen in die Schätzung einbezogen.

- Exclude cases pairwise (paarweiser Ausschluss):
  Bei der Berechnung der Korrelationskoeffizienten zwischen jeweils zwei Variablen werden alle Fälle mit gültigen Werten für dieses Variablenpaar einbezogen, unabhängig davon, ob diese Fälle Missing-Values bei anderen Variablen aufweisen. Im Ergebnis entstehen Matrizen mit paarweisen Korrelationskoeffizienten, die bei der nachfolgenden Schätzung verwendet werden. Dabei entstehen jedoch eine Reihe von Problemen, u.a. mögliche inkonsistente Schätzungen, kein einheitlicher Stichprobenumfang zur Bestimmung der Anzahl der Freiheitsgrade (es wird das Minimum der paarweisen gültigen Fälle verwendet). Es ist somit große Vorsicht bei der Verwendung dieser Möglichkeit geboten, vor allem bei der Interpretation der ausgegebenen Signifikanzniveaus für den Test.

- Replace with mean (durch Mittelwert ersetzen):
  Fehlende Werte einer Variablen werden durch den Mittelwert dieser Variablen ersetzt. Dieser Mittelwert wird jedoch unter Ausschluss der Fälle mit Missing-Werten bei dieser Variablen ermittelt. Anschließend werden alle Fälle für die Schätzung der Regressionsfunktion verwendet.

Eine weitere Selektionsmöglichkeit, die nicht mit der Wahl der Selektionsmethode für die X-Variablen zu verwechseln ist, besteht im Dialogfeld „Linear Regression“ (siehe Abb. 3.3) in der Festlegung einer Selektionsvariablen, mit der die einzubeziehenden Fälle eingeschränkt werden können. Nachdem die entsprechende Variable aus der linken Quellliste in das Feld „Selection Variable:“ gebracht wurde, muss über die Schaltfläche „Rule...“ das Dialogfeld „Linear Regression: Set Rule“ geöffnet werden.

Abbildung 3.5.: Dialogfeld „Linear Regression: Set Rule“

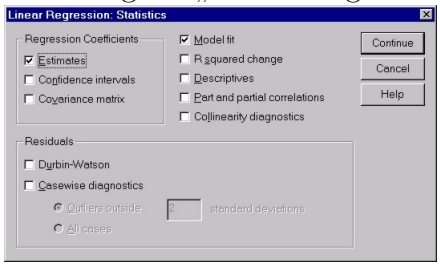


Wenn die Selektionsvariable z.B. Informationen über die Reihenfolge der Fälle enthält, kann ein Teilbereich für die Fälle festgelegt werden. Wenn die Selektionsvariable dagegen eine Faktorvariable ist, können die Fälle, die in die Regressionsschätzung einbezogen werden sollen, auf bestimmte Faktorausprägungen begrenzt werden.

Im Dialogfeld „Linear Regression“ (siehe Abb. 3.3) sind weitere Schaltflächen vorhanden, über die der Ergebnis-Output bzw. die Ausgabe von Grafiken festgelegt und die Speicherung bestimmter Ergebnisse veranlaßt werden kann.

Über die Schaltfläche „Statistics...“ gelangt man in das Dialogfeld „Linear Regression: Statistics“.

Abbildung 3.6.: Dialogfeld „Linear Regression: Statistics“



Der voreingestellte Output enthält

- die Parameterschätzungen (Estimates):
  - die Regressionsparameter b_k $(k = 0, 1, \ldots, m)$,
  - je Regressionsparameter den Standardfehler (Std. Error), d.h. die Wurzel aus dem zugehörigen Diagonalelement der Matrix $\mathbf{S_b}$ gemäß (3.17),
  - je Regressionskoeffizient den standardisierten Regressionskoeffizienten gemäß (3.18), als Beta im Output bezeichnet,
  - je Regressionsparameter den Wert der Teststatistik T gemäß (3.19) mit der zweiseitigen Überschreitungswahrscheinlichkeit (Sig.),
- die Modellgüte (Model fit):

- den multiplen Korrelationskoeffizienten (R),
- das Bestimmtheitsmaß (R Square) gemäß (3.11),
- das korrigierte Bestimmtheitsmaß (Adjusted R Square) gemäß (3.13),
- den Standardfehler der Regressionsschätzung s_u (Std. Error of Estimate), d.h. die Wurzel aus (3.16),
- eine Tabelle der Varianzanalyse (entsprechend der gezeigten Aufspaltung der Gesamtvarianz von Y) mit Sum of Squares gemäß (3.10), Anzahl der Freiheitsgrade (df - degrees of freedom), Mean Square (als Sum of Square dividiert durch df) und den Wert des F-Tests gemäß (3.12) mit der Überschreitungswahrscheinlichkeit (Sig.).

Darüber hinaus kann man sich ausgeben lassen:

- über Confidence intervals: die Konfidenzintervalle für die Regressionsparameter zum Konfidenzniveau von 95% (siehe Formel (3.22)).
- über Covariance matrix: die Varianz-Kovarianz-Matrix der geschätzten Regressionsparameter $\mathbf{S_b}$ entsprechend (3.17).
- über R squared change: die Veränderung des Bestimmtheitsmaßes (R Square Change), des F-Wertes (F Change), die zugehörigen Freiheitsgrade df1 und df2 sowie die Überschreitungswahrscheinlichkeit für die Veränderung des F-Wertes (Sig. F Change), deren Ausgabe jedoch nur Sinn macht, wenn als Selektionsmethode für die X-Variablen Stepwise, Forward oder Backward gewählt wurde.
- über Descriptives: univariate Statistiken wie Mittelwert und Standardabweichung jeder Variablen, die Matrix der Korrelationskoeffizienten zwischen jeweils zwei Variablen mit einseitigem Signifikanzniveau und Anzahl der gültigen Fälle.
- über Part and partial correlations:
  - einfache lineare Korrelationskoeffizienten zwischen der Y-Variablen und der jeweiligen X-Variablen (Zero-order),
  - die partiellen Korrelationskoeffizienten zwischen der Y-Variablen und der jeweiligen X-Variablen bei Kontrolle für die anderen in der Regressionsfunktion enthalten X-Variablen (Partial),
  - die sogenannten Part-Korrelationskoeffizienten, die sich als Wurzel aus $(R^2 - R^2_{(k)})$ ergeben, wobei $R^2_{(k)}$ das Bestimmtheitsmaß einer linearen Regressionsfunktion ohne die Variable X_k ist. Da die Differenz $R^2 - R^2_{(k)}$ die Veränderung im Bestimmtheitsmaß

durch die Herausnahme der Variablen X_k beinhaltet, sagt sie etwas über die relative Bedeutung dieser X-Variablen aus: Ein großer Wert dieser Differenz zeigt an, dass die Variable X_k einen wesentlichen Beitrag zur Erklärung der Variation in der Y-Variablen leistet.

- ▶ über Collinearity diagnostics (Multikollinearitätsdiagnose)

  - Die Toleranz einer Variablen X_k ist definiert als $1 - R^2(X_k)$, worin $R^2(X_k)$ das Bestimmtheitsmaß einer linearen Regressionsfunktion der Variablen X_k in Abhängigkeit von den anderen X-Variablen ist. Je kleiner die Toleranz ist, um so größer ist die Multikollinearität, da X_k als eine Linearkombination der anderen X-Variablen darstellbar ist.
  - Der Variance-Inflation-Faktor (VIF) ist der reziproke Wert der Toleranz:

    $$VIF_k = 1/(1 - R^2(X_k)).$$

    Je größer dieser Faktor wird, desto größer ist die Varianz der Regressionskoeffizienten b_k, woher auch der Name dieses Faktor herrührt.
  - Es werden die Eigenwerte[30] λ_k $(k = 0, 1, \ldots, m)$ der skalierten, nicht-zentrierten Kreuzprodukt-Matrix aller X-Varablen berechnet. Wenn einige der Eigenwerte deutlich größer als die anderen sind, so deutet dies auf eine Datenmatrix hin, bei der kleine Änderungen in den Variablenwerten in großen Veränderungen der Schätzungen resultieren können.
  - Mittels der Eigenwerte ist der Condition Index definiert

    $$\eta_k = \sqrt{\frac{\lambda_{max}}{\lambda_k}}. \tag{3.43}$$

    Große Werte des Condition Index zeigen enge Beziehungen zu anderen Variablen an.
  - Die Varianzanteile (Variance Proportions) ergeben sich, wenn die Varianz eines Regressionsparameters in eine Summe von Komponenten zerlegt wird, wobei jede Komponente mit einem Eigenwert verbunden ist. Variablen, die hohe Varianzanteile bei ein und demselben Eigenwert haben, sind in hohem Maße voneinander abhängig.

- ▶ im Feld Residual über Durbin-Watson: den Wert der Durbin-Watson-Teststatistik sowie unter der Überschrift „Residuals Statistics“ Minimum, Maximum, Mittelwert, Standardabweichung und Anzahl der gültigen Fälle für die Regreßwerte, die Residuen, die standardisierten Regreßwerte und die standardisierten Residuen.

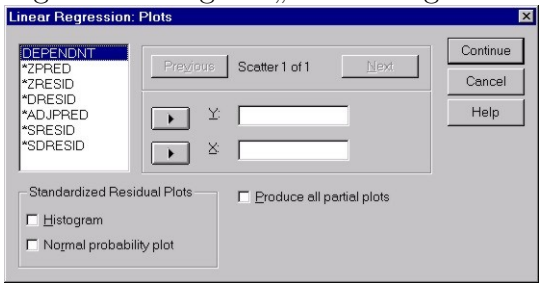
[30] Siehe Anhang H.

- im Feld Residuals über Casewise diagnostics (fallweise Diagnose): entweder eine Ausgabe
  - von Werten der standardisierten Residuen außerhalb eines (festlegbaren) Vielfachen der Standardabweichung (Outliers outside ... standard deviations) oder
  - aller Werte der standardisierten Residuen (All cases)

  zusammen mit den beobachteten Werten von Y, den Regreßwerten, den Residuen sowie der „Residual Statistics“ mit Minimum, Maximum, Standardabweichung und Anzahl der gültigen Fälle für die Regreßwerte, die Residuen, die standardisierten Regreßwerte und die standardisierten Residuen.

Zur visuellen Veranschaulichung können eine Reihe von Grafiken erstellt werden. Über die Schaltfläche „Plots...“ gelangt man in das Dialogfeld „Linear Regression: Plots“ (siehe Abb. 3.7).

Abbildung 3.7.: Dialogfeld „Linear Regression: Plots“



In der linken Quellliste stehen folgende Variablen, deren Inhalt zum Teil bereits weiter oben erläutert wurde:

- Dependent: die abhängige Variable Y,
- *ZPRED: Standardisierte Regreßwerte (standardized predicted values), definiert als

$$ZPRED = \frac{\hat{y}_i - \bar{y}}{\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{n-1}}} \tag{3.44}$$

- *ZRESID: Standardisierte Residuen (standardized residuals), gemäß (3.33),
- *DRESID: Ausgeschlossene Residuen (deleted residuals), gemäß (3.35),
- *ADJPRED: Korrigierte Regreßwerte (adjusted predicted values), definiert als $y_i - DRESID_i$,

- *SRESID: Studentisierte Residuen (studentized residuals), gemäß (3.34),
- *SDRESID: Studentisierte ausgeschlossene Residuen (studentized deleted residuals), gemäß (3.36).

Mit jeweils zwei dieser Variablen kann ein Scatterplot erzeugt werden, wobei eine Variable für die Ordinate (nach „Y:“ zu bringen) und eine Variable für die Abszisse (nach „X:“ zubringen) auszuwählen ist. Um weitere Scatterplots zu vereinbaren, ist „Next“ anzuklicken. Bis zu 9 Plots können spezifiziert werden.

Bei der Option „Produce all partial plots“ werden partielle Residuenplots erzeugt. Dabei handelt es sich um Scatterplots der Residuen, wobei die Ordinatenachse die Residuen der Regressionsfunktion der Variablen Y in Abhängigkeit von den übrigen X-Variablen (außer X_k) und die Abszissenachse die Residuen der Regressionsfunktion der Variablen X_k in Abhängigkeit von den übrigen X-Variablen aufnimmt. Die Plots werden in absteigender Ordnung des Standardfehlers der Regressionskoeffizienten der ursprünglichen Regressionsfunktion angezeigt.

Bei der Option „Standardized Residual Plots“ kann man sich für

- ein Histogramm mit eingezeichneter Normalverteilung,
- Wahrscheinlichkeitsplots (Normal probability plot) der standardisierten Residuen gegen die Normalverteilung (P-P-Plot)

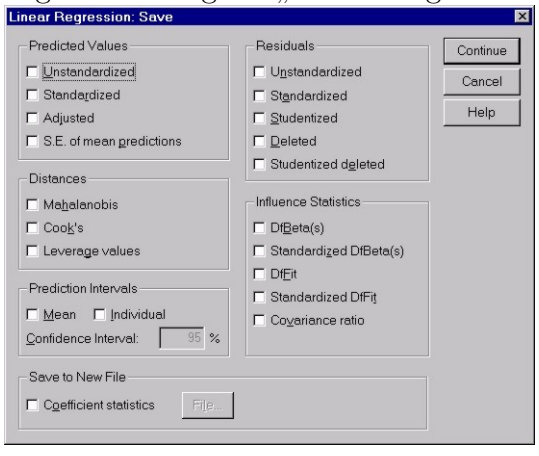
entscheiden.

Um die oben bereits angesprochene Modelldiagnostik durchführen zu können, sind bestimmte Schätzergebnisse zur weiteren Verarbeitung zu speichern. Über die Schaltfläche „Save...“ gelangt man in das Dialogfeld „Linear Regression: Save“, in dem die gewünschte Auswahl vorgenommen werden kann.
SPSS speichert die Variablen im gleichen Datenfile und vergibt automatisch neue Variablennamen, von denen man auf ihren Inhalt schließen kann. Wenn eine oder mehrere neue Variablen gespeichert werden, erscheint eine Ausgabe zusammenfassender Statistiken im Output. Die Speichermöglichkeiten sind im einzelnen:

Predicted Values (Regreßwerte):

- Unstandardized: unstandardisierte Regreßwerte;
- Standardized: standardisierte Regreßwerte;
- Adjusted: korrigierte Regreßwerte, d.h diejenigen Regreßwerte, die sich bei einer Schätzung der Regressionsfunkton unter Ausschluß des i-ten Beobachtungspunktes ergeben;

Abbildung 3.8.: Dialogfeld „Linear Regression: Save“



- S.E. of mean predictions, d.h. die Standardfehler der Regreßwerte als Wurzel aus den Diagonalelementen von (3.23).

Distances (Distanzen):

- Mahalanobis: ein verallgemeinertes Maß der Distanz nach Mahalanobis, das zur Bestimmung der Unterschiede bzw. Ähnlichkeit zwischen den n Fällen verwendet wird, die durch m verschiedene Variablen charakterisiert sind. Der Unterschied zwischen den Fällen j und k (Zeilenvektoren $\mathbf{x}_j$ und $\mathbf{x}_k$ der Matrix $\mathbf{X}$) wird durch den (empirischen) Abstandsindex $d(j,k)$ wie folgt gemessen:

$$d(jk) = (\mathbf{x}_j - \mathbf{x}_k)^T \mathbf{S}_x^{-1} (\mathbf{x}_j - \mathbf{x}_k) = (n-1)h_i. \tag{3.45}$$

  Die Matrix $\mathbf{S}_x^{-1}$ ist die Inverse der Kovarianzmatrix der m X-Variablen. Der Mahalanobis-Abstand berücksichtigt damit auch die Abhängigkeiten der m X-Variablen.

- Cook's Distanzen gemäß (3.38);
- Leverage values: gemäß (3.32).

Prediction Intervals (Konfidenzintervalle):

- Mean: untere und obere Grenzen des Konfidenzintervalls der Regreßwerte gemäß (3.25);
- Individual: untere und obere Grenzen des Konfidenzintervalls für eine einzelne Beobachtung von Y gemäß (3.28).

Die Voreinstellung für das Konfidenzniveau ist 95%. Es können aber auch andere Werte größer 0 und kleiner 100 eingegeben werden.

Residuals:

- Unstandardized: unstandardisierte Residuen gemäß (3.9);
- Standardized: standardisierte Residuen gemäß (3.33);
- Studentized: studentisierte Residuen gemäß (3.34);
- Deleted: ausgeschlossene Residuen gemäß (3.35);
- Studentized deleted: studentisierte ausgeschlossene Residuen gemäß (3.36).

Influence Statistics (Einflußgrößen):

- DfBeta(s): nach (3.39);
- Standardized DfBeta(s): standardisierte DfBeta(s) nach (3.40);
- DfFit: nach (3.41);
- Standardized DfFit: standardisierter DfFit-Wert nach (3.42);
- Covariance ratio (Kovarianzverhältnis): Verhältnis der Determinante der Kovarianz-Matrix der Residuen mit einem speziellen ausgeschlossenen Wert zur Determinante der Kovarianzmatrix der Residuen mit allen Fällen. Ein Kovarianzverhältnis nahe Eins zeigt an, dass der ausgeschlossene Fall die Kovarianzmatrix nur wenig beeinflußt.

Save to new file:
Es werden Statistiken der Regressionskoeffizienten (Coefficient statistics) unter einem festlegbaren File-Namen abgespeichert: die geschätzten Regressionsparameter (EST), ihre Standardfehler (SE), die zugehörige Überschreitungswahrscheinlichkeit (SIG), die Anzahl der Freiheitsgrade (DFE) von Residual aus der Anova-Tabelle sowie die Varianz-Kovarianz-Matrix der Regressionskoeffizienten (COV).

Schließlich sei noch die Schaltfläche „WLS>>“erwähnt (siehe Abb. 3.3). Über sie kann eine gewichtete Kleinst-Quadrate-Schätzung der Regressionsparameter erreicht werden. Ein Klick auf diese Schaltfläche öffnet in demselben Dialogfeld ein Feld „WLS Weight:“, in das diejenige Variable eingetragen wird, die die Gewichte enthält. Als Gewichtsvariable dürfen nicht die abhängige Variable und die ausgewählten X-Variablen verwendet werden. Werte der Gewichtsvariablen, die Null oder negativ oder Missing-Values sind, schließen den betreffenden Fall von der Schätzung aus.

3.3. Beispiel

Mit diesem Beispiel[31] sollen die wesentlichsten Aspekte der Regressionsanalyse demonstriert werden, wobei auch die Korrelationsanalyse einbezogen wird. Eine Vollständigkeit kann im Rahmen dieses Skriptes nicht erreicht werden.

Problemstellung

Es wird angenommen, dass der Pro-Kopf-Verbrauch von Eiscreme vom Preis, Familieneinkommen und von der Temperatur abhängt. Dazu wurden über 30 Monate die folgenden Variablen beobachtet:

Y - Pro-Kopf-Verbrauch von Eiscreme, angegeben in halbe Liter (ice cream consumption, pints per capita),

X_1 - Preis pro halbe Liter Eiscreme, angegeben in Dollar (price of ice cream per pint),

X_2 - wöchentliches Familieneinkommen, angegeben in Dollar (weekly family income),

X_3 - mittlere Temperatur in °F (mean temperature).

Für die Wiedergabe der SPSS-Outputs werden teilweise Abkürzungen icc, price, income, temp verwendet.
Die Beobachtungen dieser Variablen sind in der Datei icecream.sav enthalten.
Zur Verifizierung der obigen Annahme wird eine lineare Regressionsfunktion

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + \hat{u}_i, \quad i = 1, \ldots, n = 30$$

geschätzt. Für statistische Tests soll ein Signifikanzniveau von $\alpha = 0,05$ zugrundegelegt werden.

Erste Einschätzung der Abhängigkeit

Mittels der Korrelationsanalyse läßt sich überprüfen, inwieweit eine X-Variable einen wesentlichen Einfluß auf Y ausübt und inwieweit Multikollinearität zwischen den X-Variablen existiert, die die Schätzung der Regressionsparameter beeinträchtigt. Multikollinearität, als die korrelative lineare Abhängigkeit zwischen den erklärenden X-Variablen in einer multiplen Regressionsfunktion, bewirkt, dass diese Variablen nicht mehr unabhängig voneinander variieren. Mit zunehmender Multikollinearität wird die Identifizierung der Regressionskoeffizienten schwächer und ihre Schätzung unzuverlässiger. Möglichkeiten zur Verminderung von Multikollinearität

[31] Das Beispiel wurde entnommen aus: Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J., Ostrowski, E. (1994) S. 214

sind u.a. die Eliminierung von Variablen, Variablentransformationen, Bereinigungsverfahren und die Verwendung von externen Informationen zur Schätzung der Regressionsparameter. Über die Korrelationsanalyse kann man sich die paarweisen einfachen linearen Korrelationskoeffizienten und die partiellen Korrelationskoeffizienten zwischen Y und einer X-Variablen bei Ausschaltung des Einflusses aller restlichen X-Variablen berechnen lassen.

Unter SPSS erhält man diese Korrelationskoeffizienten über den Aufruf:

■ Analyze

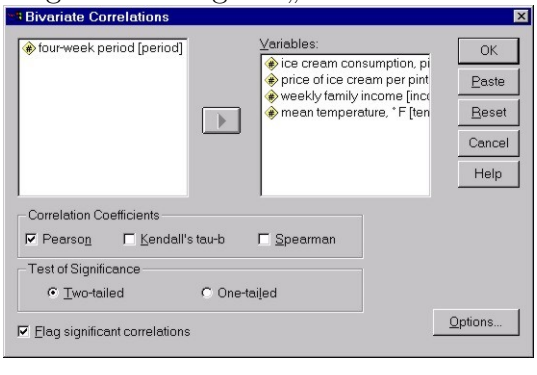
■ Correlate

Eine Matrix der einfachen linearen Korrelationskoeffizienten nach Bravais-Pearson (einschließlich der als Y gewählten Variablen) kann man sich über

■ Bivariate ...

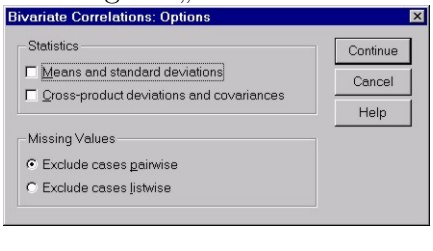
ausgeben lassen, indem in dem Dialogfeld „Bivariate Correlations“ die relevanten Variablen aus der linken Quellliste in das Feld „Variables:“ gebracht werden.

Abbildung 3.9.: Dialogfeld „Bivariate Correlations“



Die Voreinstellung mit „Pearson“ unter „Correlation Coefficients“ wird belassen. Für die Ausgabe des Signifikanzniveaus kann man sich zwischen einem zweiseitigen und einem einseitigen Test des Korrelationskoeffizienten gegen Null entscheiden. Die Anwahl von „Options...“ bietet noch die Möglichkeit der Ausgabe der Mittelwerte und Standardabweichungen für jede einbezogene Variable sowie der Kreuzproduktabweichung (Zähler der Kovarianz) und der Kovarianz für jedes Variablenpaar.

Abbildung 3.10.: Dialogfeld „Bivariate Correlations: Options“



Mittels dieser Korrelationsmatrix läßt sich zum einen die Stärke des einfachen Zusammenhanges zwischen der Variablen Y und einer der X-Variablen, zum anderen die Korrelation zwischen jeweils zwei X-Variablen (Multikollinearität) beurteilen.

Für das Beispiel erhält man folgenden Output, wobei ein zweiseitiges Signifikanzniveau gewählt und auf die Optionen verzichtet wurde.

Tabelle 3.1.: SPSS-Output der einfachen linearen Korrelationskoeffizienten

Correlation

		icc	price	income	temp
icc	Pearson Correlation	1,000	-,260	,048	,766**
	Sig. (2-tailed)	,	,166	,801	,000
	N	30	30	30	30
price	Pearson Correlation	-.260	1,000	-,107	-,108
	Sig. (2-tailed)	,166	,	,572	,569
	N	30	30	30	30
income	Pearson Correlation	,048	-,107	1,000	-,325
	Sig. (2-tailed)	,801	,572	,	,080
	N	30	30	30	30
temp	Pearson Correlation	,776**	-,108	-,325	1,000
	Sig. (2-tailed)	,000	,569	,080	,
	N	30	30	30	30

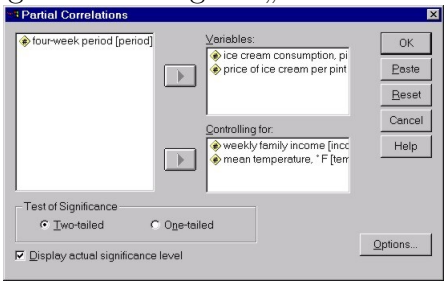
**. Correlation is significant at the 0.01 level (2-tailed).

Da bei vielen Untersuchungen jedoch die Abhängigkeit der Variablen Y von mehr als einer X-Variablen Gegenstand der Analyse ist, reicht die Prüfung der einfachen Korrelationskoeffizienten zwischen Y und X_k $(k = 1, \ldots, m)$ nicht mehr aus, da es durchaus zu Überlagerungen von Einflüssen kommen kann. Ein nächster Schritt ist deshalb die Berechnung von partiellen Korrelationskoeffizienten, die über

■ Partial...

erfolgen kann.

Abbildung 3.11.: Dialogfeld „Partial Correlations“



Die als abhängig gewählte Variable Y und eine der erklärenden X-Variablen wird in das Feld „Variables“ und alle anderen X-Variablen in das Feld „Controlling for:“ gebracht. Weitere Entscheidungen sind: Wahl eines zwei- oder einseitigen Test, unter Options die Ausgabe der Mittelwerte und Standardabweichungen für jede einbezogene Variable sowie der einfachen Korrelationskoeffizienten (Zero-order correlations) zusammen mit der Anzahl der Freiheitsgrade und der Überschreitungswahrscheinlichkeit.

Als sehr nachteilig erweist sich unter SPSS, dass nicht mit einem Durchlauf alle partiellen Korrelationskoeffizienten geprüft werden können, sondern jeder einzeln mit der angegebenen Prozedur aufgerufen werden muss.

Für das Beispiel müssen somit drei Durchläufe erfolgen, wobei wiederum ein zweiseitiger Test durchgeführt und keine der Optionen ausgegeben werden sollen:

- Pro-Kopf-Verbrauch von Eiscreme in Abhängigkeit vom Preis bei Ausschaltung der Einflüsse von Einkommen und Temperatur,
- Pro-Kopf-Verbrauch von Eiscreme in Abhängigkeit vom Einkommen bei Ausschaltung der Einflüsse von Preis und Temperatur,
- Pro-Kopf-Verbrauch von Eiscreme in Abhängigkeit von der Temperatur bei Ausschaltung der Einflüsse von Preis und Einkommen.

Im Output ist hier nicht die Anzahl der gültigen Fälle, sondern die Anzahl der Freiheitsgrade $f = n - m - 1$ enthalten (Angabe in der zweiten Zeile in Klammern). P in der dritten Zeile beinhaltet die Überschreitungswahrscheinlichkeit, die mit dem vorgegebenen Signifikanzniveau zu vergleichen ist.

Tabelle 3.2.: SPSS-Output der partiellen linearen Korrelationskoeffizienten

```
- - - P A R T I A L    C O R R E L A T I O N    C O E F F I C I E N T S - - -

Controlling for..    INCOME    TEMP

                        ICC          PRICE
 ICC                  1,0000         -,2384
                      (    0)        (   26)
                       P= ,         P= ,222
 price                -,2384         1,000
                      (   26)        (    0)
                     P= ,222         P= ,
(Coefficient / (D.F.) / 2-tailed Significance)

„ , “ is printed if a coefficient cannot be computed
```

- - - P A R T I A L C O R R E L A T I O N C O E F F I C I E N T S - - -

Controlling for.. TEMP PRICE

	ICC	INCOME
ICC	1,0000	,4845
	(0)	(26)
	P= .	P= ,009
INCOME	,4845	1,000
	(26)	(0)
	P= ,009	P= .

(Coefficient / (D.F.) / 2-tailed Significance)

„ . “ is printed if a coefficient cannot be computed

- - - P A R T I A L C O R R E L A T I O N C O E F F I C I E N T S - - -

Controlling for.. PRICE INCOME

	ICC	TEMP
ICC	1,0000	,8358
	(0)	(26)
	P= .	P= ,000
TEMP	,8358	1,000
	(26)	(0)
	P= ,000	P= .

(Coefficient / (D.F.) / 2-tailed Significance)

„ . “ is printed if a coefficient cannot be computed

Schlussfolgerungen:

- Auf der Basis der einfachen linearen Korrelationskoeffizienten weist nur die Variable Temperatur einen signifikanten Einfluss auf den Pro-Kopf-Verbrauch von Eiscreme zum 5%-Niveau auf. Wird jedoch der Einfluss der jeweils anderen X-Variablen ausgeschaltet (partielle Korrelation), so haben die Variablen Einkommen und Temperatur einen signifikanten Einfluss auf den Pro-Kopf-Verbrauch von Eiscreme.

- Signifikante Multikollinearität wird nicht angezeigt, jedoch ist die Korrelation zwischen Einkommen und Temperatur relativ hoch und sollte für die weitere Analyse beachtet werden.

Die Prüfung der einfachen bzw. partiellen Korrelationskoeffizienten zwischen Y und X_k ($k = 1, \ldots, m$) vor der Regressionschätzung ist kein zwingendes Muss, da die Regressionsprozedur unter SPSS auch Auswahlverfahren einschließt. Jedoch sollte auf jeden Fall auf Multikollinearität geprüft werden, um von vornherein unter Hinzuziehung fachwissenschaftlicher Kriterien hoch korrelierte X-Variablen auszuschließen.

Multiple lineare Regressionsschätzung mit der Methode Enter

Im nächsten Schritt soll eine multiple lineare Regressionsfunktion unter Aufnahme einer Regressionskonstanten und aller X-Variablen bei Verwendung der Methode „Enter“ (Aufnahme aller X-Variablen in einem Schritt) geschätzt werden. Dazu werden im Dialogfeld „Linear Regression“ (siehe Abb. 3.3) die Variable icc in das Feld „Dependent:“ und die Variablen price, income und temp in das Feld „Independent(s):“ gebracht:

a) Standardausgabe

Im Dialogfeld „Linear Regression: Statistics“ (siehe Abb. 3.6) wird die Voreinstellung für die Ausgabe der Schätzungen und Modellgüte belassen. Der Output ist in Tabelle 3.3 enthalten.

Tabelle 3.3.: SPSS-Output der Standardausgabe der Regressionsschätzung

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	mean temperature °F price of ice cream per pint, weekly family income[a]	,	Enter

a. All requested variables entered.

b. Dependent Variable: ice cream consumption, pints per capita

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,848[a]	,719	,687	3,6833E-02

a. Predictors: (Constant), mean temperature °F, price of ice cream per pint, weekly family income

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9,025E-02	3	3,008E-02	22,175	,000[a]
	Residual	3,527E-02	26	1,357E-03		
	Total	,126	29			

a. Predictors: (Constant), mean temperature °F, price of ice cream per pint, weekly family income

b. Dependent Variable: ice cream consumption, pints per capita

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	,197	,270		,730	,472
	price	-1,044	,834	-,132	-1,252	,222
	income	3,308E-03	,001	,314	2,824	,009
	temp	3,458E-03	,000	,863	7,762	,000

a. Dependent Variable: ice cream consumption, pints per capita

Nach der Angabe der erklärenden X-Variablen und der Methode der Aufnahme dieser X-

Variablen in die Regressionsfunktion in der ersten der vier Tabellen werden in der zweiten Tabelle Kennziffern zur Modellgüte (Model Summary) ausgegeben:

- der multiple Korrelationskoeffizient (R) als Wurzel aus (3.11),
- das Bestimmtheitsmaß (R Square) gemäß (3.11),
- das korrigierte Bestimmtheitsmaß (Adjusted R Square) gemäß (3.13),
- der Standardfehler der Regressionsschätzung s_u (Standard Error) als Wurzel aus (3.16).

Die dritte Tabelle enthält die Tabelle der Varianzanalyse (ANOVA), die die Aufspaltung der Varianz der abhängigen Variablen Y gemäß (3.10) beinhaltet. Die wohl wichtigste Information in dieser ANOVA-Tabelle ist diejenige über den Prüfwert des F-Tests gemäß (3.12) mit der Überschreitungswahrscheinlichkeit (Sig.). Da Sig $= 0,000 < \alpha = 0,05$ ist, wird die Nullhypothese $H_0 : R^2 = 0$ auf dem vorgegebenen Signifikanzniveau abgelehnt: Die in die lineare Regressionsfunktion einbezogenen X-Variablen Preis, Einkommen und Temperatur erklären *zusammen* einen wesentlichen Teil der Variation von Y. Das bedeutet jedoch nicht, dass jede X-Variable einzeln wesentlich zur Erklärung der Variation in Y beiträgt.

Der nächste Teil des Outputs (Coefficients) enthält die Schätzergebnisse:

- die (unstandardisierten) Regressionsparameter b_k (B) gemäß (3.7),
- die Standardfehler der Regressionsparameter $s(b_k)$ (Std. Error) als Wurzel aus den Diagonalelementen von (3.17),
- die standardisierten Regressionskoeffizienten b_k^* (Beta) gemäß (3.18),
- die Prüfwerte des T-Tests (t) gemäß (3.19),
- die Überschreitungswahrscheinlichkeiten (Sig.).

Letztere zeigen im Vergleich zum vorgegebenen Signifikanzniveau $\alpha = 0,05$ an, dass die Regressionskoeffizienten bei den Variablen wöchentliches Familieneinkommen und mittlere Temperatur signifikant verschieden von Null sind. Dagegen weist die Variable Preis keinen signifikanten Einfluß auf den Pro-Kopf-Verbrauch von Eiscreme auf. Dieses Ergebnis stimmt mit den Schlußfolgerungen über die partiellen Korrelationskoeffizienten überein.

b) Option Confidence intervals

Bei Anforderung der Konfidenzintervalle für die geschätzten Regressionsparameter (siehe Abb. 3.6) wird die Tabelle Coefficients um den nachstehenden Teil erweitert.

Tabelle 3.4.: Zusätzlicher Output bei Anforderung von Confidence intervals

Coefficients[a]

Model		95% Confidence Interval for B	
		Lower Bound	Upper Bound
1	(Constant)	-,358	,753
	price	-2,759	,671
	income	,001	,006
	temp	,003	,004

Die Konfidenzintervalle für die Regressionskoeffizienten werden gemäß (3.22) ermittelt. Die Konfidenzintervalle für die Regressionskoeffizienten von Einkommen und Temperatur verdeutlichen nochmals den signifikanten Einfluß dieser beiden X-Variablen auf den Pro-Kopf-Verbrauch von Eiscreme, da diese Intervalle nicht die Null überspannen.

c) Option Covariance matrix

Bei Anforderung der Kovarianz-Matrix der geschätzten Regressionsparameter (siehe Abb. 3.6) wird zusätzlich die Tabelle Coefficient Correlations ausgegeben.

Tabelle 3.5.: Zusätzlicher Output bei Anforderung von Covariance matrix

Coefficient Correlations[a]

Model			temp	price	income
1	Correlations	temp	1,000	,152	,340
		price	,152	1,000	,152
		income	,340	,152	1,000
	Covariances	temp	1,985E-07	5,657E-05	1,776E-07
		price	5,657E-05	,696	1,482E-04
		income	1,776E-07	1,482E-04	1,372E-06

a. Dependent Variable: ice cream consumption, pints per capita

Die Kovarianz-Matrix wird gemäß (3.17) berechnet. Die Wurzel aus den Diagonalelementen dieser Kovarianz-Matrix sind die Standardfehler der Regressionsparameter, die bereits in der Spalte Std. Error der Teiltabelle Coefficients (siehe Tabelle 3.3) angegeben waren.

d) Option Collinearity diagnostics

Da bei der gewählten Methode „Enter“ alle X-Variablen gleichzeitig in die Regressionsfunktion aufgenommen werden, sollte eine Kollinearitätsdiagnose durchgeführt werden. Bei Anforderung von Collinearity diagnostics (siehe Abb. 3.6) wird die Tabelle Coefficients um den ersten Teil

der Tabelle 3.6 und der Output zusätzlich um die Tabelle Collinearity Diagnostics erweitert.

Tabelle 3.6.: Zusätzlicher Output bei Anforderung von Collinearity diagnostics

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	(Constant)		
	price	,966	1,036
	income	,874	1,144
	temp	,874	1,144

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions			
				(Constant)	price	income	temp
1	1	3,917	1,000	,00	,00	,00	,01
	2	7,958E-02	7,015	,00	,00	,01	,81
	3	3,401E-03	33,938	,02	,07	,83	,11
	4	3,696E-04	102,941	,98	,93	,16	,07

a. Dependent Variable: ice cream consumption, pints per capita

Aus den Ergebnissen für die Toleranz und den Variance-Inflation-Faktor (VIF) ist nicht unbedingt auf Multikollinearität zu schließen. Jedoch signalisieren die Eigenwerte bzw. die Werte des Condition Index enge Beziehungen zwischen den X-Variablen, da ein bzw. zwei Werte deutlich größer als die anderen sind. Hohe Werte der Varianzanteile in einer Zeile deuten auf Abhängigkeiten zwischen den jeweiligen Variablen hin. In der Zeile des Eigenwertes 4 gibt es hohe Varianzanteile bei der Konstanten und dem Preis, so dass auf eine enge Beziehung zwischen der Dummy-Variablen (für die Konstante) und der Variablen Preis geschlossen werden kann. Dies könnte eine Ursache für die Nichtsignifikanz der Konstanten und des Regressionskoeffizienten von Preis sein, denn existierende Multikollinearität zwischen X-Variablen erhöht die Standardfehler der Regressionskoeffizienten, was wiederum eher die Nullhypothese begünstigt.

e) Option Durbin-Watson

Bei Anforderung von Durbin-Watson (siehe Abb. 3.6) wird die Tabelle Model Summary um eine Spalte erweitert, die im ersten Teil von Tabelle 3.7 angegeben wird. Außerdem wird im Output zusätzlich eine Tabelle Residuals Statistics ausgegeben.

Tabelle 3.7.: Zusätzlicher Output bei Anforderung von Durbin-Watson

Model Summary

Model	Durbin-Watson
1	1,021

Residual Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	,25233	,46901	,35943	5,5786E-02	30
Residuals	-6,53E-02	7,899E-02	7,216E-17	3,4876E-02	30
Std. Predicted Value	-1,920	1,964	,000	1,000	30
Std. Residual	-1,773	2,144	,000	,947	30

a. Dependent Variable: ice cream consumption, pints per capita

Der Durbin-Watson-Wert wird gemäß (3.30) bestimmt. Dieser Test-Wert deutet darauf hin, dass positive Autokorrelation in den Residuen vorhanden ist, da $d = 1,021$ deutlich kleiner als 2 (keine Autokorrelation) ist. In Tabellen mit kritischen Werten des Durbin-Watson-Tests findet man für $m = 3$ (Anzahl der X-Variablen in der Regressionsfunktion), $n = 30$ und $\alpha = 0,05$ die Werte $d_u = 1,28$ und $d_o = 1,57$, so dass signifikante Autokorrelation der Residuen zum 5%-Niveau vorliegt. Die Ursachen für das Auftreten von Autokorrelation können vielschichtig sein. Es kann sich z.B. um eine Fehlspezifikation der Regressionsfunktion im Sinne von fehlenden erklärenden Variablen oder eines nicht adäquaten Funktionstyps oder um einen Strukturbruch in dem Datenmaterial handeln.

f) Option Casewise diagnostics

Bei Anforderung von Casewise diagnostics (sieh Abb. 3.6) wird im Output zusätzlich eine Tabelle gleichen Namens ausgegeben. Es wurde Outliers outside 2 standard deviations gewählt.

Tabelle 3.8.: Zusätzlicher Output bei Anforderung von Casewise diagnostics

Casewise Diagnostics[a]

Case Number	Std. Residual	icc	Predicted Value	Residual
30	2,144	,548	,46901	7,90E-02

a. Dependent Variable: ice cream consumption, pints per capita

Außerdem wird im Output zusätzlich die Tabelle Residuals Statistics ausgegeben (siehe Tabelle 3.7).

Der Fall 30 weist einen beobachteten Wert des Pro-Kopf-Verbrauchs von Eiscreme von $y_{30} = 0,548$ (Spalte 3 der Tabelle 3.8), einen Regreßwert von $\hat{y}_{30} = 0,46901$ (Spalte 4), ein unstandardisiertes Residuum von $\hat{u}_{30} = 0,079$ (Spalte 5) und ein standardisiertes Residuum

von 2,144 (Spalte 2) auf. Letzteres fällt aus dem gewählten Bereich von Mittelwert ±2 Standardabweichungen heraus und wird somit als Ausreißer angesehen.

Multiple Regressionschätzung mit der Methode Stepwise

Da bei der Methode Enter für die Einbeziehung der X-Variablen auch diejenigen X-Variablen in die Regressionsfunktion eingehen, die letztendlich keinen signifikanten Einfluss auf die abhängige Variable Y auf dem vorgegebenen Signifikanzniveau α ausüben, kann man sich auch für eine schrittweise Methode entscheiden.

Im Dialogfeld „Linear Regression“ (siehe Abb. 3.3) werden wie vorher die Variablen ausgewählt, jetzt aber als Methode „Stepwise“ gewählt. Im Dialogfeld „Linear Regression: Options“ (siehe Abb. 3.4), das man über die Schaltfläche „Options. . . “ erreicht, wird das Kriterium für die Aufnahme und Ausschluss einer X-Variablen über das Signifikanzniveau festgelegt. Diese Festlegung ist auf Grund der Korrelationen zwischen den ausgewählten und noch nicht ausgewählten X-Variablen in den einzelnen Schritten kein einfaches Problem und erfordert möglicherweise mehrere Durchläufe mit veränderten Werten. Für das Beispiel werden die Voreinstellungen für Use probability of F Entry: 0,05, Removal: 0,10 belassen. Im Dialogfeld „Linear Regression: Statistics“ (siehe Abb. 3.6) wird neben den Voreinstellungen „Estimates“ und „Model Fit“ auch auf „R square change“ und „Part and partial correlations“ entschieden. Tabelle 3.9 enthält den dazugehörigen SPSS-Output.

Tabelle 3.9.: SPSS-Output der Regressionsschätzung mit der Methode Stepwise

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	mean temperature °F	,	Stepwise (Criteria: Probability-of-F-to-enter <=, 050, Probability-of-F-to-remove>=, 100).
2	weekly family income	,	Stepwise (Criteria: Probability-of-F-to-enter <=, 050, Probability-of-F-to-remove>=, 100).

a. Dependent Variable: ice cream consumption, pints per capita

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square change	F Change	df1	df2	Sig. F Change
1	,776[a]	,602	,587	4,2262E-02	,602	42,280	1	28	,000
2	,838[a]	,702	,680	3,7217E-02	,100	9,104	1	27	,006

a. Predictors: (Constant), mean temperature °F

b. Predictors: (Constant), mean temperature °F, weekly family income

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7,551E-02	1	7,551E-02	42,280	,000[a]
	Residual	5,001E-02	28	1,786E-03		
	Total	,126	29			
2	(Constant)	8,812E-02	2	4,406E-02	31,811	,000[b]
	Residual	3,740E-02	27	1,385E-03		
	Total	,126	29			

a. Predictors: (Constant), mean temperature °F

b. Predictors: (Constant), mean temperature °F, weekly family income

c. Dependent Variable: ice cream consumption, pints per capita

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	,207	,025		8,375	,000
	temp	3,107E-03	,000	,776	6,502	,000
2	(Constant)	-,113	,108		-1,045	,305
	temp	3,543E-03	,000	,884	7,963	,000
	income	3,530E-03	,001	,335	3,017	,006

a.Dependent Variable: ice cream consumption, pints per capita

Coefficients[a]

Model		Correlations		
		Zero-order	Partial	Part
1	(Constant)			
	temp	,776	,776	,776
2	(Constant)			
	temp	,776	,837	,837
	income	,048	,502	,317

Excluded Variables[c]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	price	-,178[a]	-1,515	,141	-,280	,988
	income	,335[a]	3,017	,006	,502	,895
2	price	-,132[b]	-1,252	,222	-,238	,966

a. Predictors: (Constant), mean temperature °F

b. Predictors: (Constant), mean temperature °F, weekly family income

c. Dependent Variable: ice cream consumption, pints per capita

Bei der Methode Stepwise wird als erste Variable Temperatur in die Regressionsfunktion aufgenommen (Model 1), da diese Variable den höchsten einfachen linearen Korrelationskoeffizienten mit Y aufweist (siehe Spalte Correlations Zero-order in der Teiltabelle Coeffcients bzw. Tabelle 3.1). Zur resultierenden Regressionsfunktion gehört ein Bestimmtheitsmaß von $R^2 = 0,602$. Zur Entscheidung über die Aufnahme der nächsten Variablen werden die partiellen Korrelationskoeffizienten zwischen der abhängigen Variablen Y und einer der noch nicht in der Regressionsfunktion enthaltenen X-Variablen bei Kontrolle des Einflusses von Temperatur herangezogen. Diese partiellen Korrelationskoeffizienten findet man in der Teiltabelle Excluded Variables in der Spalte Partial Correlation oder man kann sie sich über die oben beschriebene Prozedur Partial Correlation ausgeben lassen. Es sind r(icc, income | temp) = 0,5022 (signifikant zum 5%-Niveau) und r(icc, price | temp) = -0,28 (nicht signifikant). Als zweite Variable wird deshalb Einkommen (Model 2) aufgenommen, was zu einem Bestimmtheitsmaß von $R^2 = 0,702$ führt. Dieses Bestimmtheitsmaß ist signifikant verschieden von Null (siehe Teiltabelle ANOVA). 70,2% der Variation im Pro-Kopf-Verbrauch von Eiscreme wird durch die lineare Abhängigkeit von den Variablen mittlere Temperatur und wöchentliches Familieneinkommen erklärt. Dabei ist die bedeutungsvollere Variable die Temperatur, da sie den größeren standardisierten Regressionskoeffizienten aufweist (Spalte Beta in der Teiltabelle Coefficients).

Die Veränderung des Bestimmtheitsmaßes (R Square Change) durch die Aufnahme der Variablen Einkommen beträgt 0,1 (siehe Tabelle Model Summary). Zur Veränderung des Bestimmtheitsmaßes gehören ein F-Wert (F Change) von $F_{change} = 9,104$, die Freiheitsgrade df1 = 1 (Aufnahme einer Variablen) und df2 = 27 (n-m-1 = 30-2-1) sowie eine Überschreitungswahrscheinlichkeit des F-Wertes von 0,006. Die Aufnahme der Variablen Einkommen führt somit zu einem signifikanten Zuwachs des Bestimmtheitsmaßes.

Die Variable Preis erfüllt das Aufnahmekriterium nicht (siehe Teiltabelle Excluded Variables). Wegen $F = t^2$ ergibt sich zur Variablen Preis ein Wert $F = (-1,252)^2 = 1,5675$, zu dem eine Wahrscheinlichkeit von Sig = 0,778 gehört.
Die sich im Ergebnis der Prozedur Stepwise ergebende Regressionsfunktion lautet:

$$y_i = -0,113 + 0,003543\ temp + 0,00353\ income + \hat{u}_i.$$

Durch die Option Part and partial correlations wird in der Teiltabelle Coefficients die Spalte Correlations aufgenommen. Wie bereits erwähnt, sind die Zero-order Correlations die einfachen linearen Korrelationskoeffizienten zwischen Pro-Kopf-Verbrauch und der jeweiligen X-Variablen (vgl. auch Tabelle 3.1).

In der Spalte Partial findet man die partiellen Korrelationskoeffizienten zwischen Pro-Kopf-Verbrauch und einer X-Variablen bei Ausschaltung des Einflusses der anderen (in der Regressionsfunktion enthaltenen) X-Variablen. Da Modell 1 nur eine X-Variable (Temperatur) enthält,

ist der partielle Korrelationskoeffizient identisch mit dem einfachen Korrelationskoeffizienten. Die partiellen Korrelationskoeffizienten des Modells 2 kann man sich zum Vergleich über die SPSS-Prozedur Partial Correlation ausgeben lassen.

Die Part-Korrelationskoeffizienten ergeben sich wie folgt: Da im Modell 1 nur eine X-Variable (Temperatur) in der Regressionsfunktion enthalten ist, wird als Part-Korrelationskoeffizient der einfache Korrelationskoeffizient ausgegeben.

Der Part-Korrelationskoeffizent für die Variable Temperatur im Modell 2 ergibt sich als Wurzel aus [Bestimmtheitsmaß des Modells 2 - Bestimmtheitsmaß einer Regressionsfunktion nur mit der Variablen Einkommen (d.h. ohne Temperatur)]. Letztere Regressionsfunktion ist eine einfache Regressionsfunktion, so dass das Bestimmtheitsmaß gleich dem Quadrat des einfachen Korrelationskoeffizienten zwischen Pro-Kopf-Verbrauch von Eiscreme und Einkommen ist. Dieser Korrelationskoeffizient kann aus der Spalte Zero-order bzw. aus der Tabelle 3.1 entnommen werden: r(icc, income) = 0,048, so dass R^2(icc, income) = 0,002 (gerundet) ist. R^2(Model 2) ergibt sich aus der Teiltabelle Model Summary zu R^2(Model 2) = 0,702. Damit folgt: R^2(Model 2) - R^2(icc, income) = 0,702 - 0,002 = 0,7. Die Wurzel daraus ist (gerundet) 0,837.

Analog ergibt sich der Part-Korrelationskoeffizient für die Variable Einkommen im Modell 2. Das Bestimmtheitsmaß R^2(icc, temp) ist das des Modells 1: R(Model 1) = 0,602. Damit ist R^2(Model 2) - R^2(icc, temp) = 0,702 - 0,602 = 0,1, was dem R Square Change entspricht. Die Wurzel daraus ist (gerundet) 0,317.

Die Part-Korrelationskoeffizienten unterstreichen nochmals die Tatsache, dass die Variable Temperatur den größeren Beitrag zur Erklärung der Variation im Pro-Kopf-Verbrauch von Eiscreme leistet.

Läßt man noch den Durbin-Watson-Test durchführen, so zeigt der Wert von d = 1,003 auch für dieses Regressionsmodell positive Autokorrelation der Residuen an.
Die Casewise Diagnostics identifiziert zwei Ausreißer.

Tabelle 3.10.: Zusätzlicher Output bei Anforderung von Casewise diagnostics

Casewise Diagnostics[a]

Case Number	Std. Residual	icc	Predicted Value	Residual
1	2,111	,386	,30743	7,857E-02
30	2,469	,548	,45609	9,191E-02

a. Dependent Variable: ice cream consumption, pints per capita

Residual Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	,25073	,45609	,35943	5,5125E-02	30
Residuals	-6,54E-02	9,191E-02	5,181E-17	3,5911E-02	30
Std. Predicted Value	-1,972	1,753	,000	1,000	30
Std. Residual	-1,758	2,469	,000	,965	30

a. Dependent Variable: ice cream consumption, pints per capita

Modelldiagnose

Bezüglich der Modelldiagnose kann hier keine Vollständigkeit angestrebt werden, sondern es können nur einige Aspekte vorgestellt werden. Dafür wird von der geschätzten Regressionsfunktion nach der Methode Stepwise ausgegangen.

Für die Durchführung der Modelldiagnostik werden neue Variablen gespeichert, was über das Dialogfeld „Linear Regression: Save“ erfolgt (siehe Abb. 3.8). Im Datenfile können die nachstehenden 25 Variablen gespeichert werden:

Name	Contents
pre_1	Unstandardized Predicted Value
res_1	Unstandardized Residual
dre_1	Deleted Residual
adj_1	Adjusted Predicted Value
zpr_1	Standardized Predicted Value
zre_1	Standardized Residual
sre_1	Studentized Residual
sdr_1	Studentized Deleted Residual
sep_1	Standard Error of Predicted Value
mah_1	Mahalanobis Distance
coo_1	Cook’s Distance
lev_1	Centered Leverage Value
cov_1	Covratio
dff_1	DFFIT
sdf_1	Standardized DFFIT
dfb0_1	DFBETA Intercept
dfb1_1	DFBETA INCOME
dfb2_1	DFBETA TEMP

sdb0_1	Standardized DFBETA Intercept
sdb1_1	Standardized DFBETA INCOME
sdb2_1	Standardized DFBETA TEMP
lmci_1	95% L CI for ICC mean
umci_1	95% U CI for ICC mean
lici_1	95% L CI for ICC individual
uici_1	95% U CI for ICC individual

Darüber hinaus wird im Output die folgende Tabelle Residuals Statistics ausgegeben.

Tabelle 3.11.: Zusätzlicher Output bei Speicherung von neuer Variablen

Residuals Statistics[a]

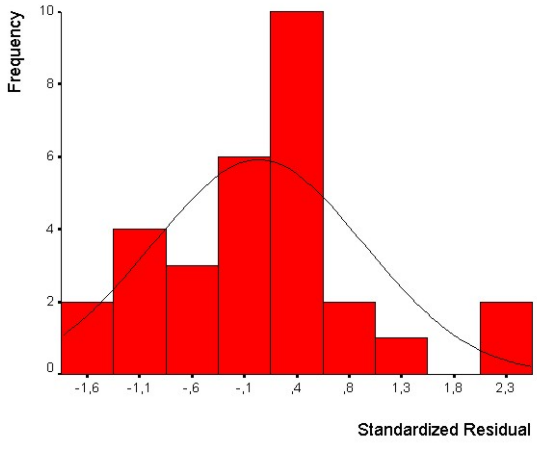
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	,25073	,45609	,35943	5,5125E-02	30
Std. Predicted Value	-1,972	1,753	,000	1,000	30
Standard Error of Predicted Value	7,1309E-03	1,6166E-02	1,1482E-02	2,6260E-03	30
Adjusted Predicted Value	,24950	,43865	,35869	5,4675E-02	30
Residual	-6,54202E-02	9,1905 E-02	5,1810E-17	3,5911E-02	30
Std. Residual	-1,758	2,469	,000	,965	30
Stud. Residual	-1,806	2,694	,010	1,018	30
Deleted Residual	-6,90599E-02	,10935	7,4822E-04	4,0010E-02	30
Stud. Deleted Residual	-1,890	3,091	,022	1,081	30
Mahal. Distance	,098	4,505	1,933	1,282	30
Cook's Distance	,000	,459	,039	,088	30
Centered Leverage Value	,003	,155	,067	,044	30

a. Dependent Variable: ice cream consumption, pints per capita

a) Prüfung der standardisierten Residuen auf Normalverteilung

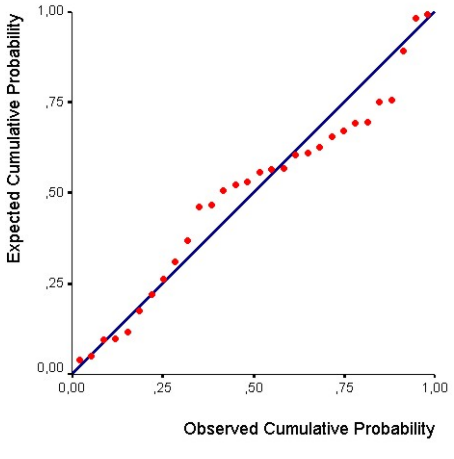
Zur Prüfung auf Normalverteilung werden die standardisierten Residuen verwendet, da sie auf die Standardabweichung normiert sind und somit große Abweichungen besser zu erkennen sind als bei den unstandardisierten Residuen. Die Prüfung kann explorativ mit einem Histogramm bei eingezeichneter Normalverteilungskurve sowie einem P-P-Plot erfolgen. Zu diesem Zweck wird über das Dialogfeld „Linear Regression: Plots“ (siehe Abb. 3.7) ein Histogramm und der Normalverteilungsplot (normal probability plot) der standardisierten Residuen angefordert.

Abbildung 3.12.: Histogramm der standardisierten Residuen



Das von SPSS ausgegebene Histogramm wurde dahingehend bearbeitet, dass die standardisierten Residuen in 9 Klassen eingeteilt werden und ihr Wertebereich auf der Abszisse von -1,8 bis 2,5 festgelegt wird.

Abbildung 3.13.: Normal P-P-Plot der standardisierten Residuen



Im P-P-Plot werden die Werte der empirischen Verteilungsfunktion der standardisierten Residuen gegen die Werte der Verteilungsfunktion der Normalverteilung abgetragen. Wenn die beiden Verteilungen übereinstimmen, liegen die Punkte auf der Winkelhalbierenden.

Im Histogramm sind vor allem am rechten Schweif und bezüglich der Kurtosis Abweichungen festzustellen. Auch im P-P-Plot ist eine gewisse Systematik um die Winkelhalbierende in dem Sinne zu beobachten, dass die standardisierten Residuen im mittleren Wertebereich oberhalb, dann unterhalb liegen.

Um festzustellen, ob es sich dabei um signifikante Abweichungen von der Normalverteilung

handelt, wird eine statistische Prüfung mit dem Kolmogorov-Smirnov-Test[32] unter Verwendung der abgespeicherten Variablen zre_1 durchgeführt. Da bei dem Testaufruf über das entsprechende Dialogfeld stets Mittelwert und Standardabweichung[33] aus den Daten verwendet wird, erfolgt der Test über die Syntax, um gegen die Normalverteilung mit Mittelwert Null und Standardabweichung Eins (d.h. die Standardnormalverteilung) zu prüfen:

```
NPAR TESTS
/K-S(NORMAL 0,1)=zre_1
/MISSING ANALYSIS.
```

Das Ergebnis enthält Tabelle 3.12. Der Test führt jedoch nicht zur Ablehnung der Nullhypothese auf dem 5%-Niveau.

Tabelle 3.12.: Ergebnis des Kolmogorov-Smirnov-Tests

One-Sample Kolmogorov-Smirnov Test

		Standardized Residuals
N		30
Normal Parameters[ab]	Mean	0
	Std. Deviation	1
Most Extreme Differences	Absolute	,145
	Positive	,145
	Negative	-,129
Kolmogorov-Smirnov Z		,793
Asymp. Sig. (2-tailed)		,556

a. Test distribution is Normal

b. User-Specified

b) Prüfung der Residuen auf Linearität

Für diese Prüfungen können Scatterplots, aber auch Line-Plots verwendet werden. Hier sollen Line-Plots genutzt werden, da sie im Gegensatz zu Scatterplots Tendenzen einfacher sichtbar werden lassen. Der Aufruf für die Line-plots erfolgt über

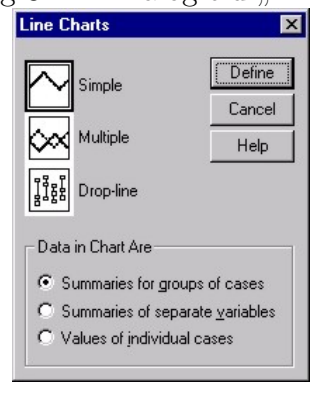
■ Graphs

  ■ Line...

[32] Der Kolmogorov-Smirnov-Test wird in jedem guten Statisitk-Lehrbuch beschrieben. Vgl. auch Rönz, B. (2001)

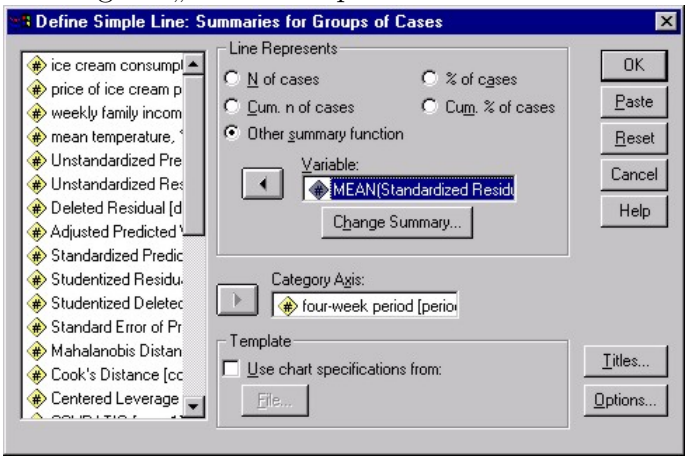
[33] Dabei ist außerdem zu beachten, dass bei der Berechnung der Standardabweichung aus den Daten, die der Kolmogorov-Smirnov-Test verwendet, durch n-1 dividiert wird, für die Standardabweichung der Residuen jedoch durch n-m-1 dividiert werden muss (m - Anzahl der erklärenden Variablen).

Abbildung 3.14.: Dialogfeld „Line Charts“



In diesem Dialogfeld wird auf „Simple“ und „Summaries for groups of cases“ entschieden. Anschließend ist die Schaltfläche „Define“ zu betätigen.

Abbildung 3.15.: Dialogfeld „Define Simple line: Summaries for Groups of Cases“



In diesem Dialogfeld wird die Variable, deren Werte auf der Abszisse abgetragen werden sollen, in das Feld „Category Axis:“ gebracht. In dem Feld „Line Represents“ wird auf „Other summary function“ entschieden und die Variable, deren Werte auf der Ordinate abgetragen werden sollen, in das Feld „Variable:“ gebracht und auf OK geklickt. Es werden Line-Plots der standardisierten Residuen gegen die Monate (four week period), die abhängige Variable Y (Pro-Kopf-Verbrauch von Eiscreme) und gegen die erklärende Variable Temperatur erzeugt.

Abbildung 3.16.: Line-Plot der standardisierten Residuen gegen die Monate

Abbildung 3.17.: Line-Plot der standardisierten Residuen gegen den Pro-Kopf-Verbrauch von Eiscreme

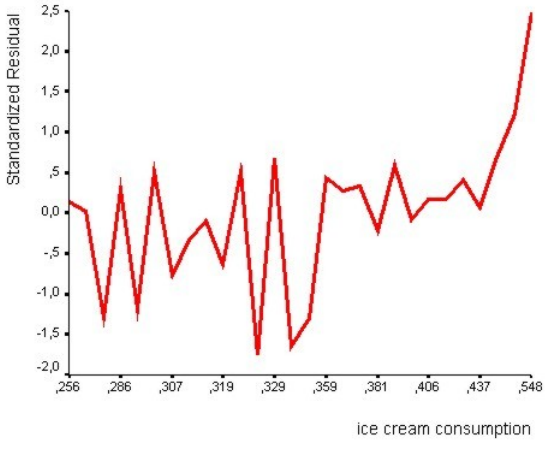
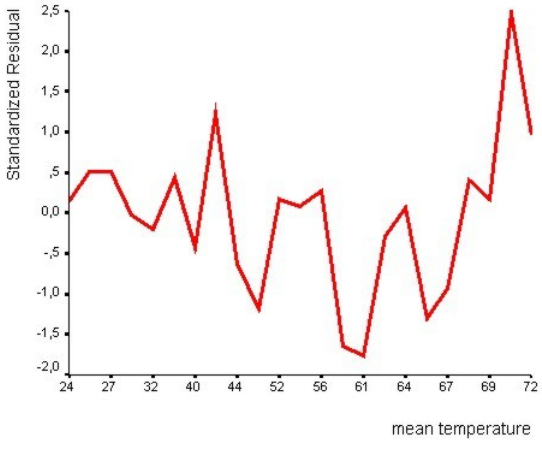
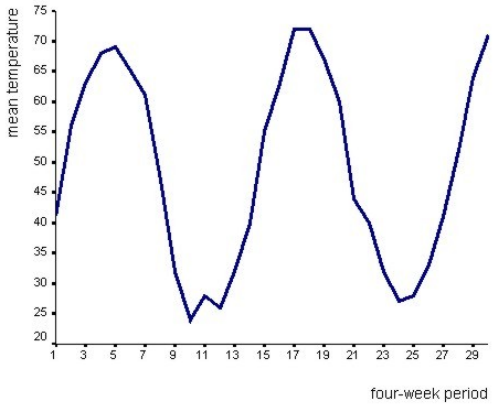


Abbildung 3.18.: Line-Plot der standardisierten Residuen gegen die Temperatur



Bei zutreffender Annahme der Linearität der standardisierten Residuen sollten keine ausgeprägten Muster auftreten. Alle drei Plots lassen jedoch Nichtlinearität in den Residuen erkennen. Um eine mögliche Ursache für diese Nichtlinearität zu finden wird ein Line-Plot der erklärenden Variablen Temperatur gegen die Monate erstellt.

Abbildung 3.19.: Line-Plot der Temperatur gegen die Monate

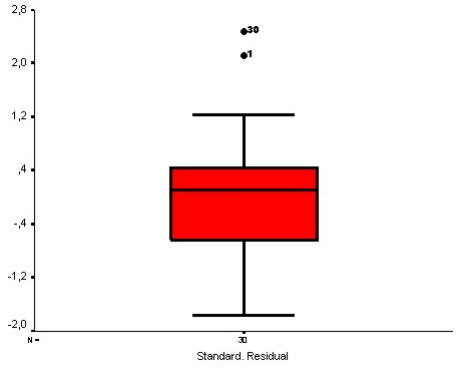


Die hier deutlich sichtbare Periodizität der Temperatur kann eine Ursache für die Nichtlinearität und die signifikante Autokorrelation der Residuen sein. Sie muss für eine erneute Regressionsschätzung berücksichtigt werden.

c) Einzeleinschätzung der Fälle

Wie bereits über die Casewise Diagnostics gezeigt (vgl. Tabelle 3.10), liegen zwei Fälle mit ihren standardisierten Residuen außerhalb $\pm 2 s_u$: Fall 1 mit einem standardisierten Residuum von 2,111 und Fall 30 mit einem standardisierten Residuum von 2,469. Dies läßt sich auch mit einem Boxplot unter Verwendung der abgespeicherten Variablen zre_1 zeigen.

Abbildung 3.20.: Boxplot der standardisierten Residuen



Während Fall 1 bei der Variablen Y keinen extremen Wert aufweist, sondern mit $y_1 = 0,386$

[pints per capita] recht nahe am Mittelwert $\bar{y} = 0,3594$ [pints per capita] liegt, weist der Fall 30 auch bei der Variablen Y den maximalen Wert von $y_{30} = 0,548$ [pints per capita] auf und ist in dieser Hinsicht als ein Ausreißer anzusehen. Eine Ursache könnte für diesen Fall die Tatsache sein, dass er bei der Variablen Temperatur einen exponierten Wert aufweist (drittgrößter Beobachtungswert mit 71°F), was sich in einem hohen Wert sowohl der Mahalanobis Distanz (viertgrößter Wert mit 3,66) als auch des Leverage (viertgrößter Wert mit 0,12623) niederschlägt.

Welche Auswirkungen die Herausnahme des Falles 1 bzw. des Falles 30 aus der Regressionsschätzung zeitigen, kann an Cook's Distanz, den SDBETAS und SDFITS beurteilt werden. Für diese Kriterien werden die Extremwerte ausgegeben, was über

■ Analyze
  ■ Descriptive Statistics
    ■ Explore...

in dem Dialogfeld „Explore: Statistics" durch Entscheidung für Outliers erreicht werden kann.

Tabelle 3.13.: Outliers für Cook's Distanz, die SDBETAS und SDFITS

Extreme Values

		Cook's Distance		Standardized DFBETA income		Standardized DFBETA temp		Standardized DFFIT	
		Case Number	Value	Case Number	Value	Case Number	Value	Case Number	Value
Highest	1	30	,45921	30	,85903	30	1,06868	30	1,34689
	2	1	,18072	8	,23715	17	,31457	1	,79946
	3	23	,08897	6	,21141	23	,16251	17	,44614
	4	17	,06461	26	,15160	18	,15970	18	,26174
	5	7	,06049	27	,09489	8	,10592	26	,20108
Lowest	1	9	,00002	1	-,60766	1	-,41764	23	-,52579
	2	15	,00008	23	-,34319	20	-,26047	7	-,44582
	3	3	,00015	20	-,15549	7	-,22475	6	-,42760
	4	14	,00017	25	-,09438	19	-,20998	20	-,42616
	5	29	,00019	18	-,09174	6	-,17801	8	-,33531

Die Fälle 1 und 30 stehen bei allen Kriterien an exponierter Stelle. Führt man eine erneute Schätzung der linearen Regressionsfunktion nach der Methode „Schrittweise" mit den gleichen Ein- und Ausschlußkriterien wie im obigen Durchlauf, jedoch bei Ausschluß der beiden Fälle 1 und 30 durch, so lässt sich eine gewisse Verbesserung der Schätzung erreichen. Auf die Wiedergabe des Outputs soll hier aus Platzgründen verzichtet werden. Deutlich wird dabei jedoch, dass nicht die Herausnahme der beiden extremen Fälle das entscheidende Kriterium für eine

Verbesserung der Schätzung ist, sondern die Berücksichtigung der Nichtlinearität und eventuell weiterer erklärender Variablen.

Mit diesen Ausführungen sollte überblicksmäßig gezeigt werden, welche Bedeutung eine eingehende Modelldiagnose hat und dass eine Regressionsanalyse nicht mit der Schätzung der Regressionsfunktion beendet ist.

3.4. Kurvenanpassung

Mit dieser weiteren unter

- Analyze
  - Regression
    - Curve Estimation...

angebotenen Option kann die Form der Beziehung zwischen zwei Variablen Y und X mittels 11 verschiedener Funktionstypen analysiert werden. Diese Funktionstypen und ihre linearisierten Gleichungen sind:

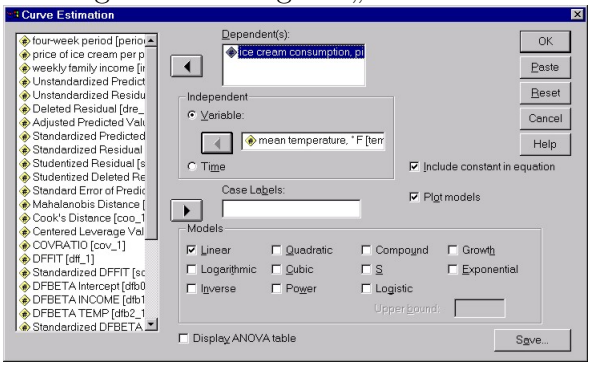
Modell	Regressionsfunktion	Linearisierte Gleichung
(1) Linear	$\hat{y} = b_0 + b_1 x$	
(2) Logarithmic	$\hat{y} = b_0 + b_1 \ln x$	
(3) Inverse	$\hat{y} = b_0 + b_1/x$	
(4) Quadratic	$\hat{y} = b_0 + b_1 x + b_2 x^2$	
(5) Cubic	$\hat{y} = b_0 + b_1 x + b_2 x^2 + b_3 x^3$	
(6) Power	$\hat{y} = b_0 \cdot x^{b_1}$	$\ln \hat{y} = \ln b_0 + b_1 \ln x$
(7) Compound	$\hat{y} = b_0 \cdot b_1^x$	$\ln \hat{y} = \ln b_0 + x \ln b_1$
(8) S	$\hat{y} = \exp(b_0 + b_1/x)$	$\ln \hat{y} = b_0 + b_1/x$
(9) Logistic	$\hat{y} = 1/(1/a + b_0 \cdot b_1^x)$	$\ln (1/\hat{y} - 1/a) = \ln b_0 + x \ln b_1$
(10) Growth	$\hat{y} = \exp(b_0 + b_1 x)$	$\ln y = b_0 + b_1 x$
(11) Exponential	$\hat{y} = b_0 \cdot \exp(b_1 x)$	$\ln \hat{y} = \ln b_0 + b_1 x.$

Darin sind b_0 die Regressionskonstante, b_1, b_2 und b_3 Regressionskoeffizienten, ln der natürliche Logarithmus, $e = \exp$ die Basis des natürlichen Logarithmus und a eine Obergrenze für die logistische Funktion. a muss eine positive Zahl größer als der größte Wert der abhängigen Variablen Y sein. Wird kein Wert spezifiziert, so verwendet SPSS unendlich und damit $1/a = 0$, wodurch dieser Term in der Funktion entfällt. Im Dialogfeld „Curve Estimation“ (Abb. 3.21) ist/sind die abhängige(n) Variable(n) in das Feld „Dependent(s)“ zu bringen. Bezüglich der unabhängigen Variablen gibt es zwei Auswahlmöglichkeiten:

- eine Variable aus der linken Quellliste oder

- die Zeit (Time), wenn für die abhängige Variable Zeitreihendaten gegeben sind und Trendfunktionen bestimmt werden sollen.

Abbildung 3.21.: Dialogfeld „Curve Estimation“



Weiterhin ist zu entscheiden, ob die Regressionskonstante b_0 in der zu schätzenden Funktion enthalten sein soll oder nicht. Bei Ausschluß wird sie in den Modellen (1) bis (5) sowie (8) und (10) gleich Null, sonst gleich 1 gesetzt.

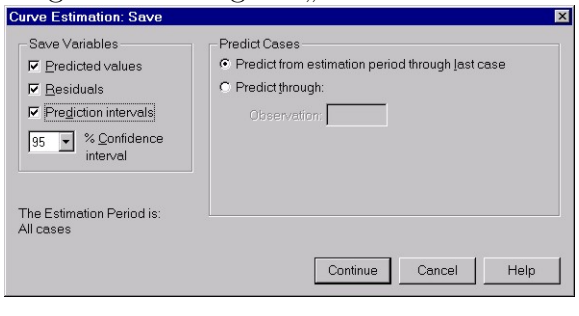
Anschließend ist das Modell bzw. sind die Modelle auszuwählen, wobei beim logistischen Modell eine Obergrenze (Upper bound) eingetragen werden kann. Einen ersten Anhaltspunkt für die Auswahl der Modelle kann ein vorheriger Scatterplot von Y und X geben, wobei jedoch für eine Trendeinschätzung eine Zeit-Variable in der Datei vorhanden sein muss.

Die Wahl von „Display ANOVA table“ erzeugt im Output für jede abhängige Variable und jedes Modell eine Tabelle der Varianzanalyse, wie bereits bei der linearen Regression beschrieben wurde.

Mit „Plot models“ wird eine Graphik je abhängiger Variable erstellt, in der die beobachteten Werte von Y sowie alle geschätzten Funktionen enthalten sind.

Über die Schaltfläche „Save...“ gelangt man in das Dialogfeld „Curve Estimation: Save“.

Abbildung 3.22.: Dialogfeld „Curve Estimation: Save“



Es können für jedes ausgewählte Modell die Regreßwerte, die Residuen sowie 95%-Konfidenzintervalle

der Regreßwerte in den Datenfile gespeichert werden, wobei die gespeicherten Werte diejenigen für die Ausgangsmodelle und nicht für die linearisierten Modelle sind.

Wurde die unabhängige Variable Zeit gewählt, können außerdem Vorhersagewerte gespeichert werden:

- Von der Schätzperiode bis zum letzten Fall vorhersagen (Predict from estimation period through last case)
  Es werden die Werte für alle Fälle der Datei entsprechend der linken unteren Angabe über die Schätzperiode (The Estimation Period is:) vorhergesagt. Diese Schätzperiode ist voreingestellt auf alle Fälle, kann jedoch vorher im SPSSS Data Editor über
  ■ Data
    ■ Select Cases...
  verändert werden.

- Vorhersagen bis (Predict through)
  Hierbei werden Vorhersagewerte der Variablen Y über den Beobachtungszeitraum hinaus ermittelt, wobei anzugeben ist, bis zu welcher Beobachtung dies geschehen soll.

Zu beachten ist bei der Kurvenanpassung:

- In allen linearisierten Modellen werden additive Störvariablen U_i verwendet, denn die Modelle werden in ihrer linearisierten Form nach der im Abschnitt 3.1 beschriebenen Methode geschätzt.

- Bezüglich der Störvariablen in allen Modellen werden die gleichen Voraussetzungen wie bei einem linearen Regressionsmodell unterstellt.

- Bei der Modelldiagnose der Residuen sind bei allen Modellen, bei denen die abhängige Variable transformiert wird, die Residuen des linearisierten Modells zu analysieren, denn diese müssen die Voraussetzungen der Schätzmethode erfüllen. Für die Modelle (6) bis (8) sowie (10) und (11) erhält man die transformierten Residuen durch $\ln(y_i) - \ln(\hat{y}_i)$, für das logistische Modell (9) mittels $\ln(1/y_i) - \ln(1/\hat{y}_i)$.

- Die gespeicherten Werte sind diejenigen für die Ausgangsmodelle und nicht für die linearisierten Modelle.

- Der Output enthält die Statistiken und geschätzten Parameterwerte des Ausgangsmodells.
  Für die Modelle (6) bis (11) ergibt sich der Standardfehler (SE) der Regressionsparameter

nach:

$$\begin{aligned} SE(b_0) &\approx \exp(\ln b_0) \cdot SE(\ln b_0) \\ SE(b_1) &\approx \exp(\ln b_1) \cdot SE(\ln b_1). \end{aligned} \tag{3.46}$$

Die 95%-Konfidenzintervalle für die Regreßwerte ergeben sich wie nachstehend:
für die Modelle (1) bis (5):

$$\begin{aligned} \hat{y}_i &\pm t_{df;0,025}\sqrt{MSE\left(1 + h_i + \frac{1}{n}\right)}, \text{wenn } b_0 \text{ enthalten} \\ \hat{y}_i &\pm t_{df;0,025}\sqrt{MSE\,(1 + h_i)} \qquad \text{sonst,} \end{aligned} \tag{3.47}$$

für die Modelle (6) bis (8) sowie (10) und (11):

$$\exp\left[\ln \hat{y}_i \pm t_{df;0,025}\sqrt{MSE\left(1 + h_i + \frac{1}{n}\right)}\right], \tag{3.48}$$

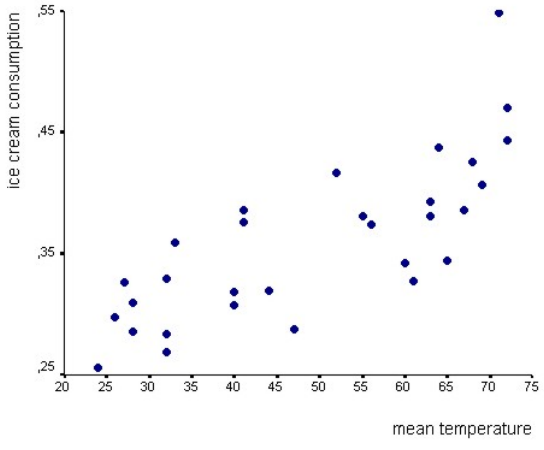
für das Modell (9):

$$\frac{1}{\exp\left[\ln \hat{y}_i \pm t_{df;0,025}\sqrt{MSE\left(1 + h_i + \frac{1}{n}\right)}\right] + \frac{1}{a}}, \tag{3.49}$$

worin $t_{df;0,025}$ das 97,5%-Quantil der t-Verteilung, *df* die sich aus MSE ergebenden Anzahl der Freiheitsgrade, MSE (mean square error) der mittlere quadratische Fehler aus der Anpassung eines linearen Modells und h_i der Leverage des i-ten Beobachtungspunktes sind.

• Beispiel aus Abschnitt 3.3 (Fortsetzung):
Für das Beispiel aus Abschnitt 3.3 soll die Abhängigkeit des Pro-Kopf-Verbrauchs an Eiscreme (Y) von der mittleren Temperatur (°F) untersucht werden. Wie bereits bei der Modelldiagnose diskutiert wurde, könnte aufgrund der Periodizität der Temperatur eine nichtlineare Abhängigkeit bestehen, auch wenn dies aus dem Scatterplot nicht so eindeutig ersichtlich ist.
Geprüft werden soll mit einem Signifikanzniveau von $\alpha = 0,10$.

Abbildung 3.23.: Scatterplot für Pro-Kopf-Verbrauch an Eiscreme und mittlere Temperatur



Zu Demonstrationszwecken werden in einem ersten Durchlauf der Kurvenanpassung alle Modelle angefordert (ohne ANOVA Tabelle und ohne Plot).

Tabelle 3.14.: Kurvenanpassung von Pro-Kopf-Verbrauch an Eiscreme und mittlere Temperatur

Independent: Temp										
DEP.	Mth	Rsq	d.f.	F	Sigf	Upper Bound	b0	b1	b2	b3
ICC	LIN	,602	28	42,28	,000		,2069	,0031		
ICC	LOG	,570	28	37,19	,000		-,1662	,1371		
ICC	INV	,528	28	31,35	,000		,4863	-5,4981		
ICC	QUA	,625	27	22,54	,000		,3212	-,0022	5,5E-05	
ICC	CUB	,683	26	18,70	,000		-,2902	,0408	-,0009	6,5E-06
ICC	COM	,624	28	46,56	,000		,2325	1,0086		
ICC	POW	,603	28	42,55	,000		,0822	,3809		
ICC	S	,569	28	36,97	,000		-,6829	-15,415		
ICC	GRO	,624	28	46,56	,000		-1,4587	,0086		
ICC	EXP	,624	28	46,56	,000		,2325	,0086		
ICC	LGS	,624	28	46,56	,000	,	4,3002	,9915		

In diesem Output bedeuten:

- Dep. - Dependent Variable,
- Mth - gewählte Methode (Funktion),
- Rsq - das Bestimmtheitsmaß,

- d.f. - die Anzahl der Freiheitsgrade, die für die Modelle verschieden sind, da eine unterschiedliche Anzahl von Parametern zu schätzen sind,
- F - der Prüfwert des F-Tests zur Prüfung der Nullhypothese $H_0 : R^2 = 0$, was identisch ist mit der Prüfung, dass alle Regressionskoeffizienten gleich Null sind,
- Sigf - die Überschreitungswahrscheinlichkeit für diesen Prüfwert,
- Upper Bound - die Obergrenze bei der logistischen Funktion, falls angegeben.

Die beste Anpassung wird durch das kubische Modell erreicht, weshalb nur für dieses Modell in einem zweiten Durchlauf die ANOVA Tabelle und der Plot ausgegeben werden soll.

Tabelle 3.15.: Kubisches Modell für Pro-Kopf-Verbrauch an Eiscreme und mittlere Temperatur

Dependent variable.. ICC Method.. Cubic

Listwise Deletion of Missing Data

Multiple R	,82665
R Square	,68335
Adjusted R Square	,64681
Standard Error	,03910

Analysis of Variance:

	DF	Sum of Squares	Mean Square
Regression	3	,08577600	,02859200
Residuals	26	,03974737	,00152875

F = 18,70292 Signif. F = ,0000

- - - - - - - - - - Variables in the Equation - - - - - - - - - -

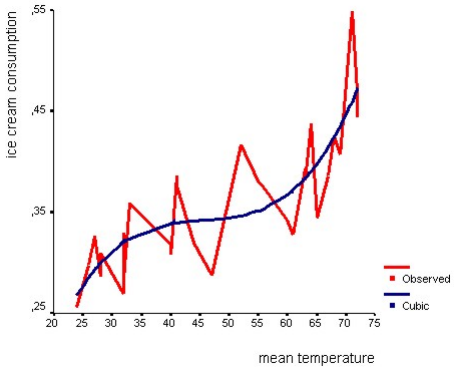
Variable	B	SE B	Beta	T	Sig T
TEMP	,040753	,020061	10,172300	2,031	,0525
TEMP**2	-,000885	,000433	-21,593149	-2,045	,0511
TEMP**3	6,46847135E-06	2,9670E-06	12,343231	2,180	,0385
(Constant)	-,290169	,292958		-,990	,3311

Nach der Nennung der abhängigen Variablen und des ausgewählten Regressionsmodells (Method) erscheinen (wie bereits bei der linearen Regression behandelt) Maßzahlen für die Güte der Anpassung der Regressionsfunktion an die empirischen Werte und die Tabelle der Varianzanalyse. Wie ersichtlich wird mit der kubischen Funktion eine relativ gute Anpassung an die empirischen Werte erreicht: $R^2 = 0,68335$. Das Bestimmtheitsmaß ist zum 10%-Niveau signifikant größer als Null, da Signif. F = 0,0000 kleiner als $\alpha = 0,10$ ist.

Der Zuwachs im Bestimmheitsmaß gegenüber dem linearen Modell ($R^2 = 0,602$) ist signifikant.[34]

Es folgt die Tabelle mit den geschätzten Regressionsparametern und den zugehörigen Gütemaßen. Außer der Konstanten sind alle Regressionskoeffizienten zum 10%-Niveau signifikant verschieden von Null.

Abbildung 3.24.: Lineplot für Pro-Kopf-Verbrauch an Eiscreme und mittlere Temperatur mit kubischem Modell



Auch wenn das kubische Regressionsmodell zu einer Verbesserung der Schätzung führte, signalisiert Abb. 3.24, dass ein anderer nichtlinearer Funktionstyp gewählt werden muss. Weiterhin ist zu berücksichtigen (wie bereits im Abschnitt 3.3 gezeigt), dass es weitere entscheidende erklärende X-Variablen gibt, die in das Regressionsmodell aufgenommen werden müssen.

• Beispiel 3.2:

In der Datei telefon.sav ist neben der Zeit-Variablen jahr die Variable telefon enthalten, die von 1900 bis 1970 die Anzahl der Telefone (in 1000) in den USA enthält.[35] Dieses Beispiel soll der Demonstration der Verwendung der Zeit (Time) dienen. Im Dialogfeld „Curve Estimation“ (siehe Abb. 3.21) wird die Variable Telefon in das Feld „Dependent(s):“ gebracht und im Feld „Independent“ Time angeklickt. Es werden bei der Kurvenanpassung zunächst alle Modelle ausgewählt, jedoch ohne ANOVA-Tabelle, Plots und Speichern.

[34] Dies läßt sich leicht prüfen, indem ein lineares Regressionsmodell des Pro-Kopf-Verbrauchs an Eiscreme in Abhängigkeit von den Variablen temp, temp**2 und temp**3 geschätzt wird, wobei jede dieser erklärenden X-Variablen in einem separaten Block gebracht und R squared change angefordert wird.

[35] Die Daten für dieses Beispiel wurden Chambers, Cleveland, Kleiner, Tukey (1983), S. 370 entnommen

Tabelle 3.16.: Kurvenanpassung von Telefon und Time

Independent: Time

DEP.	Mth	Rsq	d.f.	F	Sigf	Upper Bound	b0	b1	b2	b3
TEL	LIN	,809	69	293,14	,000		-15815	1376,59		
TEL	LOG	,490	69	66,37	,000		-46529	24282,9		
TEL	INV	,121	69	9,52	,003		39,250	,80697		
TEL	QUA	,968	68	1032,80	,000		13318,9	-1018,0	33,2581	
TEL	CUB	,997	67	7135,40	,000		-1989,1	1447,54	-51,775	,7872
TEL	COM	,943	69	1146,70	,000		3578,04	1,0506		
TEL	POW	,902	69	634,08	,000		568,113	1,0945		
TEL	S	,439	69	53,99	,000		10,3088	-5,1026		
TEL	GRO	,943	69	1146,70	,000		8,1826	,0494		
TEL	EXP	,943	69	1146,70	,000		3578,04	,0494		
TEL	LGS	,943	69	1146,70	,000	,	,0003	,9518		

Die beste Anpassung wird durch die kubische Funktion erreicht; sie unterscheidet sich jedoch nur wenig von der erreichten Anpassung durch ein quadratisches, zusammengesetztes, Wachstums-, expontielles bzw. logistisches Modell, jedoch deutlich von einem linearen Modell. Für das kubische Modell wird ein Output mit ANOVA-Tabelle und Plot erstellt. Gleichzeitig sollen die Regreßwerte und die Residuen gespeichert werden und eine Vorhersage für 10 Jahre über den Beobachtungszeitraum hinaus, d.h. bis zur 81. Beobachtung durchgeführt werden (siehe Tabelle 3.17).

Die letzten Zeilen im Output enthalten eine Information über die gespeicherten Variablen und die Anzahl der Vorhersagewerte.

Das Bestimmtheitsmaß und die Abb. 3.25 zeigen eine sehr gute Anpassung des kubischen Modells an die Beobachtungswerte. Nur im Zeitraum zwischen 1923 und 1945 zeigen sich deutlichere Abweichungen vom kubischen Modell. Die Regressionsparameter sind alle auf dem 5%-Niveau signifikant verschieden von Null.

Die gestrichelte Linie in der Abbildung kennzeichnet das Ende des Beobachtungszeitraumes. Die Vorhersagewerte sind mit besonderer Vorsicht zu behandeln, denn wie aus der Grafik ersichtlich, würde die Anzahl der Telefone in nur wenigen Jahren rasant ansteigen. Nicht jedes Regressionsmodell, das im Beobachtungszeitraum eine gute Anpassung aufweist, muss auch für die Vorhersage gut geeignet sein.

Tabelle 3.17.: Kubisches Modell für Telefon und Time

Dependent variable.. TELEFON Method.. Cubic

Listwise Deletion of Missing Data

Multiple R	,99844
R Square	,99688
Adjusted R Square	,99674
Standard Error	1803,05385

Analysis of Variance:

	DF	Sum of Squares	Mean Square
Regression	3	69591594065,0	23197198021,7
Residuals	67	217817214,1	3251003,2

F = 7135,39687 Signif. F = ,0000

- - - - - - - - - - Variables in the Equation - - - - - - - - - -

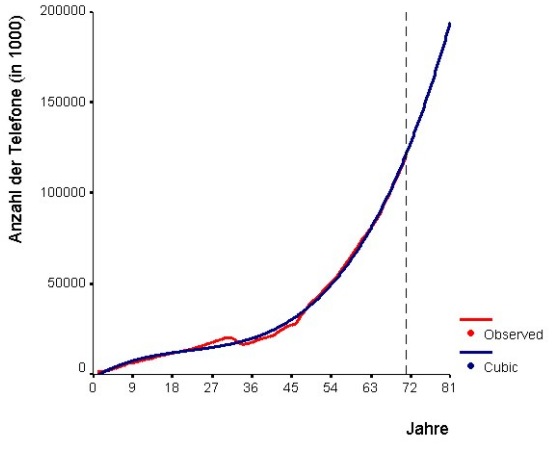
Variable	B	SE B	Beta	T	Sig T
TIME	1447,536990	107,881214	,946077	13,418	,0000
TIME**2	-51,754869	3,468575	-2,513100	-14,921	,0000
TIME**3	,787157	,031680	2,602529	24,847	,0000
(Constant)	-1989,113407	903,207080		-2,202	,0311

The following new variables are being created:

Name	Label
FIT_1	Fit for TELEFON from CURVEFIT, MOD_1 CUBIC
ERR_1	Error for TELEFON from CURVEFIT, MOD_1 CUBIC

10 new cases have been added.

Abbildung 3.25.: Line Plot für Telefon und Zeit mit kubischem Modell und Vorhersagewerten



4. Reliabilitäts- und Homogenitätsanalyse von Konstrukten

Es gibt bei vielen praktischen Untersuchungen bzw. Forschungsproblemen Merkmale, die nicht oder nur sehr schwer beobachtet werden können. Beispiele dafür sind u.a. Intelligenz, Erfahrung, Fähigkeit, Wohlfahrt, soziale Sicherheit, soziales Klima, Wohlbefinden, Arbeitszufriedenheit, Armut, Umweltbelastung, Unterentwicklung, Arbeitsleistung, Kreditwürdigkeit, Fahrtauglichkeit, Funktionsfähigkeit, Popularität, Schönheit. Mit solchen Merkmalen hat man somit vor allem in der Psychologie, der Soziologie, aber auch in den Wirtschaftswissenschaften zu tun.

Solche Merkmale werden als (theoretische) Konstrukte oder latente Variablen bezeichnet. Ein Konstrukt kann nur über die Beobachtung einer Vielzahl von Quellvariablen (auch als Items oder Indikatoren bezeichnet) erfaßt werden, die mit dem Konstrukt in enger sachlicher Beziehung stehen und verschiedene Aspekte dieses theoretischen Konstrukts beinhalten. Die Quellvariablen sind in geeigneter Weise zu einer neuen Variablen zusammenzufassen, so dass mit ihr das gemessen wird, was beabsichtigt ist. Im allgemeinen wird es jedoch nicht möglich sein, alle Quellvariablen zu erfassen, sondern man wird sich auf eine gewisse Anzahl beschränken. Um z.B. die Arbeitszufriedenheit der Mitarbeiter einer Bank festzustellen, kann eine Umfrage per Fragebogen durchgeführt werden. Um zu garantieren, dass eine große Anzahl von Befragten den Fragebogen ausfüllt, muss er überschaubar gehalten werden. Dies impliziert, dass nicht alle möglichen Fragen zur Arbeitszufriedenheit aufgenommen werden können, sondern eine Auswahl auf diejenigen Fragen erfolgt, von denen man annimmt, dass es sich um die wichtigsten handelt.

Das Ergebnis der Zusammenfassung der ausgewählten Quellvariablen (Items, Indikatoren) wird oft als Skala oder Index bezeichnet. Da beide Begriffe in der Statistik auch anderweitig verwendet werden, soll zur Vermeidung von Verwechslungen das Ergebnis der Zusammenfassung im weiteren synthetische Variable genannt werden. Eine synthetische Variable ist also das Pendant zum theoretischen Konstrukt.

Aus dem Beispiel ist bereits ersichtlich, dass in sehr vielen Fällen die Quellvariablen mittels Einschätzungen durch Personen, Befragng von Personen zu bestimmten Aspekten mit vorgege-

benen Bewertungen oder durch Tests (Tests hier nicht im statistischen Sinne) erzeugt werden müssen, jedoch kaum als tatsächlich meßbare Quellvariablen zur Verfügung stehen.

Bevor mittels solcherart erstellter synthetischer Variablen Hypothesen z.B. über Unterschiede zwischen Objekt- oder Personengruppen oder über Zusammenhänge geprüft bzw. Theorien aufgebaut oder verifiziert werden können, muss eine Einschätzung ihrer Konstruktion vorgenommen werden.[36] Die fachwissenschaftlich korrekte Auswahl der Quellvariablen zwecks Beobachtung des Konstrukts kann nicht Aufgabe der Statistik sein. Mittels der Statistik können jedoch Aussagen über ihre Reliabilität (Zuverlässigkeit) und Homogenität getroffen werden. Diese Prüfung ergibt sich aus der Notwendigkeit heraus, dass die zur Erstellung der synthetischen Variablen verwendeten Quellvariablen in der Regel nur eine Auswahl von allen möglichen Items sind bzw. sein können, jedoch auf ihrer Basis Rückschlüsse auf das theoretische Konstrukt vorgenommen werden sollen. Da jedoch immer mehrere Items einbezogen werden müssen, handelt es sich bei der Reliabilität und Homogenität um Problemkreise der multivariaten Statistik.

4.1. Reliabilitätsanalyse

Es stellt sich zunächst die Frage, ob die synthetische Variable zuverlässig in dem Sinne ist, dass sie ohne allzu große Fehler das widerspiegelt, was mit dem theoretischen Konstrukt gemessen werden soll. Jedes einbezogene Item sollte in einem gewissen Ausmaß das theoretische Konstrukt erfassen, und das unabhängig davon, ob sich die Bedingungen der Anwendung ändern, z.B. andere Objekte/Personen einbezogen werden. Jede Beobachtung eines Items kann somit in der folgenden Weise formal geschrieben werden:

$$X_{ij} = \beta_j + \varepsilon_{ij}, \tag{4.1}$$

worin X_{ij} die Beobachtung des Objektes bzw. der Person i für das Item j ($i = 1, \ldots, n; j = 1, \ldots, m$), β_j der wahre Betrag des Items j zum theoretischen Konstrukt (oft als Score bezeichnet) und ε_{ij} ein zufälliger Fehler sind. Ausgehend von dieser Gleichung kann Reliabilität in dem Sinne definiert werden, dass die Beobachtungen eines Items im wesentlichen den Beitrag zum Konstrukt beinhalten und nur einen kleinen Fehler.

Wird die synthetische Variable als Summe der Items definiert, dann kann bei Zufälligkeit der Fehlerterme unterstellt werden, dass ihr Erwartungswert über die Items gleich Null ist, während die Beiträge zum Konstrukt gleich bleiben. Je mehr zuverlässige Items, d.h. Quellvariablen, aus fachwissenschaftlicher Sicht gefunden werden können, umso größer ist der Gesamtbeitrag zum Konstrukt und umso zuverlässiger ist die synthetische Variable.

[36] Zur Konstruktion von synthetischen Variablen bzw. Skalen siehe u.a. Carmines, E.G., Zeller, R.A. (1980); Nunally, J.C. (1970), Winer, B.J. (1971), De Gruitjer, P.N.M., van der Kamp, L.J.T. (1976), Thorndyke, R.L., Hagen, E.P. (1977), Kline, P. (1979), Kline, P. (1986).

Die Feststellung der Reliabilität basiert auf der Varianzzerlegung analog zur Varianzanalyse.[37] Sie geht von der Matrix der Ausgangswerte aus, wie sie sich in der Tabelle 4.1 präsentiert, wobei im weiteren die beobachteten Personen oder Objekte als Fälle bezeichnet werden. Diese Ausgangswerte sind vor der Ziehung der Stichprobe Zufallsvariablen und somit auch die Maßzahlen, die auf ihrer Basis berechnet werden.

Tabelle 4.1.: Ausgangstabelle der Reliabilitätsanalyse

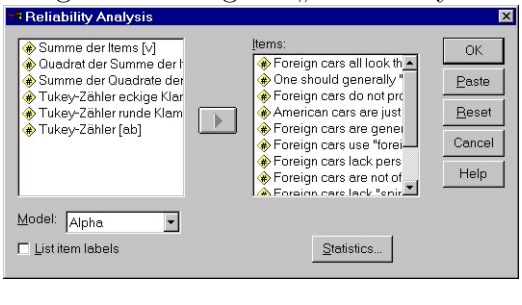
Fälle	Items				
	X_1	...	X_j	...	X_m
1	X_{11}	...	X_{1j}	...	X_{1m}
⋮	⋮	...	⋮	...	⋮
i	X_{i1}	...	X_{ij}	...	X_{im}
⋮	⋮	...	⋮	...	⋮
n	X_{n1}	...	X_{nj}	...	X_{nm}

Die Behandlung der Reliabilitätsanalyse soll im weiteren auf die durch SPSS offerierten Möglichkeiten ausgerichtet werden. Diese Darstellung wird parallel untersetzt mit einem Beispiel[38]. Der Aufruf der Reliabilitätsanalyse erfolgt über

- Analyze
  - Scale
    - Reliability Analysis...

Es öffnet sich das Dialogfeld „Reliability Analysis“.

Abbildung 4.1.: Dialogfeld „Reliability Analysis“



Die in die Analyse einzubeziehenden Items werden aus der linken Quellliste in das Feld „Items:“ gebracht. Über „List item labels“ (links unten in dem Dialogfeld) kann man sich den Inhalt der

[37] Vgl. u.a. Rönz, B. (2001).

[38] Das Beispiel wurde dem Softwarepaket „Statistica™“ entnommen. Die Beschreibung des Beispiels ist im Handbuch STATISTICA™, Volume III, S.3103, StatSoft 1994 enthalten

Items ausgeben lassen.

• Beispiel 4.1 (Teil 1):
Voreingenommenheit gegen etwas oder jemand ist ein weiteres typisches Beispiel für ein theoretisches Konstrukt, das nicht direkt, sondern nur über eine Reihe von Quellvariablen meßbar ist. Hier wird Voreingenommenheit gegenüber ausländischen Autos betrachtet. 100 Personen wird ein Fragebogen mit 10 Fragen (Items) vorgelegt, wobei die Übereinstimmung mit dem Inhalt der Fragen auf einer Rangskala von 1 (keine Zustimmung) bis 9 (Übereinstimmung) gekennzeichnet werden konnte. Die Befragung wurde in den USA durchgeführt. Die 10 Fragen stellen nur eine Auswahl aller möglichen Fragen zur Voreingenommenheit gegenüber ausländischen Autos dar, so dass im Ergebnis nicht das theoretische Konstrukt selbst, sondern nur die synthetische Variable als empirisches Pendant widergespiegelt wird. Die Ausgangswerte sind in der Datei 10items.sav enthalten. Die Beobachtungswerte in der Datei sind die Realisationen der Zufallsvariablen X_{ij} $(n = 100; m = 10; i = 1, \ldots, 100; j = 1, \ldots, 10)$ der Tabelle 4.1.

Bei der Behandlung der einzelnen Problemkreise der Reliabilitätsanalyse wird im weiteren immer nur der relevante Teil des Outputs angegeben.

Über „List item labels“ erhält man den Inhalt der 10 Fragen (Items):

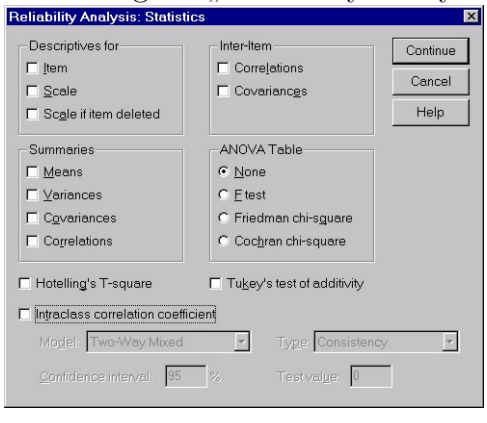
1.	ITEM1	Foreign cars all look the same
2.	ITEM2	One should generally „buy American“
3.	ITEM3	Foreign cars do not provide enough space
4.	ITEM4	American cars are just as well built
5.	ITEM5	Foreign cars are generally too expensive
6.	ITEM6	Foreign cars use „foreign technology“
7.	ITEM7	Foreign cars lack personality
8.	ITEM8	Foreign cars are not of better quality
9.	ITEM9	Foreign cars lack „spirit“
10.	ITEM10	Foreign cars are not better than American

Von diesen Items hofft man, dass sie einen großen Beitrag β_j $(j = 1, \ldots, 10)$ zu der synthetischen Variablen „Voreingenommenheit gegenüber ausländischen Autos“ leisten und die Fehler ε_{ij} in den Beobachtungen (Antworten zu den Fragen, responses) klein sind. Zur Verdeutlichung dieser Problematik sei folgende zusätzliche Frage (Item) unterstellt: Yellow foreign cars are particular ugly. Der Beitrag dieses Item zur synthetischen Variablen „Voreingenommenheit gegenüber ausländischen Autos“ wird deutlich geringer ausfallen, da bei der Einschätzung durch die befragten Personen nicht nur deren mögliche Aversion gegenüber ausländischen Autos eine Rolle spielen, sondern auch ihre persönliche Farbpräferenz eingehen wird.

4.1.1. Statistiken und Tests der Reliabilitätsanalyse

Bevor auf die verfügbaren Modelle der Reliabilitätsanalyse eingegangen wird, sollen zunächst die wählbaren Statistiken behandelt werden, die im Dialogfeld „Reliability Analysis: Statistics“ angefordert werden können.

Abbildung 4.2.: Dialogfeld „Reliability Analysis: Statistics“



Deskriptive Statistiken für (Descriptives for)

a) Item

- der Mittelwert $\bar{X}_j$ der Beobachtungswerte des j-ten Items (j=1,. . . , m)

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, \tag{4.2}$$

- die Standardabweichung des j-ten Items (j = 1,. . . ,m)

$$\begin{aligned} S_j^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij}-\bar{X}_j)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_{ij}^2 - n\bar{X}_j^2\right), \\ S_j &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_{ij}-\bar{X}_j)^2}, \end{aligned} \tag{4.3}$$

- die Anzahl der gültigen Fälle (cases) für jedes Item.

• Beispiel 4.1. (Teil 2)

		Mean	Std Dev	Cases
1.	ITEM1	4,5000	1,4460	100,0
2.	ITEM2	4,7400	1,2603	100,0
3.	ITEM3	4,7000	1,3521	100,0
4.	ITEM4	4,4800	1,3218	100,0
5.	ITEM5	4,5900	1,4777	100,0
6.	ITEM6	4,5500	1,4797	100,0
7.	ITEM7	4,6500	1,3661	100,0
8.	ITEM8	4,7800	1,3968	100,0
9.	ITEM9	4,6700	1,4217	100,0
10.	ITEM10	4,4500	1,4169	100,0

Zur besseren Einschätzung von Mittelwert und Standardabweichung sollten die Häufigkeitsverteilungen für die Items herangezogen werden, die dann jedoch über

■ Analyze
  ■ Descriptive Statistics
    ■ Frequencies

aufgerufen werden müssen. Es zeigt sich, dass die Mittelwerte und die Standardabweichung der Items sich nicht sehr stark voneinander unterscheiden.

b) Skala (Scale)

Hierunter werden deskriptive Statistiken für die synthetische Variable ausgegeben.

- der Mittelwert der synthetischen Variablen
  Nach der Voraussetzung, dass die synthetische Variable die Summe der Items ist, werden für jeden Fall die Beobachtungswerte über die Items summiert, womit sich die synthetische Variable ergibt, die mit Y symbolisiert werden soll. Für die synthetische Variable wird der Mittelwert berechnet als

$$\bar{Y} = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij} = \sum_{j=1}^{m} \bar{X}_j, \tag{4.4}$$

  was identisch ist mit der Summe der Mittelwerte der Items.

- die Varianz und die Standardabweichung der synthetischen Variablen

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\sum_{j=1}^{m} X_{ij} - \bar{Y}\right)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\sum_{j=1}^{m} X_{ij}\right)^2 - n\left(\sum_{j=1}^{m}\bar{X}_j\right)^2\right],$$

$$S_Y = \sqrt{S_Y^2}, \tag{4.5}$$

- die Anzahl der Items, die zur Bildung der synthetischen Variablen herangezogen werden.

• Beispiel 4.1. (Teil 3)

Statistics for	Mean	Variance	Std Dev	N of Variables
SCALE	46,1100	68,3009	8,2644	10

Die Summe der Beobachtungswerte über die 10 Items ist z.B. für den Fall 1 (Person 1) gleich 49, d.h., der Wert der synthetischen Variablen Y für den Fall 1 ist $y_1 = 49$. Die Summe über aller 100 y-Werte ergibt 4611, so dass $\bar{y} = 4611/100 = 46,11$ resultiert. Dieses Ergebnis erhält man auch, indem die Summe der Mittelwerte der 10 Items (siehe Means unter a) berechnet wird.
Die Summe der quadrierten y-Werte über alle 100 Fälle ergibt:

$$\sum_i y_i^2 = \sum_i (\sum_j x_{ij})^2 = 219375,$$

so dass für die Varianz von Y resultiert:

$$s_y^2 = (219375 - 100 \cdot 46,11^2)/99 = 68,3009.$$

c) Skala, wenn Item gelöscht (Scale, if item deleted)
Mit diesen Statistiken kann eine Einschätzung des Einflusses des einzelnen Items auf die synthetische Variable vorgenommen werden.

- Mittelwert der synthetischen Variablen, wenn das j-te Item nicht einbezogen wird

$$\bar{Y}(j) = \bar{Y} - \bar{X}_j, \tag{4.6}$$

- Varianz der synthetischen Variablen, wenn das j-te Item nicht einbezogen wird[39]

$$S_Y^2(j) = S_Y^2 + S_j^2 - 2Cov(Y, X_j) \tag{4.7}$$

mit

$$Cov(Y, X_j) = \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\sum_{j=1}^{m} X_{ij}\right) X_{ij} - \sum_{j=1}^{m} \bar{X}_j \sum_{i=1}^{n} X_{ij}\right], \tag{4.8}$$

[39]Es ist hier die Varianz der Differenz zweier Zufallsvariablen Y und X_j zu berechnen, wofür sich allgemein ergibt: $Var(Y) + Var(X_j) - 2Cov(Y, X_j)$.

- die (korrigierte) Korrelation zwischen dem j-ten Item und der Summe der anderen Items

$$R(j) = \frac{Cov(Y, X_j) - S_j^2}{S_j S_Y(j)}, \tag{4.9}$$

- der quadrierte multiple Korrelationskoeffizient (Bestimmtheitsmaß)
  Er ergibt sich, wenn man eine lineare Regressionsfunktion des j-ten Items als abhängige Variable mit allen anderen Items als erklärende Variablen berechnet. Diese Maßzahl wird nur ausgegeben, wenn weitere Statistiken wie z.B. die Kovarianz- und/oder die Korrelationsmatrix angefordert werden.

- der Alpha-Koeffizient, wenn das j-te Item ausgeschlossen wird
  Hierbei handelt es sich um den Reliabilitätskoeffizienten Alpha noch Cronbach, der weiter unten behandelt wird.

$$A(j) = \frac{m-1}{m-2}\left(1 - \frac{\sum\limits_{\substack{k=1 \\ k \neq j}}^{m} S_k^2}{S_Y^2(j)}\right). \tag{4.10}$$

• Beispiel 4.1. (Teil 4)

Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
ITEM1	41,6100	52,4625	,6563	,5072	,7522
ITEM2	41,3700	54,3365	,6661	,5330	,7547
ITEM3	41,4100	55,4161	,5492	,3639	,7668
ITEM4	41,6300	57,1445	,4709	,3056	,7760
ITEM5	41,5200	64,8178	,0546	,0574	,8249
ITEM6	41,5600	63,3196	,1186	,0457	,8179
ITEM7	41,4600	54,5741	,5876	,4436	,7620
ITEM8	41,3300	53,8597	,6092	,4463	,7590
ITEM9	41,4400	55,6226	,5025	,3281	,7720
ITEM10	41,6600	54,3277	,5729	,4106	,7633

Wenn z.B. Item 1 nicht in die Bildung der synthetischen Variablen einbezogen wird, dann ergibt sich als Mittelwert der synthetischen Variablen:

$$\bar{y}(1) = \bar{y} - \bar{x}_1 = 46,11 - 4,5 = 41,61.$$

Für die Varianz der synthetischen Variablen, wenn Item 1 nicht einbezogen wird, folgt:

$$s_y^2(1) = 68,3009 + 1,446^2 - 2 \cdot 8,9646 = 52,46, \quad s_y(1) = 7,2431,$$

wobei s_y^2 und s_1 aus den vorangegangenen Teilen des Outputs entnommen werden können und die Kovarianz $Cov(Y, X_1)$ über die bivariate Korrelation (siehe Kapitel 3) ermittelt werden kann.

Für die Korrelation des Items 1 mit der Summe der anderen Items folgt:

$$R(1) = [8,9646 - 1,446^2]/1,446 \cdot 7,2431 = 0,6563.$$

Schätzt man eine multiple lineare Regressionsfunktion (siehe Kapitel 3) mit Item 1 als abhängige Variable und Items 2 bis Item 10 als erklärende Variable, so erhält man im Model Summary als Bestimmtheitsmaß R Square 0,507.

Anhand dieser Bestimmtheitsmaße zeigt sich vor allem, dass zwischen Item 5 bzw. Item 6 und den restlichen Items kaum eine Beziehung besteht. Die Beobachtungswerte des Items 5 und auch des Items 6 können nur sehr unzulänglich aus den jeweils anderen Items vorhergesagt werden (Squared Multiple Correlation 0,0574 bzw. 0,0457), während dagegen z.B. 53,3% der beobachteten Variabilität in den Beobachtungswerten des Items 2 durch die anderen Items erklärt werden können.

Summaries

Hierunter werden durchschnittliche Statistiken für die Items ausgegeben.

- der durchschnittliche Items-Mittelwert (Means)

$$\bar{X} = \frac{1}{m}\sum_{j=1}^{m}\bar{X}_j = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}X_{ij}, \tag{4.11}$$

  mit der zugehörigen Varianz dieses Mittelwertes

$$S^2(\bar{X}) = \frac{1}{m-1}\sum_{j=1}^{m}(\bar{X}_j - \bar{X})^2 = \frac{1}{m-1}\left[\sum_{j=1}^{m}\bar{X}_j^2 - \frac{\left(\sum_{j=1}^{m}\bar{X}_j\right)^2}{m}\right], \tag{4.12}$$

  und außerdem Minimum ($\bar{X}_j$), Maximum ($\bar{X}_j$), das Verhältnis von Maximum ($\bar{X}_j$) zu Minimum ($\bar{X}_j$), die Spannweite (Range) = Maximum ($\bar{X}_j$) - Minimum ($\bar{X}_j$),

- die durchschnittliche Items-Varianz (Variances)
  Diese Statistiken werden analog zum durchschnittlichen Items-Mittelwert berechnet, wenn dort S_j^2 statt $\bar{X}_j$ eingesetzt wird.

- die durchschnittliche Items-Kovarianz (Covariances)

$$\overline{Cov} = \frac{\sum_{j<k} \sum Cov(X_j, X_k)}{m(m-1)} \tag{4.13}$$

  mit

$$Cov(X_j, X_k) = \frac{1}{n-1} \left[\sum_{i=1}^{n} X_{ij} X_{ik} - n \bar{X}_j \bar{X}_k \right], \quad j, k = 1, \ldots, m, \tag{4.14}$$

  und der Varianz

$$S^2(\overline{Cov}) = \frac{\sum_{j<k} \sum [Cov(X_j X_k)]^2 - \frac{[\sum_{j<k} \sum Cov(X_j X_k)]^2}{m(m-1)}}{m(m-1) - 1}. \tag{4.15}$$

- die durchschnittliche Items-Korrelation (Correlations)
  Diese Statistiken werden analog zur durchschnittlichen Items-Kovarianz berechnet, wenn dort r_{jk} statt $Cov(X_j, X_k)$ eingesetzt wird.

• Beispiel 4.1. (Teil 5)

Item Means	Mean	Minimum	Maximum	Range	Max/Min	Variance
	4,6110	4,4500	4,7800	,3300	1,0742	,0131
Item Variances	Mean	Minimum	Maximum	Range	Max/Min	Variance
	1,9474	1,5883	2,1894	,6011	1,3785	,0371
Inter-item Covariances	Mean	Minimum	Maximum	Range	Max/Min	Variance
	,5425	-,0848	1,0657	1,1505	-12,5595	,1317
Inter-item Correlations	Mean	Minimum	Maximum	Range	Max/Min	Variance
	,2863	-,0434	,5765	,6198	-13,2874	,0378

Die durchschnittliche Einschätzung über alle Items (als arithmetisches Mittel aus den Mittelwerten der Items) beträgt 4,611 bei einem Minimum von 4,45 (Item 10), einem Maximum von 4,78 (Item 8) und einer Varianz von 0,0131.

Inter-Item

Inter-Item produziert die Matrix der Korrelationskoeffizienten (Correlations) und die Matrix der Kovarianzen (Covariances), wobei in letzterer in der Hauptdiagonalen die Varianzen der Items gemäß (4.3) und abseits der Diagonalen die Kovarianzen gemäß (4.14) stehen. Die Korrelationskoeffizienten sind die einfachen linearen Bravais-Pearson-Korrelationskoeffizienten:

$$r_{jk} = Cov(X_j, X_k)/S_j S_k \tag{4.16}$$

- Beispiel 4.1. (Teil 6)

Covariance Matrix

	ITEM1	ITEM2	ITEM3	ITEM4	ITEM5	ITEM6	ITEM7	ITEM8	ITEM9	ITEM10
ITEM1	2,0909									
ITEM2	1,0505	1,5883								
ITEM3	,9596	,7899	1,8283							
ITEM4	,8182	,6008	,6505	1,7471						
ITEM5	,0859	-,0471	,2192	,1079	2,1837					
ITEM6	,2980	,2151	,0960	-,0848	,0561	2,1894				
ITEM7	1,0657	,9788	,8636	,4828	,0167	,2854	1,8662			
ITEM8	,7576	,9725	,7111	,6521	,2725	,2030	,8414	1,9511		
ITEM9	,8131	,7921	,6778	,7863	-,0761	,1429	,5601	,8156	2,0213	
ITEM10	1,0253	,8354	,5606	,6909	,0146	,1843	,8359	1,0192	,8167	2,0076

Correlation Matrix

	ITEM1	ITEM2	ITEM3	ITEM4	ITEM5	ITEM6	ITEM7	ITEM8	ITEM9	ITEM10
ITEM1	1,0000									
ITEM2	,5765	1,0000								
ITEM3	,4908	,4635	1,0000							
ITEM4	,4281	,3607	,3640	1,0000						
ITEM5	,0402	-,0253	,1097	,0552	1,0000					
ITEM6	,1393	,1154	,0480	-,0434	,0256	1,0000				
ITEM7	,5395	,5685	,4676	,2674	,0083	,1412	1,0000			
ITEM8	,3751	,5525	,3765	,3532	,1320	,0982	,4410	1,0000		
ITEM9	,3955	,4421	,3526	,4184	-,0362	,0679	,2884	,4107	1,0000	
ITEM10	,5004	,4678	,2926	,3689	,0070	,0879	,4318	,5150	,4054	1,0000

Auch anhand der Korrelationsmatrix ist zu erkennen, dass Item 5 und Item 6 nur geringe Korrelationen mit den anderen Items aufweisen. Vor allem treten auch negative Korrelationskoeffizienten bei diesen beiden Items auf. Dies widerspricht der Bildung einer synthetischen Variablen, denn dafür müssen die Korrelationen positiv sein.

ANOVA Table

Eine Varianzanalyse-Tabelle (ANOVA Tabelle) wird stets ausgegeben, wenn einer der darunter stehenden Tests angefordert wird.[40] Die Varianzzerlegung geschieht in folgender Weise:

★ Total (Gesamt)

- Summe der Abweichungsquadrate (Sum of Sq.)

$$SS_T = \sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \bar{X})^2 \tag{4.17}$$

- Anzahl der Freiheitsgrade (DF)

$$DF_T = nm - 1 \tag{4.18}$$

- Varianz (Mean Square)

$$MS_T = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \bar{X})^2}{nm - 1} \tag{4.19}$$

Die Summe der Abweichungsquadrate SS_T wird nunmehr in zwei Komponenten zerlegt:

$$\begin{aligned} SS_T &= \sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \bar{X})^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m}[(X_{ij} - \bar{X}_i) + (\bar{X}_i - \bar{X})]^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \bar{X}_i)^2 + 2\sum_{i=1}^{n}[(\bar{X}_i - \bar{X})\sum_{j=1}^{m}(X_{ij} - \bar{X}_i)] + \sum_{i=1}^{n}\sum_{j=1}^{m}(\bar{X}_i - \bar{X})^2. \end{aligned}$$

Aufgrund der Nulleigenschaft des arithmetischen Mittels ist der zweite Term auf der rechten Seite gleich Null und es folgt:

$$SS_T = \sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \bar{X})^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \bar{X}_i)^2 + \sum_{i=1}^{n} m(\bar{X}_i - \bar{X})^2. \tag{4.20}$$

[40] Die ANOVA-Tabelle und die Tests sind das Ergebnis einer zweifaktoriellen Varianzanalyse mit Meßwertwiederholung und einem Wert pro Zelle, wobei die Items und die Fälle jeweils Faktoren sind. Diese Analyse soll hier nicht weiter vertieft werden. Siehe u.a. Bosch, K. (1992), S. 521 ff.; Bortz, J. (1993), S. 266 ff.; Backhaus, K., Erichson, B., Plinke, W., Weiber, R. (1994), S. 66 ff.

★ between People (zwischen den Fällen)

- Summe der Abweichungsquadrate (Sum of Sq.)
  Die 2. Komponente auf der rechten Seite von (4.20) ist die Summe der Abweichungsquadrate zwischen den Fällen (between people), d.h. die Summe der Abweichungsquadrate der Mittelwerte $\bar{X}_i$ vom Gesamtmittelwert $\bar{X}$, und sei mit SS_{bp} symbolisiert.

$$SS_{bp} = m \sum_{i=1}^{n} (\bar{X}_i - \bar{X})^2 \tag{4.21}$$

- Anzahl der Freiheitsgrade (DF)

$$DF_{bp} = n - 1 \tag{4.22}$$

- Varianz (Mean Square)

$$MS_{bp} = \frac{m \sum_{i=1}^{n} (\bar{X}_i - \bar{X})^2}{n - 1} \tag{4.23}$$

★ within People (innerhalb der Fälle)

- Summe der Abweichungsquadrate (Sum of Sq.)
  Die 1. Komponente auf der rechten von (4.20) ist die Summe der Abweichungsquadrate innerhalb der Fälle (within people) und sei mit SS_{wp} symbolisiert.

$$SS_{wp} = \sum_{i=1}^{n} \sum_{j=1}^{m} (X_{ij} - \bar{X}_i)^2 \tag{4.24}$$

- Anzahl der Freiheitsgrade (DF)

$$DF_{wp} = n(m - 1) \tag{4.25}$$

- Varianz

$$MS_{wp} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} (X_{ij} - \bar{X}_i)^2}{n(m - 1)} \tag{4.26}$$

Die Summe der Abweichungsquadrate innerhalb der Fälle wird in der folgenden Weise weiter aufgespalten:

$$SS_{wp} = \sum_{i=1}^{n} \sum_{j=1}^{m} [X_{ij} - (\bar{X}_j - \bar{X}) + (\bar{X}_j - \bar{X}) - \bar{X}_i]^2. \tag{4.27}$$

Nach Umformen und Ausrechnen folgt:

$$\begin{aligned} SS_{wp} &= \sum_{i=1}^{n}\sum_{j=1}^{m}[\{(X_{ij}-\bar{X}_i)-(\bar{X}_j-\bar{X})\}+(\bar{X}_j-\bar{X})]^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m}[(X_{ij}-\bar{X}_i)-(\bar{X}_j-\bar{X})]^2 \\ &\quad +2\sum_{i=1}^{n}\sum_{j=1}^{m}[(X_{ij}-\bar{X}_i)-(\bar{X}_j-\bar{X})](\bar{X}_j-\bar{X})+\sum_{i=1}^{n}\sum_{j=1}^{m}(\bar{X}_j-\bar{X})^2. \end{aligned} \tag{4.28}$$

Der 2.Term auf der rechten Seite von (4.28) ist gleich Null, wie man sich leicht überzeugen kann.

□ Between Measures (zwischen den Items)

- Summe der Abweichungsquadrate (Sum of Sq.)
  Der 3. Term auf der rechten Seite von (4.28) ist inhaltlich die Summe der Abweichungsquadrate zwischen den Items (between measures), symbolisiert mit SS_{bm}.

$$SS_{bm}=\sum_{i=1}^{n}\sum_{j=1}^{m}(\bar{X}_j-\bar{X})^2 \tag{4.29}$$

- Anzahl der Freiheitsgrade (DF)

$$DF_{bm}=m-1 \tag{4.30}$$

- Varianz (Mean Square)

$$MS_{bm}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(\bar{X}_j-\bar{X})^2}{m-1} \tag{4.31}$$

□ Residual (Residuen)

- Summe der Abweichungsquadrate (Sum of Sq.)
  Der 1. Term auf der rechten Seite von (4.28) ist die restliche Summe der Abweichungsquadrate (residual), symbolisiert mit SS_r.

$$SS_r=\sum_{i=1}^{n}\sum_{j=1}^{m}[(X_{ij}-\bar{X}_i)-(\bar{X}_j-\bar{X})]^2=SS_{wp}-SS_{bm}. \tag{4.32}$$

- Anzahl der Freiheitsgrade (DF)

$$DF_r=(n-1)(m-1)=DF_{wp}-DF_{bm} \tag{4.33}$$

- Varianz (Mean Square)

$$MS_r = \frac{SS_r}{(n-1)(m-1)} \tag{4.34}$$

Die Tests prüfen die Nullhypothese auf Gleichheit der Mittelwerte der Items. Der Wert der Teststatistik und die Überschreitungswahrscheinlichkeit (Prob.) sind in den letzten beiden Spalten der ANOVA-Tabelle enthalten.
Die F-Teststatistik

$$F = \frac{MS_{bm}}{MS_r} \tag{4.35}$$

folgt unter der Nullhypothese einer F-Verteilung mit den Freiheitsgraden $f_1 = m - 1$ und $f_2 = (n-1)(m-1)$. Ist $F > F_{f1;f2;1-\alpha}$, so wird die Nullhypothese abgelehnt, wobei $F_{f1;f2;1-\alpha}$ das $(1-\alpha)$-Quantil der F-Verteilung zum vorgegebenen Signifikanzniveau α ist.

Sind die Beobachtungswerte Rangdaten, ist der Friedman-Chi-Quadrat-Test (im Output mit Chi-Square bezeichnet) anzuwenden.

Sind die Beobachtungswerte dichotom, so ist Cochrans-Chi-Quadrat-Test (im Output mit Q bezeichnet) der relevante Test.

Die Teststatistik für den Friedman-Chi-Quadrat-Test bzw. Cochrans-Chi-Quadrat-Test ist definiert als

$$\chi^2 = \frac{SS_{bm}}{MS_{wp}}, \tag{4.36}$$

die unter der Nullhypothese chi-quadrat-verteilt mit $f = m - 1$ Freiheitsgraden ist. Ist $\chi^2 > \chi^2_{f;1-\alpha}$, so wird die Nullhypothese abgelehnt, wobei $\chi^2_{f;1-\alpha}$ das $(1-\alpha)$-Quantil der χ^2-Verteilung zum vorgegeben Signifikanzniveau α ist.

Für alle Tests gilt bezüglich der Testentscheidung auch: Wenn die Überschreitungswahrscheinlichkeit (Prob.) kleiner als das vorgegebene Signifikanzniveau ist, dann wird die Nullhypothese abgelehnt.

Mit dem Friedman-Test wird gleichzeitig noch Kendall's Konkordanz-Koeffizient[41] ausgegeben, der als der Anteil der Summe der Abweichungsquadrate zwischen den Items an der Summe der Abweichungsquadrate insgesamt definiert ist:

$$W = \frac{SS_{bm}}{SS_T}. \tag{4.37}$$

- Beispiel 4.1. (Teil 7)

Da im vorliegenden Beispiel die Beobachtungen Rangdaten sind, wird der Friedman-Chi-Quadrat-Test auf einem Signifikanzniveau von $\alpha = 0,05$ durchgeführt. Die Output ist wie folgt:

[41] Zu den Konkordanzkoeffizienten siehe u.a. Bortz, J., Lienert, G.A., Boehnke, K. (1990), S. 465 ff.

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	Chi-Square	Prob.
Between People	676,1790	99	6,8301		
Within People	1263,5000	900	1,4039		
Between Measures	11,7690	9	1,3077	8,3831	,4960
Residual	1251,7310	891	1,4049		
Total	1939,6790	999	1,9416		
Grand Mean	4,6110				

Coefficient of Concordance $W = ,0061$

Um den Bezug zur oben angegebenen Varianzzerlegung herzustellen, wird diese ANOVA-Tabelle hier nochmals angegeben, wobei die konkreten Werte durch die Formelnummern ersetzt werden.

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	Chi-Square	Prob.
Between People	(4.21)	(4.22)	(4.23)		
Within People	(4.24)	(4.25)	(4.26)		
Between Measures	(4.29)	(4.30)	(4.31)	(4.36)	
Residual	(4.32)	(4.33)	(4.34)		
Total	(4.17)	(4.18)	(4.19)		
Grand Mean	(4.11)				

Coefficient of Concordance $W =$ (4.37)

Auf dem Signifikanzniveau von $\alpha = 0,05$ kann die Nullhypothese nicht verworfen werden, da Prob. $> \alpha$ ist. Die Items stimmen in ihren Mittelwerten überein, was bereits aufgrund des Outputs der einzelnen Mittelwerte zu vermuten war.

Hotellings T^2 ist ein weiterer Test auf Gleichheit der Itemsmittelwerte, der auf dem multivariaten Trennmaß[42] (auch als Hotellings Spur-Kriterium bezeichnet) T^2 basiert, auf dessen Einzeldarstellung hier verzichtet werden soll. Dieses multivariate Trennmaß drückt den diagnostischen Wert von Merkmalen (hier Items) aus und wird vor allem in der Diskriminanzanalyse angewandt. Es ist im Gegensatz zur F-Statistik auch für Analysen mit unterschiedlicher Anzahl von Items und Fällen vergleichbar. Hotellings T^2 folgt unter H_0 einer F-Verteilung mit $f_1 = m - 1$ und $f_2 = n - (m - 1)$ Freiheitsgraden.

[42] Siehe u.a. Ahrens, Läuter (1974), S. 46

• Beispiel 4.1 (Teil 8)

Hotelling's T-Squared =	9,9511	F =	1,0163	Prob. =	,4331
Degrees of Freedom:		Numerator =	9	Denominator =	91

Auch dieser Test zeigt keine signifikanten Unterschiede zwischen den Mittelwerten der Items an.

Tukey's Additivitätstest prüft die Annahme, dass keine multiplikativen Wechselwirkungen zwischen den Items existieren und somit Additivität der Items gegeben ist. Dieser Test ist von Bedeutung, weil unterstellt wird, dass sich die synthetische Variable Y als Summe der Items ergibt. Für Tukey's Additivitätstest wird die restliche Summe der Abweichungsquadrate (residual) in der ANOVA-Tabelle weiter aufgespalten:

- die auf Nichtadditivität zurückzuführende Summe der Abweichungsquadrate (non-additivity)

$$SS_{na} = \frac{\left(\sum_{i=1}^{n}(\bar{X}_i - \bar{X})\left[\sum_{j=1}^{m} X_{ij}(\bar{X}_j - \bar{X})\right]\right)^2}{\frac{1}{nm}SS_{bm}SS_{bp}} \tag{4.38}$$

  mit einer Anzahl der Freiheitsgrade $DF_{na} = 1$,

- die auf Additivität zurückzuführende Summe der Abweichungsquadrate (balance)

$$SS_{bal} = SS_r - SS_{na} \tag{4.39}$$

  mit der Anzahl der Freiheitsgrade $df_{bal} = (n-1)(m-1) - 1$.

Die mittleren Quadratsummen (Mean Square MS_{na} und MS_{bal}) ergeben sich, indem die jeweilige Summe der Abweichungsquadrate durch die zugehörige Anzahl der Freiheitsgrade dividiert wird.
Die Teststatistik lautet

$$F = \frac{MS_{na}}{MS_{bal}}, \tag{4.40}$$

die unter H_0 F-verteilt mit $f_1 = 1$ und $f_2 = (n-1)(m-1) - 1$ Freiheitsgraden ist.

Desweiteren wird ein Koeffizient (power of transformation)[43] ausgegeben, mit dem die Beobachtungen exponentiert werden müßten, um Additivität zu erreichen.

[43] Siehe u.a. Rönz, B. (2001)

• Beispiel 4.1. (Teil 9)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	Chi-Square	Prob.
Between People	676,1790	99	6,8301		
Within People	1263,5000	900	1,4039		
Between Measures	11,7690	9	1,3077	8,3831	,4960
Residual	1251,7310	891	1,4049		
Nonadditivity	1,7395	1	1,7395	1,2386	,2660
Balance	1249,9915	890	1,4045		
Total	1939,6790	999	1,9416		
Grand Mean	4,6110				

Coefficient of Concordance $W = ,0061$

Tukey estimate of power to which observations
must be raised to achieve additivity = -1,1558

Auf dem Signifikanzniveau von $\alpha = 0,05$ kann die Nullhypothese auf Additivität nicht verworfen werden, da Prob. $= 0,266 > \alpha$ ist. Es kann somit davon ausgegangen werden, dass die synthetische Variable als Summe der Items erzeugt werden kann.

Intraclass correlation coefficient (ICC)

Diese Option[44] gibt die Möglichkeit, die Konsistenz (Consistency) bzw. die absolute Übereinstimmung (Absolute Agreement) der Werte der Items zu beurteilen. Das unterliegende statistische Verfahren ist wiederum die Varianzanalyse. Da es verschiedene varianzanalytische Modelle gibt, ist eine entsprechende Wahl unter Model zu treffen.

Two-Way Random:

Für dieses varianzanalytische Modell wird unterstellt, dass die Fälle eine Zufallsstichprobe aus einer großen Grundgesamtheit sind und dass auch die Items eine Zufallsstichprobe aus einer großen Anzahl möglicher Items sind. Beide, Fälle und Items, sind Quellen der Variation in den Daten, wobei davon ausgegangen wird, dass ihre Effekte zufällig sind. Es ist das Modell der zweifaktoriellen Varianzanalyse mit zufälligen Effekten (two-way random effects model) zu wählen.

[44] Vgl. u.a. McGraw, K.O., Wong, S.P. (1996)

Two-Way Mixed:
Für dieses Modell sind die Fälle wiederum eine Zufallsstichprobe aus einer großen Grundgesamtheit und ihre Effekte auf die Daten zufällig. Die Items werden jedoch als vorgegeben betrachtet, so dass Variation in den Daten durch ihre festen Effekte hervorgerufen werden kann. Es ist das Modell der zweifaktoriellen Varianzanalyse mit gemischten Effekten (two-way mixed effects model) zu wählen.
One-Way Random:
Hierbei wird nur die Wirkung eines Faktors mit zufälligen Effekten unterstellt (one-way random effects model).

• Beispiel 4.1. (Teil 10)
Für dieses Beispiel interessiert nicht die absolute Übereinstimmung in den Einschätzungen für die Items, sondern ob die Einschätzungen ähnliche Muster aufweisen, so dass auf Konsistenz geprüft wird. Wie bereits weiter oben ausgeführt, können sowohl die Personen als auch die Fragen (Items) als eine Zufallsstichprobe angesehen werden. Es wird deshalb Two-Way Random gewählt, wobei die Voreinstellungen für ein 95%-Konfidenzintervall und den Testwert (Test value: 0) belassen werden. Der relevante Output dafür ist wie folgt:

Intraclass Correlation Coefficient
Two-Way Random Effect Model (Consistency Definition):
People and Measure Effect Random
Single Measure Intraclass Correlation = ,2786*
95,00% C.I.: Lower = ,2116 Upper = ,3609
F = 4,8618 DF = (99, 891,0) Sig. = ,0000 (Test Value = ,0000)
Average Measure Intraclass Correlation = ,7943
95,00% C.I.: Lower = ,7285 Upper = ,8495
F = 4,8618 DF = (99, 891,0) Sig. = ,0000 (Test Value = ,0000)
∗: Notice that the same estimator is used whether the interaction effect is present or not.

Abgesehen von offensichtlichen Unterschieden in den Einschätzungen zeigt der durchschnittliche Items Intraclass-Korrelationskoeffizient (Average Measure Intraclass Correlation) von 0,7943, der signifikant verschieden von Null ist (Test value: 0), an, dass die Summe der 10 Items recht zuverlässig die synthetische Variable „Voreingenommenheit gegenüber ausländischen Autos“ abbildet.

Single Measure Intraclass Correlation (Einzel Item Intraclass-Korrelationskoeffizent) gibt eine Aussage über die Reliabilität, wenn nur eine einzige Frage (Item) verwendet würde. Wie nicht anders zu erwarten, ist dieser Korrelationskoeffizient deutlich niedriger.

4.1.2. Modelle der Reliabilitätsanalyse

Die Auswahl des Modells für die Reliabilitätsanalyse wird im Dialogfeld „Reliability Analysis“ (siehe Abb. 4.1) unter Model getroffen. Von den möglichen Modellen sollen hier nur Alpha und Split-half ausführlicher diskutiert werden.

Alpha (Cronbach's Alpha)
Dieser Reliabilitätskoeffizient ist ein häufig verwendetes Maß zur Einschätzung, inwieweit das theoretische Konstrukt durch die beobachtete synthetische Variable widergespiegelt wird, d.h. ein Maß für die innere Konsistenz der synthetischen Variablen. Cronbach's Alpha basiert auf der Anzahl m der einbezogenen Items und auf dem Verhältnis der durchschnittlichen Inter-Items-Kovarianz zur durchschnittlichen Varianz der Items. Es ist wie folgt definiert:

$$A = \frac{m}{m-1}\left(1 - \frac{\sum_{j=1}^{m} S_j^2}{S_Y^2}\right) = \frac{m\overline{Cov}/\overline{Var}}{1+(m-1)\overline{Cov}/\overline{Var}}, \tag{4.41}$$

Darin sind S_j^2 gemäß (4.3), S_Y^2 gemäß (4.5) und die durchschnittliche Kovarianz zwischen den Items sowie die durchschnittliche Varianz der Items gemäß den Statistiken aus „Summaries“ zu berechnen.

Wenn die Items standardisiert sind, so dass sie gleiche Varianzen haben, reduziert sich das Verhältnis der durchschnittlichen Inter-Items-Kovarianz zur durchschnittlichen Varianz der Items auf die durchschnittliche Inter-Items-Korrelation und Cronbach's standardisiertes Alpha ergibt sich zu:

$$A_{st} = \frac{m\overline{r}}{1+(m-1)\overline{r}}, \tag{4.42}$$

worin der durchschnittliche Korrelationskoeffizient gemäß den Statistiken aus „Summaries“ ist. A_{st} wird nur ausgegeben, wenn gleichzeitig Statistiken aus „Summaries“ bzw. „Inter-Item“ angefordert werden.

Der Wertebereich von Cronbach's Alpha ist: $0 \leq A \leq 1$. Je größer sein Wert ist, desto zuverlässiger beschreiben die Items gemeinsam die synthetische Variable.

Wenn in den Items kein wahrer Beitrag zum Konstrukt, sondern nur zufällige Fehler enthalten sind (siehe Formel (4.1)), so stimmt die Summe der Varianzen der Items mit der Varianz der Summe der Items überein und A wird Null. Wird andererseits mit den Items exakt dasselbe gemessen (d.h, die Beobachtungswerte jedes Falles stimmen über die Items überein), so tritt keine Variation innerhalb der Fälle ($SS_{wp} = 0$) und damit auch keine Variation zwischen den Items ($SS_{bm} = 0$) und keine Restvariation ($SS_r = 0$), sondern nur eine Variation zwischen den

Fällen (SS_{bp}) auf. Dies impliziert perfekte paarweise Korrelation der Items ($r_{jk} = 1$ für alle j und k, j $\neq$ k) und damit einen durchschnittlichen Korrelationskoeffizienten von Eins ($\overline{r} = 1$), so dass entsprechend (4.42) $A = 1$ wird.

Anhand der Formeln (4.41) bzw. (4.42) wird deutlich, dass Cronbach's Alpha auch von der Anzahl der beobachteten Items (m) abhängt. Je größer diese Anzahl wird, desto größer wird A selbst bei gleichbleibender durchschnittlicher Korrelation. Somit sollte man Vorsicht bei der Interpretation walten lassen, da ein großer Reliabilitätskoeffizient A auch bei einer geringen durchschnittlichen Korrelation der Items auftreten kann, wenn nur die Anzahl m der Items genügend groß ist.

In diesem Sinne lohnt sich eine Analyse des Effektes jedes Items auf den Reliabilitätskoeffizienten A, was bereits weiter vorn unter „Descriptives for Scale if item deleted" mit der Formel (4.10) betrachtet wurde. A(j) ist der Wert von Cronbach's Alpha bei Ausschluß des j-ten Items. Im Vergleich mit A unter Einschluß aller Items kann der Einfluß des j-ten Items abgeschätzt werden.

•Beispiel 4.1. (Teil 11)

Reliability Coefficients 10 items

Alpha = ,7943 Standardized item alpha = ,8005

Mit dem Reliabilitätskoeffizienten soll der Anteil der Variabilität in den Einschätzungen auf die Fragen ermittelt werden, der auf die Unterschiede in den Personen zurückzuführen ist. Das heißt, es wird unterstellt, dass die Einschätzungen auf eine zuverlässige Befragung unterschiedlich sind, weil die Personen unterschiedliche Auffassungen haben und nicht weil die Befragung mehrdeutig ist oder verschiedene Interpretationen zuläßt.

Aus Teil 5 des Beispiels 4.1 kann die durchschnittliche Inter-Item-Kovarianz (0,5425), die durchschnittliche Varianz der Items (1,9474) und der durchschnittliche Korrelationskoeffizient (0,2863) entnommen werden. Einsetzen in (4.41) bzw. (4.42) ergibt:

$$A = \frac{m\overline{Cov}/\overline{Var}}{1+(m-1)\overline{Cov}/\overline{Var}} = \frac{10 \cdot \frac{0,5425}{1,9474}}{1+9 \cdot \frac{0,5425}{1,9474}} = 0,7943$$

$$A_{st} = \frac{m\overline{r}}{1+(m-1)\overline{r}} = \frac{10 \cdot 0,2863}{1+9 \cdot 0,2863} = 0,8005$$

Da die Streuungen der Items (siehe Teil 2 des Beispiels 4.1) nicht sehr voneinander abweichen, sind beide Alpha-Koeffizienten fast gleich. Würden die Streuungen der Items große Differenzen aufweisen, so wäre der Unterschied erheblich größer.

Die Zuverlässigkeit der 10 Fragen (Items) zur Messung der Voreingenommenheit gegenüber ausländischen Autos kann als recht gut angesehen werden, obwohl es offensichtlich noch Fragen

gibt, die dieses Konstrukt weiter ausfüllen könnten. Die befragten Personen schwanken in ihrem Antwortverhalten relativ wenig, antworten also tendenziell gleichermaßen.

Andererseits zeigt die Spalte „Alpha if Item Deleted“ im Teil 4 des Beispiel 4.1, dass die Frage 5 (Foreign cars are generally too expensive) bzw. 6 (Foreign cars use „foreign technology“) herausgenommen werden kann, wodurch ein Anstieg des Reliabilitätskoeffizienten eintritt. Dies erklärt sich auch aus folgendem Grund: Es wird unterstellt, dass jedes ausgewählte Item in einem bestimmten Ausmaß das theoretische Konstrukt erfaßt, so dass gleichgerichtete Variation und somit positive Korrelation zwischen den Items angenommen werden kann. Tritt negative Korrelation auf, so sind diese Items für die synthetische Variable unbrauchbar. Die Korrelationsmatrix (Teil 6 des Beispiels 4.1) zeigt bei den Items 5 und 6 negative Korrelationen zu anderen Items. Nimmt man beide Items heraus, so erhält man:

Reliability Coefficients 8 items

Alpha = ,8550 Standardized item alpha = ,8556

Split-Half Modell

Bisher wurde unterstellt, dass alle Items zu einer synthetischen Variablen zusammengefaßt werden. Beim Split-Half Modell wird die synthetische Variable in zwei Teile aufgespaltet und die Korrelation zwischen den beiden Teilen untersucht. Die ersten m_1 Items in dem Feld „Items:“ werden zum Teil 1 und die restlichen $m_2 = m - m_1$ Items zum Teil 2 zusammengefaßt. Wenn die Anzahl der Items m eine gerade Zahl ist, gilt $m_1 = m_2 = m/2$; wenn m eine ungerade Zahl ist, ist $m_1 = (m+1)/2$. Damit wird bereits ein wesentlicher Nachteil dieses Modells deutlich, indem die Reliabilitätskoeffizienten von der Reihenfolge der Items beeinflußt werden. Bei Anwendung dieses Modells kann unterstellt werden, dass zwei separate Untersuchungen mit unterschiedlichen Items an denselben Fällen durchgeführt werden.

Die oben behandelten Statistiken, die die deskriptiven Statistiken für Skala sowie die „Summaries“ betreffen (siehe Abb. 4.2), erscheinen dann für beide Teile sowie für die synthetische Variable insgesamt, alle anderen Statistiken sind wie vorher.

• Beispiel 4.1. (Teil 12)

Für den folgenden Output wurden auch die Statistiken für „Summaries“ angefordert.

R E L I A B I L I T Y A N A L Y S I S - S C A L E (S P L I T)

N of Cases = 100,0

Statistics for	Mean	Variance	Std Dev	N of Variables
Part 1	23,0100	19,9090	4,4619	5
Part 2	23,1000	21,4444	4,6308	5
Scale	46,1100	68,3009	8,2644	10

Item Means	Mean	Minimum	Maximum	Range	Max/Min	Variance
Part 1	4,6020	4,4800	4,7400	,2600	1,0580	,0135
Part 2	4,6200	4,4500	4,7800	,3300	1,0742	,0157
Scale	4,6110	4,4500	4,7800	,3300	1,0742	,0131

Item Variances	Mean	Minimum	Maximum	Range	Max/Min	Variance
Part 1	1,8877	1,5883	2,1837	,5955	1,3749	,0605
Part 2	2,0071	1,8662	2,1894	,3232	1,1732	,0141
Scale	1,9474	1,5883	2,1894	,6011	1,3785	,0371

Inter-item Covariances	Mean	Minimum	Maximum	Range	Max/Min	Variance
Part 1	,5235	-,0471	1,0505	1,0976	-22,3176	,1506
Part 2	,5704	,1429	1,0192	,8763	7,1307	,1068
Scale	,5425	-,0848	1,0657	1,1505	-12,5595	,1317

Inter-item Correlations	Mean	Minimum	Maximum	Range	Max/Min	Variance
Part 1	,2863	-,0253	,5765	,6017	-22,8076	,0454
Part 2	,2888	,0679	,5150	,4470	7,5794	,0285
Scale	,2863	-,0434	,5765	,6198	-13,2874	,0378

Reliability Coefficients 10 items

Correlation between forms =	,6521	Equal length Spearman-Brown =	,7894
Guttman Split-half =	,7891	Unequal-length Spearman-Brown =	,7894
Alpha for part 1 =	,6574	Alpha for part 2 =	,6650
5 items in part 1		5 items in part 2	

Im weiteren wird der letzte Teil des Outputs diskutiert, der die Reliabiltätskoeffizienten enthält und mit „Reliability Coefficients 10 items“ überschrieben ist. Zunächst wird der Korrelationskoeffizient zwischen den beiden Teilen (Correlation between forms) ausgegeben, der sich wie folgt berechnet:

$$r_{Y_1,Y_2} = \frac{0,5(S_Y^2 - S_{Y_1}^2 - S_{Y_2}^2)}{S_{Y_1} S_{Y_2}}, \tag{4.43}$$

worin Y_1 und Y_2 die beiden Teile der synthetischen Variablen bezeichnen, $S_Y^2, S_{Y_1}^2$ und $S_{Y_2}^2$ die Varianzen der synthetischen Variablen bzw. ihrer beiden Teile sich gemäß Formel (4.5) unter Berücksichtigung der entsprechenden Anzahl m, m_1 bzw. m_2 von Items berechnen und S_{Y_1} bzw. S_{Y_2} die Standardabweichungen sind. Er beinhaltet die Korrelation zwischen den Summen der Items in jedem Teil. Ein Korrelationskoeffizient $r_{Y_1,Y_2} = 1$ würde perfekte Zuverlässigkeit der synthetischen Variablen anzeigen und umgekehrt.

Für das Beispiel können $S_Y^2, S_{Y_1}^2, S_{Y_2}^2, S_{Y_1}$ und S_{Y_2} aus dem ersten Teil dieses Outputs entnommen werden, so dass sich ergibt:

$$r_{Y_1,Y_2} = 0,5(68,3009 - 19,909 - 21,4444)/(4,4619 \cdot 4,6308) = 0,6521.$$

Dieser Koeffizient ist eine Schätzung der Reliabilität bei Verwendung von 5 Items in jedem Teil.

Der Spearman-Brown-Koeffizient trifft eine Aussage über die Reliabilität der m Items, wenn sie aus 2 Teilen mit m_1 und m_2 Items bestehen würden, die eine Reliabilität von r_{Y_1,Y_2} aufweisen. Er wird für den Fall gleicher Anzahl von Items in jedem Teil (Equal length Spearman-Brown, SB_{EL}) als standardisiertes Alpha nach Formel (4.42) berechnet, wobei m = 2 den beiden Teilen entspricht und anstatt der durchschnittlichen Korrelation die Correlation between forms r_{Y_1,Y_2} verwendet wird:

$$SB_{EL} = \frac{2r_{Y_1,Y_2}}{1 + r_{Y_1,Y_2}}, \tag{4.44}$$

Bei einer ungleichen Anzahl von Items in den beiden Teilen (Unequal-length Spearman-Brown, SB_{UL}) ist die Berechnung wie folgt:

$$SB_{UL} = \frac{-r^2_{Y_1,Y_2} + \sqrt{r^4_{Y_1,Y_2} + 4r^2_{Y_1,Y_2}(1 - r^2_{Y_1,Y_2})m_1m_2/m^2}}{2(1 - r^2_{Y_1,Y_2})m_1m_2/m^2}. \tag{4.45}$$

Im Falle gleicher Anzahl in beiden Teilen sind beide Koeffizienten gleich. Die Anzahl der Items in jedem Teil ist am Outputende aufgeführt.

Da insgesamt 10 Fragen gestellt wurden, sind in jedem Teil 5 Fragen enthalten, so dass der Equal length Spearman-Brown Koeffizient berechnet wird:

$$SB_{EL} = 2 \cdot 0,6521/(1 + 0,6521) = 0,7894.$$

Guttman-Split-half ist ein weiterer Reliabilitätskoeffizient, der jedoch im Gegensatz zum Spearman-Brown-Koeffizient nicht gleiche Reliabilität oder gleiche Varianz in beiden Teilen voraussetzt. Er trifft eine Aussage über die Reliabilität einer synthetischen Variablen, die aus zwei Items, die jeweils einen der Teile repräsentieren, besteht. Dementsprechend ist die Berechnungsweise ein Cronbach's Alpha für zwei Items, wobei die Kovarianz zwischen den Item-Summen verwendet wird. Für das Beispiel ist:
Kovarianz zwischen den Item-Summen beider Teile = 13,474
durchschnittliche Varianz der Item-Summen = $(19,9090 + 21,4444)/2 = 20,6767$
$G_{Y_1,Y_2} = 2 \cdot (13,474/20,6767)/(1 + 13,474/20,6767) = 0,7891$
Guttman Split-half läßt sich jedoch auch nach der folgenden Formel berechnen:

$$G_{Y_1,Y_2} = \frac{2(S^2_Y - S^2_{Y_1} - S^2_{Y_2})}{S^2_Y}, \tag{4.46}$$

wofür $S^2_Y, S^2_{Y_1}$ und $S^2_{Y_2}$ aus dem ersten Teil des Outputs „RELIABILITY ANALYSIS - SCALE (SPLIT)“ entnommen werden können.
Für das Beispiel läßt sich dies leicht nachrechnen:

$$G_{Y_1,Y_2} = 2(68,3009 - 19,909 - 21,4444)/68,3009 = 0,7891.$$

Schließlich enthält der Output noch Cronbach's Alpha für jeden Teil.
Um den oben erwähnten Einfluß der Reihenfolge der Items zu demonstrieren, wird das Split-Half Modell mit einer anderen, zufällig gewählten Reihenfolge gerechnet: 10, 3, 6, 1, 4, 9, 5, 2, 8, 7. Man erhält folgenden Output:

R E L I A B I L I T Y A N A L Y S I S - S C A L E (S P L I T)

Reliability Coefficients

N of Cases = 100,0		N of Items = 10	
Correlation between forms =	,7023	Equal length Spearman-Brown =	,8251
Guttman Split-half =	,8251	Unequal-length Spearman-Brown =	,8251
Alpha for part 1 =	,6415	Alpha for part 2 =	,6452
5 items in part 1		5 items in part 2	

Zum Vergleich sei hier noch das Ergebnis des Split-Half Modells bei Ausschluß der Items 5 und 6 unter Verwendung der geordneten Reihenfolge angegeben:

R E L I A B I L I T Y A N A L Y S I S - S C A L E (S P L I T)

N of Cases = 100,0

Reliability Coefficients 8 Items

Correlation between forms =	,7319	Equal length Spearman-Brown =	,8452
Guttman Split-half =	,8451	Unequal-length Spearman-Brown =	,8452
Alpha for part 1 =	,7641	Alpha for part 2 =	,7397
4 items in part 1		4 items in part 2	

Auf die anderen unter SPSS verfügbaren Modelle soll hier nicht weiter eingegangen werden. Es sei nur soviel erwähnt, dass das Parallel Modell darüberhinaus gleiche Mittelwerte aller Items voraussetzt. Beide Methoden liefern Maximum-Likelihood-Schätzungen des Reliabilitätskoeffizienten sowie einen Chi-Quadrat-Test zur Prüfung der Anpassung des Modells an die Daten.

4.2. Homogenitätsanalyse

Nachdem die Reliabilität der synthetischen Variablen überprüft wurde, muss noch festgestellt werden, ob diese Variable homogen ist. Homogenität bedeutet hier, dass sie tatsächlich nur ein theoretisches Konstrukt (eine Dimension) widerspiegelt und nicht in breitem Maße Effekte anderer Konstrukte beinhaltet, die für die Fragestellung nicht zutreffen. Ein multivariates statistisches Verfahren, das zur Homogenitätsprüfung herangezogen werden kann, ist die Faktoranalyse.[45] Um die Homogenitätsprüfung zu verstehen, muss deshalb ein knapper Exkurs der Faktoranalyse vorausgeschickt werden.

4.2.1. Exkurs zur Faktoranalyse

Das Ziel der Faktoranalyse besteht darin, latente Strukturen aufzudecken, die sich hinter einer bestimmten Mengen von (manifesten, beobachtbaren) Variablen verbergen. Ausgehend von Beobachtungen dieser Variablen an n Fällen (Objekten, Personen usw.) sucht man mit der Faktoranalyse nach Ursachenkomplexen, die hinter diesen Variablen stehen, selbst aber nicht meßbar sind und als Faktoren bezeichnet werden. Im Kontext des vorangegangenen Abschnittes sind die Ausgangsvariablen die Quellvariablen (Items) und die Faktoren sind die theoretischen Konstrukte bzw. die synthetischen Variablen als empirische Pendants. Ein weiteres Ziel der Faktoranalyse ist die mit der Extrahierung von Faktoren einhergehende Reduktion der Datendimension, indem für weitere statistische Analysen (z.B. eine Regressionsanalyse) nicht die Ausgangsvariablen, sondern die wesentlich kleinere Zahl von Faktoren verwendet wird. Die Datengrundlage der Faktorenanalyse sind wie im Abschnitt 4.1 die Beobachtungswerte x_{ij} von m meßbaren (Zufalls-)Variablen X_j $(j = 1, \ldots, m)$ an n Fällen $(i = 1, \ldots, n)$, die in einer $(n \times m)$ Datenmatrix $\mathbf{X} = [x_{ij}]$ zusammengefäßt werden (vgl. Tabelle 4.1). Die Beobachtungswerte sollten Resultat einer einfachen Zufallsstichprobe sein, so dass sie unabhängig voneinander sind.

Die Variablen sollten metrisch skaliert und zumindestens approximativ normalverteilt sein, da auch bei diesem multivariaten Verfahren die Berechnungen auf dem Bravais-Pearson-Korrelationskoeffizienten basieren. Für nicht normalverteilte bzw. auf einem niedrigeren Skalenniveau gemessene Variablen gibt es jedoch Möglichkeiten der Vorbehandlung, auf die hier nicht weiter eingegangen werden kann. Weiterhin wird für die Faktoranalyse gefordert, dass der Stichprobenumfang n möglichst groß ist, um zufallsbedingte Extraktionen von Faktoren zu vermeiden. Der Stichprobenumfang muss auf jeden Fall größer als die Anzahl m der Variablen sein.

Um die unterschiedlichen Maßeinheiten sowie das unterschiedliche Niveau (Mittelwerte) und die verschiedene Streuung der Variablen zu eliminieren, werden die Variablen einer Standardi-

[45] Vgl. u.a. Hartung, J., Elpelt, B. (1989), Kapitel VIII; Backhaus, K. et al. (1994), S. 188 ff.; Überla, K. (1968); Bortz, J. (1993), S.472 ff.

sierung (z-Transformation)

$$z_{ij} = \frac{x_{ij} - \bar{x}}{s_j} \tag{4.47}$$

unterzogen, so dass die Variablen Z_j Mittelwert Null und Varianz Eins aufweisen. Die Matrix $\mathbf{X}$ geht in die $(n \times m)$ Matrix $\mathbf{Z} = [z_{ij}]$ über.

Der Grundgedanke zur Spezifikation des Modells der Faktoranalyse besteht darin, dass jede Z-Variable durch die Faktoren erklärt werden soll. Wenn angenommen wird, dass es Q solcher Faktoren F_q $(q = 1, \ldots, Q)$ gibt, wobei $Q \leq m$ ist, so kann das Modell der Faktoranalyse in einer ersten Version wie folgt geschrieben werden:

$$Z_j = a_{j1}F_1 + \ldots + a_{jq}F_q + \ldots + a_{jQ}F_Q = \sum_{q=1}^{Q} a_{jq}F_q; \quad j = 1, \ldots, m. \tag{4.48}$$

Notiert man (4.48) für jeden Wert z_{ij}, so folgt:

$$z_{ij} = a_{j1}f_{i1} + \ldots + a_{jq}f_{iq} + \ldots + a_{jQ}f_{iQ} = \sum_{q=1}^{Q} a_{jq}f_{iq} \tag{4.49}$$

für $j = 1, \ldots, m$ und $i = 1, \ldots, n$. Zur Vereinfachung läßt sich (4.49) in Matrizenschreibweise angeben. Mit $\mathbf{Z} = [z_{ij}], \mathbf{F} = [f_{iq}]$ und $\mathbf{A} = [a_{jq}]$ folgt

$$\mathbf{Z}_{(n\times m)} = \mathbf{F}_{(n\times Q)} \cdot \mathbf{A}^T_{(Q\times m)}. \tag{4.50}$$

Darin sind:

- **F** - die Faktorwertematrix
  Sie enthält die Werte f_{iq} jedes Faktors F_q für jeden Fall i bei jeder Variablen Z_j. Diese Faktorwerte sowie die Anzahl Q der Faktoren sind unbekannt und müssen geschätzt werden. Bezüglich der Faktoren wird gefordert, dass sie unkorreliert (orthogonal) sind und jeder Faktor mindestens auf zwei Z-Variablen wirkt, d.h., je Faktor mindestens zwei Koeffizienten a_{jq} verschieden von Null sind. In diesem Sinne spricht man dann von gemeinsamen Faktoren.

- **A** - die Faktorladungsmatrix
  Die Koeffizienten a_{jq} in (4.48) bis (4.50) repräsentieren Gewichte, mit denen die Faktoren in die Variablen Z_j eingehen. Sie werden Faktorladungen und **A** Faktorladungsmatrix genannt. Die j-te Zelle $\mathbf{a}_j = [a_{j1}, a_{j2}, \ldots, a_{jQ}]$ von **A** beinhaltet die Ladungen sämtlicher Q Faktoren auf die Variable Z_j und die q-te Spalte von **A** alle Ladungen des Faktors F_q, mit denen dieser in die Variablen $Z_1, \ldots, Z_m$ eingeht. a_{jq} als Ladung des Faktors F_q auf die standardisierte Variable Z_j stellt das Gewicht dar, mit dem der Faktor F_q auf die Variable Z_j wirkt. Die Faktorladung a_{jq} beinhaltet somit ein Maß für den linearen Zusammenhang

zwischen der Variablen Z_j und dem Faktor F_q. Sie sind ebenfalls unbekannt und müssen gleichzeitig mit der Extraktion der Faktoren aus dem Datenmaterial geschätzt werden.

Im Unterschied zur Regressionsanalyse wird hier schon ein entscheidendes Problem der Faktoranalyse deutlich: Bei der Regressionsanalyse entscheidet der Anwender über die Anzahl der einzubeziehenden erklärenden X-Variablen, ihre Anzahl ist somit bekannt. Zum anderen liegen für die erklärenden X-Variablen der Regressionsanalyse Beobachtungswerte vor. Bei der Faktoranalyse ist die Anzahl Q der Faktoren zu Beginn der Analyse unbekannt und es gibt auch keine Beobachtungswerte für die Faktoren. Die Anzahl der Faktoren und die Faktorwerte f_{iq} jedes Faktors $(q = 1, \ldots, Q; i = 1, \ldots, n)$ müssen neben den Faktorladungen a_{jq} im Rahmen der Faktoranalyse bestimmt werden.

Wie bei den anderen multivariaten statistischen Verfahren ist es auch bei der Faktoranalyse kaum möglich, eine Variable vollständig durch gemeinsame Faktoren bzw. die Zusammenhangsstruktur aller Variablen nur auf Basis gemeinsamer Faktoren zu erklären. Das Modell der Faktoranalyse (4.50) ist noch nicht vollständig spezifiziert. Analog zur Regressionsanalyse nimmt man je Variable eine Stör- oder Fehlergröße ε_j $(j = 1, \ldots, m)$ auf, die speziell nur auf eine Variable Z_j wirkende Einflüsse, Beobachtungsfehler in den Daten usw., also den durch die gemeinsamen Faktoren nicht erklärbaren Rest beinhaltet. Diese Störgröße ist als Abweichung

$$\varepsilon_j = Z_j - \sum_{q=1}^{Q} a_{jq} F_q; \quad j = 1, \ldots, m \tag{4.51}$$

definiert. Die Extraktion der Faktoren und die Schätzung der Faktorladungen soll in der Weise erfolgen, dass diese Störgröße möglichst klein sind. Um ein einheitliches Modellkonzept zu realisieren, wird davon ausgegangen, dass sich jede Störgröße ε_j aus einem sogenannten Einzelrestfaktor U_j, der ausschließlich auf die Variable Z_j wirkt, und der dazugehörigen Faktorladung e_j zusammensetzt.

$$\varepsilon_j = e_j \cdot U_j. \tag{4.52}$$

Als weitere Forderungen für das Modell der Faktoranalyse kommt nunmehr hinzu, dass diese Einzelfaktoren untereinander und mit den gemeinsamen Faktoren unkorreliert sein sollen.
Das letztendlich spezifizierte Modell lautet:

$$Z_j = a_{j1}F_1 + \ldots + a_{jq}F_q + \ldots + a_{jQ}F_Q + e_jU_j = \sum_{q=1}^{Q} a_{jq}F_q + e_jU_j; \quad j = 1, \ldots, m \tag{4.53}$$

bzw. für jeden Wert (i = 1,. . . , n; j = 1,. . . , m) notiert

$$\begin{aligned} z_{ij} &= a_{j1}f_{i1} + \ldots + a_{jq}f_{iq} + \ldots + a_{jQ}f_{iQ} + e_ju_{ij} \\ &= \sum_{q=1}^{Q} a_{jq}f_{iq} + e_ju_{ij} \end{aligned} \tag{4.54}$$

Faßt man die $n \cdot m$ Gleichungen zu einer Matrizengleichung zusammen, so lautet diese

$$\mathbf{Z}_{(n\times m)} = \mathbf{F}_{(n\times Q)}\mathbf{A}^T_{(Q\times m)} + \mathbf{U}_{(n\times m)}\mathbf{E}_{(m\times m)}, \tag{4.55}$$

worin

$\mathbf{Z}_{(n\times m)}$ die Matrix der standardisierten Beobachtungswerte,

$\mathbf{F}_{(n\times Q)}$ die Matrix der Faktorwerte der gemeinsamen Faktoren,

$\mathbf{A}^T_{(Q\times m)}$ die Matrix der Faktorladungen der gemeinsamen Faktoren (transponiertes Faktormuster),

$\mathbf{U}_{(n\times m)}$ die Matrix der Faktorwerte der Einzelrestfaktoren und

$\mathbf{E}_{(m\times m)}$ die Diagonalmatrix mit den Faktorladungen der Einzelrestfaktoren auf der Hauptdiagonalen

sind.

Zusammenfassung der Voraussetzungen der Faktoranalyse:

- unabhängige Beobachtungen,
- großer Stichprobenumfang,
- metrisch skalierte Variablen,
- (approximativ) normalverteilte Variablen,
- Wirkung der gemeinsamen Faktoren auf mindestens zwei Variablen,
- Unkorreliertheit der gemeinsamen Faktoren,
- Unkorreliertheit der Einzelfaktoren untereinander,
- Unkorreliertheit der Einzelrestfaktoren mit den gemeinsamen Faktoren.

Die Gleichung (4.55) ist jedoch nicht lösbar, da alle Matrizen **A**, **F**, **U** und **E** auf der rechten Seite der Gleichung unbekannt sind. In einem ersten Lösungsschritt werden die Bravais-Pearson-Korrelationskoeffizienten zwischen den standardisierten Variablen berechnet:

$$r_{jk} = \frac{1}{n-1}\sum_{i=1}^{n} z_{ij}z_{ik}; \quad j,k = 1,\ldots,m \tag{4.56}$$

bzw. in Matrixnotation

$$\mathbf{R}_{(m\times m)} = \mathbf{Z}^T_{(m\times n)}\mathbf{Z}_{(n\times m)}/(n-1), \tag{4.57}$$

worin $\mathbf{R} = [r_{jk}]$, $(j, k = 1, \ldots, m)$ die Korrelationsmatrix bezeichnet, in deren Hauptdiagonalen $r_{jj} = 1$ steht. (4.55) in (4.57) eingesetzt, führt zu:

$$\begin{aligned} \mathbf{R} &= [\mathbf{F}\mathbf{A}^T + \mathbf{U}\mathbf{E}]^T[\mathbf{F}\mathbf{A}^T + \mathbf{U}\mathbf{E}]/(n-1) \\ &= \mathbf{A}\mathbf{A}^T + \mathbf{E}\mathbf{E}. \end{aligned} \tag{4.58}$$

Dies läßt sich wie folgt zeigen:

$$\begin{aligned} \mathbf{R} &= [\mathbf{F}\mathbf{A}^T + \mathbf{U}\mathbf{E}]^T[\mathbf{F}\mathbf{A}^T + \mathbf{U}\mathbf{E}]/(n-1) \\ &= [(\mathbf{F}\mathbf{A}^T)^T\mathbf{F}\mathbf{A}^T + (\mathbf{F}\mathbf{A}^T)^T\mathbf{U}\mathbf{E} + (\mathbf{U}\mathbf{E})^T\mathbf{F}\mathbf{A}^T + (\mathbf{U}\mathbf{E})^T\mathbf{U}\mathbf{E}]/(n-1) \\ &= [\mathbf{A}\mathbf{F}^T\mathbf{F}\mathbf{A}^T + \mathbf{A}\mathbf{F}^T\mathbf{U}\mathbf{E} + \mathbf{E}\mathbf{U}^T\mathbf{F}\mathbf{A}^T + \mathbf{E}\mathbf{U}^T\mathbf{U}\mathbf{E}]/(n-1) \\ &= \mathbf{A}[\mathbf{F}^T\mathbf{F}/(n-1)]\mathbf{A}^T + \mathbf{A}[\mathbf{F}^T\mathbf{U}/(n-1)]\mathbf{E} + \mathbf{E}[\mathbf{U}^T\mathbf{F}/(n-1)]\mathbf{A}^T \\ &\quad + \mathbf{E}[\mathbf{U}^T\mathbf{U}/(n-1)]\mathbf{E}. \end{aligned} \tag{4.59}$$

Zu berücksichtigen ist nun, dass

- die $(Q \times Q)$-Matrix $\mathbf{F}^T\mathbf{F}/(n-1)$ die Korrelationsmatrix der gemeinsamen Faktoren $F_1, \ldots, F_Q$ ist, da Standardisierung der Faktoren vorausgesetzt wurde. Wegen der Voraussetzung der Unkorreliertheit der Faktoren untereinander muss gelten: $\mathbf{F}^T\mathbf{F}/(n-1) = \mathbf{I}_Q$, worin $\mathbf{I}_Q$ eine Einheitsmatrix mit Q Zeilen und Spalten ist.

- die $(m \times m)$-Matrix $\mathbf{U}^T\mathbf{U}/(n-1)$ die Korrelationsmatrix der Einzelrestfaktoren $U_1, \ldots, U_m$ ist. Wegen der Voraussetzung der Unkorreliertheit der Einzelfaktoren untereinander muss gelten: $\mathbf{U}^T\mathbf{U}/(n-1) = \mathbf{I}_m$, worin $\mathbf{I}_m$ eine Einheitsmatrix mit m Zeilen und Spalten ist.

- die Matrizen $\mathbf{U}^T\mathbf{F}/(n-1)$ und $\mathbf{F}^T\mathbf{U}/(n-1)$ gerade die Korrelationskoeffizienten der Einzelrestfaktoren mit den gemeinsamen Faktoren beinhalten. Wegen der vorausgesetzten Unkorreliertheit müssen diese Matrizen Nullmatrizen sein, also $\mathbf{U}^T\mathbf{F}/(n-1) = \mathbf{0}$ und $\mathbf{F}^T\mathbf{U}/(n-1) = \mathbf{0}$.

Nach Einsetzen erhält man

$$\mathbf{R} = \mathbf{A}\mathbf{I}\mathbf{A}^T + \mathbf{A}\mathbf{0}\mathbf{E} + \mathbf{E}\mathbf{0}\mathbf{A}^T + \mathbf{E}\mathbf{I}\mathbf{E} = \mathbf{A}\mathbf{A}^T + \mathbf{E}\mathbf{E}, \tag{4.60}$$

was zu beweisen war.
Die Gleichung (4.58) wird als Fundamentaltheorem der Faktoranalyse bezeichet: Unter den Voraussetzungen des Modells läßt sich die Korrelationsmatrix $\mathbf{R}$ aus den Faktorladungen a_{jq} der gemeinsamen Faktoren und den Faktorladungen e_j der Einzelrestfaktoren berechnen.

Ausgehend von (4.58) kann jeder einzelne Korrelationskoeffizient r_{jk} als

$$r_{jk} = a_{j1}a_{k1} + \ldots + a_{jq}a_{kq} + \ldots + a_{jQ}a_{kQ} + \vartheta_{jk}e_j e_k$$
$$\vartheta_{jk} = 1 \quad \text{falls} j = k; \qquad \vartheta_{jk} = 0 \quad \text{falls} j \neq k \tag{4.61}$$

geschrieben werden. Dabei gilt speziell für die Elemente der Hauptdiagonale von $\mathbf{R}$:

$$r_{jj} = a_{j1}^2 + \ldots + a_{jq}^2 + \ldots + a_{jQ}^2 + e_j^2 = \sum_{q=1}^{Q} a_{jq}^2 + e_j^2 = 1. \tag{4.62}$$

Wegen

$$r_{jk} = \frac{Cov(Z_j, Z_k)}{Var(Z_j)Var(Z_k)} = Cov(Z_j, Z_k) \tag{4.63}$$

und

$$r_{jj} = Cov(Z_j, Z_j) = Var(Z_j) = 1, \tag{4.64}$$

gilt die Aussage von (4.62) auch für die Varianz von Z_j. a_{jq}^2 ist der Erklärungsbeitrag des Faktors F_q an der Varianz der Variablen Z_j.

Die Summe der Varianzbeiträge der gemeinsamen Faktoren, symbolisiert mit h_j^2,

$$h_j^2 = \sum_{q=1}^{Q} a_{jq}^2 = 1 - e_j^2, \quad j = 1, \ldots, m \tag{4.65}$$

wird Kommunalität der standardisierten Variablen Z_j genannt. Sie beinhaltet den Teil der Varianz von Z_j, der durch alle gemeinsamen Faktoren erklärt werden kann. Die Kommunalität ist somit der quadrierte multiple Korrelationskoeffizient der Variablen Z_j mit den gemeinsamen Faktoren.

Die Varianz bestmöglich durch die zu extrahierenden gemeinsamen Faktoren zu erklären, ist das Ziel der Analyse. Kann die Varianz einer Variablen durch die extrahierten gemeinsamen Faktoren restlos erklärt werden, wird die Kommunalität gleich 1. Das wird bei praktischen Untersuchungen jedoch kaum gelingen, d.h., die Kommunalitäten werden kleiner als 1 sein. Der Restbetrag zu 1 geht auf die Wirkung des Einzelrestfaktors zurück. e_j^2 ist der Beitrag des Einzelrestfaktors U_j an der Varianz der Variablen Z_j.

Gleichzeitig ist aber, wie aus der Formel (4.65) zu ersehen ist, die Frage zu beantworten, wie viele gemeinsame Faktoren sind erforderlich, um die Varianzen der Variablen und damit die Korrelationsmatrix hinreichend genau zu reproduzieren. Gleichzeitig wird jedoch gefordert, dass die Anzahl der Faktoren möglichst klein sein soll (extrahiert man so viele Faktoren, wie es Variablen gibt, hat man nichts gewonnen). Außerdem soll erreicht werden, dass zum einen

die Beziehungen zwischen den Faktoren und andererseits die Beziehungen zwischen den Faktoren und den Variablen (d.h. die Struktur der Faktoren) einfach sein soll, damit sie einer inhaltlichen Interpretation zugänglich sind. Bereinigt man die Matrix $\mathbf{R}$ um den Einfluß der Einzelrestfaktoren, indem $\mathbf{EE}$ auf beiden Seiten subtrahiert wird:

$$\mathbf{R} - \mathbf{EE} = \mathbf{AA}^T = \mathbf{R}_h, \tag{4.66}$$

so erhält man die sogenannte reproduzierte Korrelationsmatrix $\mathbf{R}_h$, in deren Hauptdiagonale die Kommunalitäten h_j^2 stehen. Unter Verwendung der geschätzten Kommunalitäten

$$\hat{h}_j^2 = \sum_{q=1}^{Q} \hat{a}_{jq}^2 = 1 - \hat{e}_j^2 \tag{4.67}$$

kann die reproduzierte Korrelationsmatrix $\mathbf{R}_h$ wie folgt geschrieben werden, wobei zum Vergleich auch die Korrelationsmatrix $\mathbf{R}$ angegeben wird, in deren Hauptdiagonale Einsen stehen:

$$\mathbf{R}_h = \begin{pmatrix} \hat{h}_1^2 & \cdots & r_{1j} & \cdots & r_{1m} \\ \vdots & \vdots & \ddots & \cdots & \\ r_{j1} & \cdots & \hat{h}_j^2 & \cdots & r_{jm} \\ \vdots & \vdots & \ddots & \cdots & \\ r_{m1} & \cdots & r_{mj} & \cdots & \hat{h}_m^2 \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} 1 & \cdots & r_{1j} & \cdots & r_{1m} \\ \vdots & \vdots & \ddots & \cdots & \\ r_{j1} & \cdots & 1 & \cdots & r_{jm} \\ \vdots & \vdots & \ddots & \cdots & \\ r_{m1} & \cdots & r_{mj} & \cdots & 1 \end{pmatrix}. \tag{4.68}$$

Da r_{jk} der Korrelationskoeffizient zwischen Z_j und Z_k ist, die Z-Variablen wiederum die standardisierten X-Variablen sind, sind die Elemente von $\mathbf{R}_h$ außer den Diagonalelementen $\hat{h}_j^2$ bekannt.

Aus Formel (4.65) bzw. (4.67) ist ersichtlich, dass die unbekannten Kommunalitäten von den ebenfalls unbekannten Faktorladungen abhängen. Für die Gleichung (4.66) $\mathbf{R}_h = \mathbf{AA}^T$ gibt es unendlich viele Lösungen. Ihre Schätzung kann somit nur in einem iterativen Prozeß erfolgen, indem mit vorgegebenen Anfangswerten begonnen wird und diese im Verlauf der Faktorextraktion verbessert werden, bis ein Genauigkeitskriterium erfüllt ist. Je nach Wahl der zusätzlichen Kriterien zur Erreichung einer eindeutigen Lösung existieren unterschiedliche Extraktionsverfahren, die zu verschiedenen Faktorlösungen gelangen können. Zwei häufig verwendete Verfahren sind die Hauptkomponentenmethode und die Hauptachsenmethode.[46] Die beiden Methoden unterscheiden sich im modelltheoretischen Ansatz.

★ Hauptkomponentenmethode

Der Hauptkomponentenmethode unterliegt das Modell (4.50). Das beinhaltet, dass die Varianz einer Variablen vollständig und ausschließlich durch gemeinsame Faktoren erklärt wird und kein Einzelrestfaktor existiert. Das hat zur Folge, dass die Ausgangskommunalitäten stets Eins sind, da in (4.65) keine Einzelrestladungen e_j auftreten. Wenn im

[46] Vgl. u.a. Backhaus, K. et. al. (1994), S. 221 ff.

Verlauf der Faktorextraktion so viele gemeinsame Faktoren wie Variablen extrahiert werden ($Q = m$), werden auch die Kommunalitätswerte von Eins reproduziert, da in diesem Fall die Varianz einer Variablen vollständig durch die gemeinsamen Faktoren erklärt wird:

$$\hat{h}_j^2 = \sum_{q=1}^{Q=m} a_{jq}^2 = 1.$$

Das Ziel der Faktoranalyse besteht darin, bei bestmöglicher Erklärung der Varianz der Variablen weniger gemeinsame Faktoren als Variablen ($Q < m$) zu extrahieren. Im Ergebnis der Faktorextraktion werden dann Kommunalitätswerte kleiner Eins auftreten. Die nicht erklärte Varianz der Variablen $1 - \hat{h}_j^2$ wird jedoch nicht auf das Wirken des Einzelrestfaktors, sondern auf die fehlenden gemeinsamen Faktoren zurückgeführt.

★ Hauptachsenmethode

Der Hauptachsenmethode unterliegt das Modell (4.55). Das beinhaltet, dass die Varianz einer Variablen zum überwiegenden Teil durch die gemeinsamen Faktoren erklärt und die Restvariation auf den Einzelrestfaktor zurückgeführt wird. Die Konsequenz ist, dass Einzelrestladungen e_j auftreten und die Kommunalitäten gemäß (4.65) definiert sind. Somit werden bereits die Ausgangswerte der Kommunalitäten kleiner als Eins vorgegeben. Unter SPSS wird als Startwert für die Kommunalität der Wert des Bestimmtheitsmaßes der jeweiligen Variablen in Abhängigkeit von allen anderen Variablen verwendet.

Rechentechnisch erfolgt die Bestimmung der Faktoren als Lösung des Eigenwertproblems der reellen symmetrischen Matrix $\mathbf{R}_h$. Der Erklärungsbeitrag der Faktoren im Hinblick auf die Varianz aller Variablen $X_1, \ldots, X_m$ kann durch die Eigenwerte[47] λ_q $(q = 1, 2, \ldots, Q)$ der Korrelationsmatrix $\mathbf{R}_h$ beschrieben werden, d.h. der Eigenwert eines Faktors ist der Erklärungsbeitrag dieses Faktors an der Varianz aller Variablen:

$$\lambda_q = \sum_{j=1}^{m} \hat{a}_{jq}^2 \tag{4.69}$$

Die Koeffizienten a_{jq} entsprechen den Komponenten des zu λ_q gehörenden Eigenvektors $\mathbf{v}_q$. Sie sind die Koordinaten der Variablenpunkte im Faktoren-Koordinatensystem. Die Erklärungsbeiträge der Faktoren machen in den meisten Fällen deutlich, dass eine Variablenreduzierung ohne wesentliche Informationsverluste vorgenommen werden kann.

Da angestrebt wird, eine möglichst kleine Anzahl von Faktoren zu extrahieren, muss ein Abbruchkriterium festgelegt werden, d.h. wann ein Faktor nicht mehr nennenswert zur Erklärung der Gesamtvarianz beiträgt. Dafür gibt es verschiedene Kriterien:

[47] Auch als charakteristische oder latente Wurzeln bezeichnet. Siehe Anhang H.

- die Residualmatrix: Die im letzten Schritt erreichte Residualmatrix soll in etwa der Nullmatrix entsprechen. Als Faustregel gilt, dass ihre Elemente kleiner als 0,05 sein sollen.

- die Varianzprozentanteile: Hierbei legt der Anwender fest, wieviel Prozent der Gesamtvarianz der Variablen durch die extrahierten Faktoren erklärt werden sollen. Der Erklärungsbeitrag aller gemeinsamen Faktoren an der Varianz aller Variablen ist die Summe der Eigenwerte dividiert durch die Gesamtvarianz der Variablen. Da die Varianz einer standardisierten Variablen Z_j Eins ist, ist die Gesamtvarianz aller Z-Variablen identisch mit der Anzahl der Variablen.

$$\frac{\sum_{q=1}^{Q} \lambda_q}{m} = \frac{1}{m} \sum_{q=1}^{Q} \sum_{j=1}^{m} \hat{a}_{jq}^2. \tag{4.70}$$

- Auswahl von Faktoren mit Eigenwerten größer als Eins (Kaiser-Kriterium): Dieses Kriterium orientiert sich ebenfalls an den Varianzbeiträgen der Faktoren, aber es werden nur Faktoren mit Varianzbeiträgen größer als 1 in die weitere Betrachtung einbezogen, denn nur wenn der Varianzbeitrag λ_q eines Faktors größer als 1 ist, repräsentiert dieser Faktor mehr Varianz als eine standardisierte Variable.

Mit der Festlegung eines Abbruchkriteriums wird somit die Dimension des Faktorraumes und die Zuordnung der Variablen zu den Faktoren bestimmt. Dies kann bei jedem Extraktionskriterium zu anderen Ergebnissen führen.

Bei den meisten praktischen Untersuchungen erhält man Faktoren, die von der Problemstellung her nur schwer interpretierbar sind, da sehr unterschiedliche Varianzbeiträge je Faktor auftreten und in diesem Sinne keine einfache Struktur der Faktoren erreicht wurde. Einfache Struktur eines Faktors bedeutet dabei, dass alle seine Ladungen nahe bei 1 oder nahe bei 0 liegen. Aus dem sachlichen Inhalt der hochgeladenen Variablen kann bei einfacher Struktur eindeutiger auf einen möglichen sachlichen Inhalt geschlossen werden. Eine solche einfache Struktur kann oftmals durch eine Drehung des Koordinatensystems der Faktoren um seinen Ursprung gefunden werden. Durch diese Rotation werden die bereits gefundenen Zuordnungen der Variablen zu den Faktoren in keiner Weise verändert, jedoch eine Erhöhung der Faktorladungen erreicht. Eine oft verwendete Rotationsregel lautet deshalb: Man drehe das Achsensystem der Faktoren so lange, bis die Varianz der quadrierten Faktorladungen der Faktoren ein Maximum erreicht. Dieses Verfahren wird deshalb als Varimaxrotation bezeichnet.

Als letztes verbleibt noch die Schätzung der Faktorwerte f_{iq} $(i = 1, \ldots, n; q = 1, \ldots, Q)$, die eine zusätzliche Auswertung der Faktorenanalyse ermöglichen. Die Faktorwerte f_{iq} beinhalten die Beziehung zwischen den Faktoren und den Fällen, d.h. die Bedeutung eines Faktors für

einen Fall. Durch sie lassen sich die Fälle ebenfalls im Q-dimensionalen Faktorraum abbilden. Die Faktorwerte sind in der Faktorwertematrix **F** der Formel (4.50) enthalten. Die Matrix **Z** in (4.50) ist über die Beobachtungswerte der X-Variablen gegeben, die Matrix der Faktorladungen **A** wurde im Ergebnis der Faktorextraktion ermittelt. So kann **F** bestimmt werden, indem beide Seiten von (4.50) mit $\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}$ erweitert werden:

$$\begin{aligned} \mathbf{ZA}(\mathbf{A}^T\mathbf{A})^{-1} &= \mathbf{FA}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}, \\ \mathbf{F} &= \mathbf{ZA}(\mathbf{A}^T\mathbf{A})^{-1}. \end{aligned} \tag{4.71}$$

Diese Faktorwerte sind standardisierte Werte mit dem Mittelwert Null und Varianz Eins und erlauben folgende allgemeine Interpretation: Ist ein Faktorwert Null, so weist der Fall einen im Vergleich zu allen anderen Fällen durchschnittlichen Variablenwert bezüglich dieses Faktors auf. Ein positiver Faktorwert impliziert einen überdurchschnittlichen Variablenwert eines Falles bezüglich dieses Faktors im Vergleich zu allen anderen Fällen und entsprechend ein negativer Faktorwert einen unterdurchschnittlichen Variablenwert.

Sollen im Faktormodell auch Einzelrestfaktoren berücksichtigt werden (siehe Gleichung (4.55)), so können die Faktorwerte nicht als Linearkombination der standardisierten Variablen ermittelt werden, denn die Matrix **U** der Einzelrestfaktoren ist unbekannt. Sie müssen dann z.B. mittels einer Regressionsschätzung bestimmt werden.

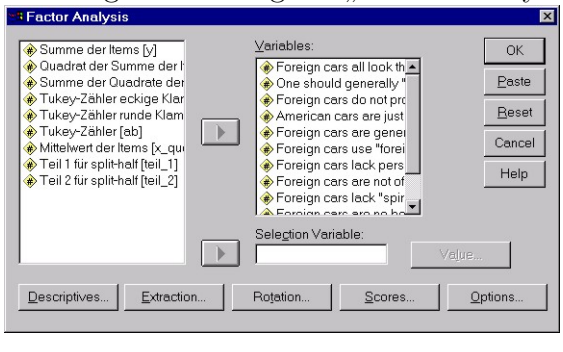
4.2.2. Die Faktoranalyse unter SPSS

Unter SPSS wird die Faktoranalyse über

- Analyze
  - Data Reduction
    - Factor...

aufgerufen, womit sich das Dialogfeld „Factor Analysis“ öffnet.

Abbildung 4.3.: Dialogfeld „Factor Analysis“

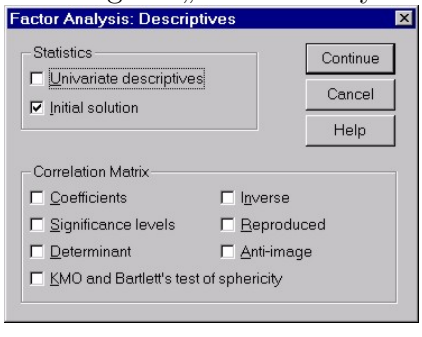


Nachdem die für die Untersuchung relevanten Variablen aus der linken Quellliste in das Feld

„Variables:“ gebracht wurden, müssen die Methoden und Kriterien gewählt, kann der Output gestaltet und Speicherung veranlaßt werden. Dabei sollen aber nur die wesentlichen Aspekte behandelt werden.

Über die Schaltfläche „Descriptives...“ gelangt man in das Dialogfeld „Factor Analysis: Descriptives“.

Abbildung 4.4.: Dialogfeld „Factor Analysis: Descriptives“



In diesem Dialogfeld können unter „Statistics“ für den Output angefordert werden:

- unter „Univariate descriptives“ der Mittelwert (Mean), die Standardabweichung (Std. Deviation) und die Anzahl die gültigen Fälle (Analysis N) für jede Variable;
- die Anfangslösung (Initial solution) mit
  - den Anfangskommunalitäten, d.h. die Diagonalelemente der reproduzierten Korrelationsmatrix $\mathbf{R}_h$ gemäß (4.66),
  - der laufenden Nummer des Faktors,
  - den (Anfangs)Eigenwerten der jeweiligen Faktoren (Initial Eigenvalues),
  - dem prozentualen Anteil der Varianz aller Variablen, der durch diesen Faktor erklärt wird (% of Variance), d.h. Eigenwert des Faktors dividiert durch die Anzahl der Variablen,
  - dem kumulierten prozentualen Anteil der erklärten Varianz (Cumulative %).

Unter „Correlation Matrix“ kann man sich u.a. ausgeben lassen:

- die Matrix der Korrelationskoeffizienten $\mathbf{R}$ zwischen allen Variablen (Coefficients);
- das einseitige Signifikanzniveau dieser Korrelationskoeffizienten (Significance levels);
- die Determinante der Korrelationsmatrix (Determinant);
- die Inverse der Korrelationsmatrix $\mathbf{R}^{-1}$ (Inverse);

- die reproduzierte Korrelationsmatrix (Reproduced), d.h. die sich nach der Faktorextraktion ergebende Korrelationsmatrix mit den reproduzierten Kommunalitäten in der Hauptdiagonalen, sowie die Differenzen zwischen den beobachteten und reproduzierten Korrelationskoeffizienten;

- die Anti-Image-Matrix
  Diese Matrix enthält außerhalb der Diagonalen die negativen partiellen Korrelationskoeffizienten zwischen zwei Variablen, die als Anti-Image-Korrelation bezeichnet werden. Wenn hinter den Variablen tatsächlich gemeinsame Faktoren stehen, dann sollten die partiellen Korrelationskoeffizienten zwischen zwei Variablen klein sein, wenn die Einflüsse der anderen Variablen ausgeschaltet wurden. Mit den partiellen Korrelationskoeffienten wird somit die Korrelation zwischen den Einzelrestfaktoren bei diesen beiden Variablen geschätzt. Hohe partielle Korrelationskoeffizienten zeigen an, dass die Beziehung zwischen zwei Variablen nicht in einem großem Ausmaß durch die anderen Variablen (und damit durch die dahinter stehenden gemeinsamen Faktoren) erklärt werden kann. Bei einer großen Anzahl hoher Koeffizienten ist die Faktoranalyse kaum sinnvoll.

  In der Diagonalen der Anti-Image-Matrix steht für jede Variable ein Maß für die Angemessenheit der Stichproben (measure of sampling adequacy, MSA). Hiermit wird das Ausmaß der einfachen linearen Korrelation r_{jk} zwischen den betrachteten Variablen X_j und den anderen Variablen X_k $(k = 1, \ldots, m; j \neq k)$ mit dem Ausmaß der partiellen Korrelation der betrachteten Variablen X_j und den anderen Variablen X_k (symbolisiert mit p_{jk}) verglichen:

$$MSA_j = \frac{\sum\limits_{j \neq k} r_{jk}^2}{\sum\limits_{j \neq k} r_{jk}^2 + \sum\limits_{j \neq k} p_{jk}^2}. \tag{4.72}$$

  MSA_j wird einen Wert nahe Eins annehmen, wenn die Summe der quadrierten partiellen Korrelationskoeffizienten der Variablen X_j und den anderen Variablen X_k klein ist. Ist der Wert von MSA_j dagegen klein, sollte die Variable X_j aus der Faktoranalyse herausgenommen werden.

- das Kaiser-Meyer-Olkin-Maß und den Bartlett-Test (KMO and Bartlett's test of sphericity)
  KMO ist ein Maß für die Angemessenheit der Stichprobe (measure of sampling adequacy)

unter Einbeziehung aller Paare von Variablen:

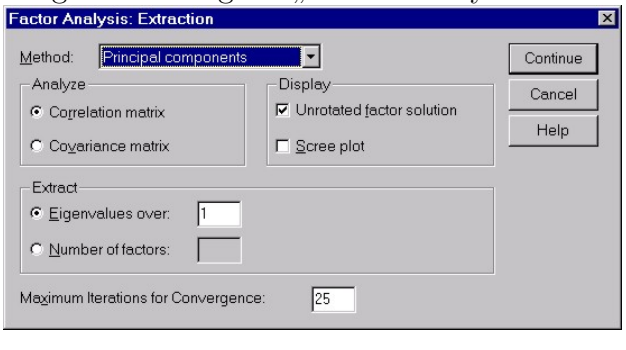
$$KMO = \frac{\sum\sum_{j \neq k} r_{jk}^2}{\sum\sum_{j \neq k} r_{jk}^2 + \sum\sum_{j \neq k} r_{jk}^2}. \tag{4.73}$$

Wenn der Wert von KMO klein ist, dann ist die Durchführung einer Faktoranalyse mit diesen Variablen nicht sinnvoll.

Der Bartlett-Test prüft die Nullhypothese, dass die Korrelationsmatrix $\mathbf{R}$ in der Grundgesamtheit eine Einheitsmatrix ist, d.h., keinerlei einfache lineare Korrelation zwischen irgendwelchen Paaren von Variablen existiert. Voraussetzung für diesen Test ist, dass die Daten (zumindest approximativ) aus einer multivariaten normalverteilten Grundgesamtheit stammen. Kann die Nullhypothese nicht abgelehnt werden, sollte keine Faktoranalyse mit diesen Variablen durchgeführt werden.

Über die Schaltfläche „Extraction…“ gelangt man in das Dialogfeld „Factor Analysis: Extraction“, in dem vor allem die Faktorextraktionsmethode und die Anzahl der Faktoren zu entscheiden sind.

Abbildung 4.5.: Dialogfeld „Factor Analysis: Extraction“



Insgesamt werden 7 Extraktionsmethoden angeboten, wobei hier nur auf die behandelten Hauptkomponentenmethode (Principal components, PC) und die Hauptachsenmethode (Principal axis factoring) zurückgegriffen werden soll.
Bei der Anzahl der zu extrahierenden Faktoren kann der Nutzer zwischen zwei Möglichkeiten entscheiden:

- Eigenvalues over
  Es werden nur Faktoren extrahiert, deren Eigenwert einen bestimmten Wert überschreitet.

Voreingestellt ist ein Wert von 1 (Kaiser-Kriterium). Dieser Wert kann aber verändert werden.

- Number of factors
  Der Nutzer gibt die Anzahl der zu extrahierenden Faktoren ein. Dies erfordert jedoch schon einiges Wissen über die Struktur der Faktoren.

Desweiteren kann der Nutzer

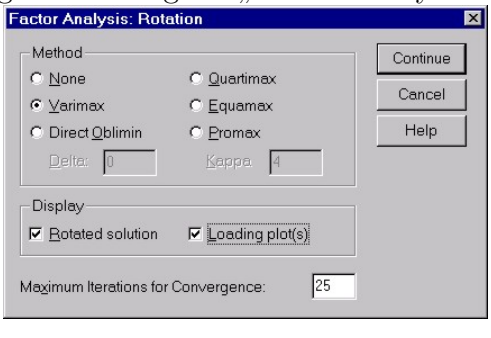
- die maximale Anzahl von Schritten (Iterationen) zur Erreichung des Konvergenzkriteriums verändern (Maximum Iterations for Convergence),
- entscheiden, ob die Korrelations- oder Kovarianzmatrix bei der Extraktion zugrunde gelegt werden soll (Analyze).

Unter „Display“ wird der Output gestaltet:

- Ausgabe der unrotierten Faktorlösung (unrotated factor solution), im Output mit Component Matrix bzw. Factor Matrix überschrieben. Sie enthält die Faktorladungen, wobei die Zeilen den Variablen und die Spalten den Faktoren entsprechen.
- Ausgabe einer Graphik (Scree plot), die ein Plot der Eigenwerte in abnehmender Folge über den Nummern der Faktoren darstellt.

Über die Schaltfläche „Rotation...“ öffnet sich das Dialogfeld „Factor Analysis: Rotation“.

Abbildung 4.6.: Dialogfeld „Factor Analysis: Rotation“



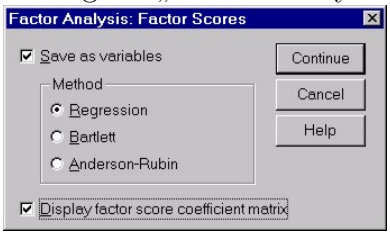
Unter „Method“ kann eine Wahl aus 5 verschiedenen Rotationsmethoden getroffen werden. Hier soll nur die Varimax-Methode betrachtet werden.
Unter „Display“ sind wiederum Optionen für die Ausgabe enthalten:

- Rotated solution
  die rotierte Faktorladungsmatrix (im Output überschrieben mit Rotated Factor Matrix);

- Loading plot(s)
  Graphik(en), die die Variablen nach der Rotation mittels der Faktorladungen im maximal dreidimensionalen Faktorraum zeigen.

Über die Schaltfläche „Scores. . . “ gelangt man in das Dialogfeld „Factor Analysis: Factor Scores“.

Abbildung 4.7.: Dialogfeld „Factor Analysis: Factor Scores“

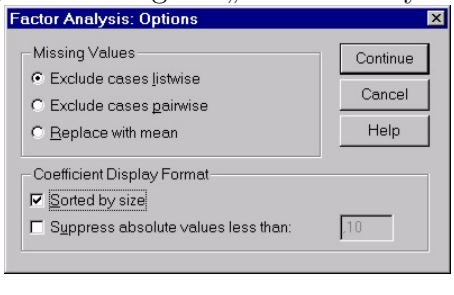


Hier kann SPSS angewiesen werden, die Faktorwerte zu speichern, wobei der Datei für jeden Faktor eine neue Variable hinzugefügt wird. Zur Berechnung der Faktorwerte stehen 3 Methoden zur Auswahl, von denen hier nur die Regressionsschätzung eine Rolle spielen soll.

Weiterhin kann die Matrix der Score-Koeffizienten der Faktorwerte (Display factor score coefficient matrix) in den Output aufgenommen werden. Sie enthält die Koeffizienten der Matrix $\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}$ aus Formel (4.71), mit denen die standardisierten Variablenwerte $\mathbf{Z}$ multipliziert werden müssen, um zu den Faktorwerten zu gelangen.

Über die Schaltfläche „Options“ sollte im Dialogfeld „Factor Analysis: Options“ auf „Sorted by size“ (sortiert nach der Größe der Faktorladungen) entschieden werden, da dann im Output die Faktorladungen und damit die Variablen gruppiert und in abnehmender Reihenfolge je Faktor angezeigt werden, was einen schnelleren Überblick über mögliche Inhalte der Faktoren erleichtert.

Abbildung 4.8.: Dialogfeld „Factor Analysis: Options“



4.2.3. Die Verwendung der Faktoranalyse zur Homogenitätsprüfung

Zurückkommend auf das Problem der Einschätzung der Homogenität und in Fortführung des Beispiels 4.1 muss nunmehr geprüft werden, ob hinter den einbezogenen Variablen (Items) nur ein einziger Faktor, d.h. nur eine synthetische Variable, steht. Im Dialogfeld „Factor Analysis“ werden die 10 Fragen (Items) aus der linken Quellliste in das Feld „Variables“ gebracht.

a) KMO und Bartlett's Test

In dem Dialogfeld „Factor Analysis: Descriptives“ (siehe Abb. 4.4) wird zunächst nur das Kaiser-Meyer-Olkin-Maß (KMO) und Bartlett's Test angefordert, um grundsätzlich die Angemessenheit der Stichprobe für die Durchführung der Faktoranalyse einzuschätzen. Den zugehörigen Teil des Outputs enthält Tabelle 4.2.

Tabelle 4.2.: KMO und Bartlett's Test

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,863
Barlett's Test of Sphericity	Approx. Chi-Square	282,850
	df	45
	Sig.	,000

Da das Kaiser-Meyer-Olkin-Maß ausreichend groß ist und mit Bartlett's Test die Nullhypothese, dass die Korrelationsmatrix in der Grundgesamtheit eine Einheitsmatrix ist, auf dem 5%-Niveau verworfen wird, kann mit der vorliegenden Stichprobe eine Faktoranalyse durchgeführt werden.
Für die weitere Bearbeitung des Beispiels wird

- im Dialogfeld „Factor Analysis: Descriptives“ (siehe Abb. 4.4) nur auf die Ausgabe der Ausgangslösung (Initial solution) entschieden; auf die Ausgabe der deskriptiven Statistiken und der umfangreichen Matrizen wird verzichtet, da dies bereits bei der Reliabilitätsanalyse erfolgte;

- im Dialogfeld „Factor Analysis: Extraction“ (siehe Abb. 4.5) die Hauptkomponentenmethode (Principal components) bei Zugrundelegung der Korrelationsmatrix gewählt, da für die Homogenitätsprüfung nur die Anzahl der gemeinsamen Faktoren entscheidend ist;

- im gleichen Dialogfeld die unrotierte Faktorlösung und der Scree Plot angefordert, wobei Faktoren mit Eigenwerten größer als 1 (Kaiser Kriterium) extrahiert werden sollen;

- im Dialogfeld „Factor Analysis: Rotation“ (siehe Abb. 4.6) als Rotationsmethode die Varimaxmethode gewählt sowie die rotierte Lösung und die Loading Plots angefordert;
- im Dialogfeld „Factor Analysis: Factor Scores“ (siehe Abb. 4.7) Regression für die Berechnung der Faktorwerte gewählt, das Speichern der Faktorwerte und die Ausgabe der Matrix der Koeffizienten der Faktorwerte veranlaßt;
- im Dialogfeld „Factor Analysis: Options“ (siehe Abb. 4.8) SPSS angewiesen, die Faktorladungen nach der Größe zu ordnen.

Obwohl im SPSS-Output die Variablenlabel erscheinen, werden hier aus Platzgründen bei der Wiedergabe des Outputs die Variablennamen verwendet. Darüber hinaus werden die Teile des Outputs einzeln wiedergegeben und kommentiert.

b) Faktorextraktion

Die Angaben zur Faktorextraktion bei Verwendung der Hauptkomponentenmethode (Principal components) enthält die Tabelle 4.3.

Ausgangslösung:

Bezüglich der Ausgangslösung wird am Beispiel deutlich, was bereits im Abschnitt 4.2.1 erläutert wurde:

- Es werden zunächst so viele Faktoren extrahiert, wie es Variablen gibt.
- Die Kommunalitäten der Ausgangslösung bei der Hauptkomponentenmethode (Spalte Initial in der Teiltabelle Communalities) sind alle Eins, da keine Einzelrestfaktoren unterstellt werden. Da die Kommunalitäten die Diagonalelemente der reproduzierten Korrelationsmatrix sind, ist bei der Ausgangslösung die reproduzierte Korrelationsmatrix $\mathbf{R}_h$ identisch mit der Korrelationsmatrix $\mathbf{R}$. Wegen (4.64) ist die Kommunalität h_j^2 bei jedem standardisierten Item Z_j gleich der Varianz von Z_j, so dass die Summe der Kommunalitäten gleich der Summe der Varianzen der Z_j bzw. gleich der Anzahl der einbezogenen Items ist.
- In der Teiltabelle Total Variance Explained stehen unter Initial Eigenvalues, Total, die Eigenwerte λ_q jedes gemeinsamen Faktors ($q = 1, \ldots, 10$). Die Summe der Eigenwerte aller Faktoren ergibt die Gesamtvarianz aller Variablen Z_j. Der Eigenwert eines Faktors beinhaltet den Erklärungsbeitrag dieses Faktors an der Varianz aller Variablen. So erklärt der 1. Faktor 40,24% der Varianz aller Variablen (Spalte: % of Variance), während der 10. Faktor nur noch einen Erklärungsbeitrag von 3,16% aufweist.

Tabelle 4.3.: Faktorextraktion mit der Hauptkomponentenmethode

Communalities

	Initial	Extraction
item 1	1,000	,603
item 2	1,000	,654
item 3	1,000	,481
item 4	1,000	,542
item 5	1,000	,953
item 6	1,000	,815
item 7	1,000	,567
item 8	1,000	,533
item 9	1,000	,475
item 10	1,000	,507

Extraction Method: Principal Component Analysis.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4,024	40,236	40,236	4,024	40,236	40,236
2	1,059	10,590	50,825	1,059	10,590	50,825
3	1,046	10,464	61,289	1,046	10,464	61,289
4	,817	8,170	69,459			
5	,741	7,409	76,868			
6	,613	6,132	83,000			
7	,514	5,144	88,144			
8	,460	4,605	92,748			
9	,410	4,097	96,845			
10	,316	3,155	100,000			

Extraction Method: Principal Analysis

Component Matrix[a]

	Component		
	1	2	3
2	,797	4,469E-02	-,128
1	,773	5,615E-02	-4,362E-02
7	,719	,197	-9,966E-02
8	,717	4,659E-02	,130
10	,706	-3,261E-02	-8,387E-02
3	,670	-1,160E-02	,179
9	,643	-,224	-,103
4	,607	-,388	,154
6	,160	,858	-,230
5	6,753E-02	,273	,935

Extraction Method: Principal Component Analysis .

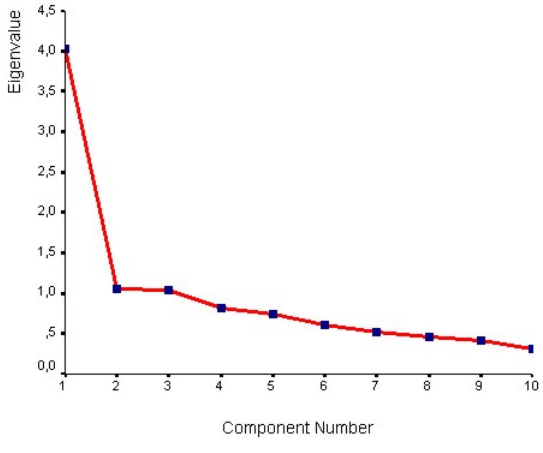
a. 3 components extracted.

Faktorextraktion:
Bei der Faktorextraktion nach dem Kaiser-Kriterium werden 3 Faktoren extrahiert (Teiltabelle Total Variance Explained unter Extraction Sums of Squared Loadings). Dies sind die Faktoren mit Eigenwerten größer als Eins. Diese 3 Faktoren erklären 61,29% der Varianz aller Variablen, wobei allein auf den 1. Faktor 40,24% entfallen. Gemäß (4.70) ergibt sich:

$$\frac{\sum_{q=1}^{3} \lambda_q}{\sum_{j=1}^{10} s_j^2} = \frac{4,024 + 1,059 + 1,046}{10} = \frac{6,129}{10} = 0,6129.$$

Abb. 4.9 enthält den Faktor-Screeplot.

Abbildung 4.9.: Faktor-Screeplot



Wenn nur diese 3 Faktoren verwendet werden, so resultieren die geschätzten Kommunalitäten $\hat{h}_j^2$ in der Spalte Extraction der Teiltabelle Communalities. Sie stehen in der Hauptdiagonalen der reproduzierten Korrelationsmatrix $\mathbf{R}_h$ (4.68), was sich leicht überprüfen läßt, indem man sich über das Dialogfeld „Factor Analysis: Descriptives“ die reproduzierte Korrelationsmatrix ausgeben läßt.

Component Matrix (dritter Teil in der Tabelle 4.3) ist die Matrix $\mathbf{A}$, die die Faktorladungen a_{jp} der unrotierten Faktorlösung enthält ($j = 1, \ldots, 10; q = 1, 2, 3$), wobei die Zeilen den Items und die Spalten den Faktoren (Components) entsprechen. Jedoch wurde in dieser Matrix nach der Größe der Faktorladungen des 1. Faktors geordnet. Die Faktorladungen sind die Koordinaten der Items im dreidimensionalen Faktorraum.

Gemäß (4.67) gilt, dass die Kommunalität jedes Items gleich der Summe der quadrierten Faktorladungen über die Faktoren ist. So folgt z.B. für die Kommunalität des ersten Items

$(j = 1)$:

$$\hat{h}_1^2 = a_{11}^2 + a_{12}^2 + a_{13}^2 = 0,773^2 + 0,05615^2 + (-0,04362)^2 = 0,603.$$

Ferner gilt gemäß (4.69), dass der Eigenwert eines Faktors gleich der Summe der quadrierten Faktorladungen über die Items ist. So folgt z.B. für den Eigenwert des ersten Faktors $(q = 1)$:

$$\begin{aligned} \lambda_1 &= 0,797^2 + 0,773^2 + 0,719^2 + 0,717^2 + 0,706^2 + \\ &\quad 0,670^2 + 0,643^2 + 0,607^2 + 0,160^2 + 0,06753^2 \\ &= 4,023 \end{aligned}$$

Aus der Component-Matrix ist zu erkennen, dass Item 6 hoch korreliert ist mit dem Faktor 2 und Item 5 den Faktor 3 hoch lädt, während alle anderen Items mit dem Faktor 1 relativ hoch korreliert sind.

c) Faktorrotation

Die mittels der Varimax-Rotation erzeugten Ergebnisse enthält Tabelle 4.4.

Tabelle 4.4.: Ergebnisse der Rotation

Total Variance Explained

Component	Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %
1	4,007	40,071	40,071
2	1,067	10,665	50,737
3	1,055	10,552	61,289

Extraction Method: Principal Component Analysis.

Rotated Component Matrix[a]

	Component		
	1	2	3
2	,797	,121	-7,016E-02
1	,769	,107	1,323E-02
10	,709	2,962E-02	-5,308E-02
7	,708	,255	-4,946E-03
8	,707	4,732E-02	,175
3	,662	-2,453E-02	,204
9	,660	-,152	-,127
4	,625	-,382	7,406E-02
6	,112	,895	2,074E-02
5	1,679-02	8,692E-03	,976

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 4 iterations.

Component Transformation Matrix

Component	1	2	3
1	,997	,054	,052
2	-,066	,960	,272
3	-,035	-,275	,961

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

Component Score Coefficient Matrix

	Component		
	1	2	3
item 1	,190	,073	-,016
item 2	,199	,085	-,096
item 3	,161	-,049	,170
item 4	,169	-,384	,050
item 5	-,031	,003	,930
item 6	-,006	,840	,011
item 7	,169	,215	-,032
item 8	,170	,018	,141
item 9	,177	-,168	-,144
item 10	,180	,002	-,076

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser
Component Scores.

Component Score Covariance Matrix

Component	1	2	3
1	1,000	,000	,000
2	,000	1,000	,000
3	,000	,000	1,000

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser
Component Scores.

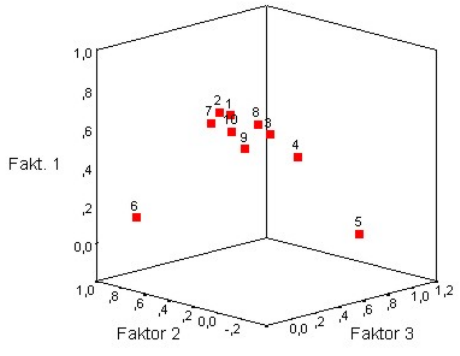
Die Rotated Component Matrix enthält die Faktorladungen a_{jq} der rotierten Faktorlösung. Es sind die Koordinaten der Variablen im rotierten dreidimensionalen Faktorraum, den die Abb. 4.10 zeigt.
Im Vergleich zur unrotierten Component-Matrix ist ersichtlich, dass für dieses Beispiel die Rotation keine substantiellen Verbesserungen gebracht hat.

Die Component Transformation Matrix ist diejenige Matrix, mit der die unrotierte Faktorladungsmatrix multipliziert werden muss, um die rotierte Faktorladungsmatrix zu erhalten (unter Berücksichtigung veränderter Zeilenabfolge):

$$\mathbf{A}_{unrotiert} \times \text{Component Transformation Matrix} = \mathbf{A}_{rotiert}$$

Die Component Score Coefficient Matrix entspricht der Matrix $\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}$ aus der Formel (4.71). Sie enthält die Koeffizienten, mit denen die standardisierten Variablenwerte z_{ij} multi-

Abbildung 4.10.: Faktor-Plot der rotierten Faktorlösung



pliziert werden müssen, um zu den Faktorwerten f_{iq} zu gelangen, die in den neuen Variablen fac1_1, fac2_1 und fac3_1 gespeichert wurden.

Die Component Score Covariance Matrix verdeutlicht, dass die Faktoren orthogonal, d.h. unkorreliert, sind.

d) Auswertung der Faktorenanalyse

Ein großes Problem bei der Faktorenanalyse stellt sich erst nach der Extraktion der Faktoren ein: ihre inhaltliche Interpretation. Dabei geht man von der Component-Matrix bzw. Faktor-Matrix (im Falle einer durchgeführten Rotation von der rotierten Component- bzw. Faktor-Matrix) aus. Aufgrund derjenigen Variablen, die einen Faktor hochladen, versucht man dem Faktor eine inhaltliche Bedeutung beizumessen. Im Prinzip steht dahinter die Formulierung von Hypothesen über Zusammenhänge zwischen den beobachteten Variablen und den Faktoren als Repräsentanten der Konstrukte.

Für das gegebene Problem der Homogenitätsanalyse stellt sich die inhaltliche Interpretation der Faktoren nicht zwangsläufig, denn es soll nur festgestellt werden, ob hinter den Items **ein** theoretisches Konstrukt und damit ein Faktor steht. Für das Beispiel muss dies verneint werden. Man sollte somit mit diesen 10 Items keine synthetische Variable bilden, die ein theoretisches Konstrukt „Voreingenommenheit gegenüber ausländischen Autos“ messen soll.

Die Auswertung der rotierten Component-Matrix zeigt jedoch, dass die bereits bei der Reliabilitätsanalyse aufgefallenen Items 5 und 6 auch hier jeweils einen anderen Faktor sehr hoch laden. Mit der Frage 6 identifizieren die befragten Personen offensichtlich eher ein „generelles Vertrauensproblem in die Technologie“ (zweiter Faktor) und mit der Frage 5 ein „generelles Preisproblem“ (dritter Faktor). Die Hoffnung besteht, dass möglicherweise durch die Herausnahme dieser beiden Items nur ein einziger Faktor extrahiert werden kann, d.h. die restlichen Items eventuell homogen sind. Es wird deshalb eine Faktoranalyse ohne diese beiden Fragen

durchgeführt. Im Ergebnis (auf die Wiedergabe des SPSS-Outputs soll aus Platzgründen verzichtet werden) erhält man nur noch einen Faktor mit einem Eigenwert größer als 1, der 50% der Varianz aller Variablen erklärt. Die synthetische Variable könnte entweder aus den Items 1-4 und 7-10 als ihre Summe gebildet werden (wie bei der Reliabilitätsanalyse unterstellt) oder man verwendet den extrahierten Faktor mit seinen Faktorwerten als synthetische Variable.

Anhang A

Beispiel für Lowess

Die Kurvenanpassung Lowess (locally weighted regression scatter plot smoothing) soll anhand eines Beispiels demonstriert werden. Die Werte der Variablen X und Y seien wie folgt, wobei die Reihe bereits nach der Größe der x_i geordnet wurde:

x_i	55	110	123	140	150	165	178	180	183	190
y_i	23	41	54	75	44	61	65	103	114	115

Beläßt man die unter SPSS gegebene Voreinstellung für den Prozentsatz der einzubeziehenden Beobachtungswerte mit 50%, so folgt für die Anzahl der Punkte $K = 10 \cdot 0,5 = 5$.

Soll nun an der Stelle x_1 ein Vorhersagewert $\hat{y}_1$ für Y ermittelt werden, so werden die Punkte (x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4) und (x_5, y_5) einbezogen. Die Abstände von x_1 zu den einzelnen Nachbarwerten enthält die Spalte d(1) der Tabelle A.1, wobei x_1 als ein Nachbar zu sich selbst betrachtet wird.

Der Abstand von $x_1 = 55$ zum K = 5ten nächsten Nachbar beträgt 95, so dass die Bereichsgrenzen (mit BG in der Tabelle A.1 gekennzeichnet) mit

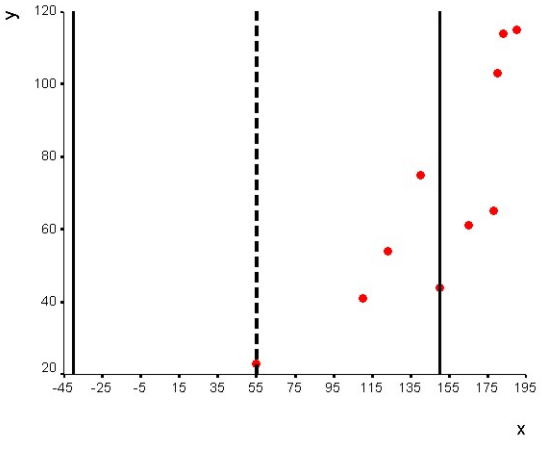
$$x_1 - \max[d(1)] = 55 - 95 = -40 \quad \text{und} \quad x_1 + \max[d(1)] = 55 + 95 = 150$$

festzusetzen sind. Abbildung A.1 zeigt diesen Bereich (durchgezogene senkrechte Linien), wobei deutlich wird, dass x_1 im Zentrum des Bereichs liegt (gestrichelte Linie).

Tabelle A.1.: Beobachtungswerte x_1 und y_1 und Abstände zu den Bereichsnachbarn

y_1	x_1	d(1)	d(2)	d(3)	d(4)	d(5)	d(6)	d(7)	d(8)	d(9)	d(10)
23	55	0	**55**								
41	110	55	0	13	**30**						
54	123	68	13	0	17	27					
75	140	85	30	17	0	10					
44	150	**95**	40	27	10	0	15				
61	165			**42**	15	15	0	**13**	**15**	**18**	**25**
65	178					**28**	13	0	2	5	12
103	180						15	2	0	3	10
114	183						**18**	5	3	0	7
115	190							12	10	7	0
BG		-40	55	81	110	122	147	165	165	165	165
		150	165	165	170	178	183	191	195	201	215

Abbildung A.1.: Bereich für die Vorhersage von Y an der Stelle x_1



Nunmehr werden sogenannte Nachbarschaftsgewichte nach

$$W(x_k) = \left(1 - \left|\frac{x_i - x_k}{d_i}\right|^3\right)^3$$

berechnet, wobei d_i der Abstand von x_i zum K-ten nächsten Nachbarpunkt ist. Diese Gewichtsfunktion ist symmetrisch, erreicht den größten Wert an der Stelle x_i und den Wert Null jeweils an den Bereichsgrenzen.

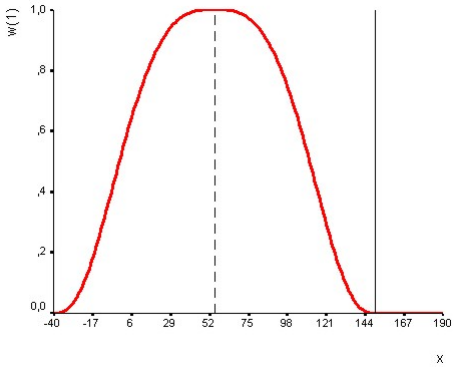
Die Nachbarschaftsgewichte $w(x_1)$ für den Bereich um x_1 enthält die Spalte w(1) der Tabelle A.2. Die näher bei x_1 liegenden X-Werte erhalten eine größere Gewichtung als die weiter entfernt liegenden Werte. Die Gestalt dieser Gewichtsfunktion zeigt die Abb. A.2, wobei die Werte der Gewichtsfunktion zur besseren Veranschaulichung nicht nur für die beobachteten x_k-Werte

$(k = 1, \ldots, 5)$ sondern auch für Zwischenwerte berechnet wurde.

Tabelle A.2.: Nachbarschaftsgewichte

x_i	w(1)	w(2)	w(3)	w(4)	w(5)
55	1,0000	0,0000			
110	0,5235	1,0000	0,9137	0,0000	
123	0,2540	0,9609	1,0000	0,5474	,0011
140	0,0228	0,5879	0,8140	1,0000	0,8695
150	0,0000	0,2330	0,3960	0,8930	1,0000
165			0,0000	0,6699	0,6060
178					0,0000
BG	-40	55	81	110	122
	150	165	165	170	178
x_i	w(6)	w(7)	w(8)	w(9)	w(10)
150	0,0748				
165	1,0000	0,0000	0,0000	0,0000	0,0000
178	0,2421	1,0000	0,9929	0,9371	0,7036
180	0,0748	0,9891	1,0000	0,9862	0,8200
183	0,0000	0,8388	0,9762	1,0000	0,9356
190		0,0097	0,7250	0,8337	1,0000
BG	147	165	165	165	165
	183	191	195	201	215

Abbildung A.2.: Nachbarschaftsgewichte im Bereich um x_1



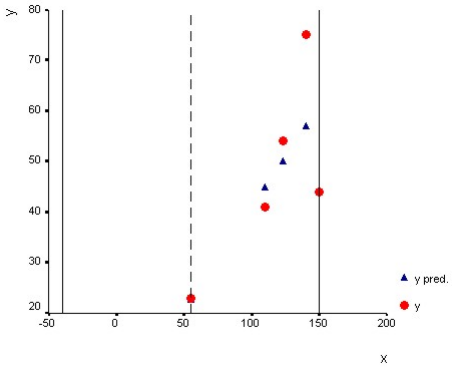
Unter Verwendung der einbezogenen Punkte (x_1, y_1) bis (x_5, y_5) wird nunmehr eine mit $w(x_k)$ gewichtete lineare Regressionsfunktion geschätzt, die

$$\sum_k w(x_k)(y_k - b_0 - b_1 x_k)^2$$

minimiert. Die geschätzte Regressionsfunktion $\hat{y}_k = b_0 + b_1 x_k = 0,376 + 0,403 x_k$ zeigt die mittlere Abhängigkeit der Variablen X in diesem ausgewählten Bereich von X-Werten.

Von den geschätzten Regreßwerten $\hat{y}_k$ (y pred. in Abbildung A.3) wird jedoch nur der Wert $\hat{y}_1$ an der Stelle x_1 verwendet: $\hat{y}_1 = 22,53$.

Abbildung A.3.: Beobachtete und geschätzte Y-Werte im Bereich um x_1



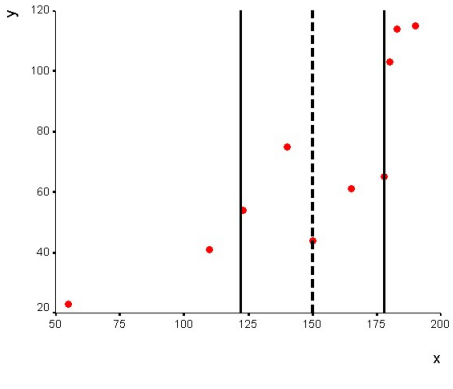
Die gleiche Prozedur wird für alle anderen X-Werte durchgeführt.

Wenn z.B. an der Stelle x_5 ein Vorhersagewert $\hat{y}_5$ für Y ermittelt wird, so werden die Punkte (x_3, y_3), (x_4, y_4), (x_5, y_5), (x_6, y_6) und (x_7, y_7) einbezogen. Die Abstände von x_5 zu den einzelnen Nachbarwerten enthält die Spalte d(5) der Taballe A.1, wobei x_5 ebenfalls wieder als ein Nachbar zu sich selbst betrachtet wird. Der Abstand von $x_5 = 100$ zum K = 5ten nächsten Nachbar beträgt 28, so dass die Bereichsgrenzen mit

$$x_5 - \max[d(5)] = 150 - 28 = 122 \qquad x_5 + \max[d(5)] = 150 + 28 = 178$$

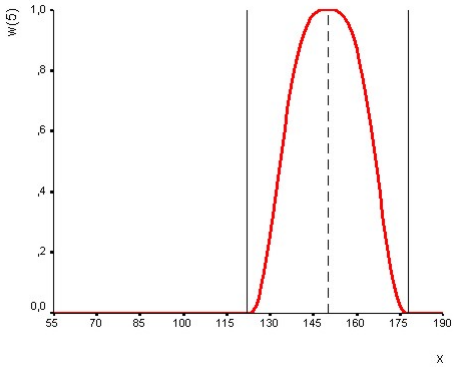
festzusetzen sind. Abbildung A.4 zeigt diesen Bereich.

Abbildung A.4.: Bereich für die Vorhersage von Y an der Stelle x_5



Die Nachbarschaftsgewichte $w(x_5)$ für den Bereich um x_5 enthält die Spalte w(5) der Tabelle A.2. Abbildung A.5 zeigt die Gewichtsfunktion.

Abbildung A.5.: Nachbarschaftsgewichte im Bereich um x_5

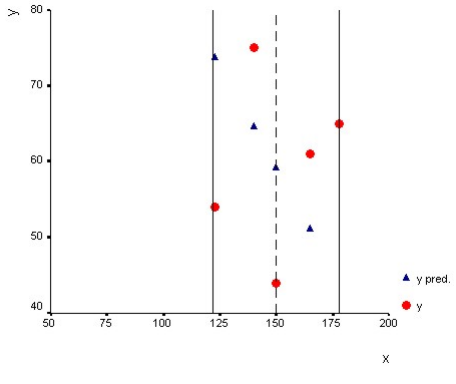


Die geschätzte Regressionsfunktion im Bereich um x_5 ist:

$$\hat{y}_k = b_0 + b_1 x_k = 140,001 - 0,539 x_k.$$

Von den geschätzten Regreßwerten $\hat{y}_k$ wird jedoch nur der Wert $\hat{y}_5$ an der Stelle x_5 verwendet: $\hat{y}_5 = 59,15$.

Abbildung A.6.: Beobachtete und geschätzte Y-Werte im Bereich um x_5



Nachdem diese Prozedur für alle $i, \ldots, n$ (mit $n = 10$ für dieses Beispiel) durchlaufen wurde, liegen für alle x_i geschätzte Werte $\hat{y}_i$ vor (siehe Tabelle A.3, Spalte 3 und Abbildung A.7).

Abbildung A.7.: Beobachtete und geschätzte Y-Werte für alle x_i

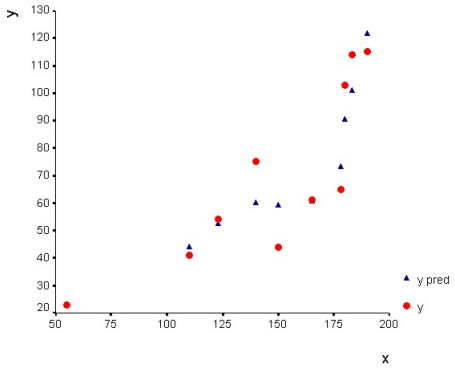


Tabelle A.3.: Beobachtungswerte x_i und y_i, geschätzte Werte $\hat{y}_i$, Residuen $\hat{u}_i$ und Robustheitsgewichte

y_i	x_i	$\hat{y}_i$	$\lvert\hat{u}_i\rvert$	$\lvert e_i\rvert$	$G(x_i)$
23	55	22,53	0,47	0,0107	0,99977
41	110	43,95	2,95	0,0669	0,99103
54	123	52,25	1,75	0,0397	0,99685
75	140	59,94	15,06	0,3420	0,77975
44	150	59,13	15,13	0,3434	0,77803
61	165	60,61	0,39	0,0088	0,99985
65	178	73,04	8,04	0,1825	0,93453
103	180	90,27	12,73	0,2890	0,83997
114	183	100,72	13,28	0,3015	0,82641
115	190	121,65	6,65	0,1509	0,95499

In der zweiten Stufe des Lowess-Verfahrens wird der Tatsache Rechnung getragen, dass die Schätzung der Y-Werte in den Bereichen um die x_i stark von Ausreißern beeinflußt sein kann. Es werden sogenannte Robustheitsgewichte bestimmt. Dazu werden die Residuen $\hat{u}_i = y_i - \hat{y}_i$ für alle $i = 1, \ldots, n$ (Spalte 4 der Tabelle A.3) sowie der Median der absoluten Residuen m = median $|\hat{u}_i|$ ermittelt. Anschließend werden die Residuen durch 6m dividiert: $e_i = \hat{u}_i/6m$ (Spalte 5 der Tabelle A.3. Unter der Voraussetzung normalverteilter Residuen ist 6m eine Schätzung von 4σ, wobei σ die Standardabweichung der Grundgesamtheit ist. Die Werte e_i bilden die Basis der Bestimmung der Robustheitsgewichte (Spalte 6 der Tabelle A.3), die nach der Formel

$$G(x_i) = \begin{cases} (1 - e_i^2)^2 & \text{für } |e| < 1 \\ 0 & \text{sonst} \end{cases}$$

berechnet werden. Je kleiner ein Residuum im Vergleich zu 6m ist, desto größer ist sein Robustheitsgewicht. Für Residuen, die größer als oder gleich 6m sind, ist das Robustheitsgewicht Null.

Abbildung A.8 zeigt den Plot der Residuen gegen die beobachteten X-Werte und Abbildung A.9 den Plot der Robustheitsgewichte gegen die Residuen.

Abbildung A.8.: Residuen-Plot

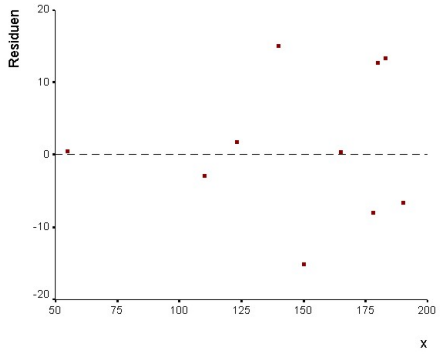
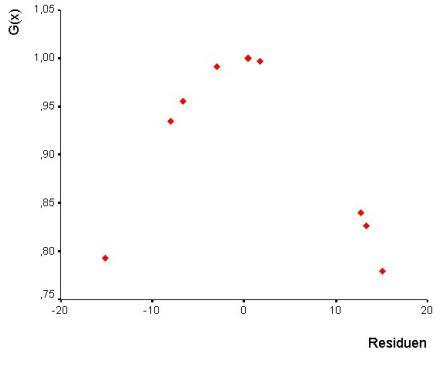
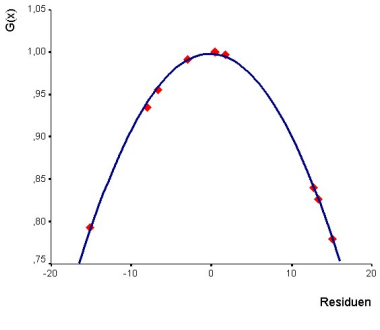


Abbildung A.9.: Robustheitsgewichte $G(x_i)$



Zur besseren Veranschaulichung der Funktion wurden Robustheitsgewichte $G(x_i)$ auch für Residuenwerte bestimmt, da sich nicht aufgrund der Werte des Beispiels ergeben (siehe Abbildung A.10).

Abbildung A.10.: Funktion der Robustheitsgewichte



Diese Robustheitsgewichte, multipliziert mit den Nachbarschaftsgewichten, werden für eine erneute Anpassung einer Regressionslinie innerhalb der einzelnen Bereiche verwendet:

$$\sum_k w(x_k)G(x_k)(y_k - b_0 - b_1 x_k)^2$$

Dadurch erhält man eine Reihe neuer geglätteter Werte $\hat{y}_i$.
Der Schritt der Bestimmung von Robustheitsgewichten kann mehrmals wiederholt werden. Je höher die vorgegebene Anzahl der Iterationen ist, desto genauer ist die Anpassung (wobei jedoch die weitere Erhöhung einer bereits hohen Iterationszahl kaum eine sichtbare Verbesserung bringt).

Anhang B

Testentscheidung unter Verwendung statistischer Software

Die allgemeine Vorgehensweise bei der Hypothesenprüfung ist wie folgt:

- Formulierung der Null- und Alternativhypothese
- Konstruktion der Teststatistik V als Funktion der Stichprobenvariablen $V = V(X_1, \ldots, X_n)$. Die Verteilungsfunktion der Teststatistik V muss unter der Annahme, dass die Nullhypothese wahr ist, zumindest approximativ bekannt sein.
- Vorgabe eines Signifikanzniveaus α $(0 < \alpha < 1)$
- Bestimmung des Ablehnungsbereiches der Nullhypothese im Wertebereich der Teststatistik V, so dass die Wahrscheinlichkeit dafür, dass V Werte aus diesem Ablehnungsbereich annimmt, nicht größer als α ausfällt, falls die Nullhypothese wahr ist. Der Wert, der den Nichtablehnungsbereich der Nullhypothese vom Ablehnungsbereich trennt, heißt kritischer Wert und kann für das vorgegebene Signifikanzniveau α aus der Verteilungsfunktion von V bestimmt werden. Bei zweiseitigen Test erhält man zwei kritische Werte und zwei Teilsegmente des Ablehnungsbereiches der Nullhypothese.
- Ziehen einer Zufallsstichprobe vom Umfang n und Berechnung der Realisation v (Testwert) der Teststatistik V
- Testentscheidung: Die Nullhypothese wird auf dem vorgegebenen Signifikanzniveau α abgelehnt, wenn der aus der Stichprobe berechnete Wert v der Teststatistik V ein Element des Ablehnungsbereiches ist, andernfalls besteht keine Veranlassung die Nullhypothese zu verwerfen.

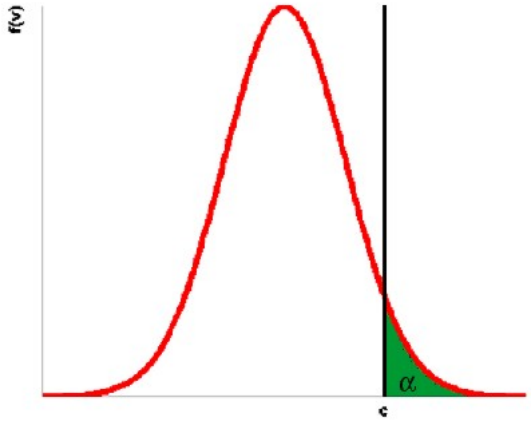
Das vorgegebene Signifikanzniveau α entspricht dabei der Wahrscheinlichkeit eines Fehlers 1. Art, d.h. der Wahrscheinlichkeit, die Nullhypothese H_0 abzulehnen, obwohl sie wahr ist.
Zur Veranschaulichung sei angenommen, dass

- ein rechtsseitiger Test für einen Parameter ϑ durchgeführt wird:
  $H_0 : \vartheta \leq \vartheta_0 \quad \text{und} \quad H_1 : \vartheta > \vartheta_0$,

- die Teststatistik V bei Gültigkeit der Nullhypothese standardnormalverteilt ist:
  $V \sim N(0; 1)$.

Der Ablehnungsbereich der H_0 wird dann durch alle Werte der Teststatistik gebildet, für die $\{v|v > c\}$ gilt. Die Wahrscheinlichkeit, eine Realisation aus dem Ablehnungsbereich der Nullhypothese H_0 zu erhalten, entspricht dem vorgegebenen Signifikanzniveau $\alpha = P(V > c|\vartheta_0)$ und ist in der folgenden Abb. B.1 durch die markierte (grüne) Fläche gekennzeichnet.

Abbildung B.1.: Signifikanzniveau α und Entscheidungsbereiche beim rechtsseitigen Test



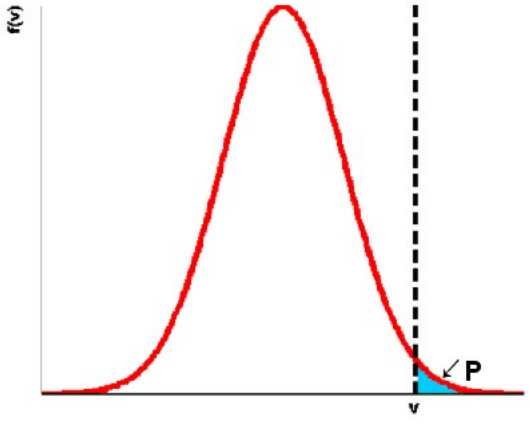
Nichtablehnungsbereich der H_0 | Ablehnungsbereich der H_0

Die Testentscheidung ist wie folgt: Ist der aus der Stichprobe berechnete Testwert v ein Element des Ablehnungsbereichs der H_0, so wird die Nullhypothese auf dem vorgegebenen Signifikanzniveau α und basierend auf der Zufallsstichprobe vom Umfang n verworfen. Andernfalls besteht keine Veranlassung, H_0 abzulehnen. Die Testentscheidung basiert somit auf einem Vergleich des Testwertes v mit den Entscheidungsbereichen.

Bei Verwendung statistischer Software (z.B. SPSS) wird ebenfalls der Testwert v auf der Grundlage der Stichprobe berechnet und im Output ausgewiesen. Zusätzlich wird die Überschreitungswahrscheinlichkeit dieses Testwertes v ausgegeben, d.h. die Wahrscheinlichkeit $P(V > v|\vartheta_0)$, dass die Teststatistik V einen Wert annimmt, der größer als dieser berechnete Testwert v ist (bei Gültigkeit der Nullhypothese H_0). Diese Überschreitungswahrscheinlichkeit wird im Output statistischer Software sehr unterschiedlich bezeichnet (z.B. als Significance, P-value, 1-tailed P bzw. 1-tailed Sig. beim einseitigen Test bzw. 2-tailed P bzw. 2 - tailed Sig beim zweiseitigen Test). Hier sei das Symbol P verwendet, so dass $P = P(V > v|\vartheta_0)$ gilt.

Abb. B.2 veranschaulicht diese Überschreitungswahrscheinlichkeit durch die markierte (himmelblaue) Fläche.

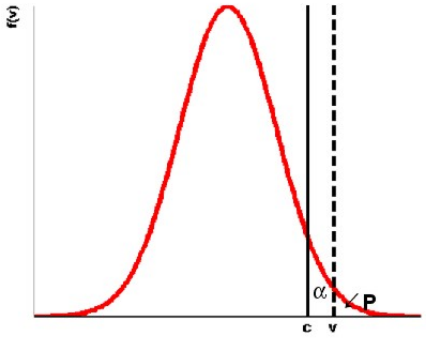
Abbildung B.2.: Überschreitungswahrscheinlichkeit $P = P(V > v|\vartheta_0)$ bei Gültigkeit der H_0



Der Nutzer der Software braucht nicht erst zu Tabellen der entsprechenden Verteilung der Teststatistik V greifen, um den bzw. die kritischen Werte und damit die Entscheidungsbereiche des Tests zu ermitteln. Im Output sind alle notwendigen Informationen für die Testentscheidung enthalten, die nunmehr auf dem Vergleich des vorgegebenen Signifikanzniveaus α und der Überschreitungswahrscheinlichkeit P beruht.
Dies sei wie folgt gezeigt.
a) Ablehnung der Nullhypothese H_0
Ergibt sich aufgrund einer konkreten Stichprobe ein Testwert v, der weit von ϑ_0 entfernt liegt, dann ist die Überschreitungswahrscheinlichkeit $P = P(V > v|\vartheta_0)$ unter der Verteilung von H_0 sehr klein. v ist ein für die Gültigkeit der Nullhypothese extremer Wert und die Nullhypothese erscheint unplausibel. Ein solcher Wert v kommt eher unter der Alternativhypothese zustande, so dass auf einen signifikanten Unterschied zwischen ϑ_0 und ϑ geschlossen wird, d.h. die Nullhypothese abgelehnt wird.
Entscheidungsregel:
Ist die im Output der Software ausgegebene Überschreitungswahrscheinlichkeit P kleiner als das vorgegebene Signifikanzniveau α ($P < \alpha$), so impliziert dies, dass der Testwert v ein Element des Ablehungsbereiches der H_0 zum vorgegebenen Signifikanzniveau α ist. Die Nullhypothese wird abgelehnt.

Bei dem hier demonstrierten rechtsseitigen Test wird diese Entscheidungsregel in der Abb. B.3 deutlich.

Abbildung B.3.: Signifikanzniveau $\alpha = P(V > c|\vartheta_0)$ und Überschreitungswahrscheinlichkeit $P = P(V > v|\vartheta_0)$ bei Gültigkeit der NullhypotheseH_0 für einen rechtsseitigen Test



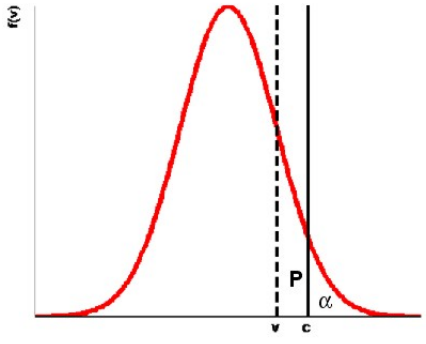
Nichtablehnungsbereich der H_0 | Ablehnungsbereich der H_0

b) Nichtablehnung der Nullhypothese
Ergibt sich aufgrund einer konkreten Stichprobe ein Testwert v, der relativ nahe bei ϑ_0 liegt, dann ist die Überschreitungswahrscheinlichkeit $P = P(V > v|\vartheta_0)$ unter der Verteilung von H_0 groß. v ist ein für die Verteilung der Nullhypothese plausibler Wert, die Abweichung zwischen v und ϑ kann als zufällig angesehen werden. Die Nullhypothese wird in diesem Fall nicht abgelehnt.
Entscheidungsregel:
Ist $P \geq \alpha$, so impliziert dies, dass der Testwert v ein Element des Nichtablehnungsbereichs der H_0 ist. Die Nullhypothese wird nicht abgelehnt.
Mit den gleichen Regeln sind die Testentscheidungen bei einem linksseitigen Test bzw. einem zweiseitigen Test zu treffen.

Wird ein linksseitiger Test für einen Parameter ϑ mit $H_0 : \vartheta \geq \vartheta_0$ und $H_1 : \vartheta < \vartheta_0$ durchgeführt, dann gilt $\alpha = P(V < c)$, wobei α vorgegeben ist und der kritische Wert c als Quantil der Ordnung α aus der Tafel der Standardnormalverteilung aufzusuchen ist. Der Ablehnungsbereich wird durch alle Werte der Teststatistik V gebildet, für die gilt $\{v|v < c\}$. Die im Output der Software ausgegebene Wahrscheinlichkeit beinhaltet nun $P = P(V < v)$.

Wird ein zweiseitiger Test für einen Parameter ϑ mit $H_0 : \vartheta = \vartheta_0$ und $H_1 : \vartheta \neq \vartheta_0$ durchgeführt, dann gilt $\alpha = P(V < -c) + P(V > c) = \alpha/2 + \alpha/2$, wobei α vorgegeben ist und der kritische Wert c als Quantil der Ordnung $1 - \alpha/2$ aus der Tafel der Standardnormalverteilung aufzusuchen ist. Der Ablehnungsbereich wird durch alle Werte der Teststatistik V gebildet, für

Abbildung B.4.: Signifikanzniveau $\alpha = P(V > c|\vartheta_0)$ und Überschreitungswahrscheinlichkeit $P = P(V > v|\vartheta_0)$ bei Gültigkeit der Nullhypothese H_0 für einen rechtsseitigen Test



Nichtablehnungsbereich der H_0 | Ablehnungsbereich der H_0

die $\{v|v < -c$ oder $v > c\}$ gilt. Die unter SPSS berechnete Wahrscheinlichkeit beinhaltet nun $P = P(V < -v) + P(V > v)$.

In beiden Fällen ist die Testentscheidung wie vorher:

a) Ablehnung der Nullhypothese H_0:

Ist $P < \alpha$, so impliziert dies, daß der berechnete Wert der Teststatistik ein Element des Ablehnungsbereiches der H_0 zum vorgegebenen Signifikanzniveau α ist. Die Nullhypothese wird abgelehnt.

b) Nichtablehnung der Nullhypothese H_0:

Ist $P \geq \alpha$, so impliziert dies, daß der berechnete Wert der Teststatistik ein Element des Nichtablehnungsbereiches der H_0 zur Wahrscheinlichkeit $1 - \alpha$ ist. Die Nullhypothese wird nicht abgelehnt.

Diese Testentscheidungen gelten entsprechend für nichtparametrische Tests.

Anhang C

Informationen zum Chi - Quadrat - Unabhängigkeitstest

Die generelle Vorgehensweise bei Unabhängigkeitstests ist im Prinzip wie bei den Parametertests. Es wird eine Teststatistik konstruiert, die die Informationen bei Gültigkeit der Nullhypothese sowie die Informationen aus der Zufallsstichprobe enthält und auf deren Basis eine Aussage über die Nullhypothese möglich ist. Die Verteilung der Teststatistik muß unter der Nullhypothese (zumindest approximativ) bekannt sein. Auch bei Unabhängigkeitstests wird stets die Nullhypothese statistisch geprüft und in Abhängigkeit von der Testentscheidung besteht die Möglichkeit einen Fehler 1. Art mit der Wahrscheinlichkeit $P(\text{„}H_1\text{“}|H_0) = \alpha$ bzw. einen Fehler 2. Art mit der Wahrscheinlichkeit $P(\text{„}H_0\text{“}|H_1) = \beta$ zu begehen, wobei „H_0“ Nichtablehnung der Nullhypothese aufgrund der Testdurchführung und „H_1“ Ablehnung der Nullhypothese aufgrund der Testdurchführung bedeuten. Mit dem vorgegebenen Signifikanzniveau α kann die Wahrscheinlichkeit eines Fehlers 1. Art niedrig gehalten werden; die Wahrscheinlichkeit eines Fehlers 2. Art ist dagegen in der Regel nicht bekannt. Man wird deshalb bestrebt sein, die Nullhypothese abzulehnen, da dann die statistische Sicherheit einer Fehlentscheidung bekannt ist.

Zur Hypothesenformulierung

Wenn die Zufallsvariablen X und Y in der Grundgesamtheit wirklich unabhängig sind, dann ist zu erwarten, dass die Tatsache im Prinzip auch in der Stichprobe zu beobachten ist. Im Prinzip bedeutet dabei, dass Abweichungen zwischen den beobachteten gemeinsamen absoluten Häufigkeiten h_{jk} und den bei Unabhängigkeit erwarteten gemeinsamen absoluten Häufigkeiten e_{jk} in der Regel immer auftreten werden. Zu entscheiden ist, ob die Abweichungen noch zufallsbedingt sind oder ob es sich um signifikante Abweichungen handelt. Da stets die Nullhypothese statistisch geprüft wird, muß die Unabhängigkeit zwischen X und Y immer als H_0 formuliert werden, um die erwarteten absoluten Häufigkeiten überhaupt ermitteln zu können. Große Abweichungen zwischen beobachteten gemeinsamen absoluten Häufigkeiten h_{jk} und den

bei Unabhängigkeit erwarteten gemeinsamen absoluten Häufigkeiten e_{jk} sprechen tendenziell gegen die Unabhängigkeit, d.h. man wird die Nullhypothese ablehnen.

Zur Teststatistik

Die Tatsache, dass die beobachteten gemeinsamen absoluten Häufigkeiten Zufallsvariablen H_{jk} sind, läßt sich wie folgt zeigen.

Aus der Grundgesamtheit wird ein Element zufällig gezogen und festgestellt, ob das Wertepaar (x_j, y_k) aufgetreten ist, d.h. ob das Ereignis $\{X = x_j\} \bigcap \{Y = y_k\}$ eingetreten ist oder nicht. Es gibt somit nur zwei mögliche Ergebnisse des Zufallsexperimentes. Die Wahrscheinlichkeit für das Eintreten des Ereignisses $\{X = x_j\} \bigcap \{Y = y_k\}$ ist p_{jk} und die Wahrscheinlichkeit für das Nichteintreten $1-p_{jk}$. Das Zufallsexperiment wird n-mal wiederholt, wobei die einzelnen Versuche unabhängig voneinander sind (da eine einfache Zufallsstichprobe vorausgesetzt wird). Damit sind die Wahrscheinlichkeiten p_{jk} konstant. Es liegt somit ein Bernoulli-Experiment vor.

Bei n-maliger Durchführung der Versuche interessiert die Gesamtzahl des Eintretens des Ereignisses $\{X = x_j\} \bigcap \{Y = y_k\}$, d.h. die absolute Häufigkeit des Wertepaares (x_j, y_k) in der Stichprobe. Diese Häufigkeit kann von Stichprobe zu Stichprobe unterschiedlich sein, so dass $H_{jk} = \{$Anzahl des Eintretens des Ereignisses $\{X = x_j\} \bigcap \{Y = y_k\}$ in einer einfachen Zufallsstichprobe vom Umfang n$\}$ eine diskrete Zufallsvariable ist, die die Werte $0, \ldots, n$ annehmen kann. Die Zufallsvariable H_{jk} ist binomialverteilt mit den Parametern n und p_{jk} : $H_{jk} \sim B(n; p_{jk})$. Der Erwartungswert von H_{jk} ist $E(H_{jk}) = n \cdot p_{jk}$.

Nur bei Gültigkeit der Nullhypothese, d.h. bei stochastischer Unabhängigkeit von X und Y, ergibt sich nach dem Multiplikationssatz für unabhängige Ereignisse, dass die gemeinsame Wahrscheinlichkeit p_{jk} (die im Erwartungswert von H_{jk} enthalten ist) das Produkt der beiden Randwahrscheinlichkeiten p_{j+} und p_{+k} ist, d.h. $p_{jk} = p_{j+} \cdot p_{+k}$. Für die bei Unabhängigkeit erwarteten gemeinsamen absoluten Häufigkeiten resultiert: $e_{jk} = n \cdot p_{jk} = n \cdot p_{j+} \cdot p_{+k}$. Dies verdeutlicht nochmals, dass die Unabhängigkeit zwischen X und Y immer als H_0 formuliert werden muß.
Diese Herleitung gilt für alle $j = 1, \ldots, J$ und $k = 1, \ldots, K$ gleichermaßen.

Da die H_{jk} Zufallsvariablen sind, ist auch die Teststatistik V (2.19) eine Zufallsvariable, die eine bestimmte Verteilung aufweist. Bei Gültigkeit der Nullhypothese, hinreichend großem Stichprobenumfang n und Einhaltung der Approximationsbedingung ist die Teststatistik V approximativ chi - quadrat - verteilt mit $DF = (J-1)(K-1)$ Freiheitsgraden.
Bestimmung der Anzahl der Freiheitsgrade: Insgesamt sind $J \cdot K$ Wahrscheinlichkeiten p_{jk} in der zweidimensionalen Verteilung der Zufallsvariablen X und Y enthalten. Ein Freiheitsgrad geht grundsätzlich verloren, weil die Wahrscheinlichkeiten untereinander nicht unabhängig sind. Wegen $\sum_j \sum_k p_{jk} = 1$ folgt, dass jede Wahrscheinlichkeit p_{jk} durch die anderen $J \cdot K - 1$ Wahr-

scheinlichkeiten bestimmt ist. $DF = j \cdot K - 1$ wäre somit die Anzahl der Freiheitsgrade, wenn sich bei Gültigkeit der Nullhypothese alle Wahrscheinlichkeiten p_{jk} aus den (bekannten) Randwahrscheinlichkeiten gemäß $p_{jk} = p_{j+} \cdot p_{+k}$ bestimmen ließen. Die Randwahrscheinlichkeiten p_{j+} und p_{+k} sind jedoch unbekannt und müssen aus der Stichprobe geschätzt werden, wodurch sich die Anzahl der Freiheitsgrade weiter verringert. Die Randverteilung von X enthält J Randwahrscheinlichkeiten p_{j+}. Wegen $\sum_j p_{j+} = 1$ sind nur $J-1$ Wahrscheinlichkeiten p_{j+} unbekannt und zu schätzen. Die Randverteilung von Y enthält K Randwahrscheinlichkeiten p_{+k}. Wegen $\sum_k p_{+k} = 1$ sind nur $K - 1$ Wahrscheinlichkeiten p_{+k} unbekannt und zu schätzen. Insgesamt sind damit $(J - 1) + (K - 1)$ Randwahrscheinlichkeiten aus der Stichprobe zu schätzen. Somit folgt für die Anzahl der Freiheitsgrade:

$$DF = J \cdot K - 1 - [(J - 1) + (K - 1)] = J \cdot K - J - K + 1 = (J - 1) \cdot (K - 1).$$

Da in der Teststatistik V die Terme $(H_{jk} - \hat{e}_{jk})^2/\hat{e}_{jk}$ nur positive Werte annehmen können, nimmt sie ebenfalls nur positive Werte an. Große Abweichungen $H_{jk} - \hat{e}_{jk}$ führen zu großen Werten von V. Somit führen nur große Werte von V zur Ablehnung der H_0, während kleine Werte von V nicht gegen die Nullhypothese sprechen. Der Chi - Quadrat - Unabhängigkeitstest ist somit ein rechtsseitiger Test.

Eine Tabelle der Chi-Quadrat-Verteilung für ausgewählte Wahrscheinlichkeiten $1 - \alpha$ und Anzahlen von Freiheitsgraden DF ist am Ende dieses Anhangs zu finden.

Testentscheidung und Interpretation

Wenn v in den Ablehnungsbereich der H_0 fällt, wird die Nullhypothese auf dem Signifikanzniveau α und basierend auf der Zufallsstichprobe vom Umfang n abgelehnt („H_1“). Es konnte statistisch gezeigt werden, dass die Zufallsvariablen X und Y nicht stochastisch unabhängig sind. Bei dieser Entscheidung besteht die Möglichkeit, einen Fehler 1. Art („H_1“$|H_0$) zu begehen, wenn in Wirklichkeit die Nullhypothese richtig ist. Die Wahrscheinlichkeit für einen Fehler 1. Art entspricht dem vorgegebenen Signifikanzniveau α.

Wenn v in den Nichtablehnungsbereich der H_0 fällt, wird die Nullhypothese basierend auf der Zufallsstichprobe vom Umfang n nicht abgelehnt („H_0“). Das Stichprobenergebnis gibt keine Veranlassung, die Unabhängigkeit der Zufallsvariablen X und Y zu verwerfen. Bei dieser Entscheidung besteht die Möglichkeit, einen Fehler zweiter Art („H_0“$|H_1$) zu begehen, wenn in Wirklichkeit die Alternativhypothese richtig ist.

Unter SPSS wird die Testentscheidung im Vergleich von vorgegebenem Signifikanzniveau α und im Output ausgegebener Überschreitungswahrscheinlichkeit für den Testwert v getroffen. Ist die im Output ausgegebene Überschreitungswahrscheinlichkeit kleiner als das vorgegebene

Signifikanzniveau α, so wird die Nullhypothese abgelehnt. Andernfalls besteht aufgrund der Stichprobe keine Veranlassung, H_0 zu verwerfen.
Achtung:
Im SPSS-Output steht als Überschrift für diese Überschreitungswahrscheinlichkeit „Asmp. Sig. (2-sided)“ und suggeriert somit, dass ein zweiseitiger Test durchgeführt wurde. Tatsächlich handelt es sich jedoch richtigerweise um eine einseitige Überschreitungswahrscheinlichkeit.

χ^2 - **Verteilung** (entnommen aus Rönz, B., Strohe, H.G. (1994), S. 70)

Quantile $\chi^2_{DF,1-\alpha}$ der Verteilungsfunktion F für die Wahrscheinlichkeit $1-\alpha$:

$F(\chi^2_{DF,1-\alpha}) = P(\chi^2 \leq \chi^2_{DF,1-\alpha}) = 1-\alpha$

DF/α	0,10	0,05	0,01	0,001	DF
1	2,71	3,841	6,635	10,827	1
2	4,61	5,991	9,210	13,815	2
3	6,25	7,815	11,345	16,268	3
4	7,78	9,488	13,277	18,465	4
5	9,24	11,070	15,086	10,517	5
6	10,6	12,592	16,812	22,457	6
7	12,0	14,067	18,475	24,322	7
8	13,4	15,507	20,090	26,125	8
9	14,7	16,919	21,666	27,877	9
10	16,0	18,307	23,209	29,588	10
11	17,3	19,675	24,725	31,264	11
12	18,5	21,026	26,217	32,909	12
13	19,8	22,362	27,688	34,528	13
14	21,1	23,685	29,141	36,123	14
15	22,3	24,996	30,578	37,697	15
16	23,5	26,296	32,000	39,252	16
17	24,8	27,587	33,409	40,790	17
18	26,0	28,869	34,805	42,312	18
19	27,2	30,144	36,191	43,820	19
20	28,4	31,410	37,566	45,315	20
21	29,6	32,671	38,932	46,797	21
22	30,8	33,924	40,289	48,268	22
23	32,0	35,172	41,638	49,797	23
24	33,2	36,415	42,980	51,179	24
25	34,4	37,652	44,314	52,620	25
26	35,6	38,885	45,642	54,052	26
27	36,7	40,113	46,963	55,476	27
28	37,9	41,337	48,278	56,893	28
29	39,1	42,557	49,588	58,302	29
30	40,3	43,773	50,892	59,703	30
40	51,8	55,8	63,7	73,4	40
50	63,2	67,5	76,2	86,7	50
60	74,4	79,1	88,4	99,6	60
70	85,5	90,5	100,4	112,3	70
80	96,6	101,9	112,3	124,8	80
90	107,6	113,1	124,1	137,2	90
100	118,5	124,3	135,8	149,4	100

Anhang D

Informationen zum Likelihood - Ratio - Test

Wie bereits im Anhang C erläutert, sind die beobachteten Zellhäufigkeiten h_{jk} $(j = 1, \ldots, J; k = 1, \ldots, K)$ Realisationen von diskreten Zufallsvariablen H_{jk}, wobei H_{jk} als {Anzahl des Eintretens des Ereignisses $\{X = x_j\} \bigcap \{Y = y_k\}$ in einer einfachen Zufallsstichprobe vom Umfang n} definiert ist. Jede dieser Zufallsvariablen H_{jk} kann die Werte 0, ...,n annehmen und ist binomialverteilt mit den Parametern n und $p_{jk} : H_{jk} \sim B(n; p_{jk})$.

Nunmehr wird jedoch die Kontingenztabelle als Ganzes betrachtet, die aus $J \cdot K$ Zufallsvariablen H_{jk} besteht. Die Wahrscheinlichkeit dafür, dass H_{11} die Realisation h_{11} und H_{12} die Realisation h_{12} ... und H_{JK} die Realisation h_{JK} annimmt, ist gegeben durch die Multinomialverteilung:

$$
\begin{aligned}
P(H_{11} = h_{11}, \ldots, H_{JK} &= h_{JK}|n; p_{11}, \ldots, p_{JK}) = f(h_{11}, \ldots, h_{JK}|n; p_{11}, \ldots, p_{JK}) \\
&= \frac{n!}{h_{11}! \cdot h_{12}! \cdot \ldots h_{JK}!} \cdot p_{11}^{h_{11}} \cdot \ldots \cdot p_{jk}^{h_{jk}} \cdot \ldots \cdot p_{JK}^{h_{JK}} \qquad \text{(D.1)} \\
&= \frac{n!}{\prod_{j=1}^{J} \prod_{k=1}^{K} h_{jk}!} \cdot \prod_{j=1}^{J} \prod_{k=1}^{K} p_{jk}^{h_{jk}}
\end{aligned}
$$

Die Multinomialverteilung[1] ergibt sich als Verallgemeinerung der Binomialverteilung, indem die Beschränkung auf zwei mögliche Realisationen bei der Durchführung des Zufallsexperiments aufgehoben wird.

Vor der Ziehung der Stichprobe gibt $f(h_{11}, \ldots, h_{JK}|n; p_{11}, \ldots, p_{JK})$ die Wahrscheinlichkeit dafür an, eine spezielle Kontingenztabelle mit den Zellhäufigkeiten $h_{11}, \ldots, h_{JK}$ bei vorgegebenem Stichprobenumfang n und festen (unbekannten) Parametern $p_{11}, \ldots, p_{JK}$ zu erhalten.

[1] Siehe u.a. Dobson, A.J. (1991), S. 126 ff.; Fahrmeir, L., Hamerle, A. (1984), S. 33 ff.; Christensen, R. (1990), S. 14 ff., 41 ff.; Santner Th.J., Duffy, D.E. (1989), S. 16 ff.

$f(h_{11},\ldots,h_{JK}|n;p_{11},\ldots,p_{JK})$ hängt sowohl von den konkreten Realisationen $h_{11},\ldots,h_{JK}$ der Zufallsvariablen $H_{11},\ldots,H_{JK}$ als auch von den unbekannten Parametern $p_{11},\ldots,p_{JK}$ ab.

Nach der Ziehung der Stichprobe liegen die Zellhäufigkeiten $h_{11},\ldots,h_{JK}$ vor. Dann hängt $P(h_{11},\ldots,h_{JK}|n;p_{11},\ldots,p_{JK})$ nur noch von den Parametern $p_{11},\ldots,p_{JK}$ ab. Um dies zu verdeutlichen schreibt man

$$\begin{aligned} L(p_{11},\ldots,p_{JK}|h_{11},\ldots,h_{JK}) &= \frac{n!}{h_{11}!\cdot h_{12}!\cdot\ldots h_{JK}!}\cdot p_{11}^{h_{11}}\cdot\ldots\cdot p_{jk}^{h_{jk}}\cdot\ldots\cdot p_{JK}^{h_{JK}} \\ &= \frac{n!}{\prod_{j=1}^{J}\prod_{k=1}^{K}h_{jk}!}\cdot\prod_{j=1}^{J}\prod_{k=1}^{K}p_{jk}^{h_{jk}} \end{aligned} \tag{D.2}$$

Die Funktion $L(p_{11},\ldots,p_{JK})$ heißt Likelihood - Funktion von $p_{11},\ldots,p_{JK}$. Für mögliche Werte $p_{11},\ldots,p_{JK}$ gibt $L(p_{11},\ldots,p_{JK})$ die Wahrscheinlichkeit für die konkret realisierte Kontingenztabelle mit $h_{11},\ldots,h_{JK}$ an.

Die Log-Likelihood-Funktion lautet [mit $p=(p_{11},\ldots,p_{JK}]$:

$$\begin{aligned} l(p) &= \ln(n!)-[\ln(h_{11}!)+\ldots+\ln(h_{JK}!)]+h_{11}\ln p_{11}+\ldots+h_{jk}\ln p_{jk}+\ldots+h_{JK}\ln p_{JK} \\ &= \ln(n!)-\sum_{j=1}^{J}\sum_{k=1}^{K}\ln(h_{jk}!)+\sum_{j=1}^{J}\sum_{k=1}^{K}h_{jk}\ln p_{jk} \end{aligned} \tag{D.3}$$

Der ML-Schätzer $\hat{p}_{jk}$ für die unbekannte Zellwahrscheinlichkeit p_{jk} ist die beobachtete Häufigkeit $f_{jk}=h_{jk}/n$, d.h. die Likelihood-Funktion bzw. die Log-Likelihood-Funktion nimmt ihr Maximum an der Stelle h_{jk}/n $(j=1,\ldots,J;k=1,\ldots,K)$ an:

$$\begin{aligned} L(\hat{p}) &= \frac{n!}{h_{11}!\cdot h_{12}!\cdot\ldots h_{JK}!}\cdot\left(\frac{h_{11}}{n}\right)^{h_{11}}\cdot\ldots\cdot\left(\frac{h_{jk}}{n}\right)^{h_{jk}}\cdot\ldots\cdot\left(\frac{h_{JK}}{n}\right)^{h_{JK}} \\ &= \frac{n!}{\prod_{j=1}^{J}\prod_{k=1}^{K}h_{jk}!}\cdot\prod_{j=1}^{J}\prod_{k=1}^{K}\left(\frac{h_{jk}}{n}\right)^{h_{jk}} \end{aligned} \tag{D.4}$$

$$\begin{aligned} l(\hat{p}) &= \ln(n!)-[\ln(h_{11}!)+\ldots+\ln(h_{JK}!)]+h_{11}\ln\left(\frac{h_{11}}{n}\right)+\ldots+h_{JK}\ln\left(\frac{h_{JK}}{n}\right) \\ &= \ln(n!)-[\ln(h_{11}!)+\ldots+\ln(h_{JK}!)]+\sum_{j=1}^{J}\sum_{k=1}^{K}h_{jk}\ln\left(\frac{h_{jk}}{n}\right) \\ &= \ln(n!)-\sum_{j=1}^{J}\sum_{k=1}^{K}\ln(h_{jk}!)+\sum_{j=1}^{J}\sum_{k=1}^{K}h_{jk}(\ln h_{jk}-\ln n) \end{aligned} \tag{D.5}$$

Bei Gültigkeit der Nullhypothese H_0 ergeben sich die Zellwahrscheinlichkeiten p_{jk} gemäß Formel (2.9) als $p_{jk} = p_{j+} \cdot p_{+k}$ und die erwarteten absoluten Häufigkeiten e_{jk} gemäß Formel (2.10) als $e_{jk} = n \cdot p_{j+} \cdot p_{+k} = n \cdot p_{jk}$. Ersetzt man die Zellwahrscheinlichkeiten p_{jk} in (D.2) bzw. (D.3) durch e_{jk}/n, so folgt für die Likelihood-Funktion und die Log-Likelihood-Funktion unter H_0:

$$\begin{aligned} L(p_0) &= \frac{n!}{h_{11}! \cdot h_{12}! \cdot \ldots h_{JK}!} \cdot \left(\frac{e_{11}}{n}\right)^{h_{11}} \cdot \ldots \cdot \left(\frac{e_{jk}}{n}\right)^{h_{jk}} \cdot \ldots \cdot \left(\frac{e_{JK}}{n}\right)^{h_{JK}} \\ &= \frac{n!}{\prod_{j=1}^{J}\prod_{k=1}^{K} h_{jk}!} \cdot \prod_{j=1}^{J}\prod_{k=1}^{K} \left(\frac{e_{jk}}{n}\right)^{h_{jk}} \end{aligned} \tag{D.6}$$

$$\begin{aligned} l(p_0) &= \ln(n!) - [\ln(h_{11}!) + \ldots + \ln(h_{JK}!)] + h_{11} \ln\left(\frac{e_{11}}{n}\right) + \ldots + h_{JK} \ln\left(\frac{e_{JK}}{n}\right) \\ &= \ln(n!) - \sum_{j=1}^{J}\sum_{k=1}^{K} \ln(h_{jk}!) + \sum_{j=1}^{J}\sum_{k=1}^{K} h_{jk}(\ln e_{jk} - \ln n) \end{aligned} \tag{D.7}$$

Da die wahren absoluten Häufigkeiten e_{jk} unbekannt sind, werden sie gemäß Formel (2.11) geschätzt, so dass die Likelihood- bzw. Log-Likelihood-Funktion bei Gültigkeit von H_0 ihr Maximum an der Stelle $\hat{e}_{jk}/n$ $(j = 1, \ldots, J; k = 1, \ldots, K)$ annimmt:

$$L(\hat{p}_0) = \frac{n!}{\prod_{j=1}^{J}\prod_{k=1}^{K} h_{jk}!} \cdot \prod_{j=1}^{J}\prod_{k=1}^{K} \left(\frac{\hat{e}_{jk}}{n}\right)^{h_{jk}} \tag{D.8}$$

$$l(\hat{p}_0) = \ln(n!) - \sum_{j=1}^{J}\sum_{k=1}^{K} \ln(h_{jk}!) + \sum_{j=1}^{J}\sum_{k=1}^{K} h_{jk}(\ln \hat{e}_{jk} - \ln n) \tag{D.9}$$

Bildet man nun das Verhältnis des Maximums der Likelihood-Funktion bei Gültigkeit von H_0 (D.8) zum Maximum der Likelihood-Funktion ohne Restriktion der H_0 (D.4), so ergibt sich das Likelihood-Ratio:

$$\frac{L(\hat{p}_0)}{L(\hat{p})} = \frac{\frac{n!}{\prod_{j=1}^{J}\prod_{k=1}^{K} h_{jk}!} \cdot \prod_{j=1}^{J}\prod_{k=1}^{K} \left(\frac{\hat{e}_{jk}}{n}\right)^{h_{jk}}}{\frac{n!}{\prod_{j=1}^{J}\prod_{k=1}^{K} h_{jk}!} \cdot \prod_{j=1}^{J}\prod_{k=1}^{K} \left(\frac{h_{jk}}{n}\right)^{h_{jk}}} = \frac{\prod_{j=1}^{J}\prod_{k=1}^{K} \left(\frac{\hat{e}_{jk}}{n}\right)^{h_{jk}}}{\prod_{j=1}^{J}\prod_{k=1}^{K} \left(\frac{h_{jk}}{n}\right)^{h_{jk}}} \tag{D.10}$$

Entsprechend ergibt sich das Log-Likelihood-Ratio als Differenz von (D.9) und (D.5):

$$\begin{aligned}
\Lambda = l(\hat{p}_0) - l(\hat{p}) &= \left[\ln(n!) - \sum_{j=1}^{J}\sum_{k=1}^{K}\ln(h_{jk}!) + \sum_{j=1}^{J}\sum_{k=1}^{K} h_{jk}(\ln \hat{e}_{jk} - \ln n)\right] \\
&\quad - \left[\ln(n!) - \sum_{j=1}^{J}\sum_{k=1}^{K}\ln(h_{jk}!) + \sum_{j=1}^{J}\sum_{k=1}^{K} h_{jk}(\ln h_{jk} - \ln n)\right] \\
&= \sum_{j=1}^{J}\sum_{k=1}^{K} h_{jk}(\ln \hat{e}_{jk} - \ln n) - \sum_{j=1}^{J}\sum_{k=1}^{K} h_{jk}(\ln h_{jk} - \ln n) \\
&= \sum_{j=1}^{J}\sum_{k=1}^{K} h_{jk}[(\ln \hat{e}_{jk} - \ln n) - (\ln h_{jk} - \ln n)] \\
&= \sum_{j=1}^{J}\sum_{k=1}^{K} h_{jk}(\ln \hat{e}_{jk} - \ln h_{jk}) \\
&= \sum_{j=1}^{J}\sum_{k=1}^{K} h_{jk}\left(\frac{\ln \hat{e}_{jk}}{\ln h_{jk}}\right)
\end{aligned} \tag{D.11}$$

Als Teststatistik des Likelihood-Ratio-Tests wird jedoch

$$\lambda = -2\sum_{j=1}^{J}\sum_{k=1}^{K} h_{jk}\left(\frac{\ln \hat{e}_{jk}}{\ln h_{jk}}\right) \tag{D.12}$$

verwendet. Der Grund für die Multiplikation mit -2 ist, dass mit der Multiplikation die Approximation $\lambda \sim \chi^2_{1-\alpha;(J-1)(K-1)}$ gültig ist, wenn H_0 wahr und der Stichprobenumfang n groß ist.

Anhang E

Informationen zu Fisher's exaktem Test

Mit Fisher's exaktem Test wird die Nullhypothese auf Unabhängigkeit der beiden dichotomen Variablen X und Y in einer 2×2 Kontingenztabelle geprüft, wenn

- der Stichprobenumfang klein ($n \leq 30$) ist,
- wenigstens eine erwartete Zellhäufigkeit kleiner als 5 ist und/oder
- die 2×2 Kontingenztabelle eine starke Asymmetrie der Zeilen- und Spaltensummen aufweist.

Tabelle E.1.: 2×2 Kontingenztabelle

Merkmal X	Merkmal Y		Randverteilung X
	y_1	y_2	
x_1	h_{11}	h_{12}	h_{1+}
x_2	h_{21}	h_{22}	h_{2+}
Randverteilung Y	h_{+1}	h_{+2}	n

In Tabelle E.1 ist die beobachtete 2×2 Kontingenztabelle angegeben. Für die weiteren Darlegungen wird davon ausgegangen, dass die kleinste der beobachteten Zellhäufigkeiten der 2×2 Kontingenztabelle h_{11} in der Zelle (1;1) ist (was durch Umstellung der Kontingenztabelle stets erreicht werden kann).

Unter der Voraussetzung fester Häufigkeiten der Randverteilungen (h_{1+}, h_{2+}, h_{+1} und h_{+2}) können alle möglichen 2×2 Kontingenztabellen aufgeschrieben werden. Jede dieser Kontingenztabellen ist durch die Häufigkeit in der Zelle (1;1) eindeutig bestimmt, denn die übrigen Zellhäufigkeiten lassen sich mit Hilfe der festen Randhäufigkeiten bestimmen. Für die beobachtete und alle anderen möglichen 2×2 Kontingenztabellen kann die Wahrscheinlichkeit ihrer Realisierung ermittelt werden. Die Berechnung der Wahrscheinlichkeiten basiert auf folgenden

Überlegungen: Einer Gesamtheit vom endlichen Umfang n, die h_{+1} Elemente mit der Eigenschaft y_1 und $n - h_{+1} = h_{+2}$ Elemente mit der Eigenschaft y_2 (d.h. Elementen, die die Eigenschaft y_1 nicht aufweisen) enthält, wird eine Zufallsstichprobe vom Umfang h_{1+} nach dem Zufallsauswahlmodell ohne Zurücklegen entnommen. Damit sind alle Randhäufigkeiten der 2×2 Kontingenztabelle eindeutig festgelegt. Des weiteren wird die Gültigkeit der Nullhypothese der Unabhängigkeit der Variablen X und Y vorausgesetzt. V sei die Anzahl der in dieser Stichprobe enthaltenen Elemente mit der Eigenschaft y_1, d.h. h_{11}. Mit diesen Festlegungen ist ein Zufallsexperiment gegeben, dass dem Modell der hypergeometrischen Verteilung zugrunde liegt. Die Wahrscheinlichkeit, dass die Zufallsvariable V eine Realisation (Zellhäufigkeit) h_{11} annimmt ergibt sich zu:

$$\begin{aligned} P(V = h_{11}) &= \frac{\binom{h_{+1}}{h_{11}} \cdot \binom{h_{+2}}{h_{1+} - h_{11}}}{\binom{n}{h_{1+}}} \\ &= \frac{\frac{h_{+1}!}{h_{11}!(h_{+1} - h_{11})!} \cdot \frac{h_{+2}!}{(h_{1+} - h_{11})!(h_{+2} - h_{1+} + h_{11})!}}{\frac{n!}{h_{1+}!(n - h_{1+})!}} \\ &= \frac{h_{+1}!h_{+2}!h_{1+}!h_{2+}!}{h_{11}!(h_{+1} - h_{11})!(h_{1+} - h_{11})!(h_{+2} - h_{1+} + h_{11})!n!} \\ &= \frac{h_{+1}!h_{+2}!h_{1+}!h_{2+}!}{h_{11}!h_{21}!h_{12}!h_{22}!n!} \end{aligned} \tag{E.1}$$

bzw.

$$P(V = h_{11}) = \frac{h_{+1}!h_{+2}!h_{1+}!h_{2+}!}{n!} \cdot \frac{1}{h_{11}!h_{21}!h_{12}!h_{22}!} \tag{E.2}$$

Fisher's exakter Test verwendet die Zufallsvariable V als Teststatistik. Als Überschreitungswahrscheinlichkeit P (Significance) wird die Wahrscheinlichkeit dafür berechnet, bei Gültigkeit der Nullhypothese die beobachtete 2×2 Kontingenztabelle und alle noch extremeren und damit unwahrscheinlicheren Kontingenztabellen (bei festen Randhäufigkeiten) zu erhalten. Diese Überschreitungswahrscheinlichkeit ergibt sich als Summe der Wahrscheinlichkeit $P(V = h_{11})$ für das Auftreten der beobachteten Kontingenztabelle und der Wahrscheinlichkeiten für das Auftreten von Kontingenztabellen, in denen die Zellhäufigkeit h_{11} durch extremere Werte ersetzt wird.

Zur Festlegung, was unter H_0 als extremere Kontingenztabelle anzusehen ist, wird die Differenz

$$D = \frac{h_{11}}{h_{1+}} - \frac{h_{21}}{h_{2+}} \tag{E.3}$$

herangezogen. Die Hypothesenformulierung als zwei- oder einseitige Nullhypothese entscheidet letztendlich darüber, was als extremer gilt.

- Bei einem zweiseitigen Test gehören dazu alle Kontingenztabellen, für die der Betrag der Differenz D einen Wert aufweist, der gleich dem für die beobachtete Kontingenztabelle oder größer ist. Dabei werden sowohl extrem kleinere Werte für h_{11} als auch extrem größere Werte für h_{11} eingeschlossen.
  Die Wahrscheinlichkeit P ergibt sich als
  $\sum P(V = h_{11})$ gemäß (E.2) für alle Kontingenztabellen, für die $|D| \geq |D_{beob.}|$ gilt.

- Bei einem einseitigen Test ist die unter H_1 festgelegte Richtung der Abweichung maßgebend.

  - Beim rechtsseitigen Test ($H_1 : p_{11} > p_{1+} \cdot p_{+1}$) werden alle Kontingenztabellen betrachtet, für die die Differenz D einen Wert aufweist, der gleich dem für die beobachtete Kontingenztabelle oder größer ist.
    Die Überschreitungswahrscheinlichkeit P ergibt sich als
    $\sum P(V = h_{11})$ gemäß (E.2) für alle Kontingenztabellen, für die $D \geq D_{beob.}$ gilt.
  - Beim linksseitigen Test ($H_1 : p_{11} < p_{1+} \cdot p_{+1}$) werden alle Kontingenztabellen als extrem betrachtet, für die die Differenz D einen Wert aufweist, der gleich dem für die beobachtete Kontingenztabelle oder kleiner ist.
    Die Überschreitungswahrscheinlichkeit P ergibt sich als
    $\sum P(V = h_{11})$ gemäß (E.2) für alle Kontingenztabellen, für die $D \leq D_{beob.}$ gilt.

Die Testentscheidung wird im Vergleich der berechneten Überschreitungswahrscheinlichkeit P und dem vorgegebenen Signifikanzniveau α getroffen. Ist $P < \alpha$, so wird die Nullhypothese abgelehnt.

Beispiel:
Die Variable X beinhalte die Zufriedenheit mit der beruflichen Situation (x_1 - zufrieden, x_2 - unzufrieden) und die Variable Y das Geschlecht (y_1 - männlich, y_2 - weiblich). Es soll auf dem 5%-Niveau die Nullhypothese geprüft werden, dass die Wahrscheinlichkeit einer zufriedenen Beurteilung der beruflichen Situation bei den Männern nicht kleiner ist als bei den Frauen, d.h. die Zufriedenheit mit der beruflichen Situation unabhängig vom Geschlecht ist. Damit ist ein linksseitiger Test durchzuführen. Die beobachtete Kontingenztabelle enthält die Tabelle E.2.

Tabelle E.2.: Beobachtete 2×2 Kontingenztabelle

Merkmal X	Merkmal Y		Randver-
	y_1	y_2	teilung X
x_1	4	9	13
x_2	8	6	14
Randverteilung Y	12	15	17

Die Wahrscheinlichkeit, bei Gültigkeit von H_0 und den gegebenen Randhäufigkeiten diese Kontingenztabelle mit $h_{11} = 4$ zu erhalten, ergibt sich nach Formel (E.2) zu:

$$P(V = h_{11}) = \frac{12!15!13!14!}{27!} \cdot \frac{1}{4!8!9!6!} = 0,123514.$$

Dies ist der Wert der Wahrscheinlichkeitsfunktion der hypergeometrischen Verteilung H(27; 12; 13) an der Stelle $h_{11} = 4$ (siehe 2. Spalte der Tabelle E.3).

Tabelle E.3.: Hypergeometrische Verteilung H(27; 12; 13)

$V = h_{11}$	Wahrscheinlichkeitsfunktion der H(27; 12; 13)	Verteilungsfunktion der H(27; 12; 13)
0	0,000005	0,000005
1	0,000272	0,000277
2	0,004491	0,004769
3	0,032937	0,037706
4	0,123514	0,161220
5	0,254085	0,415305
6	0,296433	0,711738
7	0,197622	0,909360
8	0,074108	0,983468
9	0,014971	0,998439
10	0,001497	0,999936
11	0,000063	0,999999
12	0,000001	1,000000

Für die gegebenen Randhäufigkeiten lassen sich nunmehr alle möglichen 2×2 Kontingenztabellen notieren, deren Zellhäufigkeiten in den Spalten 1 - 4 der Tabelle E.4 enthalten sind. Jede dieser Kontingenztabellen ist eindeutig durch die Zellhäufigkeit h_{11} bestimmt. Die Wahrscheinlichkeiten, bei Gültigkeit von H_0 und den gegebenen Randhäufigkeiten die Kontingenztabellen mit der jeweiligen Zellhäufigkeit h_{11} zu erhalten, entsprechen den Werten der Wahrscheinlichkeitsfunktion der H(27; 12; 13), die in der 2. Spalte der Tabelle E.3 angegeben sind. Für die

möglichen Kontingenztabellen wird die Differenz nach Formel (E.3) berechnet (Spalte 5 der Tabelle E.4).

Tabelle E.4.: Mögliche 2×2 Kontingenztabellen und Differenz gemäß (E.3)

h_{11}	h_{12}	h_{21}	h_{22}	D
0	13	12	2	-0,8571
1	12	11	3	-0,7088
2	11	10	4	-0,5604
3	10	9	5	-0,4121
4	9	8	6	-0,2637
5	8	7	7	-0,1154
6	7	6	8	0,0330
7	6	5	9	0,1813
8	5	4	10	0,3297
9	4	3	11	0,4780
10	3	2	12	0,6264
11	2	1	13	0,7747
12	1	0	14	0,9231

Wurde die Nullhypothese einseitig formuliert, so impliziert dies, dass alle Kontingenztabellen mit einer Zellhäufigkeit $h_{11} \leq 4$ zur Berechnung der Überschreitungswahrscheinlichkeit herangezogen werden müssen, denn diese sind unter H_0 als unwahrscheinliche Kontingenztabellen anzusehen. Die Überschreitungswahrscheinlichkeit P ergibt sich als Summe der Wahrscheinlichkeiten aller Kontingenztabellen, für die $D \leq D_{beob.}$ gilt:

$$P = P(V = 4) + P(V = 3) + P(V = 2) + P(V = 1) + P(V = 0) =$$
$$0,123514 + 0,032937 + 0,004491 + 0,000272 + 0,000005 = 0,16122.$$

Diese Summe der Wahrscheinlichkeiten ist gleich dem Wert der Verteilungsfunktion der H(27; 12; 13) an der Stelle $h_{11} = 4$ (siehe Spalte 3 der Tabelle E.3):

$$P = P(V \leq 4) = 0,16122.$$

Wurde die Nullhypothese zweiseitig formuliert, so impliziert dies, dass extreme Abweichungen von der beobachteten Zellhäufigkeit h_{11} nach beiden Seiten einzubeziehen sind. Zur Berechnung der Überschreitungswahrscheinlichkeit werden alle Kontingenztabellen herangezogen, für die $|D| \geq |D_{beob.}|$ gilt, denn diese sind nunmehr unter H_0 als unwahrscheinliche Kontingenztabellen anzusehen. Das sind die Kontingenztabellen mit einer Zellhäufigkeit $h_{11} \leq 4$ und $h_{11} \geq 8$:

$$\begin{aligned} P &= [P(V = 4) + P(V = 3) + P(V = 2) + P(V = 1) + P(V = 0)] \\ &\quad + [P(V = 8) + P(V = 9) + P(V = 10) + P(V = 11) + P(V = 12)] \end{aligned}$$

$$
\begin{aligned}
&= [0,123514 + 0,032937 + 0,004491 + 0,000272 + 0,000005] \\
&\quad +[0,074108 + 0,014971 + 0,001497 + 0,000063 + 0,000001] \\
&= 0,16122 + 0,09064 = 0,25186.
\end{aligned}
$$

Diese Summe der Wahrscheinlichkeiten kann auch mittels der Verteilungsfunktion der H(27; 12; 13) berechnet werden, denn es ist

$$
\begin{aligned}
P(V = 4) \quad &+ \quad P(V = 3) + P(V = 2) + P(V = 1) + P(V = 0) = P(V \leq 4) \text{ und} \\
P(V = 8) \quad &+ \quad P(V = 9) + P(V = 10) + P(V = 11) + P(V = 12) = P(V \geq 8) = 1 - P(V \leq 7),
\end{aligned}
$$

so dass sich die zweitseitige Überschreitungswahrscheinlichkeit zu

$$
P = P(V \leq 4) + [1 - P(V \leq 7)]
$$

ergibt. Aus Tabelle E.3 entnimmt man $P(V \leq 4) = 0,16122$ und $P(V \leq 7) = 0,909360$, womit

$$
P = P(V \leq 4) + 1 - P(V \leq 7) = 0,16122 + (1 - 0,909360) = 0,25186
$$

resultiert. Für dieses Beispiel wurde die Berechnung der Überschreitungswahrscheinlichkeit ausführlich demonstriert. Diese beiden Überschreitungswahrscheinlichkeiten (ein- und zweiseitig) werden im SPSS - Output auf drei Dezimalstellen gerundet ausgegeben. Der Anwender hat nun je nach der Formulierung der H_0 die entsprechende Überschreitungswahrscheinlichkeit mit dem vorgegeben Signifikanzniveau α zu vergleichen. Ist $P < \alpha$, so wird die Nullhypothese abgelehnt.

Für das Beispiel des einseitigen Tests ist $P = 0,161 > \alpha = 0,05$. Es besteht auf Basis der Stichprobe keine Veranlassung, die Nullhypothese auf Unabhängigkeit der Zufriedenheit mit der beruflichen Situation vom Geschlecht zu verwerfen.

Anhang F: t-Verteilung (entnommen aus Rönz, B., Strohe, H.G. (1994), S. 375)
Quantile t der Verteilungsfunktion F für die Wahrschinlichkeit $1 - \alpha$ und df Freiheitsgrade $F(t) = P(T \leq t) = 1 - \alpha$

df	$1-\alpha$							
	0,75	0,90	0,95	0,975	0,99	0,995	0,999	0,9995
1	1,000	3,078	6,314	12,706	31,821	63,657	318,315	636,619
2	0,816	1,886	2,920	4,303	6,965	9,925	22,327	31,598
3	0,765	1,638	2,353	3,182	4,541	5,841	10,215	12,924
4	0,741	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,727	1,476	2,015	2,571	3,365	4,032	5,894	6,869
6	0,718	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,711	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	0,706	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,703	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,700	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,697	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,695	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,694	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,692	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,691	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,690	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,689	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,688	1,330	1,734	2,101	2,552	2,878	3,611	3,922
19	0,688	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,687	1,325	1,725	2,086	2,528	2,845	3,552	3,848
21	0,686	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,686	1,321	1,171	2,074	2,508	2,819	3,505	3,792
23	0,685	1,319	1,714	2,069	2,500	2,807	3,485	3,767
24	0,685	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,684	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,684	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,684	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,683	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,683	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,683	1,310	1,697	2,042	2,457	2,759	3,385	3,646
40	0,681	1,303	1,684	2,021	2,423	2,704	3,307	3,551
60	0,679	1,296	1,671	2,000	2,390	2,660	3,232	3,460
120	0,677	1,289	1,658	1,980	2,358	2,617	3,170	3,373
∞	0,674	1,282	1,645	1,960	2,326	2,576	3,090	3,291

Anhang G

Konkordanz, Diskordanz und Ties in Kontingenztabellen

Unter Verwendung des Beispiels 2.8 soll die Bestimmung von konkordanten bzw. diskordanten Paaren sowie Ties demonstriert werden. Als Kontingenztabelle für dieses Beispiel ergab sich Tabelle 2.16.

Tabelle 2.16.: Kontingenztabelle für die Beurteilung des Studiums (X) und soziale Lage (Y)

Beurteilung des Studiums * Soziale Lage Crosstabulation

Count

			Soziale Lage			Total
			sehr gut, gut (1)	befriedigend (2)	schlecht (3)	
Beurteilung des Studiums	sehr gut, gut	(1)	17	4	3	24
	befriedigend	(2)	23	16	19	58
	schlecht	(3)	2	5	18	25
Total			42	25	40	107

Anzahl der konkordanten Paare

Für die konkordanten Paare gilt: $\{X_i < X_h; Y_i < Y_h\}$ bzw. $\{X_i > X_h; Y_i > Y_h\}$.

a) Gegeben sei ein Student i mit „sehr gut, gut (1)“ sowohl bei der Beurteilung des Studiums als auch bei der sozialen Lage, d.h. ein Student aus der Zelle (1,1) der Kontingenztabelle. Ein anderer Student h, der bei beiden Variablen einen höheren Rangwert aufweist, ist z.B. einer aus der Zelle (2,2) mit „befriedigend (2)“ sowohl bei der Beurteilung des Studiums als auch bei der sozialen Lage. Für dieses Paar von Studenten ist Konkordanz in den Rangwerten gegeben:

$$\{(X_i = 1) < (X_h = 2); (Y_i = 1) < (Y_h = 2\}.$$

Nun sind jedoch 17 Studenten in der Zelle (1,1) und 16 Studenten in der Zelle (2,2) zu finden, so dass sich insgesamt $17 \cdot 16 = 272$ Paare von Studenten ergeben, die diese konkordante Ordnungsrelation in X und Y aufweisen.

Es wird wiederum einen Student aus der Zelle (1,1) mit „sehr gut, gut (1)“ sowohl bei der Beurteilung des Studiums als auch bei der sozialen Lage gewählt. Ein anderer Student, der bei beiden Variablen einen höheren Rangwert aufweist, ist z.B. einer aus der Zelle (2,3) mit „befriedigend (2)“ bei der Beurteilung des Studiums und „schlecht (3)“ bei der sozialen Lage. Auch für dieses Paar von Studenten ist Konkordanz in den Rangwerten gegeben, da $\{(X_i = 1) < (X_h = 2); (Y_i = 1) < (Y_h = 3)\}$ ist. Nun sind jedoch 17 Studenten in der Zelle (1,1) und 19 Studenten in der Zelle (2,3) zu finden, so dass es $17 \cdot 19 = 323$ Paare von Studenten ergeben, die diese konkordante Ordnungsrelation in X und Y aufweisen.

In analoger Weise ergeben sich $17 \cdot 5 = 85$ Paare von Studenten aus Zelle (1,1) und Zelle (3,2) mit $\{(X_i = 1) < (X_h = 3); (Y_i = 1) < (Y_h = 2)\}$ und $17 \cdot 18 = 306$ Paare von Studenten aus Zelle (1,1) und Zelle (3,3) mit $\{(X_i = 1) < (X_h = 3); (Y_i = 1) < (Y_h = 3)\}$. Als Zwischensumme erhält man $17(16 + 19 + 5 + 18) = 986$ konkordante Paare. Dies läßt sich in der Kontingenztabelle wie folgt kennzeichnen.

			Soziale Lage			Total
			sehr gut, gut (1)	befriedigend (2)	schlecht (3)	
Beurteilung	sehr gut, gut	(1)	17	4	3	24
des Studiums	befriedigend	(2)	23	16	19	58
	schlecht	(3)	2	5	18	25
Total			42	25	40	107

b) Nunmehr sei ein Student mit „sehr gut, gut (1)“ bei der Beurteilung des Studiums und „befriedigend (2)“ bei der sozialen Lage, d.h. ein Student aus der Zelle (1,2), gegeben. Ein anderer Student, der bei beiden Variablen einen höheren Rangwert aufweist, ist z.B. einer mit „befriedigend (2)“ bei der Beurteilung des Studiums und „schlecht (3)“ bei der sozialen Lage, d.h. ein Sudent aus der Zelle (2,3). Für dieses Paar von Studenten ist ebenfalls Konkordanz in den Rangwerten gegeben:
$\{(X_i = 1) < (X_h = 2); (Y_i = 2) < (Y_h = 3)\}$.
Nun sind jedoch 4 Studenten in der Zelle (1,2) und 19 Studenten in der Zelle (2,3) zu finden, so dass sich insgesamt $4 \cdot 19 = 76$ Paare von Studenten ergeben, die eine konkordante

Ordnungsrelation in X und Y aufweisen.

In analoger Weise ergeben sich $4 \cdot 18 = 72$ Paare von Studenten aus Zelle (1,2) und Zelle (3,3) mit $\{(X_i = 1) < (X_h = 3); (Y_i = 2) < (Y_h = 3)\}$.
Damit resultieren zusammen $4(19 + 18) = 148$ weitere konkordante Paare. Dies läßt sich in der Kontingenztabelle wie folgt kennzeichnen.

			Soziale Lage			Total
			sehr gut, gut (1)	befriedigend (2)	schlecht (3)	
Beurteilung	sehr gut, gut	(1)	17	4	3	24
des Studiums	befriedigend	(2)	23	16	19	58
	schlecht	(3)	2	5	18	25
Total			42	25	40	107

c) Für jeden Studenten aus Zelle (2,1) mit „befriedigend (2)“ bei der Beurteilung des Studiums und „sehr gut, gut (1)“ bei der sozialen Lage, der mit einem Studenten aus Zelle (3,2) mit „schlecht (3)“ bei der Beurteilung des Studiums und „befriedigend (2)“ bei der sozialen Lage kombiniert wird, gilt $\{(X_i = 1) < (X_h = 3); (Y_i = 1) < (Y_h = 2)\}$ und es gibt somit $23 \cdot 5 = 115$ konkordante Paare.

Für jeden Studenten aus der Zelle (2,1), der mit einem Studenten aus Zelle (3,3) mit „schlecht (3)“ bei der Beurteilung des Studiums und „schlecht (3)“ bei der sozialen lage kombiniert wird, gilt $\{(X_i = 2) < (X_h = 3); (Y_i = 1) < (Y_h = 3)\}$ und somit $23 \cdot 18 = 414$ konkordante Paare.

Damit resultieren zusammen $23(5 + 18) = 529$ weitere konkordante Paare.

			Soziale Lage			Total
			sehr gut, gut (1)	befriedigend (2)	schlecht (3)	
Beurteilung	sehr gut, gut	(1)	17	4	3	24
des Studiums	befriedigend	(2)	23	16	19	58
	schlecht	(3)	2	5	18	25
Total			42	25	40	107

d) Schließlich erhält man noch konkordante Paare, wenn ein Student aus Zelle (2,2) mit „befriedigend (2)“ sowohl bei der Beurteilung des Studiums als auch bei der sozialen

Lage mit einem Studenten aus Zelle (3,3) mit „schlecht (3)“ bei der Beurteilung des Studiums als auch bei der sozialen Lage kombiniert wird, denn es gilt dann $\{(X_i = 2) < (X_h = 3); (Y_i = 2) < (Y_h = 3)\}$. Dies sind $16 \cdot 18 = 288$ konkordante Paare.

			Soziale Lage			Total
			sehr gut, gut (1)	befriedigend (2)	schlecht (3)	
Beurteilung	sehr gut, gut	(1)	17	4	3	24
des Studiums	befriedigend	(2)	23	16	19	58
	schlecht	(3)	2	5	18	25
Total			42	25	40	107

Die Gesamtzahl C konkordanter Paare ist somit:

$$C = 17(16 + 19 + 5 + 18) + 4(19 + 18) + 23(5 + 18) + 16(18) = 1951.$$

Anzahl der diskordanten Paare

Mit analogen Überlegungen läßt sich die Anzahl der diskordanten Paare ermitteln. Für diskordante Paare gilt: $\{X_i < X_h; Y_i > Y_h\}$ bzw. $\{X_i > X_h; Y_i < Y_h\}$.

a) Gegeben sei ein Student mit „sehr gut, gut (1)“ bei der Beurteilung des Studiums und „befriedigend (2)“ bei der sozialen Lage, d.h. ein Student aus der Zelle (1,2). Ein anderer Student, der bei X einen höheren und bei Y einen niedrigeren Rangwert aufweist, ist z.B. einer aus der Zelle (2,1) mit „befriedigend (2)“ bei der Beurteilung des Studiums und „sehr gut, gut (1)“ bei der sozialen Lage. Für dieses Paar von Studenten ist Diskordanz in den Rangwerten gegeben:
$\{(X_i = 1) < (X_h = 2); (Y_i = 2) > (Y_h = 1)\}$.
Es gibt $4 \cdot 23 = 92$ Paare von Studenten mit dieser diskordanten Ordnungsrelation in X und Y.

Es wird wiederum ein Student aus der Zelle (1,2) gewählt. Ein anderer Student, der bei X einen höheren und bei Y einen niedrigeren Rangwert aufweist, ist einer aus der Zelle (3,1) mit „schlecht (3)“ bei der Beurteilung des Studiums und „sehr gut, gut (1)“ bei der sozialen Lage. Auch für dieses Paar von Studenten ist Diskordanz in den Rangwerten gegeben, da $\{X_i = 1) < (X_h = 3); (Y_i = 2) > (Y_h = 1)\}$ ist. Davon gibt es $4 \cdot 2 = 8$ Paare. Als Zwischensumme erhält man $4(23 + 2) = 100$ diskordanter Paare.

			Soziale Lage			Total
			sehr gut, gut (1)	befriedigend (2)	schlecht (3)	
Beurteilung	sehr gut, gut	(1)	17	4	3	24
des Studiums	befriedigend	(2)	23	16	19	58
	schlecht	(3)	2	5	18	25
Total			42	25	40	107

b) Für die $3 \cdot 23 = 69$ Paare von Studenten aus Zelle (1,3) und Zelle (2,1), für die „sehr gut, gut (1)“ bei der Beurteilung des Studiums und „schlecht (3)“ bei der sozialen Lage bzw. „befriedigend (2)“ bei der Beurteilung des Studiums und „sehr gut, gut (1)“ bei der sozialen Lage beobachtet wurde, gilt die Relation
$\{(X_i = 1) < (X_h = 2); (Y_i = 3) > (Y_h = 1)\}$.
Für die $3 \cdot 16 = 48$ Paare von Studenten aus Zelle (1,3) und Zelle (2,2), für die „sehr gut, gut (1)“ bei der Beurteilung des Studiums und „schlecht (3)“ bei der sozialen Lage bzw. „befriedigend (2)“ bei der Beurteilung des Studiums als auch bei der sozialen Lage beobachtet wurde, gilt die Relation $\{(X_i = 1) < (X_h = 2); (Y_i = 3) > (Y_h = 2)\}$.
Für die $3 \cdot 2 = 6$ Paare von Studenten aus Zelle (1,3) und Zelle (3,1), für die „sehr gut, gut (1)“ bei der Beurteilung des Studiums und „schlecht (3)“ bei der sozialen Lage bzw. „schlecht (3)“ bei der Beurteilung des Studiums und „sehr gut, gut (1)“ bei der sozialen Lage beobachtet wurde, gilt die Relation
$\{(X_i = 1) < (X_h = 3); (Y_i = 3) > (Y_h = 1)\}$.
Für die $3 \cdot 5 = 15$ Paare von Studenten aus Zelle (1,3) und Zelle (3,2), für die „sehr gut, gut (1)“ bei der Beurteilung des Studiums und „schlecht (3)“ bei der sozialen Lage bzw. „schlecht (3)“ bei der Beurteilung des Studiums und „befriedigend (2)“ bei der sozialen Lage beobachtet wurde, gilt die Relation
$\{(X_i = 1) < (X_h = 3); (Y_i = 3) > (Y_h = 2)\}$.
Damit resultieren zusammen $3(23 + 16 + 2 + 5) = 138$ weitere diskordante Paare.

			Soziale Lage			Total
			sehr gut, gut (1)	befriedigend (2)	schlecht (3)	
Beurteilung	sehr gut, gut	(1)	17	4	3	24
des Studiums	befriedigend	(2)	23	16	19	58
	schlecht	(3)	2	5	18	25
Total			42	25	40	107

c) Für jeden Studenten aus Zelle (2,2) mit „befriedigend (2)“ bei der Beurteilung des Studiums als auch bei der sozialen Lage, der mit einem Studenten aus Zelle (3,1) mit „schlecht (3)“ bei der Beurteilung des Studiums und „sehr gut, gut (1)“ bei der sozialen Lage kombiniert wird, gilt $\{(X_i = 2) < (X_h = 3); (Y_i = 2) > (Y_h = 1)\}$, was $16 \cdot 2 = 32$ diskordante Paare ergibt. Dies läßt sich in der Kontingenztabelle wie folgt kennzeichnen:

			Soziale Lage			Total
			sehr gut, gut (1)	befriedigend (2)	schlecht (3)	
Beurteilung	sehr gut, gut	(1)	17	4	3	24
des Studiums	befriedigend	(2)	23	16	19	58
	schlecht	(3)	2	5	18	25
Total			42	25	40	107

d) Für die $19 \cdot 2 = 38$ Paare von Studenten aus Zelle (2,3) und Zelle (3,1), für die „befriedigend (2)“ bei der Beurteilung des Studiums und „schlecht (3)“ bei der sozialen Lage bzw. „schlecht (3)“ bei der Beurteilung des Studiums und „sehr gut, gut (1)“ bei der sozialen Lage beobachtet wurden, gilt die Relation
$\{(X_i = 2) < (X_h = 3); (Y_i = 3) > (Y_h = 1)\}$.
Für die $19 \cdot 5 = 95$ Paare von Studenten aus Zelle (2,3) und Zelle (3,2), für die „befriedigend“ bei der Burteilung des Studiums und „schlecht (3)“ bei der sozialen Lage bzw. „schlecht (3)“ bei der Beurteilung des Studiums und „befriedigend (2)“ bei der sozialen Lage beobachtet wurde, gilt die Relation
$\{X_i = 2) < (X_h = 3); (Y_i = 3) > (Y_h = 2)\}$.
Damit resultieren $19(2 + 5) = 133$ weitere diskordante Paare.

			Soziale Lage			Total
			sehr gut, gut (1)	befriedigend (2)	schlecht (3)	
Beurteilung	sehr gut, gut	(1)	17	4	3	24
des Studiums	befriedigend	(2)	23	16	19	58
	schlecht	(3)	2	5	18	25
Total			42	25	40	107

Die Gesamtzahl D diskordanter Paare ist somit:

$$D = 4(23 + 2) + 3(23 + 16 + 2 + 5) + 16(2) + 19(2 + 5) = 403.$$

Anzahl der Ties

Für **Ties in X** gilt $\{X_i = X_h; Y_i < Y_h\}$.

a) Gegeben sei ein Student mit „sehr gut, gut (1)“ sowohl bei der Beurteilung des Studiums als auch bei der sozialen Lage, d.h. ein Student aus der Zelle (1,1). Ein anderer Student, für den bei der Beurteilung des Studiums der gleiche Rangwert „sehr gut, gut (1)“, bei der sozialen Lage jedoch ein höherer Rangwert beobachtet wurde, muß zur Zelle (1,2) oder (1,3) gehören. Für ein solches Paar von Studenten gilt die Relation: $\{(X_i = 1) = (X_h = 1); (Y_i = 1) < (Y_h = 2)\}$ bzw. $\{(X_i = 1) = (X_h = 1); (Y_i = 1) < (Y_h = 3)\}$. Insgesamt gibt es damit $17(4 + 3) = 119$ Paare von Studenten.

			Soziale Lage			Total
			sehr gut, gut (1)	befriedigend (2)	schlecht (3)	
Beurteilung	sehr gut, gut	(1)	17	4	3	24
des Studiums	befriedigend	(2)	23	16	19	58
	schlecht	(3)	2	5	18	25
Total			42	25	40	107

Gegeben sei ein Student mit „sehr gut, gut (1)“ bei der Beurteilung des Studiums und „befriedigend (2)“ bei der sozialen Lage, d.h. ein Student aus der Zelle (1,2). Ein anderer Student, für den bei der Beurteilung des Studiums der gleiche Rangwert „sehr gut, gut

(1)“, bei der sozialen Lage jedoch ein höherer Rangwert beobachtet wurde, muß zur Zelle (1,3) gehören. Für ein solches Paar von Studenten gilt die Relation:
$\{(X_i = 1) = (X_h = 1); (Y_i = 2) < (Y_h = 3)\}$. Insgesamt gibt es davon $4 \cdot 3 = 12$ Paare von Studenten.

			Soziale Lage			Total
			sehr gut, gut (1)	befriedigend (2)	schlecht (3)	
Beurteilung	sehr gut, gut	(1)	17	4	3	24
des Studiums	befriedigend	(2)	23	16	19	58
	schlecht	(3)	2	5	18	25
Total			42	25	40	107

b) Analog ergibt sich für die zweite Zeile der Kontingenztabelle:
$23(16 + 19) = 805$ Paare von Studenten mit der Relation
$\{(X_i = 2) = (X_h = 2); (Y_i = 1) < (Y_h = 2)\}$ bzw.
$\{(X_i = 2) = (X_h = 2); (Y_i = 1) < (Y_h = 3)\}$
und $16 \cdot 19 = 304$ Paare von Studenten mit der Relation
$\{(X_i = 2) = (X_h = 2); (Y_i = 2) < (Y_h = 3)\}$.

c) Für die dritte Zeile folgt:
$2(5 + 18) = 46$ Paare von Studenten mit der Relation
$\{(X_i = 3) = (X_h = 3); (Y_i = 1) < (Y_h = 2)\}$ bzw.
$\{(X_i = 3) = (X_h = 3); (Y_i = 1) < (Y_h = 3)\}$
und $5 \cdot 18 = 90$ Paare von Studenten mit der Relation
$\{(X_i = 3) = (X_h = 3); (Y_i = 2) < (Y_h = 3)\}$.

Die Gesamtzahl der Ties in X ist:

$$T_X = 17(4 + 3) + 4(3) + 23(16 + 19) + 16(19) + 2(5 + 18) + 5(18) = 1376.$$

Für **Ties in Y** gilt $\{X_i < X_h; Y_i = Y_h\}$.

a) Gegeben sei ein Student mit „sehr gut, gut (1)“ sowohl bei der Beurteilung des Studiums als auch bei der sozialen Lage, d.h. ein Student aus der Zelle (1,1). Ein anderer Student, für den bei der sozialen Lage (Y) der gleiche Rangwert „sehr gut, gut (1)“, bei der Beurteilung des Studiums (X) jedoch ein höherer Rangwert beobachtet wurde, muß zur Zelle (2,1) oder (3,1) gehören. Für ein solches Paar von Studenten gilt die Relation:

$\{(X_i = 1) < (X_h = 2); (Y_i = 1) = (Y_h = 1)\}$ bzw.
$\{(X_i = 1) < (X_h = 3); (Y_i = 1) = (Y_h = 1)\}$. Insgesamt gibt es damit $17(23 + 2) = 425$ Paare von Studenten.

			Soziale Lage			Total
			sehr gut, gut (1)	befriedigend (2)	schlecht (3)	
Beurteilung	sehr gut, gut	(1)	17	4	3	24
des Studiums	befriedigend	(2)	23	16	19	58
	schlecht	(3)	2	5	18	25
Total			42	25	40	107

Gegeben sei ein Student mit „befriedigend (2)“ bei der Beurteilung des Studiums und „sehr gut, gut (1)“ bei der sozialen Lage, d.h. ein Student aus der Zelle (2,1). Ein anderer Student, für den bei der sozialen Lage (Y) der gleiche Rangwert „sehr gut, gut (1)“, bei der Beurteilung des Studiums (X) jedoch ein höherer Rangwert beobachtet wurde, muß zur Zelle (3,1) gehören. Für ein solches Paar von Studenten gilt die Relation: $\{(X_i = 2) < (X_h = 3); (Y_i = 1) = (Y_h = 1)\}$ Insgesamt gibt es damit $23 \cdot 2 = 46$ Paare von Studenten.

			Soziale Lage			Total
			sehr gut, gut (1)	befriedigend (2)	schlecht (3)	
Beurteilung	sehr gut, gut	(1)	17	4	3	24
des Studiums	befriedigend	(2)	23	16	19	58
	schlecht	(3)	2	5	18	25
Total			42	25	40	107

b) Analog ergibt sich für die zweite Spalte der Kontingenztabelle:
$4(16 + 5) = 84$ Paare von Studenten mit der Relation
$\{(X_i = 1) < (X_h = 2); (Y_i = 2) = (Y_h = 2)\}$ bzw.
$\{(X_i = 1) < (X_h = 3); (Y_i = 2) = (Y_h = 2)\}$
und $16 \cdot 5 = 80$ Paare von Studenten mit der Relation
$\{(X_i = 2) < (X_h = 3); (Y_i = 2) = (Y_h = 2)\}$.

c) Für die dritte Spalte folgt:
$3(19 + 18) = 111$ Paare von Studenten mit der Relation
$\{(X_i = 1) < (X_h = 2); (Y_i = 3) = (Y_h = 3)\}$ bzw.
$\{(X_i = 1) < (X_h = 3); (Y_i = 3) = (Y_h = 3)\}$
und $19 \cdot 18 = 342$ Paare von Studenten mit der Relation
$\{(X_i = 2) < (X_h = 3); (Y_i = 3) = (Y_h = 3)\}$.

Die Gesamtzahl der Ties in Y ist:

$$T_Y = 17(23 + 2) + 23(2) + 4(16 + 5) + 16(5) + 3(19 + 18) + 19(18) = 1088$$

Für **Ties in X und Y** gilt: $\{X_i = X_h; Y_i = Y_h\}$.
Bei diesen Ties müssen beide Studenten zur gleichen Zelle der Kontingenztabelle gehören, da nur in diesem Fall bei beiden Variablen die gleichen Rangwerte auftreten. Für die 17 Studenten in der Zelle (1,1) lassen sich $17 \cdot 16/2 = 136$ Paare bilden.
Allgemein kann somit für eine Zelle geschrieben werden:

$$\binom{h_{jk}}{2} = \frac{h_{jk}!}{2!(h_{jk} - 2)!} = \frac{h_{jk}(h_{jk} - 1)}{2}.$$

Für die Gesamtzahl der Ties in X und Y resultiert:

$$T_{XY} = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{h_{jk}(h_{jk} - 1)}{2}.$$

Anhang H

Eigenwerte

Jede quadratische Matrix $\mathbf{A}$ besitzt (neben einer Determinanten und eventuell einer Inversen) eine sogenannte charakteristische Gleichung:

$$|\mathbf{A} - \lambda \cdot \mathbf{I}| \cdot \mathbf{x} = \mathbf{0} \qquad \mathbf{I} - \text{Einheitsmatrix} \tag{H.1}$$

$$\begin{pmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{H.2}$$

Die Matrix $(\mathbf{A} - \lambda \cdot \mathbf{I})$ heißt charakteristische Matrix, die zugehörige Determinante charakteristische Determinante. Die Lösungen für λ gilt es derart zu bestimmen, dass die charakteristische Determinante Null wird. Mann nennt diese Lösungen Eigenwerte von $\mathbf{A}$:

$$|\mathbf{A} - \lambda \cdot \mathbf{I}| = \mathbf{0} \tag{H.3}$$

$$\begin{vmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{vmatrix} = 0 \tag{H.4}$$

Jeder zu $\lambda = \lambda_i$ $(i = 1, \dots, n)$ gehörige Lösungsvektor $\mathbf{x} = \mathbf{x}_i$ heißt Eigenvektor der Matrix $\mathbf{A}$ zum Eigenwert λ_i.

Beispiel:

$$A = \begin{pmatrix} 2 & 2 & 1 \\ 1 & 3 & 1 \\ 1 & 2 & 2 \end{pmatrix}$$

Die charakteristische Gleichung lautet:

$$\begin{pmatrix} 2-\lambda & 2 & 1 \\ 1 & 3-\lambda & 1 \\ 1 & 2 & 2-\lambda \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Die charakteristische Determinante ist:

$$\begin{vmatrix} 2-\lambda & 2 & 1 \\ 1 & 3-\lambda & 1 \\ 1 & 2 & 2-\lambda \end{vmatrix} = -\lambda^3 + 7\lambda^2 - 11\lambda + 5 = 0$$

Dies liefert die Eigenwerte $\lambda_1 = 5,\ \lambda_2 = 1,\ \lambda_3 = 1$.
Die Eigenvektoren zu $\mathbf{A}$ ermittelt man durch Einsetzen der λ-Werte in die charakteristische Gleichung:
Für $\lambda_1 = 5$:

$$\begin{pmatrix} -3 & 2 & 1 \\ 1 & -2 & 1 \\ 1 & 2 & -3 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Um den zu $\lambda_1 = 5$ gehörigen Eigenvektor zu bestimmen, ist dieses Gleichungssystem nach $\mathbf{x}$ aufzulösen.

Literaturverzeichnis

[1] Agresti, A. (1990), Categorial Data Analysis, John Wiley & Sons, New York

[2] Agresti, A. (1996), An Introduction to Categorial Data Analysis, John Wiley & Sons, New York et al.

[3] Ahrens, H., Läuter, J. (1974), Mehrdimensionale Varianzanalyse, Akedemie-Verlag, Berlin

[4] Backhaus, K., Erichson, B., Plinke, W., Weiber, R. (1994), Multivariate Analysemethoden, Springer, Berlin et al.

[5] Bamberg, G., Baur, F. (1991), Statistik, Oldenbourg, München, Wien

[6] Berry, D.A., Lindgren, B.W. (1990), Statistics: Theory and Methods, Brooks/Cole Publishing Company, Pacific Grove, California

[7] Bortz, J. (1993) Statistik, Springer, Berlin et al.

[8] Bortz, J., Lienert, G.A., Boehnke, K. (1990), Verteilungsfreie Methoden in der Biostatistik, Springer, Berlin et al.

[9] Bosch, K. (1992), Statistik-Taschenbuch, Oldenbourg, München, Wien

[10] Bühl, A., Zöfel, P. (1994), SPSS für Windows Version 6, Addison-Wesley

[11] Büning, H., Trenkler, G. (1978), Nichtparametrische statistische Methoden, Walter de Gruyter, Berlin, New York

[12] Carmines, E.G., Zeller, R.A. (1980), Reliability and Validity Assessment, Sage Publications, Beverly Hills

[13] Chambers, J.M., Cleveland, W.S., Kleiner, B., Tukey, P.A. (1983), Graphical Methods for Data Analysis, Chapman & Hall, New York, London

[14] Christensen, R. (1990), Log-Linear Models, Springer, New York et al.

[15] Clauß, G., Finze, F.-R., Partzsch, L. (1994), Statistik für Soziologen, Pädagogen, Psychologen und Mediziner, Verlag Harri Deutsch, Thun und Frankfurt am Main

[16] Cleveland, W.S. (1979), Robust locally weighted regression and smoothing scatterplots, Journal of the American Statistical Association, 74, S. 829-836

[17] Cleveland, W.S. (1985), The elements of graphing data, Wadsworth & Brooks/Cole, Advanced Books & Software, Pacific Grove, California

[18] Daniel, W.W. (1990), Applied Nonparametric Statistics, 2nd ed., The Daxbury Advanced Series in Statistics and Decision Sciences, PWS-Kent Publishing Company, Boston

[19] De Gruitjer, P. N.M., van der Kamp, L.J.T. (Eds.) (1976), Advances in psychological and educational measurement, John Wiley, New York

[20] Dobson, A.J. (1991), An Introduction to Generalized Linear Models, Chapman & Hall, London et al.

[21] Eckstein, P.P. (1997), Angewnadte Statistik mit SPSS, Betriebswirtschaftlicher Verlag Dr. Th. Gabler GmbH, Wiesbaden

[22] Fahrmeir, L., Hamerle, A. (1984), Multivariate statistische Verfahren, Walter de Gruyter, Berlin, New York

[23] Fahrmeir, L., Künstler, R., Pigeot, I., Tutz, G. (1997), Statistik, Springer Verlag, Berlin et al.

[24] Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. Ostrowski, E. (1994), A Handbook of Small Data Sets, Chapman & Hall, London et al.

[25] Hartung, J., Elpelt, B. (1989), Multivariate Statistik, 3. Auflage, Oldenbourg, München, Wien

[26] Hartung, J., Elpelt, B., Klösener, K.-H. (1993), Statistik, 9. Auflage, Oldenbourg, München, Wien

[27] Hosmer, D.W., Lemeshow, S. (1989), Applied Logistic Regression, John Wiley & Sons, New York et al.

[28] Judge, G.G. et al. (1988), Introduction to the Theory and Practice of Econometrics, John Wiley & Sons, New York et al.

[29] Kähler, W.-M. (1994), SPSS für Windows, Vieweg Wiesbaden

[30] Kleinbaum, D.G. (1994), Logistic Regression, Springer, New York et al.

[31] Kline, P. (1979), Psychometrics and psychology, Academic Press, London

[32] Kline, P. (1986), A handbook of test construction, Mathuen, New York

[33] Lienert, G.A. (1973), Verteilungsfreie Methoden in der Biostatistik, Band I, Verlag Anton Hain, Meisenheim am Glan

[34] Lindgren, B.W. (1993), Statistical Theory, 4th ed., Chapman & Hall, New York et al.

[35] Litz, H.P. (1997), Statistische Methoden in den Wirtschafts- und Sozialwissenschaften, Oldenbourg Verlag, München, Wien

[36] McGraw, K.O., Wong, S.P. (1996), Forming inferences about some intraclass correlation coefficients, Psychological Methods, 1:1, S. 30 - 46

[37] Nunally, J.C. (1970), Introduction to Psychological Measurement, McGraw-Hill, New York

[38] Rönz, B. (2001) Computergestützte Statistik I, Skript, Humboldt-Universität zu Berlin

[39] Rönz, B., Förster, E. (1992), Regressions- und Korrelationsanalyse, Gabler-Verlag, Wiesbaden

[40] Rönz, B., Strohe, H.G. (Hrsg.) (1994), Lexikon Statistik, Gabler-Verlag, Wiesbaden

[41] Sachs, L. (1992), Angewandte Statistik, Springer Verlag, Berlin et al.

[42] Santner, Th.J., Duffy, D.E. (1989), The Statistical Analysis of Discrete Data, Springer, New York et al.

[43] Schlittgen, R. (1990), Einführung in die Statistik, Oldenbourg, München, Wien

[44] Schwarze, J. (1990), Grundlagen der Statistik II, Verlag Neue Wirtschafts-Briefe, Herne, Berlin

[45] SPSS for Windows: Base System User's Guide, Release 6.0, SPSS Inc., 1993

[46] SPSS Base System Syntax Reference Guide 6.0, SPSS Inc.

[47] SPSS Statistical Alogrithms 2nd Ed., SPSS Inc.

[48] SPSS for Windows: Professional Statistics 6.0, SPSS Inc.

[49] SPSS for Windows: Advanced Statistics 6.0, SPSS Inc.

[50] SPSS Base 9.0, Applications Guide, SPSS Inc., 1999

[51] SPSS Base 9.0, User's Guide, SPSS Inc., 1999

[52] SPSS Advanced Models 9.0, SPSS Inc., 1999

[53] SPSS Regression Models 9.0, SPSS Inc., 1999

[54] SPSS Interactive Graphics 9.0, SPSS Inc., 1999

[55] SPSS Base 9.0, Syntax Reference Guide, SPSS Inc., 1999

[56] STATISTICA^{TM}, Volume III, StatSoft 1994

[57] Thorndyke, R.L., Hagen, E.P. (1977), Measurement and evaluation in psychology and education, John Wiley, New York

[58] Überla, K. (1968), Faktorenanalyse, Springer, Berlin, Heidelberg, New York

[59] Winer, B.J. (1971), Statistical Principles in Experimental Design, New York

[60] Wittenberg, R. (1991), Computergestützte Datenanalyse, Gustav Fischer Verlag Stuttgart, UTB 1603

[61] Wittenberg, R., Cramer, H. (1992), Datenanalyse mit SPSS, Gustav Fischer Verlag Stuttgart, UTB 1602

Internetadressen

http://www.spss.com
http://www.xplore-stat.de/index_js.html
http://www.splus.mathsoft.com
http://www.f3.fhtw-berlin.de/Professoren/Eckstein/download.html
http://www.statsoft.com
http://www.ise.wiwi.hu-berlin.de/statistik/glm_d.html

Stichwortverzeichnis

10items.sav 164
3D-Balkendiagramm 25, 28
3D-Scatterplot 4, 8, 19
3D-Space-Plot 22, 23
3D-Spektral-Plot 22, 23
3D-Surface-Plot 22, 23
Abbruchkriterium 194
abhängige Stichprobe 86, 90
Ablehnungsbereich 217 ff
allbus.sav 27, 75
Alpha 186, 180
Anfangskommunalität 196
ANOVA-Tabelle 135, 152, 172, 175, 176
Anti-Image-Matrix 197
Approximationsbedingung 44, 224
Assoziationsmaße 53, 62
asymmetrische Maße 56, 71, 74
ausgeschlossene Residuen 115, 128
Ausreißer 13, 142, 150
Ausschluß 119
Autokorrelation 113, 138, 142

Backward 118
Balkendiagramm 22, 25 ff
Bartlett-Test 197, 201
bedingte Häufigkeitsverteilung 31, 37, 42

bedingte Verteilung 26, 38, 39, 41
Beispiel 2.1 11, 14, 16
Beispiel 2.2 16
Beispiel 2.3 18
Beispiel 2.4 19
Beispiel 2.5 24
Beispiel 2.6 26, 35, 48, 54, 59, 64, 66
Beispiel 2.7 49
Beispiel 2.8 69, 71, 73, 241
Beispiel 2.9 75
Beispiel 2.10 78
Beispiel 2.11 81
Beispiel 2.12 84
Beispiel 2.13 101
Beispiel 2.14 96
Beispiel 3.1 129, 154
Beispiel 3.2 157
Beispiel 4.1 164, 166, 167, 168, 170, 175, 177, 178, 179, 181, 182, 201
Bestimmtheitsmaß 11, 107, 108, 121, 123, 135, 141, 142, 156, 168, 169, 193
Bindung 67
Binomialtest 90
Binomialverteilung 91, 93, 96, 229
bivariat 3
bivariate Analyse 3
Boxplot 149
Bravais-Pearson-Korrelationskoeffizient 46, 51, 53, 74, 171, 189, 186

Chi-Quadrat-Unabhängigkeitstest 4, 42 ff, 223 ff
chi-quadrat-verteilt 43ff, 88, 96, 100, 175, 224
Chi-Quadrat-Verteilung (χ^2-Verteilung) 44, 45, 48, 51, 88, 99, 225, 227
Clustered Bar Chart 25
Cochrans-Chi-Quadrat-Test 175
Cochran's Test 94 ff
Component Score Coefficient Matrix 206
Component Transformation Matrix 206
computergestützte Statistik 1

Condition Index 124
Cook's Distanz 116, 127, 150
Cramer's V 54
Cronbach's Alpha 4, 180, 185

deskriptive Statistiken 3
DFBETAS 116, 128, 143
DFFIT 116, 128, 143
dichotom 86, 90, 94
diskordant 67, 68, 70, 241 ff
Diskordanz 68, 241 ff
diskret 6, 25, 224
Durbin-Watson-Test 113, 124, 137, 142

Eigenvektor 193
Eigenwert 124, 193, 194, 196, 199, 201, 202, 151 f
einfache Zufallsstichprobe 43
einfacher Scatterplot 9
Einschluß 118
Einzelrestfaktor 188 ff
Enter 118
Entropie 65
erwartete Zellhäufigkeit 30, 46, 90
Eta 74 ff
europa.sav 11, 18, 24
explorative Datenanalyse 4
explorative Zusammenhangsanalyse 8

F-Test 120
F-Teststatistik 175
F-verteilt 177
F-Verteilung 108, 175
Faktor 5, 186 ff
Faktorenanalyse 4, 186 ff
Faktorladung 187 ff
Faktorladungsmatrix 187, 199, 206

Faktorvariable 80, 94
Faktorwerte 187, 195, 200
Faktorwertematrix 187
Fall-Kontroll-Studie 81, 83, 85
Fisher's exakter Test 4, 46, 233 ff
Forward 118
Friedmann-Chi-Quadrat-Test 175
Fundamentaltheorem 190

Gamma 68, 70
Goodman und Kruskal's Tau 62
Gruppenvariable 10, 16, 18, 25, 27
gruppiertes Balkendiagramm 25
Guttman Split-half 184

Hat-Matrix 114
Hauptachsenmethode 192, 198, 201, 203
Hebelwert 114
hke.sav 84
Homogenität 4, 162
Homogenitätsanalyse 4, 186, 207
Hotellings T^2 176
hypergeometrisch 95, 236

icecream.sav 129
induktive Statistik 3
intervallskaliert 74
Intraclass Correlation Coefficient 178
Inzidenzrate 80, 82
Item 161 ff, 165, 201

Kaiser-Kriterium 194, 199, 204
Kaiser-Meyer-Olkin 197, 201
Kappa-Koeffizient 77 ff
Kendall's Konkordanz Koeffizient 175
Kendall's Tau-Werte 72 f

Kohorten-Studie 80, 81
Kolmogorov-Smirnov-Test 146
Kombination 68
Kommunalität 191 ff
Konfidenzintervall 11, 83, 84, 111, 112, 123, 127, 136, 154, 179
Konfidenzniveau 11, 111, 112, 123, 127
konkordant 67, 68, 70, 241 ff
Konkordanz 68, 241 ff
Konstrukt 4, 161 ff, 186, 207
Kontingenzkoeffizient 4, 53, 54
Kontingenztabelle 29 ff, 46 f, 49, 54, 57, 59, 65, 69,71, 72, 77, 79, 80, 82, 85, 87, 90, 91, 95, 97, 229, 233 ff, 241 ff
Kontrollvariable 33, 39, 94
Korrelationskoeffizient 52, 121, 123, 130, 141, 171, 179, 183, 191, 196
Korrelationsmatrix 131, 171, 190, 191, 193, 196, 198, 201
korrigiertes Bestimmtheitsmaß 108, 123, 135
korrigiertes standardisiertes Residuum 32
Kovarianz 52, 109, 111, 130, 169, 184
Kovarianzmatrix 127, 128, 199
Kovarianzverhältnis 128
Kreuztabelle 29, 32, 34
kubische Regression 11, 15
Kurvenanpassung 10ff, 14, 16, 18, 24, 151 ff, 209

Lambda 57 ff
Leverage 114, 150
Likelihood-Funktion 230 f
Likelihood-Quotienten-Test (likelihood-ratio-Test) 4, 46, 229 ff
Line-Plot 146, 147, 149, 157, 159
Linear-by-Linear Association 46
lineare Regression 11, 103 ff
Loading Plot 200, 202
Log-Likelihood-Funktion 230 ff
Lowess 12, 14, 209

Mahalanobis 127, 150

Mantel-Haenszel-Test 4, 46, 94
McNemar Test 86 ff
measure of sampling adequacy 197
Methode der kleinsten Quadrate 11, 12, 105
metrisch skaliert 6, 8, 25, 29, 46, 75, 186, 189
mieten.sav 16
ML-Schätzer 230
Modelldiagnose 112 ff, 143 ff, 153
Modelldiagnostik 4, 126
Modellgüte 122, 134
Multikollinearität 104, 124, 129, 130, 131, 133, 137
Multinomialverteilung 229
multipler Korrelationskoeffizient 123, 168
multivariate Analyse 3

Nachbarschaftsgewicht 13, 210 ff, 216
Nichtablehnungsbereich 217 ff
nominalskaliert 6, 25, 29, 53, 54, 56, 74
Normalgleichungen 106
normalverteilt 52, 105, 186, 189, 198, 214
Normalverteilung 53, 111, 113, 115, 126, 144, 146
Nullhypothese 43, 45, 48, 62, 65, 80, 83 ff, 88, 92, 95 f, 97, 108, 113, 146, 175 f, 198, 201, 217 ff, 231, 233, 235
numerische Variable 7

odds 81, 83 ff
odds ratio 81, 84 ff, 96, 99
ordinalskaliert 6, 25, 29, 53, 67, 69, 71, 72
Overlay Scatterplot 24

P-P-Plot 113, 126, 144, 145
Parallelstichprobe 86
Part-Korrelationskoeffizient 123, 142
partieller Korrelationskoeffizient 197
percept_91_96.sav 97

PHI 55
Point Selection Mode 15
PRE-Maß 56 ff, 69, 74, 75
Projektion 19
Projektionsmatrix 114
Punktidentifizierung 16, 19

Q-Q-Plot 112
quadratische Regression 11

Randverteilung 29, 30, 34, 35, 39, 96, 233
Randwahrscheinlichkeit 31
Rangzahl 7, 67
rauchen.sav 81
Regressionsanalyse 4, 103, 188
Regressionskoeffizient 104, 110, 112, 126, 151
Regressionskonstante 104, 120, 134, 151, 152
Regressionsmodell 103 ff, 109, 110
Regreßwert 13, 106, 111, 112, 114, 124, 125, 152, 212, 213
relatives Risiko 80 ff, 82
Reliabilität 4, 162
Reliabilitätsanalyse 4, 161 ff, 207
Reliabilitätskoeffizient 4, 168, 180, 182, 184
Remove 119
reproduzierte Korrelationsmatrix 192, 196, 197, 202, 204
Residualmatrix 194
Residuen 13, 31, 35, 42,43, 111 ff, 114, 124 f, 128, 138, 146, 152, 174, 214, 215
Response-Variable 94
Robustheitsgewicht 13, 214 ff
Rotation 20, 194, 199, 200, 202
Rückwärts 118

Scatterplot 4, 8, 11, 14, 16, 113, 146, 155
Scatterplot-Matrix 4, 8, 17, 18
Scheinvariable 103

Schrittweise	119, 150
schwierigkeiten.sav	49
Screeplot	199, 201, 204
SDBETAS	116, 150
SDFITS	116, 150
Selektionsmethoden	118 f
Signifikanzniveau	42, 44, 47 f, 83 f, 97, 120, 129, 132, 135, 139, 175 f, 196, 217 f
Skala	161, 166
Skalenniveau	66 ff, 43, 186
Somer's D	71, 72
Spearman-Brown-Koeffizient	184
Spearman'sche Rangkorrelationskoeffizient	51, 53, 67
Split-Half Modell	4, 182
Standardabweichung	52, 110, 114, 124, 130, 132, 146, 165, 166, 196, 214
Standardfehler	32, 89, 110, 112, 123, 126, 135, 136
standardisierte Regressionskoeffizienten	110, 122, 135
standardisierte Residuen	32, 35, 37, 114, 124, 125, 128, 144, 149
Standardnormalverteilung	52
STATISTICA	22, 28, 163
statistischer Test	7, 42
Stepwise	119, 139
stetige Variable	6, 43
Stichprobenregressionsmodell	104
Stichprobenumfang	30, 42, 43, 54, 63, 87, 89, 186, 189, 224, 229, 232
Streuungsdiagramm	8
studentisierte Residuen	115, 126, 128
studium.sav	69
symmetrische Maße	56, 69, 71, 72
symmetrsche Variable	4, 162 ff, 186, 201, 207
t-Test	4, 80 ,135
t-Verteilung	52, 111, 154, 239
telefon.sav	157
Test auf Unabhängigkeit	42
Testfunktion	108, 111, 113
Teststatistik	43 ff, 52, 65, 88, 96, 122, 175, 177, 217 ff, 232, 234

Ties 67, 68, 71, 72
Toleranz 124
Tukey's Additivitätstest 177

überlagerter Scatterplot 24
Unabhängigkeit 30, 31, 34, 37 ff, 42, 48, 53, 54, 94, 99
univariate Analyse 3
Unsicherheitskoeffizient 65 ff
unstandardisiertes Residuum 32, 35, 38, 39, 128, 144

Varianz-Inflation-Faktor 124
Varianz-Kovarianz-Matrix 109, 111, 112, 123, 128, 136
Varianzanalyse 5, 123, 135, 156, 163, 172, 178
Varianzanteile 124
Varianzprozentanteil 194
Varimax-Methode 202
Varimaxrotation 194, 205
verkehr.sav 19
Vorwärts 118

Wahrscheinlichkeitsplot 126

Yateskorrektur 45, 51

Zufallsexperiment 43, 224
Zufallsstichprobe 7,217, 224
Zusammenhangsmaße 51, 67