Alexandra Nonnenmacher
Ina Berninger
Hanna Hoffmann

Einführung in die Datenanalyse mit SPSS

Fassung 1.7, Februar 2012

Alexandra Nonnenmacher
Leibniz Universität Hannover
Institut für Politische Wissenschaft
Schneiderberg 50
30167 Hannover

a.nonnenmacher@ipw.uni-hannover.de

Ina Berninger
Universität zu Köln
Forschungsinstitut für Soziologie
Greinstr. 2
50939 Köln

berninger@wiso.uni-koeln.de

Hanna Hoffmann
Heinrich-Heine-Universität Düsseldorf
Institut für Sozialwissenschaften
Universitätsstr. 1
40225 Düsseldorf

hanna.hoffmann@uni-duesseldorf.de

Fassung 1.0, Dezember 2005 (Daten: ALLBUS 2000)
Fassung 1.1, März 2006 (Daten: ALLBUS 2000)
Fassung 1.2, März 2009 (Daten: ALLBUS 2000, kumulierter ALLBUS 2006)
Fassung 1.3, September 2009 (Daten: ALLBUS 2000, kumulierter ALLBUS 2006)
Fassung 1.4, März 2010 (Daten: ALLBUS 2000, kumulierter ALLBUS 2006)
Fassung 1.5, März 2011 (Daten: ALLBUS 2000, kumulierter ALLBUS 2006)
Fassung 1.6, November 2011 (Daten: ALLBUS 2000, kumulierter ALLBUS 2006)
Fassung 1.7, Februar 2012 (Daten: ALLBUS 2000, kumulierter ALLBUS 2006)

Inhalt

1 Allgemeine Hinweise

1.1 Der Aufbau von SPSS

1.1.1 Datenfenster

- Datenansicht: Beinhaltet in der Regel die Variablen in den Spalten, die Fälle in den Zeilen.
- Variablen: Merkmale der Fälle, z.B. Alter, Geschlecht, Schulabschluss.
- Fälle: In den meisten Datensätzen Personen; jede andere Einheit ist aber auch möglich, z.B. Länder, Stadtteile, Geburtskohorten etc.
- Bei den meisten Variablen handelt es sich um numerische Variablen, d.h. jeder Ausprägung ist ein Wert zugeordnet (z.B. v3 „Erhebungsgebiet: West – Ost“: 1 = Alte Bundesländer, 2 = Neue Bundesländer).
- Wechseln zwischen Anzeige der numerischen Werte und Anzeige der Wertelabels: 4. Icon von rechts.

- Variablenansicht: Variablen in den Zeilen.
- 1. Spalte: Variablenname. Darf in den SPSS-Versionen unter Version 12 nicht mehr als 8 Zeichen umfassen. Das erste Zeichen muss ein Buchstabe sein. Erlaubt sind Zahlen und Buchstaben. sowie die Sonderzeichen _, Punkt, @ und #. Leerzeichen und andere Zeichen sind nicht erlaubt. Das letzte Zeichen darf kein Punkt sein. Groß- und Kleinschreibung sind nicht relevant. Die Schlüsselworte all, and, by, eq, ge, gt, le, lt, ne, not, or, to und with sowie Kommandonamen können nicht als Namen vergeben werden.
- 2. Spalte: Variablentyp. Standardeinstellung: numerisch. Weitere Interessante Einstellungen sind Date oder String.
- 3. Spalte: Variablenbreite. Höchste Anzahl der Stellen, die die Werte in der Variablen annehmen können. Diese Einstellung ist nur mäßig relevant, weil auch Werte in eine Variable geschrieben werden können, die mehr Stellen haben als angegeben. Die Einstellung ist nur wichtig für die Angabe der Wertelabels.
- 4. Spalte: Dezimalstellen. Anzahl der dargestellten (nicht von SPSS gespeicherten!) Dezimalstellen.
- 5. Spalte: Variablenlabel. Genauere Bezeichnung des Inhalts der Variablen, als mit dem Variablennamen möglich ist.

- 6. Spalte: Wertelabels. Gibt an, welcher numerische Wert in der Variablen welche Bedeutung hat. v3: Dem Wert 1 wurde das Label „Alte Bundesländer“ zugeordnet, dem Wert 2 das Label „Neue Bundesländer“.
- 7. Spalte: Fehlende Werte. Gibt an, welche numerischen Werte nicht in Analysen einbezogen werden sollen. Beispiel: v285 „Familienstand Befragte/r“: Der Wert 9 bedeutet „Keine Angabe“. Da der Wert 9 (und alle höheren Werte) als fehlender Wert deklariert wurde, werden die Fälle, die den Wert 9 haben, bei Analysen ausgeschlossen. Da „Keine Angabe“ keine Information zum Familienstand beinhaltet, ist dies nur sinnvoll.
- 8. Spalte: Spaltenbreite. Breite der Variable in der Datenansicht.
- 9. Spalte: Ausrichtung. In der Standardeinstellung bei numerischen Variablen rechtsbündig, bei String- (Text-)Variablen rechtsbündig.
- 10. Spalte: Skalenniveau. Im Grunde uninteressant, weil auch bei falsch eingegebenen Skalenniveau fast jede Berechnung mit allen numerischen Variablen durchgeführt werden kann. Nur relevant für die „Interaktiven Graphiken“. Bitte verlassen Sie sich nicht darauf, dass das Skaleniveau richtig angegeben ist; die meisten Nutzer von Datensätzen achten nicht darauf.

- Wechseln zwischen Daten- und Variablenansicht: Strg + T.
- Öffnen der Fenster für die Wertelabels und fehlenden Werte: Leertaste.

1.1.2 Syntaxfenster

- Öffnen eines Syntaxfensters: Datei – Neu – Syntax (File – New – Syntax).
- Dokumentation und Ausführung sämtlicher Kommandos (Datentransformation, Analysen).
- Syntaxdateien können unabhängig von der Datendatei unter einem eigenen Namen abgespeichert werden.
- In die Syntaxdateien können eigene Kommentare eingefügt werden, um später leichter nachvollziehen zu können, welche Datentransformationen und Analysen durchgeführt wurden. Die Kommentare müssen immer mit einem * beginnen und mit einem Punkt enden.
- Es empfiehlt sich dringend, jedes Kommando, auch das Öffnen und Speichern der Datendatei, per Syntax und nicht durch Klicken auf der Windows-Oberfläche zu aktivieren. Der Vorteil beim Arbeiten mit der Syntax ist die Wiederholbarkeit: Wenn Analysen ein zweites Mal

durchgeführt werden müssen, Datendateien plötzlich verschwunden sind oder Datentransformationen fehlerhaft sind, ist es über die Windows-Oberfläche entweder mühsam oder unmöglich, die Fehler zu korrigieren, Analysen zu wiederholen oder Dateien wiederherzustellen.

1.1.3 Ausgabefenster

- Öffnet sich automatisch, wenn der erste Output (das Ergebnis einer Analyse o. ä.) produziert wird.
- Beinhaltet (bei der richtigen Einstellung von SPSS, siehe Kap. 1.2.1) alle Analyse-Ergebnisse, die dazugehörige Syntax und Fehlermeldungen.
- Linker Rand: Inhaltsangabe des Gesamtinhalts der Ausgabe.
- Einzelne Teile der Ausgabe können durch Markierung in der Inhaltsangabe und Entf gelöscht werden; der gesamte Inhalt der Ausgabe durch Markieren von „Ausgabe“ (Output) und Entf.

1.2 Arbeiten mit der Syntax

1.2.1 Vorbereitung

- Sinnvolle Einstellungen: Bearbeiten – Optionen (Edit – Options):
  - Reiter Allgemein: „Syntax-Fenster beim Start öffnen” („Open syntax window at start-up“) aktivieren,
  - Reiter Viewer: „Befehle im Log anzeigen” („Display commands in the log“),
  - Reiter Beschriftung der Ausgabe (Output Labels): Jeweils die letzte angezeigte Option aktivieren („Namen und Labels“).
- Geschmackssache:
  - Reiter Allgemein: „Fenster des Viewer öffnen“ („Raise viewer Output“) und „Zur neuen Ausgabe blättern“ („Scroll to new output“) aktivieren,
  - Reiter Diagramme: „Nur Muster durchlaufen“ („Cycle through patterns”) aktivieren,
  - Reiter Pivot-Tabellen (Pivot-Tables): Design der Tabellen verändern,
  - Reiter Daten (Data): „Anzeigeformat für neue numerische Variablen“ („Display Format for New Numeric Variables“) verändern, z.B. 0 Dezimalstellen (kommt auf das Format der Variablen an).

1.2.2 Kommandos: Grundlagen

- SPSS unterscheidet nicht zwischen Groß- und Kleinschreibung.
- Jede Datentransformation / Analyse wird in der Syntax durch ein Hauptkommando eingeleitet.
- Wenn außerdem Unterkommandos spezifiziert werden müssen, werden diese durch / abgetrennt (in einer Zeile oder in der nächsten Zeile).
- Jedes Kommando (Hauptkommando plus Unterkommandos) wird durch einen Punkt abgeschlossen. Fehlende Punkte sind eine der Hauptfehlerquellen in der Syntax.
- Ausführung eines Kommandos: Stelle markieren oder mit dem Cursor an die Stelle, dann Strg + R oder Icon > oder im Menü „Ausführen“ („Run“) und die entsprechende Aktion auswählen.
- Syntaxhilfe: Im Syntaxfenster kann durch das vierte Icon von rechts die Syntaxhilfe zum aktuellen Hauptkommando aufgerufen werden. Die Syntaxhilfe zeigt sämtliche Unterkommandos, die möglich sind. Unterkommandos, die mit zwei Sternchen gekennzeichnet sind, sind die Standardeinstellung. Sie müssen nicht explizit formuliert werden.
- Eine genauere Beschreibung der einzelnen Kommandos mit Beispielen bietet das Handbuch SPSS-Base.
- (Fast) alle Kommandonamen können durch die (meist drei, manchmal vier) ersten Buchstaben abgekürzt werden.
- Wenn ein Kommando sich auf mehrere Variablen bezieht, die im Datensatz direkt aufeinander folgen, kann das Schreiben der Liste dieser Variablen durch Einsetzen des Befehls „to“ vereinfacht werden. Beispiel: „frequencies v1 to v5“ statt „frequencies v1 v2 v3 v4 v5“.
- Wenn ein Kommando für mehrere Werte gilt, die direkt aufeinander folgen, kann durch den Befehl „thru“ abgekürzt werden. Beispiel: „recode (1 thru 5=0)“ statt „ recode (1,2,3,4,5=0)“.
- Weitere zentrale Befehle/Operatoren:
  - Addition: +
  - Subtraktion: -
  - Multiplikation: *
  - Division: /

- Exponieren: **
- größer als: > oder gt (greater than)
- größer gleich: >= oder ge (greater equal)
- kleiner als: < oder lt
- kleiner gleich: <= oder le
- gleich: = oder eq
- ungleich: <> oder ~= oder ne
- logisches „und“: & oder and
- logisches „oder“: | oder or
- logisches „nicht“: not

- Alle arithmetischen und statistischen Funktionen sind in SPSS-Base beim Kommando „compute“ aufgeführt, alle logischen Operatoren unter „do if“.

2 Öffnen und Speichern einer Datendatei

2.1 Öffnen

Datensätze (und damit auch das Programm) können durch einen Doppelklick auf die jeweilige Datei geöffnet werden. Die anfangs etwas umständlichere Version, die Datei über ein Kommando in der Syntax zu öffnen, hat allerdings einen entscheidenden Vorteil: Wenn ein Datensatz, in dem neu berechnete Variablen enthalten sind, aus irgendwelchen Gründen nicht mehr auffindbar ist, oder wenn die Ergebnisse von Analysen verschwunden sind, können Datensätze oder Analysen exakt repliziert werden.

```
get file = "C:\Magisterarbeit\Datensatz_06-01-13.sav".
```

Bedeutung: Öffne das Datenfile, das auf dem Laufwerk C im Ordner „Magisterarbeit" liegt und „Datensatz_06-01-13" heißt. (Der in diesem Beispiel angegebene Pfad ist für die Arbeit im Kurs nicht korrekt und muss ersetzt werden. Alle anderen Kommandos, die im Skript enthalten sind, können dagegen ohne Änderungen auf die Beispiel-Datensätze angewendet werden.)

Die Dateiendung „sav" steht für SPSS-Datenfiles und muss immer mit angegeben werden. Beachtet werden muss außerdem, dass der gesamte Pfad angegeben ist, d. h. die Namen aller Ordner, Unterordner und Unter-Unterordner, soweit vorhanden. Das bedeutet, dass Datensätze nie an eine andere Stelle auf dem Rechner verschoben werden sollten, weil SPSS die Datei in diesem Fall nicht mehr findet.

2.2 Speichern

Die Speicherung aller Datentransformationen in einem neuen Datenfile wird durch das Kommando „save outfile" angefordert.

```
save outfile = "C:\Magisterarbeit\Datensatz_06-01-13_2.sav".
```

Bedeutung: Speichere den Datensatz auf dem Laufwerk C im Ordner „Magisterarbeit" unter dem Namen „Datensatz_06-01-13_2". (Auch dieser Pfad muss im Kurs ersetzt werden.)

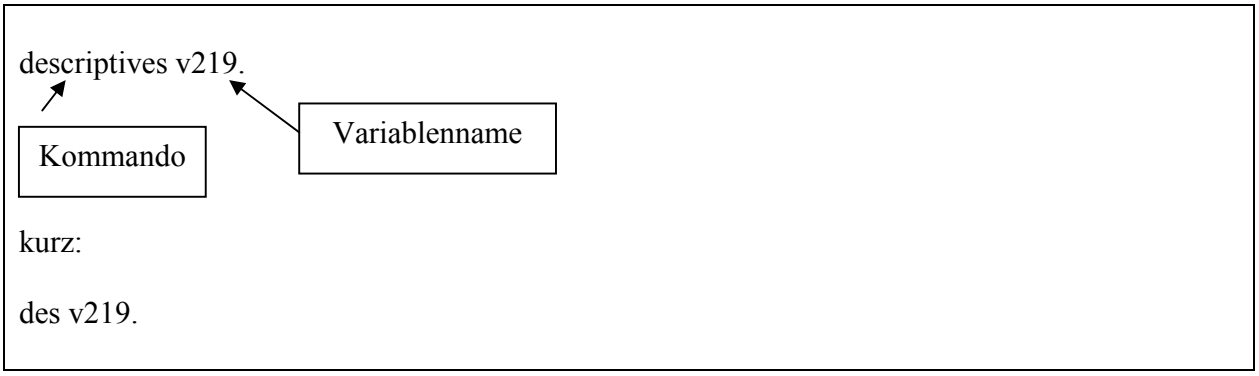
Am besten ist es, bei einem neuen Syntax-File sofort mit den beiden Kommandos „get file“ und „save outfile“ zu beginnen und alle anderen Kommandos zwischen die beiden zu schreiben. Bei regelmäßigem Speichern des Syntax-Files besteht so die Sicherheit, alle Datentransformationen etc. auch dann replizieren zu können, wenn der Rechner zwischendurch abstürzt.

3 Datentransformation

Unabhängig davon, ob man einen eigenen Datensatz erhebt oder eine Sekundäranalyse eines bestehenden Datensatzes durchführt, stehen die Variablen fast grundsätzlich nicht in der Form zur Verfügung, wie man sie braucht. Entweder müssen Ausprägungen zusammengefasst werden, oder es müssen neue Variablen auf der Basis der Werte bestehender Variablen berechnet werden, oder es müssen Informationen aus mehreren Variablen zusammengeführt werden, um eine neue zu bilden etc. Datentransformationen und die Bildung von neuen Variablen machen deshalb einen nicht unerheblichen Teil einer empirischen Analyse aus.
Die beiden wichtigsten Kommandos zur Datentransformation sind „compute“ und „recode (into)“; außerdem hilfreich sind „count“ und die logische Bedingung „do if“.

3.1 Descriptives

Bevor mit einer Variable gearbeitet werden kann, sollte man diese erst einmal genauer betrachten. Mit dem Befehl „Descriptives“ kann man sich die Anzahl der Fälle (N), Minimum und Maximum der Verteilung, den Mittelwert sowie die Standardabweichung anzeigen lassen. Dies wird im Folgenden an der Variable v219 (Alter des Befragten) demonstriert.

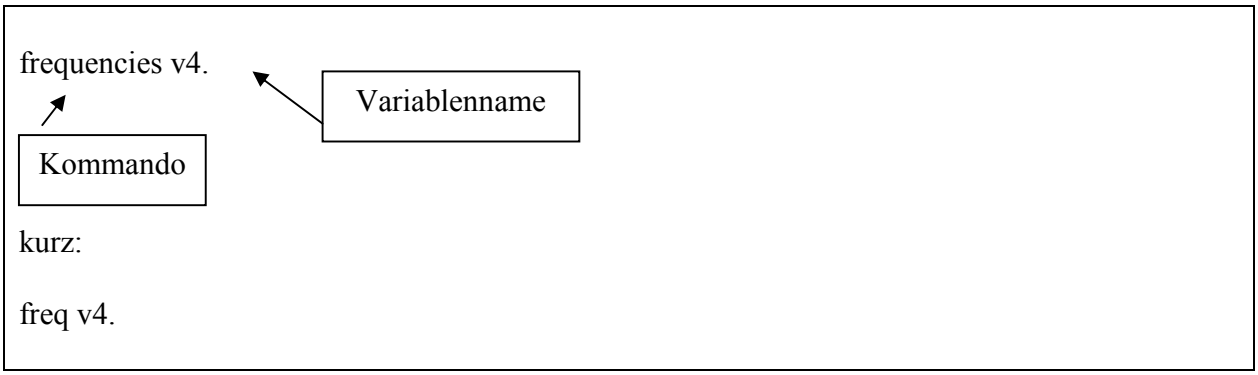


Bedeutung: Berechne die deskriptive Statistik für die Variable „v219“ (Alter des Befragten).

Anhand der deskriptiven Statistik wird nun ersichtlich wie viele der Befragten ihr Alter in der Befragung angegeben haben (N). Das Minimum der Verteilung zeigt wie alt der Jüngste unter den Befragten ist und das Maximum wie alt der Älteste ist. Der Mittelwert gibt das Durchschnittsalter der Befragten an und die Standardabweichung gibt an in welchem Bereich sich das Alter hauptsächlich bewegt.

3.2 Frequencies

Im nächsten Schritt des Kennenlernens der Variable ist es möglich, sich eine Häufigkeitsauszählung ausgeben zu lassen. Wenn aus dieser Variable dann noch eine neue Variable berechnet werden soll, empfiehlt es sich grundsätzlich, sich die Werte der Ausgangsvariablen durch eben eine solche Häufigkeitsauszählung anzeigen zu lassen.



Bedeutung: Gib eine Häufigkeitsauszählung der Variablen „v4“ aus.

Die Ausgabe besteht aus einer Übersicht der gültigen und fehlenden Fälle in der Variable v4 (Deutsche Staatsangehörigkeit?) und der Häufigkeitsauszählung. 1. Spalte: absolute Häufigkeiten, 2. Spalte: relative (prozentuale) Häufigkeiten, 3. Spalte: relative Häufigkeiten bezogen auf die Anzahl der gültigen Fälle, 4. Spalte: kumulierte relative Häufigkeiten, bezogen auf die Anzahl der gültigen Fälle.

Die Tabellen, die SPSS ausgibt, können in einigen Punkten verändert werden. Nach einem Doppelklick auf die Tabelle kann z.B. die Spaltenbreite verändert werden, können Spalten ganz gelöscht werden, können einzelne Zahlen fett oder kursiv gesetzt werden etc.

3.3 Compute

3.3.1 Beispiel 1: Alter

Das Geburtsjahr wird in Analysen fast nie benutzt, sondern in den meisten Fällen in das Alter des Befragten umgerechnet.

```
compute alter = 2000 – v218.
execute.

kurz:

comp alter = 2000 – v218.
exe.
```

Bedeutung: Berechne eine neue Variable mit dem Namen „alter“. Der Wert der Variable „alter“ soll für jeden Fall die Differenz von 2000 (dem Jahr der Befragung) und dem Geburtsjahr sein.

Datentransformations-Kommandos müssen grundsätzlich mit dem Kommando „execute“ (kurz: exe) abgeschlossen werden, anderenfalls wird der Befehl von SPSS nicht ausgeführt. Auch das Kommando execute muss mit einem Punkt beendet werden.

3.3.2 Beispiel 2: Äquivalenzeinkommen

In Analysen, in denen das Einkommen eine Rolle spielt, wird meist nicht das Haushalts-Nettoeinkommen verwendet, sondern das Äquivalenzeinkommen, d.h. das Haushalts-Nettoeinkommen, bezogen auf die Anzahl der Haushaltsmitglieder. Der Grundgedanke hierbei ist, dass Personen, die gemeinsam in einem Haushalt leben, insgesamt ein geringeres Einkommen als ein Single benötigen, um denselben Lebensstandard erreichen zu können. Eine Wohnungseinrichtung kostet z. B. für zwei Personen nicht doppelt soviel wie für eine Person.
Es gibt verschiedene Formeln, mit denen ein Äquivalenzeinkommen berechnet werden kann. Bei den meisten benötigt man Angaben zum Alter der Haushaltsmitglieder, weil Kindern, Jugendlichen und Erwachsenen unterschiedliche Gewichte zugeordnet werden. Eine einfache und brauchbare Formel (die sog. „neue OECD-Formel“) ist dagegen:

```
comp aeq = v544 / sqrt(v488).
exe.
```

Bedeutung: Berechne eine neue Variable „aeq“ (Äquivalenzeinkommen), indem das Haushalts-Nettoeinkommen durch die Wurzel der Haushaltsgröße (sqrt(Variablenname)) dividiert wird.

3.3.3 Beispiel 3: Institutionenvertrauen

Es gibt theoretische Konstrukte, die durch mehrere Items (Fragen) erhoben werden. Dies trifft häufig auf Einstellungskonstrukte zu. Beispielsweise wird das generelle Institutionenvertrauen der Befragten durch das Vertrauen in viele einzelne Institutionen gemessen. Wenn aber eine Aussage über das durchschnittliche Institutionenvertrauen gemacht werden soll, können die Antworten für die einzelnen Institutionen zu einer Variable zusammengefasst werden.

```
comp instver= mean(v82, v83, v84, v85, v86, v87, v88, v89, v90, v91, v92, v93, v94, v95, v96,
v97, v98).
exe.

kürzer:

comp instver= mean(v82 to v98).
exe.
```

Bedeutung: Berechne eine neue Variable „instver“ (Institutionenvertrauen), welche das durchschnittliche (mean) Vertrauen in die einzelnen Institutionen zusammenfasst.

3.4 Recode into

Mit dem Befehl „recode into“ können bestehende Variablen in neue, veränderte Variablen umkodiert werden. Hauptsächlich werden Ausprägungen der Variable oder ganze Wertebereiche zusammengefasst und neu definiert.

Das Kommando „compute“ gibt für all die Fälle (Zeilen), die in der/den Ausgangsvariablen einen systemdefiniert fehlenden Wert („System-Missings“) oder benutzerdefiniert fehlenden Wert („User-Missing“) aufweisen, einen System-Missing aus. Bei der neu berechneten Variablen „alter“ ist das unerheblich, weil die Ausgangsvariable keine fehlenden Werte hat. Beim Äquivalenzeinkommen zeigt sich dagegen eine Reihe von leeren Zellen. Eine

Häufigkeitsauszählung zeigt, dass für 775 Fälle kein Äquivalenzeinkommen berechnet werden konnte, weil das Haushaltseinkommen nicht angegeben wurde (und die Variable v544 die benutzerdefiniert fehlenden Werte 0, 99997 oder 99999 hat) und/oder die Zahl der Haushaltsmitglieder nicht angegeben wurde (und die Variable v488 den benutzerdefiniert fehlenden Wert 99 hat).
Man kann diese System-Missings so belassen und darauf hoffen, dass man auch nach einigen Wochen oder Monaten noch weiß, warum in diesen Zellen keine gültigen Werte stehen. In der Regel ist es allerdings so, dass man genau das nach einiger Zeit nicht mehr weiß. Deshalb sollten die benutzerdefiniert fehlenden Werte aus der/den Ausgangsvariablen in die neu berechnete Variable übertragen werden.

```
recode v544 (0,99997,99999=99999) into aeq.
recode v488 (99=99998) into aeq.
execute.

kürzer:

recode v544 (0,99997,99999=99999) into aeq
/ v488 (99=99998) into aeq.
execute.
```

Bedeutung: Gib den Fällen, bei denen die Ausgangsvariable v544 (Haushalts-Nettoeinkommen) die Werte 0, 99997 oder 99999 hat, der Zielvariablen aeq (Äquivalenzeinkommen) den Wert 99999. Gib den Fällen, bei denen die Ausgangsvariable v488 (Haushaltsgröße) den Wert 99 hat, der Zielvariablen aeq ebenfalls den Wert 99998.
Eine erneute Häufigkeitsauszählung zeigt, dass die Variable aeq jetzt keine systemdefiniert fehlenden Werte mehr hat.

3.4.1 Beispiel: Altersgruppen

Mit „thru“ können die Wertebereiche einer Variablen einer neuen Variablen zugewiesen werden. Dies bietet sich beispielsweise für Altersgruppen oder Einkommensklassen an. Beispielhaft wird das Bilden von Altersgruppen gezeigt.

```
recode v219 (18 thru 27 = 1) (28 thru 37 = 2) (38 thru 47 = 3) (48 thru 57 = 4) (58 thru 67 = 5)
(68 thru 95 = 6) into altgrup.
exe.
```

Bedeutung: Fasse die Befragten der Ausgangsvariable v219 (Alter) zwischen 18 und 27 Jahren zur Altersgruppe 1 zusammen, Befragte zwischen 28 und 37 Jahren zur Gruppe 2, Befragte zwischen 38 und 47 Jahren zur Altersgruppe 3, diejenigen zwischen 48 und 57 Jahren zur Gruppe 4, Personen zwischen 58 und 67 Jahren zur Altersgruppe 5 und alle zwischen 68 und 95 Jahren zur Altersgruppe 6. Bilde hierfür die neue Variable „altgrup“ (Altersgruppen).

3.4.2 Beispiel: Kirchgangshäufigkeit

Für die Berechnung von Zusammenhängen kann es sinnvoll sein die Skalierung einer Variablen neu anzuordnen. Dies könnte beispielsweise für die Variable Kirchgangshäufigkeit gelten. Im Datensatz ist diese durch die Ausprägungen „Über 1X die Woche“, „1X pro Woche“, „1-3X im Monat“, „Mehrmals im Jahr“, „Seltener“ und „Nie“ operationalisiert. Für die Untersuchung eines Zusammenhangs der von einem Einfluss von Religiosität ausgeht, ist es sinnvoller die Variable als zunehmende Kirchgangshäufigkeit miteinzubeziehen. Hierfür kann die Antwortskala durch das Kommando „recode into“ umgedreht werden.

```
recode v622 (6=1) (5=2) (4=3) (3=4) (2=5) (1=6) (else=sysmis) into kirch.
exe.
```

Bedeutung: Verändere die Ausprägungen der Ursprungsvariablen v622 (Kirchgangshäufigkeit) wie folgt, die Ausprägung „Nie“ bekommt den Wert 1, dies steigert sich bis „Über 1X die Woche“ welches die höchste Ausprägung 6 bekommt. Bilde hieraus eine neue Variable „kirch“ (ansteigende Kirchgangshäufigkeit), in der die weiteren Ausprägungen „Trifft nicht zu“, „Verweigert“ und „Keine Angabe“ als fehlende Werte angegeben werden.

3.4.3 Beispiel: Erwerbstätigkeit

Angenommen es ist in einem ersten Schritt von Interesse, ob ein Befragter erwerbstätig ist oder nicht. Zu diesem Zweck lässt sich eine Variable berechnen, welche nur erwerbstätig ja/nein beinhaltet (Dummy-Variable).

```
recode v234 (1 thru 3=1) (4=0) (99=sysmis) into erwerb.
exe.
```

Bedeutung: Bilde aus der ursprünglichen Variable „v234“ eine neue Variable „erwerb“ (Erwerbstätigkeit), in der alle Befragten mit irgendeiner Form der Erwerbstätigkeit die Ausprägung 1 bekommen und diejenigen die nicht erwerbstätig sind die Ausprägung 0 und gebe die Ausprägung „keine Angabe“ (99) als fehlenden Wert an.

3.5 Variablenlabels, Wertelabels, Fehlende Werte

Jede neue Variable sollte direkt, nachdem sie berechnet wurde, gelabelt werden. Später weiß man unter Umständen nicht mehr, welchen Inhalt die Variable hat und welche inhaltliche Bedeutung die numerischen Werte der Variablen haben. Das gilt insbesondere für den Fall, dass bestimmte Werte in späteren Analysen nicht mit einbezogen werden sollen, also fehlende Werte sein sollen.

```
variable labels aeq “Äquivalenzeinkommen“.
kurz:
var lab aeq “Äquivalenzeinkommen“.
```

Bedeutung: Gib der Variable mit dem Namen „aeq“ das Variablenlabel „Äquivalenzeinkommen“.

```
value labels aeq 99998 “k.A. Haushaltsgröße“ 99999 “k.A. Einkommen“.
kurz:
val lab aeq 99998 “k.A. Haushaltsgröße“ 99999 “k.A. Einkommen“.
```

Bedeutung: Der Wert 99998 in der Variable “aeq” erhält das Wertelabel “k.A. Haushaltsgröße” und der Wert 99999 erhält das Wertelabel “k.A. Einkommen“.

Wenn eine Variable schon mit Wertelabels ausgestattet ist und nur ein oder mehrere Labels zugefügt werden sollen, ist das Kommando val lab nicht geeignet, weil die alten Labels mit ihm gelöscht werden. Für das Hinzufügen von Wertelabels ist der Befehl add value labels (kurz: add val lab) geeignet.

Da SPSS durch die Rekodierung nicht mehr „weiß“, dass der Wert 99999 der Variable „aeq“ ein fehlender Wert ist, muss dieser zusätzlich noch einmal definiert werden.

```
missing values aeq (99998, 99999).
kurz:
mis val aeq (99998, 99999).
```

Bedeutung: Die Werte 99998 und 99999 in der Variable „aeq“ werden zum benutzerdefiniert fehlenden Wert deklariert, d.h. sie werden nicht in Analysen einbezogen.

Wenn, wie beim Äquivalenzeinkommen, die systemdefiniert fehlenden Werte durch einen User-Missing ersetzt werden sollen, sollte man darauf aufpassen, welchen Wert man wählt. Es wäre unsinnig, den User-Missings einen Wert zu geben, der schon in der Variablen als gültiger Wert vorkommt. Um sicherzugehen, sollte man eine Häufigkeitsauszählung der Variablen machen und prüfen, welche Werte sie hat. Im Fall des Äquivalenzeinkommens ist der höchste gültige Wert 56.569. Es bietet sich also an, für die User-Missings die 99999 zu wählen. In der Regel erhalten User-Missings die höchsten numerischen Werte, d.h. bei Variablen, deren Werte einstellig sind, die 9, bei zweistelligen Variablen die 99 usw.

3.6 Die „if“-Anweisung

Manchmal müssen bei der Berechnung einer neuen Variablen Informationen aus zwei oder drei verschiedenen Variablen zusammengefügt werden. Dazu kann die „if“-Anweisung nützlich sein. Aus der Variable v285 wird beispielsweise lediglich der offizielle Familienstand ersichtlich, nicht aber, ob die Person z.B. in einer nichtehelichen Lebensgemeinschaft lebt. Aus zwei weiteren Variablen ist zu entnehmen, ob die Person einen festen Lebenspartner hat (v403) und ob sie mit diesem einen gemeinsamen Haushalt führt (v438). Wenn eine Variable berechnen werden soll, welche die drei Ausprägungen „verheiratet“, „nichteheliche Lebensgemeinschaft“ und „Single“ aufweist, müssen zur Berechnung die Variablen v285, v403 und v438 herangezogen werden.

```
comp famstand = 99.
if v285 = 1 famstand = 1.
if v403 =1 & v438 = 1 famstand = 2.
if (v285 ~= 1 & v403 ~= 1) | (v403 = 1 & v438 = 2) famstand =3.
exe.
```

Bedeutung: Der Variablen wird zunächst ein Wert (99) zugewiesen, der später als fehlender Wert deklariert werden sollte. Die Variable erhält die Ausprägungen 1, wenn eine Person verheiratet ist, die Ausprägung 2, wenn sie mit einem festen Lebenspartner in einem gemeinsamen Haushalt lebt und die Ausprägung 3, wenn sie weder verheiratet ist noch einen festen Lebenspartner hat bzw. mit diesem nicht zusammen lebt.

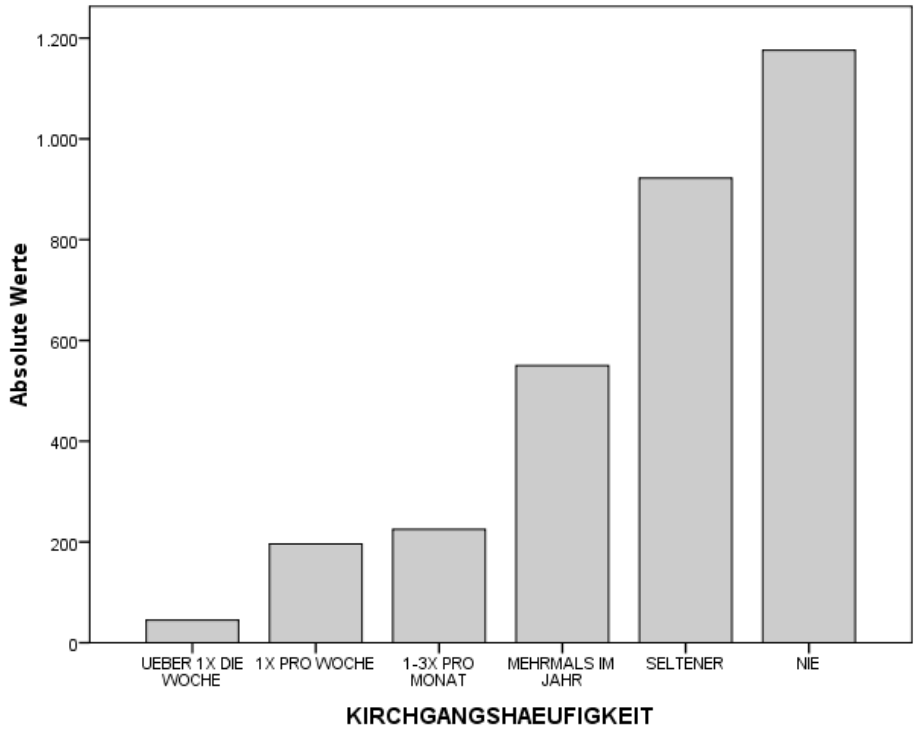
3.7 Grafische Darstellung von Häufigkeiten

Grafiken können statistische Informationen anschaulich darstellen, da optische Information kognitiv oftmals besser verarbeitet werden können als Zahlenreihen. Es muss jedoch darauf geachtet werden, dass die Informationen einer oder mehrerer Variablen sinnvoll visualisiert werden. Einen Anhaltspunkt, welche Grafik für welche Art von Variablen geeignet ist, gibt das Skalenniveau, da dieses zeigt, wie viele Werte die Variable annehmen kann.

3.7.1 Säulendiagramme

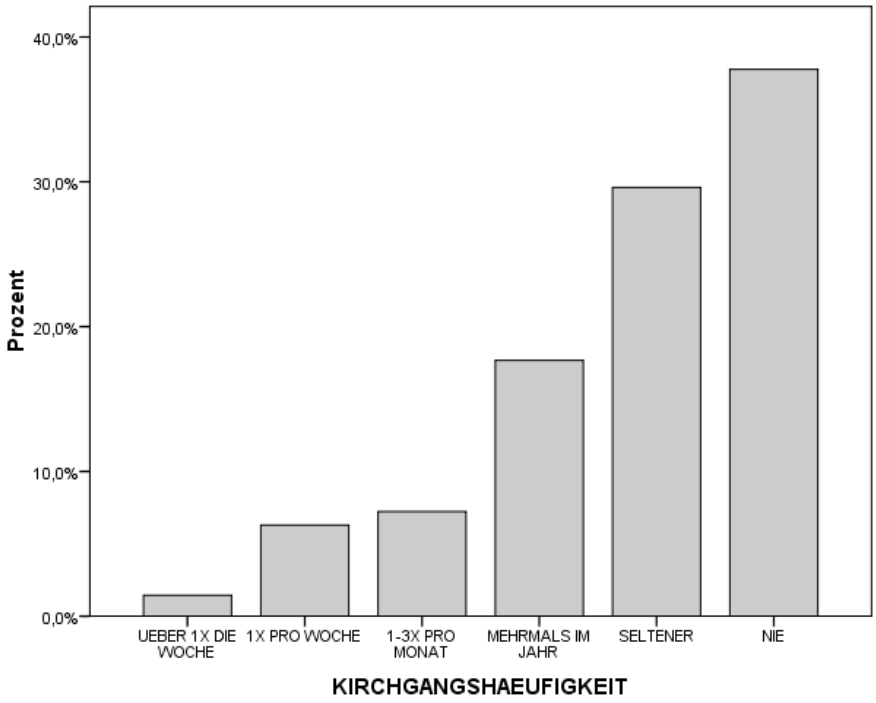
Nominal- und ordinalskalierte Variablen können sinnvollerweise mit einem Säulendiagramm visualisiert werden. Mit dem folgenden Befehl kann man die absoluten Häufigkeiten der ordinalskalierten Variablen v622 „Kirchgangshäufigkeit“ grafisch darstellen.

```
graph
/bar = v622.
```



Man kann beispielsweise an dieser Grafik ablesen, dass die meisten Befragten (nahe zu 1200) nie in die Kirche gehen. Die wenigsten Befragten (ca. 50) gehen häufiger als ein Mal pro Woche in die Kirche. Mit den Informationen über die absoluten Häufigkeiten ist jedoch erstmal nicht viel anzufangen, da sich immer die Frage anschließt: 1200 Personen von wie vielen? Aus diesem Grund ist es zumeist sinnvoller, u.a auch, wenn man Variablen oder Gruppen vergleichen will, sich nicht die absoluten, sondern die relativen Häufigkeiten einer Variablen ausgeben zu lassen. Dies lässt sich mit folgendem Befehl realisieren.

```
graph
/bar = pct by v622.
```

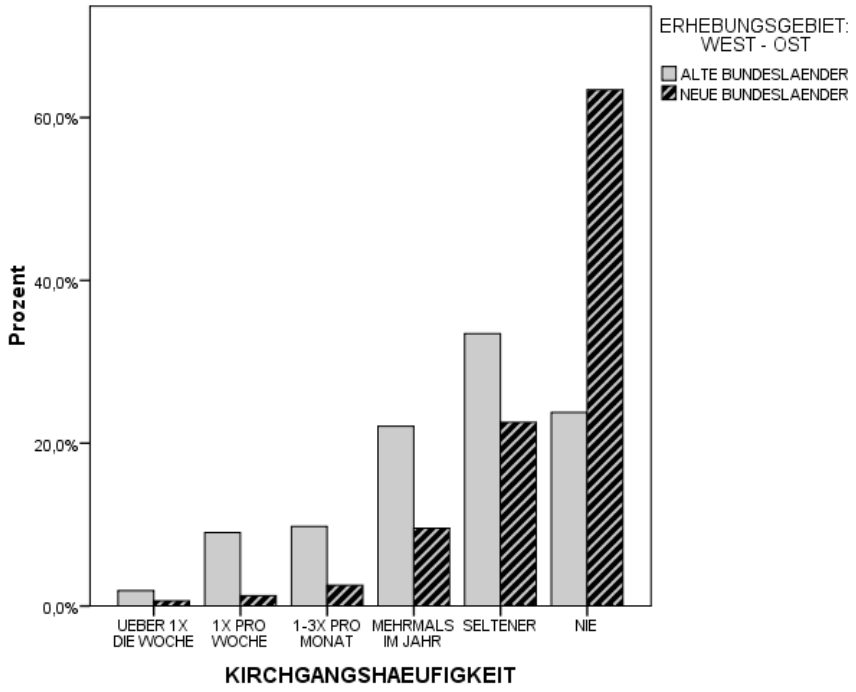


Nun wird aus der Grafik ersichtlich, dass jene Personen, die nie zur Kirche gehen ca. 38 Prozent der Befragten ausmachen. Häufiger als ein Mal pro Woche gehen lediglich ungefähr 2 Prozent der Befragten in die Kirche.

Will man wissen, ob sich Ost- und Westdeutsche in ihrer Kirchgangshäufigkeit unterscheiden, kann dies ebenfalls mit einem Säulendiagramm anschaulich visualisiert werden.

```
graph
/bar = pct by v622 by v3.
```

Bedeutung: Erstelle eine Grafik („graph"), genauer ein Balkendiagramm („bar"), in dem die prozentualen Anteile („pct by") der Variablen v622 (Kirchgangshäufigkeit) in Abhängigkeit von der Variablen v3 (Ost/Westdeutschland) abgebildet werden.



Man sieht, dass sich Ost- und Westdeutsche wesentlich bezüglich ihrer Kirchgangshäufigkeit unterscheiden. Dies überrascht allerdings nicht besonders, da die Religiosität in Ostdeutschland bei weitem nicht so ausgeprägt ist wie in Westdeutschland. Anhand dieses Vergleichs wird deutlich, dass der Anteil der Befragten die nie zur Kirche gehen in den neuen Bundesländern um fast 40 Prozentpunkte über dem Anteil derjenigen in den alten Bundesländern liegt.

Durch einen Doppelklick auf das Balkendiagramm im SPSS-Output, kann der Editor aktiviert und dann die Grafik verändert werden. Außerdem kann es durch einen Klick mit der rechten Maustaste „Objekte kopieren" in ein Word-Dokument eingefügt werden.

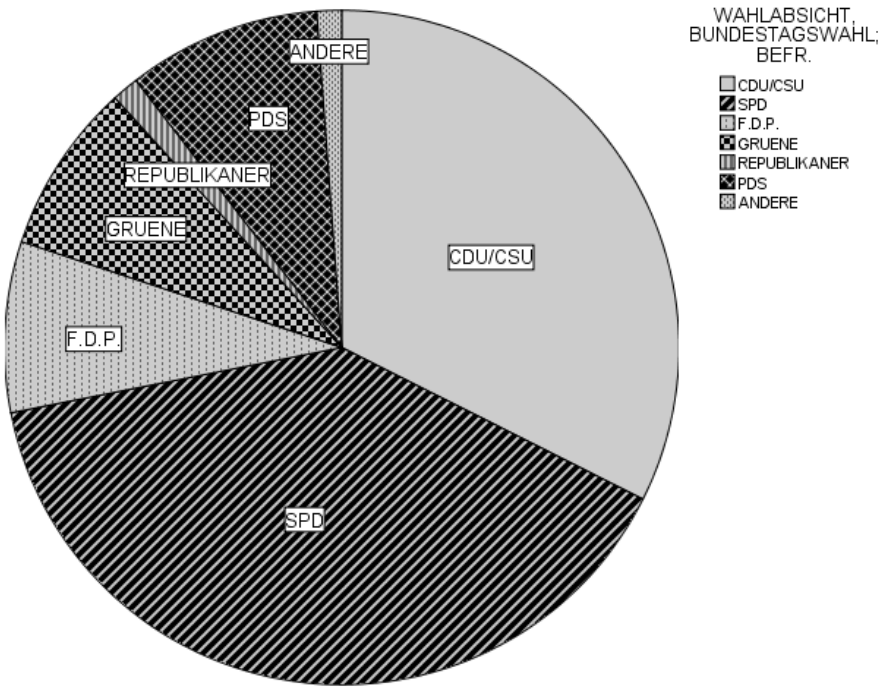
3.7.2 Kuchendiagramme

Kuchen- oder Kreisdiagramme werden in der Sozialforschung eher selten verwendet. Sie eigenen sich hauptsächlich dazu, relative Häufigkeiten dazustellen, die sich zu einer Gesamtsumme von 100 Prozent addieren lassen. Ein erheblicher Nachteil von Kuchendiagrammen ist, dass Anteile, die in etwa denselben Wert aufweisen, nur schwer vergleichbar sind. Wenn zudem noch für einen Anteil eine kräftige Farbe, für einen anderen dagegen eine sanfte verwendet wird, können sich optische Täuschungen ergeben. In Säulendiagrammen ist dieser Vergleich hingegen erheblich leichter.

Wie sehen die Wahlabsichten bezüglich der nächsten Bundestagswahl bei den Befragten des ALLBUS 2000 aus?

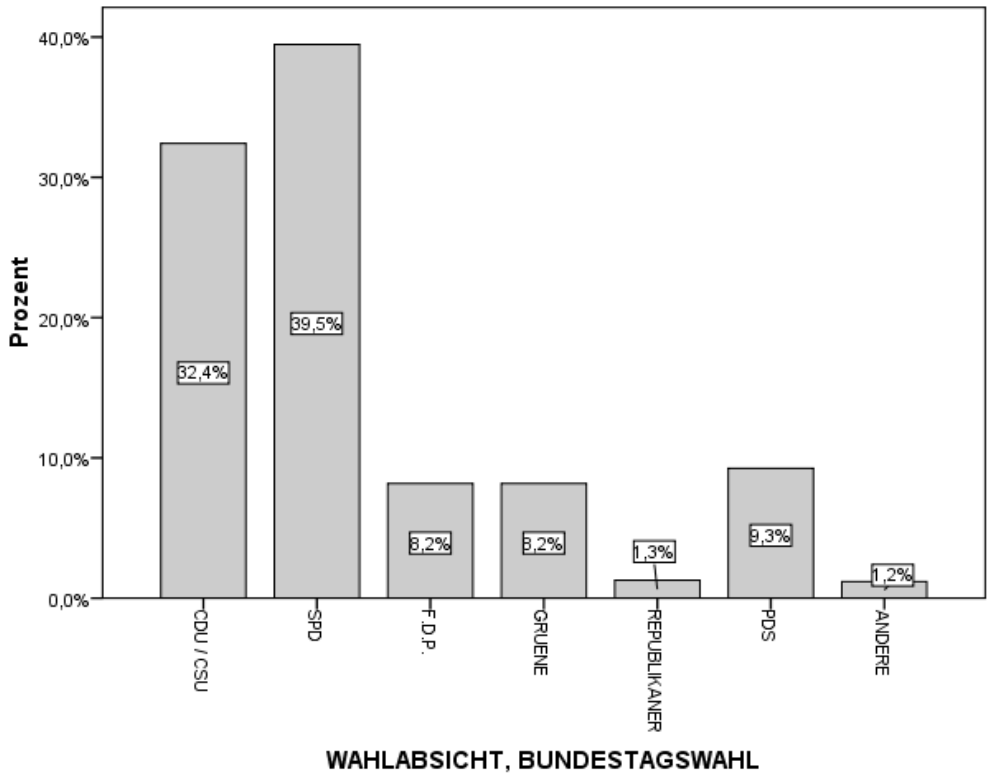
```
graph
/pie =  v627.
```

Bedeutung: Erstelle ein Kuchendiagramm („pie“) der Variablen v627 (Wahlabsichten).



Eine ganz knappe Mehrheit der ALLBUS-Befragten hatte im Jahr 2000 vor, bei der nächsten Bundestagswahl die SPD zu wählen. Der Anteil der CDU-Wähler macht einen nahezu ähnlich großen Anteil aus und bildet somit die zweitgrößte Gruppe. Die Anteile der FDP, der GRÜNEN und der PDS scheinen ebenfalls annähernd gleich groß zu sein. Darüber hinaus stellt sich auch noch die Frage ob mehr Personen die Republikaner oder eine der „anderen Parteien“ wählen möchten. Ein Säulendiagramm kann diese Anteile, wie bereits erwähnt, besser vergleichen.

Aus der nachstehenden Grafik wird deutlich, dass die PDS vor den GRÜNEN und der FDP geringfügig mehr Stimmen erhalten würde und dass die GRÜNEN und die FDP, wie schon im Kuchendiagramm angedeutet, zu gleichen Anteilen gewählt werden würden. Die Republikaner zeigen nur einen um 0,1 Prozentpunkte höheren Stimmenanteil als die „anderen Parteien“.

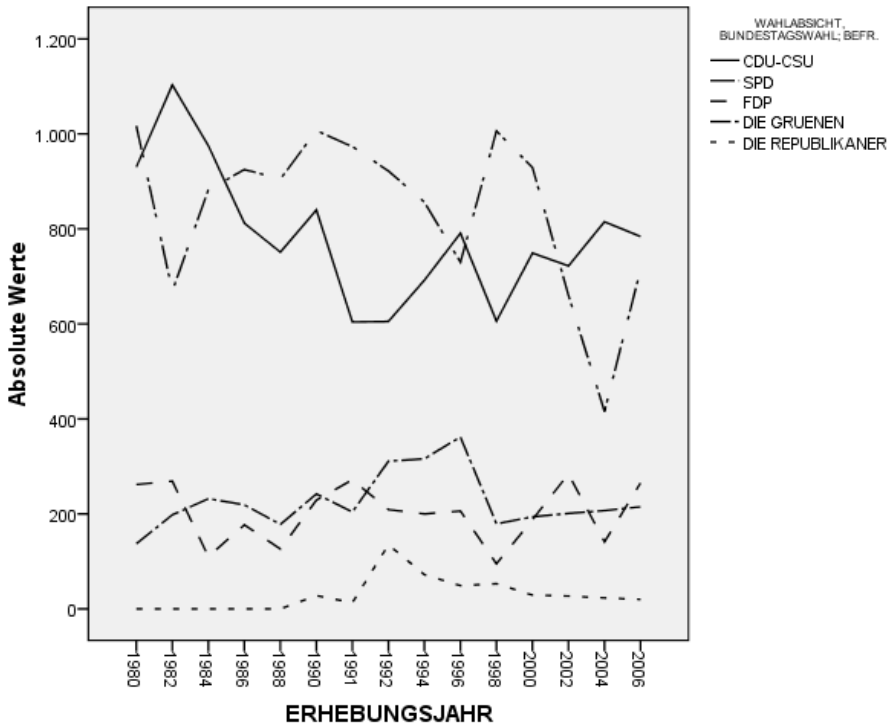


3.7.3 Liniendiagramme

Eine klassische Anwendung eines Liniendiagramms ist die Visualisierung von Zeitreihen, wie beispielsweise die Entwicklung der SPD-Wählerschaft oder die des Bruttoinlandsproduktes. Solche Entwicklungsverläufe können auch unter verschiedenen Gruppen verglichen werden. Wie entwickelte sich z.B. das BIP in Deutschland und den USA in den letzten 20 Jahren? Wie veränderte sich die SPD-Wählerschaft im Vergleich zu jener der CDU? Da für die Darstellung von Zeitverläufen Längsschnittsdaten von Nöten sind, wird für die folgenden Analysen der kumulierte ALLBUS 2006 herangezogen. In diesem befinden sich alle ALLBUS-Wellen von 1980 bis zum Jahr 2006.

```
graph
/line= v2 by v24.
```

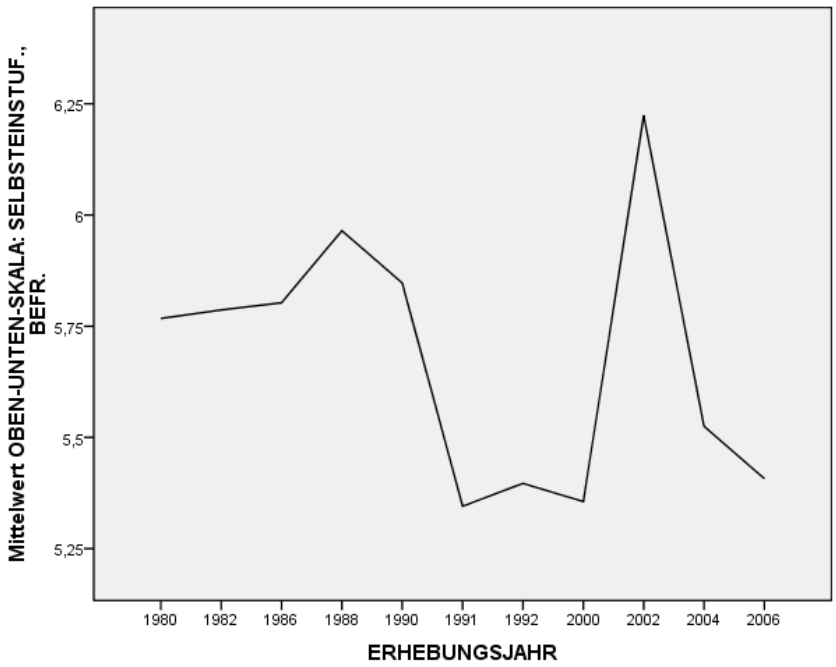
Bedeutung: Erstelle eine Grafik (“graph”), genauer ein Liniendiagramm („line“), in dem die absoluten Häufigkeiten der Variablen v24 (Wahlabsichten) über alle Erhebungszeitpunkte (v2) abgebildet werden.



Mit einem Liniendiagramm kann ebenso die Veränderung von Mittelwerten (oder auch anderen Maßzahlen) im Zeitverlauf dargestellt werden. Hier soll nun die Entwicklung des arithmetischen Mittels einer Selbsteinstufung der Befragten auf einer Oben-Unten-Skala abgebildet werden. Die Fragestellung hierzu lautet: „In unserer Gesellschaft gibt es Bevölkerungsgruppen, die eher oben stehen und solche, die eher unten stehen. Wir haben hier eine Skala, die von oben nach unten verläuft. Wenn Sie an sich selbst denken: Wo würden Sie sich auf dieser Skala einordnen?“

```
graph
/line= mean(v112) by v2.
```

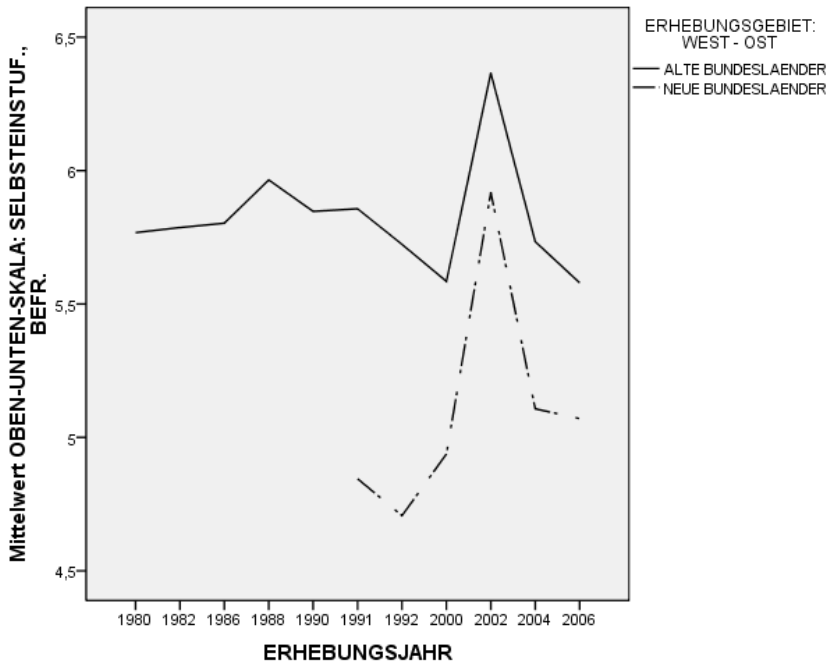
Bedeutung: Erstelle ein Liniendiagramm über die Mittelwerte der Variablen v112 (Oben-Unten-Selbsteinstufung) im Zeitverlauf (v2).



Augenfällig ist in dieser Grafik zunächst ein vergleichsweise extremer Abfall zwischen den Jahren 1990 und 1991, also direkt nach der Wiedervereinigung. Eine Aufklärung dieser Entwicklung kann eine grafische Darstellung getrennt nach Ost- und Westdeutschland liefern.

```
graph
/line= mean(v112) by v2 by v5.
```

Bedeutung: Erstelle ein Liniendiagramm über die Mittelwerte der Variablen v112 (Oben-Unten-Selbsteinstufung) im Zeitverlauf (v2) über die beiden Gruppen Ost- und Westdeutsche (v5).



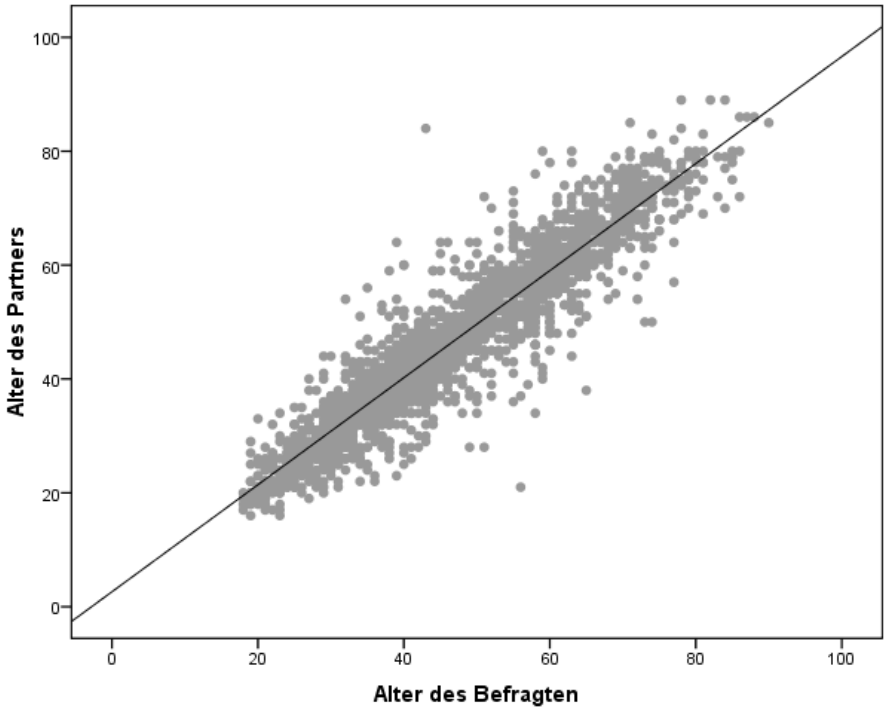
Wird Westdeutschland separat betrachtet, ergibt sich zwischen den Jahren 1990 und 1991 sogar ein leichter Anstieg des Mittelwertes zur Selbsteinstufung. Der starke Abfall in diesem Zeitintervall in der gesamtdeutschen Betrachtung war insofern den Ostdeutschen, die in diesem Jahr zum ersten Mal in die ALLBUS-Erhebung integriert wurden, geschuldet. Diese stufen sich im Vergleich zum Westen weiter unten in der Gesellschaft ein.

3.7.4 Streudiagramme

Ein Streudiagramm bietet die Möglichkeit zwei metrisch skalierte Variablen miteinander in Beziehung zu setzen. Wie stellt sich z.B. der Zusammenhang zwischen dem Alter des Befragten und dem Alter des Partners dar?

```
graph
/scatter = alter with alterpar.
```

Bedeutung: Erstellt ein Streudiagramm (scatter) mit der Variable „alter“ (Alter des Befragten) auf der x-Achse und der Variablen alterpar (Alter des (Ehe-)Partners) auf der y-Achse.



Auf der x-Achse sollte stets die unabhängige Variable abgetragen werden. Mit einem Doppelklick auf die Grafik kann man diese verändern und sich z.B. eine lineare Anpassungslinie im Diagramm anzeigen lassen (fünftes Icon von rechts in der unteren Menüleiste). Ihr steigender Verlauf zeigt hier an, dass mit steigendem Alter des Befragten das Alter des (Ehe-)Partners steigt.

3.7.5 Beschriftung einer Grafik

Grafiken sollten sinnvoll beschriftet werden. Die Beschriftung der beiden Achsen richtet sich nach den vergebenen Variablenlabels. Diese können also entsprechend der Erfordernisse verändert werden. Weiter können über die SPSS-Syntax einer Grafik z.B. Titel, Untertitel und Fußnoten gegeben werde.

```
graph
/scatter = alter with alterpar
/ title 'Alter und Alter des Partners'
/ footnote 'Quelle: ALLBUS 2000, eigene Berechnungen'.
```

Bedeutung: Erstellt das o.a. Streudiagramm mit dem Titel 'Alter und Alter des Partners' und der Fußnote 'Quelle: ALLBUS 2000, eigene Berechnungen'.

Die Position der Beschriftung sowie der Legende, Schriftgröße und -art können nach Aktivierung des Chart-Editors (Doppelklick auf die Grafik) verändert werden.

4 Kreuztabellen

Für die bivariate Analyse von Variablen mit nominalem Skalenniveau sind Kreuztabellen geeignet. Sie zeigen, kurz gesagt, welche prozentualen Anteile die Kategorien der abhängigen Variablen für die einzelnen Kategorien der unabhängigen Variablen haben. Eine wichtige Voraussetzung für die übersichtliche Darstellung der Ergebnisse einer Kreuztabellierung ist, dass die beiden Variablen nur wenige Ausprägungen (Kategorien) haben. Aus diesem Grund ist es häufig notwendig, die Kategorien einer Variablen vor der Kreuztabellierung zusammenzufassen. Ein weiterer Grund für diesen vorbereitenden Schritt ist die Zellenbesetzung, d. h. die Anzahl an Fällen in jeder Zelle der Kreuztabelle. Ist die Zellenbesetzung zu gering, können erstens die dargestellten prozentualen Anteile häufig nicht mehr sinnvoll interpretiert werden, und zweitens werden Signifikanztests unzuverlässig.

4.1 Vorbereitung: Zusammenfassung von Kategorien mit „recode into“

Gibt es in den neuen Bundesländern weniger Ausländer als in den alten Bundesländern? In der DDR gab es nur geringen Zuzug aus dem Ausland, in der alten BRD dagegen durch die Anwerbung von „Gastarbeitern“ und anderen Ursachen einen deutlich höheren Ausländeranteil in der Bevölkerung. Nun kann man sich die Frage stellen, ob diese unterschiedlich hohen Ausländeranteile auch mehr als 15 Jahre nach der Wiedervereinigung noch bestehen.

Im ALLBUS 2000 stehen für die Beantwortung dieser Frage zwei geeignete Variablen zur Verfügung: v3 (Erhebungsgebiet: West – Ost) und v4 (Deutsche Staatsangehörigkeit?). v4 hat allerdings vier Ausprägungen: 1=Ja; 2=Ja, neben 2. Staatsangehörigkeit; 3=Nein; 4=Staatenlos (und 9=Keine Angabe). So differenziert müssen die Informationen, die die Variable gibt, für die Beantwortung der Frage nicht sein. Im Gegenteil, es würde die Interpretation der Kreuztabelle erschweren, die Variable in dieser Form zu verwenden.

Da nur die Information gebraucht wird, ob die Befragten die deutsche Staatsangehörigkeit haben oder nicht, empfiehlt es sich hier, die Ausprägungen 1 (Ja, deutsche Staatsangehörigkeit) und 2 (Ja, neben 2. Staatsangehörigkeit) zusammenzufassen und die Ausprägungen 3 (Nein, keine deutsche Staatsangehörigkeit) und 4 (Nein, keine deutsche Staatsangehörigkeit, weil staatenlos) ebenfalls zusammenzufassen.

```
recode v4 (1,2 = 1) (3,4 = 0) (9 = 9) into staat.
exe.

oder:

recode v4 (1,2 = 1) (3,4 = 0) (9 = copy) into staat.
exe.

oder:

recode v4 (1,2 = 1) (3,4 = 0) (miss = copy) into staat.
exe.
```

Bedeutung: Die Fälle, die in der Variablen v4 die Werte 1 oder 2 haben, sollen in der neu zu bildenden Variablen „staat“ den Wert 1 erhalten. Die Werte 3 und 4 sollen analog den Wert 0 erhalten. Alle Fälle, die den Wert 9 haben, sollen in der neuen Variablen ebenfalls den Wert 9 erhalten.

Da der Wert 9 (Keine Angabe) in der Ausgangsvariablen als fehlender Wert deklariert ist, kann statt des konkreten Werts auch das Schlüsselwort „missing“ (kurz: miss) verwendet werden. Mit der Angabe „miss = copy“ werden alle Werte der Ausgangsvariablen, die als fehlender Wert deklariert sind, in die neue Variable kopiert. Wenn die Ausgangsvariable mehrere User-Missings hat, kann es passieren, dass man den Überblick verliert und einen dieser Werte schon vorher als gültigen Wert vergeben hat.

Die neue Variable „staat“ muss jetzt noch gelabelt werden, und die fehlenden Werte müssen deklariert werden.

```
var lab staat “deutsche Staatsangehörigkeit“.
val lab staat 0 “nicht deutsch“ 1 “deutsch“ 9 “k.A.“.
miss val staat (9).
```

4.2 Erstellung einer Kreuztabelle

Kreuztabellen werden mit dem Kommando „crosstabs“ erstellt. Die Grundform des Befehls sieht folgendermaßen aus:

```
crosstabs staat by v3.
```

Bedeutung: Erstelle eine Kreuztabelle, in der die Häufigkeiten der Ausprägungen der Variablen „staat“ für die Ausprägungen der Variablen „v3“ dargestellt werden. v3 ist also die unabhängige Variable (und steht damit hinter dem „by“), „staat“ ist die abhängige Variable.

Das Ergebnis sieht folgendermaßen aus:

STAAT deutsche Staatangehörigkeit * V3 ERHEBUNGSGEBIET: WEST - OST Crosstabulation

Count

		V3 ERHEBUNGSGEBIET: WEST - OST		
		1 ALTE BUNDESLAENDER	2 NEUE BUNDESLAENDER	Total
STAAT deutsche Staatangehörigkeit	0 nicht deutsch	210	5	215
	1 deutsch	2338	583	2921
Total		2548	588	3136

In den alten Bundesländern haben 210 der Befragten nicht die deutsche Staatsangehörigkeit und 2338 Befragte haben sie; in den neuen Bundesländern sind es 5 bzw. 583 Befragte. Insgesamt ergibt dies 2548 Fälle in den alten und 588 Fälle in den neuen Bundesländern. Diejenigen, die keine Angabe zu ihrer Staatsangehörigkeit gemacht haben (und den Wert 9 haben), werden nicht mit in die Analyse einbezogen.

Mit dieser Tabelle kann die Frage, ob es im Westen mehr Ausländer gibt als im Osten, nicht beantwortet werden. Es geht bei dieser Frage nicht um die absoluten Häufigkeiten, sondern darum, ob es im Westen einen höheren Ausländer*anteil* gibt als im Osten. Die absoluten

Häufigkeiten in den Zellen müssen also auf die Gesamthäufigkeiten bezogen werden. Kurz: Es muss prozentuiert werden.

```
crosstabs staat by v3
/cells = count column.

kurz:

cross staat by v3
/cells = cou col.
```

Bedeutung: Erstelle eine Kreuztabelle mit „v3" als unabhängiger und „staat" als abhängiger Variable. Gib für jede Zelle die absoluten Häufigkeiten („count") und die Spaltenprozente („column") an.

Die Ausgabe sieht jetzt folgendermaßen aus:

STAAT deutsche Staatangehörigkeit * V3 ERHEBUNGSGEBIET: WEST - OST Crosstabulation

			V3 ERHEBUNGSGEBIET: WEST - OST		
			1 ALTE BUNDESLAENDER	2 NEUE BUNDESLAENDER	Total
STAAT deutsche Staatangehörigkeit	0 nicht deutsch	Count	210	5	215
		% within V3 ERHEBUNGSGEBIET: WEST - OST	8,2%	,9%	6,9%
	1 deutsch	Count	2338	583	2921
		% within V3 ERHEBUNGSGEBIET: WEST - OST	91,8%	99,1%	93,1%
Total		Count	2548	588	3136
		% within V3 ERHEBUNGSGEBIET: WEST - OST	100,0%	100,0%	100,0%

Die 210 Bewohner der alten Bundesländer, die nicht die deutsche Staatsangehörigkeit haben, entsprechen 8,2 % aller Befragten aus dem Westen. In den neuen Bundesländern gibt es dagegen nur 0,9 % Nicht-Deutsche. Die eingangs gestellte Frage kann also mit „ja" beantwortet werden.

4.3 Signifikanztest bei Kreuztabellen

Die Aussage, dass es im Westen einen größeren Ausländeranteil gibt als im Osten, gilt mit der durchgeführten Analyse nur für die 3136 Personen, die in die Analyse eingegangen sind. Die weitaus interessantere Frage ist jetzt natürlich, ob dieser Zusammenhang zwischen Region und Staatsangehörigkeit auch für die Grundgesamtheit, d. h. für alle Bewohner von Deutschland, gilt. Diese Frage kann mit Hilfe eines Signifikanztests beantwortet werden.

```
cross staat by v3
/cells = cou col
/stat = chisq.
```

Bedeutung: Wie oben, zusätzlich soll der Chi²-Wert berechnet werden.

Die SPSS-Ausgabe wird mit diesem Unterkommando um eine Tabelle erweitert:

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	40,873[b]	1	,000		
Continuity Correction[a]	39,724	1	,000		
Likelihood Ratio	59,211	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	40,860	1	,000		
N of Valid Cases	3136				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 40,31.

In der Regel sind die meisten der Angaben, die SPSS in dieser Tabelle ausgibt, nicht von Interesse. Interessant ist meistens nur die erste Zeile, also der Chi²-Wert, die Anzahl der Freiheitsgrade und das Signifikanzniveau. Bei 2 x 2-Tabellen, also bei Kreuztabellen, in denen die beiden Variablen jeweils genau zwei Ausprägungen haben, wird außerdem der exakte Test nach Fischer angegeben.

Der Chi²-Wert wird auf der Basis der beobachteten absoluten Häufigkeiten für jede Zelle und der bei Unabhängigkeit der beiden Variablen erwarteten absoluten Häufigkeiten für jede Zelle berechnet. Ein bestimmter Chi²-Wert entspricht bei einer gegebenen Anzahl an Freiheitsgraden (die Spalte „df" (degress of freedom) in der Ausgabe) einem bestimmten Signifikanzniveau. Im oben abgebildeten Fall wird p mit ,000 angegeben. Ein Doppelklick auf die Tabelle und die Zelle zeigt, dass es eigentlich 0,000000000162 sind.

Die genaue Interpretation des Signifikanzniveaus lautet wie folgt: Gesetzt den Fall, dass die Nullhypothese (in unserem Beispiel: Es gibt keinen Unterschied in den Ausländeranteilen in Ost und West) in der Grundgesamtheit zutrifft, wie groß ist dann die Wahrscheinlichkeit, eine Stichprobe aus dieser Grundgesamtheit (hier: die Befragten des ALLBUS 2000) zu haben, deren Daten einen so großen (oder größeren) Unterschied zwischen Ost und West zeigen.

Es wird also eine Aussage über die Wahrscheinlichkeit der Daten gemacht unter der Voraussetzung, dass die Nullhypothese zutrifft. Es wird keine Aussage über die Nullhypothese auf der Basis der Daten gemacht! Diese falsche Interpretation ist relativ weit verbreitet. Eine einfache Formulierung wäre die zu sagen, dass Ost und West sich hinsichtlich des Ausländeranteils statistisch signifikant voneinander unterscheiden ($p < ,001$). Der Unterschied beträgt 7,3 Prozentpunkte.

Das Signifikanzniveau wird in der Regel nicht exakt angegeben, sondern es gibt Schwellen, die sich in der Literatur eingebürgert haben: $p < ,05$; $p < ,01$ und $p < ,001$. Alle Analyseergebnisse, die eine Irrtumswahrscheinlichkeit von mehr als 5 % (d. h. $p > ,05$) aufweisen, werden meist als statistisch nicht signifikant (abgekürzt n.s.) angesehen. Dies ist aber nicht mehr als eine Konvention. Es ist mit Sicherheit unsinnig, ein Ergebnis mit $p = ,05$ als statistisch signifikant, ein anderes mit $p = ,051$ aber als nicht signifikant anzusehen. In diesem Fall würde man vielleicht noch die Ergebnisse markieren, bei denen $p < ,06$ oder $p < ,1$ ist. Außerdem ist das Ergebnis eines Signifikanztests von der Stichprobengröße abhängig: Je größer die Stichprobe, desto geringer ist die Irrtumswahrscheinlichkeit. Bei sehr großen Stichproben ist oft jeder noch so geringe Unterschied statistisch signifikant und damit auf diesem Weg nicht mehr interpretierbar. Bei aller Betonung der Ergebnisse von Signifikanztests in der Literatur sollte man also daran denken, dass vor allem von Interesse ist, wie groß der Unterschied zwischen zwei (oder mehreren) Gruppen hinsichtlich des untersuchten Merkmals ist. Im obigen Beispiel ist der Ausländeranteil im Westen mehr als neunmal größer als im Osten (8,2 % zu 0,9 %), was einen beträchtlichen Unterschied

darstellt. Man könnte auch sagen, dass in den alten Bundesländern jeder Zwölfte ein Ausländer ist, und in den neuen Ländern weniger als jeder 110-te.

4.4 Zusammenhangsmaße bei Kreuztabellen

Der Chi²-Test zeigt, ob es einen statistisch signifikanten Zusammenhang zwischen zwei Merkmalen gibt; im Beispiel zwischen der Region (Ost/West) und der Staatsangehörigkeit. Er zeigt aber nicht, wie stark der Zusammenhang zwischen den beiden Variablen ist. Um dies zu quantifizieren, sind Zusammenhangsmaße geeignet. Die am stärksten verbreiteten Maße sind wahrscheinlich Phi für 2 x 2-Tabellen und Cramers V für Tabellen mit mehr als vier Zellen (bei 2 x 2-Tabellen sind beide Maße identisch). Beide Maße liegen zwischen 0 und 1. 0 bedeutet, dass kein Zusammenhang zwischen den beiden Variablen besteht; 1 bedeutet einen perfekten Zusammenhang.

Beide Maße werden über das Unterkommando „stat = phi“ angefordert:

```
cross staat by v3
/cells = cou col
/stat = chisq phi.
```

Der SPSS-Output wird damit um eine weitere Tabelle ergänzt:

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,114	,000
	Cramer's V	,114	,000
N of Valid Cases		3136	

a. Not assuming the null hypothesis.

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

Der Zusammenhang zwischen Region (Ost/West) und Staatsangehörigkeit ist mit Phi = ,114 nur schwach. Dabei muss aber bedacht werden, dass ein perfekter Zusammenhang, also Phi = 1, nur vorliegen kann, wenn beispielsweise im Westen ausschließlich Ausländer und im Osten nur Deutsche lebten.

4.5 Darstellung der Ergebnisse von Kreuztabellierungen

4.5.1 Tabellarische Darstellung

Die SPSS-Ausgabe für Kreuztabellen ist erstens nicht besonders schön, zweitens überfrachtet und drittens in mehrere Einzeltabellen aufgesplittet. In der Regel wird man versuchen, alle relevanten Ergebnisse in einer übersichtlichen Tabelle unterzubringen. Zwei Möglichkeiten zeigen zum Beispiel die folgenden Tabellen:

Tabelle 1: Familienstand in Abhängigkeit von der Staatsangehörigkeit, Anteile in Prozent

	Deutsche	**Nicht-Deutsche**
Verheiratet & zusammenlebend	60,7	65,1
Verheiratet & getrennt lebend	1,4	2,8
Verwitwet	9,3	0,0
Geschieden	6,1	7,4
Ledig	22,6	24,7
N	2.913	215

chi² = 24,37; df = 4; p < ,001; Cramers V = ,088

Die einzelnen Zellen enthalten nicht die Absolutzahlen, sondern nur die prozentualen Anteile (Spaltenprozente) pro Bundesland. Um deutlich zu machen, auf wie viele Fälle sich die Prozentuierungen jeweils beziehen, sollte der Umfang der Teilstichprobe jeweils angegeben werden. Der Chi²-Wert, die Anzahl der Freiheitsgrade, das sich daraus ergebende Signifikanzniveau und das Zusammenhangsmaß können in die Fußzeile der Tabelle geschrieben werden.

Das SPSS-Komando für diese Tabelle lautet:

```
cross v285 by staat
/cells = cou col
/stat = chisq phi.
```

Bei den prozentualen Anteilen ist es meistens völlig ausreichend, eine Nachkommastelle darzustellen. Die Information, die die zweite Dezimalstelle beinhaltet, ist in den meisten Fällen irrelevant und überfrachtet die Tabelle nur.
Der Titel der Tabelle sollte möglichst genau beschreiben, was in der Tabelle abgebildet ist und welche Maße dargestellt sind (Prozentuale Anteile, Mittelwerte, etc.).

Tabelle 2: Besitzverbreitung von Konsumgütern, in Prozent, N in Klammern

	Haupt-schüler	**Realschüler**	**Gymna-siasten**	**Gesamt**	**Signifikanz[a]**
Fernseher	45,6 (2.545)	28,5 (2.367)	31,2 (1.985)	33,4 (6.897)	Chi² = 47,36 p < ,001 Cramers V = ,314
Spielekonsole	26,1 (2.438)	23,7 (2.289)	17,6 (2.011)	22,0 (6.738)	Chi² = 10,91 p < ,01 Cramers V = ,216
Mobiltelefon	87,3 (2.489)	87,4 (2.354)	83,9 (1.879)	86,3 (6.722)	Chi² = 2,63 n.s. Cramers V = ,031
Pkw	1,2 (2.501)	2,5 (2.332)	6,0 (1.853)	2,8 (6.686)	Chi² = 7,30 p < ,05 Cramers V = ,153

[a] Für alle Zeilen gilt: df = 2

Tabelle 2 besteht eigentlich aus vier Kreuztabellen, die zusammengefasst wurden. In diesem fiktiven Beispiel wären die Variablen „Schulform“ und „Besitz eines Fernsehers ja/nein“, „Besitz einer Spielekonsole ja/nein“ usw. einzeln miteinander in Beziehung gesetzt worden. Abgebildet werden dann nur die Anteile derjenigen Fälle, die jeweils die Antwort „ja“ gegeben haben. Der Anteil derjenigen, die die Antwort „nein“ gegeben haben, muss dann die Differenz zu 100 % betragen. Er kann also bei der Darstellung weggelassen werden.
Das Einfügen einer Spalte „Gesamt“, in der die Anteile der Besitzer von Konsumgütern ohne die Trennung nach Schulformen dargestellt wird, ist dann sinnvoll, wenn man deutlich machen will, wie sich die einzelnen Kategorien der unabhängigen Variablen im Vergleich zur gesamten Stichprobe verhalten. Im vorliegenden Beispiel kann man erkennen, dass Hauptschüler mit einem Anteil von 45,6 % Fernseher-Besitzern deutlich über dem durchschnittlichen Anteil von 33,4 % in der Gesamt-Stichprobe liegen.

Da in Tabelle 2 vier Kreuztabellen zusammengefasst sind, gibt es auch vier Signifikanztests und Zusammenhangsmaße, die getrennt dargestellt werden müssen. Informationen, die für alle Zeilen gleich sind, können auch in eine Fußnote geschrieben werden. Im Beispiel ist die Anzahl der Freiheitsgrade für alle Signifikanztests gleich.
Die Größe der Teilstichproben kann sich in den Einzelanalysen unterscheiden, z. B. durch eine unterschiedlich hohe Anzahl an Verweigerungen. Es gilt also für jede Zelle ein anderes N, das wie in Tabelle 2 zusammen mit dem prozentualen Anteil in Klammern angegeben werden kann o.ä. Anregungen dafür, welche Möglichkeiten der Darstellung sinnvoll sein könnten, finden sich in vielen empirischen Analysen in der Literatur.

4.5.2 Graphische Darstellung der Ergebnisse einer Kreuztabelle

In manchen Fällen zeigt die graphische Darstellung der Ergebnisse von Kreuztabellierungen deutlicher, welche Informationen in den Daten stecken, als die tabellarische Darstellung. Sie hat allerdings auch ihre Grenzen: Wenn zu viele Einzelinformationen in einer Graphik untergebracht werden, wird sie häufig unübersichtlich.
Wie in Abschnitt 3.6 schon gezeigt wurde können Graphiken mit SPSS erstellt werden, allerdings genügt das Ergebnis häufig nicht den ästhetischen Ansprüchen:

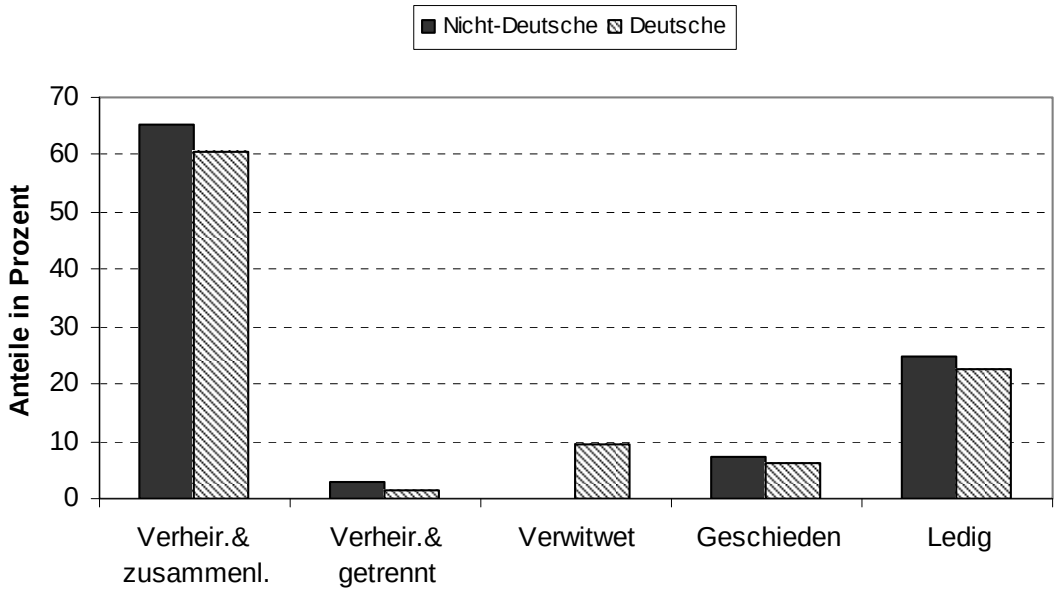
```
graph
/bar=pct by v285 by staat.
```

Bedeutung: Erstelle eine Graphik (“graph”), genauer ein Balkendiagramm („bar“), in dem die prozentualen Anteile („pct by“) der Variablen v285 (Familienstand) in Abhängigkeit von der Variablen staat (Staatsangehörigkeit deutsch ja/nein) abgebildet werden.

Mit Excel-Graphiken hat man noch einige Möglichkeiten mehr, die Abbildung den eigenen Wünschen anzupassen. Um ein Excel-Balkendiagramm zu erstellen, ist es am besten, sich zuerst eine Kreuztabelle ausgeben zu lassen, die nur die prozentualen Anteile und nicht zusätzlich die Absolutzahlen angibt:

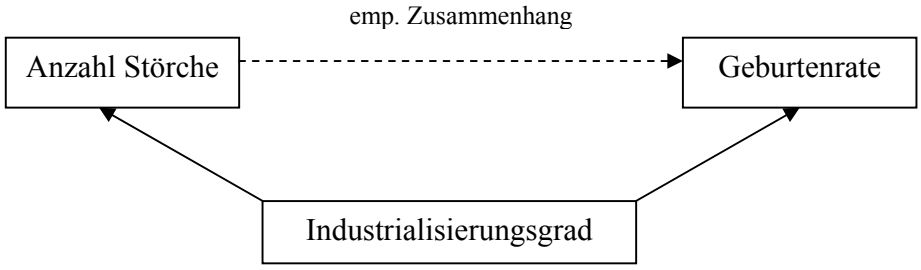
```
cross v285 by staat
/cells=col.
```

Die Kreuztabelle kann dann einfach von SPSS in Excel kopiert werden. Ein auf diese Weise erstelltes Balkendiagramm kann z. B. so aussehen:



4.6 Drittvariablenkontrolle

Wenn ein empirischer Zusammenhang zwischen zwei Variablen vorliegt, muss das nicht bedeuten, dass auch ein kausaler Zusammenhang (tatsächlicher Ursache-Wirkungs-Zusammenhang) besteht. Es kann sich um eine so genannte Scheinkorrelation handeln. Ein Beispiel, das häufig verwendet wird, handelt von Störchen und Geburtenraten: In Ländern, in denen es viele Störche gibt, ist die Geburtenrate in der Regel höher als in Ländern, in denen es wenig Störche gibt. Es gibt also einen empirischen Zusammenhang: Je größer also die Anzahl der Störche, desto höher die Geburtenrate. Daraus zu schließen, dass die Störche die Kinder bringen, die Störche also die Ursache für die Geburtenrate sind (kausaler Zusammenhang), wäre falsch. Es handelt sich um eine „Scheinkorrelation“, die durch eine Variable verursacht wird, die sich sowohl auf die Anzahl der Störche als auch auf die Geburtenrate auswirkt: den Industrialisierungsgrad einen Landes.



Solche Scheinzusammenhänge können bei nominal (oder ordinal) skalierten Variablen durch geschichtete Kreuztabellen nachgewiesen werden.

Tabelle 3: Einschätzung von Abtreibung in Abhängigkeit vom Erwerbsstatus, Anteile in Prozent

Abtreibung, wenn Frau es will …	**Erwerbstätige**	**Nicht-Erwerbstätige**	**Arbeitslose**
… soll möglich sein	47,7	40,2	62,7
… soll nicht möglich sein	52,3	59,8	37,3
N	792	634	67

chi² = 16,66; df = 2; p < ,001; Cramers V = ,106

Tabelle 3 zeigt die Ergebnisse einer Kreuztabellierung. Gefragt wurde, ob Abtreibung (gesetzlich) möglich sein soll, wenn die betroffene Frau eine Abtreibung wünscht. Die Tabelle zeigt, dass Arbeitslose zu einem deutlich höheren Anteil zustimmen als Erwerbstätige und Nicht-Erwerbstätige (Rentner, Hausfrauen, Studenten, etc.)[1]. Man könnte vermuten, dass Arbeitslose aus irgendwelchen Gründen liberaler eingestellt sind, wenn es um Abtreibung geht, also tatsächlich ein kausaler Zusammenhang besteht. Eine andere Möglichkeit ist allerdings, dass es sich hier um eine Scheinkorrelation handelt: Der Arbeitslosenanteil in den neuen Bundesländern ist deutlich höher als in den alten, und in den neuen Bundesländern bestehen aus historischen Gründen weniger Vorbehalte gegen Abtreibung. Wenn das Erhebungsgebiet (Ost/West) in die Analyse einbezogen wird, könnte sich also zeigen, dass es keinen kausalen Zusammenhang zwischen Erwerbsstatus und der Einstellung zur Abtreibung gibt.
Geschichtete Kreuztabellen werden erstellt, indem die Schichtungsvariable an das bekannte Kommando angehängt wird:

```
crosstabs v38 by erwerb_r by v3
/cells=count col
/stat=chi phi.
```

Bedeutung: Erstelle eine Kreuztabelle, in der die Variable v38 (Abtreibung ja/nein) in Abhängigkeit von den Variablen erwerb_r (Erwerbsstatus) und v3 (Erhebungsgebiet: West/Ost) dargestellt wird.

SPSS erstellt eine geschichtete Kreuztabelle, d. h. im Grunde handelt es sich um zwei Kreuztabellen, für die jeweils ein Signifikanztest durchgeführt wird: Eine Tabelle für die alten und eine für die neuen Bundesländer. Wenn es keinen kausalen Zusammenhang zwischen dem Erwerbsstatus und der Einstellung zur Abtreibung gibt, sondern der empirische Zusammenhang nur durch das Erhebungsgebiet zu Stande gekommen ist, dürfte es innerhalb eines Erhebungsgebiets (Alte/Neue Bundesländer) keine statistisch signifikanten Unterschiede zwischen den Ausprägungen des Erwerbsstatus mehr geben. Genau dieses Ergebnis ist in Tabelle 4 dargestellt.

[1] Die Variable lässt sich aus den Informationen der Variablen v234 und v265 errechnen.

Tabelle 4: Einschätzung von Abtreibung in Abhängigkeit vom Erwerbsstatus und dem Erhebungsgebiet, Anteile in Prozent

Abtreibung, wenn Frau es will …	Erwerbstätige	Nicht-Erwerbstätige	Arbeitslose	Signifikanz
Alte Bundesländer				
… soll möglich sein	41,7	35,2	42,3	Chi² = 5,42 n.s. V = ,067
… soll nicht möglich sein	58,3	64,8	57,7	
N	654	534	26	
Neue Bundesländer				
… soll möglich sein	75,5	67,0	75,0	Chi² = 2,28 n.s. V = ,090
… soll nicht möglich sein	24,5	33,0	25,0	
N	139	100	40	

4.7 Ordinales Skalenniveau bei Kreuztabellen

Variablen, die ordinales Skalenniveau haben, können ebenfalls mit Hilfe von Kreuztabellen analysiert werden. Die Syntax unterscheidet sich nicht von der bisher dargestellten. Für ordinalskalierte Daten eigenen sich allerdings andere Zusammenhangsmaße. Das gängigste Maß für den Zusammenhang zweier ordinalskalierter Variablen ist Kendalls Tau-b. Es kann Werte zwischen -1 und 1 annehmen. Ein Wert von -1 würde bedeuten, dass alle Fälle, die bei der einen Variablen den höchsten Wert haben, bei der anderen Variablen den niedrigsten Wert haben und umgekehrt. Allgemein bedeutet ein negatives Tau-b: Je höher der Wert bei a, desto niedriger der Wert bei b. Ein positives Tau-b bedeutet entsprechend: Je mehr a, desto mehr b; ein Tau-b von 0 bedeutet, dass es keinen Zusammenhang zwischen den beiden Variablen gibt. Eine genauere Besprechung des Maßes ist unter www.lrz-muenchen.de/~wlm/ilmes.htm zu finden. Das Lexikon ist auch darüber hinaus empfehlenswert, wenn es um die Erklärung statistischer Begriffe geht.

Ein Zusammenhangsmaß für ordinalskalierte Daten darf nur verwendet werden, wenn beide Variablen mindestens dieses Niveau haben. Setzt man eine nominal- und eine ordinalskalierte Variable in Beziehung, ist Cramers V geeignet.

5 Korrelationen

5.1 Nichtparametrische Korrelationen

Es kann sein, dass für eine Analyse nicht die prozentualen Anteile in jeder Zelle einer Kreuztabelle, sondern nur das Zusammenhangsmaß von Interesse ist. Da bei ordinalskalierten Variablen die Richtung des Zusammenhangs sinnvoll bestimmt und durch negative oder positive Werte des Maßes ausgedrückt werden kann, ist dies oft ausreichend. Tau-b wird in diesem Fall eher selten verwendet, üblicher ist der Korrelationskoeffizient „Spearmans Rho“ (kurz rho). Er kann nicht über das Kommando „crosstabs“ angefordert werden, sondern muss im Rahmen einer nichtparametrischen Korrelation bestimmt werden:

```
nonpar corr altgrup v144.
```

Bedeutung: Führe eine Korrelationsanalyse mit den Variablen alter_r (kategorisiertes Alter) und v144 (Stolz, Deutscher zu sein) durch. Als Standardeinstellung gibt dieser Befehl rho aus. Wenn Kendalls Tau-b angegeben werden soll oder wenn beide Maße angefordert werden sollen, muss das Kommando erweitert werden:

```
nonpar corr altgrup v144
/print=kendall.

bzw.

nonpar corr altgrup v144
/print=both.
```

Das Ergebnis ist eine Korrelationsmatrix, in der jede Variable in die Zeilen und in die Spalten aufgenommen wird. In der Diagonalen stehen die Korrelationen der Variablen mit sich selbst, die logischerweise perfekt und damit gleich 1 sind. Die Werte, die über der Diagonalen stehen, wiederholen sich spiegelverkehrt in der unteren Hälfte, weshalb in der Regel nur das untere Dreieck berichtet wird.

Correlations

			ALTER_R kat. Alter	V144 GENERELLER STOLZ, DEUTSCHER ZU SEIN
Spearman's rho	ALTER_R kat. Alter	Correlation Coefficient	1,000	-,182
		Sig. (2-tailed)	,	,000
		N	3138	1379
	V144 GENERELLER STOLZ, DEUTSCHER ZU SEIN	Correlation Coefficient	-,182	1,000
		Sig. (2-tailed)	,000	,
		N	1379	1379

Wenn bei mehr als zwei Variablen der Zusammenhang zwischen allen Variablen von Interesse ist, können im Kommando „nonpar corr“ alle Variablen nacheinander aufgeführt werden. Ist dagegen nur von Interesse, wie hoch eine Variable mit einer Reihe anderer korreliert, kann die Ausgabe auch verkleinert werden:

```
nonpar corr altgrup with v144 bildung[2].
```

Bedeutung: Erstelle eine Korrelationsmatrix, bei der die Variable „altgrup“ in den Zeilen steht und die Variablen v144 und „bildung“ in den Spalten stehen. Ergebnis:

Correlations

			V144 GENERELLER STOLZ, DEUTSCHER ZU SEIN	BILDUNG Bildungsniveau
Spearman's rho	ALTER_R kat. Alter	Correlation Coefficient	-,182	-,370
		Sig. (2-tailed)	,000	,000
		N	1379	3097

Die Ausgabe kann außerdem dadurch erweitert werden, dass das Signifikanzniveau nicht nur in Zahlen angegeben wird, sondern statistisch signifikante Korrelationen außerdem mit Sternchen markiert werden. Eigentlich sollte dies die Standardeinstellung bei SPSS sein, es scheint dort aber ein Fehler im Programm vorzuliegen.

[2] Die Variable „bildung“ ergibt sich aus der Variable v221 (Allgemeiner Schulabschluss).

```
nonpar corr altgrup with v144 bildung
/print = nosig.
```

Bedeutung: Markiere in der Korrelationsmatrix statistisch signifikante Korrelationen („nosig“ bedeutet eigentlich „no significance level“ oder ähnliches, hier liegt aber der Fehler bei SPSS vor). Ergebnis:

Correlations

				V144 GENERELLER STOLZ, DEUTSCHER ZU SEIN	BILDUNG Bildungsniveau
Spearman's rho	ALTER_R kat. Alter	Correlation Coefficient		-,182**	-,370**
		Sig. (2-tailed)		,000	,000
		N		1379	3097

**. Correlation is significant at the .01 level (2-tailed).

Nichtparametrische Korrelationen müssen außer bei ordinalskalierten Daten auch bei metrischen, aber nicht normalverteilten Daten angewendet werden. Außerdem empfiehlt es sich bei Variablen, die Ausreißer beinhalten (Werte, die extrem niedriger oder höher als alle anderen Werte sind), nichtparametrische Korrelationen durchzuführen.

ACHTUNG: Bei gewichteten Daten sind nichtparametrische Korrelationen nicht geeignet, weil die Gewichtung ignoriert wird.

5.2 Produkt-Moment-Korrelationen

Wenn es sich bei den Variablen, die analysiert werden sollen, um metrisch skalierte, (einigermaßen) normalverteilte oder um dichotome Variablen (Dummy-Variablen) handelt, sind Produkt-Moment-Korrelationen besser geeignet. Bei ihrer Berechnung wird nicht nur die Rangfolge der Werte für die Berechnung des Korrelationskoeffizienten herangezogen, sondern die tatsächlichen Werte.
Das Kommando ist dem vorhergehenden sehr ähnlich:

```
correlations alter aeq v3.

kurz:

corr alter aeq v3.
```

Bedeutung: Korreliere jeder der Variablen mit jeder anderen.

Auch beim Kommando „corr“ können durch das Schlüsselwort „with“ einzelne Variablen ausschließlich in die Zeilen bzw. Spalten der Korrelationsmatrix geschrieben werden. Die Markierung des Signifikanzniveaus kann wie vorher durch das Unterkommando „print = nosig“ angefordert werden.
Die Auswahl verschiedener Korrelationskoeffizienten ist bei Produkt-Moment-Korrelationen nicht möglich. Bei dem ausgegebenen Koeffizienten handelt es sich immer um den Korrelationskoeffizienten nach Pearson ($r_{x,y}$).

5.3 Voraussetzungen

5.3.1 Normalverteilung

Eine Voraussetzung für die Anwendung von Produkt-Moment-Korrelationen ist die (annähernde) Normalverteilung der Variablen. Verletzungen dieser Voraussetzung können zu verzerrten Koeffizienten führen, eine Prüfung ist also sinnvoll.
Die einfachste Möglichkeit hierfür ist, ein Histogramm der Variablen zu erstellen:

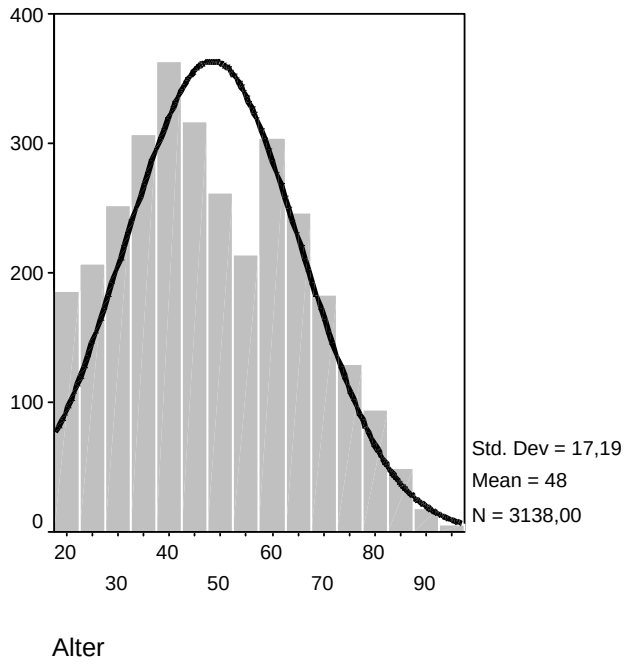
graph /histogram (normal) = alter.

kurz:

graph /hist (nor) = alter.

Bedeutung: Erstelle ein Histogramm der Variablen „alter“, über das eine Normalverteilungskurve gelegt wird.

Das Histogramm zeigt in den seltensten Fällen eine wirkliche Normalverteilung, es gibt fast immer Abweichungen. Ob diese Abweichungen so groß sind, dass die Normalverteilungs-Annahme verletzt ist und entsprechend keine Produkt-Moment-Korrelation berechnet werden sollte, bleibt dem Betrachter überlassen. Ist man sich unsicher, kann man mit der Variablen zur Probe sowohl eine nicht-parametrische Korrelation als auch eine Produkt-Moment-Korrelation berechnen. Je mehr sich die beiden Koeffizienten unterscheiden, desto eher wird eine Verletzung der Normalverteilungs-Annahme vorliegen.



5.3.2 Linearität

Produkt-Moment-Korrelationen messen lineare Zusammenhänge. Das bedeutet, dass für jede Einheit, um die der Wert der einen Variablen ansteigt, der Wert der anderen Variablen um einen konstanten Wert ansteigt. Viele Zusammenhänge sind allerdings nicht-linear. Beispiele hierfür sind kurvilineare Zusammenhänge (quadratische, kubische, etc.) und Schwellenwerte. Ob eine solche Nicht-Linearität vorliegt, ist aus dem Korrelationskoeffizienten nicht zu erkennen.

Die Frage, ob zwei Variablen in einem linearen Zusammenhang stehen, kann mittels eines Streudiagramms untersucht werden:

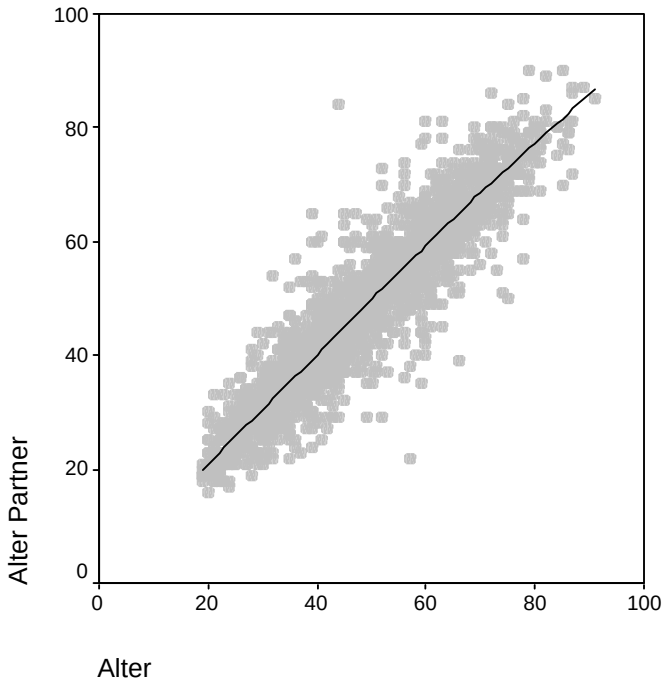
```
graph /scatterplot = alter with alterpar.
```

kurz:

```
graph /scat = alter with alterpar.
```

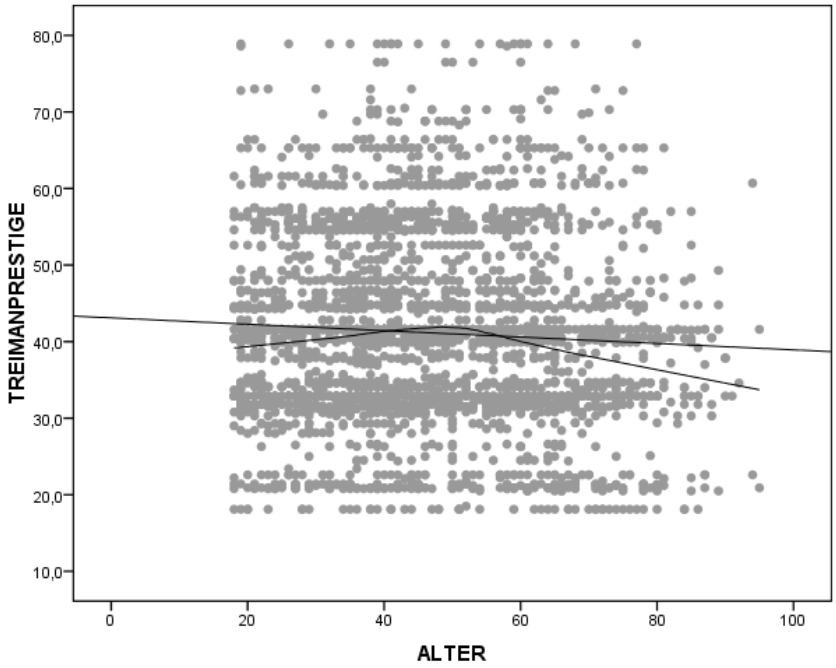
Bedeutung: Erstelle ein Streudiagramm, bei dem die Variable „alter“ auf der X-Achse (vor dem „with“) und die Variable „alterpar“ (Alter des (Ehe-)Partners) auf der Y-Achse (nach dem „with“) abgebildet wird.

Nach einem Doppelklick auf das Diagramm in der Ausgabe und einem weiteren Doppelklick auf den weißen Teil des Diagramms öffnet sich das Fenster „Optionen“ (Scatterplot Options). Die Bestätigung der Option „Anpassungslinie“ (Fit Line) aktiviert den Schalter mit den „Anpassungs-Optionen“ (Fit Options). Hier hat man die Möglichkeit, eine Regressionslinie durch die Punktewolke legen zu lassen. Die letzte Wahlmöglichkeit, die „Lowess-Anpassung“, geht im Gegensatz zu den anderen drei Möglichkeiten nicht von einer bestimmten Art der Anpassung aus, sondern orientiert sich ausschließlich an der Form der Punktewolke. Mit ihr kann also überprüft werden, ob die beiden Variablen (wenigstens annährend) in einem linearen Zusammenhang stehen. Wenn dies so ist, muss die Regressionslinie (ungefähr) eine Gerade darstellen.



Streudiagramme wie das obige, in denen die Richtung und Form eines Zusammenhangs sofort deutlich wird, kommen zwar in Lehrbüchern häufig vor, in der Realität aber selten. Das nächste Beispiel ist realitätsnäher. Hier wird auch deutlich, dass die unreflektierte Anwendung einer Korrelationsanalyse zu falschen Ergebnissen führen kann. Die Abbildung unten zeigt die Lowess-Anpassung (gebogene Linie) und eine lineare Anpassung (durchgehende Linie) für die Variablen Alter und Berufsprestige (Treiman-Skala). Mit der Lowess-Anpassung wird deutlich, dass es zwei unterschiedlich gerichtete Zusammenhänge gibt: Bis zum Alter von ca. 50 Jahren steigt das Berufsprestige mit dem Alter, danach fällt es. Für die Altersgruppe der bis 50-Jährigen liegt also ein positiver Zusammenhang vor, für die Älteren dagegen ein negativer Zusammenhang.

Würde man einen (linearen) Korrelationskoeffizienten berechnen, also eine lineare Anpassung vornehmen, ergäbe sich für die gesamte Stichprobe ein schwacher negativer Zusammenhang, der allerdings nicht der Realität entspricht.



5.4 Exkurs: Auswahl von Fällen

Um einen Zusammenhang wie den zwischen dem Alter und dem Berufsprestige zu beschreiben, kann man (neben der Berechnung von quadratischen oder kubischen Zusammenhängen) die Stichprobe teilen und für beide Hälften eine getrennte Analyse durchführen. Der Selektionsbefehl lautet „select if“:

```
temporary.
select if (alter le 50).
corr alter v825.

kurz:

temp.
select if (alter le 50).
corr alter v825.
```

Bedeutung: Wähle aus allen Fällen im Datensatz diejenigen aus, die die angegebene Bedingung (Alter kleiner/gleich 50) erfüllen. Führe nur mit diesen Fällen eine Korrelationsanalyse zwischen dem Alter und dem Berufsprestige (v825) durch.
Das Kommando „temporary“ zum Anfang ist hierbei unbedingt notwendig. „select if“ allein löscht alle Fälle, die die Bedingung nicht erfüllen. Durch den Zusatz „temporary“ werden die Fälle nur für den nächsten Befehl herausgefiltert, aber nicht gelöscht.

Die oben stehende Korrelation ergibt r = ,075 (p < ,01; N = 1.692) für die höchstens 50-Jährigen. Die Korrelation für die Älteren beträgt r = -,152 (p < ,001; N = 1.302). Für die gesamte Stichprobe beträgt der Korrelationskoeffizient r = -,058 (p < ,01; N = 2.994).

Der Befehl „select if“ ist immer dann notwendig, wenn danach eine Datenanalyse durchgeführt wird. Soll eine Datenmodifikation, also ein „compute“ oder „recode into“, nur auf eine Auswahl von Fällen bezogen werden, lautet das Kommando „do if / end if“. Immer dann, wenn der Befehl durch ein „execute“ abgeschlossen werden muss, ist also „do if / end if“ richtig:

```
do if (v3 = 1).
recode alter (18 thru 44 = 1) (45 thru 59 = 2) (60 thru 95 = 3) (miss = 9) into alter_west.
end if.
exe.
```

Sowohl “do if” als auch “end if” sind einzelne Befehle und müssen entsprechend mit einem Punkt abgeschlossen werden.

Übersicht Zusammenhangsmaße

	Dichotom	***sonstige Nominal***	***Ordinal***	***Intervall***
Dichotom	Phi Wertebereich [0;1]	Cramer's V Wertebereich [0;1]	Cramer's V Wertebereich [0;1]	- Pearson's Korrelationskoeffizient - (Eta) Wertebereich [-1;1]
sonstige Nominal		Cramer's V Wertebereich [0;1]	Cramer's V Wertebereich [0;1]	Eta Mittelwertvergleich
Ordinal			- Spearmans Rho - Kendall's tau-b/tau-c[a] Wertebereich [-1;1]	- Spearmans Rho - Kendall's tau-b/tau-c[a] Wertebereich [-1;1]
Intervall				Pearson's Korrelationskoeffizient[b] Wertebereich [-1;1]

a) tau-b bei zweistufigen Variablen und tau-c bei Variablen mit unterschiedlicher Anzahl von Ausprägungen.
b) Bei nicht-normalverteilten Variablen wird Spearmans Rho berechnet.

5.5 Drittvariablenkontrolle bei Korrelationen

Bei Korrelationen kann die Drittvariablenkontrolle mit Hilfe von Partialkorrelationen durchgeführt werden. Regressionsanalysen sind nichts anderes als eine Reihe von Partialkorrelationen. Auch bei ihnen wird der Zusammenhang zwischen der unabhängigen und der abhängigen Variablen unter Kontrolle des Einflusses der anderen Variablen berechnet.
Zuerst wird logischerweise eine Korrelationsanalyse durchgeführt:

```
corr v261 aeq.
```

Bedeutung: Korreliere die Variablen „v261“ (Dauer der Arbeitslosigkeit in Wochen) und „aeq“ (Äquivalenzeinkommen) miteinander.

Das Ergebnis zeigt, dass die Dauer der Arbeitslosigkeit (in Wochen) und das Äquivalenzeinkommen mit r = -,106 korrelieren. Die Irrtumswahrscheinlichkeit beträgt hier zwar 5,4%, da aber nur 333 Fälle in die Analyse eingegangen sind, könnte dieses Ergebnis auch als statistisch signifikant betrachtet werden. Je länger also jemand in den letzten 10 Jahren arbeitslos war, desto geringer ist zum Zeitpunkt der Befragung das Äquivalenzeinkommen.
Hier könnte es sich allerdings um einen Scheinzusammenhang handeln. Da in den neuen Bundesländern eine höhere Arbeitslosigkeit herrscht und da das Einkommen im Osten durchschnittlich geringer ist als im Westen, könnte der genannte Zusammenhang nicht mehr nachweisbar sein, wenn das Erhebungsgebiet kontrolliert wird.
Der Befehl für diese Partialkorrelation lautet:

```
part corr v261 aeq by v3.
```

Bedeutung: Führe eine Partialkorrelation durch, bei der der Zusammenhang zwischen der Dauer der Arbeitslosigkeit und dem Äquivalenzeinkommen für den Einfluss des Erhebungsgebiets (v3) kontrolliert wird.
Das Ergebnis wird in den SPSS-Versionen bis 11 einschl. in einer altmodischen Form dargestellt:

Korrelationen

Kontrollvariablen			DAUER DER ARBEITSLOSIGKEIT IN WOCHEN	aeq
ERHEBUNGSGEBIET: WEST - OST	DAUER DER ARBEITSLOSIGKEIT IN WOCHEN	Korrelation	1,000	-,072
		Signifikanz (zweiseitig)	.	,193
		Freiheitsgrade	0	330
	aeq	Korrelation	-,072	1,000
		Signifikanz (zweiseitig)	,193	.
		Freiheitsgrade	330	0

Die Korrelation zwischen den beiden Variablen beträgt unter Kontrolle (bei Konstanthaltung) des Erhebungsgebiets nur noch r = -,072 und ist mit p = ,193 deutlich nicht statistisch signifikant. Der eingangs festgestellte Zusammenhang zwischen der Dauer der Arbeitslosigkeit und dem Äquivalenzeinkommen war also eine Scheinkorrelation.

Bei Partialkorrelationen kann nicht nur für eine, sondern auch für mehrere Variablen kontrolliert werden. Die zu kontrollierenden Variablen können in dem entsprechenden Befehl hinter dem „by“ aufgelistet werden.

6 Mittelwertvergleiche

Wenn man zwei oder mehr Gruppen hinsichtlich einer metrisch skalierten Variablen vergleichen will, sind Mittelwertvergleiche angemessen. SPSS bietet für die bivariate Varianzanalyse drei verschiedene Befehle an: t-test, oneway, means. Für multivariate Varianzanalysen (ANOVA = ANalysis Of VAriance) gibt ebenfalls mehrere Kommandos, von denen hier nur „unianova“ beschrieben wird.

6.1 Kommando „t-test groups“

Der Befehl „t-test“ ist nur für den Vergleich zweier Gruppen, also für dichotome unabhängige Variablen geeignet. Für alle anderen empfiehlt sich „oneway“ oder „means“. Der T-Test gibt Aufschluss darüber, wie hoch die Mittelwerte der abhängigen Variablen für die beiden Gruppen der unabhängigen Variablen sind und ob sich diese Mittelwerte statistisch signifikant voneinander unterscheiden.

```
t-test groups = v3 (1,2)
/var = aeq.
```

Bedeutung: Führe einen Mittelwertvergleich für unabhängige Stichproben (groups) durch, dessen unabhängige Variable „v3“ (Erhebungsgebiet West/Ost) und dessen abhängige Variable „aeq“ (Äquivalenzeinkommen) ist. Die Werte der beiden Gruppen der unabhängigen Variablen müssen in Klammern hinter dem Variablennamen angegeben werden.

Die erste Tabelle in der Ausgabe berichtet die Mittelwerte, Standardabweichungen und den Standardfehler der Mittelwerte:

Group Statistics

	V3 ERHEBUNGSGEBIET:	N	Mean	Std. Deviation	Std. Error Mean
AEQ Äquivalenzeinkommen	1 ALTE BUNDESLAENDER	1905	2830,4028	2145,79971	49,16800
	2 NEUE BUNDESLAENDER	458	2209,9377	1007,48470	47,06783

In der zweiten Tabelle kann abgelesen werden, ob sich diese beiden Mittelwerte statistisch signifikant voneinander unterscheiden. Der erste Schritt für diese Prüfung ist der Levene-Test. Mit ihm wird geprüft, ob Varianzgleichheit bei den Werten der abhängigen Variablen für die beiden Gruppen vorliegt. Die Nullhypothese für den Levene-Test lautet, dass Varianzgleichheit besteht. Das Signifikanzniveau gibt die Irrtumswahrscheinlichkeit für H_1 an. Wenn der Signifikanztest also ein Signifikanzniveau von p = ,05 oder weniger angibt, liegt keine Varianzgleichheit vor.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
AEQ Äquivalenzeinkommen	Equal variances assumed	31,806	,000	6,031	2361	,000	620,4651	102,88375	418,71322	822,21699
	Equal variances not assumed			9,116	1554,716	,000	620,4651	68,06521	486,95581	753,97440

Der rechte Teil der Tabelle zeigt die Ergebnisse des T-Tests. Wenn Varianzgleichheit vorliegt, der Levene-Test also ein nicht-signifikantes Ergebnis zeigt, muss das Ergebnis des T-Tests in der oberen Zeile abgelesen werden. Wenn Varianzungleichheit vorliegt, gelten entsprechend die Eintragungen in der unteren Zeile. Im vorliegenden Fall liegt Varianzungleichheit vor. Der T-Wert beträgt 9,116, was bei 1554,716 Freiheitsgraden (df = degress of freedom) mit $p < ,001$ statistisch signifikant ist.

6.2 Kommando „oneway“

Wenn mehr als zwei Gruppen miteinander verglichen werden sollen, kommen die Kommandos „oneway“ oder „means“ in Betracht. Jeder dieser Befehle führt einen Mittelwertvergleich durch. Die Unterschiede zwischen den Kommandos sind die folgenden:

- Post-Hoc-Tests: Können nur bei „oneway“ angefordert werden,
- Ausgabe nicht nur der Mittelwerte, sondern auch anderer Maße wie Median, gruppierter Median, Standardabweichung, Varianz, Schiefe etc.: nur bei „means“,
- Berechnung von eta²: nur bei „means“,

Der Befehl „oneway“ ist ähnlich aufgebaut wie das Kommando „crosstabs“:

```
oneway v261 by bildung.

kurz:

one v261 by bildung.
```

Bedeutung: Führe eine Varianzanalyse (ANOVA) durch, bei der „v261“ (Dauer der Arbeitslosigkeit in Wochen; vor dem Schlüsselwort „by“) die abhängige und „bildung“ (rekodiertes Bildungsniveau) die unabhängige Variable ist.

Die Ausgabe dieser Grundform des Befehls liefert allerdings nur das Ergebnis eines F-Tests:

ANOVA

V261 DAUER DER ARBEITSLOSIGKEIT IN WOCHEN

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	38038,392	2	19019,196	6,498	,002
Within Groups	1036206	354	2927,136		
Total	1074245	356			

Es wird keine Tabelle mit den Mittelwerten der abhängigen Variablen für die Ausprägungen der unabhängigen Variablen ausgegeben. Außerdem ist mit dem statistisch signifikanten Ergebnis des F-Tests nur bekannt, dass sich mindestens zwei der Kategorien der unabhängigen Variablen hinsichtlich ihrer Mittelwerte statistisch signifikant voneinander unterscheiden. Man weiß nicht, wie viele und welche Gruppen dies sind. Um diese beiden Informationen in der Ausgabe zu erhalten, müssen zwei Unterkommandos spezifiziert werden:

```
oneway v261 by bildung
/statistics = descriptives
/posthoc = scheffe.

kurz:

one v261 by bildung
/stat = des
/post = schef.
```

Bedeutung: Gib zusätzlich zum Ergebnis des F-Tests die Mittelwerte (stat = des) und das Ergebnis des Post-Hoc-Tests nach Scheffé (post = schef) aus.

Die erste der beiden jetzt berechneten Tabellen enthält die gewünschten Mittelwerte:

Descriptives

V261 DAUER DER ARBEITSLOSIGKEIT IN WOCHEN

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1 Höchstens HS	117	64,11	59,655	5,509	53,20	75,02	3	323
2 MR	144	57,43	59,096	4,926	47,69	67,16	1	361
3 F/HR	96	38,04	36,034	3,679	30,74	45,35	4	209
Total	357	54,41	54,920	2,906	48,70	60,13	1	361

Die zweite Tabelle enthält das Ergebnis des Post-Hoc-Tests. Mit Hilfe dieser Tests kann geprüft werden, welche der einzelnen Mittelwerte sich statistisch signifikant voneinander unterscheiden. Der Scheffé-Test ist der gebräuchlichste der von SPSS angebotenen Post-Hoc-Tests. Er gilt als relativ konservativ, d. h. die Standardfehler der Mittelwertsdifferenzen müssen im Vergleich zu den absoluten Mittelwertsdifferenzen relativ klein sein, damit der Scheffé-Test ein statistisch signifikantes Ergebnis ausweist. SPSS erlaubt übrigens nicht die Eingabe mehrerer Post-Hoc-Tests in einem Unterkommando. Wenn die Ergebnisse verschiedener Post-Hoc-Tests verglichen werden sollen, muss für jeden Test eine neue Varianzanalyse durchgeführt und jeweils ein Test angefordert werden.

Die Ausgabe des Post-Hoc-Tests besteht aus zwei Tabellen, wobei in der Regel nur die erste Tabelle von Interesse ist (und nicht die Tabelle „Homogene Untergruppen“/“Homogeneous Subsets“):

Post Hoc Tests

Multiple Comparisons

Dependent Variable: V261 DAUER DER ARBEITSLOSIGKEIT IN WOCHEN

Scheffe

(I) BILDUNG Bildungsniveau	(J) BILDUNG Bildungsniveau	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1 Höchstens HS	2 MR	6,68	6,731	,612	-9,86	23,22
	3 F/HR	26,06*	7,448	,002	7,76	44,37
2 MR	1 Höchstens HS	-6,68	6,731	,612	-23,22	9,86
	3 F/HR	19,38*	7,130	,026	1,86	36,91
3 F/HR	1 Höchstens HS	-26,06*	7,448	,002	-44,37	-7,76
	2 MR	-19,38*	7,130	,026	-36,91	-1,86

*. The mean difference is significant at the ,050 level.

In ihr werden die Mittelwerte aller Kategorien der unabhängigen Variablen miteinander verglichen, indem jeweils die Differenz zwischen zwei Mittelwerten gebildet wird. Mit Hilfe der Mittelwertsdifferenz und ihres Standardfehlers wird bestimmt, ob die Differenz statistisch signifikant ist. Da dieser Vergleich für jede mögliche Kombination von Ausprägungen der unabhängigen Variablen gemacht wird, taucht jede Differenz zweimal in der Tabelle auf.

Der oben abgebildete Scheffé-Test zeigt, dass Personen, die höchstens einen Hauptschulabschluss haben, sich nicht statistisch signifikant von denen mit Mittlerer Reife unterscheiden, was die Dauer der Arbeitslosigkeit angeht. Zwischen (höchstens) Hauptschülern und Personen mit (Fach-)Hochschulreife besteht dagegen ein statistisch signifikanter Unterschied. Dasselbe gilt für den Unterschied zwischen denen mit Mittlerer Reife und den (Fach-)Abiturienten. Kurz zusammengefasst unterscheiden sich also (Fach-)Abiturienten statistisch signifikant von den niedriger Gebildeten; diese beiden Bildungsgruppen weisen aber keinen statistisch signifikanten Unterschied auf.

Die Darstellung der Ergebnisse eines Post-Hoc-Tests kann vereinfacht werden, indem in einer Tabelle die sich signifikant unterscheidenden Mittelwerte mit unterschiedlichen Suffixen gekennzeichnet werden:

Tabelle 3: Mittlere Dauer der Arbeitslosigkeit (in Wochen) in Abhängigkeit vom Bildungsniveau*

Bildungsniveau	**Dauer der Arbeitslosigkeit / Wochen**	**N**
Höchstens Hauptschulabschluss	64,11[a]	117
Mittlere Reife / Realschulabschluss	57,43[a]	144
(Fach-)Abitur	38,04[b]	96
Gesamt	54,41	357

* Mittelwerte mit ungleichen Suffixen unterscheiden sich mit $p < ,05$.

6.3 Kommando „means“

6.3.1 Zur Berechnung einer Varianzanalyse

Das Kommando „means“ ist in seiner Grundstruktur eigentlich dafür gedacht, eine Tabelle mit Lagemaßen zu erstellen. Mit der Standardeinstellung werden Mittelwert und Standardabweichung der abhängigen Variablen für jede Kategorie der unabhängigen Variablen ausgegeben, außerdem kann die Anzahl der Fälle (N) pro Kategorie abgelesen werden. Die Durchführung einer Varianzanalyse muss über ein Unterkommando angefordert werden:

```
means v209 by v216
/statistics = anova.

kurz:

means v209 by v216
/stat = anova.
```

Bedeutung: Erstelle eine Tabelle, in der die Mittelwerte der abhängigen Variablen „v209“ (politische Links-Rechts-Selbsteinstufung) für die Ausprägungen der unabhängigen Variablen „v216“ (Geschlecht) berichtet werden. Führe zusätzlich eine Varianzanalyse (ANOVA) durch.

Das Ergebnis besteht aus mehreren Tabellen. Die erste zeigt lediglich, wie viele Fälle in die Analyse eingegangen und wie viele aufgrund fehlender Werte ausgeschlossen worden sind. Die zweite, mit „Bericht“ überschriebene Tabelle, berichtet die Mittelwerte, Standardabweichungen und Anzahl der Fälle:

Report

V209 LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR.

V216 GESCHLECHT,	Mean	N	Std. Deviation
1 MANN	5,11	1400	1,740
2 FRAU	4,95	1395	1,721
Total	5,03	2795	1,732

Männer haben mit einem mittleren Wert von 5,11 auf der Links-Rechts-Skala einen höheren Wert als Frauen. Da der Wert 1 auf dieser Skala „extrem links“ und der Wert 10 „extrem rechts“ bedeutet, sind Männer politisch also etwas mehr „rechts“ eingestellt als Frauen. Da die Mitte der

Skala beim Wert 5,5 liegt, haben sowohl Männer als auch Frauen eine politische Einstellung, die leicht links von der Mitte liegt.
Die nächste Tabelle zeigt das Ergebnis einer Varianzanalyse:

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
V209 LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR. * V216 GESCHLECHT, BEFRAGTE<R>	Between Groups	(Combined)	18,079	1	18,079	6,039	,014
	Within Groups		8361,878	2793	2,994		
	Total		8379,956	2794			

Der Unterschied von 0,16 Punkten auf der Links-Rechts-Skala (5,11 – 4,95) ist statistisch signifikant. Die Irrtumswahrscheinlichkeit beträgt 1,4 % (dasselbe Ergebnis würde natürlich auch ein T-Test bringen, den man bei einer dichotomen unabhängigen Variablen wie dem Geschlecht ebenfalls anwenden kann).

Das Unterkommando „stat = anova“ produziert außer der Varianzanalyse auch die Ausgabe des Zusammenhangsmaßes eta²:

Measures of Association

	Eta	Eta Squared
V209 LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR. * V216 GESCHLECHT, BEFRAGTE<R>	,046	,002

Eta² (nicht eta!) ist ein Zusammenhangsmaß, das zu den PRE-Maßen gehört. PRE-Maße (Proportional Reduction of Error, Verringerung der Fehlerquote) drücken aus, wie gut die Werte der abhängigen Variablen durch die Kenntnis der Werte der unabhängigen Variablen vorhergesagt werden können. Konkret am Beispiel „Geschlecht / Politische Links-Rechts-Einstufung“: Wenn man für einen beliebigen Fall im Datensatz den Skalenwert der politischen Einstellung schätzen müsste, würde man zuerst einmal den Mittelwert der gesamten Stichprobe nehmen (5,03). Damit würde man in einigen Fällen richtig liegen, in den meisten aber falsch. Wenn es nun einen Zusammenhang zwischen dem Geschlecht und der politischen Einstellung gibt, d. h. wenn es einen Unterschied zwischen Frauen und Männern gibt, ist es möglich, die

Schätzung (oder Vorhersage) zu verbessern, indem man für einen Mann den Mittelwert der Gruppe aller Männer nimmt und für eine Frau entsprechend den Mittelwert der Gruppe aller Frauen. Damit würde man mit seiner Schätzung in mehr Fällen richtig liegen, als wenn man den Gesamt-Mittelwert der Stichprobe heranziehen würde. Das Maß, in dem man mit diesen gruppenspezifischen Mittelwerten häufiger richtig liegen würde, wird durch eta² ausgedrückt. Die Vorhersage verbessert sich im obigen Beispiel um 0,2 % (= eta² * 100; das Komma beim Ergebnis für eta² um zwei Stellen nach rechts verschieben).

Je größer die Mittelwertsunterschiede zwischen den Kategorien der unabhängigen Variablen sind und je ähnlicher sich die Fälle innerhalb einer Kategorie sind, desto größer wird eta². Im vorliegenden Fall wird die Fehlerquote bei der Vorhersage nur um 0,2 % verbessert, was sehr wenig ist. Da der Mittelwertsunterschied zwischen Frauen und Männern aber nur 0,16 Skalenpunkte beträgt, ist dies auch nicht anders zu erwarten.

6.3.2 Zur Berechnung von Mittelwerten für Merkmalskombinationen

Mit dem Befehl „means“ können auch die Mittelwerte für Merkmalskombinationen mehrerer unabhängiger Variablen berechnet werden:

```
means v209 by v216 by v3.
```

Bedeutung: Erstelle eine Tabelle, in der die Mittelwerte der Variablen „v209“ (Politische Links-Rechts-Selbsteinstufung) für jede Ausprägung von „v216“ (Geschlecht) und für jede Ausprägung von „v3“ (Erhebungsgebiet: West/Ost) berichtet werden.

Das Ergebnis ähnelt dann einer geschichteten Kreuztabelle:

Report

V209 LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR.

V216 GESCHLECHT,	V3	Mean	N	Std. Deviation
1 MANN	1 ALTE BUNDESLAENDER	5,19	1137	1,746
	2 NEUE BUNDESLAENDER	4,79	263	1,677
	Total	5,11	1400	1,740
2 FRAU	1 ALTE BUNDESLAENDER	5,07	1118	1,739
	2 NEUE BUNDESLAENDER	4,47	277	1,558
	Total	4,95	1395	1,721
Total	1 ALTE BUNDESLAENDER	5,13	2255	1,743
	2 NEUE BUNDESLAENDER	4,63	540	1,624
	Total	5,03	2795	1,732

Die in der Syntax zuerst genannte unabhängige Variable bildet die äußere Schicht, und alle anderen unabhängigen Variablen werden innerhalb der Ausprägungen der äußeren Schicht dargestellt. Die Mittelwerte beziehen sich damit auf die Kombination von Merkmalsausprägungen: Männer in den alten Bundesländern haben einen Skalenmittelwert von 5,19, Männer in den neuen Bundesländern haben einen Mittelwert von 4,79 usw. An der oben abgebildeten Tabelle kann man erkennen, dass:

- Männer etwas „rechter“ sind als Frauen, unabhängig vom Erhebungsgebiet (Männer: 5,11; Frauen: 4,95),
- die Bewohner der alten Bundesländer etwas „rechter“ sind als die Bewohner der neuen Bundesländer (Westen: 5,13; Osten: 4,63),
- Der Unterschied zwischen West und Ost damit größer ist als der Unterschied zwischen den Geschlechtern (Differenz von 0,5 Skalenpunkten beim West-Ost-Vergleich; Differenz von 0,16 Skalenpunkten beim Geschlechtervergleich),
- Männer im Westen am stärksten „rechts“ (oder anders ausgedrückt: am wenigsten „links“) sind,
- Frauen im Osten am meisten „links” sind.

Wird bei mehreren unabhängigen Variablen eine Varianzanalyse angefordert, bezieht sich ihr Ergebnis nur auf die erstgenannte unabhängige Variable. Es kann also mit „means“ nur eine bivariate Analyse durchgeführt werden; für eine multivariate Varianzanalyse ist der Befehl „unianova“ geeignet.

6.3.3 Zur Berechnung anderer Lagemaße

Mit dem Kommando „means“ können nicht nur Mittelwerte bestimmt werden, sondern auch eine Reihe von anderen (Lage-)Maßen: Mediane, gruppierte Mediane, Schiefe, Kurtosis, Minimum, Maximum, prozentualer Anteil der Fälle in den Kategorien der unabhängigen Variablen, etc. Das Kommando

```
means alter by v216
/cells = mean median skew npct.
```

erstellt eine Tabelle, in der der Mittelwert (mean), der Median (median) und die Schiefe (skew) der abhängigen Variablen „alter“ für jede Ausprägung der unabhängigen Variablen „v216“ (Geschlecht) dargestellt werden sollen. Außerdem soll der prozentuale Anteil der Ausprägungen von v216 (Geschlecht) ausgegeben werden.

Die Ausgabe sieht dann folgendermaßen aus:

Case Processing Summary

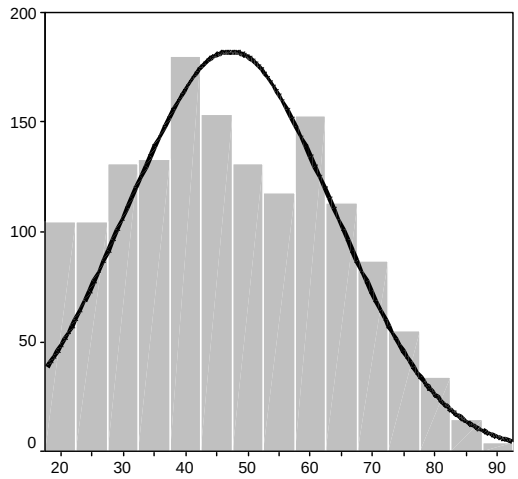
	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
ALTER Alter * V216 GESCHLECHT, BEFRAGTE<R>	3138	100,0%	0	,0%	3138	100,0%

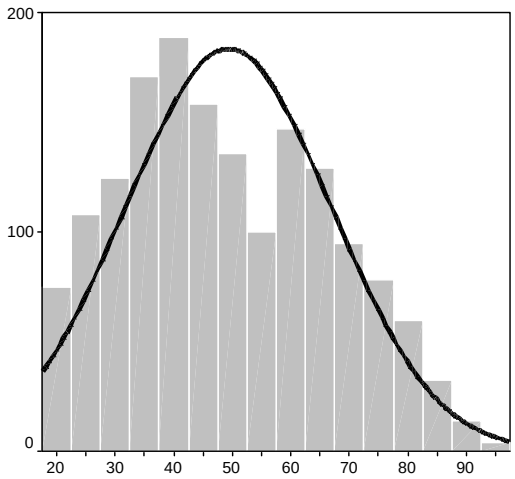
Report

ALTER Alter

V216 GESCHLECHT,	Mean	Skewness	Median	% of Total N
1 MANN	47,19	,199	46,00	48,4%
2 FRAU	49,35	,278	47,00	51,6%
Total	48,31	,253	47,00	100,0%

Die Männer in der Stichprobe sind im Mittel etwas mehr als zwei Jahre jünger als die Frauen. Beide Verteilungen sind leicht rechtsschief (oder linkssteil): Bei beiden Geschlechtern hat die Schiefe einen positiven Wert. Die Histogramme zeigen diese Asymmetrie deutlich: Bei beiden Geschlechtern ragt eine Reihe von Fällen links vom Scheitelpunkt über die Normalverteilungskurve hinaus. Der Mittelwert ist bei rechtsschiefen Verteilungen größer als der Median, wie die Tabelle auch zeigt. Insgesamt sind von den 3.138 Fällen, die in die Analyse eingegangen sind, 48,4 % Männer und 51,6 % Frauen.





Welche Maße mit dem Unterkommando „cells“ angefordert werden können, ist im Handbuch „SPSS-Base“ genau beschrieben.

7 Multivariate Varianzanalyse mit „unianova“

Mit dem Befehl „unianova“ können multivariate Varianzanalysen durchgeführt werden, d. h. es kann geprüft werden, welchen Einfluss mehrere unabhängige Variablen auf den Mittelwert der abhängigen Variablen haben. Dabei wird für jede unabhängige Variable der Einfluss aller anderen UVs kontrolliert (konstant gehalten). Korrelationen zwischen den unabhängigen Variablen werden also „herausgerechnet“, d. h. es wird im Grunde eine Drittvariablenkontrolle durchgeführt, wie sie schon von den geschichteten Kreuztabellen und Partialkorrelationen bekannt ist.

Die Grundform des Kommandos lautet:

```
unianova aeq by v3 staat with v825.

kurz:

uni aeq by v3 staat with v825.
```

Bedeutung: Berechne eine Varianzanalyse, bei der "aeq" (Äquivalenzeinkommen) die abhängige und „v3" (Erhebungsgebiet: West/Ost), „staat" (Staatsangehörigkeit deutsch: ja/nein) und „v825" (Berufsprestige nach Treiman) die unabhängigen Variablen sind. Die unabhängigen Variablen müssen entweder durch das Schlüsselwort „by" oder durch „with" abgetrennt werden: Faktoren, also kategoriale Variablen (nominal und ordinal skalierte Variablen) stehen hinter dem „by"; Kovariaten bzw. kontinuierliche Variablen (metrisch skalierte Variablen) stehen hinter dem „with". Es können auch ausschließlich Faktoren oder ausschließlich Kovariaten als unabhängige Variablen angegeben werden.

Bei gewichteten Datensätzen muss grundsätzlich die Gewichtungsvariable noch einmal in der Syntax angegeben werden! In unserem Fall lautet das Kommando also

```
uni aeq by v3 staat with v825
/regwgt = v836.

kurz:

uni aeq by v3 staat with v825
/reg = v836.
```

Bedeutung: Wie oben; zusätzlich sollen die Fälle mit der Variablen „v836" gewichtet werden.

Die Grundform des Kommandos liefert die folgende Ausgabe:

Univariate Analysis of Variance

Between-Subjects Factors

		Value Label	N
V3 ERHEBUNGSGEBIET: WEST - OST	1	ALTE BUNDESLAENDER	1465
	2	NEUE BUNDESLAENDER	843
STAAT deutsche Staatangehörigkeit	0	nicht deutsch	124
	1	deutsch	2184

Tests of Between-Subjects Effects[b]

Dependent Variable: AEQ Äquivalenzeinkommen

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	958631779[a]	4	239657944,9	66,459	,000
Intercept	2524055,777	1	2524055,777	,700	,403
V825	743682814	1	743682813,8	206,230	,000
V3	9449078,982	1	9449078,982	2,620	,106
STAAT	6522099,663	1	6522099,663	1,809	,179
V3 * STAAT	1628964,552	1	1628964,552	,452	,502
Error	8304819997	2303	3606087,710		
Total	2,635E+10	2308			
Corrected Total	9263451777	2307			

a. R Squared = ,103 (Adjusted R Squared = ,102)

b. Weighted Least Squares Regression - Weighted by V836 PERSONENBEZOGENES OST-WEST-GEWICHT

Die erste Tabelle gibt einen Überblick über die Verteilung der Faktoren: 1.465 Fälle haben die Ausprägung 1 = „Alte Bundesländer“, 843 Fälle die Ausprägung 2 = „Neue Bundesländer“. 2.184 der Befragten haben die deutsche Staatsangehörigkeit, 124 sind Nicht-Deutsche.

Die zweite Tabelle zeigt die Ergebnisse der Varianzanalyse. Das Berufsprestige (v825) hat einen statistisch signifikanten Einfluss auf das Äquivalenzeinkommen, alle anderen Variablen sowie die Konstante (Intercept) dagegen nicht. Außerdem wurde eine Interaktion zwischen den beiden Faktoren (v3 * staat) in das Modell aufgenommen, die nicht statistisch signifikant ist.
In der Fußnote der Tabelle wird das korrigierte r^2 (nicht nur r^2!) angegeben (Englisch: adjusted r squared). Die unabhängigen Variablen erklären 10,2 % der Varianz des Äquivalenzeinkommens.

Diese Ausgabe bringt zwei Probleme mit sich: Erstens ist zwar bekannt, welche der unabhängigen Variablen einen statistisch signifikanten Einfluss auf die abhängige Variable haben – aber nicht, wie groß dieser Einfluss ist. Zweitens sollen die Interaktionen zwischen den Faktoren unter Umständen gar nicht mit ins Modell aufgenommen werden; die Standardeinstellung des Befehls „unianova“ sieht dies aber vor.

Um diese beiden Probleme zu lösen, sind zwei Unterkommandos notwendig:

```
uni aeq by v3 staat with v825
/reg = v836
/print = parameter
/design = intercept v3 staat v825.

kurz:

uni aeq by v3 staat with v825
/reg = v836
/print = par
/des = intercept v3 staat v825.
```

Bedeutung: Wie oben; das Unterkommando „print = parameter“ gibt an, dass eine zusätzliche Tabelle mit den Koeffizienten für die UVs ausgegeben werden soll. Das Unterkommando „design = intercept v3 staat v825“ bedeutet, dass nur eine Konstante und die Haupteffekte der genannten drei Variablen berechnet werden sollen. Wenn (bei mehreren Variablen) eine bestimmte Interaktion in das Modell aufgenommen werden soll, kann dies z. B. durch „des = intercept v3 staat v825 v3*staat“ angefordert werden. Es sind aber nur Interaktionseffekte zwischen Faktoren (kategorialen Variablen) möglich. Das Schlüsselwort „intercept“ im Unterbefehl „design“ darf nicht abgekürzt werden.

Die Ausgabe für die o.a. Syntax liefert die folgende Ausgabe (die erste Tabelle wird hier nicht noch einmal abgebildet):

Tests of Between-Subjects Effects[b]

Dependent Variable: AEQ Äquivalenzeinkommen

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	957002815[a]	3	319000938,3	88,483	,000
Intercept	33800687,5	1	33800687,47	9,375	,002
V3	112059480	1	112059479,7	31,082	,000
STAAT	23265611,6	1	23265611,65	6,453	,011
V825	742055684	1	742055684,4	205,828	,000
Error	8306448962	2304	3605229,584		
Total	2,635E+10	2308			
Corrected Total	9263451777	2307			

a. R Squared = ,103 (Adjusted R Squared = ,102)

b. Weighted Least Squares Regression - Weighted by V836 PERSONENBEZOGENES OST-WEST-GEWICHT

Parameter Estimates[b]

Dependent Variable: AEQ Äquivalenzeinkommen

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	388,141	156,076	2,487	,013	82,077	694,205
[V3=1]	563,182	101,016	5,575	,000	365,090	761,275
[V3=2]	0[a]	,	,	,	,	,
[STAAT=0]	-411,161	161,853	-2,540	,011	-728,555	-93,768
[STAAT=1]	0[a]	,	,	,	,	,
V825	46,467	3,239	14,347	,000	40,116	52,819

a. This parameter is set to zero because it is redundant.

b. Weighted Least Squares Regression - Weighted by V836 PERSONENBEZOGENES OST-WEST-GEWICHT

Der Interaktionseffekt zwischen v3 und staat ist nicht mehr im Modell enthalten. Durch seine Entfernung sind jetzt auch die Konstante und die Einflüsse von „staat“ und „v3“ deutlich statistisch signifikant. Die Erklärungskraft des Modells ist mit r^2 (korr.) = ,102 gleich geblieben.

Die Parameterschätzer in der zweiten Tabelle zeigen, wie stark der Einfluss der UVs auf das Äquivalenzeinkommen ist. Die Konstante (intercept) gibt den Wert der abhängigen Variablen für den Fall an, dass alle kategorialen UVs den Wert der Referenzkategorie und alle metrischen UVs den Wert 0 haben. Im vorliegenden Fall kann die Konstante nicht interpretiert werden, da die

Variable v825 (Berufsprestige) keinen sinnvollen Nullpunkt hat (es gibt keinen Beruf mit dem Prestigewert 0 auf der Treiman-Skala). Um diesen Missstand zu beheben, wurde für die nächste Tabelle das Berufsprestige zentriert, d.h. der Mittelwert berechnet und dieser Mittelwert von jedem individuellen Wert abgezogen (neuer Variablenname: v825_2). Positive Werte in dieser Variablen „v825_2" bedeuten dann, dass die Person ein höheres Berufsprestige hat als der Durchschnitt der Stichprobe; negative Werte bedeuten, dass die Person ein geringeres Berufsprestige hat als der Durchschnitt. Je höher also der Wert von „v825_2", desto mehr liegt das Berufsprestige einer Person über dem Durchschnitt.

Das Ergebnis der Varianzanalyse mit der neuen Variablen unterscheidet sich größtenteils nicht von dem ersten Ergebnis. Das einzige, was sich verändert, ist die Konstante:

Parameter Estimates[b]

Dependent Variable: AEQ Äquivalenzeinkommen

					95% Confidence Interval	
Parameter	B	Std. Error	t	Sig.	Lower Bound	Upper Bound
Intercept	2313,568	89,721	25,786	,000	2137,625	2489,510
[V3=1]	563,182	101,016	5,575	,000	365,090	761,275
[V3=2]	0[a]	,	,	,	,	,
[STAAT=0]	-411,161	161,853	-2,540	,011	-728,555	-93,768
[STAAT=1]	0[a]	,	,	,	,	,
V825_2	46,467	3,239	14,347	,000	40,116	52,819

a. This parameter is set to zero because it is redundant.

b. Weighted Least Squares Regression - Weighted by V836 PERSONENBEZOGENES OST-WEST-GEWICHT

Wenn alle unabhängigen Variablen entweder bei ihrer Referenzkategorie konstant gehalten werden (gilt für kategoriale Variablen) oder den Wert 0 haben (gilt für metrische Variablen), beträgt das Äquivalenzeinkommen 2.313,57 €. Die Referenzkategorie ist diejenige, für die kein Koeffizient, sondern der Eintrag „0^{a}" in der Tabelle zu sehen ist. Bei v3 ist die Referenzkategorie „Neue Bundesländer" (v3 = 2) und bei staat „deutsch" (staat = 1). Der Wert 0 bedeutet bei der zentrierten metrischen Variablen v825_2 „durchschnittliches Berufsprestige". Eine Person, die also in den neuen Bundesländern wohnt, deutsch ist und ein durchschnittliches Berufsprestige hat, verfügt über ein Äquivalenzeinkommen von 2.313,57 €.

Die mit „B" überschriebenen (unstandardisierten) Regressionskoeffizienten geben an, um wie viel sich der Wert der abhängigen Variablen ändert, wenn die jeweilige unabhängige Variable um eine Einheit verändert wird. Mit jedem Punkt auf der Berufsprestige-Skala steigt das

Äquivalenzeinkommen also um 46,47 €. Anders ausgedrückt: Eine Person, die 50 Punkte auf der Berufsprestige-Skala hat, hat (bei gleicher Staatsangehörigkeit und im gleichen Erhebungsgebiet) ein um rund 46 € höheres Äquivalenzeinkommen als eine Person mit einem Skalenwert von 49.
Die Koeffizienten der dichotomen Variablen (Dummy-Variablen) müssen auf die Referenzkategorie bezogen werden. Der Koeffizient für v3 bedeutet, dass Bewohner der alten Bundesländer (v3 = 1) ein um 563,18 € höheres Äquivalenzeinkommen haben als die Bewohner der neuen Bundesländer (die Referenzkategorie). Der Koeffizient für v3 bedeutet entsprechend, dass Nicht-Deutsche (staat = 0) ein um 411,16 € geringeres Äquivalenzeinkommen haben als Deutsche (die Referenzkategorie). Beides gilt „unter sonst gleichen Umständen", d. h. für den Fall, dass der Wert aller anderen Variablen nicht verändert wird.

Wenn außerdem das Zusammenhangsmaß eta² für jede UV ausgegeben werden soll, ist eine Erweiterung des Unterkommandos „print" notwendig:

```
uni aeq by v3 staat with v825
/reg = v836
/print = par etasq
/design = intercept v3 staat v825.

kurz:

uni aeq by v3 staat with v825
/reg = v836
/print = par eta
/des = intercept v3 staat v825.
```

Bedeutung: Wie oben; zusätzlich wird eta² ausgegeben, d. h. der prozentuale Anteil der Gesamtvarianz der abhängigen Variablen, der durch die jeweilige unabhängige Variable erklärt wird.

In der Ausgabe werden die beiden Tabellen, die das Ergebnis der Varianzanalyse und die Parameterschätzer wiedergeben, je um eine Spalte erweitert:

Tests of Between-Subjects Effects[b]

Dependent Variable: AEQ Äquivalenzeinkommen

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	957002815[a]	3	319000938,3	88,483	,000	,103
Intercept	33800687,5	1	33800687,47	9,375	,002	,004
V3	112059480	1	112059479,7	31,082	,000	,013
STAAT	23265611,6	1	23265611,65	6,453	,011	,003
V825	742055684	1	742055684,4	205,828	,000	,082
Error	8306448962	2304	3605229,584			
Total	2,635E+10	2308				
Corrected Total	9263451777	2307				

a. R Squared = ,103 (Adjusted R Squared = ,102)

b. Weighted Least Squares Regression - Weighted by V836 PERSONENBEZOGENES OST-WEST-GEWICHT

Das Berufsprestige erklärt 8,2 % der Gesamtvarianz des Äquivalenzeinkommens. Es ist damit wesentlich bedeutsamer für die Höhe des Äquivalenzeinkommens als die anderen beiden Variablen, die nur 1,3 % (Erhebungsgebiet) bzw. 0,3 % (Staatsangehörigkeit) der Varianz erklären.

Ein abschließender Hinweis: Multivariate Varianzanalysen sind mathematisch in multivariate lineare Regressionen überführbar. Die gezeigten Ergebnisse sind also dieselben, wenn eine Regressionsanalyse durchgeführt wird – mit einer Ausnahme: Die Konstante bei einer linearen Regression gibt den Wert der abhängigen Variablen an, wenn alle unabhängigen Variablen den Wert 0 haben (auch die kategorialen). Um die Konstante bei einer Regression interpretieren zu können, muss also auch bei Dummies die 0 ein sinnvoller Wert sein. Am besten ist es, alle Dummy-Variablen in 0/1-kodierte Variablen zu transformieren.

8 Skalenbildung

Die Messung von theoretischen Konstrukten wird nur teilweise durch Indikatoren vorgenommen, die auf einer einzigen Frage (im Fragebogen oder Interview) beruhen. Indikatoren wie das Geschlecht, das Geburtsjahr o. ä. können eindeutig mit genau einer Frage erhoben werden. Vor allem bei Einstellungen und Präferenzen werden im Gegensatz dazu meist mehrere Fragen gestellt, deren Antworten anschließend zu einem Index zusammengefasst werden. Der Gedanke bei diesem Verfahren ist, dass Einstellungen und Präferenzen etc. meist so komplex sind, dass sie mit nur einer Frage nicht reliabel erhoben werden können. Es wäre z. B. nicht ausreichend, die Einstellung der Befragten zu erwerbstätigen Frauen nur mit einer Frage wie „Finden Sie, dass Frauen erwerbstätig sein sollten?“ zu erheben. Die Befragten können entweder an alleinstehende oder an verheiratete Frauen mit einem erwerbstätigen Mann denken; sie können an Mütter denken oder an Frauen ohne Kinder; sie können an Mütter von Kleinkindern oder an Mütter von erwachsenen Kindern denken etc. Die Befragten würden also unter Umständen auf völlig unterschiedliche Fragen antworten, obwohl sie dem Wortlaut nach dieselbe Frage gestellt bekommen. Um dies zu vermeiden, werden mehrere präzise formulierte Fragen stellt. Der ALLBUS 2000 gibt zum Thema „Erwerbstätige Frauen“ folgende Aussagen vor, für die die Befragten (auf einer Ordinalskala) angeben müssen, inwieweit sie ihnen zustimmen:

- Eine berufstätige Mutter kann ein genauso herzliches und vertrauensvolles Verhältnis zu ihren Kindern finden wie eine Mutter, die nicht berufstätig ist.
- Für eine Frau ist es wichtiger, ihrem Mann bei seiner Karriere zu helfen, als selbst Karriere zu machen.
- Ein Kleinkind wird sicherlich darunter leiden, wenn seine Mutter berufstätig ist.
- Es ist für alle Beteiligten viel besser, wenn der Mann voll im Berufsleben steht und die Frau zu Hause bleibt und sich um den Haushalt und die Kinder kümmert.
- Es ist für ein Kind sogar gut, wenn seine Mutter berufstätig ist und sich nicht nur auf den Haushalt konzentriert.
- Eine verheiratete Frau sollte auf eine Berufstätigkeit verzichten, wenn es nur eine begrenzte Anzahl von Arbeitsplätzen gibt, und wenn ihr Mann in der Lage ist, für den Unterhalt der Familie zu sorgen.

Die numerischen Werte der vier Antwortmöglichkeiten (1 = Stimme voll und ganz zu, 2 = Stimme eher zu, 3 = Stimme eher nicht zu, 4 = Stimme überhaupt nicht zu) werden addiert und so zu einer Gesamt-Skala zusammengefasst. Obwohl die einzelnen Items ordinal skaliert sind, geht man bei diesem Skalentyp, der sogenannten Likert-Skala, davon aus, dass die Gesamt-Skala metrisch skaliert ist. Beim Zusammenfassen von Items zu einer Skala muss darauf geachtet werden, dass die Items dieselbe Anzahl und Abstufungen von Antwortkategorien haben.

8.1 Reliabilitätsanalyse

Bevor ein solcher Index allerdings überhaupt berechnet wird, muss überprüft werden, ob die einzelnen Items tatsächlich in der Form „zusammengehören", dass sie ein einziges gemeinsames theoretisches Konstrukt abbilden. Es besteht die Möglichkeit, dass eines oder mehrere der Items so formuliert waren, dass sie inhaltlich in eine andere Richtung zielen als alle anderen Items. Im oben genannten Beispiel würde die Antwort auf eine Aussage wie „Eine Frau muss im Beruf mehr leisten als ein Mann, um genauso viel zu verdienen" wahrscheinlich nicht dasselbe Konstrukt abbilden wie die anderen sechs.

Ob die Items, die zu einer Skala gehören sollen, tatsächlich eine gemeinsame Skala darstellen, kann mit Hilfe einer Reliabilitätsanalyse festgestellt werden. Im Grunde wird bei dieser Analyse geprüft, wie hoch jedes einzelne Item mit allen anderen und mit der gesamten Skala korreliert. Die Grundform des Kommandos lautet:

```
reliability variables = v10 v11 v12 v13 v14 v15.

kurz:

rel var = v10 v11 v12 v13 v14 v15.
```

Bedeutung: Führe eine Reliabilitätsanalyse mit den Variablen v10, v11, v12, v13, v14 und v15 durch. Wenn die Variablen im Datensatz aufeinander folgen, kann die Variablenliste auch durch „v10 to v15" angekürzt werden.

Die Grundform des Befehls liefert folgendes Ergebnis:

Warnungen

Es wurde kein Unterbefehl SCALE gefunden. Die Skalierung erfolgt für alle angegebenen Variablen.

Skala: ALLE VARIABLEN

Zusammenfassung der Fallverarbeitung

		Anzahl	%
Fälle	Gültig	2795	89,1
	Ausgeschlossen[a]	343	10,9
	Insgesamt	3138	100,0

a. Listenweise Löschung auf der Grundlage aller Variablen in der Prozedur.

Reliabilitätsstatistiken

Cronbachs Alpha	Anzahl der Items
,244	6

Die Fehlermeldung zum Anfang der Ausgabe hat keine Bedeutung. Sie sagt einfach nur, dass ein überflüssiges Unterkommando (scale) nicht angegeben wurde. Das Ergebnis der Reliabilitätsanalyse wird dadurch nicht beeinflusst.

Daran schließt sich die Liste der Variablennamen und -labels an. Unter der Überschrift „Reliabilitätsstatistiken“ ist die hier interessierende Information angegeben: der Wert für (Cronbachs) Alpha. Neben diesem Gütemaß (die Standardeinstellung) können auch andere angefordert werden; Cronbachs Alpha ist allerdings das am häufigsten verwendete Maß. Cronbachs Alpha (kann auch einfach mit α bezeichnet werden) kann bei perfekter Konsistenz der Skala den Maximalwert 1 erreichen, negative Werte sind möglich. Werte ab ,8 werden meist als ausreichend bezeichnet, in der Praxis sieht man allerdings auch häufig Skalen, deren α nur bei ,7 liegt. Da Cronbachs Alpha mit der Anzahl der Items, mit denen eine Reliabilitätsanalyse durchgeführt wird, steigt, ist es teilweise schwierig zu entscheiden, ob eine Skala „gut“ ist.

Im vorliegenden Fall ist Alpha mit ,244 viel zu niedrig. Das liegt, wenn man sich die Variablen näher ansieht, die in die Reliabilitätsanalyse eingegangen sind, aber daran, dass Items falsch

gepolt sind, d. h. dass einige Items inhaltlich in die entgegengesetzte Richtung zu den anderen gehen.

Die Items v10 und v14 sind so formuliert, dass jemand, der ihnen zustimmt, die Erwerbstätigkeit von Frauen positiv bewertet. Die anderen vier Items sind so formuliert, dass Zustimmung bedeutet, dass der/die Befragte gegen die Erwerbstätigkeit von Frauen ist. Die Variablen v10 und v14 müssen also inhaltlich umgedreht werden, d. h. so umkodiert werden, dass die bisher höchstens Werte die niedrigsten werden und umgekehrt (oder die anderen vier Items werden umgedreht):

```
recode v10 (1 = 4) (2 = 3) (3 = 2) (4 = 1) (miss = 9) into v10_2
/v14 (1 = 4) (2 = 3) (3 = 2) (4 = 1) (miss = 9) into v14_2.
exe.

var lab v10_2 “v10 rekodiert (Berufst. Frau: Herzl. Verhältnis zum Kind)“
/v14_2 “v14 rekodiert (Frau, bessere Mutter bei Berufst.)“.
val lab v10_2 v14_2
  1 “Stimme voll und ganz zu“
  2 “Stimme eher zu“
  3 “Stimme eher nicht zu“
  4 “Stimme gar nicht zu“
  9 “wn/kA“.
miss val v10_2 v14_2 (9).
```

Eine neue Reliabilitätsanalyse mit der Syntax

```
rel var = v10_2 v11 v12 v13 v14_2 v15.
```

führt zu folgendem Ergebnis:

Warnungen

Es wurde kein Unterbefehl SCALE gefunden. Die Skalierung erfolgt für alle angegebenen Variablen.

Skala: ALLE VARIABLEN

Zusammenfassung der Fallverarbeitung

		Anzahl	%
Fälle	Gültig	2795	89,1
	Ausgeschlossen[a]	343	10,9
	Insgesamt	3138	100,0

a. Listenweise Löschung auf der Grundlage aller Variablen in der Prozedur.

Reliabilitätsstatistiken

Cronbachs Alpha	Anzahl der Items
,779	6

Cronbachs Alpha ist mit ,779 zwar relativ hoch, es verfehlt aber die o. a. ,8-Grenze. Um die Skala zu verbessern, gibt es unter Umständen die Möglichkeit, einzelne Items aus der Skala zu entfernen. Es ist wie oben beschrieben möglich, dass einer oder mehrere der Items nicht zu den anderen passen, also nicht dasselbe Konstrukt abbilden. Für diese Prüfung ist ein Unterkommando notwendig:

```
rel var = v10_2 v11 v12 v13 v14_2 v15
/summary = total.

kurz:

rel var = v10_2 v11 v12 v13 v14_2 v15
/sum = tot.
```

Damit wird die Ausgabe erweitert:

Zusammenfassung der Fallverarbeitung

		Anzahl	%
Fälle	Gültig	2795	89,1
	Ausgeschlossen[a]	343	10,9
	Insgesamt	3138	100,0

a. Listenweise Löschung auf der Grundlage aller Variablen in der Prozedur.

Reliabilitätsstatistiken

Cronbachs Alpha	Anzahl der Items
,779	6

Item-Skala-Statistiken

	Skalenmittelwert, wenn Item weggelassen	Skalenvarianz, wenn Item weggelassen	Korrigierte Item-Skala-Korrelation	Cronbachs Alpha, wenn Item weggelassen
"v10 rekodiert (Berufst. Frau: Herzl. Verhältnis zum Kind)"	13,2211	12,894	,431	,767
FRAU, LIEBER MANN BEI D.KARRIERE HELFEN?	13,7649	12,191	,460	,762
FRAU, NICHT ARBEITEN BEI KLEINKIND?	14,3893	11,269	,541	,742
FRAU, ZU HAUSE BLEIBEN+KINDER VERSORGEN?	13,9581	10,526	,677	,704
"v14 rekodiert (Frau, bessere Mutter bei Berufst.)"	14,0351	12,083	,496	,753
FRAU, NACH HEIRAT ARBEITSPL. FREIMACHEN?	13,7871	11,213	,551	,739

Es kommen vier Spalten dazu, die unterschiedliche Informationen liefern: Den Skalenmittelwert, wenn das Item aus der Skala entfernt wird (Scale Mean if Item Deleted), die Varianz der Skala, wenn das Item aus der Skala entfernt wird (Scale Variance if Item Deleted), den Trennschärfekoeffizienten (Corrected Item-Total Correlation) und den Wert für Alpha, wenn das Item aus der Skala entfernt wird (Alpha if Item Deleted).

Vor allem die letzten beiden Spalten sind von Interesse: „Alpha if Item Deleted" gibt an, wie hoch Alpha wäre, wenn der Item in der jeweiligen Zeile nicht in der Skala enthalten wäre. Steht hier ein Wert, der höher ist als das derzeitige Alpha von, 779, sollte die Skala ohne diesen Item gebildet werden. Im vorliegenden Fall gibt es kein Item, dessen Entfernen die Konsistenz der Skala verbessern würde.

Die zweite interessante Information ist der Trennschärfekoeffizient (Corrected Item-Total Correlation). Items mit hoher Trennschärfe unterscheiden (trennen) Befragte mit hoher Merkmalsausprägung gut von Befragten mit niedriger Merkmalsausprägung, können also gut zwischen den Befragten differenzieren. Der Trennschärfekoeffizient liegt zwischen -1 und 1. Er sollte größer als ,4 sein. Wenn der Koeffizient negativ ist, bedeutet dies meistens, dass das Item inhaltlich in die entgegengesetzte Richtung zu den anderen geht und damit umkodiert werden muss. Weist das Item nach dem Umkodieren immer noch einen geringen Trennschärfekoeffizienten (zwischen 0 und ,4) auf, muss es aus der Skala entfernt werden. Im obenstehenden Beispiel gibt es kein Item mit einer zu geringen Trennschärfe. Die Skala sollte also aus den sechs Items gebildet werden, mit denen die Reliabilitätsanalyse durchgeführt wurde.

8.2 Berechnung der Skala und Problembehandlung

Wenn feststeht, dass die Items ein gemeinsames theoretisches Konstrukt abbilden, können sie einfach durch Addition zu einer Skala zusammengefasst werden:

```
comp frau = v10_2 + v11 + v12 + v13 + v14_2 + v15.
exe.

var lab frau "Einstellung Frauen/Berufstätigkeit".
```

Eine Häufigkeitsauszählung dieser neuen Skala bringt dann das folgende Ergebnis:

FRAU Einstellung Frauen/Berufstätigkeit

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	6	18	,6	,6	,6
	7	24	,8	,9	1,5
	8	47	1,5	1,7	3,2
	9	58	1,9	2,1	5,3
	10	87	2,8	3,1	8,4
	11	128	4,1	4,6	13,0
	12	190	6,0	6,8	19,9
	13	193	6,1	6,9	26,8
	14	228	7,3	8,2	35,0
	15	248	7,9	8,9	43,9
	16	229	7,3	8,2	52,2
	17	239	7,6	8,6	60,8
	18	236	7,5	8,5	69,3
	19	238	7,6	8,6	77,9
	20	176	5,6	6,3	84,2
	21	174	5,5	6,3	90,5
	22	107	3,4	3,9	94,3
	23	94	3,0	3,4	97,7
	24	64	2,0	2,3	100,0
	Total	2777	88,5	100,0	
Missing	System	361	11,5		
Total		3138	100,0		

In dieser Tabelle sind drei Probleme zu erkennen: Erstens reicht die Range der Skala von 6 bis 24. Das ist logisch, weil sechs Items addiert worden sind, die die Werte 1 bis 4 haben. Es wäre aber schöner, eine Skala zu haben, die bei dem Wert 0 anfängt (siehe Kap. Multivariate Varianzanalyse). Zweitens ist nicht klar, was niedrige und hohe Werte bei dieser Skala bedeuten. Drittens gibt es für 11,5 % der Befragten keinen Skalenwert, weil sie mindestens einen der Items nicht beantwortet haben (zur Erinnerung: Der Befehl „compute“ berechnet nur dann ein Ergebnis, wenn alle Variablen einen gültigen Wert aufweisen).

8.2.1 Anpassung der Skalenrange und Vergabe von Wertelabels

Die Skalenrange wird einfach angepasst, indem der niedrigst-mögliche Wert vom Skalenwert abgezogen wird:

```
comp frau = frau – 6.
exe.
```

Damit reichen die Werte der Skala nun von 0 bis 18. Was die Werte 0 und 18 als Endpunkte der Skala bedeuten, muss aus der Kodierung und dem Inhalt der Items abgelesen werden. Die Items v11, v12, v13 und v15 wurden in ihrer ursprünglichen Kodierung verwendet. Sie beinhalten alle Aussagen, die gegen die Berufstätigkeit von Frauen sprechen, z.B. „Für eine Frau ist es wichtiger, ihrem Mann bei seiner Karriere zu helfen, als selbst Karriere zu machen“. Die Items sind so kodiert, dass niedrige Werte bedeuten, dass ein Befragter zustimmt, und hohe Werte bedeuten eine Ablehnung der Aussage. Wer also einen hohen Wert bei diesen Items und damit auch bei der gesamten Skala hat, lehnt eine Aussage gegen die Berufstätigkeit von Frauen ab; er oder sie stimmt der Berufstätigkeit von Frauen zu. Hohe Werte bedeuten demnach, dass der Befragte eine progressive Einstellung zur Erwerbstätigkeit von Frauen hat; niedrige Werte bedeuten entsprechend eine traditionelle Einstellung.
(Wenn Unsicherheit über die Bedeutung der Skalenendpunkte besteht, hilft es häufig, die Skala mit einer Variablen zu korrelieren, bei der man sich sicher ist, welches Vorzeichen der Koeffizient haben müsste. Die vorliegende Skala korreliert z. B. mit r = -,300 mit dem Alter. Da man erwarten kann, dass mit steigendem Alter eher eine traditionellere Einstellung zu Frauen und Erwerbstätigkeit einhergeht und die Korrelation negativ ist, müssen höhere Werte bei der Skala „nicht-traditionell“ bzw. progressiv bedeuten.)

Die Endpunkte der Skala sollten mit Wertelabels versehen werden, um bei späteren Analysen Missverständnissen vorzubeugen:

```
val lab frau 0 “traditionelle Einstellung“ 18 “progressive Einstellung“.
```

8.2.2 Ersetzung von fehlenden Werten

Bei 11,5 % fehlenden Werten würde ein relativ großer Teil der Befragten aus Analysen ausgeschlossen werden. Um diesen Anteil so gering wie möglich zu halten, kann ein Teil der fehlenden Werte für die einzelnen Befragten ersetzt werden. Dazu gibt es die Möglichkeit, die Missings durch Informationen aus den Items, die gültige Werte haben, zu ersetzen.

Für eine solche Ersetzung gilt allerdings eine Bedingung: Es müssen gültige Werte für eine genügende Anzahl an Items vorliegen. Es wäre z. B. relativ unsinnig, für einen Befragten, der nur einen der Items beantwortet hat, diesen Wert für die Berechnung der Skala zu nutzen. Die Information aus einer Skala, die auf dieser Basis gebildet würde, wäre praktisch gleich Null.

Vor der Ersetzung, egal mit welchem Verfahren, muss also eine Bedingung formuliert werden: „Führe die Ersetzung nur durch, wenn mindestens vier der sechs Items gültige Werte haben“. Um diese Bedingung später in der Syntax anwenden zu können, muss zuerst eine neue Variable gebildet werden, die zählt, bei wie vielen Items gültige Werte vorliegen. Dafür eignet sich der Befehl „count“:

```
count beding = v10_2 v11 v12 v13 v14_2 v15 (1,2,3,4).
exe.
```

Bedeutung: Zähle in den Variablen v10_2, v11, v12, v13, v14_2 und v15, wie häufig die (gültigen) Werte 1, 2, 3 oder 4 vorkommen. Schreibe das Ergebnis in die neue Variable „beding“.

Mit dieser Bedingungs-Variablen kann jetzt eine „do if“-Bedingung formuliert werden, nach der die fehlenden Skalenwerte durch die Informationen, die in den gültigen Items stecken, ersetzt werden:

```
do if (beding = 4).
comp frau = 6/4 * sum (v10_2, v11, v12, v13, v14_2, v15) - 6.
end if.
exe.

oder kürzer:

if (beding = 4) frau = 6/4 * sum (v10_2, v11, v12, v13, v14_2, v15) - 6.
exe.
```

Bedeutung: Wenn genau vier der sechs Items einen gültigen Wert haben (do if (beding = 4)), dann berechne als Skalenwert für die Skala „frau“ die gewünschte Anzahl an Items (6) geteilt durch die vorhandene Anzahl an Items (4) und multipliziere dies mit der Summe der Items, die einen gültigen Wert haben. Ziehe von dem Ergebnis 6 ab, weil vorher auch bei der gesamten Skala der Wert 6 abgezogen wurde, um ihre Range bei 0 anfangen zu lassen.
Das Schlüsselwort „sum“ berechnet genau wie z. B. „v10_2 + v11“ usw. eine Summe. Im Gegensatz zur Addition mit dem +-Zeichen gibt „sum“ aber auch dann ein Ergebnis aus, wenn einer oder mehrere der Variablen einen fehlenden Wert aufweisen. Das Kommando berechnet dann einfach die Summe der gültigen Werte.

Danach muss auch die Ersetzung für den Fall vorgenommen werden, dass fünf gültige Werte vorliegen:

```
do if (beding = 5).
comp frau = 6/5 * sum (v10_2, v11, v12, v13, v14_2, v15) - 6.
end if.
exe.

kürzer:

if (beding = 5) frau = 6/5 * sum (v10_2, v11, v12, v13, v14_2, v15) - 6.
exe.
```

Nach diesen Ersetzungen zeigt eine neue Häufigkeitsauszählung, dass der Anteil der fehlenden Werte in der Skala deutlich geringer geworden ist:

FRAU Einstellung Frauen/Berufstätigkeit

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,0 traditionelle Einstellung	19	,6	,6	,6
	1,0	24	,8	,8	1,4
	1,2	3	,1	,1	1,4
	1,5	3	,1	,1	1,5
	2,0	47	1,5	1,5	3,0
	2,4	6	,2	,2	3,2
	3,0	61	1,9	2,0	5,2
	3,6	14	,4	,4	5,6
	4,0	87	2,8	2,8	8,4
	4,5	3	,1	,1	8,5
	4,8	13	,4	,4	8,9
	5,0	128	4,1	4,1	13,1
	6,0	217	6,9	7,0	20,0
	7,0	193	6,1	6,2	26,2
	7,2	31	1,0	1,0	27,2
	7,5	11	,4	,4	27,6
	8,0	228	7,3	7,3	34,9
	8,4	28	,9	,9	35,8
	9,0	259	8,3	8,3	44,1
	9,6	35	1,1	1,1	45,3
	10,0	229	7,3	7,4	52,6
	10,5	7	,2	,2	52,9
	10,8	21	,7	,7	53,5
	11,0	239	7,6	7,7	61,2
	12,0	277	8,8	8,9	70,1
	13,0	238	7,6	7,7	77,8
	13,2	18	,6	,6	78,3
	13,5	5	,2	,2	78,5
	14,0	176	5,6	5,7	84,2
	14,4	18	,6	,6	84,7
	15,0	180	5,7	5,8	90,5
	15,6	7	,2	,2	90,8
	16,0	107	3,4	3,4	94,2
	16,8	13	,4	,4	94,6
	17,0	94	3,0	3,0	97,6
	18,0 progressive Einstellung	74	2,3	2,4	100,0
	Total	3110	99,1	100,0	
Missing	System	28	,9		
Total		3138	100,0		

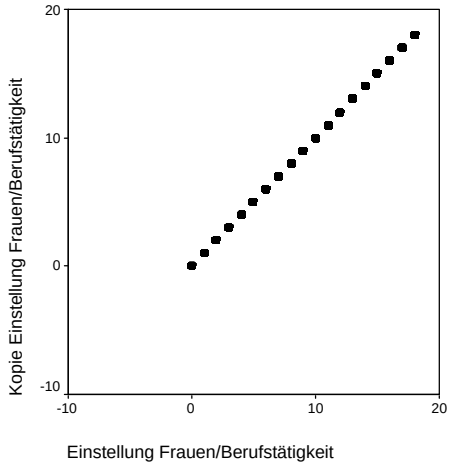
Es gibt nur noch 28 Fälle mit fehlenden Werten, was 0,9 % entspricht. Durch die Multiplikation mit Brüchen sind die ersetzten Werte häufig Dezimalwerte, was aber kein Problem ist.

Um zu prüfen, ob man keinen Fehler bei der Ersetzung gemacht hat, ist es sinnvoll, die Skala mit allen fehlenden Werten zuerst einmal zu kopieren (z. B. comp frau2 = frau), die Ersetzungen an der Kopie vorzunehmen und dann beide Variablen miteinander zu korrelieren und/oder ein Streudiagramm zu machen. Die beiden Skalen müssen perfekt miteinander korrelieren, da die fehlenden Werte in der Ursprungsvariablen gar nicht mit in die Korrelation eingehen:

Correlations

		FRAU Einstellung Frauen/Berufstätigkeit	FRAU2 Kopie Einstellung Frauen/Berufstätigkeit
FRAU Einstellung Frauen/Berufstätigkeit	Pearson Correlation	1	1,000
	Sig. (2-tailed)	,	,
	N	2777	2777
FRAU2 Kopie Einstellung Frauen/Berufstätigkeit	Pearson Correlation	1,000	1
	Sig. (2-tailed)	,	,
	N	2777	3110

Ein Streudiagramm muss zeigen, dass die Werte der beiden Variablen auf einer perfekten Geraden liegen:



Abschließend müssen die immer noch fehlenden Werte in der Skala durch einen recode-Befehl ersetzt und gelabelt werden und die User-Missings deklariert werden. Für die Vergabe des Wertelabels für die fehlenden Werte muss „add value labels“ verwendet werden, da sonst die schon existierenden Wertelabels überschrieben werden und verschwinden.

9 Explorative Faktorenanalyse

Manche Itemlisten bilden nicht eine gemeinsame Dimension ab, sondern zwei oder mehr Dimensionen. Im ALLBUS 2000 sind dies z. B. die Variablen v145 bis v156 mit der Frage:
„Wie kommt man in unserer Gesellschaft am ehesten nach oben? Beurteilen Sie bitte die Wichtigkeit der Eigenschaften und Umstände auf diesen Karten. Bitte sagen Sie mir zu jeder Aussage, wie wichtig dieses Ihrer Meinung nach für den Aufstieg in unserer Gesellschaft gegenwärtig ist.“:

- Opportunismus, Rücksichtslosigkeit
- Bildung, Ausbildung
- Politische Betätigung
- Zufall, Glück
- Intelligenz, Begabung
- Beziehungen, Protektion
- Leistung, Fleiß
- Geld, Vermögen
- Initiative, Durchsetzungsvermögen
- Soziale Herkunft, aus der „richtigen“ Familie stammen
- Bestechung, Korruption
- Kooperativer Führungsstil, Offenheit

Die Befragten sollen angeben, ob die angegebenen Eigenschaften und Umstände ihrer Meinung nach „Sehr wichtig“, „Wichtig“, „Weniger wichtig“ oder „Unwichtig“ für den gesellschaftlichen Aufstieg sind.
Eine Reliabilitätsanalyse zeigt, dass Cronbachs Alpha für eine Skala mit diesen 12 Items ,636 beträgt, was für diese Anzahl an Items deutlich zu wenig ist (zur Erinnerung: Alpha steigt mit der Anzahl der Variablen in der Skala). Das liegt in diesem Fall daran, dass die Variablen zwei unterschiedliche Dimensionen repräsentieren: Items wie z. B. „Opportunismus, Rücksichtslosigkeit“, „Politische Betätigung“ und „Beziehungen, Protektion“ stellen eine andere Dimension der Voraussetzungen für einen gesellschaftlichen Aufstieg dar als Items wie „Bildung, Ausbildung“, „Intelligenz, Begabung“, „Leistung, Fleiß“ usw.

Um zu prüfen, wie viele Dimensionen eine Reihe von Items abbildet, ist eine explorative Faktorenanlyse geeignet. Sie beruht im Grunde wieder auf den Korrelationen der Variablen untereinander. Der Grundgedanke ist also auch hier wieder, dass ein Befragter bei Variablen, die zu einer Dimension gehören, den Aussagen etwa im gleichen Maße zustimmen oder sie ablehnen müsste.

9.1 Durchführung einer Faktorenanalyse / Standardeinstellungen

Es empfiehlt sich, vor einer Reliabilitätsanalyse eine Faktorenanalyse durchzuführen. Mit ihren Ergebnissen kann dann geprüft werden, wie hoch die interne Konsistenz (Cronbachs Alpha) einer Skala ist, deren Items eine gemeinsame Dimension abbilden.

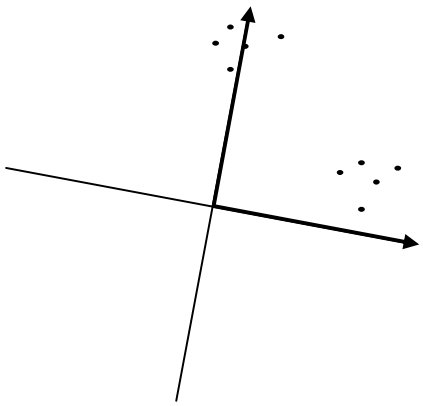
Die Syntax für eine Faktorenanalyse sieht folgendermaßen aus:

```
factor variables = v145 to v156.

kurz:

fac var = v145 to v156.
```

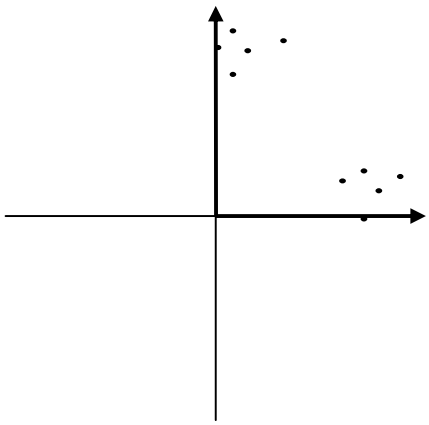
Bedeutung: Führe eine Faktorenanalyse mit den Variablen „v145“ bis „v156“ (im Datensatz aufeinanderfolgend) durch.

Mit dieser Syntax, d. h. ohne weitere Unterkommandos, werden die Standardeinstellungen des Befehls ausgeführt: Extraktion der Faktoren nach dem „Kaiser-Guttman-Kriterium“ und die Varimax-Rotation.
Die Extraktion nach dem Kaiser-Guttman-Kriterium (oft auch kurz „Kaiser-Kriterium“ genannt) bedeutet, dass alle Faktoren extrahiert werden, deren Eigenwert größer als 1 ist. Der Eigenwert (Englisch: „Eigenvalue“) eines Faktors gibt an, welcher Betrag der Gesamtstreuung aller Variablen des Faktorenmodells durch diesen Faktor erklärt wird. Ein Eigenwert, der größer als 1 ist, bedeutet, dass der entsprechende Faktor mehr Varianz erklärt als eine einzelne Variable. Da die Faktorenanalyse ja gerade dafür gedacht ist, Variablen durch einen Faktor zusammenzufassen, erfüllt das Eigenwert-Kriterium damit eine Mindestanforderung.

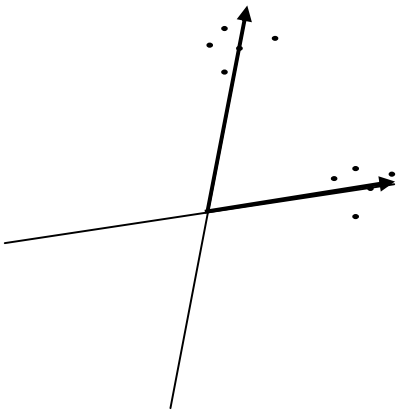
Die Rotation ist ein notwendiges Verfahren, um die Ergebnisse einer Faktorenanalyse interpretieren zu können. Was passiert, ist folgendes:



Der erste Faktor, d. h. der Faktor mit dem größten Eigenwert, wird so ausgerichtet, dass der einen möglichst geringen Abstand zu den Punkten (Variablen) hat, die er repräsentiert.
Der zweite Faktor wird dann rechtwinklig zum ersten Faktor angesetzt. Dies bedeutet, dass er unter Umständen nicht so nah an die Punkte herankommt, wie es möglich wäre.



Um eine optimale Lösung für beide Faktoren zu erreichen, lässt man sie rotieren. Bei einer orthogonalen Rotation wird der rechte Winkel zwischen den Faktoren beibehalten. Mathematisch bedeutet dies, dass die beiden Faktoren nicht miteinander korrelieren. SPSS führt unter der Standardeinstellung die orthogonale „Varimax“-Rotation durch.



Bei einer nicht-orthogonalen Rotation dürfen die Faktoren den rechten Winkel aufgeben. Sie können somit die Variablenwerte unter Umständen besser abbilden. Die Faktoren korrelieren bei einer solchen Lösung allerdings miteinander. SPSS bietet z. B. die nicht-orthogonale „Oblimin“- und die „Promax“-Rotation an.

Ohne Rotation kann es passieren, dass die Ladungen, die SPSS für die einzelnen Items auf jedem Faktor angibt, zu falschen Interpretationen führen. Die Ausgabe, die das auf Seite 84 dargestellte Kommando produziert, macht dies klarer:

Factor Analysis

Communalities

	Initial	Extraction
V145 WEG Z.ERFOLG:OPPORTUNISM.,RUECKSICHTSLOS	1,000	,407
V146 WEG ZUM ERFOLG: BILDUNG, AUSBILDUNG	1,000	,441
V147 WEG ZUM ERFOLG: POLITISCHE BETAETIGUNG	1,000	,271
V148 WEG ZUM ERFOLG: ZUFALL, GLUECK	1,000	,191
V149 WEG ZUM ERFOLG: INTELLIGENZ, BEGABUNG	1,000	,508
V150 WEG ZUM ERFOLG: BEZIEHUNGEN, PROTEKTION	1,000	,465
V151 WEG ZUM ERFOLG: LEISTUNG, FLEISS	1,000	,581
V152 WEG ZUM ERFOLG: GELD, VERMOEGEN	1,000	,434
V153 WEG Z.ERFOLG: INITIATIVE, DURCHSETZUNG	1,000	,343
V154 WEG Z.ERFOLG: HERKUNFT, RICHTIGE FAMILIE	1,000	,482
V155 WEG Z.ERFOLG: BESTECHUNG, KORRUPTION	1,000	,574
V156 WEG Z.ERFOLG: KOOPERATION, OFFENHEIT	1,000	,359

Extraction Method: Principal Component Analysis.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,708	22,567	22,567	2,708	22,567	22,567	2,696	22,471	22,471
2	2,347	19,562	42,129	2,347	19,562	42,129	2,359	19,658	42,129
3	,964	8,037	50,166						
4	,918	7,652	57,818						
5	,857	7,142	64,960						
6	,758	6,317	71,277						
7	,667	5,559	76,835						
8	,633	5,274	82,110						
9	,617	5,139	87,249						
10	,554	4,617	91,867						
11	,512	4,267	96,133						
12	,464	3,867	100,000						

Extraction Method: Principal Component Analysis.

Component Matrix[a]

	Component	
	1	2
V145 WEG Z.ERFOLG:OPPORTUNISM.,RUECKSICHTSLOS	,632	-,085
V146 WEG ZUM ERFOLG: BILDUNG, AUSBILDUNG	-,131	,651
V147 WEG ZUM ERFOLG: POLITISCHE BETAETIGUNG	,485	,187
V148 WEG ZUM ERFOLG: ZUFALL, GLUECK	,256	,355
V149 WEG ZUM ERFOLG: INTELLIGENZ, BEGABUNG	-,084	,708
V150 WEG ZUM ERFOLG: BEZIEHUNGEN, PROTEKTION	,650	,207
V151 WEG ZUM ERFOLG: LEISTUNG, FLEISS	-,222	,729
V152 WEG ZUM ERFOLG: GELD, VERMOEGEN	,629	,195
V153 WEG Z.ERFOLG: INITIATIVE, DURCHSETZUNG	,088	,579
V154 WEG Z.ERFOLG: HERKUNFT, RICHTIGE FAMILIE	,683	,127
V155 WEG Z.ERFOLG: BESTECHUNG, KORRUPTION	,752	-,092
V156 WEG Z.ERFOLG: KOOPERATION, OFFENHEIT	-,277	,531

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Rotated Component Matrix[a]

	Component	
	1	2
V145 WEG Z.ERFOLG:OPPORTUNISM.,RUECKSICHTSLOS	,607	-,196
V146 WEG ZUM ERFOLG: BILDUNG, AUSBILDUNG	-,013	,664
V147 WEG ZUM ERFOLG: POLITISCHE BETAETIGUNG	,511	,097
V148 WEG ZUM ERFOLG: ZUFALL, GLUECK	,315	,303
V149 WEG ZUM ERFOLG: INTELLIGENZ, BEGABUNG	,044	,712
V150 WEG ZUM ERFOLG: BEZIEHUNGEN, PROTEKTION	,676	,088
V151 WEG ZUM ERFOLG: LEISTUNG, FLEISS	-,088	,757
V152 WEG ZUM ERFOLG: GELD, VERMOEGEN	,654	,080
V153 WEG Z.ERFOLG: INITIATIVE, DURCHSETZUNG	,190	,554
V154 WEG Z.ERFOLG: HERKUNFT, RICHTIGE FAMILIE	,694	,003
V155 WEG Z.ERFOLG: BESTECHUNG, KORRUPTION	,724	-,224
V156 WEG Z.ERFOLG: KOOPERATION, OFFENHEIT	-,178	,572

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Die Kommunalitäten in der ersten Tabelle geben an, welchen Betrag der Streuung einer Variablen alle Faktoren zusammen erklären. Interessant ist hier nur die rechte Spalte: Die beiden extrahierten Faktoren erklären z.B. 40,7 % der Varianz der Variablen v145 (Weg zum Erfolg: Opportunismus, Rücksichtslosigkeit). Die Kommunalitäten werden in der Regel nicht berichtet, weil normalerweise die einzelnen Faktoren und nicht die Gesamtheit aller Faktoren von Interesse ist.

Die zweite Tabelle zeigt, wie viele Faktoren (nach dem Kaiser-Kriterium) extrahiert wurden und welchen Anteil der Gesamtvarianz aller Variablen sie erklären. Im vorliegenden Fall gibt es zwei Faktoren, die einen Eigenwert von mehr als 1 haben. Der erste Faktor erklärt (anfänglich, d. h.

vor der Rotation) 22,567 % der Gesamtvarianz, der zweite Faktor erklärt 19,562 %. Zusammen erklären die beiden Faktoren 42,129 % der Gesamtvarianz, der Rest geht auf die restlichen 10 Faktoren zurück. Nach der Rotation ist der Anteil der durch beide Faktoren erklärten Varianz gleich geblieben, aber die durch die einzelnen Faktoren erklärte Varianz hat sich geringfügig verändert: beim ersten Faktor auf 22,471 %, beim zweiten auf 19,658 %. Durch die orthogonale Rotation ist der erste Faktor also etwas „schlechter" geworden, der zweite etwas „besser". Genau dies kann man nach den Abbildungen auf Seite 85 auch erwarten. Normalerweise wird, wenn die Ergebnisse von Faktorenanalysen berichtet werden, der anfängliche Eigenwert der Faktoren und die nach der Rotation durch die einzelnen Faktoren erklärte Varianz oder die gemeinsam erklärte Varianz berichtet.

Die dritte Tabelle stellt die (anfängliche) Komponentenmatrix dar, die vierte Tabelle die rotierte Komponentenmatrix. Interpretiert wird in der Regel nur die rotierte Lösung. Für jede Variable wird für die beiden Faktoren eine „Faktorladung" angegeben, die die Korrelation der Variablen mit dem Faktor angibt; bildlich gesprochen also die Nähe der Variablen zum Faktor. Hohe Ladungen, d. h. Ladungen über ,5 auf dem ersten Faktor haben die Items:

- Opportunismus, Rücksichtslosigkeit
- Politische Betätigung
- Beziehungen, Protektion
- Geld, Vermögen
- Herkunft, richtige Familie
- Bestechung, Korruption

Hohe Ladungen auf dem zweiten Faktor haben die Items:

- Bildung, Ausbildung
- Intelligenz, Begabung
- Leistung, Fleiß
- Initiative, Durchsetzung
- Kooperation, Offenheit

Items, die auf dem einen Faktor hoch laden, haben auf dem anderen Faktor eine niedrige Ladung. Die Faktorenlösung ist damit „sauber“, d. h. die Variablen können genau einem Faktor zugeordnet werden. Dies wird auch als „Einfachstruktur“ bezeichnet. Die Einfachstruktur ist kein Muss, macht die Interpretation von Faktoren aber deutlich einfacher.

Das einzige Item, das zu keiner der beiden Dimensionen zu passen scheint, ist v148 (Zufall, Glück). Es lädt auf beiden Faktoren nur mit ca. ,3. In solchen Fällen kann man ausprobieren, ob das Item bei einer nicht-orthogonalen Rotation höher auf einem der beiden Faktoren lädt. Wenn dies nicht der Fall ist, muss das Item aus der Skalenbildung ausgeschlossen werden.

Bei der „Zuordnung“ von Items zu einem Faktor ist übrigens das Vorzeichen nicht von Bedeutung. „Hohe Ladungen“ sind Ladungen über ,5 oder unter -,5. Für die Interpretation des Faktors ist es allerdings von erheblicher Bedeutung, ob ein Item positiv oder negativ mit dem Faktor korreliert. Hat man z. B. hohe positive Ladungen für vier Items, die den „Anteil Kollegen im Freundeskreis“, den „Anteil Verwandte im Freundeskreis“, den „Anteil Vereinsmitglieder im Freundeskreis“ und den „Anteil Nachbarn im Freundeskreis“ beinhalten, würde dieser Faktor auf einen sehr breit gefächerten Freundeskreis hindeuten. Wäre dagegen die Ladung für „Anteil Verwandte im Freundeskreis“ positiv, die anderen drei dagegen negativ, würde dies bedeuten, dass eine Person, die einen hohen Anteil Verwandte im Freundeskreis hat, gleichzeitig einen niedrigen Anteil Kollegen, Vereinsmitglieder und Nachbarn im Freundeskreis hat. Der Faktor würde also auf einen eng begrenzten Freundeskreis hindeuten, d. h. je höher die Werte des Faktors, desto mehr ist der Freundeskreis auf Verwandte beschränkt und desto geringer ist der Anteil an Kollegen, Vereinsmitgliedern und Nachbarn.

Die Interpretation der beiden Faktoren ist im vorliegenden Beispiel relativ einfach. Auf dem ersten Faktor haben die Items eine hohe Ladung, die einen gesellschaftlichen Aufstieg durch „unlautere“ Ressourcen und Handlungen beinhalten. Die höchsten Ladungen haben „Bestechung, Korruption“ und „Herkunft, richtige Familie“; diese Items werden also durch den Faktor am besten repräsentiert (Markiervariablen). Auf dem zweiten Faktor laden dagegen Items hoch, die einen Aufstieg durch normativ erwünschte Ressourcen und Handlungen umschreiben. Die höchsten Ladungen haben hier „Leistung, Fleiß“ und „Intelligenz, Begabung“. Den ersten Faktor könnte man „gesellschaftlicher Aufstieg durch unlautere Mittel und Herkunft“, den zweiten „gesellschaftlicher Aufstieg durch Leistung und Talent“ nennen.

Die Items, die auf einem Faktor hoch und auf dem anderen niedrig laden, können wahrscheinlich in einer Skala zusammengefasst werden. Mit den Items v145, v147, v150, v152, v154 und v155

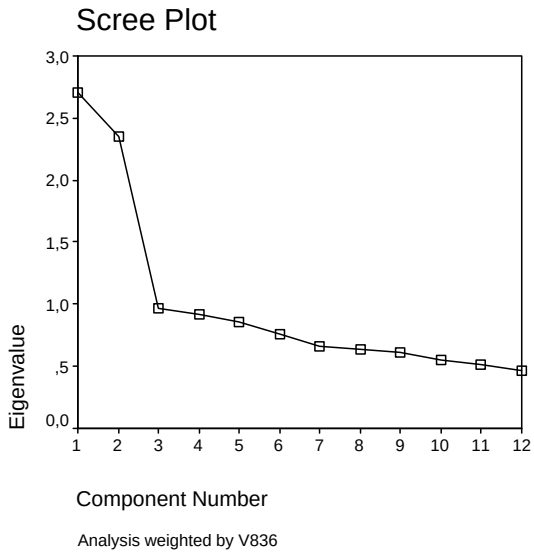
(gesellschaftlicher Aufstieg durch unlautere Mittel) sollte jetzt eine Reliabilitätsanalyse durchgeführt werden, genauso mit den Items v146, v149, v151, v153 und v156 (gesellschaftlicher Aufstieg durch Talent und Leistung). Cronbachs Alpha beträgt für diese beiden Skalen ,726 bzw. ,668 und kann auch nicht mehr durch den Ausschluss einzelner Items verbessert werden. Es ist möglich, dass die Variablen, die nach den Ergebnissen einer Faktorenanalyse eine Dimension abbilden, trotzdem eine Skala mit einem relativ niedrigen Alpha ergeben. In solchen Fällen sollten einzelne Items aus der Skala ausgeschlossen werden, um sie zu verbessern.

9.2 Alternative Extraktionsmethoden: Eigenwert und Faktorenanzahl

Die Extraktion von Faktoren nach dem Kaiser-Kriterium kann zu viele Faktoren extrahieren. Wenn ein Faktor einen Eigenwert hat, der genau 1 beträgt oder nur geringfügig mehr, wird er nach dieser Methode extrahiert, obwohl er unter Umständen nur wenig mehr Erklärungskraft hat als alle nachfolgenden Faktoren. Ob einer oder mehrere Faktoren einen so geringen Eigenwert haben und ob die nachfolgenden Faktoren kaum weniger Varianz erklären, kann an der Tabelle „Erklärte Varianz“ abgelesen werden. Es ist aber auch möglich, einen sogenannten Scree-Plot (Scree (engl.) = Geröll) zu machen:

```
fac var = v145 to v156
/plot = eigen.
```

Bedeutung: Führe eine Faktorenanalyse durch und gib zusätzlich eine Darstellung der Eigenwerte in einem Diagramm (Scree-Plot) aus.



Der Scree-Plot trägt die Eigenwerte der Faktoren auf der Y-Achse ab. Es sollten nur Faktoren extrahiert werden, die vor dem deutlichen Abfall der Eigenwerte liegen (hier: 2). Die restlichen Faktoren sehen aus, als ob Steine einen Abhang hinuntergerollt und unten als Geröll (scree) liegengeblieben wären. Wendet man dieses Kriterium an, kann es sein, dass man zu der Entscheidung kommt, einen (oder mehrere)

Faktoren nicht zu extrahieren, obwohl ihr Eigenwert größer als 1 ist.
In diesem Fall muss eine neue Faktorenanalyse durchgeführt werden, bei der mit Hilfe eines Unterkommandos angegeben wird, welche bzw. wie viele Faktoren extrahiert werden sollen. Das Kommando gibt z. B. an, dass nur Faktoren extrahiert werden sollen, deren Eigenwert mindestens 1,3 beträgt.

```
fac var = v145 to v156
/criteria = mineigen(1.3).

kurz:

fac var = v145 to v156
/crit = min(1.3).
```

Es gibt auch die Möglichkeit, die Anzahl der zu extrahierenden Faktoren festzulegen:

```
fac var = v145 to v156
/criteria = factors(2).

kurz:

fac var = v145 to v156
/crit = fac(2).
```

Bedeutung: Es werden genau zwei Faktoren bestimmt, die extrahiert werden sollen.

9.3 Nicht-orthogonale Rotation

Bei der orthogonalen Rotation, die die Standardeinstellung in SPSS ist, werden die beiden Faktoren dazu gezwungen, nicht miteinander zu korrelieren. Dies hat Vorteile, weil z. B. in multivariaten Analysen keine Drittvariableneffekte auftreten können. Der Nachteil ist, dass solche Faktoren unter Umständen die (Daten-)Realität nicht genau abbilden (siehe Seite 71).
Um eine nicht-orthogonale Rotation statt der standardmäßig durchgeführten Varimax-Rotation durchzuführen, ist ein Unterkommando notwendig:

```
fac var = v145 to v156
/rotation = oblimin.

kurz:

fac var = v145 to v156
/rot = obl.
```

oder

```
fac var = v145 to v156
/rotation = promax.

kurz:

fac var = v145 to v156
/rot = pro.
```

Oblimin und Promax sind nicht-othogonale Rotations-Verfahren, die ähnliche Ergebnisse produzieren. Oblimin ist das gebräuchliste; Promax ist vor allem für große Datensätze geeignet.

SPSS gibt bei nicht-orthogonaler Rotation drei Matrizen aus: die unrotierte Komponentenmatrix, eine „Mustermatrix“ (Pattern Matrix) und eine „Strukturmatrix“ (Structure Matrix). Für die Interpretation der Faktoren ist die Mustermatrix geeignet. Die Faktorladungen entsprechen den standardisierten Partialkorrelationen zwischen dem Item und dem Faktor (d.h. die Korrelation des Items mit dem anderen Faktor ist herauspartialisiert). Die Strukturmatrix gibt die bivariaten Korrelationen zwischen Faktor und Item wieder.

Bei einer nicht-orthogonalen Rotation (obliquen Rotation) wird als letzte Tabelle eine Korrelationsmatrix zwischen den beiden Faktoren ausgegeben. Für die Oblimin-Rotation beträgt diese Korrelation r = ,035, bei der Promax-Rotation r = -,051. Die beiden Faktoren korrelieren also so gering miteinander, dass eine orthogonale Rotation zu einem sehr ähnlichen Ergebnis geführt hätte. Dies wird auch deutlich, wenn man die drei Faktorlösungen vergleicht:

Varimax-Rotation:

Rotated Component Matrix[a]

	Component	
	1	2
V145 WEG Z.ERFOLG:OPPORTUNISM.,RUECKSICHTSLOS	,607	-,196
V146 WEG ZUM ERFOLG: BILDUNG, AUSBILDUNG	-,013	,664
V147 WEG ZUM ERFOLG: POLITISCHE BETAETIGUNG	,511	,097
V148 WEG ZUM ERFOLG: ZUFALL, GLUECK	,315	,303
V149 WEG ZUM ERFOLG: INTELLIGENZ, BEGABUNG	,044	,712
V150 WEG ZUM ERFOLG: BEZIEHUNGEN, PROTEKTION	,676	,088
V151 WEG ZUM ERFOLG: LEISTUNG, FLEISS	-,088	,757
V152 WEG ZUM ERFOLG: GELD, VERMOEGEN	,654	,080
V153 WEG Z.ERFOLG: INITIATIVE, DURCHSETZUNG	,190	,554
V154 WEG Z.ERFOLG: HERKUNFT, RICHTIGE FAMILIE	,694	,003
V155 WEG Z.ERFOLG: BESTECHUNG, KORRUPTION	,724	-,224
V156 WEG Z.ERFOLG: KOOPERATION, OFFENHEIT	-,178	,572

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Oblimin-Rotation:

Pattern Matrix[a]

	Component	
	1	2
V145 WEG Z.ERFOLG:OPPORTUNISM.,RUECKSICHTSLOS	,610	-,208
V146 WEG ZUM ERFOLG: BILDUNG, AUSBILDUNG	-,024	,665
V147 WEG ZUM ERFOLG: POLITISCHE BETAETIGUNG	,510	,088
V148 WEG ZUM ERFOLG: ZUFALL, GLUECK	,310	,298
V149 WEG ZUM ERFOLG: INTELLIGENZ, BEGABUNG	,032	,711
V150 WEG ZUM ERFOLG: BEZIEHUNGEN, PROTEKTION	,675	,075
V151 WEG ZUM ERFOLG: LEISTUNG, FLEISS	-,101	,759
V152 WEG ZUM ERFOLG: GELD, VERMOEGEN	,653	,068
V153 WEG Z.ERFOLG: INITIATIVE, DURCHSETZUNG	,180	,551
V154 WEG Z.ERFOLG: HERKUNFT, RICHTIGE FAMILIE	,695	-,009
V155 WEG Z.ERFOLG: BESTECHUNG, KORRUPTION	,728	-,238
V156 WEG Z.ERFOLG: KOOPERATION, OFFENHEIT	-,187	,576

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 4 iterations.

Promax-Rotation:

Pattern Matrix[a]

	Component	
	1	2
V145 WEG Z.ERFOLG:OPPORTUNISM.,RUECKSICHTSLOS	,602	-,180
V146 WEG ZUM ERFOLG: BILDUNG, AUSBILDUNG	,003	,664
V147 WEG ZUM ERFOLG: POLITISCHE BETAETIGUNG	,514	,111
V148 WEG ZUM ERFOLG: ZUFALL, GLUECK	,323	,312
V149 WEG ZUM ERFOLG: INTELLIGENZ, BEGABUNG	,061	,714
V150 WEG ZUM ERFOLG: BEZIEHUNGEN, PROTEKTION	,679	,106
V151 WEG ZUM ERFOLG: LEISTUNG, FLEISS	-,070	,755
V152 WEG ZUM ERFOLG: GELD, VERMOEGEN	,656	,097
V153 WEG Z.ERFOLG: INITIATIVE, DURCHSETZUNG	,203	,560
V154 WEG Z.ERFOLG: HERKUNFT, RICHTIGE FAMILIE	,695	,022
V155 WEG Z.ERFOLG: BESTECHUNG, KORRUPTION	,719	-,205
V156 WEG Z.ERFOLG: KOOPERATION, OFFENHEIT	-,164	,568

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

9.4 Sortierung der Ausgabe und Ausblenden niedriger Ladungen

Da bei der Faktorenlösung im Grunde nur die hohen Ladungen von Interesse sind und da die Ausgabe häufig unübersichtlich ist, kann es für die Interpretation der Faktoren hilfreich sein, niedrige Ladungen auszublenden und sich die Items in der Reihenfolge der Höhe der Ladungen anzeigen zu lassen.

```
fac var = v145 to v156
/format = sort blank(.5).

kurz:

fac var = v145 to v156
/form = sort blank(.5).
```

Das Unterkommando gibt an, dass erstens die Ausgabe nach der Höhe der Faktorladungen sortiert werden soll (format = sort) und zweitens alle Ladungen nicht angezeigt werden sollen, die kleiner als ,5 sind (format = blank(.5)). Da kein Kommando für die Art der Rotation angegeben wurde, wird eine Varimax-Rotation durchgeführt.

Die Ausgabe verändert sich folgendermaßen:

Rotated Component Matrix[a]

	Component	
	1	2
V155 WEG Z.ERFOLG: BESTECHUNG, KORRUPTION	,724	
V154 WEG Z.ERFOLG: HERKUNFT, RICHTIGE FAMILIE	,694	
V150 WEG ZUM ERFOLG: BEZIEHUNGEN, PROTEKTION	,676	
V152 WEG ZUM ERFOLG: GELD, VERMOEGEN	,654	
V145 WEG Z.ERFOLG:OPPORTUNISM.,RUECKSICHTSLOS	,607	
V147 WEG ZUM ERFOLG: POLITISCHE BETAETIGUNG	,511	
V148 WEG ZUM ERFOLG: ZUFALL, GLUECK		
V151 WEG ZUM ERFOLG: LEISTUNG, FLEISS		,757
V149 WEG ZUM ERFOLG: INTELLIGENZ, BEGABUNG		,712
V146 WEG ZUM ERFOLG: BILDUNG, AUSBILDUNG		,664
V156 WEG Z.ERFOLG: KOOPERATION, OFFENHEIT		,572
V153 WEG Z.ERFOLG: INITIATIVE, DURCHSETZUNG		,554

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Mit dieser Ausgabe ist es wesentlicher einfacher, anhand der Variablenlabels die inhaltliche Bedeutung der Faktoren zu erfassen. Es wird außerdem deutlich, dass die Variable v148 (Zufall, Glück) zu keinem der beiden Faktoren passt. Ihre Ladungen sind auf beiden Faktoren zu niedrig, um angezeigt zu werden.

9.5 Darstellung der Ergebnisse einer Faktorenanalyse

SPSS gibt bei auch bei der Faktorenanalyse mehr Tabellen und Informationen aus, als in der Regel berichtet werden. In den meisten Fällen beschränkt man sich darauf, die rotierte Matrix bzw. die Mustermatrix zu berichten. Was außerdem angegeben werden sollte, ist das Extraktions-Verfahren (Kaiser-Kriterium = Eigenwerte größer 1, Höhe des Eigenwerts, Festlegung der Faktorenzahl), u. U. der Grund für die Ablehnung des (am häufigsten angewendeten) Kaiser-Kriteriums (z. B. Ergebnis des Scree-Plots), die anfängliche Höhe der Eigenwerte der extrahierten Faktoren und die nach der Rotation durch die einzelnen Faktoren erklärte Varianz bzw. insgesamt erklärte Varianz. Diese Angaben können im Text gemacht oder innerhalb der Tabelle dargestellt werden. Für das Beispiel könnte eine Tabelle folgendermaßen aussehen:

Tabelle 4: Ergebnisse der Faktorenanalyse: Gesellschaftlicher Aufstieg durch…

	Faktor 1: Unlautere Mittel und Herkunft	**Faktor 2: Leistung und Talent**
Bestechung, Korruption	,724	
Herkunft, richtige Familie	,694	
Beziehungen, Protektion	,676	
Geld, Vermögen	,654	
Opportunismus, Rücksichtslosigkeit	,607	
Politische Betätigung	,511	
Leistung, Fleiß		,757
Intelligenz, Begabung		,712
Bildung, Ausbildung		,664
Kooperation, Offenheit		,572
Initiative, Durchsetzung		,554

Extraktionsmethode: Hauptkomponantenanalye mit Kaiser-Kriterium; Rotationsmethode: Varimax
Anfängliche Eigenwerte: Faktor 1 = 2,708, Faktor 2 = 2,347
Erklärte Varianz nach Rotation: Faktor 1 = 22,47 %, Faktor 2 = 19,66 %

Eine abschließende Bemerkung: Bisher wurde unterschlagen, dass mit der Standardeinstellung von SPSS eine bestimmte Art der Faktorenanalyse durchgeführt wird, die „Hauptkomponentenanalyse“ (Principal Component Analysis, kurz PCA). Um genau zu sein, sollte man diesen Begriff verwenden. In Tabelle 4 könnte man also in der Überschrift von einer Hauptkomponentenanalyse sprechen und sich den Zusatz in der Fußnote sparen. Wenn in empirischen Analysen in der Literatur von „Faktorenanalysen“ die Rede ist, sind meistens PCAs gemeint, solange keine anderen Angaben gemacht werden.

10 Lineare Regression

In Kapitel 5.5. wurden Partialkorrelationen beschrieben. Mit ihnen ist es möglich, die Ergebnisse bivariater Korrelationen daraufhin zu überprüfen, ob es sich um Scheinkorrelationen handelt. Der Nachteil bei Partialkorrelationen ist, dass zwar für mehrere Drittvariablen kontrolliert werden kann, dies aber immer nur für den Zusammenhang zwischen zwei Variablen möglich ist.
Regressionsanalysen gehen darüber hinaus, indem sie für jede unabhängige Variable den Einfluss aller anderen UVs herauspartialisieren. Ein (unstandardisierter) Regressionskoeffizient gibt also an, wie stark der Zusammenhang zwischen einer aller UVs und der AV ist, wenn der Einfluss aller anderen UVs kontrolliert (konstant gehalten) wird.
Für lineare Regressionsanalysen gelten einige Einschränkungen:

- Die abhängige Variable muss metrisch skaliert sein,
- Die unabhängigen Variablen müssen ebenfalls metrisch skaliert oder Dummy-Variablen (dichotome Variablen) sein. Ordinal oder nominal skalierte Variablen müssen in Dummies aufgesplittet und einzeln in die Regression eingeführt werden,
- Die unabhängigen Variablen dürfen nicht zu hoch miteinander korrelieren (Gefahr der Multikollinearität),
- Die Residuen müssen für alle Werte der abhängigen Variablen die gleiche Varianz aufweisen (Homoskedastizität)[3],
- die Residuen müssen normalverteilt sein.

Was diese Voraussetzungen bedeuten und wie man sie überprüft, wird in den folgenden Kapiteln behandelt. Das lineare Regressionsmodell gilt allgemein als relativ robust gegen Verletzungen dieser Voraussetzungen, d. h. es wird häufig davon ausgegangen, dass die Ergebnisse der

[3] Für die Überprüfung von Homoskedastizität bietet SPSS keinen expliziten Test an, möglich ist aber ein optischer Test anhand eines Residuenplots.

Regressionsanalyse auch dann richtig sind, wenn diese Voraussetzungen nicht erfüllt sind. Es wird im Folgenden allerdings deutlich werden, dass die Robustheits-Annahme nicht immer gerechtfertigt ist.

10.1 Durchführung einer linearen Regressionsanalyse

Die grundlegende Syntax für eine lineare Regression sieht folgendermaßen aus:

```
regression
/dependent = frau
/method = enter v209 v216 v3.

kurz:

reg
/dep = frau
/meth = enter v209 v216 v3.
```

Bedeutung: Führe eine lineare Regressionsanalyse aus, bei der „frau“ (Einstellung Frauen/Berufstätigkeit) die abhängige und v209 (Politische Links-Rechts-Selbsteinstufung), v216 (Geschlecht) und v3 (Erhebungsgebiet: West/Ost) die unabhängigen Variablen sind.

SPSS gibt folgende Ausgabe:

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	V3 ERHEBUNGSGEBIET: WEST - OST V216 GESCHLECHT, BEFRAGTE<R> , V209[a] LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR.	,	Enter

a. All requested variables entered.

b. Dependent Variable: FRAU Einstellung Frauen/Berufstätigkeit

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,348[a]	,121	,120	3,7278

a. Predictors: (Constant), V3 ERHEBUNGSGEBIET: WEST - OST , V216 GESCHLECHT, BEFRAGTE<R> , V209 LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5319,317	3	1773,106	127,596	,000[a]
	Residual	38629,979	2780	13,896		
	Total	43949,296	2783			

a. Predictors: (Constant), V3 ERHEBUNGSGEBIET: WEST - OST , V216 GESCHLECHT, BEFRAGTE<R> , V209 LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR.

b. Dependent Variable: FRAU Einstellung Frauen/Berufstätigkeit

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7,999	,390		20,506	,000
	V209 LINKS-RECHTS-SELBSTEINSTUFUNG, BEFR	-,428	,041	-,186	-10,398	,000
	V216 GESCHLECHT, BEFRAGTE<R>	1,068	,141	,134	7,550	,000
	V3 ERHEBUNGSGEBIET: WEST - OST	2,363	,180	,234	13,099	,000

a. Dependent Variable: FRAU Einstellung Frauen/Berufstätigkeit

Die erste Tabelle zeigt einfach, welche unabhängigen Variablen in die Regression eingegangen sind. In der zweiten Tabelle (Modellzusammenfassung, Model Summary) ist das korrigierte r^2 (Adjusted R Square) von Interesse. Es zeigt an, wie groß der Anteil der Varianz der abhängigen Variablen ist, der durch die unabhängigen Variablen erklärt wird. Das korrigierte r^2 ist ein Maß für die Anpassungsgüte des Regressionsmodells. Wenn z. B. vier Variablen nur 0,2 % der Varianz erklären könnten, wäre das Modell schlecht, auch wenn sich einige oder alle UVs statistisch signifikant auf die AV auswirken. In der Soziologie sind korrigierte r^2-Werte über ,4 sehr selten, wenn nicht völlig offensichtliche Variablen in Beziehung gesetzt werden. Im vorliegenden Fall ist die Modellgüte mit 12,0 % erklärter Varianz nicht berauschend, aber im Rahmen des Normalen.

In der dritten Tabelle (ANOVA) kann abgelesen werden, ob das Modell überhaupt den Daten angemessen ist. Der F-Test gibt Auskunft darüber, ob sämtliche UVs gleichzeitig einen statistisch signifikanten Einfluss auf die AV haben. Die Aussagekraft des F-Tests ist insofern beschränkt, als das er auch dann ein statistisch signifikantes Ergebnis zeigen kann, wenn sich nur eine von mehreren UVs statistisch signifikant auf die AV auswirkt. Umgekehrt ist es meist so, dass ein nicht-signifikanter F-Wert nur dann ausgewiesen wird, wenn keine der UVs einen statistisch signifikanten Einfluss auf die AV hat. In diesem Fall weiß man aber auch ohne den F-Test durch die t-Werte der einzelnen Koeffizienten, dass das Modell offensichtlich ungeeignet ist.

Das zentrale Ergebnis der Regressionsanalyse wird in der vierten Tabelle (Koeffizienten) ausgegeben. Auf die Konstante wird später eingegangen. Die unstandardisierten Koeffizienten in der ersten Spalte geben an, um welchen Betrag die AV steigt (bei positiven Koeffizienten) bzw. sinkt (bei negativen Koeffizienten), wenn die jeweilige UV um eine Einheit ansteigt. Mit jedem Punkt auf der Links-Rechts-Skala sinkt also der Skalenwert bei „Einstellung Frauen/Berufstätigkeit“ um 0,428 Punkte. Da auf der Links-Rechts-Skala hohe Werte „rechts“ bedeuten und auf der Einstellungsskala hohe Werte „progressiv“ bedeuten, zeigt der negative Koeffizient also, dass mit steigender „rechter“ politischer Einstellung eine geringere progressive Einstellung zur Berufstätigkeit von Frauen einhergeht (bzw. eine stärker traditionelle Einstellung).

Der Koeffizient für das Geschlecht beträgt 1,068. Koeffizienten von Dummy-Variablen müssen immer auf die Referenzkategorie bezogen werden, und dies ist immer die numerisch kleinere Ausprägung. Bei der Variablen „Geschlecht“ bedeutet der Wert 1 „Mann“, der Wert 2 „Frau“. „Mann“ ist also die Referenzkategorie. Frauen haben im Vergleich zu Männern 1,068 Punkte mehr auf der Einstellungsskala, sind also um nur einen Punkt progressiver als Männer. Ähnliches gilt für die Variable „Erhebungsgebiet“: 1 bedeutet „Alte Bundesländer“, 2 bedeutet „Neue Bundesländer“. Die Bewohner der neuen Bundesländer sind um 2,363 Punkte progressiver als die Bewohner der alten Bundesländer. Alle diese Angaben gelten „unter sonst gleichen Umständen“, d. h. für den Fall, dass der Wert aller anderen Variablen konstant gehalten wird.

In der zweiten Spalte ist der Standardfehler des (unstandardisierten) Regressionskoeffizienten abgetragen. Der Quotient aus Koeffizient und Standardfehler ergibt den t-Wert (4. Spalte), und zusammen mit der Anzahl der Freiheitsgrade (nicht in der Tabelle dargestellt) ergibt sich ein Signifikanzniveau (letzte Spalte) für den Koeffizienten. Im oben angeführten Beispiel wirken sich alle UVs statistisch signifikant auf die AV aus.

In der 3. Spalte werden außerdem die standardisierten Regressionskoeffizienten angegeben. Aus ihnen kann abgelesen werden, wie stark der Einfluss einer Variablen im Vergleich zu den anderen Variablen im Modell ist. Aus den unstandardisierten Koeffizienten kann dies nicht abgelesen werden, weil sie unterschiedlich skaliert sind (Politische Einstellung: 1-10; Geschlecht: 1-2, Erhebungsgebiet: 1-2). Im Beispiel hat das Erhebungsgebiet mit β (beta) = ,234 den stärksten Einfluss auf die Einstellung zur Berufstätigkeit von Frauen. Für diese Einstellung ist es also am stärksten entscheidend, ob jemand im Westen oder im Osten wohnt. Danach folgt die politische Links-Rechts-Einstellung mit β=-,186 (welches Vorzeichen die Koeffizienten haben, ist für die Frage ihrer Einflussstärke nicht entscheidend). Den relativ schwächsten Einfluss hat das Geschlecht mit β=,134. Das in der letzten Spalte dargestellte Signifikanzniveau gilt natürlich für den unstandardisierten (b) und den standardisierten (beta) Koeffizienten.

10.2 Hinzunahme weiterer Variablen

In manchen Analysen ist es sinnvoll, nicht alle Variablen auf einmal in die Regression einzuführen, sondern bestimmte (z. B. thematisch geordnete) Variablen-Blöcke nacheinander in das Modell aufzunehmen. In solchen Fällen kann man einfach mehrere getrennte Analysen rechnen, wobei die Variablenblöcke nach und nach aufgenommen werden; z. B.

```
1. Regression:
reg
/dep = aeq
/meth = enter v257 v3 staat.

2. Regression:
reg
/dep = aeq
/meth = enter v257 v3 staat v825.
```

Die Ausgabe (nur Koeffizienten) der beiden Regressionen kann dann miteinander verglichen werden. Stellt man solche Vergleiche in Tabellenform dar, empfiehlt sich ein Format wie das folgende:

Tabelle 5: Regression auf monatliches Äquivalenzeinkommen
(unstandardisierte Koeffizienten, Standardfehler in Klammern)

	Modell 1	**Modell 2**
	b	**b**
Konstante	2.317,479***	822,965*
	(372,545)	(401,489)
Arbeitsstunden pro Woche	26,527***	18,465**
	(5,716)	(5,644)
Erhebungsgebiet (Ref.: West)	-907,128***	-733,261***
	(169,494)	(166,949)
Staatsangehörigkeit (Ref.: Nicht-Deutsche)	831,941**	402,349
	(250,248)	(251,197)
Berufsprestige (Treiman-Skala)		46,834***
		(5,199)
korr. r^2	,040	,098
N	1.237	1.225

*** p < ,001; ** p < ,01; * p < ,05

An der Tabelle kann man verschiedene Vergleiche anstellen:

- Modell 1 erklärt 4,0 % der Varianz, Modell 2 dagegen 9,8 %. Mit der Hinzunahme der Variablen „Berufsprestige“ hat sich die Vorhersagekraft des Modells also deutlich verbessert.
- Mit der Hinzunahme der Variablen „Berufsprestige“ ist der Koeffizient für die Staatsangehörigkeit gesunken und nicht mehr statistisch signifikant. Staatsangehörigkeit und Berufsprestige korrelieren also miteinander, und wenn man für diesen Zusammenhang kontrolliert, zeigt sich, dass die Staatsangehörigkeit keinen direkten Einfluss auf das Äquivalenzeinkommen ausübt. Das Berufsprestige korreliert ebenfalls mit den anderen Variablen (ihre Koeffizienten verändern sich), deren Einfluss bleibt aber weiterhin statistisch signifikant.

Die Anzahl der in die Regression eingegangenen Fälle wird übrigens nicht direkt ausgegeben. Sie kann aber aus der ANOVA-Tabelle abgelesen werden: N beträgt immer die Gesamtzahl der Freiheitsgrade + 1.

Bei so großen Einheiten wie dem Äquivalenzeinkommen kann es übrigens übersichtlicher sein, die Variable (egal ob als UV oder als AV) in anderen Einheiten umzurechnen. Berechnet man beispielsweise ein neues Äquivalenzeinkommen, indem man das Ä.-Einkommen durch 1000 teilt, werden die Koeffizienten kleiner, bedeuten aber inhaltlich dasselbe:

Tabelle 6: Regression auf monatliches Äquivalenzeinkommen (in 1.000er-Schritten), (unstandardisierte Koeffizienten, Standardfehler in Klammern)

	Modell 1	**Modell 2**
	b	**b**
Konstante	2,317***	,823*
	(,373)	(,401)
Arbeitsstunden pro Woche	,027***	,018**
	(,006)	(,006)
Erhebungsgebiet (Ref.: West)	-,907***	-,733***
	(,169)	(,167)
Staatsangehörigkeit (Ref.: Nicht-Deutsche)	,832**	,402
	(,250)	(,251)
Berufsprestige (Treiman-Skala)		,047***
		(,005)
korr. r^2	,040	,098
N	1.237	1.225

*** p < ,001; ** p < ,01; * p < ,05

Das Umrechnen von Variablen in kleinere Einheiten empfiehlt sich vor allem da, wo eine UV sehr hohe Werte erreicht (z. B. das Einkommen), die AV aber nur sehr niedrige Werte erreicht (z. B. eine Skala, die von 1 bis 10 geht). Die Koeffizienten für solche UVs sind häufig so gering, dass sie nur als „,000“ angezeigt werden, weil erst bei der fünften oder sechsten Dezimalstelle andere Werte als 0 folgen. Solche Koeffizienten sind natürlich wenig aussagekräftig.

10.3 Ordinal oder nominal skalierte unabhängige Variablen

Ordinal oder nominal skalierte Variablen können nicht in dieser Form in eine Regression aufgenommen werden. Sie müssen in Dummy-Variablen aufgesplittet werden, d. h., sie werden in mehrere dichotome 0/1 Variablen umkodiert. Die Variable „bildung“ (Bildungsniveau) hat beispielsweise drei Ausprägungen: 1 = (Höchstens) Hauptschulabschluss; 2 = Mittlere Reife und 3 = (Fach-)Abitur. Eine der drei Ausprägungen wird nicht in einen Dummy umkodiert; sie stellt

in der folgenden Regression die Referenzkategorie dar. (Meist wird bei ordinal skalierten Variablen die niedrigste oder die höchste Ausprägung genommen; andere sind aber auch möglich und teilweise sinnvoll, wie ein Bespiel zeigen wird.) Die anderen beiden Ausprägungen werden so umkodiert, dass eine Variable die Information „Mittlere Reife ja/nein“ und eine Variable die Information „(Fach-)Abitur ja/nein“ beinhaltet:

```
recode bildung (2=1) (1,3=0) (9=9) into bild_m
/bildung (3=1) (1,2=0) (9=9) into bild_a.
exe.

var lab
bild_m "Mittlere Reife"
bild_a "(Fach-)Abitur".

val lab
bild_m 0 "Andere" 1 "MR" 9 "kA"
/bild_a 0 "Andere" 1 "Abi" 9 "kA".

miss val bild_m bild_a (9).
```

Die neue Variable „bild_m“ (Mittlere Reife) hat den Wert 1 für all diejenigen Fälle, die in der Variablen „bildung“ den Wert 2 (Mittlere Reife) haben. Alle anderen Fälle haben den Wert 0 (Nicht Mittlere Reife / Andere). Entsprechend haben in der Variablen „bild_a“ ((Fach-)Abitur) all diejenigen den Wert 1, die in „bildung“ den Wert 3 ((Fach-)Abitur) haben, und alle anderen den Wert 0.

Diese beiden Dummies werden zusätzlich zu den UVs aus dem letzten Beispiel in die Regression aufgenommen:

```
reg
/dep = aeq
/meth = enter v257 v3 staat v825 bild_m bild_a.
```

Der SPSS-Output für die Koeffizienten sieht folgendermaßen aus:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	984,997	411,094		2,396	,017
	V257 BEFRAGTER: ARBEITSSTUNDEN PRO WOCHE	18,780	5,661	,092	3,317	,001
	V3 ERHEBUNGSGEBIET: WEST - OST	-835,583	176,221	-,138	-4,742	,000
	STAAT deutsche Staatangehörigkeit	400,835	251,139	,044	1,596	,111
	V825 TREIMANPRESTIGE, EINORDNUNG NACH TERWEY	40,225	6,226	,217	6,461	,000
	BILD_M Mittlere Reife	340,128	164,050	,069	2,073	,038
	BILD_A (Fach-)Abitur	390,092	187,528	,077	2,080	,038

a. Dependent Variable: AEQ Äquivalenzeinkommen

Die Mittlere Reife hat einen b-Koeffizienten von 340,128; das (Fach-)Abitur hat einen b-Koeffizienten von 390,092. Diese beiden Werte müssen auf die Referenzkategorie bezogen werden, d.h. auf diejenigen mit (höchstens) Hauptschulabschluss. Befragte mit Mittlerer Reife haben ein um gut 340€ höheres Äquivalenzeinkommen als Befragte, die höchstens einen Hauptschulabschluss haben. Befragte mit (Fach-)Abitur haben ein um rund 390€ höheres Einkommen als Hauptschüler. Beide Koeffizienten sind auf dem 5%-Niveau statistisch signifikant. Ob sich Befragte mit Mittlerer Reife und solche mit (Fach-)Abitur ebenfalls statistisch signifikant voneinander unterscheiden, kann allerdings nicht abgelesen werden. Aus den Koeffizienten kann man berechnen, dass (Fach-)Abiturienten 390,092 – 340,128 = 49,964€ mehr im Monat haben als diejenigen mit Mittlerer Reife. Ob diese Differenz statistisch signifikant ist, kann nur berechnet werden, indem statt „(höchstens) Hauptschulabschluss“ eine dieser beiden Ausprägungen als Referenzkategorie genommen wird. Nimmt man beispielsweise die „Mittlere Reife“ als Referenzkategorie (und berechnet einen neuen Dummy für „(höchstens) Hauptschulabschluss“), sieht das Ergebnis der Regressionsanalyse folgendermaßen aus:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1325,124	452,998		2,925	,004
	V257 BEFRAGTER: ARBEITSSTUNDEN PRO WOCHE	18,780	5,661	,092	3,317	,001
	V3 ERHEBUNGSGEBIET: WEST - OST	-835,583	176,221	-,138	-4,742	,000
	STAAT deutsche Staatangehörigkeit	400,835	251,139	,044	1,596	,111
	V825 TREIMANPRESTIGE, EINORDNUNG NACH TERWE	40,225	6,226	,217	6,461	,000
	BILD_H (Höchstens) Hauptschulabschluss	-340,128	164,050	-,070	-2,073	,038
	BILD_A (Fach-)Abitur	49,964	176,893	,010	,282	,778

a. Dependent Variable: AEQ Äquivalenzeinkommen

„Mittlere Reife“ ist jetzt die Referenzkategorie. Befragte mit (höchstens) Hauptschulabschluss haben rund 340€ weniger als diejenigen mit Mittlerer Reife. Dies ist einfach die Umkehrung der Berechnung aus der vorherigen Regression. Befragte mit (Fach-)Abitur haben knapp 50€ mehr als solche mit Mittlerer Reife. Diesen Wert konnte man schon anhand der Koeffizienten im letzten Modell berechnen; jetzt ist aber klar, dass diese Differenz mit einer Irrtumswahrscheinlichkeit von 77,8% nicht statistisch signifikant ist. Die Mittlere Reife und das (Fach-)Abitur führen also (unter sonst gleichen Umständen) zu einem höheren Äquivalenzeinkommen als der Hauptschulabschluss; zwischen Mittlerer Reife und (Fach-)Abitur gibt es aber keinen statistisch signifikanten Unterschied. (Dies gilt, um die Regression noch etwas mehr zu interpretieren, erstaunlicherweise unter Kontrolle des Berufsprestiges. Wenn also eine Person einen Hauptschulabschluss und eine andere die Mittlere Reife hat, verdient der erste 340€ weniger als der zweite, obwohl die beiden Personen Berufe mit gleich hohem Berufsprestige haben!)

Ein abschließender Hinweis zur Wahl der Referenzkategorie: In manchen Fällen gibt eine Theorie, eine Begriffsdefinition oder eine Forschungsfrage schon vor, welche der Ausprägungen einer ordinal oder nominal skalierten Variablen als Referenzkategorie gewählt werden sollte. Wenn beispielsweise die Auswirkung der Zugehörigkeit zu irgendeiner Konfession auf die Spendenbereitschaft untersucht werden soll und die Variable „Konfession“ die Ausprägungen „katholisch“, „evangelisch“, „andere“ und „konfessionslos“ hat, sollte „konfessionslos“ die Referenzkategorie sein. Die Koeffizienten würden dann angeben, um wie viel die Spendenbereitschaft von Katholiken, Protestanten bzw. einer anderen Konfession Zugehörigen die von Konfessionslosen übersteigt (oder unterschreitet). Wenn dagegen die Frage ist, ob Katholiken mehr spenden als andere Konfessionen, wäre „katholisch“ die geeignete Referenzkategorie.

10.4 Multikollinearität

Multikollinearität bedeutet, dass die unabhängigen Variablen in einer Regression zu hoch miteinander korrelieren. In solchen Fällen kann mathematisch nicht mehr eindeutig zugeordnet werden, auf welche der miteinander korrelierenden UVs ein Einfluss auf die AV zurückgeht. In extremen Fällen (also bei sehr hohen Interkorrelationen) schließt SPSS eine oder mehrere der Variablen aus der Regression aus. Bei weniger hohen Korrelationen wird zwar für alle Variablen

ein Koeffizient berechnet, dieser ist aber unzuverlässig. Die Standardfehler und damit das Signifikanzniveau sind dagegen auch in solchen Fällen korrekt. Man weiß also bei solchen Variablen, ob sie einen statistisch signifikanten Einfluss auf die AV haben, aber nicht, wie stark dieser Einfluss ist.

Um auf Multikollinearität zu prüfen, ist ein Unterkommando notwendig:

```
reg
/statistics = default tol
/dep = aeq
/meth = enter v257 v3 staat v825 bild_m bild_a.

kurz:

reg
/stat = def tol
/dep = aeq
/meth = enter v257 v3 staat v825 bild_m bild_a.
```

Bedeutung: Berechne eine Regressionsanalyse. Gib in der Ausgabe die Standardtabellen (statistics = default), d. h. die ohne Unterbefehl produzierten Tabellen aus und zusätzlich die Statistiken für den Test auf Multikollinearität (statistics = tol).

Die Tabelle mit den Koeffizienten wird mit dem Unterbefehl „tol“ um eine Doppelspalte erweitert, in der zwei Werte abgetragen sind: Die „Toleranz“ und der „VIF“ (Variance Inflation Factor). Beides sind Skalen, die Multikollinearität anzeigen. Die „Toleranz“-Skala reicht von 0 bis 1. Werte unter 0,3 weisen auf Multikollinearität hin. Der VIF ist einfach der Quotient aus 1/Toleranz. Er reicht von 1 bis unendlich. Bei ihm weisen Werte über 4 auf Multikollinearität hin.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	1786,667	496,671		3,597	,000		
	v257 BEFRAGTER: ARBEITSSTUNDEN PRO WOCHE	18,780	5,661	,092	3,317	,001	,965	1,037
	v3 ERHEBUNGSGEBIET: WEST - OST	-835,583	176,221	-,138	-4,742	,000	,872	1,147
	staat Staatsangehörigkeit	-400,835	251,139	-,044	-1,596	,111	,950	1,052
	v825 TREIMANPRESTIGE, EINORDNUNG NACH TERWEY	40,225	6,226	,217	6,461	,000	,653	1,532
	bild_m Mittlere Reife	340,128	164,050	,069	2,073	,038	,670	1,492
	bild_a (Fach-) Abitur	390,092	187,528	,077	2,080	,038	,532	1,879

a. Dependent Variable: aeq Äquivalenzeinkommen

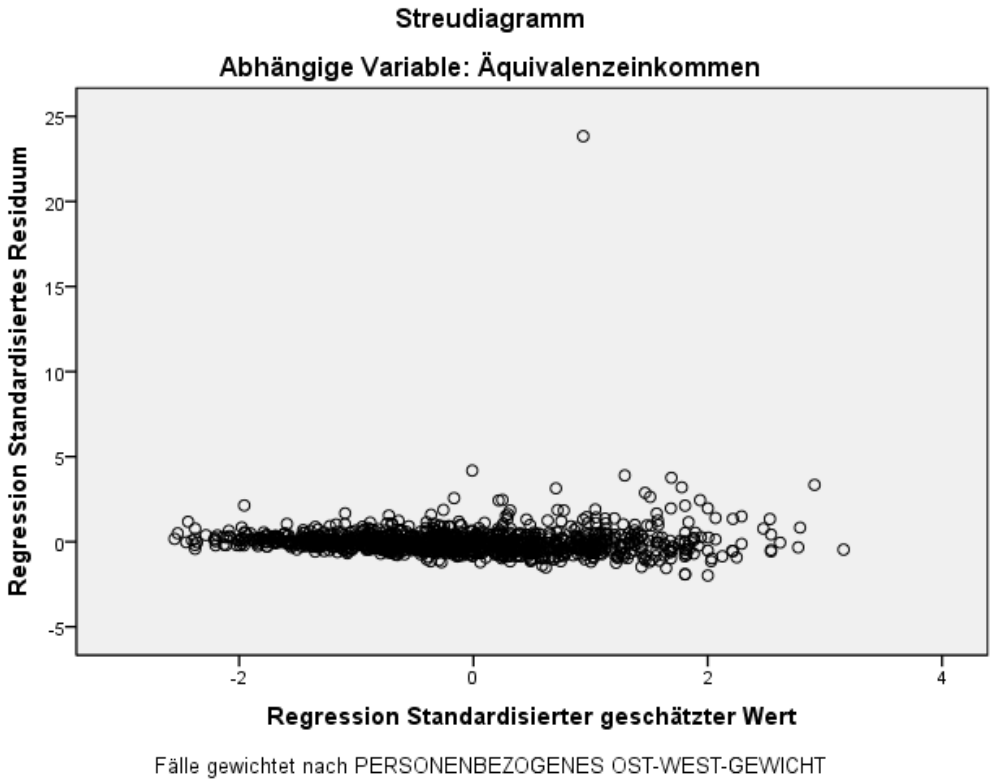
Wenn zwei Variablen hoch miteinander korrelieren, bleibt nichts anderes übrig, als eine der beiden Variablen aus der Regression zu entfernen. Inhaltlich ist dies in der Regel nicht weiter schlimm, weil die Variablen ja hoch miteinander korrelieren. Der Koeffizient für die eine Variable wird also dem für die andere sehr ähnlich sein, und Drittvariableneffekte sind ebenfalls ähnlich.

10.5 Homoskedastizität - Heteroskedastizität

Eine weitere Voraussetzung für eine Regressionsanalyse ist, dass Homoskedastizität vorliegt, d.h. dass die Varianz der Residuen konstant ist. Bei Heteroskedastizität, dem Gegenteil, muss angenommen werden, dass entweder eine Fehlspezifikation des Modells, Messfehler oder eine schiefe Verteilung der abhängigen Variablen vorliegt. Heteroskedastizität führt zu ineffizienten Schätzern und verfälschten Standardfehlern. Um diese aufzudecken, sind eigentlich Tests wie der Goldfeld/Quandt-Test, der Breusch-Pagan-Test oder der White-Test vorgesehen, welche jedoch nicht in SPSS implementiert sind. Man kann sich allerdings mit einem visuellen Test behelfen. Hierfür müssen die Residuen in einem Streudiagramm gegen die geschätzten Werte der abhängigen Variable geplottet werden. Ergibt sich im Diagramm ein Dreiecksmuster bzw. eine trichterförmige Streuung, ist das ein Zeichen für Heteroskedastizität.
Das Streudiagramm muss ebenfalls über ein Unterkommando angefordert werden:

```
reg
/dep = aeq
/meth = enter v257 v3 staat v825 bild_m bild_a
/scatterplot=(*zresid, *zpred).
```

Bedeutung: Gib mir, zusätzlich zu den Regressionsergebnissen ein Streudiagramm der standardisierten Residuen und der standardisierten geschätzten Werte für die AV aus.



Anhand der Grafik kann abgelesen werden, dass keine Heteroskedastizität vorliegt. Es zeigt sich eine homoskedastische Varianz der Residuen (diese stellt sich im Idealfall elipsenförmig dar), die geringfügig stärkere Streuung am rechten Rand kann hier vernachlässigt werden.

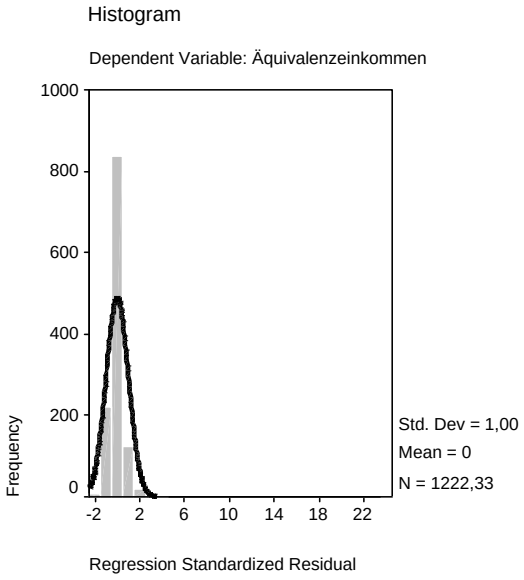
10.6 Normalverteilung der Residuen

Eine weitere Voraussetzung der Regression ist, dass die Residuen normalverteilt sind. Die Residuen sind die Abweichungen der tatsächlichen individuellen Werte von den Werten, die für jeden Fall durch die Regressionsgleichung geschätzt werden. Wenn die Residuen normalverteilt sind, bedeutet dies, dass keine für die Schätzung des Wertes der abhängigen Variablen wichtige unabhängige Variable im Modell fehlt. Anders ausgedrückt: Wenn alle wichtigen Variablen im Modell enthalten sind, müssten die Abweichungen vom Schätzwert zufällig verteilt sein – und damit normalverteilt. Wenn die Residuen deutlich von einer Normalverteilung abweichen, fehlt mindestens eine wichtige UV im Modell.

Die Normalverteilung der Residuen kann mit Hilfe eines Histogramms überprüft werden:

```
reg
/dep = aeq
/meth = enter v257 v3 staat v825 bild_m bild_a
/residuals = histogram.

kurz:

reg
/dep = aeq
/meth = enter v257 v3 staat v825 bild_m bild_a
/res = his.
```

SPSS gibt mit diesem Unterkommando ein Histogramm der standardisierten Residuen aus:



Die Residuen sind für das vorliegende Modell weit davon entfernt, normalverteilt zu ein. In solchen Fällen ist es möglich, dass die Signifikanztests für die Koeffizienten und der F-Test falsche Ergebnisse anzeigen. Bei großen Stichproben (ca. N > 40) geht man allerdings davon aus, dass trotz nicht-normalverteilter Residuen die Tests zuverlässige Ergebnisse liefern.

11 Logistische Regression

Für die Analyse dichotomer abhängiger Variablen muss die logistische Regression angewendet werden. Die abhängige Variable ist so aufgebaut, dass sie zwischen dem Eintreten eines Ereignisses und dem Nicht-Eintreten dieses Ereignisses unterscheidet. Anhand der Ergebnisse kann daraufhin abgelesen werden mit welcher Wahrscheinlichkeit das Ereignis eintritt und welche Einflussfaktoren diese Wahrscheinlichkeit beeinflussen.

Als unabhängige Variablen können sowohl metrische als auch kategoriale (nominale Variablen mit zwei und mehr Ausprägungen) Variablen einbezogen werden. Wobei die unabhängigen Variablen in SPSS häufig in Kovariate (nur metrische Variablen) und Faktoren (gemischte Skalenniveaus) unterschieden werden. Für die logistische Regression liegen zwei bedeutende Anwendungsvoraussetzungen vor. Zum einen darf, wie bei der linearen Regression auch, keine Multikollinearität zwischen den unabhängigen Variablen vorliegen. Zum anderen sollten die Fallzahlen in den Kategorien der abhängigen Variable nicht zu extrem ungleich verteilt sein. Beispielsweise sollte nicht 95 Prozent der Fälle in der einen Kategorie und 5 Prozent der Fälle in der anderen Kategorie sein.

11.1 Durchführung einer logistischen Regressionsanalyse

Die grundlegende Syntax für eine logistische Regression sieht folgendermaßen aus:

```
logistic regression „abhängige Variable“
/method = enter „Auflistung der unabhängigen Variablen“
/contrast (kategoriale Variable) = indicator (1).
```

Mit dem Unterbefehl ‘/contrast‘ werden die kategorialen Variablen als solche definiert und es wird festgelegt welche der Ausprägungen die Referenzkategorie darstellt. Im Befehl oben wird beispielsweise die erste Ausprägung als Referenzkategorie festgelegt, für diese wird demnach keine eigene Indikatorvariable gebildet. Für jede kategoriale Variable muss ein eigner Unterbefehl geschrieben werden. Bei dichotomen 0/1-kodierten Variablen wird von SPSS automatisch die Null als Referenzkategorie festgelegt, wodurch eine Definition per Unterbefehl nicht mehr notwendig ist.

Für die beispielhafte Betrachtung des Outputs wird im Folgenden auf Grundlage des ALLBUS 2000 eine logistische Regression für die Wahl der SPD gerechnet. Als unabhängige Variablen fließen die soziodemographischen Variablen Geschlecht, Alter und Bildung mit ein. Darüber hinaus wird der Einfluss der Parteiidentifikation, der politischen Links-Rechts-Selbsteinstufung und der Gewerkschaftsmitgliedschaft untersucht. Einige der Variablen müssen für die Analyse rekodiert werden.

Abhängige Variable „Wahlabsicht für die SPD“:

```
recode v627 (2=1) (1, 3 thru 96=0) into spd.
exe.
```

Unabhängige Variable “Parteiidentifikation”:

```
recode v44 (2=1) (1, 3 thru 7=0) into pi_spd.
exe.
```

Unabhängige Variable “Gewerkschaftsmitgliedschaft”:

```
recode v624 (1=1) (2=0) into gew.
exe.
```

Der Regressionsbefehl von oben sieht dann wie folgt aus:

```
logistic regression spd
/method = enter v216 alter bildung pi_spd gew v209
/contrast (bildung) = indicator (1).
```

Bedeutung: Berechne eine logistische Regression für die Wahl der SPD, füge als unabhängige Variablen das Geschlecht, das Alter, die Bildung, die Parteiidentifikation für die SPD, die Links-Rechts Selbsteinstufung und die Gewerkschaftsmitgliedschaft (rekodiert) hinzu.

SPSS produziert für die logistische Regression eine Fülle an Tabellen, von denen hier nur die wichtigsten besprochen werden sollen. In der ersten Tabelle „Zusammenfassung der Fallverarbeitung“ kann abgelesen werden wie viele Fälle in die Analyse mit einbezogen werden (N des Modells) und wie viele Fälle bei den unabhängigen Variablen fehlende Werte aufweisen. In den folgenden beiden Tabellen werden die Kodierungen der abhängigen Variablen und, wenn vorhanden, die der kategorialen unabhängigen Variablen angegeben.

Zusammenfassung der Fallverarbeitung

Ungewichtete Fälle[a]		N	Prozent
Ausgewählte Fälle	Einbezogen in Analyse	919	29,3
	Fehlende Fälle	2219	70,7
	Gesamt	3138	100,0
Nicht ausgewählte Fälle		0	,0
Gesamt		3138	100,0

a. Wenn die Gewichtung wirksam ist, finden Sie die Gesamtzahl der Fälle in der Klassifizierungstabelle.

Im Anfangsblock (Block 0) werden die Ergebnisse für ein Modell dargestellt in das nur die Konstante eingeht. Aus der Klassifizierungstabelle geht vorerst der Anteil von Fällen mit der Ausprägung 0 und 1 hervor. Darauf folgen die Ergebnisse der Regression, an denen die Chance für die Zugehörigkeit zur Gruppe 1 (hier den SPD Wählern) abgelesen werden kann, wenn keine weiteren Informationen über die Befragten vorhanden sind. Im nächsten Block (Methode = Einschluss) befinden sich daraufhin die Informationen zum Modell mit den einbezogenen unabhängigen Variablen. Die ersten Tabellen in diesem Abschnitt umfassen die Statistiken zur Beurteilung der Modellgüte. Darunter fallen der „Omnibus-Test der Modellkoeffizienten“ und die Tabelle „Modellzusammenfassung“, die den Wert für die Devianz sowie die Koeffizienten zur erklärten Varianz enthält.

Die Interpretation des Omnibus-Tests ähnelt der des F-Tests innerhalb der linearen Regressionsanalyse. So lässt sich sagen, dass wenn der Chi²-Test für den Modellschritt signifikant ist, hat zumindest eine der unabhängigen Variablen einen Einfluss auf die abhängige Variable. In der Tabelle darunter kann zum einen der -2LL-Wert für die Devianz und das Pseudo R^2 nach Cox & Snell sowie das nach Nagelkerke abgelesen werden. Nagelkerkes R^2 hat zwar den

Vorteil, dass es auf den gewohnten Wertebereich von [0;1] normiert ist, kann jedoch, wie alle Pseudo R^2-Werte, nicht, wie in der linearen Regressionsanalyse, als prozentualer Anteil an erklärter Varianz interpretiert werden. Im vorliegenden Beispiel zeigt sich ein Pseudo R^2 nach Nagelkerke von 0,498, welches auf eine sehr gute Modellanpassung hindeutet. Die Klassifizierungstabelle zeigt darunter dann noch die Fälle, die unter Berücksichtigung der unabhängigen Variablen richtig klassifiziert werden konnten.

Modellzusammenfassung

Schritt	-2 Log-Likelihood	Cox & Snell R-Quadrat	Nagelkerkes R-Quadrat
1	824,201[a]	,369	,498

a. Schätzung beendet bei Iteration Nummer 5, weil die Parameterschätzer sich um weniger als ,001 änderten.

In der Tabelle „Variablen in der Gleichung“ können abschließend die Koeffizienten der unabhängigen Variablen abgelesen werden. In der ersten Spalte werden die unabhängigen Variablen sowie die Konstante aufgelistet, in der zweiten Spalte können die B-Koeffizienten (die Logits) abgelesen werden und die dritte Spalte beinhaltet die Standardfehler der Koeffizienten. In der vierten Spalte werden die Ergebnisse der Wald-Statistik abgetragen. Auf Grund dieser Werte wird das Signifikanzniveau berechnet, in dem der Koeffizient durch den Standardfehler geteilt und dieser Wert quadriert wird. Die Werte der Wald-Statistik entsprechen somit in ihrer Bedeutung den T-Werten, welche aus der linearen Regression bekannt sind. In der fünften Spalte finden sich die Freiheitsgrade, die ebenfalls für die Berechnung des Signifikanzniveaus (sechste Spalte) notwendig sind. In der siebten Spalte werden abschließend die Odds (Exp(B)) dargestellt, für die ebenfalls das Signifikanzniveau in Spalte sechs gilt.

Variablen in der Gleichung

		Regressions-koeffizient B	Standardfehler	Wald	df	Sig.	Exp(B)
Schritt 1[a]	v216	-,326	,181	3,228	1	,072	,722
	alter	-,009	,006	2,859	1	,091	,991
	bildung			11,916	2	,003	
	bildung(1)	-,216	,221	,957	1	,328	,806
	bildung(2)	-,779	,230	11,490	1	,001	,459
	pi_spd	3,257	,232	196,565	1	,000	25,959
	v209	-,224	,053	17,677	1	,000	,799
	v624_r	,542	,233	5,413	1	,020	1,720
	Konstante	,986	,526	3,519	1	,061	2,681

a. In Schritt 1 eingegebene Variablen: v216, alter, bildung, pi_spd, v209, v624_r.

Die Ergebnisse der Regressionsanalyse zeigen, dass nicht alle unabhängigen Variablen einen signifikanten Einfluss auf die Wahlabsicht der SPD haben. Für die Bildung gilt jedoch, dass die Chance für die SPD zu stimmen bei Befragten mit Abitur (oder Fachabitur) um den Faktor 0,459 oder um 117,86 % geringer ist als bei denjenigen mit höchstens einem Hauptschulabschluss. Ein eindeutiger Effekt zeigt sich anhand der Parteiidentifikation für die SPD. Dies entspricht den allgemeinen Erwartungen, da diese inhaltlich sehr nah an der zu erklärenden Variable Wahlabsicht liegt. Einen ebenfalls sehr klaren Einfluss hat die Gewerkschaftsmitgliedschaft. Hier zeigt sich, dass die Chance der Wahlabsicht für die SPD bei einem Gewerkschaftsmitglied um 72 % größer ist als bei einem Befragten der kein Mitglied in einer Gewerkschaft ist. Darüber hinaus gilt für die politische Links-Rechts-Selbsteinstufung, dass die Chance der SPD-Wahl mit jedem Skalenpunkt um 25,16 % sinkt. Dieser Zusammenhang sollte sich so darstellen, da der Wert 1 auf der Skala „Links“ bedeutet und der Wert 10 „Rechts“.

Index der SPSS-Befehle

MATCH FILES	Zusammenführung von Datensätzen
MEANS	Tabelle von Mittelwerten
MISSING VALUES	Fehlende Werte deklarieren (benutzerdefiniert)
NONPAR CORRELATIONS	Nichtparametrische Korrelationsanalyse
ONEWAY	Mittelwertvergleich (Post-hoc-Test)
PAR CORRELATIONS	Partialkorrelation
RECODE (INTO)	Umkodierung einer Variablen (in eine neue Variable)
REGRESSION	Lineare Regressionsanalyse
RELIABILITY VARIABLES	Reliabilitätsanalyse
RENAME VARIABLES	Umbenennen einer Variablen
SAVE OUTFILE	Speichern des Datenfiles
SELECT IF	Auswahl bestimmter Fälle (Datenanalyse)
SORT CASES	Sortierung der Fälle
SPLIT FILE	Aufteilung des Datensatzes in Subgruppen
T-TEST	Berechnung eines T-Tests
TEMPORARY	Ausführung nur für den folgenden Befehl
UNIANOVA	Multivariate Varianzanalyse
VARIABLE LABELS	Variable einen Namen zuweisen
VALUE LABELS	Variable Werte zuweisen
WEIGHT	Gewichtung