Greifswald Comparative Politics

Working Paper No. 9
March 2015

Henrike Englert, Jan Helmdag & Kati Kuitto

Einführung in die Statistik und das Arbeiten mit Stata

Henrike Englert, Jan Helmdag & Kati Kuitto (2015). Einführung in die Statistik und das Arbeiten mit Stata. Greifswald Comparative Politics Working Paper No. 9.

Corresponding author: Jan Helmdag (helmdag@uni-greifswald.de)

University of Greifswald
Department of Political Science and Communication Studies
Chair of Comparative Politics
Baderstraße 6/7
17487 Greifswald
Germany

http://comparativepolitics.uni-greifswald.de

Greifswald Comparative Politics
ISSN 2195-6502

Vorwort

Das vorliegende Skript bietet eine erste Einführung in die Statistik und das Arbeiten mit dem Statistikprogramm Stata. Es ist insbesondere für die Bedürfnisse der Studierendenden der Bachelor- und Masterstudiengänge der Politikwissenschaft am Institut für Politik- und Kommunikationswissenschaft an der Universität Greifswald konzipiert, aber es bietet darüber hinaus einen guten Ausgangspunkt für jeden mit anwendungsorientiertem Interesse an sozialwissenschaftlichen statistischen Analysen. Das Skript beinhaltet die Grundlagen der deskriptiven und der Inferenzstatistik mit entsprechenden univariaten Kennzahlen, bivariaten Zusammenhangsmaßen, sowie der bi- und multivariaten linearen Regressionsanalyse. Neben den Grundlagen wird anhand von Syntaxbefehlen und Screenshots detailliert beschrieben, wie sich die Analysen mit Stata durchführen lassen.

Die Idee, ein solches Skript zu verfassen, entstand aus unserer langjährigen Lehrpraxis heraus, da wir immer wieder auf Probleme der Anwendung von Statistik und entsprechenden Programmen seitens der Studierenden in Seminaren mit empirisch-analytischen Lernzielen gestoßen sind. Zum einen haben die Studierenden unseres Bachelorstudiengangs zwar eine Statistikvorlesung und ein begleitendes Tutorium besucht, aber wenn es dann um die konkrete Anwendung auf eine bestimmte Forschungsfrage in den einzelnen Seminaren ging, war dieses Vorwissen entweder teilweise in Vergessenheit geraten oder zu unspezifisch. Zum anderen kommen innerhalb unseres forschungsorientierten Masterstudiengangs Studierende mit sehr unterschiedlichen Vorkenntnissen im Bereich Statistik zusammen, was eine deutliche Herausforderung für die erfolgreiche und zielorientierte Lehre und das Lernen in den empirisch orientierten Modulen des Studiengangs darstellt. Das vorliegende Skript soll dabei helfen die angesprochenen Defizite anzugehen und den Studierenden als ein Nachschlagewerk mit anwendungsorierentierten Beispielen dienen.

Mit Unterstützung aus dem Programm interStudies der Universität Greifswald, das auf eine Verbesserung der Studierbarkeit der Studienangebote, eine breitere Kompetenzentwicklung der Studierenden, sowie eine weitere Professionalisierung in Lehre und Prüfungswesen zielt, konnten und können wir in den Jahren 2014 und 2015 einen Intensivkurs „Einstiegshilfe für den Masterstudiengang Politikwissenschaft“ realisieren. In diesem Kurs, der als Blockveranstaltung in den ersten Wochen des ersten Fachsemesters der MA-Studierenden stattfindet, vermitteln wir die in diesem Skript enthaltenen Grundlagen der statistischen Analyse und den Umgang mit der Statistikprogramm Stata. Das Skript ist größtenteils aus diesem Kontext heraus entstanden und von der zweiten Kohorte des Kurses auf seine Praxistauglichkeit getestet. Wir bedanken uns für das Feedback der KursteilnehmerInnen im Wintersemester 2014/2015, sowie für die Mitarbeit von Sebastian Laacke und den Statistiktutorinnen und -tutoren am IPK. Wir hoffen, dass die vorliegende Einführung Ihnen den Zugang zu der spannenden Welt der statistischen Analyse erleichtern wird.

Inhaltsverzeichnis

1 Kurze Einführung in die Statistik

In diesem einleitenden Kapitel wird zunächst auf das Verhältnis von Theorie und Daten [1.1] eingegangen, um anschließend die aus dem Erhebungsvorgang entstehenden Skalenniveaus [1.2] der jeweiligen Variablen zu besprechen. Abschließend werden die einzelnen Teilbereiche der Statistik [1.3] vorgestellt, die es auf unterschiedliche Weise ermöglichen die generierten Daten zu untersuchen.

1.1 Das Verhältnis von Theorie und Daten

Statistische Analysen sind immer theoriebezogen: In der Regel werden mit ihrer Hilfe bestehende Theorien getestet. Am Anfang eines deduktiven Forschungsprozesses stehen konzeptionelle Überlegungen, also ein Erklärungsmodell, das in ein statistisches Modell übersetzt und mithilfe statistischer Verfahren getestet wird. Unter Umständen ist aber auch die Entwicklung von Theorien möglich, wenn man im Zusammenhang mit statistischen Analysen auf neue, evtl. überraschende Zusammenhänge stößt. Daher ist Theorie auch bei empirischen Untersuchungen und statistischen Analysen so wichtig. Es gilt: keine Interpretation ohne theoretische Überlegungen!

Bei empirischen Untersuchung ist es zunächst wichtig, sich anhand theoretischer Überlegungen bewusst zu machen, welches Konzept genau untersucht werden soll. Dafür ist die sogenannte Konzeptspezifikation von zentraler Bedeutung: Bei ihr wird festgelegt, welche Bedeutung die in der Theorie verwendeten Begriffe haben. Dies ist die Voraussetzung für die anschließende Operationalisierung, bei der mithilfe von Korrespondenzregeln die theoretischen Begriffe mit empirisch beobachtbaren Sachverhalten verknüpft werden, indem Messregeln formuliert werden. Diese sind nichts anderes als Erläuterungen der zur Messung notwendigen Operationen, so dass sichergestellt wird, dass das theoretische Konzept adäquat erfasst wird. Schließlich werden die Informationen über den interessierenden Sachverhalt erhoben: Der Schritt des Messens im engeren Sinne (Kodierung) besteht darin, gemäß der festgelegten Messregeln den unterschiedlichen Zuständen des beobachteten Phänomens (Merkmalen) systematisch und eindeutig Zahlen (Messwerte) zuzuordnen. Das Ergebnis sind quantitative Daten, die bestimmten Skalenniveaus zugeordnet und anschließend statistisch ausgewertet werden können.

1.2 Skalenniveaus

Die erhobenen Daten werden auf unterschiedlichen Mess- oder Skalenniveaus erfasst. Welches Niveau vorliegt, wird durch das empirische Relativ bestimmt. Skalenniveau ist daher umgangssprachlich übersetzbar mit „Informationsgehalt“. Es verrät uns, welche Aussagen wir beim Vergleich von zwei empirischen Phänomenen treffen können: Wie genau können wir diese unterscheiden? Welche Aussagen über deren Vergleich sind zulässig? Diese Eigenschaften werden dann in die Kodierung der Variablen übersetzt. Unterschieden wird in aufsteigendem Hierarchieverhältnis zwischen Nominalskala, Ordinalskala, Intervallskala und Ratioskala. Dabei handelt es sich bei den beiden ersten Skalenniveaus um sogenannte kategoriale Maßskalen, während die letzten beiden dem metrischen Meßniveau zugeordnet werden.

Die Nominalskala erlaubt es uns nur, zwischen gleich und ungleich zu unterscheiden. Lassen sich verschiedene Ausprägungen eines Phänomens in eine Rangordnung bringen, liegt zumindest ein ordinales Skalenniveau vor. Sind darüber hinaus die Abstände zwischen den Ausprägungen eindeutig bestimm- und damit interpretierbar, sind die Daten intervallskaliert. Intervallskalen haben aber keinen natürlichen Nullpunkt, dies ist erst bei der Ratio-

Tabelle 1: Skalenniveaus und deren Eigensachften

	Skalenniveau			
	Kategorial		Metrisch	
	Nominalskala	**Ordinalskala**	**Intervallskala**	**Ratioskala**
Erläuterung	Kategoriale Unterscheidung, ohne hierarchisches Verhältnis	Ausprägungen werden in Rangfolge gebracht, keine quantifizierbaren Abstände	Rangfolge der Merkmale mit gleichen Abständen, aber keinem natürlichen Nullpunkt	Natürlicher Nullpunkt vorhanden
Rechenoperationen	$=, \neq$	$<, >, \leq, \geq$	$+, -$	$\times, \div$
Beispiele	Geschlecht, Nationalität, Personennamen, Ortsbezeichnung	Schulabschluss, Soziale Schichtung, Bundesligatabelle	Temperatur in °C, Intelligenzquotient, Kalenderzeit	Temperatur in °K, Einkommen, Größe in m, Stimmenanteile in %

oder Verhältnisskala der Fall. Die Ausprägungen einer Ratioskala lassen sich daher zueinander und zum Nullpunkt in ein interpretierbares Verhältnis setzen. Grundsätzlich gilt, dass sämtliche Rechenoperationen eines niedrigeren Skalenniveaus auf ein höheres Skalenniveau angewendet werden dürfen. Auch dürfen höhere in niedrigere Skalenniveaus transformiert werden – jedoch nicht umgekehrt. In Tabelle 1 sind die angesprochenen Eigenschaften tabellarisch zusammengefasst und um einzelne Beispiele ergänzt, um einen Überblick über die verschiedenen Skalenniveaus zu erhalten

Die Unterscheidung verschiedener Skalenniveaus ist in der Statistik wichtig: Erstens bestimmt es, welche Transformationen im Rahmen einer Umkodierungen der Daten zulässig sind, ohne dass es zu einem Informationsverlust kommt. (Allerdings ist es unter Inkaufnahme von Informationsverlusten auch möglich, Daten von einem höheren Skalenniveau in ein niedrigeres umzuwandeln. Andersherum geht das natürlich nicht.) Zweitens, und für das weitere Arbeiten besonders wichtig, legt das Skalenniveau fest, welche statistischen Verfahren angewendet werden dürfen. Die verschiedenen statistischen Verfahren setzen bestimmte Dateneigenschaften hinsichtlich des Informationsgehaltes oder Messniveaus voraus. Als Daumenregel gilt: Je höher das Skalenniveau, desto mehr statistische Verfahren sind zulässig.

1.3 Teilbereiche der Statistik

Aus Sicht von Politikwissenschaftlerinnen und Politikwissenschaftlern ist Statistik eine Hilfswissenschaft, der in der empirischen Forschung aber eine ganz zentrale Funktion zukommt: Statistik hilft uns dabei, große Menge von quantitativen (bzw. quantifizierbaren) Informationen über politische und gesellschaftliche Phänomene nach wissenschaftlichen Kriterien zu verdichten und auszuwerten. Je nachdem, wie wir diese Daten statistisch auswerten, also welche Art analytischer Verfahren wir anwenden, unterscheiden wir zwischen deskriptiver, inferentieller und explorativer Statistik. Diesen verschiedenen Teilbereichen der Statistik liegen zum Teil ganz verschiedene Logiken und Verfahren zugrunde.

1.3.1 Deskriptive oder beschreibende Statistik

Wenn wir lediglich beschreibende Aussagen über die (entweder im Rahmen einer Vollerhebung oder einer Stichprobe erhobenen und) von uns analysierten Daten treffen, handelt es sich um deskriptive Statistik. Ziel der Deskription ist es meist, die Komplexität der Daten zu reduzieren, um aufschlussreiche und ggf. auch vergleichbare Informationen zu erhalten (bspw. über die Verteilung eines einzelnen Merkmals oder die gemeinsame Verteilung mehrerer Merkmale). Dafür werden die Daten strukturiert und bestimmte deskriptive Maß- oder Kennzahlen berechnet.

1.3.2 Inferentielle oder schließende Statistik

Häufig können wir in der Politikwissenschaft Daten nicht als Vollerhebung der Grundgesamtheit sondern nur in Form von Stichproben gewinnen. In den meisten Fällen wollen wir uns jedoch nicht auf Aussagen über die Stichprobendaten begrenzen, sondern Aussagen formulieren, die allgemeingültig sind, d.h. für alle Merkmalsträger der Grundgesamtheit zutreffen. Dies ist ein Einsatzbereich der Inferenzstatistik. Ein weiterer möglicher Einsatzbereich ist, wenn wir im Falle einer Vollerhebung annehmen, dass die bei der Datenerhebung entstehende Fehler zufallsverteilt sind. Dann suchen wir mittels statistischer Verfahren eine Absicherung der von uns beobachteten Phänomene und gefundenen Zusammenhänge. Die Inferenzstatistik umfasst die Summe der statistischen Schätz- und Testverfahren, die es uns erlauben – natürlich stets unter Inkaufnahme einer bestimmten Fehler- oder Irrtumswahrscheinlichkeit – Aussagen zu treffen, die über die von uns analysierten Daten hinausgehen.

1.3.3 Explorative oder strukturentdeckende Statistik

Explorative Statistik liegt dann vor, wenn wir innerhalb unserer vorliegenden Daten neue Strukturen oder Muster entdecken wollen. Diese Form der Statistik bezieht sich zunächst auch nur auf die Daten unserer Stichprobe. Zum Teil können neue Strukturen und Auffälligkeiten bereits aufgrund deskriptiver Kennzahlen vermutet werden. Der Kern der explorativen Statistik geht aber mit speziellen Analyseverfahren über die bloße Beschreibung der Daten hinaus. Dabei werden die Daten in Bezug zueinander gesetzt und so versucht, auffällige Verteilungen und damit ggf. neue Zusammenhänge aufzuspüren.

1.3.4 Uni-, bi- und multivariate Statistik

Die horizontale Unterscheidung zwischen deskriptiver, inferentieller und explorativer Statistik kann in vertikaler Richtung um die Unterscheidung zwischen uni-, bi- und multivariater Statistik ergänzt werden: Je nachdem, ob wir in unserer statistischen Auswertung nur ein einzelnes Merkmal oder den Zusammenhang zwischen zwei oder mehr als zwei Merkmalen betrachten, sprechen wir von uni-, bi- oder multivariater Statistik.

2 Einführung in Stata

In den folgenden Teilabschnitten werden das Öffnen des Programms über des Uninetzwerk [2.1], die Programmoberfläche [2.2], der Syntaxeditor [2.3], das Laden und Speichern von Datensätzen [2.4], sowie die Bearbeitung von Variablen [2.5] und Beobachtungen [2.6] vorgestellt und erläutert.

2.1 Öffnen von Stata

Zunächst sind drei Schritte notwendig, um aktiv dem hier vorgestellten Skript zu folgen und die Aufgaben sowohl innerhalb des Netzwerkes der Universität, als auch bequem von Zuhause aus zu bearbeiten.

1. WLAN-Zugang einrichten

   Eine ausführliche Beschreibung zur Aktivierung des WLAN-Zugangs befindet sich auf der Webseite des Universitätsrechenzentrums `http://www.rz.uni-greifswald.de/dienste/zugang-zum-uni-netz/eduroam.html`. Das Einrichten des WLAN-Zuganges ist notwendig, um innerhalb der Universität auf Stata zugreifen zu können. (*Ergänzung: In manchen Räumen der Universität und der Bibliothek gibt es LAN-Kabel und entsprechende Anschlüsse, die eine schnellere und stabilere Verbindung mit dem Universitätsnetzwerk herstellen lässt, als dies über WLAN der Fall ist.)

2. Stata einrichten/öffnen

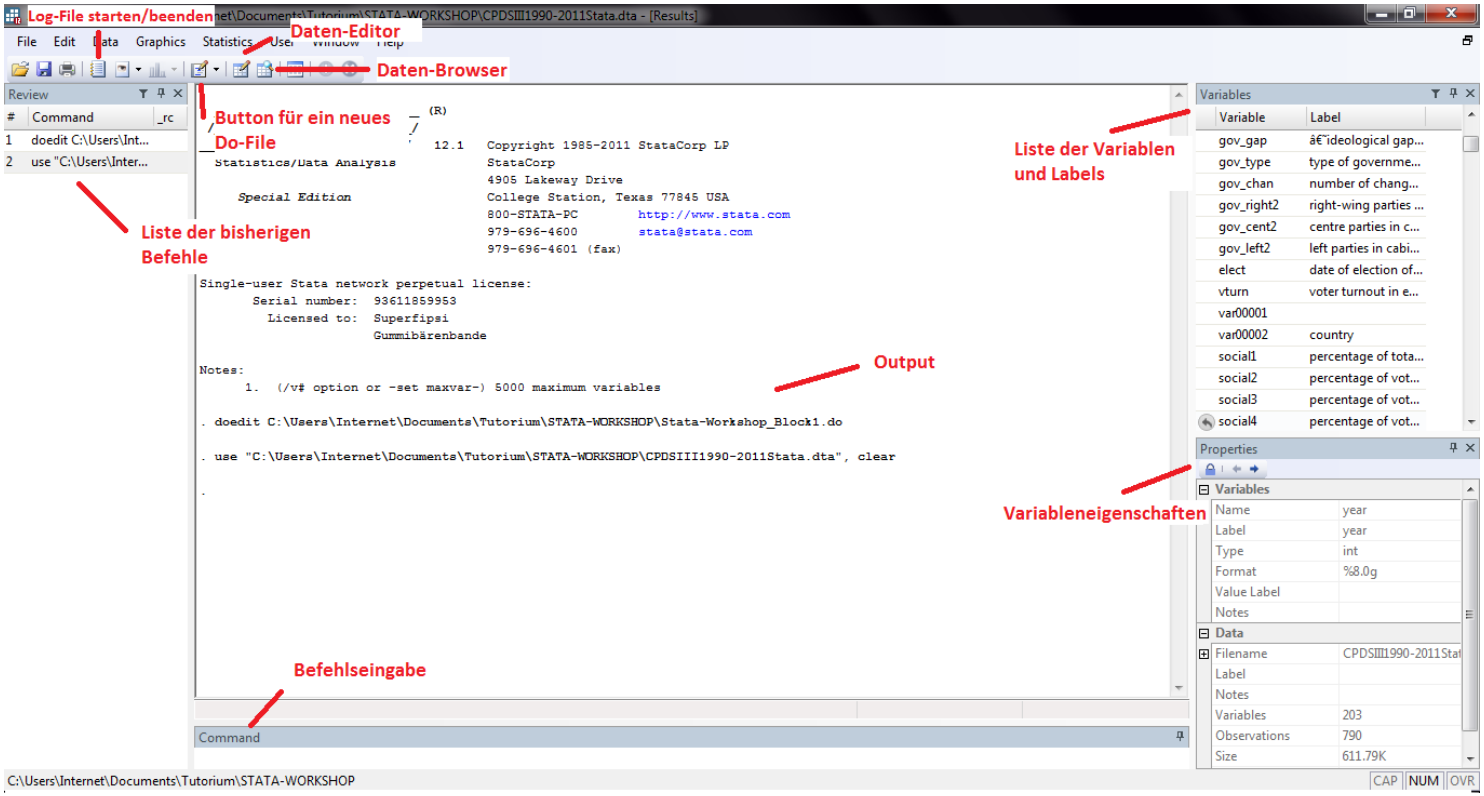
   Eine Erklärung zum Zugang zur Stata-Lizenz befindet sich ebenfalls auf der Seite des Rechenzentrums `http://www.rz.uni-greifswald.de/software/lizenzvertraege/stata.html`. Innerhalb dieses Skriptes wird die Stata-Version IC 13 genutzt, welche völlig ausreichend für die vorgenommenen Operationen ist (erst bei Analysen größerer Datensätze ist es sinnvoll auf die SE-Version zu wechseln). Es sind Versionen für unterschiedliche Betriebssysteme (Windows 32/64 Bit, Mac OS X, Linux) auswählbar.

3. VPN-Zugang einrichten

   Um Sata außerhalb des Uni-Netzwerkes zu nutzen (bspw. von Zuhause aus), ist es notwendig einen VPN-Zugang zu erstellen. Eine ausführliche Beschreibung zur Aktivierung des VPN-Clients findet sich hier: `http://www.rz.uni-greifswald.de/dienste/zugang-zum-uni-netz/vpn-zugang.html`.

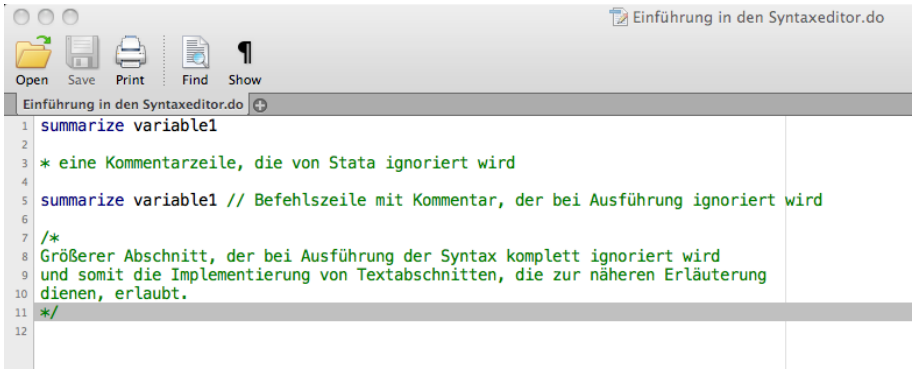
2.2 Die Programmoberfläche

Wenn wir Stata starten öffnet sich das Programmfenster, welches in mehrere Teilfenster aufgeteilt ist, die wiederum mit entsprechenden Werkzeugleisten versehen sind. Neben *Output*-Fenster, Befehlsleiste, sowie -historie, Variablenliste und -eigenschaften gibt es eine Menü- und Werkzeugleiste.



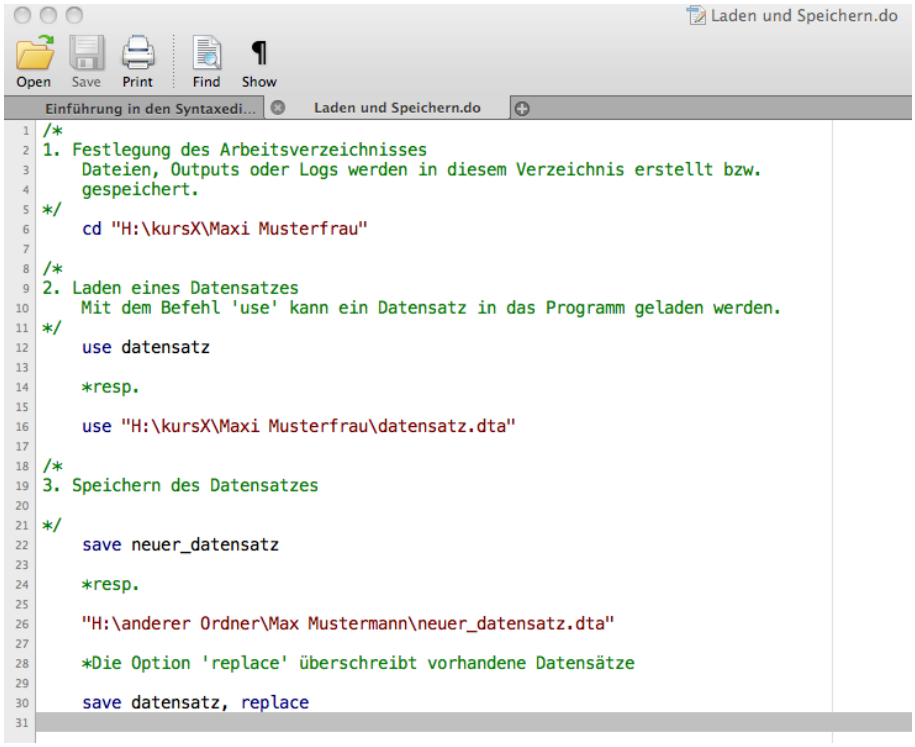
2.3 Der Syntaxeditor

Grundsätzlich enthält in Stata eine Zeile genau einen Befehl. Um die Syntax zu dokumentieren, können Zeilen mit einem Sternsymbol (*) auskommentiert werden. Stata ignoriert diese Kommentierungen bei Ausführung der entsprechenden Befehlszeile. Alternativ dazu lassen sich Kommentare in derselben Zeile durch zwei Schrägstriche (//) abgrenzen. Größere Abschnitte über mehrere Zeilen lassen sich mit /* resp. */ auskommentieren.

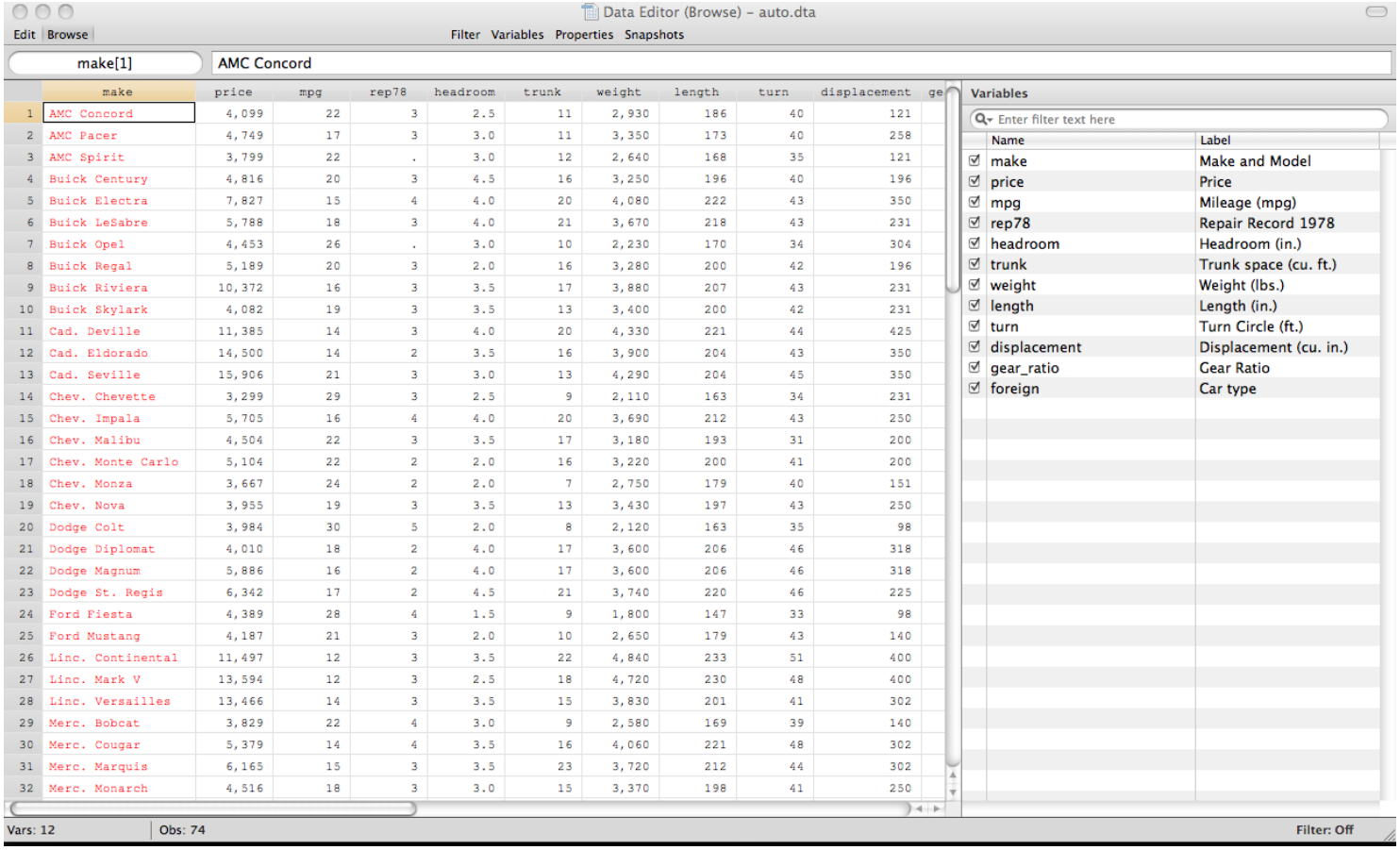


2.4 Laden und Speichern von Datensätzen

Um in Stata mit einem Datensatz zu arbeiten, muss dieser in das Programm geladen werden. Dazu wird zuerst der Pfad des Arbeitsverzeichnisses angegeben (Befehl 'cd', engl. *change directory*) und anschließend der Datensatz geladen. Alternativ dazu ist es auch möglich mit Pfadangaben zu arbeiten, ohne das Arbeitsverzeichnis zu wechseln.



Beispiel für einen geöffneten Datensatz in der Datensatzansicht (*browse*)



	make	price	mpg	rep78	headroom	trunk	weight	length	turn	displacement	ge
1	AMC Concord	4,099	22	3	2.5	11	2,930	186	40	121	
2	AMC Pacer	4,749	17	3	3.0	11	3,350	173	40	258	
3	AMC Spirit	3,799	22	.	3.0	12	2,640	168	35	121	
4	Buick Century	4,816	20	3	4.5	16	3,250	196	40	196	
5	Buick Electra	7,827	15	4	4.0	20	4,080	222	43	350	
6	Buick LeSabre	5,788	18	3	4.0	21	3,670	218	43	231	
7	Buick Opel	4,453	26	.	3.0	10	2,230	170	34	304	
8	Buick Regal	5,189	20	3	2.0	16	3,280	200	42	196	
9	Buick Riviera	10,372	16	3	3.5	17	3,880	207	43	231	
10	Buick Skylark	4,082	19	3	3.5	13	3,400	200	42	231	
11	Cad. Deville	11,385	14	3	4.0	20	4,330	221	44	425	
12	Cad. Eldorado	14,500	14	2	3.5	16	3,900	204	43	350	
13	Cad. Seville	15,906	21	3	3.0	13	4,290	204	45	350	
14	Chev. Chevette	3,299	29	3	2.5	9	2,110	163	34	231	
15	Chev. Impala	5,705	16	4	4.0	20	3,690	212	43	250	
16	Chev. Malibu	4,504	22	3	3.5	17	3,180	193	31	200	
17	Chev. Monte Carlo	5,104	22	2	2.0	16	3,220	200	41	200	
18	Chev. Monza	3,667	24	2	2.0	7	2,750	179	40	151	
19	Chev. Nova	3,955	19	3	3.5	13	3,430	197	43	250	
20	Dodge Colt	3,984	30	5	2.0	8	2,120	163	35	98	
21	Dodge Diplomat	4,010	18	2	4.0	17	3,600	206	46	318	
22	Dodge Magnum	5,886	16	2	4.0	17	3,600	206	46	318	
23	Dodge St. Regis	6,342	17	2	4.5	21	3,740	220	46	225	
24	Ford Fiesta	4,389	28	4	1.5	9	1,800	147	33	98	
25	Ford Mustang	4,187	21	3	2.0	10	2,650	179	43	140	
26	Linc. Continental	11,497	12	3	3.5	22	4,840	233	51	400	
27	Linc. Mark V	13,594	12	3	2.5	18	4,720	230	48	400	
28	Linc. Versailles	13,466	14	3	3.5	15	3,830	201	41	302	
29	Merc. Bobcat	3,829	22	4	3.0	9	2,580	169	39	140	
30	Merc. Cougar	5,379	14	4	3.5	16	4,060	221	48	302	
31	Merc. Marquis	6,165	15	3	3.5	23	3,720	212	44	302	
32	Merc. Monarch	4,516	18	3	3.0	15	3,370	198	41	250	

Bei den Zeilen handelt es sich um die einzelnen Fälle, die der Datensatz umfasst. Die Spalten stellen die verschiedenen Variablen dar. Daraus ergibt sich eine $N \times M$-Matrix – der Datensatz. Die einzelnen Zellen sind somit den Fällen zugewiesene Werte/Informationen. Wenn eine Information nicht verfügbar ist, dann ist dies durch einen Punkt gekennzeichnet (sog. *missing value*).

2.5 Bearbeiten von Variablen

Oftmals ist es der Fall, dass wir innerhalb von Datensätzen Variablen erstellen, umbenennen, umkodieren, usw. wollen. Die dazugehörigen Befehle funktionieren immer nach demselben Schema:

1. Wir teilen Stata mit, was wir machen wollen

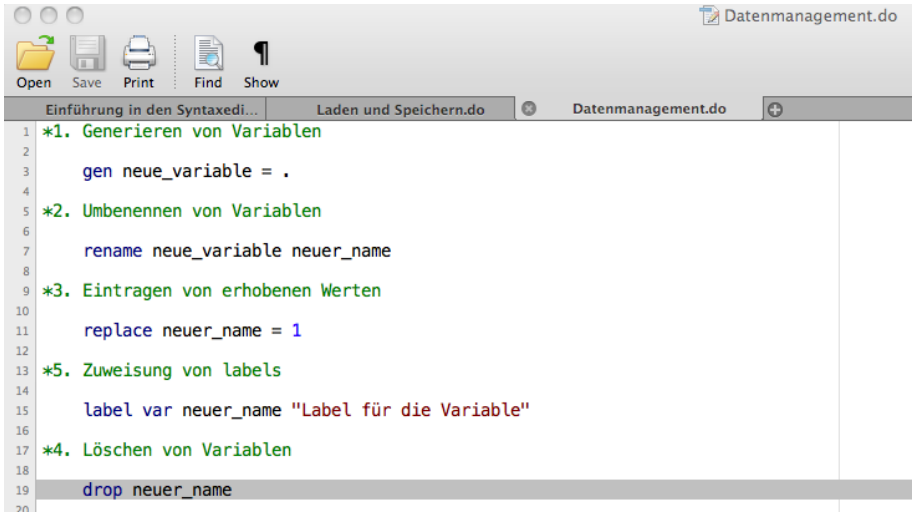
   rename

2. Wir geben an, worauf dieser Befehl angewandt wird

   rename **alter_variablenname**

3. Wir bestimmen das Resultat der Anwendung

   rename alter_variablenname **neuer_Variablenname**



2.6 Bearbeiten von Beobachtungen

Zusätzlich können Prä- und Suffixe verwendet werden, um Befehle zu spezifizieren und die durchzuführende Operation zu spezifizieren. Dies führt in der Konsequenz dazu, dass die Operation nicht für sämtliche Ausprägungen einer Variable, sondern nur für eine festgelegte Teilmenge der Variable durchgeführt wird.

1. Das **in**-Kommando

ifinby.do

```
*1. in-Kommando

    replace var1 = 5 in 1
    /*
    Dieser Befehl bewirkt, dass in der ersten Beobachtung von Variable
        var1 der Wert 5 gesetzt wird.
    */
/*
```

2. Die **if**-Bedingung

```
2. if-Bedingung

    Die if-Bedingung kann dafür genutzt werden, um über konditionelle Setzungen
        zu entscheiden, die für mehrere Werte der Variable zutreffen.
*/
    replace var1 = . if var1 == 99
    /*
    Dieser Befehl bewirkt, dass sämtliche Ausprägungen der Variable var1 mit
        dem Wert 99 auf fehlend ('missing', .) gesetzt werden. Dabei ist darauf
        zu achten, dass folgende Operatoren nach der if-Bedingung verwendet
        werden können:

        == '..ist-gleich..'/'..ist-identisch..'
        != '..ist-ungleich..'/'..nicht-identisch..'
        >, >= '..größer..', '..größer-gleich..'
        <, <= '..kleiner..', '..kleiner-gleich..'

    Mehrere dieser if-Bedingungen können auch miteinander verknüpft werden. Dazu
        können folgende Operatoren benutzt werden:

        & und (Konjunktor)
        | oder (Adjunktor)
        ! nicht (Negator)

    Folgendes Beispiel zeigt die Adjunktion zweier Prädikate:
    */
    replace var1 = . if var1 == 99 | var1 == 98
```

3. Das **by**-Kommando

```
*3. by-Kommando

    by var2: replace var1 = var1*var2
    /*
    Das by-Kommando wird als Präfix an den eigentlichen Befehl vorangestellt,
        um einen Befehl mehrfach für unterschiedliche Ausprägungen einer
        Variable durchzuführen. Bspw. kann es dazu verwendet werden, um mit
        einzelnen Ländern oder Berufsgruppen innerhalb eines Datensatzes eine
        spezielle Operation durchgeführt werden - ohne das dies für jeden
        einzelnen Fall durchgeführt werden muss.
    */
```

3 Univariate Deskription

Bei der univariaten Deskription geht es darum, die Verteilung einer Variable zu beschreiben. Dafür werden bestimmte Kennzahlen berechnet, die die Charakteristika der Verteilung in einem numerischen Wert zusammenfassen. Klassischerweise interessieren wir uns

1. für typische Vertreter der Verteilung, die wir mittels Lagemaßen oder Maßen der zentralen Tendenz zusammenfassen,
2. die Unterschiedlichkeit der Merkmalsausprägungen in der Verteilung, die wir mit Hilfe von Streuungsmaßen darstellen,
3. sowie die Form der Verteilung, die wir mit Maßen der Gestalt erfassen.

Welche univariaten Kennzahlen jeweils berechnet und sinnvoll interpretiert werden können, ist abhängig vom Skalenniveau der interessierenden Variable. Im folgenden werden die Eigenschaften der univariaten Deskription von kategorialen [3.1] und metrischen Variablen [3.2], sowie deren Anwendung in Stata vorgestellt.

3.1 Kategoriale Variablen

Im Falle kategorialer Variablen bietet es sich zunächst an, die Häufigkeitsverteilung der Merkmalsausprägungen einer Variable zu betrachten (im Falle metrischer Variablen müssen die Ausprägungen erst zu Klassen gruppiert werden, um eine sinnvoll interpretierbare Häufigkeitsverteilung zu erhalten). Die absolute Häufigkeit gibt zunächst an, wie viele der Merkmalsträger eine bestimmte Merkmalsausprägung aufweisen. Da die reine Angabe der Anzahl der Fälle mit gleicher Merkmalsausprägung ohne Bezug auf die Gesamtfallzahl zu Vergleichen nicht taugt, wird in der Regel die relative Häufigkeit bestimmt. Darunter verstehen wir den Anteil oder Prozentsatz der Merkmalsträger in einer bestimmten Kategorie. Schließlich lassen sich kumulierte Häufigkeiten berechnen, in dem die absoluten oder relativen Häufigkeiten mehrerer Ausprägungen addiert werden. Diese sind besonders dann von Interesse, wenn sich die Ausprägungen einer Variable ordnen lassen, oft also erst ab ordinalem Skalenniveau.

In Stata lassen sich absolute, relative und kumulierte Häufigkeiten einer Variable mittels folgendem Befehl berechnen und tabellarisch aufbereiten:

```
tabulate [variable]
```

```
. tabulate v8

   erhebungsgebiet
<wohngebiet>: west
            - ost |      Freq.     Percent        Cum.
------------------+-----------------------------------
alte bundeslaender |      2,358       67.76       67.76
neue bundeslaender |      1,122       32.24      100.00
------------------+-----------------------------------
             Total |      3,480      100.00
```

(Exkurs: Stata erlaubt es uns mit dem Befehl tabout, direkt publikationsfähige Tabellen zu erzeugen und in ein von uns gewähltes Datenformat zu exportieren, bspw.:

```
tabout [variable] using [Dateiname].txt, replace,

tabout [variable] using [Dateiname].xls, replace,
```

```
tabout [variable] using [Dateiname].rtf, replace,
```

Häufigkeiten lassen sich natürlich mit Stata nicht nur tabellarisch, sondern auch graphisch aufbereiten. Um die übliche Darstellungsform der Häufigkeitsverteilung kategorialer Variablen, das Balken- oder Säulendiagramm zu erzeugen, ist zunächst das Zusatzprogramm catplot.ado herunterzuladen:

```
ssc install catplot
```

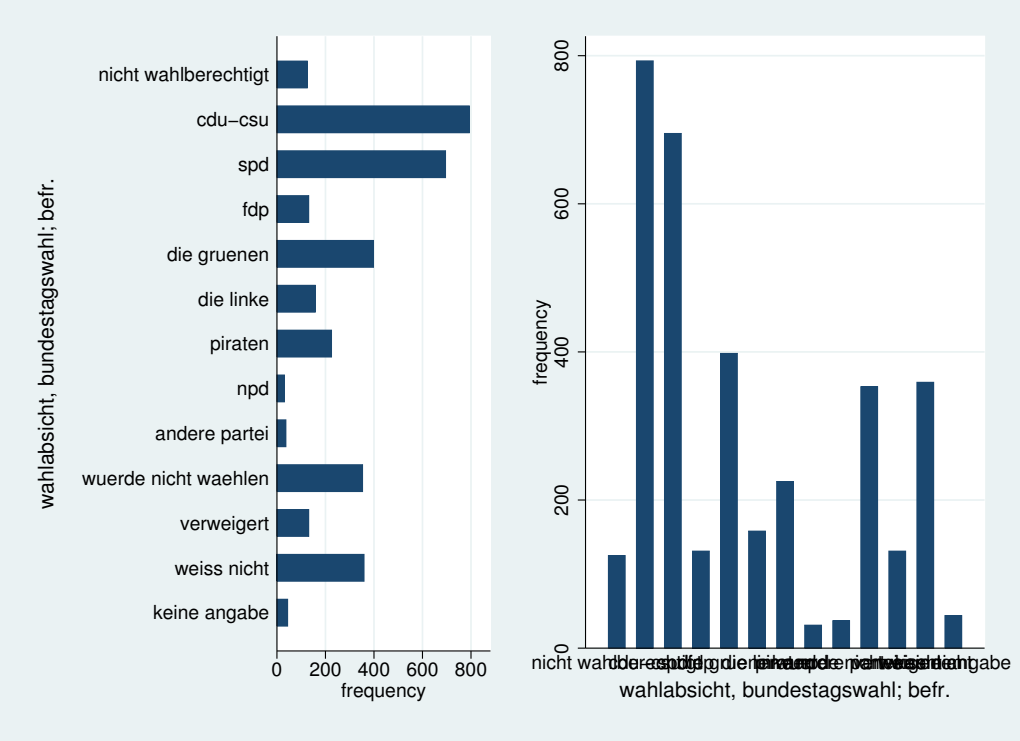
Dann kann mit dem Befehl

```
catplot [variable]
```

ein Balkendiagramm erzeugt werden, mit

```
catplot [variable], recast(bar)
```

ein Säulendiagramm. Weitere Optionen sind percent (um relative Häufigkeiten zu erhalten) und missing, um auch die Häufigkeit fehlender Werte mit zu erfassen.



Das Lagemaß, das sich für alle Skalenniveaus sinnvoll berechnen und interpretieren lässt und sich daher als Kennzahl auch für kategoriale Variablen eignet, ist der Modus oder Modalwert. Dabei handelt es sich um die am häufigsten vorkommende Merkmalsausprägung. (Im Falle einer Variablen mit vielen Ausprägungen, bspw. bei metrischem Skalenniveau, ist es i.d.R. vor der Bestimmung des Modalwertes sinnvoll, die einzelnen Ausprägungen zu gruppieren.) Gibt es nur einen Modalwert, sprechen wir von einer unimodalen Verteilung; sind hingegen mehrere Kategorien gleichhäufig besetzt, von einer multimodalen Verteilung. Der Modus lässt/die Modalwerte lassen sich entweder einfach aus der Häufigkeitstabelle ablesen oder mittels folgender Syntax in Stata berechnen:

```
egen modus=mode([variable])
replace modus = . if modus!=[variable]
table modus
```

```
. egen modus=mode(v506)

. replace modus = . if modus!=v506
(2687 real changes made, 2687 to missing)

. table modus

----------------------
   modus |      Freq.
---------+------------
       1 |        793
----------------------
```

Streuungsmaße für kategoriale Variablen bzw. ab nominalem Skalenniveau zu berechnen, ist eher unüblich. In der Regel wird die Homo- bzw. Heterogenität der Verteilung der Fälle auf die Kategorien einfach aus der Verteilungstabelle oder dem Diagramm abgelesen. In Anlehnung an entsprechende Maße für metrische Variablen gibt es aber zwei Maßzahlen: den Index qualitativer Variation (IQV) oder die Devianz. Mit diesen Kennzahlen können wir komprimiert darstellen, ob die Kategorien einer Variable (einigermaßen) gleichmäßig besetzt sind (maximale Streuung) oder ob die Fallzahlen variieren und auf bestimmte Merkmalsausprägungen besonders viele oder gar alle Fälle entfallen.

Ab ordinalem Skalenniveau lässt sich als Lagemaß der Median berechnen (Voraussetzung ist, dass sich die Datenmenge nach Größe/Höhe der Merkmalsausprägung ordnen lässt). Der Median ist der Wert einer geordneten Datenmenge, oberhalb und unterhalb dessen je 50% der Merkmalswerte liegen. Er wird auch als mittlerer Wert (oder 50%-Quantil) bezeichnet. Ist die Fallzahl ungerade, gibt es genau einen mittleren Merkmalswert. Bei einer geraden Fallzahl wird der Median als arithmetisches Mittel (s.u.) der beiden mittleren Messwerte berechnet. Mit Stata lässt sich der Median folgenderweise ermitteln:

```
tabstat [variable], statistics (med)
```

```
. tabstat v220, statistics (med)

    variable |       p50
-------------+----------
        v220 |        50
----------------------
```

Zudem lässt er sich auch aus der Menge der Informationen herauslesen, die nach dem Befehl summarize [variable], detail von Stata ausgegeben wird (s.u.).

3.2 Metrische Variablen

Für metrische Variablen lässt sich (zusätzlich zu den vorgenannten Maßzahlen) als Lagemaß das arithmetische Mittel berechnen (umgangssprachlich auch als Durchschnitt bezeichnet). Es berechnet sich als Summe aller Merkmalswerte, dividiert durch die Fallzahl. Im Vergleich zu Modus und Median ist das arithmetische Mittel ausreißeranfällig (um der Verzerrung durch Ausreißer zu begegnen, kann das arithmetische Mittel „getrimmt“ werden, indem ein bestimmter Anteil der jeweils kleinsten und größten Werte nicht in die Berechnung einbezogen wird). Folgende Eigenschaften des arithmetischen Mittels sind für weitere statistische Kennzahlen (→ Varianz) und Verfahren (→ lineare Regression) von zentraler Bedeutung: Die Summe der Differenzen der einzelnen Messwerte zum arithmetischen Mittel beträgt Null, die Summe der quadrierten Abweichungen ist minimal (also kleiner als die Summe der

quadrierten Abweichungen von allen anderen (hypothetisch möglichen) Ausprägungen). In Stata lässt sich das arithmetische Mittel mit folgenden Befehlen ermitteln:

```
mean [variable]
```

```
. mean v220

Mean estimation                   Number of obs    =      3480

--------------------------------------------------------------
             |       Mean   Std. Err.      [95% Conf. Interval]
-------------+------------------------------------------------
        v220 |   51.65431   .7811583       50.12274   53.18589
--------------------------------------------------------------
```

Mit dem Befehl summarize [variable] werden zudem auch der geringste in der Verteilung realisierte Wert der Variable, das Minimum, oder der höchste Wert, das Maximum ausgeben. Zudem gibt der Befehl ein Streuungsmaß für metrisch skalierte Variablen aus, die Standardabweichung. Die Standardabweichung berechnet sich aus der Varianz. Letztere ergibt sich, wenn die Summe der quadrierten Abweichungen der einzelnen Messwerte vom arithmetischen Mittel (s.o.) durch die Fallzahl dividiert wird. Die Varianz gibt zwar Auskunft über die Heterogenität der Messwerte einer Verteilung, ist aber aufgrund der Quadrierung inhaltlich schwer zu interpretieren, da der Bezug zur ursprünglichen Maßeinheit fehlt. Um dem zu begegnen, wird die Wurzel aus der Varianz gezogen, es ergibt sich die Standardabweichung. Diese lässt sich (näherungsweise) als mittlere oder durchschnittliche Abweichung der Messwerte vom arithmetischen Mittel interpretieren. Wenn die Daten einer Stichprobe „normalverteilt“ sind, gilt, dass in dem Intervall Mittelwert +/- 1 Standardabweichung 68% aller Daten liegen, in dem Intervall Mittelwert +/- 2 Standardabweichung 95% aller Daten und in dem Intervall Mittelwert +/- 3 Standardabweichung 99% aller Daten.

```
summarize [variable]
```

```
. summ v220

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        v220 |      3480    51.65431    46.08172        18        999
```

Mit dem Befehl summarize [variable], detail können wir uns noch weitere Kennzahlen der Verteilung der Variable ausgeben, so bspw. verschiedene Perzentile und Quartile (unter anderem auch den Median), neben der Standardabweichung auch die Varianz, sowie die beiden wichtigsten Maße der Gestalt: Schiefe und Wölbung. Mit der Schiefe wird die Abweichung der Verteilung von einer idealtypischen symmetrischen Verteilung erfasst: Bei rechtsschiefen (auch als linkssteil bezeichnet) Verteilungen entfallen mehr Fälle auf geringe Merkmalsausprägungen, die Maßzahl für die Schiefe nimmt positive Werte an; bei linksschiefen (auch als rechtssteil bezeichnet) Verteilungen weisen mehr Beobachtungen einen hohen Variablenwert auf, abzulesen an einem negativen Schiefewert. Die Wölbung gibt Auskunft darüber, ob auf mittlere Werte sehr viele Fälle entfallen, was eine einer schmalgipfligen, hochgewölbten Verteilung entspricht (einhergehend mit Werten über 3 für die Kennzahl für die Wölbung), oder ob die Fälle sich (gleichmäßiger) über einen breiteren Wertbereich verteilen (breitgipflige, gering gewölbte Verteilung, einhergehend mit Wölbungswerten unter 3.

```
summarize [variable], detail
```

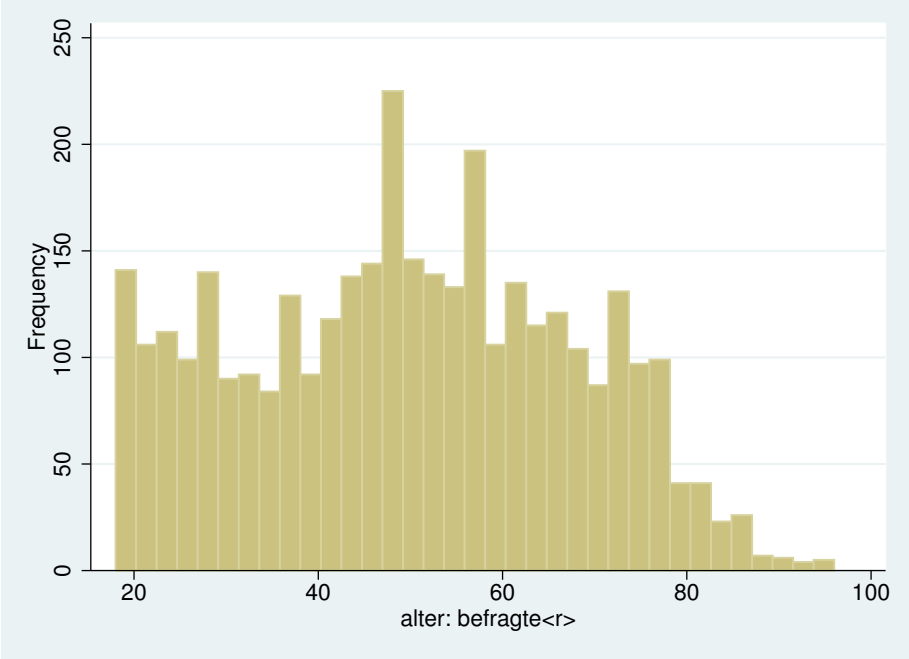
```
. summ v220, detail

                    alter: befragte<r>
-------------------------------------------------------------
      Percentiles      Smallest
 1%           19            18
 5%           21            18
10%           24            18       Obs                3480
25%           36            18       Sum of Wgt.        3480

50%           50                     Mean              51.65431
                      Largest         Std. Dev.         46.08172
75%           63           999
90%           74           999       Variance          2123.525
95%           78           999       Skewness          17.46735
99%           86           999       Kurtosis          359.5383
```

Die Verteilung der Merkmalsausprägungen metrische Variablen lassen sich ebenfalls graphisch aufbereiten. Eine klassische Darstellungsweise der Häufigkeitsverteilung metrischer Variablen ist das Histogramm (Achtung: Das Histogramm ist nicht zu verwechseln mit dem Balken- oder Säulendiagramm. Im Gegensatz zum Balken- oder Säulendiagramm schließen beim Histogramm die Werteklassen direkt aneinander, zudem gilt, dass die Fläche proportional zur Häufigkeit der Klassen ist). In Stata lässt sich ein Histogramm mit dem Befehl

```
histogram [variable], freq
```



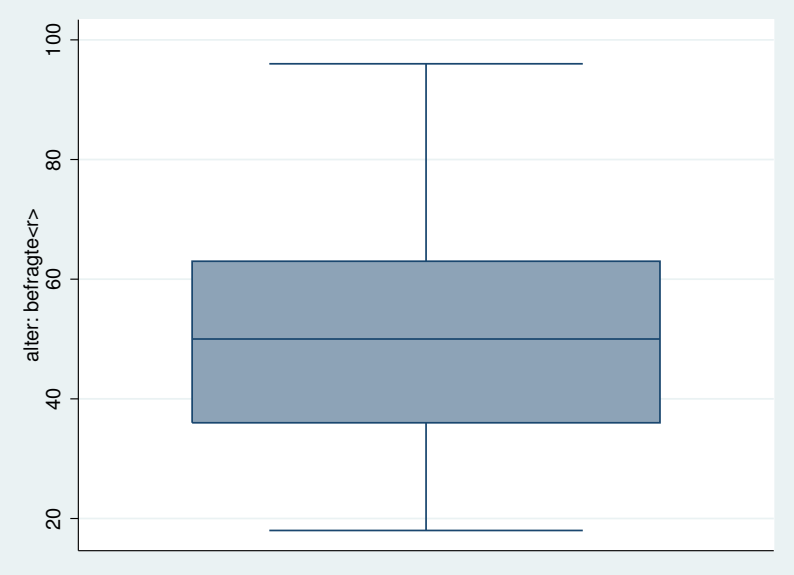
erzeugen und anschließend mittels

```
graph export [Dateiname].[Dateiendung/-format]
```

in ein gewünschtes Dateiformat exportieren um es später in Texte oder Präsentationen einzufügen (mögliche Dateiformate u.a.: .emf, .eps, .pdf, .png oder .tiff; am besten geeignet sind die beiden ersteren, da es sich hierbei um Vektorgraphiken handelt, denen bei einer Reskalierung keine Verzerrung/Verpixelung widerfährt).

Eine weitere graphische Aufbereitungsmöglichkeit der Verteilung metrischer Variablen ist der Boxplot. Dabei werden die mittleren 50% der Fälle durch eine Box abgebildet; deren Untergrenze entspricht folgerichtig dem 25%-Quantil und deren Obergrenze dem 75%-Quantil. Die Linie in der Mitte ist der Median (das 50%-Quantil). Die von der Box ausgehenden Linien verbinden diese mit dem Minimum und dem Maximum, vorausgesetzt diese liegen nicht mehr als das 1,5-fache der Boxenlänge von den Quantilsgrenzen entfernt. Weiter entfernt liegende Fälle werden gesondert gekennzeichnet: Kreise symbolisieren sog. Ausreißer (1,5-3fache Distanz), Sternchen sog. Extremwerte (über 3-fache Distanz). Boxplots können gut bei der explorativen Datenanalyse sowie bei der vergleichenden Deskription von Variablen und Verteilungen eingesetzt werden. Der Stata-Befehl zum Erzeugen eines Boxplots lautet:

```
graph box [variable]
```



4 Inferenzstatistik: Grundlagen und Anwendung auf univariate Kennzahlen

Bisher haben wir nur deskriptive Statistiken behandelt, die dazu dienen, die Verteilung der Variablen in den von uns analysierten Daten zu beschreiben. Dies mag im Falle von Vollerhebungen ausreichend sein. Wenn wir aber mit Stichprobendaten arbeiten oder annehmen, dass der Prozess der Datenerhebung im Falle einer Vollerhebung einem zufallsverteilten Fehler unterliegt, brauchen wir statistische Verfahren, die uns eine Absicherung oder Generalisierung der Aussagen erlaubt, die wir auf Basis der Kennzahlen treffen können. Diese Verfahren bietet die Inferenzstatistik. Sie umfasst eine Summe verschiedener Schätz- und Testverfahren, die es uns erlauben sollen, Aussagen zu treffen, die über die von uns analysierten Daten hinaus gehen. Zunächst wird eine vereinfachte Einführung in die Inferenzstatistik [4.1] gegeben und anschließend das statistische Schätzen am Beispiel der Konfidenzintervalle [4.2] aufgezeigt. Abschließend wird das statistische Schätzen und dessen Anwendung in Stata [4.3] vorgestellt

4.1 Vereinfachte Einführung in die Inferenzstatistik

Im folgenden soll erklärt werden, was Wahrscheinlichkeitstheorie und Stichprobenverteilungen mit statistischem Schätzen und Testen zu tun hat. Der Prozess des Schätzens und Testens unterliegt einer gewissen Irrtumswahrscheinlichkeit. Die Wahrscheinlichkeitstheorie erlaubt es, mit dieser Irrtumswahrscheinlichkeit analytisch umzugehen. Von zentraler Bedeutung sind Stichprobenverteilungen bzw. Wahrscheinlichkeitsverteilungen. Aus der Wahrscheinlichkeitsrechnung ist bezüglich des Verhältnisses von Zufallsexperimenten und empirischer Realität bekannt, dass die Merkmalsausprägungen, die man in einzelnen Zufallsexperimenten beobachtet, durchaus von der Verteilung der Merkmalsausprägungen in der Realität abweichen können. Wird das Zufallsexperiment aber (näherungsweise) unendlich oft wiederholt, ist zu beobachten, dass in der Mehrzahl der Zufallsexperimente die Verteilung der Merkmale derjenigen in der Realität gleicht oder stark ähnelt. Abweichende Verteilungen sind weniger häufig, wobei sie umso seltener sind, je stärker die Verteilung von der Realität abweicht. Daraus ergibt sich eine typische Häufigkeitsverteilung von Zufallsexperimenten. Die Eigenschaften von Zufallsverteilungen nutzen wir für das statistische (parametrische) Schätzen und Testen, da wir Stichprobenziehungen (und ggf. auch die Fehlerverteilung bei einer Vollerhebung) als Resultat eines Zufallsexperimentes betrachten. Aufgrund des Wissens über Zufallsverteilungen können wir erwarten, dass es wahrscheinlicher ist, einen Wert in unseren Daten zu haben, der dem in der Realität ähnlich ist, als einen, der stark von diesem abweicht. Darauf aufbauend können wir unter Nutzung der sich aus den Stichprobenverteilungen ableitenden Kennwerteverteilungen und ihrer Eigenschaften mit dem Irrtum bzw. Fehler umgehen, den wir beim statistischen Schätzen und Testen begehen werden. Die zugrunde zu legende Kennwerteverteilung bestimmt sich durch die Art der zu schätzenden oder testenden Kennzahl und die Anzahl der Freiheitsgrade. Im Folgenden werden nun die gängigsten Schätz- und Testverfahren beschrieben, die auf dieser Logik basieren, und ihre Anwendung anhand univariater Kennzahlen veranschaulicht. Dies ist dann auf die inferenzstatistische Behandlung bi- und multivariater Analysen zu übertragen.

4.2 Statistisches Schätzen: Das Konfidenzintervall

Wenn wir auf Basis eines Stichprobenwertes (bspw. eines Anteilswertes für die Merkmalsausprägung einer kategorialen Variable oder eines Mittelwertes einer stetigen Variable) den realen Wert in der Grundgesamtheit schätzen wollen, können wir dies nicht mit 100 %iger Sicherheit, sondern nur einer bestimmten Irrtumswahrscheinlichkeit tun. Daher berechnen wir ein Werteintervall um die Stichprobenkennzahl, bei dem wir davon ausgehen können, dass dieses den wahren Wert der Grundgesamtheit mit einer bestimmten Wahrscheinlichkeit überdeckt, das sogenannte Konfidenzintervall. Je größer der Bereich des Konfidenzintervalls, desto sicherer (aber auch ungenauer) kann die Lage des tatsächlichen Populationswertes bestimmt werden. Die Intervallgrenzen berechnen sich aus dem Stichprobenwert $\pm$ dem Produkt aus dem Standardfehler des Stichprobenwertes und dem $(1-\alpha/2)$-Quantil der Kennwerteverteilung (wobei α der von uns als maximal akzeptabel angesehen Irrtumswahrscheinlichkeit entspricht).

Mithilfe des Befehls ci lassen sich die Konfidenzintervalle, die den Bereich des wahren Polulationswertes abdecken, sowohl für binomiale als auch normalverteilte Kennwerteverteilungen ermitteln. Für **kategoriale** Variablen:

```
ci [variable], binomial [level(#)]
```

```
. tabulate v151, gen(geschlecht_)

GESCHLECHT,
BEFRAGTE<R> |      Freq.     Percent        Cum.
------------+-----------------------------------
       MANN |      1,712       49.35       49.35
       FRAU |      1,757       50.65      100.00
------------+-----------------------------------
      Total |      3,469      100.00

. ci geschlecht_1, binomial level(95)

                                                 -- Binomial Exact --
    Variable |       Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
geschlecht_1 |      3469    .493514    .0084885        .4767441    .5102948
```

In dem Beispiel wurde die Variable v151, die Informationen über das Geschlecht der/des Befragten enthält, mithilfe des tabulate-Befehls in zwei dichotome Variablen (geschlecht_1=mann; geschlecht_2=frau) zerlegt. Anschließend wurde das Konfidenzintervall für die Eintrittswahrscheinlichkeit des Merkmals *mann* ermittelt. Es zeigt sich, dass wir mit einer Sicherheit von 95 % davon ausgehen können, dass das Intervall von etwa 0,47 bis 0,51 die tatsächliche Eintrittswahrscheinlichkeit in der Population überdeckt.
Für **metrische** Variablen:

```
ci [variable], [level(#)]
```

```
. ci v629, level(95)

    Variable |       Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
        v629 |      1673    171.1662    .2220611        170.7306    171.6017
```

In dem Beispiel wurden die Konfidenzintervalle für die Körpergröße in cm aller Befragten ermittelt. Es zeigt sich, dass wir mit einer Sicherheit von 95 % davon ausgehen können, dass das Intervall von ca. 170,7 cm bis 171,6 cm den wahren Populationswert (die durchschnittliche Körpergröße aller Deutschen) überdeckt. Wenn wir die Vertrauenswahrscheinlichkeit unserer Schätzung ändern, dann verändern sich folglich auch die Konfidenzintervalle. Dabei gilt: je größer die Vertrauenswahrscheinlichkeit, desto größer das Intervall.

```
. ci v629, level(90)

    Variable |       Obs        Mean    Std. Err.       [90% Conf. Interval]
-------------+---------------------------------------------------------------
        v629 |      1673    171.1662    .2220611        170.8007    171.5316

. ci v629, level(99.9)

    Variable |       Obs        Mean    Std. Err.      [99.9% Conf. Interval]
-------------+---------------------------------------------------------------
        v629 |      1673    171.1662    .2220611        170.4342    171.8982
```

Wie an den beiden Berechnungen zu erkennen ist, beträgt die Spannweite des Konfidenzintervalls in der ersten Berechnung bei einer gewählten Vertrauenswahrscheinlichkeit von 90 % etwa 0,73 cm (also weniger als ein Zentimeter für die gesamte Population). Jedoch ist die Wahrscheinlichkeit, dass wir uns bei der Bestimmung des Intervalls für den Populationswert geirrt haben, verglichen mit der vorigen Schätzung, mit 10 % nun doppelt so hoch.

Wenn wir bei unserer Schätzung des Intervalls aber nun sehr sicher sein wollen, dann muss die Vertrauenswahrscheinlichkeit höher gewählt werden. In dem unteren Beispiel ist die Vertrauenswahrscheinlichkeit auf 99,9 % angehoben worden. Die Wahrscheinlichkeit, dass man mit der Schätzung des Intervalls nun nicht den Populationswert abdeckt, beträgt nun 0,1 %. Allerdings ist die Spannweite des Intervalls mit ca. 1,46 cm nun auch doppelt so hoch.

4.3 Statistisches Testen

Wir wollen mit statistischen Verfahren jedoch nicht nur mithilfe von Kennwerten, die wir mit unseren Daten berechnet haben, wahre Merkmalswerte schätzen. In der Regel haben wir im Vorfeld theoretische Annahmen über bestimmte Merkmale der uns interessierenden Population bzw. über Zusammenhänge zwischen zwei oder mehreren dieser Merkmale getroffen, die wir inferenzstatistisch testen wollen. Auch hier gilt, dass wir nicht mit 100 %iger Sicherheit feststellen können, ob unsere Annahmen richtig oder falsch sind. Wir können lediglich mit einer bestimmten Irrtumswahrscheinlichkeit feststellen, ob unserer empirische Beobachtung unsere Erwartung eher stützt oder nicht. Um damit formalisiert umgehen zu können, formulieren wir unsere theoretische Erwartung als Arbeitshypothese und stellen dieser die sogenannte Nullhypothese, die Negierung der Arbeitshypothese, entgegen. Wir legen dann fest, mit welcher Wahrscheinlichkeit wir bereit sind, den sogenannten α-Fehler zu begehen. Dieser besteht darin, die Nullhypothese aufgrund der Stichprobenbeobachtung fälschlicherweise abzulehnen, obwohl diese in der Grundgesamtheit gilt. Statistische Tests basieren immer auf dem α-Fehler.

Für kategoriale Variablen erfolgt der (ein- und zweiseitige) Test auf binomialverteilte Wahrscheinlichkeit des Eintrittes mithifle des bitest-Befehls:

```
bitest [variable] = [angenommene Wahrscheinlichkeit]
```

```
. tabulate v217, gen(geschl_)

geschlecht,
befragte<r>      Freq.     Percent        Cum.
------------------------------------------------
       mann      1,725       49.57       49.57
       frau      1,755       50.43      100.00
------------------------------------------------
      Total      3,480      100.00

. bitest geschl_1 = 0.5

   Variable |        N   Observed k   Expected k   Assumed p   Observed p
------------+-------------------------------------------------------------
   geschl_1 |     3480         1725         1740     0.50000     0.49569

  Pr(k >= 1725)                 = 0.700379  (one-sided test)
  Pr(k <= 1725)                 = 0.311506  (one-sided test)
  Pr(k <= 1725 or k >= 1755)    = 0.623011  (two-sided test)
```

Es werden Teststatistiken für ein- und zweiseitige Tests ausgegeben. Dabei ist es entscheidend, dass der emprische Wert (1725 Männer) gegen den Erwarungswert (bei einer vermuteten Eintrittswahrscheinlichlichkeit von 0,5 bei 3480 Fällen beträgt dieser 1740) getestet wird. Die zu testende Nullhypothese ist stets konservativ gewählt und besagt, dass kein Unterschied zwischen angenommener und empirischer Eintrittswahrscheinlichkeit besteht. Die jeweiligen Tests befinden sich im unteren Teil des Outputs und geben die Richtung des Tests an. In unserem Beispiel muss die Nullhypothese sowohl für die einseitigen Tests ($k \geq 1725$; $k \leq 1725$), als auch für den zweiseitigen Test ($k \leq 1725 \, or \, k \geq 1755$) angenommen werden. Daraus folgt, dass die Gesamtabweichung von 15 Männern weniger in der Stichprobe aufgrund von stochastischen Eigenschaften zustande gekommen sein muss.

Für metrische Variablen erfolgt der (ein- und zweiseitige) Mittelwertvergleich über den ttest-Befehl:

```
ttest [variable] = [angenommener Mittelwert]
```

```
. ttest v220 = 45

One-sample t test

Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    v220 |    3480    51.65431    .7811583    46.08172    50.12274    53.18589

    mean = mean(v220)                                         t =   8.5185
Ho: mean = 45                                 degrees of freedom =     3479

    Ha: mean < 45               Ha: mean != 45                 Ha: mean > 45
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Der erzeugte Output ist prinzipiell ähnlich zu dem Test auf binomialverteilte Wahrscheinlichkeit, jedoch mit dem Unterschied, dass nun Maßzahlen für metrische Variablen zum Einsatz kommen. Dabei werden arithmetisches Mittel, Standardfehler, Standardabweichung und Konfidenzintervall benutzt um diese gegen einen angenommenen Populationswert (in dem vorliegenden Fall das Durchschnittsalter von 45 Jahren) zu testen. Der vorliegende Test zeigt, dass das Durchschnittsalter innerhalb der Stichprobe 51,65 Jahre beträgt. Das sich aus Mittelwert und Standardfehler ergebene Konfidenzintervall erstreckt sich von 50,12 bis 53,19 Jahren. Da das von uns vermutete Durchschnittsalter von 45 Jahren nicht von diesem Intervall überdeckt wird, zeigen die Hypothesentests, dass die Nullhypothesen zur Ungleichheit der Mittelwerte ($mean! = 45$), sowie zur einseitigen Abweichung ($mean > 45$) des empirischen Wertes abgelehnt werden müssen. Daraus folgt, dass der Stichprobenmittelwert signifikant von einem angenommenen Durchschnittsalter von 45 Jahren abweicht.

Dieser Test kann auch verwendet werden um die Mittelwerte zweier Gruppen zu vergleichen. Beispielsweise kann man testen, ob das Durchschnittsgewicht beider Geschlechter signifikant voneinander abweicht. Dies kann man durch Spezifizierung einer Gruppierungsvariable erreichen:

```
ttest [variable], [by(variable)]
```

```
. ttest v631, by(v151)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    MANN |     789    83.24461    .4905215    13.77833    82.28173     84.2075
    FRAU |     805    68.25714     .458449    13.00735    67.35724    69.15704
---------+--------------------------------------------------------------------
combined |    1594    75.67566    .3843602    15.34556    74.92175    76.42956
---------+--------------------------------------------------------------------
    diff |            14.98747    .6710188                13.6713    16.30364

    diff = mean(MANN) - mean(FRAU)                            t =  22.3354
Ho: diff = 0                                  degrees of freedom =     1592

    Ha: diff < 0                Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Das vorliegende Ergebnis ist ähnlich dem des Tests bei einem vorgegebenen Wert zu betrachten, jedoch entstammen die Mittelwerte diesmal direkt den jeweiligen Gruppen. Es zeigt sich, dass das Durchschnittsgewicht von Männern bei 83,24 kg und das von Frauen bei 68,26 kg liegt. Die Kennwerte der gesamten Stichprobe lassen sich aus der Zeile *combined* entnehmen. Die zu testende Nullhypothese nimmt keinen Unterschied bei der Differenz der Mittelwert beider Gruppen an. Das Ergebnis ist, dass diese Nullhypothese verworfen werden muss. Die Konfidenzintervalle beider Gruppen überschneiden sich in dem Beispiel nicht. Deshalb kann davon ausgegangen werden, dass beide Populationsparameter tatsächlich unterschiedlich voneinander abweichen und diese Differenz von überzufälliger Natur ist.

5 Bivariate Zusammenhangsanalyse

Manchmal scheint es uns, als ob zwei Phänomene, die wir in der Realität beobachten, irgendwie zusammenhängen. Ob dies der Fall ist, können wir überprüfen, indem wir die gemeinsame Verteilung zweier Variablen analysieren. So lässt sich feststellen, ob ein solcher Zusammenhang zwischen zwei Merkmalen überhaupt vorliegt, und ggf. auch beschreiben, wie stark und wie er gerichtet ist. Es gibt eine ganze Reihe verschiedener Zusammenhangsmaße. Welches Maß wir für die Analyse welcher Variablen wählen sollten, hängt zunächst vom Skalenniveau der Variablen ab. Das passende Zusammenhangsmaß berücksichtigt deren Messeigenschaften und stellt so sicher, dass wir die maximal mögliche Information über die Art des Zusammenhangs erhalten. Zusammenhangsmaße, die ein niedriges Skalenniveau voraussetzen, können für höher skalierte Daten verwendet werden (gehen dann allerdings mit einem Informationsverlust einher). Umgekehrt ist dies nicht zulässig. Weisen die zwei Variablen ein unterschiedliches Skalenniveau auf, bestimmt das niedrigere die Wahl des Zusammenhangsmaßes. Außerdem gibt es einige Maße, deren Anwendung voraussetzt, dass wir zwischen Ursache und Wirkung unterscheiden. Vor der Berechnung solch asymmetrischer Maße müssen wir zumindest ein einfaches theoretisches Erklärungsmodell formulieren und festlegen, was abhängige und was unabhängige Variable ist. Bei symmetrischen Maßen ist diese Unterscheidung zunächst nicht notwendig (kann aber im Rahmen der Interpretation durchaus sinnvoll sein). Schließlich ist zwischen standardisierten und nichtstandardisierten Maßen zu unterscheiden. Erstere haben einen definierten Wertebereich, so dass wir die Stärke des Zusammenhangs beurteilen können. Bei dichotomen und nominalen Daten liegt der Wertebereich zwischen 0 (kein Zusammenhang) und 1 (vollkommener Zusammenhang).

Liegt ordinales oder metrisches Skalenniveau vor, kann zusätzlich noch zwischen positivem und negativem Zusammenhang unterschieden werden: Ein Wert von –1 steht dann für einen perfekten negativen Zusammenhang, und ein Wert von +1 für einen perfekten positiven Zusammenhang. Je näher der Wert sich 0 annähert, desto schwächer die gemeinsame Verteilung der Werte, bei einem Wert von genau 0 liegt kein Zusammenhang vor. Messeigenschaften, Symmetrie und Standardisierung der Maße sind bei der Interpretation zu berücksichtigen. In Tabelle 2 werden die wichtigsten Zusammenhangsmaße und deren Eigenschaften zusammengefasst.

Die Aussagen, die wir auf Basis der Zusammenhangsmaße treffen, sind zunächst jedoch auf die analysierten Daten beschränkt. Uns als Forscherinnen und Forscher interessiert jedoch in der Regel auch, ob wir den Befund generalisieren können. Daher werden die Zusammenhangsmaße um statistische Tests, die uns sagen, ob der Zusammenhang auf einem bestimmten Niveau signifikant ist, ergänzt. Im Rahmen dieses Skriptes können wir nicht die Masse an Zusammenhangsmaßen und Testverfahren behandeln, die die Statistik kennt (ein Überblick findet sich auf den folgenden zwei Seiten). Stattdessen werden wir uns im Folgenden auf einige wenige, sehr verbreitete Maße für nominalskalierte [5.1], ordnialskalierte [5.2] und metrischskalierte Variablen [5.3] beschränken und deren Anwendung sowie Berechnung mit Stata beispielhaft beschreiben.

Tabelle 2: Übersicht über bivariate Zusammenhangsmaße und deren Eigenschaften

Zusammenhangsmaß	Symmetrie	Richtung	Unabhängige Variable	Abhängige Variable	Wertebereich	Stichworte zur Berechnung	Interpretation	Anmerkung
Chi-Quadrat χ^2	Ja	Nein	Dichotom	Dichotom	0 bis ∞	Vergleich von Kontingenz- und Indifferenztabelle	Abweichung von statistischer Unabhängigkeit	Abhängig von der Fallzahl; absoluter Wert hat mitunter wenig Aussagekraft
Phi ϕ	Ja	Nein	Dichotom	Dichotom	0 bis 1	Auf χ^2 basierend	Ebenso	-
Kontingenzkoeffizient C	Ja	Nein	Nominal	Nominal	0 bis 1	Auf χ^2 basierend	Ebenso	-
Cramérs V	Ja	Nein	Nominal	Nominal	0 bis 1	Auf χ^2 basierend	Ebenso	Entspricht bei einer 2x2-Kreuztabelle dem Phi-Koeffizienten
Odds Ratio	Ja	Ja	Dichotom	Dichotom	0 bis ∞	Verhältnis der Häufigkeiten des Eintretens und Nicht-Eintretens eines Ereignisses	Chance, wie viel häufiger das Ereignis in der einen Gruppe auftritt als in einer anderen	Der Wert 1 steht für statistische Unabhängigkeit
Prozensatzdifferenz d	Nein	Nein	Dichotom	Dichotom	0 bis 100%	Größe der Subgruppendifferenz	-	-
Goodmans und Kruskals Lambda λ	Nein	Nein	Nominal	Nominal	0 bis 1	Kontingenztabelle und Schätzfehler für den Modus	PRE-Maß	Manchmal 0, trotz Zusammenhangs; verschiedene Ausprägungen möglich
Goodmans und Kruskals Gamma γ	Ja	Nein	Ordinal	Ordinal	0 bis 1	Vergleich konkordanter und diskonkordanter Paare	PRE-Maß	-
Kendalls Tau τ	Ja	Ja	Ordinal	Ordinal	$-$ bis 1	Vergleich konkordanter und diskonkordanter Paare	Positiv: mehrheitlich konkordante Paare; negativ: mehrheitlich diskonkordante Paare	Verschiedene τ-Koeffizienten möglich; auch als PRE-Maß geeignet
Somers d	Nein	Ja	Ordinal	Ordinal	-1 bis 1	Vergleich konkordanter und diskonkordanter Paare, Berücksichtigung verbundener Paare	-	-

Zusammenhangsmaß	Symmetrie	Richtung	Unabhängige Variable	Abhängige Variable	Wertebereich	Stichworte zur Berechnung	Interpretation	Anmerkung
Rangkorrelationskoeffizient/Spearmans rho ρ	Ja	Ja	Ordinal	Ordinal	-1 bis 1	Quadrierte Rangplatzdifferenzen	-	-
Effektstärke eta-Quadrat η^2	Nein	Nein	Beliebig	Metrisch	0 bis 1	Verhältnis der Summe der Abweichungsquadrate	PRE-Maß	Normalverteilung für metrische Variable vorausgesetzt
Kovarianz	Ja	Ja	Metrisch	Metrisch	$-\infty$ bis $+\infty$	Durchschnittliche Summe der Abweichungsprodukte aller Messpaarwerte	Gemeinsame Streuung zweier Merkmale, linearer Zusammenhangß	Maßstabsabhängig, Normierung erforderlich; absoluter Wert hat mitunter wenig Aussagekraft; Normalverteilung vorausgesetzt
Korrelationskoeffizient/ Pearsons r	Nein	Ja	Metrisch	Metrisch	-1 bis 1	Standardisierte Kovarianz	Linearer Zusammenhang	Normalverteilung vorausgesetzt
Determinationskoeffizient r^2	Nein	Nein	Metrisch	Metrisch	0 bis 1	Quadrierter Korrelationskoeffizient r	PRE-Maß	Relative Fehlerreduktion, abhängig von Varianz; Normalverteilung vorausgesetzt

Anmerkungen: Dichotom: Nominalskala mit genau zwei Ausprägungen; PRE-Maß: *proportional reduction in error*, Fehlerreduktion (Vorhersagekraft des Koeffizienten); Richtung: positive Zusammenhänge sind gleichgerichtet, negative Zusammenhänge sind gegensätzlich; Symmetrie: keine Unterscheidung von abh. und unabh. Variable.

5.1 Nominalskalierte Variablen

Zusammenhangsanalysen für diskrete Merkmale basieren häufig auf einer tabellarischen Auswertung ihrer gemeinsamen Verteilung: Die Häufigkeitsverteilung zweier kategorialer Variablen wird in einer sogenannten Kreuztabelle ausgewiesen. In dieser werden zunächst die als $fb_{(ij)}$ bezeichneten beobachteten (absoluten oder relativen) Häufigkeiten abgebildet (Kontingenztabelle). Mithilfe einer Indifferenztabelle können diesen die erwarteten Häufigkeiten gegenübergestellt werden, die sich bei Gleichverteilung bzw. statistischer Unabhängigkeit der Merkmale ergeben würden. Die erwarteten Häufigkeiten für jede Tabellenzelle lassen sich wie folgt bestimmten: (wobei i die Zeile und j die Spalte bezeichnet, in der sich die entsprechende Zelle in der Kreuztabelle befindet). Weichen beobachtete und erwartete Häufigkeiten systematisch voneinander ab, ist von einem Zusammenhang der Merkmale auszugehen. Kreuztabellen mit beobachteten und erwarteten Häufigkeiten lassen sich in Stata mit den Befehl

```
tabulate [variable1] [variable2]
```

erstellen. Damit wird eine Kontingenztabelle mit absoluten Häufigkeiten erstellt. Ergänzt man den Befehl um die Option , expected werden die erwarteten Häufigkeiten ausgegeben. Aus dem Output für den Befehl

```
tabulate [variable1] [variable2], expected
```

kann man also die Differenz von beobachteten und erwarteten Häufigkeiten ablesen und so ggf. schon auf eine systematische Verteilung schließen.

```
. tabulate v3 v151, expected

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| expected frequency|
+-------------------+

   ERHEBUNGSGEBIET |      GESCHLECHT, BEFRAGTE<R>
<WOHNGEBIET>: WEST |
             - OST |      MANN       FRAU |     Total
-------------------+----------------------+----------
ALTE BUNDESLAENDER |     1,178      1,214 |     2,392
                   |   1,180.5    1,211.5 |   2,392.0
-------------------+----------------------+----------
NEU BUNDESLAENDER  |       534        543 |     1,077
                   |     531.5      545.5 |   1,077.0
-------------------+----------------------+----------
             Total |     1,712      1,757 |     3,469
                   |   1,712.0    1,757.0 |   3,469.0
```

Die *key*-Tabelle im oberen Teil der Ausgabe gibt an, wofür die jeweiligen Kennziffern innerhalb der Tabelle stehen. In diesem Falle stehen demnach innerhalb der Zellen die absolute Häufigkeit oben und die erwartete Häufigkeit unten.

Eine Zusammenhangsanalyse erschöpft sich jedoch nicht nur in dem Ablesen von Differenzen, sondern beinhaltet das Berechnen leicht interpretierbarer Zusammenhangsmaße und ggf. Teststatistiken. Für dichotome und nominale Merkmale sind diese häufig χ^2-basiert (sprich: Chi-Quadrat): Zum einen kann der χ^2-Wert als symmetrische Maßzahl für das Vorhandensein eines Zusammenhangs interpretiert werden. Er bestimmt sich aus der Summe der quadrierten Differenzen zwischen beobachteten und erwarteten Häufigkeiten, geteilt durch die erwarteten Häufigkeiten. Ist der χ^2-Wert größer als 0, liegt ein Zusammenhang der beiden Merkmale vor. Aus dem χ^2 werden weitere, standardisierte

Zusammenhangsmaße abgeleitet, deren Interpretation aufgrund des normierten Wertebereichs von 0 bis 1 einfacher ist. Das sicherlich bekannteste und am häufigsten verwendete ist Cramér's V. Es berechnet sich wie folgt:

$$V = \sqrt{\frac{\chi^2}{n(min_{[Zeile,Spalte]} - 1)}}, \tag{1}$$

hierbei wird korrigiert, dass die Höhe des χ^2-Werts von der Anzahl der Beobachtungen beeinflusst ist. Darüber hinaus ist der χ^2-Wert bereits als solcher für das Testverfahren geeignet: Auf Basis der χ^2-Verteilung lässt sich der Prüfwert bestimmen, der dem α-Wert für den Ablehnungsbereich der Nullhypothese entspricht, und sich mit dem errechneten χ^2-Wert (der Prüfgröße) vergleichen lässt. In Stata lassen wir uns Cramér's V und den χ^2-Test ergänzend zur Kreuztabelle ausgeben:

```
tabulate [variable1] [variable2], chi V
```

erzeugt eine Kontingenztabelle,

```
tabulate [variable1] [variable2], expected chi V
```

eine kombinierte Kontingenz- und Indifferenztabelle mit absoluten Häufigkeiten. Unterhalb der Tabelle werden das Zusammenhangsmaß, die Prüfgröße und das Signifikanzniveau ausgegeben.

```
. tabulate v105 v151, chi V expected
```

Key
frequency *expected frequency*

INGLEHART-INDEX	GESCHLECHT, BEFRAGTE<R> MANN	FRAU	Total
POSTMATERIALISTEN	317 289.8	267 294.2	584 584.0
PM-MISCHTYP	541 525.4	518 533.6	1,059 1,059.0
M-MISCHTYP	523 514.5	514 522.5	1,037 1,037.0
MATERIALISTEN	301 352.3	409 357.7	710 710.0
Total	1,682 1,682.0	1,708 1,708.0	3,390 3,390.0

Pearson chi2(3) = 21.0885 Pr = 0.000
Cramér's V = 0.0789

In dem vorliegenden Beispiel wurde getestet, ob Frauen und Männer sich bezüglich ihrer persönlichen Einstellungen, die sich auf dem Kontinuum Materialismus–Postmaterialismus (ein Index entwickelt vom amerikanischem Politikwissenschaftler Ronald Inglehart) verorten lassen, unterscheiden. Das Ergebnis zeigt auf, dass sich beide Geschlechter tatsächlich signifikant voneinander unterscheiden (Pearson chi2(3) = 21.0885 Pr = 0.000) und die Abweichungen mit hoher Vertrauenswahrscheinlichkeit nicht von stochastischer Natur sind. Die Nullhypothese, die keinen Unterschied zwischen den Geschlechtern bezüglich

dieser Verortung postuliert, muss demnach abgewiesen werden. Um den Zusammenhang zwischen den beiden Merkmalen näher zu analysieren, müssen wir die Abweichungen der empirischen und der erwarteten Häufigkeiten betrachten. Es zeigt sich, dass Frauen in der Gruppe Materialisten tendenziell überrepräsentiert und in den anderen Gruppen unterrepräsentiert sind. Männer hingegen sind in den anderen Gruppen überrepräsentiert und häufiger als die jeweiligen Erwartungswerte angeben, vertreten. Nun wissen wir also, dass es einen Zusammenhang gibt und in welche Richtung dieser Zusammenhang wirkt. Jedoch wissen wir nicht, wie stark der Zusammenhang denn nun eigentlich ist. Zwar weist χ^2 auf eine Abhängigkeit hin, jedoch kann dieses Werte von null bis $+\infty$ annehmen, was uns bei der Interpretation der Stärke des Zusammenhangs nicht hilft. Dafür können wir Cramérs V heranziehen, dass χ^2 in Relation zu Fallzahl setzt. Es zeigt auf, dass die Stärke des Zusammenhangs mit einem Wert von 0,0789 sehr gering ist.

5.2 Ordinalskalierte Variablen

Für ordinale Merkmale sollte das Zusammenhangsmaß zudem der Rangordnung der Ausprägung beider Variablen Rechnung tragen, so dass die Richtung des Zusammenhangs zu erfassen ist. Ein häufig verwendetes, symmetrisches Zusammenhangsmaße für zwei ordinale Variablen ist γ (sprich: gamma). Dieses basiert auf der Logik des Paarvergleichs der Zellenbesetzung in einer Kreuztabelle: Unterschieden werden (im Sinne der Richtung des Zusammenhangs) konkordante (gleichgerichtete) und diskonkordante (gegengerichtete) Paare. Es berechnet sich als Differenz zwischen konkordanten und diskonkordanten Paare, geteilt durch die Summe der Paare:

$$\gamma = \frac{P_c - P_d}{P_c + P_d}, \tag{2}$$

(wobei c: konkordante Paare, d: diskonkordante Paare). Ist die Anzahl der konkordanten Paare größer als die der diskonkordanten Paare, besteht ein positiver Zusammenhang und γ nimmt einen Wert zwischen 0 und +1 ein. Ist die Anzahl der konkordanten Paare kleiner als die der diskonkordanten Paare, besteht entsprechend ein negativer Zusammenhang und γ nimmt einen Wert zwischen 0 und −1 ein. Es gilt: Je höher der Absolutwert von γ, desto stärker ist der Zusammenhang. Zudem ist der Absolutwert auch als PRE-Maß (proportional reduction of error) zu interpretieren: Durch Berücksichtigung der zweiten Variable wird der Fehler bei der Vorhersage der zweiten um soundsoviel Prozent reduziert. In Stata kann γ mit der zugehörigen Teststatistik ebenfalls als Ergänzung zur Kreuztabelle über die Option , gamma beim tabulate-Befehl ausgegeben werden:

```
tabulate [variable1] [variable2], gamma
```

```
. tabulate v21 v144, gamma

      EINWANDERER ZU |
           ANPASSUNG |          ZUSTIMM.: NATIONALSOZ. HATTE GUTE SEITEN
     VERPFLICHTEN? | STIMME GA  STIMME EH  WEDER NOC  STIMME EH  STIMME VO |     Total
-------------------+-------------------------------------------------------+----------
    STIMME VOLL ZU |       925         237         228         223          55 |     1,668
    STIMME EHER ZU |       596         153          73          54           9 |       885
       WEDER NOCH |       148          29          23           8           2 |       210
STIMME EHER NICHT ZU |       190          31           7          13           2 |       243
 STIMME GAR NICHT ZU |        66           4           7           7           2 |        86
-------------------+-------------------------------------------------------+----------
             Total |     1,925         454         338         305          70 |     3,092

                gamma =   -0.2844  ASE =  0.027
```

In dem vorliegenden Beispiel wurden zwei Variablen, die anhand einer Likert-Skala erfasst wurden, in einer Tabelle abgebildet. Zum einen wurde gefragt, ob man Einwanderer zu einer Anpassung verpflichten solle, zum anderen wurde gefragt, wie sehr man der Aussage, dass der Nationalsozialismus auch seine guten Seiten gehabt hätte, zustimmt. Das Ergebnis ist, dass die Anzahl konkordanter und diskonkordanter Paare einen γ-Wert von -0,2844 ergibt. Dies bedeutet, je eher die Personen der Aussage zustimmten, dass der Nationalsozialismus seine guten Seiten gehabt hätte, desto eher waren sie der Meinung, dass Einwanderer sich auch anpassen müssten. (Interpretationshilfe: Hierbei ist es wichtig zu schauen, wie die Items innerhalb der beiden ordinalen Variablen skaliert sind. Die Zeilen und Spalten, in denen die einzelnen Ausprägungen ausgeschrieben sind, werden als Interpretationshilfe herangezogen. Sollten diese Werte nicht mit Wertelabeln versehen sein, müssen wir ins Codebook schauen um die Codierung der Variable zu erfahren.) Ob das Ergebnis signifikant ist, wird von Stata nicht direkt mittels einer Teststatistik ausgegeben. Stattdessen wird der asymptotische Standardfehler (ASE) ausgegeben, mit dessen Hilfe man den z-Wert ermitteln kann. Wenn man den Wert für γ durch ASE dividiert, erhält man den z-Wert: $\frac{-0{,}2844}{0{,}027} = -10,53$. Es gilt:

$z \leq 1,96$: signifikant auf dem $p < 0,05$-Niveau
$z \leq 2,60$: signifikant auf dem $p < 0,01$-Niveau
$z \leq 3,32$: signifikant auf dem $p < 0,001$-Niveau

Wir können demnach davon ausegehen, dass wir einen leichten/moderaten Zusammenhang innerhalb der Stichprobe vorliegen haben, den wir auf die Grundgesamtheit übertragen können.

5.3 Metrischskalierte Variablen

Für stetige Merkmale, also Variablen mit metrischem Skalenniveau, ist aufgrund der hohen Anzahl möglicher Ausprägungen der Variable eine tabellenbasierte Analyse nicht sinnvoll. Mittels einer Korrelationsanalyse lässt sich jedoch analysieren, ob die stetige Veränderung einer Variablen mit der stetigen Veränderung einer anderen verknüpft ist. Mit Pearson's r (auch als Produkt-Moment-Korrelation bezeichnet), kann festgestellt werden, ob ein solch linearer Zusammenhang zwischen zwei metrischen Variablen vorliegt und wie stark dieser ist. Es handelt sich um ein symmetrisches Korrelationsmaß, das einen Wert zwischen -1 (perfekter negativer Zusammenhang), über 0 (kein Zusammenhang) bis zu +1 (perfekter positiver Zusammenhang) annehmen kann. Pearson's r wird über die Kovarianz, d.h. die gemeinsame Streuung zweier Merkmale, hergeleitet. Diese berechnet sich als Summe der Abweichungsprodukte aller Messwerte geteilt durch die Fallzahl. Die Kovarianz mit ihrem Wertebereich von $-\infty$ bis $+\infty$ kann auch schon als Korrelationsmaß interpretiert werden, kann aber nur Auskunft über die Richtung, nicht die Stärke des linearen Zusammenhangs geben. Um das Zusammenhangsmaß zu standardisieren wird Pearson's r berechnet, indem die Kovarianz durch das Produkt der Standardabweichungen beider Variablen geteilt wird:

$$r = \frac{Cov_{xy}}{S_x \cdot Sy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{x})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{3}$$

Der Standardbefehl für die Korrelationsanalyse in Stata ist correlate [variable1] [variable2]. Damit wird eine Korrelationsmatrix und Pearsons's r ausgegeben. Wird correlate

[variable1] [variable2], covariance eingegeben, erhält man die Kovarianz. Um die Zusammenhangsanalyse inferenzstatistisch abzusichern, wird ein t-Test durchgeführt. Dazu muss man in Stata einen anderen Befehl für die Korrelationsanalyse nutzen: Pearson's r erhält man auch mit dem Befehl

```
pwcorr [variable1] [variable2]
```

Die Teststatistik erhält man, in dem man den Befehl um entprechende Optionen erweitert:

```
pwcorr [variable1] [variable2], sig
```

gibt den p-Wert aus,

```
pwcorr [variable1] [variable2], star(#)
```

ersetzt die Ausgabe des p-Wertes durch Angabe des Signifikanzniveaus mit Sternchen (wobei # für das Level des Ablehnungsbereichs steht).

```
. pwcorr v631 v629

             |     v631     v629
-------------+------------------
        v631 |   1.0000
        v629 |   0.5292   1.0000

. pwcorr v631 v629, sig

             |     v631     v629
-------------+------------------
        v631 |   1.0000
             |
        v629 |   0.5292   1.0000
             |   0.0000

. pwcorr v631 v629, star(95)

             |     v631     v629
-------------+------------------
        v631 |   1.0000
        v629 |   0.5292*  1.0000
```

Sollen mehrere Niveaus unterschiedlich ausgezeichnet werden (bspw. 90%-, 95%- und 99%iges Signifikanzniveau), geht dies erst beim Export der Tabelle, bei dem die star-Option durch Angabe von Sternchen und Niveaus zu erweitern ist. Dazu benutzen wir das estout-ado und den Befehl esttab.

```
ssc install estout
pwcorr [variable1] [variable2], star(#)
esttab, star(+ 0.1 * 0.05 ** 0.01 *** 0.001)
```

```
. pwcorr v631 v629, star(95)

             |     v631     v629
-------------+------------------
        v631 |   1.0000
        v629 |   0.5292*  1.0000

. esttab, star(+ 0.1 * 0.05 ** 0.01 *** 0.001)

------------------------------------
                      (1)
                     Mean
------------------------------------
v631                75.68***
                   (196.89)
------------------------------------
N                    1594
------------------------------------
t statistics in parentheses
+ p<0.1, * p<0.05, ** p<0.01, *** p<0.001
```

Graphisch lässt sich die gemeinsame Verteilung zweier metrischer Variablen mit einem sogenannten Streudiagramm (scatterplot) darstellen:

```
scatter [variable1] [variable2]
```

oder vollständig:

```
graph twoway (scatter [variable1] [variable2])
```

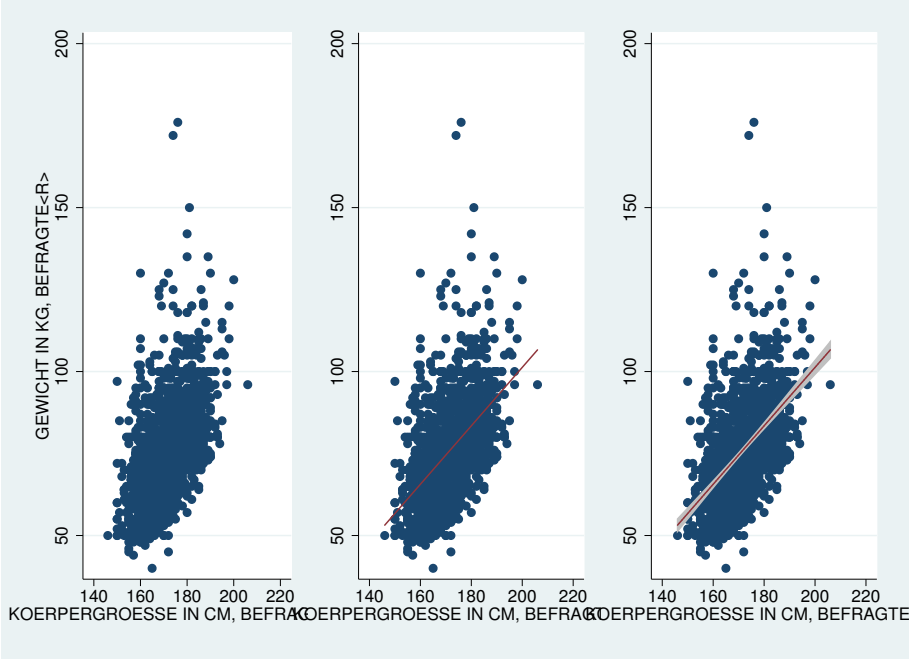
Die Graphik lässt sich um eine Linie ergänzen, die den Zusammenhang als lineare Schätzfunktion veranschaulicht:

```
graph twoway (scatter [variable1] [variable2]) (lfit [variable1][variable2])
```

bzw.

```
graph twoway (scatter [variable1] [variable2]) (lfitci [variable1][variable2])
```

mit 95%igem Konfidenzintervall.



6 Bivariate Regressionsanalyse

In diesem Kapitel wird die bivariate, lineare Regressionsanalyse vorgestellt. Diese stellt eines der wichtigsten Analyseverfahren der empirischen Sozialforschung dar und ist gleichzeitig Grundlage für die im folgenden Kapitel 7 vorgestellte multivariate Regressionsanalyse. Nach einer kurzen Einleitung [6.1] widmen wir uns den Grundannahmen [6.2] der Regression. Anschließend wird die Schätzfunktion [6.3], sowie die vier Schritte einer Regressionsanalyse [6.4] vorgestellt, bevor ein empirisches Beispiel [6.5], sowie die dazugehörige Regressionsdiagnostik [6.6] dieses Beispiels diskutiert wird.

6.1 Einleitendes

Die lineare Regressionsanalyse ist ein asymmetrisches Analyseverfahren zur Bestimmung von Zusammenhängen zwischen einer metrisch skalierten abhängigen und einer oder mehreren unabhängigen Variable(n). Bei einer bivariaten Regression wird immer der Einfluss von genau einer unabhängigen Variable auf die abhängige Variable geschätzt. Da es sich um ein asymmetrisches Verfahren handelt, eignet es sich sehr gut, um die Einflussstärke und -richtung eines Regressors x innerhalb einer Korrelation zu ermitteln. Hierbei wird eine Schätzfunktion für einen linearen Zusammenhang ermittelt. Der Zusammenhang wird also quantitativ beschreibbar und die Werte der abhängigen Variable prognostizierbar. (Wagschal 1999, 209f.; Backhaus 2008, 52f.) y ist daher also eine Funktion von x.

$$y = f(x) \tag{4}$$

Die abhängige Variable muss bei der linearen Regression immer metrisch skaliert sein. Die unabhängige Variable kann sämtliche Skalenniveaus annehmen, dies bedarf aber besonderer Rücksicht bei der Interpretation ihrer Koeffizienten. Die lineare Regressiongleichung wird mithilfe der *ordinary least square* (OLS)-Methode ermittelt, da mittels quadrierter Abstände die Regressionsgerade eindeutig bestimmt werden kann. Bei einer Regression werden die Abstände entlang der Ausprägungen der abhängigen Variable minimiert. Das Ziel besteht

darin, eine Gerade zu ermitteln, die insgesamt den niedrigsten Abstand (Residuum) zu sämtlichen Punkten (Fällen) hat, d.h. die Summe aller Abstände minimal ist.

Der Koeffizient b, der den Einfluss der unabhängigen Variable auf die abhängige Variable angibt, wird mithilfe der Kovarianz von x und y dividiert durch die Varianz von x – sämtliche Maßzahlen, die wir bereits aus der univariaten Deskription und der bivariaten Analyse kennen – ermittelt.

6.2 Grundannahmen der linearen Regression

Eine lineare Regression enthält nur dann effiziente und erwartungstreue Schätzungen der Parameter, wenn die sog. Gauss-Markov-Annahmen nicht verletzt werden:

- Lineare Beziehung zwischen abhängiger und unabhängigen Variablen
- Normalverteilung der Fehler
- Gleich bleibende Varianz der Fehler (Homoskedastizität, keine Heteroskedastizität)
- Unabhängigkeit der Fehler (keine systematische Fehlerkorellation)
- Korrekte Spezifikation des Modells. Modell hat alle relevanten Variablen miteinbezogen (Omitted Variable Bias)
- Kein Vorliegen von hoher Korrelation der unabhängigen Variablen untereinander (Multikollinearität).
- Einflussreiche Beobachtungen (Ausreißer, Extremwerte) können unter Umständen einen zu hohen Einfluss auf die Lage der Gerade haben.

6.3 Schätzfunktion

Die Schätzfunktion bildet zugleich die Regressionsgerade, die sich aus der Steigung der Funktion ableitet. Die Konstante b_0 ist hierbei der Schnittpunkt der y-Achse und b_1 ist die Steigung der Geraden.

$$\hat{Y} = b_0 + b_1 X + \varepsilon \tag{5}$$

wobei:

1. $\hat{Y}$: Schätzung der abhängigen Variable Y
2. b_0: Konstante
3. b_1: Regressionskoeffizient
4. X: unabhängige Variable
5. ε: Fehlerterm (Espilon)

6.4 Schritte einer Regressionsanalyse

1. Bestimmung des Funktionstyps
2. Bestimmung der Regressionsfunktion/der Koeffizienten
3. Bewertung der Modellgüte
4. Signifikanztest für die einzelnen Parameter und das Gesamtmodell

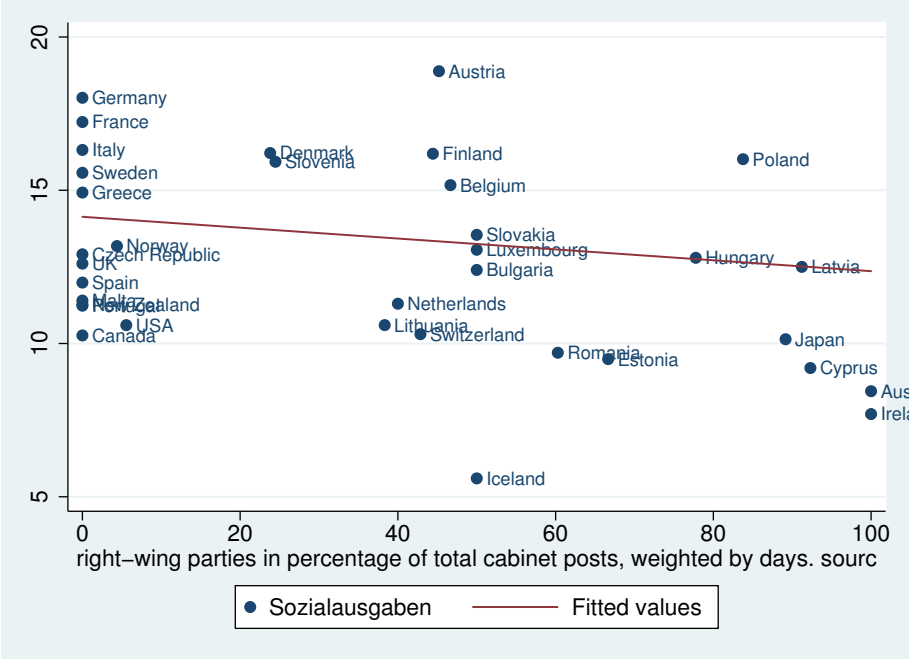
6.5 Empirisches Beispiel

Erklärt werden soll die Höhe der Sozialausgaben in Abhängigkeit vom Anteil rechter Parteien an der Regierung. Der Funktionstyp ist linear und bivariat, da zwischen einem graduellen Zusammenhang zwischen zwei Variablen ausgegangen wird.

6.5.1 Graphische Darstellung

Um einen Scatterplot mit OLS-Gerade darzustellen, bedarf es folgender Syntax:

```
graph twoway (scatter sstran gov_right1 if year==2000, ///
mlabel(country))(lfit sstran gov_right1)
```



Die Graphik zeigt auf, dass es einen negativen Zusammenhang zwischen den beiden Variablen gibt und manche Fälle näher an der Geraden liegen als andere.

6.5.2 Berechnung

Mithilfe des CPDSI berechnen wir die Höhe der Sozialausgaben (in Relation zum BIP) in Abhängigkeit vom Sitzanteil rechter Parteien in der Regierung für das Jahr 2000. Der Regressionsbefehl in Stata lautet wie folgt:

```
regress depvar [indepvars] [if] [in] [weight] [, options]
```

Für unser konkretes Beispiel lautet der Befehl bei Einsetzen der entsprechenden Variablen wie folgt:

```
regress sstran gov_right1 if year==2000
```

Bei Eingabe dieses Befehl erhalten wir folgenden Output:

Source	SS	df	MS	Number of obs =	35
				F(1, 33) =	5.41
Model	46.0246206	1	46.0246206	Prob > F =	0.0264
Residual	280.97055	33	8.51425909	R-squared =	0.1408
				Adj R-squared =	0.1147
Total	326.995171	34	9.61750502	Root MSE =	2.9179

sstran	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
gov_right1	-.0332353	.0142948	-2.32	0.026	-.0623184	-.0041523
_cons	13.86044	.7178038	19.31	0.000	12.40006	15.32082

Als Ergebnis erhalten wir folgende Schätzgleichung:

$$\hat{\text{Sozialausgaben}} = 13,86 + (-0.0332) \cdot \text{Anteil rechter Parteien} \tag{6}$$

Die Gerade schneidet die Achse demnach bei einem Anteil der Sozialausgaben von 13,86 % in Relation zum BIP. Dies wiederum bedeutet, dass ein hypothetischer Fall (in unserem Beispiel ein Mitgliedsstaat der OECD) mit einem Anteil von 0 % rechter Parteien innerhalb der Regierung – laut dem Schätzverfahren – Sozialausgaben in Höhe von 13,86 % hätte. Mit steigendem Anteil rechter Parteien würde dieser geschätzte Wert linear um den Faktor 0,0332 sinken. Das heißt, dass mit jedem Prozentpunkt Zunahme des Anteils rechter Parteien die Sozialausgaben um jeweils 0,0332 % sinkt. Dies bedeutet, dass ein Land mit einer hypothetischen Regierungsbeteiligung rechter Parteien von 100 % Sozialausgaben von 10,53 % hätte.

6.5.3 Bewertung der Modellgüte

Determinationskoeffizient Der Determinationskoeffizient R^2 ist ein Bestimmtheitsmaß, welches die Güte des Modells wiedergibt. Hierbei werden nicht erklärte Streuung und Gesamtstreuung miteinander in Relation gesetzt. Je kleiner dabei die nicht erklärte Streuung ist, desto höher ist die Varianzaufklärung für die abhängige Variable im Modell. R^2 hat einen Wertebereich von $[0; 1]$. Würden bspw. alle geschätzten Punkte auf der Regressionsgraden liegen, ergäbe sich der Wert 1, der einen perfekten linearer Zusammenhang zwischen abhängiger und unabhängiger Variable beschreibt. Ist der Wert 0, liegt kein linearer Zusammenhang zwischen den Variablen vor. Werte zwischen 0 und 1 geben Auskunft über den Anteil der erklärten Streuung und damit über die Erklärungskraft des Modells. Multipliziert man das Ergebnis des Koeffizienten R^2 mit 100, so erhält man den Prozentsatz der erklärten Varianz. In unserem Beispiel beträgt der Wert für R^2 0,1408. Dies bedeutet, dass 14,08 % der Varianz der Sozialausgaben durch das Modell erklärt werden.

Root MSE Der sog. *root mean squared error* (Root MSE) steht für die Wurzel aus den mittleren quadratischen Abweichungen. Je kleiner dieser Wert ist, desto geringer ist die allgemeine Abweichung vom Schätzwert und desto geringer ist die Fehlervarianz. In unserem Beispiel beträgt der Wert der mittleren quadratischen Abweichung $\sqrt{8.5142} = 2.9179$. Ein alternatives Modell zur Beschreibung der Sozialausgaben mit geringeren Abweichungen wäre demnach besser geeignet einen linearen Zusammenhang zwischen den Variablen zu beschreiben.

6.5.4 Signifikanztest für die einzelnen Parameter und das Gesamtmodell

F-Test Ob unser Modell und dessen Schätzfunktion auch über unsere Stichprobe hinaus Gültigkeit für die Grundgesamtheit hat, kann mithilfe des F-Tests ermittelt werden. Dieser bezieht zusätzlich die Fallzahl der Analyse mit ein, um die Gültigkeit des in dem Modell postulierten Zusammenhang zwischen abhängiger und abhängiger Variable zu prüfen (ob der Koeffizient der unabhängigen Variable ungleich Null ist). In unserem Beispiel zeigt die F-Statistik, dass die Irrtumswahrscheinlichkeit bei einem F-Wert von 5,41 gleich 0,0264 beträgt. Wir können die Nullhypothese, die besagt, dass der Koeffizient unserer unabhängigen Variable gleich null ist, mit einer Vertrauenswahrscheinlichkeit von 97,36 % verwerfen.

t-Test Analog zum F-Test für das Gesamtmodell verläuft der t-Test für die unabhängige Variable im Modell. Hier wird ebenfalls die Hypothese getestet, ob der Einfluss der einzelnen Variable gleich null ist. Der t-Wert ergibt sich aus der Division des Koeffizienten durch dessen Standardfehler (in unserem Beispiel $\frac{-0{,}0332}{0{,}1429} = -2,32$). Die t-Verteilung um den Wert Null bei entsprechender Vertrauenswahrscheinlichkeit (in diesem Fall 95 %) wird nun zugrunde gelegt um zu prüfen, ob die Nullhypothese angenommen oder verworfen werden muss. In dem Fallbeispiel beträgt die Irrtumswahrscheinlichkeit 0,026 (die Vertrauenswahrscheinlichkeit ist demnach 97,4 %. Die Konfidenzintervalle geben auch hier den Bereich an, in dem sich der tatsächliche Koeffizient, der den Zusammenhang innerhalb der Grundgesamtheit beschreibt, befindet.

6.6 Regressionsdiagnostik

Mithilfe von Stata ist es möglich die im Abschnitt 6.2 aufgelisteten Grundannahmen der Regression zu testen. In den folgenden Unterabschnitten werden die dazu benötigten Befehle vorgestellt und eine Interpretationshilfe gegeben.

6.6.1 Test auf Heteroskedastizität

Der Breusch-Pagan/Cook-Weisberg-Test testet, ob Heteroskedastizät vorliegt. Wenn dies der Fall ist, dann sind die Residuen nicht normalverteilt und es liegt unter Umständen eine systematische Verzerrung innerhalb der Schätzung vor. Die Nullhypothese ist konservativ gewählt: Es liegt keine Heteroskedastizität vor (*constant variance*).

```
estat hettest
```

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of sstran

         chi2(1)      =     0.05
         Prob > chi2  =   0.8282
```

In dem vorliegenden Fall kann die Nullhypothese nicht verworfen werden. Es kann davon ausgegangen werden, dass keine Heteroskedastizität vorliegt.

6.6.2 Ramsey Test

Der Ramsey Test überprüft, ob in dem Modell wichtige erklärende Variablen ausgelassen wurden (sog. *omitted variable bias*. Wenn *bias* vorliegt, muss eventuell das Modell respezifiziert werden (diese Respezifizierung sollte anhand von theoretischen Überlegungen

erfolgen). Die Nullhypothese dieses Tests ist ebenfalls konservativ gewählt: Es liegt kein *bias* vor.

```
estat ovtest
```

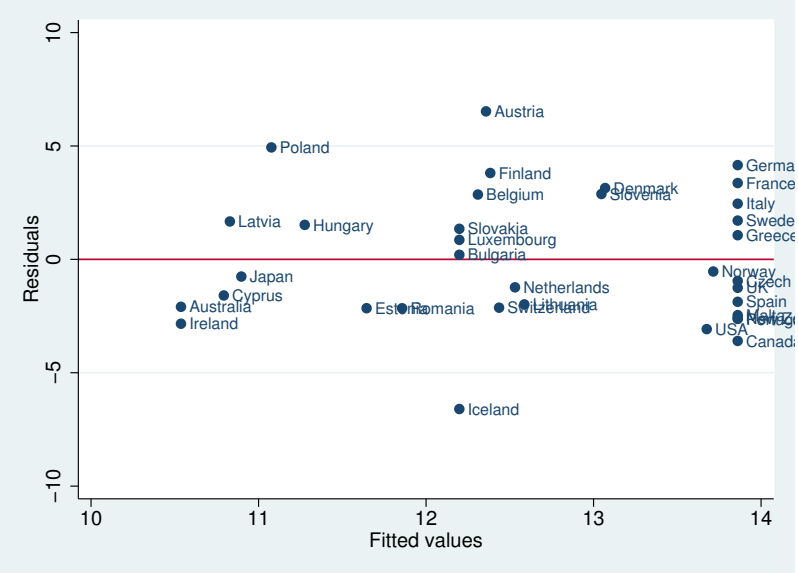
```
Ramsey RESET test using powers of the independent variables
       Ho:  model has no omitted variables
                  F(3, 30) =      0.90
                  Prob > F =      0.4547
```

In dem vorliegenden Fall kann die Nullhypothese nicht verworfen werden. Es kann davon ausgegangen werden, dass kein *bias* vorliegt.

6.6.3 Residual versus fitted plot

Das *residual versus fitted plot* stellt die Schätzwerte der Fälle gegenüber deren Residuen dar. Diese Darstellungsmöglichkeit eignet sich, um graphisch das Vorhandensein von Heteroskedastizität, sowie eines linearen Zusammenhangs zu kontrollieren. Wenn eine Punktwolke vorliegt, bei der die einzelnen Fälle stochastisch verteilt sind, kann davon ausgegangen werden, dass keine Gauss-Markov-Bedingung verletzt wurde. Wenn der Lage der Punkte jedoch eine Systematik unterliegt (unterschiedliche Varianz der Residuen, nicht-lineare Verteilung), dann sind Grundannahmen der linearen Regression verletzt.

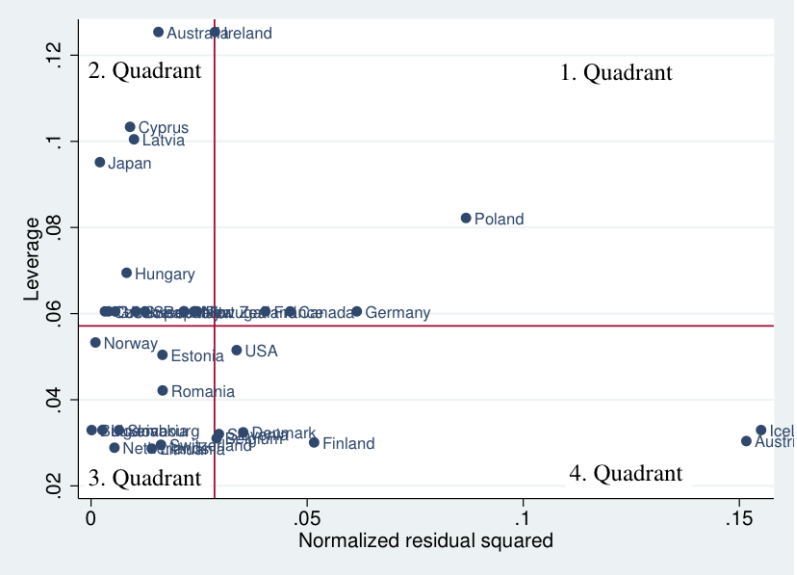
```
rvfplot, yline(0) mlabel(country)
```



6.6.4 Leverage plot für Ausreißer

Das leverage plot für Ausreißer ist ein wichtiges Instrument um einflussreiche Fälle mit einer hohen Hebelwirkung auf die Regressionsgerade graphisch zu identifizieren.

```
lvr2plot, yline(0) mlabel(country)
```



1. Quadrant Problematische Fälle sind in diesem Quadranten verortet. Diese Fälle besitzen ein großes Residuum und eine große Hebelwirkung auf die Ausgleichsgerade. Typischerweise weisen diese Fälle extreme Werte in x auf.

2. Quadrant Fälle innerhalb dieses Quadranten haben eine große Hebelwirkung auf die Ausgleichsgerade, jedoch ein geringes Residuum (diese Fälle entsprechen demnach unserer Erwartung, können aber auch zur Überschätzung des Einflusses führen).

3. Quadrant Fälle innerhalb des dritten Quadranten besitzen weder ein großes Residuum, noch eine große Hebelwirkung auf die Ausgleichsgerade ausüben und sind demnach völlig unproblematisch.

4. Quadrant Sämtliche Fälle innerhalb dieses Quadranten besitzen ein großes Residuum, jedoch keine Hebelwirkung. Typischerweise sind dies Fälle, die keine starke Ausprägung in x aufweisen und nahe dem arithmetischen Mittel von x liegen.

Eine Auflistung der Cook's Distanzen (D) für die gilt: $D > \frac{4}{n}$, zeigt problematische Ausreißer auf. Diese Fälle sollten evtl. aus der Regression entfernt werden und das Modell sollte neu spezifiziert werden.

```
predict cooksd, cooksd
sort cooksd
list country year sstran gov_right1 cooksd ///
if cooksd>4/e(N) & year==2000
```

country	year	sstran	gov_ri~1	cooksd
Poland	2000	16.01364	83.76	.1397055

Wie auch schon bei der graphischen Inspektion zeigt die numerische Diagnostik, dass Polen einen Ausreißer innerhalb der linearen Regression darstellt. Demzufolge sollte eine Respezifizierung des Modells ohne Hereinnahme Polens in Erwägung gezogen werden. Jedoch sollte sowohl die Herausnahme, als auch die Beibehaltung des Falles aufgrund theoretischer Abwägungen erfolgen und begründet werden.

7 Multivariate Regression

In diesem letzten Kapitel wird die multivariate Regressionsanalyse vorgestellt. Diese baut auf die im vorigen Kapitel 6 vorgestellte bivariate Regressionsanalyse auf und ist eines der am häufigsten angewandten Analyseverfahren in der quantitativ-empirischen Forschung der Sozialwissenschaften. Nach einer kurzen Einleitung [7.1] widmen wir uns einem empirischen Beispiel [7.2]. Daran anschließend wird eine der gängigsten Formen der tabellarischen Darstellung und deren Generierung mithilfe von Stata [7.3] vorgestellt. Abschließend werden speziell für die multivariate Regression geltende Regressionsdiagnostiken [7.4] vorgestellt, die zusätzlich zu den im vorigen Kapitel vorgestellten Diagnoseverfahren durchgeführt werden sollten.

7.1 Einleitendes

Das multivariate lineare Regressionsmodell ist ein Schätzmodell, bei dem der Einfluss mehrerer unabhängiger Variablen auf die abhängige Variable bestimmt wird. Dies hat den Vorteil, dass der Einfluss mehrerer Faktoren bestimmt und kontrolliert werden kann. In den Sozialwissenschaften ist es oft der Fall, dass Zusammenhänge multikausal geartet sind und mehrere Bedingungen zwar notwendig sind, aber erst in ihrem Zusammenwirken zu hinreichenden Bedingungen werden, die ein Phänomen auslösen bzw. die Intensität des Phänomens bestimmen.

Das Schätzmodell des multivariaten linearen Regressionsmodell unterscheidet sich im Gegensatz zum bivariaten Modell in der Anzahl der Determinanten b_k:

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k + \varepsilon \tag{7}$$

Die Ermittlung der Koeffizienten erfolgt mittels der *ordinary least squares*-Methode. Die relative Effektstärke der einzelnen Variablen ist mithilfe standardisierter beta-Regressionskoeffizienten zu ermitteln. Die Bewertung der Modellgüte erfolgt wie bei dem bivariaten Modell über den Determinationskoeffizienten r^2. Jedoch beinhaltet dies eine gewisse Problematik, da mit jeder zusätzlich in das Modell integrierten unabhängigen Variable auch die Varianzaufklärung – und sei sie auch noch so gering – steigt. Deshalb wird mit dem sog. korrigierten R-Quadrat ($adj.\ r^2$)ein neues Gütemaß eingeführt, dass auf die Anzahl der Freiheitsgrade des Modells kontrolliert. Der Vorteil an diesem Maß ist, dass es nur größer wird, wenn die Summe der Abweichungsquadrate überproportional zu der Anzahl der Freiheitsgraden sinkt – also zusätzliche Variablen maßgeblich zur Reduktion der Abweichungen beitragen. Der Nachteil dieses Gütemaßes ist jedoch, dass es nicht mehr als

Anteil erklärter Varianz interpretiert werden kann, weshalb oftmals beide Maße betrachtet werden.

7.2 Empirisches Beispiel

Der Stata-Befehl zur multivariaten Regression folgt dabei derselben Logik wie dies schon beim Übergang von univariater zu bivariater Deskription der Fall war, indem weitere unabhängige Variablen hinter dem regress-Befehl und der anhängigen Variablen angegeben werden:

```
regress [variable1][variable2][variable3]...[variablek]
```

```
. reg sstran gov_right1 effpar_leg if year==2000

      Source |       SS       df       MS              Number of obs =      35
-------------+------------------------------           F(  2,    32) =    5.11
       Model |  79.1850744     2  39.5925372            Prob > F      =  0.0118
    Residual |  247.810096    32  7.74406551           R-squared     =  0.2422
-------------+------------------------------           Adj R-squared =  0.1948
       Total |  326.995171    34  9.61750502           Root MSE      =  2.7828

------------------------------------------------------------------------------
      sstran |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  gov_right1 |  -.0374479   .0137841    -2.72   0.011    -.0655251   -.0093706
  effpar_leg |   .6955855   .3361436     2.07   0.047     .0108835    1.380288
       _cons |     11.3425   1.39615     8.12   0.000     8.498637    14.18637
------------------------------------------------------------------------------
```

In dem vorliegenden Beispiel wurde der Einfluss rechter Parteien in der Regierung und der effektiven Anzahl an Parteien im Parlament auf die Sozialausgaben in Relation zum BIP geschätzt. Es zeigt sich, dass je stärker rechte Parteien in der Regierung vertreten sind, desto geringer sind die Sozialausgaben. Mit jedem zusätzlichen Prozentpunkt rechter Parteien in der Regierung sinken die Sozialausgaben um durchschnittlich 0,04%. Bei der Anzahl der effektiven Parteien ist der Zusammenhang jedoch positiv: je mehr effektive Parteien im Parlament vertreten sind, desto höher sind die Sozialausgaben. Mit jeder effektiven Partei mehr im Parlament steigen die Sozialausgaben um durchschnittlich 0,7% des BIP.

Die nicht-standardisierten Koeffizienten geben keinen Anhaltspunkt darüber, wie stark in Relation zu einer anderen Variable der jeweilige Einfluss der Variable auf das Regressionsgewicht ist. Aufgrund der unterschiedlichen Einheiten der beiden unabhängigen Variablen (Anteil rechter Parteien, Anzahl effektiver Parteien) können wir die Effektstärken nicht direkt miteinander vergleichen, da die Einheiten zu unterschiedlich sind. Um diese jedoch in Relation bringen zu können und eine Aussage darüber zu treffen, welcher der beiden Faktoren denn nun einen größeren Einfluss auf das Regressionsgewicht hat, müssen standardisierte beta-Koeffizienten angegeben werden. Dies geschieht mit Anfügen der beta-Option an den regress-Befehl:

```
regress [variable1][variable2][variable3]...[variable_k], beta
```

```
. reg sstran gov_right1 effpar_leg if year==2000, beta
```

Source	SS	df	MS		
Model	79.1850744	2	39.5925372	Number of obs	= 35
Residual	247.810096	32	7.74406551	F(2, 32)	= 5.11
				Prob > F	= 0.0118
Total	326.995171	34	9.61750502	R-squared	= 0.2422
				Adj R-squared	= 0.1948
				Root MSE	= 2.7828

sstran	Coef.	Std. Err.	t	P>|t|	Beta
gov_right1	-.0374479	.0137841	-2.72	0.011	-.4227186
effpar_leg	.6955855	.3361436	2.07	0.047	.3219795
_cons	11.3425	1.39615	8.12	0.000	.

Ein Vergleich der Beträge der standardisierten beta-Koeffizienten zeigt, dass der relative Einfluss rechter Parteien (-0,42) stärker ist als der Einfluss der Anzahl effektiver Parteien im Parlament (0,32).

7.3 Vergleich mehrerer multivariater Regressionen und Erstellung publikationsfähiger Tabellen

Mithilfe des esttab-Befehls lassen sich mehrere Regressionsmodelle schnell und übersichtlich miteinander vergleichen. Mit Anfügen des eststo:-Präfixes teilen wir Stata mit, dass die aus der folgenden Regressionen berechneten Koeffizienten gespeichert werden. Diese können dann später mithilfe des esttab-Befehls wieder aufgerufen werden.

```
. quietly eststo: reg sstran gov_right1 if year==2000

. quietly eststo: reg sstran gov_right1 effpar_ele if year==2000

. quietly eststo: reg sstran gov_right1 effpar_ele debt deficit womenpar if year==2000

. esttab est1 est2 est3, ///
> alignment(c) width(100%) b(%9.3f) se(%9.3f) ///
> nonumbers mtitles("Modell \#1" "Modell \#2" "Modell \#3") /// Titel der Modelle
> order(gov_right1 effpar_ele debt deficit womenpar) /// Reihenfolge der Variablen
> drop() /// Entfernen einzelner Variablen (bspw. Fixed Effects in einer TSCS)
> star(* 0.05 ** 0.01 *** 0.001) ///
> stats(r2 r2_a N, fmt(%9.3f %9.3f %9.0f) labels("R2" "Korr. R2" "N")) /// Zusätzliche Statistiken
> title("Der Einfluss exogener Faktoren auf die Sozialausgaben") /// Überschrift
> nonote addn("Nicht-standardisierte Koeffizienten einer linearen Regression" ///
> "Standardfehler in Klammern." ///
> "* p < 0.05, ** p < 0.01, *** p < 0.001") // Anmerkungen am Tabellenende
```

⋆ Anmerkung: Durch den quietly-Präfix wurden im Output-Fenster von Stata die Regressionstabellen nicht angezeigt

Der Einfluss exogener Faktoren auf die Sozialausgaben

	Modell \#1	Modell \#2	Modell \#3
gov_right1	-0.033*	-0.039**	-0.037*
	(0.014)	(0.014)	(0.015)
effpar_ele		0.664*	0.640*
		(0.298)	(0.313)
debt			0.014
			(0.016)
deficit			-0.029
			(0.137)
womenpar			0.019
			(0.057)
_cons	13.860***	10.989***	9.917***
	(0.718)	(1.456)	(2.088)
R2	0.141	0.256	0.275
Korr. R2	0.115	0.210	0.150
N	35	35	35

Nicht-standardisierte Koeffizienten einer linearen Regression
Standardfehler in Klammern.
* p < 0.05, ** p < 0.01, *** p < 0.001

Der Vergleich des bivariaten Modells aus dem vorherigen Abschnitt mit den beiden multivariaten Modellen zeigt, dass die Hinzunahme von unabhängigen Variablen immer auch mit einer höheren Varianzaufklärung einhergeht. Allerdings zeigt der Vergleich der beiden multivariaten Modelle, dass die Hinzunahme der ökonomischen Variablen und dem Anteil der Frauen im Parlament das korrigierte R-Quadrat wieder sinken lässt, da diese nicht überproportional zur Varianzaufklärung beitragen.

7.4 Regressionsdiagnostik

Bei der multivariaten Regression gibt es neben den bekannten Tests zur Regressionsdiagnostik zusätzliche Testverfahren, die auf Eigenschaften der unabhängigen Variablen untereinander testen (bspw. Multikollinearität). Der sog. *variance inflation factor* gibt an, wie stark die unabhängigen Variablen untereinander korrelieren. Die jeweiligen Faktoren für einzelne Variablen und das Gesamtmodell können Werte von null bis ∞ annehmen. Grundsätzlich gilt die Daumenregel, dass absolute Werte unter zehn unproblematisch sind. Wenn jedoch der Wert von zehn sowohl für einzelne Variablen oder gar das Gesamtmodell überschritten wird, sollte das Modell respezifiziert werden.

```
estat vif
```

```
. estat vif

    Variable |      VIF       1/VIF
-------------+----------------------
    womenpar |     1.46    0.684493
     deficit |     1.42    0.705531
  gov_right1 |     1.21    0.826103
  effpar_ele |     1.07    0.938220
        debt |     1.05    0.949705
-------------+----------------------
    Mean VIF |     1.24
```

Die Diagnostik des Modells zeigt auf, dass keine der unabhängigen Variable – und damit auch das Gesamtmodell – problematisch für das Schätzverfahren ist.
Wenn der vif-Test anzeigt, dass eine oder mehrere Variablen hoch mit anderen Variablen korrelieren, dann kann mithilfe von Stata eine Korrelationsmatrix erstellt werden, die die einzelnen Korrelationskoeffizienten der unabhängigen Variablen angibt.

```
estat vce, correlation
```

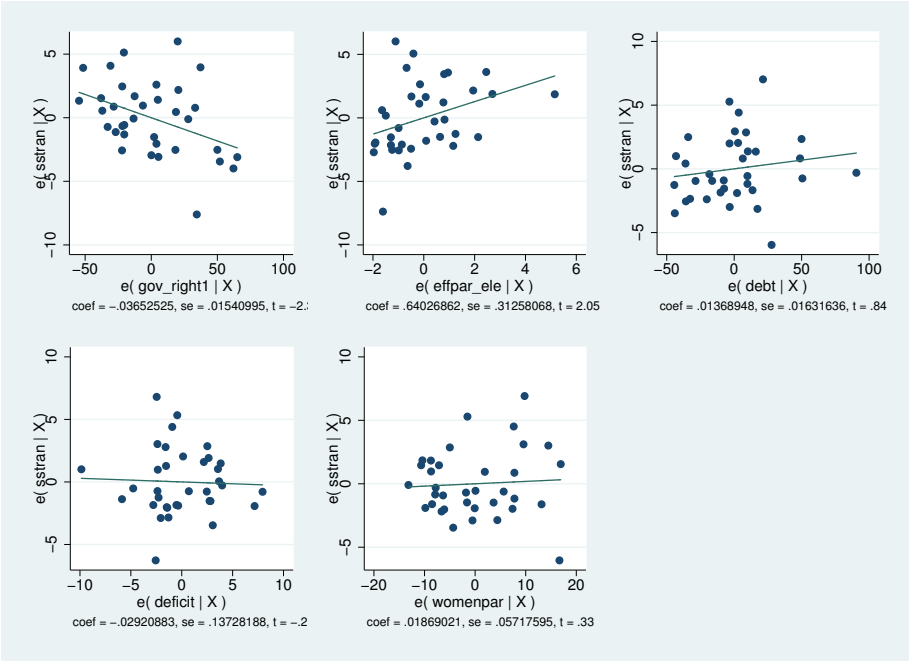
```
. estat vce, corr

Correlation matrix of coefficients of regress model

      e(V) | gov_ri~1  effpar~e      debt    deficit  womenpar     _cons
-----------+------------------------------------------------------------
gov_right1 |   1.0000
effpar_ele |  -0.2400   1.0000
      debt |   0.1969  -0.0760   1.0000
   deficit |   0.1252  -0.0338  -0.0641   1.0000
  womenpar |   0.2060  -0.0962   0.1310  -0.4884   1.0000
     _cons |  -0.3163  -0.5434  -0.5006   0.1482  -0.5218   1.0000
```

Da es bei der multivariaten Regression im Gegensatz zur bivariaten Regression nicht mehr möglich ist, sich den Zusammenhang zwischen den unabhängigen Variablen und der abhängigen Variablen innerhalb eines Scatterplots zu betrachten, gibt es unter Stata einen *added variable plot*-Befehl. Die Teilgraphiken zeigen den Einfluss der Variablen auf den Schätzwert der abhängigen Variablen.

```
avplots
```



http://comparativepolitics.uni-greifswald.de