Analysen politikwissenschaftlicher Datensätze mit Stata

Formales: Scheinerwerb

- Unbenoteter Schein:
  - Teilnahme am Seminargespräch
  - Hausaufgaben
  - Die Lektüre der Pflichttexte ist für alle Teilnehmer *verbindlich*, gelegentlich überprüfe ich Ihren Kenntnisstand.
  - Sie dürfen *maximal zwei* Sitzungen versäumen.
- Benoteter Schein: zusätzlich Hausarbeit
  - thematischer Zusammenhang mit dem Seminar
  - eigenständige Analyse
  - explizite Fragestellung

Formales: Hausarbeit

- Gliederung, Zitierweise, Literaturverzeichnis etc. bitte entsprechend den üblichen Standards (vgl. Homepage)
- Schriftgröße etc.: Formatvorlage
- Umfang ca. 7.000-9.000 Worte, entspricht etwa 20-25 reinen Textseiten
- Beginn bereits während des Semesters
- **Letzter Abgabetermin: Montag, 23. August 2004. Eine Verschiebung des Termins ist *nicht* möglich**

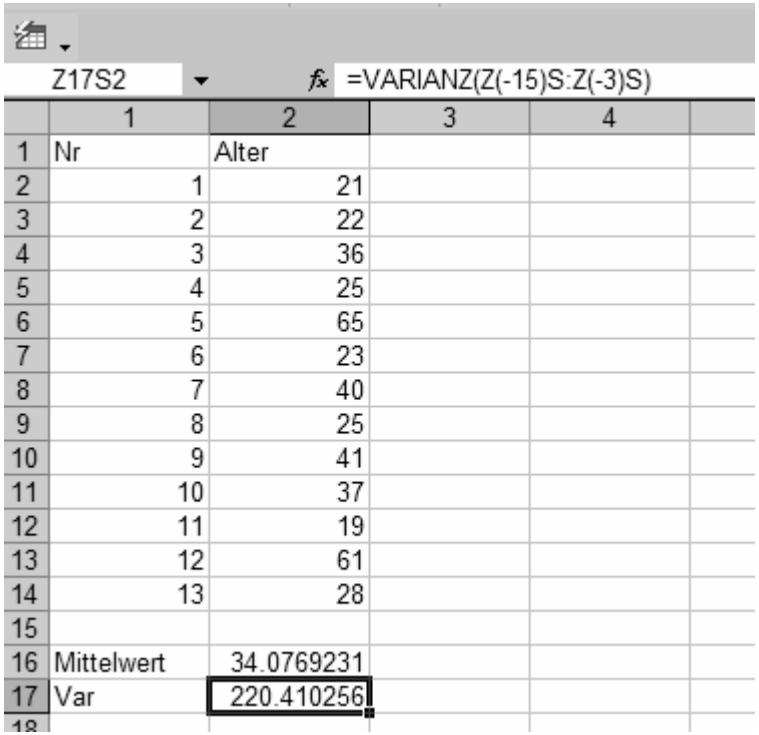
Weitere Informationen zum Seminar

- http://www.politik.uni-mainz.de/kai.arzheimer/Lehre-Stata/

Was sind und wozu braucht man Datenanalyseprogramme?

- Im Grundstudium haben Sie gelernt, *einfache* statistische Analysen *weniger Fälle* mit Papier und Bleistift und evtl. Hilfstabellen durchzuführen
- In sozialwissenschaftlichen Untersuchungen fallen Datenbestände an, die oft Tausende von Fällen mit Hunderten von Variablen umfassen
- Datenanalyseprogramme lösen zwei Probleme:
  - Daten müssen zuverlässig aufbewahrt und wiedergefunden werden
  - Bei großen Fallzahlen und/oder komplizierten Verfahren sind Analysen ohne maschinelle Hilfe nicht mehr durchzuführen

Excel als Alternative?



Z17S2 | =VARIANZ(Z(-15)S:Z(-3)S)

	1	2	3	4
1	Nr	Alter		
2	1	21		
3	2	22		
4	3	36		
5	4	25		
6	5	65		
7	6	23		
8	7	40		
9	8	25		
10	9	41		
11	10	37		
12	11	19		
13	12	61		
14	13	28		
15				
16	Mittelwert	34.0769231		
17	Var	220.410256		

- Bei *etwas* komplexeren Analysen behilft man sich oft mit Excel

Excel als Alternative?

- Vorteile:
  - intuitiver Zugang
- Probleme:
  - Fallzahl beschränkt
  - kaum/keine Rekodierungsmöglichkeiten
  - wenig Analysemöglichkeiten
  - Geschwindigkeit
  - Keine Trennung von Daten und Analyse ➜ großes Potential für fatale Fehler
- Eher eine Art luxuriöser Taschenrechner

Was leisten Datenanalyseprogramme (1) ?

- Speichern und verwalten die Daten in einem kompakten, binären Format (*.sav bei SPSS, *.dta bei Stata)
- Präsentieren dem Benutzer die Daten als Matrix (~Tabelle)
  - deren Zeilen (in der Regel) einem *Fall* (Person, Partei, Staat etc.) entsprechen
  - deren Spalten jeweils eine *Variable* enthalten
  - deren Zellen jeweils einen Meßwert für einen Fall enthalten
- Dies entspricht der Form der Datensammlung, die Sie im Grundkurs kennengelernt haben

Konvention für die Darstellung von Daten

- Eine gegebene Variable (z.B. Alter oder Geschlecht) wird in Formeln durch den Buchstaben x abgekürzt
- Die zu messenden Objekte (z.B. Befragte) erhalten eine laufende Nummer (ID)
- Diese Nummer wird durch den Buchstaben i repräsentiert
- Dieser Buchstabe dient als Laufindex: x_i ist der am i-ten Befragten gemessene Wert der Variablen x
- Die Zahl der Fälle in der Stichprobe wird mit dem Buchstaben n abgekürzt
- Eine Datenmatrix ist eine Tabelle, in der jede Zeile einem Befragten entspricht. Die Meßwerte für diesen Befragten werden in die Spalten der Tabelle eingetragen
- Die Zeilen dieser Tabelle können beliebig umsortiert werden, ohne daß Information verlorengeht

Datenmatrix

ID	Geschlecht	HF	FS	Alter
1	w	n	7	23,00
2	m	j	3	22,00
3	w	j	3	21,00
...				

Beispiel Datenmatrix EB 32 (Querschnitt Herbst 1989)

- Ca. 1387 Variablen, 23.397 Fälle
- v1 ICPR-Studiennummer, v4 ID, v5 Land(1), v6 Gewichtungsvariable, v7 Land(2)
- v10 Eintrag Wählerregister, v11 Entwicklung ind. Wirtschaftlage, v12 Lebenszufriedenheit, v13 Demokratiezufriedenheit, v14/15 Häufigkeit politischer Diskussionen, v16/17 Inglehart-Index erstes/zweites Ziel

v1	v4	v5	v6	v7	v10	v11	v12	v13	v14	v15	v16	v17
1752	1	belgium	1.0908	belgium	present a	same	very sati	not very	from time	occasiona	fight ris	maintaini
1752	2	belgium	.8069	belgium	present a	same	fairly sa	not at al	often	never	fight ris	maintaini
1752	3	belgium	.834	belgium	present a	same	fairly sa	fairly sa	from time	frequentl	fight ris	freedom o
1752	4	belgium	1.2368	belgium	present a	same	fairly sa	fairly sa	from time	occasiona	fight ris	freedom o
1752	5	belgium	.7549	belgium	present a	better	very sati	fairly sa	rarely	occasiona	freedom o	fight ris
1752	6	belgium	1.0432	belgium	present a	better	dk	fairly sa	often	never	fight ris	giving pe
1752	7	belgium	.7885	belgium	present a	same	fairly sa	fairly sa	never	occasiona	fight ris	freedom o
1752	23390	united ki	2.3624	great bri	present a	inap., fo	very sati	very sati	never	never	freedom o	maintaini
1752	23391	united ki	.9488	great bri	present a	inap., fo	very sati	fairly sa	from time	never	maintaini	freedom o
1752	23392	united ki	1.769	great bri	present a	inap., fo	fairly sa	very sati	from time	occasiona	fight ris	maintaini
1752	23393	united ki	1.1469	great bri	present a	inap., fo	not at al	not at al	rarely	occasiona	giving pe	fight ris
1752	23394	united ki	1.9318	great bri	present a	inap., fo	fairly sa	not very	often	occasiona	giving pe	freedom o
1752	23395	united ki	1.8181	great bri	present a	inap., fo	very sati	not very	from time	occasiona	giving pe	freedom o
1752	23396	united ki	1.4571	great bri	present a	inap., fo	very sati	fairly sa	from time	occasiona	giving pe	fight ris
1752	23397	united ki	2.0249	great bri	present a	inap., fo	very sati	fairly sa	often	frequentl	giving pe	freedom o

Was leisten Datenanalyseprogramme (2) ?

- Können *regelbasiert*
  - bestehende Variablen verändern
  - neue Variablen erzeugen
  - Berechnungen mit bestehenden Variablen vornehmen
  - Aus diesen Ergebnissen wiederum neue Variablen erzeugen

Wie unterscheiden sich Datenanalyseprogramme?

- spezialisierte Programme für bestimmte Verfahren (z.B. LISREL, EQS, MLWin)
  - nur für bestimmte statistische Modelle geeignet
  - oft sehr kompliziert in der Handhabung
- allgemeine Programmpakete „eierlegende Wollmilchsäue“
  - SPSS, SAS, Stata
  - relativ leicht zugänglich, breites Anwendungsspektrum
  - je unterschiedlichen Benutzerschnittstellen, Stärken, Schwächen etc.

SPSS vs. Stata

- SPSS unbestrittener Marktführer im sozialwissenschaftlichen Bereich
- Vorzüge von Stata
  - Lizenzierungspolitik und Updates
  - Geschwindigkeit (nicht im Kurs...)
  - (relativ) konsistente, extrem leistungsfähige und durchdachte Syntax
  - Kommandozeilenorientierung, motiviert zum strukturierten und nachvollziehbaren Arbeiten
  - extrem breites Spektrum an Regressionsmodellen
  - hervorragende Dokumentation
  - leichte Erweiterbarkeit, aktive Benutzergemeinschaft

Was kann Stata?

- Schwächen:
  - Faktor- und Clusteranalyse: nur „Grundausstattung“ vorhanden
  - Skalierungsverfahren kaum unterstützt (nur Cronbachs alpha)
- Stärken
  - Rekodierung und Auswahl von Untermengen
  - Vielfältige Möglichkeiten der Programmierung
  - Post-Analyse für alle statistischen Modelle (vorhergesagte Werte, Tests für Koeffizienten, lineare und nicht-lineare Kombinationen von Koeffizienten etc.)
  - Regressionsmodelle u.a. für intervallskalierte, dichotome, nominale, ordinale abhängige Variablen
  - Modellierung des Auswahlverfahrens / der Datenstruktur: Paneldaten, Zeitreihen, Klumpenstichproben, mehrstufige Zufallsauswahlen
  - Mit freiem Zusatzmodul: Strukturgleichungsmodelle und Mehrebenenanalyse
  - Druckreife Grafiken

Wo finde ich Hilfe?

- Eingebaute Hilfe (z.B. `help regression` eingeben)
- Kohler/Kreuter
- Anwendungsorientierte Einführungen zu multivariaten Verfahren
  - Backhaus et al. (diverse Auflagen)
  - Hair et al. (diverse Auflagen)
- Regression
  - Berry 1993 (knappe Einführung in die Annahmen des klassischen Regressionsmodells und deren Konsequenzen)
  - Fox 1997 (sehr umfassende Darstellung der linearen Regression und zahlreicher verwandter Modelle; eindeutig für Fortgeschrittene)

Was erwartet mich?

- Abweichung vom normalen Seminarkonzept: Keine Referate
- Das bedeutet *nicht*: Vorlesungscharakter
- Vielmehr: Präsentation + Diskussion + Übungen am PC (in Zweiergruppen) + Hausaufgaben (in Maßen)
- Dreiteilung
  - Einführung in Datenanalyse / Stata
  - Anwendungsorientierte Einführung in deskreptive Statistik und Regression
  - Replikation von Studien aus dem Bereich der empirischen Politikforschung

Was erwartet mich?

I. Einführung

29.04. Einführung und Seminarüberblick

06.05. Stata kennenlernen: *Kohler/Kreuter Kapitel 1+2 (zur Nachbereitung)*

13.05. Elemente der Syntax, Variablen erstellen und verändern: *Kohler/Kreuter Kapitel 3-5*

Was erwartet mich?

II. Typische Anwendungsfälle

27.05. Deskriptive Statistik mit Stata: Verteilungen, Tabellen, Zusammenhangsmaße. Zur Wiederholung: *Gehring und Weins, Kapitel 5-7*

17.06. Konfidenzintervalle, Hypothesentests, lineare Regression. Zur Wiederholung: *Gehring und Weins, Kapitel 8, 11, 12.*

24.06. Regression mit abhängigen kategorialen Variablen und ihre Umsetzung in Stata. *Long/Freese 2001: Kapitel 3 und 4*

01.07. Fortsetzung: Regression mit abhängigen kategorialen Variablen

Was erwartet mich?

III. Replikationen

08.07. Johnson/Martin (Einfluß des Supreme Court auf öffentliche Meinung)

15.07. Fearon/Laitin (Korrelate von Bürgerkriegen)

22.07. Arzheimer/Carter (international vergleichende Wahlforschung: Wähler der extremen Rechten)

29.07. Nadeau/Niemi/Yoshinaka (international vergleichende Wahlforschung: ökonomisches Wählen)