Analysen politikwissenschaftlicher Datensätze mit Stata

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Sitzung 2: Stata kennenlernen

Konventionen

- Die Menüs benötigen Sie nur in Ausnahmefällen. File>Exit bedeutet: Öffnen Sie das „File“-Menü (ganz links) und wählen Sie dann den Punkt „Exit“ ganz unten
- Etwas häufiger benötigen Sie Tastenkombinationen. Alt-F4 bedeutet: Halten Sie die linke Alt-Taste (neben der Leertaste) gedrückt und betätigen Sie dann die Taste F4
- Kommandos, die Sie in das Kommandofenster von Stata eintippen sollen erkennen Sie an der Schreibmaschinenschrift: `exit,clear`

Dateitypen

- *.dta-Dateien sind Datensätze.
  - enthalten Daten in binärer Form, z.B. die Ergebnisse einer Wahlstudie
- *.do-Dateien
  - enthalten Stata-Befehle
  - dienen dazu, Datensätze in eine definierte Form zu bringen (Rekodierungen etc.)
  - sowie Analysen zu dokumentieren und zu reproduzieren
  - profile.do legt notwendige Verzeichnisse an und wählt passende Voreinstellungen
- *.ado-Dateien definieren Stata-Kommandos
- *.gph-Dateien enthalten Grafiken
- Außerdem kann Stata auch Text-Dateien (*.txt), z.B. mit Ausgaben oder Befehlen erzeugen

Stata starten

- Loggen Sie sich, falls noch nicht geschehen, auf dem Terminalserver ein
- Verbinden Sie falls nötig das Netzlaufwerk z: mit dem Kurs-Share
- Starten Sie das Programm durch Doppelklick auf z:\profile.do

Bildschirmansicht

- Vergrößern Sie das Stata-Fenster maximal, um den Bildschirm optimal auszunutzen
- Das Stata-Fenster beinhaltet vier Sub-Fenster: Eingabe, Ergebnisse, Variablen und vorherige Eingaben (review)
- Schließen Sie die beiden letztgenannten

Bildschirmansicht

- Richten Sie die anderen Fenster so ein, daß möglichst viel Raum für die Ausgabe bleibt
- Mit dem Window-Menü oder den dort angegebenen Tasten-Kombinationen können Sie die geschlossenen Fenster wieder öffnen
- Beim Verlassen des Programms wird die Ansicht gespeichert, beim Starten neu geladen
- Mit Prefs>Default Windowing könne Sie die Werkseinstellung wieder herstellen

Befehlseingabe

- Das Ergebnisfenster zeigt
  - die Befehle, die in profile.do aufgerufen wurden, in weißer Schrift
  - deren Ausgabe in gelber und grüner Schrift
  - ganz unten einen weißen Punkt (command prompt)
- Tippen Sie in das Eingabefenster: `describe`
- `describe` fordert eine Beschreibung des Speicherinhalts an (derzeit keiner)
- Stata ist „case-sensitive“: Tippen Sie `Describe` oder `DESCRIBE` ein
- Die meisten Befehle lassen sich abkürzen, solange sie dabei eindeutig bleiben. Probieren Sie `desrib`, `desc`, `de` oder `d` aus

Aktuelles Verzeichnis

- Ähnlich wie früher MS/Dos verwaltet Stata ein „aktuelles Verzeichnis“
  - Datensätze aus diesem Verzeichnis können ohne Pfadangabe geladen werden
  - Dateien ohne Pfadangabe werden in dieses Verzeichnis geschrieben
- Nach dem Starten ist dies der Ordner StataSeminar auf Ihrem Laufwerk U:

Aktuelles Verzeichnis

- Tippen Sie `pwd` („print working directory“) um sich anzeigen zu lassen, welches das aktuelle Verzeichnis ist
- Mit `dir` können Sie sich dessen Inhalt anzeigen lassen
- Mit `cd` („change directory“) wechseln Sie das Verzeichnis.
- Tippen Sie jetzt `cd z:\daten\kohler-kreuter`, um in das Verzeichnis mit den Beispieldaten zu K&K zu wechseln. Alternativ können Sie auch `cd z:/daten/kohler-kreuter` eintippen

Eingabeerleichterungen

- mit ctrl-5 öffnen Sie das review-Fenster, in dem Ihre bisherigen Eingaben aufgelistet sind
  - einfacher Klick: Kommando wird in Eingabe übernommen
  - Doppelklick: Kommando wird außerdem ausgeführt
- schnellere Lösung: mit den Bild-auf und –ab-Tasten blättern Sie in Ihren Eingaben vor und zurück
- die Tab-Taste expandiert Variablennamen, sofern diese eindeutig sind

Einen Datensatz laden

- Tippen Sie `dir *.dta`, für eine Liste aller Datensätze in diesem Verzeichnis
- Falls die Liste nicht in das Fenster paßt, wird die Ausgabe angehalten, und es erscheint die Aufforderung – more –
  - Drücken Sie die Leertaste. Die Ausgabe rollt dann um eine Fensterhöhe weiter
  - Mit der Eingabetaste rollt die Ausgabe um eine Zeile weiter
  - Mit q, Strg-Pause oder dem weiß-roten Symbol unterbrechen Sie die Ausgabe
  - Mit dem Rollbalken oder dem Mausrad können Sie zurückblättern
  - Dies funktioniert bei *allen* Ausgaben von Stata

Einen Datensatz laden

- Mit `use data1` oder `use data1.dta` laden Sie den Datensatz (eine Auswahl aus dem SOEP) in den Speicher
- Tippen Sie wieder d ein
- Sie erhalten Informationen über die Nutzung des Speichers sowie über die Variablen
- Mit `drop einzug – np9507` löschen Sie alle Variablen vom Jahr des Einzugs bis zu den Sorgen über den Arbeitsplatz aus dem Datensatz

Einen Datensatz anschauen

- Mit `browse` können Sie sich wie bei SPSS die Datenmatrix anzeigen lassen, die aber meist von geringem Interesse ist
- `list` zeigt nacheinander die Ausprägung *aller Variablen* für alle Fälle. Brechen Sie die Ausgabe ab
- `list sex eink` listet nacheinander Geschlecht und Einkommen Ihrer Befragten

Auswahl von Fällen: in

- Sie interessieren sich für Geschlecht und Wohnort der ältesten Befragten
  - sortieren Sie den Datensatz nach dem Geburtsjahr: sort gebjahr
  - Lassen Sie sich für die Fälle 1 bis 5 die entsprechenden Variablen zeigen: `list gebjahr sex bul in 1/5`
- Die in-Bedingung
  - kann mit den meisten Kommandos verwendet werden
  - beschränkt das Kommando auf die Fälle mit der entsprechenden Ordnungsnummer
  - Ordnungsnummer bezieht sich auf *aktuelle Sortierung*
  - in 1/5: Fälle 1-5
  - in 202: Fall 202
  - in -10/-1: zehnter Fall vom Ende der Datei bis letzter Fall

Einfache Maßzahlen

- `summarize sex eink` berechnet Fallzahl, Mittelwert, Standardabweichung, Minimum und Maximum für beide Variablen
- Mittelwert und Standardabweichung für Geschlecht nicht interpretierbar
- Fallzahlen unterscheiden sich: Warum?

Kodierung

- Sie möchten wissen, wie die Variable sex kodiert ist
  - `tabulate sex` zeigt Ihnen, wie viele Männer und Frauen der Datensatz enthält, nicht aber, welche numerischen Werte für „männlich“ und „weiblich“ stehen
  - `d sex` zeigt ihnen, daß der Variable ein „value label“ gleichen Namens zugeordnet ist
- Jetzt haben Sie zwei Möglichkeiten
  - `label list sex` zeigt Ihnen die Kodierung
  - `numlabel sex,add` fügt dem „value label“ die numerischen Werte hinzu, so daß Sie bei `tab sex` nun gleich erkennen können, wie die Variable kodiert ist

Auswahl von Fällen: if

- Sie interessieren sich für Einkommensunterschiede zwischen Männern und Frauen
  - Mit `summ eink if sex==1` berechnen Sie den Mittelwert für die Männer
  - `summ eink if sex==2` liefert das Ergebnis für die Frauen
- `if` wählt Fälle aufgrund einer *logischen Bedingung* aus
- == steht für Gleichheit (ein einfaches = bedeutet eine Zuweisung und führt hier zu einer Fehlermeldung)

Auswahl von Fällen: if

- Sie wollen die Analyse auf Personen mit einem eigenen Einkommen beschränken
  - `summ eink if sex==1 & eink ~=0`
  - `summ eink if sex==2 & eink ~=0`
- & steht für eine logische Verknüpfung: beide Bedingungen müssen erfüllt sein, damit der Fall in die Analyse eingeht
- ~= bedeutet ungleich: das Einkommen ist *nicht* gleich null

Auswahl von Fällen: if

- Sie wollen die Analyse auf Personen mit einem eigenen Einkommen von mindestens 500 Mark beschränken
  - `summ eink if sex==1 & eink >=500`
  - `summ eink if sex==2 & eink >=500`
- >= steht für größer/gleich, d.h., das Einkommen hat einen Wert von 500 oder mehr

Das Präfix by

- Die Einkommensunterschiede zwischen Männern und Frauen lassen sich auch mit dem by-Präfix analysieren
- by führt ein Kommando für Subgruppen aus, die durch eine oder mehrere Variablen definiert sind
- `by sex: summ eink` führt zu einem Fehler, da die Daten vorher mit `sort sex` sortiert werden müssen
- alternativ `bysort sex: summ eink`
- Auch hier können Sie logische Bedingungen angeben: `by sex: summ eink if eink>=500`

Optionen

- Die meisten Stata-Befehle akzeptieren zusätzliche Optionen, die mit einem Komma abgetrennt werden
- Beispielsweise zeigt `summarize eink, detail` die Perzentile etc. an
- Viele Optionen können ebenfalls abgekürzt werden: `summ eink,det`

Häufigkeitsverteilungen

- `tab sex` zeigt Ihnen die Zahl der Männer und der Frauen
- mit `tab fam sex` erzeugen Sie eine Kreuztabelle für die Merkmale Familienstand (Zeilen) und Geschlecht (Spalten)

Label

- Datenanalyse Programme kennen Variablen für
  - Buchstabenketten („Strings“)
  - Zahlen (in verschiedenen Varianten)
- In der Praxis werden fast nur numerische Variablen verwendet
- Diese haben meist kurze und wenig intuitive Namen (v101)
- Dichotome, nominale oder ordinale Merkmale werden in der Regel durch kleine ganze Zahlen repräsentiert
- Um Eingabefehler zu reduzieren und die Ausgabe lesbarer zu machen, vergibt man für Variablen (v101 ˜ Konfession) und Werte (1 ˜ katholisch, 2 ˜ protestantisch etc.) *Labels*

Label

- Stata kennt u.a.
  - Label für Variablen, denen damit eine Erläuterung hinzugefügt wird
  - Label für *Werte von Variablen*, deren Bedeutung dadurch dokumentiert wird (z.B. ja=1, nein=0)
  - Anders als bei SPSS können solche Werte-Labels auf beliebig viele Variablen angewendet werden

Label

- Mit `d est` die Variable est anzeigen
- `label var est "Erwerbstaetigkeit 97"` verändert das Label der Variablen
- Den Werten der Variablen ist kein Label zugeordnet
- Mit `label def estlb 1 "Vollzeit" 2 "Teilzeit" 3 "Umschulung" 4 "unregelmäßig" 5 "arbeitslos" 6 "Wehrdienst"  7 "nicht erwerbstätig"` definieren Sie ein Wertelabel
- Schon bei der Eingabe sehen Sie, daß deutsche Sonderzeichen immer noch Probleme machen
- Mit `lab val est estlb` ordnen Sie das neue Werte-Label estlb der Variablen est zu

Einfache Rekodierungen

- Geschlecht wird üblicherweise als Dummy (Variable mit den Ausprägungen 0/1) kodiert
- Um die vorhandene Variable sex zu rekodieren, erzeugen Sie zuerst eine Variable „male“ die den Wert 1 annimmt, wenn sex gleich 1 ist:
- `generate male=1 if sex==1`
  (Achten Sie auf die Gleichheitszeichen!)
- Setzen Sie diese neue Variable auf 0, wenn sex gleich 2 ist: `replace male=0 if sex==2`

Einfache Rekodierungen

- generate erzeugt eine Variable, der zugleich ein Wert zugewiesen werden kann
- replace ersetzt den Wert einer bestehenden Variablen durch einen neuen Wert
- Beide Kommandos können mit if (oder in) eingeschränkt werden

Einfache Rekodierungen

- Wird generate eingeschränkt, bleibt die Variable für alle Fälle, die nicht unter die if-Klausel fallen oder nachträglich verändert werden, undefiniert (missing)
- Undefinierte, d.h. fehlende Werte, werden durch einen Punkt repräsentiert (entspricht sysmis bei SPSS). Weitere missing-Werte sind möglich, werden aber selten verwendet

Einfache Rekodierungen

- Der Erwerbsstatus soll in eine dichotome Variable „Vollzeit“ (vs. Teilzeit) rekodiert werden. Andere Erwerbsverhältnisse sind hier nicht von Interesse
- `generate vollzeit=1 if est==1`
- `replace vollzeit = 0 if est ==2`
- `tab voll`
- `tab voll,miss`

Einfache statistische Kontrolle

- Hat das Geschlecht auch dann einen Effekt auf das Einkommen, wenn der Faktor „Vollzeiterwerb“ kontrolliert wird?
- `bysort voll male: summ eink if eink>0`

Hilfe!!!

- Vor allem am Anfang werden Sie mit Sicherheit Syntax, Optionen oder sogar den Namen eines Befehls vergessen
- Mit `search` können Sie nach Stichworten suchen: `search label`
- Mit `help` rufen Sie die Dokumentation eines Befehls auf. Der Name darf abekürzt werden: `help desc`
- In beiden Fällen kann es sinnvoll sein, vorher über Help>Contents oder Help>Search ein eigenes Fenster zu öffnen

Reproduzierbare Analysen: Do-Files

- Ernsthafte Analysen müssen nachvollziehbar sein
- Dazu brauchen Sie zunächst ein Protokoll Ihrer interaktiven Sitzung
- `log using u:\StataSeminar\log.txt,replace` (in der letzten Zeile von profile.do) fordert Stata auf, alle Ein- und Ausgaben der laufenden Sitzung in eine Textdatei zu speichern
- `cmdlog using u:\StataSeminar\kommandos.txt`,replace speichert zusätzlich Ihre Eingaben in einer separaten Datei
- Beide Dateien können Sie z.B. im Wordpad oder in WinWord öffnen
- Innerhalb von Stata können Sie die Dateien mit `view u:\StataSeminar\kommandos.txt` bzw. `view u:\StataSeminar\log.txt` betrachten

Reproduzierbare Analysen: Do-Files

- Zum Betrachten von Log-Dateien und zum Erstellen von Do-Dateien empfiehlt sich die Verwendung des Editors Emacs
- Emacs
  - extrem mächtig
  - unterstützt Programmiersprachen, auch Stata
  - tastaturorientiert, in diesem Kurs keine grundlegende Einführung möglich, alle wichtigen Funktionen auch über Menüs erreichbar

Reproduzierbare Analysen: Do-Files

- Installiert in z:\emacs-20-7\
  - dort ins Unterverzeichnis bin gehen
  - Mit der rechten Maustaste einmal auf runemacs.exe klicken und das Symbol auf den Desktop ziehen
  - Unter den drei Optionen „Verknüpfung hier erstellen“ wählen
  - Wenn Sie wollen: Mit der rechten Maustaste auf das neue Symbol klicken, „Eigenschaften“ wählen und unter „Arbeitsverzeichnis“ u:\StataSeminar eintragen
  - Durch Doppelklick auf das neue Symbol starten Sie Emacs in Ihrem privaten Stata-Verzeichnis

Reproduzierbare Analysen: Do-Files

- Um Ihre letzte Analyse zu dokumentieren
  - Starten Sie Emacs wie beschrieben
  - Öffnen Sie die Do-File-Vorlage z:\muster.do
    - ctrl-x ctrl-f oder Files>Open File
    - Ganz unten öffnet sich eine Kommado-Zeile
    - Mit ctrl-a an den Anfang springen, mit ctrl-k den Vorschlag löschen und statt dessen z:\muster.do eintippen (drücken Sie versuchsweise nach z:\m die Tab-Taste!)
    - Return

Reproduzierbare Analysen: Do-Files

- Speichern Sie die Vorlage unter u:\StataSeminar\sexeink.do
  - Wählen Sie Files>Save Buffer As (oder ctrl-x ctrl-w
  - Löschen Sie wie vorher den vorgeschlagenen Namen
  - Geben Sie statt dessen u:\StataSeminar\sexeink.do an (Probieren Sie aus, was passiert, wenn Sie nach Eingabe von u:\St die Tab-Taste drücken)

Reproduzierbare Analysen: Do-Files

- Teilen Sie das Emacs-Fenster mit ctrl-x 2 oder Files>Split Window
- Laden Sie die Datei kommandos.txt. Da diese im selben Verzeichnis liegt, können Sie den Pfad beibehalten. Tippen Sie einfach k und drücken die Tab-Taste

Reproduzierbare Analysen: Do-Files

- Ziemlich weit oben in kommando.txt sehen Sie den cd-Befehl, mit dem Sie das Verzeichnis gewechselt haben, und danach den use-Befehl
- Die Do-Datei soll aber in Ihrem eigenen Verzeichnis ablaufen und Dateien schreiben können
- Tippen Sie deshalb in die neue do-Datei `use z:\daten\kohler-kreuter\data1, replace`
- Die Option „replace“ bedeutet, das evtl. vorhandene Daten im Speicher ersetzt werden sollen

Reproduzierbare Analysen: Do-Files

- Ziemlich am Ende von kommandos.txt steht der eigentliche Analysebefehl `bysort voll male: summ eink if eink>0`
- Kopieren Sie diesen, und fügen Sie ihn in die Do-Datei ein
  - zu kopierenden Text mit der Maus oder mit ctrl-Leertaste plus Cursor-Tasten markieren
  - mit Edit>Copy oder Alt-W kopieren
  - mit Edit>Paste oder Ctrl-Y einfügen

Reproduzierbare Analysen: Do-Files

- Vor der Auswertung benötigen Sie noch die Rekodierungen für die Variablen male und vollzeit, die Sie zwischen beiden Befehlen einfügen müssen
- Mit Pos1 und Ende springen Sie an den Anfang und das Ende der Datei
- Mit Ctrl-a und Ctrl-e springen Sie an den Anfang und das Ende einer Zeile (explizite Zeilenumbrüche beachten)

Reproduzierbare Analysen: Do-Files

- Mit Search>Search und Search>Repeat Backwards können Sie vorwärts und rückwärts suchen
- Schneller und sinnvoller ist meist die inkrementelle Suche
  - Tippen Sie ctrl-S und beginnen Sie, das gesuchte Wort einzutippen. Meist genügen wenige Buchstaben
  - Durch erneutes Drücken von ctrl-S springen Sie zur nächsten Fundstelle
  - Mit ctrl-R suchen Sie in gleicher Weise rückwärts

Reproduzierbare Analysen: Do-Files

- In der Do-Datei markiert Emacs Schlüsselwörter und rückt die Zeilen sinnvoll ein, sobald Sie Tab oder return drücken
- Wenn Sie die Do-Datei fertiggestellt haben, speichern Sie sie mit Files>Save Buffer oder ctrl-x ctrl-s
- Wechseln Sie zu Stata und geben Sie ein: `do u:\stataseminar\sexeink`

Reproduzierbare Analysen: Do-Files

- Sie können Do-Files auch innerhalb von Stata bearbeiten: `doedit u:\stataseminar\sexeink`
- Der Editor ist nicht so leistungsfähig und kann jeweils nur eine Datei öffnen
- Dafür können Sie einzelne Befehle markieren und ausführen lassen (Tools>Do Selection).
  - sinnvoll, wenn Sie die interaktiv die optimale Variante eines Befehls suchen
  - Evtl. doedit ohne Dateinamen starten. Auf diese Weise können Sie interaktiv mehrere Befehle oder besonders komplexe Befehle ausprobieren, die Sie dann mit Copy&Paste in eine Do-Datei einfügen

Hausaufgabe

- Erzeugen Sie unter Verwendung von Muster.do eine Datei nrw-sexeink50.do, die
  - die vorherige Analyse wiederholt
  - aber auf das Bundesland NRW beschränkt ist und
  - nur Einkommen von mindestens 50 Mark berücksichtigt
- Schicken Sie die Lösung bis zum nächsten Mittwoch an do-files@politik.uni-mainz.de
- Verwenden Sie dafür folgendes Schema:
  - als Betreff nrw-sexeink50.do verwenden
  - erste Zeile: Ihr Name, zweite Zeile leer lassen
  - ab der dritten Zeile: Text des do-Files per copy&paste einfügen (bitte nicht als attachement schicken)
  - Ansonsten bitte kein Text