

GSOEP: Generieren von Analyse-Files im Querschnitt und Längsschnitt

Blockseminar 18. Januar 2004

Zusammenführung von Datensätzen mit Stata

Prinzipiell gibt es drei Kommandos:

1. *append*: Datenfiles untereinander, gleiche Var-Zahl, größere Fallzahl

-> '*lange*' Form

append using filename [, nolabel]

-einfache Struktur, problemlose Anwendung

-für SOEP-Daten nicht geeignet, da jedes Jahr
andere Var-Names

2. *merge*: Datenfiles nebeneinander: größere Var-Zahl, gleiche Fallzahl

-> '*weite*' Form

merge [varlist] using filename [filename...]

-komplexer und riskanter in der Anwendung

-Für SOEP-Daten notwendig

-wenn 'langer' Datensatz notwendig -> ***reshape***

(3. *joinby*: wird hier nicht näher erläutert)

Der Befehl *MERGE*

Beschreibung:

- *merge* fügt korrespondierende Beobachtungen (Variablen) zusammen
- Dem derzeit geöffneten Datensatz (*master dataset*) werden Variablen eines gespeicherten Datensatzes hinzugefügt (*using dataset*)

Options:

- *keep(varlist)* spezifiziert diejenigen Variablen, die aus dem *Using Dataset* beibehalten werden sollen. Wenn *keep* nicht spezifiziert ist, werden alle Variablen beibehalten

Besonderheiten:

- Variablen müssen genau benannt werden
- Es kann kein Bereich spezifiziert werden

Übung: einfaches Zusammenfügen (1)

```
. cd "c:\data\kkdata"  
. copy blqm.dta "c:\data\mo14-16\project7\blqm.dta"  
. copy blbev.dta "c:\data\mo14-16\project7\blbev.dta"  
. cd "c:\data\mo14-16\project7"  
. use blbev.dta, clear  
. merge using blqm.dta  
. list land bev95 bila qm
```

Zwei weitere Variablen im Datensatz, jedoch unbrauchbares Ergebnis:

- Flächen und Bevölkerungszahlen gehören zu unterschiedlichen Bundesländern
- *merge using* ohne Zusatz setzt voraus, dass in beiden Files die Fälle in der gleichen Reihenfolge untereinander stehen -> vorher sortieren!!!

```
. use blqm.dta, clear  
. sort bula  
. save blqm2.dta  
. use blbev.dta, clear  
. sort land  
. merge using blqm2.dta  
. list land bev95 bila qm
```

Übung: einfaches Zusammenfügen (2)

- Für die ersten 8 Bundesländer gewünschter Erfolg, dann verschieben sich die Spalten
- Tritt auf, wenn die Files eine unterschiedliche Fallzahl haben
- Lösung Schlüsselvariable: gewährleistet eine eindeutige Zuordnung
- Hier "Bundesland" als Schlüsselvariable -> muss in beiden Files den gleichen Namen haben:

```
. use blqm.dta, clear
. generate str22 land = bul
. sort land
. save blqm2.dta, replace
. use blbev.dta, clear
. sort land
. merge land using blqm2.dta
. list land bev95 bula qm
```

Ergebniskontrolle der *merge*-Prozedur

Zur Ergebniskontrolle wird bei einem *merge*-Befehl automatisch eine Kontrollvariable gebildet:

_merge

- Steht immer in der letzten Spalte des neuen Datensatzes
- Gibt an, aus welchem der beiden Datensätze die Informationen kommen

| | |
|------------------|---|
| <i>_merge==1</i> | <i>obs. from master data</i> |
| <i>_merge==2</i> | <i>obs. from using data</i> |
| <i>_merge==3</i> | <i>obs. from both master and using data</i> |

. tab *_merge*

-> zeigt, dass für 14 der 16 Bundesländer die Flächendaten zugespielt wurden.

Längsschnitt - 'lange' und 'weite' Form

'Lange' Form:

| | t_1 | | t_2 | | |
|-------|----------|----------|----------|----------|----------|
| | x_{11} | x_{12} | x_{21} | x_{22} | x_{nt} |
| p_1 | | | | | |
| p_2 | | | | | |
| p_3 | | | | | |

'Weite' Form:

| | x_1 | x_2 |
|-------|----------|-------|
| t_1 | p_{11} | |
| | p_{21} | |
| | p_{31} | |
| t_2 | p_{12} | |
| | p_{22} | |
| | p_{32} | |
| | \vdots | |
| t | p_{1t} | |
| | p_{2t} | |
| | p_{3t} | |

reshape-Kommando - Struktur und kleine Übung

- Spezielle Befehle zur Analyse von Paneldaten werden in Stata als "xt"-Kommandos bezeichnet
- Für die Anwendung von Panelanalysen in Stata ("xt"-Kommandos) müssen Daten im *langen* Format vorliegen, die Umformung erfolgt mit *reshape*
reshape long *varnames, i(varlist) [j(varname [values]) ...]*

i(varlist) spezifiziert die Variablen, deren eindeutige Werte eine logische Beobachtung bezeichnen

j(varname [values]) spezifiziert die Variablen, deren eindeutige Werte eine Teilbeobachtungen bezeichnen

Beispiel: *Kopiere data2.dta aus 'kkstata' in 'project7'*

```
. use data2.dta, clear
. keep persnr gebjahr sex wohngr* hhgr*
. reshape long wohngr hhgr, i(persnr) j(welle)
```

Damit "xt"-Kommandos verwendet werden können, muss Stata bekannt sein, welche Daten die Personen und welche die Zeit definieren

```
. iis persnr
. tis welle
```

GSOEP-Datenstruktur: Querschnitt ↔ Längsschnitt

Querschnittsdaten

nur zu einem Zeitpunkt erhoben oder keine zeitspezifischen Informationen

Längsschnittdaten

zu mehreren Zeitpunkten erhoben oder zeitspezifische Informationen

Im GSOEP jährliche Interviews, deshalb:

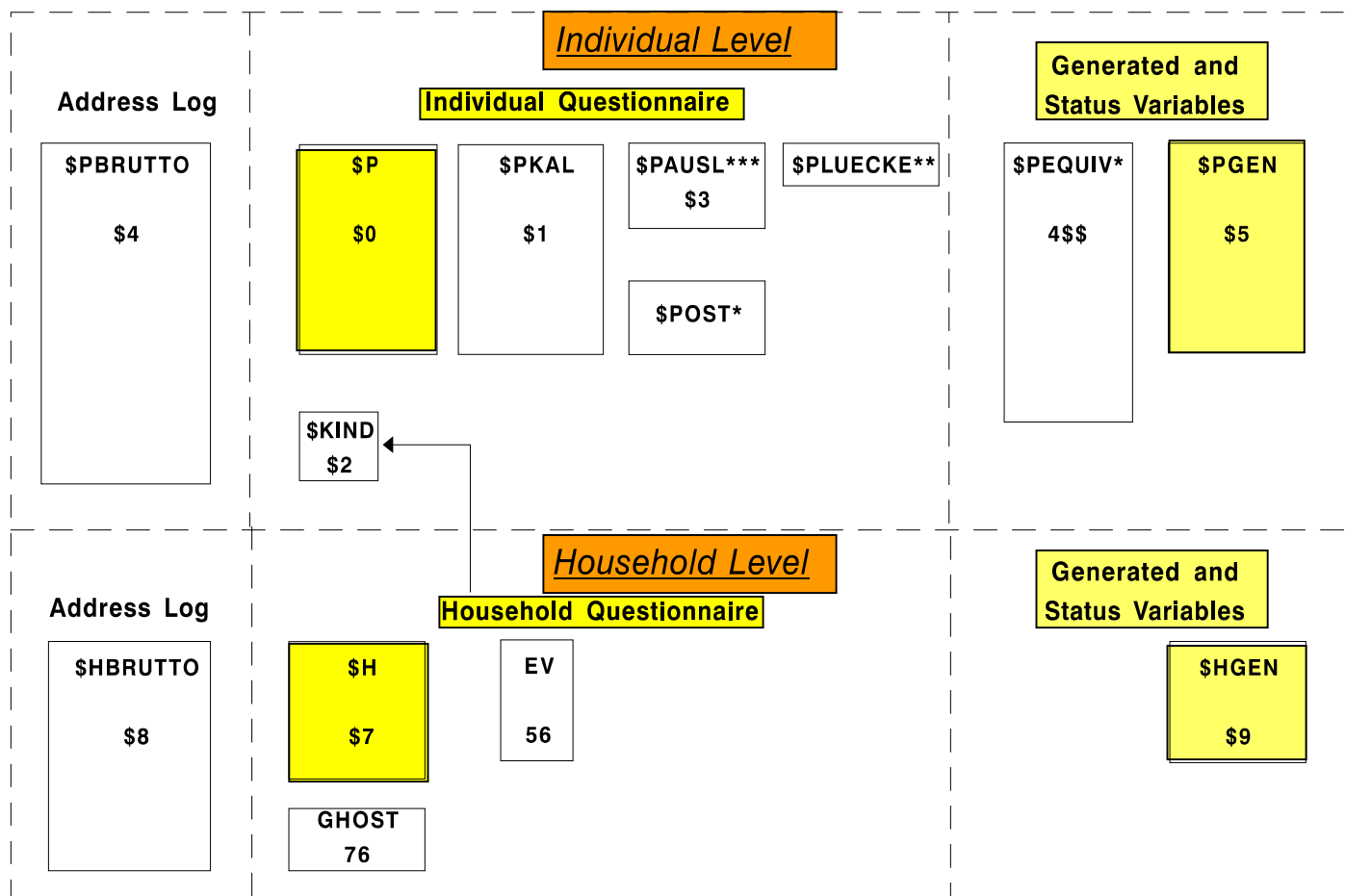
grundlegende Datenstruktur der Rohdaten: Querschnitt

In jeder Welle einzelne Datensätze für

- *Haushaltsdaten (\$H)*
- *Personendaten (\$P)*
- *Kinder (\$KIND)*
- *Häufig verwendete Variablen \$PGEN, \$HGEN*

=> Sowohl für Querschnitts- als auch für Längsschnittanalysen sind "match-" und "merge"-Prozeduren erforderlich

Die SOEP Datenstruktur (Querschnitt)



\$: Wave specification: A,B,C ... N for file names; 1,2,3 ... 14 for file numbers.

* Waves G and H only; ** Waves B through M only; *** Waves A through L only

Generieren eines Längsschnitt-Analyse-Files im GSOEP

Kann als zweistufiger Prozess betrachtet werden:

1. Generieren von Querschnitt-Files
 - für jedes Analysejahr
 - aus verschiedenen Rohdatensätzen eines Jahres
2. Zusammenfügen der Querschnitt-Files zu einem Längsschnitt-File

Fallzuordnung in *merge*- und *match*-Prozeduren

1. Eineindeutige Fallzuordnung: PERSNR
2. Jährlicher Haushalts-Identifizier: HHNRAKT
(genau die gleiche Information in \$HHNR)
3. Jeder Person kann ein "Ursprungshaushalt" zugeordnet werden

... Diese Informationen sind im Master-File PPFAD abgelegt

***PPFAD* enthält:**

- Alle Mitglieder aller Haushalte, die zu irgend einem Zeitpunkt erfasst wurden (interviewte Personen, Kinder und auch alle, die nie geantwortet haben)
- Überprüfte Informationen zu Geschlecht und Geburtsjahr

Spezifikation eines Analyse-Files

1. Über Fragebogen - Gibt einen genauen Überblick über die Struktur der Fragen. Sinnvoll u.a. bei der Hypothesenbildung und bei der Auswertung (z.B. bei "überraschenden" Ergebnissen)
2. Mit SOEP-Info / Item Correspondence online auf der DIW-Webseite
 - Nach Variablen-Name
 - Topics
 - Auch Online über Fragebogen
3. DOS-Version von SOEP-Info
 - Nachteil: unkomfortabler, mangelnde Aktualität
 - Vorteil: netzunabhängig
 - (zip-File ist im temp-Verzeichnis)

Variablenauswahl über SOEP-Info

WWW-SOEPInfo [SOEP-DB 2002] - Microsoft Internet Explorer

Adresse <http://panel.gsoep.de/soepinfo2002/>

[SOEPInfo-Main-Actions] [Basic Information] [Help]

Basket: 0 Vars

[Basket-Actions] [Select all] [Clear] [Delete] [Files]

[SOEPInfo-Main-Actions]

- Vaname Search
- Word Search
- Topics**
- Topics (long)
- Questionnaires
- Load Basket
- Language

SOEPInfo - Themengebiete

Haupt-Themen

- Demographie, Bevölkerung und Biographie
- Arbeitsmarkt und Beschäftigung
- Einkommen, Steuern und Soziale Sicherung
- Wohnen
- Gesundheit
- Leistungen und Ausgaben privater Haushalte
- Bildung und Qualifikation
- Grundorientierungen, Partizipation und Integration
- Verkehr und Umwelt
- Bruttoinformation und Methode
- Jugendbiographie
- Sozialisation
- Berufsbiographie

(top)

1 Demographie, Bevölkerung und Biographie

- 1.1 Bevölkerungsveränderung
- 1.2 Haushaltsstruktur
- 1.3 Familienstruktur
- 1.4 Netzwerke
- 1.5 Migration

[Variables] [Topics] [Questionnaires]

Fertig

Start Explorer - t... Stata-GSD... WWW-S... Intercooled... gen-p-q do... Internet

16:40

Variablenauswahl über SOEP-Info

The screenshot shows the SOEP-Info website interface. The browser title is "www-SOEPinfo (SOEP-DB 2002) - Microsoft Internet Explorer". The address bar shows "http://panel.gsoep.de/soepinfo2002/". The main content area displays a list of variables with their descriptions and response scales. A red arrow points to the "Load Basket" button in the left-hand menu.

Variables listed:

- [SP0101] mit Ihrer Gesundheit?
0 1 2 3 4 5 6 7 8 9 10
- [SP0102] (falls Sie erwerbstätig sind) mit Ihrer Arbeit?
0 1 2 3 4 5 6 7 8 9 10
- [SP0103] (falls Sie im Haushalt tätig sind) mit Ihrer Tätigkeit im Haushalt?
0 1 2 3 4 5 6 7 8 9 10
- [SP0104] mit dem Einkommen Ihres Haushalts?
0 1 2 3 4 5 6 7 8 9 10
- [SP0105] mit Ihrer Wohnung?
0 1 2 3 4 5 6 7 8 9 10
- [SP0106] mit Ihrer Freizeit?
0 1 2 3 4 5 6 7 8 9 10
- [SP0107] (falls Sie Kinder im Vorschulalter haben) mit den vorhandenen Möglichkeiten der Kinderbetreuung?
0 1 2 3 4 5 6 7 8 9 10
- [SP0108] mit der Krankenversicherung, der Arbeitslosen-, der Renten- und der Pflegeversicherung in der Bundesrepublik, also mit dem, was man das Netz der sozialen Sicherung nennt?
0 1 2 3 4 5 6 7 8 9 10
- [SP0109] mit dem Zustand der Umwelt hier in der Region?
0 1 2 3 4 5 6 7 8 9 10

Variablenauswahl und erstellen eines DO-Files

www.SOEPinfo [SOEP-DB 2002] - Microsoft Internet Explorer

Adresse http://panel.gsoep.de/soepinfo2002/

[SOEPinfo-Main-Actions] [Basic Information] [Help]

Basket: 3 Vars

| | | | |
|---------|----|------|---------------|
| PP13002 | PP | 1999 | Gebur ts Jahr |
| QP13902 | QP | 2000 | Gebur ts Jahr |
| RP13002 | RP | 2001 | Gebur ts Jahr |

[Basket-Actions] Select all Clear Delete [Files]

- [Basket-Actions]
- Standards -
- Frequencies
- Items
- Items & Frequencies
- Questionnaires
- List
- Generators -
- SPSS
- SAS
- STATA**

Meta: GEBJAHR (PPFAD)

Files: xP

| | | | | | | | | | | | | | |
|---------|---------|---------|---------|---------|---------|---------|---------|------|------|------|------|------|------|
| 1984 | 1985 | 1986 | 1987 | 1988 | 1989 | 1990 | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 |
| AF12Z | BP81 | CP8802 | DP9002 | EP8102 | FP1000 | | | | | | | | |
| 1991 W. | 1991 O. | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 | | | | | | |
| HP10002 | - | IP10002 | JP10002 | KP10002 | LP10002 | MP10502 | NP11202 | | | | | | |
| 1998 | 1999 | 2000 | 2001 | 2002 | | | | | | | | | |
| OP11802 | PP13002 | QP13902 | RP13002 | SP13002 | | | | | | | | | |

Geburtsjahr

Meta: GEBJAHR (PPFAD)

Files: xKIND

| | | | | | | | |
|----------|----------|----------|---------|---------|---------|---------|---------|
| 1984 | 1985 | 1986 | 1987 | 1988 | 1989 | 1990 W. | 1990 O. |
| AKGEBURT | BKGEBURT | CKGEBURT | DKGJAHR | EKGJAHR | FKGJAHR | GKGJAHR | - |
| 1991 W. | 1991 O. | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 |
| HRGJAHR | - | IRGJAHR | JRGJAHR | KRGJAHR | LKGJAHR | MKGJAHR | NRGJAHR |
| 1998 | 1999 | 2000 | 2001 | 2002 | | | |
| ORGJAHR | PKGJAHR | QKGJAHR | RKGJAHR | SKGJAHR | | | |

Geburtsjahr (4-Steller)

Meta: GEBJAHR (PPFAD)

[Variables] [Topics] [Questionnaires]

Spezifizierung des DO-Files

www.SOEPinfo [SOEP-DB 2002] - Microsoft Internet Explorer

Adresse: http://panel.gsoep.de/soepinfo2002/

[SOEPinfo-Main-Actions] [Basic Information] [Help]

Basket: 3 Vars

| | | | |
|---------|----|------|-------------|
| PP13002 | PP | 1999 | Geburtsjahr |
| QP13902 | QP | 2000 | Geburtsjahr |
| RP13002 | RP | 2001 | Geburtsjahr |

[Basket-Actions] Select all Clear Delete [Files]

STATA Options

Data Files Path: Save

Temp Path: Save

Level: Individuals Households

Panel Data Design: Balanced Unbalanced

Unit of Analysis: Only Adult Respondents All Sample Members

Gender: Both Male Female

Original Sample:

| | | |
|---|---|--|
| <input type="checkbox"/> A German West | <input type="checkbox"/> B Foreigner West | <input type="checkbox"/> C German East |
| <input type="checkbox"/> D 84-93 Immigrant | <input type="checkbox"/> E Refreshment 1998 | <input type="checkbox"/> F ISOEP 2000 |
| <input type="checkbox"/> G High Income 2002 | | |

Geographic Region: Both West East

Generate STATA Code

[Variables] [Topics] [Questionnaires]

Taskbar: Start, Internet, Explorer - t..., Stata-GSO..., www-S..., Intercooled..., gen-p-q.do..., 16:48