

SOEP and SAS*

U. Rendtel

October 1996

Contents

1 SAS File Organisation	2
2 Linking Files	2
3 Creating Index-Files	3
4 Matching individual and household information in a cross-section	3
5 Matching individual information across time	5
6 Aggregation of individual information on household level	7
7 Use of the Macro Language	7
8 Preparing files for longitudinal analysis	9
9 Pooled regression analysis	10
10 Random effects models	10
11 Fixed effects models	11
12 A useful SAS macro for the estimation of fixed and random effects models	13
13 Variance estimation for population totals	17
13.1 Bounds assuming independent selection of initial households . . .	17
13.2 A SAS-macro for the computation of the bounds	19
13.3 Variance estimates by random groups	19
14 Duration analysis	22
15 Exercises	23
16 Solutions for exercises	25

*This is a revised version of the corresponding to Section 4 in the SOEP Desktop Companion

1 SAS File Organisation

The directory in which the data are stored has to be specified in the SAS programs:

```
LIBNAME soep 'c:\soepdir';
```

Value labels are stored in an extra file (in our case : *FORMATS.SC2*). The SAS system looks for this file whenever a dataset with value labels is opened. Only the directory in which this file is stored has to be specified in the form:

```
LIBNAME library 'c:\soepdir';
```

A convenient way is the use of a macro variable `path` that contains the actual path where the soep files and its formats are located:

```
%let path='/home/FB/fb02/poetter/soep/kurs/';  
/* This is the location in Frankfurt */  
libname soep &path;  
libname library &path;  
run;
```

The set up of the libnames has to be done only at the beginning of a SAS session.

Data access:

```
data soep.ap; set soep.ap;
```

2 Linking Files

There are several methods of linking files together:

- linking household files of one wave:
link-variables: HHNR, HHNRAKT
example: *BH, BHBRUTTO*
- linking individual files of one wave:
link-variables: HHNR, HHNRAKT, PERSNR
(data are sorted this way)
example: *BP, BPBRUTTO, BPGEN*
- linking all individuals of different waves who stem from the same original household :
one link-variable: HHNR
- linking individual files of different waves:
only one link-variable : PERSNR
(individual can live in another household now, thus HHNRAKT is not necessarily the same in different waves)
example : *AP, BP, CP*
Data are not sorted this way.

In the last example, where the individual files *AP*, *BP*, and *CP* are linked, the data have to be sorted first.

```
PROC SORT DATA=soep.ap out=ap; BY persnr;
PROC SORT DATA=soep.bp out=bp; BY persnr;
PROC SORT DATA=soep.cp out=cp; BY persnr;
DATA wave1to3;
MERGE ap bp cp; BY persnr; RUN;
```

An alternative is to create Index-files.

3 Creating Index-Files

```
PROC DATASETS LIBRARY=soep;
MODIFY ap; INDEX CREATE persnr; RUN;          ==> creates ap.si2
MODIFY bp; INDEX CREATE persnr; RUN;          ==> creates bp.si2
MODIFY cp; INDEX CREATE persnr; RUN;          "

MODIFY apbrutto; INDEX CREATE persnr; RUN;    "
MODIFY bpbrutto; INDEX CREATE persnr; RUN;    "
MODIFY cpbrutto; INDEX CREATE persnr; RUN;    "

MODIFY apgen; INDEX CREATE persnr; RUN;       "
MODIFY bpgen; INDEX CREATE persnr; RUN;       "
MODIFY cpgen; INDEX CREATE persnr; RUN;       "
```

Now the data can be linked by *PERSNR* without first sorting. Whenever a *BY*-variable is used the SAS system looks automatically for the corresponding index-file. The link of household and individual data of one wave, can still be done without an index by using the original sorting order (*HHNR HHNRAKT PERSNR*).

The gain in computation time by using index variables can be substantial. Besides the programs become easier to read and are shorter. You get information on existing index files by the command `PROC DATASETS LIBRARY=soep;`.

4 Matching individual and household information in a cross-section

Example: Determinants of life satisfaction (Cross-sectional analysis)

Dependent variable: Score of life satisfaction

Covariates: Gender, age, per-capita income in household, nationality

Step 1: Use `soepinfo` Where is the information and what are the names of the variables ?

- Life satisfaction: BP9301 in *BP*
- sex: BSEX in *BPBRUTTO*
- age (year of birth): BGEBURT in *BPBRUTTO*
- household size: BHHGR in *BHBRUTTO*
- household income: BH39 in *BH*

- nationality (language version of questionnaire): BPERFAS in *BP*

There are different levels with different case numbers: Household vs. Individuals, and Gross vs. Net.

Step 2: Linking gross and net information at household level:

```
/* file soep07.sas */
DATA bhh;
  MERGE
    soep.bhhbrutto (KEEP=hhnr hhnrakt bhhgr RENAME= (bhhgr=hhsz) )
    soep.bh (KEEP=hhnr hhnrakt bh39 RENAME= (bh39=hhinc) IN=net ) ;
  BY hhnr hhnrakt; /* original sorting is used here */
  IF net ; /* only households with net information */
```

Step 3: Linking gross and net information at individual level:

```
/* File soep07.sas continued */
DATA bperson;
  MERGE
    soep.bpbrutto
      (KEEP=hhnr hhnrakt persnr bgebur bsex
       RENAME=(bgebur=birth bsex=sex) )
    soep.bp
      (KEEP=hhnr hhnrakt persnr bp9301 bperfas
       RENAME=(bp9301=satisf bperfas=nat)
       IN=net);
  BY hhnr hhnrakt persnr;
  IF net;
```

Note: Merging by “hhnrakt persnr” keeps the individuals file in an adequate ordering to merge it with a household file.

Step 4: Linking household and individual information:

```
/* File soep07.sas continued */
DATA b;
  MERGE bperson bhh;
  BY hhnr hhnrakt;
  percapt=hhinc/hhsz;
  age=1985-birth;
  dsex=0; IF ( sex EQ 2 ) THEN dsex=1;
  dnat=0; IF ( nat EQ 1 ) THEN dnat=1;
```

Step 5: Data analysis

```
/* File soep07.sas continued */
PROC REG DATA=b;
  MODEL satisf=dsex dnat age percapt ;
  RUN;
```

Example: Comparison of the life satisfaction and percapita income for heads of the household and spouses.

```

/* individual satisfaction with household income (wave B) */
/* file: soep05.sas */

data bh;

merge soep.bhbrutto (keep=hhnr hhnrakt bhgr rename=(bhgr=hhsz))
      soep.bh (keep=hhnr hhnrakt bh39 rename=(bh39=hhinc) in=netto);
by hhnr hhnrakt; /* original sorting is used here */
if netto; /* only household with net information */

/* merge individual information from questionnaire (bp)
and additional individual information (bpbrutto) */

data bperson;

merge soep.bpbrutto (keep=hhnr hhnrakt persnr bstell
                    rename=(bstell=relrp))
      soep.bp (keep=hhnr hhnrakt persnr bp0102
              rename=(bp0102=sat) in=netto);
by hhnr hhnrakt persnr; /* original sorting is used here */
if netto; /* only persons with net information */

/* merge household and individual information */

data b;

merge bh bperson;
by hhnr hhnrakt; /* note unequal sample sizes */
percapit=hhinc/hhsz;
if relrp <2; /* select reference person and spouses only*/

/* create a table with life satisfaction and percapita income
for heads and spouses */

proc tabulate ;
var sat percapit;
class relrp;
table relrp, (sat percapit)*(mean n);
run;

```

5 Matching individual information across time

Example: Comparison of prospective and retrospective life satisfaction.

```

/* life satisfaction across time: Use of index files. */
/* file: soep06.sas */

/* comparison of satisfaction */
data compsat;

/*      01 : satisfaction with your life today
...     02 : satisfaction with your life last year
...     03 : satisfaction with your life next year */

merge

soep.ap (keep=persnr ap6801 ap6802 ap6803
         rename=(ap6801=sact84 ap6802=sly84 ap6803=sny84)in=y1)
soep.bp (keep=persnr bp9301 bp9302 bp9303
         rename=(bp9301=sact85 bp9302=sly85 bp9303=sny85)in=y2)
soep.cp (keep=persnr cp9601 cp9602 cp9603
         rename=(cp9601=sact86 cp9602=sly86 cp9603=sny86)in=y3);
by persnr; /* index files (sorting by persnr ) are necessary here */

in1=y1; /* store inclusion indicators y1 y2 y3 */
in2=y2;
in3=y3;

/* look what you got is always a good strategy */
proc contents; run;

/* prepare data for comparison of:
1) yearly actual satisfaction
2) actual satisfaction with last year's prediction for this year
(prospective)

```

```

3)      actual satisfaction with next year's retrospective judgement
4)      prospective and retrospective judgement
*/

data compsat; set compsat;

      d8584=Sact85-sact84;
      d8685=Sact86-sact85;

      d85pr = sny84-sact85;   * prospective;
      d85re = sly86-sact85;   * retrospective;

      dprre      = sny84-Sly86;
      if in1 and in2 and in3; * select only complete cases;

      proc tabulate;

var sact84 sact85 sact86 d8584 d8685
    sny84 sly86 d85pr d85re dprre;

table (sact84 sact85 sact86 d8584 d8685)
      *(mean*f=5.2 std*f=5.2 n*f=5.0);

table (sny84 sly86 d85pr d85re dprre)
      *(mean*f=5.2 std*f=5.2 n*f=5.0); run;

title1 'Prospective discrepancy';
title2
      "In 84 predicted satisfaction for 85 - actual satisfaction 85";

proc chart ;

      vbar d85pr /
      midpoints=-10 to 10 by 1
      type=FREQ ;

run;

title1 ' Retrospective discrepancy';
title2
      "In 86 remembered satisfaction for 85 - actual satisfaction 85";

proc gchart ;

      vbar d85re /
      midpoints=-10 to 10 by 1
      type=FREQ ;

run; quit;

```

6 Aggregation of individual information on household level

Example: Computation of equivalence income and per-capita income in wave 1.

Equivalence scale

Household Member	Scale Value
Head of the household	1.0
Spouse	0.8
All other adults older than 15	0.6
Child up to 15 years	0.5

```

/* Aggregation of individual information on household level      */
/* Example: Computation of equivalence income and per-capita    */

/* ***** Equivalence scale ***** */
/* Household Member          Scale Value      */
/* Head of the household      1.0             */
/* Spouse                     0.8             */
/* All other adults older than 15  0.6         */
/* Child up to 15 years       0.5             */
/* ***** */

DATA hhweight(KEEP= hhnrakt wsum);
  SET soep.apbrutto(KEEP=hhnrakt astell ageburt
    RENAME=(astell=relation ageburt=birth));
  BY hhnrakt; /* index files (sorting by hhnrakt) is used here */
  IF FIRST.hhnrakt THEN wsum=0;
  /* On the FIRST entry of each household wsum is reset to 0.0 */
  IF (relation eq . or relation eq 0) THEN wsum+1;
  IF (relation eq 1 or relation eq 2) THEN wsum+0.8;
  IF (relation ge 3 ) THEN DO;
    IF (1984-birth ge 16) THEN wsum+0.6;
    IF (1984-birth lt 16) THEN wsum+0.5;
  END;
  IF LAST.hhnrakt THEN OUTPUT;
  /* On the LAST entry of each household wsum is written to file hhweight */
proc print data=hhweight(obs=40); run;

DATA hh;
  MERGE hhweight
    soep.ah (KEEP=hhnrakt ahgr ah46
    RENAME=(ahgr=size ah46=hhinc));
  BY hhnrakt;
  percapit=. ; IF (hhinc GT 1) THEN percapit=hhinc/size;
  equiv=. ;    IF (hhinc GT 1) THEN equiv=hhinc/wsum;

PROC MEANS DATA=hh N NMIS MEAN MIN MAX MAXDEC=2;
  CLASS size;
  VAR percapit equiv;
  RUN;

```

7 Use of the Macro Language

Doing the same for different waves!

Input variables in macro `example3`

Description	Macro Variable
letter of wave	<code>lwave</code>
name of household income variable	<code>hhinc</code>
name of satisfaction variable	<code>satisf</code>
name of the nationality variable	<code>nat</code>
value of survey year	<code>year</code>

```

/* use of macro language */
/* file: soep08.sas */
/* ***** */
/* Meaning of macro variables */
/* ***** */
letter of wave :lwave
name of household income variable :hhinc
name of satisfaction variable :satisf
name of the nationality variable :nat
value of survey year :year
*****
*/

%MACRO example3(lwave, hhinc, satisf, nat, year);
DATA &lwave.hh;
MERGE
  soep.&lwave.hbrutto (KEEP=hhnr hhnrakt &lwave.hhgr
                     RENAME= (&lwave.hhgr=hhsizes) )
  soep.&lwave.h (KEEP=hhnr hhnrakt &hhinc RENAME= (&hhinc=hhinc) IN=net ) ;
BY hhnr hhnrakt;
IF net ;
DATA &lwave.person;
MERGE
  soep.&lwave.pbrutto
  (KEEP=hhnr hhnrakt persnr &lwave.geburt &lwave.sex
   RENAME= (&lwave.geburt=birth &lwave.sex=sex) )
  soep.&lwave.p
  (KEEP=hhnr hhnrakt persnr &satisf &nat
   RENAME= (&satisf=satisf &nat=nat)
   IN=net);
BY hhnr hhnrakt persnr;
IF net;
DATA &lwave;
MERGE &lwave.person &lwave.hh;
BY hhnr hhnrakt;
percapit=hhinc/hhsizes;
age=&year-birth;
dsex=0; IF ( sex EQ 2 ) THEN dsex=1;
dnat=0; IF ( nat EQ 1 ) THEN dnat=1;
%mend example3; /* end of macro program */

```

Use of macro for waves 2–6

```

/* invocation of macro example3 for waves b,c,d,e and f */
%example3(b, bh39, bp9301, bperfas, 1985);
%example3(c, ch51, cp9601, cperfas, 1986);
%example3(d, dh51, dp9801, dperfas, 1987);
%example3(e, eh42, ep89, eperfas, 1988);
%example3(f, fh42, fp108, fperfas, 1989);
/* Temporary output files are named: b,c,d,e and f */
/* They are used in program file soep09.sas */
proc contents data=b ;run;
proc contents data=c ;run;
proc contents data=d ;run;
proc contents data=e ;run;
proc contents data=f ;run;

```

Note: Macro `example3` cannot be used for wave A since there is no file *AH-BRUTTO*!

8 Preparing files for longitudinal analysis

Observations of a person (unit of analysis) in one row. There are two versions of this format:

Balanced design: Only units with complete observations at all points of time are included.

Unbalanced design: All units are included. This causes missing values for some points of time.

Task: All time dependent variables have to be distinguished by a time index! (Here variables DSEX DNAT are assumed to be constant over time.)

Input variables in `indtime`

Description	Macro Variable
dataset where variables shall be indexed value of time index	dataset time

```

/* Generation of a file in longitudinal format II          */
/* file soep12.sas                                       */
/* Temporary files b,c,d,e and f are needed             */
/* ***** */

/* Macro indtime for time indexing varvariables          */
/* dataset where variables shall be indexed : dataset    */
/* value of time index      : time                      */
/* ***** */
%MACRO indtime(dataset,time);
DATA dat&TIME;
  SET &dataset (RENAME=(satisf=satis&time
                        percapit=percap&time
                        age=age&time) );
PROC SORT DATA= dat&time;
  BY persnr;
/* there is probably no presort file. Sorting by persnr will needed */
%MEND indtime;

%indtime(b, 85);
%indtime(c, 86);
%indtime(d, 87);
%indtime(e, 88);
%indtime(f, 89);
run;

```

Generation of a file for unbalanced analysis:

```

DATA unbal;
  MERGE dat85 dat86 dat87 dat88 dat89;
  BY persnr;
proc print data=unbal(obs=5); run;

```

Generation of a file for balanced analysis:

```

DATA balanced;
  MERGE dat85(IN=in85) dat86(IN=in86) dat87(IN=in87) dat88(IN=in88) dat89(IN=in89);
  BY persnr;
  IF (in85 AND in86 AND in87 AND in88 AND in89);
proc print data=balanced(obs=5); run;

```

9 Pooled regression analysis

Generation of a file for “pooled” analysis: Observations of the same persons at different points of time are regarded as independent observations.

```
DATA pool;
  SET b c(IN=c) d(IN=d) e(IN=e) f(IN=f);
  d86=0; IF (c) THEN d86=1 ;
  d87=0; IF (d) THEN d87=1 ;
  d88=0; IF (e) THEN d88=1 ;
  d89=0; IF (f) THEN d89=1 ;
PROC REG DATA=pool;
  MODEL satisf=d86 d87 d88 d89 dsex dnat age percapit;
RUN;
```

10 Random effects models

Observations of a person (unit of analysis) in different rows. Identification of units by variable PERSNR. Identification of time by variable TIME. Observations should be ordered by PERSNR and TIME to facilitate analysis and data checks.

```
/* ***** */
/* Generation of a file for longitudinal analysis, Format I: */
/*
/* Observations of a person (unit of analysis) in different rows. */
/* Identification of units by variable PERSNR. Identification of */
/* time by variable TIME. */
/* Observations should be ordered by PERSNR and TIME} to */
/* facilitate analysis and data checks. */
/* ***** */

/* file: soep10.sas */
/* Temporary files b,c,d,e,f are needed */

DATA long;
  SET b(IN=b) c(IN=c) d(IN=d) e(IN=e) f(IN=f);
  IF (b) THEN time=85 ;
  IF (c) THEN time=86 ;
  IF (d) THEN time=87 ;
  IF (e) THEN time=88 ;
  IF (f) THEN time=89 ;
  d86=0; IF (c) THEN d86=1 ; /* dummy variables for calender years */
  d87=0; IF (d) THEN d87=1 ;
  d88=0; IF (e) THEN d88=1 ;
  d89=0; IF (f) THEN d89=1 ;
PROC SORT DATA=long;
  BY persnr time;
RUN;

/* Estimation of a Random effects model */
/* See: SAS/Stat: Changes and Enhancements p. 591 */
/* Note that PROC MIXED is resoucre consuming */
/* therefore the number of observations is restricted here */

PROC MIXED DATA=long(obs=3000) METHOD=ML;
  CLASS persnr ;
  MODEL satisf=d86 d87 d88 d89 dsex dnat age percapit /S ;
  RANDOM persnr;
RUN;

/* Compare the result with the pooled regression estimate for
the same observations */
PROC REG DATA=long(OBS=3000);
  MODEL satisf=d86 d87 d88 d89 dsex dnat age percapit;
RUN;
```

Use of procedure TSCSREG in the ETS module:

Observations are assumed to be sorted by person (unit of analysis) and time. Note that the number of observations per unit is to be constant in the TSCSREG procedure. The procedure treats units and time as random effects.

```
/* ***** */
/* file: soep10a.sas */
/* Temporary file long from previous example is needed */

/* Data management: Skip all observations from persons with less than
5 valid observations */

PROC MEANS NOPRINT DATA=long( WHERE=(percapit GT 0 AND satisf ge 0) );
  BY persnr ;
  VAR persnr ;
  OUTPUT OUT=pmeans N=nvalid ;
/* dataset pmeans contains for each person the number of valid observations
on variable nvalid */
run;

Proc print data=pmeans(OBS=20); run;

DATA long2;
  MERGE long(IN=inlong) pmeans(WHERE=(nvalid eq 5) IN =valid);
  BY persnr ; /* long and pmeans (unequal length) are sorted by persnr */
/* if valid and inlong; /* only valid cases */
run;

PROC TSCSREG DATA=long2;
  ID persnr time;
  /* first ID variable: identifier for units */
  /* second ID variable: identifier for replications */
  MODEL satisf=age dnat dsex percapit ;
run;

PROC REG DATA=long2 ;
  MODEL satisf=age dnat dsex percapit ;
  RUN;
```

11 Fixed effects models

Example: Estimation of individual intercepts. Note that all time constant covariates have to be removed from the model.

```
/* Estimation of a fixed effects model */
/* Note: constant covariates dsex and dnat have to be excluded !! */
/* file: soep11.sas */
/* Temporary file long is needed */

/* The simple approach will in general not work!! */
/* Reason: Persons with only 1 valid observation */

/* ***** Will probably not work !!! ***** */
PROC MIXED DATA=long(obs=3000) METHOD=ML;
  CLASS persnr ;
  MODEL satisf= persnr d86 d87 d88 d89 age percapit /S ;
  RUN;
  ***** */

/* Data management: Skip all observations from persons with only one
valid observation */

PROC MEANS NOPRINT DATA=long(obs=1000 WHERE=(percapit GT 0) );
  BY persnr ;
  VAR persnr ;
```

```

OUTPUT OUT=pmeans N=nvalid ;
/* dataset pmeans contains for each person the number of valid observations
on variable nvalid */
run;

Proc print data=pmeans(OBS=20); run;

DATA long2;
MERGE long(obs=1000 in=inlong) pmeans(WHERE=(nvalid gt 1) IN =valid);
BY persnr ; /* long and pmeans (unequal length) are sorted by persnr */
if valid and inlong; /* only valid cases */
/* age is badly scaled: it is almost a constant */
age2=age-40;
run;

proc print data=long2(OBS=20); run;

PROC REG DATA=long2 ;
MODEL satisf=d86 d87 d88 d89 age2 percaptit ;
RUN;

PROC MIXED DATA=long2 order=data ;
CLASS persnr ;
MODEL satisf= d86 d87 d88 d89 age2 percaptit persnr/S noint ;
RUN;

```

Example: Subtraction of individual means. Here, only the coefficients of the time variable variables are estimate. Note the effect of badly scaled variables like age which is quasi constant for elderly persons.

```

/* file: soep11a.sas */
/*****/
/* Note: identifier persnr will be needed from the
data set
*****/
/*
/* specification of the model:
/* _y = metric dependent variable
/* _x = time-dependent variables
/* _z = time-constant variables
/* _nx = number of _x variables
/* _nz = number of _z variables
/*****/
/* specify model here: ..... */

%let _y = satisf ;
%let _x = age percaptit d86 d87 d88 d89 ;
%let _z = dsex dnat ;

%let _nx = 6 ;
%let _nz = 2 ;

%macro fixed(indat);

%let _within = wx1-wx&_nx ; /* parameter within estimate */

/* step 1 : Pooled regression */
PROC REG DATA=&indat ;
MODEL &_y= &_x &_z ;
TITLE ' ***** Pooled Regression *****';
RUN;

/* step 2 : within means */

/* Computation of means over time per persnr */
PROC MEANS DATA=&indat NOPRINT ;
BY persnr;
ID persnr; /* displayes variable persnr with means */
VAR &_y &_x ;
OUTPUT OUT=quer MEAN= my &_within ;

```

```

/* individual means of covariates assigned to &_within */
RUN;

/* step 3: subtraction of means from the original data */
DATA diff;
MERGE &indat quer ;
BY persnr ;
yd= &_y-my; /* within--transformation of y */
ARRAY awx &_x;
ARRAY abx &_within;
ARRAY ax &_x;
DO OVER awx;
    awx = ax - abx; /* within--transformation of covariates */
END;
KEEP yd &_x ;

/* step 4 : within estimate */
PROC REG DATA=diff ;
MODEL yd=&_x /NOINT ;
TITLE '*** Within estimate *** ' ;
RUN;
%mend;

%fixed(long2);

```

12 A useful SAS macro for the estimation of fixed and random effects models

The Random Effects Model:

$$y_{i,t} = X'_{i,t}\beta + \alpha_i + u_{i,t}$$

Where α_i and $u_{i,t}$ are independent disturbances with variances σ_α^2 and σ_u^2 . Let T_i be the number of observations from unit i.

It can be shown (see for example Amemiya "Advanced Econometrics" 1985, p. 181) that by a simple transformation of the data one can achieve a situation where serial correlation is eliminated.

This transformation is given by:

$$Y_{i,t}^* = Y_{i,t} - (1 - \psi_i^{0.5})\bar{Y}_{i,\bullet}$$

where:

$$\psi_i = \frac{\sigma_u^2}{\sigma_u^2 + T_i\sigma_\alpha^2}$$

$\bar{Y}_{i,\bullet}$ = individual mean of unit i

For the covariates the same transformation applies:

$$X_{i,t}^* = X_{i,t} - (1 - \psi_i^{0.5})\bar{X}_{i,\bullet}$$

Therefore estimation of the model parameters is efficient and the standard errors are correct, if we apply the OLS (i.e. the standard regression procedure) with the transformed data.

In order to do this one has to estimate the variances σ_α^2 and σ_u^2 in advance.

Estimation of σ_u^2

Using the "within"-transformation we get:

$$Y_{i,t} - \bar{Y}_{i,\bullet} = (X_{i,t} - \bar{X}_{i,\bullet})'\beta + u_{i,t} - \bar{u}_{i,\bullet}$$

Because of

$$E((u_{i,t} - \bar{u}_{i,\bullet})^2) = \sigma_u^2(T_i - 1)/T_i$$

the residual squares $w_{i,t}$ of the within estimate, i.e. OLS applied to the above transformed data, serve for an estimate of σ_u^2 :

$$\hat{\sigma}_u^2 = \frac{1}{R - N - p} \sum_{i=1}^N \sum_t w_{i,t}^2$$

where $R = \sum_i T_i$ and p is the number of parameters of the within estimate.

Estimation of σ_α^2

Averaging over time yields the "between"-transformation:

$$\bar{Y}_{i,\bullet} = \bar{X}'_{i,\bullet}\beta + \alpha_i + \bar{u}_{i,\bullet}$$

It holds that:

$$E((\alpha_i + \bar{u}_{i,\bullet})^2) = \sigma_\alpha^2 + \frac{T_i}{T_i^2} \sigma_u^2$$

Summation over units gives:

$$\sum_{i=1}^N E((\alpha_i + \bar{u}_{i,\bullet})^2) = N\sigma_\alpha^2 + \sum_{i=1}^N \frac{1}{T_i} \sigma_u^2$$

Therefore the sum of residual squares of the between estimate may be used to estimate the left side of the above equation. Inserting $\hat{\sigma}_u^2$ on the right side, we can derive an estimate for σ_α^2 .

A useful SAS macro

The following SAS macro prepares its user with the estimation of the random effects model. The steps follow the above lines.

Important note: The macro assumes a data file in longitudinal format I, i.e. observations of a unit in different rows. Therefore it is necessary to identify observations by variables "unit" and "time".

```
/* *****  
/* Note: identifiers UNIT and TIME will be needed from the  
/* data set  
/* *****  
/*  
/* specification of the model:
```

```

/*      _y = metric dependent variable
/*      _x = time-dependent variables
/*      _z = time-constant variables
/*      _nx = number of _x variables
/*      _nz = number of _z variables
/*****
/* specify model here: ..... */

%let _y      = <name of dependent variable> ;
%let _x      = <names of time-dependent variables> ;
%let _z      = <names of time-constant variables> ;

%let _nx     = <number of x-variables> ;
%let _nz     = <number of z-variables> ;

%macro ec(indat);

/*
*****
program: ec sas a
purpose: estimation of error component model
         for unbalanced panel data
input:   sas file with identifiers unit and time
output:  (1) pooled regression
         (2) within estimate (fixed effects model)
         (3) error component estimate (random effects model)
authors: ulrich rendtel , johannes schwarze (DIW, Berlin)
*****
*/
DATA work1;
  SET &indat;
%let _betwen = bx1-bx&_nx bz1-bz&_nz; /* parameters between estimate */
%let _within = wx1-wx&_nx ;         /* parameter within estimate */
PROC SORT; BY unit; RUN;

PROC CONTENTS; /* Documentation of variable names and cases numbers */
  TITLE '***** dataset and variables *****';
  RUN;
PROC MEANS N NMISS MEAN MIN MAX STD MAXDEX=2;
  TITLE '***** descriptive statistics *****';
  RUN;

/* step 1 : Pooled regression */
PROC REG ;
  MODEL &_y= &_x &_z ;
  TITLE '***** Pooled Regression *****';
  RUN;

/* step 2 : within estimate and estimation of within variance */

/* Computation of means over time per unit */
PROC MEANS NOPRINT ;
  BY unit;
  ID unit; /* displayes variable unit with means */
  VAR &_y &_x &_z;
  OUTPUT OUT=quer MEAN= my &_betwen N=ti ;
  /* individual means of covariates assigned to &_betwen */
  RUN;

/* store sumti = sum of ti (observation of unit i ) */
/* store sumedti = sum of 1/ti */
/* store n = number of units */
DATA work2 ;
  SET quer ;
  edti=1.0/ti ;
PROC MEANS NOPRINT ;
  VAR ti edti ;
  OUTPUT OUT=merk SUM= sumti sumedti N= n ;
  RUN;
DATA merk1 ;
  SET merk ;
  KEEP sumti sumedti n ;

/* next step: subtraction of means from the original data */
DATA diff;
  MERGE work1 quer ;
  BY unit ;

```

```

yd= &y-my; /* within--transformation of y */
ARRAY awx &x; /* This is a trick to force SAS to use the original
              variable labels for the transformed variables */
ARRAY abx &_betwen;
ARRAY ax &x;
DO OVER awx;
    awx = ax - abx; /* within--transformation of covariates */
END;
KEEP yd &x      ;

/* step 3 : within estimate */
PROC REG DATA=diff OUTEST=within ;
MODEL yd=&x /NOINT ;
TITLE '*** Within estimate *** t values not corrected *****';
RUN;

/* estimation of sigu */
DATA merk2 ;
MERGE merk1 within ;
sigw=_rmse_ ;
p=&_nx;
korr= (sumti-p)/(sumti-n-p);
sigu=sigw*sqrt(korr);
tkorr=1.0/sqrt(korr);
KEEP sigw sigu tkorr sumti sumedti n ;
PROC PRINT;
TITLE '***** tkorr corrects the t values *****';
RUN;

/* computation of between estimate */
PROC REG DATA=quer OUTEST=between;
MODEL my=&_betwen ;
TITLE '***** Between estimate *****';
RUN;

DATA merk3 ;
MERGE merk2 between;
sigb=_rmse_ ;

/* Computation of siga (std. deviation of time-constant
              variance component) */
siga2=sigb*sigb-sigu*sigu*sumedti/n ;
siga=sqrt(max(siga2,0));
KEEP sigu siga ;
PROC PRINT;
TITLE '***** estimates for sigu and siga *****';
RUN;

/* computation of vector gam=1-SQRT( sigu2/(sigu2+ti*siga2) ) */
DATA quer2 ;
IF _n_ =1 THEN SET merk3 ; /* merk3 contains 1 observation */
SET quer ; /* quer contains n observations */
gam=1-sqrt(sigu*sigu/(sigu*sigu+ti*siga*siga));
KEEP unit gam my &_betwen;

/* step 4 : EC estimate */
/* gam-transformation : yt=y-gam*my (other covariates the same way) */

DATA trans ;
MERGE work1 quer2 ;
BY unit ;
ty=&y-gam*my ;
tconst=1.0-gam ;
ARRAY aex &x &z;
ARRAY abx &_betwen;
ARRAY trans &x &z ;
DO OVER aex;
    trans = aex - gam*abx;
END;
KEEP unit time ty tconst &x &z      ;

PROC REG;
MODEL ty= tconst &x &z / NOINT;
TITLE '***** EC-Estimate *****';
run;
%MEND;

```

13 Variance estimation for population totals

13.1 Bounds assuming independent selection of initial households

A population estimator of the form $\hat{P} = \sum_{i \in G} \frac{C_i}{\pi_i} Y_i$ has the following variance (cf. Särndal et al., 1992, p.43):

$$\begin{aligned} V(\hat{P}) &= \sum_{i,j \in G} \left(\frac{E_{\pi}(C_i C_j)}{\pi_i \pi_j} - 1 \right) Y_i Y_j \\ &= \sum_{i \in G} \left(\frac{1}{\pi_i} - 1 \right) Y_i^2 + \sum_{i \in G} \sum_{j \neq i} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) Y_i Y_j \end{aligned} \quad (1)$$

where π_{ij} is the probability of the i -th and j -th units being jointly selected. If $\pi_{ij} > 0$ applies to all $i, j \in G$, then

$$\hat{V}(\hat{P}) = \sum_{i \in S} \left(\frac{1}{\pi_i} - 1 \right) \frac{Y_i^2}{\pi_i} + \sum_{i \in S} \sum_{j \neq i} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) \frac{Y_i Y_j}{\pi_{ij}} \quad (2)$$

is an unbiased estimator for $V(\hat{P})$.

With (1), the following is obtained in wave 1 for the estimation of household-related attributes:

$$V(\hat{P}) = \sum_{h \in G} \left(\frac{1}{\pi_h} - 1 \right) Y_h^2 \quad (3)$$

$V(\hat{P})$ can consequently be estimated without bias using:

$$\hat{V}(\hat{P}) = \sum_{h \in S} \left(\frac{1}{\pi_h} - 1 \right) \frac{Y_h^2}{\pi_h} \quad (4)$$

For the estimation of person-related cross-sectional attributes, let H_i be the number of all persons living together with person i in a household. Considering $\pi_{ij} = \pi_i$ for $j \in H_i$, we obtain the following from (1):

$$\begin{aligned} V(\hat{P}) &= \sum_{i \in G} \left(\frac{1}{\pi_i} - 1 \right) Y_i^2 + \sum_{i \in G} \sum_{j \in H_i} \left(\frac{1}{\pi_i} - 1 \right) Y_i Y_j \\ &= \sum_{i \in G} \left(\frac{1}{\pi_i} - 1 \right) Y_i \left(Y_i + \sum_{j \in H_i} Y_j \right) \end{aligned} \quad (5)$$

An unbiased estimate of $V(\hat{P})$ is:

$$\begin{aligned} \hat{V}(\hat{P}) &= \sum_{i \in S} \left(\frac{1}{\pi_i} - 1 \right) \frac{Y_i^2}{\pi_i} + \sum_{i \in S} \sum_{j \in H_j} \left(\frac{1}{\pi_i} - 1 \right) \frac{Y_i Y_j}{\pi_{ij}} \\ &= \sum_{i \in S} \left(\frac{1}{\pi_i} - 1 \right) \frac{Y_i}{\pi_i} \left(Y_i + \sum_{j \in H_j} Y_j \right) \end{aligned} \quad (6)$$

Here, the last equals sign again uses $\pi_{ij} = \pi_i$.

The specific effect of the panel design on $V(\hat{P})$ consists in the fact that in later waves:

- The selection of particular households is dependent.
- The selection of persons living in different households is dependent.

Lower and upper bounds can be derived:

$$\begin{aligned}
\hat{V}_L(\hat{P}) &= \sum_{i \in S_t} \left(\frac{1}{\pi_{i,t}} - 1 \right) \frac{Y_i^2}{\pi_{i,t}} + \sum_{i \in S_t} \sum_{j \in H_{i,t} \cap S_t} \left(\frac{1}{\pi_{j,t}} - 1 \right) \frac{Y_i Y_j}{\pi_{i,t}} \\
&\leq \hat{V}(\hat{P}) \\
&\leq \sum_{i \in S_t} \left(\frac{1}{\pi_{i,t}} - 1 \right) \frac{Y_i^2}{\pi_{i,t}} + \sum_{i \in S_t} \sum_{j \in R_{i,t} \cap S_t} \left(\frac{1}{\pi_{j,t}} - 1 \right) \frac{Y_i Y_j}{\pi_{i,t}} \\
&= \hat{V}_U(\hat{P})
\end{aligned} \tag{7}$$

where:

$H_{i,t} \cap S_t$ = set of persons in the sample living together with person i .

$R_{i,t} \cap S_t$ = set of persons in the sample attributed to the root-household of person i .

13.2 A SAS-macro for the computation of the bounds

```

/* *****
input: indata = sas file with variables hrf y and id
      hrf     = weight
      Y       = characteristic
      id      = identifier for dependent units
output: outdata=sas file (one observation!) with output variables
      phat    = estimated population total
      vhat    = estimated variance bound of that estimate
*****
*/;
%MACRO varb(indata,hrf,y,id,outdata);
PROC SORT DATA=&indata;
BY &id;
RUN;
DATA wprod;
SET &indata;
hrfy=&y*(&hrf-1);

/* compute sum of (hrf-1)*y within dependent unit */
PROC UNIVARIATE DATA=wprod NOPRINT;
BY &id;
VAR hrfy;
OUTPUT OUT=work SUM=hsum;
RUN;

/* combine sum of (hrf-1)*y with every unit and compute
hrf*y X sum of (hrf-1)*y */
DATA work2 ;
MERGE &indata work ;
BY &id;
p=&hrf*&y;
prod=&hrf*&y*hsum;

/* compute the sum over the above expressions */
PROC UNIVARIATE DATA=work2 NOPRINT;
VAR p prod;
OUTPUT OUT=&outdata SUM=phat vhat;
%MEND;

/* example with party preference for social democrats (SPD) in
wave 5 */
DATA spd;
MERGE ep(KEEP=hnr hnrakt persnr ep7702
RENAME=(ep7702=pref) )
phrf(KEEP=ephrf persnr
RENAME=(ephrf=hrf) );
BY persnr;
y=0; IF (pref eq 1) y=1;

/* compute lower and upper bounds */
%varb(spd,hrf,y,hnrakt,lower);
%varb(spd,hrf,y,hnr ,upper);

/* compute coefficient of variation cv and std deviation sig */
DATA work;
SET lower upper;
cv=sqrt(vhat/(phat*phat));
sig=sqrt(vhat);

/* results */
PROC PRINT DATA=work;

```

13.3 Variance estimates by random groups

The methodology The idea of the method is based on producing a situation in which there are R independent and identical replications of the sampling experiment. If \hat{P}_r is an estimation of the population parameter on the basis of the r -th sampling experiment ($r = 1, \dots, R$), then:

$$\hat{\hat{P}} = \frac{1}{R} \sum_{r=1}^R \hat{P}_r \quad (8)$$

is also an estimate for P and

$$\hat{\hat{V}} = \frac{1}{R(R-1)} \sum_{r=1}^R (\hat{P}_r - \hat{\hat{P}})^2 \quad (9)$$

is, under the above assumption, an unbiased estimator for $V(\hat{\hat{P}})$: cf. Wolter (1985, p. 21). If \hat{P} is a linear estimator, then $\hat{\hat{P}} = \hat{P}$ applies. $\hat{\hat{V}}$ is consequently also a meaningful estimator for $V(\hat{P})$.

The basic idea of the procedure proposed here consists in finding a subdivision of the original sample into R sub-samples (‘random groups’), so that each of the sub-samples can be regarded as a realization of the original sampling experiment with a reduced sample size: cf. Wolter (1985, p. 30 ff).

A SAS macro for the random group estimator

```

/* Variance estimation for cross--sectional totals and          */
/* proportions by random groups: Party preference for          */
/* social democrats (SPD)                                     */
/* *****                                                    */
/* file: soep14.sas                                          */

/* *****                                                    */
input: indata = sas file with variables hrf y and rg
      hrf      = weight
      Y        = characteristic
      rg       = random group
output: outdata=sas file (one observation!) with
      output variables phat and vhat
      phat     = estimated population total
      vhat     = estimated variance of that estimate
      *****
*/;

%MACRO rg(indata,hrf,y,rg,outdata);
PROC SORT DATA=&indata;
BY &rg; /* sort data by random groups */

DATA &indata;
SET &indata;
p=8*&hrf*&y;

PROC UNIVARIATE DATA=&indata NOPRINT;
BY &rg;
VAR p;
OUTPUT OUT=work SUM =p_r;
RUN;
/* data file work has 8 observations:
   one for each random group */

PROC UNIVARIATE DATA=work NOPRINT;
VAR p_r;
OUTPUT OUT=work2 SUM=phat VAR=vhat;
/* data file has one observation */

DATA &outdata;
SET work2;
phat=phat/8.0;
vhat=vhat/8.0;
/* compute coefficient of variation cv and std deviation sig */
cv=sqrt(vhat/(phat*phat));
sig=sqrt(vhat);

```

```

KEEP phat vhat cv sig;
PROC PRINT DATA=&outdata;
%MEND;

/* example with party preference for social democrats (SPD) in
   wave 5 */
DATA spd;
MERGE soep.ep(KEEP=hnr hnrakt persnr ep7702
              RENAME=(ep7702=pref)
              IN=inep
              )
      soep.phrf(KEEP=hphrf prgroup persnr
                RENAME=(hphrf=hrf) );
BY persnr;
y=0; IF (pref eq 1) THEN y=1;
IF ineq;
/* compute random group estimate */
%rg(spd,hrf,y,prgroup,result);

```

14 Duration analysis

Prepare a file that contains all unemployment spells that are not left censored. Compute the duration of spells and estimate the survivor function by procedure LIFETEST. Compare the risk of staying in unemployment for men and woman.

```
/* ***** */
/* file: soep15.sas */
/* ***** */

/* Create a file that contains all unemployment spells which are
not left censored */

DATA unempl;
  MERGE soep.artkalen(KEEP= hhnr persnr begin end zensor spelltyp
                    WHERE=( (spelltyp EQ 5) AND (zensor le 3) )
                    IN=inspell )
        soep.ppfad(KEEP= hhnr persnr sex) ;
  BY hhnr persnr ; /* note the original sorting of files artkalen and ppfad */
  IF inspell;
  length=end-begin;
  run;
Proc print DATA=unempl(obs=10);run;

PROC LIFETEST DATA=unempl plots=(s) graphics ;
  TIME length*zensor(2,3); /* Right censoring is indicated by values 2,3 */
  STRATA sex;
  SYMBOL1 c=blue h=1 v=square;
  SYMBOL2 c=red h=1 v=circle;
  note 'comparison of unemployment length';
run;
```

15 Exercises

1. Use the **GLOBALS** window to get an overview of the SOEP files by opening the **LIB** and the **VAR** window. What is name of the variable with the label ‘‘Schulabschluss’’ in file *APGEN*?
2. Use the **ACCESS** window to view the content of file *APGEN*; (Option **BL=Browse+List**). Notice the sorting of the data file. Use the **Search** window to examine subpopulations.
3. Use the procedure calls:

```
Proc DATASETS LIBRARY=soep; run;  
Proc CONTENTS DATA=soep.apgen; run;
```

Notice the information about index files.

4. Prepare a file with the following variables from wave 1: Monthly net income of the head of the household, Monthly rent to be paid, satisfaction with household income. Analyse the joint correlation matrix of the 3 variables. (Ignore the ordinal level of satisfaction scores).
5. Write a macro that prepares the same variables for a specified wave. Use a time index for the names of the variables.
6. Prepare a longitudinal file (one record per person), which is balanced (waves 1 and 5) with help of the macro from the previous exercise. Compute the relative increase of incomes and rents. Establish a scatter plot of these increases.
7. Compute for each child (up to 16 years) the number of siblings (up to 16 years) living in the household. Identify siblings by their relation to their mother.
 - a) Write a program for an individual cross-section.
 - b) Re-write this program as a macro. Produce time-indexed variables.
 - c) Give a frequency table of the observed longitudinal patterns (i.e. histories concerning siblings) along waves 1,3 and 5.
8. Apply macro **EC** (on file **ecmacro.sas**) for the estimation of a random effects model.

Dependent variable: log of the gross monthly income
Covariates:

age and sex (in *PPFAD*)

institutional years necessary to receive current degree of education (in *PGEN*)

Years with the current employer (in *PGEN*)

Use the first five soep waves (A to E). Use the **include** statement to avoid a lengthy program.

9. Estimate the net change between 1984 (wave 1) and 1987 (wave 4) in the population total for persons with SPD party preference. Use the random group estimator for the std. error of this estimate.
10. Estimate the population total of persons with permanent party preference for social democrats period: 1984–1988 period: 1986–1988 Calculate also the variances of these estimates Use the random group estimator (macro rg)
11. Prepare a file that contains all persons who experienced exactly two spells of unemployment. The file shall include the following variables: Gender of person, Duration of first recorded unemployment spell, Duration of second recorded unemployment spell, Right Censoring status of the second spell.

16 Solutions for exercises

Exercise 4: Prepare a file with the following variables from wave 1: Monthly net income of the head of the household, Monthly rent to be paid, Satisfaction with household income. Analyse the joint correlation matrix of the 3 variables. (Ignore the ordinal level of satisfaction scores).

```
/* ***** */
/* file: ex01.sas */
/* ***** */

/* Step 1: Selection of heads of the household,
their satisfaction with income and their net income */
DATA aperson;
  MERGE soep.ap(KEEP=hhnr hhnrakt persnr ap0302 ap3302
              RENAME=(ap0302=satisf ap3302=income)
              IN=ina)
        soep.apbrutto(KEEP=hhnr hhnrakt persnr astell
                     WHERE=(astell eq . or astell eq 0)
                     in=head );
  BY hhnr hhnrakt persnr;
  IF ina and head; /* not all persons in the gross sample gave an interview */

/* Step 2: Selection of information on household level */
DATA a;
  MERGE soep.ahgen(KEEP=hhnr hhnrakt amiete
                 RENAME=(amiete=rent) )
        aperson;
  BY hhnr hhnrakt;

/* Step 3: Computation of correlation matrix */
PROC CORR DATA=a;
  VAR income rent satisf;
  RUN;
```

Exercise 5: Write a macro that prepares the same variables for a specified wave. Use a time index for the names of the variables.

```
/* ***** */
/* file: ex02.sas */
/* ***** */

%macro exerc2(lwave,satisf,income,time);
DATA &lwave.person;
  MERGE soep.&lwave.p(KEEP=hhnr hhnrakt persnr &satisf &income
                   RENAME=(&satisf=satisf&time
                           &income=income&time)
                   IN=iners)
        soep.&lwave.pbrutto(KEEP=hhnr hhnrakt persnr &lwave.stell
                          WHERE=(&lwave.stell eq . or &lwave.stell eq 0)
                          IN=head );
  BY hhnr hhnrakt persnr;
  IF iners and head ;
  /* not all persons in the gross sample gave an interview */
DATA &lwave;
  MERGE soep.&lwave.hgen(KEEP=hhnr hhnrakt &lwave.miete
                      RENAME=(&lwave.miete=rent&time) )
        &lwave.person ;
  BY hhnr hhnrakt;

/* Sort file by persnr to ease matching */
```

```

PROC SORT DATA=&lwave;
  BY persnr;
  RUN;
%mend;

%exerc2(a,ap0302,ap3302,84);
%exerc2(e,ep0102,ep4402,88);

```

Exercise 6: Prepare a longitudinal file (one record per person), which is balanced (waves 1 and 5) with help of the above macros. Compute the relative increase of incomes and rents. Establish a scatter plot of these increases.

```

/* ***** */
/* file: ex03.sas */
/* temporary files needed from exercise 5: a and e */
/* ***** */

/* preparing a file for balanced analysis */

DATA balanced;
  MERGE a(in=a)
        e(in=e);
  BY persnr;
  IF ( a AND e);
  IF ( income84 gt 500 and rent84 gt 100 );
/* compute relative changes */
  increl=(income88-income84)/income84;
  rentrel=(rent88-rent84)/rent84;

/* Produce a scatter plot */

PROC GPLOT DATA=balanced(where=(increl le 3 AND rentrel le 3)) ;
  PLOT increl*rentrel/ grid;
  TITLE3 'Relative increases of income vs. rent';
  RUN;
quit;

```

Exercise 7: Compute for each child (up to 16 years) the number of siblings (up to 16 years) living in the household. Identify siblings by their relation to their mother.

- Write a program for an individual cross-section.
- Re-write this program as a macro. Produce time-indexed variables.
- Give a frequency table of the observed longitudinal patterns (i.e. histories concerning siblings) along waves 1,3 and 5.

```

/* ***** */
/* Exercise 7 : Suggested solution */
/*      part a) */
/*      file ex07.sas */
/* ***** */

DATA child;
/* CHildinformation in record akind */
  SET soep.akind(KEEP=hhnrakt persnr akmutti
                RENAME=(akmutti=idmama)

```

```

        WHERE=(idmama gt 0) );
/* idmama = Persnr of mother of the child. In some cases it is missing */
PROC SORT;
BY idmama; /* Sorting is important here */
RUN;
DATA sibling(KEEP=nkid idmama);
SET child;
BY idmama; /* For each mother the number of children in the household
            is computed on variable nkid */
IF FIRST.idmama THEN nkid=0;
nkid+1;
IF LAST.idmama THEN OUTPUT;

/* Merging child and sibling information */
DATA all;
MERGE child sibling;
BY idmama;

PROC FREQ DATA=all;
TABLE nkid ;

/* *****
* Exercise 7 : Suggested solution
*           part b) : macro producing time--indexed variables
*           file ex07.sas
* *****
*/

%MACRO SIBLING(lwave,indtime);
DATA child;
SET soep.&lwave.kind(KEEP=hhnrakt persnr &lwave.kmutti
RENAME=(&lwave.kmutti=idmama)
WHERE=(idmama gt 0) );

PROC SORT;
BY idmama;
RUN;
DATA sibling(KEEP=nkid&indtime idmama);
SET child;
BY idmama;
IF FIRST.idmama THEN nkid&indtime=0;
nkid&indtime+1;
IF LAST.idmama THEN OUTPUT;
DATA all&indtime;
MERGE child sibling;
BY idmama;
PROC SORT DATA=all&indtime;
BY persnr;
%MEND ;

/* *****
* Exercise 7 : Suggested solution
*           part c) : Sibling history along waves 1, 3 and 5
*           file: ex07.sas
* *****
*/

%sibling(a,1);
%sibling(c,2);
%sibling(e,3);

DATA LONGITUD;
MERGE all1 all2 all3;
BY persnr;
/* kid(t)=0: at time t not in the household or in the sample */
ARRAY nkid_t nkid1-nkid3;
DO OVER nkid_t;
IF nkid_t eq . THEN nkid_t=0;
END;
pattern=100*nkid1+10*nkid2+nkid3;

PROC FREQ DATA=longitud;
TABLES nkid1 nkid2 nkid3;
TABLES nKid1*nkid2 nkid2*nkid3;
TABLES pattern;

```

run;

Exercise 8: Apply macro EC for the estimation of a random effects model.

Dependent variable: log of the gross monthly income Covariates:

age and sex (in *PPFAD*)

institutional years necessary to receive current degree of education (in *PGEN*)

Years with the current employer (in *PGEN*)

Use the first five soep waves (A to E).

```

/* Apply macro EC for the estimation of a random effects model. */
/* (Use the include statement) */
/* Dependent variable: log of the gross monthly income */
/* Covariates: sex age ( in PPFAD ) */
/* institutional years necessary to receive */
/* current degree of education ( in PGEN ) */
/* Years with the current employer ( in PGEN ) */
/* ***** */
/* file ex08.sas */
/* ***** */

/* macro for extracting variables in each wave */

%MACRO extract(lwave,inc,time);
/* inc= monthly gross income */
/* bilzeit=edu= years of education */
/* erwzeit=empl= years with the same employer */
DATA &lwave;
MERGE soep.&lwave.p(KEEP=persnr &inc
RENAME=(&inc=inc)
WHERE=(inc gt 100)
IN=valinc )
soep.&lwave.pgen(KEEP=persnr &lwave.bilzeit &lwave.erwzeit
RENAME=(&lwave.bilzeit=edu &lwave.erwzeit=empl)
WHERE=(edu gt 0 AND empl gt 0)
IN=valedu );
BY persnr;
IF valedu AND valinc;
unit=persnr;
time=&time;
%MEND;

%extract(a,ap3301,84);
%extract(b,bp4301,85);
%extract(c,cp5201,86);
%extract(d,dp4401,87);
%extract(e,ep4401,88);

DATA long1;
SET a(IN=a)
b(IN=b)
c(IN=c)
d(IN=d)
e(IN=e) ;
d85=0; if b then d85=1;
d86=0; if c then d86=1;
d87=0; if d then d87=1;
d88=0; if e then d88=1;

PROC SORT DATA=long1;
BY unit time;

/* control */
PROC PRINT DATA=long1(OBS=20);

DATA long2;
MERGE soep.ppfad(KEEP=persnr sex gebjahr)
long1(in=inlong1);

```

```

by persnr;
if inlong1;
Dsex=0; IF (sex EQ 2) THEN dsex=1;
age=1900+time-gebjahr;
exp=age-edu-6;
lninc=log(inc);
/* definition of the model */

%let _y      = lninc ;
%let _x      = exp edu empl ;
%let _z      = d85 d86 d87 d88 dsex ;

%let _nx     = 3 ;
%let _nz     = 5 ;

%include 'kurs\ecmacro.sas';
/* Read in source code of macro ec on file ecmacro.sas */
%ec(long2);

```

Exercise 9: Estimate the net change between 1984 (wave 1) and 1987 (wave 4) in the population total for persons with SPD party preference. Use the random group estimator.

```

/* ***** */
/* File: ex05.sas */
/* ***** */

/* ***** */
input: indat1 = sas file with variables hrf1 y1 rg
      hrf1 = weight at time 1
      Y1 = attribute at time 1
      rg = random-group
      indat2 = sas file with variables hrf2 y2 rg
      hrf2 = weight at time 2
      Y2 = attribute at time 2
output: outdata=sas file with variables phat vhat
      phat = estimated difference (net) of stock
      vhat = estimated variance of phat
      *****
*/;

%MACRO rgdiff(indat1,indat2,hrf1,y1,hrf2,y2,rg,outdat);
PROC SORT DATA=&indat1;
BY &rg; /* sort data by random groups */
RUN;

DATA dat1;
SET &indat1;
p1=8*&hrf1*&y1;
/* compute 8 random group estimates for the first point in time */

PROC UNIVARIATE DATA=dat1 NOPRINT;
BY &rg; /* results in 8 estimates */
VAR p1 ;
OUTPUT OUT=work1 SUM =y1_r;
/* variable y1_r gives population estimate for random groups */
RUN;

/* the same for the second point in time */

PROC SORT DATA=&indat2;
BY &rg;

DATA dat2;
SET &indat2;
p2=8*&hrf2*&y2;

PROC UNIVARIATE DATA=dat2 NOPRINT;
BY &rg;
VAR p2 ;

```

```

OUTPUT OUT=work2 SUM=y2_r ;
RUN;

/* computation of differences within random groups */

DATA workd;
MERGE work1 work2;
BY &rg ;
d_r=y1_r - y2_r;

/* computation of the variance of differences */

PROC UNIVARIATE DATA=workd NOPRINT;
VAR d_r;
OUTPUT OUT=&outdat MEAN=phat VAR=vhat;
RUN;

DATA &outdat;
SET &outdat;
vhat=vhat/8.0;

/* compute coefficient of variation cv and std deviation sig */
cv=sqrt(vhat/(phat*phat));
sig=sqrt(vhat);

PROC PRINT;

%MEND;

DATA spd1;
MERGE soep.ap(KEEP=hnr hnrakt persnr ap5602
              RENAME=(ap5602=pref)
              IN=inap )
      soep.phrf(KEEP=aphrf prgroup persnr
                RENAME=(aphrf=hrf1) );
BY persnr;
y1=0; IF (pref eq 1) then y1=1;
if inap;
run;

DATA spd2;
MERGE soep.fp(KEEP=hnr hnrakt persnr fp9302
              RENAME=(fp9302=pref)
              IN=infp )
      soep.phrf(KEEP=fphrf prgroup persnr
                RENAME=(fphrf=hrf2) );
BY persnr;
y2=0; IF (pref eq 1) then y2=1;
if infp;
run;

%rgdiff(spd1,spd2,hrf1,y1,hrf2,y2,prgroup,resultd);

```

Exercise 10: Estimate the population total of persons with permanent party preference for social democrats period: 1984–1988 period: 1986–1988 Calculalate also the variances of these estimates Use the random group estimator (macro rg)

```

/* ***** */
/* file: ex06.sas */
/* ***** */

%MACRO rg(indata,hrf,y,rg,outdata);
PROC SORT DATA=&indata;
BY &rg; /* sort data by random groups */

DATA &indata;
SET &indata;

```

```

p=8*&hrf*&y;
PROC UNIVARIATE DATA=&indata NOPRINT;
BY &rg;
VAR p;
OUTPUT OUT=work SUM =p_r;
RUN;
/* data file work has 8 observations:
   one for each random group */

PROC UNIVARIATE DATA=work NOPRINT;
VAR p_r;
OUTPUT OUT=work2 SUM=phat VAR=vhat;
/* data file has one observation */

DATA &outdata;
SET work2;
phat=phat/8.0;
vhat=vhat/8.0;
/* compute coefficient of variation cv and std deviation sig */
cv=sqrt(vhat/(phat*phat));
sig=sqrt(vhat);
KEEP phat vhat cv sig;
PROC PRINT DATA=&outdata;run;
%MEND;

DATA long1_5;
MERGE soep.ap(KEEP=persnr ap5602
              WHERE=(ap5602 EQ 1)
              IN=a)
      soep.bp(KEEP=persnr bp7902
              WHERE=(bp7902 EQ 1)
              IN=b)
      soep.cp(KEEP=persnr cp7902
              WHERE=(cp7902 EQ 1)
              IN=c)
      soep.dp(KEEP=persnr dp8802
              WHERE=(dp8802 EQ 1)
              IN=d)
      soep.ep(KEEP=persnr ep7702
              WHERE=(ep7702 EQ 1)
              IN=e)
      soep.phrf(KEEP=persnr aephfrf prgroup
                RENAME=(aephfrf=hrf) );
BY persnr;
IF (a AND b AND c AND d AND e);
y=1; /* only persons with permanent SPD preference in long1_5 */
%rg(long1_5,hrf,y,prgroup,res1_5);

/* now for period 3--5 */
DATA long3_5;
MERGE soep.cp(KEEP=persnr cp7902
              WHERE=(cp7902 EQ 1)
              IN=c)
      soep.dp(KEEP=persnr dp8802
              WHERE=(dp8802 EQ 1)
              IN=d)
      soep.ep(KEEP=persnr ep7702
              WHERE=(ep7702 EQ 1)
              IN=e)
      soep.phrf(KEEP=persnr cphrf dpbleib epbleib prgroup);
BY persnr;
IF (c AND d AND e);
hrf=cphrf*dpbleib*epbleib;
y=1; /* only persons with permanent SPD preference in long3_5 */
%rg(long3_5,hrf,y,prgroup,res3_5);

```

Exercise 11: Prepare a file that contains all persons who experienced exactly two spells of unemployment. The file shall include the following variables: Gender of person, Duration of first recorded unemployment spell, Duration of second recorded unemployment spell, Right Censoring status of the second spell.

```

/* ***** */
/* Exercise 11 suggested solution */
/* file: ex09.sas */
/* ***** */
/* Create a file that contains all unemployment spells which are
not left censored */

DATA unempl;
  SET artkalen(KEEP= persnr spellnr begin end zensor
              WHERE=( spelltyp EQ 5) AND (zensor le 3) );

/* Create a file that contains for each person the number of
unemployment spells */

DATA unemplc(KEEP= persnr count);
  SET soep.unempl;
  BY persnr;
  IF FIRST.persnr THEN count=0;
  count+1;
  IF LAST.persnr THEN OUTPUT;

/* Create a file that contains exactly two unemployment spells
per person */

DATA unempl2;
  MERGE unemplc(WHERE= (count EQ 2)
              IN=ge2 )
        unempl ;
  BY persnr;
  IF ge2 ;

/* Create a file with the first spell */

DATA first(KEEP=persnr length1 censor1);
  SET unempl2 ;
  BY persnr;
  IF FIRST.persnr THEN DO;
    length1=end-begin;
    censor1=censor;
  OUTPUT;
END;

/* Create a file with the second spell */

DATA second(KEEP=persnr length2 censor2);
  SET unempl2 ;
  BY persnr;
  IF LAST.persnr THEN DO;
    length2=end-begin;
    censor2=censor;
  OUTPUT;
END;

/* Create a longitudinal file where the first spell is uncensored */

DATA long;
  MERGE first(WHERE= (censor1 EQ 1)
              IN=infirst);
        second;
  BY persnr;

/* Add variable sex */

DATA long;
  MERGE long(IN=inlong)
        soep.ppfad(KEEP= persnr sex);
  BY persnr;
  IF inlong;
  dsex=0; IF (sex eq 2) THEN dsex=1;

```