Short course: Panel surveys in social and economic research and the treatment of nonresponse

Ulrich Rendtel

Freie Universität Berlin
Economic Department
Institute for Statistics and Econometrics

Lecture held at University of Helsinki
Universidad de Columbia a Bogota
Freie Universität
Spring 2011
Final version: 23. Feb. 2011

What is a Panel?

Wikipedia search for "Panel Data" : http://en.wikipedia.org
There are different panel units:

- Persons: Newborns; entrants into educational system; entrants into firms,entrants into unemployment, poverty, etc.
- Households (unstable units!): poverty measurement at household level; marriage, divorce, child-birth at household level
- Families (unstable units!): (Intergenerational) stability, time to first child after marriage.
- Firms (unstable units!): Investments, R&D activities at firm level
- Employer/Employees files(unstable relationship): Employee data at individual level
- Towns, states: aggregates, (international) comparisons

Panel data

From Wikipedia, the free encyclopedia

In statistics and econometrics, the term **panel data** refers to two-dimensional data. In marketing, *panel data* refers to data collected at the point-of-sale (also called *scanner data*).

Data are broadly classified according to the number of dimensions. A data set containing observations on a single phenomenon observed over multiple time periods is called time series. In time series data, both the *values* and the *ordering* of the data points have meaning. A data set containing observations on multiple phenomena observed at a single point in time is called cross-sectional. In cross-sectional data sets, the *values* of the data points have meaning, but the *ordering* of the data points does not. A data set containing observations on multiple phenomena observed over multiple time periods is called **panel data**. Alternatively, the second dimension of data may be some other than time. For example, when there is a sample of groups, like siblings or families, and several observations from every group, the data is panel data. Whereas time series and cross-sectional data are both *one-dimensional*, panel data sets are *two-dimensional*.

Data sets with more than two dimensions are typically called *multi-dimensional panel data*.

Contents [hide]

Example

[edit]

balanced panel:

persnr	*year*	*income*	*age*	*sex*
1	2003	1500	27	1
1	2004	1700	28	1
1	2005	2000	29	1
2	2003	2100	41	2
2	2004	2100	42	2
2	2005	2200	43	2

unbalanced panel:

persnr	*year*	*income*	*age*	*sex*
1	2003	1500	27	1
1	2004	1700	28	1
2	2003	2100	41	2
2	2004	2100	42	2
2	2005	2200	43	2
3	2004	3000	35	1

In the example above, two data sets with a two-dimensional panel structure are shown. Individual characteristics (income, age, sex) are collected for different persons and different years. In the left data set two persons (1, 2) are observed over three years (2003, 2004, 2005). Due to the fact that *each* person is observed *every* year, the left-hand data set is called an **balanced panel**, whereas the data set on the right hand is called an **unbalanced panel**, since Person 1 is not observed in year 2005 and person 3 only in 2004.

Data sets which have a panel design

[edit]

- German Socio-Economic Panel (SOEP)
- Household, Income and Labour Dynamics in Australia Survey (HILDA)

Information on panels from the internet

- List of Panel projects: http://www.paneldata.eu
- Mentioned in the course:
  - ECHP (CHINTEX project):
    http://www.destatis.de/CHINTEX/
  - SOEP:
    http://www.diw.de/en/soep
  - German Micro Census Panel:
    http://www.forschungsdatenzentrum.de/bestand/mikrozensus-panel/
- European Union Statistics of Income and Living Conditions (EU-SILC):
  http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/eu_silc/
- National Education Panel Survey (NEPS):
  http://www.uni-bamberg.de/neps/

Zurück Suchen Favoriten

Adresse http://www.paneldata.eu/

Google Los geht's! Lesezeichen 1 blockiert Rechtschreibprüfung Senden an

PanelData

- Home
- - Europe
- - North America
- - Asia/Oceania
- - Transcontinental
- PanelWhiz
- Tools
- About the Author
- Contact

World Panel Data Sets for Economics Researchers

• **Europe**:

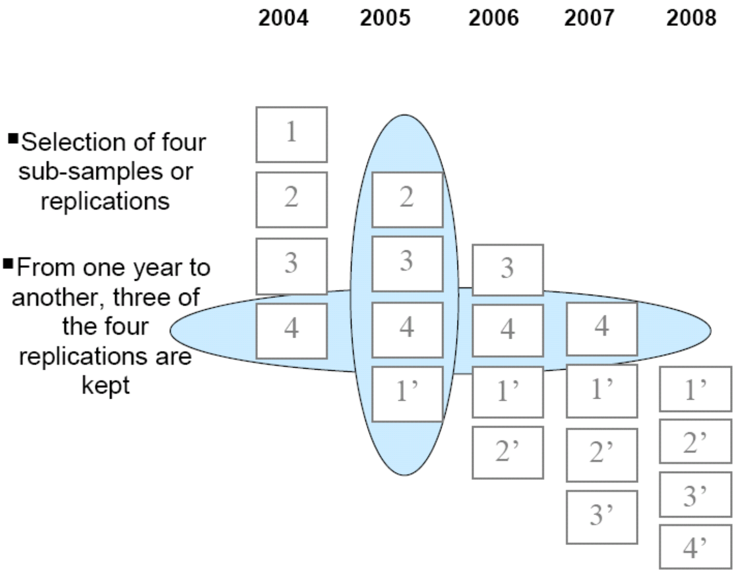
- [BE] PSBH / Belgian Household Panel
- [BH] BaHPS / Bosnia and Herzegovina Household Panel
- [CH] SHP / Swiss Household Panel
- [DE] SOEP / German Socio-Economic Panel (**PanelWhiz supported!**)
- [DE] IAB-BP / German Firm Panel from IAB-Nuernberg
- [DE] IAB-Matched / Matched Employer-Employee from IAB-Nuernberg
- [DE] MZ Panel / German Mikrozensus Panel
- [HU] HHP / Hungarian Household Panel
- [LU] PSELL / Luxembourg Household Panel
- [RU] RLMS / Russian Longitudinal Monitoring Survey
- [SE] LINDA / Longitudinal Individual Data for Sweden
- [UK] BHPS / British Household Panel Study

• **Pan Europe**:

- [EU] ECHP / European Community Household Panel
- [EU] EPAG / European Panel Analysis Group
- [EU] CHINTEX / ECHP User Group

Formats of panels

- Simple panel: unlimited participation (from the cradle to the grave), all household panels.
- Cohort sample: Sample from selected cohorts with unlimited participation (NEPS)
- Rotation panel: fixed, limited participation duration, for example 4 waves (EU-SILC), German MC Panel, all labour force surveys (LFS)
- Split panel: Simple panel + series of independent cross-sections


2004
2005
2006
2007
2008
■Selection of four sub-samples or replications
■From one year to another, three of the four replications are kept
1
2
3
4
2
3
4
1'
3
4
1'
2'
4
1'
2'
3'
1'
2'
3'
4'

Selection strategies

- Selection from register frame: known individual identifiers, possibility to use stratified sampling, known selection probabilities, automatic tracing.
- Selection from access panel (mostly commercial use): known individual identifiers, possibility to use stratified sampling, mostly quota sampling, automatic tracing.
- Multi-stage sampling from population: unknown individual identifiers, stratified sampling, known design selection probabilities, tracing only if intended.

Follow-up rules

- No follow-up in case of residential mobility (Area sampling): save field costs
- Follow-up of residential movers via telephone mode.
- Follow-up of only first wave panelists ("Sample persons") (PSID, ECHP, EU-SILC, BHPS, ...)
  Consequence: loss of all "Non-sample persons" who separate from "sample persons".
- Follow-up of all interviewed persons in households.
  Consequence: Additional information about household nets, over-sampling of persons who live in households with fusions, possibility of exploding sample size. (SOEP)
- Follow-up of firms in case of fusions, change of branch, etc. even more complicated.

Refreshment samples (1/2)

- Inclusion of new units entering the population: start-ups (firms), newborns, immigrants.
  - Sampling of population gains feasible with register (otherwise not).
  - Immigrant samples in the SOEP: Cumulation of households with immigrants after the first wave of the SOEP (1984) by screening interviews. Over-sampling of mixed households. (Mainly immigrants from eastern Europe after the fall of the "iron curtain" (1989).
  - Sample of newborns taken from the panel parents (easy to manage in a household panel). Advantage: Intergenerational analysis becomes possible.

Refreshment samples (2/2)

- Start of a "fresh" second panel in order to include population gains, increase sample sizes and counteract panel attrition (SOEP2 (Subsample F) starting in wave 2000) Over-sampling of persistent population.
- Inclusion of a new cohort
- Selective sample to counteract panel attrition: selection of "statistical twins" from an access panel.
  Correction of cross-sectional distributions at best. Statistical properties not clear.

Relationship of Panel Analysis and Time Series Analysis

Number of units: N ; Number of points in time T

- Panel analysis: N large and T small.
- Time series analysis: N small and T large.
- 2-dim asymptotics :
  - $\lim N/T \rightarrow \infty$
  - $\lim N/T \rightarrow 0$
  - $\lim N/T \rightarrow const$

What are the aims of panel analysis?

- Estimation of statistical models ("Model based approach"):
  - Causal effects: Change of X causes change of Y (before and after treatment measurement)
  - Variation of growth curves (for example in nutrition surveys)
  - Duration of episodes (for example duration of unemployment)
  - Transitions between states (for example labour force states in successive years)
- Population counts (Inclusion probabilities according to a sampling design ("Design based approach") :
  - Number of persons with specified longitudinal profiles (for example, persons in persistent poverty)
  - Separation of gross and net change (Gross change = flows between labour force states, net change = change of marginal distribution over labour force states)
  - Trend analysis: trends in the marginal population counts over panel waves.

Poverty Analysis from Finish ECHP (1/2)

Table 4: Register and survey based estimates of inequality and poverty in 1995 and 1999

	1995		1999	
	Survey	Register	Survey	Register
Measures of inequality				
- d90/d10 decile ratio	2.92	2.58	3.23	2.86
- Coefficient of variation	0.467	0.599	0.581	0.603
- Gini coefficient	0.238	0.226	0.265	0.251
Measures of poverty				
- Head count ratio	0.071	0.045	0.084	0.059
- Poverty gap ratio	0.020	0.012	0.026	0.016

Note: Poverty line= 50 percent of median income.

Note: Survey weights used.

Poverty Analysis from Finish ECHP (2/2)

Table 5: Transitions between the states "Poor" and "Non-poor" for survey and register income. Time interval: 1995 and 1999. (Un-weighted results)

	Transitions in percent	
	Poor	Non-Poor
	Register	
Poor	31.65	68.34
Non-Poor	5.34	94.65
	Survey	
Poor	30.40	69.59
Non-Poor	8.66	91.33

Need for meta information and data management

- Information is typically stored in a wave-based scheme. Household files + person files. Gross-sample information + net-sample information (10 files per wave). The SOEP is a collection of about 250 single flat files that must be combined!
  Web support of the SOEP: http://panelgsoep.de/soepinfo2009/
- Meta data: Link to "PanelWhiz": http://www.panelwhiz.eu
  Charity ware (20 Euro): Generates Stata-Files for the management of several household panels.

The 2 formats of panel files (1/2)

The Compressed or Flat-File Format:

Output 18.5.1 Compressed Data Set

Obs	i	cs	num	X_1	X_2	X_3	X_4	X_5	X_6	Y_1	Y_2	Y_3	Y_4	Y_5	Y_
1	1	CS1	-1.56058	0.40268	0.91951	0.69482	-2.28899	-1.32762	1.92348	2.30418	2.11850	2.66009	-4.94104	-0.83053	5.0135
2	2	CS2	0.30989	1.01950	-0.04699	-0.96695	-1.08345	-0.05180	0.30266	4.50982	3.73887	1.44984	-1.02996	2.78260	1.7385
3	3	CS3	0.85054	0.60325	0.71154	0.66168	-0.66823	-1.87550	0.55065	4.07276	4.89621	3.90470	1.03437	0.54598	5.0146
4	4	CS4	-0.18885	-0.64946	-1.23355	0.04554	-0.24996	0.09685	-0.92771	2.40304	1.48182	2.70579	3.82672	4.01117	1.9763
5	5	CS5	-0.04761	-0.79692	0.63445	-2.23539	-0.37629	-0.82212	-0.70566	3.58092	6.08917	3.08249	4.26605	3.65452	0.8182

The 2 formats of panel files (2/2)

The Long Format:

Output 18.5.2 Uncompressed Data Set

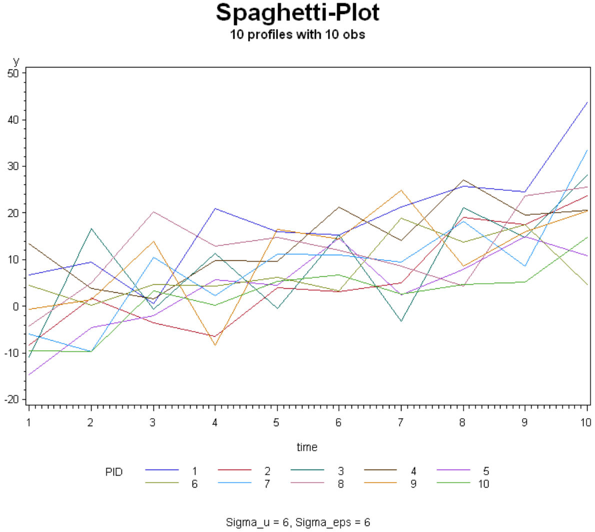
Obs	I	t	X	Y	CS	NUM
1	1	1	0.40268	2.30418	CS1	-1.56058
2	1	2	0.91951	2.11850	CS1	-1.56058
3	1	3	0.69482	2.66009	CS1	-1.56058
4	1	4	-2.28899	-4.94104	CS1	-1.56058
5	1	5	-1.32762	-0.83053	CS1	-1.56058
6	1	6	1.92348	5.01359	CS1	-1.56058

Transformation into Long Format

```
proc panel data=mysas.flat;
 flatdata indid=persnr tsname="t"
         transform=(income satis tenure)
 keep=(sex prgroup) /out=mysas.long;
 id persnr t;
 run;
```

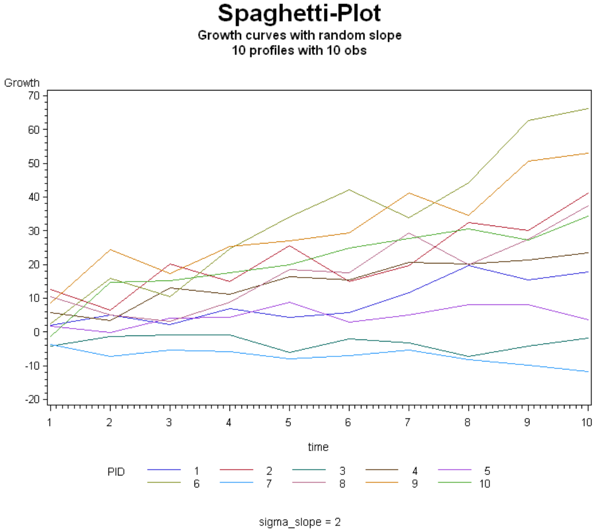
Useful descriptive statistics: Spaghetti Plots (1/3)

Trend plus large unit variation and large shocks:



Useful descriptive statistics: Spaghetti Plots (2/3)

Random slope plus moderate unit variation and moderate shocks:



Useful descriptive statistics: Spaghetti Plots (3/3)

```
Data Sim;
Do PID=1 to &n;
   alpha=&sigma_alpha*rannor(0) ;
   Beta_ran=&beta+ &sigma_beta*rannor(0); *Random slopes;
 Do time=1 to &t;
   u=&sigma_u*rannor(0);     * Variance Components;
   X=time+rannor(0);         * x strongly correlated with time
   y=alpha +&beta*x+ u;      * RE model;
   xx=x+&rho* alpha;         * xx correlated with alpha;
   yy=alpha+ &beta*xx +u;    * FGLM inconsistent ;
   YYY=alpha+ Beta_ran*x +u;* Mixed model with random slope;
   output;
 end;
end; run;
 symbol I=j v=none r=100; * join obs, no values, 100 repli.;
 proc gplot data=sim;  plot Growth*time=pid;
```

Useful descriptive statistics: Spaghetti Plots (3/3)

```
/* Programm simulates panel with n units and t waves */
/* Setting of the parameters via Macro variables: */
%Let n=20;              * Number of units;
%let t=10;               * Number of points in time;
%let sigma_alpha=3;      * Std. dev. of constants;
%let sigma_u=1;          * Std. dev. of shocks;
%let beta=2;             * Fixed effect of x;
%let rho=2;              * Covariance(X,alpha) inflation factor;
%let sigma_beta=1;       * Std. dev. of random slope of X;
```

Literature and further reading on general aspects

- Kasprzyk et al. (eds)(1989): Panel surveys, Wiley, New York.
- Lynn, P. (ed) (2009): Methodology of Longitudinal Surveys, Wiley, New York

Two linear models : The Fixed Effects (FE) Model

Index for units $i = 1, \ldots, N$, index for time $t = 1, \ldots, T$
Outcome variable $Y_{i,t}$ and covariate vector $X_{i,t}$ for each unit at each point in time.
For each unit there is a specific constant $\alpha_i \quad i = 1, \ldots, N$ and for each point in time there is a specific intercept γ_t in the linear model:

$$y_{i,t} = \alpha_i + \gamma_t + \beta' X_{i,t} + u_{i,t}$$

- Interpretation: the model parameters refer explicitly to the units and time periods. Hence we condition on these units and time periods.
- Makes sense in the case of state panels, for example, all federal states of the US or Germany or the member states of the EU.
- The number of coefficients may increase considerably.
- Alternative naming: Two-way model, because of the similarity with the two-way ANOVA model. Factor 1 identifies the units and factor 2 identifies the points in time.

Two linear models : The Random Effects (RE) Model (1/2)

For each unit there is a specific variance component $\alpha_i \quad i = 1, \dots, N$ that is independent from the shock component $u_{i,t}$ and follows a Normal distribution with expectation 0 and variance σ^2_α.

$$y_{i,t} = \alpha + \gamma_t + \beta' X_{i,t} + \alpha_i + u_{i,t}$$

- Interpretation: the model does not condition on the single units. It is a model that refers to the whole population. However the time periods are considered as fixed.
- Makes sense in the case of household panels.
- The number of coefficients increases by 1 (the variance σ^2_α) at the price of a distributional assumption (Normality of the α_i)

Two linear models (3/3): The Random Effects (RE) Model

- Alternative naming: Variance Component Model because for the random component $\epsilon_{i,t} = \alpha_i + u_{i,t}$ we get:

$$Cov(\epsilon_{i,t}, \epsilon_{j,s}) = \begin{cases} \sigma_\alpha^2 + \sigma_u^2, & \text{if } i = j \text{ and } t = s; \\ \sigma_\alpha^2, & \text{if } i = j \text{ and } t \neq s; \\ 0, & i \neq j. \end{cases}$$

  Matrix notation for Cov-matrix of $\epsilon_i = (\epsilon_{i,1}, \ldots, \epsilon_{i,T})'$:

$$\begin{aligned} Cov(\epsilon_i) &= \sigma_\alpha^2 \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & & \vdots \\ 1 & \ldots & 1 \end{pmatrix} + \sigma_u^2 \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} \\ &= \sigma_\alpha^2 \mathbf{1}\mathbf{1}' + \sigma_u^2 \mathbb{E} \end{aligned}$$

  where $\mathbf{1}$ is a row vector of T ones and $\mathbb{E}$ is the unit matrix of dimension T.
- If time dependence is omitted: One-way model Random Effects model.

The Kronecker Product notation

Econometric textbooks often use the Kronecker product notation.
Let A a matrix of Dimension $I \times J$ and let B a matrix of dimension $M \times N$ then the Kronecker product of the two matrices A and B is defined as a matrix of dimension $(IM) \times (JN)$ with:

$$A \otimes B = \begin{pmatrix} & \vdots & \\ \dots & a_{(i,j)}B & \dots \\ & \vdots & \end{pmatrix}$$

Then Σ the covariance matrix of $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$ can be written as:

$$\Sigma = \mathbb{E}_I \otimes \Sigma_i \qquad \text{where } \Sigma_i = \sigma_\alpha^2 \mathbf{1}\mathbf{1}' + +\sigma_u^2 \mathbb{E}_T$$

Thus Σ is a block diagonal matrix with diagonal elements Σ_i.

5 different panel estimators (1–3)

- The Pooled Estimator: OLS applied to $y_{i,t}$ and $x_{i,t}$ and time dummies.
  FE-Model: inconsistent (because of missing α_i's)
  RE-Model: consistent but wrong significance results (because of independence assumption)
- Dummy Variable (DV)-Estimator: OLS applied to $y_{i,t}$ and $x_{i,t}$ and unit and time dummies.
  FE-Model: Efficient
  RE-Model: Does not apply to model
- Within-Estimator: OLS applied to $y_{i,t} - y_{i,t-1}$ and $x_{i,t} - x_{i,t-1}$
  FE-Model without time dummies: consistent for β
  RE-Model without time dummies: Consistent

5 different panel estimators (4)

- Feasible Generalized Least Squares Estimator (FGLS): The covariance of error terms for unit i is the $T \times T$ Matrix:

$$\Sigma_i = \begin{pmatrix} \sigma_\alpha^2 + \sigma_u^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_u^2 & \dots & \sigma_\alpha^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \sigma_u^2 \end{pmatrix}$$

If we use $\hat{\sigma}_u^2$ and $\hat{\sigma}_\alpha^2$ as appropriate estimators for the respective variance components we obtain the estimated covariance $\hat{\Sigma}_i$. The GLS estimate with the estimated variance components is given by:

$$\hat{\beta}_{FGLS} = \left(\sum_i \mathbf{x}_i' \hat{\Sigma}_i^{-1} \mathbf{x}_i \right)^{-1} \left(\sum_i \mathbf{x}_i' \hat{\Sigma}_i^{-1} \mathbf{y}_i \right)$$

5 different panel estimators (4+5)

- FGLS estimator:
  FE-Model: does not apply
  RE-Model without time constants: asymptotical efficient
- The ML-estimate: can be derived by standard calculations (See Hsiao (1986 p.38 ff). Iterative Computation is necessary.
  The FGLS-estimator can be shown to be the first step in an iterative procedure to solve the ML-estimates. Therefore it is asymptotically efficient.

Two different ways for the computation of the FGLS estimator

The FGLS estimator can be shown to have the following two representations (see Màtyàs (1996, p.56)):

- OLS applied to $\tilde{y}_{i,t} = y_{i,t} - \theta \bar{y}_{i,.}$ and $\tilde{x}_{i,t} = y_{i,t} - \theta \bar{x}_{i,.}$ where $\theta = 1 - \sqrt{\frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + T\hat{\sigma}_\alpha^2}}$
- Between Estimator: OLS applied to $\bar{y}_{i,.}$ and $\bar{x}_{i,.}$
  The FGLS estimator can be shown to a linear combination of the Within- and the Between-Estimator

The Hausman Test in panel analysis (1/3)

- A frequent argument in econometric textbooks about panels:
  In the RE model the α_i represent unobserved variables that are specific to the unit, for example, intelligence if the outcome variable is the log of earned income. If education level is a covariate then the α_i are correlated with one of the covariates.
- In this case the FGLS Estimator $\hat{\beta}_{FGLS}$ is no longer consistent, as the estimated education level effect includes the intelligence effect.
- However: The Within Estimator $\hat{\beta}_W$ remains consistent, because the α_i are eliminated.
- Under the RE-Model (=Null Hypothesis) $\hat{\beta}_{FGLS}$ is asymptotically efficient and $\hat{\beta}_W$ is consistent. The test alternative ("Some of the covariates is correlated with the α_i") is not explicitly formulated.

The Hausman Test in panel analysis (2/3)

The Hausman Test uses a general asymptotic result on the covariance of $\hat{\beta}_{consistent} - \hat{\beta}_{efficient}$:

$$Cov(\hat{\beta}_{consistent} - \hat{\beta}_{efficient}) = Cov(\hat{\beta}_{consistent}) - Cov(\hat{\beta}_{efficient})$$

The Hausman Test:

$$\begin{aligned} T_{Hausman} &= (\hat{\beta}_W - \hat{\beta}_{FGLS})'(Cov(\hat{\beta}_W) - Cov(\hat{\beta}_{FGLS}))^{-1}(\hat{\beta}_W - \hat{\beta}_{FGLS}) \\ &\sim \chi^2_{DF} \end{aligned}$$

The number of degrees of freedom DF is equal to the number of estimated parameters.

The Hausman Test in panel analysis (3/3)

- $(cov(\hat{\beta}_W) - cov(\hat{\beta}_{FGLS}))$ is sometimes not invertible.
- Time-constant variables have to be removed from the model.
- Often covariates, like education level, are time-constant for the large majority of the sample. The Within estimator of education level then depends only on those few, who changed their education during the panel. Instability of Within estimate!

Example: SOEP data from (1984–2008) Dependent variable: Logearning

```
Proc Panel data=mysas.human_cap(obs=1000);
class svyyear ;
id pid svyyear;
model logearning = svyyear education_years marital_status
                   experience experience_q
      /noint fixone ranone pooled;
run;
```

- *id* Person identifier "pid" then time identifier "svyyear"!
- Model options: pooled, fixone, fixtwo, ranone, rantwo,
- Class statement generates dummies for each survey year.
- The data should be in the "Long"-format.

The General Mixed Model

- In the RE model the intercept has a random variation over the population.
- Maybe, some of the slope coefficients vary over the population.
- Furthermore one is interested in the impact of other covariates on these random slope coefficients.
- The Mixed Model (in matrix notation):

$$Y = X\beta + Z\gamma + \epsilon$$

  - $\beta =$ parameter vector of the fixed effects with known design matrix X
  - $\gamma =$ parameter vector of the random effects with known design matrix Z
  - The vector of errors ϵ and the random effects γ are independent and multivariate Normal distributed with expectations $\mathbf{0}$ and covariances $Cov(\epsilon) = R$ and $Cov(\gamma) = G$
  - $Cov(Y) = ZGZ' + R$

An example: growth curves of children (1/2)

```
data pr;
 input Person Gender $ y1 y2 y3 y4;
 y=y1; Age=8;  output;
 y=y2; Age=10; output;
 y=y3; Age=12; output;
 y=y4; Age=14; output;
 drop y1-y4;
 datalines;
1   F   21.0    20.0    21.5    23.0
2   F   21.0    21.5    24.0    25.5
    ...
;
```

Transformation into Long-format!

An example: growth curves of children (2/2)

```
proc mixed data=pr method=ml;
   class Person Gender;
   model y = Gender Age Gender*Age / s;
   random intercept Age / type=un sub=Person g;
run;
```

- Model option "s": display FE solution vector.
- "Type=un" requests unstructured covariance matrix for the random effects.
- Option "g": display the estimated G matrix.

Overview of several SAS-Mixed procedures

HPmixed: High case numbers of fixed and random effects can decrease the efficiency of Proc Mixed considerably. Proc HPmixed is specialized for a few Mixed models with simple covariance structures but more efficient in handling of the covariance structures.

GLIMmix: The linear Mixed model assumes a multivariate Normal distribution for the error terms. Proc GLIMmix deals with Non-Gaussian distributions.

NLmixed: Nonlinear models, like the Logit model, can be estimated by Proc NLmixed (Non-Linear Mixed).

Literature and further reading

- Hsiao, Ch. (1986): Analysis of Panel Data, Cambridge University Press, Cambridge
- Wooldridge, J (2002): Econometric analysis of cross-section and panel data. MIT Press
- Baltagi, B. (2001): Econometric Analysis of Panel Data. Second Edition, Wiley, New York.
- Verbeke,G., Molenberghs, G. (2000): Linear mixed models for longitudinal data, Springer, New York. (Biometrical textbook)

The representation of state sequences by Loglinear Models (1/2)

- Let the state space be given by the set $\{$e(mployed),u(memployed),n(ot in labour force) $\}$.
- Z_t indicates the state at wave $t = 1, 2, 3$.The state sequence (Z_1, Z_2, Z_3) generates a $3 \times 3 \times 3$ contingency table.
- In the cells there are the observed numbers $N_{Z_1=z1,Z_2=z2,Z_3=z3}$ in the panel.
- In order to simplify the notation we write $Z_1 = A$,$Z_2 = B$ and $Z_3 = C$.
- The expected number of cell counts $N_{A=a,B=b,C=c}$ is denoted by $\mu_{a,b,c}^{A,B,C}$

The representation of state sequences by Loglinear Models (2/2)

A Loglinear Model the expected cell counts is given by:

$$\log(\mu_{a,b,c}^{A,B,C}) = \beta_0 + \beta_a^A + \beta_b^B + \beta_c^C + \beta_{a,b}^{A,B} + \beta_{b,c}^{B,C} + \beta_{a,c}^{A,C} + \beta_{a,b,c}^{A,B,C}$$

β_a^A is the main effect of A. (Notation A).
$\beta_{a,b}^{A,B}$ is the interaction term of A and B. (Notation A*B).
$\beta_{a,b,c}^{A,B,C}$ is the (3-fold) interaction term of A,B and C. (Notation A*B*C).

Hierarchical Loglinear Models (1/2)

A Loglinear Model is called **hierarchical**, if the model contains for each interaction term of higher order all lower corresponding interaction terms. By dropping higher order interaction terms, one can formulate statements about independence and conditional independence:

- Joint independence:

  $$\underline{\text{Def.}}: \quad \pi_{a,b,c}^{A,B,C} = \pi_a^A \, \pi_b^B \, \pi_c^C \qquad \text{for all } a, b, c$$

  Model representation: $A + B + C$

- C is independent from A and B:

  $$\text{Def.}: \quad \pi_{a,b,c}^{A,B,C} = \pi_{ab}^{AB} \, \pi_c^C \qquad \text{for all } a, b, c$$

  Model representation: $A + B + A * B + C$

Hierarchical Loglinear Models (2/2)

- Conditional independence: A and C are independent for fixed values of B

$$\pi^{AC|B}_{ac|b} = \frac{\pi^{ABC}_{abc}}{\pi^{B}_{b}} = \pi^{A|B}_{a|b}\,\pi^{C|B}_{c|b} = \frac{\pi^{AB}_{ab}}{\pi^{B}_{b}}\,\frac{\pi^{CB}_{cb}}{\pi^{B}_{b}}$$

Model representation: $A + B + A * B + C + B * C$

A Markov Chain Model over 4 panel waves

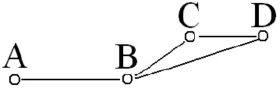
- Markov Chain for $Z_1 = A, Z_2 = B, Z_3 = C$ and $Z_4 = D$ is given by:

$$\pi_{abcd}^{ABCD} = \pi_{d|cba}^{D|CBA}\pi_{c|ba}^{C|BA}\pi_{b|a}^{B|A}\pi_{a}^{A}$$
$$= \pi_{d|c}^{D|C}\pi_{c|b}^{C|B}\pi_{b|a}^{B|A}\pi_{a}^{A}$$

Model representation: $A + B + C + D + A * B + B * C + C * D$
Note that there is no interaction between A and C (and A and D) because there is no direct impact of state A on state C (and D). The same holds for the direct impact of B on D.

Graphical Models

Graphical models are special hierarchical Loglinear models where the conditional independence relations can be directly read from a graph that connects the variables.



- Interpretation: A influences C and D only thru B
- Conditional independence: $A \otimes (C, D) \mid B$
- The cliques (Direct connections of all members) of the graph: $\{A, B\}$ and $\{B, C, D\}$
- Graphical model: The cliques of the graph generate the highest interaction terms in the hierarchical model.
- Hierarchical model representation.
  $A + B + A * B + C + D + B * C + B * D + C * D + B * C * D$

Loglinear Models with SAS

```
  PROC CATMOD DATA=mysas.Divorce;
   WEIGHT number;
   MODEL sex*sex_b*sex_o*mstatus=_Response_ ;
   LOGLIN   sex|sex_b  mstatus|sex_o|sex_b;
RUN; QUIT;
```

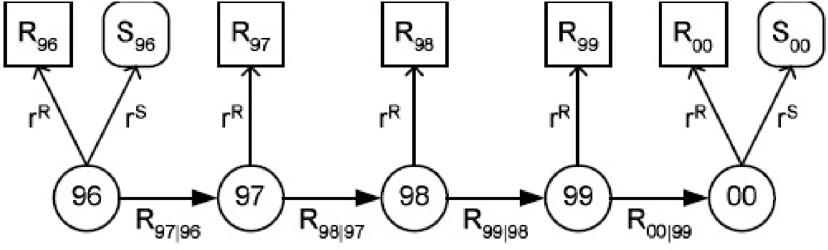
- You have to choose the "WEIGHT" statement for the counts ("number") of the table.
- The "MODEL" statement generates the contingency table.
- The "LOGLIN" statement specifies the cliques of the graph. $A|B|C$ means all 3-interaction terms of variables A,B and C plus all lower terms.

A latent Markov model (1/3)

- Model transitions between poverty states for the years (19)96. (19)97, (19)98, (19)99 and (20)00.
- For 1996 and 2000 two measurements for each person: one measurement from (ECHP) survey and one measurement from Finnish national register.
- Survey measurement indicated by S_{year}. Register measurement indicated by R_{year}.
- Years in between only register measurement.
- **Assumption 1**: Measurements depend only on the true but latent poverty state (indicated by circles).
- **Assumption 2**: The transitions between latent Markov states follow a Markov chain.

A latent Markov model (2/3)

The graphical representation of the latent Markov model:



A latent Markov model (3/3)

Observed and estimated transitions between the states "Poor" and "Non-poor". Time interval: 1996 and 2000

		Transitions in percent	
	Start	Poor	Non-Poor
	Register		
Poor	3.91	31.65	68.34
Non-Poor	96.8	5.34	94.65
	Survey		
Poor	7.56	30.40	69.59
Non-Poor	92.44	8.66	91.33
	True		
Poor	8.20	70.04	29.95
Non-Poor	91.79	3.06	96.93

Literature & Software (1/2)

- Hierarchical Models: every standard statistical package
  In SAS Proc Catmod with "loglin" statement
- Latent and Mixed Markov Models: PANMARK Package by v.d. Pol
  Useful but a little bit old.
  See. http://www.john-uebersax.com/stat/soft.htm
- However, Latent Markov models may be also estimated by LEM, which is freeware.
- Example can be found in Rendtel, U. / Nordberg, L. / Jäntti, M./ Hanisch, J. / Basic, E.(2004): Report on quality of income data CHINTEX Working Paper No.21, Statistisches Bundesamt, Wiesbaden. see http://www.destatis.de/CHINTEX/

Literature & Software (2/2)

- Langeheine, R., Pol F., v.d.(1990):A Unifying Framework for Markov Modeling in Discrete Space and Discrete Time, Sociological Methods Research, Vol. 18, 416-441.
- Pol, F.,v.d., R. Langeheine and W. de Jong (1991): PANMARK User Manual, Panel Analysis Using Markov Chains, Netherlands Central Bureau of Statistics, Voorburg.
- Pol, F., v.d., and J. de Leeuw (1986): A latent Markov Model to Correct for Measurement Error, Sociological Methods and Research, 15, 118-141.
- Rendtel, U., R. Langeheine and R. Berntsen (1998): "The estimation of poverty dynamics using different measurements of household income", Review of Income and Wealth, 44, 81-97.

Basic considerations (1/4)

- After the begin of an episode (spell), say unemployment, one is interested in the duration of this period.
- The exit from unemployment may result in different events, say employment, out-of-the-labour-force or some kind of training. The exits are regarded as competing risks. There are two types of analysis: One ignores the exit while the other makes inferences with respect to the exit.
- A new feature: the censoring of episodes (spells).
    - **Right Censoring:** The begin of a spell is observed, however, the end was not observed. Reasons: Spell continues after survey ends or person left the survey (not followed or discontinued cooperation)
    - **Left Censoring:** The start of the spell is not observed, however the end is observed. Reasons: The spell has begun, before the person entered the panel. Retrospective interviewing is imprecise.
    - **Left and right Censoring:** Start and end of the spell are unknown.

Basic considerations (2/4)

- Units of duration measurement:
  - Days (register)
  - Month (survey, register)
  - year (Survey)
- The three clocks: Calender time, process time and age
  - Calender time: often month 0 is the start of the panel.
  - Process time: elapse of time since the beginning of a spell, for example no. of month since the beginning of an unemployment.
  - Age: elapse of time since birth.

Basic considerations (3/4)

- Statistical analysis of the distribution of T, duration of spell (episode), time to event, ...
- Survival time: $S(t) = P(T > t) = 1 - F(t)$
  Contribution to likelihood in case of right censored spells!
- The hazard rate $h(t)$:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t (1 - F(t))} \\ &= \frac{f(t)}{1 - F(t)} \end{aligned}$$

  where $f(t)$ is the density of T and $F(t)$ is the distribution function of T.
- The hazard rate is measures the instant risk to stop the episode at time t, if the episode lasts at least until time t.

Basic considerations (4/4)

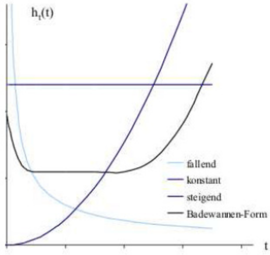
There are unique relationships between these 3 descriptions:

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

$$S(t) = \exp\left(\int_0^t h(u)du\right)$$

$$f(t) = -\frac{d(S(t)}{dt}$$

Typical hazard curves



- Declining (Infant mortality)
- Constant (electronic equipment without attrition)
- increasing (mechanical components with attrition)
- Bath tub shape (human life)

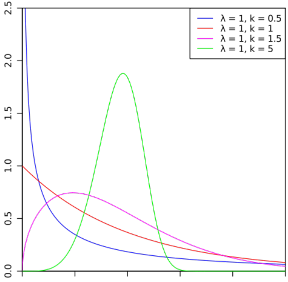
The hazard of some distributions

- The exponential distribution is a distribution "without memory": $F(t) = 1 - e^{-\lambda t}$ and $f(t) = F'(t) = \lambda e^{-\lambda t}$:

$$h(t) = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

- Weibull distribution: Hazard is a polynomial!

$$h(t) = \frac{k}{\lambda}(\frac{t}{\lambda})^{k-1}$$



3 different estimators of the survival function

- Parametric model, for example Exponential or Weibull, estimate model parameters, compute $\hat{S}(t)$ from estimated parameters. Handling of censored observations necessary!
- Nonparametric model:
    - Life table method:
      Subdivision of time axis into fixed, typically even spaced, time intervals
    - Kaplan-Meier (or Product Limit) estimate:
      Observations are ordered with respect to ascending duration or censoring times.
      Intervals are given by time-spans between the ordered data. Even spaced time intervals of the Life table method are regarded as restrictive!

The Kaplan-Meier estimate

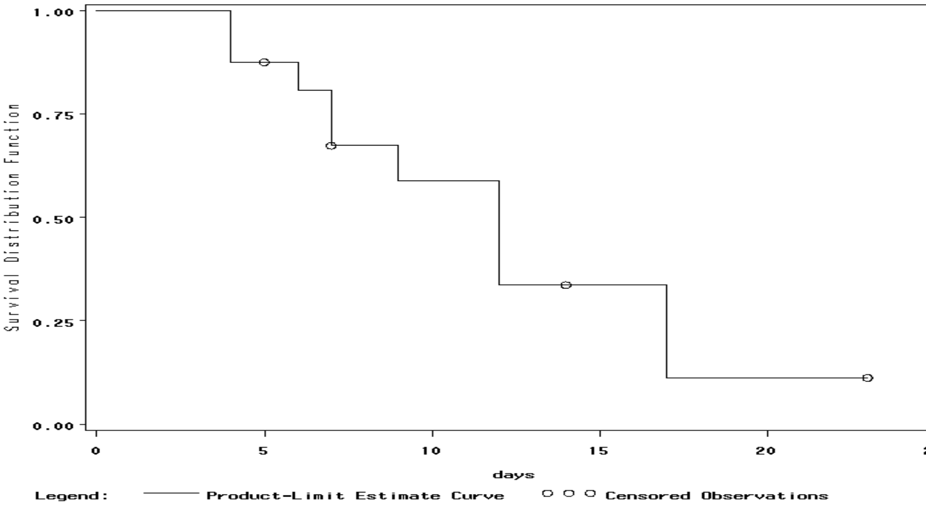
- $t_1 \leq t_2 \leq \ldots \leq t_n$ ordered set n durations
- $R_i =$ number of episodes under risk in time interval (t_{i-1}, t_i)
- $E_i =$ number of episodes with termination in time interval (t_{i-1}, t_i)
- $r_i = \frac{R_i - E_i}{R_i}$ estimated risk of survival in time interval (t_{i-1}, t_i)
- $$\hat{S}(t) = \prod_{t_i \leq t} r_i = r_1 \times r_2 \times \ldots \times r_i$$
- $\hat{S}(t)$ is a monotone decreasing step function that is constant on the intervals (t_{i-1}, t_i).
- The largest time value with defined $\hat{S}(t)$ is $t_{max} =$ maximum over all durations and censoring times.

A numerical example

Data with censorings (indicated by $+$):
$4,4,5^+,6,7,7,7^+,7^+,9,12,12,14^+,17,17,23^+$

i	t_i	E_i	C_i	R_i	$\hat{S}(t_i)$
1	4	2	0	16	$1 \cdot \frac{14}{16} = 0.8750$
2	5	0	1	14	$\frac{14}{16} \cdot \frac{14}{14} = 0.8750$
3	6	1	0	13	$\frac{14}{16} \cdot \frac{12}{13} = 0.8077$
4	7	2	2	12	$0.8077 \cdot \frac{10}{12} = 0.6731$
5	9	1	0	8	$0.6731 \cdot \frac{7}{8} = 0.5889$
6	12	3	0	7	$0.5889 \cdot \frac{4}{7} = 0.3365$
7	14	0	1	4	$0.3365 \cdot \frac{4}{4} = 0.3365$
8	17	2	0	3	$0.3365 \cdot \frac{1}{3} = 0.1122$
9	23	0	1	1	$0.1122 \cdot \frac{1}{1} = 0.1122$

The resulting Kaplan-Meier Plot



The use of Kaplan-Meier Plots

- The main use of Kaplan-Meier Plots is the comparison of survival curves between groups (or strata), for example comparison of treated vs control.
- The Log-Rank test is the standard test of group comparisons. It tests:

$$H_0 : S_1(t) = S_2(t) \quad \text{vs.} \quad H_1 : S_1(t) \neq S_2(t)$$

- The test bases on a comparison of observed ranks with the ranks that are expected under the NULL-hypothesis.
- Extensions to $k > 2$ groups are possible

The SAS-code

- Survival time of HIV-patients

  Variables:

  TIME: Survival time in months

  CENSOR: 1:deceased, not censored; 0: censored

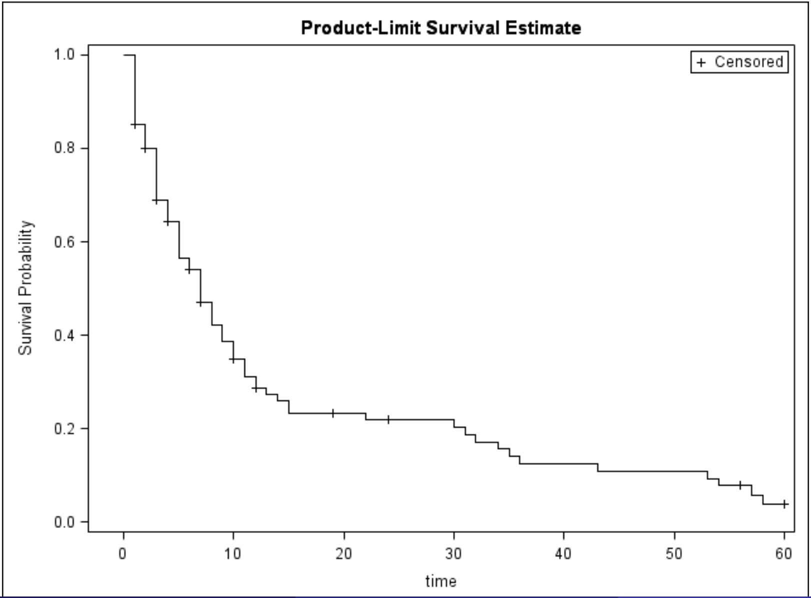
  DRUG: Drug consumption (1:yes; 0:no)

  AGE: Age at start of the study

- Generation of the Kaplan-Meier plots

```
ODS GRAPHICS ON;
PROC LIFETEST DATA=hmohiv PLOTS=(s);
TIME time*censor(0);
RUN;
ODS GRAPHICS OFF;
```

Plot of a survival functions



Local confidence limits

- An estimate of the variance of $\hat{S}(t)$ is given by the Greenwood formula:

$$\widehat{Var}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_{(i)} \leq t} \frac{E_i}{R_i(R_i - E_i)}$$

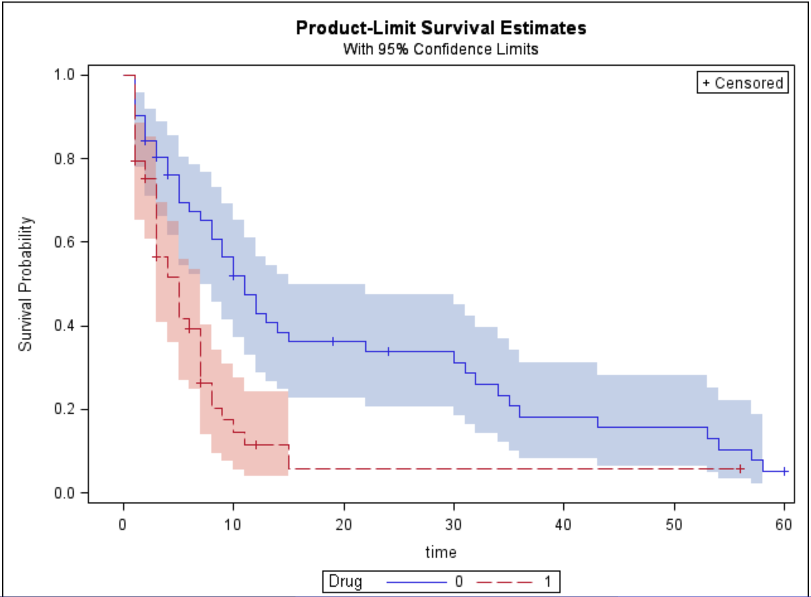
- The local confidence interval at time t is given by:

$$\hat{S}(t) \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{Var}(\hat{S}(t))}$$

- SAS code:

```
ODS GRAPHICS ON;
 PROC LIFETEST DATA=mysas.hmohiv plots=survival(cl);
 TIME time*censor(0);
 STRATA drug; RUN;
ODS GRAPHICS OFF;
```

Comparison of the survival functions



Different Hazard models

- Parametric model, for example Exponential or Weibull, estimate model parameters, compute $\hat{h}(t)$ from estimated parameters. Handling of censored observations necessary!
- Nonparametric model: the semi-parametric of Cox

$$h(t, \mathbf{x}) = h_0(t) exp(\mathbf{x}'\beta)$$

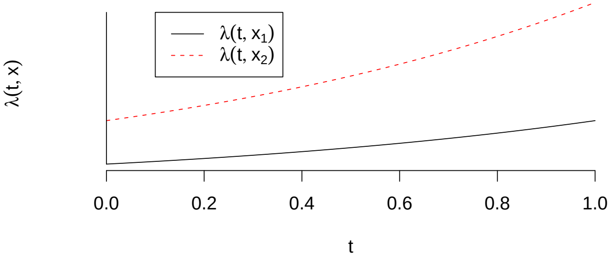
where $h_0(t)$ is an unrestricted **baseline hazard** function (nonparametric part). $\exp(\beta_p x_p)$ displays the effect of covariate x_p on the hazard function (parametric part).

Proportionality of Hazard rates

Proportionality of the Hazard rates $h(t, x_1)$, $h(t, x_2)$ for 2 individuals with covariates x_1, x_2:

$$\frac{h(t, x_1)}{h(t, x_2)} = \frac{h_0(t)e^{x_1\beta}}{h_0(t)e^{x_2\beta}} = e^{(x_1 - x_2)\beta}$$

Therefore the resulting Hazard curves are proportional (not necessary parallel!)



The SAS code for the Proprotional Hazard model

Have drug use and age an effect on the hazard rate of the HIV survival time?

```
proc phreg data=mysas.hmohiv;
model time*censor(0)=age drug;
run;
```

Note, the values after variable censor indicate Right-censored spells.

Testing the proportionality of the model

- Check of the proportionality assumption by an extra interaction term of the covariate with $\log(t)$. The interaction is computed for every time of event.
  This is automatically done by the PROC PHREG.
- Example:

```
proc phreg data=mysas.hmohiv;
model time*censor(0)=age drug drugtime;
drugtime=drug*time;
run;
```

Output of the survival function

- SAS code:
  BASELINE OUT=SAS-data-set COVARIATES=SAS-data-set SURVIVAL=s;
  Calculates for each covariate pattern listed in data set after the COVARIATES statement the survival function. Values in the data set after the OUT statement. The values of the survival function are written under a variable named by "s"
- Example:

```
proc phreg data=mysas.hmohiv;
model time*censor(0)=age drug ;
baseline out=test covariates=mysas.hmohiv survival=s;
run;
```

Literature & References

- Lawless. J.F. (2003): Statistical models and Methods for Lifetime Data, Second Edition, Wiley, New York.
- Allison, P. (1995): Survival Analysis using SAS, SAS Institute, Cary, NC. USA

The basics of design-based reasoning (1/3)

- A sample s is taken from a finite universe U
- The sampling follows a probability distribution over the set of possible samples. Thus S is a random set with realisation s and $Pr(S = s) = p(s)$.
- For each unit $k \in U$ the selection is indicated by a variable I_k:

$$I_k = \begin{cases} 1, & \text{if } k \in s; \\ 0, & \text{else} . \end{cases}$$

- Inclusion probabilities $Pr(I_k = 1) = Pr(k \in s) = \pi_k$
- Twofold inclusion probabilities $Pr(I_k = 1, I_j = 1) = Pr(k, j \in s) = \pi_{k,j}$

The basics of design-based reasoning(2/3)

- Characteristic of interest of unit k y_k is **not a random variable**!
- Population totals $t_y = \sum_U y_k$ are to be estimated by sample s.
- The π-estimator of t_y: $\hat{t}_y = \sum_U \frac{I_k}{\pi_k} y_k = \sum_s \frac{1}{\pi_k} y_k$
  Note: $\pi_k > 0$ for all $k \in U$ must hold.
  The π-estimator is often called Horvitz-Thompson (HT) estimator.
- The design weights $d_k = 1/\pi_k$.
  Design-weighted sample results: $\hat{t}_y = \sum_s d_k y_k$
  In official statistics "Weighting" is mostly associated with the use of a linear estimator with weights for the observations.
- Under random sampling $\hat{t}_y$ is unbiased: $E_\pi(\hat{t}_y) = t_y$
- Notice: No statistical model for y is assumed! The only randomness is the randomness of S!

The basics of design-based reasoning (3/3)

- The variance of the π-estimator:

$$V(\hat{t}_y) = \sum\sum_U Cov(I_k, I_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

  with $Cov(I_k, I_l) = \pi_{k,l} - \pi_k \pi_l$ and $\sum\sum_U$ as a shorthand for $\sum_{k \in U}\sum_{l \in U}$
- The general task in the design-based approach is to find sampling designs to keep the variance of the population estimates small.
- Often the coefficient of variation $\sqrt{V(\hat{t}_y)}/t_y$ is used as quality criterion.
- The variance of $\hat{t}_y$ has to be estimated on the basis of the sample:

$$\hat{V}(\hat{t}_y) = \sum\sum_s \frac{Cov(I_k, I_l)}{\pi_{k,l}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

Some sampling strategies (1/3)

- Simple (SI) random sampling with or without replacement: the classical urn experiment.
  - Fixed sample size n
  - $\pi_k = \frac{n}{N} \quad \pi_{k,l} = \frac{n(n-1)}{N(N-1)}$
  - $\hat{t}_y = \sum_s \frac{N}{n} y_k = N\bar{y}_s$ where $\bar{y}_s$ is the mean of the y-values in s.
  - Variance of π-estimator $V(\hat{t}_y) = (N^2/n)(1 - n/N)\sigma^2_{y,U}$ where $\sigma^2_{y,U}$ is the population variance of the y-values in U. Mind the difference to the model-based calculation of the variance of $N\bar{y}_s$!
- There are other sampling strategies than SI-sampling: sampling proportional to size (PPS), Bernoulli sampling (BE) with unequal sampling probabilities for the units.

Some sampling strategies (2/3)

- Stratified (ST) sampling: SI sampling within non-overlapping strata $U_h \quad (h = 1, \ldots, H)$, for example cross-classification of regions with age-sex groups. Sampling is independent, strata sizes N_h are known.
  - The strata sample sizes n_h can be used to minimize the variance of the population estimate (**Neyman allocation**): $n_h \propto N_h \sigma_{y,U_h}$
    where σ_{y,U_h} is the standard deviation of the y-values in stratum h.
    The result is intuitively appealing as it proposes to allocate sample size in those strata where the variation of the y-values is large. It marks the end of 'representative sampling'!
  - Population estimate: $\hat{t}_{y,ST} = \sum_{h=1}^{H} \hat{t}_{y,h}$
    where $\hat{t}_{y,h}$ is the π-estimate of the y-total in stratum h.
  - Because of the independence of sampling between strata we have:
    $V(\hat{t}_{y,ST}) = \sum_{h=1}^{H} V(\hat{t}_{y,h})$
  - Stratification can reduce the variance of population estimates considerably in case of large between strata variance of y values!

Some sampling strategies (3/3)

- In cases where no register for the original units exists, for example pupils, one switches to larger units, for example schools, with a register. The schools form clusters of pupils.
  - Cluster (CL) sampling: all units of the selected clusters are selected. The German micro census uses area sampling: all households of a selected area form a cluster of households. Increases variance!
  - 2-Stage (2ST) sampling: the clusters form the primary sampling units (PSU's). From each PSU a sample of secondary sampling units (SSU's) is selected.
    - Second stage sampling is independent between PSU's.
    - Inclusion probabilities: $\pi_k = \pi_i \pi_{k|i}$ if SSU k lies in PSU i where π_i is the inclusion probability of PPS i and $\pi_{k|i}$ is the conditional probability to include SSU k if PSU i is selected.
    - Often we have: $\pi_i \propto N_i$ and $\pi_{k|i} = n_{SSU}/N_i$ where N_i is the number of SSU's in PSU i.
      This convenient for the field organisation (fixed sample size n_{SSU} in every PSU). The result is an **equal probability sample** which is not SI. This selection scheme was used for the first wave of the SOEP (Subsample A).

Selection of samples with Proc Surveyselect

```
 proc surveyselect data=mysas.universe
    out=mysas.sample
    method=SRS
    sampsize=1000
    stats ;  * stats generates weights;
run;
```

The HT-estimator with Proc Surveymeans

```
 Proc surveymeans data=mysas.human_cap_sample(where=(svyyear=2
                  sum total=13119; * Total=number of elements
var earnings ;
weight samplingWeight;
run;
```

Literature & Software

About the Design-based Approach: Särndal, C.-E., Swensson, B., Wretman, J. (1992): Model Assisted Suvey sampling, Springer, New York.

A practical textbook: Lehtonen, R; Pahkinen, E. (2004): Practical Methods for Design and Analysis of Complex Surveys, Second Edition, Wiley, New York.

Sampling of the SOEP: Haisken-DeNew, J.; Frick, J. (Eds.) (2005) Desktop Companion to the German Socio-Economic Panel (SOEP), Download under: http://www.diw.de/en/diw_02.c.222846.en/desktop_companion

SAS Procedures:

- Proc SURVEYSELECT: Sampling from a frame.
- Proc SURVEYMEANS: Estimation with Survey weights.
- Proc SURVEYFREQ, SURVEYREG, SURVEYLOGISTIC: Frequency, Regression and Logistic Regression with survey data.

The generalized regression (GREG) estimator (1/4)

Idea:

- Step 1: Take a good prediction $\hat{y}_k$ of y_k on the basis of a covariate vector x_k.

$$\hat{\mathrm{B}} = \left(\sum_s d_k x_k x_k'\right)^{-1} \left(\sum_s d_k x_k y_k\right)$$

  Calculate $\hat{y}_k = x_k'\hat{\mathrm{B}}$ and the sample residuals $e_k = y_k - \hat{y}_k \quad k \in s$

- Step 2: Calculate the prediction total for U! Estimate the residual total by $\sum_s \frac{e_k}{\pi_k}$
- Step 3: In order to calculate the prediction total one has to know the total t_x of the covariate vector

$$\sum_U \hat{y}_k = t_x'\hat{\mathrm{B}}$$

The generalized regression (GREG) estimator (2/4)

The GREG estimator can be written as:

$$
\begin{aligned}
\hat{t}_{y,\mathrm{GREG}} =& \sum\nolimits_U \hat{y}_k + \sum\nolimits_s \frac{e_k}{\pi_k} \\
=& t_x' \hat{\mathrm{B}} + \sum\nolimits_s d_k (y_k - \hat{y}_k) \\
=& t_x' \hat{\mathrm{B}} + \hat{t}_{y,\pi} - \hat{t}_{x,\pi}' \hat{\mathrm{B}} \\
=& \hat{t}_{y,\pi} + (t_x - \hat{t}_{x,\pi})' \hat{\mathrm{B}}
\end{aligned}
$$

where $\hat{t}_{y,\pi}$ is the π-estimator of the y-total.
Thus $\hat{t}_{y,\mathrm{GREG}}$ can be read as an correction of the π-estimator.

The generalized regression (GREG) estimator (3/4)

Properties of the GREG

- The GREG is asymptotically design unbiased!
  Important: The unbiasedness holds whether or not the prediction model is correct!
  Therefore the Särndal et al. (1992) call this approach "model assisted" in contrast to "model based".
- The GREG weights w_k may be written as corrections of the design weights d_k:

$$w_k = d_k g_k = d_k(1 + x_k' \lambda)$$

where:

$$\lambda = \left(\sum_s d_k x_k x_k' \right)^{-1} (t_x - \hat{t}_{x,\pi})$$

The generalized regression (GREG) estimator (4/4)

- The variance of the GREG may be approximated by:

$$\hat{V}(\hat{t}_{y,\mathrm{GREG}}) = \sum\sum_{s} \frac{Cov(I_k, I_l)}{\pi_{k,l}} \frac{g_k e_k}{\pi_k} \frac{g_l e_l}{\pi_l}$$

  Notice: This is very similar to the variance formula of the π-estimate, however the y's are replaced by the residuals.

- If the y's are a linear combination of the covariate vectors x the variance of the GREG is 0!
- The GREG fulfills the calibration property:

$$\hat{t}_{x,\mathrm{GREG}} = t_x$$

The GREG estimator with Proc Surveyreg (1/2)

```
ods output estimates=mysas.total_predicted;
proc surveyreg data=mysas.human_cap_sample(where=(svyyear=200
class  marital_status gender_cohort;
model  earnings=gender_cohort marital_status /noint;
weight samplingweight;
output out=mysas.from_reg r=residual_reg;
estimate   'total of predicted values in 2000 population'
  gender_cohort  331 1320 1892 2318 1208 77 156 1033 1689 182
  marital_status 8591 3123 214 1191 /e;
run;
ods output close;
```

The GREG estimator with Proc Surveyreg (2/2)

```
ods output statistics=mysas.HT_residuals;
Proc surveymeans data=mysas.from_reg sum total=13119;
var residual_reg ;
weight samplingWeight;
run;
ods output close;

data greg;
  merge mysas.Total_predicted mysas.ht_residuals ;
  T_GREG=estimate+sum ;
  std_Greg=stddev ; std_REG=stdErr;
  keep T_Greg std_greg std_Reg;
  run; proc print; run;
```

Extensions of the GREG

- Ignore the residual term: $\rightsquigarrow$ 'synthetic' estimators (model dependent); small variances but possible bias.
- Model inhomogeneity on subgroups: $\rightsquigarrow$ Small Area estimators, Fixed and Random Effects Models for Areas (see Lehtonen/Pahkinen textbook)
- Departure from the linear model, for example use of the Logit model (see Lehtonen/Pahkinen textbook)

The general idea of calibration (1/3)

- Modify the design-based weights in such a way, that with the modified weights for some variables the known population totals are met:

$$\sum_{k \in s} d_k g_k x_k = \sum_{k \in U} x_k$$

where d_k is the design weight, g_k is the correction factor and X_k is a vector with known population totals.

- The calibration estimator for variable y is then given by:

$$\hat{T}_{y,CAL} = \sum_{k \in s} d_k g_k y_k$$

- The correction factors are not well-defined unless:
  - we have specified a distance function to the design-weights that is to be minimized .
  - we have restricted the functional form of the correction factors g_k

The general idea of calibration (2/3)

General approach of Deville/Särndal (1992): Select $w_k = d_k g_k$ such that:

$$\sum_s d_k G\left(\frac{w_k}{d_k}\right) = \text{minimum}$$

and G fulfills:

1. $G(x) \geq 0$ is strictly convex
2. $G(1) = 0, \Rightarrow 1$ is the absolute minimum of G.
3. $G'(1) = 0, \Rightarrow 1$ is the only absolute minimum of G.
4. $G''(1) = 1, \Rightarrow G$ behaves until the second derivative like a parabola $1/2(x-1)^2$

The general idea of calibration (3/3)

The Lagrange multiplier gives:

$$\sum_s d_k G\left(\frac{w_k}{d_k}\right) - \lambda'\left(\sum_s w_k x_k - \sum_U x_k\right)$$

Derivative for w_k:

$$d_k G'\left(\frac{w_k}{d_k}\right)\frac{1}{d_k} - \lambda' x_k = 0$$

With $F = (G')^{-1}$ one obtains:

$$g_k = F(\lambda' x_k)$$

Calibration with quadratic distances

$$
\begin{aligned}
G(x) &= \frac{1}{2}(x-1)^2 && x \in \mathbb{R} \\
G'(x) &= x - 1 \\
F(u) &= u + 1 && u \in \mathbb{R}
\end{aligned}
$$

$$w_k = d_k(1 + x_k'\lambda)$$

This results in the GREG!

Logarithmic distances

$$
\begin{aligned}
G(x) &= x\ln(x) - x + 1 & x \in \mathbb{R} \\
G'(x) &= \ln(x) \\
F(u) &= exp(u) & u \in \mathbb{R}
\end{aligned}
$$

$$w_k = d_k \exp(x_k'\lambda)$$

This results in the Iterative Proportional Fitting (IPF) solution!
(Fitting-to-Margins, Raking, ...)
Here x_k is a vector consisting of groups of dummy variables like:

$$
\begin{aligned}
x' &= (\text{Agegroup-Dummy}_1, \ldots, \text{Agegroup-Dummy}_L, \\
&= \text{Edu.group-Dummy}_1, \ldots, \text{Edu.group-Dummy}_M \\
&= \ldots)
\end{aligned}
$$

The racking procedure (1/5)

2 discrete variables: A (r values) B (c values)
Joint distribution of $A \star B$ unknown,

Marginal distr. of A known: N_{i+} ($i = 1, \ldots, r$)
Marginal distr. of B known: N_{+j} ($j = 1, \ldots, c$)

π-estimator of N_{ij}:

$$\hat{N}_{ij} = \sum_{s_{ij}} \frac{1}{\pi_k} \quad \text{where } s_{ij} = \text{subsample with} A = i, B = j$$

		B			
		1	$\ldots$	c	
	1				N_{1+}
A	$\vdots$		$\hat{N}_{ij}$		$\vdots$
	r				N_{r+}
		N_{+1}	$\ldots$	N_{+c}	

The racking procedure (2/5)

1. Step (Fit to *A*-margins):

Compute: $\hat{N}_{i+} = \sum_{j=1}^{c} \hat{N}_{ij}$ (Estimated total of A)

Compare: $\frac{\text{Total}}{\text{Estimate}} = \frac{N_{i+}}{\hat{N}_{i+}} \quad (i = 1, \ldots, r)$

Proportional correction factor for A:

$$\tilde{N}_{ij} = \frac{\text{Total}}{\text{Estimate}} \hat{N}_{ij} = \frac{N_{i+}}{\hat{N}_{i+}} \hat{N}_{ij}$$

guarantees $\sum_{j=1}^{c} \tilde{N}_{ij} = N_{i+} \quad (i = 1, \ldots, r)$

Replace $\hat{N}_{ij}$ by $\tilde{N}_{ij}$

The racking procedure (3/5)

2. Step (Fit to B-margins):
Proportional correction factor for factor B:

$$\tilde{N}_{ij} = \frac{\text{Total}}{\text{Estimate}} \hat{N}_{ij} = \frac{N_{+j}}{\hat{N}_{+j}} \hat{N}_{ij}$$

guarantees $\sum_{i=1}^{r} \tilde{N}_{ij} = N_{+j} \quad (j = 1, \ldots, c)$

Replace $\hat{N}_{ij}$ by $\tilde{N}_{ij}$
3. Step: Fit to A margins!

4. Step: Fit to B margins!

⋮

until $\frac{\text{Total}}{\text{Estimate}} \approx 1$ holds for A **and** B.

The racking procedure (4/5)

- For the solution N^*_{ij} it holds:

$$N^*_{ij} = \sum\nolimits_{s_{ij}} \frac{1}{\pi_k}\alpha_i\beta_j = \sum\nolimits_s \frac{1}{\pi_k} g_k$$

where

$$g_k = \begin{cases} \alpha_i\beta_j & \text{if } A = i, B = j \\ 0 & \text{else} \end{cases}$$

- $x'_k = (\delta_{1\bullet k}, \ldots, \delta_{r\bullet k}, \delta_{\bullet 1k}, \ldots, \delta_{\bullet ck})$
  where

$$\delta_{i\bullet k} = \begin{cases} 1 & A = i \text{ for unit } k \\ 0 & \text{else} \end{cases}$$

$$\delta_{\bullet jk} = \begin{cases} 1 & B = j \text{ for unit } k \\ 0 & \text{else} \end{cases}$$

The racking procedure (5/5)

The **calibration constraint** is fulfilled:

$$\sum\nolimits_U x_k = (N_{1+}, \ldots, N_{r+}, N_{+1}, \ldots, N_{+c})'$$

$$\sum\nolimits_s \frac{1}{\pi_k} g_k x_k =$$

$$(\sum_{j=1}^{c} N^*_{1j}, \ldots, \sum_{j=1}^{c} N^*_{rj}, \sum_{i=1}^{r} N^*_{i1}, \ldots, \sum_{i=1}^{r} N^*_{ic})'$$

$$= (N_{1+}, \ldots, N_{r+}, N_{+1}, \ldots, N_{+c})' = \sum\nolimits_U x_k$$

Model assisted estimation vs Calibration

- Statisticians are quite experienced in statistical modeling.
- Statistical agencies are more familiar with the calibration idea. There are some non-statistical benefits from calibration:
  - Calibration increases comparability across countries in European surveys.
  - Calibration increases comparability across panel waves in a panel survey.
- Negative weights may result from the GREG.
- Extensive Fitting-to-Margins may result in large variations of the sample weights.

Literature

Review Article Särndal, C.-E. (2007): The calibration approach in survey theory and practice, Survey Methodology, Vol. 33, 99–119

Calibration in panels

- Initial calibration: Initial wave.
- Final calibration: Last wave.
- Sequential calibration: First initial wave, then last wave. (Example: ECHP)
- Simultaneous calibration: First and last wave.
- Longitudinal calibration: Simultaneous calibration + calibration on known population changes (births, deceased persons, divorces) (Example: German MC Panel)

Use of linear panel models for prediction

- Random and Fixed Effects models may be estimated from the panel sample
- However, the predictions for the whole population are in general not feasible:

$$\sum\nolimits_U \hat{y}_k = \sum\nolimits_U (x_k'\beta + \alpha_k)$$

A simple example

$$y_{k,t} = \alpha_0 + x_{k,t}\beta' + \alpha_k + \epsilon_{k,t} \quad t = 1, 2$$

and $\alpha_k \sim N(0, \sigma^2_\alpha)$ and $\epsilon_{k,t} \sim N(0, \sigma^2_\epsilon)$

- Take ML estimator of α_0, β, obtain $\hat{\alpha}_0, \hat{\beta}$
- $\hat{\alpha}_k = \bar{y}_k - \hat{\alpha}_0 - \bar{x}_k\hat{\beta}$
- For $k \in s$ calculate $y_{k,1} - \hat{y}_{k,1}$:

$$\begin{aligned} y_{k,1} - \hat{y}_{k,1} &= y_{k,1} - \hat{\alpha}_k - \hat{\alpha}_0 - x_{k,1}\hat{\beta} \\ &= y_{k,1} - \bar{y}_k - (x_{k,1} - \bar{x}_k)\hat{\beta} \end{aligned}$$

- $\sum_U \hat{y}_{k,1} = (\sum_U x_{k,1})\hat{\beta} + \sum_U \hat{\alpha}_k$
  However by model assumption $\sum_U \hat{\alpha}_k = 0$
- $\Rightarrow$ Gain in precision over the cross-sectional estimator!

Inclusion probabilities for household panels (1/7)

- Household context is important for many analyses (poverty defined via household equivalence income) despite persons are the natural units of longitudinal analysis
- A simple example: persons i and j live in different households at wave 1 and move together in wave 2.
  Inclusion probabilities in wave 1 for person i: π_i and for person j: π_j
  Inclusion probability for persons i and j in wave 2:

  $$P(i \text{ selected in wave 1 or } j \text{ selected in wave 1 }) = \pi_i + \pi_j - \pi_{ij}$$

  If i selected in wave 1 and j not selected in wave 1: π_i known, π_j and π_{ij} often unknown!
  $\Rightarrow$ unknown design inclusion probabilities in wave 2!

Inclusion probabilities for household panels (2/7)

- A stupid rule: do not use information from the so-called "non-sample" persons, loss in efficiency!
- A better alternative: I_i, I_j inclusion indicators wave 1 for i and j; $0 \leq \lambda_i \leq 1$ and $\lambda_j = 1 - \lambda_i$ fixed (!) numbers.
  Compute: $w_{i,j} = w(I_i, I_j) = \lambda_i \frac{I_i}{\pi_i} + \lambda_j \frac{I_j}{\pi_j}$
  Then: $E(w) = 1 \Rightarrow$ Use of weight w produces unbiased population total estimates in wave 2 without knowledge of inclusion probability!
- Selection of λ ?

Inclusion probabilities for household panels (3/7)

- A famous rule ("Fair share"): w = average of all individual weights of adult sample persons. number$=n_{h,adult}$

$$\text{individual weight}_i = \begin{cases} 1/\pi_i, & \text{if } I_i = 1 \text{ (i.e. i is a sample person);} \\ 0, & \text{if } I_i = 0 \text{ (i.e. i is not a sample person).} \end{cases}$$

What is the corresponding λ–representation?

$$\begin{aligned} w_h &= \frac{1}{n_{h,adult}} \sum_{i \in \text{household h}} \frac{I_i}{\pi_i} \\ &= \frac{1}{n_{h,adult}} \sum_{i \in \text{household h}} I_i d_i \end{aligned}$$

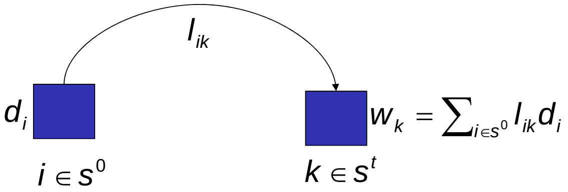
Inclusion probabilities for household panels (4/7)

A more formal approach:

$U^0, U^1, U^2, \ldots, U^t$ = Universe of persons at wave $0, 1, 2, \ldots, t$
$s^0 \subset U^0$ sample of persons with $I_i = 1$
y_k^t = variable of interest for person $k \in U^t$
Total of interest: $T_{y^t} = \sum_{k \in U^t} y_k^t$
Design weights: $d_i = 1/\pi_i$
Link function $l_{j,k}$: mapping $U^0 \times U^t \rightarrow \mathbb{R}^+$ reflects tracing from person j in wave 0 to person k in wave k.

Link Functions

Redistribute initial weights d_i of $i \in s^0$ onto $k \in s^t$.



Example: Link Functions

Typically defined from wave to wave:

NO WEIGHT SHARE

$$l_{ik} = \begin{cases} 1 & i \text{ identical to } k \\ 0 & \text{otherwise} \end{cases}$$

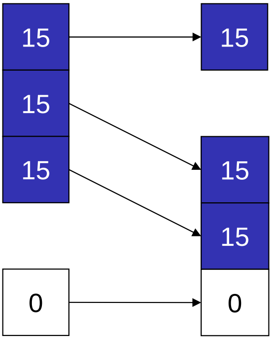
EQUAL WEIGHT SHARE

$$l_{ik} = \begin{cases} 1/N_h & i \text{ in household } h \\ 0 & \text{otherwise} \end{cases}$$

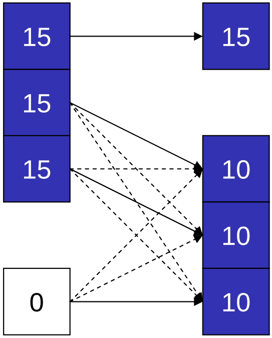
k lives in household h, size N_h.

Example: Link Functions

NO WEIGHT SHARE



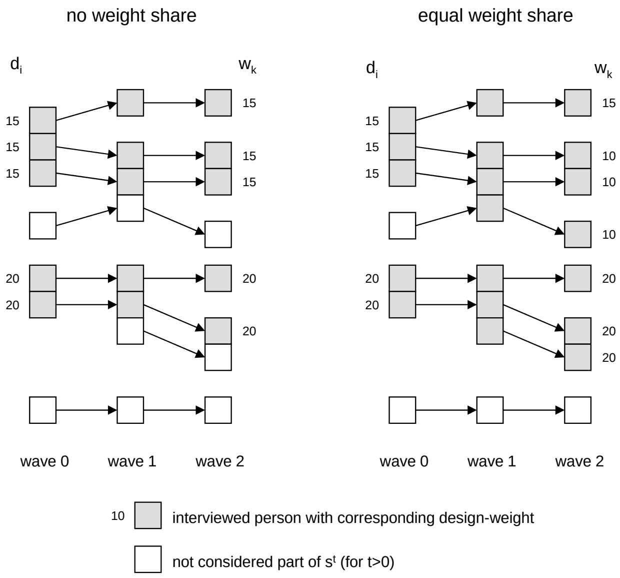
EQUAL WEIGHT SHARE



Inclusion probabilities for household panels (5/7)

- Usually link functions are constructed between persons j and k from the populations U^{t-1} and U^t
- Connecting link functions:

$$l_{ik}^{0,t} = \sum_{j \in U^{t-1}} l_{jk}^{t-1,t} \sum_{i \in U^0} l_{ij}^{0,t-1}$$



no weight share
equal weight share
d_i
w_k
15
15
15
15
15
15
20
20
20
20
15
15
15
15
10
10
10
20
20
20
20
wave 0
wave 1
wave 2
wave 0
wave 1
wave 2
10
interviewed person with corresponding design-weight
not considered part of s^t (for $t>0$)

Inclusion probabilities for household panels (6/7)

The estimator of T_{y^t}:

$$\hat{T}_{y^t} = \sum_{k\in s^t} w_k y_k^t = \sum_{k\in s^t} y_k^t \sum_{i\in s^0} l_{ik} d_i$$

- l_{ik} is to be known only for $i \in s^0$ and $k \in s^t$
- Convexity condition for the link : for all $k \in U^t$: $\sum_{i\in U^0} l_{ik} = 1$

$$\begin{aligned} E(\hat{T}_{y^t}) &= \sum_{k\in U^t} \sum_{i\in U^0} l_{ik} d_i E(I_i) y_k^t \\ &= \sum_{k\in U^t} y_k^t \sum_{i\in U^0} l_{ik} \\ &= \sum_{k\in U^t} y_k^t \end{aligned}$$

Inclusion probabilities for household panels (7/7)

- Variance:

$$\hat{T}_{y^t} = \sum_{i \in U^0} d_i I_i \sum_{k \in U^t} l_{ik} y_k^t = \sum_{i \in s^0} d_i \tilde{y}_i^t$$

where $\tilde{y}_i^t = \sum_{k \in U^t} l_{ik} y_k^t$ can be seen as the "future" contribution y^t of person $i \in U^0$ to the estimation of the total of T_{y^t}.

$$V(\hat{T}_{y^t}) = \sum_{i \in U^0} \sum_{i' \in U^0} Cov(I_i, I_j) d_i d_{i'} \tilde{y}_i^t \tilde{y}_{i'}^t$$

- Variance estimation

$$\hat{V}(\hat{T}_{y^t}) = \sum_{i \in s^0} \sum_{i' \in s^0} \frac{Cov(I_i, I_j)}{\pi_{i,j}} \tilde{y}_i^t \tilde{y}_{i'}^t$$

Literature on weighting in household panels

- Lavallée, P. (1995): Cross-sectional Weighting of Longitudinal Surveys of Individual Households Using the Weight Share Method. *Survey Methodology*, 21, 25-32.
- Kalton, G., Brick, J. (1995): *Weighting Schemes for Household Panel Surveys*. Survey Methodology, 21, 33-44.
- Lavallée, P., Deville, J.-C. (2002): Theoretical Foundations of the Generalised Weight Share Method. *Proceedings of the International Conference on Recent Advances in Survey Sampling 2002*. Carleton University, Ottawa.
- Rendtel, U., Harms,T. (2009): Weighting and Calibration for Household Panels, In: Lynn (ed.), Methodology of Longitudinal Surveys, Wiley, New York, 265–286.

Causes for nonresponse in surveys

- Latest Compilation Book: Groves et al. (eds) (2002): Survey Nonresponse, Wiley.
- Groves, R. (1998): Nonresponse in Household Interview Surveys. Wiley
- Causes for nonresponse:
  - Invalid address (if selection via register)
  - No contact
  - Unable to respond
  - Unwilling to cooperate (last stage of sequential model)
  - Nonresponse on sensitive items
  - Nonresponse by design (Rotating out respondents, no tracing of residential movers)

Panel attrition

Definition:

- Successive nonresponse of
- eligible persons/households
- after start of the panel

This is **not** panel attrition:

- Demographic losses
  - Identification of deceased persons
  - Identification of emigrants
- Restricted statistical/software ability to analyze unbalanced panels

Panel attrition in the SOEP

Figure 9: **All first wave persons (subsample A+B). Development until wave 22.**

Panel attrition in the ECHP

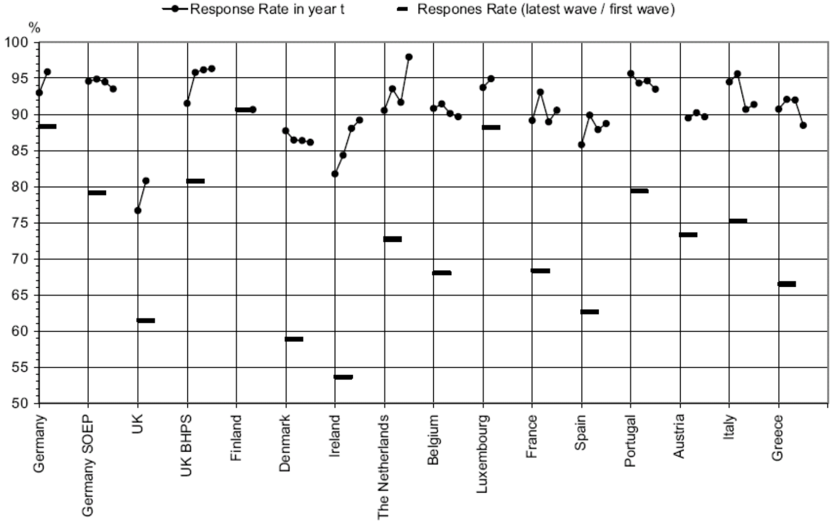


Figure 2 Response rates across countries for wave 2 to wave 5 and the overall response rate

Panel attrition in the FIN ECHP (1/2)

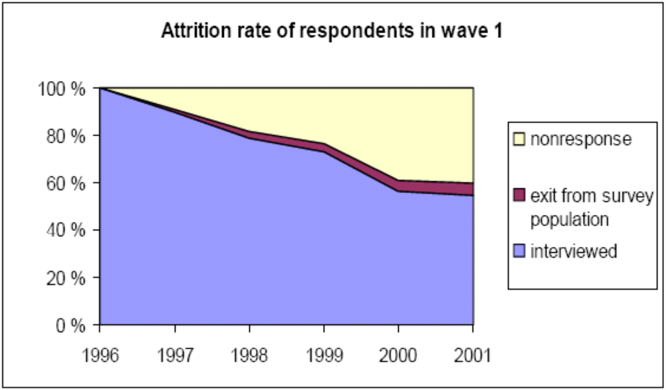
Statistics Finland

Outcome of households

	1996	1997	1998	1999	2000	2001
Old households	-	4 441	4 254	4 213	4 159	3 238
New households	5 732	244	320	295	169	144
Overcoverage	81	26	40	32	48	63
Net sample	5 651	4 659	4 534	4 476	4 280	3 319
Interviewed	4 139	4 104	3 920	3 822	3 104	3 115
Nonresponse	1 512	555	614	654	1 176	204
non-contact	199	135	103	95	81	64
refusal	1 288	402	353	315	969	134
language	13	3	2	1	1	-
technical loss	12	15	156	243	125	6

Panel attrition in the FIN ECHP (2/2)

Attrition in FI ECHP



Specific causes for panel attrition

- Tracing failure of residential movers (but follow-up via telephone!)
- Unwillingness to cooperate
  - Late unit nonresponse after previous item nonresponse
  - Change of the interviewer
  - "No time" at new residence / household (perception of household as unit of survey)
  - Changes in the household composition may exhibit private details (for example change of partner)
- Changes in field work conditions
  - Change of interview mode (switch to telephone/CAPI/postal)
  - Changes in the questionnaire (SOEP wave 5: balance of assets)
  - Cumulative response burden

Impact of some variables of ECHP attrition

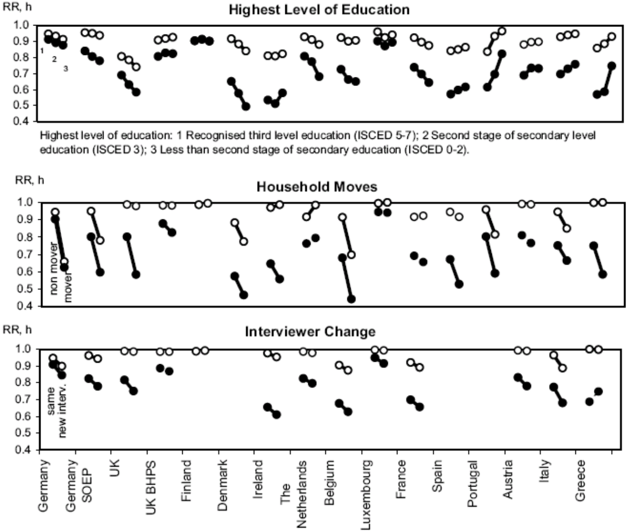


Figure 5 Response rates for subsamples

Literature on panel attrition:

Some theoretical results:

- "Toward a theory of nonresponse in panel surveys" see Lepkowski, J.;Couper, M. (2002): Nonresponse in the Second wave of Longitudinal Household Surveys. In: Groves et al. (eds): Survey Nonresponse, Wiley, 259–272.
- Rendtel, U. (2002): Attrition in Household Panels: A Survey. CHINTEX Working Paper No. 4, URL: www.destatis.de/chintex/download/paper4.pdf
- An econometric view: Verbeek, M.; Nijman, Th. (1996): Incomplete Panels and Selection Bias. In: Matyas, L; Sevestre, P. (eds), The Econometrics of Panel Data (Second edition), Kluwer, Dordrecht, 449–490.

Literature on panel attrition

Some empirical results:

- Unit Nonresponse in the ECHP: Behr, A., et al. (2005):Extent and Determinants of Panel Attrition in the European Community Household Panel. European Sociological Review, 21,489–512.
- Unit Nonresponse in the Finnish subsample of the ECHP: Rendtel et al. (2004): Report on Panel effects, CHINTEX Working paper 22, URL: www.destatis.de/chintex/download/paper22.pdf
- Item Nonresponse in the ECHP: Buck, N. (2004): Item Nonresponse in the ECHP, In: Ehling/Rendtel (eds): Harmonisation of Panel Surveys and Data Quality, Statistisches Bundesamt, Wiesbaden, 188–209.

Literature on panel attrition

Some empirical results:

- PSID: Fitzgerald et al.(1998): An Analysis of Sample Attrition in Panel Data - The Michigan Panel Study of Income Dynamics. Journal of Human Resources, 33, 251-299.

Initial bias fade-away (1/4)

Nonresponse is thought to:

- Reduce case numbers: poor significance results
- Distort sample distributions: Normal to Non-normal
- Lead to invalid statistical inference: Bias and/or Variance
- Initial nonresponse and panel attrition may cumulate in their distorting effects

The last hypothesis can be checked for variables that are known from a population register for all eligible persons, like in Finland!

Initial bias fade-away (2/4)

The direct approach in the Finnish subsample of the ECHP linked at person level with records from the Finnish population register.

- Merge the wave 1 gross–sample with information from the population register files
- In later waves (2–6) : Add information on dwelling units to calculate household based figures
- Compare results for 3 samples:
  - Full: gross–sample wave 1
  - RESP: net–sample wave 1
  - OBS: net–sample wave t

Initial bias fade-away (3/4)

- Difference FULL – RESP : Effect of initial nonresponse
- Difference RESP – OBS : Effect of panel attrition
- Difference FULL – OBS : Total effect of nonresponse

Initial bias fade-away (4/4)

- Column "Full": Income Quintiles (Household equivalence Income) defined for the gross-sample FIN-ECHP Wave 1 (=1996) with 14616 persons; Bounds in FIM: 57924, 73136, 88899, 114579
- Column "RESP": Respondents of the first wave grouped according above bounds
- ⇒ High incomes are under-represented in first wave.

	t=1996	
Sample	Full	Resp
size	14616	7809
Distr. on states	(1)	(2)
$\pi(1)$	20.0	21.8
$\pi(2)$	20.0	20.7
$\pi(3)$	20.0	21.8
$\pi(4)$	20.0	20.1
$\pi(5)$	20.0	15.6

	t=1996		t=2000					
Sample size	Full 14616	Resp 7809	Full 14616		Resp 7809		Obs 5192	
Distr. on states	(1)	(2)	Markov (3)	emp. (4)	Markov (5)	emp. (6)	Markov (7)	emp. (8)
$\pi(1)$	20.0	21.8	23.3	23.9	23.9	22.2	23.9	22.4
$\pi(2)$	20.0	20.7	17.8	16.9	18.2	16.6	18.2	17.4
$\pi(3)$	20.0	21.8	18.4	18.3	18.6	17.9	18.6	17.6
$\pi(4)$	20.0	20.1	21.3	20.6	21.0	21.4	21.0	21.8
$\pi(5)$	20.0	15.6	19.3	20.4	18.1	22.0	18.1	20.9

Initial nonresponse has almost vanished! Is this a singular result?

Year	FULL Sample ALL (1)	FULL Sample RESP (2)	OBS (3)
1996	0.253	0.228	0.228
1997	0.236	0.232	0.231
1998	0.255	0.243	0.243
1999	0.252	0.246	0.246
2000	0.273	0.255	0.256

Table 6: **Comparison of Gini-coefficients of the OECD equivalence income for different years.**

Year	FULL Sample ALL (1)	FULL Sample RESP (2)	OBS (3)
	Less than 50 percent of median		
1996	4.9	4.4	4.4
1997	5.6	5.2	5.0
1998	6.0	5.7	5.3
1999	6.0	5.7	5.4
2000	6.5	6.4	6.0
	Less than 50 percent of mean		
1996	7.3	5.8	5.8
1997	7.1	6.9	6.7
1998	8.1	7.7	7.2
1999	8.1	7.8	7.5
2000	9.7	9.2	8.9

Table 8: **Comparison of percentages of poor defined by having 50 percent or less than the median or the mean of OECD equivalence income.**

	Year			
	1997	1998	1999	2000
Transition 1 → 1				
ALL	65.8	61.3	57.1	54.8
RESP	68.2	63.2	57.6	54.3
OBS	69.0	66.4	61.9	55.7
Transition 2 → 2				
ALL	51.2	46.5	42.3	37.9
RESP	55.2	50.4	42.2	38.1
OBS	54.8	51.3	42.8	36.9
Transition 3 → 3				
ALL	43.8	39.6	35.1	33.6
RESP	50.4	45.8	38.1	35.3
OBS	49.6	46.3	39.2	36.4
Transition 4 → 4				
ALL	44.7	40.7	37.2	35.5
RESP	54.4	46.6	38.9	35.8
OBS	55.2	48.2	42.1	37.0
Transition 5 → 5				
ALL	66.2	62.3	58.7	56.3
RESP	73.2	67.4	61.9	58.1
OBS	74.6	68.2	62.0	56.4

Table 10: **Transition rates between quintiles of the OECD equivalence income. Starting period is 1996. Ending period varies between 1997 and 2000.**

A Markov Chain Approach (1/3)

- Variable of interest $X_t \quad (t = 1, 2, \ldots)$ follow a Markov chain with a finite state space $\mathcal{S} = \{1,2,..., k\}$
- Transition matrix P between subsequent states time-homogeneous
- There exists a steady state distribution π^* of P: $\pi_t = P^t \pi_0 \rightarrow \pi^*$ for $t \rightarrow \infty$
- Initial nonresponse results in different starting distributions $\pi_{0,\text{FULL}}$ and $\pi_{0,\text{RESP}}$
- Transition matrix P is the same for both samples!
- Then $\pi_{t,\text{RESP}} \rightarrow \pi_{t,\text{FULL}}$ for $t \rightarrow \infty$
  In a non-formal saying: the effect of the initial nonrespose "fades" out!

A Markov Chain Approach (2/3)

The estimated transition matrix between income quintiles:

$$P = \begin{pmatrix} 72.2 & 18.3 & 5.4 & 2.5 & 1.6 \\ 20.6 & 49.9 & 21.4 & 6.3 & 1.9 \\ 6.9 & 16.7 & 49.1 & 23.2 & 4.1 \\ 4.5 & 5.1 & 16.3 & 57.1 & 17.0 \\ 4.0 & 2.6 & 4.0 & 16.0 & 73.4 \end{pmatrix}$$

A Markov Chain Approach (3/3)

	t=1996		t=2000					
Sample size	Full 14616	Resp 7809	Full 14616		Resp 7809		Obs 5192	
Distr. on states	(1)	(2)	Markov (3)	emp. (4)	Markov (5)	emp. (6)	Markov (7)	emp. (8)
$\pi(1)$	20.0	21.8	23.3	23.9	23.9	22.2	23.9	22.4
$\pi(2)$	20.0	20.7	17.8	16.9	18.2	16.6	18.2	17.4
$\pi(3)$	20.0	21.8	18.4	18.3	18.6	17.9	18.6	17.6
$\pi(4)$	20.0	20.1	21.3	20.6	21.0	21.4	21.0	21.8
$\pi(5)$	20.0	15.6	19.3	20.4	18.1	22.0	18.1	20.9

Concluding remarks on bias fade-away

- Results from the PSID (started in 1968!) indicated in the same direction. Here Fitzgerald et al. (1998) investigated the panel attrition in the PSID. They stated that the difference of the un-calibrated PSID distribution to the Census results becomes smaller with increasing number of waves!
- The result depends on the speed of the Markov chain to swing into the steady state.
- Therefore one may conjecture that a refreshment sample incurs also a fresh nonresponse bias to the panel.
- Further results from other countries may be necessary!

Rubin's likelihood approach

- Distribution of interest: $f(Y|\theta) = f(Y_{obs}, Y_{mis}|\theta)$
- Joint distribution of Y and R: $f(Y, R|\theta, \psi) = f(Y|\theta) f(R|Y, \psi)$
- Likelihood of observed data:

$$f(Y_{obs}, R|\theta, \psi) = \int f(Y_{obs}, Y_{mis}|\theta) f(R|Y_{obs}, Y_{mis}, \psi) dY_{mis}$$

- **Missing at random (MAR):** $f(R|Y_{obs}, Y_{mis}, \psi) = f(R|Y_{obs}, \psi)$
- Under MAR:

$$\begin{aligned} f(Y_{obs}, R|\theta, \psi) &= f(R|Y_{obs}, \psi) \int f(Y_{obs}, Y_{mis}|\theta) dY_{mis} \\ &= f(R|Y_{obs}, \psi) f(Y_{obs}|\theta) \end{aligned}$$

MAR in regression analysis (1/4)

- Covariates X_i **always** observed, Y_i observed with missings indicated by $R_i = 1$.
- Model of interest: $Y_i = \beta' X_i + \epsilon_i$
- MAR holds if: $P(R_i = 1 | Y_i, X_i) = P(R_i = 1 | X_i)$
- The selection of units is a simple random sample within the strata formed by the covariates of the model.
- Relationship to conditional independence: MAR $\Leftrightarrow R \otimes Y | X$
- The MAR condition cannot be tested from the observed data!
- OLS on the basis of the complete units is consistent.

MAR in regression analysis (2/4)

- Covariates X_i and $Y_{i,t=1}$ always observed, $Y_{i,t=2}$ observed with missings indicated by $R_i = 1$.
- Model of interest: $Y_{i,t=2} = \beta'_{t=2} X_i + \epsilon_{i,t=2}$
- R_i depends on $Y_{i,t=1}$, for example by stochastic censoring model: $R_i^* = \gamma'_0 + \gamma'_1 Y_{i,t=1} + \delta_i$ and $R_i = 1$ if $R_i^* > 0$
- Note that $Y_{i,t=1}$ does not enter the likelihood for $\beta_{t=2}$, the model of interest! In order to factorize the likelihood, one has to assume:

$$P(R_i = 1 | Y_{i,t=2}, X_i) = P(R_i = 1 | X_i)$$

  This does not hold unless $\epsilon_{i,t=1} \equiv 0$:

$$R_i^* = \gamma'_0 + \gamma'_1 (\beta'_{t=1} X_{i,t=1}) + \delta_i$$

- "Missing on observables" (MO, Fitzgerald et al. 1998): $P(R_i = 1 | Y_{i,t=2}, Y_{i,t=1}, X_i) = P(R_i = 1 | Y_{i,t=1})$

MAR in regression analysis (3/4)

2–wave panel (Continued)

- Controversy: All observed variables should be included in the likelihood (Rubin):

$$f(Y_{t=2}|X) = \int f(Y_{t=2}|Y_{t=1}, X) f(Y_{t=1}|X) dY_{t=1}$$

  Then MO=MAR.
  Note that we need note formulate a model for the response!
- However, the above model equation does not look like a simple regression model.
- One has to formulate two models one is not interested in!
  This is the consequence of Rubin's approach to formulate a likelihood of **all** observed variables

MAR in regression analysis (4/4)

- Multiple Imputation (MI):
  - Estimate the distribution $f(Y_{t=2}|Y_{t=1}, X)$ from the observed wave 2 data.
  - For each unit i with value of $y_{t=2}$ generate M imputations according $f(Y_{t=2}|Y_{t=1}, X)$
  - Regress the $y_{t=2}$-values (imputed and observed) on X. For each version of the imputed values one obtains an estimate $\hat{\beta}_{t=2}^{(m)} \quad (m = 1, \ldots, M)$.
  - The MI-estimate of $\beta_{t=2}$ is the mean of the $\hat{\beta}_{t=2}^{(m)}$.
  - The multiple replication serves as a means the compute the correct variance of the estimate. Let V_m the variance of $\hat{\beta}_{t=2}^{(m)}$ and compute the between variance B of the $\hat{\beta}_{t=2}^{(m)}$ as:

$$B = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\beta}_{t=2}^{(m)} - \bar{\beta}_{t=2})^2$$

Then the variance of $\bar{\beta}_{t=2}$ can be estimated by:

$$V(\bar{\beta}_{t=2}) = \frac{1}{M} \sum V_m + B$$

NMAR missing data pattern in a $A \times B \times R$ contingency table (1/5)

Analysis of transitions between the labour force states: Employed (E), Unemployed (U), Not in the labour force (N).
Empirical analysis for the German MC: does not cover residential mobility!
Hypothesis: getting into employment may cause residential mobility.

- A labour force state at time 1, B labour force state at time 1,
- Quantity of interest: $P(B|A)$
- A always observed, B observed for residential stayers $R = 1$

$$P(R = 1|B, A) = \begin{cases} P(R = 1|A) & \text{MAR;} \\ P(R = 1|B) & \text{Restricted NMAR;} \\ P(R = 1|A, B, A * B) & \text{Unrestricted NMAR.} \end{cases}$$

Comparison MC panel and SOEP

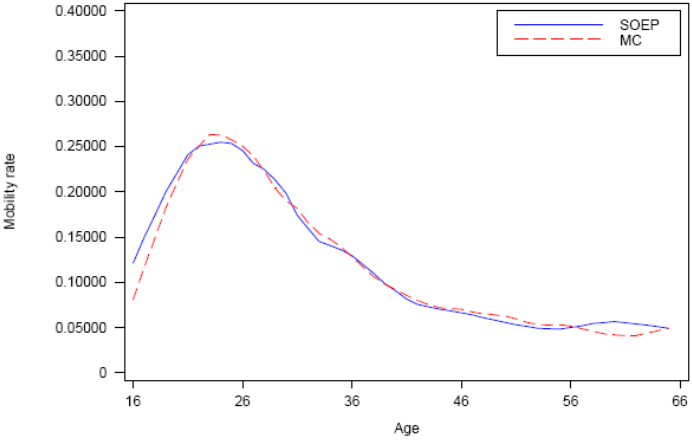
Table 1: **The cumulative extent of residential mobility in the MC and the SOEP. Percentage and cases of individuals with residential mobility after 1996.**

Sample	Transition					
	1996-1997		1996-1998		1996-1999	
	%	cases	%	cases	%	cases
MC	11.13	12594	19.30	21719	25.87	28968
SOEP	10.51	1520	20.23	2836	26.64	3524
SOEP*	9.94		19.62		26.15	

Data base: MC, SOEP = unweighted results, SOEP* = design-based results using the design weights and the attrition correction factors

Comparison MC panel and SOEP

Figure 1: **Mobility rates from 1996 to 1997 calculated from the SOEP and the MC. Rates computed from a scatter plot smoother (cubic spline interpolation) according to SAS procedure LOESS.**



Comparison MC panel and SOEP

Table 4: **Bias estimates for flows between labour force states (unweighted results). Δ = estimate of absolute Bias. Boldface figures: Significant differences $\hat{P}_{ALL} - \hat{P}_{IMMO}$**

Flows from 96 to		E FULL	E IMMO	E Δ	U FULL	U IMMO	U Δ	N FULL	N IMMO	N Δ
	97	91.02	91.16	-0.14	4.92	4.86	0.06	4.05	3.97	0.08
E	98	87.82	88.03	-0.21	6.32	6.04	**0.28**	5.86	5.93	-0.07
	99	87.01	86.37	**0.64**	6.04	6.30	-0.26	6.96	7.33	-0.37
	97	32.83	30.85	**1.98**	48.39	49.83	**-1.44**	18.78	19.32	-0.54
U	98	34.92	31.79	**3.13**	40.13	41.20	-1.07	24.95	27.01	-2.06
	99	41.37	37.46	**3.91**	28.91	29.10	-0.19	29.71	33.44	-3.73
	97	12.74	11.64	**1.10**	5.48	4.97	0.51	81.77	83.39	-1.62
N	98	19.66	16.07	**3.59**	5.09	4.40	**0.69**	75.25	79.54	-4.29
	99	25.89	21.13	**4.76**	4.53	3.71	**0.82**	69.58	75.15	-5.57

Source: Authors' calculations, Data base: SOEP, Waves: 1996-1999

Comparison MC panel and SOEP

Table 2: **Probability of residential immobility over the period 1996-1997 (GEE analysis with household clusters)**

Variable		MC	SOEP	Diff.
Intercept		**1.6114**	**2.0190**	**-0.4076**
		(0.0596)	(0.1265)	(0.1399)
Age ≤ 30		**-0.5132**	**-0.5354**	0.0221
		(0.0281)	(0.0657)	(0.0714)
Age > 45		**0.7191**	**0.5402**	0.1788
		(0.0338)	(0.0880)	(0.0943)
Household size	1 person	**-0.5452**	**-0.4675**	-0.0776
		(0.0388)	(0.1122)	(0.1187)
	2 persons	**-0.1379**	-0.0726	-0.0653
		(0.0377)	(0.0945)	(0.1017)
Sex	male	**0.0337**	0.0024	0.0313
		(0.0133)	(0.0308)	(0.0335)
Region	East-Germany	-0.0196	-0.0800	0.0684
		(0.0350)	(0.0879)	(0.0946)
Education	vocational	0.0251	0.1075	-0.0824
		(0.0255)	(0.0561)	(0.0616)
	tertiary level	**-0.1561**	**-0.1987**	0.0427
		(0.0290)	(0.0688)	(0.0746)
Nationality	German	**0.5114**	0.1570	**0.3543**
		(0.0415)	(0.0845)	(0.0941)
Marital Status	Married	**0.3265**	**0.3389**	-0.0124
		(0.0352)	(0.0796)	(0.0870)
Labour Force Status	Employment	**-0.1621**	**-0.1251**	-0.0371
		(0.0239)	(0.0554)	(0.0603)
	Unemployment	**-0.2631**	-0.0437	**-0.2194**
		(0.0394)	(0.0870)	(0.0955)
Observations (Individuals)		76'835	11'955	
Log Likelihood		-24'876	-3'918	
Pseudo R^2		0.1166	0.2366	

Dependent Variable: indicator of mobility

coefficients for logarithm of odds ratio $\mathrm{P}(R=1)/\mathrm{P}(R=0)$

Standard deviations in paranthesis

Comparison MC panel and SOEP

Table 7: **Alternative models for residential immobility in the SOEP. Period 1996-1997. GEE analysis with household clusters. Model 2: recent and current labour force states included (Main effects). Model 3+4: Model 2 + different indicators for transitions**

Variable		Model 2	Model 3	Model 4
Intercept		**2.1183**	**2.1674**	**2.1961**
		(0.1310)	(0.1326)	(0.1325)
Age $\leq$ 30		**-0.4658**	**-0.4532**	**-0.4495**
		(0.0685)	(0.0685)	(0.0683)
Age $>$ 45		**0.4927**	**0.4918**	**0.4918**
		(0.0741)	(0.0739)	(0.0740)
Household size	1 person	**-0.4211**	**-0.4334**	**-0.4311**
		(0.1213)	(0.1213)	(0.1212)
	2 persons	-0.0396	-0.0452	-0.0438
		(0.1016)	(0.1016)	(0.1013)
Sex	male	-0.0051	-0.0045	-0.0034
		(0.0302)	(0.0302)	(0.0301)
Region	East-Germany	-0.0944	-0.0990	-0.0978
		(0.0931)	(0.0934)	(0.0932)
Education	vocational	**0.1406**	**0.1419**	**0.1440**
		(0.0569)	(0.0570)	(0.0568)
	tertiary level	**-0.1698**	**-0.1733**	**-0.1762**
		(0.0708)	(0.0703)	(0.0703)
Nationality	German	0.1445	0.1540	0.1549
		(0.0841)	(0.0842)	(0.0842)
Marital Status	Married	**0.3242**	**0.3197**	**0.3250**
		(0.0852)	(0.0853)	(0.0850)
Labour Force Status	Employment (96)	0.0018	-0.0061	0.1638
		(0.0862)	(0.1897)	(0.1963)
	Unemployment (96)	0.0152	0.0465	0.2527
		(0.1074)	(0.1318)	(0.1410)
	Employment (97)	**-0.1811**	0.1154	0.1076
		(0.0856)	(0.1145)	(0.0982)
	Unemployment (97)	-0.1591	0.1063	0.1211
		(0.1020)	(0.1488)	(0.1353)
	Δ_{middle}		-0.2670	**-0.5229**
			(0.2307)	(0.2355)
	Δ_{high}		**-0.4856**	**-0.6008**
			(0.1578)	(0.1464)
Observations (Individuals)		11'955	11'156	11'156

NMAR missing data pattern in a $A \times B \times R$ contingency table (2/5)

	$R=1$			R=0
	B			
A	E	U	N	
E	$n(EE)$	$n(EU)$	$n(EN)$	$n(E.)$
U	$n(UE)$	$n(UU)$	$n(UN)$	$n(U.)$
N	$n(NE)$	$n(NU)$	$n(NN)$	$n(N.)$

The likelihood:

$$L = \prod_{i \in R=1} P(A,B)P(R=1|A,B) \times \prod_{i \in R=0} \sum_{B} P(A,B)P(R=0|A,B) \quad (1)$$

NMAR missing data pattern in a $A \times B \times R$ contingency table (3/5)

- A standard NMAR model: Mobility depends on the last wave labour force state B

$$(P(R|A,B) = P(R|B)) = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}$$

- Restrictions taken from the SOEP:

$$(P(R=1|A=a,B=b)) = \begin{pmatrix} m & m & m \\ h & m & l \text{ or } m \\ h & h & l \end{pmatrix}$$

- Observed cells: $3 \times 3 + 3 = 12$, Unrestricted model parameters: $9(R|A,B) + 6(B|A) + 2(A) = 17$, Model restrictions 6 ($+$ 1 for size of sample in Loglinear Model), DF$=$0.

Bias correction (relative bias)

$$B_{rel} = \frac{\hat{P}_{\mathrm{COR,SAMPLE}}(B|A) - \hat{P}_{\mathrm{IMMO,SAMPLE}}(B|A)}{\hat{P}_{\mathrm{ALL,SOEP}}(B|A) - \hat{P}_{\mathrm{IMMO,SOEP}}(B|A)}$$

(1 $\Rightarrow$ perfect correction, $< 0 \Rightarrow$ "correction" in the wrong direction, $> 1 \Rightarrow$ correction beyond the correct value)

t	Bias	(Bias correction)/Bias							
		SOEP			MC			SOEP	MC
		Mod1	Mod2	NMAR	Mod1	Mod2	NMAR	Mod3	Mod3
		Transition $U \rightarrow E$							
1997	1.98	0.62	0.20	-0.09	1.80	1.30	-0.37	0.49	0.46
1998	3.13	0.89	0.15	-0.07	1.78	1.09	-0.59	0.54	0.64
1999	3.91	1.03	0.07	-0.05	1.68	0.94	-0.61	0.87	0.69
		Transition $N \rightarrow E$							
1997	1.10	0.82	0.52	0.32	2.26	1.78	0.21	0.55	0.52
1998	3.59	0.68	0.33	0.27	1.29	0.93	0.08	0.41	0.48
1999	4.56	0.85	0.35	0.28	1.26	0.89	0.01	0.67	0.69

- Minor changes of NMAR nonresponse model can have dramatic consequences for the bias reduction.
- The standard NMAR model may even "correct" into the wrong direction or not even indicate a bias.
- A standard weighting approach (Mod3) performs reasonably well (see later).

The variances of different NMAR estimates

96 to	$U \to E$ ALL	IMMO	alt_1	alt_2	alt_3	$N \to E$ ALL	IMMO	alt_1	alt_2
97	32.84	30.85	32.08	31.24	30.68	12.74	11.64	12.54	12.21
	(1.49)	(1.55)	(1.83)	(1.78)	(1.60)	(0.68)	(0.68)	(0.93)	(0.87)
98	34.92	31.79	34.55	32.26	31.57	19.66	16.07	18.48	17.26
	(1.57)	(1.72)	(2.41)	(2.16)	(1.86)	(0.86)	(0.87)	(1.71)	(1.39)
99	41.37	37.46	41.74	37.74	37.25	25.89	21.13	25.19	22.78
	(1.66)	(1.94)	(2.46)	(2.45)	(2.24)	(1.00)	(1.06)	(2.54)	(1.84)

alt_1 : transitions $U \to N$ attributed to the low mobility group

alt_2 : transitions $U \to N$ attributed to the mean mobility

alt_3 : Main effect model for B

Standard deviations in parenthesis.

NMAR missing data pattern in a $A \times B \times R$ contingency table (4/5)

- Despite more data plus identifying restrictions twice as high standard errors of estimates!
- $\Rightarrow$ Flat likelihood!
- Often substantive over-corrections!
- Easy estimation with LEM Package (Freeware)

LEM: A useful program

LEM stands for: **L**oglinear and event history analysis with missing data using the **EM** algorithm.

Free download + documentation from:
http://www.uvt.nl/faculteiten/fsw/
organisatie/departementen/mto/software2.html

LEM: Example 1 with SOEP data

$$P(R|A,B) = P(R|B)$$

LEM for Windows

File Edit Tools Window Examples

Log

Output

Input - Example_1.inp

```
res 1                    * No. response variables
man 2                    * No. of manifest variables
dim 2 3 3                * No. of values of resp. + manifest vars
lab R A B                * Labels of resp. manifest vars.
sub AB A                 * Observed tables
mod A B|A  {AB} R|AB {RB} *  Models for tables. Here: R depends only on B
dat [4221 308 358 233 181 208 313 55 1113 * Table AB
    2278 294 558]                         * Table A
```

LEM: Example 2 with SOEP data

Medium mobility group: $A = 1(e)$ and $B = 1, 2, 3(e, u, n)$ and $A = 2(u), B = 2(u)$

High mobility group: $A = 2(u), B = 1(e)$ and $A = 3(n), B = 1, 2(e, u)$

Low mobility group: $A = 2(u), B = 3(n)$ and $A = 3(n), B = 3(n)$

Input - Example_2.inp

```
res 1
man 2
dim 2 3 3
lab R A B
sub AB A
mod A B|A
R|AB {fac(ABR,3)}       * 3 Restrictions for the ABR table
des [0 0 0 0 0 0 0 0 0  * No restrictions for the AB table
     1 1 1 2 1 3 2 2 3] * Parameters with 1 set to be equal, 2 and 3 similar
dat [4221 308 358 233 181 208 313 55 1113
     2278 294 558]
```

Control by age-groups

- Age turned out to be the most important variable for regional mobility
- Control for age by using a break down of tables with respect to age-group

Transition	$U \rightarrow E$			$N \rightarrow E$		
	ALL	IMMO	Δ	ALL	IMMO	Δ
	Age≤30					
97	52.43	52.12	0.31	25.98	24.16	**1.82**
98	55.09	56.02	0.93	37.86	33.33	**4.53**
99	65.69	64.05	1.64	50.07	46.28	**3.79**
	Age>30					
97	24.02	22.04	**1.98**	6.36	6.13	**0.23**
98	25.90	23.25	**2.75**	10.13	8.81	1.32*
99	30.28	28.78	1.50	12.72	11.28	**1.44**
	Total					
97	32.84	30.85	**1.99**	12.74	11.64	**1.10**
98	34.92	31.79	**3.13**	19.66	16.07	**3.59**
99	41.37	37.46	**3.89**	25.89	21.13	**4.76**

Δ = estimate of absolute Bias

Boldface figures: Significant differences $\hat{P}_{ALL} - \hat{P}_{IMMO}$

* indicates: the Hausman test did not apply because of negative difference of variances

Weighting by inverse response probabilities (1/3)

- There are often more observable variables for the explanation of nonresponse than in the model of interest.
  - Y_i outcome variable of interest, for example whether a change $E \Rightarrow U$ occurs or not.
  - X_i a set of covariates to explain $P(Y_i = 1|X_i)$.
  - Z_i a set of covariates to explain $P(R_i = 1|Z_i)$. Some covariates of X_i may also belong to Z_i.
  - Missing on observables is needed: $P(R_i = 1|Y_i, X_i, Z_i) = P(R_i = 1|Z_i)$
- Idea: Weight observations with R_i with $\pi_i = 1/P(R_i = 1)$ in the score equation!

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln l_i(\theta) = 0$$

$$\Rightarrow \sum_{i=1}^{n} \frac{R_i}{\pi_i} \frac{\partial}{\partial\theta} \ln l_i(\theta) = 0$$

Weighting by inverse response probabilities (2/3)

- Example Transition of labour states explained by a Logit model. Missingness due to residential mobility (NMAR!). Evaluation data from the SOEP.

$$\ln \frac{P(Y_i = 1|X_i)}{1 - P(Y_i = 1|X_i)} = \beta' X_i$$

Score equation for the Logit model:

$$U_\beta = \sum_i X_i \left(Y_i - P(Y_i = 1|X_i)\right) = \sum_i X_i \left(Y_i - \mu_i\right)$$

The weighted score equation is:

$$U_\beta(\pi) = \sum_i \frac{R_i}{\pi_i} X_i \left(Y_i - \mu_i\right)$$

Weighting by inverse response probabilities (3/3)

$$E_{R,Y|X,Z}\left[\sum_i \frac{R_i}{\pi_i(Z_i)} X_i\,(Y_i-\mu_i)\right]$$

$$=E_{Y|X,Z}\left[\sum_i E_{R|Y,X,Z}\left(\frac{R_i}{\pi_i(Z_i)} X_i\,(Y_i-\mu_i)\middle|\, Y_i, Z_i, X_i\right)\right]$$

$$=E_{Y|X,Z}\left[\sum_i X_i\,(Y_i-\mu_i)\,\frac{1}{\pi_i(Z_i)} E_{R|Y,X,Z}\,(R_i|\,Y_i, Z_i, X_i)\right]$$

$$=E_{Y|X,Z}\left[\sum_i X_i\,(Y_i-\mu_i)\,\frac{P(R_i=1|\,Y_i, Z_i, X_i)}{\pi_i(Z_i)}\right]$$

$$=E_{Y|X,Z}\left[\sum_i X_i\,(Y_i-\mu_i)\right] \quad \text{(original score equation!)}$$

Bias reduction of Inverse Probability Weighting (IPW)

$$IR = \frac{\hat{P}_{\text{IPW,MC}}(B|A) - \hat{P}_{\text{IMMO,MC}}(B|A)}{\hat{P}_{\text{FULL,SOEP}}(B|A) - \hat{P}_{\text{IMMO,SOEP}}(B|A)}$$

Table 8: **Bias reduction expressed by ratio (bias –correction)/bias (SOEP and MC data)**

t	Bias	(Bias–correction)/Bias		
		SOEP	MC	MC*
		$U \to E$		
1997	1.98	0.49	0.46	0.59
1998	3.13	0.54	0.64	0.80
1999	3.91	0.87	0.69	0.80
		$N \to E$		
1997	1.10	0.55	0.52	1.00
1998	3.59	0.41	0.48	0.70
1999	4.56	0.67	0.69	0.97

Pattern Mixture models (1/5)

1. Different factorizations:
   - $P(R, Y, X) = P(R|Y, X) \times P(Y|X) \times P(X)$
   - $P(R, Y, X) = P(Y|X, R) \times P(X|R) \times P(R)$
2. Pattern mixture models assume that the relationship between Y and X is different for responders and non-responders. The sample before nonresponse is a mixture. Nonresponse acts like a segregation of the two populations.
3. However only one part of the mixture is observed! Therefore identification restrictions are necessary.

Pattern Mixture models (2/5)

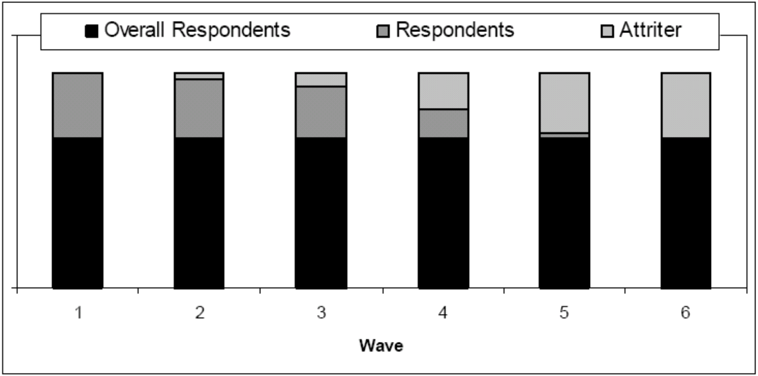
The MAR condition and pattern mixture models:

$$
\begin{aligned}
f(y|x,r) &= \frac{f(y,x,r)}{f(x,r)} \\
&= \frac{f(r|y,x)f(y,x)}{f(x,r)} \\
&= \frac{f(r|x)f(y,x)}{f(x,r)} \\
&= \frac{f(y,x)}{f(x)} \\
&= f(y|x)
\end{aligned}
$$

Pattern Mixture models (3/5)

A useful routine in panel analysis: Subdivide the wave-1 respondents according attrition in later waves:

Figure 12: **The division of a panel according to future attrition.**



Pattern Mixture models (4/5)

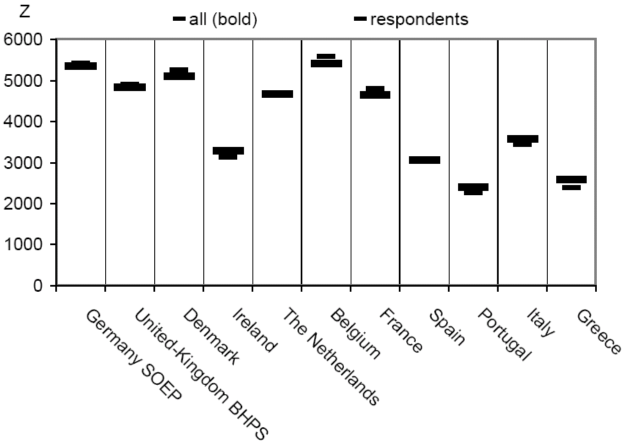
- The idea is that attrition acts like a segregation of the wave-one respondents.
- Compare the estimation results for the FULL first wave sample with the results for the permanent responders.
- H_0 states that conditioning on R is irrelevant.
- Under H_0 the restriction to the subsample of permanent responders affects only the efficiency of the model estimate.
  If the estimator on the basis of the full sample is efficient, one may apply the Hausman test for the difference of the full and the restricted sample.
- If H_0 is rejected, one would conjecture that attrition is de-mixing also in future waves.

Pattern Mixture models (5/5)

- Some results for the ECHP User Data Base (UDB): Period 1994 – 1999 (6 waves)
- Does panel attrition disturb comparative analysis, for example, the ranking of the member states?
- Details in: Behr et al. (2003): Comparing poverty, income inequality and mobility under panel attrition. A cross country comparison based on the European Community Household Panel. CHINTEX Working Paper No.12, URL: www.destatis.de/chintex/download/paper12.pdf

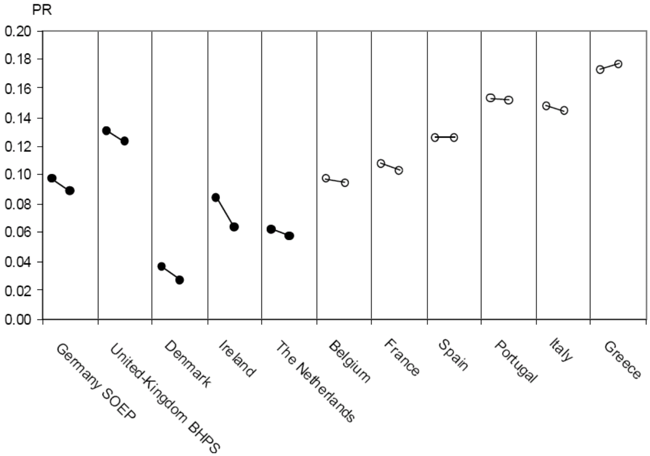
Testing the poverty line

Figure 13: **Comparison of poverty lines in the ECHP.**



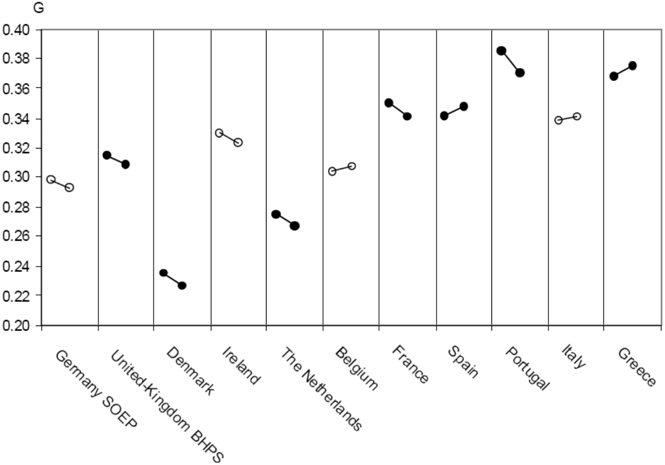
Testing the poverty rate

Figure 14: **Poverty rates and significance of the attrition bias**



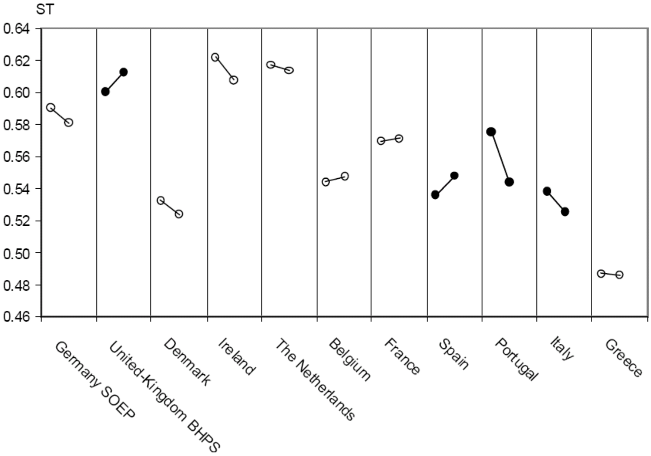
Testing the Gini coefficient

Figure 15: **Comparison of Gini-Coefficients**



Testing the proportion of stayers in income position

Figure 17: The proportion of stayers in the same income quintile.



Stability of rank position

Measure	Rank Correlation
Poverty Rate	0.99
Average Poverty Gap	0.95
Gini	0.98
SST-Index	0.98
Stayer	0.88
Average Rang Difference	0.96
Rank Correlation	0.98
Ratio Ups/Downs	0.93

Table 15: **The correlation of the rank position of the 11 countries for different measures of poverty and income stability**

The rule of imputation

- Inverse Probability Approach: Find a good model for R_i. Use only the weighted **complete cases**.
- Now: Find a good prediction for the missing values without formulating a model for response (MAR)!
- Analyse the **full** sample with the imputed values!

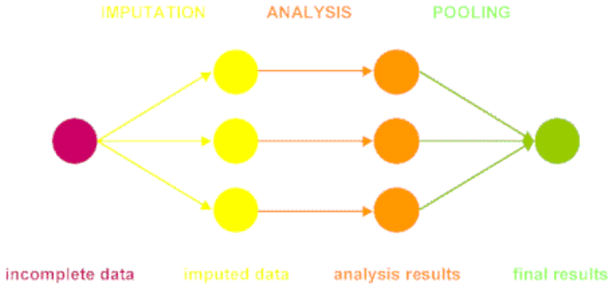
Naive imputation in panels

In panel surveys there are some naive approaches for imputation:

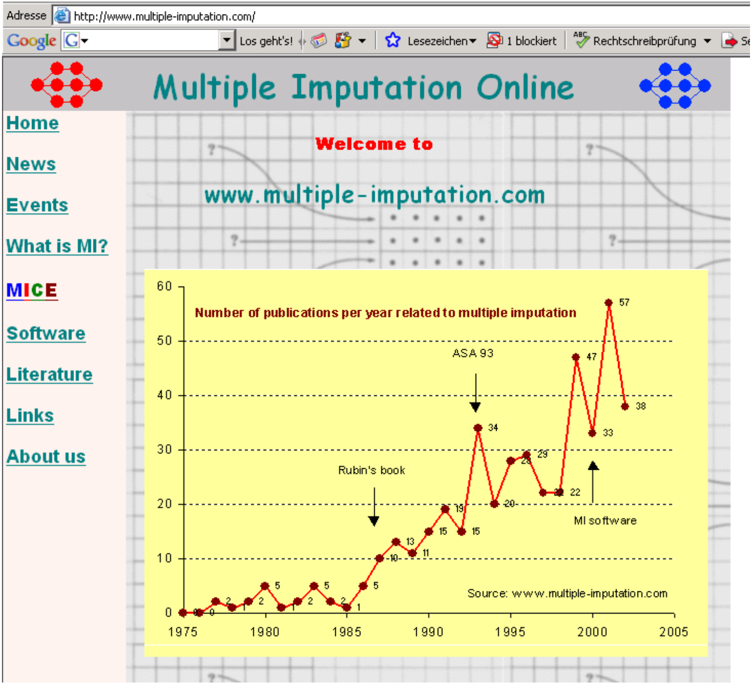
- Mean of observed values ($\Rightarrow$ biased level)
- Conditional mean of observed values ($\Rightarrow$ biased variance)
- Carry forward last observation ($\Rightarrow$ biased serial correlation)
- Conditional mean plus error (Single imputation) ($\Rightarrow$ biased inference)

Solution: Multiple Imputation!

Multiple imputation is a statistical technique for analyzing incomplete data sets, that is, data sets for which some entries are missing. Application of the technique requires three steps: *imputation*, *analysis* and *pooling*. The figure illustrates these steps.



IMPUTATION	Impute (=fill in) the missing entries of the incomplete data sets, not once, but *m* times (*m*=3 in the figure). Imputed values are drawn for a distribution (that can be different for each missing entry). This step results is *m* complete data sets.
ANALYSIS	Analyze each of the *m* completed data sets. This step results in *m* analyses.
POOLING	Integrate the *m* analysis results into a final result. Simple rules exist for combining



Adresse
http://www.multiple-imputation.com/
Google
Los geht's!
Lesezeichen
1 blockiert
Rechtschreibprüfung
Multiple Imputation Online
Home
News
Events
What is MI?
MICE
Software
Literature
Links
About us
Welcome to
www.multiple-imputation.com
Number of publications per year related to multiple imputation
ASA 93
Rubin's book
MI software
Source: www.multiple-imputation.com
0
10
20
30
40
50
60
1975
1980
1985
1990
1995
2000
2005

Multiple Imputation (General 1/3)

Likelihood + Prior: $E(\theta|Y_{OBS}$ Expected posterior value

Complete data posterior:
$p(\theta|Y_{obs}, Y_{Mis}) \propto p(\theta)L(\theta|Y_{obs}, Y_{Mis})$

Link observed and complete data posterior:

$$\begin{aligned} p(\theta|Y_{OBS}) &= \int P(\theta, Y_{MIS}|Y_{OBS})dY_{MIS} \\ &= \int p(\theta|Y_{MIS}, Y_{OBS})P(Y_{MIS}|Y_{OBS})dY_{MIS} \end{aligned}$$

$E(\theta|Y_{OBS}) = E[E(\theta|Y_{MIS}, Y_{OBS})|Y_{OBS}]$
$Var(\theta|Y_{OBS}) =$
$E[Var(\theta|Y_{MIS}, Y_{OBS})|Y_{OBS}] + Var[E(\theta|Y_{MIS}, Y_{OBS})|Y_{OBS}]$

Multiple Imputation (General 2/3)

Generate M independent draws:
$Y_{MIS}^{(m)} \sim p(Y_{MIS}|Y_{OBS}) \quad m = 1, \ldots, M$

Estimate $E(\theta|Y_{MIS}^{(m)}, Y_{OBS})$ by $\hat{\theta}_{(m)}$

Estimate $E(\theta|Y_{OBS})$ by $\hat{\theta} = 1/M \sum_{m=1}^{M} \hat{\theta}_{(m)}$

Estimate $Var(\hat{\theta}|Y_{OBS})$ by:

$$Var(\hat{\theta}|Y_{OBS}) \approx \frac{1}{M} \sum_{m=1}^{M} V_m + \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_{(m)} - \hat{\theta})^2 = \bar{V} + B$$

where V_m is the complete data posterior Variance of θ calculated for the m^{th} complete data set

An improved Variance estimation is:
$Var(\theta|Y_{OBS}) \approx \bar{V} + (1 + M)^{-1} B$

Multiple Imputation (General 3/3)

- Generation of $Y_{MIS}^{(m)} \sim p(Y_{MIS}|Y_{OBS})$ may be difficult! Use Markov Chain Monte Carlo -technique!
- Selection of a "non-informative prior"!
- Case of multidimensional normal data: Package NORM or SAS routine PROC MI
- SAS: Proc MIANAYSE computes the correct standard errors.
- Problem: The imputer and the analyst use different models!
- Recommendation: The imputer's model should the contain the model of the analyst.
- Automatic sequential procedure MICE (Multiple Imputation Conditional Expectation). See also Ragunathan's IVEware Package (Imputation and Variance estimation)
- Up to now: No special approach or program for panels. Prediction of level or change, serial correlation! MICE etc. use level models.

Literature on Multiple Imputation

- Schafer, J. (1997). Analysis of Incomplete Multivariate Data. New York: Chapman and Hall.
- Little/ Rubin (2002): Statistical Analysis with Missing Data, Second Edition, Chapter 10, Wiley
- Allison, P. (2002): Missing Data, Sage
- General information on MI: www.multiple-imputation.com
- Meng (1994): Multiple imputation inferences with uncongenial sources of input (with discussion) Stat. Science, 10, 538–573.

Model: Stochastic censoring destroys the Normal distribution of the variable of interest. By value of the model one can make conclusions about the Normal distribution.

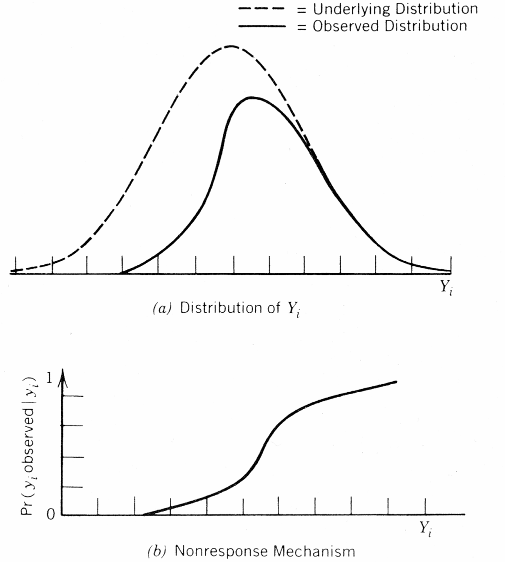


Figure 15.1. The normal stochastic censoring model for a single sample.

Sample selection models (1/3)

- Regression analysis: X, V always observed, dependent variable Y is observed if $R = 1$:

$$\begin{aligned} Y &= \beta' X + \epsilon \quad \text{observed, if } R^* > 0 \text{ where} \\ R^* &= \gamma_1' X + \gamma_2' V + \delta \end{aligned}$$

- Normality assumption for residual terms

$$\begin{pmatrix} \epsilon \\ \delta \end{pmatrix} = \mathrm{N}\left(0, \begin{pmatrix} \sigma_\epsilon^2 & \rho\sigma_\epsilon \\ \rho\sigma_\epsilon & 1 \end{pmatrix}\right)$$

- Correction for the expected value:

$$E(Y|X, V, R = 1) = \beta' X + \rho\sigma_\epsilon \frac{\phi(\gamma_1' X + \gamma_2' V)}{\Phi(\gamma_1' X + \gamma_2' V)}$$

Sample selection models

- ϕ/Φ (inverse Mill's ratio) is an almost linear function.
- Without V in R-eq. β is only identified from the non-linearity of ϕ/Φ.
- The instrument variable V must have no impact on distribution of $Y|X$ but must have an impact on the distribution $R|X$. The assumption of an **a-priori zero coefficient** of V is crucial for the model!
- MAR is equivalent to $\rho_{\sigma\epsilon} = 0$
- If the a-priori zero assumption is wrong, severe biases and wrong std. errors may occur, Rendtel (1992). See also the critique of the approach in Little/Rubin (2002, Chapter 15).

The Heckman 2-stage estimator

- Response equation for wave 2 (or later), regression equation for wave 2 (or later), joint Normality for ϵ and δ. Covariates X and V from wave 1 and always known.
- Estimation by ML or Heckman's two step procedure:
  1. Estimate response Probit on the basis of wave 1 and calculate from $\hat{\gamma}$ the estimated Mill's ratio $\hat{H} = H(\hat{\gamma}_1' X + \hat{\gamma}_2' V)$
  2. Estimate the regression equation for all wave 2 respondents with $\hat{H}$ as an augmented variable. Use OLS.
- Use of the model for multi-period panels by collapsing attrition intervals. X may become poor indicators for attrition.

Ridder's (1990) multi-period extension of the sample selection model

$$Y_{it} = \beta' X_{i,t} + \mu_i + \nu_{i,t} \qquad 1 \leq t \leq T$$
$$R^*_{it} = \gamma_1' X_{i,t} + \gamma_2' V_{i,t} + \xi_i + \eta_{it} \qquad 1 \leq t \leq T$$

- ML estimation in Verbeek/Nijman (1992): For each person evaluation of a twofold integral!
- No such simple procedure as the Heckman procedure
- Missing covariates may occur in case of time-dependent covariates. Model is not suited for such a case.

General concepts

- Response set $r \subset s$; Response indicator $R_k = 1$ if unit $k \in r$.
- Interpretation: r is sampled from s via Poisson sampling with selection probabilities θ_k for unit k.
  Response is independent across units and response probabilities between units.
- Response Homogeneity Groups (RHG): Population is divided into G response homogeneity groups $U_1, \ldots, U_g, \ldots, U_G$.
  Within group g the response probability is estimated by $\theta_k = m_g / n_g$ for $k \in U_g$
  where $n_g =$ number of $s \bigcap U_g$ and $m_g =$ number of $r \bigcap U_g$
- The corrected $\pi-$estimator is defined by:

$$\hat{t}_{\pi^*} = \sum\nolimits_U \frac{R_k I_k}{\theta_k \pi_k} y_k = \sum\nolimits_r \frac{1}{\theta_k \pi_k} y_k$$

General concepts

- Properties of $\hat{t}_{\pi^*}$ in the framework of 2-stage sampling.
  - Realisation of random sample s according to design.
  - Realisation of Poisson sampling r from s.
- Bias estimation:

$$B(\hat{t}_{\pi^*}) = E_D[E_R(\hat{T}_y|s)] - \sum_U y_k$$

- If the correct response probabilities are used, $B(\hat{t}_{\pi^*}) = 0$
  Important note: Under nonresponse the design-based approach has lost its ability to produce unbiased estimates independent from a statistical model!
- Bethlehem (200x) has derived the following Bias approximation $\tilde{B}$, see also Lundström/Sarndal (2005,pp 106–108) :

$$\tilde{B} = -\sum_U (1 - \theta_k) y_k$$

  $\tilde{B}$ can be interpreted as a population covariance of the response probabilities θ_k and y_k.

Calibration levels under Nonresponse (1/2)

So far calibration has been a tool for variance reduction. In the case of nonresponse it can be also a tool for bias reduction. Form of corrected weights $w_k = d_k g_k$

- A1 Calibration to sample: $\sum_r g_k x_k = \sum_s x_k$
- A2 Calibration to population estimates: $\sum_r d_k g_k x_k = \sum_s d_k x_k$
- A3 Calibration to population totals: $\sum_r d_k g_k x_k = \sum_U x_k$
- B1 ML-estimation of θ_k: $\sum_r x_k = \sum_s g_k^{-1} x_k$
  $g_k^{-1} = e^{x_k' \hat{\lambda}} / (1 + e^{x_k' \hat{\lambda}}) = \hat{\theta}_k$ with score function of the Logit model for the R_k explained by x_k: $\sum_s (R_k - \hat{\theta}_k) x_k = 0$

Calibration levels under Nonresponse (2/2)

- Functional restriction of $g_k = \hat{\theta}_k^{-1} = f(x_k'\hat{\lambda})$ with f known monotonic real-valued function and $\hat{\lambda}$ chosen to fill calibration constraints.
- Standard calibration: $f(x_k'\hat{\lambda}) = 1 + x_k'\hat{\lambda}$ and $(A3)$
- $f(x_k'\hat{\lambda}) = x_k'\hat{\lambda}$ and $(A1)$ yields: $\hat{\lambda} = (\sum_s x_k x_k')^{-1}(\sum_s x_k - \sum_r x_k)$
- Raking weights: $f(x_k'\hat{\lambda}) = e^{-x_k'\hat{\lambda}}$ and $(A3)$
- The post-stratification estimator is obtained by:
  Population is divided into G response homogeneity groups $U_1, \dots, U_g, \dots, U_G$. $x_k = (I_1(k), \dots, I_G(k))$ indicates for each unit $k \in U$ the membership to the response groups.
  With $f(x_k'\hat{\lambda}) = x_k'\hat{\lambda}$ and $(A1)$ or $(B1)$ one obtains: $g_k = n_g/m_g$ for $k \in U_g$
  where $n_g =$ size of $s \bigcap U_g$ and $m_g =$ size of $r \bigcap U_g$

A general calibration estimator

Lundström/Särndal discuss a general calibration estimator $\hat{T}_{GCAL} = \sum_r w_k y_k$:

$$w_k = d_{\alpha,k} g_k \qquad g_k = 1 + \hat{\lambda}' z_k$$

where:

1. $d_{\alpha,k}$ initial weights, often a general correction of nonresponse by setting $d_{\alpha,k} = (n/m) d_k$
2. z_k vector of instrument variables, often $z_k = x_k$
3. Calibration to U and to population estimates: $\mathbf{X} = (\sum_U x_k^1, \sum_s d_k x_k^2)'$
4. $\hat{\lambda} = (\mathbf{X} - \sum_r d_{\alpha,k} x_k)' (\sum_r z_k x_k')^{-1}$

A bias approximation

The bias of can then be approximated by $\tilde{B}$, see Lundström/Sarndal (2005,pp 106–108) :

$$\tilde{B} = -\sum_U (1-\theta_k) e_{\theta,k}$$

where:

$$e_{\theta,k} = y_k - x_k' \mathbf{B}_{U,\theta} \qquad \mathbf{B}_{U,\theta} = \left(\sum_U \theta_k z_k x_k'\right)^{-1} \sum_U \theta_k z_k y_k$$

- $\tilde{B}$ can be interpreted as a population covariance of the response probabilities θ_k and some regression residuals $e_{\theta,k}$.
- The approximation gets better with increasing size of r.

Conclusions from bias approximation

- Bias is independent from sampling design!
- Whether we calibrate by some vector x up to population or to population estimates, does not affect the size of the bias approximation.
- $\tilde{B} = 0$, if there is some vector λ with:

$$\frac{1}{\theta_k} = \phi_k = 1 + \lambda' z_k \quad \text{for all } k \in U$$

because we then have: $1 - \theta_k = \theta_k \lambda' z_k$ for $k \in U$. Therefore:

$$\sum_U (1 - \theta_k) e_{\theta,k} = \lambda' \sum_U \theta_k z_k (y_k - x_k' \mathbf{B}_{U,\theta}) = 0$$

- $\tilde{B} = 0$, if $y_k = \beta' x_k$ for all $k \in U$. Because then $e_{\theta,k} = 0$ for all $k \in U$

Selection of auxiliary information

See Chapter 10 of Lundström/Särndal (2005) for an extended discussion!

Principle 1 The auxiliary vector should explain the inverse response probability.
Keeps bias small for **all** study variables. May inflate the variance of the weights and hence the variance of the estimates. Often bias is regarded as more important in survey sampling!

Principle 2 The auxiliary vector should explain the main study variables.
Specific weights might be a good idea, although unusual in practice.

Principle 3 The auxiliary vector should identify the most important domains.
Regional stratification often unknown to users.

Further remarks on calibration

- More is better ?.
  But avoid: Negative weights, extreme variation of weights.
- Number of constraints may depend on the sample size of the survey.
- Software: CLAN, CALMAR (SAS based macros, not very user friendly!)
- A variance estimator of the general calibration estimator is given Lundström/Särndal (2005, p.136)

Calibration for nonresponse in a panel

- Initial wave: similar to every cross-sectional survey, but calibrations are transferred to later waves.
- Later waves: the number of possible control variables (at the level of the previous sample) is **very** large!
- Variables like "Change of the interviewer" are probably unrelated to many variables of interest, but a powerful for the prediction of attrition.
- Lagged metric variables are powerful predictors for the current value.
- The process of participation in a panel survey is sequential, wave by wave. Variance formulas for such multi-phase surveys are intractable. Need to variance estimation by other means!
- Lump together different waves: $\Rightarrow$ reduces number of stages. For example, in PSID: attrition after 5 years.

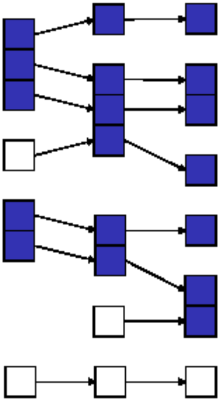
Initial Calibration:

Up to now:

$$\hat{T}_{y^0} = \sum_{i \in s^0} d_i y_i^0$$

$$\hat{T}_{y^t} = \sum_{k \in s^t} w_k y_k^t = \sum_{k \in s^t} y_k^t \sum_{i \in s^0} l_{ik} d_i$$

Further variables x^0 with T_{x^0} known.



Initial Calibration:

Up to now:

$$\hat{T}_{y^0} = \sum_{i \in s^0} d_i y_i^0$$

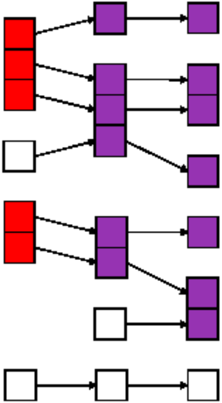
$$\hat{T}_{y^t} = \sum_{k \in s^t} w_k y_k^t = = \sum_{k \in s^t} y_k^t \sum_{i \in s^0} l_{ik} d_i$$

Further variables x^0 with T_{x^0} known.

Modification: $d_i \rightsquigarrow d_i g_i^0$

$$g_i^0 = \text{argmin} \sum_{i \in s^0} (d_i g_i^0 - d_i)^2 / d_i$$

s.t. $T_{x^0} = \sum_{i \in s^0} d_i g_i^0 x_i^0$



Initial Calibration

Properties:

- $\hat{T}^{IC}_{y^0} = \sum_{i \in s^0} d_i g_i^0 y_i^0$
  regular calibration $\rightsquigarrow$ variance estimator known

- $\hat{T}^{IC}_{y^t} = \sum_{k \in s^t} y_k^t \sum_{i \in s^0} l_{ik} d_i g_i^0$

  – variance sources: s^0 , $g_i^0 = 1 + x_i^{0\prime} \hat{\lambda}_0$
  $\hat{\lambda}_0$ via $C(\hat{\lambda}_0) = T_{x^0} - \sum_{i \in s^0} d_i g_i^0 x_i^0 = 0$

  – separation via Taylor: $\hat{T}^{IC}_{y^t}(\hat{\lambda}_0) \approx \hat{T}^{IC}_{y^t}(\lambda_0) + H_1(\hat{\lambda}_0 - \lambda_0)$
  $$0 \approx C(\lambda_0) + H_2(\hat{\lambda}_0 - \lambda_0)$$

  – linearised version : $\hat{T}^{IC}_{y^t} \approx \hat{T}^{IC}_{y^t}(\lambda_0) - H_1 H_2^{-1} C(\lambda_0)$
  depends only on s^0

Future contributions towards y^t and x^t that comes from person $i \in U^0$:

$$\tilde{y}_i^t = \sum_{k \in s^t} l_{ik} y_k^t \qquad \text{and} \qquad \tilde{x}_i^t = \sum_{k \in s^t} l_{ik} x_k^t$$

Redistribution of weights $d_i g_i^0$ for $i \in s^0$ onto the persons $k \in s^t$ according to the follow-up rule: $w_k^C = \sum_{i \in s^0} l_{ik} d_i g_i^0$:

$$\hat{T}_{y^t}^{IC} = \sum_{k \in s^t} y_k^t w_k^C$$

Taylor linearisation leads to:

$$\hat{T}_{y^t}^{IC} = T_{x^0}{}'\hat{B}^{IC} + \sum_{i \in s^0} d_i \tilde{e}_i^{IC}$$

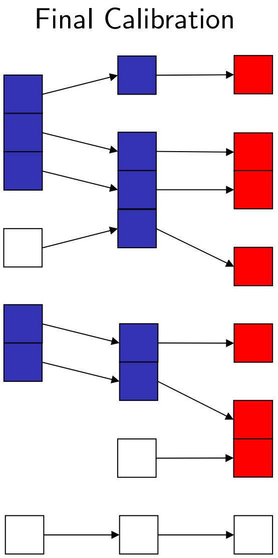
with

$$\begin{aligned} \tilde{e}_i^{IC} &= \tilde{y}_i^t - x_i^{0\prime}\hat{B}^{IC} \\ \hat{B}^{IC} &= (\sum_{i \in s^0} d_i x_i^0 x_i^{0\prime})^{-1}(\sum_{i \in s^0} d_i x_i^0 \tilde{y}_i^t) \end{aligned}$$

Interpretation: regression of the future contributions $\tilde{y}_i^t$ versus the values x_i^0 at wave 0.
Variance estimator:

$$\hat{V}(\hat{T}_{y^t}^{IC}) = \sum_{i \in s^0} \sum_{i' \in s^0} (\pi_i^{-1}\pi_{i'}^{-1} - \pi_{ii'}^{-1}) g_i^0 g_{i'}^0 \tilde{e}_i^{IC} \tilde{e}_{i'}^{IC}$$

Final Calibration



g_k^t: $\sum_{k \in s^t} w_k g_k^t y_k^t = T_{x^t}$

$$\hat{T}_{y^t}^{FC} = \sum_{k \in s^t} w_k g_k^t y_k^t$$

Taylor linearisation leads to:

$$\hat{T}_{y^t}^{FC} = T_{x^t}{}'\hat{B}^{FC} + \sum\nolimits_{i \in s^0} d_i \tilde{e}_i^{FC}$$

with

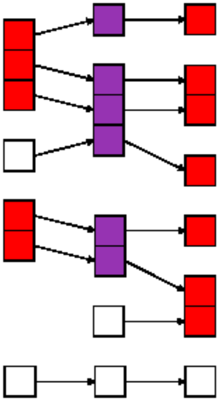
$$\begin{array}{rcl} \tilde{e}_i^{FC} & = & (\tilde{y}_i^t - \tilde{x}_i^{t\prime}\hat{B}^{FC})(\sum\nolimits_{k \in s^t} l_{ik} g_k^t) \\ \hat{B}^{FC} & = & (\sum\nolimits_{k \in s^t} w_k x_k^t x_k^{t\prime})^{-1}(\sum\nolimits_{k \in s^t} w_k x_k^t y_k^t) \end{array}$$

Interpretation: regression of the future contributions y^t on x^t for the persons $i \in s^0$.
Variance estimator:

$$\hat{V}(\hat{T}_{y^0}^{FC}) = \sum\nolimits_{i \in s^0} \sum\nolimits_{i' \in s^0} (\pi_i^{-1}\pi_{i'}^{-1} - \pi_{ii'}^{-1})\tilde{e}_i^{FC}\tilde{e}_{i'}^{FC}$$

Other sorts of Calibration:

Initial and Final Calibration



$$g_i^0: \quad \sum_{i\in s^0} d_i g_i^0 x_i^0 = T_{x^0}$$

$$w_k^C = \sum_{i\in s^0} l_{ik} d_i g_i^0$$

$$g_k^t: \quad \sum_{k\in s^t} w_k^C g_k^t x_k^t = T_{x^t}$$

$$\hat{T}_{y^t}^{FC} = \sum_{k\in s^t} w_k^C g_k^t y_k^t$$

Two adjustment factors g_i^0 for $i \in U^0$ and g_k^t for $k \in U^t$. With $w_k^C = \sum_{i \in s^0} l_{ik} d_i g_i^0$ we have:

$$\hat{T}_{y^t}^{IFC} = \sum_{k \in U^t} y_k^t g_k^t w_k^C$$

Taylor linearisation leads to:

$$\hat{T}_{y^t}^{IFC} \approx T_{x^t}{}' \hat{B}^t + T_{x^0}{}' \hat{B}^0 + \sum_{i \in s^0} d_i e_i^{IFC}$$

Variance estimator:

$$\hat{V}(\hat{T}_{y^t}^{IFC}) = \sum_{i \in s^0} \sum_{i' \in s^0} (\pi_i^{-1} \pi_{i'}^{-1} - \pi_{ii'}^{-1}) g_i^0 g_{i'}^0 e_i^{IFC} e_{i'}^{IFC}$$

with

$$\begin{aligned} e_i^{IFC} &= -x_i^{0\prime} \hat{B}^0 + \sum_{k \in U^t} l_{ik} g_k^t (y_k^t - x_k^{t\prime} \hat{B}^t) \\ \hat{B}^t &= (\sum_{k \in s^t} w_k x_k^t x_k^{t\prime})^{-1} (\sum_{k \in s^t} w_k x_k^t y_k^t) \\ \hat{B}^0 &= (\sum_{i \in s^0} d_i x_i^0 x_i^{0\prime})^{-1} (\sum_{i \in s^0} d_i x_i^0 \sum_{k \in s^t} l_{ik} (y_k^t - x_k^t \hat{B}^t)) \end{aligned}$$

Further reading

- Estevao,V.M., Särndal, C.-E. (2000): A Functional Form Approach to Calibration. *Journal of Official Statistics*, 16, 379-399.
- Estevao,V.M., Särndal, C.-E. (2002): The Ten cases of auxiliary information for calibration in two phase sampling. *Journal of Official Statistics*, 18, 233–255.
- Estevao, V.; Särndal, C.-E. (2004): Borrowing strength is not the best technique within a wide class of design consistent domain estimators. *Journal of Official Statistics*,20, 645-660.
- Estevao, V.; Särndal, C.-E. (2006):Survey Estimates by Calibration on Complex Auxiliary Information. *International Statstical Review* 74, 127-147.
- Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2003) The effect of model choice in estimation for domains, including small domains. *Survey Methodology* 29, 33-44.
- Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2005) Does the model matter? Comparing model-assisted and model-dependent estimators