Issues of the analysis of Socio-Economic Panel Surveys

Ulrich Rendtel

Freie Universität Berlin
Economic Department
Institute for Statistics and Econometrics

Lecture held at Economic Department of Univ. of Pisa
29. + 30. November 2017

Gains from panel data

- Observation of individual change:
  - Identification of gross-change and net-change.
  - Identification of individual risks.
  - Only small observed net-changes.
  - Motivation for panel surveys: "Is it always the same part of the population that stays in poverty or does it exchange rapidly ?"
- Observation of the duration, say of unemployment: Better than retrospective questionnaire mode
- Cumulation of events of interest, say a divorce.
- Prospective study: health risks before death or a disease
- Control of unobserved heterogeneity in regression models via panel models
- Treatment of causality: individual reactions on individuals changes

Poverty Risks (Analysis from Finnish ECHP)

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

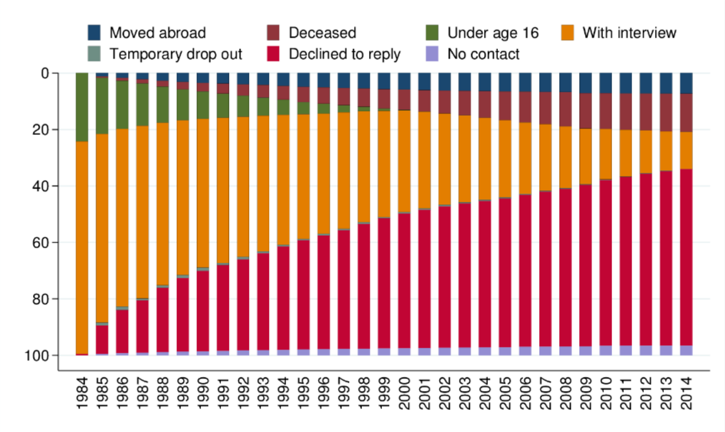
	Transitions in percent	
	Poor	Non-Poor
	Register	
Poor	31.65	68.34
Non-Poor	5.34	94.65
	Survey	
Poor	30.40	69.59
Non-Poor	8.66	91.33

Advantages for field work

- Low nonresponse rates after start: saves field costs!
- Less item nonresponse: more precise data!
- Respondents learn to use the questionnaire instruments.

Problems with panel surveys: Attrition after start

Panel attrition in the SOEP



Panel attrition is thought to ...

- aggravate a potential initial bias in a cumulative fashion
- like the cumulative down-melting of the case numbers.
- be compensated by "refreshment" sample.
- be the reason to stop a panel.
- be the reason for the use of rotation panel schemes with limited duration.

Discussion here:

- Scenario I: Static relationship of the variables of interest
  Attrition acts like a de-mixing of populations.
- Scenario II: Dynamic situation
  Attrition is no longer selective **after** the start of the panel

Pattern Mixture models (1/4)

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 (2/4)

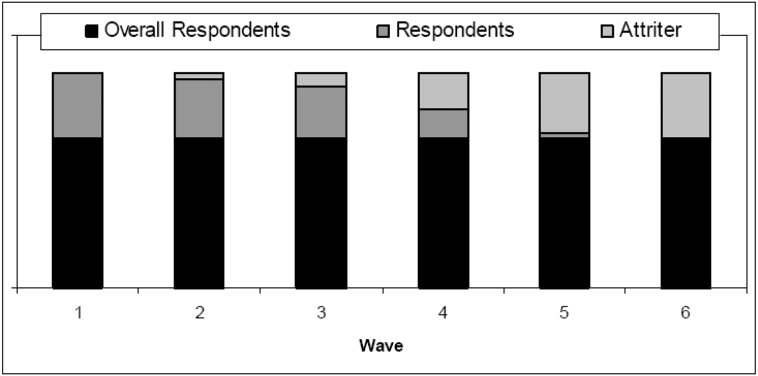
The MAR condition and pattern mixture models:

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

Pattern Mixture models (3/4)

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 (4/4)

- 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: A simulation study

- Sample size $N = 1000$ with two groups of n_1 (Proportion h_1=2/3) and $n_2 = N - n_1$ persons (Proportion $h_2 = 1/3$)
- No of waves: $T = 10$
- Nonresponse rate in group 1 $r_1 = 0.05$ and in group 2 $r_2 = 0.25$
- Lin. model for Y_k with covariates $\mathbf{X}'_k = (1; X_{k,1}; X_{k,2}; X_{k,3})$

$$\begin{aligned} Y_k &= \mathbf{X}'_k\beta_1 + \epsilon_k \text{ for } k = 1, \ldots, n_1 \\ Y_k &= \mathbf{X}'_k\beta_2 + \epsilon_k \text{ for } k = n_1 + 1, \ldots, N \end{aligned}$$

- Distribution of covariates and errors:

$$X_1 \sim N(45, 400) \quad X_2 \sim N(10, 20) \quad X_3 \sim B(0.51) \quad \epsilon_k \sim N(0, 5)$$

- Parameter for group 1: $\beta'_1 = (500, 1, 3, 50)$
- Parameter for group 2: $\beta'_2 = (500, f * 1, f * 3, f * 50) \quad f \in \{0.8, 0.9, 1.01, 1.05, 1.5, 2.0\}$

Pattern Mixture models: Power of the Hausman test for different values of f

Analyse	Welle 2	3	4	5	6	7	8	9	10
Basis ($f = 1{,}2$)									
Teststärke (in %)	17	97	100	100	100	100	100	100	100
$f = 0{,}8$									
Teststärke (in %)	21	100	100	100	100	100	100	100	100
$f = 0{,}9$									
Teststärke (in %)	13	96	100	100	100	100	100	100	100
$f = 1{,}01$									
Teststärke (in %)	19	99	100	100	100	100	100	100	100
$f = 1{,}05$									
Teststärke (in %)	19	98	100	100	100	100	100	100	100
$f = 1{,}5$									
Teststärke (in %)	18	100	100	100	100	100	100	100	100
$f = 2{,}0$									
Teststärke (in %)	17	99	100	100	100	100	100	100	100

Pattern Mixture models: Power of the Hausman test for different attrition rates ($f = 1.2$)

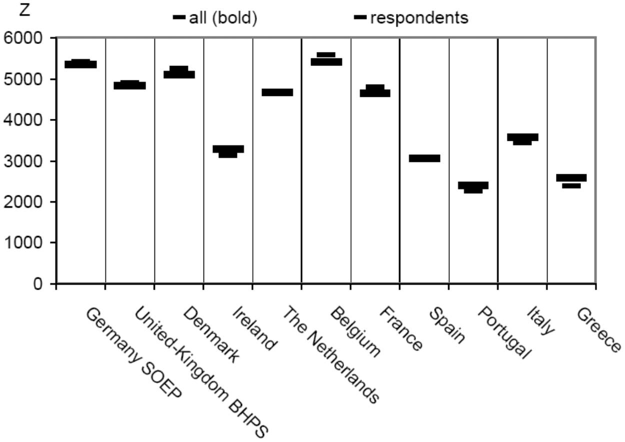
Analyse	Welle								
	2	3	4	5	6	7	8	9	10
Basis ($r_1 = 0{,}05$; $r_2 = 0{,}25$)									
Teststärke (in %)	17	97	100	100	100	100	100	100	100
$r_1 = 0{,}05$; $r_2 = 0{,}1$									
Teststärke (in %)	0	0	0	7	24	53	81	91	96
$r_1 = 0{,}05$; $r_2 = 0{,}5$									
Teststärke (in %)	100	100	100	100	100	100	100	100	100
$r_1 = 0{,}01$; $r_2 = 0{,}25$									
Teststärke (in %)	47	100	100	100	100	100	100	100	100
$r_1 = 0{,}1$; $r_2 = 0{,}25$									
Teststärke (in %)	3	86	100	100	100	100	100	100	100
$r_1 = 0{,}2$; $r_2 = 0{,}25$									
Teststärke (in %)	0	0	0	6	17	16	22	35	36
$r_1 = 0{,}01$; $r_2 = 0{,}5$									
Teststärke (in %)	100	100	100	100	100	100	100	100	100

Pattern Mixture models: Empirical Results

- 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

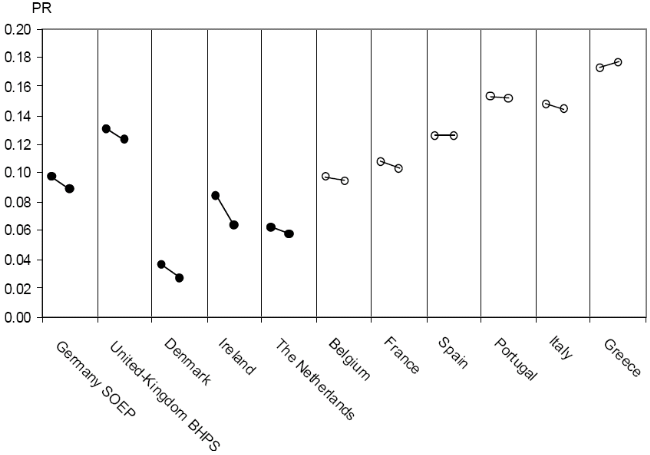
Testing the poverty line

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



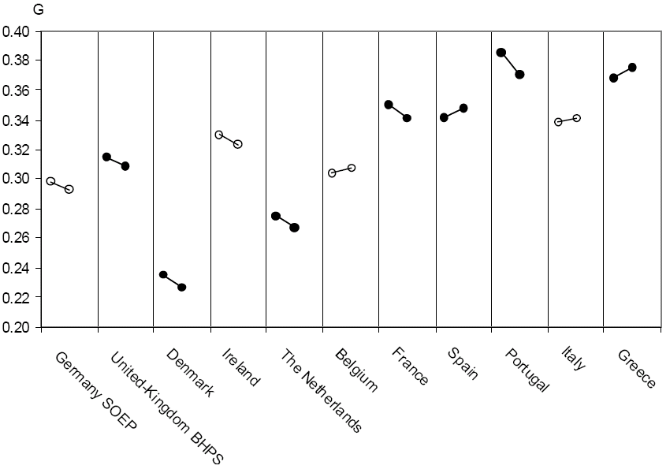
Testing the poverty rate

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



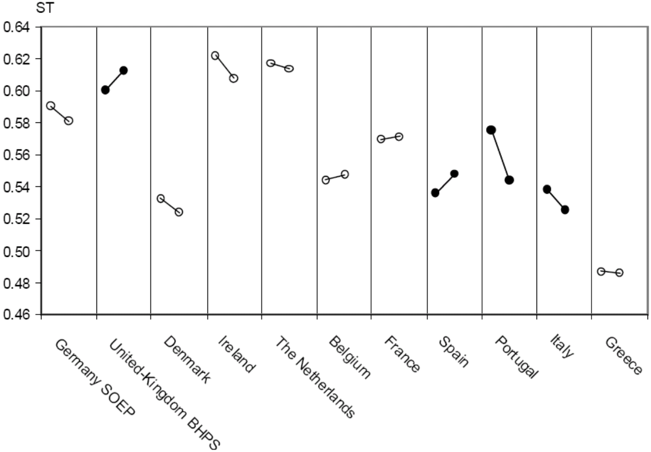
Testing the Gini coefficient

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

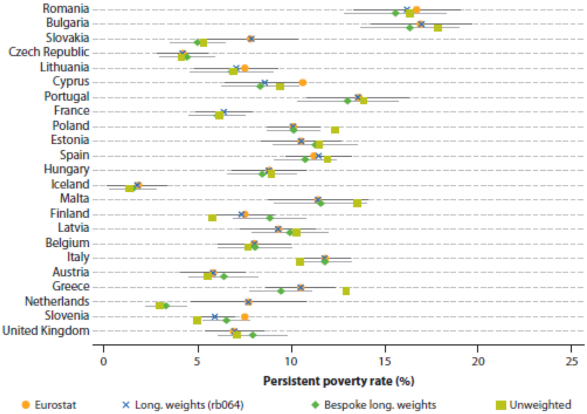
Table 15: **The correlation of the rank position of the 11 countries for different measures of poverty and income stability**

Discussion

- Empirically the differences revealed by the test routine are small.
- If case of large differences: Use a longitudinal weighting scheme! Give higher weights to permanent responders.
- Jenkins, St., v. Kerm, Ph. (2017): How does attrition affect estimates of persistent poverty rates? The case of EU-SILC. In: Atkinson/Guio/Marlier Monitoring social inclusion in Europe, Chapter 22, Eurostat, Luxembourg, DOI 102785/60152
  Retention rates (2008 - 2011)for poverty status, income quintile group, age-sex classes, household type, labour force status and education of head of household, proxy interview
  Weights are different from longitudinal weights (sequential wave by wave), cross-sectional weights (Fair share base weights), Eurostat weights (cross-sectional + calibrations)

Persistent poverty rates with different sampling weights

Figure 22.4: Estimates of 2008-2011 persistent poverty rates with different sampling weights (%)



NB: Countries ordered as in Figure 22.1.
Reading note: Persistent poverty rates calculated by the authors use alternative sampling weights (defined in the text) and are surrounded by 95 % confidence intervals. They are contrasted with rates published by Eurostat (circles) and plain unweighted estimates (squares).
Source: Authors' computation, UDB August 2013. Eurostat estimates are from Eurostat (2014).

The fade-away of initial nonresponse bias in panel surveys (1/2)

The basic idea:

- Suppose the distribution of $Y_{k,t=1}$ is highly selective in the respondent sample s_1 of the panel at wave one, i.e. nonresponse is not ignorable for the estimation in the population.
- Suppose $Y_{k,t}$ evolves independently over time (=waves) for every unit k.
- Suppose that attrition from wave 1 to wave 2 is ignorable.
- **Then**: the distribution of the $y_{k,t}$ is no longer biased at $t = 2, 3 \ldots$

The fade-away of initial nonresponse bias in panel surveys (2/2)

These conditions have to be relaxed!

- The $Y_{k,t}$ follow a Markov chain, not necessarily time-homogeneous.
- The Markovian law is the same for respondents and nonrespondents.
- Every state of the Markov chain can be reached from each other with positive probability for a fixed number of transitions.

A useful contraction theorem (1/2)

Let $\pi_F(0)$ the distribution on the state space for the *FULL* sample (respondents and nonrespondents)
Let $\pi_R(0)$ the distribution on the state space for the *RESP* sample (respondents of wave 1).
Let $P(t)$ the transition law at wave t.
The distribution on the state space at subsequent wave computes as:
$\pi_F(t) = P'(t)\pi_F(t-1)$ and $\pi_R(t) = P'(t)\pi_R(t-1)$ for $t = 1, 2, \ldots$

A useful contraction theorem (2/2)

When all entries of $\pi_R(t)$ are strictly positive, we have the inequalities

$$m_t \equiv \min_i \frac{\pi_{F,i}(t)}{\pi_{R,i}(t)} \leq \frac{\pi_{F,j}(t)}{\pi_{R,j}(t)} \leq \max_i \frac{\pi_{F,i}(t)}{\pi_{R,i}(t)} \equiv M_t, \tag{1}$$

for all $j = 1, \ldots, I$.

Theorem

Suppose that there is lower bound $0 < p_L \leq p_{i,j}(t)$ for all t. Then $\pi_F(t)$ and $\pi_R(t)$ converge uniformly in the sense that

$$\lim_{t\to\infty} (M_t - m_t) = 0. \tag{2}$$

The speed of the fade away effect

Assume that the transition law is stable over time, i.e. $P(t) = P$
Let $p_{ij}^{(t)}$ be the t-fold transition probability from state i to state j.
If all entries of P are positive, then there exists a steady state distribution with $\pi^* = P'\pi^*$
Furthermore the convergence to the steady state distribution follows a uniform geometric pattern where λ_2 is given by the second largest eigenvalue of P.

Theorem

Let λ_2 the second largest eigenvalue of P. Then

$$|p_{ij}^{(t)} - \pi_j^*| = O(|\lambda_2|^t) \text{ for all } i, j \in S. \tag{3}$$

Extensions to longitudinal profiles

- The contraction theorem gives a basis for the convergence of cross-sectional distributions.
- In panels one is more interested in longitudinal profiles!
- Extension of the state space: $(i, j) =$ "i is followed by j"
- Note: The transition $(i, j) \rightsquigarrow (j, z)$ covers three waves!
- Usually the convergence is somewhat slower than in the cross-sectional case.

Analysis with register-based panel surveys

- Register based panel surveys offer the possibility to check the assumption of a homogeneous transition law for respondents and nonrespondents for the register variables.
- Direct measurement of a bias and its decline is possible with access to register records of nonrespondents
- Here: PASS (Panel Arbeitsmarkt u. Soziale Sicherung) recipient sample (unemployment benefit II) based on the register files of German Federal Employment Agency

Nonresponse and attrition in the PASS recipient sample

Wave	gross sample	net sample	attrition rate
wave 1	23,773	6,798	71.40 %
wave 2	6,444	3,468	46.20 %
wave 3	5,737	3,665	36.12 %
wave 4	3,760	2,697	28.27 %
wave 5	3,199	2,257	29.44 %

Results

- For respondents and nonrespondents there is virtually no difference in the transition rates between UBII and no UBII payments.
- Despite a nonresponse rate of 71.40 % there is no over- or under-estimation of the proportion of persons with UBII-payments!
- The assumption of completely missing at random attrition does not hold exactly, although there is no permanent trend to prefer the participation of persons with UBII payments (saliency effect)
- To demonstrate the power of the contraction theorem an artificial starting distribution was used with 95 % (instead of 79 %) receiving UBII-payments.
- Also attrition was simulated according the empirical rates which are not uniform with respect to UBII-status

Results of the simulation experiment

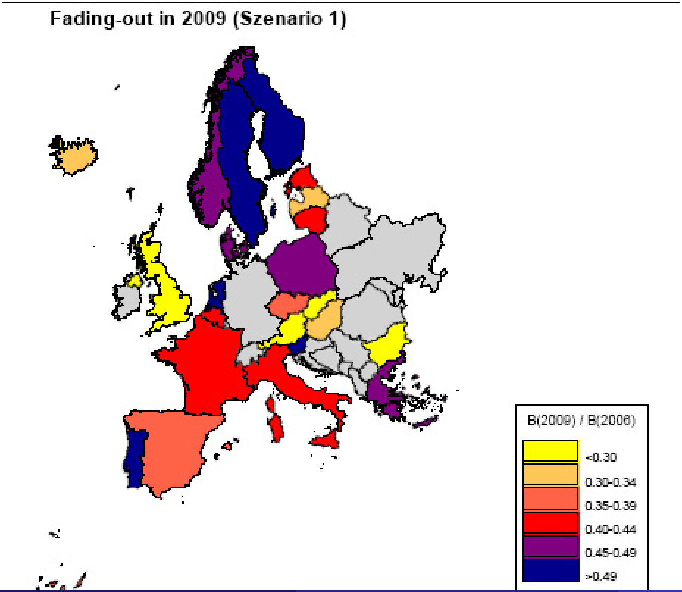
	OBS Sample	*FULL* Sample	$Bias_t$	$Bias_t / Bias_{t-1}$
Wave 1	95.0	79.0	16.0	
Wave 2	79.0	68.8	10.2	0.64
Wave 3	68.9	61.8	7.1	0.70
Wave 4	62.0	57.9	4.1	0.58
Wave 5	56.0	52.8	3.2	0.79

The second eigenvalue of the pooled transition matrix is $\lambda_2 = 0.69$.
Thus time-inhomogeneity of transitions and small deviations from ignorable attrition does not affect the approximation according λ_2!

Similar results

- PASS: Longitudinal profiles (2 waves)
- FIN-ECHP: Transitions between income quintiles
- FIN-SILC: Transitions between income quintiles
- EU-SILC: Comparison of speed across EU-countries (see map!)
- Rendtel, U. (2015): Is there a fade-away effect of initial nonresponse bias in EU-SILC? Discussion Paper Economics 2015/25 FB Wirtschaftswissenschaft FUB, Berlin.
- PSID: Fitzgerald et al. (1998): PSID distributions reduce with increased duration the χ^2-distance to census distributions without calibration and other nonresponse adjustments.
  Fitzgerald, J., Gottschalk, P., Moffitt, R. (1998): An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics, Journal of Human Resources, 33, 251–299.

Reduction of initial bias in 2005 until 2009



True change or fade-away of a bias? (1/2)

The population sample (no link to register data!) of the PASS reveals the following results:

	wave 1	wave 2	wave 3	wave 4	wave 5	steady state
UBII	13.4	12.2	11.8	11.5	10.5	8.6
no UBII	86.6	87.8	88.2	88.5	89.5	91.4

True change or fade-away of a bias? (2/2)

Discussion:

- The UBII was introduced in 2005. The PASS started in 2006. The population distribution was not yet in it's steady state and decreased. The absolute figures of UBII households decreased from 3.773 Mio (2007)to 3.476 Mio (2011) by a factor of 0.921.
- However: the decrease in the population sample was much stronger (by factor 0.783)
- Use of the convergence speed: The second eigenvalue of the transition matrix is $\lambda_2 = 0.78$.
  Take (Percentage at wave t - Percentage at steady state) $\approx \text{Bias}_t$.
  One obtains:

Wave	2	3	4	5
$Bias_t / Bias_{t-1}$	0.75	0.88	0.90	0.66

  The average decrease factor is $0.80 \approx \lambda_2$!

Models with covariates

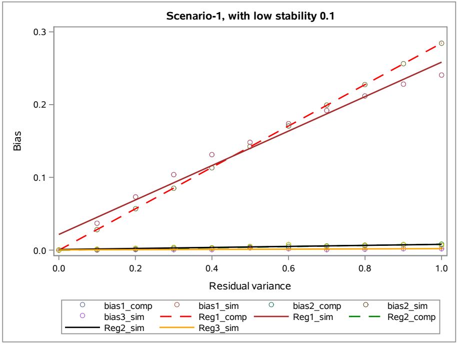
- $Y_{i,t} = a_t + b_t X_{i,t} + e_{i,t}$, regression at time t.
- Variance components for covariate X_{it} and error term e_{it}:
  - $X_{i,t} = M_i + Z_{i,t}$, with $Var(M_i) = \kappa$, $Var(Z_{i,t}) = 1 - \kappa$
  - $e_{i,t} = V_i + U_{i,t}$, with $Var(V_i) = \gamma\sigma^2$, $Var(U_{i,t}) = (1-\gamma)\sigma^2$
- For the shock components a first order autoregressive model is assumed:
  - $Z_{i,t} = \rho Z_{i,(t-1)} + \epsilon_{i,t}$
  - $U_{i,t} = \phi U_{i,(t-1)} + \xi_{i,t}$
- Nonresponse at start of the panel is supposed to be nonignorable: $P(I_i = 1 | Y_{i,1}) = \alpha + \beta Y_{i,1}$

The fade-away effect for the bias of OLS-estimates

- If nonresponse depends only on covariates there is no bias. The larger σ the larger is the bias.
- If there is a large change of the residual component (small values of γ or ϕ): Large fade-away effect!
- If there is a large change of the covariate component (small values of κ or ρ: Large fade-away effect!
- There may be intermixed cases.
- For the bias of OLS estimate of b_t one may derive an approximation formula:

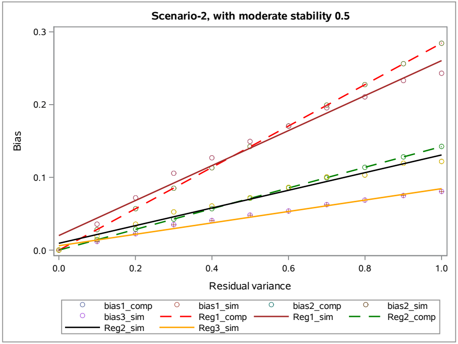
$$p\lim(\hat{b}_2) \approx b_2 - b_1\sigma^2\beta^2 \frac{(\kappa+\rho(1-\kappa))(\gamma+\phi(1-\gamma))}{(\alpha+\beta a_1)^2 - \beta^2 b_1^2(\kappa+\rho(1-\kappa))^2}$$

Scenario A, with low stability $\kappa = \gamma = \rho = \phi = 0.1$.



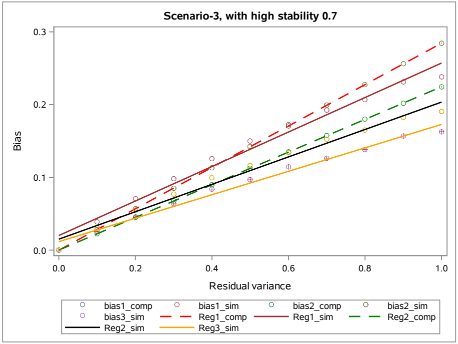
The impact of residual variance σ^2 on the bias of OLS estimates of $b_t = 1$ bias. Number of observations in the sample $n = 1000$. Nonresponse rate is 15%. Dotted Line: Bias computed by approximation formula.

Scenario B, with moderate stability $\kappa = \gamma = \rho = \phi = 0.5$.



The impact of residual variance σ^2 on the bias of OLS estimates of $b_t = 1$ bias. Number of observations in the sample $n = 1000$. Nonresponse rate is 15%. Dotted Line: Bias computed by approximation formula.

Scenario C, with high stability $\kappa = \gamma = \rho = \phi = 0.7$.



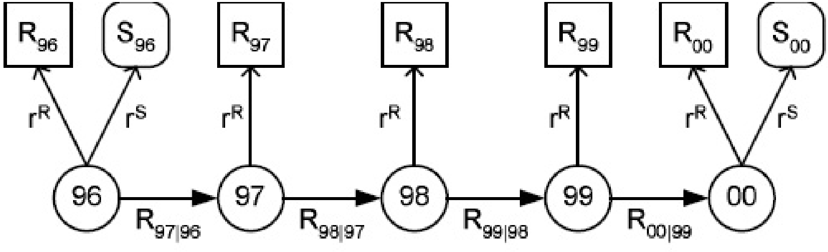
The impact of residual variance σ^2 on the bias of OLS estimates of $b_t = 1$ bias. Number of observations in the sample $n = 1000$. Nonresponse rate is 15%. Dotted Line: Bias computed by approximation formula.

Discussion

- Treatment of nonresponse: Convex combination of design-based estimator and the estimated steady state distribution. Convex factor depending on λ_2: Low (high) weight of steady state distribution in case of slow (fast) fade-away.
- Design of panel surveys:
  - Results from longer running panels may be more reliable.
  - Refreshment samples may incur a fresh nonresponse bias.
  - Prolongate duration of rotation groups. In EU-SILC from 4 to 6 waves!
- Longitudinal Surveys with uncontrolled starting conditions:
  - Control of slowly changing variables like gender, age, family status,... by stratification.
  - Quickly changing attitudes become representative after some time by fade-away!

Measurement of change by a latent Markov model

The graphical representation of the latent Markov model:



Joint measurements of survey and register income in Waves 1996 and 2000 of Finnish ECHP.
Restrictions over time for measurement error.

Change at the latent level

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

	Start	Transitions in percent 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
	Latent		
Poor	8.20	70.04	29.95
Non-Poor	91.79	3.06	96.93

Nordberg L.; Rendtel, U.; Basic, E. (2004): Measurement Error of Survey and Register Income. In Ehling/Rendtel (Hrsg.): Harmonisation of Panel Surveys and Data Quality, Statistisches Bundesamt, Wiesbaden, S. 65-88

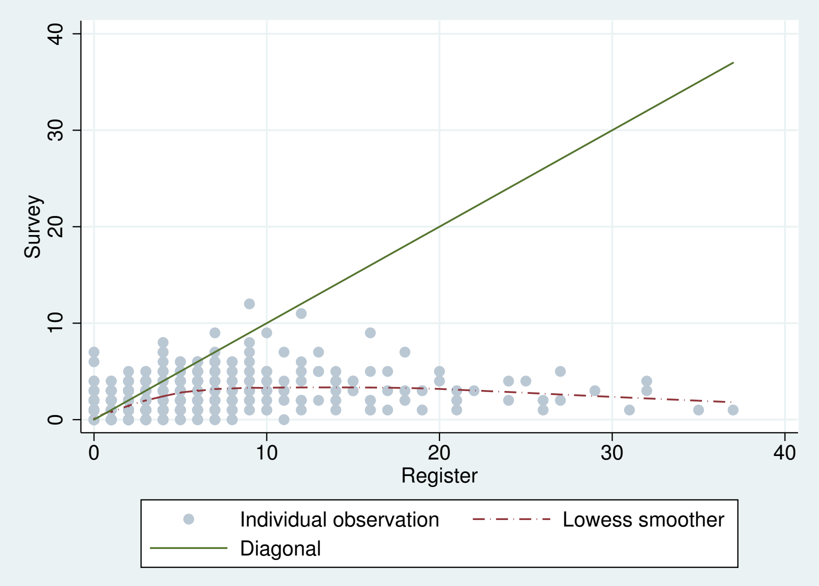
Measurement of unemployment spells

The gold standard: individual verification data from the labour force administration (Finland, ECHP)

- Omission and over-reporting of unemployment
- Heaping at start and end of reporting scheme
- Failure to report exit from unemployment
- Precision of time unit: day (register) vs month (survey)

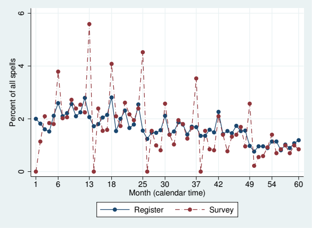
Source: Pyy-Martikainen, M., Rendtel, U. (2009): Measurement Errors in Retrospective Report of Event Histories. A Validation Study with Finnish Register Data. Survey Research Methods, 3, 139–155

Omissions and over-reporting of unemployment spells

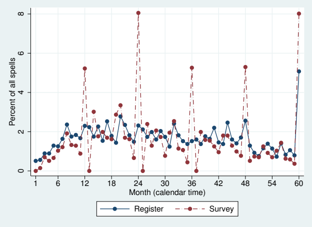


Begin and end of register and survey spells of unemployment

Begin of Spell

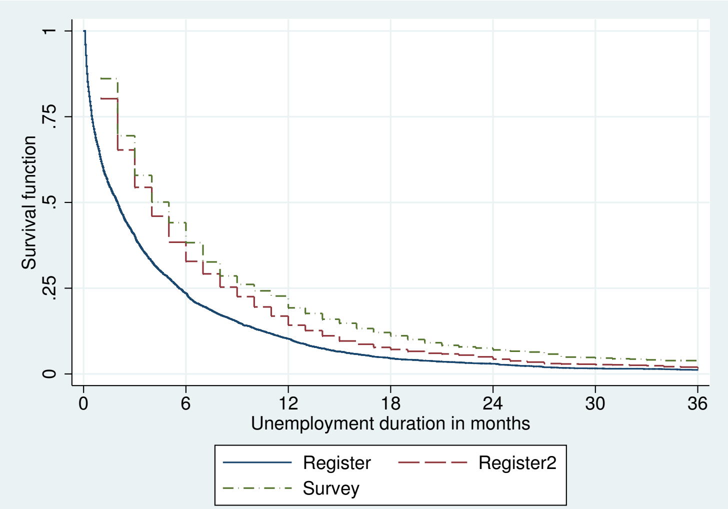


End of Spell



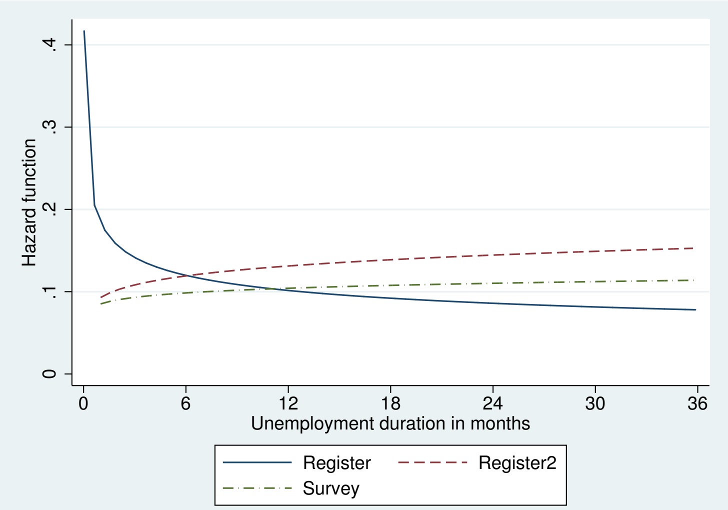
Comparison of Kaplan-Meier plots

Register 2 spells = month counts as unemployed if more than 28 days of unemployment



Comparison of parametric Weibull modell

Register 2 spells = month counts as unemployed if more than 28 days of unemployment



Conclusions

- Dummy variables for heaping points don't reduce a bias in Cox- or Weibull models.
- It is only possible to match unemployment duration at the individual level, not at spell level!
- At the individual level measurement error is correlated with unemployment duration and other important covariates.
- No attenuation bias: independence assumptions of measurement error models are violated!
- Avoid "Chopping of spells" by appropriate questionnaire design (Jäckle/Lynn (2004) "Dependent interviewing")
- Register data for evaluation are very useful.

Final remarks

- The importance of attrition problems has been over-stated for panel surveys.
- As long as case numbers are large enough panel attrition does not show up to have serious consequences for the analysis results.
- Attrition analysis should not pick for significances among a large set of explaining variables from previous waves.
- There are serious measurement problems in panel surveys that deserve more attention.
- Register information can be very powerful for nonresponse analysis, measurement analysis and calibration.