Graduate Programme
Institute for Employment Research
and University Erlangen-Nuremberg,
School of Business and Economics

Linking survey data with administrative employment data: The case of the IAB-ALWA survey

4th User Conference of the FDZ of the BA at the IAB

April 09, 2011

Manfred Antoni
(GradAB, FB E1)

Motivation I: research opportunities

- Various research questions and methods of inference require rich data sets.
- Survey and administrative data sets have their respective comparative advantages.
- Shortcomings of administrative data:
    - No information on civil servant or self-employed individuals
    - Incomplete or inconsistent educational information
    - No information on social cultural background, motivation, ability etc.
- Shortcomings of survey data:
    - Potential for bias due to unit or item non-response, recall error, non-compliance
    - Lack of detailed or reliable information on earnings

⇒ Combination of both data sources increases potential for research.

Motivation I: research opportunities

- Various research questions and methods of inference require rich data sets.
- Survey and administrative data sets have their respective comparative advantages.
- Shortcomings of administrative data:
  - No information on civil servant or self-employed individuals
  - Incomplete or inconsistent educational information
  - No information on social cultural background, motivation, ability etc.
- Shortcomings of survey data:
  - Potential for bias due to unit or item non-response, recall error, non-compliance
  - Lack of detailed or reliable information on earnings

⇒ Combination of both data sources increases potential for research.

Motivation II: survey methodology

- Omitting questions, thereby reducing respondent burden and survey costs
- Less interview terminations or panel attrition
- Validation and improvement of (both survey and administrative) data quality possible
- Lessons for questionnaire and survey design

Motivation III

- Utility of combined data depends on a successful link between data sources.
- Selectivity may arise on different stages of the process of linkage:
  - Selective consent to linkage by respondents
  - Differences in success of record linkage between groups
- Potential for research might be reduced, results could be biased. (cf. Hartmann and Krug, 2009)
- Thorough analysis of selectivity is necessary before a given data set is made available.

⇒ Do characteristics of respondents or the interview situation influence consent to record linkage?

⇒ Do successfully linked respondents differ from the overall survey respondents?

Motivation III

- Utility of combined data depends on a successful link between data sources.
- Selectivity may arise on different stages of the process of linkage:
  - Selective consent to linkage by respondents
  - Differences in success of record linkage between groups
- Potential for research might be reduced, results could be biased. (cf. Hartmann and Krug, 2009)
- Thorough analysis of selectivity is necessary before a given data set is made available.

⇒ Do characteristics of respondents or the interview situation influence consent to record linkage?

⇒ Do successfully linked respondents differ from the overall survey respondents?

Motivation III

- Utility of combined data depends on a successful link between data sources.
- Selectivity may arise on different stages of the process of linkage:
  - Selective consent to linkage by respondents
  - Differences in success of record linkage between groups
- Potential for research might be reduced, results could be biased. (cf. Hartmann and Krug, 2009)
- Thorough analysis of selectivity is necessary before a given data set is made available.

⇒ Do characteristics of respondents or the interview situation influence consent to record linkage?

⇒ Do successfully linked respondents differ from the overall survey respondents?

Outline

- Data sets and their linkage
  - IAB-ALWA survey
  - Administrative data
  - The process of record linkage
- State of research and hypotheses
  - Previous results
  - Determinants of consent
  - Determinants of linkage success
- Empirical results
- Preliminary summary and outlook

Work and Learning in a Changing World (ALWA) I

- Retrospective interviews $>$ 10,000 German residents (cf. Antoni et al., 2010)
- Birth cohorts 1956-1988 (aged 18-52 at time of interview)
- Monthly longitudinal information on, e.g.
  - residential,
  - formal and non-formal educational,
  - (self-)employment and unemployment histories,
  - military and alternative services,
  - partnership histories,
  - times of child care and parental leave.
- Aided recall techniques reduce recall error during interviews. (cf. Drasch and Matthes, forthcoming)

Work and Learning in a Changing World (ALWA) II

- Cross-sectional information on, e.g.
  - place and date of birth,
  - immigrant background,
  - religiousness,
  - language skills,
  - family background,
  - importance of different domains of life,
  - self-reported measures of cognitive skills and personality traits,
  - informal learning and cultural activities.
- Data access: Scientific Use File provided by the FDZ (cf. `http://fdz.iab.de/en/FDZ_Individual_Data/ALWA.aspx`)

Integrated Employment Biographies (IEB)

Administrative data of the German Federal Employment Agency contain

- daily information on histories of dependent employment and registered unemployment since 1975,
- information on benefit receipt and earnings,
- as from 2000: information on the participation in different active labour market policy measures.
- Additional data sources are added with each new version.

Establishment History Panel (BHP)

Employment spells can be supplemented by firm data on

- economic sector,
- qualification structure,
- age structure,
- wage distribution inside the firm,
- worker flows for different subgroups of employees,
- founding and closing of firms under consideration.

Identifiers of respondents

- ALWA lacks unique identifiers for a direct link to administrative records.
- Identifiers for matching:
  - first and last name
  - gender
  - day, month and year of birth
  - postal code
  - place name
  - street name
  - house number
- Sources of identifiers:
  - Field information (infas Institute for Applied Social Sciences)
  - Personal register data (IAB department IT Services and Information Management)

Standardization

- Extensive standardization of identifiers before records are compared:
  - minimizing variation in spelling of names, places and street names,
  - filling in missing information in postal codes or place names,
  - deleting blanks and special characters,
  - standardizing or deleting abbreviations,
  - deleting suffixes of house numbers.

Probabilistic record linkage

- Record linkage based on exact matches: even smallest variations in spelling lead to a rejection of a potential match.
- String comparator metrics penalize deviations between identifiers but do not reject record pairs directly.
- Comparison with blocking on the postal code was done with software Merge ToolBox (v0.7) (cf. Schnell, Bachteler, and Reiher, 2005)
- Parameters for Jaro-Winkler metric according to prior experience with IAB data (cf. Bachteler, 2008)
- Model based on Fellegi and Sunter (1969) classifies record pairs into links, possible links and non-links after comparing both files.
- Pairs that are classified as possible links are subsequently coded as either links or non-links by hand.

Number of observations over the stages of record linkage

	N	$\frac{N}{N_c}$	$\frac{N}{N_r}$
CATI respondents (N_r)	10404		100%
Consenting CATI respondents (N_c)	9531	100%	91.61%
Exact matches	5035	52.83%	48.39%
Exact and probabilistic matches (Jaro-Winkler)	7919	83.09%	76.11%

- Data sets and their linkage
  - IAB-ALWA survey
  - Administrative data
  - The process of record linkage
- State of research and hypotheses
  - Previous results
  - Determinants of consent
  - Determinants of linkage success
- Empirical results
- Preliminary summary and outlook

Previous results

- Numerous studies on linkage of survey and medical or health records. (cf. Dunn et al., 2004; Huang et al., 2007; Kho et al., 2009)
- Not applicable due to different survey populations and data to be linked.
- Studies on comparable populations also linking survey data with administrative employment data rather few in number.
  - Germany: Beste (2011), Hartmann and Krug (2009)
  - UK: Jenkins et al. (2006), Sala, Burton, and Knies (2010)
  - USA: Gustman and Steinmeier (1999), Haider and Solon (2000), Olson (1999), Singer, van Hoewyk, and Neugebauer (2003)
- Low number of observations, small sets of control variables or different procedure of record linkage render generalization of results doubtful if not impossible.

Possible determinants of consent: interview situation

- elapsed duration of interview (**+/–**)
- share of refused answers (esp. to sensitive questions) (**–**)
- share of answers with recall problems (**+**)
- consent to follow-up interview or paper-and-pencil tests (**+**)
- weekday, time of interview
- disturbances, comprehension problems during interview

Possible determinants of consent: respondent characteristics

- foreign nationality, migration background (**–**)
- native language German (**+/–**)
- cognitive abilities (**+/–**)
- qualification (**+**)
- employment status (**+/–**)
- income (∩)
- sex, age
- household composition, marital status

Possible determinants of linkage success

- foreign nationality (**–**)
- employment status
  - registered as unemployed (**+**)
  - employed (**+**)
  - self-employed (**–**)
  - in formal education (except dual vocational training) (**–**)
  - out of the labor force (**–**)

- Data sets and their linkage
  - IAB-ALWA survey
  - Administrative data
  - The process of record linkage
- State of research and hypotheses
  - Previous results
  - Determinants of consent
  - Determinants of linkage success
- Empirical results
- Preliminary summary and outlook

Determinants of consent and linkage success, separate univariate probit regression I

	consent		match	
Foreign nationality	0.866	(0.103)	1.264∗∗	(0.134)
Native language German	0.852	(0.102)	0.954	(0.097)
25-34	0.815∗∗	(0.076)	0.867∗∗	(0.059)
35-44	0.865	(0.088)	0.797∗∗∗	(0.057)
45-52	0.841	(0.092)	0.753∗∗∗	(0.052)
Born in East Germany	1.174∗∗	(0.078)	1.111∗∗∗	(0.044)
Training + lower secondary	1.055	(0.101)	1.174∗∗	(0.083)
Training + intermediate	0.999	(0.080)	1.052	(0.058)
Training + upper secondary	1.119	(0.101)	1.148∗∗	(0.077)
Master craftsman	1.050	(0.127)	0.820∗∗	(0.066)
Higher Education	1.020	(0.085)	0.906	(0.062)
Prose literacy score	0.966	(0.022)	0.987	(0.015)
Document literacy score	0.972	(0.019)	0.972∗∗	(0.012)
Numeracy score	0.963	(0.022)	1.001	(0.015)
High-cultural activity	0.926∗∗∗	(0.018)	0.910∗∗∗	(0.016)
(Self-)Employed	1.309∗∗∗	(0.113)	1.061	(0.082)
In formal education	1.415∗∗∗	(0.150)	0.823∗∗	(0.070)
Other	1.337∗∗∗	(0.131)	0.812∗∗	(0.070)
Personal net income $<$ 500EUR	1.014	(0.078)	1.008	(0.062)
500-999EUR	0.939	(0.069)	1.156∗∗	(0.068)
1000-1499EUR	0.953	(0.062)	0.995	(0.051)
2000-2999EUR	0.998	(0.074)	0.872∗∗	(0.050)
More than 3000EUR	1.026	(0.082)	0.779∗∗∗	(0.043)
Income refused	0.504∗∗∗	(0.052)	0.489∗∗∗	(0.040)

Determinants of consent and linkage success, separate univariate probit regression II

	consent		match
Share of refused answers	0.000∗∗∗	(0.000)	
Share of 'dont't know'	0.005∗∗∗	(0.007)	
Duration before consent quest. (m)	0.998	(0.002)	
Interview on weekend	1.045	(0.065)	
Disturbance during int.	1.035	(0.082)	
Comprehension problems during int.	1.029	(0.079)	
Other problems during int.	0.853∗∗	(0.060)	
Consent to follow-up survey	1.965∗∗∗	(0.148)	
Consent to cognitive tests	1.513∗∗∗	(0.072)	
pseudoR^2	0.103		0.040

Notes: ALWA, own calculations; 9789 observations; 210 interviewers; cluster-robust standard errors in parentheses; ***, **, * denote significance at 1%, 5%, 10%; reference category: aged 18-24, no training, unemployed, net household income of 1500-1999 EUR; additional dummies: sex, married, partner in household, children in household.

Preliminary summary

- Respondent characteristics:
  - Foreign nationality or native language don't influence consent, foreigners are even matched more successfully than Germans.
  - Qualification not relevant for consent, influence on match success inversely u-shaped.
  - Employment status: unemployed show lowest consent, are matched most successfully.
  - Personal net income not relevant for consent, match the least likely for 2 highest income brackets.
- Interviewer situation:
  - Refusal of income information coincides with non-consent and a lack of matching success.
  - The higher the share of refused answers, the less likely is consent.
  - Interview duration plays no role for consent.

Outlook

- Improving record linkage:
    - Reviewing and classifying possible links by hand
    - Retrieving IEB-spells with standard variables for respondents matched so far for validation
    - Considering spells from before 2007 for ALWA respondents not linked so far
- Empirical strategy:
    - Differentiating between dependent employment and self-employment
    - Re-estimation of models on determinants of consent with interviewer characteristics
    - Calculation of marginal effects
    - Implementing the specification as multilevel model

Thank you for your attention

Manfred Antoni
(GradAB, FB E1)
Contact: Manfred.Antoni@iab.de