FDZ

Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)

FDZ·Methodenreport

04/2013
EN

Methodological aspects of labour market data

Example programs for data preparation of the Sample of Integrated Labour Market Biographies for Stata

Creating cross-sectional data and biographical variables

August 2013 (second, updated version)

Johanna Eberle,
Alexandra Schmucker,
Stefan Seth

Bundesagentur für Arbeit

Example programs for data preparation of the Sample of Integrated Labour Market Biographies for Stata

Creating cross-sectional data and biographical variables

Johanna Eberle (Institute for Employment Research)
Alexandra Schmucker (Institute for Employment Research)
Stefan Seth (Institute for Employment Research)

Contents

Abstract 3

Zusammenfassung 3

1 Intorduction 4

2 Outline of the Stata do files 5

3 Generated biographical variables 6

3.1 First day in employment (ein_erw) 6

3.2 Number of days in employment (tage_erw) 6

3.3 First day in establishment (ein_bet) 7

3.4 Number of days in establishment (tage_bet) 7

3.5 First day in job (ein_job) 8

3.6 Numbers of days in job (tage_job) 8

3.7 Number of benefit receipts (anz_lst) 9

3.8 Number of days of benefit receipts (tage_lst) 10

4 Description of the generated variables of the parallel status at the reference date 11

4.1 Type of second job (nb) 11

4.2 Secondary employment in the same establishment as main occupation (nb_betr) 11

4.3 Parallel observation: LeH (leh) 12

4.4 Parallel observation: (X)ASU (asu) 12

4.5 Parallel observation: (X)LHG (lhg) 12

4.6 Total income, all sources (gtentgelt) 13

4.7 Cutoff date of the cross section (stichtag) 13

References 14

Appendix 15

Abstract

This FDZ-Methodenreport (including codes for Stata) outlines an approach to construct cross-sectional data at arbitrary reference dates when working with the Sample of Integrated Labour Market Biographies. In addition, the construction of biographic variables is described.

Zusammenfassung

Der vorliegende FDZ-Methodenreport (einschließlich der Programmierbeispiele für Stata) beschreibt die Erstellung von Querschnittdaten zu frei wählbaren Stichtagen und die Generierung von biografischen Merkmalen auf Basis der Stichprobe der Integrierten Arbeitsmarktbiografien.

Keywords: Sample of Integrated Labour Market Biographies (SIAB), data preparation, cross-sectional data, data management

1 Intorduction

This FDZ-Methodenreport, which includes programs for Stata, demonstrates some ways to prepare SIAB data. The paper provides examples of data re techniques that simplify the data structure. This will hopefully ease access to the SIAB dataset, especially for re-searchers unfamiliar with analyzing spell data.

The main goal of the data preparation shown here is to create cross-sectional data sets at arbitrary reference dates. Also, we simplify the SIAB data structure by keeping only one 'main' observation per person/date. (In the SIAB data, in contrast, there may be concurrent information.) We mitigate the drawback of this procedure – the loss of information - by generating biographical variables, such as a count of days in employment, or the date of entry into the current job. In addition, we show how to create variables recovering important information from simultaneous observations that are deleted.

The provided Stata do files are examples which can be adjusted to specific user needs[1]. The files have been developed using the Sample of Integrated Labour Market Biographies 7510. However, as individual micro-data provided by the FDZ are standardized they can with minor modifications be used for other FDZ data products like, e. g., ALWA-ADIAB as well.

The steps presented in the following sections merely simplify the data. We do not consider techniques to improve data quality or to impute missing values. For this we refer the reader to existing FDZ-Methodenreporte, e. g., Gartner 2005, concerning imputation of wages above the contribution limit and Fitzenberger et al. 2005 or Drews 2006, dealing with the improvement of the education variable.

[1] There are, for instance, numerous definitions of unemployment which affect the calculation of unemployment durations (Kruppe et al. 2007).

2 Outline of the Stata do files

The included Stata do files create several biographical variables and cross-sectional data sets from the longitudinal data at arbitrary points in time. They are structured as follows.

master.do: Here we define Stata macros for directories, file names, and reference dates. Users will have to customize the macros accordingly. Then, the do files that create the biographical variables and cross-sectional data are called. Finally, temporary data sets are deleted.

SIAB_bio.do: This file creates the biographical variables (see chapter 3) from the longitudinal data. Durations are calculated based on the end date of each observation. The temporary data set which is saved at the end of the routine comprises all information from the input data plus the biographical variables.

SIAB_quer.do: In a first step, this program creates, for each reference date, variables indicating the labor market status recorded at the same time as the identified main observation (cf. chapter 4). Note that this do file uses the data set which was created by SIAB_bio.do. In the second step, all observations except the main observations are deleted so that there is only one observation per person/reference date. The duration variables calculated in SIAB_bio.do are adjusted to reflect the durations as of the selected reference date. Finally, a data set is saved for each of the defined reference dates. It is possible to construct a panel data through linking the separate files by person ID (persnr) and reference date (stichtag), if required.

Disclaimer: The attached Stata do files have been tested with SIAB 7510 v1 using Stata 11. Prior to applying them to data products other than SIAB 7510 users should consult the respective FDZ data documentation to make sure that this is appropriate. This is especially important when dealing with “non-7510” data, that is, data not covering the years 1975-2010, because the meaning of underlying variables might differ.

The data version, as well as the data documentation can be obtained from the respective FDZ-Datenreport. The FDZ does not guarantee that the specifications chosen in the provided codes can be applied to all research interests. We strongly advise users to check if the specifications can be transferred to their research project before adopting the routines.

Users who are unfamiliar with processing longitudinal data of the IAB may consult the FDZ-Methodenreport 6/2007 (Drews et al. 2007, only in German available). For a general introduction in data analysis with Stata we recommend Kohler und Kreuter (2012a und 2012b).

3 Generated biographical variables

3.1 First day in employment (ein_erw)

Variable label	first day in employment
Variable name	ein_erw
Category	Generated biographical variables
Origin	Generated from BeH
Data type	Date
Hierarchy	None
Detailed description	This variable specifies the start date of employment subject to social security in the SIAB. Training periods are not included (occupational status = 0). Persons always have a missing value if they pass a training period in the SIAB but do not have an employment covered by the social security system. Episodes before the first employment subject to social security have a missing. The start date of first employment (ein_erw) might occur a long time after the first day in establishment (ein_bet) and the first day in job (ein_job) because in the latter cases training periods are included.
Notes of quality	For West Germans the variable is left censored on 1.1.1975. For East Germans the censoring is not so clear. Entries on 1.1.1990 are censored for sure, but often also entries on 1.1.1991 and 1.1.1992 may be affected because in 1990 and 1991 many employment notifications are missing.

3.2 Number of days in employment (tage_erw)

Variable label	Number of days in employment
Variable name	tage_erw
Category	Generated biographical variables
Origin	Generated from BeH
Data type	Date
Hierarchy	None
Detailed description	This variable sums up the number of days a person has been employed up to the end date of the current observation or respectively after the creation of the cross-sectional data up to the respective reference date, not counting training periods (occupational status = 0). If an individual was just in training, the variable adopts the value 0.
Notes of quality	For West Germans the variable is left censored on 1.1.1975. For East Germans the censoring is not so clear. Entries on 1.1.1990 are censored for sure, but often also entries on 1.1.1991 and 1.1.1992 may be affected because in 1990 and 1991 many employment notifications are missing.

3.3 First day in establishment (ein_bet)

Variable label	first day in establishment
Variable name	ein_bet
Category	Generated biographical variables
Origin	Generated from BeH
Datatype	Date
Hierarchy	none
Detailed description	This variable contains the start date of the first employment notification in the current establishment in the SIAB. This might also be a training period. An interruption of the employment in the establishment does not change the start date, i.e. it is constant for each combination of person number and establishment number. In the case of a missing or invalid establishment number, the variable contains a missing value. The start date of first employment (ein_erw) can occur a long time after the first day in establishment (ein_bet) and the first day in job (ein_job) because in the latter cases training periods are included.
Notes of quality	For West Germans the variable is left censored on 1.1.1975. For East Germans the censoring is not so clear. Entries on 1.1.1990 are censored for sure, but often also entries on 1.1.1991 and 1.1.1992 may be affected because in 1990 and 1991 many employment notifications are missing.

3.4 Number of days in establishment (tage_bet)

Variable label	number of days in establishment
Variable name	tage_bet
Category	Generated biographical variables
Origin	Generated from BeH
Datatype	numeric
Hierarchy	none

Detailed description	The variable counts how many days a person has been working in the establishment till the end date of the present episode or respectively after the creation of the cross-sectional data up to the respective reference date. Training periods in the establishment are included, employment gaps not.
	If the number of days in the establishment is alternatively calculated with the first day in establishment variable (ein_bet),
	`gen tage_bet_neu = {d(30.6.2005), endepi} – ///`
	`ein_bet + 1`
	the values obtained are larger or equal than tage_bet because tage_bet does not include interruptions of employment.

Notes of quality	For West Germans the variable is left censored on 1.1.1975. For East Germans the censoring is not so clear. Entries on 1.1.1990 are censored for sure, but often also entries on 1.1.1991 and 1.1.1992 may

	be affected because in 1990 and 1991 many employment notifications are missing.

3.5 First day in job (ein_job)

Variable label	first day in job
Variable name	ein_job
Category	Generated biographical variables
Origin	Generated from BeH
Datatype	numeric
Hierarchy	none
Detailed description	This variable contains the start date of the first employment notification in the current job. Training periods (occupational status = 0) in the same establishment are treated as separate jobs, even if they follow directly or are followed directly by a job in the same establishment. An employment in the same establishment after a gap is considered a new job if - the reason for notification of the last employment record before the gap indicates the end of the last job (reason of notification = 30, 34, 40, or 49) and the gap is longer than 92 days or - the reason for notification of the last employment record before the gap does not indicate the end of the last job and the gap is longer than 366 days. The first day in new job (ein_job) can not occur before first day in establishment (ein_bet), but it can occur before first day in employment (ein_erw).
Notes of quality	For West Germans the variable is left censored on 1.1.1975. For East Germans the censoring is not so clear. Entries on 1.1.1990 are censored for sure, but often also entries on 1.1.1991 and 1.1.1992 may be affected because in 1990 and 1991 many employment notifications are missing.

3.6 Numbers of days in job (tage_job)

Variable label	numbers of days in job
Variable name	tage_job
Category	Generated biographical variables
Origin	Generated from BeH
Datatype	numeric
Hierarchy	none
Detailed description	The variable counts how many days a person has been working in the current job to the end date of the present Episode or respectively after the creation of the cross-sectional data up to the respective reference

	date of the cross-sectional model.
	Training periods (stib = 0) in the same establishment are treated as separate jobs, even if they follow directly or are followed directly by a job in the same establishment.
	An employment in the same establishment after a gap is considered a new job if
	- the reason for notification of the last employment record before the gap indicates the end of the last job (grund = 30, 34, 40, or 49) and the gap is longer than 92 days or
	- the reason for notification of the last employment record before the gap does not indicate the end of the last job and the gap is longer than 366 days.
	
	If the number of days in the current job is alternatively calculated with the first day in job variable (ein_job),
	`gen tage_job_neu = {d(30.6.2005), endepi} - ///`
	`ein_job + 1`
	the values obtained are larger or eqal than tage_ job because tage_job does not include interruptions of employment.

Notes of quality	For West Germans the variable is left censored on 1.1.1975. For East Germans the censoring is not so clear. Entries on 1.1.1990 are censored for sure, but often also entries on 1.1.1991 and 1.1.1992 may be affected because in 1990 and 1991 many employment notifications are missing.

3.7 Number of benefit receipts (anz_lst)

Variable label	number of benefit receipts
Variable name	anz_lst
Category	Generated biographical variables
Origin	Generated from LEH/LHG/XLHG
Datatype	numeric
Hierarchy	none
Detailed description	The variable contains the number of benefit receipts spells of a person up to the end date of the current observation. Social Code II and Social Code III benefits are treated the same. Hence, the meaning of the variable changes in 2005. The variable is not incremented if a benefit receipt spell is interrupted by a period of less than 10 days or if the type of benefit changes.
Notes of quality	For West Germans the variable is left censored on 1.1.1975. For East Germans the censoring is not so clear. Entries on 1.1.1990 are censored for sure, but often also entries on 1.1.1991 and 1.1.1992 may be affected because in 1990 and 1991 many employment notifications are missing.

3.8 Number of days of benefit receipts (tage_lst)

Variable label	number of days of benefit receipt
Variable name	tage_lst
Category	Generated biographical variables
Origin	Generated from LEH/LHG/XLHG
Datatype	Numeric
Hierarchy	none
Detailed description	The variable contains the number of days of benefit receipt of a person up to the end date of the current observation or respectively after the creation of the cross-sectional data up to the respective reference date of the cross-sectional model. Social Code II and Social Code III benefits are treated the same. Hence, the meaning of the variable changes in 2005. For various reasons, it is possible that someone is employed (employment subject to social security or marginal part-time employees) and also receives benefit receipts at the same time. In this case the benefit receipts spells are counted in tage_lst.
Notes of quality	For West Germans the variable is left censored on 1.1.1975. For East Germans the censoring is not so clear. Entries on 1.1.1990 are censored for sure, but often also entries on 1.1.1991 and 1.1.1992 may be affected because in 1990 and 1991 many employment notifications are missing.

4 Description of the generated variables of the parallel status at the reference date

4.1 Type of second job (nb)

Variable label	Type of second job
Variable name	nb
Category	Generated biographical variables
Origin	Generated from BeH
Datatype	Numeric
Hierarchy	none
Detailed description	The variable indicates if there exists secondary employment and of what kind of secondary employment it is about up to the respective reference date. In this case only the first secondary employment is taken into account. Information about further parallel employment will be lost. A distinction was primarily made between full-time and part-time employment. From 1999 marginal part-time employees are recorded and depicted. Secondary employment which has no valid data to the variable „employment status“ or to the variable „occupational status“ and therefore cannot be classified as full-time, part-time or marginal part-time employees are listed as „secondary employment not specified“. If persons did not work in secondary employment up to the reference date then the variable contains a missing. Values and Labels: 1 full-time job 2 part-time job 3 marginal part-time job 4 not specified second job
Notes of quality	-

4.2 Secondary employment in the same establishment as main occupation (nb_betr)

Variable label	Secondary employment in the same establishment as main occupation
Variable name	nb_betr
Category	Generated biographical variables
Origin	Generated from BeH
Datatype	Numeric
Hierarchy	none
Detailed description	This variable indicates if secondary employment is notified in the same establishment as the main occupation up to the respective reference date. If there is no valid establishment number for the main

	occupation or the secondary employment, then the variable contains missings. Values and Labels: 0 other establishment 1 same establishment
Notes of quality	-

4.3 Parallel observation: LeH (leh)

Variable label	Parallel observation: LeH
Variable name	leh
Category	Generated biographical variables
Origin	Generated from LeH
Datatype	Numeric
Hierarchy	none
Detailed description	This variable indicates if in addition to the main observation an observation from the LeH is present up to the reference date.
Notes of quality	-

4.4 Parallel observation: (X)ASU (asu)

Variable label	Parallel observation: (X)ASU
Variable name	asu
Category	Generated biographical variables
Origin	Generated from ASU/XASU
Datatype	Numeric
Hierarchy	none
Detailed description	The variable indicates if in addition to the main observation an observation from the job seeker History (ASU) or XSozial-BA-Book II (XASU) is present up to the respective reference date.
Notes of quality	-

4.5 Parallel observation: (X)LHG (lhg)

Variable label	Parallel observation: (X)LHG
Variable name	lhg
Category	Generated biographical variables
Origin	Generated from LHG/XLHG
Datatype	Numeric
Hierarchy	none

Detailed description	The variable indicates if in addition to the main observation an observation from Unemployment Benefit II Recipient History A2LL(LHG) or XSozial-BA-SGB II (XLHG) is present up to the reference date.
Notes of quality	-

4.6 Total income, all sources (gtentgelt)

Variable label	Total income, all sources
Variable name	gtentgelt
Category	Generated biographical variables
Origin	Generated from BeH/LeH
Datatype	Numeric
Hierarchy	none
Detailed description	This variable contains the sum of all income from employment notifications and benefit receipt observations up to the respective reference date.
Notes of quality	-

4.7 Cutoff date of the cross section (stichtag)

Variable label	Cutoff date of the cross section
Variable name	stichtag
Category	Generated technical variables
Origin	Generated
Datatype	Date
Hierarchy	none
Detailed description	This variable contains the date of the respective reference date for which the cross-sectional data was created.
Notes of quality	-

References

Berge, Philipp vom; König, Marion; Seth, Stefan (2013): Sample of Integrated Labour Market Biographies (SIAB) 1975 - 2010. FDZ-Datenreport, 01/2013 (en)

Drews, Nils; Groll, Dominik; Jacobebbinghaus, Peter (2007): Programmierbeispiele zur Aufbereitung von FDZ Personendaten in STATA. FDZ-Methodenreport, 06/2007

Drews, Nils (2006): Qualitätsverbesserung der Bildungsvariable in der IAB-Beschäftigtenstichprobe 1975-2001. FDZ-Methodenreport, 05/2006

Fitzenberger, Bernd; Osikominu, Aderonke; Völter, Robert (2005): Imputation rules to improve the education variable in the IAB employment subsample. FDZ-Methodenreport, 03/2005

Gartner, Hermann (2005): The imputation of wages above the contribution limit with the German IAB employment sample. FDZ-Methodenreport, 02/2005

Kohler, Ulrich; Kreuter, Frauke (2012a): Datenanalyse mit Stata * allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung. 4. Auflage. München: Oldenbourg Wissenschaftsverlag

Kohler, Ulrich; Kreuter, Frauke (2012b): Data Analysis Using Stata. Third Edition. Stata Press

Kruppe, Thomas; Müller, Eva; Wichert, Laura; Wilke, Ralf A. (2007): On the definition of unemployment and its implementation in register data * the case of Germany. FDZ-Methodenreport, 03/2007

Appendix

Download of the Stata do files

- English version:
  http://doku.iab.de/fdz/reporte/2013/MR_04-13_EN_programs.ZIP
- German version
  http://doku.iab.de/fdz/reporte/2013/MR_04-13_Programme.ZIP

Impressum

FDZ-Methodenreport 04/2013

Herausgeber
Forschungsdatenzentrum (FDZ)
der Bundesagentur für Arbeit
im Institut für Arbeitsmarkt- und Berufsforschung
Regensburger Str. 104
90478 Nürnberg

Redaktion
Stefan Bender, Heiner Frank

Technische Herstellung
Heiner Frank

Rechte
Nachdruck - auch auszugsweise - nur mit
Genehmigung des FDZ gestattet

Bezugsmöglichkeit
http://doku.iab.de/fdz/reporte/2013/MR_04-13.pdf

Internet
http://fdz.iab.de/

Rückfragen zum Inhalt an:

Alexandra Schmucker
Institut für Arbeitsmarkt und
Berufsforschung (IAB)
Regensburger Str. 104
90478 Nürnberg
Phone: 0911 / 179-1762
mailto:Alexandra.Schmucker@iab.de